
On the Sampling Problem for Kernel Quadrature

François-Xavier Briol^{1,2} Chris J. Oates^{3,4} Jon Cockayne¹ Wilson Ye Chen⁵ Mark Girolami^{2,4}

Abstract

The standard Kernel Quadrature method for numerical integration with random point sets (also called Bayesian Monte Carlo) is known to converge in root mean square error at a rate determined by the ratio s/d , where s and d encode the smoothness and dimension of the integrand. However, an empirical investigation reveals that the rate constant C is highly sensitive to the distribution of the random points. In contrast to standard Monte Carlo integration, for which optimal importance sampling is well-understood, the sampling distribution that minimises C for Kernel Quadrature does not admit a closed form. This paper argues that the practical choice of sampling distribution is an important open problem. One solution is considered; a novel automatic approach based on adaptive tempering and sequential Monte Carlo. Empirical results demonstrate a dramatic reduction in integration error of up to 4 orders of magnitude can be achieved with the proposed method.

1. INTRODUCTION

Consider approximation of the Lebesgue integral

$$\Pi(f) = \int_{\mathcal{X}} f d\Pi \quad (1)$$

where Π is a Borel measure defined over $\mathcal{X} \subseteq \mathbb{R}^d$ and f is Borel measurable. Define $\mathcal{P}(f)$ to be the set of Borel measures Π' such that $f \in L_2(\Pi')$, meaning that $\|f\|_{L_2(\Pi')}^2 = \int_{\mathcal{X}} f^2 d\Pi' < \infty$, and assume $\Pi \in \mathcal{P}(f)$. In situations where $\Pi(f)$ does not admit a closed-form, Monte

Carlo (MC) methods can be used to estimate the numerical value of Eqn. 1. A classical research problem in computational statistics is to reduce the MC estimation error in this context, where the integral can, for example, represent an expectation or marginalisation over a random variable of interest.

The default MC estimator comprises of

$$\hat{\Pi}_{\text{MC}}(f) = \frac{1}{n} \sum_{j=1}^n f(\mathbf{x}_j),$$

where \mathbf{x}_j are sampled identically and independently (i.i.d.) from Π . Then we have a root mean square error (RMSE) bound

$$\sqrt{\mathbb{E}[\hat{\Pi}_{\text{MC}}(f) - \Pi(f)]^2} \leq \frac{C_{\text{MC}}(f; \Pi)}{\sqrt{n}},$$

where $C_{\text{MC}}(f; \Pi) = \text{Std}(f; \Pi)$ and the expectation is with respect to the joint distribution of the $\{\mathbf{x}_j\}_{j=1}^n$. For settings where the Lebesgue density of Π is only known up to normalising constant, Markov chain Monte Carlo (MCMC) methods can be used; the rate-constant $C_{\text{MC}}(f; \Pi)$ is then related to the asymptotic variance of f under the Markov chain sample path.

Considerations of computational cost place emphasis on methods to reduce the rate constant $C_{\text{MC}}(f; \Pi)$. For the MC estimator, this rate constant can be made smaller via importance sampling (IS): $f \mapsto f \cdot d\Pi/d\Pi'$ where an optimal choice $\Pi' \in \mathcal{P}(f \cdot d\Pi/d\Pi')$, that minimises $\text{Std}(f \cdot d\Pi/d\Pi'; \Pi')$, is available in explicit closed-form (see Robert and Casella, 2013, Thm. 3.3.4). However, the RMSE remains asymptotically gated at $O(n^{-1/2})$.

The default Kernel Quadrature (KQ) estimate comprises of

$$\hat{\Pi}(f) = \sum_{j=1}^n w_j f(\mathbf{x}_j), \quad (2)$$

where the $\mathbf{x}_j \sim \Pi'$ are independent (or arise from a Markov chain) and $\text{supp}(\Pi) \subseteq \text{supp}(\Pi')$. In contrast to MC, the weights $\{w_j\}_{j=1}^n$ in KQ are in general non-uniform, real-valued and depend on $\{\mathbf{x}_j\}_{j=1}^n$. The KQ nomenclature derives from the (symmetric, positive-definite) kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ that is used to construct an interpolant $\hat{f}(\mathbf{x}) = \sum_{j=1}^n \beta_j k(\mathbf{x}, \mathbf{x}_j)$ such that

¹University of Warwick, Department of Statistics. ²Imperial College London, Department of Mathematics. ³Newcastle University, School of Mathematics and Statistics ⁴The Alan Turing Institute for Data Science ⁵University of Technology Sydney, School of Mathematical and Physical Sciences. Correspondence to: François-Xavier Briol <f-x.briol@warwick.ac.uk>.

$\hat{f}(\mathbf{x}_j) = f(\mathbf{x}_j)$ for $j = 1, \dots, n$. The weights w_j in Eqn. 2 are implicitly defined via the equation $\hat{\Pi}(f) = \int_{\mathcal{X}} \hat{f} d\Pi$. The KQ estimator is identical to the posterior mean in Bayesian Monte Carlo (O’Hagan, 1991; Rasmussen and Ghahramani, 2002), and its relationship with classical numerical quadrature rules has been studied (Diaconis, 1988; Särkkä et al., 2015).

Under regularity conditions, Briol et al. (2015b) established the following RMSE bound for KQ:

$$\sqrt{\mathbb{E}[\hat{\Pi}(f) - \Pi(f)]^2} \leq \frac{C(f; \Pi')}{n^{s/d-\epsilon}}, \quad (s > d/2)$$

where both the integrand f and each argument of the kernel k admit continuous mixed weak derivatives of order s and $\epsilon > 0$ can be arbitrarily small. An information-theoretic lower bound on the RMSE is $O(n^{-s/d-1/2})$ (Bakhvalov, 1959). The faster convergence of the RMSE, relative to MC, can lead to improved precision in applications. Akin to IS, the samples $\{\mathbf{x}_j\}_{j=1}^n$ need not be draws from Π in order for KQ to provide consistent estimation (since Π is encoded in the weights w_j). Importantly, KQ can be viewed as post-processing of MC samples; the kernel k can be reverse-engineered (e.g. via cross-validation) and does not need to be specified up-front.

One notable disadvantage of KQ methods is that little is known about how the rate constant $C(f; \Pi')$ depends on the choice of sampling distribution Π' . In contrast to IS, no general closed-form expression has been established for an optimal distribution Π' for KQ (the technical meaning of ‘optimal’ is defined below). Moreover, limited practical guidance is available on the selection of the sampling distribution (an exception is Bach, 2015, as explained in Sec. 2.4) and in applications it is usual to take $\Pi' = \Pi$.

This choice is convenient but leads to estimators that are not efficient, as we demonstrate in dramatic empirical examples in Sec. 2.3.

The main contributions of this paper are twofold. First, we formalise the problem of optimal sampling for KQ as an important and open challenge in computational statistics. To be precise, our target is an optimal sampling distribution for KQ, defined as

$$\Pi^* \in \arg \min_{\Pi'} \sup_{f \in \mathcal{F}} \sqrt{\mathbb{E}[\hat{\Pi}(f) - \Pi(f)]^2}. \quad (3)$$

for some functional class \mathcal{F} to be specified. In general a (possibly non-unique) optimal Π^* will depend on \mathcal{F} and, unlike for IS, also on the kernel k and the number of samples n .

Second, we propose a novel and automatic method for selection of Π' that is rooted in approximation of the unavailable Π^* . In brief, our method considers candidate sampling

distributions of the form $\Pi' = \Pi_0^{1-t}\Pi^t$ for $t \in [0, 1]$ and Π_0 a reference distribution on \mathcal{X} . The exponent t is chosen such that Π' minimises an empirical upper bound on the RMSE. The overall approach is facilitated with an efficient sequential MC (SMC) sampler and called SMC-KQ. In particular, the approach (i) provides practical guidance for selection of Π' for KQ, (ii) offers robustness to kernel misspecification, and (iii) extends recent work on computing posterior expectations with kernels obtained using Stein’s method (Oates et al., 2017).

The paper proceeds as follows: Empirical results in Sec. 2 reveal that the RMSE for KQ is highly sensitive to the choice of Π' . The proposed approach to selection of Π' is contained in Sec. 3. Numerical experiments, presented in Sec. 4, demonstrate that dramatic reductions in integration error (up to 4 orders of magnitude) can be achieved with SMC-KQ. Lastly, a discussion is provided in Sec. 5.

2. BACKGROUND

This section presents an overview of KQ (Sec. 2.1 and 2.2), empirical (Secs. 2.3) and theoretical (Sec. 2.4) results on the choice of sampling distribution, and discusses kernel learning for KQ (Sec. 2.5).

2.1. Overview of Kernel Quadrature

We now proceed to describe KQ: Recall the approximation \hat{f} to f ; an explicit form for the coefficients β_j is given as $\beta = \mathbf{K}^{-1}\mathbf{f}$, where $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $f_j = f(\mathbf{x}_j)$. It is assumed that \mathbf{K}^{-1} exists almost surely; for non-degenerate kernels, this corresponds to Π having no atoms. From the above definition of KQ,

$$\hat{\Pi}(f) = \sum_{j=1}^n \beta_j \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}_j) \Pi(d\mathbf{x}).$$

Defining $z_j = \int_{\mathcal{X}} k(\cdot, \mathbf{x}_j) d\Pi$ leads to the estimate in Eqn. 2 with weights $\mathbf{w} = \mathbf{K}^{-1}\mathbf{z}$. Pairs (Π, k) for which the z_j have closed form are reported in Table 1 of Briol et al. (2015b). Computation of these weights incurs a computational cost of at most $O(n^3)$ and can be justified when either (i) evaluation of f forms the computational bottleneck, or (ii) the gain in estimator precision (as a function in n) dominates this cost (i.e. whenever $s/d > 3 + 1/2$).

Notable contributions on KQ include Diaconis (1988); O’Hagan (1991); Rasmussen and Ghahramani (2002) who introduced the method and Huszar and Duvenaud (2012); Osborne et al. (2012a;b); Gunter et al. (2014); Bach (2015); Briol et al. (2015a;b); Särkkä et al. (2015); Kanagawa et al. (2016); Liu and Lee (2017) who provided consequent methodological extensions. KQ has been applied to a wide range of problems including probabilistic ODE

solvers (Kersting and Hennig, 2016), reinforcement learning (Paul et al., 2016), filtering (Prüher and Šimandl, 2015) and design of experiments (Ma et al., 2014).

Several characterisations of the KQ estimator are known and detailed below. Let \mathcal{H} denote the Hilbert space characterised by the reproducing kernel k , and denote its norm as $\|\cdot\|_{\mathcal{H}}$ (Berlinet and Thomas-Agnan, 2011). Then we have the following: (a) The function \hat{f} is the minimiser of $\|g\|_{\mathcal{H}}$ over $g \in \mathcal{H}$ subject to $g(\mathbf{x}_j) = f(\mathbf{x}_j)$ for all $j = 1, \dots, n$. (b) The function \hat{f} is the posterior mean for f under the Gaussian process prior $f \sim \text{GP}(0, k)$ conditioned on data \mathbf{f} and $\hat{\Pi}(f)$ is the mean of the implied posterior marginal over $\Pi[f]$. (c) The weights \mathbf{w} are characterised as the minimiser over $\gamma \in \mathbb{R}^n$ of

$$e_n(\gamma; \{\mathbf{x}_j\}_{j=1}^n) = \sup_{\|f\|_{\mathcal{H}}=1} \left| \sum_{j=1}^n \gamma_j f(\mathbf{x}_j) - \Pi(f) \right|,$$

the maximal error in the unit ball of \mathcal{H} . These characterisations connect KQ to (a) non-parametric regression, (b) probabilistic integration and (c) quasi-Monte Carlo (QMC) methods (Dick and Pillichshammer, 2010). The scattered data approximation literature (Sommariva and Vianello, 2006) and the numerical analysis literature (where KQ is known as the ‘empirical interpolation method’; Eftang and Stamm, 2012; Kristoffersen, 2013) can also be connected to KQ. However, our search of all of these literatures did not yield guidance on the optimal selection of the sampling distribution Π' (with the exception of Bach (2015) reported in Sec. 2.4).

2.2. Over-Reliance on the Kernel

In Osborne et al. (2012a); Huszar and Duvenaud (2012); Gunter et al. (2014); Briol et al. (2015a), the selection of \mathbf{x}_n was approached as a greedy optimisation problem, wherein the maximal integration error $e_n(\mathbf{w}; \{\mathbf{x}_j\}_{j=1}^n)$ was minimised, given the location of the previous $\{\mathbf{x}_j\}_{j=1}^{n-1}$. This approach has demonstrated considerable success in applications. However, the error criterion e_n is strongly dependant on the choice of kernel k and the sequential optimisation approach is vulnerable to kernel misspecification. In particular, if the intrinsic length scale of k is “too small” then the $\{\mathbf{x}_j\}_{j=1}^n$ all cluster around the mode of Π , leading to poor integral estimation (see Fig. 5 in the Appendix). Related work on sub-sample selection, such as leverage scores (Bach, 2013), can also be non-robust to mis-specified kernels. The partial solution of online kernel learning requires a sufficient number n of data and is not always practicable in small- n regimes that motivate KQ.

This paper considers sampling methods as a robust alternative to optimisation methods. Although our method also makes use of k to select Π' , it reverts to $\Pi' = \Pi$ in the

limit as the length scale of k is made small. In this sense, sampling offers more robustness to kernel mis-specification than optimisation methods, at the expense of a possible (non-asymptotic) decrease in precision in the case of a well-specified kernel. This line of research is thus complementary to existing work. However, we emphasise that robustness is an important consideration for general applications of KQ in which kernel specification may be a non-trivial task.

2.3. Sensitivity to the Sampling Distribution

To date, we are not aware of a clear demonstration of the acute dependence of the performance of the KQ estimator on the choice of distribution Π' . It is therefore important to illustrate this phenomenon in order to build intuition.

Consider the toy problem with state space $\mathcal{X} = \mathbb{R}$, target distribution $\Pi = \text{N}(0, 1)$, a single test function $f(x) = 1 + \sin(2\pi x)$ and kernel $k(x, x') = \exp(-(x - x')^2)$. For this problem, consider a range of sampling distributions of the form $\Pi' = \text{N}(0, \sigma^2)$ for $\sigma \in (0, \infty)$. Fig. 1 plots

$$\hat{R}_{n,\sigma} = \sqrt{\frac{1}{M} \sum_{m=1}^M (\hat{\Pi}_{n,m,\sigma}(f) - \Pi(f))^2},$$

an empirical estimate for the RMSE where $\hat{\Pi}_{n,m,\sigma}(f)$ is the m th of M independent KQ estimates for $\Pi(f)$ based on n samples drawn from the distribution Π' with standard deviation σ ($M = 1000$). In this case $\Pi(f) = 1$ is available in closed-form. It is seen that the ‘obvious’ choice of $\sigma = 1$, i.e. $\Pi' = \Pi$, is sub-optimal. The intuition here is that ‘extreme’ samples \mathbf{x}_i from the tails of Π are rather informative for building the interpolant \hat{f} underlying KQ; we should therefore over-sample these values via a heavier-tailed Π' . The same intuition is used for column sampling and to construct leverage scores (Mahoney, 2011; Drineas et al., 2012).

2.4. Established Results

Here we recall the main convergence results to-date on KQ and discuss how these relate to choices of sampling distribution. To reduce the level of detail below, we make several assumptions at the outset:

Assumption on the domain: The domain \mathcal{X} will either be \mathbb{R}^d itself or a compact subset of \mathbb{R}^d that satisfies an ‘interior cone condition’, meaning that there exists an angle $\theta \in (0, \pi/2)$ and a radius $r > 0$ such that for every $\mathbf{x} \in \mathcal{X}$ there exists $\|\boldsymbol{\xi}\|_2 = 1$ such that the cone $\{\mathbf{x} + \lambda \mathbf{y} : \mathbf{y} \in \mathbb{R}^d, \|\mathbf{y}\|_2 = 1, \mathbf{y}^T \boldsymbol{\xi} \geq \cos \theta, \lambda \in [0, r]\}$ is contained in \mathcal{X} (see Wendland, 2004, for background).

Assumption on the kernel: Consider the integral operator $\Sigma : L_2(\Pi) \rightarrow L_2(\Pi)$, with $(\Sigma f)(\mathbf{x})$ defined as the

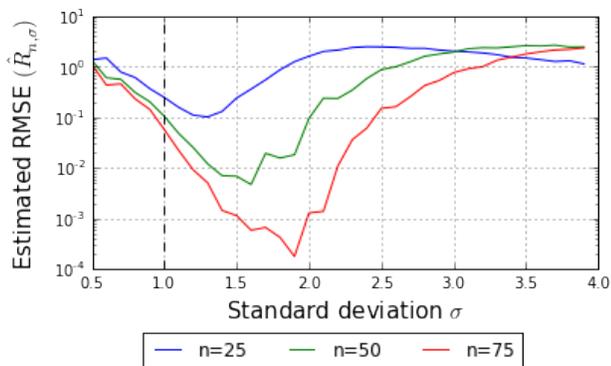


Figure 1. The performance of kernel quadrature is sensitive to the choice of sampling distribution. Here the test function was $f(x) = 1 + \sin(2\pi x)$, the target measure was $N(0, 1)$, while n samples were generated from $N(0, \sigma^2)$. The kernel $k(x, x') = \exp(-(x - x')^2)$ was used. Notice that the values of σ that minimise the root mean square error (RMSE) are uniformly greater than $\sigma = 1$ (dashed line) and depend on the number n of samples in general.

Bochner integral $\int_{\mathcal{X}} f(x')k(x, x')\Pi(dx')$. Assume that $\int_{\mathcal{X}} k(x, x)\Pi(dx) < \infty$, so that Σ is self-adjoint, positive semi-definite and trace-class (Simon, 1979). Then, from an extension of Mercer's theorem (König, 1986) we have a decomposition $k(x, x') = \sum_{m=1}^{\infty} \mu_m e_m(x)e_m(x')$, where μ_m and $e_m(x)$ are the eigenvalues and eigenfunctions of Σ . Further assume that \mathcal{H} is dense in $L_2(\Pi)$.

The first result is adapted and extended from Thm. 1 in Oates et al. (2016).

Theorem 1. Assume that Π' admits a density π' defined on a compact domain \mathcal{X} . Assume that $\pi' > c$ for some $c > 0$. Let $\mathbf{x}_1, \dots, \mathbf{x}_m$ be fixed and define the Euclidean fill distance

$$h_m = \sup_{\mathbf{x} \in \mathcal{X}} \min_{j=1, \dots, m} \|\mathbf{x} - \mathbf{x}_j\|_2.$$

Let $\mathbf{x}_{m+1}, \dots, \mathbf{x}_n$ be independent draws from Π' . Assume k gives rise to a Sobolev space $\mathbb{H}_s(\Pi)$. Then there exists $h_0 > 0$ such that, for $h_m < h_0$,

$$\sqrt{\mathbb{E}[\hat{\Pi}(f) - \Pi(f)]^2} \leq C(f)n^{-s/d+\epsilon}$$

for all $\epsilon > 0$. Here $C(f) = c_{k, \Pi', \epsilon} \|f\|_{\mathcal{H}}$ for some constant $0 < c_{k, \Pi', \epsilon} < \infty$ independent of n and f .

All proofs are reserved for the Appendix. The main contribution of Thm. 1 is to establish a convergence rate for KQ when using importance sampling distributions. A similar result appeared in Thm. 1 of Briol et al. (2015b) for samples from Π (see the Appendix) and was extended to MCMC samples in Oates et al. (2016). An extension to

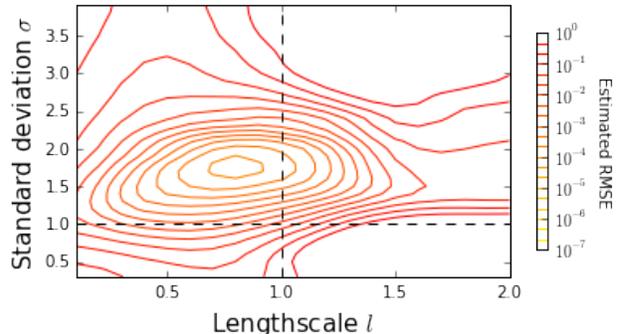


Figure 2. The performance of kernel quadrature is sensitive to the choice of kernel. Here the same set-up as Fig. 1 was used with $n = 75$. The kernel $k(x, x') = \exp(-(x - x')^2/\ell^2)$ was used for various choices of parameter $\ell \in (0, \infty)$. The root mean square error (RMSE) is sensitive to choice of ℓ for all choices of σ , suggesting that online kernel learning could be used to improve over the default choice of $\ell = 1$ and $\sigma = 1$ (dashed lines).

the case of a mis-specified kernel was considered in Kana-gawa et al. (2016). However a limitation of this direction of research is that it does not address the question of how to select Π' .

The second result that we present is a consequence of the recent work of Bach (2015), who considered a particular choice of $\Pi' = \Pi_B$, depending on a fixed $\lambda > 0$, via the density $\pi_B(\mathbf{x}; \lambda) \propto \sum_{m=1}^{\infty} \frac{\mu_m}{\mu_m + \lambda} e_m^2(\mathbf{x})$. The following is adapted from Prop. 1 in Bach (2015):

Theorem 2. Let $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \Pi_B$ be independent and $\lambda > 0$. For $\delta \in (0, 1)$ and $n \geq 5d(\lambda) \log \frac{16d(\lambda)}{\delta}$, $d(\lambda) = \sum_{m=1}^{\infty} \frac{\mu_m}{\mu_m + \lambda}$, we have that

$$|\hat{\Pi}(f) - \Pi(f)| \leq 2\lambda^{1/2} \|f\|_{\mathcal{H}},$$

with probability greater than $1 - \delta$.

Some remarks are in order: (i) Bach (2015, Prop. 3) showed that, for Π_B , integration error scales at an optimal rate in n up to logarithmic terms and, after n samples, is of size $\sqrt{\mu_n}$. (ii) The distribution Π_B is obtained from minimising an upper bound on the integration error, rather than the error itself. It is unclear to us how well Π_B approximates an optimal sampling distribution for KQ. (iii) In general Π_B is hard to compute. For the specific case $\mathcal{X} = [0, 1]^d$, \mathcal{H} equal to $\mathbb{H}_s(\Pi)$ and Π uniform, the distribution Π_B is also uniform (and hence independent of n ; see Sec. 4.4 of Bach (2015)). However, even for the simple example of Sec. 2.3, Π_B does not appear to have a closed form (details in Appendix). An approximation scheme was proposed in Sec. 4.2 of Bach (2015) but the error of this scheme was not studied.

Optimal sampling for approximation in $\|\cdot\|_{L_2(\Pi)}$ with weighted least squares (not in the kernel setting) was

considered in Hampton and Doostan (2015); Cohen and Migliorati (2016).

2.5. Goals

Our first goal was to formalise the sampling problem for KQ; this is now completed. Our second goal was to develop a novel automatic approach to selection of Π' , called SMC-KQ; full details are provided in Sec. 3.

Also, observe that the integrand f will in general belong to an infinitude of Hilbert spaces, while for KQ a single kernel k must be selected. This choice will affect the performance of the KQ estimator; for example, in Fig. 2, the problem of Sec. 2.3 was reconsidered based on a class of kernels $k(x, x') = \exp(-(x - x')^2/\ell^2)$ parametrised by $\ell \in (0, \infty)$. Results showed that, for all choices of σ parameter, the RMSE of KQ is sensitive to choice of ℓ . In particular, the default choice of $\ell = 1$ is not optimal. For this reason, an extension that includes kernel learning, called SMC-KQ-KL, is proposed in Sec. 3.

3. METHODS

In this section the SMC-KQ and SMC-KQ-KL methods are presented. Our aim is to explain in detail the main components (SMC, temp, crit) of Alg. 1. To this end, Secs. 3.1 and 3.2 set up our SMC sampler to target tempered distributions, while Sec. 3.3 presents a heuristic for the choice of temperature schedule. Sec. 3.4 extends the approach to kernel learning and Sec. 3.5 proposes a novel criterion to determine when a desired error tolerance is reached.

3.1. Thermodynamic Ansatz

To begin, consider f , k and n as fixed. The following ansatz is central to our proposed SMC-KQ method: An optimal distribution Π^* (in the sense of Eqn. 3) can be well-approximated by a distribution of the form

$$\Pi_t = \Pi_0^{1-t} \Pi^t, \quad t \in [0, 1] \quad (4)$$

for a specific (but unknown) ‘inverse temperature’ parameter $t = t^*$. Here Π_0 is a reference distribution to be specified and which should be chosen to be un-informative in practice. It is assumed that all Π_t exist (i.e. can be normalised). The motivation for this ansatz stems from Sec. 2.3, where $\Pi = N(0, 1)$ and $\Pi_t = N(0, \sigma^2)$ can be cast in this form with $t = \sigma^{-1}$ and Π_0 an (improper) uniform distribution on \mathbb{R} . In general, tempering generates a class of distributions which over-represent extreme events relative to Π (i.e. have heavier tails). This property has the potential to improve performance for KQ, as demonstrated in Sec. 2.3.

The ansatz of Eqn. 4 reduces the non-parametric sampling problem for KQ to the one-dimensional parametric prob-

lem of selecting a suitable $t \in [0, 1]$. The problem can be further simplified by focusing on a discrete temperature ladder $\{t_i\}_{i=0}^T$ such that $t_0 = 0$, $t_i < t_{i+1}$ and $t_T = 1$. Discussion of the choice of ladder is deferred to Sec. 3.3. This reduced problem, where we seek an optimal index $i^* \in \{0, \dots, T\}$, is still non-trivial as no closed-form expression is available for the RMSE at each candidate t_i . To overcome this *impasse* a novel approach to estimate the RMSE is presented in Sec. 3.5.

3.2. Convex Ansatz (SMC)

The proposed SMC-KQ algorithm requires a second ansatz, namely that the RMSE is convex in t and possesses a global minimum in the range $t \in (0, 1)$. This second ansatz (borne out in numerical results in Fig. 1) motivates an algorithm that begins at $t_0 = 0$ and tracks the RMSE until an increase is detected, say at t_i ; at which point the index $i^* = i - 1$ is taken for KQ.

To realise such an algorithm, this paper exploited SMC methods (Chopin, 2002; Del Moral et al., 2006). Here, a particle approximation $\{(w_j, \mathbf{x}_j)\}_{j=1}^N$ to Π_{t_0} is first obtained where \mathbf{x}_j are independent draws from Π_0 , $w_j = N^{-1}$ and $N \gg n$. Then, at iteration i , the particle approximation to $\Pi_{t_{i-1}}$ is re-weighted, re-sampled and subject to a Markov transition, to deliver a particle approximation $\{(w'_j, \mathbf{x}'_j)\}_{j=1}^N$ to Π_{t_i} . This ‘re-sample-move’ algorithm, denoted SMC, is standard but, for completeness, pseudo-code is provided as Alg. 2 in the Appendix.

At iteration i , a subset of size n is drawn from the unique¹ elements in $\{\mathbf{x}'_j\}_{j=1}^N$, from the particle approximation to Π_{t_i} , and proposed for use in KQ. A criterion `crit`, defined in Sec. 3.5, is used to determine whether the resultant KQ error has increased relative to $\Pi_{t_{i-1}}$. If this is the case, then the distribution $\Pi_{t_{i-1}}$ from the previous iteration is taken for use in KQ. Otherwise the algorithm proceeds to t_{i+1} and the process repeats. In the degenerate case where the RMSE has a minimum at t_T , the algorithm defaults to standard KQ with $\Pi' = \Pi$.

Both ansatz of the SMC-KQ algorithm are justified through the strong empirical results presented in Sec. 4.

3.3. Choice of Temperature Schedule (temp)

The choice of temperature schedule $\{t_i\}_{i=0}^T$ influences several aspects of SMC-KQ: (i) The SMC approximation to Π_{t_i} is governed by the ‘distance’ (in some appropriate metric) between $\Pi_{t_{i-1}}$ and Π_{t_i} . (ii) The speed at which the minimum t^* can be reached is linear in the number of

¹This ensures that kernel matrices have full rank. It does *not* introduce bias into KQ, since in general Π' need not equal Π . However, to keep notation clear, we do not make this operation explicit.

temperatures between 0 and t^* . (iii) The precision of KQ depends on the approximation $t^* \approx t_{i^*}$. Factors (i,iii) motivate the use of a fine schedule with T large, while (ii) motivates a coarse schedule with T small.

For this work, a temperature schedule was used that is well suited to both (i) and (ii), while a strict constraint $t_i - t_{i-1} \leq \Delta$ was imposed on the grid spacing to acknowledge (iii). The specific schedule used in this work was determined based on the conditional effective sample size of the current particle population, as proposed in the recent work of Zhou et al. (2016). Full details are presented in Algs. 4 and 5 in the Appendix.

3.4. Kernel Learning

In Sec. 2.5 we demonstrated the benefit of kernel learning for KQ. From the Gaussian process characterisation of KQ from Sec. 2.1, it follows that kernel parameters θ can be estimated, conditional on a vector of function evaluations \mathbf{f} , via maximum marginal likelihood:

$$\theta' \leftarrow \arg \max_{\theta} p(\mathbf{f}|\theta) = \arg \min_{\theta} \mathbf{f}^{\top} \mathbf{K}_{\theta}^{-1} \mathbf{f} + \log |\mathbf{K}_{\theta}|.$$

In SMC-KQ-KL, the function evaluations \mathbf{f} are obtained at the first² n (of N) states $\{\mathbf{x}_j\}_{j=1}^n$ and the parameters θ are updated in each iteration of the SMC. This demands repeated function evaluation; this burden can be reduced with less frequent parameter updates and caching of all previous function evaluations. The experiments in Sec. 4 assessed both SMC-KQ and SMC-KQ-KL in terms of precision per *total* number of function evaluations, so that the additional cost of kernel learning was taken into account.

3.5. Termination Criterion (`crit`)

The SMC-KQ-KL algorithm is designed to track the RMSE as t is increased. However, the RMSE is not available in closed form. In this section we derive a tight upper bound on the RMSE that is used for the `crit` component in Alg. 1.

From the worst-case characterisation of KQ presented in Sec. 2.1, we have an upper bound

$$|\hat{\Pi}(f) - \Pi(f)| \leq e_n(\mathbf{w}; \{\mathbf{x}_j\}_{j=1}^n) \|f\|_{\mathcal{H}}. \quad (5)$$

The term $e_n(\mathbf{w}; \{\mathbf{x}_j\}_{j=1}^n)$, denoted henceforth as $e_n(\{\mathbf{x}_j\}_{j=1}^n)$ (since \mathbf{w} depends on $\{\mathbf{x}_j\}_{j=1}^n$), can be computed in closed form (see the Appendix). This motivates

²This is a notational convention and is without loss of generality. In this paper these states were a random sample (without replacement) of size n , though stratified sampling among the N states could be used. More sophisticated alternatives that also involve the kernel k , such as leverage scores, were **not** considered, since in general these (i) introduce a vulnerability to mis-specified kernels and (ii) require manipulation of a $N \times N$ kernel matrix (Patel et al., 2015).

Algorithm 1 SMC Algorithm for KQ

function SMC-KQ($f, \Pi, k, \Pi_0, \rho, n, N$)
input f (integrand)
input Π (target disn.)
input k (kernel)
input Π_0 (reference disn.)
input ρ (re-sample threshold)
input n (num. func. evaluations)
input N (num. particles)
 $i \leftarrow 0; t_i \leftarrow 0; R_{\min} \leftarrow \infty$
 $\mathbf{x}'_j \sim \Pi_0$ (initialise states $\forall j \in 1 : N$)
 $w'_j \leftarrow N^{-1}$ (initialise weights $\forall j \in 1 : N$)
 $R \leftarrow \text{crit}(\Pi, k, \{\mathbf{x}'_j\}_{j=1}^N)$ (est'd error)
while `test`($R < R_{\min}$) and $t_i < 1$ **do**
 $i \leftarrow i + 1; R_{\min} \leftarrow R$
 $\{(w_j, \mathbf{x}_j)\}_{j=1}^N \leftarrow \{(w'_j, \mathbf{x}'_j)\}_{j=1}^N$
 $t_i \leftarrow \text{temp}(\{(w_j, \mathbf{x}_j)\}_{j=1}^N, t_{i-1}, \rho)$ (next temp.)
 $\{(w'_j, \mathbf{x}'_j)\}_{j=1}^N \leftarrow \text{SMC}(\{(w_j, \mathbf{x}_j)\}_{j=1}^N, t_i, t_{i-1}, \rho)$
 (next particle approx.)
 $R \leftarrow \text{crit}(\Pi, k, \{\mathbf{x}'_j\}_{j=1}^N)$ (est'd error)
end while
 $\mathbf{f}_j \leftarrow f(\mathbf{x}_j)$ (function eval. $\forall j \in 1 : n$)
 $z_j \leftarrow \int_{\mathcal{X}} k(\cdot, \mathbf{x}_j) d\Pi$ (kernel mean eval. $\forall j \in 1 : n$)
 $\mathbf{K}_{j,j'} \leftarrow k(\mathbf{x}_j, \mathbf{x}_{j'})$ (kernel eval. $\forall j, j' \in 1 : n$)
 $\hat{\Pi}(f) \leftarrow \mathbf{z}^{\top} \mathbf{K}^{-1} \mathbf{f}$ (eval. KQ estimator)
return $\hat{\Pi}(f)$

the following upper bound on MSE:

$$\mathbb{E}[\hat{\Pi}(f) - \Pi(f)]^2 \leq \underbrace{\mathbb{E}[e_n(\{\mathbf{x}_j\}_{j=1}^n)^2]}_{(*)} \underbrace{\|f\|_{\mathcal{H}}^2}_{(**)} \quad (6)$$

The term $(*)$ can be estimated with the bootstrap approximation

$$\mathbb{E}[e_n(\{\mathbf{x}_j\}_{j=1}^n)^2] = \sum_{m=1}^M \frac{e_n(\{\tilde{\mathbf{x}}_{m,j}\}_{j=1}^n)^2}{M} =: R^2$$

where $\tilde{\mathbf{x}}_{m,j}$ are independent draws from $\{\mathbf{x}_j\}_{j=1}^n$. In SMC-KQ the term $(**)$ is an unknown constant and the statistic R , an empirical proxy for the RMSE, is monitored at each iteration. The algorithm terminates once an increase in this statistic occurs. For SMC-KQ-KL the term $(**)$ is non-constant as it depends on the kernel hyper-parameters; then $(**)$ can in addition be estimated as $\|f\|_{\mathcal{H}}^2 = \mathbf{w}^{\top} \mathbf{K}_{\theta} \mathbf{w}$ and we monitor the product of R and $\|f\|_{\mathcal{H}}$, with termination when an increase is observed (c.f. `test`, defined in the Appendix).

Full pseudo-code for SMC-KQ is provided as Alg. 1, while SMC-KQ-KL is Alg. 9 in the Appendix. To summarise, we have developed a novel procedure, SMC-KQ (and an extension SMC-KQ-KL), designed to approximate the optimal

KQ estimator based on the unavailable optimal distribution in Eqn. 3 where \mathcal{F} is the unit ball of \mathcal{H} . Earlier empirical results in Sec. 2.3 suggest that SMC-KQ has potential to provide a powerful and general algorithm for numerical integration. The additional computational cost of optimising the sampling distribution does however have to be counterbalanced with the potential gain in error, and so this method will mainly be of practical interest for problems with expensive integrands or complex target distributions. The following section reports experiments designed to test this claim.

4. RESULTS

Here we compared SMC-KQ (and SMC-KQ-KL) against the corresponding default approaches KQ (and KQ-KL) that are based on $\Pi' = \Pi$. Sec. 4.1 below reports an assessment in which the true value of integrals is known by design, while in Sec. 4.2 the methods were deployed to solve a parameter estimation problem involving differential equations.

4.1. Simulation Study

To continue our illustration from Sec. 2, we investigated the performance of SMC-KQ and SMC-KQ-KL for integration of $f(x) = 1 + \sin(2\pi x)$ against the distribution $\Pi = N(0, 1)$. Here the reference distribution was taken to be $\Pi_0 = N(0, 8^2)$. All experiments employed SMC with $N = 300$ particles, random walk Metropolis transitions (Alg. 3), the re-sample threshold $\rho = 0.95$ and a maximum grid size $\Delta = 0.1$. Dependence of the subsequent results on the choice of Π_0 was investigated in Fig. 10 in the Appendix.

Fig. 3 (top) reports results for SMC-KQ against KQ, for fixed length-scale $\ell = 1$. Corresponding results for SMC-KQ-KL against KQ-KL are shown in the bottom plot. It was observed that SMC-KQ (resp. SMC-KQ-KL) outperformed KQ (resp. KQ-KL) in the sense that, on a per-function-evaluation basis, the MSE achieved by the proposed method was lower than for the standard method. The largest reduction in MSE achieved was about 8 orders of magnitude (correspondingly 4 orders of magnitude in RMSE). A fair approximation to the $\sigma = 2$ method, which is approximately optimal for $n = 75$ (c.f. results in Fig. 1), was observed. The termination criterion in Sec. 3.5 was observed to be a good approximation to the optimal temperature t^* (Fig. 9 in Appendix). As an aside, we note that the MSE was gated at 10^{-16} for all methods due to numerical condition of the kernel matrix \mathbf{K} (a known feature of the Gaussian kernel used in this experiment).

The investigation was extended to larger dimensions ($d = 3$ and $d = 10$) and more complex integrands f in the Ap-

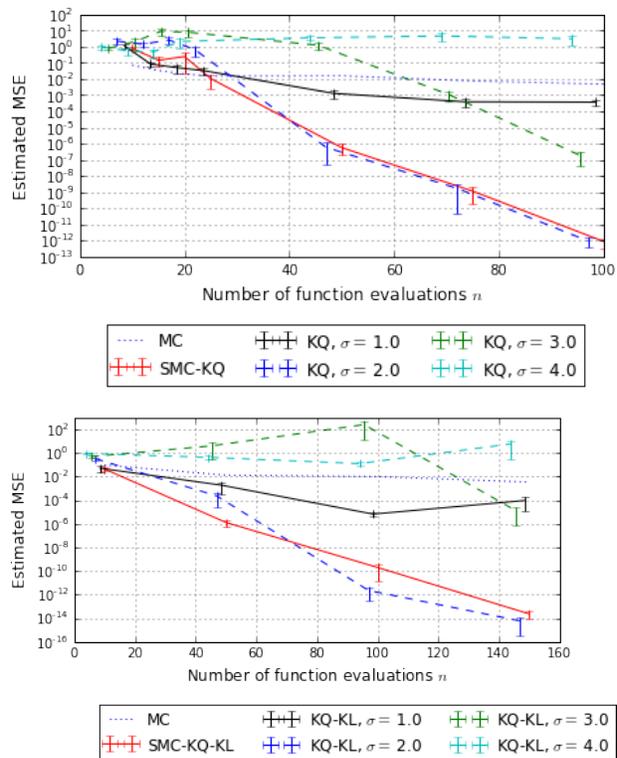


Figure 3. Performance on for the running illustration of Figs. 1 and 2. The top plot shows SMC-KQ against KQ, whilst the bottom plot illustrates the versions with kernel learning.

pendix. In all cases, considerable improvements were obtained using SMC-KQ over KQ.

4.2. Inference for Differential Equations

Consider the model given by $dx/dt = f(t|\theta)$ with solution $x(t|\theta)$ depending on unknown parameters θ . Suppose we can obtain observations through the following noise model (likelihood): $y(t_i) = x(t_i|\theta) + e_i$ at times $0 = t_1 < \dots < t_n$ where we assume $e_i \sim N(0, \sigma^2)$ for known $\sigma > 0$. Our goal is to estimate $x(T|\theta)$ for a fixed (potentially large) $T > 0$. To do so, we will use a Bayesian approach and specify a prior $p(\theta)$, then obtain samples from the posterior $\pi(\theta) := p(\theta|y)$ using MCMC. The posterior predictive mean is then defined as: $\Pi(x(T|\cdot)) = \int x(T|\theta)\pi(\theta)d\theta$, and this can be estimated using an empirical average from the posterior samples. This type of integration problem is particularly challenging as the integrand requires simulating from the differential equation at each iteration. Furthermore, the larger T or the smaller the grid, the longer the simulation will be and the higher the computational cost.

For a tractable test-bed, we considered Hooke's law, given

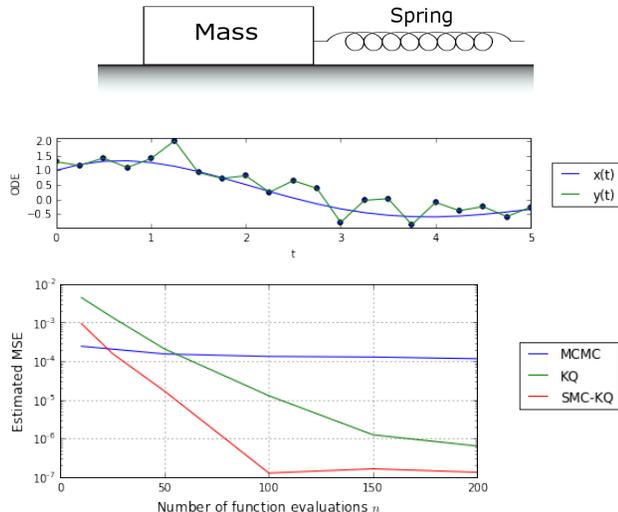


Figure 4. Comparison of SMC-KQ and KQ on the ODE inverse problem. The top plot illustrates the physical system, the middle plot shows observations of the ODE, whilst the bottom plot illustrates the superior performance of SMC-KQ against KQ.

by the following second order homogeneous ODE given by

$$\theta_5 \frac{d^2x}{dt^2} + \theta_4 \frac{dx}{dt} + \theta_3 x = 0,$$

with initial conditions $x(0) = \theta_1$ and $x'(0) = \theta_2$. This equation represents the evolution of a mass on a spring with friction (Robinson, 2004, Chapter 13). More precisely, θ_3 denotes the spring constant, θ_4 the damping coefficient representing friction and θ_5 the mass of the object. Since this differential equation is an overdetermined system we fixed $\theta_5 = 1$. In this case, if $\theta_4^2 \leq 4\theta_3$, we get a damped oscillatory behaviour as presented in Fig. 4 (top). Data were generated with $\sigma = 0.4$, $(\theta_1, \theta_2, \theta_3, \theta_4) = (1, 3.75, 2.5, 0.5)$, with log-normal priors with scale equal to 0.5 for all parameters.

To implement KQ under an unknown normalisation constant for Π , we followed Oates et al. (2017) and made use of a Gaussian kernel that was adapted with Stein’s method (see the Appendix for details). The reference distribution Π_0 was an wide uniform prior on the hypercube $[0, 10]^4$. Brute force computation was used to obtain a benchmark value for the integral. For the SMC algorithm, an independent lognormal transition kernel was used at each iteration with parameters automatically tuned to the current set of particles. Results in Fig. 4 demonstrate that SMC-KQ outperforms KQ for these integration problems. These results improve upon those reported in Oates et al. (2016) for a similar integration problem based on parameter estimation for differential equations.

5. DISCUSSION

In this paper we formalised the optimal sampling problem for KQ. A general, practical solution was proposed, based on novel use of SMC methods. Initial empirical results demonstrate performance gains relative to standard approach of KQ with $\Pi' = \Pi$. A more challenging example based on parameter estimation for differential equations was used to illustrate the potential of SMC-KQ for Bayesian computation in combination with Stein’s method.

Our methods were general but required user-specified choice of an initial distribution Π_0 . For compact state spaces \mathcal{X} we recommend taking Π_0 to be uniform. For non-compact spaces, however, there is a degree of flexibility here and default solutions, such as wide Gaussian distributions, necessarily require user input. However, the choice of Π_0 is easier than the choice of Π' itself, since Π_0 is not required to be optimal. In our examples, improved performance (relative to standard KQ) was observed for a range of reference distributions Π_0 .

A main motivation for this research was to provide an alternative to optimisation-based KQ that alleviates strong dependence on the choice of kernel (Sec. 2.2). This paper provides essential groundwork toward that goal, in developing sampling-based methods for KQ in the case of complex and expensive integration problems. An empirical comparison of sampling-based and optimisation-based methods is reserved for future work.

Two extensions of this research are identified: First, the curse of dimension that is intrinsic to standard Sobolev spaces can be alleviated by demanding ‘dominating mixed smoothness’; our methods are compatible with these (essentially tensor product) kernels (Dick et al., 2013). Second, the use of sequential QMC (Gerber and Chopin, 2015) can be considered, motivated by further orders of magnitude reduction in numerical error observed for deterministic point sets (see Fig. 13 in the Appendix).

ACKNOWLEDGEMENTS

FXB was supported by the EPSRC grant [EP/L016710/1]. CJO & MG we supported by the Lloyds Register Foundation Programme on Data-Centric Engineering. WYC was supported by the ARC Centre of Excellence in Mathematical and Statistical Frontiers. MG was supported by the EPSRC grants [EP/J016934/3, EP/K034154/1, EP/P020720/1], an EPSRC Established Career Fellowship, the EU grant [EU/259348], a Royal Society Wolfson Research Merit Award. FXB, CJO, JC & MG were also supported by the SAMSI working group on Probabilistic Numerics.

REFERENCES

- F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *Proc. I. Conf. Learn. Theory*, 2013.
- F. Bach. On the equivalence between kernel quadrature rules and random features. *arXiv:1502.06800*, 2015.
- N. S. Bakhvalov. On approximate computation of integrals. *Vestnik MGU, Ser. Math. Mech. Astron. Phys. Chem.*, 4: 3–18, 1959. In Russian.
- A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.
- F-X. Briol, C. J. Oates, M. Girolami, and M. A. Osborne. Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Adv. Neur. Inf. Proc. Sys.*, 2015a.
- F-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role for statisticians in numerical analysis? *arXiv:1512.00933*, 2015b.
- N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.
- A. Cohen and G. Migliorati. Optimal weighted least-squares methods. *arXiv:1608.00512*, 2016.
- P. Del Moral, A. Doucet, and A. Jasra. Sequential monte carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 68:411–436, 2006.
- P. Diaconis. *Bayesian Numerical Analysis*, volume IV of *Statistical Decision Theory and Related Topics*, pages 163–175. Springer-Verlag, New York, 1988.
- J. Dick and F. Pillichshammer. *Digital nets and sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, 2010.
- J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.
- P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13:3475–3506, 2012.
- J. L. Eftang and B. Stamm. Parameter multi-domain ‘hp’ empirical interpolation. *I. J. Numer. Methods in Eng.*, 90(4):412–428, 2012.
- M. Gerber and N. Chopin. Sequential quasi Monte Carlo. *J. R. Statist. Soc. B*, 77(3):509–579, 2015.
- T. Gunter, R. Garnett, M. Osborne, P. Hennig, and S. Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Adv. Neur. Inf. Proc. Sys.*, 2014.
- J. Hampton and A. Doostan. Coherence motivated sampling and convergence analysis of least squares polynomial Chaos regression. *Comput. Methods Appl. Mech. Engrg.*, 290:73–97, 2015.
- A. Hinrichs. Optimal importance sampling for the approximation of integrals. *J. Complexity*, 26(2):125–134, 2010.
- F. Huszar and D. Duvenaud. Optimally-weighted herding is Bayesian quadrature. In *Uncert. Artif. Intell.*, 2012.
- M. Kanagawa, B. Sriperumbudur, and K. Fukumizu. Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Adv. Neur. Inf. Proc. Sys.*, 2016.
- H. Kersting and P. Hennig. Active uncertainty calibration in bayesian ode solvers. In *Proc. Conf. Uncert. Artif. Intell.*, 2016.
- H. König. Eigenvalues of compact operators with applications to integral operators. *Linear Algebra Appl.*, 84: 111–122, 1986.
- S. Kristoffersen. The empirical interpolation method. Master’s thesis, Department of Mathematical Sciences, Norwegian University of Science and Technology, 2013.
- Q. Liu and J. D. Lee. Black-Box Importance Sampling. *I. Conf. Artif. Intell. Stat.*, 2017.
- Y. Ma, R. Garnett, and J. Schneider. Active Area Search via Bayesian Quadrature. *I. Conf. Artif. Intell. Stat.*, 33, 2014.
- M. W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, 2011.
- C. J. Oates, J. Cockayne, F-X. Briol, and M. Girolami. Convergence Rates for a Class of Estimators Based on Stein’s Identity. *arXiv:1603.03220*, 2016.
- C. J. Oates, M. Girolami, and N. Chopin. Control Functionals for Monte Carlo Integration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 2017. To appear.
- A. O’Hagan. Bayes-Hermite quadrature. *J. Statist. Plann. Inference*, 29:245–260, 1991.
- M. A. Osborne, D. Duvenaud, R. Garnett, C. E. Rasmussen, S. Roberts, and Z. Ghahramani. Active learning of model evidence using Bayesian quadrature. In *Adv. Neur. Inf. Proc. Sys.*, 2012a.

- M. A. Osborne, R. Garnett, S. Roberts, C. Hart, S. Aigrain, and N. Gibson. Bayesian quadrature for ratios. In *Proc. I. Conf. Artif. Intell. Stat.*, 2012b.
- R. Patel, T. A. Goldstein, E. L. Dyer, A. Mirhoseini, and R. G. Baraniuk. OASIS: Adaptive Column Sampling for Kernel Matrix Approximation. *arXiv:1505.05208*, 2015.
- S. Paul, K. Ciosek, M. A. Osborne, and S. Whiteson. Alternating Optimisation and Quadrature for Robust Reinforcement Learning. *arXiv:1605.07496*, 2016.
- L. Plaskota, G.W. Wasilkowski, and Y. Zhao. New averaging technique for approximating weighted integrals. *J. Complexity*, 25(3):268–291, 2009.
- J. Průher and M. Šimandl. Bayesian Quadrature in Non-linear Filtering. In *12th I. Conf. Inform. Control Autom. Robot.*, 2015.
- C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In *Adv. Neur. Inf. Proc. Sys.*, 2002.
- C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.
- J. C. Robinson. *An introduction to ordinary differential equations*. Cambridge University Press, 2004.
- S. Särkkä, J. Hartikainen, L. Svensson, and F. Sandblom. On the relation between Gaussian process quadratures and sigma-point methods. *arXiv:1504.05994*, 2015.
- T. Shi, M. Belkin, and B. Yu. Data spectroscopy: Eigenspaces of convolution operators and clustering. *Ann. Statist.*, 37(6):3960–3984, 2009.
- B. Simon. *Trace Ideals and Their Applications*. Cambridge University Press, 1979.
- A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learn. Theor.*, pages 13–31, 2007.
- A. Sommariva and M. Vianello. Numerical cubature on scattered data by radial basis functions. *Computing*, 76(3-4):295–310, 2006.
- N. M. Temme. *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*. Wiley, New York, 1996.
- H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2004.
- Y. Zhou, A. M. Johansen, and J. A. D. Aston. Towards Automatic Model Comparison: An Adaptive Sequential Monte Carlo Approach. *J. Comput. Graph. Statist.*, 25(3):701–726, 2016.

A. Appendix

This appendix complements the paper ‘‘On the sampling problem for kernel quadrature’’. Section A.1 discusses the potential lack of robustness of greedy optimization methods, which motivated the development of SMC-KQ. Sections A.2 and A.3 discuss some of the theoretical aspects of KQ, whilst Section A.4 and A.5 presents additional numerical experiments and details for implementation. Finally, Section A.6 provides detailed pseudo-code for all algorithms used in this paper.

A.1. Lack of Robustness of Optimisation Methods

To demonstrate the non-robustness to mis-specified kernels, that is a feature of optimisation-based methods, we considered integration against $\Pi = N(0, 1)$ for functions that can be approximated by the kernel $k(x, x') = \exp(-(x - x')^2/\ell^2)$. An initial state x_1 was fixed at the origin and then for $n = 2, 3, \dots$ the state x_n was chosen to minimise the error criterion $e_n(\mathbf{w}; \{x_j\}_{j=1}^n)$ given the location of the $\{x_j\}_{j=1}^n$. This is known as ‘sequential Bayesian quadrature’ (SBQ; Huszar and Duvenaud, 2012; Gunter et al., 2014; Briol et al., 2015a). The kernel length scale was fixed at $\ell = 0.01$ and we consider (as a thought experiment, since it does not enter into our selection of points) a more regular integrand, such as that shown in Fig. 5 (top). The location of the states $\{x_j\}_{j=1}^n$ obtained in this manner are shown in Fig. 5 (bottom). It is clear that SBQ is not an efficient use of computation for integration of the integrand against $N(0, 1)$. Of course, a bad choice of kernel length scale parameter ℓ can in principle be alleviated by kernel learning, but this will not be robust the case where n is very small.

This example motivates sampling-based methods as an alternative to optimisation-based methods. Future work will be required to better understand when methods such as SBQ can be reliable in the presence of unknown kernel parameters, but this was beyond the scope of this work.

A.2. Additional Definitions

The space $L_2(\Pi)$ is defined to be the set of Π -measurable functions $f : \mathcal{X} \rightarrow \mathbb{R}$ such that the Lebesgue integral

$$\int_{\mathcal{X}} f^2 d\Pi$$

exists and is finite.

For a multi-index $\alpha = (\alpha_1, \dots, \alpha_d)$ define $|\alpha| = \alpha_1 + \dots + \alpha_d$. The (standard) Sobolev space of order $s \in \mathbb{N}$ is denoted

$$\mathbb{H}_s(\Pi) = \{f : \mathcal{X} \rightarrow \mathbb{R} \text{ s.t. } (\partial x_1)^{\alpha_1} \dots (\partial x_d)^{\alpha_d} f \in L_2(\Pi) \forall |\alpha| \leq s\}.$$

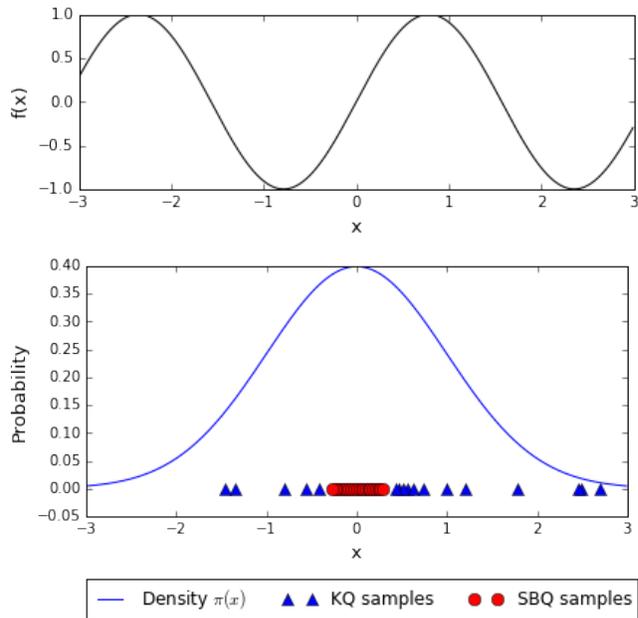


Figure 5. Sequential minimisation of the error criterion $e_n(\mathbf{w}; \{x_j\}_{j=1}^n)$, denoted SBQ, does not lead to adequate placement of points $\{x_j\}_{j=1}^n$ when the kernel is mis-specified. [Here the kernel length scale was fixed to $\ell = 0.01$. Selected points x_j are represented as red. For comparison, a collection of draws from Π , as used in KQ, are shown as blue points.]

This space is equipped with norm

$$\|f\|_{\mathbb{H}_s(\Pi)} = \left(\sum_{|\alpha| \leq s} \|(\partial x_1)^{\alpha_1} \dots (\partial x_d)^{\alpha_d} f\|_{L_2(\Pi)}^2 \right)^{1/2}.$$

Two normed spaces $(\mathcal{F}, \|\cdot\|)$ and $(\mathcal{F}, \|\cdot\|')$ are said to be ‘norm equivalent’ if there exists $0 < c < \infty$ such that

$$c^{-1}\|f\|' \leq \|f\| \leq c\|f\|'$$

for all $f \in \mathcal{F}$.

A.3. Theoretical Results

A.3.1. PROOF OF THEOREM 1

Proof. From Thm. 11.13 in Wendland (2004) we have that there exist constants $0 < c_k < \infty, h_0 > 0$ such that

$$|\hat{f}(\mathbf{x}) - f(\mathbf{x})| \leq c_k h_n^s \|f\|_{\mathcal{H}} \quad (7)$$

for all $\mathbf{x} \in \mathcal{X}$, provided $h_n < h_0$, where

$$h_n = \sup_{\mathbf{x} \in \mathcal{X}} \min_{i=1, \dots, n} \|\mathbf{x} - \mathbf{x}_i\|_2.$$

Under the hypotheses, we can suppose that the deterministic states $\mathbf{x}_1, \dots, \mathbf{x}_m$ ensure $h_m < h_0$. Then Eqn. 7 holds

for all $n > m$, where the $\mathbf{x}_{m+1}, \dots, \mathbf{x}_n$ are independent draws from Π' . It follows that

$$\begin{aligned} |\hat{\Pi}(f) - \Pi(f)| &\leq \sup_{\mathbf{x} \in \mathcal{X}} |\hat{f}(\mathbf{x}) - f(\mathbf{x})| \\ &\leq c_k h_n^s \|f\|_{\mathcal{H}}. \end{aligned}$$

Next, Lem. 1 in Oates et al. (2016) establishes that, under the present hypotheses on \mathcal{X} and Π' , there exists $0 < c_{\Pi', \epsilon} < \infty$ such that

$$\mathbb{E}[h_n^{2s}] \leq c_{\Pi', \epsilon} m^{-2s/d+\epsilon}$$

for all $\epsilon > 0$, where $c_{\Pi', \epsilon}$ is independent of n .

Combining the above results produces

$$\begin{aligned} \mathbb{E}[\hat{\Pi}(f) - \Pi(f)]^2 &\leq c_k^2 \mathbb{E}[h_n^{2s}] \|f\|_{\mathcal{H}}^2 \\ &\leq c_k^2 c_{\Pi', \epsilon} m^{-2s/d+\epsilon} \|f\|_{\mathcal{H}}^2 \end{aligned}$$

as required, with $c_{k, \Pi', \epsilon} = c_k c_{\Pi', \epsilon}^{1/2}$. \square

A.3.2. PROOF OF THEOREM 2

Proof. The Cauchy-Schwarz result for kernel mean embeddings (Smola et al., 2007) gives

$$\begin{aligned} &|\hat{\Pi}(f) - \Pi(f)| \tag{8} \\ &\leq \left\| \sum_{i=1}^n w_i k(\cdot, \mathbf{x}_i) - \int_{\mathcal{X}} k(\cdot, \mathbf{x}) \Pi(d\mathbf{x}) \right\|_{\mathcal{H}} \|f\|_{\mathcal{H}}. \end{aligned}$$

Consider the first term above. Since \mathcal{H} is dense in $L_2(\Pi)$, it follows that $\Sigma^{1/2}$ (the unique positive self-adjoint square root of Σ) is an isometry from $L_2(\Pi)$ to \mathcal{H} . Now, since $k(\cdot, \mathbf{x}) \in \mathcal{H}$, there exists a unique element $\psi(\cdot, \mathbf{x}) \in L_2(\Pi)$ such that $\Sigma^{1/2} \psi(\cdot, \mathbf{x}) = k(\cdot, \mathbf{x})$. Then we have that

$$\begin{aligned} &\left\| \sum_{i=1}^n w_i k(\cdot, \mathbf{x}_i) - \int_{\mathcal{X}} k(\cdot, \mathbf{x}) \Pi(d\mathbf{x}) \right\|_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^n w_i \Sigma^{1/2} \psi(\cdot, \mathbf{x}_i) - \int_{\mathcal{X}} \Sigma^{1/2} \psi(\cdot, \mathbf{x}) \Pi(d\mathbf{x}) \right\|_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^n w_i \psi(\cdot, \mathbf{x}_i) - \int_{\mathcal{X}} \psi(\cdot, \mathbf{x}) \Pi(d\mathbf{x}) \right\|_{L_2(\Pi)}. \end{aligned}$$

For $f \in L_2(\Pi)$, we have $f \in \mathcal{H}$ if and only if

$$f = \int_{\mathcal{X}} g(\mathbf{x}) \psi(\cdot, \mathbf{x}) \Pi(d\mathbf{x}) \tag{9}$$

for some $g \in L_2(\Pi)$, in which case $\|f\|_{\mathcal{H}}$ is equal to the infimum of $\|g\|_{L_2(\Pi)}$ under all such representations g . In particular, it follows that $\|f\|_{\mathcal{H}} = 1$ for the particular choice with $g(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathcal{X}$.

Under the hypothesis on n , Prop. 1 of Bach (2015) established that when $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \Pi_{\mathbf{B}}$ are independent, then

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \inf_{\|\beta\|_2^2 \leq \frac{4}{n}} \left\| \sum_{i=1}^n \frac{\beta_i}{\pi_{\mathbf{B}}(\mathbf{x}_i)^{1/2}} \psi(\cdot, \mathbf{x}_i) - f \right\|_{L_2(\Pi)}^2 \leq 4\lambda$$

with probability at least $1 - \delta$. Fixing the function f in Eqn. 9 leads to the statement that

$$\inf_{\|\beta\|_2^2 \leq \frac{4}{n}} \left\| \sum_{i=1}^n \frac{\beta_i}{\pi_{\mathbf{B}}(\mathbf{x}_i)^{1/2}} \psi(\cdot, \mathbf{x}_i) - \int_{\mathcal{X}} \psi(\cdot, \mathbf{x}) \Pi(d\mathbf{x}) \right\|_{L_2(\Pi)}^2$$

is at most 4λ with probability at least $1 - \delta$. The infimum over $\|\beta\|_2^2 \leq 4/n$ can be replaced with an unconstrained infimum over \mathbb{R}^n to obtain the weaker statement that

$$\inf_{\beta \in \mathbb{R}^n} \left\| \sum_{i=1}^n \frac{\beta_i}{\pi_{\mathbf{B}}(\mathbf{x}_i)^{1/2}} \psi(\cdot, \mathbf{x}_i) - \int_{\mathcal{X}} \psi(\cdot, \mathbf{x}) \Pi(d\mathbf{x}) \right\|_{L_2(\Pi)}^2$$

is at most 4λ with probability at least $1 - \delta$. Now, recall from Sec. 2.1 that the KQ weights w are characterised through the solution β^* to this optimisation problem as $w_i = \beta_i^* \pi_{\mathbf{B}}(\mathbf{x}_i)^{-1/2}$. It follows that

$$\left\| \sum_{i=1}^n w_i \psi(\cdot, \mathbf{x}_i) - \int_{\mathcal{X}} \psi(\cdot, \mathbf{x}) \Pi(d\mathbf{x}) \right\|_{L_2(\Pi)}^2 \leq 4\lambda$$

with probability at least $1 - \delta$. Combining this fact with Eqn. 8 completes the proof. \square

A.3.3. $\Pi_{\mathbf{B}}$ FOR THE EXAMPLE OF FIGURE 1

In this section we consider scope to derive $\Pi_{\mathbf{B}}$ in closed-form for the example of Fig. 1. The following will be used:

Proposition 1 (Prop. 1 in Shi et al. (2009)). *Let $\mathcal{X} = \mathbb{R}$, $\Pi = \mathcal{N}(\mu, \sigma^2)$ and $k(x, x') = \exp(-(x - x')^2 / \ell^2)$. Define $\beta = 4\sigma^2 / \ell^2$ and denote the j th Hermite polynomial as $H_j(x)$. Then the eigenvalues μ_j and corresponding eigenfunctions e_j of the integral operator Σ are*

$$\mu_j = \sqrt{\frac{2}{(1 + \beta + \sqrt{1 + 2\beta})}} \times \left(\frac{\beta}{1 + \beta + \sqrt{1 + 2\beta}} \right)^j$$

and

$$\begin{aligned} e_j(x) &= \frac{(1 + 2\beta)^{1/8}}{\sqrt{2^j j!}} \exp\left(-\frac{(x - \mu)^2 \sqrt{1 + 2\beta} - 1}{2\sigma^2}\right) \\ &\quad \times H_j\left(\left(\frac{1}{4} + \frac{\beta}{2}\right)^{1/4} \frac{x - \mu}{\sigma}\right) \end{aligned}$$

for $j \in \{0, 1, 2, \dots\}$.

Proposition 2 (Ex. 6.8 in Temme (1996), p.167). *The bilinear generating function for Hermite polynomials is*

$$\begin{aligned} \sum_{j=0}^{\infty} \frac{t^j}{j!} H_j(x) H_j(z) \\ = \frac{1}{\sqrt{1-4t^2}} \exp\left(x^2 - \frac{(x-2zt)^2}{1-4t^2}\right). \end{aligned}$$

Proposition 3. *For the example in Fig. 1 we have*

$$\begin{aligned} \pi_B(x; \lambda) \propto \\ \exp(-x^2) \sum_{j=0}^{\infty} \frac{1}{1+\lambda 2^{j+1}} \frac{1}{2^j j!} H_j^2\left(\sqrt{\frac{3}{2}}x\right). \end{aligned}$$

Proof. For the example of Fig. 1, in the notation of Prop. 1, we have $\mu = 0$, $\sigma = 1$, $\ell = 1$ and $\beta = 4$. Thus

$$\begin{aligned} \mu_j &= \left(\frac{1}{2}\right)^{j+1} \\ e_j(x)^2 &= \sqrt{3} \exp(-x^2) \frac{1}{2^j j!} H_j^2\left(\sqrt{\frac{3}{2}}x\right) \end{aligned}$$

and so

$$\begin{aligned} \pi_B(x; \lambda) \propto \sum_j \frac{\mu_j}{\mu_j + \lambda} e_j^2(x) \\ \propto \exp(-x^2) \sum_{j=0}^{\infty} \frac{1}{1+\lambda 2^{j+1}} \frac{1}{2^j j!} H_j^2\left(\sqrt{\frac{3}{2}}x\right) \end{aligned}$$

as required. \square

To the best of our knowledge, the expression for Π_B in Prop. 3 does not admit a closed form. This poses a practical challenge. However, some limited insight is available through basic approximations:

- For large values of λ we have $1 + \lambda 2^{j+1} \approx \lambda 2^{j+1}$ for all $j \in \{0, 1, 2, \dots\}$, from which we obtain

$$\begin{aligned} \pi_B(x; \lambda) &\approx \exp(-x^2) \sum_{j=0}^{\infty} \frac{1}{4^j j!} H_j^2\left(\sqrt{\frac{3}{2}}x\right) \\ &\propto \exp(-x^2) \exp(x^2) = 1, \end{aligned}$$

where the second step made use of Prop. 2. Thus when large integration errors are tolerated, Π_B requires that we take the states x_i to be approximately uniform over \mathcal{X} (of course, this limiting distribution is improper and serves only for illustration).

- For small values of λ , the series in Prop. 3 is dominated by the first m terms such that $j < m$ if and only if $\lambda 2^{j+1} < 1$. Indeed, for $j \leq m$ we have

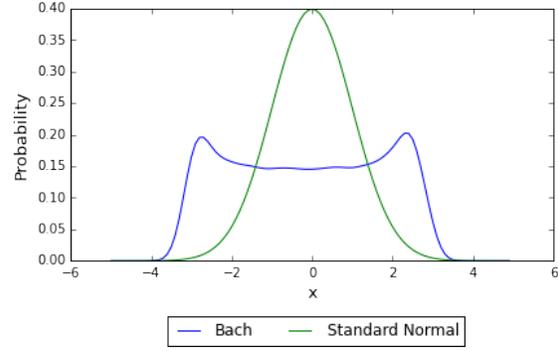


Figure 6. Numerical approximation of Π_B for the running illustration. Here the regularisation parameter was $\lambda = 10^{-15}$.

$1 + \lambda 2^{j+1} \approx 1$. Thus we have a computable approximation

$$\pi_B(x; \lambda) \approx \exp(-x^2) \sum_{j=0}^m \frac{1}{2^j j!} H_j^2\left(\sqrt{\frac{3}{2}}x\right)$$

where $m = \lceil -\log_2(\lambda) \rceil$. Empirical results (not shown) indicate that this is not a useful approximation from a practical standpoint, since at finite m the tails of the approximation are explosive (due to the use of a polynomial basis).

The approximation method in Bach (2015) was also used to obtain the numerical approximation to Π_B shown in Fig. 6. This appears to support the intuition that it is beneficial to over-sample from the tails of Π .

To finish, we remark that Prop. 3 implies that the integration error in this example scales as

$$\sqrt{\mu_n} \sim 2^{-n/2}$$

as $n \rightarrow \infty$ when samples are drawn from Π_B . This agrees with both intuition and empirical results that concern approximation with exponentiated quadratic kernels.

A.3.4. ADDITIONAL THEORETICAL MATERIAL

As mentioned in the Main Text, the worst-case error $e_n(\{\mathbf{x}_j\}_{j=1}^n)$ can be computed in closed form:

$$e_n(\{\mathbf{x}_j\}_{j=1}^n)^2 = \Pi \otimes \Pi(k) - 2\mathbf{w}^\top \mathbf{K} \mathbf{z} + \mathbf{w}^\top \mathbf{K} \mathbf{w}$$

Here we have defined

$$\Pi \otimes \Pi(k) = \iint_{\mathcal{X} \times \mathcal{X}} k(\mathbf{x}, \mathbf{x}') \Pi \otimes \Pi(d\mathbf{x} \times d\mathbf{x}')$$

where $\Pi \otimes \Pi$ is the product measure of Π with itself.

Next, we report a result which does not address KQ itself, but considers importance sampling methods for integration

of functions in a Hilbert space. The following is due to Plaskota et al. (2009); Hinrichs (2010) and we provide an elementary proof of their result:

Theorem 3. *The assumptions of Sec. 2.4 are taken to hold. In addition, we assume that distributions Π, Π' admit densities π, π' . Introduce importance sampling estimators of the form*

$$\hat{\Pi}_{\text{IS}}(f) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i) \frac{\pi(\mathbf{x}_i)}{\pi'(\mathbf{x}_i)},$$

where $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \Pi'$ are independent, and consider the distribution Π' that minimises

$$\sup_{f \in \mathcal{F}} \sqrt{\mathbb{E}[\hat{\Pi}_{\text{IS}}(f) - \Pi(f)]^2}.$$

For $\mathcal{F} = \{f\}$ we have that Π' is $\pi'(\mathbf{x}) \propto |f(\mathbf{x})|\pi(\mathbf{x})$, while for $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$ we have that Π' is $\pi'(\mathbf{x}) \propto \sqrt{k(\mathbf{x}, \mathbf{x})}\pi(\mathbf{x})$.

Proof. The first result, for $\mathcal{F} = \{f\}$ is well-known; e.g. Thm. 3.3.4 in Robert and Casella (2013).

For the second case, where \mathcal{F} is the unit ball in \mathcal{H} , we start by establishing a (tight) upper bound for the supremum of f^2 over $f \in \mathcal{F}$:

$$\begin{aligned} |f(\mathbf{x})| &= |\langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}| \\ &\leq \|f\|_{\mathcal{H}} \|k(\cdot, \mathbf{x})\|_{\mathcal{H}} \\ &= \|f\|_{\mathcal{H}} \sqrt{\langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}} \\ &= \|f\|_{\mathcal{H}} \sqrt{k(\mathbf{x}, \mathbf{x})} \end{aligned}$$

where the inequality here is Cauchy-Schwarz. Squaring both sides and taking the supremum over $f \in \mathcal{F}$ gives

$$\sup_{f \in \mathcal{F}} f(\mathbf{x})^2 \leq \sup_{f \in \mathcal{F}} \|f\|_{\mathcal{H}}^2 k(\mathbf{x}, \mathbf{x}) = k(\mathbf{x}, \mathbf{x}). \quad (10)$$

This is in fact an equality, since for given $\mathbf{x} \in \mathcal{X}$ we can take $f(\mathbf{x}') = k(\mathbf{x}', \mathbf{x})/\sqrt{k(\mathbf{x}, \mathbf{x})}$ which has $\|f\|_{\mathcal{H}} = 1$ and $f(\mathbf{x})^2 = k(\mathbf{x}, \mathbf{x})$.

Our objective is expressed as

$$\sup_{f \in \mathcal{F}} \sqrt{\mathbb{E}[\hat{\Pi}_{\text{IS}}(f) - \Pi(f)]^2} = \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \text{Std}\left(\frac{f\pi}{\pi'}; \Pi'\right)$$

and since

$$\text{Std}\left(\frac{f\pi}{\pi'}; \Pi'\right)^2 = \Pi'\left(\left(\frac{f\pi}{\pi'}\right)^2\right) - \Pi'\left(\frac{f\pi}{\pi'}\right)^2$$

we thus aim to minimise

$$\sup_{f \in \mathcal{F}} \Pi'\left(\left(\frac{f\pi}{\pi'}\right)^2\right)$$

over $\Pi' \in \mathcal{P}(\mathcal{F} \cdot d\Pi/d\Pi')$. (Here $\mathcal{F} \cdot d\Pi/d\Pi'$ denotes the set of functions of the form $f \cdot d\Pi/d\Pi'$ such that $f \in \mathcal{F}$.)

Combining Eqns. 10 and A.3.4, we have

$$\begin{aligned} \sup_{f \in \mathcal{F}} \Pi'\left(\left(\frac{f\pi}{\pi'}\right)^2\right) &\leq \Pi'\left(\sup_{f \in \mathcal{F}} \left(\frac{f\pi}{\pi'}\right)^2\right) \\ &= \Pi'\left(k(\cdot, \cdot) \left(\frac{\pi(\cdot)}{\pi'(\cdot)}\right)^2\right) \end{aligned}$$

As before, this is in fact an equality, as can be seen from $f(\mathbf{x}) = \sqrt{k(\mathbf{x}, \mathbf{x})}$.

From Jensen's inequality,

$$\begin{aligned} \Pi'\left(k(\cdot, \cdot) \left(\frac{\pi(\cdot)}{\pi'(\cdot)}\right)^2\right) &\geq \left(\Pi'\left(\sqrt{k(\cdot, \cdot)} \frac{\pi(\cdot)}{\pi'(\cdot)}\right)\right)^2 \quad (11) \\ &= \left(\Pi(\sqrt{k(\cdot, \cdot)})\right)^2. \end{aligned}$$

Since the right hand side is independent of Π' , a choice of Π' for which Eqn. 11 is an equality must be a minimiser of Eqn. A.3.4. It remains just to verify this fact for $\pi'(\mathbf{x}) = \sqrt{k(\mathbf{x}, \mathbf{x})}\pi(\mathbf{x})/C$, where the normalising constant is $C = \Pi(\sqrt{k(\cdot, \cdot)})$. For this choice

$$\begin{aligned} \Pi'\left(k(\cdot, \cdot) \left(\frac{\pi(\cdot)}{\pi'(\cdot)}\right)^2\right) &= \Pi'(C^2) \\ &= \left(\Pi(\sqrt{k(\cdot, \cdot)})\right)^2 \end{aligned}$$

as required. \square

A.4. Implementation of `test` ($R < R_{\min}$)

Here we provide details for how the criterion $R < R_{\min}$ was tested. The problem with the naive approach of comparing R estimated at t_{i-1} directly with R estimated at t_i is that Monte Carlo error can lead to an incorrect impression that R is increasing, when it is in fact decreasing, and cause the algorithm to terminate when estimation is poor (see Fig. 7 and note the jaggedness of the estimated R curve as a function of inverse temperature t). Our solution was to apply a least-squares linear smoother to the estimates for R over 5 consecutive temperatures. This approach, denoted `test`, illustrated in Fig. 7, determines whether the gradient of the linear smoother is positive or negative, and in this way we are able to provide robustness to Monte Carlo error in the termination criterion. To be precise, the algorithm requires at least 5 temperature evaluations before termination is considered (Fig. 7; left) and terminates when the gradient of the linear smoother becomes positive for the first time (Fig. 7; right). The success of this strategy was established in Fig. 9 later in the Appendix.

A.5. Experimental Results

A.5.1. IMPLEMENTATION OF SIMULATION STUDY

Denote by $N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the p.d.f. of the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Furthermore, we denote by $\boldsymbol{\Sigma}_{\sigma}$ the diagonal covariance matrix

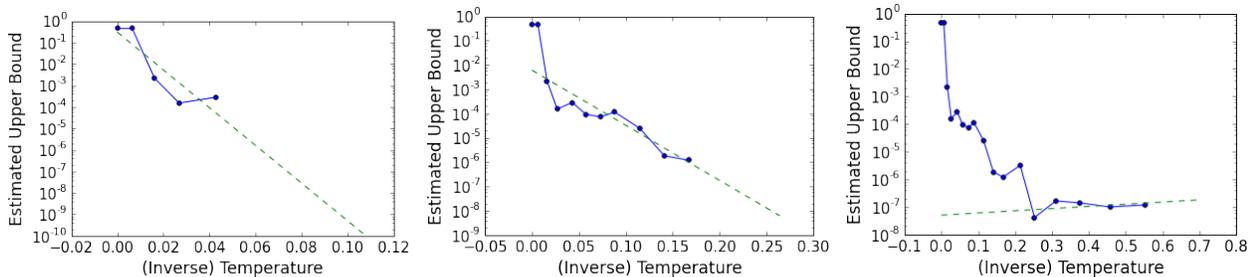


Figure 7. Implementation of $\text{test}(R < R_{\min})$. A linear smoother (dashed line) was based on 5 consecutive (inverse) temperature parameters $t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}, t_i$. To begin it is required that 5 temperatures are considered (left panel). The algorithm terminates on the first occasion when the linear smoother takes a positive gradient (right panel).

with diagonal element σ^2 . Then elementary manipulation of Gaussian densities produces:

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &:= \exp\left(-\frac{\sum_{j=1}^d (x_j - y_j)^2}{l^2}\right) \\ &= (\sqrt{\pi}l)^d \phi(\mathbf{x}|\mathbf{y}, \Sigma_{l/\sqrt{2}}) \\ \nabla_l k(x, y) &:= \frac{2 \sum_{j=1}^d (x_j - y_j)^2}{l^3} k(\mathbf{x}, \mathbf{y}) \\ \Pi[k(\cdot, \cdot)] &:= (\sqrt{\pi}l)^d \mathbf{N}(\mathbf{0}|\mathbf{0}, \Sigma_\sigma + \Sigma_{l/\sqrt{2}}) \\ \Pi \otimes \Pi(k) &:= (\sqrt{\pi}l)^d \mathbf{N}(\mathbf{0}|\mathbf{0}, \Sigma_{\sqrt{2}\sigma} + \Sigma_{l/\sqrt{2}}) \end{aligned}$$

A.5.2. DEPENDENCE ON PARAMETERS FOR THE SIMULATION STUDY

For the running illustration with $f(x) = 1 + \sin(x)$, $\Pi = \mathbf{N}(0, 1)$, $\Pi' = \mathbf{N}(0, \sigma^2)$ and $k(x, x') = \exp(-(x - x')^2/l^2)$, we explored how the RMSE of KQ depends on the choice of both σ and l . Here we go beyond the results presented in Fig. 2, which considered fixed n , to now consider the simultaneous choice of both σ, l for varying n . Note that in these numerical experiments the kernel matrix inverse \mathbf{K}^{-1} was replaced with the regularised inverse $(\mathbf{K} + \lambda \mathbf{I})^{-1}$ that introduces a small ‘nugget’ term $\lambda > 0$ for stabilisation. Results, shown in Fig. 8, demonstrate two principles that guided the methodological development in this paper:

- Length scales l that are ‘too small’ to learn from n samples do not permit good approximations \hat{f} and lead in practice to high RMSE. At the same time, if l is taken to be ‘too large’ then efficient approximation at size n will also be sacrificed. This is of course well understood from a theoretical perspective and is borne out in our empirical results. These results motivated extension of SMC-KQ to SMC-KQ-KL.
- In general the ‘sweet spot’, where σ and l lead to minimal RMSE, is quite small. However, the problem of optimal choice for σ and l does not seem to become

more or less difficult as n increases. This suggests that a method for selection of σ (and possibly also of l) ought to be effective regardless of the number n of states that will be used.

A.5.3. ADDITIONAL RESULTS FOR THE SIMULATION STUDY

To understand whether the termination criterion of Sec. 3.5 was suitable (and, by extension, to examine the validity of the convexity ansatz in Sec. 3.2), in Fig. 9 we presented histograms for both estimated and actual optimal (inverse) temperature parameter t^* . Results supported the use of the criterion, in the form described above for test .

In Fig. 10 reports the dependence of performance on the choice of initial distribution Π_0 . There was relatively little influence on the RMSE obtained by the method for this wide range of initial distribution, which supports the purported robustness of the method.

We also test the method on more complex integrands in Fig. 11: $f(x) = 1 + \sin(4\pi x)$ and $f(x) = 1 + \sin(8\pi x)$. These are more challenging for KQ compared to the illustration in the Main Text, since they are more difficult to interpolate due to their higher periodicity. However, SMC-KQ still manages to adapt to the complexity of the integrand and performs as well as the best importance sampling distribution ($\sigma = 2$).

As an extension, we also study the robustness to the dimensionality to the problem. In problem, we consider the generalisation of our main test function to $f : \mathbb{R}^d \rightarrow \mathbb{R}$ given by $f(\mathbf{x}) = 1 + \prod_{j=1}^d \sin(2\pi x_j)$. Notice that the integral can still be computed analytically and equals 1. We present results for $d = 2$ and $d = 3$ in Fig. 12. These two cases are more challenging for both the KQ and SMC-KQ methods, since the higher dimension implies a slower convergence rate. Once again, we notice that SMC-KQ manages to adapt to the complexity of the problem at hand, and provides improved performance on simpler sampling distributions.

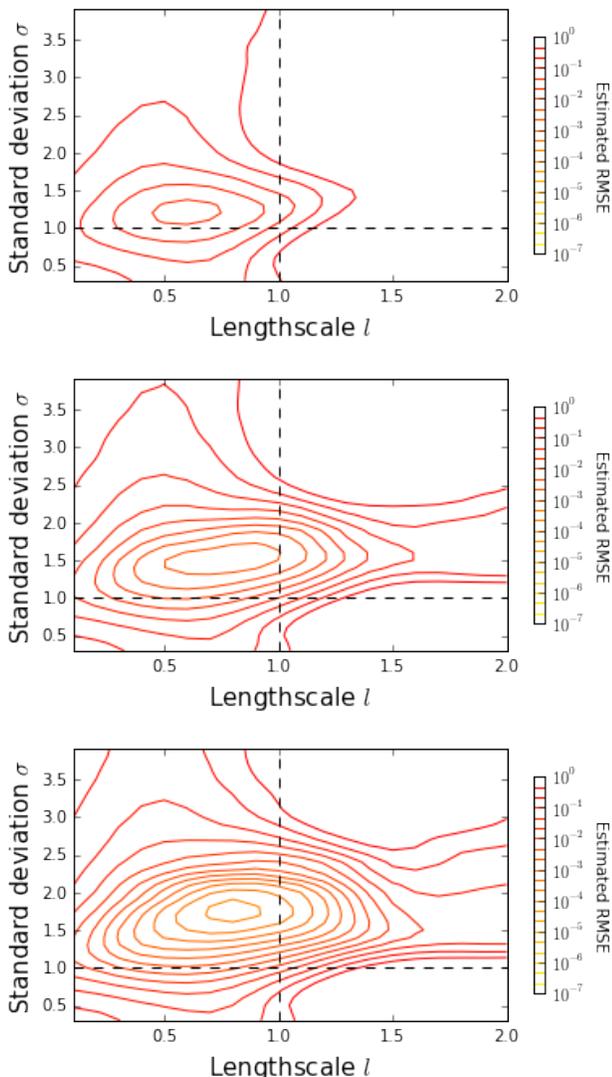


Figure 8. Example of Fig. 2, continued. Here we consider the simultaneous choice of sampling standard deviation σ and kernel length-scale ℓ , reporting empirical estimates for the estimated root mean square integration error (over $M = 300$ repetitions) in each case for sample size (a) $n = 25$ (top), (b) $n = 50$ (middle) and (c) $n = 75$ (bottom).

Finally, we considered replacing the independent samples $x_j \sim \Pi$ with samples drawn from a quasi-random point sequence. Fig. 13 reports results where draws from $N(0, 1)$ were produced based on a Halton quasi-random number generator. In this case, the performance is improved by up to 10 orders of magnitude in MSE when the sampling is done with respect to a range of tempered sampling distribution (here $N(0, 3^2)$). This suggests that a SQMC approach (Gerber and Chopin, 2015) could provide further improvement and this suggested for future work.

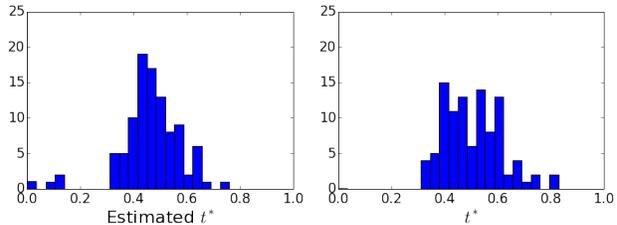


Figure 9. Histograms for the optimal (inverse) temperature parameter t^* . Left: Estimate of t^* provided under the termination criterion of Sec. 3.5. Right: Estimate of t^* obtained by estimating R over a grid for $t \in [0, 1]$ and returning the global minimum. The similarity of these histograms is supportive of the convexity ansatz in Sec. 3.2.

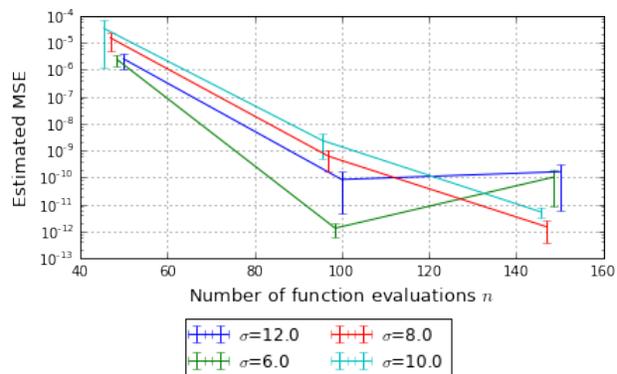


Figure 10. Comparison of the performance of SMC-KQ on the running illustration of Figs. 1 and 2 for varying initial distribution $\Pi_0 = N(0, \sigma^2)$.

A.5.4. IMPLEMENTATION OF STEIN'S METHOD

Following Oates et al. (2017) we considered the Stein operator

$$\mathbb{S}[f](\boldsymbol{\theta}) := [\nabla_{\boldsymbol{\theta}} + \nabla \log \pi(\boldsymbol{\theta})][f](\boldsymbol{\theta})$$

and denote the score function by $u_j(\boldsymbol{\theta}) = \nabla_{\theta_j} \log \pi(\boldsymbol{\theta})$. Here π is the p.d.f. for Π . Applying the Stein operator to each argument of a base kernel k_b , and adding a constant, gives produces the new kernel:

$$k(\boldsymbol{\theta}, \boldsymbol{\phi}) := 1 + \sum_{j=1}^d \begin{aligned} & [\nabla_{\theta_j} \nabla_{\phi_j} k_b(\boldsymbol{\theta}, \boldsymbol{\phi}) \\ & + u_j(\boldsymbol{\theta}) \nabla_{\phi_j} k_b(\boldsymbol{\theta}, \boldsymbol{\phi}) \\ & + u_j(\boldsymbol{\phi}) \nabla_{\theta_j} k_b(\boldsymbol{\theta}, \boldsymbol{\phi}) \\ & + u_j(\boldsymbol{\theta}) u_j(\boldsymbol{\phi}) k_b(\boldsymbol{\theta}, \boldsymbol{\phi}) \end{aligned}$$

which we will use for our KQ estimator. Using integration by parts, we can easily check that $\Pi[k(\cdot, \boldsymbol{\theta})] = 1$ and $\Pi \otimes \Pi(k) = 1$. In this experiment, the base kernel was taken to be Gaussian: $k_b(\boldsymbol{\theta}, \boldsymbol{\phi}) = \exp(-\sum_{j=1}^d (\theta_j - \phi_j)^2 / \ell_j^2)$. We

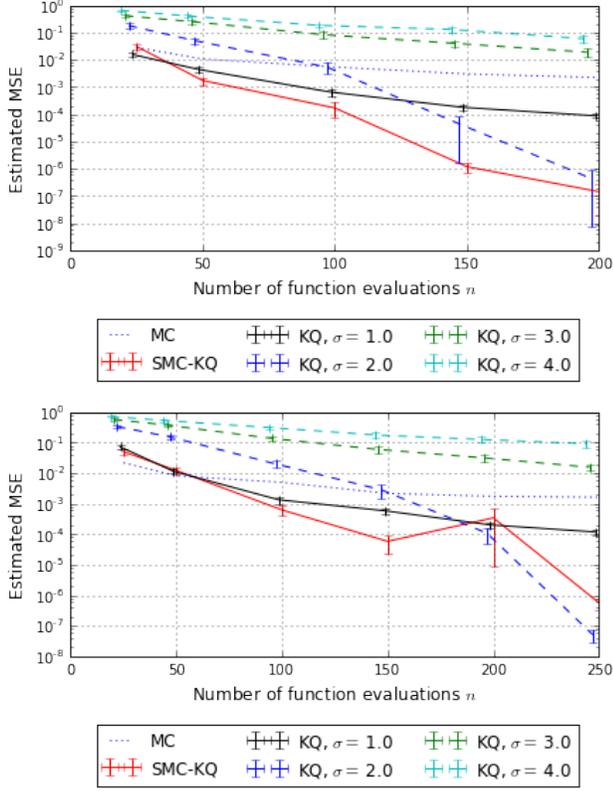


Figure 11. Performance of KQ and SMC-KQ on the integration problem with $f(x) = 1 + \sin(4\pi x)$ (top) and $f(x) = 1 + \sin(8\pi x)$ (bottom) integrated against $N(0, 1)$. The SMC sampler was initiated with a $N(0, 8^2)$ distribution. The kernel used was Gaussian with length scales $\ell = 0.25$ (top) and $\ell = 0.15$ (bottom) each chosen to reflect the complexity of the functions.

obtained the derivatives:

$$\begin{aligned} \frac{dk(\boldsymbol{\theta}, \boldsymbol{\phi})}{d\theta_j} &= -\frac{2}{\ell_j^2}(\theta_j - \phi_j)k(\boldsymbol{\theta}, \boldsymbol{\phi}) \\ \frac{dk(\boldsymbol{\theta}, \boldsymbol{\phi})}{d\phi_j} &= \frac{2}{\ell_j^2}(\theta_j - \phi_j)k(\boldsymbol{\theta}, \boldsymbol{\phi}) \\ \frac{dk(\boldsymbol{\theta}, \boldsymbol{\phi})}{d\theta_j d\phi_j} &= \frac{(2\ell_j^2 - 4(\theta_j - \phi_j)^2)}{\ell_j^4}k(\boldsymbol{\theta}, \boldsymbol{\phi}) \end{aligned}$$

Furthermore, we can obtain expressions for the score function for posterior densities as follows:

$$u_j(\boldsymbol{\theta}) = \frac{d}{d\theta_j} \log \pi(\boldsymbol{\theta}) + \frac{d}{d\theta_j} \log \pi(\mathbf{y}|\boldsymbol{\theta}).$$

A.6. Algorithms and Implementation

A.6.1. SMC SAMPLER

In Alg. 2 the standard SMC scheme is presented. Resampling occurs when the effective sample size, $\|\mathbf{w}\|_2^{-2}$

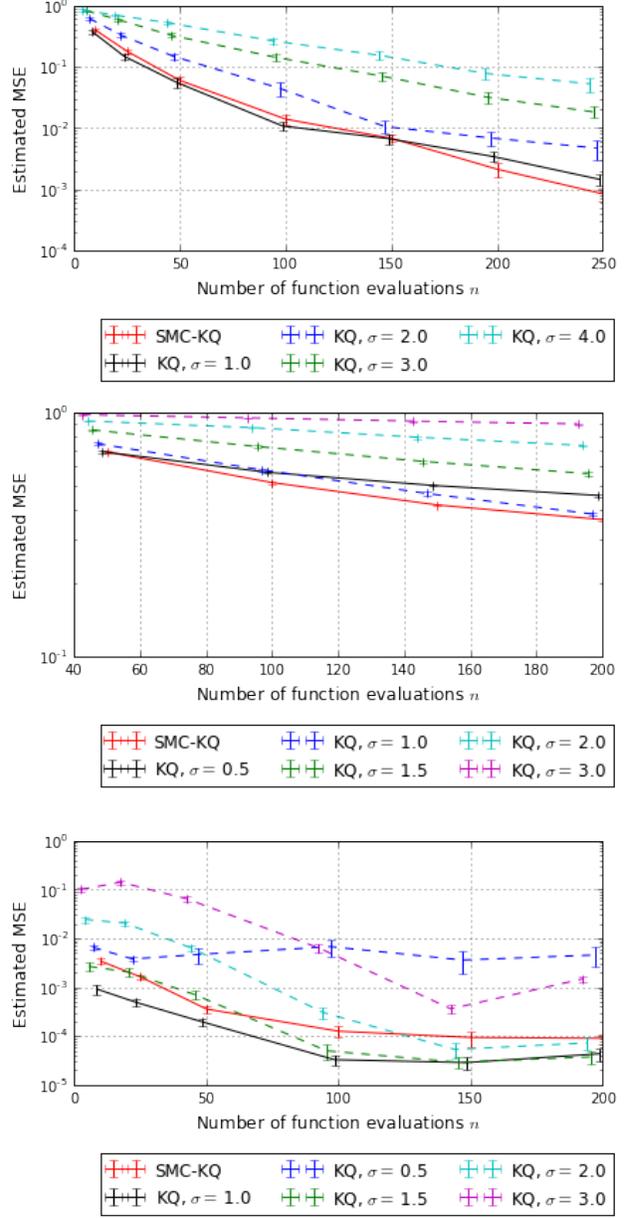


Figure 12. Performance of KQ and SMC-KQ on the integration problem with $f(\mathbf{x}) = 1 + \prod_{j=1}^d \sin(2\pi x_j)$ integrated against a $N(\mathbf{0}, \mathbf{I})$ distribution for $d = 2$ (top), $d = 3$ (middle) and $d = 10$ (bottom). The SMC sampler was initiated with a $N(\mathbf{0}, 8^2 \mathbf{I})$ distribution. The kernel used was a (multivariate) Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\sum_{j=1}^d (x_j - y_j)^2 / \ell_j^2)$ with the length scales $\ell_1 = \dots = \ell_d = 0.25$ were used.

drops below a fraction ρ of the total number N of particles. In this work we took $\rho = 0.95$ which is a common default.

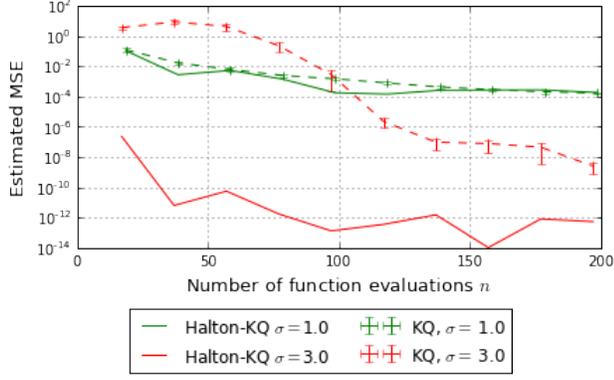


Figure 13. Comparison between KQ with $x_j \sim N(0, 1)$ independent and KQ with $x_j = \Phi^{-1}(u_j)$ where the $\{u_j\}_{j=1}^n$ are the first n terms in the Halton sequence and Φ is the standard Gaussian cumulative density function.

Algorithm 2 Sequential Monte Carlo Iteration

```

function SMC( $\{(w_j, \mathbf{x}_j)\}_{j=1}^N, t_i, t_{i-1}, \rho$ )
input  $\{(w_j, \mathbf{x}_j)\}_{j=1}^N$  (particle approx. to  $\Pi_{i-1}$ )
input  $t_i$  (next inverse-temperature)
input  $t_{i-1}$  (previous inverse-temperature)
input  $\rho$  (re-sample threshold)
 $w'_j \leftarrow w_j \times [\pi(\mathbf{x}_j)/\pi_0(\mathbf{x}_j)]^{t_i - t_{i-1}}$  ( $\forall j \in 1 : N$ )
 $\mathbf{w}' \leftarrow \mathbf{w}' / \|\mathbf{w}'\|_1$  (normalise weights)
if  $\|\mathbf{w}'\|_2^{-2} < N \cdot \rho$  then
     $\mathbf{a} \sim \text{Multinom}(\mathbf{w}')$ 
     $\mathbf{x}'_j \leftarrow \mathbf{x}_{a(j)}$  (re-sample  $\forall j \in 1 : N$ )
     $w'_j \leftarrow N^{-1}$  (reset weights  $\forall j \in 1 : N$ )
end if
 $\mathbf{x}'_j \sim \text{Markov}(\mathbf{x}'_j; \Pi_i, \{(w_j, \mathbf{x}_j)\}_{j=1}^N)$  (Markov update  $\in 1 : N$ )
return  $\{(w'_j, \mathbf{x}'_j)\}_{j=1}^N$  (particle approx. to  $\Pi_i$ )
    
```

Denote

$$\begin{aligned}
 q(\mathbf{x}, \cdot; \{(w_j, \mathbf{x}_j)\}_{j=1}^N) &= N(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
 \boldsymbol{\mu} &= \sum_{j=1}^N w_j \mathbf{x}_j \\
 \boldsymbol{\Sigma} &= \sum_{j=1}^N w_j (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^\top.
 \end{aligned}$$

The above standard adaptive independence proposal was used within a Metropolis-Hastings Markov transition:

Algorithm 3 Markov Iteration

```

function Markov( $\mathbf{x}, \pi, \{(w_j, \mathbf{x}_j)\}_{j=1}^N$ )
input  $\mathbf{x}$  (current state)
input  $\pi$  (density of invar. dist.)
 $\mathbf{x}^* \sim q(\mathbf{x}, \mathbf{x}^*; \{(w_j, \mathbf{x}_j)\}_{j=1}^N)$  (propose)

 $r \leftarrow \frac{\pi_i(\mathbf{x}^*)q(\mathbf{x}^*, \mathbf{x}; \{(w_j, \mathbf{x}_j)\}_{j=1}^N)}{\pi_i(\mathbf{x})q(\mathbf{x}, \mathbf{x}^*; \{(w_j, \mathbf{x}_j)\}_{j=1}^N)}$ 

 $u \sim \text{Unif}(0, 1)$ 
if  $u < r$  then
     $\mathbf{x} \leftarrow \mathbf{x}^*$  (accept)
end if return  $\mathbf{x}$  (next state)
    
```

A.6.2. CHOICE OF TEMPERATURE SCHEDULE

Following Zhou et al. (2016) we employed an adaptive temperature schedule construction. This was based on the conditional effective sample size of the SMC particle set, estimated as follows:

Algorithm 4 Conditional Effective Sample Size

```

function CESS( $\{(w_j, \mathbf{x}_j)\}_{j=1}^N, t$ )
input  $\{(w_j, \mathbf{x}_j)\}_{j=1}^N$  (particle approx.  $\Pi_{i-1}$ )
input  $t$  (candidate next inverse-temperature)
 $z_j \leftarrow [\pi(\mathbf{x}_j)/\pi_0(\mathbf{x}_j)]^{t_i - t_{i-1}}$  ( $\forall j \in 1 : N$ )
 $E \leftarrow N \left( \frac{\sum_{j=1}^N w_j z_j}{\sum_{j=1}^N w_j z_j^2} \right)^2$ 
return  $E$  (est'd. cond. ESS)
    
```

The specific construction for the temperature schedule is detailed in Alg. 5 below and makes use of a Sequential Least Squares Programming algorithm:

Algorithm 5 Adaptive Temperature Iteration

```

function temp( $\{(w_j, \mathbf{x}_j)\}_{j=1}^N, t_{i-1}, \rho, \Delta$ )
input  $\{(w_j, \mathbf{x}_j)\}_{j=1}^N$  (particle approx.  $\Pi_{i-1}$ )
input  $t_{i-1}$  (current inverse-temperature)
input  $\rho$  (re-sample threshold)
input  $\Delta$  (max. grid size, default  $\Delta = 0.1$ )
 $t \leftarrow \text{solve}(\text{CESS}(\{(w_j, \mathbf{x}_j)\}_{j=1}^N, t) = N \cdot \rho)$ 
(binary search in  $[t_{i-1}, 1]$ )
 $t_i \leftarrow \min\{t_{i-1} + \Delta, t\}$  return  $t_i$  (next inverse-temperature)
    
```

A.6.3. TERMINATION CRITERION

For SMC-KQ we estimated an upper bound on the worst case error in the unit ball of the Hilbert space \mathcal{H} . This was computed as follows, using a bootstrap algorithm:

Algorithm 6 Termination Criterion

```

function crit( $\Pi, k, \{\mathbf{x}_j\}_{j=1}^N$ )
input  $\Pi$  (target disn.)
input  $k$  (kernel)
input  $\{\mathbf{x}_j\}_{j=1}^N$  (collection of states)
 $R^2 \leftarrow 0$ 
 $e_0 \leftarrow \iint_{\mathcal{X} \times \mathcal{X}} k(\mathbf{x}, \mathbf{x}') \Pi \otimes \Pi(d\mathbf{x} \times d\mathbf{x}')$  (in'l error)
for  $m = 1, \dots, M$  do
     $\tilde{\mathbf{x}}_j \sim \text{Unif}(\{\mathbf{x}_j\}_{j=1}^N)$  ( $\forall j \in 1:n$ )
     $z_j \leftarrow \int_{\mathcal{X}} k(\cdot, \tilde{\mathbf{x}}_j) d\Pi$  (k'l mean eval.  $\forall j \in 1:n$ )
     $K_{j,j'} \leftarrow k(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_{j'})$  (kernel eval.  $\forall j, j' \in 1:n$ )
     $\mathbf{w} \leftarrow \mathbf{z}^T \mathbf{K}^{-1}$  (KQ weights)
     $e_n^2 \leftarrow \mathbf{w}^T \mathbf{K} \mathbf{w} - 2\mathbf{w}^T \mathbf{z} + e_0^2$ 
     $R^2 \leftarrow R^2 + e_n^2 M^{-1}$ 
end for
return  $R$  (est'd error)
    
```

Note that this could be slightly improved using a weighted bootstrap approach.

For SMC-KQ-KL an empirical upper bound on integration error was estimated. This requires that the norm $\|f\|_{\mathcal{H}}$ be estimated, which was achieved as follows:

Algorithm 7 Termination Crit. + Kernel Learning

```

function crit-KL( $f, \Pi, k, \{\mathbf{x}_j\}_{j=1}^N$ )
input  $f$  (integrand)
input  $\Pi$  (target disn.)
input  $k$  (kernel)
input  $\{\mathbf{x}_j\}_{j=1}^N$  (collection of states)
 $R^2 \leftarrow 0$ 
 $e_0 \leftarrow \iint_{\mathcal{X} \times \mathcal{X}} k(\mathbf{x}, \mathbf{x}') \Pi \otimes \Pi(d\mathbf{x} \times d\mathbf{x}')$  (in'l error)
for  $m = 1, \dots, M$  do
     $\tilde{\mathbf{x}}_j \sim \text{Unif}(\{\mathbf{x}_j\}_{j=1}^N)$  ( $\forall j \in 1:n$ )
     $f_j \leftarrow f(\tilde{\mathbf{x}}_j)$  (function eval.  $\forall j \in 1:n$ )
     $z_j \leftarrow \int_{\mathcal{X}} k(\cdot, \tilde{\mathbf{x}}_j) d\Pi$  (k'l mean eval.  $\forall j \in 1:n$ )
     $K_{j,j'} \leftarrow k(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_{j'})$  (kernel eval.  $\forall j, j' \in 1:n$ )
     $\mathbf{w} \leftarrow \mathbf{z}^T \mathbf{K}^{-1}$  (KQ weights)
     $e_n^2 \leftarrow \mathbf{w}^T \mathbf{K} \mathbf{w} - 2\mathbf{w}^T \mathbf{z} + e_0^2$ 
     $R^2 \leftarrow R^2 + e_n^2 M^{-1}$ 
end for
 $z_j \leftarrow \int_{\mathcal{X}} k(\cdot, \mathbf{x}_j) d\Pi$  (kernel mean eval.  $\forall j \in 1:n$ )
 $K_{j,j'} \leftarrow k(\mathbf{x}_j, \mathbf{x}_{j'})$  (kernel eval.  $\forall j, j' \in 1:n$ )
 $\mathbf{w} \leftarrow \mathbf{z}^T \mathbf{K}^{-1}$  (KQ weights)
 $S^2 \leftarrow R^2 \times \mathbf{w}^T \mathbf{K} \mathbf{w}$  return  $S$  (est'd error bound)
    
```

In Alg. 7 the literal interpretation, that f is re-evaluated on values of \mathbf{x}_j which have been previously examined, is clearly inefficient. In practice such function evaluations were cached and then do not contribute further to the total number of function evaluations that are required in the algorithm.

A.6.4. KERNEL LEARNING

A generic approach to select kernel parameters is the maximum marginal likelihood method:

Algorithm 8 Parameter Update

```

function kern-param( $\mathbf{f}, \{\mathbf{x}_j\}_{j=1}^n, k_\theta$ )
input  $\mathbf{f}$  (integrand evals.)
input  $\{\mathbf{x}_j\}_{j=1}^n$  (associated states)
input  $k_\theta$  (parametric kernel)
 $\theta' \leftarrow \arg \min_{\theta} \mathbf{f}^T \mathbf{K}_\theta^{-1} \mathbf{f} + \log |\mathbf{K}_\theta|$  (numer. opt.)
(s.t.  $K_{\theta,j,j'} = k_\theta(\mathbf{x}_j, \mathbf{x}_{j'})$ ) return  $\theta'$  (optimal params)
    
```

A.6.5. IMPLEMENTATION OF SMC-KQ-KL

Our final algorithm to present is the full implementation for SMC-KQ-KL:

Algorithm 9 SMC for KQ with Kernel Learning

```

function SMC-KQ-KL( $f, \Pi, k_\theta, \Pi_0, \rho, n, N$ )
input  $f$  (integrand)
input  $\Pi$  (target disn.)
input  $k_\theta$  (parametric kernel)
input  $\Pi_0$  (reference disn.)
input  $\rho$  (re-sample threshold)
input  $n$  (num. func. evaluations)
input  $N$  (num. particles)
 $i \leftarrow 0; t_i \leftarrow 0; R_{\min} \leftarrow \infty$ 
 $\mathbf{x}'_j \sim \Pi_0$  (initialise states  $\forall j \in 1 : N$ )
 $w'_j \leftarrow N^{-1}$  (initialise weights  $\forall j \in 1 : N$ )
 $\theta' \leftarrow \text{kern-param}(f, \{\mathbf{x}'_j\}_{j=1}^n)$  (kernel params)
 $R \leftarrow \text{crit-KL}(f, \Pi, k_{\theta'}, \{\mathbf{x}'_j\}_{j=1}^N)$  (est'd error)
while  $\text{test}(R < R_{\min})$  and  $t_i < 1$  do
     $i \leftarrow i + 1; R_{\min} \leftarrow R; \theta \leftarrow \theta'$ 
     $\{(w_j, \mathbf{x}_j)\}_{j=1}^N \leftarrow \{(w'_j, \mathbf{x}'_j)\}_{j=1}^N$ 
     $t_i \leftarrow \text{temp}(\{(w_j, \mathbf{x}_j)\}_{j=1}^N, t_{i-1})$  (next temp.)
     $\{(w'_j, \mathbf{x}'_j)\}_{j=1}^N \leftarrow \text{SMC}(\{(w_j, \mathbf{x}_j)\}_{j=1}^N, t_i, t_{i-1}, \rho)$ 
    (next particle approx.)
     $\theta' \leftarrow \text{kern-param}(f, \{\mathbf{x}'_j\}_{j=1}^n)$  (kernel params)
     $R \leftarrow \text{crit-KL}(f, \Pi, k_{\theta'}, \{\mathbf{x}'_j\}_{j=1}^N)$  (est'd error)
end while
 $\mathbf{f}_j \leftarrow f(\mathbf{x}_j)$  (function eval.  $\forall j \in 1 : n$ )
 $z_j \leftarrow \int_{\mathcal{X}} k_\theta(\cdot, \mathbf{x}_j) d\Pi$  (kernel mean eval.  $\forall j \in 1 : n$ )
 $\mathbf{K}_{j,j'} \leftarrow k_\theta(\mathbf{x}_j, \mathbf{x}_{j'})$  (kernel eval.  $\forall j, j' \in 1 : n$ )
 $\hat{\Pi}(f) \leftarrow \mathbf{z}^\top \mathbf{K}^{-1} \mathbf{f}$  (eval. KQ estimator) return  $\hat{\Pi}(f)$ 
    (estimator)
    
```

As stated here, Alg. 9 is inefficient as function evaluations that are produced in the `kern-param` and `crit-KL` components are not included in the KQ estimator $\hat{\Pi}(f)$. Thus a trivial modification is to store all function evaluations (f_j, \mathbf{x}_j) that are produced and to include all of these in the ultimate KQ estimator. This was the approach taken in our experiments that involved SMC-KQ-KL. However, since it is somewhat cumbersome to include in the pseudo-code, we have not made this explicit in the notation. Our reported results are on a per-function-evaluation basis and so we **do** adjust for this detail in our reported comparisons.