Conditional Bayesian Quadrature

Zonghao Chen
Department of Computer Science
University College London

Masha Naslidnyk
Department of Computer Science
University College London

Arthur Gretton
Gatsby Computational Neuroscience Unit
University College London

François-Xavier Briol
Department of Statistical Science
University College London

Abstract

We propose a novel approach for estimating conditional expectations uniformly over classes of functions, which is able to incorporate prior information about the integrands. We employ the framework of probabilistic numerical methods such as Bayesian quadrature. As a result, our method provides a way of quantifying our uncertainty, and leads to a fast convergence rate which is confirmed both theoretically and empirically on challenging tasks in mathematical finance, Bayesian sensitivity analysis and health economics.

1 Introduction

This paper considers the computational challenge of estimating certain intractable expectations which arise in machine learning, statistics, and beyond. Given a function $f : \mathcal{X} \times \Theta \to \mathbb{R}$, we are interested in estimating certain *conditional expectations* (sometimes also called parametric expectations) $I : \Theta \to \mathbb{R}$ uniformly over the parameter space Θ, where:

$$I(\theta) = \mathbb{E}_{X \sim \mathbb{P}_\theta}[f(X, \theta)] = \int_{\mathcal{X}} f(x, \theta) \mathbb{P}_\theta(\mathrm{d}x),$$

and $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ is a family of distributions on the integration domain $\mathcal{X}$. We will assume that $I(\theta)$ is sufficiently smooth in θ, but that it is not available in closed form and must be approximated through samples and function evaluations.

Conditional expectations arise when calculating tail probabilities in rare-event simulation [72], computing moment generating, characteristic, or cumulative distribution functions [27, 45]. It also arises when computing the conditional value at risk or various valuations of options [49, 5], for Bayesian sensitivity analysis [50, 37], or even more broadly for scientific sensitivity analysis; see for example Sobol indices [68]. Parametric expectations $I(\theta)$ are also often computed as an intermediate quantity. For example, given $\phi : \mathbb{R} \to \mathbb{R}$ and some probability distribution $\mathbb{Q}$ on Θ, we are often interested in *nested expectation* given by $\mathbb{E}_{\theta \sim \mathbb{Q}}[\phi(I(\theta))]$ [35, 64]. These arise when computing the expected information gain in Bayesian experimental design [16], and for computing the expected value of partial perfect information in health economics [31].

Methods for computing $I(\theta)$ generally select T parameter values $\theta_1, \ldots, \theta_T \in \Theta$, then simulate N realisations from each corresponding probability distribution $\mathbb{P}_{\theta_1}, \ldots, \mathbb{P}_{\theta_T}$ at which they evaluate the integrand f, leading to a total of NT evaluations. Classical Monte Carlo can be used to estimate $I(\theta_1), \ldots, I(\theta_T)$, but in many applications we are also interested in estimating either $I(\theta^*)$ for a fixed $\theta^* \notin \{\theta_1, \ldots, \theta_T\}$, or $I(\theta)$ uniformly over $\theta \in \Theta$. As a result, a second step to combine $I(\theta_t)$ is required for the estimate.

Preprint. Under review.

The most straightforward approach to estimate conditional expectation is importance sampling [28, 51, 72, 19], where $I(\theta)$ is estimated by weighting function evaluations to account for the fact that the samples were not obtained from $\mathbb{P}_\theta$. Unfortunately, this approach is only applicable when f does not depend on θ, and it is usually difficult to identify an appropriate importance distribution. Alternatively, least-squares Monte Carlo [49, 5] or regression-based kernel mean shrinkage estimators [55, 17] first estimate $I(\theta_1), \ldots, I(\theta_T)$ through Monte Carlo, then estimate $I(\theta)$ through either linear, polynomial or kernel ridge regression based on these T Monte Carlo estimators. These methods are therefore dependent on the accuracy of the Monte Carlo estimators and of the regression method.

In addition, there are two main limitations which all of these methods suffer from. Firstly, they are very sample-intensive; i.e. they require a large number of function evaluations to reach a given level of accuracy, which makes them infeasible if sampling or evaluating the integrand is expensive. Secondly, obtaining a good, finite-sample, quantification of uncertainty for $I(\theta)$ is often infeasible. This is a significant limitation for challenging integration problems, where we would ideally like to know how accurate our estimator is likely to be.

To tackle these limitations, we propose a novel algorithm called *conditional Bayesian quadrature* (CBQ). The name comes from the fact that our approach extends the Bayesian quadrature algorithm [21, 63, 66, 14] to the computation of conditional expectations. As such, CBQ falls in the line of work on probabilistic numerical methods [32, 18, 62, 33]. Our algorithm is based on a hierarchical Bayesian model consisting of two-stages of Gaussian process regression, and leads to a univariate Gaussian posterior distribution on $I(\theta)$ whose mean and variance are parametrised by θ.

This approach allows us to mitigate the two main limitations of existing methods. Firstly, we show both theoretically and empirically that our method is more sample efficient than alternatives under mild smoothness conditions on f and $I(\theta)$ whenever the dimension of $\mathcal{X}$ and Θ is not too large. As a result, smaller N and T are needed to achieve a desired accuracy, and the method will therefore be preferable for expensive problems. Secondly, the fact that we have an entire posterior distribution on $I(\theta)$ allows us to provide finite-sample Bayesian quantification of uncertainty.

The remainder of the paper is structured as follows: In Section 2, we review existing methods for computing conditional expectations and Bayesian quadrature. In Section 3, we formalize our algorithm *conditional Bayesian quadrature*. In Section 4, we prove the convergence rate of our method. In Section 5, we provide empirical results and compare with other baseline methods.

2 Background

We aim to compute conditional expectations of the form $I(\theta) = \mathbb{E}_{X\sim\mathbb{P}_\theta}[f(X,\theta)]$, where we will assume that $\mathcal{X} \subseteq \mathbb{R}^d$, $\Theta \subseteq \mathbb{R}^p$, and the integrand $f(\cdot,\theta)$ is in $L^2(\mathbb{P}_\theta) := \{h : \mathcal{X} \to \mathbb{R} : \mathbb{E}_{X\sim\mathbb{P}_\theta}[h^2(X)] < \infty\}$, the space of square-integrable functions with respect to $\mathbb{P}_\theta$ for all $\theta \in \Theta$. The latter is a minimal assumption which ensures that Monte Carlo estimators satisfy the central limit theorem. We will assume that our observations are parameters, points, and function values:

$$\theta_{1:T} := [\theta_1 \cdots \theta_T]^\top \in \Theta^T, \quad x^t_{1:N} := [x^t_1 \cdots x^t_N]^\top \in \mathcal{X}^N \quad \text{and}$$
$$f(x^t_{1:N}, \theta_t) = [f(x^t_1, \theta_t) \cdots f(x^t_N, \theta_t)]^\top \in \mathbb{R}^N \quad \text{for all } t \in \{1, \ldots, T\},$$

where we use square brackets to indicate vectors. Our method could straightforwardly be extended to allow a different number of samples N_t per parameter value θ_t, but we do not consider this case in order to simplify notation throughout. In this section, we will review existing methods for conditional expectations and the core ingredient for our method: the Bayesian quadrature algorithm.

2.1 Methods for Conditional Expectations

Sampling-based Methods. Assume that $x^t_{1:N} \sim \mathbb{P}_{\theta_t}$ for all $t \in \{1, \ldots, T\}$. Then, we can construct a *Monte Carlo* (MC) estimator for $I(\theta_t)$ through $\hat{I}_{\text{MC}}(\theta_t) := \frac{1}{N}\sum_{i=1}^N f(x^t_i, \theta_t)$ [67]. The disadvantages of this approach are that we cannot estimate $I(\theta)$ for $\theta \notin \{\theta_1 \ldots, \theta_T\}$, and that we can only use N, rather than NT, samples per estimator.

If we assume $\mathbb{P}_\theta$ has a Lebesgue density $p_\theta : \mathcal{X} \to \mathbb{R}$ which has full support on $\mathcal{X}$ for all $\theta \in \Theta$, and the integrand does not depend on θ (i.e. $f(x,\theta) = f(x)$), then the *importance sampling* (IS) estimator is able to make use of all NT samples and can estimate $I(\theta)$ for any parameter

$\theta \in \Theta$: $\hat{I}_{\text{IS}}(\theta) := \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}(p_{\theta_t}(x_i^t)/p_\theta(x_i^t))f(x_i^t)$. In this case, the importance distributions are selected to be the conditional distributions $\mathbb{P}_{\theta_1}, \ldots, \mathbb{P}_{\theta_T}$. There are other choices of importance distributions that allow for a more flexible form [19] and that can minimize the variance [28, 51, 72].

Regression-based Methods. Once again, assume that $x_{1:N}^t \sim \mathbb{P}_{\theta_t}$ for all $t \in \{1, \ldots, T\}$. Regression-based methods such as least-squares Monte Carlo (LSMC) [49] and kernel mean shrinkage estimators (KMS) [17] are also two-stage approaches. The first stage consists of computing MC estimators for the parameter values at which samples are generated: $\hat{I}_{\text{MC}}(\theta_1), \ldots, \hat{I}_{\text{MC}}(\theta_T)$. The second stage, then consists of estimating $I(\theta)$ by interpolating based on the estimators from the first stage: $\arg\min_{\phi \in \mathcal{F}(\Theta)} \frac{1}{T}\sum_{t=1}^{T}(\phi(\theta_t) - \hat{I}_{\text{MC}}(\theta_t))^2$.

The LSMC estimator $\hat{I}_{\text{LSMC}}(\theta)$ solves the problem for $\mathcal{F}(\Theta)$ being a space of order$-p$ polynomials, whereas the KMS estimator $\hat{I}_{\text{KMS}}(\theta)$ solves it for $\mathcal{F}(\Theta)$ being a ball in a reproducing kernel Hilbert space (RKHS) [9][1]. Clearly, both the performance and computational cost of these estimators will depend on the choice of family. LSMC costs $\mathcal{O}(TN + p^3)$, whereas KMS costs $\mathcal{O}(TN + T^3)$. On the other hand, KMS will lead to a smaller error than LSMC when $I(\theta)$ cannot be well approximated by a low-order polynomial.

Other Related Work. Alternative approaches for estimating conditional expectation are based on sub-fields of transfer learning, such as multi-task learning [79, 25, 70] or meta-learning [71]. This line of research tends to assume that several related integrals need to be computed, and the relationship between these integrals is encoded through a vector-valued RKHS or by assuming they are independent draws from an environment of integration tasks. They do not explicitly utilise the property that the conditional expectation is a smooth function of the parameter. Multilevel Monte Carlo methods are also popular in estimating the expectation of random variables, by combining samples from multiple levels of resolution [27]. However, they are not able to estimate new integrals $I(\theta^*)$ or $I(\theta)$ uniformly over $\theta \in \Theta$.

2.2 Bayesian Quadrature

Consider the expectation $I = \mathbb{E}_{X\sim\mathbb{P}}[f(X)]$ of some function $f : \mathcal{X} \to \mathbb{R}$, where we emphasise that neither f nor $\mathbb{P}$ depend on θ in this subsection. In Bayesian quadrature (BQ) [21, 63, 66, 14], we begin by positing a Gaussian process (GP) prior on f; a GP [65] is a distribution over functions such that every finite dimensional distribution (i.e. evaluations of the functions at a finite number of points) is Gaussian. We will denote this prior $\mathcal{GP}(m_\mathcal{X}, k_\mathcal{X})$, where $m_\mathcal{X} : \mathcal{X} \to \mathbb{R}$ is known as the mean function and $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the covariance (or reproducing kernel) function. These two functions fully characterise the distribution, and can be used to encode prior knowledge about smoothness, periodicity, or sparsity of f. Once a GP prior has been identified, we condition this prior on data $f(x_{1:N}) = [f(x_1), \ldots, f(x_N)]^\top$ for $x_{1:N} \in \mathcal{X}^N$. This leads to a posterior GP on f, which induces a univariate Gaussian posterior distribution $\mathcal{N}(\hat{I}_{\text{BQ}}, \sigma_{\text{BQ}}^2)$ on I, where:

$$
\begin{aligned}
\hat{I}_{\text{BQ}} &= \mathbb{E}_{X\sim\mathbb{P}}[m_\mathcal{X}(X)] + \mu(x_{1:N})^\top(k_\mathcal{X}(x_{1:N}, x_{1:N}) + \lambda_\mathcal{X}\text{Id}_N)^{-1}(f(x_{1:N}) - m_\mathcal{X}(x_{1:N})),\\
\sigma_{\text{BQ}}^2 &= \mathbb{E}_{X,X'\sim\mathbb{P}}[k_\mathcal{X}(X, X')] - \mu(x_{1:N})^\top(k_\mathcal{X}(x_{1:N}, x_{1:N}) + \lambda_\mathcal{X}\text{Id}_N)^{-1}\mu(x_{1:N}).
\end{aligned}
$$

The function $\mu(x) = \mathbb{E}_{X\sim\mathbb{P}}[k_\mathcal{X}(X, x)]$ is known as the kernel mean embedding of the distribution $\mathbb{P}$ [54] and $\mathbb{E}_{X,X'\sim\mathbb{P}}[k_\mathcal{X}(X, X')]$ is known as the initial error. Here $\lambda_\mathcal{X} \geq 0$ is a regularisation parameter, often called “nugget” or “jitter”, which, although not essential from a statistical viewpoint, is often used to ensure the matrix can be numerically inverted [1, 7] and is known not to impact the asymptotic convergence rate of the GP [76].

The posterior mean $\hat{I}_{\text{BQ}}$ provides a point estimate for I, and is known as the quadrature rule since it is an affine combination of function evaluations. It is noteworthy that BQ does not impose restriction on how $x_{1:N}$ is selected, and as such does not require independent realisations from $\mathbb{P}$. In fact, a number of active learning approaches have been proposed, see [30, 25]. The posterior variance σ_{BQ}^2 gives us a notion of epistemic uncertainty for I, where the uncertainty arises due to the fact that we have only observed f at N points. For our model to be well-calibrated and the posterior variance σ_{BQ}^2 to be

[1] A more appropriate name for this algorithm would be “kernel regression Monte Carlo", but we follow the terminology in [17] for simplicity and to avoid confusing readers familiar with this literature.

meaningful, we need to select the GP prior and all associated hyperparameters carefully. Doing so a-priori can be challenging, and the most common approach is therefore empirical Bayes [15]. A detailed discussion is provided in Appendix B.2.

The convergence rate of the BQ estimator has been studied extensively [14, 38, 78] and is particularly fast for low- to mid-dimensional problems where f is smooth. This has to be contrasted with the computational cost, which is inherited from GP regression and is $\mathcal{O}(N^3)$. For this reason, BQ has prominently been applied to problems where sampling or evaluating the integrand is expensive, or N is otherwise small. Examples range from differential equation solvers [44], variational inference [2] and simulator-based inference [10] to problems in engineering and the sciences including computer graphics [52, 79], cardiac modelling [59] and tsunami modelling [48]. For cheaper problems, [41, 42] and [36] propose BQ methods where the computational cost is much lower, but these are applicable only with specific point sets $x_{1:N}$ and measures $\mathbb{P}$.

We note that the computing the BQ mean and variance requires that μ and its integral are known in closed-form; see Table 1 in [14], [56] or the `ProbNum` package [77] for examples. It is a rather strong requirement as it is easy to encounter pairs of kernel and distribution for which this does not hold. Fortunately, there are multiple solutions for it and a detailed discussion is provided in Appendix B.1. Here, we briefly mention a popular solution based on Stein reproducing kernels [6].

Stein reproducing kernels For any reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we can obtain a Stein reproducing kernel by applying a Langevin Stein operator on both arguments of the kernel k [6]:

$$\begin{aligned} k_p(x, x') := &\nabla_x \log p(x)^\top k(x, x') \nabla_{x'} \log p(x') + \nabla_x \log p(x)^\top \nabla_{x'} k(x, x') \\ &+ \nabla_{x'} \log p(x')^\top \nabla_x k(x, x') + \nabla_x \cdot \nabla_{x'} k(x, x') \end{aligned}$$

where $\nabla_x = (\partial/\partial x_1, \cdots, \partial/\partial x_d)^\top$. The main advantage of this construction is that the mean embedding $\mu(x') = \int_{\mathcal{X}} k_p(x, x') p(x) dx = 0$ by construction. However, this means our GP prior on f encodes beliefs that the function has mean zero, which we do not have. Therefore, we add a constant $c \in \mathbb{R}$ to the Stein kernel k_p, i.e $\tilde{k}_p(x, x') = k_p(x, x') + c$, so that the kernel mean embedding becomes $\mu(x') = c$. The constant c can be treated as a kernel hyperparameter and estimated alongside all other parameters. Stein reproducing kernels are also non-stationary, which implies prior beliefs that f has different properties across different parts of $\mathcal{X}$. Therefore, using a GP prior with Stein kernel as covariance function requires additional caution. Fortunately, our experiments in Section 5 and Appendix C do not exhibit a huge difference between Stein kernel and traditional kernels.

3 Methodology

Conditional Bayesian quadrature (CBQ) provides a Bayesian hierarchical model for $I(\theta^*)$ for any $\theta^* \in \Theta$, and the posterior mean of this hierarchical model is called the CBQ estimator. The algorithm can be expressed in two stages:

- **Stage 1:** Compute $\hat{I}_{\mathrm{BQ}}(\theta_{1:T}), \sigma^2_{\mathrm{BQ}}(\theta_{1:T})$ to obtain the BQ posteriors on $I(\theta_1), \ldots, I(\theta_T)$.
- **Stage 2:** Perform GP regression over $I(\theta)$ using the outputs of stage 1. The posterior mean $\hat{I}_{\mathrm{CBQ}}(\theta^*)$ is the CBQ estimator for $I(\theta^*)$, and the variance $k_{\mathrm{CBQ}}(\theta^*, \theta^*)$ quantifies uncertainty.

This can be summarised using the direct acyclic graph in Figure 1, where the first stage corresponds to the part of the model inside the plate over $t \in \{1, \ldots, T\}$, and the second stage corresponds to the remainder of the graph. The CBQ posterior mean and covariance are given by

$$\hat{I}_{\mathrm{CBQ}}(\theta) := m_\Theta(\theta) + k_\Theta(\theta, \theta_{1:T})^\top \left(k_\Theta(\theta_{1:T}, \theta_{1:T}) + (\lambda_\Theta + \sigma^2_{\mathrm{BQ}}(\theta_{1:T}))\mathrm{Id}_T\right)^{-1} \hat{I}_{\mathrm{BQ}}(\theta_{1:T}),$$

$$k_{\mathrm{CBQ}}(\theta, \theta') := k_\Theta(\theta, \theta') - k_\Theta(\theta, \theta_{1:T})^\top (k_\Theta(\theta_{1:T}, \theta_{1:T}) + (\lambda_\Theta + \sigma^2_{\mathrm{BQ}}(\theta_{1:T}))\mathrm{Id}_T)^{-1} k_\Theta(\theta_{1:T}, \theta')$$

where $\lambda_\Theta \geq 0$ is a regularisation parameter, $\hat{I}_{\mathrm{BQ}}(\theta_t)$ and $\sigma^2_{\mathrm{BQ}}(\theta_t)$ are the BQ posterior mean and variance for $I(\theta_t)$, and $m_\Theta : \Theta \to \mathbb{R}$ and $k_\Theta : \Theta \times \Theta \to \mathbb{R}$ are the prior mean and covariance for the GP in the second stage. Similarly to BQ, the “quadrature" terminology is justified since $\hat{I}_{\mathrm{CBQ}}(\theta) := \sum_{t=1}^T \sum_{i=1}^N w_{ti}^{\mathrm{CBQ}} f(x_i^t, \theta_t)$ for some weights $w_{it}^{\mathrm{CBQ}} \in \mathbb{R}$ when $m_\Theta(\theta) = 0$.

The first stage corresponds to the BQ procedure highlighted in Section 2.2: we model $f(\cdot, \theta_t)$ with independent GP$(m^t_{\mathcal{X}}, k^t_{\mathcal{X}})$ priors, condition on observations $f(x^t_{1:N}, \theta_t)$, and consider the posterior

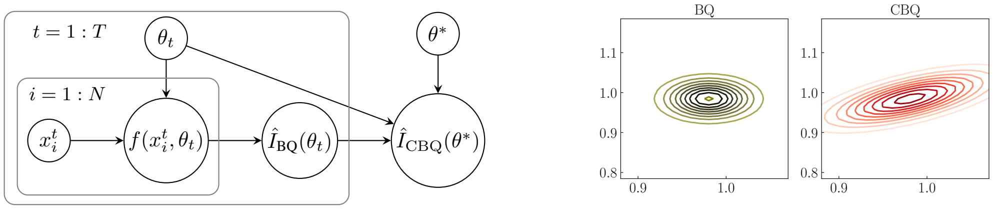


Figure 1: *Illustration of conditional Bayesian quadrature (CBQ).* **Left:** Directed acyclic graph representation of CBQ. Circle nodes indicate random variables and large rectangles correspond to independent replications over indices. **Right:** Posteriors on $I(\theta_{1:2}) = [I(\theta_1), I(\theta_2)]^\top$ for $\theta_1 \approx \theta_2$. Unlike BQ, the CBQ posterior takes into account the relation between the two quantities.

distribution on $I(\theta_t)$ for all $t \in \{1, \dots, T\}$. We therefore require access to closed-form expressions for each of the T kernel mean embeddings and initial errors. Note that at this stage, we do not share any samples across the estimators of $I(\theta_1), \dots, I(\theta_T)$.

In the second stage, we place a GP(m_Θ, k_Θ) prior on $I : \Theta \to \mathbb{R}$, and assume $\hat{I}_{\mathrm{BQ}}(\theta_t)$ are noisy evaluations of $I(\theta_t)$: $\hat{I}_{\mathrm{BQ}}(\theta_t) = I(\theta_t) + \varepsilon_t$, where the noise terms $\varepsilon_{1:T}$ are independent zero-mean Gaussian noise with variance $\sigma^2_{\mathrm{BQ}}(\theta_1), \dots, \sigma^2_{\mathrm{BQ}}(\theta_T)$ respectively. Since the variance is input-dependent, this corresponds to heteroscedastic GP regression [46]. We now briefly comment on the choice of prior and likelihood in this second stage:

- The GP(m_Θ, k_Θ) prior can be used to encode prior knowledge about how the expectation $I(\theta)$ varies with the parameter θ. Typically, the stronger this prior information, the faster the CBQ estimator's convergence rate will be; this statement will be made formal in Section 4.
- The likelihood for the heteroscedastic GP is directly inherited from the BQ posteriors in the first stage: the posterior on $I(\theta_t)$ is a univariate normal with mean $\hat{I}_{\mathrm{BQ}}(\theta_t)$ and variance $\sigma^2_{\mathrm{BQ}}(\theta_t)$. As expected, when the number of samples N grows, the BQ variance $\sigma^2_{\mathrm{BQ}}(\theta_t)$ will decrease, indicating that we are more certain about $I(\theta_t)$. This is then directly taken into account in stage 2.

The total computational cost of our approach is $\mathcal{O}(TN^3 + T^3)$ due to the need to compute T BQ estimators in the first stage and heteroscedastic GP regression in the second stage. Fast GP approaches could be used to significantly reduce this cost (see e.g. [74]). The requirement for a closed form kernel mean has been discussed in Section 2.2 and a detailed discussion is in Appendix B.1.

Interestingly, CBQ also provides us with a straightforward way of obtaining a joint posterior on the expectation at $\theta_1^*, \dots, \theta_{T_{\text{Test}}}^* \in \Theta$. This posterior is a multivariate Gaussian with mean vector $\hat{I}_{\mathrm{CBQ}}(\theta^*_{1:T_{\text{Test}}})$ and covariance matrix $k_{\mathrm{CBQ}}(\theta^*_{1:T_{\text{Test}}}, \theta^*_{1:T_{\text{Test}}})$, which can be computed at an additional $\mathcal{O}(T^2 T_{\text{test}})$ cost. This is illustrated in the right plot of Figure 1 on a synthetic example from Section 5; as observed, CBQ takes into account that the expectation will be similar for similar parameter values.

CBQ is closely related to LSMC and KMS as it simply corresponds to different choices for the two stages. The main difference is in stage 1, where we use BQ rather than MC. This is where we expect the greatest gains for our approach due to the fast convergence rate of BQ estimators (this will be confirmed in Section 4). However, one disadvantage of our approach is that it will require cubic operations in N and optimisation of hyperparameters at each of the T parameter values, whereas MC has a linear cost in N and no hyperparameters. For stage 2, we use heteroscedastic GP regression rather than polynomial or kernel ridge regression. As such, the second stage of KMS and CBQ is identical up to a minor difference in the way in which the Gram matrix $k_\Theta(\theta_{1:T}, \theta_{1:T})$ is regularised before inversion. Finally, one significant advantage of CBQ over LSMC and KMS is that it is a fully Bayesian model, meaning that we obtain a posterior distribution on $I(\theta)$ for any $\theta \in \Theta$.

A natural alternative would be to place a GP prior directly on $(x, \theta) \mapsto f(x, \theta)$ and condition on observations. The implied distribution on $I(\theta_1), \dots, I(\theta_T)$ would also be a multivariate Gaussian distribution. This approach coincides with the multi-output Bayesian quadrature approach of [79]. However, the computational cost is $\mathcal{O}(N^3 T^3)$, due to fitting a GP on NT observations, which quickly becomes intractable as N or T grow. The same holds true if f does not depend on θ (but $\mathbb{P}_\theta$ does), in which case the task reduces to the conditional mean process with NT observations as studied in Proposition 3.2 of [17], and when $T = 1$, we then recover standard BQ.

4 Theoretical Results

Our main result is Theorem 1 below, which guarantees that CBQ is able to approximate conditional expectations when T grows. To derive the result, we combine existing results on the convergence of GP interpolation from [78], with results on importance-weighted kernel ridge regression from [29]; see Appendix A for the proof.

The theorem will depend on the smoothness of the problem. We will say a function has smoothness s if it is in the Sobolev space of functions with at least s (weak) derivatives that are square Lebesgue-integrable. Similarly, we will say a kernel has smoothness s whenever its corresponding RKHS is a space of functions of smoothness s. This is for example the case of the Matérn$-\nu$ kernel in dimension d whenever $s = \nu + d/2$, defined as $k_\nu(x,y) = \frac{\eta}{\Gamma(\nu)2^{\nu-1}}(\frac{\sqrt{2\nu}}{l}\|x-y\|_2)^\nu K_\nu(\frac{\sqrt{2\nu}}{l}\|x-y\|_2)$ where K_ν is the modified Bessel function of the second kind. Furthermore, to quantify smoothness in a way specific to the task, we use the *source condition* of [29] for $r \in [1/2, 1]$. This is a standard condition in the kernel methods literature that compares the smoothness of $I(\theta)$ to the least smooth function in the RKHS of k_Θ: if k_Θ is not smoother than $I(\theta)$, the condition holds for $r = 1/2$, and a larger r implies smoother $I(\theta)$; see Appendix A. Lastly, we denote $\|f\|^2_{\mathcal{L}^2(\mathcal{X})} = \int_{\mathcal{X}} f^2(x)\mathrm{d}x$.

Theorem 1. *Suppose the following assumptions hold:*

1. *The domains $\mathcal{X} \subset \mathbb{R}^d$ and $\Theta \subset \mathbb{R}^p$ are open, convex, and bounded.*
2. *The parameters and samples satisfy: $\theta_{1:T} \sim \mathrm{Unif}(\Theta)$, and $x^t_{1:N} \sim \mathbb{P}_{\theta_t}$ for all t.*
3. *$\mathbb{P}_\theta$ has density p_θ for any $\theta \in \Theta$, $\inf_{\theta\in\Theta, x\in\mathcal{X}} p_\theta(x) > 0$ and $\sup_{\theta\in\Theta} \|p_\theta\|_{\mathcal{L}^2(\mathcal{X})} < \infty$.*
4. *The kernels $k_\mathcal{X}$ and k_Θ are Matérn kernels of smoothness $s_\mathcal{X} > d/2$ and $s_\Theta > p/2$ respectively.*
5. *The function $x \mapsto f(x,\theta)$ is of smoothness at least $s_\mathcal{X}$, $\theta \mapsto I(\theta)$ is of smoothness at least s_Θ.*
6. *The source condition for $I(\theta)$ and k_Θ holds with $r \in [1/2, 1]$.*

Then, letting $\lambda_\mathcal{X} = 0$ and $\lambda_\Theta = \mathcal{O}(T^{\frac{1}{2r+1}})$, we have that for any $\delta \in (0,1)$ there is a $T_0(\delta) > 0$ and an $N_0 > 0$ such that for any $N \geq N_0$ and $T \geq T_0$, with probability at least $1-\delta$ it holds that

$$\|\hat{I}_{\mathrm{CBQ}}(\theta) - I(\theta)\|_{\mathcal{L}^2(\Theta)} \leq C_1(\delta)T^{-\frac{r}{2r+1}} + C_2(\delta)T^{-\frac{r+1}{2r+1}}N^{-\frac{s_\mathcal{X}}{d}+\varepsilon},$$

for any arbitrarily small $\varepsilon > 0$, and $C_1(\delta) = \mathcal{O}(\log(1/\delta))$ and $C_2(\delta) = \mathcal{O}((1/\delta)\log(1/\delta))$ that are independent of N, T, ε.

The result is expressed in probability to account for randomness in $\theta_{1:T}$ and $x^t_{1:N}$ for all t. It indicates that growing N will only help up to some extent (by making the second term approach zero), but that growing T is essential to ensure convergence. This is intuitive since we cannot expect to approximate $I(\theta)$ uniformly simply by increasing N at some fixed points in Θ. Despite this, we will see in our experiments in Section 5 that increasing N will be essential to improving performance. For example, although we are not aware of any formal results, we can expect a similar bound for LSMC and KMS but with $N^{-\frac{1}{2}}$ instead of $N^{-\frac{s_\mathcal{X}}{d}+\varepsilon}$ (due to the MC integration rate); this explains why our method will outperform these approaches when $s_\mathcal{X}$ is large relative to d. Since the cost of CBQ is $\mathcal{O}(TN^3 + T^3)$, we note that we can take $N = \mathcal{O}(T^{\frac{2}{3}})$ without increasing the overall cost.

The rate in T will depend on the smoothness of $I(\theta)$ through the smoothness parameter $r \in [1/2, 1]$, which in turns depends on the smoothenss of the integrand f and the density p_θ in θ. The smoother these are, the faster the convergence rate will be. Although we are not aware of such a result, we can expect the same rate in T to hold for KMS since it is based on kernel ridge regression. On the other hand, LSMC will be inherently limited due to the use of linear or polynomial regression, and we expect it may not be possible to show consistency when $I(\theta)$ is not a polynomial in θ.

We now briefly discuss our assumptions. Firstly, the assumptions on $\mathcal{X}$ and Θ are used for simplicity of the presentation, and could be straightforwardly generalised to any bounded domain with Lipschitz boundary satisfying an interior cone condition; see [40, 78]. Secondly, the smoothness assumptions are used to guarantee that the problem is regular enough for approximation with a GP in both the first and second stage of the problem; note that we could weaken these assumptions following [40]. For simplicity, we also assume that the kernel parameters are known, but this could be extended to estimation in bounded sets; see [73]. Thirdly, our assumptions on the point sets ensure the points $\theta_{1:T}$ and $x^t_{1:N}$ cover $\mathcal{X}$ and Θ sufficiently fast in probability as N and T grow. Finally, the assumption on the regulariser $\lambda_X = 0$ may be relaxed to provide greater stability of the interpolation at the cost of slowing down convergence (we give the more general result for $\lambda_X > 0$ in Appendix A); in contrast, growing $\lambda_\Theta > 0$ in T is both natural and necessary: since we work in a bounded domain, we are in

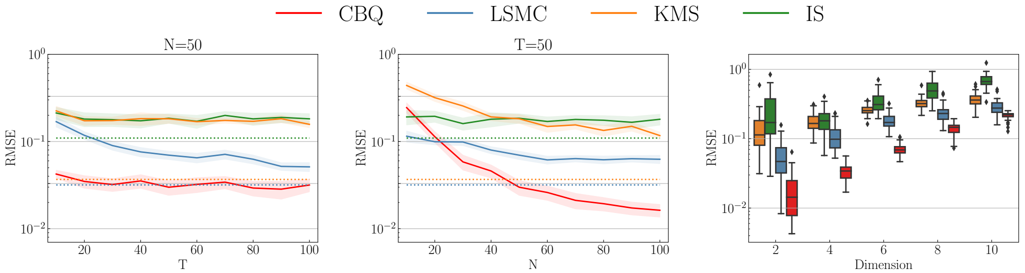


Figure 2: *Bayesian sensitivity analysis.* **Left:** RMSE of all methods when $d = 2$ and $N = 50$. **Middle:** RMSE of all methods when $d = 2$ and $T = 50$. **Right:** RMSE of all methods when $N = T = 100$.

an in-fill asymptotic regime and so should expect the conditioning of the Gram matrix to become worse as $T \to \infty$. As previously discussed, we will take N large but finite, and drive $T \to \infty$. For this reason, we only need λ_Θ to grow whilst $\lambda_\mathcal{X}$ may remain constant or be taken to be zero.

5 Experiments

We will now evaluate the empirical performance of CBQ against baselines including IS, KMS and LSMC. For the first three experiments, we focus on the case where f does not depend on θ (i.e. $f(x, \theta) = f(x)$), and for the fourth experiment we focus on the case where f depends on both x and θ. All methods are given access to the same samples and parameter values to ensure a fair comparison. We therefore use $\mathbb{P}_{\theta_1}, \ldots, \mathbb{P}_{\theta_T}$ as our importance distributions in IS. More detailed descriptions of the experimental settings and hyperparameter selection can be found in Appendix B and Appendix C. The code to reproduce our experiments is available at `https://github.com/Anonymous65536/cbq`.

Synthetic Experiment: Bayesian Sensitivity Analysis for Linear Models. Bayesian inference requires the derivation of a posterior distribution from a prior and a likelihood. Determining the sensitivity of this posterior to the hyperparameters is a critical step in assessing the robustness of the outcomes of Bayesian models [58, 37]. One approach to do so is to define some quantity of interest, and compute its expected value under the posterior with different hyperparameters.

We will demonstrate this approach using Bayesian linear regression. Our observations are $\mathcal{D} = \{Y \in \mathbb{R}^{m \times d}, z \in \mathbb{R}^m\}$ where m is the number of observations and d is the dimension including the intercept. We use a $\mathcal{N}(w; 0, \theta \mathrm{Id}_d)$ prior on the regression weights $w \in \mathbb{R}^d$, where the covariance matrix is a diagonal matrix with values $\theta \in \mathbb{R}^p$ (and here $p = d$). Using a Gaussian likelihood, we can obtain (via conjugacy) a multivariate Gaussian posterior $p_\theta(w|\mathcal{D})$ whose mean and variance have a closed form expression. We can then analyse the sensitivity of the posterior to θ by computing $I(\theta) = \int_{\mathbb{R}^d} f(w) p_\theta(w|\mathcal{D}) dw$ where f represents the quantity of interest. For example, if $f(w) = w^\top w$, then $I(\theta)$ is the second moment of the posterior, whereas if $f(w) = w^\top y^*$ for some new observation y^*, then $I(\theta)$ is the predictive mean. In this simple setting, the value of $I(\theta)$ can be computed analytically, making it a good synthetic example to benchmark our approach.

The results for $f(w) = w^\top w$ are in Figure 2, whilst the results for $f(w) = w^\top y^*$ are in Appendix C.1. In all the plots, we measure performance in terms of root mean squared error (RMSE). We sample parameter values $\theta_{1:T}$ from a $\mathrm{Unif}(\Theta)$ where $\Theta = (1, 3)^d$, and for each such parameter θ_t, obtain N observations from $p_{\theta_t}(\cdot|\mathcal{D})$. For CBQ, $k_\mathcal{X}$ is selected to be Gaussian Radial Basis Function (RBF) so tha kernel mean embedding μ has a closed form, and k_Θ is selected to be Matérn-3/2. We provide an ablation study on different kernels in Appendix C.1.5.

Figure 2 shows the performance of CBQ against baselines with varying N, T and d. In the left and middle plots, we can see that CBQ clearly outperforms baselines. Specifically, its rate of convergence is initially much faster in N than in T, which confirms the intuition from Theorem 1. The dotted lines also give the performance of baselines under a very large number of samples $N = T = 1000$. The performance of CBQ is either comparable or better than these under a much smaller sample size. In the right-most panel, we report results when d increases. The baselines gradually catch up with CBQ as d grows, which is again expected given the dependence of Theorem 1 on p (implicitly through r) and d.

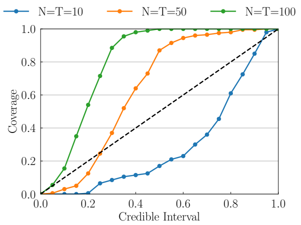


Figure 3: *Calibration for Bayesian sensitivity analysis of linear model in* $d = 2$.

Figure 3 shows the calibration of the CBQ posterior for this same integrand when $d = 2$. The coverage is the percentage of times a credible interval contains $I(\theta)$ under repetitions of the experiment. The black diagonal line represents the ideal case, with any curve lying above the black line indicating underconfidence and any curve lying below indicating overconfidence. It is generally more preferable to be underconfident than overconfident. We observe that when the number of samples is as small as 10, CBQ is overconfident, which can be explained by a poor performance of empirical Bayes in the small sample regime. On the other hand, when N and T increase, CBQ becomes underconfident, meaning that our posterior variance is more inflated than needed from a frequentist viewpoint. Calibration plots for the rest of the experiments can be found in Appendix C.

Finally, we highlight that since CBQ does not impose restrictions on how the samples are selected, it is straightforward to combine it with other point sets such as in quasi-Monte Carlo [34]. In Appendix C.1.4, we show the performance of CBQ and other baselines when samples are generated through this approach. Additionally, we previously mentioned that BQ conditions a GP on all NT samples and is more computationally expensive than CBQ. This claim is demonstrated through an empirical comparison of BQ and CBQ in Appendix C.1.3.

Butterfly Call Option with the Black-Scholes Model. Insurance companies are typically interested in computing the expected loss of their portfolios if a shock were to occur in the economy, and computing this quantity often requires the computation of conditional expectations.

In this example, we consider a butterfly call option whose price at time $\tau \in [0, \infty)$ is denoted by $S(\tau)$ and follows the Black-Scholes formula $S(\tau) = S_0 \exp(\sigma W(\tau) - \sigma^2 \tau/2)$ with $\sigma > 0$ being the underlying volatility, $W(\tau)$ being the standard Brownian motion and $S_0 = S(0)$ being the initial price. The financial derivative we are interested in is a butterfly call option whose payoff at time τ can be expressed as $\psi(S(\tau)) = \max(S(\tau) - K_1, 0) + \max(S(\tau) - K_2, 0) - 2\max(S(\tau) - (K_1 + K_2)/2, 0)$. We follow the set-up in [4, 5] assuming that a shock occurs at time η, at which time the price is $S(\eta) = \theta$, and this shock multiplies the option price by $1 + s$. As a result, the expected loss of the option can be expressed as $\mathcal{L} = \mathbb{E}[\max(I(\theta), 0)]$, where $I(\theta) = \int_0^\infty f(x) p_\theta(x) dx$, $x = S(\zeta)$ is the price at the time ζ at which the option matures, $f(x) = \psi(x) - \psi((1 + s)x)$, and p_θ is the density of a log-normal distribution induced from the Black-Scholes model (see Appendix C.2 for full details).

For this class of financial option, the ground truth $\mathcal{L}$ has a closed form so we can easily compare the performance of our methods against baselines. Note that $\mathcal{L}$ is a double expectation and only the inner expectation $I(\theta)$ is implemented with CBQ, LSMC, IS, KMS, and standard MC is used for the outer expectation. We take the initial price to be $S_0 = 100$, the volatility to be $\sigma = 0.3$, the strikes to be $K_1 = 50$ and $K_2 = 150$, the option maturity to be $\zeta = 2$, and a shock at $\eta = 1$ with strength $s = 0.2$. The observations $\theta_{1:T}$ are obtained by simulating according to the Black-Scholes formula starting with price S_0 at time 0, whilst $x^t_{1:N}$ are obtained by simulating according to the Black-Scholes formula conditioned on a price θ_t at time η. For CBQ, $k_\mathcal{X}$ is chosen as a logarithmic RBF kernel to ensure a closed-form kernel mean embedding (see Appendix C.2). A Stein kernel, employing a Matérn-3/2 as the base, is also utilized as $k_\mathcal{X}$. k_Θ is selected to be Matérn-3/2.

Results are presented in the left-most panel of Figure 4. Similar to the previous example, CBQ exhibits performance that is comparable, if not superior, to baseline methodologies even if baselines have a substantial sample size of $N = T = 1000$ (as plotted through dotted lines). We can also see that CBQ with log-RBF kernel and CBQ with Stein kernel have similar performance. Further experimental outcomes and calibration results are available in Appendix C.2.

Bayesian Sensitivity Analysis for Susceptible-Infectious-Recovered (SIR) Model. The SIR model is commonly used to simulate the dynamics of infectious diseases through a population [43]. In this model, the dynamics are governed by a system of ordinary differential equations (ODE) parametrised by a positive infection rate x and a recovery rate γ (see Appendix C.3). The accuracy of the numerical solution to this system typically hinges on the selection of a step size parameter. While smaller step sizes yield more accurate solutions, they are also associated with a much higher computational cost. For example, using a step size of 0.1 days for simulating a 150-day period would require a computation time of 7 seconds for generating a single sample. The cost would become even larger as the step size gets smaller, as depicted in the right-most panel of Figure 4. Consequently, there is a clear necessity for more data-efficient algorithms such as CBQ in such circumstances.

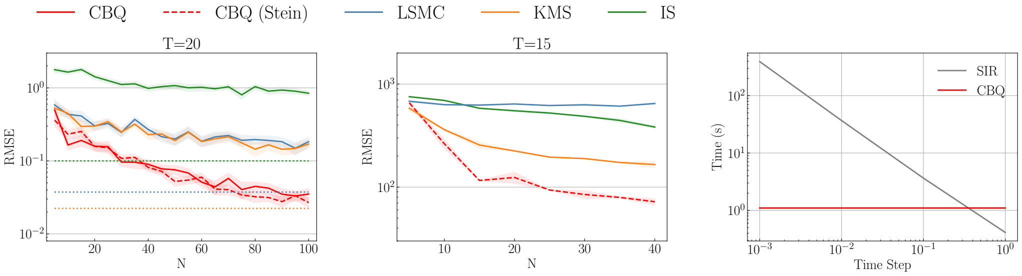


Figure 4: *Black-Scholes and SIR model.* **Left:** RMSE of all methods for the Black-Scholes example with $T = 20$. **Middle:** RMSE of all methods for the SIR example with $T = 15$. **Right:** The computational cost (in wall clock time) for CBQ and for obtaining one single numerical solution from SIR. In practice, the process of obtaining samples from SIR equations is repeated NT times.

We once again perform a Bayesian sensitivity analysis and place a $\mathrm{Gamma}(X; \theta, \xi)$ prior on the infection rate x, where θ represent the initial belief of the infection rate deduced from the study of the virus in the laboratory at the beginning of the outbreak, and ξ represents the amount of uncertainty. We are interested in the expected peak number of infected individuals: $I_{\max}(\theta) = \int_{\mathcal{X}} \max_r N_I^r(x) p_\theta(x) dx$, where p_θ is the density of the Gamma prior and N_I^r is the solution to the SIR ODEs and represents the number of infections at day r. It is important to study the sensitivity of initial estimate of the infection rate θ to $I_{\max}(\theta)$. In this experiment, we fix the rate parameter $\xi = 10$, and alter the shape parameter θ. The total population is set to be 10^6. The observations $\theta_{1:T}$ are sampled from the uniform distribution $\mathrm{Unif}\ (2, 9)$ and then N observations of $x^t_{1:N}$ are sampled from $\mathrm{Gamma}(X; \theta_t, \xi)$. We use a MC estimator with 5000 samples as the pseudo ground truth and evaluate the RMSE across all methods. For CBQ, we employ a Stein kernel for $k_{\mathcal{X}}$, with the Matérn-3/2 as the base kernel and k_Θ is selected to be Matérn-3/2.

We can see in the middle panel of Figure 4 that CBQ clearly outperforms baselines. Although the CBQ estimator exhibits a higher computational cost compared to baselines, we demonstrate in the right-most panel of Figure 4 that, due to the increased computational expense of obtaining samples with smaller step size, using CBQ is ultimately more efficient overall within the same period of time. Additional experimental results can be found in Appendix C.3.

Uncertain Decision Making in Health Economics. In the medical world, it is important to compare the financial cost and benefits of conducting additional experiments on patients. The expected value of partial perfect information (EVPPI) quantifies the expected gain from conducting experiments to obtain precise knowledge of some unknown variables [13]. EVPPI can be expressed as $\mathbb{E}[\max_c I_c(\theta)] - \max_c \mathbb{E}[f_c(X, \theta)]$ where f_c represents some measure of patient outcome (such as quality-adjusted life-years) under treatment c in a set of potential treatments $\mathcal{C}$, θ denotes the additional variables we could measure, and $I_c(\theta) = \int_{\mathcal{X}} f_c(x, \theta) p(x|\theta) dx$ denotes the expected patient outcome given our measurement of θ.

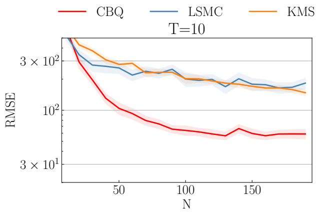


Figure 5: *Uncertain decision making.* We study RMSE for different estimators of EVPPI.

We adopt the same experimental setup as delineated in [26], wherein X and θ have a joint 19-dimensional Gaussian distribution (see Appendix C.4 for the mean and covariance), meaning $p(x|\theta)$ is a Gaussian density. Our target of interest is EVPPI under a binary decision-making problem ($\mathcal{C} = \{1, 2\}$) with $f_1(x, \theta) = 10^4(\theta_1 x_5 x_6 + x_7 x_8 x_9) - (x_1 + x_2 x_3 x_4)$ and $f_2(x, \theta) = 10^4(\theta_2 x_{13} x_{14} + x_{15} x_{16} x_{17}) - (x_{10} + x_{11} x_{12} x_4)$. We estimate $I_c(\theta)$ with CBQ and baselines, and use MC for the outer expectation. Note that IS is no longer applicable here because f depends on both x and θ, so we only comparing CBQ against KMS and LSMC. We draw 10^7 samples from the joint distribution to generate a pseudo ground truth, and evaluate the RMSE across different methods. For CBQ, we select Matérn-3/2 for $k_{\mathcal{X}}$ and also Matérn-3/2 for k_Θ. In Figure 5, we can see that CBQ outperforms baselines with much smaller RMSE. Additional experimental results can be found in Appendix C.4.

6 Conclusion

We proposed CBQ, a novel algorithm which is tailored for the computation of conditional expectations where obtaining samples or function evaluations is costly. We showed that CBQ exhibits an accelerated convergence rate, and provides the additional benefit of Bayesian quantification of uncertainty. Looking forward, we believe further gains in accuracy could be obtained by developing active learning schemes to N, T, and the location of $\theta_{1:T}$ and $x^t_{1:N}$ for all t in an adaptive manner. Additionally, CBQ could be extended for nested expectation problems by using a second level of BQ based on the second stage heteroscedastic GP, potentially leading to a further increase in accuracy.

References

[1] Rachid Ababou, Amvrossios C Bagtzoglou, and Eric F Wood. On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. *Mathematical Geology*, 26:99–133, 1994.

[2] Luigi Acerbi. Variational Bayesian Monte Carlo. *Advances in Neural Information Processing Systems*, 31, 2018.

[3] Robert A Adams and John JF Fournier. *Sobolev spaces*. Elsevier, 2003.

[4] Aurélien Alfonsi, Adel Cherchali, and Jose Arturo Infante Acevedo. Multilevel Monte Carlo for computing the SCR with the standard formula and other stress tests. *Insurance: Mathematics and Economics*, 100:234–260, 2021.

[5] Aurélien Alfonsi, Bernard Lapeyre, and Jérôme Lelong. How many inner simulations to compute conditional expectations with least-square Monte Carlo? *arXiv preprint arXiv:2209.04153*, 2022.

[6] Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, et al. Stein's method meets computational statistics: a review of some recent developments. *Statistical Science*, 38(1):120–139, 2023.

[7] Ioannis Andrianakis and Peter G Challenor. The effect of the nugget on Gaussian process emulators of computer models. *Computational Statistics & Data Analysis*, 56(12):4215–4228, 2012.

[8] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[9] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[10] Ayush Bharti, Masha Naslidnyk, Oscar Key, Samuel Kaski, and François-Xavier Briol. Optimally-weighted estimators of the maximum mean discrepancy for likelihood-free inference. *arXiv preprint arXiv:2301.11674*, 2023.

[11] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[12] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[13] Alan Brennan, Samer Kharroubi, Anthony O'Hagan, and Jim Chilcott. Calculating partial expected value of perfect information via Monte Carlo sampling algorithms. *Medical Decision Making*, 27(4):448–470, 2007.

[14] François-Xavier Briol, Chris J Oates, Mark Girolami, Michael A Osborne, and Dino Sejdinovic. Probabilistic integration. *Statistical Science*, 34(1):1–22, 2019.

[15] George Casella. An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.

[16] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[17] Siu Lun Chau, Shahine Bouabid, and Dino Sejdinovic. Deconditional downscaling with Gaussian processes. *Advances in Neural Information Processing Systems*, 34:17813–17825, 2021.

[18] Jon Cockayne, Chris J Oates, Timothy John Sullivan, and Mark Girolami. Bayesian probabilistic numerical methods. *SIAM review*, 61(4):756–789, 2019.

[19] Julien Demange-Chryst, François Bachoc, and Jérôme Morio. Efficient estimation of multiple expectations with the same sample by adaptive importance sampling and control variates. *arXiv preprint arXiv:2212.00568*, 2022.

[20] Ronald A DeVore and Robert C Sharpley. Besov spaces on domains in $\mathbb{R}^d$. *Transactions of the American Mathematical Society*, 335(2):843–864, 1993.

[21] Persi Diaconis. Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV*, 1:163–175, 1988.

[22] David Eric Edmunds and Hans Triebel. *Function spaces, entropy numbers, differential operators*, volume 120. Cambridge Univ Press, 1996.

[23] David Heaver Fremlin. *Measure theory*, volume 4. 2000.

[24] Mathieu Gerber and Nicolas Chopin. Sequential quasi Monte Carlo. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(3):509–579, 2015.

[25] Alexandra Gessner, Javier Gonzalez, and Maren Mahsereci. Active multi-information source Bayesian quadrature. In *Uncertainty in Artificial Intelligence*, pages 712–721, 2020.

[26] Michael B Giles and Takashi Goda. Decision-making under uncertainty: using MLMC for efficient estimation of EVPPI. *Statistics and Computing*, 29:739–751, 2019.

[27] Michael B Giles, Tigran Nagapetyan, and Klaus Ritter. Multilevel Monte Carlo approximation of distribution functions and densities. *SIAM/ASA journal on Uncertainty Quantification*, 3(1):267–295, 2015.

[28] Peter Glynn and Donald Igelhart. Importance sampling for stochastic simulations. *Management Science*, 35(1367-1392), 1989.

[29] Davit Gogolashvili, Matteo Zecchin, Motonobu Kanagawa, Marios Kountouris, and Maurizio Filippone. When is importance weighting correction needed for covariate shift adaptation? *arXiv preprint arXiv:2303.04020*, 2023.

[30] Tom Gunter, Michael A Osborne, Roman Garnett, Philipp Hennig, and Stephen J Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. *Advances in Neural Information Processing Systems*, 27, 2014.

[31] Anna Heath, Ioanna Manolopoulou, and Gianluca Baio. A review of methods for analysis of the expected value of information. *Medical Decision Making*, 37(7):747–758, 2017.

[32] Philipp Hennig, Michael A Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142, 2015.

[33] Philipp Hennig, Michael A Osborne, and Hans P Kersting. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, 2022.

[34] Fred Hickernell. A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, 67(221):299–322, 1998.

[35] L Jeff Hong and Sandeep Juneja. Estimating the mean of a non-linear function of conditional expectation. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pages 1223–1236. IEEE, 2009.

[36] Rathinavel Jagadeeswaran and Fred J Hickernell. Fast automatic Bayesian cubature using lattice sampling. *Statistics and Computing*, 29(6):1215–1229, 2019.

[37] Noa Kallioinen, Topi Paananen, Paul-Christian Bürkner, and Aki Vehtari. Detecting and diagnosing prior and likelihood sensitivity with power-scaling. *arXiv preprint arXiv:2107.14054*, 2021.

[38] Motonobu Kanagawa and Philipp Hennig. Convergence guarantees for adaptive Bayesian quadrature methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[39] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

[40] Motonobu Kanagawa, Bharath K Sriperumbudur, and Kenji Fukumizu. Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *Foundations of Computational Mathematics*, 20:155–194, 2020.

[41] Toni Karvonen and Simo Sarkka. Fully symmetric kernel quadrature. *SIAM Journal on Scientific Computing*, 40(2):A697–A720, 2018.

[42] Toni Karvonen, Simo Särkkä, and Chris Oates. Symmetry exploits for Bayesian cubature methods. *arXiv preprint arXiv:1809.10227*, 2018.

[43] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, 1927.

[44] Hans Kersting and Philipp Hennig. Active uncertainty calibration in Bayesian ODE solvers. *arXiv preprint arXiv:1605.03364*, 2016.

[45] Sebastian Krumscheid and Fabio Nobile. Multilevel Monte Carlo approximation of functions. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1256–1293, 2018.

[46] Quoc V Le, Alex J Smola, and Stéphane Canu. Heteroscedastic gaussian process regression. In *International Conference on Machine Learning*, pages 489–496, 2005.

[47] Christiane Lemieux. Randomized quasi-Monte Carlo: A tool for improving the efficiency of simulations in finance. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, volume 2, pages 1565–1573. IEEE, 2004.

[48] Kaiyu Li, Daniel Giles, Toni Karvonen, Serge Guillas, and François-Xavier Briol. Multilevel Bayesian quadrature. *arXiv preprint arXiv:2210.08329*, 2022.

[49] Francis A Longstaff and Eduardo S Schwartz. Valuing american options by simulation: a simple least-squares approach. *The Review of Financial Studies*, 14(1):113–147, 2001.

[50] Hedibert F. Lopes and Justin L. Tobias. Confronting prior convictions: On issues of prior sensitivity and likelihood robustness in Bayesian analysis. *Annual Review of Economics*, 3:107–131, 2011.

[51] Neal Madras and Mauro Piccioni. Importance sampling for families of distributions. *The Annals of Applied Probability*, 9(4):1202–1225, 1999.

[52] Ricardo Marques, Christian Bouville, Mickaël Ribardière, Luís Paulo Santos, and Kadi Bouatouch. A spherical Gaussian framework for Bayesian Monte Carlo rendering of glossy surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 19(10):1619–1632, 2013.

[53] Deyu Ming and Serge Guillas. Linked Gaussian Process Emulation for Systems of Computer Models using Matérn Kernels and Adaptive Design, 2021.

[54] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[55] Krikamol Muandet, Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, and Bernhard Schölkopf. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 17, 2016.

[56] Yu Nishiyama and Kenji Fukumizu. Characteristic kernels and infinitely divisible distributions. *Journal of Machine Learning Research*, 17(180):1–28, 2016.

[57] Ziang Niu, Johanna Meier, and François-Xavier Briol. Discrepancy-based inference for intractable generative models using quasi-Monte Carlo. *Electronic Journal of Statistics*, 17(1):1411–1456, 2023.

[58] Jeremy E Oakley and Anthony O'Hagan. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society: Series B*, 66(3):751–769, 2004.

[59] Chris Oates, Steven Niederer, Angela Lee, François-Xavier Briol, and Mark Girolami. Probabilistic models for integration error in the assessment of functional cardiac models. *Advances in Neural Information Processing Systems*, 30, 2017.

[60] Chris J Oates, Jon Cockayne, François-Xavier Briol, and Mark Girolami. Convergence rates for a class of estimators based on Stein's method. *Bernoulli*, 2019.

[61] Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society. Series B*, pages 695–718, 2017.

[62] Chris J Oates and Timothy John Sullivan. A modern retrospective on probabilistic numerics. *Statistics and Computing*, 29(6):1335–1351, 2019.

[63] Anthony O'Hagan. Bayes–hermite quadrature. *Journal of Statistical Planning and Inference*, 29:245–260, 1991.

[64] Tom Rainforth, Rob Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting Monte Carlo estimators. In *International Conference on Machine Learning*, pages 4267–4276, 2018.

[65] Carl E. Rasmussen and Chris Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006.

[66] Carl Edward Rasmussen and Zoubin Ghahramani. Bayesian Monte Carlo. *Advances in Neural Information Processing Systems*, pages 505–512, 2003.

[67] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[68] Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1-3):271–280, 2001.

[69] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12(7), 2011.

[70] Zhuo Sun, Alessandro Barp, and François-Xavier Briol. Vector-valued control variates. *arXiv preprint arXiv:2109.08944*, 2021.

[71] Zhuo Sun, Chris J Oates, and François-Xavier Briol. Meta-learning control variates: Variance reduction with limited data. *arXiv preprint arXiv:2303.04756*, 2023.

[72] Xiaojin Tang. *Importance sampling for efficient parametric simulation*. PhD thesis, Boston University, 2013.

[73] Aretha L Teckentrup. Convergence of Gaussian process regression with estimated hyperparameters and applications in Bayesian inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4):1310–1337, 2020.

[74] Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

[75] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[76] Holger Wendland and Christian Rieger. Approximate interpolation with applications to selecting smoothing parameters. *Numerische Mathematik*, 101(4):729–748, 2005.

[77] Jonathan Wenger, Nicholas Krämer, Marvin Pförtner, Jonathan Schmidt, Nathanael Bosch, Nina Effenberger, Johannes Zenn, Alexandra Gessner, Toni Karvonen, François-Xavier Briol, et al. ProbNum: Probabilistic Numerics in Python. *arXiv preprint arXiv:2112.02100*, 2021.

[78] George Wynne, François-Xavier Briol, and Mark Girolami. Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness. *The Journal of Machine Learning Research*, 22(1):5468–5507, 2021.

[79] Xiaoyue Xi, François-Xavier Briol, and Mark Girolami. Bayesian quadrature for multiple related integrals. In *International Conference on Machine Learning*, pages 5373–5382, 2018.

Supplementary Material

Table of Contents

Appendix A Convergence Analysis for the CBQ Estimator

To validate our methodology, we establish a rate at which the CBQ estimator $\hat{I}_{\mathrm{CBQ}}$ converges to the true I in the $\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{tr}})$ norm, $\|\hat{I}_{\mathrm{CBQ}} - I\|_{\mathcal{L}^2(\Theta,\mathbb{P}_{\mathrm{tr}})} = \int_\Theta (\hat{I}_{\mathrm{CBQ}}(\theta) - I(\theta))^2 \mathbb{P}_{\mathrm{tr}}(\mathrm{d}\theta)$, for $\mathbb{P}_{\mathrm{tr}}$ such that $\theta_t \sim \mathbb{P}_{\mathrm{tr}}$ for $t \in \{1, \dots, T\}$. This result is presented in the main text in Theorem 1. In this section, we prove the more general version of Theorem 1 presented in the main text and intermediate results required, and expand on the technical background.

The proof primarily builds on two results: [78, Theorem 4] is used to get a bound in Stage 1, and [29, Theorem 4] is used to get a bound in Stage 2. Specifically, [29, Theorem 4] is used to establish a bound on $\|\hat{I}_{\mathrm{CBQ}} - I\|_{\mathcal{L}^2(\Theta,\mathbb{P}_{\mathrm{tr}})}$ in terms of T (the number of samples in Θ), and the largest value for BQ variance, $\max_{t\in\{1,\dots,T\}} \sigma^2_{\mathrm{BQ}}(\theta_t)$. Then, [78, Theorem 4] is used to bound the variance $\sigma^2_{\mathrm{BQ}}(\theta_t)$ for any $t \in \{1, \dots, T\}$ in terms of N (the number of samples in $\mathcal{X}$).

In Appendix A.1, we define the weight function $w(\theta)$ that establishes a connection between Stage 2 of the method to the setting of importance-weighted kernel ridge regression in [29, Theorem 4]. Then in Appendix A.2 we present technical assumptions under which both [78, Theorem 4] and [29, Theorem 4] for the defined $w(\theta)$ hold. Finally, in Appendix A.3 we prove a more general form of Theorem 1 for $\lambda_{\mathcal{X}} \geq 0$ and $\theta_{1:T}$ sampled from distribution other than uniform.

A.1 Connection to Importance-Weighted Kernel Ridge Regression

Recall the CBQ estimator proposed in Section 3,

$$\hat{I}_{\mathrm{CBQ}}(\theta) = k_\Theta(\theta, \theta_{1:T})^\top \left(k_\Theta(\theta_{1:T}, \theta_{1:T}) + (\lambda_\Theta + \sigma^2_{\mathrm{BQ}}(\theta_{1:T}))\mathrm{Id}_T\right)^{-1} \hat{I}_{\mathrm{BQ}}(\theta_{1:T}),$$

where $\lambda_\Theta \geq 0$ is the regularisation parameter, and $\hat{I}_{\mathrm{BQ}}(\theta_t)$ and $\sigma^2_{\mathrm{BQ}}(\theta_t)$, for $t \in \{1, \dots, T\}$, are BQ posterior mean and variance obtained in the first stage,

$$\begin{aligned}\hat{I}_{\mathrm{BQ}}(\theta_t) &= \mu_\theta(x^t_{1:N})^\top \left(k_{\mathcal{X}}(x^t_{1:N}, x^t_{1:N}) + \lambda_{\mathcal{X}}\mathrm{Id}_N\right)^{-1} f(x^t_{1:N}, \theta_t),\\ \sigma^2_{\mathrm{BQ}}(\theta_t) &= \mathbb{E}_{X,X'\sim\mathbb{P}_\theta}[k_{\mathcal{X}}(X, X')] - \mu_\theta(x^t_{1:N})^\top \left(k_{\mathcal{X}}(x^t_{1:N}, x^t_{1:N}) + \lambda_{\mathcal{X}}\mathrm{Id}_N\right)^{-1} \mu_\theta(x^t_{1:N}).\end{aligned}$$

It was pointed out in [29, Remark 2], (and can be seen through straightforward differentiation) that the estimator $\hat{I}_{\mathrm{CBQ}}$ is the minimiser of the importance weighted kernel ridge regression loss over functions in the RKHS $\mathcal{H}_\Theta$ induced by the kernel k_Θ,

$$\hat{I}_{\mathrm{CBQ}} = \underset{F\in\mathcal{H}_\Theta}{\arg\min} \Big\{ \sum_{t=1}^T \frac{\tau}{1 + \lambda_\Theta^{-1}\sigma^2_{\mathrm{BQ}}(\theta_t)} \left(F(\theta_t) - \hat{I}_{\mathrm{BQ}}(\theta_t)\right)^2 + \tau\lambda_\Theta^{-1}\|F\|^2_{\mathcal{H}_\Theta} \Big\},$$

for any $\tau > 0$.[2] Suppose θ_i are sampled from a probability measure $\mathbb{P}_{\mathrm{tr}}$ on Θ. Then,

$$\mathbb{P}_{\mathrm{te}}(A) = \int_A w(\theta)\mathbb{P}_{\mathrm{tr}}(\mathrm{d}\theta) \tag{A.1}$$

defines a positive measure $\mathbb{P}_{\mathrm{te}}$ on Θ for any positive $w(\theta) > 0$ for which the integral exists [23, Proposition 232D]; further, if $w(\theta)$ is bounded, the measure is finite. Suppose we construct a $w(\theta)$ that satisfies these requirements, and is such that $w(\theta_t) = \tau(1 + \lambda_\Theta^{-1}\sigma^2_{\mathrm{BQ}}(\theta_t))^{-1}$. Then, since $\mathbb{E}[\hat{I}_{\mathrm{BQ}}(\theta_i)] = I(\theta_i)$, the importance-weighted loss can be considered an unbiased finite-sample approximation of

$$\int_\Theta (F(\theta) - I(\theta))^2 \mathbb{P}_{\mathrm{te}}(\mathrm{d}\theta) + \frac{1}{n}\|F\|^2_{\mathcal{H}_\Theta}.$$

Under a further assumption that the problem is well-specified, meaning $I(\theta) \in \mathcal{H}_\Theta$, an upper bound on $\|\hat{I}_{\mathrm{CBQ}} - I\|_{\mathcal{L}^2(\Theta,\mathbb{P}_{\mathrm{te}})}$ in terms of T and $\sup_{\theta\in\Theta} w(\theta)$ was established in [29, Theorem 4]. To apply the result, we define $w(\theta)$ of convenient form that satisfies the requirements mentioned above,

[2] We will keep τ as a free parameter for now, and use it in Appendix A.3 to ensure constraints on λ_Θ imposed by [29, Theorem 4] are satisfied

specifically $w(\theta) \in (0, W_0]$ for some $W_0 < \infty$ and any $\theta \in \Theta$, and $w(\theta_t) = \tau(1 + \lambda_\Theta^{-1}\sigma^2_{\mathrm{BQ}}(\theta_t))^{-1}$ for some $t \in \{1, \ldots, T\}$.[3] Take $\sigma^2_{\mathrm{BQ}}(\theta_{t'}) = \max_{t\in\{1,\ldots,T\}}\{\sigma^2_{\mathrm{BQ}}(\theta_t)\} > 0$, and define

$$w(\theta) = \begin{cases} \tau(1 + \lambda_\Theta^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1} & \text{if } \|\theta - \theta_t\|_\Theta \geq \varepsilon' \text{ for all } t \in \{1, \ldots, T\} \\ \tau A_t - \tau B_t \frac{\|\theta - \theta_t\|_\Theta}{\varepsilon'}, & \text{for } t \text{ such that } \|\theta - \theta_t\|_\Theta < \varepsilon' \end{cases} \tag{A.2}$$

for $\|\cdot\|_\Theta$ the Euclidean norm on Θ, some fixed $0 < \varepsilon' \leq \min_{i,j\in\{1,\ldots,T\},\, i\neq j} \|\theta_i - \theta_j\|_\Theta$, and

$$A_t = (1 + \lambda_\Theta^{-1}\sigma^2_{\mathrm{BQ}}(\theta_t))^{-1} \qquad \text{and} \qquad B_t = (1 + \lambda_\Theta^{-1}\sigma^2_{\mathrm{BQ}}(\theta_t))^{-1} - (1 + \lambda_\Theta^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1}.$$

For $\Theta \subset \mathbb{R}$, such $w(\theta)$ is easily visualised, as can be seen in Figure 6.

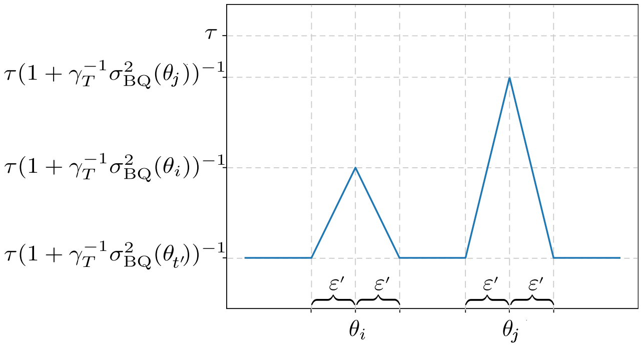


Figure 6: Illustration of $w(\theta)$ for $\Theta \subset \mathbb{R}$

It is easy to see that $w(\theta)$ is bounded above by $\tau \max_{t\in\{1,\ldots,T\}}(1 + \lambda_\Theta^{-1}\sigma^2_{\mathrm{BQ}}(\theta_t))^{-1} < \tau$, and below by $\tau(1 + \lambda_\Theta^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1} > 0$ for any $\theta \in \Theta$, and $w(\theta_t) = \tau(1 + \lambda_\Theta^{-1}\sigma^2_{\mathrm{BQ}}(\theta_t))^{-1}$ as required.

Note that the weight $w(\theta)$ constructed here is by no means a unique way to establish a useful connection between our setting of heteroscedastic GP regression, and importance-weighted kernel ridge regression. As will become evident in the proofs in Appendix A.3, one could use any $w(\theta)$ provided it satisfies $w(\theta_t) = \tau(1 + \lambda_\Theta^{-1}\sigma^2_{\mathrm{BQ}}(\theta_t))^{-1}$ for any t, is bounded below by some function of $\sigma^2_{\mathrm{BQ}}(\theta_{t'})$, and is bounded above by an expression that does not grow in T or N. Our proposed construction is simple and easy to visualise, and the parameter ε' has no impact on the speed of convergence as we shall see in the results in Appendix A.3.

A.2 Technical Assumptions

Prior to presenting our findings, we present and justify the assumptions we have made. Throughout we use Sobolev spaces to quantify function smoothness. A Sobolev space $\mathcal{W}^{2,s}(\mathcal{X}, \mu)$, with $s > d/2$ and a measure μ on $\mathcal{X} \subseteq \mathbb{R}^d$, consists of functions that satisfy certain conditions: they are square integrable under the measure μ, and all weak derivatives up to and including order s are also square integrable under μ. Weak derivatives are a generalization of ordinary derivatives, allowing for functions that are not necessarily differentiable everywhere.

Further, we assume the kernels $k_\Theta, k_\mathcal{X}$ to be Matérn. This choice is key: the RKHS of a Matérn kernel is norm-equivalent to a Sobolev spaces $\mathcal{W}^{2,s}$ for some s [76, Corollary 10.48], which allows us to directly compare smoothness of the kernel with smoothness of the true function (provided we know it lies in a Sobolev space $\mathcal{W}^{2,s'}$ for some s'). We refer to [3] for an in-depth treatment of Sobolev spaces and [9] for general RKHS theory.

The following is a more general form of the assumptions in Theorem 1: specifically, we allow for the case when $\theta_{1:T}$ came from a distribution other than uniform, and do not assume $\lambda_\mathcal{X} = 0$.

(a) $\mathcal{X} \subset \mathbb{R}^d$ is open, convex, and bounded.
(b) $\Theta \subset \mathbb{R}^p$ is open, convex, and bounded.

[3] The integrability requirement is specific to $\mathbb{P}_{\mathrm{tr}}$ and will be assumed at a later stage.

(c) θ_t were sampled i.i.d. from some $\mathbb{P}_{\text{tr}}$, and $\mathbb{P}_{\text{tr}}$ is equivalent to the uniform distribution on Θ, meaning $\mathbb{P}_{\text{tr}}(A) = 0$ for a set $A \subset \Theta$ if and only if $\text{Unif}(A) = 0$.

(d) $x^t_{1:N} \sim \mathbb{P}_{\theta_t}$ for all $t \in \{1, \cdots, T\}$.

(e) $\mathbb{P}_\theta$ has a density p_θ for any $\theta \in \Theta$, and the densities are such that $\inf_{\theta\in\Theta, x\in\mathcal{X}} p_\theta(x) = \eta > 0$ and $\sup_{\theta\in\Theta} \|p_\theta\|_{\mathcal{L}^2(\mathcal{X})} = \eta_0 < \infty$.

(f) $k_\mathcal{X}$ is a Matérn kernel of order $\nu_\mathcal{X}$ such that $s_\mathcal{X} \geq \nu_\mathcal{X} + d/2$.

(g) k_Θ is a Matérn kernel of order ν_Θ such that $s_\Theta \geq \nu_\Theta + p/2$.

(h) For any θ, $f(x, \theta)$ lies in the Sobolev space $\mathcal{W}^{2,s_\mathcal{X}}(\mathcal{X})$.

(i) $I(\theta)$ lies in the Sobolev space $\mathcal{W}^{2,s_\Theta}(\Theta)$.

(j) For the integral operator $L' : \mathcal{L}^2(\Theta, \mathbb{P}_{\text{tr}}) \to \mathcal{H}_\Theta$, there is a $g \in \mathcal{L}^2(\Theta, \mathbb{P}_{\text{tr}})$ such that $I = L^r g$ for some $r \in [1/2, 1]$. We denote $R_0 = \|g\|_{\mathcal{L}^2(\Theta,\mathbb{P}_{\text{tr}})}$.

(k) $\lambda_\Theta = cT^{1/(2r+1)}$, for $c > 0$ and $\alpha \in (0, 1)$.

(l) $\lambda_\mathcal{X} \geq 0$.

These correspond to assumptions in Theorem 1 as follows: (a) and (b) form 1, (c) and (d) form a more general form of 2, (e) is 3, (f) and (g) form 4, (h) and (i) form 5, and (j) is 6. Assumption (k) is mentioned in the text of Theorem 1 following the list of assumptions, and (l) is the more general form of the condition $\lambda_\mathcal{X} = 0$ in Theorem 1.

Crucially, in the proofs in the next section we will see that (a) to (l) imply that the setting of the model in Stage 1 satisfies the assumptions of [78, Theorem 4], and the setting of the model in Stage 2 satisfies the assumptions of [29, Theorem 4]—the two key results we will use to prove the convergence rate of the estimator.

A.3 Proof of Theorem 1

We are now ready to state the bound on $\|\hat{I}_{\text{CBQ}} - I\|_{\mathcal{L}^2(\Theta,\mathbb{P}_{\text{tr}})}$, which is essentially a corollary of [29, Theorem 4]. This bound depends on the largest BQ variance $\max_{t\in\{1,\ldots,T\}} \sigma^2_{\text{BQ}}(\theta_t)$; we obtain a bound on BQ variance $\sigma^2_{\text{BQ}}(\theta_t)$ for any t in Theorem 2. Combining the two results gives Theorem 3, which is the generalised version of Theorem 1.

Before proving our corollary, we point out that $\mathbb{P}_{\text{te}}$ as defined in Equation (A.1) is a finite positive measure and not necessarily a probability measure; meanwhile in the statement of [29, Theorem 4] $\mathbb{P}_{\text{te}}$ is asked to be a probability measure. This is not an issue since $\mathbb{P}_{\text{te}}$ being a probability measure is never used in the proof of [29, Theorem 4]—instead, the proof only asks that $\mathbb{P}_{\text{te}}$ be a finite positive measure. We will therefore make use of Theorem 4 for a finite positive measure $\mathbb{P}_{\text{te}}$.

Corollary 1. *Suppose Assumptions (b), (c), (g) and (i) to (k) hold. Then, for a fixed δ, there is a $T_0(\delta) > 0$ such that for all $T \geq T_0(\delta)$ with probability at least $1 - \delta/2$ it holds that*

$$\|\hat{I}_{\text{CBQ}} - I\|_{\mathcal{L}^2(\Theta,\mathbb{P}_{\text{tr}})} \leq K_0(\log(12/\delta) + K_3)(1 + c^{-1}T^{-1/(2r+1)}\sigma^2_{\text{BQ}}(\theta_{t'}))T^{-r/(2r+1)},$$

for $\sigma^2_{\text{BQ}}(\theta_{t'}) = \max_{t\in\{1,\ldots,T\}} \sigma^2_{\text{BQ}}(\theta_t)$, $K_0 = 16(M + \|I\|_{\mathcal{H}_\Theta})(1 + \tau^{-1/2}\|\Theta\|^{1/2})c^{-1/2}$, and $K_3 = R_0 c^{r+1/2}(M + \|I\|_{\mathcal{H}_\Theta})^{-1}(1 + \tau^{-1/2}\|\Theta\|^{1/2})^{-1}/16$.

Proof. First, we show the assumptions in the Theorem hold for $R \leq \tau R_0$, the r in Assumption (j), $q = 1$, $W = \tau$, and $\sigma^2 = \|\Theta\|\tau$.

Assumption 1 (Existence of the target function): As discussed in [29], Assumption 1 holds if the model is well-specified, $I(\theta) \in \mathcal{H}_\Theta$, and the kernel k_Θ is universal. The latter holds for Matérn kernels by [69]. The former is trivial as well: it is well-known that $\mathcal{H}_\Theta \simeq \mathcal{W}^{2,\nu_\Theta+p/2}(\Theta)$: the RKHS of a Matérn kernel of order ν_Θ over an open, convex and bounded $\Theta \subset \mathbb{R}^p$ is norm-equivalent to the Sobolev space $W^{2,\nu_\Theta+p/2}(\Theta)$ [76, Corollary 10.48][4]. By (i) $I(\theta) \in \mathcal{W}^{2,s_\Theta}(\Theta)$, and finally by the inclusion of Sobolev spaces $I(\theta) \in \mathcal{W}^{2,\nu_\Theta+p/2}(\Theta)$.

Assumption 2 (The smoothness of the target function): We assumed that the Assumption holds for $\mathbb{P}_{\text{tr}}$ in (j). By definition of $\mathbb{P}_{\text{te}}$, for any $\mathbb{P}_{\text{te}}$-integrable $g' : \Theta \to \mathbb{R}$ it holds that $\int_\Theta g'(\theta)\mathbb{P}_{\text{te}}\mathrm{d}\theta =$

[4] Strictly speaking, the result in [76, Corollary 10.48] is stated for integer-order spaces only, meaning $\nu_\Theta + p/2 \in \mathbb{Z}$. However, it can be straightforwardly extended to fractional order Sobolev-Slobodeckij spaces using [76, Corollary 10.13] that says the RKHS of a Matérn kernel on $\mathbb{R}^p$ is norm-equivalent to a Bessel potential space, which in turn is norm-equivalent to the Sobolev-Slobodeckij space by [3, Section 7.62], and finally using an extension operator in [20, Theorems 6.1 and 6.7] to restrict this to open, convex and bounded subsets of $\mathbb{R}^p$.

$\int_\Theta g'(\theta) w(\theta) \mathbb{P}_{\mathrm{tr}} \mathrm{d}\theta$. Since $w(\theta)$ is bounded above and below away from zero, $\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{tr}})$ is norm-equivalent to $\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{te}})$. Therefore, $g \in \mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{tr}})$ and the Assumption holds for $R \leq \tau R_0$ as $w(\theta) \leq \tau$ by construction.

Assumption 3 (Importance-weighting function): For $w(\theta)$ as defined in Equation (A.2) and $q = 1$, $W = \tau$, and $\sigma^2 = \|\Theta\|\tau$ it holds for all $m \in \mathbb{N}$, $m \geq 2$, that

$$\left(\int_\Theta w(\theta)^{\frac{q+m-1}{q}} \mathbb{P}_{\mathrm{tr}}(\mathrm{d}\theta) \right)^q \leq \frac{1}{2} m! W^{m-2} \sigma^2$$

since $w(\theta)$ is bounded from above, $\int_\Theta w(\theta)^m \mathbb{P}_{\mathrm{tr}}(\mathrm{d}\theta) < \|\Theta\| \tau^m \max_t (1 + \lambda_\Theta^{-1} \sigma^2_{\mathrm{BQ}}(\theta_t))^{-m} < \|\Theta\|\tau$.

Assumption 4 (Effective dimension): It is a standard result (see, for instance, [22, Section 3.3.4]) that for k_Θ being a Matérn kernel of order ν_Θ, the i-th eigenvalue decays at the rate of $i^{-\frac{2\nu_\Theta + p}{p}}$. As pointed out in the discussion after Assumption 4 in [29], this implies Assumption 4 holds for $s' = p/(2\nu_\Theta + p)$.

Therefore by [29, Theorem 4], for $\lambda = \tau c \lambda_\Theta T^{-1} = \tau c T^{-(1-1/(2r+1))}$ and the weight function $w(\theta)$ defined in (A.2), we have that with probability at least $1 - \delta/2$

$$\|\hat{I}_{\mathrm{CBQ}} - I\|_{\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{te}})} \leq T^{-r\beta} \left(16(M + \|I\|_{\mathcal{H}_\Theta})(W + \sigma E_{s'}^{1-q}) c^{-A/2} \log(12/\delta) + c^r R \right) \quad \text{(A.3)}$$

provided[5]

$$\lambda = \tau c T^{-(1-1/(2r+1))} \leq 1, \qquad \tau c \geq \left(64(W + \sigma^2) E_{s'}^{2(1-q)} \log^2(12/\delta) \right)^{1/(1+A)}, \quad \text{(A.4)}$$

for the constants $W = \tau$, $\sigma^2 = \|\Theta\|\tau$, $q = 1$, $r \in [1/2, 1]$, $R \leq \tau R_0$, and $A = 1$, $\beta = 1/(2r+1)$. Then, the conditions on c in (A.4) become

$$c \in [8\tau^{-1/2} \log(12/\delta)\, (1 + \|\Theta\|)^{1/2}, \tau^{-1} T^{1-1/(2r+1)}]. \quad \text{(A.5)}$$

We denote the smallest T for which this holds by T_0. The rate in (A.3) becomes

$$\|\hat{I}_{\mathrm{CBQ}} - I\|_{\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{te}})} \leq T^{-\frac{r}{2r+1}} \left(16(M + \|I\|_{\mathcal{H}_\Theta})(\tau + \tau^{1/2} \|\Theta\|^{1/2}) c^{-1/2} \log(6/\delta) + \tau c^r R_0 \right)$$

Since $\mathbb{P}_{\mathrm{te}}(\mathrm{d}\theta) = w(\theta) \mathbb{P}_{\mathrm{tr}}(\mathrm{d}\theta)$, and $w(\theta) \geq \tau(1 + \lambda_\Theta^{-1} \sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1} > 0$ for all θ, it holds that

$$\begin{aligned} \|\hat{I}_{\mathrm{CBQ}} - I\|_{\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{te}})} &\geq \min_{\theta \in \Theta} w(\theta) \|\hat{I}_{\mathrm{CBQ}} - I\|_{\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{tr}})} \\ &\geq \tau(1 + \lambda_\Theta^{-1} \sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1} \|\hat{I}_{\mathrm{CBQ}} - I\|_{\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{tr}})}, \end{aligned}$$

and therefore

$$\begin{aligned} \|\hat{I}_{\mathrm{CBQ}} - I\|_{\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{tr}})} &\leq (1 + \lambda_\Theta^{-1} \sigma^2_{\mathrm{BQ}}(\theta_{t'})) T^{-r/(2r+1)} \\ &\quad \times \left(16(M + \|I\|_{\mathcal{H}_\Theta})(1 + \tau^{-1/2} \|\Theta\|^{1/2}) c^{-1/2} \log(6/\delta) + c^r R_0 \right), \end{aligned}$$

and we arrive at the statement of the theorem. □

The need to introduce τ in Appendix A.1 is clear now: without it, the condition on c in Equation (A.5) may not hold. Since $\tau > 0$ can be selected at will, we may set it to the smallest value for which Equation (A.5) holds.

Next, we establish a bound on $\sigma^2_{\mathrm{BQ}}(\theta_t)$ for any $t \in \{1, \dots, T\}$.

Theorem 2. *Suppose Assumptions (a), (d) to (f), (h) and (l) hold. Then there is a $N_0 > 0$ such that for all $N \geq N_0$ with probability at least $1 - \delta/2$ it holds that*

$$\sigma^2_{\mathrm{BQ}}(\theta_t) \leq \lambda_{\mathcal{X}} + \frac{2}{\delta} \eta_0 K_1 K_2^{d/2} N^{-1/2+\varepsilon} \left(K_2^{\nu_{\mathcal{X}}} N^{-\nu_{\mathcal{X}}/d+\varepsilon} + \lambda_{\mathcal{X}} \right)$$

for any $t \in \{1, \dots, T\}$, any arbitrarily small $\varepsilon > 0$, and K_1, K_2 independent of N, t, ε.

[5] We omit the definition of E_s intentionally as, since $q = 1$, it is always raised to the power of zero in this work.

Proof. Recall

$$\hat{I}_{\mathrm{BQ}}(\theta_t) = \mu_\theta(x^t_{1:N})^\top \left(k_\mathcal{X}(x^t_{1:N}, x^t_{1:N}) + \lambda_\mathcal{X}\mathrm{Id}_N\right)^{-1} f(x^t_{1:N}, \theta_t),$$
$$\sigma^2_{\mathrm{BQ}}(\theta_t) = \mathbb{E}_{X,X'\sim\mathbb{P}_\theta}[k_\mathcal{X}(X, X')] - \mu_\theta(x^t_{1:N})^\top \left(k_\mathcal{X}(x^t_{1:N}, x^t_{1:N}) + \lambda_\mathcal{X}\mathrm{Id}_N\right)^{-1} \mu_\theta(x^t_{1:N}).$$

We seek to bound $\sigma^2_{\mathrm{BQ}}(\theta_t)$. [39, Proposition 3.8] pointed out that the Gaussian noise posterior is the worst-case error in the $\mathcal{H}^{\lambda_\mathcal{X}}_\mathcal{X}$, the RKHS induced by the kernel $k^{\lambda_\mathcal{X}}_\mathcal{X}(x, x') = k_\mathcal{X}(x, x') + \lambda_\mathcal{X}$. Through straightforward algebraic manipulations and using the reproducing property, one can show that

$$\sigma^2_{\mathrm{BQ}}(\theta_t) - \lambda_\mathcal{X} = \mathrm{MMD}^2(\hat{\mathbb{P}}^N_\theta, \mathbb{P}_\theta; \mathcal{H}^{\lambda_\mathcal{X}}_\mathcal{X}) = \sup_{\|f\|_{\mathcal{H}^{\lambda_\mathcal{X}}_\mathcal{X}}\leq 1} \left| w^{\lambda_\mathcal{X}}_t f(x^t_{1:N}) - \int_\mathcal{X} f(x)\mathbb{P}_\theta(\mathrm{d}x)\right|, \quad \text{(A.6)}$$

for the empirical measure $\hat{\mathbb{P}}^N_\theta = w^{\lambda_\mathcal{X}}_t \delta_{x^t_{1:N}}$, where $\delta_{x^t_i}$ for all i is the Dirac delta distribution, $\delta_{x^t_{1:N}} = [\delta_{x^t_1} \ldots \delta_{x^t_N}]^\top$ is our usual vector notation used throughout this work, and the weights are the optimal BQ weights $w^{\lambda_\mathcal{X}}_t = \left(k_\mathcal{X}(x^t_{1:N}, x^t_{1:N}) + \lambda_\mathcal{X}\mathrm{Id}_N\right)^{-1} \mu_\theta(x^t_{1:N})$.

Since $\mathcal{H}^{\lambda_\mathcal{X}}_\mathcal{X}$ is induced by the sum of kernels, $k^{\lambda_\mathcal{X}}_\mathcal{X}(x, x') = k_\mathcal{X}(x, x') + \lambda_\mathcal{X}$, it holds that $\mathcal{H}_\mathcal{X} \subseteq \mathcal{H}^{\lambda_\mathcal{X}}_\mathcal{X}$, and $\|f\|_{\mathcal{H}^{\lambda_\mathcal{X}}_\mathcal{X}} \leq \|f\|_{\mathcal{H}_\mathcal{X}}$ [8, Theorem I.13.IV]. Therefore, the class of functions f for which $\|f\|_{\mathcal{H}_\mathcal{X}} \leq 1$ is larger than that for which $\|f\|_{\mathcal{H}^{\lambda_\mathcal{X}}_\mathcal{X}} \leq 1$, and

$$\sup_{\|f\|_{\mathcal{H}^{\lambda_\mathcal{X}}_\mathcal{X}}\leq 1} \left| w^{\lambda_\mathcal{X}}_t f(x^t_{1:N}) - \int_\mathcal{X} f(x)\mathbb{P}_\theta(\mathrm{d}x)\right| \leq \sup_{\|f\|_{\mathcal{H}_\mathcal{X}}\leq 1} \left| w^{\lambda_\mathcal{X}}_t f(x^t_{1:N}) - \int_\mathcal{X} f(x)\mathbb{P}_\theta(\mathrm{d}x)\right|. \quad \text{(A.7)}$$

Next, note that for $\hat{f}_t(x) = k(x, x^t_{1:N})^\top \left(k_\mathcal{X}(x^t_{1:N}, x^t_{1:N}) + \lambda_\mathcal{X}\mathrm{Id}_N\right)^{-1} f(x^t_{1:N})$,

$$\begin{aligned}\left| w^{\lambda_\mathcal{X}}_t f(x^t_{1:N}) - \int_\mathcal{X} f(x)\mathbb{P}_\theta(\mathrm{d}x)\right| = \left|\int_\mathcal{X} \left(\hat{f}_t(x) - f(x)\right)\mathbb{P}_\theta(\mathrm{d}x)\right| &\leq \int_\mathcal{X} \left|\hat{f}_t(x) - f(x)\right| \mathbb{P}_\theta(\mathrm{d}x) \\ &\leq \|\hat{f}_t - f\|_{\mathcal{L}^2(\mathcal{X})}\|p_\theta\|_{\mathcal{L}^2(\mathcal{X})},\end{aligned} \quad \text{(A.8)}$$

where the last inequality is an application of Hölder inequality. By Assumption (e), $\|p_\theta\|_{\mathcal{L}^2(\mathcal{X})}$ is bounded above by η_0. In order to apply [78, Theorem 4] to bound $\|\hat{f}_t - f\|_{\mathcal{L}^2(\mathcal{X})}$, we show the assumptions in the Theorem hold.

Assumption 1 (Assumptions on the Domain): An open, bounded, and convex $\mathcal{X}$ satisfies the assumption, as discussed in [78].

Assumption 2 (Assumptions on the Kernel Parameters) and Assumption 3 (Assumptions on the Kernel Smoothness Range): Our setting is more specific than the one [78, Theorem 4]: the kernel $k_\mathcal{X}$ is Matérn, and therefore all smoothness constants mentioned in Assumptions 2 and 3 have the same value, $\nu_\mathcal{X} + d/2$.

Assumption 4 (Assumptions on the Target Function and Mean Function): The target function f was assumed to have higher smoothness than $k_\mathcal{X}$ in (f) and (h); the mean function was taken to be zero.

Assumption 5 (Additional Assumptions on Kernel Parameters): By (f) and (h) the smoothness of the true function $s_\mathcal{X} \geq \nu_\mathcal{X} + d/2 > d/2$, which verifies both statements in the Assumption since all smoothness constants of the kernel are equal to $\nu_\mathcal{X} + d/2$.

Therefore [78, Theorem 4] holds, and for $\mathcal{W}^0_2(\mathcal{X}) = \mathcal{L}^2(\mathcal{X})$

$$\|\hat{f}_t - f\|_{\mathcal{L}^2(\mathcal{X})} \leq K_1 h^{d/2}_{x^t_{1:N}} \left(h^{\nu_\mathcal{X}}_{x^t_{1:N}} + \lambda_\mathcal{X}\right),$$

for any N for which the fill distance $h_{x^t_{1:N}} \leq h_0$ for some h_0, and K_1 and h_0 that depend on $\mathcal{X}, s_\mathcal{X}, \nu_\mathcal{X}$. Since $x^t_i \sim \mathbb{P}_{\theta_t}$, we can guarantee that $h_{x^t_{1:N}} \leq h_0$ with high probability using [60, Lemma 2], which says that provided the density $\inf_x p_{\theta_t}(x) > 0$, there is a K_2 such that $\mathbb{E}\, h_{x^t_{1:N}} \leq C_t N^{-1/d+\varepsilon}$ for an arbitrarily small $\varepsilon > 0$, for C_t that depends on t through $\inf_x p_{\theta_t}(x)$: the smaller

$\inf_x p_{\theta_t}(x)$, the larger C_t. Since we assumed $\inf_{x,\theta} p_{\theta_t}(x) = \eta > 0$ there is a K_2 such that $C_t \leq K_2$ for any t. Therefore, we may take N_0 to be the smallest N for which $\mathbb{E}\, h_{x^t_{1:N}} \leq K_2 N^{-1/d+\varepsilon}$ holds, and have for all $N \geq N_0$

$$\mathbb{E}_{x_i^t \sim \mathbb{P}_\theta} \|\hat{f}_t - f\|_{\mathcal{L}^2(\mathcal{X})} \leq K_1 K_2^{d/2} N^{-1/2+\varepsilon} \left(K_2^{\nu_{\mathcal{X}}} N^{-\nu_{\mathcal{X}}/d+\varepsilon} + \lambda_{\mathcal{X}} \right)$$

By Markov's inequality, for any $\delta/2 \in (0,1)$ it holds with probability at least $1 - \delta/2$ that

$$\|\hat{f}_t - f\|_{\mathcal{L}^2(\mathcal{X})} \leq \frac{2}{\delta} K_1 K_2^{d/2} N^{-1/2+\varepsilon} \left(K_2^{\nu_{\mathcal{X}}} N^{-\nu_{\mathcal{X}}/d+\varepsilon} + \lambda_{\mathcal{X}} \right) \tag{A.9}$$

Putting together Equations (A.6) to (A.9) and Assumption (e), we get the result,

$$\begin{aligned}
\sigma^2_{\mathrm{BQ}}(\theta_t) - \lambda_{\mathcal{X}} &= \sup_{\|f\|_{\mathcal{H}^{\lambda_{\mathcal{X}}}_{\mathcal{X}}} \leq 1} \left| w_t^{\lambda_{\mathcal{X}}} f(x^t_{1:N}) - \int_{\mathcal{X}} f(x) \mathbb{P}_\theta(\mathrm{d}x) \right| \\
&\leq \sup_{\|f\|_{\mathcal{H}_{\mathcal{X}}} \leq 1} \left| w_t^{\lambda_{\mathcal{X}}} f(x^t_{1:N}) - \int_{\mathcal{X}} f(x) \mathbb{P}_\theta(\mathrm{d}x) \right| \\
&\leq \sup_{\|f\|_{\mathcal{H}_{\mathcal{X}}} \leq 1} \|\hat{f}_t - f\|_{\mathcal{L}^2(\mathcal{X})} \|p_\theta\|_{\mathcal{L}^2(\mathcal{X})} \\
&\leq \frac{2}{\delta} \eta_0 K_1 K_2^{d/2} N^{-1/2+\varepsilon} \left(K_2^{\nu_{\mathcal{X}}} N^{-\nu_{\mathcal{X}}/d+\varepsilon} + \lambda_{\mathcal{X}} \right).
\end{aligned}$$

□

We are now ready to state our main convergence result, which is a more general version of Theorem 1.

Theorem 3. *Suppose all technical assumptions in Appendix A.2 hold. Then for any $\delta \in (0,1)$ there is a $T_0(\delta) > 0$ and an $N_0 > 0$ such that for any $N \geq N_0$ and $T \geq T_0$, with probability at least $1 - \delta$ it holds that*

$$\|\hat{I}_{\mathrm{CBQ}} - I\|_{\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{tr}})} \leq K_0(\log(12/\delta) + K_3) \left(1 + c^{-1} T^{-1/(2r+1)} \left(\lambda_{\mathcal{X}} + \frac{2}{\delta} \eta_0 K_1 K_2^{d/2} N^{-1/2+\varepsilon} \left(K_2^{\nu_{\mathcal{X}}} N^{-\nu_{\mathcal{X}}/d+\varepsilon} + \lambda_{\mathcal{X}} \right) \right) \right) T^{-r/(2r+1)},$$

for any arbitrarily small $\varepsilon > 0$, and constants K_0, K_1, K_2, K_3 independent of $N, T, \delta, \varepsilon$.

Proof. Recall that for any two events A and B,

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}(\neg A \cup \neg B) \geq 1 - \mathbb{P}(\neg A) - \mathbb{P}(\neg B) = \mathbb{P}(A) + \mathbb{P}(B) - 1.$$

Taking A to be the event in Corollary 1, and B to be the event in Theorem 2,

$$\begin{aligned}
A &= \left\{ \|\hat{I}_{\mathrm{CBQ}} - I\|_{\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{tr}})} \leq K_0(\log(12/\delta) + K_3)(1 + c^{-1} T^{-1/(2r+1)} \sigma^2_{\mathrm{BQ}}(\theta_{t'})) T^{-r/(2r+1)} \right\}, \\
B &= \left\{ \text{for all } t,\ \sigma^2_{\mathrm{BQ}}(\theta_t) \leq \lambda_{\mathcal{X}} + \frac{2}{\delta} \eta_0 K_1 K_2^{d/2} N^{-1/2+\varepsilon} \left(K_2^{\nu_{\mathcal{X}}} N^{-\nu_{\mathcal{X}}/d+\varepsilon} + \lambda_{\mathcal{X}} \right) \right\},
\end{aligned}$$

we get the result. □

As discussed in the main text, convergence is fastest when the regulariser $\lambda_{\mathcal{X}}$ is set to 0; $\lambda_{\mathcal{X}} > 0$ ensures greater stability at the cost of a lower speed of convergence. For clarity we show how Theorem 1 in the main text follows from the more general Theorem 3 by setting $\lambda_{\mathcal{X}} = 0$.

Proof of Theorem 1. Take $\lambda_{\mathcal{X}} = 0$, $C_1(\delta) = K_0(\log(12/\delta) + K_3) = O(\log(1/\delta))$, and $C_2(\delta) = 2c^{-1} K_0(\log(12/\delta) + K_3) \eta_0 K_1 K_2^{d/2+\nu_{\mathcal{X}}} / \delta = O((1/\delta)\log(1/\delta))$ in Theorem 3. The result follows. □

Appendix B Practical Considerations

B.1 Tractable Kernel Means

In the main text, we have shown that both BQ and our method CBQ require that the kernel mean embedding μ and its integral are known in closed-form; see Table 1 in [14] or the `ProbNum` package [77] for pairs of kernels and distributions. When the pair of kernels and distributions does not produce a closed-form embedding, there are multiple other solutions as well.

First, when the embedding of $\mathbb{P}$ is intractable but the embedding of $\mathbb{Q}$ is known, we can use the 'importance sampling trick' which consists of writing the integral as $I = \mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{X\sim\mathbb{Q}}[g(X)]$ where $g(x) = f(x)p(x)/q(x)$ and p, q are the densities of $\mathbb{P}, \mathbb{Q}$. Alternatively, assuming that we know the quantile function Φ^{-1} of $\mathbb{P}$, we can use the 'inverse transform trick' which consists of writing $I = \mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{U\sim\mathbb{U}}[g(U)]$ where $g(u) = f(\Phi^{-1}(u))$ and $\mathbb{U}$ is a uniform distribution on some hypercube. Additionally, if the distribution $\mathbb{P}$ is only known up to the normalization constant, for example the posterior distribution of Bayesian neural networks, then we can use Stein reproducing kernels [6] which provides more flexible closed-form kernel mean embeddings.

Stein Reproducing Kernels Suppose we have a distribution with density $p(x)$ and a function $f(x)$ with the property that $\lim_{x\to\infty} p(x)f(x) = 0$. We can define the Stein operator T_p acting on function f and obtain the Stein identity.

$$T_p[f](x) = f(x)\nabla_x \log p(x) + \nabla_x f(x), \quad \mathbb{E}_p[T_p[f](x)] = 0$$

As a result, for any positive definite kernel $k_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the base kernel, we can obtain a Stein kernel by applying the Stein operator on both arguments of the kernel k_0.

$$\begin{aligned} k_p(x, x') := T_p^x[T_p^{x'}[k_0(x, x')]] =& \nabla_x \log p(x)^\top k_0(x, x')\nabla_{x'} \log p(x') + \nabla_x \log p(x)^\top \nabla_{x'} k_0(x, x') \\ &+ \nabla_{x'} \log p(x')^\top \nabla_x k_0(x, x') + \nabla_x \cdot \nabla_{x'} k_0(x, x') \end{aligned}$$

where $\nabla_x = (\partial/\partial x_1, \cdots, \partial/\partial x_d)^\top$ and $\nabla_x \cdot \nabla_{x'} k_0(x, x') = \sum_{i=1}^d \frac{\partial k_0(x,x')}{\partial x_i \partial x_i'}$. It is noteworthy to mention that when taking the derivative of the logarithm, the normalization constant of $p(x)$ is no longer required.

Stein identity indicates that the kernel mean embedding equals to 0 by construction, i.e $\mu(x') = \int_{\mathcal{X}} k_p(x, x')p(x)dx = 0$. However, this means our GP prior on f encodes the belief that the function has mean zero, which we do not have. Therefore, we add a learnable constant c to the Stein kernel k_p, i.e $\tilde{k}_p(x, x') = k_p(x, x') + c$, so that the kernel mean embedding $\mu(x') = \int_{\mathcal{X}} \tilde{k}_p(x, x')p(x)dx = c$. The constant c for the Stein kernel is selected jointly via maximizing marginal log-likelihood as other hyperparameters. A similar technique has been used in [61] to select control functionals.

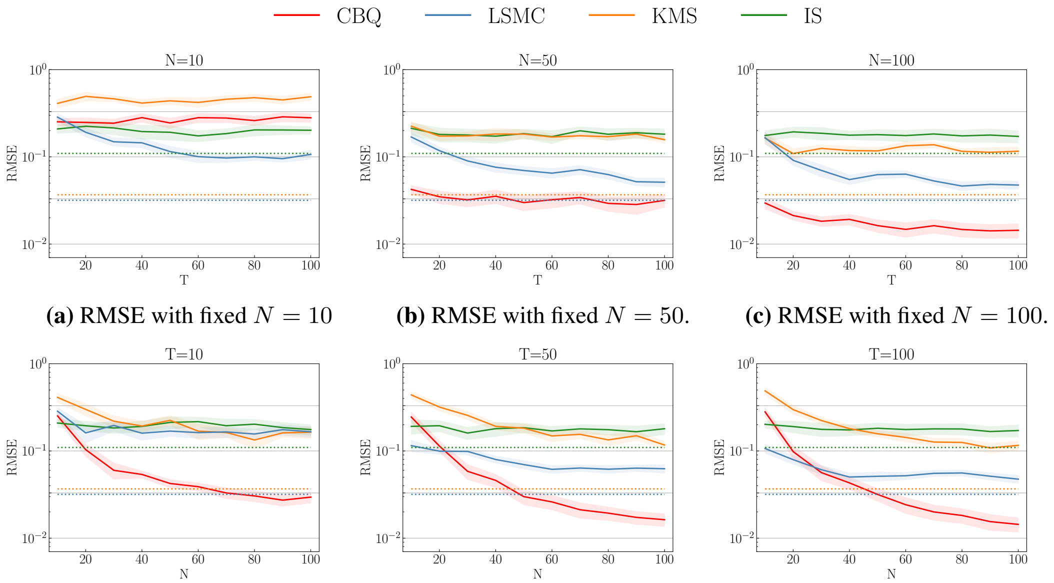
Unlike traditional kernels like RBF or Matérn kernels, Stein reproducing kernels are non-stationary, which implies a prior belief that f has different properties across different parts of $\mathcal{X}$. Therefore for practitioners, using a GP prior with Stein kernel as covariance requires extra caution. Fortunately, our experiments do not exhibit huge differences in performance between Stein and traditional kernels.

B.2 Hyperparameter Selection

In the entirety of the experiments presented within this paper, the Gaussian Process (GP) prior mean functions, denoted as $m_\Theta(\theta)$ and $m_{\mathcal{X}}(x)$, are consistently considered to be zero functions. In accordance with this, the regression target is correspondingly normalized.

Covariance functions determine the properties of samples from a Gaussian process, so the hyperparameters of both kernel $k_{\mathcal{X}}$ and k_Θ needs to be carefully selected. Normally for CBQ, that would include four hyperparameters: lengthscale $l_{\mathcal{X}}$, ℓ_Θ and amplitude $A_{\mathcal{X}}$ and A_Θ. In principle, all hyperparameters are selected via maximising the log-marginal likelihood. For $k_{\mathcal{X}}$, suppose the GP mean $m(x_{1:N})$ is 0, the log-marginal likelihood can be written as [65]:

$$\begin{aligned} L(l_{\mathcal{X}}, A_{\mathcal{X}}) = &-\frac{1}{2} f(x_{1:N})^\top \Big(k_{\mathcal{X}}(x_{1:N}, x_{1:N}; l_{\mathcal{X}}, A_{\mathcal{X}}) + \lambda_{\mathcal{X}} \mathrm{Id}_N\Big)^{-1} f(x_{1:N}) \\ &- \frac{1}{2} \log |k_{\mathcal{X}}(x_{1:N}, x_{1:N}; l_{\mathcal{X}}, A_{\mathcal{X}})| - \frac{N}{2} \log(2\pi). \end{aligned}$$



(a) RMSE with fixed $N = 10$ **(b)** RMSE with fixed $N = 50$. **(c)** RMSE with fixed $N = 100$.

(d) RMSE with fixed $T = 10$ **(e)** RMSE with fixed $T = 50$. **(f)** RMSE with fixed $T = 100$.

Figure 7: *Bayesian sensitivity analysis*. **First Row:** RMSE of all methods when dimension $d = 2$ with fixed $N = 10, 50, 100$ and increasing T. **Second Row:** RMSE of all methods when dimension $d = 2$ with fixed $T = 10, 50, 100$ and increasing N. The intergral is $f(w) = w^\top w$.

We do a grid search over $[1.0, 10.0, 100.0, 1000.0]$ for amplitude $A_\mathcal{X}$ and a grid search over $[0.1, 0.3, 1.0, 3.0, 10.0]$ for lengthscale $l_\mathcal{X}$ and we select the value that gives the largest log-marginal likelihood.

If $k_\mathcal{X}$ is Stein kernel, we have an extra hyperparameter c along with lengthscale $l_\mathcal{X}$ and amplitude $A_\mathcal{X}$. For Stein kernel, we use gradient based optimization like stochastic gradient descent on the log-marginal likelihood to find the optimal value for $c, l_\mathcal{X}, A_\mathcal{X}$. The optimization is implemented with `JAX` autodiff library [12].

For kernel k_Θ, we also optimize the hyperparameters via maximizing log-marginal likelihood.

$$L(l_\Theta, A_\Theta) = -\frac{1}{2}\hat{I}_{\mathrm{BQ}}(\theta_{1:T})^\top \Big(k_\Theta(\theta_{1:T}, \theta_{1:T}; l_\Theta, A_\Theta) + \big(\lambda_\Theta + \sigma^2_{\mathrm{BQ}}(\theta_{1:T})\big)\,\mathrm{Id}_T\Big)^{-1} \hat{I}_{\mathrm{BQ}}(\theta_{1:T}) \\ - \frac{1}{2}\log|k_\Theta(\theta_{1:T}, \theta_{1:T}; l_\Theta, A_\Theta)| - \frac{T}{2}\log(2\pi).$$

Similar to above, we also do a grid search over $[1.0, 10.0, 100.0, 1000.0]$ for amplitude A_Θ and a grid search over $[0.1, 0.3, 1.0, 3.0, 10.0]$ for lengthscale l_Θ and we select the value that gives the largest log-marginal likelihood.

Appendix C Experiments

In this section, we provide more detailed description of the settings in all experiments in the main text, and we provide further results and ablation studies. All figures reported in the paper are created using the median values obtained from 20 separate runs with different random seeds. Standard error is shown as shaded area around the median.

C.1 Synthetic Experiment: Bayesian Sensitivity Analysis for Linear Models

C.1.1 Experimental Setting

In the toy experiment, we do sensitivity analysis on the hyperparameters in Bayesian linear regression. Our observations are $\mathcal{D} = \{Y \in \mathbb{R}^{m \times d}, z \in \mathbb{R}^m\}$ where m is the number of observations and d

is the dimension including the intercept. We use a $\mathcal{N}(w; 0, \theta \mathrm{Id}_d)$ prior on the regression weights $w \in \mathbb{R}^d$, where the covariance matrix is a diagonal matrix with values $\theta \in \mathbb{R}^p$ (and here $p = d$). Using a Gaussian likelihood $p(z \mid w) = \mathcal{N}(z; w^\top y, \eta)$, we can obtain (via conjugacy) a multivariate Gaussian posterior $p_\theta(w|\mathcal{D})$ whose mean and variance have a closed form expression [11].

$$p_\theta(w \mid \mathcal{D}) = \mathcal{N}(\tilde{m}, \tilde{\Sigma}), \quad \tilde{\Sigma}^{-1} = \operatorname{diag}(\theta)^{-1} + \eta Y^\top Y, \quad \tilde{m} = \eta \tilde{\Sigma} Y^\top z$$

We can then analyse the sensitivity of the posterior to θ by computing $I(\theta) = \int_{\mathbb{R}^d} f(w) p_\theta(w|\mathcal{D}) dw$ where f represents the quantity of interest. For example, if $f(w) = w^\top w$, then $I(\theta)$ is the second moment of the posterior and the results are already reported in the main text. If $f(w) = w^\top y^*$ for some new observation y^*, then $I(\theta)$ is the predictive mean. Both integrals are known to have closed form expression in Bayesian linear regression [11], so it is a good synthetic example to benchmark our approach. We sample parameter values $\theta_{1:T}$ from a $\operatorname{Unif}(\Theta)$ where $\Theta = (1, 3)^d$, and for each such parameter θ_t, we obtain N observations from $p_{\theta_t}(w|\mathcal{D})$. In total, we have $N \times T$ samples.

For conditional Bayesian quadrature (CBQ), we need to specify two kernels. First, we choose the kernel on the space of parameter $w \in \mathbb{R}^d$ (corresponds to $k_{\mathcal{X}}$ in Appendix 3) to be a Gaussian kernel with lengthscale l and amplitude A

$$k(w, w') = A \exp(-\frac{1}{2l^2}(w - w')^\top (w - w')) \tag{C.10}$$

So as a result we can have a closed from kernel mean embedding under the Gaussian posterior distribution.

$$\mu_\theta(w) = A |\, \mathrm{I} + l^{-2} \tilde{\Sigma} \,|^{-1/2} \exp\left(-\frac{1}{2}(w - \tilde{m})^\top (\tilde{\Sigma} + l^2 \mathrm{Id}_d)^{-1} (w - \tilde{m})\right) \tag{C.11}$$

And the integral of μ (known as the initial error) also has a closed form

$$\int \mu_\theta(w) p_\theta(w) dw = \frac{Al}{\sqrt{|\, l^2 \mathrm{Id}_d + 2\tilde{\Sigma} \,|}} \tag{C.12}$$

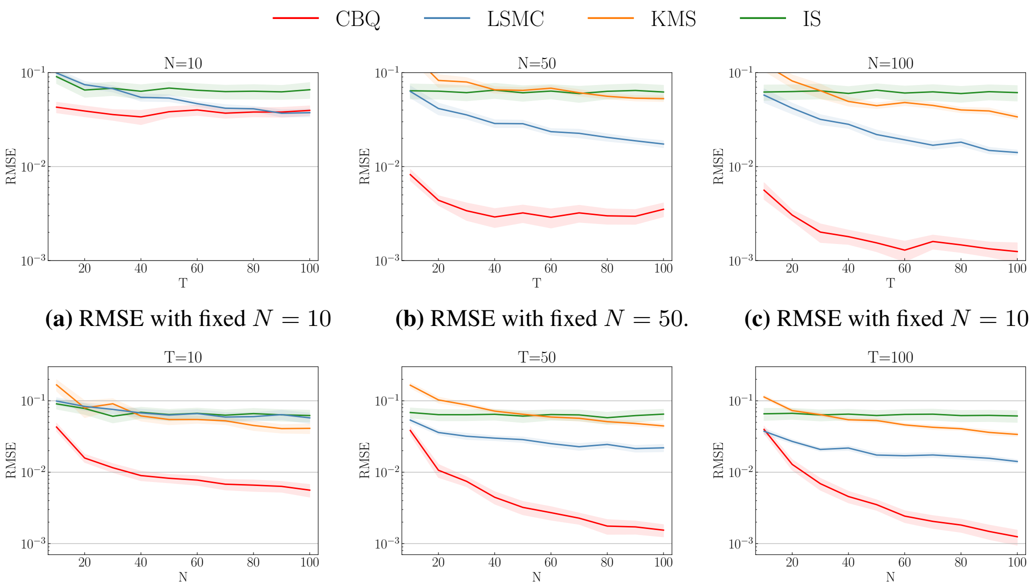
Then we choose the kernel on the space of θ to be Matérn-3/2 kernel. The hyperparameters for both kernels are selected according to Appendix B.2.

There are hyperparameters in baseline methods as well. For importance sampling (IS) estimator, we use $p_{\theta_1}(w), \cdots, p_{\theta_T}(w)$ as our importance distributions in IS. For kernel mean shrinkage (KMS) estimator, we also use Matérn-3/2 kernel on the space of θ and select hyperparameters according to Appendix B.2. For least square Monte Carlo (LSMC), the hyperparameter is the order of polynomials, which is chosen among the set $\{1, 2, 3, 4\}$ that returns the best performance.

C.1.2 More Experimental Results

We provide more experimental results for Bayesian sensitivity analysis here. In Figure 7, the integrand is chosen to be $f(w) = w^\top w$ and the dimension d is 2. In the first row of Figure 7, we fix $N = 10, 50, 100$ showing the performance of RMSE with increasing T. In the second row of Figure 7, we fix $T = 10, 50, 100$ showing the performance of RMSE with increasing N. In Figure 8, the integrand is chosen to be $f(w) = w^\top y^*$ and the dimension d is 2. In the first row of Figure 8, we fix $N = 10, 50, 100$ showing the performance of RMSE with increasing T. In the second row of Figure 8, we fix $T = 10, 50, 100$ showing the performance of RMSE with increasing N. We can see that CBQ has demonstrated consistent smaller RMSE for both integrands under the same number of samples and faster convergence rate compared to all other baseline methods. Also, we can confirm the theory that CBQ has a faster convergence rate in N than in T.

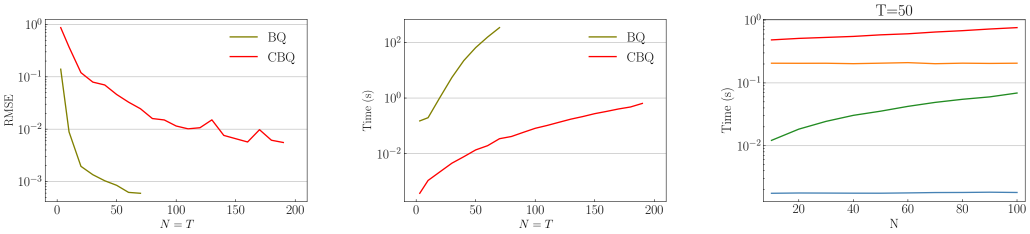
In Figure 9c, we show the computational cost of different methods in Bayesian sensitivity analysis for fixed $T = 50$. It is clear from the figure that CBQ is more computationally expensive, so in this simple setting it is more efficient to spend more budget on obtaining more samples. Nonetheless, in scenarios where the expense of sample collection constitutes a significant fraction of the computational budget, or when the evaluation of the integrand proves to be highly costly, it becomes more cost-effective to spend a larger share of the budget towards CBQ.



(a) RMSE with fixed $N = 10$ **(b)** RMSE with fixed $N = 50$. **(c)** RMSE with fixed $N = 100$.

(d) RMSE with fixed $T = 10$ **(e)** RMSE with fixed $T = 50$. **(f)** RMSE with fixed $T = 100$.

Figure 8: *Bayesian sensitivity analysis*. **First Row:** RMSE of all methods when dimension $d = 2$ with fixed $N = 10, 50, 100$ and increasing T. **Second Row:** RMSE of all methods when dimension $d = 2$ with fixed $T = 10, 50, 100$ and increasing N. The intergral is $f(w) = w^\top y^*$.



(a) CQ and CBQ RMSE **(b)** CQ and CBQ time **(c)** Computational time

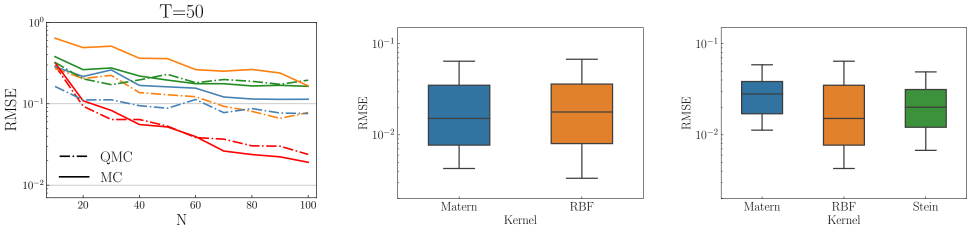
Figure 9: Left: Comparison of BQ and CBQ in terms of time (wall clock time) and RMSE in Bayesian sensitivity analysis. **Right:** Computational time (wall clock time) for different methods in Bayesian sensitivity analysis with increasing T under dimension $d = 2$ and fixed $T = 50$.

C.1.3 Comparison of BQ and CBQ

In the main text, we mentioned the comparison of multioutput BQ and CBQ in terms of their computational complexity and convergence rate. For T parameter values $\theta_1, \cdots, \theta_T$ and N samples from each probability distribution $\mathbb{P}_{\theta_1}, \ldots, \mathbb{P}_{\theta_T}$, the computational cost is $\mathcal{O}(N^3T^3)$ for multioutput BQ and $\mathcal{O}(TN^3)$ for CBQ. In Figure 9a and Figure 9b, we fix $N = T$ and demonstrate that BQ has a much faster convergence rate but the computational time grows unbearable quickly as the number of samples increases.

C.1.4 Quasi Monte Carlo

Quasi Monte Carlo (QMC) is another line of research on improving the precision of approximating intractable integrals. QMC aims to cover the integration domain more uniformly than random sampling used in standard Monte Carlo methods [57, 34, 24]. Sobol sequences are a type of low-discrepancy sequence commonly used in Quasi-Monte Carlo (QMC) methods, which are able to cover the multidimensional space more uniformly than random sequences, resulting in a faster convergence rate. However, since Sobol sequences are deterministic, we follow the technique introduced in randomized QMC [47] to shift the Sobol sequence by a random amount so that we can combine the



(a) Quasi Monte Carlo **(b)** Ablation on kernel k_Θ **(c)** Ablation on kernel $k_\mathcal{X}$

Figure 10: Left: Comparison of all methods with standard uniform sampling and Quasi-Monte Carlo methods. **Middle:** Ablation study for CBQ with different $k_\mathcal{X}$ kernels in Bayesian sensitivity analysis. **Right:** Ablation study for CBQ with different k_θ kernels in Bayesian sensitivity analysis.

advantages of deterministic sampling from QMC and the robustness of randomness from standard Monte Carlo methods.

For our method CBQ, we put no restrictions on how the data are being generated and we do not require i.i.d sampling. Therefore, we implement QMC sampling and compare the performances of all methods with random sampling in Figure 10a. It can be observed that Quasi-Monte Carlo (QMC) significantly enhances the performance of baseline methods, such as Kernel Mean Shrinkage (KMS) and Least Squares Monte Carlo (LSMC), while subtly improves the performance of CBQ. The limited degree of improvement seen in CBQ with QMC sampling can be attributed to the fact that CBQ already yields a remarkably low RMSE. Consequently, the margin of improvement offered by QMC sampling is not as evident in CBQ as in the baseline methods.

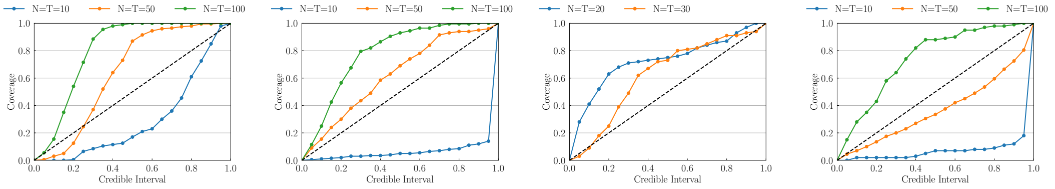
C.1.5 Ablations

We present an ablation study evaluating the impact of distinct kernel choices $k_\mathcal{X}$ and k_Θ within the framework of Bayesian sensitivity analysis. The Matérn-3/2 kernel and Gaussian Radial Basis Function (RBF) kernel are selected for k_Θ. As illustrated in Figure 10b, the performance of the CBQ remains consistent across these different k_Θ kernels.

Subsequently, we opt for Matérn-3/2, Gaussian RBF, and Stein kernel (with Matérn-3/2 as the base kernel) as choices for $k_\mathcal{X}$. When $k_\mathcal{X}$ is the RBF kernel, the formula for kernel mean embedding $\mu_\theta(w)$ is presented in (C.11). In the scenario where $k_\mathcal{X}$ is the Matérn-3/2 kernel, a closed form expression for the kernel mean embedding does not exist for the non-isotropic Gaussian distribution $\mathcal{N}(\tilde{m}, \tilde{\Sigma})$. Consequently, we employ the reparameterization trick, initially sampling ϵ from $\mathcal{N}(0, \mathrm{Id}_d)$, then calculating $w = \tilde{m} + L^\top \epsilon$ where L is the lower triangular matrix derived from the Cholesky decomposition of the covariance matrix $\tilde{\Sigma}$. In essence, $I(\theta) = \int_{\mathbb{R}^d} f(w)\mathcal{N}(w; \tilde{m}, \tilde{\Sigma})dw = \int_{\mathbb{R}^d} f(\tilde{m} + L^\top \epsilon)\mathcal{N}(\epsilon; 0, \mathrm{Id}_d)d\epsilon$. The closed form expression of kernel mean embedding for Matérn-3/2 kernel and the Gaussian distribution can be found in the Appendix S.3 of [53]. When $k_\mathcal{X}$ is Stein kernel, we choose Matérn-3/2 kernel k_0 as the base kernel and then apply Stein operator on both arguments of kernel k_0. In Figure 10c, we can see that CBQ performance is consistent under different types of kernels k_Θ. All kernel hyperparameters are chosen according to Appendix B.2.

C.1.6 Calibration

In Figure 11a, we show the calibration of the CBQ posterior for the integrand $f(w) = w^\top w$ when $d = 2$. The coverage is the percentage of times a credible interval contains $I(\theta)$ under repetitions of the experiment. The black diagonal line represents the ideal case, with any curve lying above the black line indicating underconfidence and any curve lying below indicating overconfidence. It is generally more preferable to be underconfident than overconfident. We observe that when the number of samples is as small as 10, CBQ is overconfident, which can be explained by a poor performance of empirical Bayes in the small sample regime. On the other hand, when N and T increase, CBQ becomes underconfident, meaning that our posterior variance is more inflated than needed from a frequentist viewpoint. The calibration plots for other experiments are all demonstrated in Figure 11, and the conclusions are consistent across different experiments.



(a) Calibration **(b)** Calibration **(c)** Calibration **(d)** Calibration

Figure 11: Left: Calibration plot for CBQ in Bayesian sensitivity analysis. **Middle Left:** Calibration plot for CBQ in Black-Scholes model. **Middle Right:** Calibration plot for CBQ in SIR sensitivity analysis. **Right:** Calibration plot for CBQ in uncertainty decision making.

C.2 Butterfly Call Option with the Black-Scholes Model

C.2.1 Experimental Setting

In this experiment, we consider specifically an asset whose price $S(\tau)$ at time τ follows the Black-Scholes formula

$$S(\tau) = S_0 \exp\Big(\sigma W(\tau) - \sigma^2 \tau/2\Big), \quad \text{for} \quad \tau \geq 0$$

where σ is the underlying volatility, S_0 is the initial price and W is the standard Brownian motion. The financial derivative we are interested in is a butterfly call option whose payoff at time ζ can be expressed as

$$\psi(S(\tau)) = \max(S(\tau) - K_1, 0) + \max(S(\tau) - K_2, 0) - 2\max(S(\tau) - (K_1 + K_2)/2, 0)$$

In addition to the expected payoff, insurance companies are interested in computing the expected loss of their portfolios if a shock would occur in the economy. We follow the setting in [4, 5] assuming that a shock occur at time η, at which time the price is $S(\eta) = \theta$, and this shock multiplies the option price by $1 + s$. The expected loss caused by the shock can be expressed as

$$\mathcal{L} = \mathbb{E}\Big[\max\Big(\mathbb{E}\left[\psi\left(S(\zeta)\right) - \psi\left((1+s)S(\zeta)\right) \mid S(\eta)\right], 0\Big)\Big]$$

where ζ is the maturity time of the option. Equivalently, the expected loss of the option can be expressed as $\mathcal{L} = \mathbb{E}[\max(I(\theta), 0)]$, where $I(\theta) = \int_0^\infty f(x) p_\theta(x) dx$, $x = S(\zeta)$ is the price at the time ζ at which the option matures, $f(x) = \psi(x) - \psi((1+s)x)$, and p_θ is the density of a log-normal distribution induced from the Black-Scholes model.

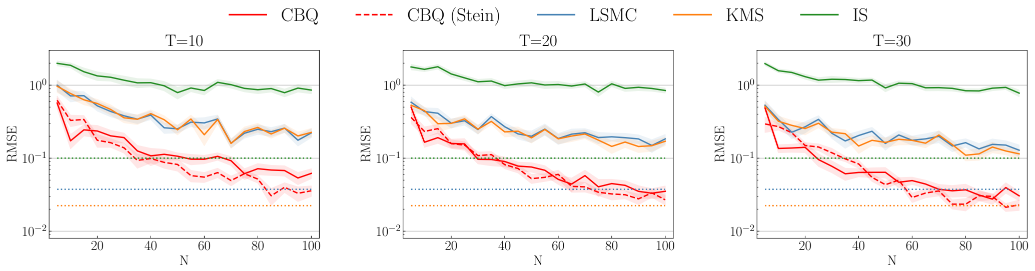
We consider the initial price $S_0 = 100$, the volatility $\sigma = 0.3$, the strikes $K_1 = 50, K_2 = 150$, the option maturity $\zeta = 2$ and the shock happens at $\eta = 1$ with strength $s = 0.2$. The observations $\theta_{1:T}$ are sampled from the log normal distribution deduced from the Black-Scholes formula $\theta_{1:T} \sim \text{Lognormal}\left(\log S_0 - \frac{\sigma^2}{2}\eta, \sigma^2\eta\right)$. Then N observations of $x_{1:N}^t$ are sampled from the log normal distribution deduced from the Black-Scholes formula $x_{1:N}^t \sim \text{Lognormal}\left(\log\theta_t - \frac{\sigma^2}{2}(\zeta - \eta), \sigma^2(\zeta - \eta)\right)$.

For conditional Bayesian quadrature (CBQ), we need to specify two kernels. First we choose the kernel $k_\mathcal{X}$. Since in Black-Scholes model, $p_\theta(x)$ is a log normal distribution, we use a log RBF kernel so that we can have a closed form mean embedding μ. We know that $p_\theta(x)$ is the log normal distribution derived from the Black-Scholes model Lognormal$(\bar{m}, \bar{\sigma}^2)$ with $\bar{m} = \log\theta - \frac{\sigma^2}{2}(\zeta - \eta)$ and $\bar{\sigma}^2 = \sigma^2(\zeta - \eta)$. The log RBF kernel is defined as

$$k(x, x') = A\exp(-\frac{1}{2l^2}(\log x - \log x')^2)$$

and the kernel mean embedding has a closed form

$$\mu_\theta(x) = A \exp\left(-\frac{\bar{m}^2 + (\log x)^2}{2(\bar{\sigma}^2 + l^2)}\right) x^{\frac{\bar{m}}{\bar{\sigma}^2 + l^2}} \Bigg/ \sqrt{1 + \frac{\bar{\sigma}^2}{l^2}}$$



(a) RMSE with fixed $T = 10$ **(b)** RMSE with fixed $T = 20$. **(c)** RMSE with fixed $T = 30$.

Figure 12: *Black-Scholes model.* RMSE of all methods with fixed $T = 10, 20, 30$ and increasing N.

The integral of kernel mean μ does not have a closed form expression, so we use the empirical average as an approximation.

For CBQ with Stein kernel, we use Matérn-3/2 as the base kernel and then apply Stein operator to both arguments of the base kernel to obtain $k_{\mathcal{X}}$. Then we choose the kernel k_Θ as Matérn-3/2 kernel. All hyperparameters in $k_{\mathcal{X}}$ and k_Θ are selected according to Appendix B.2.

There are hyperparameters in other baseline methods as well. For importance sampling (IS) estimator, we use $p_{\theta_1}(x), \cdots, p_{\theta_T}(x)$ as our importance distributions in IS. For kernel mean shrinkage (KMS) estimator, we also use Matérn-3/2 kernel on the space of θ and select hyperparameters according to Appendix B.2. For least square Monte Carlo (LSMC), the hyperparameter is the order of polynomials. We choose the order among the set $\{1, 2, 3, 4\}$ that returns the best performance.

C.2.2 More Experimental Results

We report more experimental results for computing the expected loss in butterfly call option with the Black-Scholes model in Figure 12 with fixed $T = 10, 20, 30$ and increasing N. We can see that the CBQ consistently shows smaller RMSE. Also, the performance is similar between $k_{\mathcal{X}}$ being Stein kernel and $k_{\mathcal{X}}$ being log RBF kernel. In Figure 11b, we also show the calibration of CBQ uncertainty in this experiment.

C.3 Bayesian Sensitivity Analysis for Susceptible-Infectious-Recovered (SIR) Model

C.3.1 Experimental Setting

The SIR model is commonly used to simulate the dynamics of infectious diseases through a population. It divides the population into three sections. Susceptibles (S) represents people who are not infected but can be infected after getting contact with an infectious individual. Infectious (I) represents people who are currently infected and can infect susceptible individuals. Recovered (R) represents individuals who have been infected and then removed from the disease, either by recovering or dying. The dynamics are governed by a system of ordinary differential equations (ODE) as below.

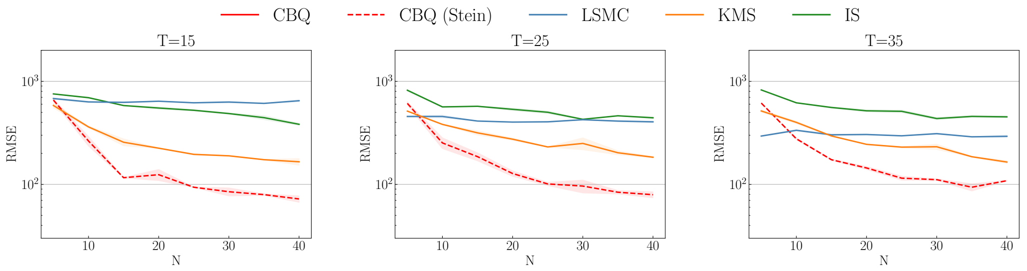
$$\frac{\mathrm{d}S}{\mathrm{d}t} = -xSI, \quad \frac{\mathrm{d}I}{\mathrm{d}t} = xSI - \gamma I, \quad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I$$

x is the infection rate and γ is the recovery rate. The solution to the SIR model would be a vector of (N_I^r, N_S^r, N_R^r) representing the number of infectious, susceptible and recovered at day r. In this experiment, we use scipy `odeint` [75] to numerically solve the ODEs.

In this experiment, we assume that the recovery rate γ is fixed and x follows a gamma prior distribution $x \sim \text{Gamma}(\theta, \xi)$ where θ represents the initial belief of the infection rate deduced from the study of the virus in the laboratory at the beginning of the outbreak, and ξ represents the amount of uncertainty. The target of interest is the expected peak number of infected individuals under the prior distribution on x:

$$I_{\text{max}}(\theta) = \mathbb{E}_x \left[\max_r N_I^r(x) \mid \theta \right] = \int_{\mathcal{X}} \max_r N_I^r(x) p_\theta(x) dx$$

It is important to know how different initial estimate (different θ) of the infection rate will lead to different final estimate of I_{max}. In this experiment, we fix the rate parameter $\xi = 10$, and alter



(a) RMSE with fixed $T = 15$ **(b)** RMSE with fixed $T = 25$. **(c)** RMSE with fixed $T = 35$.

Figure 13: *SIR model.* RMSE of all methods with fixed $T = 15, 25, 35$ and increasing N.

the shape parameter θ. The total population is set to be 10^6 and the recovery rate $\gamma = 0.05$. The observations $\theta_{1:T}$ are sampled from the uniform distribution $\mathrm{Unif}\,(2, 9)$ and then N observations of $x^t_{1:N}$ are sampled from $\mathrm{Gamma}(\theta_t, \xi)$.

For conditional Bayesian quadrature (CBQ), we need to specify two kernels. First we choose $k_{\mathcal{X}}$, we use Matérn-3/2 as the base kernel and then apply Stein operator to both arguments of the base kernel to obtain $k_{\mathcal{X}}$. Then we choose k_{Θ} as Matérn-3/2 kernel. All hyperparameters in $k_{\mathcal{X}}$ and k_{Θ} are selected according to Appendix B.2. We use a MC estimator with 5000 samples as the pseudo ground truth and evaluate the RMSE across all methods.

There are hyperparameters in other baseline methods as well. For importance sampling (IS) estimator, we use $p_{\theta_1}(x), \cdots, p_{\theta_T}(x)$ as our importance distributions in IS. For kernel mean shrinkage (KMS) estimator, we also use Matérn-3/2 kernel on the space of θ and select hyperparameters according to Appendix B.2. For least square Monte Carlo (LSMC), the hyperparameter is the order of polynomials. We choose the order among the set $\{1, 2, 3, 4\}$ that returns the best performance.

C.3.2 More Experimental Results

We report more experimental results for computing the expected peak number of infections from SIR model in Figure 13 with fixed $T = 15, 25, 35$. We can see that the CBQ consistently shows smaller RMSE. In Figure 11c, we also show the calibration of CBQ uncertainty in this experiment.

C.4 Uncertainty Decision Making in Health Economics

C.4.1 Experimental Settings

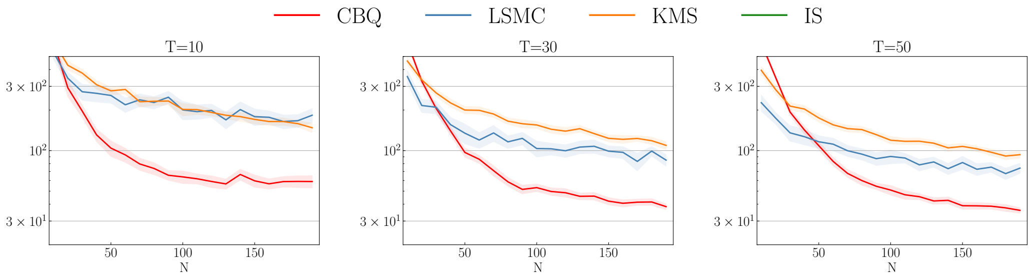
In the medical world, it is important to compare the cost and the relative advantage of conducting an extra medical experiment [13]. In the area of oil and gas reservoir, an cost analysis is necessary before deciding whether to drill additional wells. The expected value of partial perfect information (EVPPI) quantifies the expected gain from conducting extra experiments to obtain precise knowledge of some unknown variables [13]. EVPPI can be expressed as

$$\mathrm{EVPPI} = \mathbb{E}\Big[\max_c \mathbb{E}[f_c(X, \theta) \mid \theta]\Big] - \max_c \mathbb{E}\Big[f_c(X, \theta)\Big]$$

where $c \in \mathcal{C}$ is a set of potential treatments and f_c measures the potential outcome of treatment c. Our method is applicable for estimating the inner conditional expectation of the first term. After denoting $I(\theta) = \mathbb{E}[f_c(X, \theta) \mid \theta] = \int_{\mathcal{X}} f_c(x, \theta) p(x|\theta) dx$, we have

$$\mathrm{EVPPI} = \mathbb{E}\Big[\max_c I_c(\theta)\Big] - \max_c \mathbb{E}\Big[f_c(X, \theta)\Big]$$

We adopt the same experimental setup as delineated in [26], wherein X and θ have a joint 19-dimensional Gaussian distribution, meaning that $p(x \mid \theta)$ is a Gaussian distribution. The specific meanings of all X and θ are outlined in Table 1. All these variables are independent except that $\theta_1, \theta_2, X_6, X_{14}$ are pairwise correlated with a correlation coefficient 0.6. The observations $\theta_{1:T}$ are sampled from the marginal Gaussian distribution $p(\theta)$ and then N observations of $x^t_{1:N}$ are sampled from $p(x \mid \theta_t)$.



(a) RMSE with fixed $T = 10$ **(b)** RMSE with fixed $T = 30$. **(c)** RMSE with fixed $T = 50$.

Figure 14: *Uncertainty decision making.* RMSE of all methods with fixed $T = 10, 30, 50$ and increasing N.

Variables	Mean	Std	Meaning
X_1	1000	1.0	Cost of treatment
X_2	0.1	0.02	Probability of admissions
X_3	5.2	1.0	Days of hospital
X_4	400	200	Cost per day
X_5	0.3	0.1	Utility change if response
X_6	3.0	0.5	Duration of response
X_7	0.25	0.1	Probability of side effects
X_8	-0.1	0.02	Change in utility if side effect
X_9	0.5	0.2	Duration of side effects
X_{10}	1500	1.0	Cost of treatment
X_{11}	0.08	0.02	Probability of admissions
X_{12}	6.1	1.0	Days of hospital
X_{13}	0.3	0.05	Utility change if response
X_{14}	3.0	1.0	Duration of response
X_{15}	0.2	0.05	Probability of side effects
X_{16}	-0.1	0.02	Change in utility if side effect
X_{17}	0.5	0.2	Duration of side effects
θ_1	0.7	0.1	Probability of responding
θ_2	0.8	0.1	Probability of responding

Table 1: Variables in the health economics experiment.

Our target of interest is EVPPI under a binary decision-making problem ($\mathcal{C} = \{1, 2\}$) with $f_1(x, \theta) = 10^4(\theta_1 x_5 x_6 + x_7 x_8 x_9) - (x_1 + x_2 x_3 x_4)$ and $f_2(x, \theta) = 10^4(\theta_2 x_{13} x_{14} + x_{15} x_{16} x_{17}) - (x_{10} + x_{11} x_{12} x_4)$. We estimate $I_c(\theta)$ with CBQ and baselines, and use MC for the outer expectation.

For conditional Bayesian quadrature (CBQ), we select Matérn-3/2 for $k_{\mathcal{X}}$ and also Matérn-3/2 for k_Θ. The kernel mean embedding has a closed form expression according to the discussion in Appendix C.1.5. All hyperparameters in $k_{\mathcal{X}}$ and k_Θ are selected according to Appendix B.2. Note that IS is no longer applicable here because f depends on both x and θ, so we only comparing CBQ against KMS and LSMC. We draw 10^7 samples from the joint distribution to generate a pseudo ground truth, and evaluate the RMSE across different methods.

There are hyperparameters in other baseline methods as well. For kernel mean shrinkage (KMS) estimator, we also use Matérn-3/2 kernel on the space of θ and select hyperparameters according to Appendix B.2. For least square Monte Carlo (LSMC), the hyperparameter is the order of polynomials. We choose the order among the set $\{1, 2, 3, 4\}$ that returns the best performance.

C.4.2 More Experimental Results

We report more experimental results for computing the EVPPI in Figure 14 with fixed $T = 10, 20, 50$ and increasing N. We can see that the CBQ consistently shows smaller RMSE. In Figure 11d, we also show the calibration of CBQ uncertainty in this experiment.