SUPPLEMENTARY MATERIAL

Active Multi-Information Source Bayesian Quadrature

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 $\boldsymbol{y} = \boldsymbol{f}(\mathbf{x}) + \boldsymbol{\epsilon}$, where $\boldsymbol{y} \in \mathbb{R}^L$. In multi-source models, we wish to observe only elements of $\boldsymbol{f}$, i.e., $y_l = f_l(\mathbf{x}) + \epsilon_l$. These observations can be written as projections of the vector-valued observations,

$$y_l = \mathbf{h}_l^\intercal \boldsymbol{y} \tag{11}$$

where $\mathbf{h}_l$ denotes a vector with a 1 in the l^{th} coordinate and zero elsewhere. Let $\mathbf{Y} \in \mathbb{R}^{NL}$ denote the vector of N stacked vector-valued noisy observations $[\boldsymbol{y}_1, \dots, \boldsymbol{y}_N]$. Then the corresponding N observations of elements $\boldsymbol{\ell} = [l_1 \ \dots \ l_N]^\intercal$ is

$$\mathbf{y}_{\boldsymbol{\ell}} = \begin{bmatrix} \mathbf{h}_{l_1}^\intercal & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{h}_{l_N}^\intercal \end{bmatrix} \mathbf{Y} \ =: \ \mathbf{H}^\intercal \mathbf{Y}, \tag{12}$$

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

$$\operatorname{cov}[\mathbf{y}_{\boldsymbol{\ell}}, \mathbf{y}_{\boldsymbol{\ell}}] = \mathbf{H}^\intercal \left(\underbrace{\begin{bmatrix} \mathbf{K}(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \mathbf{K}(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \mathbf{K}(\mathbf{x}_N, \mathbf{x}_1) & \cdots & \mathbf{K}(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}}_{\mathbf{K}(\mathbf{X}, \mathbf{X})} + \boldsymbol{\Sigma} \otimes \mathbf{1}_{N \times N} \right) \mathbf{H} \tag{13}$$

where $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2, \dots, \sigma_L^2) \in \mathbb{R}^{L \times L}$ and $\mathbf{1}_{N \times N}$ is an $N \times N$ matrix with every element a 1. Also, $\mathbf{K}(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{NL \times NL}$. With the following mappings, we arrive at the multi-source notation introduced in Section 2;

$$\begin{aligned} \mathbf{H}^\intercal \mathbf{K}_{\mathbf{XX}} \mathbf{H} &= \mathbf{K}_{\boldsymbol{\ell}\boldsymbol{\ell}}(\mathbf{X}, \mathbf{X}); \\ \mathbf{H}^\intercal (\boldsymbol{\Sigma} \otimes \mathbf{1}_{N \times N}) \mathbf{H} &= \boldsymbol{\Sigma}_{\boldsymbol{\ell}}; \\ \mathbf{H}\mathbf{Y} &= \mathbf{y}_{\boldsymbol{\ell}}; \quad \text{etc.} \end{aligned} \tag{14}$$

Hence, the notational detour over vector-valued observations $\mathbf{Y}$ 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.

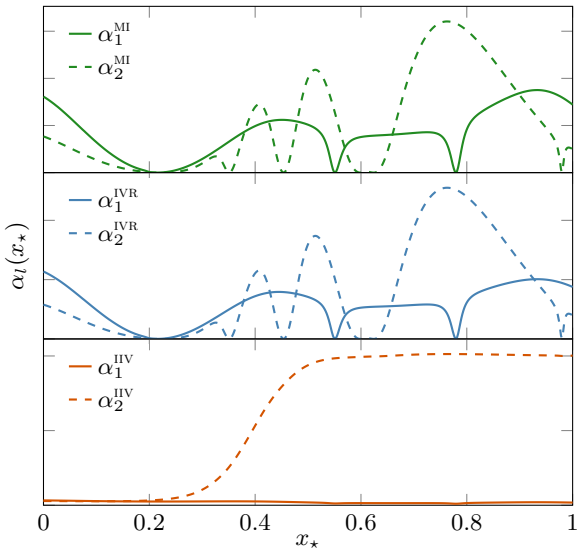

Figure 7: MI, IVR, and IP acquisitions for the top row of Figure 3. MI and IVR do not differ a lot, i.e., the correlation ρ is rarely large enough for MI to leave the linear regime. MI puts slightly more emphasis on the primary source where $x_\star$ is close to 1. This indicates that the correlation between Z and $y_\star$ 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.

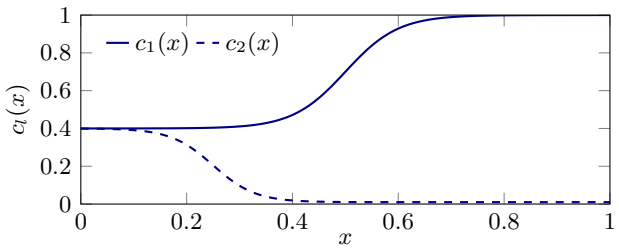

Figure 8: The cost used for the experiment in Figure 3.

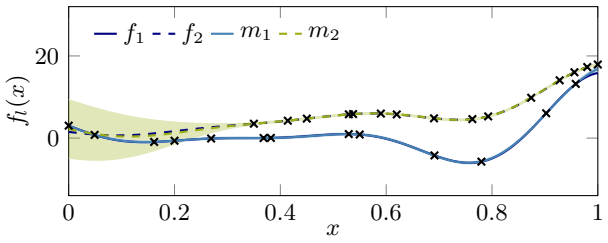

Figure 9: A later state for the experiment shown in Figure 3. Note the absence of f_2 evaluations for small x where c_1 and c_2 are similar.

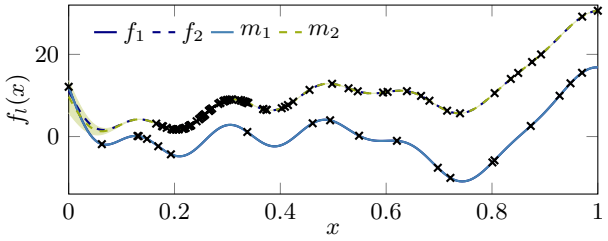

Figure 10: Final state of the GP for the second experiment explained in Section 4.1 and shown in Figure 4 ('wigglified Forrester'). Note the increasing density of evalutions of the secondary source where the cost is minimal, and the lack of f_2 queries where $c_1(x) \simeq c_2(x)$. The leftmost evaluation is at the primary source. See Figure 4 for the cost functions. In this experiment, the IP acquisition exclusively evaluates at the location of the minimum of the secondary cost function and is thus stuck.

C DETAILS FOR THE INFECTIONS MODEL

C.1 THE SIR MODEL AND EXTENSIONS

When the population size is large, the SIR (Susceptible, Infected, Recovered) model can be described by the following system of ordinary differential equations,

$$\begin{aligned}\frac{\mathrm{d}\,N_S}{\mathrm{d}\,t} &= -a\frac{N_S N_I}{N},\\ \frac{\mathrm{d}\,N_I}{\mathrm{d}\,t} &= a\frac{N_S N_I}{N} - bN_I,\\ \frac{\mathrm{d}\,N_R}{\mathrm{d}\,t} &= bN_I,\end{aligned} \tag{15}$$

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., $N_I \to N_R$).

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 $\mathbb{E}_a[\max_t N_I(t)]$ of simultaneously infected individuals 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 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

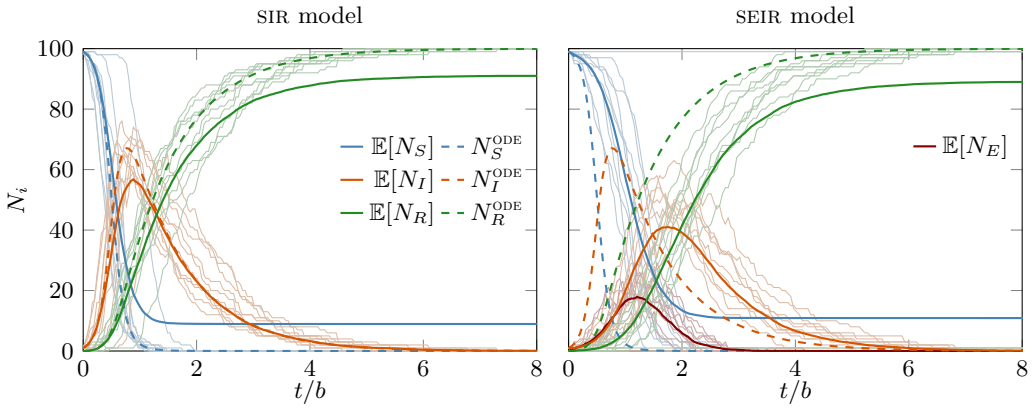

Figure 11: Demonstration of the SIR and SEIR models for $a/b = 10$. See text for details.

but not yet infectious. The modified system of ODEs that also account for assumed known incubation time γ^{-1} are

$$\begin{aligned}
\frac{\mathrm{d}\, N_S}{\mathrm{d}\, t} &= -a\frac{N_S N_I}{N}, \\
\frac{\mathrm{d}\, N_E}{\mathrm{d}\, t} &= a\frac{N_S N_I}{N} - \gamma N_E, \\
\frac{\mathrm{d}\, N_I}{\mathrm{d}\, t} &= \gamma N_E - b N_I, \\
\frac{\mathrm{d}\, N_R}{\mathrm{d}\, t} &= b N_I,
\end{aligned} \tag{16}$$

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 individuum 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.

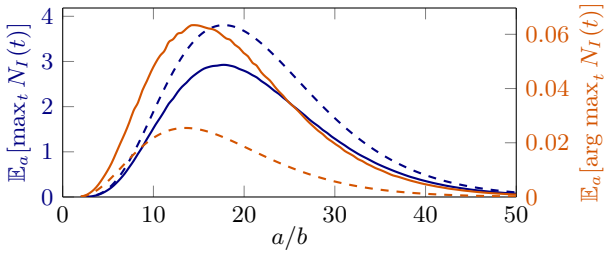

Figure 12: Integrands used for the epidemiological model. Solid lines denote the primary source (i.e., stochastic simulations), dashed lines indicate the secondary source (solving the system of ODEs). It is apparent from the function that simply integrating the cheap source introduces a significant bias.

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

$$\Phi_k^1(\mathbf{x}) = (2\pi|\mathbf{A}_k^1|)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k^1)^\intercal (\mathbf{A}_k^1)^{-1}(\mathbf{x}-\mathbf{m}_k^1)}, \tag{17}$$

i.e., $f_1(\mathbf{x}) = \sum_{k=1}^K z_k^1 \Phi_k^1(\mathbf{x})$. For this, we sample $K = 20$ means uniformly $\mathbf{m}_k^1 \sim \text{Uniform}[-3, 3]^2$ in the 2D domain. We then sample corresponding covariance matices $\mathbf{A}_k^1$ according to $\mathbf{u}_k^1 \sim \mathcal{N}(0, \boldsymbol{I})$, $\boldsymbol{\kappa}_k^1 \sim \text{Uniform}[0, 1]^2$, and $\mathbf{A}_k^1 := \text{diag}(\boldsymbol{\kappa}_k^1) + \mathbf{u}_k^1(\mathbf{u}_k^1)^\intercal$. 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

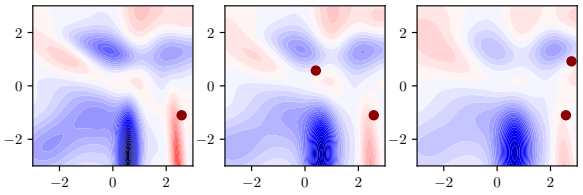

Figure 13: Integrands used for the bivariate linear combination of Gaussians. From left to right: primary source f_1 and secondary sources f_2 and f_3. Initial evaluations marked as red dots.

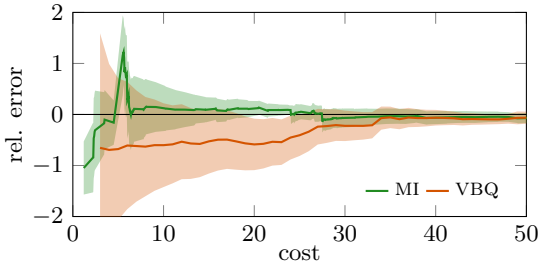

Figure 14: Relative error vs. budget spent for vanilla-BQ and AMS-BQ.

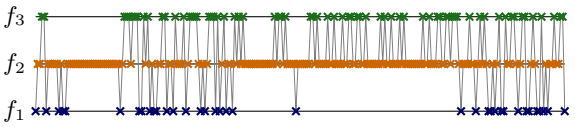

Figure 15: Evaluation sequence of primary and secondary sources in 2D experiment (250 evaluations shown).

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 $\Phi_k^i(\mathbf{x})$, $i = 2, 3$ 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 choses 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.

Variable	Shape	Description
L		number of sources, indexed by l where $l = 1$ is the primary source
D		dimension of the input space, indexed by d
N		number of source-input-evaluation triplets, indexed by n
$N_\star$		number of potential new source-input-evaluation triplets
$\mathbf{x}$	$D \times 1$	input location
$\boldsymbol{f}(\mathbf{x})$	$L \times 1$	$[f_1(\mathbf{x}) \ \dots \ f_L(\mathbf{x})]^\intercal$, where $f_l(\mathbf{x})$ is the l^{th} source
$\langle f_l \rangle$		$\int_\Omega f_l(\mathbf{x}) \ \mathrm{d}\pi(\mathbf{x})$ integral of the l^{th} source
$\pi(\mathbf{x})$		integration measure on Ω
Ω		domain that is integrated over, $\Omega \subseteq \mathbb{R}^D$
Z		random variable for integral of interest $Z \sim \mathcal{N}(\mathbb{E}[Z \mid \mathcal{D}], \mathbb{V}[Z \mid \mathcal{D}])$
$(l_n, \mathbf{x}_n)$	$(1, D \times 1)$	n^{th} source-location pair where $\boldsymbol{f}$ is evaluated
$(\boldsymbol{\ell}, \mathbf{X})$	$(N \times 1, N \times D)$	N source-location-pairs $([l_1 \ \dots \ l_N]^\intercal, [\mathbf{x}_1 \ \dots \ \mathbf{x}_N]^\intercal)$
$\mathbf{f}_{\boldsymbol{\ell}}$	$N \times 1$	$[f_{l_1}(\mathbf{x}_1) \ \dots \ f_{l_N}(\mathbf{x}_N)]^\intercal$ noise-free function evaluations
$\mathbf{y}_{\boldsymbol{\ell}}$	$N \times 1$	$[f_{l_1}(\mathbf{x}_1) + \epsilon_{l_1} \ \dots \ f_{l_N}(\mathbf{x}_N) + \epsilon_{l_N}]^\intercal$ noisy function evaluations
$\boldsymbol{y}$	$L \times 1$	$\boldsymbol{y} = \boldsymbol{f}(\mathbf{x}) + \boldsymbol{\epsilon}$ simultaneous evaluation of all sources
$\boldsymbol{\epsilon}$	$L \times 1$	$\boldsymbol{\epsilon} = [\epsilon_1 \ \dots \ \epsilon_L]$ noise vector, $\epsilon_l \sim \mathcal{N}(0, \sigma_l^2)$
$\boldsymbol{\Sigma}_{\boldsymbol{\ell}}$	$N \times N$	$= \mathrm{diag}(\sigma_{l_1}^2, \dots, \sigma_{l_N}^2)$ diagonal noise matrix with noise per level $\sigma_{l_n}^2$
$\mathcal{D}$		N collected data triplets $\{(l_n, \mathbf{x}_n, f_{l_n}(\mathbf{x}_n))\}_{n=1}^N$
$\mathbf{K}$	$L \times L$	$\mathbf{K} = \mathrm{cov}[\boldsymbol{f}, \boldsymbol{f}]$ matrix-valued covariance matrix
$k_{ll'}(\mathbf{x}, \mathbf{x}')$		covariance function $\mathrm{cov}[f_l(\mathbf{x}), f_{l'}(\mathbf{x}')]$
$\mathbf{k}_{l\boldsymbol{\ell}}(\mathbf{x}, \mathbf{X})$	$1 \times L$	vector-valued covariance $\mathrm{cov}[f_l(\mathbf{x}), \mathbf{f}_{\boldsymbol{\ell}}(\mathbf{X})]$
$\mathbf{K}_{\boldsymbol{\ell}\boldsymbol{\ell}}(\mathbf{X}, \mathbf{X})$	$N \times N$	$\mathrm{cov}[\mathbf{f}_{\boldsymbol{\ell}}(\mathbf{X}), \mathbf{f}_{\boldsymbol{\ell}}(\mathbf{X})]$
$\mathbf{G}_{\boldsymbol{\ell}}(\mathbf{X})$	$N \times N$	Gram matrix $\mathbf{K}_{\boldsymbol{\ell}\boldsymbol{\ell}}(\mathbf{X}, \mathbf{X}) + \boldsymbol{\Sigma}_{\boldsymbol{\ell}}$
$\mathbf{m}(\mathbf{x})$	$L \times 1$	GP prior mean for multi-output GP
$\mathbf{m}_{\boldsymbol{\ell}}(\mathbf{X})$	$N \times 1$	prior mean evaluated at source-location pairs $(\boldsymbol{\ell}, \mathbf{X})$
$m_{l \mid \mathcal{D}}$		posterior mean at source l
$k_{ll' \mid \mathcal{D}}$		posterior covariance of sources l, l'
$\langle m_l \rangle$		$\int_\Omega m_l(\mathbf{x}) \ \mathrm{d}\pi(\mathbf{x})$ integrated prior mean
$\langle \mathbf{k}_{\boldsymbol{\ell}l}(\mathbf{X}, \cdot) \rangle$	$1 \times L$	kernel mean of l^{th} source at source-location pairs $(\boldsymbol{\ell}, \mathbf{X})$
$\langle\langle k_{ll'} \rangle\rangle$		$\iint_\Omega k_{ll'}(\mathbf{x}, \mathbf{x}') \ \mathrm{d}\pi(\mathbf{x}) \, \mathrm{d}\pi(\mathbf{x}')$ initial error
$\mathbf{B}$	$L \times L$	coregionalization matrix for the kernel used in the ICM
$\kappa(\mathbf{x}, \mathbf{x}')$		kernel encoding purely spatial correlation in the ICM
$(\boldsymbol{\ell}_\star, \mathbf{X}_\star, \mathbf{y}_{\boldsymbol{\ell}_\star})$	$(N_\star \times 1, N_\star \times D, N_\star \times 1)$	potential new source-location-evaluation triplets
$c_{\boldsymbol{\ell}_\star}(\mathbf{X}_\star)$		cost of evaluating at $(\boldsymbol{\ell}_\star, \mathbf{X}_\star)$; $c_{\boldsymbol{\ell}_\star}(\mathbf{X}_\star) = \sum_{i=1}^{N_\star} c_{l_i}(\mathbf{x}_i)$
$\mathbf{V}_{\boldsymbol{\ell}_\star \mid \mathcal{D}}(\mathbf{X}_\star)$	$N_\star \times N_\star$	$= \mathbf{K}_{\boldsymbol{\ell}_\star\boldsymbol{\ell}_\star \mid \mathcal{D}}(\mathbf{X}_\star, \mathbf{X}_\star) + \boldsymbol{\Sigma}_{\boldsymbol{\ell}_\star}$; in the myopic case denoted as $v_{l_\star \mid \mathcal{D}}(\mathbf{x}_\star)$
$\rho^2_{1\boldsymbol{\ell}_\star \mid \mathcal{D}}(\mathbf{X}_\star)$		scalar correlation function for $(\boldsymbol{\ell}_\star, \mathbf{X}_\star)$, defined in Eq. (6)
$\alpha_{\boldsymbol{\ell}_\star}(\mathbf{X}_\star)$		non-myopic acquisition function, $\alpha_{l_\star}(\mathbf{x}_\star)$ in the myopic case

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.