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’ 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.

1See 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 resources 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, \( f_1 : \Omega \rightarrow \mathbb{R}, x \mapsto f_1(x) \) and integrated against the probability measure \( \pi \) on \( \Omega \subseteq \mathbb{R}^D \),

\[
\langle f_1 \rangle =: \int_{\Omega} f_1(x) \, d\pi(x) \quad (1)
\]

in presence of \( L - 1 \) not necessarily ordered or ordered secondary information sources \( f_2, \ldots, f_L \), with \( f_l : \Omega \rightarrow \mathbb{R} \). Each source \( l \in \mathcal{L} = \{1, \ldots, L\} \) comes with an input-dependent cost \( c_l(x) \) which must be invested to query \( f_l \) at location \( x \). For ease of interpretation and numerical stability we set \( c : \mathcal{L} \times \Omega \rightarrow [\delta, 1] \) and \( 0 < \delta \leq 1 \). This is equivalent to assuming there exists a \( c_{\text{min}} > 0 \) and a \( c_{\text{max}} < \infty \) s.t. \( c_{\text{min}} \leq c_l(x) \leq c_{\text{max}} \) and then normalizing w.r.t. \( c_{\text{max}} \), i.e., \( \delta = \frac{c_{\text{min}}}{c_{\text{max}}} \). 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 \( c_{\text{max}} \) nor \( c_{\text{min}} \) 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 \( f : \Omega \rightarrow \mathbb{R}, x \mapsto f(x) \) be a function and \( \pi \) a probability measure on \( \Omega \subseteq \mathbb{R}^D \) that has an intractable integral \( \langle f \rangle = \int_{\Omega} f(x) \, d\pi(x) \). In vanilla Bayesian quadrature (VBQ), we express our epistemic uncertainty about the value of \( f \) 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 \( m : \Omega \rightarrow \mathbb{R}, x \mapsto m(x) \) denotes the prior mean function and \( k : \Omega \times \Omega \rightarrow \mathbb{R}, (x, x') \mapsto k(x, x') \) the covariance function or kernel. Observations come in form of potentially noisy function evaluations \( y = f(x) + \epsilon \).
with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Let $X$ denote the matrix of $N$ input locations $X = [x_1 \ldots x_N]^{\top}$ and $y = f(X) + \epsilon$ the set of corresponding observations, summarized in $D = \{X, y\}$ (see Rasmussen & Williams, 2006 for an introduction to GP inference). With the closure property of GPs, the posterior over $Z$ when conditioning on $D$ is a univariate Gaussian distribution with posterior mean $\mathbb{E}[Z | D] = \langle m_D \rangle$ and variance $\mathbb{V}[Z | D] = \int_D \int_D k_D(\mathbf{x}) \, d\mathbf{z}' = : \langle k_D \rangle$ that are integrals over the GP’s posterior mean $m_D(\mathbf{x})$ and covariance $k_D(\mathbf{x}, \mathbf{x}')$. These expressions are detailed below for the general multi-source case that VBO 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 $f = [f_1, \ldots, f_L]$, $f : \Omega \to \mathbb{R}^L$. Non-linear models for multi-source modeling exist and have been considered by Perdikaris et al. (2017). These do however come with the additional technical difficulty that the model may not be integrable analytically—a sensible pre-requisite for VBO—and are thus another beast altogether. The notation mimics the single-output case, that is, $f \sim \mathcal{GP}(m, K)$, where $K$ 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 $\text{cov}[f_l(\mathbf{x}), f_{l'}(\mathbf{x}')] = k_{ll'}(\mathbf{x}, \mathbf{x}')$. 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 $\epsilon_l \sim \mathcal{N}(0, \sigma^2_l)$ as usually only one element of $f$ is being observed (see supplementary material (supp. mat.) for alternative representation as linear observations).

The dataset $D = \{\ell, X, y_\ell\}$ contains $N$ data triplets from evaluating elements of $f$ at $N$ locations $X = [x_1 \ldots x_N]^{\top}$ with corresponding sources $\ell = [l_1 \ldots l_N]^{\top}$ and observations $y_\ell = [f_{l_1}(x_1) + \epsilon_1 \ldots f_{l_N}(x_N) + \epsilon_N]^{\top}$. The GP posterior over $f$ has mean and covariance

$$m_D(\mathbf{x}) = m_\ell(\mathbf{x}) + k_{\ell}(\mathbf{x}, \mathbf{x}) G_\ell(\mathbf{x})^{-1}(y_\ell - m_\ell(\mathbf{x})),
$$

$$k_{D}(\mathbf{x}, \mathbf{x}') = k_{ll'}(\mathbf{x}, \mathbf{x}') - k_{\ell}(\mathbf{x}, \mathbf{x}) G_\ell(\mathbf{x})^{-1} k_{\ell}(\mathbf{x}, \mathbf{x}'),$$

(2)

with the kernel Gram matrix $G_\ell(X) = K_{\ell}(X, X) + \Sigma_\ell \in \mathbb{R}^{N \times N}$ and $\Sigma_\ell = \text{diag}(\sigma^2_1, \ldots, \sigma^2_N)$. 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 VBO (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 $D$ is a univariate Gaussian with mean and variance

$$E[Z | D] = \langle m_1 \rangle + (k_{l_1}(\mathbf{x}, \mathbf{x})) G_1(X)^{-1}(y_1 - m_1(\mathbf{x})),
$$

$$\mathbb{V}[Z | D] = \mathbb{V}[k_{l_1} | D] = \mathbb{V}[k_{l_1}] = \int_D (k_{l_1}(\mathbf{x}, \mathbf{x})) G_1(X)^{-1}(k_{l_1}(\mathbf{x}, \mathbf{\cdot})),$$

(3)

where $(k_{l_1}(\mathbf{x}, \mathbf{x})) = \int_D k_{l_1}(\mathbf{x}, \mathbf{x}') \, d\pi(\mathbf{x})$ is the kernel mean and $\mathbb{V}[k_{l_1}] = \int_D (k_{l_1})^{\top} G_1(X)(k_{l_1}) \, d\pi(\mathbf{x})'$ the initial error, both of source 1. Just as in VBO, the kernel is required to be integrable analytically.

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

$$k_{ll'}(\mathbf{x}, \mathbf{x}') = B_{ll'} \kappa(\mathbf{x}, \mathbf{x}'),$$

(4)

where $B \in \mathbb{R}^{L \times L}$ 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 $\kappa$ is the squared-exponential, aka. RBF kernel $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-l^2 \|x - x'\|_2^2)$ 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 lengthscales 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
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* new observations \((l_*, x_*, y_{l_*})\) in the general multi-step look-ahead (non-myopic) approach. A myopic approximation is instead optimize for a single new observation triplet \((l_*, x_*, y_{l_*})\) 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 \((l_*, x_*)\) can be chosen such that they jointly maximize the mutual information (MI) \(I[Z; y_{l_*}]\) between the integral Z and a set of new but yet unobserved observations \(y_{l_*}\) with \(y_{l_*} = f_{l_*}(x_*) + \epsilon_{l_*}\). In terms of the individual and joint differential entropies over Z and \(y_{l_*}\), \(I[Z; y_{l_*}] = \mathcal{H}(Z) + \mathcal{H}(y_{l_*}) - \mathcal{H}[Z, y_{l_*}]\). Sections 2.2 and 2.3 imply that both Z and \(y_{l_*}\) are normally distributed and so is their joint. The differential entropy of a multivariate normal distribution with covariance matrix \(A \in \mathbb{R}^{M \times M}\) is \(\mathcal{H} = \frac{1}{2} \log(2\pi e) + \frac{1}{2} \log|A|\). Since there is no explicit dependence on the value of \(y_{l_*}\), we (sloppily) express the mutual information as a function of the new source-location pairs \((l_*, x_*)\),

\[
I[Z; l_*, x_*] = -\frac{1}{2} \log \left(1 - \rho^2_{l_*, D}(x_*)\right),
\]

where we introduce the scalar correlation

\[
\rho^2_{l_*, D}(x_*) := \frac{\langle k_{l_*, D}(x_*) \rangle V_{l_*, D}(x_*)^\dagger \langle k_{l_*, D}(x_*) \rangle}{\sqrt{\mathcal{H}[Z | D] V_{l_*, D}(x_*)^\dagger V_{l_*, D}(x_*)}},
\]

\(\in [0, 1]\), with the noise-corrected posterior covariance matrix \(V_{l_*, D}(x_*) = K_{l_*, D}(x_*, x_*) + \Sigma_{l_*} \in \mathbb{R}^{N_* \times N_*}\). In the one-step look-ahead case \((N_* = 1)\),

\[
\rho_{l_*, D}(x_*) = \frac{\langle k_{l_*, D}(x_*) \rangle}{\sqrt{\mathcal{H}[Z | D] V_{l_*, D}(x_*)}},
\]
is the bivariate correlation between $Z$ and $y_l$.

### 3.1.2 Variance-Based Acquisitions

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

$$\frac{\Delta \mathbb{V}[Z; \ell_*, \mathbf{x}_*]}{\mathbb{V}[Z | \mathcal{D}]} = \frac{\mathbb{V}[Z | \mathcal{D}] - \mathbb{V}[Z | \mathcal{D} \cup (\ell_*, \mathbf{x}_*, y_{\ell_*})]}{\mathbb{V}[Z | \mathcal{D}]} = \rho^2_{\ell_*, |\mathcal{D}|} (\mathbf{x}_*).$$  \hspace{1cm} (8)

Eq. (8) is a monotonic transformation of Eq. (5) and therefore, both utility functions share the same global maximizer $\mathbf{x}_*$. 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 been not 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 $\alpha_{\ell_*}(\mathbf{x}_*) = \sum_{i=1}^{N_s} c_i(x_i)$. 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 \textit{bits per 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 \textit{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

$$\alpha_{\ell_*}^{\text{MI}}(\mathbf{x}_*) := \frac{-\log \left( 1 - \rho^2_{\ell_*, |\mathcal{D}|}(\mathbf{x}_*) \right)}{c_{\ell_*}(\mathbf{x}_*)},$$

$$\alpha_{\ell_*}^{\text{IVR}}(\mathbf{x}_*) := \frac{\rho^2_{\ell_*, |\mathcal{D}|}(\mathbf{x}_*)}{c_{\ell_*}(\mathbf{x}_*)},$$

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 $\mathbf{x}_*$ 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 same 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
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 $\rho^2$. Among the same acquisition rates that are zero at $\rho^2 = 0$, the differences in the corresponding policy can also be understood from the functional dependence on $\rho$. MI diverges at perfect correlation $\rho^2 \to 1$. Therefore, and since the cost $c$ lies in $[0, 1]$, MI will always take a ‘perfect step’ to learn the integral exactly, i.e., it will always select the points $x_*$ with correlation $\rho^2(x_*) = 1$, 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 $x_*$ with $\rho^2(x_*) = 1$ exists. These interpretations are reinforced by the left three plots of Figure 2, in which we plot MI, IVR, and IP versus a univariate $x_*$ for the synthetic choice $\rho^2(x_*) = 0.95 \sin^2(10x_*)$, $x_* \in [0, 0.2]$ and a myopic step ($N_* = 1$). In the pure VBQ situation, the locations of all their maxima coincide, but as soon as a non-constant cost $c(x_*)$ is applied, the shapes of the acquisition functions become relevant which discriminates their $X_*$ 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 $\rho^2$ 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 2: right: VBQ acquisitions $\alpha_{\text{VBQ}}$ as functions of the squared correlation $\rho^2(X_*)$; left: VBQ acquisitions as a function of univariate $x_*$ and myopic step ($N_* = 1$) for a synthetic $\rho^2(x_*)$. Without cost, their maximizers coincide (top), but when divided by an input-dependent cost $c(x_*)$ (bottom), the maximizers disperse (indicated by the dashed vertical lines) (middle). For implications, cf. Section 3.2.](image-url)
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 \(x_*=0\); for larger \(x_*,\) \(f_2\) 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 \(x_* \leq 0.2\) (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 \(x_*\) 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 length scale 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 \(B\) (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

\[^3\text{Due to the construction of } B = WW' + \text{diag}(\eta)\]
We now consider multi-source models in which and and are given the same initial value of and are further apart than the prior lengthscale from the locations of the initial primary and secondary source, respectively. The vertical jumps in the AMS-BQ methods occur when \( f_2 \) is evaluated at cheaper cost.

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 \( E[|Z| - (f_1)]/(f_1) \) 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 individual 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 \( E_a \max_i N_i(t) \) and the time this maximum occurs \( E_a [\arg \max_t N_i(t)] \), 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 \( \sim 16 \) 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 \( 8 \times 10^{-3} \) s to evaluate. As in previous experiments, we set a gamma prior on the lengthscale, a prior on the coregionalization matrix \( B \), 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 \( E_a[\max_t N_1(t)] \) (left) and \( E_a[\arg \max_t N_1(t)] \) (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 \( y_\ell^* \) 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


A  MULTI-SOURCE MODELS AND MULTI-OUTPUT GPS

We have seen in Section 2.2 that linear multi-source models can be phrased in terms of multi-output GPS. Typically, the goal of multi-output GPS is to model a vector-valued function and observations come as a vector \(y = f(x) + \epsilon\), where \(y \in \mathbb{R}^L\). In multi-source models, we wish to observe only elements of \(f\), i.e., \(y_l = f_l(x) + \epsilon_l\). These observations can be written as projections of the vector-valued observations,

\[ y_l = h_l^T y \]  

where \(h_l\) denotes a vector with a 1 in the \(l\)th coordinate and zero elsewhere. Let \(Y \in \mathbb{R}^{NL}\) denote the vector of \(N\) stacked vector-valued noisy observations \([y_1, \ldots, y_N]\). Then the corresponding \(N\) observations of elements \(\ell = [l_1 \ldots l_N]^T\) is

\[ y_\ell = \begin{bmatrix} h^T_{l_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & h^T_{l_N} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} =: H^T Y, \]  

where \(H\) is a sparse \(NL \times N\) matrix. Note the delicate notational difference between the \(N\) observations of single elements of \(f\), \(y_\ell \in \mathbb{R}^N\), and a single evaluation of the vector-valued function \(y \in \mathbb{R}^L\). The covariance matrix between all of the observations is

\[ \text{cov}[y_\ell, y_{\ell'}] = H^T \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_N) \\ \vdots & \ddots & \vdots \\ K(x_N, x_1) & \cdots & K(x_N, x_N) \end{pmatrix} + \Sigma \otimes 1_{N \times N} \]  

where \(\Sigma = \text{diag}(\sigma^2_1, \ldots, \sigma^2_N) \in \mathbb{R}^{L \times L}\) and \(1_{N \times N}\) is an \(N \times N\) matrix with every element a 1. Also, \(K(X, X) \in \mathbb{R}^{NL \times NL}\). With the following mappings, we arrive at the multi-source notation introduced in Section 2;

\[ H^T K_{XX} H = K_{\ell \ell}(X, X); \]
\[ H^T (\Sigma \otimes 1_{N \times N}) H = \Sigma_\ell; \]
\[ HY = y_\ell; \]  

Hence, the notational detour over vector-valued observations \(Y\) is not required and evaluations of individual sources can be incorporated easily in the multi-source model. From the mappings Eq. (14) follow the posterior mean and covariance of the multi-source model Eq. (2).

B  ADDITIONAL PLOTS FOR SECTION 4.1

Section 4.1 showed two examples to demonstrate the behavior of our derived acquisition functions. All relevant details and cross-references are in the captions.

Figure 7: MI, IVR, and IP acquisitions for the top row of Figure 3. MI and IVR do not differ a lot, i.e., the correlation \(\rho\) is rarely large enough for MI to leave the linear regime. MI puts slightly more emphasis on the primary source where \(x_s\) is close to 1. This indicates that the correlation between \(Z\) and \(y_s\) is quite large there. The bottom plot displays the pathology of IP, where the acquisition for the secondary source essentially follows the inverse cost \(c_2\).
When the population size is large, the SIR (Susceptible, Infected, Recovered) model can be described by the following system of ordinary differential equations,

\[
\begin{align*}
\frac{d N_S}{dt} &= -a \frac{N_S N_I}{N}, \\
\frac{d N_I}{dt} &= a \frac{N_S N_I}{N} - b N_I, \\
\frac{d N_R}{dt} &= b N_I,
\end{align*}
\]

in which \(a\) is the rate of infection and \(b\) the rate of recovery. It is the most basic of a series of compartmental epidemiological models. Various extensions exist to accommodate additional effects e.g., vital dynamics, immunity, incubation time (cf. e.g., Hethcote, 2000). Some of these extensions serve as a general model refinement, others are relevant to specific diseases.

Statistical properties, however, are not captured by the description through ODEs and call for a stochastic model. The Gillespie algorithm (Gillespie, 1976; Gillespie, 1977) enables discrete and stochastic simulations in which every trajectory is an exact sample of the solution of the ‘master equation’ that defines a probability distribution over solutions to a stochastic equation. In the SIR model, the rate constants are time-independent and thus, the underlying process is Markovian in which the event times are Poisson distributed. Here, an event denotes the transition of one individual from one compartment to another (e.g., \(N_I \rightarrow N_R\)).

\section{C.2 Experimental Setup}

For the AMS-BQ experiment, we assume that we know the recovery rate \(b\), but we are uncertain about the infection rate \(a\). Therefore, we rescale the ODEs and place a shifted gamma prior on \(a/b\) that starts at \(a/b = 1\) and has shape and scale parameters 5 and 4 respectively. With this prior we encode our belief that the infection rate is significantly larger than the recovery rate so an offset of the epidemic is very likely. Also, we set the population size to \(N = 100\) to be well below the thermodynamic limit and set one individual to be infected initially. We are interested in the expected maximum number \(E_a[\max_t N_I(t)]\) of simultaneously infected individuals and the time this maximum occurs \(E_a[\arg \max_t N_I(t)]\), 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 using the Gillespie algorithm (Gillespie, 1976; Gillespie, 1977); in addition to the base model (SIR), we include the state ‘exposed’, in which individuals are infected.
but not yet infectious. The modified system of ODEs that also account for assumed known incubation time $\gamma^{-1}$ are

\begin{align}
\frac{d N_S}{dt} &= -a N_S N_I / N, \\
\frac{d N_E}{dt} &= a N_S N_I / N - \gamma N_E, \\
\frac{d N_I}{dt} &= \gamma N_E - b N_I, \\
\frac{d N_R}{dt} &= b N_I,
\end{align}

(16)

where we set $\gamma = 10b$. We absorb the prior on $a/b$ in the black-box function for all methods.

Figure 11 shows the SIR and SEIR models (Eq. (15) and (16), respectively) with 10 stochastic trajectories (thin lines). The solid lines indicate the mean of 100 of these stochastic realizations, and the dashed lines show the solution of the ODEs, in both cases for the SIR model. We also use the SIR model for solving the ODEs even though the stochastic model simulates the SEIR model. The purpose of this is to mimic a case where secondary sources are simplified simulations in that minor components are deprecated. In the stochastic case, there is not always an outbreak of the disease, i.e., the initially infected individual recovers before infecting someone else. This causes the average $N_R$ to level off significantly below 1. For the integrals, only outbreaks are taken into account. The corresponding integrands for the quantities of interest are shown in Figure 12.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure11.png}
\caption{Demonstration of the SIR and SEIR models for $a/b = 10$. See text for details.}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{figure12.png}
\caption{Integrands used for the epidemiological model. Solid lines denote the primary source (i.e., stochastic simulations), dashed lines indicate the secondary source (solving the system of ODEs). It is apparent from the function that simply integrating the cheap source introduces a significant bias.}
\end{figure}

D BIVARIATE LINEAR COMBINATION OF GAUSSIANS

We construct an integrand (primary source) $f_1$ in the 2D-domain $[-3,3]^2$ as a linear combination of $K = 20$ normalized Gaussian basis functions

$$
\Phi_k(x) = (2\pi|A_k^0|)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-m_k^0)^T (A_k^0)^{-1}(x-m_k^0)},
$$

i.e., $f_1(x) = \sum_{k=1}^{K} z_k^1 \Phi_k(x)$. For this, we sample $K = 20$ means uniformly $m_k^1 \sim \text{Uniform}[-3,3]^2$ in the 2D domain. We then sample corresponding covariance matrices $A_k^1$ according to $u_k^1 \sim \mathcal{N}(0, I)$, $\kappa_k^1 \sim \text{Uniform}[0,1]^2$, and $A_k^1 := \text{diag}(\kappa_k^1) + u_k^1 (u_k^1)^T$. The scalar weights $z_k^1$ are sampled from a standard Gaussian $z_k^1 \sim \mathcal{N}(0, 1)$ and can be negative. Thus $f_1$ is not a probability density function but rather a linear combination of Gaussians with varying location, shape, and weight. We then construct secondary
sources $f_2$ and $f_3$ consecutively by adding uniform noise to the means, and additive uniform noise to the diagonal of the covariance matrices. Thus, with each additional source, each of the $K$ means get randomly but consecutively shifted up and right, and the basis functions $\Phi_i^k(x)$, $i = 2, 3$ randomly become wider and flatter. Additionally we consecutively add Gaussian random noise to the weights $z_k$ which ensures that the true integrals of the secondary sources differ from the integral of the primary source. All sources are depicted in Figure 13: the primary source $f_1$ on the left, and secondary sources $f_2$ and $f_3$ in the middle and right respectively. The cost for evaluating the primary source is 1 everywhere, the cost of evaluating $f_2$ and $f_3$ are 5% of the primary cost each.

The priors on the kernel lengthscale and coregionalization matrix $\mathbf{B}$ are set analogously to the other experiments already described in Section 4.1. AMS-BQ is initialized with one evaluation of the primary source and two evaluations each of the secondary sources which amounts to a total initial cost of 1.2 (initial evaluations shown as red dots in Figure 13). Vanilla-BQ is initialized with three evaluations which are needed to get an initial guess for its hyperparameters (initial cost=3). The result is shown in Figure 14 which plots relative error of the integral estimate versus the budget spent as well as two standard deviations of the relative error as returned by the model. It is apparent that AMS-BQ finds a good solution faster than vanilla-BQ.

Figure 15 illustrates the sequence of sources chosen by AMS-BQ. Secondary source $f_2$ is chosen more often than secondary source $f_3$ at equal evolution cost of 0.05. This is intuitive since $f_2$, by construction, provides more information about $f_1$ than $f_3$, but both secondary sources shrink the budget equally when queried. The percentage of number of evaluations for each source after spending a total budget of 50 is 15%, 57%, 28% for sources $f_1$, $f_2$, $f_3$ respectively.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Shape</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td></td>
<td>number of sources, indexed by $l$ where $l = 1$ is the primary source</td>
</tr>
<tr>
<td>$D$</td>
<td></td>
<td>dimension of the input space, indexed by $d$</td>
</tr>
<tr>
<td>$N$</td>
<td></td>
<td>number of source-input-evaluation triplets, indexed by $n$</td>
</tr>
<tr>
<td>$N_s$</td>
<td></td>
<td>number of potential new source-input-evaluation triplets</td>
</tr>
<tr>
<td>$x$</td>
<td>$D \times 1$</td>
<td>input location</td>
</tr>
<tr>
<td>$f(x)$</td>
<td>$L \times 1$</td>
<td>$[f_1(x) \ldots f_L(x)]^T$, where $f_l(x)$ is the $l$th source</td>
</tr>
<tr>
<td>$\langle f \rangle_l$</td>
<td>$L \times 1$</td>
<td>$\int_{\Omega} f_l(x) , d\pi(x)$ integral of the $l$th source</td>
</tr>
<tr>
<td>$\pi(x)$</td>
<td></td>
<td>integration measure on $\Omega$</td>
</tr>
<tr>
<td>$\Omega$</td>
<td></td>
<td>domain that is integrated over, $\Omega \subseteq \mathbb{R}^D$</td>
</tr>
<tr>
<td>$Z$</td>
<td></td>
<td>random variable for integral of interest $Z \sim \mathcal{N}(\mathbb{E}[Z</td>
</tr>
<tr>
<td>$(l_n, x_n)$</td>
<td>$(1, D \times 1)$</td>
<td>$n$th source-location pair where $f$ is evaluated</td>
</tr>
<tr>
<td>$(\ell, X)$</td>
<td>$(N \times 1, N \times D)$</td>
<td>$N$ source-location-pairs ($(l_1 \ldots l_N)^T, [x_1 \ldots x_N]^T$)</td>
</tr>
<tr>
<td>$f_\ell$</td>
<td>$N \times 1$</td>
<td>$[f_{l_1}(x_1) \ldots f_{l_N}(x_N)]^T$ noise-free function evaluations</td>
</tr>
<tr>
<td>$y_\ell$</td>
<td>$N \times 1$</td>
<td>$[f_{l_1}(x_1) + \epsilon_{l_1} \ldots f_{l_N}(x_N) + \epsilon_{l_N}]^T$ noisy function evaluations</td>
</tr>
<tr>
<td>$y$</td>
<td>$L \times 1$</td>
<td>$y = f(x) + \epsilon$ simultaneous evaluation of all sources</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>$L \times 1$</td>
<td>$\epsilon = [\epsilon_1 \ldots \epsilon_L]$ noise vector, $\epsilon_i \sim \mathcal{N}(0, \sigma^2_i)$</td>
</tr>
<tr>
<td>$\Sigma_\ell$</td>
<td>$N \times N$</td>
<td>diagonal noise matrix with noise per level $\sigma^2_i$, $(N\times N)$</td>
</tr>
<tr>
<td>$\mathcal{D}$</td>
<td></td>
<td>$N$ collected data triplets ${(l_n, x_n, f_{l_n}(x_n))}_{n=1}^N$</td>
</tr>
<tr>
<td>$\mathbf{K}$</td>
<td>$L \times L$</td>
<td>$\mathbf{K} = \text{cov}[f, f]$ matrix-valued covariance matrix</td>
</tr>
<tr>
<td>$k_{WW^\prime}(x, x^\prime)$</td>
<td></td>
<td>covariance function $\text{cov}[f_l(x), f_l^\prime(x^\prime)]$</td>
</tr>
<tr>
<td>$k_{\ell\ell}^\prime(x, X)$</td>
<td>$1 \times L$</td>
<td>vector-valued covariance $\text{cov}[f_\ell(x), f_\ell(x^\prime)]$</td>
</tr>
<tr>
<td>$\mathbf{G}_\ell(X)$</td>
<td>$N \times N$</td>
<td>$\text{Gram matrix } \mathbf{K}<em>{\ell\ell}(X, X) + \Sigma</em>\ell$</td>
</tr>
<tr>
<td>$m(x)$</td>
<td>$L \times 1$</td>
<td>GP prior mean for multi-output GP</td>
</tr>
<tr>
<td>$m_{l_{\mathcal{D}}}$</td>
<td>$N \times 1$</td>
<td>prior mean evaluated at source-location pairs $(\ell, X)$</td>
</tr>
<tr>
<td>$\text{posterior mean at source } l$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k_{WW^\prime \mid \mathcal{D}}$</td>
<td>$L \times L$</td>
<td>posterior covariance of sources $l, l^\prime$</td>
</tr>
<tr>
<td>$\langle k_m \rangle$</td>
<td>$1 \times L$</td>
<td>kernel mean of $l$th source at source-location pairs $(\ell, X)$</td>
</tr>
<tr>
<td>$\langle\langle k_m \rangle\rangle$</td>
<td>$L \times L$</td>
<td>$\int_{\Omega} k_{WW^\prime}(x, x^\prime) , d\pi(x) , d\pi(x^\prime) \text{ initial error}$</td>
</tr>
<tr>
<td>$\mathbf{\kappa}(x, x^\prime)$</td>
<td></td>
<td>coregionalization matrix for the kernel used in the ICM</td>
</tr>
<tr>
<td>$(\ell_s, X_s, y_{\ell_s})$</td>
<td>$(N_s \times 1, N_s \times D, N_s \times 1)$</td>
<td>potential new source-location-evaluation triplets</td>
</tr>
<tr>
<td>$c_{\ell_s}(x_s)$</td>
<td></td>
<td>cost of evaluating at $(\ell_s, x_s)$: $c_{\ell_s}(x_s) = \sum_{i=1}^{N_s} c_i(x_i)$</td>
</tr>
<tr>
<td>$\mathbf{V}_{\ell_s \mid \mathcal{D}}(X_s)$</td>
<td>$N_s \times N_s$</td>
<td>$\mathbf{K}<em>{\ell_s \mid \mathcal{D}}(X_s, X_s) + \Sigma</em>{\ell_s}$: in the myopic case denoted as $v_{l_s \mid \mathcal{D}}(x_s)$</td>
</tr>
<tr>
<td>$\rho^2_{\ell_s \mid \mathcal{D}}(X_s)$</td>
<td></td>
<td>scalar correlation function for $(\ell_s, X_s)$, defined in Eq. (6)</td>
</tr>
<tr>
<td>$\alpha_{\ell_s}(x_s)$</td>
<td></td>
<td>non-myopic acquisition function, $\alpha_{\ell_s}(x_s)$ in the myopic case</td>
</tr>
</tbody>
</table>

Table 1: Summary of the notation used. Generally, vector-valued quantities are denoted by lower case bold letters and matrices are upper case bold letters. Normal font denotes scalars.