Active Multi-Information Source Bayesian Quadrature

Alexandra Gessner*
University of Tübingen
MPI for Intelligent Systems
Tübingen, Germany
agessner@tue.mpg.de

Javier Gonzalez
Amazon Research
Cambridge, UK
gojav@amazon.com

Maren Mahsereci
Amazon Research
Cambridge, UK
mahserec@amazon.com

Abstract

Bayesian quadrature (BQ) is a sample-efficient probabilistic numerical method to solve integrals of expensive-to-evaluate black-box functions, yet so far, *active* BQ learning schemes focus merely on the integrand itself as information source, and do not allow for information transfer from cheaper, related functions. Here, we set the scene for active learning in BQ when multiple related information sources of variable cost (in input and source) are accessible. This setting arises for example when evaluating the integrand requires a complex simulation to be run that can be approximated by simulating at lower levels of sophistication and at lesser expense. We construct meaningful cost-sensitive multi-source acquisition *rates* as an extension to common utility functions from vanilla BQ (VBQ), and discuss pitfalls that arise from blindly generalizing. In proof-of-concept experiments we scrutinize the behavior of our generalized acquisition functions. On an epidemiological model, we demonstrate that active multi-source BQ (AMS-BQ) allocates budget more efficiently than VBQ for learning the integral to a good accuracy.

1 INTRODUCTION

Integrals of expensive-to-evaluate functions arise in many scientific and industrial applications, for example when expected values need to be computed and each evaluation of the integrand requires the run of a complex computer simulation where an input is only known by its distribution e.g., in meteorology, astrophysics, fluid dynamics, biology, operations research, et cetera. This complex simulation could be a Monte Carlo simulation, a finite-element or finite-volume simulation, or a stochastic model. Within reasonable budget, integration using Monte Carlo may not be feasible and alternative numerical integration schemes are needed that require fewer function evaluations.

Bayesian quadrature (BQ)—a means of constructing posterior measures over the unknown value of the integral (O'Hagan, 1991; Diaconis, 1988; Briol et al., 2019)—mitigates a high sample demand by encoding known or assumed structure of the integrand such as smoothness or regularity, usually via a Gaussian process (GP). With its increased 'data'[1] efficiency, BQ is a natural choice when function evaluations are precious (Rasmussen & Ghahramani, 2003). In the past, BQ has been applied in reinforcement learning (Paul et al., 2018), filtering (Kersting & Hennig, 2016), and has been extended to probabilistic integrals (Osborne et al., 2012a; Osborne et al., 2012b; Gunter et al., 2014; Chai & Garnett, 2019).

A complementary approach to sample efficiency is to make use of related, cheaper *secondary* information sources. The task of finding approximations to computationally demanding numerical models is an area of active research all on its own (see e.g., Benner et al., 2017). Secondary information sources of reduced cost and quality include numerical models that are run at a lower resolution (e.g., a coarser grid in a fluid dynamics application), model simplifications by neglecting details or by using an approximate model that is easier to solve numerically, and analytic approximations. A primary source could be an elaborate Earth system model to simulate anthropogenic

*work primarily performed during an internship at Amazon Research, Cambridge, UK.

[1] See e.g., Hennig et al. (2015) and Cockayne et al. (2017) for a discussion on 'data' in numerical solvers.

climate change. There exist a plethora of such models and secondary sources might parametrize important effects like albedo or neglect detailed land surface processes or ocean biogeochemistry (Flato et al., 2013).

Multi-source modeling is a statistical technique for harvesting information from related functions by constructing correlated surrogates over multiple sources. When the information sources are hierarchical in that they are ordered from most to least informative, this concept is known as multi-fidelity modeling (Kennedy & O'Hagan, 2000; Peherstorfer et al., 2018; Forrester et al., 2007; Le Gratiet & Garnier, 2014). The notion of *multi-source* models is more overarching and includes settings in which sources do not exhibit an easily identifiable order, if any. Each of the sources has its own *cost* function that quantifies the cost of evaluating the source at a certain input. An input-dependent cost might arise when the simulation run to query the integrand needs to be refined for certain values of the input to ensure numerical stability. A linear instance of a multi-source model is a multi-output GP aka. co-kriging (Alvarez et al., 2012). BQ with multi-output GPs to integrate several related functions has been studied by Xi et al. (2018), who impose properties on data—which they assume given—to prove theoretical guarantees and consistency of the Bayes estimator.

BQ leverages active learning schemes similar to Bayesian optimization (Shahriari et al., 2016) or experimental design (Atkinson et al., 2007; Yu et al., 2006). Through its argmax, an acquisition function identifies optimal future locations to query the integrand according to a user-defined metric. Metrics of interest in BQ are information gain on the value of the integral or its predictive variance. By optimizing the target *per cost*, active multi-source BQ (AMS-BQ) trades off improvement on the target (the integral) and ressources spent. In Bayesian optimization, this setting has been explored by Poloczek et al. (2017).

We summarize our contributions:

- We lay the foundations for active BQ for the task of integrating an expensive function that comes with cheaper approximations. We assign cost to function evaluations and generalize VBQ acquisition functions to acquisition *rates* that trade off improvement on the integral against cost.
- We find that some rates induce sane, others pathological acquisition policies. Pathologies were not present in the common VBQ acquisition schemes that all give rise to the same degenerate policy, regardless of the acquisition's *value*. Cost-adapted rate policies do depend on these values and are thus intricately tied to the meaning of the acquisition function that encodes progress on the quadrature task. Simply put, *all* considered (even pathological) multi-source acquisition policies collapse onto a single policy for VBQ, as a corner case of AMS-BQ.
- We conduct proof-of-concept experiments which show that AMS-BQ improves upon VBQ in that it spends less budget on learning the integral to an acceptable precision.

2 MODEL

We wish to estimate the integral over the information source of interest (the *primary* source), w.l.o.g. indexed by 1, $f_1 : \Omega \mapsto \mathbb{R}$, $\mathbf{x} \mapsto f_1(\mathbf{x})$ and integrated against the probability measure π on $\Omega \subseteq \mathbb{R}^D$,

$$\langle f_1 \rangle =: \int_\Omega f_1(\mathbf{x}) \, \mathrm{d}\pi(\mathbf{x}) \tag{1}$$

in presence of $L-1$ not necessarily ordered or orderable *secondary* information sources $f_2, \dots, f_L$, with $f_l : \Omega \mapsto \mathbb{R}$. Each source $l \in \mathcal{L} = \{1, \dots, L\}$ comes with an input-dependent cost $c_l(\mathbf{x})$ which must be invested to query f_l at location $\mathbf{x}$. For ease of interpretation and numerical stability we set $c : \mathcal{L} \times \Omega \mapsto [\delta, 1]$ and $0 < \delta \leq 1$. This is equivalent to assuming there exists a $c_{\min} > 0$ and a $c_{\max} < \infty$ s.t. $c_{\min} \leq c_l(\mathbf{x}) \leq c_{\max}$ and then normalizing w.r.t. $c_{\max}$, i.e., $\delta = \frac{c_{\min}}{c_{\max}}$. In other words, no query takes an infinite amount of resources, nor does any evaluation come for free. Normalization is not required and in practice, neither $c_{\max}$ nor $c_{\min}$ need to be known.

In this section we review the tools for building a statistical model that allows us to harvest information from both the primary and the secondary sources for learning the integral $\langle f_1 \rangle$ of Eq. (1), before turning to the decision theoretic problem of how to actively select locations and sources to query next in Section 3.

2.1 VANILLA BQ

Let $f : \Omega \mapsto \mathbb{R}$, $\mathbf{x} \mapsto f(\mathbf{x})$ be a function and π a probability measure on $\Omega \subseteq \mathbb{R}^D$ that has an intractable integral $\langle f \rangle = \int_\Omega f(\mathbf{x}) \, \mathrm{d}\pi(\mathbf{x})$. In vanilla Bayesian quadrature (VBQ), we express our epistemic uncertainty about the value of $\langle f \rangle$ through a random variable Z. The distribution over Z is obtained by integrating a Gaussian process (GP) prior that is placed over the integrand f, i.e. $f \sim \mathcal{GP}(m, k)$, where $m : \Omega \mapsto \mathbb{R}$, $\mathbf{x} \mapsto m(\mathbf{x})$ denotes the prior mean function and $k : \Omega \times \Omega \mapsto \mathbb{R}$, $(\mathbf{x}, \mathbf{x}') \mapsto k(\mathbf{x}, \mathbf{x}')$ the covariance function or kernel. Observations come in form of potentially noisy function evaluations[2] $y = f(\mathbf{x}) + \epsilon$

with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Let $\mathbf{X}$ denote the matrix of N input locations $\mathbf{X} = [\mathbf{x}_1 \ \dots \ \mathbf{x}_N]^\intercal$ and $\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\epsilon}$ the set of corresponding observations, summarized in $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ (see Rasmussen & Williams, 2006 for an introduction to GP inference). With the closure property of GPs, the posterior over Z when conditioning on $\mathcal{D}$ is a univariate Gaussian distribution with posterior mean $\mathbb{E}[Z \,|\, \mathcal{D}] = \langle m_\mathcal{D} \rangle$ and variance $\mathbb{V}[Z \,|\, \mathcal{D}] = \int_\Omega \int_\Omega k_\mathcal{D} \, \mathrm{d}\pi(\mathbf{x}) \, \mathrm{d}\pi(\mathbf{x}') =: \langle\langle k_\mathcal{D} \rangle\rangle$ that are integrals over the GP's posterior mean $m_\mathcal{D}(\mathbf{x})$ and covariance $k_\mathcal{D}(\mathbf{x}, \mathbf{x}')$. These expressions are detailed below for the general multi-source case that VBQ is a subset of and further derivations can be found in Briol et al. (2019). So as not to replace an intractable integral by another intractable integral, the kernel $k(\mathbf{x}, \mathbf{x}')$ is chosen to be integrable against $\pi(\mathbf{x})$.

2.2 MULTI-SOURCE MODELS

We consider linear multi-source models, which can equally be phrased as multi-output Gaussian processes (Alvarez et al., 2012) over the vector-valued function $\boldsymbol{f} = [f_1, \dots, f_L]$, $\boldsymbol{f} : \Omega \mapsto \mathbb{R}^L$. Non-linear models for multi-source modeling exist and have been considered by Perdikaris et al. (2017). They do however come with the additional technical difficulty that the model may not be integrable analytically—a sensible pre-requisite for BQ—and are thus another beast altogether. The notation mimics the single-output case, that is, $\boldsymbol{f} \sim \mathcal{GP}(\mathbf{m}, \mathbf{K})$, where $\mathbf{K}$ is an $L \times L$ matrix-valued covariance function. More precisely, the covariance between two sources f_l and $f_{l'}$ at inputs $\mathbf{x}$ and $\mathbf{x}'$ is $\mathrm{cov}[f_l(\mathbf{x}), f_{l'}(\mathbf{x}')] = k_{ll'}(\mathbf{x}, \mathbf{x}')$. The kernel $k_{ll'}(\mathbf{x}, \mathbf{x}')$ encodes not only characteristics of the individual sources (e.g., smoothness), but crucially the correlation between them. In the multi-source setting, observations come in source-location-evaluation triplets $(l, \mathbf{x}, y_l)$ with $y_l = f_l(\mathbf{x}) + \epsilon_l$ and source-dependent observation noise $\epsilon_l \sim \mathcal{N}(0, \sigma_l^2)$ as usually only one element of $\boldsymbol{f}$ is being observed (see supplementary material (supp. mat.) for alternative representation as linear observations).

The dataset $\mathcal{D} = \{\boldsymbol{\ell}, \mathbf{X}, \mathbf{y}_{\boldsymbol{\ell}}\}$ contains N data triplets from evaluating elements of $\boldsymbol{f}$ at N locations $\mathbf{X} = [\mathbf{x}_1 \ \dots \ \mathbf{x}_N]^\intercal$ with corresponding sources $\boldsymbol{\ell} = [l_1 \ \dots \ l_N]^\intercal$ and observations $\mathbf{y}_{\boldsymbol{\ell}} = [f_{l_1}(\mathbf{x}_1) + \epsilon_{l_1} \ \dots \ f_{l_N}(\mathbf{x}_N) + \epsilon_{l_N}]^\intercal$. The GP posterior over $\boldsymbol{f}$ has mean and covariance

$$
\begin{aligned}
m_{l|\mathcal{D}}(\mathbf{x}) &= m_l(\mathbf{x}) + \mathbf{k}_{l\boldsymbol{\ell}}(\mathbf{x}, \mathbf{X}) \mathbf{G}_{\boldsymbol{\ell}}(\mathbf{X})^{-1} (\mathbf{y}_{\boldsymbol{\ell}} - \mathbf{m}_{\boldsymbol{\ell}}(\mathbf{X})), \\
k_{ll'|\mathcal{D}}(\mathbf{x}, \mathbf{x}') &= k_{ll'}(\mathbf{x}, \mathbf{x}') - \mathbf{k}_{l\boldsymbol{\ell}}(\mathbf{x}, \mathbf{X}) \mathbf{G}_{\boldsymbol{\ell}}(\mathbf{X})^{-1} \mathbf{k}_{\boldsymbol{\ell}l}(\mathbf{X}, \mathbf{x}'),
\end{aligned} \tag{2}
$$

with the kernel Gram matrix $\mathbf{G}_{\boldsymbol{\ell}}(\mathbf{X}) = \mathbf{K}_{\boldsymbol{\ell}\boldsymbol{\ell}}(\mathbf{X}, \mathbf{X}) + \boldsymbol{\Sigma}_{\boldsymbol{\ell}} \in \mathbb{R}^{N \times N}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\ell}} = \mathrm{diag}(\sigma_{l_1}^2, \dots, \sigma_{l_N}^2)$. A summary of the notation used can be found in Table 1 in the supplementary material.

2.3 MULTI-SOURCE BQ

The multi-source model of Section 2.2 can be integrated and gives rise to a quadrature rule similar to VBQ (cf. sec. 2.1). Let Z denote the random variable representing the integral of interest $\langle f_1 \rangle$ of Eq. (1). The posterior over Z given data triplets $\mathcal{D}$ is a univariate Gaussian with mean and variance

$$
\begin{aligned}
\mathbb{E}[Z \,|\, \mathcal{D}] &= \langle m_1 \rangle + \langle \mathbf{k}_{1\boldsymbol{\ell}}(\cdot, \mathbf{X}) \rangle \mathbf{G}_{\boldsymbol{\ell}}(\mathbf{X})^{-1} (\mathbf{y}_{\boldsymbol{\ell}} - \mathbf{m}_{\boldsymbol{\ell}}(\mathbf{X})), \\
\mathbb{V}[Z \,|\, \mathcal{D}] &= \langle\langle k_{11} \rangle\rangle - \langle \mathbf{k}_{1\boldsymbol{\ell}}(\cdot, \mathbf{X}) \rangle \mathbf{G}_{\boldsymbol{\ell}}(\mathbf{X})^{-1} \langle \mathbf{k}_{\boldsymbol{\ell}1}(\mathbf{X}, \cdot) \rangle,
\end{aligned} \tag{3}
$$

where $\langle \mathbf{k}_{1\boldsymbol{\ell}}(\cdot, \mathbf{X}) \rangle = \int_\Omega \mathbf{k}_{1\boldsymbol{\ell}}(\mathbf{x}, \mathbf{X}) \, \mathrm{d}\pi(\mathbf{x})$ is the kernel mean and $\langle\langle k_{11} \rangle\rangle = \int_\Omega \int_\Omega k_{11}(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\pi(\mathbf{x}) \, \mathrm{d}\pi(\mathbf{x}')$ the initial error, both of source 1. Just as in VBQ, the kernel is required to be integrable analytically.

We choose an intrinsic coregionalization model (ICM) (Alvarez et al., 2012) with kernel

$$
k_{ll'}(\mathbf{x}, \mathbf{x}') = \mathbf{B}_{ll'} \kappa(\mathbf{x}, \mathbf{x}'), \tag{4}
$$

where $\mathbf{B} \in \mathbb{R}^{L \times L}$ is a positive definite matrix. Eq. (4) is a simple extension of a standard kernel $\kappa(\mathbf{x}, \mathbf{x}')$ to the multi-source case which factors the correlation between the sources and input locations. If $\kappa(\mathbf{x}, \mathbf{x}')$ is integrable analytically, $k_{ll'}(\mathbf{x}, \mathbf{x}')$ will be, too, and thus retains the favorable property of a BQ-kernel. A typical choice for κ is the squared-exponential, aka. RBF kernel $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2 / 2\lambda^2)$ with no dependence on the sources l and l'. This model can easily be extended e.g., to a linear model of coregionalization (LMC) without challenging integrability of k. This would untie the lenthscales between sources, but would also introduce $L - 1$ additional generally unknown kernel parameters. The simpler ICM is also used by Xi et al. (2018) to establish convergence rates for a multi-output BQ rule.

3 ACTIVE LEARNING

Active learning describes the automation of the decision-making about prospective actions to be taken by an algorithm in order to achieve a certain learning objective. A heuristic measure of improvement towards the specified goal (here: learning the value of an integral) is defined through a *utility function*. It transfers the decision-theoretic problem to an optimization problem, but usually an unfeasible one. Therefore, the utility is commonly approximated by

[2] Noise free evaluations are usually assumed in BQ, but this might not be true for a black-box integrand.

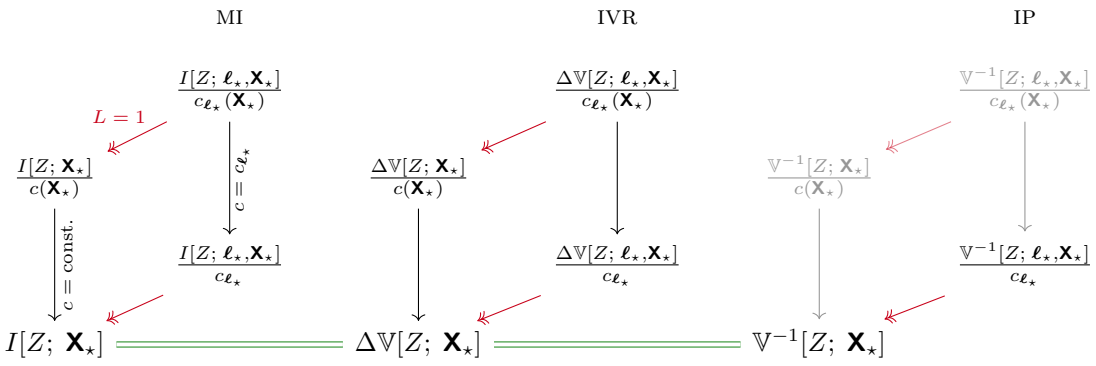

Figure 1: The multi-source acquisition-cube for a few of the possible acquisition functions. MI, IVR, and IP stand for 'mutual information', 'integral variance reduction', and 'integral precision', respectively. The forward arrows (↙) denote the special case of one source only ($L = 1$) as in the case of VBQ. The downward facing arrows (↓) denote the special case where the cost c is not dependent on the locations $\mathbf{X}_\star$. The double-lines (═) between nodes denote that these acquisition functions are equivalent in the sense that they yield the same optimal $\mathbf{X}_\star$. The two grayed-out acquisitions for IP highlight that they exhibit non-favorable behavior (cf. Section 3.2). The bottom front row in the cube denotes the special case of VBQ ($L = 1$ and $c(\mathbf{X}_\star) = \text{const.}$) where all three acquisition policies (MI, IVR, IP) coincide.

an *acquisition function*. Optimizing the acquisition function induces an *acquisition policy* that pins down what action to take next. To obtain a sequence of actions, the considered method (here: BQ) is placed in a loop where it is iteratively fed with $N_\star$ new observations $(\boldsymbol{\ell}_\star, \mathbf{X}_\star, \mathbf{y}_{\boldsymbol{\ell}_\star})$ in the general multi-step look-ahead (*non-myopic*) approach. A *myopic* approximation is to instead optimize for a single new observation triplet $(l_\star, \mathbf{x}_\star, y_{l_\star})$ at a time. Besides feasibility, the lack of exact model knowledge motivates a loop in which the model is repeatedly updated with new observations.

In Section 3.1, we recapitulate several utilities that are commonly used for VBQ. All these utilities give rise to the same acquisition policy in the absence of cost and are thus not greatly differentiated between in the literature. Intriguingly, the policies do not coincide for AMS-BQ if cost is accounted for in the acquisition functions, as will be shown and discussed in Section 3.2.

3.1 MULTI-SOURCE BQ ACQUISITIONS

In the absence of any notion of evaluation cost (or if all sources come at the same cost), the utility functions from VBQ generalize straightforwardly to the multi-source case. The VBQ case can be recovered by setting the number of sources to one.

3.1.1 Mutual Information

From an information theoretic perspective, new source-location pairs $(\boldsymbol{\ell}_\star, \mathbf{X}_\star)$ can be chosen such that they jointly maximize the mutual information (MI) $I[Z; \mathbf{y}_{\boldsymbol{\ell}_\star}]$ between the integral Z and a set of new but yet unobserved observations $\mathbf{y}_{\boldsymbol{\ell}_\star}$ with $y_{l_\star^i} = f_{l_\star^i}(\mathbf{x}_\star^i) + \epsilon_{l_\star^i}$. In terms of the individual and joint differential entropies over Z and $\mathbf{y}_{\boldsymbol{\ell}_\star}$, $I[Z; \mathbf{y}_{\boldsymbol{\ell}_\star}] = H[Z] + H[\mathbf{y}_{\boldsymbol{\ell}_\star}] - H[Z, \mathbf{y}_{\boldsymbol{\ell}_\star}]$. Sections 2.2 and 2.3 imply that both Z and $\mathbf{y}_{\boldsymbol{\ell}_\star}$ are normally distributed and so is their joint. The differential entropy of a multivariate normal distribution with covariance matrix $\mathbf{A} \in \mathbb{R}^{M \times M}$ is $H = \frac{M}{2}\log(2\pi e) + \frac{1}{2}\log|\mathbf{A}|$. Since there is no explicit dependence on the value of $\mathbf{y}_{\boldsymbol{\ell}_\star}$, we (sloppily) express the mutual information as a function of the new source-location pairs $(\boldsymbol{\ell}_\star, \mathbf{X}_\star)$,

$$I[Z; \boldsymbol{\ell}_\star, \mathbf{X}_\star] = -\frac{1}{2}\log\left(1 - \rho^2_{1\boldsymbol{\ell}_\star|\mathcal{D}}(\mathbf{X}_\star)\right), \quad (5)$$

where we introduce the scalar correlation

$$\rho^2_{1\boldsymbol{\ell}_\star|\mathcal{D}}(\mathbf{X}_\star) := \frac{\langle \mathbf{k}_{1\boldsymbol{\ell}_\star|\mathcal{D}}(\cdot, \mathbf{X}_\star)\rangle \, \mathbf{V}^{-1}_{\boldsymbol{\ell}_\star|\mathcal{D}}(\mathbf{X}_\star) \, \langle \mathbf{k}_{\boldsymbol{\ell}_\star 1|\mathcal{D}}(\mathbf{X}_\star, \cdot)\rangle}{\mathbb{V}[Z \,|\, \mathcal{D}]}, \quad (6)$$

$\in [0, 1]$, with the noise-corrected posterior covariance matrix $\mathbf{V}_{\boldsymbol{\ell}_\star|\mathcal{D}}(\mathbf{X}_\star) = \mathbf{K}_{\boldsymbol{\ell}_\star\boldsymbol{\ell}_\star|\mathcal{D}}(\mathbf{X}_\star, \mathbf{X}_\star) + \boldsymbol{\Sigma}_{\boldsymbol{\ell}_\star} \in \mathbb{R}^{N_\star \times N_\star}$. In the one-step look-ahead case ($N_\star = 1$),

$$\rho_{1l_\star|\mathcal{D}}(\mathbf{x}_\star) = \frac{\langle k_{1l_\star|\mathcal{D}}(\cdot, \mathbf{x}_\star)\rangle}{\sqrt{v_{l_\star|\mathcal{D}}(\mathbf{x}_\star)\,\mathbb{V}[Z \,|\, \mathcal{D}]}} \quad (7)$$

is the bivariate correlation between Z and $y_{l_\star}$.

3.1.2 Variance-Based Acquisitions

Variance-based approaches attempt to select $(\boldsymbol{\ell}_\star, \mathbf{X}_\star)$ such that the variance on Z shrinks maximally. As MI, the integral variance reduction (IVR) normalized by the current integral variance $\mathbb{V}[Z \mid \mathcal{D}]$ can be written in terms of correlation ρ as

$$\frac{\Delta\mathbb{V}[Z; \boldsymbol{\ell}_\star, \mathbf{X}_\star]}{\mathbb{V}[Z \mid \mathcal{D}]} = \frac{\mathbb{V}[Z \mid \mathcal{D}] - \mathbb{V}[Z \mid \mathcal{D} \cup (\boldsymbol{\ell}_\star, \mathbf{X}_\star, \mathbf{y}_{\boldsymbol{\ell}_\star})]}{\mathbb{V}[Z \mid \mathcal{D}]} = \rho^2_{1\boldsymbol{\ell}_\star|\mathcal{D}}(\mathbf{X}_\star). \tag{8}$$

Eq. (8) is a monotonic transformation of Eq. (5) and therefore, both utility functions share the same global maximizer $\mathbf{X}_\star$. In fact, any monotonic transformation of Eq. (6), whether interpretable or not, gives rise to the same acquisition policy. This is because the policy only depends on the *locations*, but not the *value* of the utility function's global maximum. Hence, in VBQ it is equivalent to consider maximal shrinkage of the variance of the integral, minimization of the integral's standard deviation, or maximal increase of the integral's precision (IP), to name a few—they all lead to the same active learning scheme and have thus not been greatly distinguished between in previous work on active VBQ.

3.2 COST-SENSITIVE ACQUISITIONS

When there is a location and/or source-dependent cost associated to evaluating the information sources (cf. Section 2), the utility function should trade off the improvement made on the integral against the budget spent for function evaluations. This is achieved by considering the ratio of a cost-insensitive BQ utility and the cost function $c_{\boldsymbol{\ell}_\star}(\mathbf{X}_\star) = \sum_{i=1}^{N_\star} c_{l_i}(\mathbf{x}_i)$. Such a ratio can be interpreted as an acquisition *rate* and bears the units of the utility function divided by units of cost. The notion of a rate becomes clearer when considering for example the mutual information utility Eq. (5) with cost measured in terms of evaluation time: the unit is $\frac{\text{bits}}{\text{second}}$, i.e., a rate of information gain.

This construction has an important consequence: Modification of the VBQ utility function (i.e., the numerator), even by a monotonic transformation, changes the maximizer of the cost-adapted acquisition rate and hence, also the acquisition *policy*. In other words, the degeneracy of BQ acquisition functions in terms of the policy they induce in the absence of cost is lifted when evaluation cost is included, firstly, because the argmax of each acquisition is shifted differently with cost, and, secondly, because acquisition *values* from different sources are discriminated against each other now. As will be discussed below, not all monotonic transformations yield a sensible acquisition policy; indeed, some display pathological behavior.

The adapted non-myopic acquisition rates for the BQ utilities mutual information (MI Eq. (5)) and integral variance reduction (IVR Eq. (8)) are

$$\alpha^{\text{MI}}_{\boldsymbol{\ell}_\star}(\mathbf{X}_\star) \coloneqq \frac{-\log\left(1 - \rho^2_{1\boldsymbol{\ell}_\star|\mathcal{D}}(\mathbf{X}_\star)\right)}{c_{\boldsymbol{\ell}_\star}(\mathbf{X}_\star)} \tag{9}$$

$$\alpha^{\text{IVR}}_{\boldsymbol{\ell}_\star}(\mathbf{X}_\star) \coloneqq \frac{\rho^2_{1\boldsymbol{\ell}_\star|\mathcal{D}}(\mathbf{X}_\star)}{c_{\boldsymbol{\ell}_\star}(\mathbf{X}_\star)}, \tag{10}$$

where we have dropped the factor $1/2$ in MI as an arbitrary scaling factor. It is evident that these acquisition rates do no longer share their maximizer; yet they still induce a meaningful acquisition scheme. Both MI and IVR have the property to be zero at $\rho^2 = 0$ and thus never select points $\mathbf{X}_\star$ that are uncorrelated with the integral Z, no matter the cost, e.g., locations that have already been observed exactly (with $\sigma^2 = 0$). Such points do not update the posterior of the integral Z when conditioned on. In VBQ these locations are the minimizers of all acquisition functions and thus excluded no matter their value. This is not ensured for the cost-adapted acquisition rates and therefore, they additionally require the numerator to be zero at $\rho^2 = 0$. Hence, not every monotonic transformation of the BQ utility produces a sane acquisition policy in the presence of cost.

Consider for example the valid transformation $\rho^2 \mapsto \rho^2 - 1$, which is -1 at $\rho = 0$. Maximizing this utility function corresponds to maximizing the negative integral variance, i.e., minimizing the integral variance, which is very commonly done in VBQ. Since $\rho^2 \in [0, 1]$, $\rho^2 - 1$ is negative everywhere and gets larger (takes a smaller negative value) with larger cost. Hence when maximized, this acquisition would favor expensive evaluations.

More subtle is the misbehavior of the integral precision (IP) which is positive everywhere and has the desired behavior of favoring low-cost evaluations. In terms of the squared correlation $\rho^2 \in [0, 1]$ from Eq. (6) (with simplified notation for convenience), the numerator of the IP acquisition rate can be written as $(1 - \rho^2)^{-1}$. This expression is non-zero at $\rho^2 = 0$ and therefore, it does not exclude points of zero correlation when they come at sufficiently cheap cost, and in experiments we observe it getting stuck re-evaluating at the location of minimum cost ad infinitum. We conjecture that this is because IP only encodes an absolute scale of the integral variance but

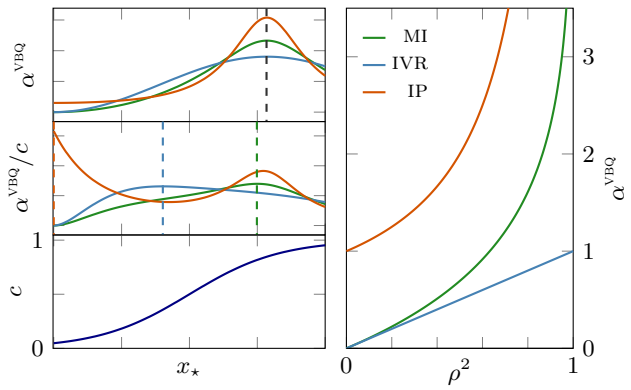

Figure 2: *right:* VBQ acquisitions α^{VBQ} as functions of the squared correlation $\rho^2(\mathbf{X}_\star)$; *left:* VBQ acquisitions as a function of univariate $x_\star$ and myopic step ($N_\star = 1$) for a synthetic $\rho^2(x_\star)$. Without cost, their maximizers coincide (top), but when divided by an input-dependent cost $c(x_\star)$ (bottom), the maximizers disperse (indicated by the dashed vertical lines) (middle). For implications, cf. Section 3.2.

does not quantify any "improvement" on the integral value.

Figure 1 illustrates the augmentation of utility functions from VBQ with multiple information sources and cost. Figure 2 displays the behavior of a few acquisitions, MI, IVR, and IP. The right plot shows these acquisitions as used in VBQ in terms of the squared correlation $\rho^2 \in [0, 1]$ (Eq. (6)) in the absence of cost. All acquisitions are strictly monotonically increasing functions of ρ^2. Among the sane acquisition rates that are zero at $\rho^2 = 0$, the differences in the corresponding policy can also be understood from the functional dependence on ρ. MI diverges at perfect correlation $\rho^2 \to 1$. Therefore, and since the cost c lies in $[\delta, 1]$, MI will always take a 'perfect step' to learn the integral exactly, i.e., it will always select the points $\mathbf{X}_\star$ with correlation $\rho^2(\mathbf{X}_\star) = 1$, if the step is available and no matter the cost. IVR, however, is finite at $\rho^2 = 1$ and trades off cost against correlation even if the perfect $\mathbf{X}_\star$ with $\rho^2(\mathbf{X}_\star) = 1$ exists. These interpretations are reinforced by the left three plots of Figure 2, in which we plot MI, IVR, and IP versus a univariate $x_\star$ for the synthetic choice $\rho^2(x_\star) = 0.95 \sin^2(10x_\star)$, $x_\star \in [0, 0.2]$ and a myopic step ($N_\star = 1$). In the pure VBQ situation, the locations of all their maxima coincide, but as soon as a non-constant cost $c(x_\star)$ is applied, the shapes of the acquisition functions become relevant which discriminates their $\mathbf{X}_\star$ and lifts the degeneracy in policies. MI tends more towards higher correlation than IVR, the maximizer of which moves further towards locations of lower cost. While MI and IVR act differently, they are both sensible choices for acquisition functions in AMS-BQ. In fact for low to mid-ranged values of $\rho^2 \lesssim 0.5$ where MI is approximately a linear function of ρ^2 they roughly coincide.

The choice of acquisition ultimately depends on the application and the user, who may choose which measures of improvement on the integral and cost to trade off.

4 EXPERIMENTS

The key practical applications for AMS-BQ is solving integrals of expensive-to-evaluate black-box functions that are accompanied by cheaper approximations, potentially in a setting where a finite budget is available. Typical applications are models of complex nonlinear systems that need to be tackled computationally. With evaluations being precious, the goal is to get a decent estimate of the integral with as little budget as possible, rather than caring about floating-point precision. In the experiments, we focus on the rear vertices of the acquisition cube Figure 1, i.e., multiple sources with source and input-dependent or only source-dependent cost, and separate them into two main experiments:

1. A synthetic multi-source setting with cost that varies in source and location for the purpose of exploring and demonstrating the behavior of the acquisition functions derived in Section 3.
2. An epidemiological model of the spread of a disease with uncertain input, in which two sources correspond to simulations that differ in cost as well as quality of the quantity of interest.

Additionally, we present a bivariate experiment with three sources in the supp. mat. Section D. We take a myopic approach to all scenarios in that we optimize the acquisition for a single source-location pair a time. The implementation of the GP-model uses `GPy` (`GPy`, since 2012) in `Python 3.7`.

4.1 MULTI-SOURCE, VARIABLE COST

We initially consider a synthetic two-source setting with univariate input. The cost functions depend on both source and location. The experiment's purpose is to demonstrate our findings from Section 3 and convey intuition about the behavior of the novel acquisition functions. The sources we consider have been suggested by Forrester et al. (2007) with the primary source $f_1(x) = (6x-2)^2 \sin(12x-4)$ and the secondary source $f_2(x) = \frac{1}{2} f_1(x) + 10x$ for $x \in [0, 1]$.

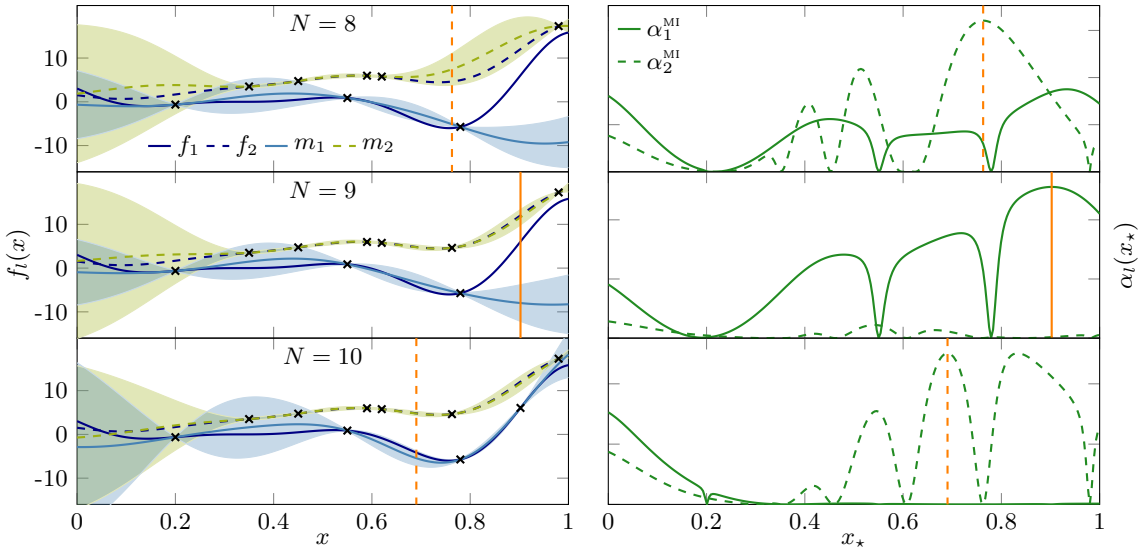

Figure 3: Demonstration of the sequential selection of new source-location pairs to query $\boldsymbol{f}$ using the MI acquisition in a two-source setting with source and location dependent cost (Figure 8). *Left column:* The GP; *right column:* the acquisition function for the primary (solid) and secondary source (dashed) for three consecutive iterations. Vertical orange lines indicate the location and source of the new query.

The cost functions both take the form of a scaled and shifted logistic function in a way that the cost lies in $(0, 1]$ (cf. Figure 8 in the supp. mat.). The costs of both sources converge to the same value close to $x_\star = 0$; for larger $x_\star$, f_2 is two orders of magnitude cheaper than f_1. Figure 3 shows snapshots of three consecutive query decisions taken by the MI multi-source acquisition. The GP model (depicted in the left column) has been initialized with 3 datapoints in the primary and 5 in the secondary source and merely the noise variance was constrained to 10^{-2}. The MI acquisition given the current state of the GP is shown on the right—the top left frame is shown for MI, IVR, and IP in Figure 7 in the supp. mat. to emphasize the pathology of IP and to highlight the subtle difference between MI and IVR in practice. The acquisition function is optimized using the L-BFGS-B optimizer in `scipy`. We observe that AMS-BQ does not query f_2 where the source costs are almost identical for $x_\star \lesssim 0.2$ (see Figure 9 in supp. mat.). This is because the two sources are not perfectly correlated and evaluating f_1 always conveys more information about Z than f_2. The fact that c_2 decreases with increasing $x_\star$ is nicely represented in the increasing height of the maxima of the dashed acquisition function for the secondary source in the top left frame of Figure 3.

For assessing the performance of AMS-BQ, we compare against VBQ and a percentile estimator (PE) that both operate on the primary source. The latter is obtained by separating the domain into intervals that contain the same probability mass and summing up the function values at these nodes. For the uniform integration measure used here, this is equivalent to a right Riemann sum. We assume that GP inference comes at negligible cost as compared to the function evaluations and thus consider cost to be incurred purely by querying the information sources.

To render the integration problem more difficult, we modify the Forrester functions to vary on a smaller length scale by adding a sinusoidal term and adapting some parameters, s.t. $f_1(x) = (6x-2)^2 \sin(12x-4) - (2-x)^2 \sin(36x)$ and $f_2(x) = \frac{3}{4} f_1(x) + 16\left(x - \frac{1}{2}\right) + 10$ which we integrate from 0 to 1 against a uniform measure (cf. Figure 4, top left). To avoid over- or underfitting, we set a conservative gamma prior on the lengthscale with a mode at a small fraction of the domain $[0, 1]$ for both VBQ and AMS-BQ, and assume zero observation noise. With six[3] more hyperparameters than VBQ, AMS-BQ is more prone to over-/underfitting, and we further set a prior on the coregionalization matrix $\mathbf{B}$ (cf. Section 2.3) with parameters estimated from the initial three data points using empirical Bayes. This is to avoid initial over- or under-estimation of the correlation between sources, which would either cause the active scheme to select only f_2 or only f_1, respectively. Compared to the

[3]Due to the construction of $\mathbf{B} = \mathbf{W}\mathbf{W}^\intercal + \mathrm{diag}(\boldsymbol{\eta})$

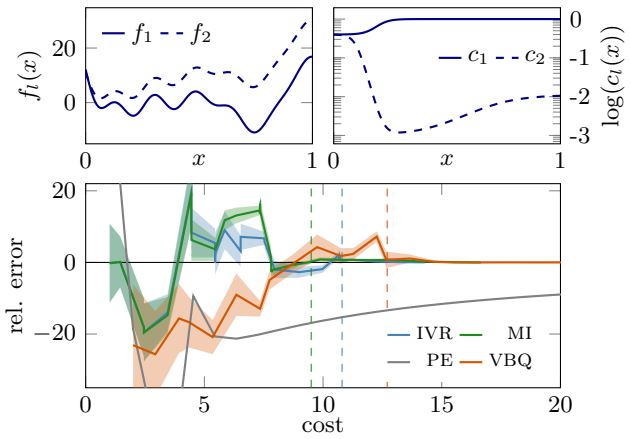

Figure 4: *Top left:* the wigglified Forrester functions with f_1 and f_2 primary/secondary source, respectively; *top right:* the cost functions used; *bottom:* relative error $\mathbb{E}[Z]-\langle f_1\rangle/\langle f_1\rangle$ with two std. deviations (shaded) as a function of normalized cost for the AMS-BQ acquisitions MI and IVR compared to VBQ and a percentile estimator (PE). Vertical dashed lines are a visual help to indicate the cost spent to achieve acceptable accuracy.

previous experiment, the cost is changed to have a minimum, but still composed of a sum of logistic functions and normalized to be in $(0, 1]$ (Figure 4, top right). The effect of these cost functions on the final state is depicted in the supp. mat., Figure 10. Furthermore, this setting reveals the pathology of the IP acquisition (cf. Section 3.2) that everlastingly re-evaluates the secondary source at the location of minimal cost. The convergence behavior of the well-behaved acquisition functions MI and IVR are displayed in Figure 4 (bottom) in comparison to VBQ and PE. The hyperparameters of the GP are optimized after every newly acquired node, both for VBQ and AMS-BQ. Figure 4 shows the superior performance of both AMS-BQ methods in arriving close the true integral with little budget. The vertical jumps in the AMS-BQ methods occur when f_2 is evaluated at cheaper cost.

4.2 A SIMULATION OF INFECTIONS

We now consider multi-source models in which sources come with input-independent cost, a.k.a. multi-fidelity models (bottom rear MI vertex in Figure 1). We choose an epidemiological model in which evaluating the primary source requires running numerous stochastic simulations and the secondary source solves a system of ordinary differential equations. Epidemiological models deal with simulating the propagation of an infectious disease through a population. The SIR model forms the base for many compartmental models and assumes a population of fixed size N where at any point in time, each individuum is in one of three states—susceptible, infected, and recovered (SIR)—with sizes N_S, N_I, and N_R (Kermack & McKendrick, 1927). The dynamics are determined by stochastic discrete-time events of individuals changing infection state, for which Poisson processes (i.e., exponentially distributed inter-event times) are commonly assumed (see e.g., Daley & Gani, 1999). In the thermodynamic limit where N is large, the average dynamics is governed by a system of ODEs that does not admit a generic analytic solution. There are two parameters in the SIR model: the infection rate a, and the recovery rate b. Model details and experiment setup can be found in Section C (supp. mat.).

For the AMS-BQ experiment, we assume that we know b, but we are uncertain about a. We are interested in the expected maximum number of simultaneously infected individuals $\mathbb{E}_a[\max_t N_I(t)]$ and the time this maximum occurs $\mathbb{E}_a[\arg\max_t N_I(t)]$, which might be relevant for vaccination planning. Querying the primary source f_1 for the quantities of interest as a function of a requires numerous realizations of a stochastic four-compartments epidemic model (an extension to the SIR model) using the Gillespie algorithm (Gillespie, 1976; Gillespie, 1977). For each trajectory, the maximum value and time are computed and henceforth averaged over. In our implementation, each query of f_1 takes $\sim 16\,\mathrm{s}$ on a laptop's CPU. The secondary source f_2 solves the system of ODEs for given a and computes the maximum value and time for the resulting function $N_I(t)$, which takes about $8 \cdot 10^{-3}$s to evaluate. As in previous experiments, we set a gamma prior on the lengthscale, a prior on the coregionalization matrix $\mathbf{B}$, and the noise variance to zero as in Section 4.1. Both VBQ and AMS-BQ are given the same initial value of f_1, and AMS-BQ additionally gets the value of f_2 at the same location, as well as one more random datum from f_2. This is justified since AMS-BQ needs to learn more hyperparameters than VBQ and secondary source evaluations are very cheap. Otherwise, if the initial evaluations of f_2 were further apart than the prior lengthscale from the locations of the initial primary datum, virtually zero correlation would be inferred between the sources, and the primary source would be evaluated until a sampled location roughly coincides with locations where the secondary sources have been evaluated.

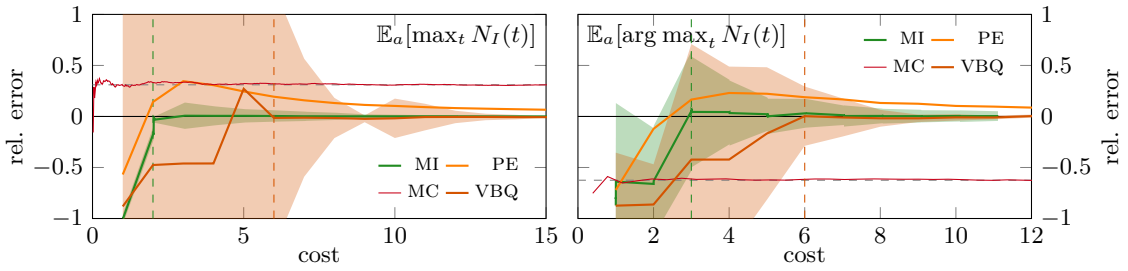

Figure 5: Relative error vs. budget spent for the SIR model for the max number of simultaneously infected individuals (left) and for the time after which the maximum occurs (right). Primary source has cost 1.

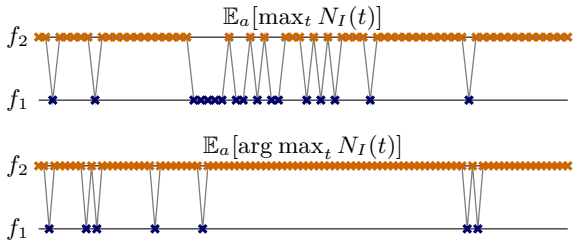

Figure 6: Evaluation sequences of primary and secondary source in the SIR experiment.

Figure 5 shows the relative error of the AMS-BQ estimator against normalized cost as compared to VBQ and PE for $\mathbb{E}_a[\max_t N_I(t)]$ (left) and $\mathbb{E}_a[\arg\max_t N_I(t)]$ (right). The horizontal dashed line shows $\langle f_2 \rangle$, i.e., the integral of the secondary source with one evolution of a Monte Carlo estimator of f_2. This illustrates that simply using the secondary source for the integral estimate might be computationally cheap, but results in an unknown bias. In the left plot, AMS-BQ achieves a good estimate with one additional evaluation of f_1 only, while VBQ takes another six evaluations. Again, the vertical jumps for AMS-BQ are caused by evaluations of f_2. The initial high confidence on the integral is caused by the choice of prior on the output scale from the initial data, which is located in the tail of the gamma prior on a. Figure 6 displays the order in which AMS-BQ evaluates primary and secondary source.

5 DISCUSSION

The multi-source model presented in Section 2.2 can be extended in various ways to increase its expressiveness by using a more elaborate kernel (e.g., one lengthscale per source), or by encoding knowledge about the functions to be integrated, e.g., a probabilistic integrand. Other applications might come with the complication that the cost function c is unknown a priori and needs to be learned during the active BQ-loop from measurements of the amount of resource required during the queries. A simple example was presented in Section 4.2 where the cost was parameterized by a constant, estimated during the initial observations. A probabilistic (in contrast to parametric) model upon the cost would induce an acquisition function which is not only conditioned on the uncertain model predictions but also on the uncertain cost predictions. Furthermore, as in other active learning schemes, non-myopic steps for acquiring multiple observations $\mathbf{y}_{\ell_\star}$ at once might be beneficial especially when the multi-source model is already known, and does not benefit from being re-fitted to new data; or when multiple evaluations of sources come at lower cost than evaluating sequentially. On the experimental side, more elaborate applications of AMS-BQ in areas of active research are reserved for future work.

5.1 CONCLUSION

We have placed multi-source BQ in a loop and thus enabled active learning to infer the integral of a primary source while including information from cheaper secondary sources. We discovered that utilities that yield redundant acquisition policies in VBQ give rise to various policies, some desirable and others pathological, when evaluation cost is accounted for. Our experiments illustrate that with the sensible acquisition functions, the AMS-BQ algorithm allocates budget to information retrieval more efficiently than traditional methods do for solving expensive integrals.

Acknowledgements

We thank Andrei Paleyes for software-related support, as well as Philipp Hennig, Motonobu Kanagawa, and Aaron Klein for useful discussions. AG acknowledges support by the IMPRS-IS.

References

Alvarez, M. A., L. Rosasco, & N. D. Lawrence (2012). "Kernels for vector-valued functions: a review". In: *Foundations and Trends® in Machine Learning* 4.3, pp. 195–266.

Atkinson, A., A. Donev, & R. Tobias (2007). *Optimum experimental designs, with SAS*. Oxford Statistical Science Series. OUP Oxford.

Benner, P., A. Cohen, M. Ohlberger, & K. Willcox (2017). *Model reduction and approximation: theory and algorithms*. Vol. 15. SIAM.

Briol, F.-X., C. J. Oates, M. Girolami, M. A. Osborne, & D. Sejdinovic (Feb. 2019). "Probabilistic integration: a role in statistical computation?" In: *Statistical Science* 34.1, pp. 1–22.

Chai, H. & R. Garnett (2019). "Improving quadrature for constrained integrands". In: *Proceedings of Machine Learning Research*. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 2751–2759.

Cockayne, J., C. Oates, T. Sullivan, & M. Girolami (2017). "Bayesian probabilistic numerical methods". In: *Arxiv:1702.03673 [stat.me]*.

Daley, D. J. & J. Gani (1999). *Epidemic modelling: an introduction*. Cambridge Studies in Mathematical Biology. Cambridge University Press.

Diaconis, P. (1988). "Bayesian numerical analysis". In: *Statistical Decision Theory and Related Topics IV* 1, pp. 163–175.

Flato, G. et al. (2013). "Evaluation of climate models". In: *Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. Chap. 9, pp. 741–866.

Forrester, A. I. J., A. Sóbester, & A. J. Keane (2007). "Multi-fidelity optimization via surrogate modelling". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 463.2088, pp. 3251–3269.

Gillespie, D. T. (1976). "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions". In: *Journal of Computational Physics* 22.4, pp. 403–434.

Gillespie, D. T. (1977). "Exact stochastic simulation of coupled chemical reactions". In: *The Journal of Physical Chemistry* 81.25, pp. 2340–2361.

`GPy` (since 2012). *GPy: a Gaussian process framework in python*. `http://github.com/SheffieldML/GPy`.

Gunter, T., M. A. Osborne, R. Garnett, P. Hennig, & S. J. Roberts (2014). "Sampling for inference in probabilistic models with fast Bayesian quadrature". In: *Advances in Neural Information Processing Systems 27*, pp. 2789–2797.

Hennig, P., M. A. Osborne, & M. Girolami (2015). "Probabilistic numerics and uncertainty in computations". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 471.2179.

Hethcote, H. W. (2000). "The mathematics of infectious diseases". In: *SIAM Review* 42.4, pp. 599–653.

Kennedy, M. C. & A. O'Hagan (2000). "Predicting the output from a complex computer code when fast approximations are available". In: *Biometrika* 87.1, pp. 1–13.

Kermack, W. O. & A. G. McKendrick (1927). "A contribution to the mathematical theory of epidemics". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 115.772, pp. 700–721.

Kersting, H. & P. Hennig (2016). "Active uncertainty calibration in Bayesian ODE solvers". In: *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016)*. AUAI Press, pp. 309–318.

Le Gratiet, L. & J. Garnier (2014). "Recursive co-kriging model for design of computer experiments with multiple levels of fidelity". In: *International Journal for Uncertainty Quantification* 4.5, pp. 365–386.

O'Hagan, A. (1991). "Bayes-Hermite quadrature". In: *Journal of Statistical Planning and Inference* 29, pp. 245–260.

Osborne, M. et al. (2012a). "Active learning of model evidence using Bayesian quadrature". In: *Advances in Neural Information Processing Systems 25*, pp. 46–54.

Osborne, M. et al. (2012b). "Bayesian quadrature for ratios". In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Vol. 22. Proceedings of Machine Learning Research. PMLR, pp. 832–840.

Paul, S. et al. (2018). "Alternating optimisation and quadrature for robust control". In: *AAAI Conference on Artificial Intelligence*.

Peherstorfer, B., K. Willcox, & M. Gunzburger (2018). "Survey of multifidelity methods in uncertainty propagation, inference, and optimization". In: *SIAM Review* 60.3, pp. 550–591.

Perdikaris, P., M. Raissi, A. Damianou, N. D. Lawrence, & G. E. Karniadakis (2017). "Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 473.2198.

Poloczek, M., J. Wang, & P. Frazier (2017). "Multi-information source optimization". In: *Advances in Neural Information Processing Systems 30*, pp. 4288–4298.

Rasmussen, C. E. & Z. Ghahramani (2003). "Bayesian Monte Carlo". In: *Advances in Neural Information Processing Systems 15*. Max-Planck-Gesellschaft. Cambridge, MA, USA: MIT Press, pp. 489–496.

Rasmussen, C. E. & C. K. I. Williams (2006). *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press, p. 248.

Shahriari, B., K. Swersky, Z. Wang, R. P. Adams, & N. de Freitas (2016). "Taking the human out of the loop: a review of Bayesian optimization". In: *Proceedings of the IEEE* 104.1, pp. 148–175.

Xi, X., F.-X. Briol, & M. Girolami (2018). "Bayesian quadrature for multiple related integrals". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 5373–5382.

Yu, K., J. Bi, & V. Tresp (2006). "Active learning via transductive experimental design". In: *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, Pennsylvania, USA: ACM, pp. 1081–1088.