Conditional Bayesian Quadrature

Anonymous Author(s)
Affiliation
Address
email

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 a smooth function of the parameter 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 [Tang, 2013], moment generating, characteristic, or cumulative distribution functions [Giles et al., 2015, Krumscheid and Nobile, 2018], the conditional value at risk or various valuations of options [Longstaff and Schwartz, 2001, Alfonsi et al., 2022], for Bayesian sensitivity analysis [Lopes and Tobias, 2011, Kallioinen et al., 2021], or even more broadly for scientific sensitivity analysis; see for example Sobol indices [Sobol, 2001]. 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 the *nested expectation* given by $\mathbb{E}_{\theta \sim \mathbb{Q}}[\phi(I(\theta))]$ [Hong and Juneja, 2009, Rainforth et al., 2018]. These arise widely when computing the expected information gain in Bayesian experimental design [Chaloner and Verdinelli, 1995], and for computing the expected value of partial perfect information in health economics [Heath et al., 2017].

Existing methods for computing $I(\theta)$ 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. These can be used to estimate $I(\theta_1), \ldots, I(\theta_T)$ with classical Monte Carlo methods, but in most applications we will also be 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, existing algorithm have a second step where these estimators are combined to make them valid for any θ.

The simplest approach is importance sampling [Glynn and Igelhart, 1989, Madras and Piccioni, 1999, Tang, 2013, Demange-Chryst et al., 2022]), where $I(\theta)$ is estimated by weighting function

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

evaluations to account for the fact that the samples were not obtained from $\mathbb{P}_\theta$. Despite its simplicity, this approach is only applicable when f does not depend on θ, and it is usually difficult to identify an appropriate importance distribution which will work well uniformly over $\theta \in \Theta$. Alternatively, least-squares Monte Carlo [Longstaff and Schwartz, 2001, Alfonsi et al., 2022] or regression-based kernel mean shrinkage estimators Muandet et al. [2016], Chau et al. [2021] 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 highly dependent on the accuracy of the Monte Carlo estimators and of the regression method.

In addition, there are two main limitations to these existing methods. Firstly, they are very sample-intensive: a total of NT function evaluations are required, and the corresponding convergence rate is relatively slow: $\mathcal{O}(N^{-1/2}T^{-1/2})$ for importance sampling, and $\mathcal{O}((C + N^{-1/2})T^{-\beta})$ for some $C > 0$ for least-squares Monte Carlo. This means that both N and T need to be large to obtain a good accuracy, which makes the method infeasible if sampling or evaluating the integrand is expensive. Secondly, obtaining a good, finite-sample, quantification of uncertainty for $I(\theta)$ is often infeasible.

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 [Diaconis, 1988, O'Hagan, 1991, Rasmussen and Ghahramani, 2002, Briol et al., 2019] algorithm to the computation of parametric or conditional expectations. As such, CBQ falls in the line of work on probabilistic numerical methods [Hennig et al., 2015, Cockayne et al., 2019, Oates and Sullivan, 2019, Hennig et al., 2022]. 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 is parametrised by θ. This posterior provides a finite-sample Bayesian quantification of uncertainty for $I(\theta)$.

In Section 4, we show that our method is more sample efficient than least-squares Monte Carlo in that it achieves a faster convergence rate of $\mathcal{O}((C' + N^{-\alpha})T^{-\beta})$ under mild smoothness conditions on f and $\mathbb{P}_\theta$, where $C' > 0$ and the parameters $\alpha > \frac{1}{2}$ and $\frac{1}{2} < \beta < 1$ depend on the Gaussian process prior and the dimensionality of $\mathcal{X}$ and Θ. As a result, smaller N and T are needed to achieve a desired accuracy, and the method will therefore be preferable for expensive problems.

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 a central limit theorem. We will assume that our observations are any 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, as well as 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 $\hat{I}_{\text{MC}}(\theta_t) := \frac{1}{N}\sum_{i=1}^N f(x^t_i, \theta_t)$ [Robert and Casella, 2000]. The disadvantage of this approach is that we cannot estimate $I(\theta)$ for any $\theta \notin \{\theta_1 \ldots, \theta_T\}$, 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 is independent of 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 $\theta \in \Theta$: $\hat{I}_{\text{IS}}(\theta) := \frac{1}{N}\sum_{t=1}^T\sum_{i=1}^N (p_{\theta_t}(x^t_i)/p_\theta(x^t_i))f(x^t_i)$. In this case, $\mathbb{P}_{\theta_1}, \ldots, \mathbb{P}_{\theta_T}$ are called importance distributions, and their choice is made to minimise the variance of $\hat{I}_{\text{IS}}(\theta)$ [Glynn and Igelhart, 1989,

Madras and Piccioni, 1999, Tang, 2013]. We could also select importance distributions outside the parametric family, but this makes this choice even more challenging [Demange-Chryst et al., 2022].

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) [Longstaff and Schwartz, 2001] and kernel mean shrinkage estimators (KMS) [Chau et al., 2021] 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) [Berlinet and Thomas-Agnan, 2011]. Clearly, both the performance and computational cost of these estimators will depend on the choice of family. LSMC will cost $\mathcal{O}(TN + p^3)$, whereas KMS has a computational cost of $\mathcal{O}(TN + T^3)$. On the other hand, KMS will clearly lead to a much smaller error than LSMC when $I(\theta)$ cannot be approximated well by a low-order polynomial.

Other Related Work Another possible approach are methods based on sub-fields of transfer learning, such as multi-task learning [Xi et al., 2018, Gessner et al., 2020, Sun et al., 2021] or meta-learning [Sun et al., 2023]. These methods tend to assume that several related integrals need to be computed, and the relationship between these is encoded through a vector-valued RKHS or by assuming they are independent draws from an environment of integration tasks. As such, they do not explicitly assume that the conditional expectation is a smooth function of some parameter, and are therefore not able to make use of this property. On the other hand, some approaches use more structure: for example, multi-level methods assume that θ is a parameter encoding the accuracy, but also computational cost, of an integrand [Giles et al., 2015].

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) [Diaconis, 1988, O'Hagan, 1991, Rasmussen and Ghahramani, 2002, Briol et al., 2019], we begin by positing a Gaussian process (GP) prior on f; a GP [Rasmussen and Williams, 2006] 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 distributed. 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.

[Hudson: If we add a constant for CBQ, then we also need a λ here.] [Hudson: We can use the result from Goerge's paper, that Theorem 4, misspecified setting. We assume Gaussian likelihood, and actually we have no noise, so ϵ is 0, and the rate is the same as the interpolation case. Check the paragraph 'Nugget likelihood' and Collary 5.] Once a GP prior has been selected, we condition this prior on data $f(x_{1:N}) = [f(x_1), \ldots, f(x_T)]^\top$ for $x_{1:N} \in \mathcal{X}^N$. This leads to a posterior GP on f, which itself induces a univariate Gaussian posterior distribution $\mathcal{N}\big(\hat{I}_{\text{BQ}}, \sigma^2_{\text{BQ}}\big)$ on I, where the mean and variance are:

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

and $\mu(x) = \mathbb{E}_{X \sim \mathbb{P}}[k_{\mathcal{X}}(X, x)]$ is known as the kernel mean embedding of the distribution $\mathbb{P}$ [Muandet et al., 2017]. The function μ is assumed to be known in closed-form; see Table 1 in Briol et al. [2019] or the `ProbNum` package [Wenger et al., 2021] for pairs of kernels and distributions. Further discussions on obtain closed-form expressions is provided in Section 6.1.

The posterior mean provides a natural point estimate for I. The mean is a quadrature rule (sometimes called cubature rule when $d > 1$) since it is an affine combination of function evaluations. We note that a significant advantage of BQ is that it does not impose restriction on how $x_{1:N}$ is selected, and as such doesn't require independent realisations from $\mathbb{P}$. In fact, a number of active learning approaches have been proposed, see Gunter et al. [2014], Gessner et al. [2020].

The posterior variance gives us a notion of epistemic uncertainty for I, where the uncertainty is 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^2_{BQ} to be 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. A detailed discussion on hyperparameters selection is provided in Section 6.2.

The convergence rate of the BQ estimator has been studied extensively [Briol et al., 2019, Kanagawa and Hennig, 2019, Wynne et al., 2021] and is particularly fast for low- to mid-dimensional problems where the integrand is smooth. This has to be contrasted with the computational cost, which is inherited from GP regression and is $\mathcal{O}(N^3)$ (due to matrix inversion). 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 [Kersting and Hennig, 2016], variational inference [Acerbi, 2018] and simulator-based inference [Bharti et al., 2023] to problems in engineering and the sciences including computer graphics [Marques et al., 2013, Xi et al., 2018], cardiac modelling [Oates et al., 2017a] and tsunami modelling [Li et al., 2022b]. For cheaper problems, [Karvonen and Särkkä, 2018, Karvonen et al., 2019] and [Jagadeeswaran and Hickernell, 2019] proposed BQ methods where the computational cost is much lower, but these are applicable only with specific point sets $x_{1:N}$ and measures $\mathbb{P}$.

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. Computation of the CBQ posterior can be expressed in two stages:

- **Stage 1:** Compute $\{\hat{I}_{\mathrm{BQ}}(\theta_t), \sigma^2_{\mathrm{BQ}}(\theta_t)\}_{t=1}^T$ 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.

[Hudson: Add a constant γ to the noise, for numerical stability reason. $\gamma + \sigma^2_{BQ} = \lambda * \sigma^2_{BQ}$, and λ is increasing with T. $\lambda > 1$.] This can be summarised using the direct acyclic graph in Figure 1a, 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

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

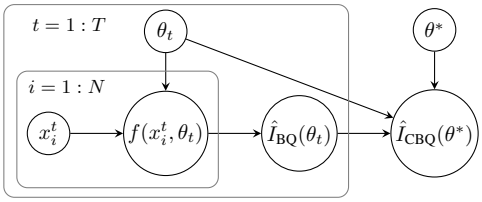
$$k_{\mathrm{CBQ}}(\theta, \theta') := k_\Theta(\theta, \theta') - k_\Theta(\theta, \theta_{1:T})^\top (k_\Theta(\theta_{1:T}, \theta_{1:T}) + \sigma^2_{\mathrm{BQ}}(\theta_{1:T})\mathrm{Id}_T)^{-1} k_\Theta(\theta_{1:T}, \theta') \tag{3}$$

where $\hat{I}_{\mathrm{BQ}}(\theta_t)$ 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 $\hat{I}_{\mathrm{CBQ}}(\theta) := \sum_{t=1}^T \sum_{i=1}^N w_{it}^{\mathrm{CBQ}} f(x_i^t, \theta_t)$ for some weights $w_{it}^{\mathrm{CBQ}} \in \mathbb{R}$ when $m_\Theta(\theta) = 0$.

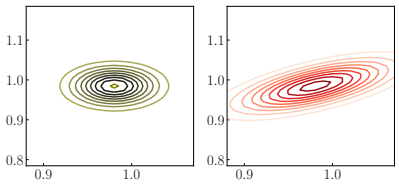
The first stage corresponds to the BQ procedure highlighted in Section 2.2: we model $f(\cdot, \theta_t)$ with independent $\mathrm{GP}(m^t_{\mathcal{X}}, k^t_{\mathcal{X}})$ priors, condition on observations $f(x^t_{1:N}, \theta_t)$, and consider the posterior distribution on $I(\theta_t)$ for all $t \in \{1, \ldots, T\}$. 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 $I(\theta_1), \ldots, I(\theta_T)$.

In the second stage, we place a $\mathrm{GP}(m_\Theta, k_\Theta)$ 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 $\sigma^2_{\mathrm{BQ}}(\theta_1), \ldots, \sigma^2_{\mathrm{BQ}}(\theta_T)$ respectively. Since the variance is input-dependent, this corresponds to heteroscedastic GP regression [Le et al., 2005]. We now briefly comment on the choice of prior and likelihood in this second stage:

- The $\mathrm{GP}(m_\Theta, k_\Theta)$ 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.



(a) Directed acyclic graph representation of CBQ. Circle nodes indicate random variables and large rectangles correspond to independent replications over indices.



(b) 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.

Figure 1: Illustration of conditional Bayesian quadrature (CBQ)

- 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. From a purely computational viewpoint, we can therefore take $N = \mathcal{O}(T^{2/3})$ without significantly increasing our computational cost. Fast GP algorithms like sparse GP [Titsias, 2009] can be used to further reduce the computational cost. Another major bottleneck for CBQ as well as for all BQ-related methods is that the closed form kernel mean embedding and initial error is required in Equation (1). We provide a detailed discussion on different ways to get around this restriction in practice in Section 6.1.

Interestingly, CBQ also provides us with a straightforward way of obtaining a joint posterior on the expectation at $\theta^*_1, \ldots, \theta^*_{T_{\mathrm{Test}}} \in \Theta$. This posterior is a multivariate Gaussian with mean vector $\hat{I}_{\mathrm{CBQ}}(\theta^*_{1:T_{\mathrm{Test}}})$ and covariance matrix $k_{\mathrm{CBQ}}(\theta^*_{1:T_{\mathrm{Test}}}, \theta^*_{1:T_{\mathrm{Test}}})$, which can be computed at an additional $\mathcal{O}(T^2 T_{\mathrm{test}})$ cost. This is illustrated in Figure 1b 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 $I(\theta_1), \ldots, I(\theta_T)$ would also be a multivariate Gaussian distribution. This approach coincides with the multi-output Bayesian quadrature approach of Xi et al. [2018], which uses multi-task GPs and was further studied by Karvonen et al. [2019], Gessner et al. [2020]. However, the computational cost is $\mathcal{O}(N^3T^3)$, 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 Chau et al. [2021], and when $T = 1$, we then recover standard BQ. Interestingly, if f does not depend on θ CBQ can b derived from another perspective of conditional kernel mean embedding detailed in Section A.

4 Theoretical Results

Our main result is the following theorem, which guarantees that CBQ is able to approximate $I(\theta)$ uniformly when T grows. To derive the result, we combine existing results on the convergence of GP

interpolation from Wynne et al. [2021], with results on importance-weighted kernel ridge regression from Gogolashvili et al. [2023]; see **??** for the proof.

Theorem 1. *Suppose $\mathcal{X} = [0,1]^d$, $\Theta = [0,1]^p$, and $\mathbb{P}_\theta$ has a density $0 < p_\theta < \infty$ on $\mathcal{X}$. Furthermore assume that $x \mapsto f(x,\theta)$ has smoothness XXX and $\theta \mapsto f(x,\theta)$ has smoothness XXX. Finally, assume that $k_\mathcal{X}$ has smoothness $\alpha > d/2$ and k_Θ has smoothness $\beta > XXX$. Then, $\exists C_1, C_2 > 0$ independent of N and T such that*

$$\sup_{\theta \in \Theta} |I(\theta) - \hat{I}_{CBQ}(\theta)| \leq (C_1 + C_2 N^{-\frac{\alpha}{d}}) T^{-\beta}.$$

The result indicates that growing N will only help up to some extent, but that growing T is essential to ensure convergence. This is intuitive since we cannot expect to approximate $I(\theta)$ uniformly simply by increasing the number of integrand evaluations 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 $N^{-\frac{\alpha}{d}}$ (due to the MC integration rate); this explains why our method will outperform these approaches when α is large relative to d.

The rate in T will itself depend on the smoothness 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 once again 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, 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. Secondly, the constraints on the smoothness of the integrand are used to guarantee that it is regular enough to be approximated at a fast rate by a GP. For simplicity, we also assume that the kernel parameters are known, but this could be straightforwardly extend to estimation in bounded sets; see **?**. We could also do misspecified smoothness following the work of Kanagawa et al. [2020] and misspecified likelihoods using the work of Wynne et al. [2021].

5 Experiments

In this section, we evaluate the empirical performance of CBQ against baseline methods including IS, KMS and LSMC. For the first three experiments in this section, we focus on the case that the integrand is independent of θ (i.e. $f(x,\theta) = f(x)$), and for the fourth experiment we focus on the case that the integrand depends on both x and θ. More detailed descriptions of the experimental settings and hyperparameter selection can be found in Section 6.2 and Section 7. The code to reproduce our experiments is available at `github.com/Anonymous65536/cbq`.

Synthetic Experiment: Bayesian Sensitivity Analysis for Linear Models Bayesian inference requires the derivation of the posterior from a prior and a likelihood. Determining the sensitivity of posterior to the hyperparameters in the prior and the likelihood is a critical step in assessing the robustness of the outcomes of Bayesian models [Oakley and O'Hagan, 2004, Kallioinen et al., 2021].

We use Bayesian linear regression as a simple toy experiment. The 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 multivariate normal distribution as the prior $p(w) = \mathcal{N}(w; 0, \bar{\Sigma}), w \in \mathbb{R}^D$ and we use Gaussian likelihood so the posterior $p(w \mid \mathcal{D}; \bar{\Sigma})$ is known to have a closed form expression. We are interested in doing the sensitivity analysis on the effect of hyperparameter $\bar{\Sigma}$ to the target of interest $I(\bar{\Sigma})$ with all other hyperparameters fixed: $I(\bar{\Sigma}) = \int f(w) p(w \mid \mathcal{D}; \bar{\Sigma}) dw$. For simplicity, here we only consider $\bar{\Sigma}$ as a diagonal matrix.

If the integrand $f(w) = w^\top w$, then $I(\bar{\Sigma})$ describes the second moment (energy) of the posterior distribution, if the integrand $f(w) = w^\top y^*$ and y^* is a new observation , then $I(\bar{\Sigma})$ describes the predictive mean. In this simple setting, the ground truth of the integral $I(\bar{\Sigma})$ can be computed analytically so we can easily compare the performance of our methods against other baseline methods. The observations $\bar{\Sigma}_{1:T}$ are sampled from a uniform distribution on each entry, i.e $\bar{\Sigma}_{1:T} \sim [1,3]^d$. For each $\bar{\Sigma}_t$, N observations are sampled from the posterior $w^t_{1:N} \sim p(w \mid \mathcal{D}, \bar{\Sigma}_t)$. In total, we

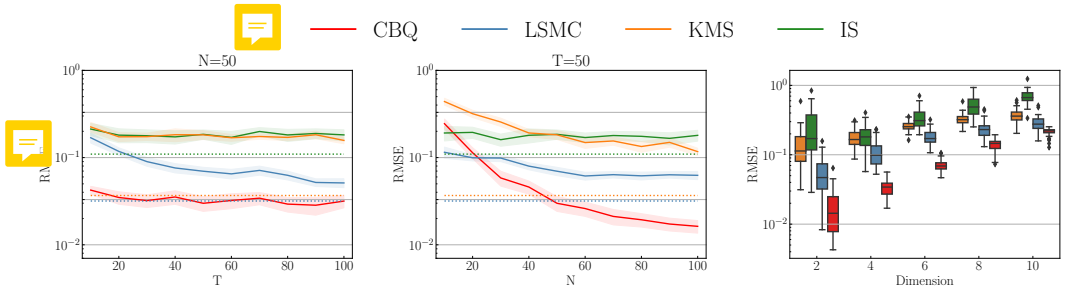


Figure 2: Experimental results of Bayesian sensitivity analysis. **Left:** RMSE of all methods under dimension $d = 2$ with fixed $N = 50$ and increasing T. **Middle:** RMSE of all methods under dimension $d = 2$ with fixed $T = 50$ and increasing N. **Right:** RMSE of all methods with increasing dimension d and fixed $N = T = 100$.

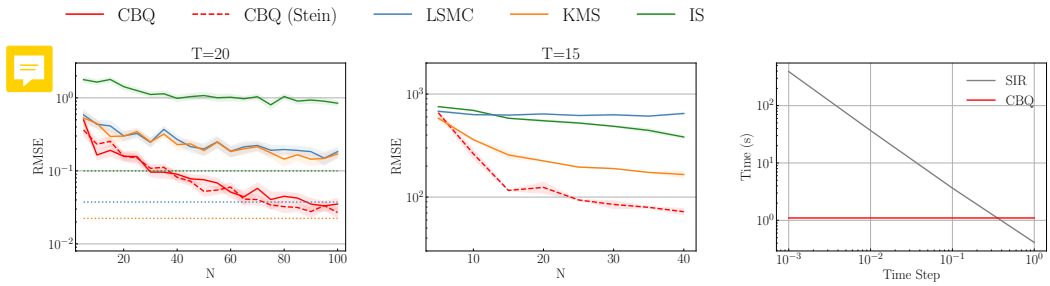


Figure 3: Experimental results for Black-Scholes and SIR. **Left:** RMSE of all methods with fixed $T = 20$ and increasing N. **Middle:** RMSE of all methods with fixed $T = 15$ and increasing N. **Right:** The computational cost (in wall clock time) for CBQ and for obtaining one single numerical solution from SIR. In practice, the process of obtaining samples from SIR equations is repeated $N \times T$ times.

have $N \times T$ samples. We show the results of $f(w) = w^\top w$ in Figure 2 and we show the results of $f(w) = w^\top y^*$ in Appendix 7.1. For CBQ, $k_{\mathcal{X}}$ is selected to be Gaussian Radial Basis Function (RBF) so that the kernel mean embedding μ has a closed form and k_Θ is selected to be Matern-3/2. We provide an ablation study on using different kernels in Section 7.1.5.

In Figure 2, we show the performance of our method in terms of root mean squared error (RMSE) against other baseline methods with varying N, T and dimension d. In the left and middle plots of Figure 2, we can see that our method clearly outperforms other baseline methods. And specifically for CBQ, the rate of convergence is much faster in N than in T, which confirms the intuition in the theory that the error bound has a faster convergence rate in N than in T. In the left and middle panels of Figure 2, we also show in dotted line the performance of baseline methods under very large number of samples $N = T = 1000$. The performance of CBQ is comparable or even better under much smaller sample size.

In the right-most panel of Figure 2, we report the performance of all methods when the underlying dimension d increases. It shows that the performance of other methods gradually catch up with CBQ as dimension d increases, mainly because in high dimensions BQ does not work very well since Gaussian process tend to suffer in high dimensions. We also provide a study on the calibration of the uncertainty provided by CBQ in Section 7.1.6.

Butterfly Call Option with the Black-Scholes Model Conditional expectations are common in risk management. For example, the payoff of a financial asset at time η can be expressed as $\mathbb{E}[\psi(S_\zeta) \mid S_\eta]$ where ζ is the expiration time. In this experiment, we consider specifically an asset whose price S_η at time η follows the Black-Scholes formula $S_\eta = S_0 \exp(\sigma W_\eta - \sigma^2 \eta/2)$ for $\eta \geq 0$, with σ being the underlying volatility and W being 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

$$\psi(S_\zeta) = \max(S_\zeta - K_1, 0) + \max(S_\zeta - K_2, 0) - 2\max\left(S_\zeta - \frac{K_1 + K_2}{2}, 0\right)$$

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 Alfonsi et al. [2021, 2022] assuming that a shock occur at time η that multiplies the price of the butterfly option by $1+s$, so the expected loss caused by the shock can be expressed as $\mathcal{L} = \mathbb{E}\Big[\max\left(\mathbb{E}\left[\psi(S_\zeta) - \psi((1+s)S_\zeta] \mid S_\eta\right), 0\Big]$. 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 $\{S_\eta\}_{1:T}$ and $\{S_\zeta\}_{1:N}^t$ are sampled from the Black-Scholes formula. In total, we have $N \times T$ samples. The ground truth $\mathcal{L}$ has a closed form so we can easily compare the performance of our methods against other baseline methods. For CBQ, $k_{\mathcal{X}}$ is chosen as a logarithmic Radial Basis Function (RBF) kernel, thereby ensuring that the kernel mean embedding possesses a closed form solution, as outlined in Section 7.2. A Stein kernel, employing a Matern-3/2 kernel as the base, is also utilized as $k_{\mathcal{X}}$. The choice of k_Θ is a Matern-3/2 kernel.

In the left-most panel of Figure 3, we can see that both CBQ with log RBF kernel and CBQ with Stein kernel achieve much lower RMSE compared against other baselines. Similar to Bayesian sensitivity analysis, the CBQ exhibits performance that is comparable, if not superior, to baseline methodologies when dealing with a substantial sample size of $N = T = 1000$. Further experimental outcomes and calibration results are available in Section 7.2..

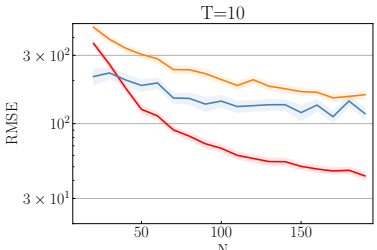
Bayesian Sensitivity Analysis for Susceptible-Infectious-Recovered (SIR) Model The SIR model is commonly used to simulate the dynamics of infectious diseases through a population. The dynamics are governed by a system of ordinary differential equations (ODE) parametrised by a positive infection rate $\beta > 0$ and a recovery rate $\gamma > 0$ (as detailed in Section 7.3). The precision of the numerical solution to this system typically hinges on the selection of a step size parameter, which guides the discretization process. While smaller step sizes yield more precise solutions, they are also associated with a higher computational cost. For example, employing a step size of 0.1 days for simulating a 150-day period would necessitate a computation time of 7 seconds for generating a solitary data sample. The computational cost would become even larger as the step size gets lower, as depicted in the right-most panel of Figure 3. Consequently, there is a clear necessity for a more data-efficient algorithm in such circumstances.

We assume that the infection rate β follows a gamma prior distribution $\beta \sim \mathrm{Gamma}(\beta; \bar{\beta}, \xi)$ where $\bar{\beta}$ 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 of the initial belief. The target of interest is the expected peak number of infected individuals under the prior distribution on β: $I_{\text{max}}(\bar{\beta}) = \mathbb{E}_\beta\left[\max_r I_r(\beta) \mid \bar{\beta}\right]$. I_r is the solution to SIR ODEs which represents the number of infections at day r. It is always important to know how different initial estimate (different $\bar{\beta}$) 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 the shape parameter $\bar{\beta} \in [2, 9]$. The total population is set to be 10^6. The observations $\bar{\beta}_{1:T}$ are sampled from the uniform distribution $\mathrm{U}[2, 9]$ and then N observations of $\beta_{1:N}^t$ are sampled from the gamma distribution $\mathrm{Gamma}(\beta; \bar{\beta}_t, \xi)$ for a fixed t. In total, we have $N \times T$ observations. As we do not have the ground truth in closed form in this experiment, we use the average of 5000 Monte Carlo simulations as the pseudo ground truth and evaluate the RMSE across different methods. For CBQ, we employ a Stein kernel for $k_{\mathcal{X}}$, with the Matern-3/2 kernel as the base kernel and k_Θ is selected to be Matern-3/2 kernel.

We can see in the middle panel of Figure 3 that CBQ clearly outperforms other baseline methods. Although the CBQ estimator exhibits a higher computational complexity compared to baseline methods, we demonstrate in the right-most panel of Figure 3 that, due to the increased computational expense of obtaining samples when the step size is set to be very small, using CBQ is ultimately more efficient overall within the same period of time. More experimental results can be found in Section 7.3. For future work, it would be interesting to find out the potential performance gain by applying CBQ to more complicated computer simulations like cardiac modelling [Oates et al., 2017a] and tsunami modelling [Li et al., 2022a].

Uncertainty Decision Making in Health Economics In the medical world, it is important to compare the cost and the relative advantage of conducting an extra medical ex-



periment. The expected value of partial perfect information (EVPPI) quantifies the average advantage gained from conducting extra experiment to obtain precise knowledge of the uncertain variables θ. EVPPI can be expressed as $\mathbb{E}[\max_c \mathbb{E}[f_c(X,\theta) \mid \theta]] - \max_c \mathbb{E}[f_c(X,\theta)]$ where $c \in \mathcal{C}$ is a set of potential treatments. A more detailed explanation of EVPPI can be found in Section 7.4.

We adopt the same experimental setup as delineated in Giles and Goda [2019], Brennan et al. [2007], wherein $X \cup \theta = (X_1, \cdots, X_{19})$ comprise 19 independent Gaussian random variables. Detailed information on the practical interpretations, as well as the joint mean and covariance of these Gaussian random variables, can be found in Section 7.4. The problem under investigation is a binary decision-making problem ($\mathcal{C} = 1, 2$), characterized by $f_1(X,\theta) = \lambda\left(X_5X_6X_7 + X_8X_9X_{10}\right) - \left(X_1 + X_2X_3X_4\right)$ and $f_2(X,\theta) = \lambda\left(X_{14}X_{15}X_{16} + X_{17}X_{18}X_{19}\right) - \left(X_{11} + X_{12}X_{13}X_4\right)$, with $\lambda = 10^4$.

In this experiment, we select $\theta = (X_5, X_{14})$. So the observations $\theta_{1:T}$ are sampled from the marginal distribution $p(X_5, X_{14})$ and then N observations $x^t_{1:N}$ are sampled from the conditional distribution $p(X \mid \theta_t)$ for a fixed t. In total, we have $N \times T$ observations. We draw 10^7 samples of (X, θ) to generate a pseudo ground truth and evaluate the RMSE across different methods.

Note that importance sampling (IS) is no longer applicable here because the integrand f depends on both X and θ, so we are only comparing against KMS and LSMC. For CBQ, we select Matern-3/2 for $k_{\mathcal{X}}$ and also Matern-3/2 for k_Θ. In Figure 4, we can see that CBQ outperforms other baseline methods with much lower RMSE. More experimental results can be found in Section 7.4.

6 Conclusions

In this study, we ~~have~~ proposed a novel algorithm, the Conditional Bayesian Quadrature (CBQ), tailored for the computation of conditional expectations, espeically in the area where sample acquisition is expensive or the evaluation of the integrand is costly. The CBQ algorithm is grounded on a two-tiered hierarchical strategy: the primary stage provides a BQ approximation for $I(\theta_1), \cdots, I(\theta_t)$, while the subsequent stage employs a heteroscedastic Gaussian Process (GP) regression based on the means and variances from the initial stage. Our methodology, as corroborated both theoretically and empirically, exhibits an accelerated convergence rate and provides the additional benefit of uncertainty quantification.

Looking forward, we are excited about the possibility of applying CBQ to more complex computer simulation tasks. Additionally, the extension of CBQ to nested expectation seems relatively straightforward. The comparison between convergence rate of CBQ and nested Monte Carlo methods suggests another promising venue for exploration in the future.

Ackowledgments: The authors would like to thank Motonobu Kanagawa for some helpful pointers to the literature on kernel ridge regression. [Hudson: Comment out before submission.]

[FXB: The references need to be tidied - see the "paper writing guide" on our Teams channel] [Hudson: Discussion]

References

L. Acerbi. Variational Bayesian Monte Carlo. In *Neural Information Processing Systems*, pages 8223–8233, 2018.

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.

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.

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

A. Bharti, M. Naslidnyk, O. Key, S. Kaski, and F-X. Briol. Optimally-weighted estimators of the maximum mean discrepancy for likelihood-free inference. *arXiv:2301.11674*, 2023.

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

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.

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.

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

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

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

J. Cockayne, C. Oates, T. Sullivan, and M. Girolami. Bayesian probabilistic numerical methods. *SIAM Review*, 61(4):756–789, 2019.

J. Demange-Chryst, F. Bachoc, and J. Morio. Efficient estimation of multiple expectations with the same sample by adaptive importance sampling and control variates. *arXiv:2212.00568*, 2022.

P. Diaconis. Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV*, pages 163–175, 1988.

Seth Flaxman, Dino Sejdinovic, John P Cunningham, and Sarah Filippi. Bayesian learning of kernel embeddings. *arXiv preprint arXiv:1603.02160*, 2016.

David Heaver Fremlin. *Measure theory*, volume 4. Torres Fremlin, 2000.

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

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

M. B. Giles, T. Nagapetyan, and K. Ritter. Multilevel monte carlo approximation of distribution functions and densities. *SIAM-ASA Journal on Uncertainty Quantification*, 3:267–295, 2015. doi: 6.

Michael B. Giles and Takashi Goda. Decision-making under uncertainty: using MLMC for efficient estimation of EVPPI. *Statistics and Computing*, 29(4):739–751, 2019. ISSN 15731375. doi: 10.1007/s11222-018-9835-1.

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

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.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors-supplementary. *arXiv preprint arXiv:1205.4656*, 2012.

T. Gunter, R. Garnett, M. Osborne, P. Hennig, and S. Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Advances in Neural Information Processing Systems*, pages 2789–2797, 2014.

A. Heath, I. Manolopoulou, and G. Baio. A Review of Methods for Analysis of the Expected Value of Information. *Medical Decision Making*, 37(7):747–758, 2017.

P. Hennig, M. A. Osborne, and H. Kersting. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, 2022.

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.

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

L. J. Hong and S. Juneja. Estimating the mean of a non-linear function of conditional expectation. In *Proceedings of the 2009 Winter Simulation Conference*, pages 1223–1236, 2009. ISBN 2013206534.

R. Jagadeeswaran and F. J. Hickernell. Fast automatic Bayesian cubature using lattice sampling. *Statistics and Computing*, 29(6):1215–1229, 2019.

N. Kallioinen, T. Paananen, P.-C. Bürkner, and A. Vehtari. Detecting and diagnosing prior and likelihood sensitivity with power-scaling. *arXiv:2107.14054*, 2021.

M. Kanagawa and P. Hennig. Convergence guarantees for adaptive Bayesian quadrature methods. In *Neural Information Processing Systems*, pages 6237–6248, 2019.

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.

T. Karvonen and S. Särkkä. Fully symmetric kernel quadrature. *SIAM Journal on Scientific Computing*, 40(2):697–720, 2018.

T. Karvonen, S. Särkkä, and C. J. Oates. Symmetry exploits for Bayesian cubature methods. *Statistics and Computing*, 29:1231–1248, 2019.

H. Kersting and P. Hennig. Active uncertainty calibration in Bayesian ODE solvers. In *Uncertainty in Artificial Intelligence*, pages 309–318, 2016.

S. Krumscheid and F. Nobile. Multilevel monte carlo approximation of functions. *SIAM-ASA Journal on Uncertainty Quantification*, 6(3):1256–1293, 2018. ISSN 21662525. doi: 10.1137/17M1135566.

Q. V. Le, A. J. Smola, and S. Canu. Heteroscedastic Gaussian Process regression. In *International Conference on Machine Learning*, pages 489–496, 2005.

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.

Kaiyu Li, Daniel Giles, Toni Karvonen, Serge Guillas, and François-Xavier Briol. Multilevel bayesian quadrature. *arXiv preprint arXiv:2210.08329*, 2022a.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. *arXiv preprint arXiv:2208.01711*, 2022b.

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.

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. ISSN 19411383. doi: 10.1146/annurev-economics-111809-125134.

N. Madras and M. Piccioni. Importance sampling for families of distributions. *The Annals of Applied Probability*, 9(4):1202–1225, 1999.

R. Marques, C. Bouville, M. Ribardiere, P. Santos, and K. 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.

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

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.

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.

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

C. J. Oates and T. J. Sullivan. A modern retrospective on probabilistic numerics. *Statistics and Computing*, 29:1335–1351, 2019.

C. J. Oates, S. Niederer, A. Lee, F-X. Briol, and M. Girolami. Probabilistic models for integration error in the assessment of functional cardiac models. In *Neural Information Processing Systems*, pages 110–118, 2017a.

Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for monte carlo integration. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 695–718, 2017b.

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

T. Rainforth, R. Cornish, H. Yang, A. Warrington, and F. Wood. On nesting Monte Carlo estimators. In *International Conference on Machine Learning*, volume 10, pages 6789–6817, 2018. ISBN 9781510867963.

C. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 489–496, 2002.

CE. Rasmussen and CKI. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2000.

I. M. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55:271–280, 2001.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35:363–417, 2012.

Z. Sun, A. Barp, and F-X. Briol. Vector-valued control variates. *arXiv:2109.08944*, 2021.

Z. Sun, C. J. Oates, and F-X. Briol. Meta-learning control variates: Variance reduction with limited data. *arXiv:2303.04756*, 2023.

X. Tang. *Importance sampling for efficient parametric simulation*. PhD thesis, Boston University, 2013.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009.

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. doi: 10.1038/s41592-019-0686-2.

J. Wenger, N. Krämer, M. Pförtner, J. Schmidt, N. Bosch, N. Effenberger, J. Zenn, A. Gessner, T. Karvonen, F-X Briol, M. Mahsereci, and P. Hennig. ProbNum: Probabilistic numerics in Python. *arXiv:2112.02100*, 2021. URL `http://arxiv.org/abs/2112.02100`.

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.

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

Supplementary Material

Table of Contents

A CBQ from the Perspective of Conditional Kernel Mean Embedding **15**

B Convergence rate **17**

2.1 Technical Assumptions 18

2.2 Convergence 18

3 Attempt at a better q TODO delete pre submission **21**

4 Convergence rate—OLD TODO delete pre submission **22**

4.1 Technical Assumptions 23

4.2 Convergence 24

5 Notation (TODO delete pre submission) **25**

6 Practical Considerations **25**

6.1 Tractable Kernel Means 25

6.2 Hyperparameter Selection 26

7 Experiments **26**

7.1 Synthetic Experiment: Bayesian Sensitivity Analysis for Linear Models 26

7.2 Butterfly Call Option with the Black-Scholes Model 30

7.3 Bayesian Sensitivity for a Susceptible-Infectious-Recovered (SIR) Model 31

7.4 Uncertainty Decision Making 32

7.5 Experimental Settings 32

Appendix A CBQ from the Perspective of Conditional Kernel Mean Embedding

In the main text, CBQ is mainly derived from the perspective of probabilistic numerics and Gaussian process regression is the main probabilistic tool used in the derivation. In this section, we offer a distinct perspective, viewing CBQ through a frequentist approach. This alternative viewpoint reveals an intriguing connection to the concept of conditional kernel mean embedding, thus expanding our understanding of the CBQ method.

Recall that we have assumed two positive definite kernels $k_\Theta : \Theta \times \Theta \to \mathbb{R}$ and $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with corresponding reproducing kernel Hilbert spaces $\mathcal{H}_\Theta$ and $\mathcal{H}_\mathcal{X}$. The notation Θ and $\mathcal{X}$ is quite rare in the literature of kernel mean embedding, but we stick to this notation to keep consistency with the main text. $\mathcal{U}_{X|\theta} \in \mathcal{H}_\mathcal{X}$ is the conditional mean embedding of conditional distribution $P(X \mid \Theta = \theta)$ such that the conditional expectation can be written in the form of inner product:

$$\mathbb{E}[f(X) \mid \Theta = \theta] = \int_\mathcal{X} f(x) p(x \mid \theta) = \left\langle f, \mathcal{U}_{X|\theta} \right\rangle_{\mathcal{H}_\mathcal{X}} \tag{A.1}$$

Note that in the main text, a more general form of function f is allowed that depends on both θ and x and the distribution $\mathbb{P}_\theta$ has a more flexible parametric form. In this section, we only focus on the setting that the function f depends solely on x and $\mathbb{P}_\theta$ is a conditional distribution with density $p(x \mid \theta)$.

Since for every $\theta \in \Theta$, there exists an embedding $\mathcal{U}_{X|\theta}$, so we are interested in finding an operator $\mathcal{U}_{X|\Theta} : \Theta \to \mathcal{H}_\mathcal{X}$ such that

$$\mathcal{U}_{X|\theta} = \mathcal{U}_{X|\Theta} k_\Theta(\theta, \cdot)$$

$\mathcal{U}_{X|\Theta}$ is the conditional mean embedding of conditional distribution $P(X \mid \Theta)$ and also belongs to the product reproducing kernel Hilbert space: $\mathcal{U}_{X|\Theta} \in \mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\Theta$. If we find a good approximation of $\mathcal{U}_{X|\Theta}$, then the conditional expectation can be written as

$$\mathbb{E}[f(X) \mid \Theta = \theta] = \left\langle f, \mathcal{U}_{X|\theta} \right\rangle_{\mathcal{H}_\mathcal{X}} = \left\langle f, \mathcal{U}_{X|\Theta} k_\Theta(\theta, \cdot) \right\rangle_{\mathcal{H}_\mathcal{X}}$$

A more detailed discussion can be found in Muandet et al. [2016].

The finding of the optimal approximating operator $F \in \mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\Theta$ can be written as the minimizer of the following objective [Grünewälder et al., 2012].

$$\mathcal{E}[F] = \sup_{\|f\|_{\mathcal{H}_\mathcal{X}} \leq 1} \mathbb{E}_\Theta \left[\left(\mathbb{E}_X[f(X) \mid \Theta] - \langle f, F(X) \rangle_{\mathcal{H}_\mathcal{X}} \right)^2 \right] \tag{A.2}$$

which can be further upper bounded by

$$\begin{aligned} \mathcal{E}[F] &= \sup_{\|f\|_{\mathcal{H}_\mathcal{X}} \leq 1} \mathbb{E}_\Theta \left[\left\langle f, \mathbb{E}_X \left[k_\mathcal{X}(X, \cdot) \mid \Theta \right] - F(\Theta) \right\rangle_{\mathcal{H}_\mathcal{X}}^2 \right] \\ &\leq \sup_{\|f\|_{\mathcal{H}_\mathcal{X}} \leq 1} \|f\|_{\mathcal{H}_\mathcal{X}}^2 \, \mathbb{E}_\Theta \left\| \mathbb{E}_X \left[k_\mathcal{X}(X, \cdot) \mid \Theta \right] - F(\Theta) \right\|_{\mathcal{H}_\mathcal{X}}^2 \\ &= \mathbb{E}_\Theta \left\| \mathbb{E}_X \left[k_\mathcal{X}(X, \cdot) \mid \Theta \right] - F(\Theta) \right\|_{\mathcal{H}_\mathcal{X}}^2 \\ &\leq \mathbb{E}_{\Theta, X} \left[\left\| k_\mathcal{X}(X, \cdot) - F(\Theta) \right\|_{\mathcal{H}_\mathcal{X}}^2 \right] \end{aligned} \tag{A.3}$$

Normally, in the literature of conditional kernel mean embedding, only sample pairs $\{x_{1:T}, \theta_{1:T}\}$ are observed and the objective above is replaced by the empirical estimate with an extra regularization [Grünewälder et al., 2012, Li et al., 2022b].

$$\mathcal{E}_{\text{original}}[F] = \sum_{t=1}^{T} \left\| k_\mathcal{X}(x_t, \cdot) - F(\theta_t) \right\|_{\mathcal{H}_\mathcal{X}}^2 + \lambda \|F\|_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\Theta}^2 \tag{A.4}$$

We use the term 'original' to indicate that this is the objective that is commonly used in the literature. Later on in the section we are going to show other different objectives $\mathcal{E}$ with different subscripts.

With the aid of the Riesz representer theorem, the minimizer to (A.4) has a closed form expression. The derivations can be found in Grünewälder et al. [2012], Li et al. [2022b].

$$F_{\text{original}}(\theta)(\cdot) = k_\Theta(\theta, \theta_{1:T})(k_\Theta(\theta_{1:T}, \theta_{1:T}) + T\lambda I)^{-1} k_\mathcal{X}(x_{1:T}, \cdot) \tag{A.5}$$

And therefore, the conditional expectation in Equation (A.1) can be expressed as

$$\begin{aligned} \mathbb{E}\left[f(\widehat{Y) \mid \Theta} = \theta\right]_{\text{original}} &= \langle f, F_{\text{original}}(\theta)\rangle_{\mathcal{H}_\mathcal{X}} \\ &= k_\Theta(\theta, \theta_{1:T})(k_\Theta(\theta_{1:T}, \theta_{1:T}) + T\lambda I)^{-1} f(x_{1:T}) \end{aligned} \tag{A.6}$$

However, we have mentioned in the main text that in our setting we observe N samples of x from the conditional distribution $P(X \mid \Theta = \theta_t)$ for any given t. Recall from Section 2 that our observations are:

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

As a result, the original objective $\mathcal{E}_{\text{original}}$ requires some minor modifications to adapt to our setting, so we come up with a new objective

$$\mathcal{E}_{\text{multiple}}[F] = \sum_{t=1}^{T} \|\mathbb{E}_X\left[k_\mathcal{X}(X, \cdot) \mid \Theta = \theta_t\right] - F(\theta_t)\|^2_{\mathcal{H}_\mathcal{X}} + \lambda \|F\|^2_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\Theta} \tag{A.7}$$

The new objective $\mathcal{E}_{\text{multiple}}[F]$ differs from $\mathcal{E}_{\text{original}}[F]$ only by the inner expectation $\mathbb{E}_X\left[k_\mathcal{X}(X, \cdot) \mid \Theta = \theta_t\right]$. In the new objective $\mathcal{E}_{\text{multiple}}[F]$, the inner expectation is kept till later. With the same derivation as above, the minimizer to (A.7) has a closed form expression

$$F_{\text{multiple}}(\theta)(\cdot) = k_\Theta(\theta, \theta_{1:T})(k_\Theta(\theta_{1:T}, \theta_{1:T}) + T\lambda I)^{-1} \begin{bmatrix} \mathbb{E}_X\left[k_\mathcal{X}(X, \cdot) \mid \Theta = \theta_1\right] \\ \vdots \\ \mathbb{E}_X\left[k_\mathcal{X}(X, \cdot) \mid \Theta = \theta_T\right] \end{bmatrix} \tag{A.8}$$

And then we take the inner product to have

$$\begin{aligned} \mathbb{E}\left[f(\widehat{Y) \mid \Theta} = \theta\right]_{\text{multiple}} &= \langle f, F_{\text{multiple}}(\theta)\rangle_{\mathcal{H}_\mathcal{X}} \\ &= k_\Theta(\theta, \theta_{1:T})(k_\Theta(\theta_{1:T}, \theta_{1:T}) + T\lambda I)^{-1} \begin{bmatrix} \mathbb{E}_X\left[f(X) \mid \Theta = \theta_1\right] \\ \vdots \\ \mathbb{E}_X\left[f(X) \mid \Theta = \theta_T\right] \end{bmatrix} \end{aligned} \tag{A.9}$$

If we compare Equation (A.9) with Equation (A.5), we can see that in Equation (A.5) a one sample Monte Carlo estimate is applied to approximate the conditional expectations $\mathbb{E}_X\left[f(X) \mid \Theta = \theta_t\right]$ due to the limitation that only one sample observed from the conditional distribution $P(X \mid \Theta = \theta_t)$. Since now we have multiple samples $x^t_{1:n}$ available, the first thing to come to mind is to use again Monte Carlo to take an empirical average over function evaluations, which gives us

$$\mathbb{E}\left[f(\widehat{Y) \mid \Theta} = \theta\right]_{\text{shrinkage}} = k_\Theta(\theta, \theta_{1:T})(k_\Theta(\theta_{1:T}, \theta_{1:T}) + T\lambda I)^{-1} \begin{bmatrix} \frac{1}{N}\mathbf{1}^\top f(x^1_{1:N}) \\ \vdots \\ \frac{1}{N}\mathbf{1}^\top f(x^T_{1:N}) \end{bmatrix} \tag{A.10}$$

where $\mathbf{1}$ is a vector of ones.

This approach is known as the kernel mean shrinkage estimator [Muandet et al., 2016] and the application of it has been considered in Chau et al. [2021] . It also has a Bayesian interpretation in Flaxman et al. [2016].

Our approach CBQ proposed in the main text uses Bayesian quadrature to estimate $\mathbb{E}_X\left[f(X) \mid \Theta = \theta_i\right]$ instead of simply averaging the function evaluations. In Section 2.2, if the GP prior on f is chosen to have zero mean, then the BQ estimate is $\mathbb{E}_X\left[f(\widehat{X) \mid \Theta} = \theta_t\right]_{\text{BQ}} = \mu_{\theta_t}(x^t_{1:N})^\top k_\mathcal{X}(x^t_{1:N}, x^t_{1:N})^{-1} f(x^t_{1:N})$.

And aftering plugging in the BQ estimate of $\mathbb{E}_X\left[f(X) \mid \Theta = \theta_t\right]$ into Equation (A.8)

$$\mathbb{E}\left[f(\widehat{Y) \mid \Theta} = \theta\right]_{\text{CBQ}} = k_\Theta(\theta, \theta_{1:T})(k_\Theta(\theta_{1:T}, \theta_{1:T}) + T\lambda I)^{-1} \begin{bmatrix} \mu_{\theta_1}(x^1_{1:N})^\top k_{\mathcal{X}}(x^1_{1:N}, x^t_{1:N})^{-1} f(x^t_{1:N}) \\ \vdots \\ \mu_{\theta_T}(x^T_{1:N})^\top k_{\mathcal{X}}(x^T_{1:N}, x^T_{1:N})^{-1} f(x^T_{1:N}) \end{bmatrix} \tag{A.11}$$

Now Equation (A.11) has the same form as the mean estimate in Equation (2) in the main text, and the only difference is that the regularization constant λ here is replaced by the diagonal heteroskedastic noise $\sigma^2_{BQ}(\theta_{1:T})\text{Id}_T$ in the main text. So far from a second perspective, we have developed the conditional Bayesian quadrature estimator. However, since this is a frequentist perspective, it is not able to provide uncertainty quantification.

Appendix B Convergence rate

Recall the CBQ estimator proposed in (2),

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

where $\gamma_T > 0$, $\hat{I}_{\text{BQ}}(\theta_t)$ and $\sigma^2_{\text{BQ}}(\theta_t)$, for $t \in \{1, \ldots, T\}$, are BQ posterior mean and variance obtained in the first stage as given in (1)

$$\begin{aligned} \hat{I}_{\text{BQ}}(\theta_t) &= \mu_\theta(x^t_{1:m})^\top k_\Theta(x^t_{1:m}, x^t_{1:m})^{-1} f(x^t_{1:m}, \theta_t), \\ \sigma^2_{\text{BQ}}(\theta_t) &= \mathbb{E}_{X,X' \sim \mathbb{P}_\theta}[k_\Theta(X, X')] - \mu_\theta(x^t_{1:m})^\top k_\Theta(x^t_{1:m}, x^t_{1:m})^{-1} \mu_\theta(x^t_{1:m}). \end{aligned}$$

It was pointed out in Gogolashvili et al. [2023, Remark 2], (and can be seen through straightforward differentiation) that the estimator $\hat{I}_{\text{CBQ}}(\theta)$ is the minimiser of the importance weighted kernel ridge regression loss over functions in the RKHS $\mathcal{H}_\Theta$,

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

for any $\tau > 0$. Suppose θ_i were sampled from a probability measure $\mathbb{P}_{\text{tr}}$ on Θ. Then,

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

defines a positive measure on Θ for any positive $w(\theta) > 0$ for which the integral exists [Fremlin, 2000, 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 $w(\theta_t) = \tau(1 + \gamma_T^{-1} \sigma^2_{\text{BQ}}(\theta_t))^{-1}$. Then, since $\mathbb{E}[\hat{I}_{\text{BQ}}(\theta_i)] = I(\theta_i)$, this (TODO proper reference) loss can be considered an unbiased finite-sample approximation of

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

Under a further assumption that the problem is well-specified, meaning $I(\theta) \in \mathcal{H}_\Theta$, an upper bound on the rate of convergence of $\hat{I}_{\text{CBQ}}(\theta)$ to $I(\theta)$ as $n \to \infty$ was established in Gogolashvili et al. [2023, Theorem 4]. Specifically, [TODO summarise once it's more clear.]

To apply the result, we define $w(\theta)$ of convenient form that satisfies the requirements mentioned above, specifically $w(\theta) \in (0, A]$ for some $A < \infty$ and any $\theta \in \Theta$, and $w(\theta_t) = \tau(1 + \gamma_T^{-1} \sigma^2_{\text{BQ}}(\theta_t))^{-1}$ for some $t \in \{0, \ldots, T\}$.[1] Take $t' = \arg\min_{t \in \{0,\ldots,T\}} \{\sigma^{-2}_{\text{BQ}}(\theta_t)\} > 0$, and define

(12) for

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

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

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

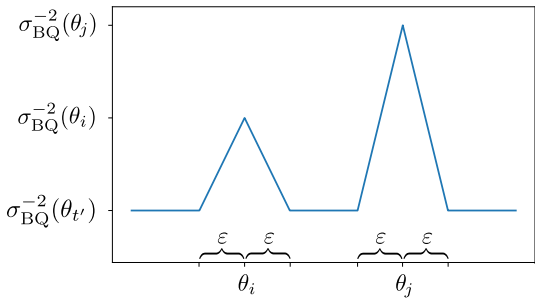


Figure 5: Illustration of $w(\theta)$ for $\Theta \subset \mathbb{R}$

TODO change figure. or remove It is easy to see that $w(\theta)$ is bounded above by $\tau \max_{t\in\{0,\dots,T\}}(1 + \gamma_T^{-1}\sigma_{\mathrm{BQ}}^2(\theta_t))^{-1} < \tau$, and below by $\tau(1 + \gamma_T^{-1}\sigma_{\mathrm{BQ}}^2(\theta_{t'}))^{-1} > 0$ for any $\theta \in \Theta$, and $w(\theta_t) = \tau(1 + \gamma_T^{-1}\sigma_{\mathrm{BQ}}^2(\theta_t))^{-1}$ as required.

2.1 Technical Assumptions

- $I(\theta)$ lies in the Sobolev space $\mathcal{W}^{2,s}(\Theta)$.
- k_Θ is a Matérn kernel of order ν such that $s \geq \nu + d/2$
- $\gamma_T = cT^\alpha$, for $c > 0$ and $\alpha \in (0, 1)$.
- $|I(\theta)| \leq M$, for all $\theta \in \Theta$
- θ_t were sampled i.i.d. from some $\mathbb{P}_{\mathrm{tr}}$, and

2.2 Convergence

Lemma 1 (Assumption 2 in Gogolashvili et al. [2023])**.** *Under technical assumptions in TODO, $I(\theta) = L^r g$ for $r = \min\{s/(2\nu + d), 1\}$ and some $g \in \mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{te}})$ of norm $R = \|g\|_{\mathcal{L}^2(\Theta,\mathbb{P}_{\mathrm{te}})}$ for some $R > 0$.*

Proof. By Steinwart and Scovel [2012, Theorem 4.6], the range of the integral operator $\mathrm{ran}L^r$ coincides with the $2r$-interpolation space between $\mathcal{H}_\Theta$ and $\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{te}})$. Therefore, we ought to show that $I(\theta)$ lies in this $2r$-interpolation space. First, we note that $\mathcal{H}_\Theta \simeq \mathcal{W}^{2,\nu+d/2}(\Theta)$, and given condition TODO and the boundedness of $w(x)$, $\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{te}})$ is norm-equivalent to $\mathcal{L}^2(\Theta)$. This implies norm-equivalence of the interpolations spaces, therefore we will be establishing if $I(\theta)$ lies in this $2r$-interpolation space between $\mathcal{W}^{2,\nu+d/2}(\Theta)$ and $\mathcal{L}^2(\Theta)$.

The $2r$-interpolation space between $\mathcal{W}^{2,\nu-d/2}(\Theta)$ and $\mathcal{L}^2(\Theta)$ is the space $\mathcal{W}^{2,2r(\nu-d/2)}(\Theta)$ (TODO ref?). Since $I(\theta)$ lies in $\mathcal{W}^{2,s}$, by inclusion of Sobolev spaces we have that it lies in $\mathcal{W}^{2,s_0}$ for any $s_0 \leq s$. Therefore, $I(\theta)$ lies in the aforementioned interpolation space whenever $2r(\nu + d/2) \leq s$, meaning $r \leq s/(2\nu + d)$. The result follows. □

Note that, under Assumption TODO, $s/(2\nu + d) \geq 1/2$, and we have $r \in [1/2, 1]$, as is required by Assumption 2 in Gogolashvili et al. [2023].

Lemma 2 (Assumption 3 in Gogolashvili et al. [2023])**.** *Under technical assumptions in TODO, for $q = 1$, $W = \tau$, and $\sigma^2 = \|\Theta\|\tau$ it holds for all $m \in \mathbb{N}$, $m \geq 2$, that*

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

Proof. By definition of $w(\theta)$,

$$\int_\Theta w(\theta)^m \mathrm{d}\theta < \|\Theta\|\tau \max_t (1 + \gamma_T^{-1}\sigma_{\mathrm{BQ}}^2(\theta_t))^{-m} < \|\Theta\|\tau, \tag{13}$$

and the result follows. □

Lemma 3 (Assumption 4 in Gogolashvili et al. [2023])**.** *Under technical assumptions in TODO, for* $s = d/(2\nu + d)$ *it holds that*

$$E_s = \max\left(1, \sup_{\lambda \in (0,1]} \sqrt{\sum_{i=1}^{\infty} \frac{\mu_i \lambda^s}{\mu_i + \lambda}}\right) < \infty.$$

Proof. It is a standard result (see, for instance, [**?**, Section 3.3.4]) that for k_Θ being a Matérn kernel of order ν, the i-th eigenvalue decays at the rate of $i^{-\frac{2\nu+d}{d}}$. As pointed out in the discussion after Assumption 4 in Gogolashvili et al. [2023], this implies $E_s < \infty$ holds for $s = d/(2\nu + d)$. □

Theorem 2. *Suppose technical assumptions in TODO hold,* T *is large enough so that* $c\tau T^{-(1-\alpha)} < 1$*, and* $c \geq 8\tau^{-1/2}|\log(6/\delta)|(\|\Theta\| + 1)^{1/2}$*. Then*

$$\|\hat{I}_{\mathrm{CBQ}}(\theta) - I(\theta)\|_{\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{tr}})} \leq C_0(1 + \gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))T^{-r\beta},$$

for $C_0 = \tau\left(4(M + \|I\|_{\mathcal{H}_\Theta})(1 + \sqrt{\|\Theta\|})|\log(6/\delta)|^{1/2} + (c\tau)^r R\right)$.

Proof. By Gogolashvili et al. [2023, Theorem 4], for $\lambda = \tau c T^{-(1-\alpha)}$ and the weight function $w(\theta)$ defined in (**??**), we have that

$$\|\hat{I}_{\mathrm{CBQ}}(\theta) - I(\theta)\|_{\mathcal{L}^2(\Theta, \mathbb{P}_{\mathrm{te}})} \leq T^{-r\beta}\left(16(M + \|I\|_{\mathcal{H}_\Theta})(W + \sigma E_s^{1-q})c^{-A/2}\log(6/\delta) + c^r R\right)$$

provided $\lambda = \tau c T^{-(1-\alpha)} \leq 1$ and $\tau c \geq \left(64(W + \sigma^2)E_s^{2(1-q)}\log^2(6/\delta)\right)^{1/(1+A)}$, for the constants[2] $A = 1/(s(1-q) + q)$, and W, σ^2, q, r, R specified in Lemmas 1 to 3, meaning.

$$W, \sigma^2, q, r = \min\{s/(2\nu + d), 1\}, R, A = 1$$

$\lambda T/w = \gamma_T(1 + \gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta))$

$\lambda = T^{-1}\gamma_T$. Recall $\gamma_T = cT^\alpha$, $\alpha \in (0, 1)$

$\lambda = cT^{-(1-\alpha)}$

$\beta = (1 - \alpha) = \frac{1}{2r + s(1-q) + q} = \frac{1}{2\min\{s/(2\nu+d),1\} + d(1-q)/(2\nu+d) + q}$

$$|I(\theta)| \leq M, \text{ for any } \theta \in \Theta$$

$q = 1$, $W = \sup_t(1 + \gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_t))^{-1}$, and $\sigma^2 = \|\Theta\|(1 + \gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_t))^{-2}$, $s = d/(2\nu + d)$, $r = \min\{s/(2\nu + d), 1\}$

$$c \geq 8\log(6/\delta)\left(\frac{1}{1 + c^{-1}T^{-\alpha}\sigma^2_{\mathrm{BQ}}(\theta_{t''})} + \frac{\|\Theta\|}{(1 + c^{-1}T^{-\alpha}\sigma^2_{\mathrm{BQ}}(\theta_{t''}))^2}\right)^{1/2}$$

□

Assumptions are

- $\mathcal{X}$ is bounded and satisfies an (R, δ) interior cone condition with a Lipschitz boundary, also known as $\mathcal{L}(R, \delta)$ domain.
- $f(\cdot, \theta_t) \in W_2^{\tau_f}(\mathcal{X})$ for some $\tau_f > d/2$.
- The kernel $k_\mathcal{X}$ is τ_k smooth in that the RKHS $\mathcal{H}_\mathcal{X}$ is norm equivalent to $W_2^{\tau_k}$.
- $\tau_f = \tau_k$

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

- fill distance h_X. Given a collection of points $X \in \mathcal{X}$, the fill distance h_X is defined as $h_X := \sup_{x\in\mathcal{X}} \inf_{y\in X} \|x-y\|_2$.
- The distribution $\mathbb{P}_\theta(x)$ has density $p_\theta(x)$ for any $\theta \in \Theta$.

Lemma 4. *[Corollary 5 in Wynne et al. [2021]] For Gaussian process regression on observations $\{x_{1:N}, \hat{f}(x_{1:N},\theta)\}$ with corruption $\varepsilon_{1:N}$. Let $s \in [0, \tau_f]$, under some assumptions above and $h_{X_N} \le C_1 N^{-\frac{1}{d}}$ for some $C_1 > 0$. Then, $\exists C, h_0 > 0$ such that with $h_{X_N} < h_0$, we have*

$$\mathbb{E}\left[\left\| f(\cdot,\theta) - k_{\mathcal{X}}(\cdot, x_{1:N}) \left(k_{\mathcal{X}}(x_{1:N}, x_{1:N}) + \gamma_N\right)^{-1} \hat{f}(x_{1:N},\theta)\right\|_{W_q^s(\mathcal{X})}\right]$$
$$\le C n^{-\frac{1}{2}+\frac{s}{d}} \left(\mathbb{E}\left[\|\varepsilon\|_2\right] + n^{-\frac{\tau_f}{d}+\frac{1}{2}} \left(\|f\|_{W_2^\tau(\mathcal{X})} + \|m\|_{W_2^{\tau_f}(\mathcal{X})}\right)\right),$$

where the constant C depends on $\mathcal{X}, d, q, \tau_f$ and h_0 depends on R, d, τ_f.

Proof. It is a slight modification of the original Corollary 5 in Wynne et al. [2021] by taking $q = 2$ and taking the GP prior mean to have fixed hyperparameters. □

Theorem 3. *Under Assumptions blah blah blah, for any θ_t, $t = \{0, \dots, T\}$, we have that*

$$\sigma^2_{\mathrm{BQ}}(\theta_t) \le C N^{-\frac{\tau_f}{d}},$$

where the constant $C = C_0 \|f\|_{W_2^{\tau_f}(\mathcal{X})} \|p_{\theta_t}(x)\|_{L^2(\mathcal{X})}$ and C_0 depends on $\mathcal{X}, d, \tau_f$.

Proof. Recall from Section 3 that BQ variance has the form

$$\sigma^2_{\mathrm{BQ}}(\theta_t) = \mathbb{E}_{X,X'\sim\mathbb{P}}\left[k_{\mathcal{X}}(X,X')\right] - \mu(x_{1:N})^\top \left(k_{\mathcal{X}}(x_{1:N},x_{1:N}) + \gamma_N\right)^{-1} \mu(x_{1:N}) \tag{14}$$

where γ_N is a small jitter added to the Gram matrix to ensure invertibility.

In the literature of BQ, it is commonly assumed that the function evaluations are noise-free. For example, the function evaluations are the result of very expensive computer simulations. However, in Equation (14), adding a small jitter γ_N to the Gram matrix indicates that the likelihood is assumed to be Gaussian with very small variance. Therefore, we fall in the misspecified case as Lemma 4 so we have

$$\mathbb{E}\left[\left\| f(\cdot,\theta_t) - k_{\mathcal{X}}(\cdot, x^t_{1:N}) \left(k_{\mathcal{X}}(x^t_{1:N}, x^t_{1:N}) + \gamma_N\right)^{-1} \hat{f}(x^t_{1:N},\theta_t)\right\|_{W_q^s(\mathcal{X})}\right]$$
$$\le C n^{-\frac{1}{2}+\frac{s}{d}} \left(\mathbb{E}\left[\|\varepsilon\|_2\right] + n^{-\frac{\tau_f}{d}+\frac{1}{2}} \left(\|f\|_{W_2^\tau(\mathcal{X})} + \|m\|_{W_2^{\tau_f}(\mathcal{X})}\right)\right),$$

Since we are in the noiseless case so $\hat{f} = f$ and $\varepsilon = 0$, and the expectation on the left hand side is removed. The GP prior mean is taken to be a zero function so $m = 0$. Then let $s = 0$, we arrive at

$$\left\| f(\cdot,\theta_t) - k_{\mathcal{X}}(\cdot, x^t_{1:N}) \left(k_{\mathcal{X}}(x^t_{1:N}, x^t_{1:N}) + \gamma_N\right)^{-1} f(x^t_{1:N},\theta_t)\right\|_{L^2(\mathcal{X})} \le C N^{-\frac{\tau_f}{d}} \|f\|_{W_2^{\tau_f}(\mathcal{X})} \tag{15}$$

With Holder inequality, we have

$$\int \left| f(x,\theta_t) - k_{\mathcal{X}}(x, x^t_{1:N}) \left(k_{\mathcal{X}}(x^t_{1:N}, x^t_{1:N}) + \gamma_N\right)^{-1} f(x^t_{1:N},\theta_t) \right| \left| p_{\theta_t}(x) \right| dx$$
$$\le \left\| f(\cdot,\theta_t) - k_{\mathcal{X}}(\cdot, x^t_{1:N}) \left(k_{\mathcal{X}}(x^t_{1:N}, x^t_{1:N}) + \gamma_N\right)^{-1} f(x^t_{1:N},\theta_t)\right\|_{L^2(\mathcal{X})} \|p_{\theta_t}(x)\|_{L^2(\mathcal{X})}$$
$$\le C N^{-\frac{\tau_f}{d}} \|f\|_{W_2^{\tau_f}(\mathcal{X})} \|p_{\theta_t}(x)\|_{L^2(\mathcal{X})}$$

From Section 2.3 in Briol et al. [2019], we know that the BQ posterior variance is equivalently the worst case error in the reproducing kernel Hilbert space $\mathcal{H}_{\mathcal{X}}$

$$\sigma^2_{BQ}(\theta_t) = \sup_{\substack{g(\cdot,\theta_t)\in\mathcal{H}_{\mathcal{X}} \\ \|g(\cdot,\theta_t)\|_{\mathcal{H}_{\mathcal{X}}}\le 1}} \int \left| g(x,\theta_t) - k_{\mathcal{X}}(x, x^t_{1:N}) \left(k_{\mathcal{X}}(x^t_{1:N}, x^t_{1:N}) + \gamma_N\right)^{-1} g(x^t_{1:N},\theta_t) \right| \left| p_{\theta_t}(x) \right| dx$$

And since Equation (2.2) holds for any $f(\cdot,\theta_t)$, and therefore we have

$$\sigma^2_{BQ}(\theta_t) \le C N^{-\frac{\tau_f}{d}} \|f\|_{W_2^{\tau_f}(\mathcal{X})} \|p_{\theta_t}(x)\|_{L^2(\mathcal{X})}$$

And hence Theorem 2 is proved. [Hudson: not sure about the final step.] □

3 Attempt at a better q TODO delete pre submission

Lemma 5 (Assumption 3 in Gogolashvili et al. [2023])**.** *Under technical assumptions in TODO, there exist constants $q \in [0,1]$, $W > 0$, and $\sigma > 0$ such that, for all $m \in \mathbb{N}$, $m \geq 2$, it holds that*

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

Proof.

$$w(\theta) = \begin{cases} A_t - B_t \frac{\|\theta - \theta_t\|_\Theta}{\varepsilon}, & t \text{ such that } \|\theta - \theta_t\|_\Theta < \varepsilon \\ A_t - B_t & \text{otherwise} \end{cases}$$

$$A_t = (1 + \gamma_T^{-1} \sigma_{\mathrm{BQ}}^2(\theta_t))^{-1}$$

$$B_t = (1 + \gamma_T^{-1} \sigma_{\mathrm{BQ}}^2(\theta_t))^{-1} - (1 + \gamma_T^{-1} \sigma_{\mathrm{BQ}}^2(\theta_{t'}))^{-1}$$

$$w(\theta)^{\frac{q+m-1}{q}} = \begin{cases} \left(A_t - B_t \frac{\|\theta - \theta_t\|_\Theta}{\varepsilon}\right)^{\frac{q+m-1}{q}}, & t \text{ such that } \|\theta - \theta_t\|_\Theta < \varepsilon \\ (A_t - B_t)^{-\frac{q+m-1}{q}} & \text{otherwise} \end{cases}$$

Then,

$$\int_\Theta w(\theta)^{\frac{q+m-1}{q}} \mathrm{d}\theta = (\|\Theta\| - \varepsilon T V_d)\,(A_t - B_t)^{-\frac{q+m-1}{q}} + \sum_{t=1}^T \int_{B_\varepsilon(\theta_t)} \left(A_t - B_t \frac{\|\theta - \theta_t\|_\Theta}{\varepsilon}\right)^{\frac{q+m-1}{q}} \mathrm{d}\theta \tag{16}$$

where V_d is the volume of a d-dimensional unit ball, $V_d = \pi^{d/2}/\Gamma(1 + d/2)$.

$$\int_{B_\varepsilon(\theta_t)} \left(A_t - B_t \frac{\|\theta - \theta_t\|_\Theta}{\varepsilon}\right)^{\frac{q+m-1}{q}} \mathrm{d}\theta = 2\pi \int_0^\varepsilon (A_t - B_t r/\varepsilon)^{\frac{q+m-1}{q}} r^{d-1} \mathrm{d}r \prod_{i=1}^{d-2} \int_0^\pi \sin^{d-2-i+1}(\phi_i) \mathrm{d}\phi_i$$

$$2\pi \int_0^1 r^{d-1} \mathrm{d}r \prod_{i=1}^{d-2} \int_0^\pi \sin^{d-2-i+1}(\phi_i) \mathrm{d}\phi_i = V_d = \pi^{d/2}/\Gamma(1 + d/2)$$

$$2\pi \prod_{i=1}^{d-2} \int_0^\pi \sin^{d-2-i+1}(\phi_i) \mathrm{d}\phi_i = d\pi^{d/2}/\Gamma(1 + d/2)$$

$$\frac{d\pi^{d/2}}{\Gamma(1 + d/2)} \int_0^\varepsilon (A_t - B_t r/\varepsilon)^{\frac{q+m-1}{q}} r^{d-1} \mathrm{d}r = \frac{\pi^{d/2}}{\Gamma(1 + d/2)} \left[\frac{A_t^{\frac{q+m-1}{q}+d} - (A_t - B_t)^{\frac{q+m-1}{q}+d}}{B_t(\frac{q+m-1}{q} + d)} \prod_{j=1}^{d-1} \frac{d - j + 1}{\frac{q+m-1}{q} + j} \right.$$

$$\left. - \sum_{i=1}^{d-1} \frac{1}{B_t^i} \varepsilon^{d-i} (A_t - B_t)^{\frac{q+m-1}{q}+i} \prod_{j=1}^{i} \frac{d - j + 1}{\frac{q+m-1}{q} + j} \right]$$

The latter equality is

$$\int_0^\varepsilon (a - br)^c r^{d-1} \mathrm{d}r = \frac{1}{d} \frac{a^{c+d} - (a - b\varepsilon)^{c+d}}{b(c + d)} \prod_{j=1}^{d-1} \frac{d - j + 1}{c + j} - \frac{1}{d} \sum_{i=1}^{d-1} \frac{1}{b^i} \varepsilon^{d-i} (a - b\varepsilon)^{c+i} \prod_{j=1}^{i} \frac{d - j + 1}{c + j}$$

which comes from integration by parts,

$$\begin{aligned} \int_0^\varepsilon (a - br)^c r^{d-1} \mathrm{d}r &= -\frac{1}{b(c+1)} \int_0^\varepsilon r^{d-1} \mathrm{d}\,(a - br)^{c+1} \\ &= -\frac{1}{b(c+1)} \varepsilon^{d-1} (a - b\varepsilon)^{c+1} + \frac{1}{b(c+1)} \int_0^\varepsilon (a - br)^{c+1} \mathrm{d}r^{d-1} \\ &= -\frac{1}{b(c+1)} \varepsilon^{d-1} (a - b\varepsilon)^{c+1} \\ &\quad - \frac{d-1}{b^2(c+1)(c+2)} \varepsilon^{d-2} (a - b\varepsilon)^{c+2} \\ &\quad + \frac{d-1}{b^2(c+1)(c+2)} \int_0^\varepsilon (a - br)^{c+2} \mathrm{d}r^{d-2} = \dots \end{aligned}$$

and

$$\int_0^{\varepsilon} (a - br)^{c+d-1} \, \mathrm{d}r = \frac{a^{c+d} - (a - b\varepsilon)^{c+d}}{b(c+d)}.$$

So (16) becomes

$$\begin{aligned}\int_\Theta w(\theta)^{\frac{q+m-1}{q}} \mathrm{d}\theta = (\|\Theta\| - \varepsilon T V_d) \left(1 + \gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'})\right)^{-\frac{q+m-1}{q}} \\ + \sum_{t=1}^{T} \frac{\pi^{d/2}}{\Gamma(1+d/2)} \left[\frac{A_t^{\frac{q+m-1}{q}+d} - (A_t - B_t)^{\frac{q+m-1}{q}+d}}{B_t(\frac{q+m-1}{q}+d)} \prod_{j=1}^{d-1} \frac{d-j+1}{\frac{q+m-1}{q}+j} \right. \\ \left. - \sum_{i=1}^{d-1} \frac{1}{B_t^i} \varepsilon^{d-i} (A_t - B_t)^{\frac{q+m-1}{q}+i} \prod_{j=1}^{i} \frac{d-j+1}{\frac{q+m-1}{q}+j} \right]\end{aligned}$$

blah

$$w(\theta) = (1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1} + \max\left\{ \left(1 - \frac{\|\theta - h(\theta)\|_\Theta}{\varepsilon}\right) \left((1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(h(\theta)))^{-1} - (1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1}\right), 0 \right\}$$

$$w(\theta) = \max\left\{ \left(1 - \frac{\|\theta - h(\theta)\|_\Theta}{\varepsilon}\right) (1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(h(\theta)))^{-1} + \frac{\|\theta - h(\theta)\|_\Theta}{\varepsilon} (1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1}, (1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1} \right\}$$

$$w(\theta) = \max\left\{ \left(1 - \frac{\|\theta - h(\theta)\|_\Theta}{\varepsilon}\right) (1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(h(\theta)))^{-1} + \frac{\|\theta - h(\theta)\|_\Theta}{\varepsilon} (1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1}, (1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1} \right\}$$

$$\left(\int_\Theta w(\theta)^{\frac{q+m-1}{q}} \mathbb{P}_{\mathrm{tr}}(\mathrm{d}\theta)\right)^{\frac{q}{q+m-1}} \leq (1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1} + \left(\int_{\Theta_{T,\varepsilon}} B(\theta)^{\frac{q+m-1}{q}} \mathbb{P}_{\mathrm{tr}}(\mathrm{d}\theta)\right)^{\frac{q}{q+m-1}}$$

for

$$B(\theta) = \left(1 - \frac{\|\theta - h(\theta)\|_\Theta}{\varepsilon}\right) \left((1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(h(\theta)))^{-1} - (1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1}\right)$$

So

$$\left(\int_\Theta B(\theta)^{\frac{q+m-1}{q}} \mathbb{P}_{\mathrm{tr}}(\mathrm{d}\theta)\right)^{\frac{q}{q+m-1}} = \left(\int_{\Theta_{T,\varepsilon}} \left((1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(h(\theta)))^{-1} - (1+\gamma_T^{-1}\sigma^2_{\mathrm{BQ}}(\theta_{t'}))^{-1}\right)^{\frac{q+m-1}{q}} \mathbb{P}_{\mathrm{tr}}(\mathrm{d}\theta)\right)^{\frac{q}{q+m-1}}$$

By the construction of $w(\theta)$,

$$\int_\Theta w(\theta)^m \mathrm{d}\theta \leq \|\Theta\| \sigma^{-2m}_{\mathrm{BQ}}(\theta_{t'}) + \varepsilon_0$$

□

4 Convergence rate—OLD TODO delete pre submission

Recall the CBQ estimator proposed in (2),

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

where $\alpha > 0$, $\hat{I}_{\mathrm{BQ}}(\theta_t)$ and $\sigma^2_{\mathrm{BQ}}(\theta_t)$, for $t \in \{1, \ldots, T\}$, are BQ posterior mean and variance obtained in the first stage as given in (1)

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

It was pointed out in Gogolashvili et al. [2023, Remark 2], (and can be seen through straightforward differentiation) that the estimator $\hat{I}_{\mathrm{CBQ}}(\theta)$ is the minimiser of the importance weighted kernel ridge regression loss over functions in the RKHS $\mathcal{H}_\Theta$,

$$\hat{I}_{\mathrm{CBQ}}(\theta) = \operatorname*{arg\,min}_{F \in \mathcal{H}_\Theta} \left\{ \sum_{t=1}^{T} \frac{\left(F(\theta_t) - \hat{I}_{\mathrm{BQ}}(\theta_t)\right)^2}{\sigma^2_{\mathrm{BQ}}(\theta_t)} + T^{-\alpha} \|F\|^2_{\mathcal{H}_\Theta} \right\}.$$

Suppose θ_i were sampled from a probability measure $\mathbb{P}_{\mathrm{tr}}$ on Θ. Then,

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

defines a positive measure on Θ for any positive $w(\theta) > 0$ for which the integral exists [Fremlin, 2000, 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 $w(\theta_t) = 1/\sigma^2_{\mathrm{BQ}}(\theta_t) = \sigma^{-2}_{\mathrm{BQ}}(\theta_t)$. Then, since $\mathbb{E}[\hat{I}_{\mathrm{BQ}}(\theta_i)] = I(\theta_i)$, this (TODO proper reference) 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 the rate of convergence of $\hat{I}_{\mathrm{CBQ}}(\theta)$ to $I(\theta)$ as $n \to \infty$ was established in Gogolashvili et al. [2023, Theorem 4]. Specifically, [TODO summarise once it's more clear.]

To apply the result, we define $w(\theta)$ of convenient form that satisfies the requirements mentioned above, specifically $w(\theta) \in (0, A]$ for some $A < \infty$ and any $\theta \in \Theta$, and $w(\theta_t) = \sigma^{-2}_{\mathrm{BQ}}(\theta_t)$ for some $t \in \{0, \ldots, T\}$.[3] Take $t' = \arg\min_{t \in \{0,\ldots,T\}} \{\sigma^{-2}_{\mathrm{BQ}}(\theta_t)\} > 0$, and define

$$w(\theta) = \sigma^{-2}_{\mathrm{BQ}}(\theta_{t'}) + \max\left\{ \left(1 - \frac{\|\theta - h(\theta)\|_\Theta}{\varepsilon}\right) \left(\sigma^{-2}_{\mathrm{BQ}}(h(\theta)) - \sigma^{-2}_{\mathrm{BQ}}(\theta_{t'})\right), 0 \right\}$$

where $h(\theta) = \arg\min_{\theta' \in \{\theta_0,\ldots,\theta_T\}} \{\|\theta - \theta'\|_\Theta\}$ is the point in the set $\{\theta_0, \ldots, \theta_T\}$ that is closest to θ. For $\Theta \subset \mathbb{R}$, such $w(\theta)$ is easily visualised, as can be seen in Figure 6. Crucially, for a bounded Θ, the volume $\int_\Theta w(\theta)\mathrm{d}\theta$ can be made arbitrarily close to $\|\Theta\| \sigma^{-2}_{\mathrm{BQ}}(\theta_{t'})$ for a small enough ε.

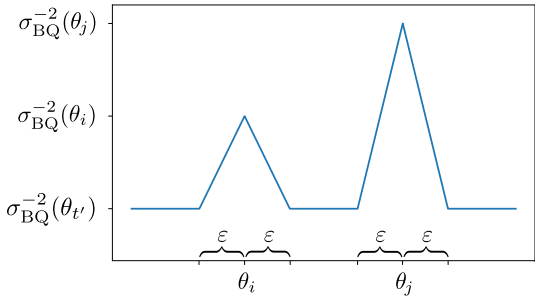


Figure 6: Illustration of $w(\theta)$ for $\Theta \subset \mathbb{R}$

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

4.1 Technical Assumptions

- $I(\theta)$ lies in the Sobolev space $\mathcal{W}^{2,s}(\Theta)$.
- k_Θ is a Matérn kernel of order ν such that $s \geq \nu + d/2$

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

4.2 Convergence

Lemma 6 (Assumption 2 in Gogolashvili et al. [2023])**.** *Under technical assumptions in TODO,* $I(\theta) = L^r g$ *for* $r = \min\{s/(2\nu + d), 1\}$ *and some* $g \in \mathcal{L}^2(\Theta)$.

Proof. By Steinwart and Scovel [2012, Theorem 4.6], the range of the integral operator $\mathrm{ran}L^r$ coincides with the $2r$-interpolation space between $\mathcal{H}_\Theta \simeq \mathcal{W}^{2,\nu+d/2}(\Theta)$ and $\mathcal{L}^2(\Theta)$. Therefore, we ought to show that $I(\theta)$ lies in this $2r$-interpolation space.

The $2r$-interpolation space between $\mathcal{W}^{2,\nu-d/2}(\Theta)$ and $\mathcal{L}^2(\Theta)$ is the space $\mathcal{W}^{2,2r(\nu-d/2)}(\Theta)$ (TODO ref?). Since $I(\theta)$ lies in $\mathcal{W}^{2,s}$, by inclusion of Sobolev spaces we have that it lies in $\mathcal{W}^{2,s_0}$ for any $s_0 \leq s$. Therefore, $I(\theta)$ lies in the aforementioned interpolation space whenever $2r(\nu + d/2) \leq s$, meaning $r \leq s/(2\nu + d)$. The result follows. □

Note that, under Assumption TODO, $s/(2\nu + d) \geq 1/2$, and we have $r \in [1/2, 1]$, as is required by Assumption 2 in Gogolashvili et al. [2023].

Lemma 7 (Assumption 3 in Gogolashvili et al. [2023])**.** *Under technical assumptions in TODO, there exist constants* $q \in [0, 1]$, $W > 0$, *and* $\sigma > 0$ *such that, for all* $m \in \mathbb{N}$, $m \geq 2$, *it holds that*

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

Proof.

$$w(\theta) = \sigma_{\mathrm{BQ}}^{-2}(\theta_{t'}) + \max\left\{\left(1 - \frac{\|\theta - h(\theta)\|_\Theta}{\varepsilon}\right)\left(\sigma_{\mathrm{BQ}}^{-2}(h(\theta)) - \sigma_{\mathrm{BQ}}^{-2}(\theta_{t'})\right), 0\right\}$$

By the construction of $w(\theta)$,

$$\int_\Theta w(\theta)^m \mathrm{d}\theta \leq \|\Theta\| \sigma_{\mathrm{BQ}}^{-2m}(\theta_{t'}) + \varepsilon_0$$

□

$n^{-1/2}$, $n^{-(1-A)/2}$, n^{-r}

As $0 \leq A \leq 1$, $-1/2 \leq -(1-A)/2 \leq 0$, the second term dominates the first term always. Therefore, the rate is $n^{-\min\{r,(1-A)/2\}}$. Recall $r = \min s/(2\nu + d), 1 \in [1/2, 1]$, and if we take $q = 1$ (easiest), then $A = d/(2\nu + d)$. That gives the rate of $n^{-\min\{s/(2\nu+d),\nu/(2\nu+d),1\}}$. Since we assumed $s \geq \nu + d/2$, the rate is $n^{-\min\{\nu/(2\nu+d),1\}}$

so long as

$$c^{1+q} n^{-q} \geq 64(V + \gamma^2) N(\lambda)^{1-q} \log^2(6/\delta)$$

for $q = 1$,

$$c^2 n^{-1} \geq 64(W + \sigma^2) \log^2(6/\delta)$$

$c^{1+q} n^{-q} \geq 64(V + \gamma^2) c^{s(q-1)} n^{s(1-q)} (E_s)^{2(1-q)} \log$

$c^{1+q-s(q-1)} n^{-q-s(1-q)} \geq 64(V + \gamma^2)(E_s)^{2(1-q)} \log$

5 Notation (TODO delete pre submission)

$$
\begin{aligned}
&\theta_{1:T} = [\theta_1 \cdots \theta_T]^\top \in \Theta^T, \Theta \subseteq \mathbb{R}^p \\
&x^t_{1:T} = [x^t_1 \cdots x^t_N]^\top \in \Theta^N \text{ for all } t \in \{1, \dots, T\}, \Theta \subseteq \mathbb{R}^d \\
&m_\Theta, k_\Theta, \mathcal{H}_\Theta, m_\Theta, k_\Theta, \mathcal{H}_\Theta \\
&\hat{I}_{\text{MC}}(\theta), \hat{I}_{\text{IS}}(\theta), \hat{I}_{\text{LSMC}}(\theta), \hat{I}_{\text{KMS}}(\theta), \\
&(\hat{I}_{\text{BQ}}, \sigma^2_{\text{BQ}}) \text{ or later } (\hat{I}_{\text{BQ}}(\theta), \sigma^2_{\text{BQ}}(\theta)), \\
&\mu_\theta(x) = \mathbb{E}_{X \sim \mathbb{P}_\theta}[k_\Theta(X, x)], \mathbb{E}_{X, X' \sim \mathbb{P}_\theta}[k_\Theta(X, X')] \\
&\sigma^2(\theta), \qquad |I(\theta) - \hat{I}_{\text{BQ}}(\theta)| < \|f(\cdot, \theta)\|_{\mathcal{H}_\Theta} \sigma^2_{\text{BQ}}(\theta)
\end{aligned}
$$

$$
\begin{aligned}
g(x, \theta, \omega), I_g(\theta, \omega) = \mathbb{E}_{X \sim \mathbb{P}_\theta}[g(X, \theta, \omega)], k_{\Theta,\Theta} = k_\Theta \times k_\Theta \\
k_\Theta(\theta, \theta') = \int \int k_\Theta(x, x') p(x' \mid \theta') p(x \mid \theta) dx dx'
\end{aligned}
$$

6 Practical Considerations

6.1 Tractable Kernel Means

One of the limitations for CBQ and for all BQ-related methods is that the kernel mean embedding μ is assumed to be known in closed-form; see Table 1 in Briol et al. [2019] or the `ProbNum` package [Wenger et al., 2021] 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 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 kernels 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 $\lim_{n \to \infty} p(x)f(x) = 0$. 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.

$$
\begin{aligned}
k_p(x, x') = T^x_p[T^{x'}_p[k_0(x, x')]] = &\nabla_x \log p(x) k_0(x, x') \nabla_{x'} \log p(x') + \nabla \log p(x) \nabla_{x'} k_0(x, x') \\
&+ \nabla \log p(x') \nabla_x k_0(x, x') + \nabla_x \nabla_{x'} k_0(x, x')
\end{aligned}
$$

It is noteworthy to mention that when taking the derivative of the logarithm, the requirement for the normalization constant of $p(x)$ is effectively eliminated.

Stein identity indicates that the kernel mean embedding of equals to 0, i.e $\mu(x') = \int k_p(x, x') p(x) dx = 0$. However, it is unreasonable to have $\mu(x_{1:N}) = 0$ as that would suggest the BQ estimate in Equation (1) is always equal to 0, and hence the CBQ estimate in Equation (2) is also 0. Therefore, we add a small 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 $\mu(x') = \int \tilde{k}_p(x, x') p(x) dx = c$. 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 Oates et al. [2017b] to select control functionals.

6.2 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, 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 is $m(x_{1:N})$, the log-marginal likelihood can be written as [Rasmussen and Williams, 2006]:

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

Fortunately, the optimal amplitude parameter $A_\mathcal{X}$ is known in closed-form by taking the derivative:

$$A^*_\mathcal{X} := \sqrt{\frac{f(x_{1:N})^\top \tilde{k}_\mathcal{X}(x_{1:N}, x_{1:N}; l_\mathcal{X})^{-1} f(x_{1:N})}{N}},$$

where $\tilde{k}_\mathcal{X}$ denotes $k_\mathcal{X}$ with the amplitude parameter equals to 1. Therefore, for $k_\mathcal{X}$ we only need to select the optimal lengthscale $l_\mathcal{X}$, and we use a grid search over $[0.1, 0.3, 1.0, 3.0, 10.0]$ and select the value that gives the largest log-marginal likelihood.

If $k_\mathcal{X}$ is a Stein kernel, it has another 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 [Bradbury et al., 2018].

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

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

For k_Θ, both amplitude A_Θ and lengthscale l_Θ does not have a closed form expression. We 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.

7 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.

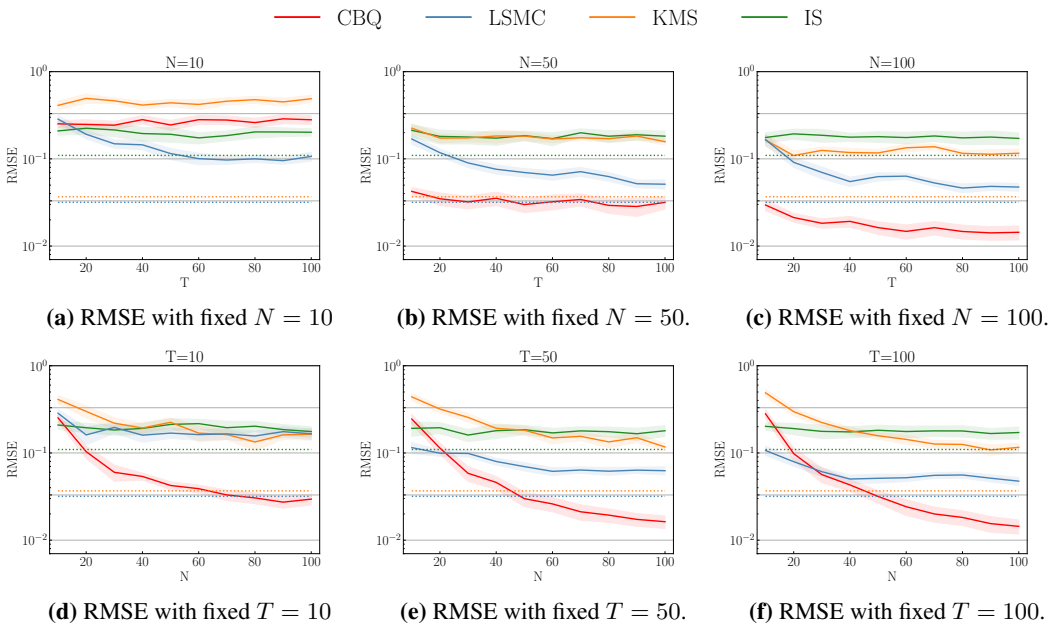
7.1 Synthetic Experiment: Bayesian Sensitivity Analysis for Linear Models

7.1.1 Experimental Setting

In the toy experiment, we do sensitivity analysis on the hyperparameters in Bayesian linear regression. The 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. The prior is chosen as a multivariate Gaussian $p(w) = \mathcal{N}(0, \bar{\Sigma})$ and the likelihood is also Gaussian $p(\mathcal{D} \mid w) = \mathcal{N}(z; w^\top y, \eta)$. The posterior is known to have a closed form [Bishop, 2006]

$$p(w \mid \mathcal{D}) = \mathcal{N}(\tilde{m}, \tilde{\Sigma}), \quad \tilde{\Sigma}^{-1} = \bar{\Sigma}^{-1} + \eta Y^\top Y, \quad \tilde{m} = \eta\tilde{\Sigma}Y^\top Z$$

We are interested in the integral of two functions against the posterior distribution $I(\bar{\Sigma}) = \int f(w)p(w \mid \mathcal{D}; \bar{\Sigma})dw$. The first integrand is $f(w) = w^\top w$ and the integral describes the sum



(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: More experimental results for Bayesian sensitivity analysis. The integrand is $f(w) = w^\top w$ and the dimension $d = 2$.

of posterior variance. The second integrand is $f(w) = w^\top y^*$ for a new observation y^* and the integral describes the predictive mean. Both integrals are known to have closed form expression in Bayesian linear regression [Bishop, 2006], so it is easier for us to compare the performances of different methods. We are interested in the analysis on the sensitivity of $I(\bar{\Sigma})$ towards $\bar{\Sigma}$. The prior covariance $\bar{\Sigma}$ is chosen to be a diagonal matrix for simplification. The observations $\bar{\Sigma}_{1:T}$ are sampled from a uniform distribution on each entry, i.e $\bar{\Sigma}_{1:T} \sim [1, 3]^d$. For each $\bar{\Sigma}_t$, we have N samples from the posterior $w^t_{1:N} \sim p(w \mid \mathcal{D}, \bar{\Sigma}_t)$. 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 Section 3) to be a Gaussian kernel with lengthscale l and amplitude A

$$k(w, w') = A \exp(-\frac{1}{2l^2}(w - w')^\top (w - w')) \tag{17}$$

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

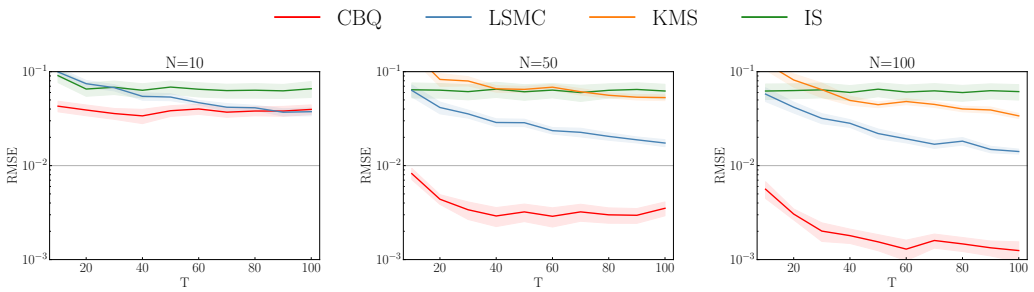
$$\mu_{\bar{\Sigma}}(w) = A|\, \mathrm{I} + l^{-2}\tilde{\Sigma}\,|^{-1/2} \exp\left(-\frac{1}{2}(w - \tilde{m})^\top (\tilde{\Sigma} + l^2 \mathrm{I})^{-1}(w - \tilde{m})\right) \tag{18}$$

Then we choose the kernel on the space of $\bar{\Sigma}$ (corresponds to k_Θ in Section 3). Since we assume $\bar{\Sigma}$ to be diagonal, we use product Matern kernel where the kernel on the space of each entry of $\bar{\Sigma}$ is chosen to be a Matern-3/2 kernel. The hyperparameters for both kernels are selected according to Section 6.2.

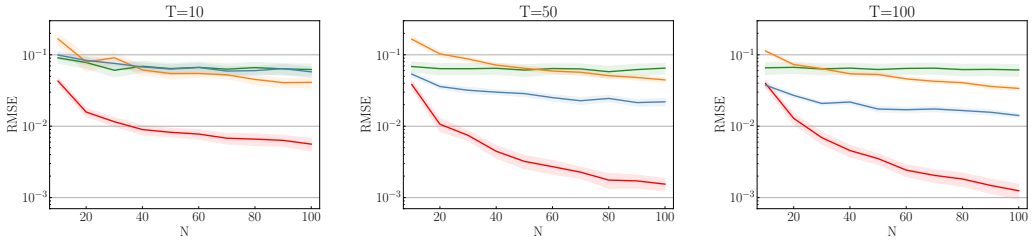
There are hyperparameters in baseline methods as well. For importance sampling (IS) estimator, there are no hyperparameters. For kernel mean shrinkage (KMS) estimator, we also use product Matern-3/2 kernel on the space of $\bar{\Sigma}$ and select hyperparameters according to Section 6.2. For least square 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.

7.1.2 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

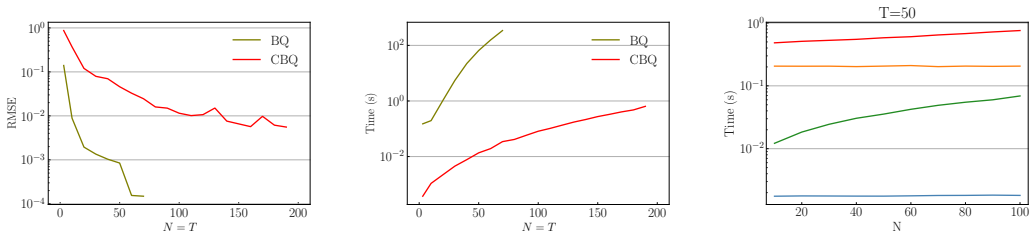


(a) RMSE with fixed $N = 10$ **(b)** RMSE with fixed $N = 50$. **(c)** RMSE with fixed $N = 100$.



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

Figure 8: More experimental results for Bayesian sensitivity analysis. The integrand is $f(w) = w^\top y^*$ and the dimension $d = 2$.

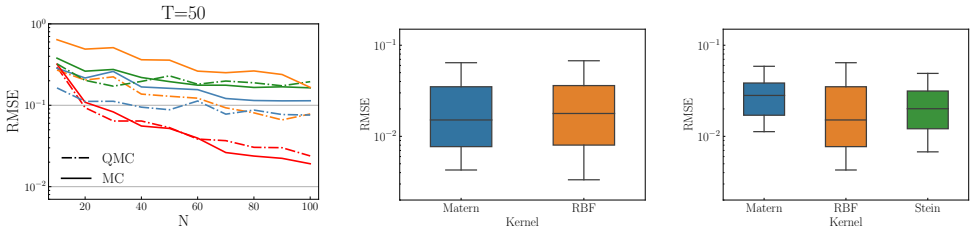


(a) CQ and CBQ RMSE **(b)** CQ and CBQ time **(c)** Computational time

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 $T = 50$. [FXB: I dont know how easy it is to try this or not, so this might be an annoying idea (in which case we dont need to do it). But what about trying a curve with $N = 2T$ or $N = 5T$ or perhaps even $N = T3/2$. From previous experiments, it seems like CBQ's performance improves faster when N grows than when T grows. For BQ, I think it wouldnt change too much, but for CBQ things might look much better. Just a thought]

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 lower 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 the application of the CBQ method.



(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 sampling methods 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.

7.1.3 Comparison of BQ and CBQ

In Section 3, we mentioned the comparison of 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 BQ and $\mathcal{O}(TN^3)$ for CBQ and the convergence rate is [Hudson: add convergence rate]. In Figure 9a and Figure 9b, we fix $N = T$ and demonstrate that BQ has a much faster convergence rate but the computational time gets unbearable quickly as the number of samples increases.

7.1.4 Quasi Monte Carlo

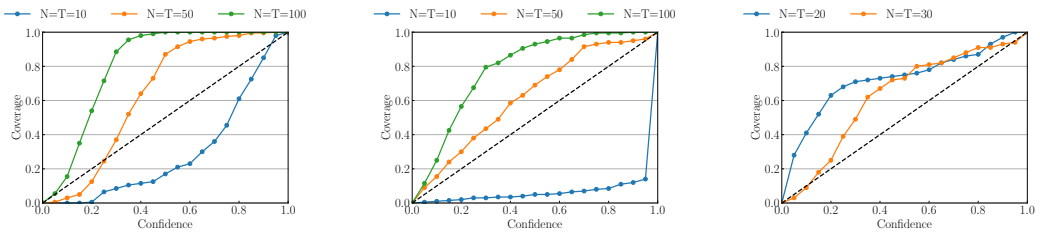
Quasi Monte Carlo (QMC) is another line of research on improving the precision of approximating intractable integrals. QMC aims to cover the integration domain more uniformly than random sampling used in standard Monte Carlo methods [Niu et al., 2023, Hickernell, 1998, Gerber and Chopin, 2015]. Sobol sequences are a type of low-discrepancy sequence commonly used in Quasi-Monte Carlo (QMC) methods, and Sobol sequences 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 Lemieux [2004] to shift the Sobol sequence by a random amount so that we can combine the 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 observations are being generated - 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 Conjugate Bayesian Quadrature (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.

7.1.5 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 Matern-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 Matern-3/2, Gaussian RBF, and Stein kernel (with Matern-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 $\mu_{\tilde{\Sigma}}(w)$ is presented in (18). In the scenario where $k_\mathcal{X}$ is the Matern-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 u from $\mathcal{N}(0, \mathrm{I})$, then calculating $w = \tilde{m} + L^\top u$ where L is the lower triangular matrix derived from the Cholesky decomposition of the covariance matrix $\tilde{\Sigma}$. In essence, $I(\tilde{\Sigma}) = \int f(w)\mathcal{N}(w; \tilde{m}, \tilde{\Sigma})dw = \int f(\tilde{m} +$



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

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

$L^\top u)\mathcal{N}(u;0,I)du$. When $k_{\mathcal{X}}$ is Stein kernel, we choose Matern-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 Section 6.2.

7.1.6 Calibration

In Section 3, we mentioned that CBQ provides uncertainty quantification for the integral of interest, but in the main text we only use root mean squared error (RMSE) as the performance metric where the uncertainty information is not fully utilized. In Figure 11a, we show the calibration of CBQ in Bayesian sensitivity analysis with integrand $f(w) = w^\top w$ and dimension $d = 2$. The black diagonal line represents the ideal case, the curve lying above the black line represents underconfidence and the curve lying below the black line represents overconfidence. It is generally more preferable to be underconfident than overconfident. In Figure 11a, we can see that when the number of samples are small as 10, CBQ is overconfident probably because we use empirical Bayes in selecting the hyperparameters that will tend to give more confident estimates. When the number of samples increases to 50, CBQ tends to become well calibrated and when the number increases to 100, CBQ tends to become underconfident.

7.2 Butterfly Call Option with the Black-Scholes Model

7.2.1 Experimental Setting

In this experiment, we consider specifically an asset whose price S_η at time η follows the Black-Scholes formula

$$S_\eta = S_0 \exp(\sigma W_\eta - \sigma^2 \eta/2), \quad \text{for} \quad \eta \geq 0$$

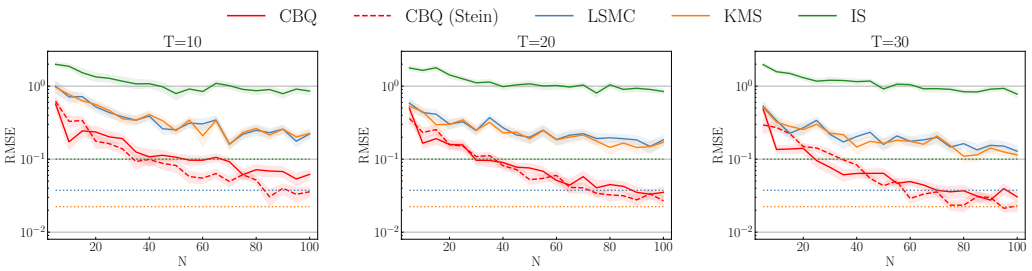
with σ being the underlying volatility and W being the standard Brownian motion. The financial derivative we are interested in is a butterfly call option whose payoff at time ζ can be expressed as

$$\psi\left(S_\zeta\right) = \max\left(S_\zeta - K_1, 0\right) + \max\left(S_\zeta - K_2, 0\right) - 2\max\left(S_\zeta - \frac{K_1 + K_2}{2}, 0\right)$$

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 Alfonsi et al. [2021, 2022] assuming that a shock occur at time η that multiplies the price of the butterfly option by $1 + s$, so the expected loss caused by the shock can be expressed as

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

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 $\{S_\eta\}_{1:T}$ are sampled from the log normal distribution deduced from the Black-Scholes formula $\{S_\eta\}_{1:T} \sim \text{Lognormal}\left(\log S_0 - \frac{\sigma^2}{2}\eta, \sigma^2\eta\right)$. Then N observations of $\{S_\zeta\}^t_{1:N}$ are sampled from the log normal distribution deduced from the Black-Scholes formula $\{S_\zeta\}^t_{1:N} \sim \text{Lognormal}\left(\log S_\eta - \frac{\sigma^2}{2}(\zeta - \eta), \sigma^2(\zeta - \eta)\right)$.



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

Figure 12: Butterfly call option with the Black-Scholes model

For conditional Bayesian quadrature (CBQ), we need to specify two kernels. First we choose the kernel on the space of S_ζ (corresponding to $k_\mathcal{X}$ in Section 3). Since in Black-Scholes model, $p(S_\zeta \mid S_\eta)$ is a log normal distribution, we use a log RBF kernel so that we can have a closed form mean embedding μ. $p(S_\zeta \mid S_\eta)$ is the log normal distribution derived from the Black-Scholes model $S_\zeta \sim \text{Lognormal}(\bar{m}, \bar{\sigma}^2)$ with $\bar{m} = \log S_\eta - \frac{\sigma^2}{2}t$ and $\bar{\sigma}^2 = \sigma^2 t$. The log RBF kernel is defined as

$$k(S_\zeta, S'_\zeta) = \lambda \exp(-\frac{1}{2l^2}(\log S_\zeta - \log S'_\zeta)^2)$$

and the kernel mean embedding has a closed form

$$\mu_{S_\eta}(S_\zeta) = \exp\left(-\frac{\bar{m}^2 + (\log S_\zeta)^2}{2(\bar{\sigma}^2 + l^2)}\right) S_\zeta^{\frac{\bar{m}}{\bar{\sigma}^2 + l^2}} \Big/ \bar{\sigma}\sqrt{\frac{1}{\bar{\sigma}^2} + \frac{1}{l^2}}$$

For CBQ with Stein kernel, we use Matern-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 on the space of S_η (corresponding to k_Θ in Section 3) as Matern-3/2 kernel. All hyperparameters in $k_\mathcal{X}$ and k_Θ are selected according to Section 6.2.

There are hyperparameters in other baseline methods as well. For importance sampling (IS) estimator, there are no hyperparameters. For kernel mean shrinkage (KMS) estimator, we also use product Matern-3/2 kernel on the space of $\bar{\Sigma}$ and select hyperparameters according to Section 6.2. 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.

7.2.2 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 $T = 10, 20, 30$. We can see that the outstanding performance of CBQ is consistent.

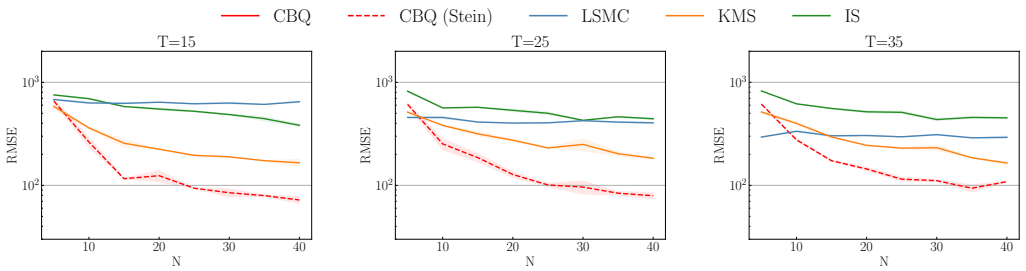
In Figure 11b, we also show the calibration of CBQ uncertainty using the log RBF kernel. The conclusion is almost consistent with the conclusion from Bayesian sensitivity analysis in that CBQ is overconfident when the number of observations are small and tend to become underconfident as the number of observation increases.

7.3 Bayesian Sensitivity for a Susceptible-Infectious-Recovered (SIR) Model

7.3.1 Experimental Settings

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) represent people who are not infected but can be infected after getting contact with an infectious individual. Infectious (I) represent people who are currently infected and can infect susceptible individuals. Recovered (R) represent 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.

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



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

Figure 13: Bayesian Sensitivity analysis for Susceptible-Infectious-Recovered (SIR) Model

β is the infection rate and γ is the recovery rate. The solution to the SIR model would be a vector of (I_d, S_d, R_d) indexed by time step d. In this experiment, we use scipy odeint function [Virtanen et al., 2020] to solve the ODEs.

In this experiment, we assume that the recovery rate γ is fixed and β follows a gamma prior distribution $\beta \sim \text{Gamma}(\beta; \bar{\beta}, \xi)$ where $\bar{\beta}$ 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 β: $I_{\max}(\bar{\beta}) = \mathbb{E}_\beta \left[\max_d I_d(\beta) \mid \bar{\beta}\right]$. It is always important to know how different initial estimate (different $\bar{\beta}$) of the infection rate will lead to different final estimate of $I_{\max}$.

For conditional Bayesian quadrature (CBQ), we need to specify two kernels. First we choose the kernel on the space of β (corresponding to $k_{\mathcal{X}}$ in Section 3). We use Matern-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 on the space of $\bar{\beta}$ (corresponding to k_Θ in Section 3) as Matern-3/2 kernel. All hyperparameters in $k_{\mathcal{X}}$ and k_Θ are selected according to Section 6.2.

There are hyperparameters in other baseline methods as well. For importance sampling (IS) estimator, there are no hyperparameters. For kernel mean shrinkage (KMS) estimator, we also use product Matern-3/2 kernel on the space of $\bar{\Sigma}$ and select hyperparameters according to Section 6.2. 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.

7.3.2 More Experimental Results

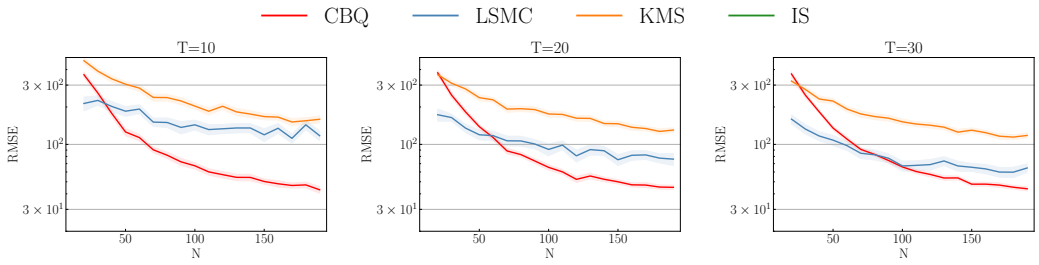
We report more experimental results for computing the expected peak number of infections from SIR model in Figure 13 with fixed $T = 15, 25, 35$. We can see that the outstanding performance of CBQ is consistent.

In Figure 11c, we also show the calibration of CBQ uncertainty. The conclusion is almost consistent with the conclusion from Bayesian sensitivity analysis in that CBQ is overconfident when the number of observations are small and tend to become underconfident as the number of observation increases.

7.4 Uncertainty Decision Making

7.5 Experimental Settings

7.5.1 More Experimental Results



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

Figure 14: Uncertainty decision making in health economics [Hudson: Larger N]