Nonmyopic Gaussian Process Optimization with Macro-Actions

Dmitrii Kharkovskii
National University of Singapore

Chun Kai Ling
Carnegie Mellon University

Bryan Kian Hsiang Low
National University of Singapore

Abstract

This paper presents a multi-staged approach to nonmyopic adaptive *Gaussian process optimization* (GPO) for *Bayesian optimization* (BO) of unknown, highly complex objective functions that, in contrast to existing nonmyopic adaptive BO algorithms, exploits the notion of macro-actions for scaling up to a further lookahead to match up to a larger available budget. To achieve this, we generalize GP upper confidence bound to a new acquisition function defined w.r.t. a nonmyopic adaptive macro-action policy, which is intractable to be optimized exactly due to an uncountable set of candidate outputs. The contribution of our work here is thus to derive a nonmyopic adaptive ϵ-*Bayes-optimal macro-action GPO* (ϵ-Macro-GPO) policy. To perform nonmyopic adaptive BO in real time, we then propose an asymptotically optimal anytime variant of our ϵ-Macro-GPO policy with a performance guarantee. We empirically evaluate the performance of our ϵ-Macro-GPO policy and its anytime variant in BO with synthetic and real-world datasets.

1 Introduction

Recent advances in *Bayesian optimization* (BO) have delivered a promising suite of tools for optimizing an unknown (possibly noisy, non-convex, with no closed-form expression/derivative) objective function with a finite budget of function evaluations, as demonstrated in a wide range of applications like automated machine learning, robotics, sensor networks, environmental monitoring, among others (Shahriari *et al.*, 2016). Conventionally, a BO algorithm relies on some choice of acquisition function (e.g., probability of improvement or *expected improvement* (EI) over currently found maximum, information-based (Hennig and Schuler, 2012; Hernández-Lobato *et al.*, 2014), or *upper confidence bound* (UCB) (Srinivas *et al.*, 2010)) as a heuristic to guide its search for the global maximum. To do this, the BO algorithm exploits the chosen acquisition function to repeatedly select an input for evaluating the unknown objective function that trades off between observing a likely maximum based on a GP belief of the unknown objective function (exploitation) vs. improving the GP belief (exploration) until the budget is expended.

Unfortunately, such a conventional BO algorithm is greedy/myopic and hence performs suboptimally with respect to the given finite budget[1]. To be nonmyopic, its policy to select the next input has to additionally account for its subsequent selections of inputs for evaluating the unknown objective function. Perhaps surprisingly, this can be partially achieved by batch BO algorithms capable of *jointly*[2] optimizing a batch of inputs (Chevalier and Ginsbourger, 2013; Daxberger and Low, 2017; Shah and Ghahramani, 2015; Wu and Frazier, 2016) because their selection of each input has to account for that of all other inputs of the batch[3]. However, since the batch size is typically set to be much smaller than the given budget, they have to repeatedly select the next batch greedily. Unlike the conventional BO algorithm described above, their selection of each input is independent of the outputs observed from evaluating the objective function at the other selected inputs of the batch, thus sacrificing some degree of adaptivity. Hence, they also perform suboptimally with respect to the given budget.

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

[1]Acquisition functions like EI (Bull, 2011; Vazquez and Bect, 2010) and UCB (Srinivas *et al.*, 2010) offer theoretical guarantees for the convergence rate of their BO algorithms (i.e., in the limit) via regret bounds. In practice, since the budget is limited, such bounds are suboptimal as they cannot be specified to be arbitrarily small.

[2]In contrast, a *greedy* batch BO algorithm (Contal *et al.*, 2013; Desautels *et al.*, 2014; González *et al.*, 2016a) selects the inputs of a batch one at a time myopically.

[3]Batch BO is traditionally considered when resources are available to evaluate the objective function in parallel. We suggest a further possibility of using batch BO for a non-myopic selection of inputs of the batch here.

Some nonmyopic adaptive BO algorithms (Lam and Willcox, 2017; Lam *et al.*, 2016; Ling *et al.*, 2016; Marchant *et al.*, 2014; Osborne *et al.*, 2009) have been developed to combine the best of both worlds. But, they have been empirically demonstrated to be effective and tractable for at most a lookahead of 5 observations which is usually much less than the size of the available budget in practice, thus causing them to behave myopically in this case. To increase the lookahead, the work of (González *et al.*, 2016b) has proposed a two-staged approach that utilizes a *greedy* batch BO algorithm[2] in its second stage to efficiently but myopically optimize all but the first input afforded by the budget. Note that the above works on nonmyopic adaptive BO do not provide theoretical performance guarantees except for that of (Ling *et al.*, 2016). The challenge therefore remains in devising a multi-staged approach to nonmyopic adaptive BO that can empirically scale well to a further lookahead (and hence match up to a larger budget) and still be amenable to a theoretical analysis of its performance, which is the focus of our work here.

To address this challenge, we exploit the notion of macro-actions (i.e., each denoting a sequence of primitive actions executed in full without considering any observation taken after performing each primitive action in the sequence) inherent to the structure of several real-world task environments/applications such as environmental sensing and monitoring, mobile sensor networks, and robotics. Some examples are given below and described in detail in Section 4 (see Fig. 1 for more examples): (a) In monitoring of algal bloom in the coastal ocean, an *autonomous underwater vehicle* (AUV) is deployed on board a research vessel in search for a hotspot of peak phytoplankton abundance and tasked to take dives from the vessel to gather "Gulper" water samples for on-deck testing that can be cast as macro-actions (Pennington *et al.*, 2016), and (b) in servicing the mobility demands within an urban city, an autonomous robotic vehicle in a mobility-on-demand system cruises along different road trajectories abstracted as macro-actions to find a hotspot of highest mobility demand to pick up a user (Chen *et al.*, 2015b). Macro-actions have in fact been well-studied and used by the planning community[4] to scale up algorithms for planning under uncertainty to a further lookahead (He *et al.*, 2010, 2011; Lim *et al.*, 2011), which is realized from a much reduced space of possible sequences of primitive actions (i.e., macro-actions) induced by the structure of the task environment/application.

The use of macro-actions in the context of nonmyopic adaptive BO poses an interesting research question: *How can an acquisition function be defined with respect to a nonmyopic adaptive macro-action*[5] *policy and optimized tractably to yield such a policy with a provable performance guarantee for a given finite budget?* The main technical difficulty in answering this question stems from the need to account for the correlation of outputs to be observed from evaluating the unknown objective function at inputs found within a macro-action and between different macro-actions (Section 3). Such a correlation structure is the chief ingredient to be exploited for selecting informative observations to find the global maximum.

This paper presents a principled multi-staged Bayesian sequential decision problem framework for nonmyopic adaptive *GP optimization* (GPO) (Section 3) that, in particular, exploits macro-actions inherent to the structure of several real-world task environments/applications for scaling up to a further lookahead (as compared to the existing nonmyopic adaptive BO algorithms discussed above) to match up to a larger available budget. To achieve this, we first generalize GP-UCB to a new acquisition function defined with respect to a nonmyopic adaptive macro-action policy, which, unfortunately, is intractable to be optimized exactly due to an uncountable set of candidate outputs. The key novel contribution of our work here is to show that it is in fact possible to solve for a nonmyopic adaptive ϵ-*Bayes-optimal macro-action GPO* (ϵ-Macro-GPO) policy given an arbitrarily user-specified loss bound ϵ via stochastic sampling in each planning stage which requires only a polynomial number of samples in the length of macro-actions[6]. To perform nonmyopic adaptive BO in real time, we then propose an asymptotically optimal anytime variant of our ϵ-Macro-GPO policy with a performance guarantee. We empirically evaluate the performance of our ϵ-Macro-GPO policy and its anytime variant in BO with synthetic and real-world datasets (Section 4).

2 Modeling Spatially Varying Phenomena with Gaussian Processes

To simplify exposition of our work here, we will assume the task environment to be a spatially varying phenomenon (e.g., indoor environmental quality of an office environment, plankton bloom in the ocean, mobility demand within an urban city, as described in Section 1). A mobile sensing agent utilizes our pro-

[4]Macro-actions are also studied in reinforcement learning community but named as options instead (Konidaris and Barto, 2007; Stolle and Precup, 2002).

[5]In BO, each macro-action denotes a sequence of inputs for evaluating the unknown objective function.

[6]In contrast, though the nonmyopic adaptive BO algorithm of (Ling *et al.*, 2016) based on deterministic sampling can be naively generalized to exploit macro-actions, it requires an exponential number of samples per planning stage, as detailed in (Kharkovskii *et al.*, 2020).

posed nonmyopic adaptive ϵ-Macro-GPO policy or its anytime variant to select and gather observations from the task environment for finding the global maximum.

Notations and Preliminaries. Let $\mathcal{S}$ be the domain of a spatially varying phenomenon corresponding to a set of input locations. In every stage $t > 0$, the agent executes one of the available macro-actions of length κ at its current input location by deterministically moving through a sequence of κ input locations, denoted by a vector $s_t \in \mathcal{A}(s_{t-1})$, and observes the corresponding output measurements $z_t \in \mathbb{R}^\kappa$, where $\mathcal{A}(s_{t-1}) \subseteq \mathcal{S}^\kappa$ denotes a finite set of available macro-actions at the agent's current input location[7] (see visual illustration in Fig. 1 and its caption b). The state of the agent at its initial starting input location is represented by prior observations/data $d_0 \triangleq \langle s_0, z_0\rangle$ available before planning where s_0 and z_0 denote, respectively, vectors comprising input locations visited and corresponding output measurements observed by the agent prior to planning. The agent's initial starting input location is the last component of s_0. In stage $t > 0$, the state of the agent is represented by observations/data $d_t \triangleq \langle \mathbf{s}_t, \mathbf{z}_t\rangle$ where $\mathbf{s}_t \triangleq s_0 \oplus \ldots \oplus s_t$ and $\mathbf{z}_t \triangleq z_0 \oplus \ldots \oplus z_t$ denote, respectively, vectors comprising input locations visited and corresponding output measurements observed by the agent up till stage t and '$\oplus$' denotes vector concatenation.

Gaussian Process (GP). The spatially varying phenomenon is modeled as a realization of a GP: Each input location $s \in \mathcal{S}$ is associated with an output measurement y_s. Let $y_\mathcal{S} \triangleq \{y_s\}_{s\in\mathcal{S}}$ denote a GP, that is, every finite subset of $y_\mathcal{S}$ has a multivariate Gaussian distribution. Then, the GP is fully specified by its *prior* mean $\mu_s \triangleq \mathbb{E}[y_s]$ (we assume w.l.o.g. that $\mu_s = 0$ for all $s \in \mathcal{S}$) and covariance $\sigma_{ss'} \triangleq \mathrm{cov}[y_s, y_{s'}]$ for all $s, s' \in \mathcal{S}$, the latter of which characterizes the spatial correlation structure of the phenomenon. For example, $\sigma_{ss'}$ can be defined by the commonly-used squared exponential covariance function $\sigma_{ss'} \triangleq \sigma_y^2 \exp\{-0.5(s-s')^\top \Gamma^{-2}(s-s')\}$ where σ_y^2 is the signal variance controlling the intensity of output measurements and Γ is a diagonal matrix with length-scale components ℓ_1 and ℓ_2 controlling the spatial correlation or "similarity" between output measurements in the respective east-west and north-south directions of the 2D phenomenon.

All output measurements observed by the agent are corrupted by an additive noise ε, i.e., $z_{i,j} \triangleq y_{s_{i,j}} + \varepsilon$ for stage $i = 0, \ldots, t$ and $j = 1, \ldots \kappa$ where $s_{i,j}$ is the j-th input location of macro-action s_i at stage i, $z_{i,j}$ is the corresponding output measurement and $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$ with the noise variance σ_n^2. Supposing the agent has gathered observations $d_t = \langle \mathbf{s}_t, \mathbf{z}_t\rangle$ from stages 0 to t the GP model can exploit these observations d_t to perform probabilistic regression by providing a Gaussian posterior belief $p(z_{t+1}|s_{t+1}, d_t) = \mathcal{N}(\mu_{s_{t+1}|d_t}, \Sigma_{s_{t+1}|\mathbf{s}_t})$ of noisy output measurements for any κ input locations $s_{t+1} \subset \mathcal{S}$ with the following *posterior* mean vector and covariance matrix, respectively:

$$\begin{aligned} \mu_{s_{t+1}|d_t} &\triangleq \Sigma_{s_{t+1}\mathbf{s}_t}\Sigma_{\mathbf{s}_t\mathbf{s}_t}^{-1}\mathbf{z}_t^\top \ , \\ \Sigma_{s_{t+1}|\mathbf{s}_t} &\triangleq \Sigma_{s_{t+1}s_{t+1}} - \Sigma_{s_{t+1}\mathbf{s}_t}\Sigma_{\mathbf{s}_t\mathbf{s}_t}^{-1}\Sigma_{\mathbf{s}_t s_{t+1}} \end{aligned} \tag{1}$$

where $\Sigma_{s_{t+1}\mathbf{s}_t}$ is a matrix with covariance components $\sigma_{ss'}$ for every input location s of s_{t+1} and s' of $\mathbf{s}_t$, $\Sigma_{\mathbf{s}_t s_{t+1}}$ is the transpose of $\Sigma_{s_{t+1}\mathbf{s}_t}$, and $\Sigma_{\mathbf{s}_t\mathbf{s}_t}$ ($\Sigma_{s_{t+1}s_{t+1}}$) is a matrix with covariance components $\sigma_{ss'} + \sigma_n^2\delta_{ss'}$ for every pair of input locations s, s' of $\mathbf{s}_t$ (s_{t+1}) and $\delta_{ss'}$ is a Kronecker delta of value 1 if $s = s'$, and 0 otherwise. A key property of the GP model is that, different from $\mu_{s_{t+1}|d_t}$, $\Sigma_{s_{t+1}|\mathbf{s}_t}$ is independent of the output measurements $\mathbf{z}_t$.

3 ϵ-Bayes-Optimal Macro-GPO

Problem Formulation. To cast nonmyopic adaptive *macro-action GP optimization* (Macro-GPO) as a Bayesian sequential decision problem, we define a nonmyopic adaptive macro-action policy π to sequentially decide in each stage t the next macro-action $\pi(d_t) \in \mathcal{A}(s_t)$ to be executed for gathering κ new observations based on the current observations d_t over a finite planning horizon of H stages (i.e., a lookahead of κH observations). The goal of the agent is to plan/decide its macro-actions to visit input locations $\mathbf{s}_H \triangleq s_1 \oplus \ldots \oplus s_H$ with the maximum total corresponding output measurements $\mathbf{1}^\top\mathbf{z}_H = \sum_{t=1}^H \mathbf{1}^\top z_t = \sum_{t=1}^H\sum_{i=1}^\kappa z_{t,i}$ or, equivalently, minimum cumulative regret where $\mathbf{z}_H \triangleq z_1 \oplus \ldots \oplus z_H$ and $z_t \triangleq (z_{t,1}, \ldots, z_{t,\kappa})$. However, since only the prior observations/data d_0 are known, the Macro-GPO problem involves finding a macro-action policy π to select input locations $\mathbf{s}_H$ to be visited by the agent with the maximum *expected* total corresponding output measurements $\mathbb{E}_{\mathbf{z}_H|d_0,\pi}[\mathbf{1}^\top\mathbf{z}_H]$ instead.

Supposing the size of the available budget in a real-world task environment exceeds the lookahead of κH observations, it can afford a *stronger exploration behavior* by including an additional weighted exploration term $\beta\, \mathbb{I}[y_\mathcal{S}; \mathbf{z}_H|d_0, \pi]$; its effect on BO performance is empirically investigated in Section 4 (see Fig. 4c). The conditional mutual information $\mathbb{I}[y_\mathcal{S}; \mathbf{z}_H|d_0, \pi]$ here can be interpreted as the information gain on the phenomenon over the entire domain $\mathcal{S}$ (i.e., equivalent to $y_\mathcal{S}$) from gathering observations $\langle \mathbf{s}_H, \mathbf{z}_H\rangle$ selected according to the macro-action policy π given the prior data d_0. Then, the acquisition function w.r.t. a non-

[7]Note that $\mathcal{A}(s_{t-1})$ depends on the agent's current input location which corresponds to the last component of macro-action s_{t-1} executed in the previous stage $t-1$.

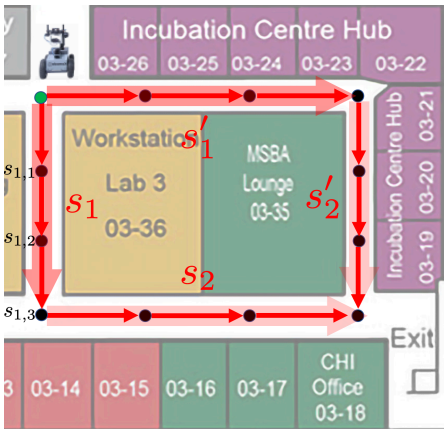


Figure 1: Example of monitoring indoor environmental quality of an office environment (Choi *et al.*, 2012): (a) A mobile robot mounted with a weather board is tasked to find a hotspot of peak temperature by exploring different stretches of corridors that can be naturally abstracted into macro-actions. (b) In stage $t = 1$, the robot is at its initial starting input location (green dot). It can decide to execute macro-action s_1 (translucent red arrow), which is a sequence of $\kappa = 3$ primitive actions (opaque red arrows) moving it through a sequence of $\kappa = 3$ input locations (black dots) to arrive at input location $s_{1,3}$. So, $s_1 \triangleq (s_{1,1}, s_{1,2}, s_{1,3})$. (c) To derive a myopic Macro-GPO or ϵ-Macro-GPO policy with $H = 1$, the last stages of Bellman equations in (5)-(8) require macro-actions s_1 and s'_1 as inputs. To derive a nonmyopic one with $H = 2$, they require macro-action sequences $s_1 \oplus s_2$ and $s'_1 \oplus s'_2$ as inputs instead.

myopic adaptive macro-action policy π when starting in d_0 and following π thereafter can be defined as

$$V_0^\pi(d_0) \triangleq \mathbb{E}_{\mathbf{z}_H|d_0,\pi}[\mathbf{1}^\top \mathbf{z}_H] + \beta\, \mathbb{I}[y_\mathcal{S}; \mathbf{z}_H|d_0, \pi] \ . \quad (2)$$

Applying the chain rule for mutual information and a few other information-theoretic results to (2) yields the following H-stage Bellman equations (see (Kharkovskii *et al.*, 2020) for the proof):

$$\begin{aligned} V_t^\pi(d_t) &\triangleq Q_t^\pi(\pi(d_t), d_t) \ , \\ Q_t^\pi(s_{t+1}, d_t) &\triangleq R(s_{t+1}, d_t) + \\ &\quad \mathbb{E}_{z_{t+1}|s_{t+1},d_t}[V_{t+1}^\pi(\langle \mathbf{s}_{t+1}, \mathbf{z}_t \oplus z_{t+1}\rangle)] \end{aligned} \quad (3)$$

for stages $t = 0, \ldots, H-1$ where $V_H^\pi(d_H) \triangleq 0$ and

$$R(s_{t+1}, d_t) \triangleq \mathbf{1}^\top \mu_{s_{t+1}|d_t} + 0.5\beta \log |I + \sigma_n^{-2} \Sigma_{s_{t+1}|\mathbf{s}_t}| \ . \quad (4)$$

To solve the Macro-GPO problem, Bayes-optimality is exploited to select input locations to be visited by the agent that maximize the expected total corresponding output measurements (and, if the budget can afford, the additional weighted exploration term representing the information gain on the phenomenon) with respect to all possible induced sequences of future GP

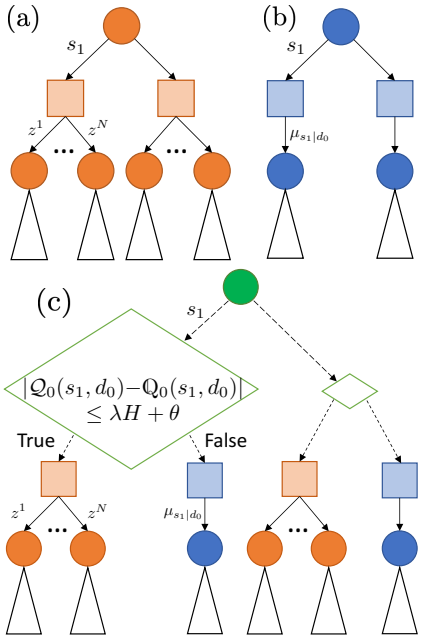


Figure 2: Visual illustrations of policies induced by (a) stochastic sampling (6), (b) most likely observations (7), and (c) our ϵ-Macro-GPO policy π^ϵ (8). Circles denote nodes d_t. Squares denote nodes $\langle \mathbf{s}_{t+1}, \mathbf{z}_t \rangle$.

posterior beliefs $p(z_{t+1}|s_{t+1}, d_t)$ for $t = 0, \ldots, H-1$. Formally, this involves choosing a nonmyopic adaptive macro-action policy π to maximize $V_0^\pi(d_0)$, which we call the Bayes-optimal Macro-GPO policy π^*. That is, $V_0^*(d_0) \triangleq V_0^{\pi^*}(d_0) = \max_\pi V_0^\pi(d_0)$. Plugging π^* into $V_t^\pi(d_t)$ and $Q_t^\pi(s_{t+1}, d_t)$ (3) gives

$$\begin{aligned} V_t^*(d_t) &\triangleq \max_{s_{t+1} \in \mathcal{A}(s_t)} Q_t^*(s_{t+1}, d_t) \ , \\ Q_t^*(s_{t+1}, d_t) &\triangleq R(s_{t+1}, d_t) + \\ &\quad \mathbb{E}_{z_{t+1}|s_{t+1},d_t}[V_{t+1}^*(\langle \mathbf{s}_{t+1}, \mathbf{z}_t \oplus z_{t+1}\rangle)] \end{aligned} \quad (5)$$

for stages $t = 0, \ldots, H-1$ where $V_H^*(d_H) \triangleq 0$.[8] When the lookahead of κH observations matches up to the available budget, the Bayes-optimal Macro-GPO policy π^* can naturally trade off between exploration vs. exploitation without needing the additional weighted exploration term in (2) or (4) (i.e., $\beta = 0$): Its selected macro-action $\pi^*(d_t) = \text{argmax}_{s_{t+1} \in \mathcal{A}(s_t)} Q_t^*(s_{t+1}, d_t)$ in each stage t has to trade off between exploiting the current GP posterior belief $p(z_{t+1}|\pi^*(d_t), d_t)$ to maximize the expected total corresponding output measurements $R(\pi^*(d_t), d_t) = \mathbf{1}^\top \mu_{\pi^*(d_t)|d_t}$ vs. improving the GP posterior belief of the phenomenon (i.e., exploration) so as to maximize the expected total output measurements $\mathbb{E}_{z_{t+1}|\pi^*(d_t),d_t}[V_{t+1}^*(\langle \mathbf{s}_t \oplus \pi^*(d_t), \mathbf{z}_t \oplus z_{t+1}\rangle)]$ in the later stages.

When the available budget is larger than the looka-

[8] To understand the effect of H on how much macro-action sequence information are required as inputs to the Bellman equations in (5)-(8), refer to Fig. 1 and its caption c for a visual illustration.

head of κH observations, it can afford a *stronger exploration behavior* by setting a positive weight $\beta > 0$ on the exploration term $0.5\log|I+\sigma_n^{-2}\Sigma_{\pi^*(d_t)|\mathbf{s}_t}|$ in (4); its effect on BO performance is empirically investigated in Section 4 (see Fig. 4c). This exploration term can be interpreted as the information gain $\mathbb{I}[y_{\mathcal{S}};z_{t+1}|d_t,\pi^*(d_t)]$ on the phenomenon (see (Kharkovskii *et al.*, 2020)) from executing the macro-action $\pi^*(d_t)$ to gather κ new observations. As such, $\pi^*(d_t)$ can gain more information on the phenomenon (larger exploration term) by gathering observations with higher uncertainty (larger individual posterior variance) but lower correlation (smaller magnitude of posterior covariance) between them.

ϵ-Bayes-Optimal Macro-GPO (ϵ-Macro-GPO). In general, the Macro-GPO policy π^* cannot be derived exactly because the expectation term in (5) (and hence Q_t^* and V_t^*) often cannot be evaluated in closed form due to an uncountable set of candidate output measurements. To overcome this difficulty, we will derive a nonmyopic adaptive ϵ-Macro-GPO policy π^ϵ whose expected performance loss is theoretically guaranteed to be within an arbitrarily user-specified loss bound ϵ. Preliminary to its design is the approximation of the expectation term in (5) for each candidate macro-action s_{t+1} in every stage using *stochastic* sampling of N i.i.d. multivariate Gaussian vectors $z^1,\ldots,z^N$ from the GP posterior belief $p(z_{t+1}|s_{t+1},d_t)$ (1), as illustrated in Fig. 2a:

$$\begin{aligned}\mathcal{V}_t(d_t) &\triangleq \max_{s_{t+1}\in\mathcal{A}(s_t)}\mathcal{Q}_t(s_{t+1},d_t)\ ,\\ \mathcal{Q}_t(s_{t+1},d_t) &\triangleq R(s_{t+1},d_t)+\frac{1}{N}\sum_{\ell=1}^{N}\mathcal{V}_{t+1}(\langle\mathbf{s}_{t+1},\mathbf{z}_t\oplus z^\ell\rangle)\end{aligned} \tag{6}$$

for stages $t=0,\ldots,H-1$ where $\mathcal{V}_H(d_H)\triangleq 0$.[8] We prove in (Kharkovskii *et al.*, 2020) that $\mathcal{Q}_t(s_{t+1},d_t)$ (6) can approximate $Q_t^*(s_{t+1},d_t)$ (5) arbitrarily closely (i.e., $|\mathcal{Q}_t(s_{t+1},d_t)-Q_t^*(s_{t+1},d_t)|\leq\lambda H$ given a user-specified $\lambda>0$) for all s_{t+1} with a high probability of at least $1-\delta$ requiring only a polynomial number N of samples in the macro-action length κ (9) per planning stage. Such a result, however, only entails probabilistic bounds on how far $\mathcal{V}_t(d_t)$ (6) is from $V_t^*(d_t)$ (5) (see (Kharkovskii *et al.*, 2020)) and on the resulting policy loss. We will prove a stronger non-trivial result: In the unlikely event (with an arbitrarily small probability of at most δ) that $\mathcal{Q}_t(s_{t+1},d_t)$ (6) is unboundedly far from $Q_t^*(s_{t+1},d_t)$ (5) for some s_{t+1}, we instead rely on the κ most likely observations

$$\begin{aligned}\mathbb{V}_t(d_t) &\triangleq \max_{s_{t+1}\in\mathcal{A}(s_t)}\mathbb{Q}_t(s_{t+1},d_t)\ ,\\ \mathbb{Q}_t(s_{t+1},d_t) &\triangleq R(s_{t+1},d_t)+\mathbb{V}_{t+1}(\langle\mathbf{s}_{t+1},\mathbf{z}_t\oplus\mu_{s_{t+1}|d_t}\rangle)\end{aligned} \tag{7}$$

for stages $t=0,\ldots,H-1$ where $\mathbb{V}_H(d_H)\triangleq 0$.[8] Unlike $\mathcal{Q}_t(s_{t+1},d_t)$ (6), the approximation quality of $\mathbb{Q}_t(s_{t+1},d_t)$ (7) can be *deterministically* bounded but cannot be user-specified to be arbitrarily good (see (Kharkovskii *et al.*, 2020)): $|\mathbb{Q}_t(s_{t+1},d_t)-Q_t^*(s_{t+1},d_t)|\leq\theta$ for all s_{t+1} where $\theta\triangleq\mathcal{O}(\kappa^{H+1/2})$. To ease understanding, we visually illustrate in Fig. 2 how the policies induced by stochastic sampling (6) vs. most likely observations (7) differ and are used to design our ϵ-Macro-GPO policy π^ϵ (8). Essentially, we design π^ϵ to strictly follow the policy induced by stochastic sampling (6) only if $\mathcal{Q}_t(s_{t+1},d_t)$ (6) is boundedly close to $\mathbb{Q}_t(s_{t+1},d_t)$ (7) for all s_{t+1}:

$$\begin{aligned}\pi^\epsilon(d_t) &\triangleq \operatorname{argmax}_{s_{t+1}\in\mathcal{A}(s_t)}Q_t^\epsilon(s_{t+1},d_t)\ ,\\ Q_t^\epsilon(s_{t+1},d_t) &\triangleq \begin{cases}\mathcal{Q}_t(s_{t+1},d_t) & \text{if } |\mathcal{Q}_t(s_{t+1},d_t)-\\ & \quad \mathbb{Q}_t(s_{t+1},d_t)|\leq\lambda H+\theta,\\ \mathbb{Q}_t(s_{t+1},d_t) & \text{otherwise;}\end{cases}\end{aligned} \tag{8}$$

for stages $t=0,\ldots,H-1$.[8] Like the Macro-GPO policy π^*, π^ϵ can also naturally trade off between exploration vs. exploitation, by the same reasoning as earlier. Unlike the deterministic policy π^*, π^ϵ is stochastic due to its use of stochastic sampling in $\mathcal{Q}_t$ (6).

To understand the rationale/implications of our choice of if condition in (8), refer to Fig. 3. These implications are central to establishing our main result deterministically bounding the *expected* performance loss of π^ϵ relative to that of π^*, i.e., π^ϵ is ϵ-Bayes-optimal (see proof in (Kharkovskii *et al.*, 2020)):

Theorem 1. *Suppose that the observations d_0, $H\in\mathbb{Z}^+$, a budget of κH input locations, and a user-specified loss bound $\epsilon>0$ are given. Then, $V_0^*(d_0)-\mathbb{E}_{\pi^\epsilon}[V_0^{\pi^\epsilon}(d_0)]\leq\epsilon$ by setting $\theta\triangleq\mathcal{O}(\kappa^{H+1/2})$ (see (Kharkovskii* et al.*, 2020)), $\delta=\epsilon/(8\theta H)$, and $\lambda=\epsilon/(4H^2)$ to yield*

$$N=\mathcal{O}((\kappa^{2H}/\epsilon^2)\log(\kappa A/\epsilon)) \tag{9}$$

where A denotes the largest number of candidate macro-actions available at any input location in $\mathcal{S}$.

Remark 1. It can be observed from Theorem 1 that the number N of stochastic samples increases[9] with (a) a tighter user-specified loss bound ϵ, (b) a larger number A of candidate macro-actions at any input location in $\mathcal{S}$, and (c) a greater macro-action length κ.

Anytime ϵ-Macro-GPO. Unlike the Bayes-optimal policy π^*, our policy π^ϵ can be derived exactly since its incurred time does not depend on the size of the uncountable set of candidate output measurements. But, deriving π^ϵ (8) requires expanding an entire search tree of $\mathcal{O}(N^H)$ nodes to solve the H-stage Bellman equations of $\mathcal{V}_t$ (6), which is not always needed to achieve

[9] In fact, N also increases when a larger H is available and the spatial phenomenon varies with more intensity and less noise, i.e., larger σ_y^2/σ_n^2 (see (Kharkovskii *et al.*, 2020)). These constants are omitted from (9) to ease clutter.

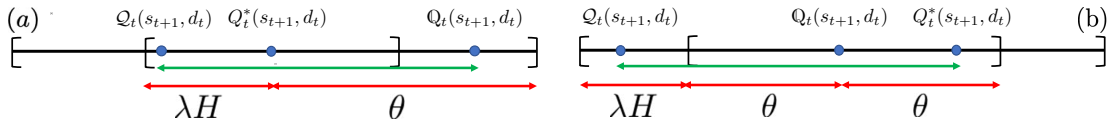


Figure 3: (a) When $|\mathcal{Q}_t(s_{t+1}, d_t) - Q_t^*(s_{t+1}, d_t)| \leq \lambda H$, $|\mathcal{Q}_t(s_{t+1}, d_t) - \mathbb{Q}_t(s_{t+1}, d_t)|$ (green) is at most $\lambda H + \theta$ (red) and hence $Q_t^\epsilon(s_{t+1}, d_t) = \mathcal{Q}_t(s_{t+1}, d_t)$. (b) When $|\mathcal{Q}_t(s_{t+1}, d_t) - Q_t^*(s_{t+1}, d_t)| > \lambda H$ and $|\mathcal{Q}_t(s_{t+1}, d_t) - \mathbb{Q}_t(s_{t+1}, d_t)| \leq \lambda H + \theta$, $Q_t^\epsilon(s_{t+1}, d_t) = \mathcal{Q}_t(s_{t+1}, d_t)$ due to (8) and $|Q_t^\epsilon(s_{t+1}, d_t) - Q_t^*(s_{t+1}, d_t)|$ (green) is at most $\lambda H + 2\theta$ (red). A rigorous analysis of the if condition in (8) is covered in (Kharkovskii *et al.*, 2020).

ϵ-Bayes optimality in practice. To ease this computational burden (e.g., for real-time planning), we propose an asymptotically optimal anytime variant of our ϵ-Macro-GPO policy that can attain good BO performance quickly and improve its approximation quality over time, as briefly discussed here and detailed along with the pseudocode in (Kharkovskii *et al.*, 2020).

The intuition behind our anytime ϵ-Macro-GPO algorithm is to incrementally expand a search tree by iteratively simulating greedy exploration paths down the partially constructed tree and expanding the sub-trees rooted at nodes with the largest uncertainty of their corresponding values $V_t^*(d_t)$ so as to improve their approximation quality. Such an uncertainty at each encountered node d_t is quantified by the gap between its maintained upper and lower heuristic bounds $\overline{V}_t^*(d_t)$ and $\underline{V}_t^*(d_t)$ that are (a) tightened via backpropagation from the leaves up through node d_t to the root d_0 and (b) subsequently used to refine that at its siblings by exploiting the Lipschitz continuity of V_t^*. Consequently, each iteration of our anytime ϵ-Macro-GPO algorithm only incurs linear time in N. The formulation of our anytime variant resembles that of ϵ-Macro-GPO policy π^ϵ (8) except that it utilizes the lower heuristic bound instead of $\mathcal{Q}_t$ (6) and a modified if condition to bound its expected performance loss likewise, as detailed in (Kharkovskii *et al.*, 2020).

4 Experiments and Discussion

This section empirically evaluates the performance of our nonmyopic adaptive ϵ-Macro-GPO policy and its anytime variant for a given finite budget with three datasets featuring simulated plankton density phenomena, a real-world traffic phenomenon, and a real-world temperature phenomenon over an office environment (Kharkovskii *et al.*, 2020) whose results are consistent with that here. The performances of our ϵ-Macro-GPO policy and its anytime variant are compared with that of state-of-the-art (a) nonmyopic GP-UCB (Marchant *et al.*, 2014) generalized to handle macro-actions that coincides with our deterministic policy (7) exploiting the most likely observations during planning, (b) *distributed batch GP-UCB* (DB-GP-UCB) (Daxberger and Low, 2017) that casts a macro-action as a batch to be optimized and is thus equivalent to ϵ-Macro-GPO with $H = 1$, (c) q-EI (Chevalier and Ginsbourger, 2013) that does likewise, and (d) greedy batch BO algorithms[10] such as GP-BUCB (Desautels *et al.*, 2014), GP-UCB-PE (Contal *et al.*, 2013), and BBO-LP (González *et al.*, 2016a) whose implementations are detailed in (Kharkovskii *et al.*, 2020). Four performance metrics are used: (a) average normalized[11] output measurements observed by the agent (larger average output measurements imply less average/cumulative regret (Section 3)), (b) simple regret (i.e., difference between global maximum and currently found maximum), (c) no. of explored nodes in all constructed search trees (more nodes incur more time), and (d) average time per stage.

Simulated plankton density phenomena. An *autonomous underwater vehicle* (AUV) is deployed on board of a *research vessel* (RV) in search for a hotspot of peak phytoplankton abundance (i.e., algal bloom) in coastal ocean. The AUV and RV are initially positioned near the center of the plankton density (mg/m^3) phenomenon spatially distributed over a 5 km by 5 km region that is discretized into a 50×50 grid of input locations. The phenomenon is modeled as a realization of a GP and simulated using the GP hyperparameters $\mu_s = 0$, $\ell_1 = \ell_2 = 0.5$ km, $\sigma_y^2 = 1$, and $\sigma_n^2 = 10^{-5}$. The AUV is tasked to execute the selected macro-action of a straight dive (due to limited maneuverability) along one of the 4 cardinal directions from the RV to gather "Gulper" water samples/observations over $\kappa = 4$ input locations for precise on-deck testing (Pennington *et al.*, 2016); given a budget of 20 observations, this is repeated for 5 times from the input location that it has previously surfaced.

Figs. 4a and 4b show results of the performances of ϵ-Macro-GPO with $H = 2, 3, 4$ (lookahead of, respectively, 8, 12, 16 observations), $\beta = 0$, and $N = 100$,[12] and the other tested BO algorithms averaged over

[10] Unlike DB-GP-UCB and q-EI, a greedy batch BO algorithm cannot exploit the full informativeness of any candidate macro-action for its macro-action selection: Since it selects the inputs of a batch one at a time myopically[2], its first few selected input locations immediately decide its chosen macro-action and consequently the remaining sequence of input locations found within.

[11] To ease interpretation of results, the prior mean is subtracted from each output measurement to normalize it.

[12] Specifying the value of N (instead of ϵ) may yield a loose ϵ based on Theorem 1. Nevertheless, the resulting

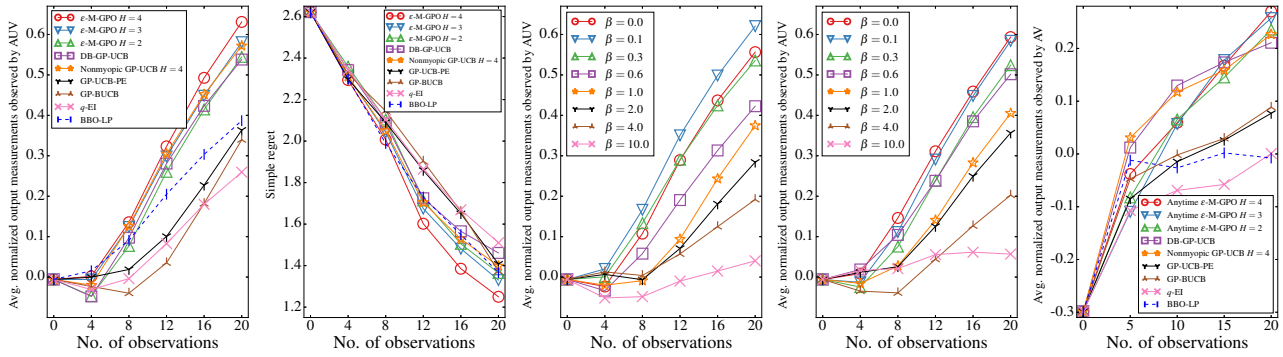


Figure 4: Graphs of (a) average normalized[11] output measurements observed by AUV, (b) simple regrets achieved by tested BO algorithms, average normalized output measurements achieved by ϵ-Macro-GPO (ϵ-M-GPO in the graphs) with (c) $H = 2$ and (d) $H = 3$ and varying exploration weights β vs. no. of observations for simulated plankton density phenomena and (e) average normalized[11] output measurements observed by the AV vs. no. of observations for real-world traffic phenomenon. Standard errors are given in (Kharkovskii *et al.*, 2020).

250 independent realizations of the simulated phenomena. It can be observed that as the number of observations increases, the nonmyopic adaptive BO algorithms generally outperform the myopic ones. In particular, the performance of ϵ-Macro-GPO improves considerably by increasing H: ϵ-Macro-GPO with the furthest lookahead (i.e., $H = 4$) achieves the largest average normalized output measurements observed by the AUV and smallest simple regret after 20 observations at the cost of a larger number of explored nodes (see (Kharkovskii *et al.*, 2020)). For example, the nonmyopic ϵ-Macro-GPO with $H = 4$ achieves $0.093\sigma_y$ ($0.059\sigma_y$) more average output measurements and $0.211\sigma_y$ ($0.148\sigma_y$) less simple regret than myopic DB-GP-UCB (nonmyopic GP-UCB with the same horizon $H = 4$ but assuming most likely observations during planning), which are expected.

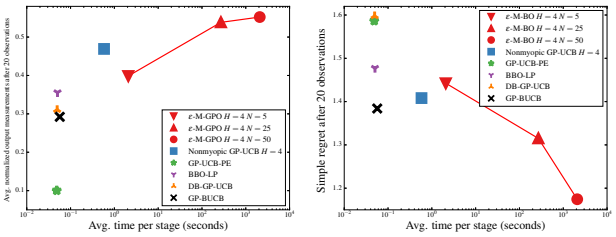
Figs. 4c and 4d show the effect of varying exploration weights β on the performance of ϵ-Macro-GPO with $H = 2$ and $H = 3$, respectively. It can be observed from Fig. 4c that when $H = 2$, ϵ-Macro-GPO with $\beta = 0.1$ achieves $0.064\sigma_y$ more average normalized output measurements than that with $\beta = 0$ after 20 observations, which indicates the need of a slightly stronger exploration behavior. Fig. 4d shows that by increasing to a lookahead of 12 observations (i.e., $H = 3$), ϵ-Macro-GPO no longer needs the additional weighted exploration term in (4) (i.e., $\beta = 0$) since it can naturally trade off between exploration vs. exploitation, as explained previously (Section 3). From Figs. 4c and 4d, $\beta = 10$ greatly hurts its performance due to an overly aggressive exploration.

We also investigate the effect of varying the number N of stochastic samples on the behavior of ϵ-Macro-GPO. To this end, ϵ-Macro-GPO with a fixed horizon H offers an advantage of being able to trade off its performance for time efficiency by decreasing N. This observation is theoretically validated in Theorem 1 and empirically illustrated in Fig. 5.

Figs. 5a and 5b show results of the performances of ϵ-Macro-GPO with $H = 4$ (lookahead of 16 observations), $\beta = 0$, and $N = 5, 25, 50$, and the other tested BO algorithms averaged over 35 independent realizations of the simulated plankton density phenomena. It can be observed that the performance of ϵ-Macro-GPO improves considerably by increasing N: ϵ-Macro-GPO with the largest number of samples (i.e., $N = 50$) achieves the largest average normalized output measurements and smallest simple regret after 20 observations at the cost of larger average time per iteration. For example, ϵ-Macro-GPO with $N = 50$ achieves $0.26\sigma_y$ more average output measurements and $0.21\sigma_y$ less simple regret than myopic GP-BUCB, but needs 2085.37 more seconds per iteration.

Real-world traffic phenomenon. To service the mobility demands within the central business district of an urban city, an *autonomous vehicle* (AV) in a mobility-on-demand system cruises along different road trajectories to find a hotspot of highest mobility demand to pick up a user. The 29.4 km by 11.9 km service area is gridded into 100×50 input regions, of which only 2506 input regions are accessible to the AV via the road network. The AV can cruise from input region s to an adjacent input region s' using one primitive action iff at least one road segment in the road network starts in s and ends in s'; the maximum outdegree from any input region is 8. In any input region, a surrogate demand measurement is obtained by counting the number of pickups from all historic taxi trajectories generated by a major taxi company during

[11] ϵ-Macro-GPO with $H = 3{,}4$ empirically outperforms other tested BO algorithms.



(a) (b)

Figure 5: Graphs of (a) average normalized[11] output measurements observed by AUV and (b) simple regrets achieved by tested BO algorithms vs. average time per stage for simulated plankton density phenomena.

9:30-10 p.m. on August 2, 2010 (Chen *et al.*, 2015b); the resulting mobility demand pattern is visualized in Appendix A.5. The original demand measurements are log-transformed to remove skewness and extremity for stabilizing the GP covariance structure and the GP hyperparameters $\mu_s = 1.5673$, $\ell_1 = 0.1689$ km, $\ell_2 = 0.1275$ km, $\sigma_y^2 = 0.7486$, and $\sigma_n^2 = 0.0111$ are then learned using maximum likelihood estimation (Rasmussen and Williams, 2006); note that the length-scales and signal-to-noise ratio are relatively smaller than that of the simulated plankton density phenomena. The AV is tasked to execute the selected macro-action of a cruising trajectory along $\kappa = 5$ adjacent input regions to observe their corresponding demand measurements; given a budget of 20 observations, this will be repeated for 4 times from the input region that it has previously cruised to. Since every input region s has a large number of available macro-actions (i.e., with an average of 178 and maximum of 1193), 20 of them are randomly selected to form its representative set of candidate macro-actions.

Fig. 4e shows results of the performances of *anytime* ϵ-Macro-GPO with $H = 2, 3, 4$ (lookahead of, respectively, 10, 15, 20 observations), $\beta = 0$, and $N = 300$ after running for 1500 iterations[12], and the other tested BO algorithms averaged over 35 random starting input regions of the AV. Similar to the results for simulated plankton density phenomena, the performance of anytime ϵ-Macro-GPO improves considerably by increasing H: Anytime ϵ-Macro-GPO with the furthest lookahead (i.e., $H = 4$) achieves the largest average normalized output measurements observed by the AV and among the least simple regret (see (Kharkovskii *et al.*, 2020)) after 20 observations at the cost of a larger number of explored nodes ((Kharkovskii *et al.*, 2020)). For example, the nonmyopic anytime ϵ-Macro-GPO with $H = 4$ achieves $0.069\sigma_y$ ($0.05\sigma_y$) more average output measurements and $0.188\sigma_y$ ($0.219\sigma_y$) less simple regret than myopic DB-GP-UCB (nonmyopic GP-UCB with $H = 4$), which are expected. The effect of varying exploration weights β on the performance of anytime ϵ-Macro-GPO is similar to that for the simulated plankton density phenomena and reported in (Kharkovskii *et al.*, 2020).

Lastly, we investigate the effect of downsampling the number of available macro-actions per input region to 20 on the performance of anytime ϵ-Macro-GPO. The results are reported in (Kharkovskii *et al.*, 2020) and show that anytime ϵ-Macro-GPO with $H = 4$ and 20 randomly selected macro-actions outperforms that with $H = 2$ and all available macro-actions at the cost of a larger number of explored nodes.

5 Conclusion and Future Work

This paper describes ϵ-Macro-GPO and its anytime variant for nonmyopic adaptive BO that have been empirically shown to scale up to a lookahead of 20 observations by exploiting macro-actions and consequently achieve superior BO performance. Different from the asymptotic no-regret performance[1] typical of GP-UCB and its variants, we theoretically guarantee the *expected* performance loss of ϵ-Macro-GPO and its anytime variant that can be specified to be arbitrarily small given a *limited* budget. Though this requires a polynomial number of stochastic samples in the macro-action length κ in each planning stage (Theorem 1), our experiments reveal that a relatively small sample size (N=100-300) is needed for ϵ-Macro-GPO and its anytime variant to outperform state-of-the-art BO algorithms. Though a sufficiently large exploration weight β is usually needed to guarantee asymptotic no-regret performance[1] for GP-UCB and its variants, we have observed in our experiments that their performances are highly sensitive to the chosen value of β given a finite/limited budget and can be greatly hurt by an often unknowingly "large" value of β due to excessive exploration. To sidestep this, ϵ-Macro-GPO can eliminate the need of β (i.e., $\beta = 0$) by utilizing a further lookahead, that is, if computational resources permit or are more affordable than the cost of function evaluations. For future work, we plan to generalize ϵ-Macro-GPO and its anytime variant to nonmyopic batch active learning (Cao *et al.*, 2013; Hoang *et al.*, 2014a,b; Low *et al.*, 2008, 2009, 2011, 2012, 2014a; Ouyang *et al.*, 2014; Zhang *et al.*, 2016), high-dimensional BO (Hoang *et al.*, 2018), and multi-fidelity BO (Zhang *et al.*, 2017, 2019) settings. For applications with a huge budget of function evaluations, we like to couple ϵ-Macro-GPO and its anytime variant with the use of distributed/decentralized (Chen *et al.*, 2012, 2013a,b, 2015a; Hoang *et al.*, 2016, 2019b,a; Low *et al.*, 2015; Ouyang and Low, 2018) or online/stochastic (Hoang *et al.*, 2015, 2017; Low *et al.*, 2014b; Xu *et al.*, 2014; Teng *et al.*, 2020; Yu *et al.*, 2019a,b) sparse GP models to represent the belief of the unknown objective function efficiently.

Acknowledgments

This research is supported by the Singapore Ministry of Education Academic Research Fund Tier 2, MOE2016-T2-2-156.

References

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence.*. Oxford University Press.

Bull, A. D. (2011). Convergence rates of efficient global optimization algorithms. *JMLR*, **12**, 2879–2904.

Cao, N., Low, K. H., and Dolan, J. M. (2013). Multi-robot informative path planning for active sensing of environmental phenomena: A tale of two algorithms. In *Proc. AAMAS*, pages 7–14.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, **12**(6), 805–849.

Chen, J., Low, K. H., Tan, C. K.-Y., Oran, A., Jaillet, P., Dolan, J. M., and Sukhatme, G. S. (2012). Decentralized data fusion and active sensing with mobile sensors for modeling and predicting spatiotemporal traffic phenomena. In *Proc. UAI*, pages 163–173.

Chen, J., Cao, N., Low, K. H., Ouyang, R., Tan, C. K.-Y., and Jaillet, P. (2013a). Parallel Gaussian process regression with low-rank covariance matrix approximations. In *Proc. UAI*, pages 152–161.

Chen, J., Low, K. H., and Tan, C. K.-Y. (2013b). Gaussian process-based decentralized data fusion and active sensing for mobility-on-demand system. In *Proc. RSS*.

Chen, J., Low, K. H., Jaillet, P., and Yao, Y. (2015a). Gaussian process decentralized data fusion and active sensing for spatiotemporal traffic modeling and prediction in mobility-on-demand systems. *IEEE Trans. Autom. Sci. Eng.*, **12**, 901–921.

Chen, J., Low, K. H., Jaillet, P., and Yao, Y. (2015b). Gaussian process decentralized data fusion and active sensing for spatiotemporal traffic modeling and prediction in mobility-on-demand systems. *IEEE T-ASE*, **12**(3), 901–921.

Chevalier, C. and Ginsbourger, D. (2013). Fast computation of the multi-points expected improvement with applications in batch selection. In *Proc. 7th International Conference on Learning and Intelligent Optimization*, pages 59–69.

Choi, J.-H., Loftness, V., and Aziz, A. (2012). Post-occupancy evaluation of 20 office buildings as basis for future IEQ standards and guidelines. *Energy and Buildings*, **46**, 167–175.

Contal, E., Buffoni, D., Robicquet, A., and Vayatis, N. (2013). Parallel Gaussian process optimization with upper confidence bound and pure exploration. In *Proc. ECML/PKDD*, pages 225–240.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience, second edition.

Daxberger, E. A. and Low, K. H. (2017). Distributed batch Gaussian process optimization. In *Proc. ICML*, pages 951–960.

Desautels, T., Krause, A., and Burdick, J. W. (2014). Parallelizing exploration-exploitation tradeoffs in Gaussian process bandit optimization. *JMLR*, **15**, 4053–4103.

Golub, G. H. and Van Loan, C.-F. (1996). *Matrix Computations*. Johns Hopkins Univ. Press, third edition.

González, J., Dai, Z., Hennig, P., and Lawrence, N. D. (2016a). Batch Bayesian optimization via local penalization. In *Proc. AISTATS*, pages 648–657.

González, J., Osborne, M., and Lawrence, N. D. (2016b). GLASSES: Relieving the myopia of Bayesian optimisation. In *Proc. AISTATS*, pages 790–799.

He, R., Brunskill, E., and Roy, N. (2010). PUMA: Planning under uncertainty with macro-actions. In *Proc. AAAI*, pages 1089–1095.

He, R., Brunskill, E., and Roy, N. (2011). Efficient planning under uncertainty with macro-actions. *JAIR*, **40**, 523–570.

Hennig, P. and Schuler, C. J. (2012). Entropy search for information-efficient global optimization. *JMLR*, **13**, 1809–1837.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. In *Proc. NIPS*, pages 918–926.

Hoang, Q. M., Hoang, T. N., and Low, K. H. (2017). A generalized stochastic variational Bayesian hyperparameter learning framework for sparse spectrum Gaussian process regression. In *Proc. AAAI*, pages 2007–2014.

Hoang, Q. M., Hoang, T. N., Low, K. H., and Kingsford, C. (2019a). Collective model fusion for multiple black-box experts. In *Proc. ICML*, pages 2742–2750.

Hoang, T. N., Low, K. H., Jaillet, P., and Kankanhalli, M. (2014a). Active learning is planning: Nonmyopic ϵ-Bayes-optimal active learning of Gaussian

processes. In *Proc. ECML/PKDD Nectar Track*, pages 494–498.

Hoang, T. N., Low, K. H., Jaillet, P., and Kankanhalli, M. (2014b). Nonmyopic ϵ-Bayes-optimal active learning of Gaussian processes. In *Proc. ICML*, pages 739–747.

Hoang, T. N., Hoang, Q. M., and Low, K. H. (2015). A unifying framework of anytime sparse Gaussian process regression models with stochastic variational inference for big data. In *Proc. ICML*, pages 569–578.

Hoang, T. N., Hoang, Q. M., and Low, K. H. (2016). A distributed variational inference framework for unifying parallel sparse Gaussian process regression models. In *Proc. ICML*, pages 382–391.

Hoang, T. N., Hoang, Q. M., and Low, K. H. (2018). Decentralized high-dimensional Bayesian optimization with factor graphs. In *Proc. AAAI*, pages 3231–3238.

Hoang, T. N., Hoang, Q. M., Low, K. H., and How, J. P. (2019b). Collective online learning of Gaussian processes in massive multi-agent systems. In *Proc. AAAI*.

Kharkovskii, D., Ling, C. K., and Low, K. H. (2020). Nonmyopic Gaussian process optimization with macro-actions. arXiv:2002.09670.

Konidaris, G. and Barto, A. G. (2007). Building portable options: Skill transfer in reinforcement learning. In *Proc. IJCAI*, pages 895–900.

Lam, R. R. and Willcox, K. E. (2017). Lookahead Bayesian optimization with inequality constraints. In *Proc. NIPS*.

Lam, R. R., Willcox, K. E., and Wolpert, D. H. (2016). Bayesian optimization with a finite budget: An approximate dynamic programming approach. In *Proc. NIPS*.

Leonard, N. E., Paley, D. A., Lekien, F., Sepulchre, R., Fratantoni, D. M., and Davis, R. E. (2007). Collective motion, sensor networks, and ocean sampling. *Proceedings of the IEEE*, **95**(1), 48–74.

Lim, Z., Lee, W. S., and Hsu, D. (2011). Monte Carlo value iteration with macro-actions. In *Proc. NIPS*, pages 1287–1295.

Ling, C. K., Low, K. H., and Jaillet, P. (2016). Gaussian process planning with Lipschitz continuous reward functions: Towards unifying Bayesian optimization, active learning, and beyond. In *Proc. AAAI*, pages 1860–1866.

Low, K. H., Dolan, J. M., and Khosla, P. (2008). Adaptive multi-robot wide-area exploration and mapping. In *Proc. AAMAS*, pages 23–30.

Low, K. H., Dolan, J. M., and Khosla, P. (2009). Information-theoretic approach to efficient adaptive path planning for mobile robotic environmental sensing. In *Proc. ICAPS*, pages 233–240.

Low, K. H., Dolan, J. M., and Khosla, P. (2011). Active Markov information-theoretic path planning for robotic environmental sensing. In *Proc. AAMAS*, pages 753–760.

Low, K. H., Chen, J., Dolan, J. M., Chien, S., and Thompson, D. R. (2012). Decentralized active robotic exploration and mapping for probabilistic field classification in environmental sensing. In *Proc. AAMAS*, pages 105–112.

Low, K. H., Chen, J., Hoang, T. N., Xu, N., and Jaillet, P. (2014a). Recent advances in scaling up Gaussian process predictive models for large spatiotemporal data. In S. Ravela and A. Sandu, editors, *Dynamic Data-Driven Environmental Systems Science: First International Conference, DyDESS 2014*, pages 167–181. LNCS 8964, Springer International Publishing.

Low, K. H., Xu, N., Chen, J., Lim, K. K., and Özgül, E. B. (2014b). Generalized online sparse Gaussian processes with application to persistent mobile robot localization. In *Proc. ECML/PKDD Nectar Track*, pages 499–503.

Low, K. H., Yu, J., Chen, J., and Jaillet, P. (2015). Parallel Gaussian process regression for big data: Low-rank representation meets Markov approximation. In *Proc. AAAI*, pages 2821–2827.

Marchant, R., Ramos, F., and Sanner, S. (2014). Sequential Bayesian optimisation for spatial-temporal monitoring. In *Proc. UAI*, pages 553–562.

Osborne, M. A., Garnett, R., and Roberts, S. J. (2009). Gaussian processes for global optimization. In *Proc. 3rd International Conference on Learning and Intelligent Optimization*.

Ouyang, R. and Low, K. H. (2018). Gaussian process decentralized data fusion meets transfer learning in large-scale distributed cooperative perception. In *Proc. AAAI*, pages 3876–3883.

Ouyang, R., Low, K. H., Chen, J., and Jaillet, P. (2014). Multi-robot active sensing of non-stationary Gaussian process-based environmental phenomena. In *Proc. AAMAS*, pages 573–580.

Pennington, J. T., Blum, M., and Chavez, F. P. (2016). Seawater sampling by an autonomous underwater vehicle: "Gulper" sample validation for nitrate, chlorophyll, phytoplankton, and primary production. *Limnol. Oceanogr.: Methods*, **14**(1), 14–23.

Petersen, K. B. and Pedersen, M. S. (2012). *The Matrix Cookbook*.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning.*. MIT Press.

Shah, A. and Ghahramani, Z. (2015). Parallel predictive entropy search for batch global optimization of expensive objective functions. In *Proc. NIPS*, pages 3312–3320.

Shahriari, B., Swersky, K., Wang, Z., Adams, R., and de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, **104**(1), 148–175.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. ICML*, pages 1015–1022.

Stewart, G. W. and Sun, J.-G. (1990). *Matrix Perturbation Theory*. Academic Press.

Stolle, M. and Precup, D. (2002). Learning options in reinforcement learning. In *Proc. International Symposium on Abstraction, Reformulation, and Approximation*, pages 212–223.

Taboga, M. (2017). *Lectures on probability theory and mathematical statistics*. CreateSpace Independent Publishing Platform. `http://www.statlect.com`.

Teng, T., Chen, J., Zhang, Y., and Low, K. H. (2020). Scalable variational bayesian kernel selection for sparse Gaussian process regression. In *Proc. AAAI*.

Vazquez, E. and Bect, J. (2010). Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *J. Statistical Planning and Inference*, **140**(11), 3088–3095.

Wu, J. and Frazier, P. (2016). The parallel knowledge gradient method for batch Bayesian optimization. In *Proc. NIPS*, pages 3126–3134.

Xu, N., Low, K. H., Chen, J., Lim, K. K., and Özgül, E. B. (2014). GP-Localize: Persistent mobile robot localization using online sparse Gaussian process observation model. In *Proc. AAAI*, pages 2585–2592.

Yu, H., Chen, Y., Dai, Z., Low, K. H., and Jaillet, P. (2019a). Implicit posterior variational inference for deep Gaussian processes. In *Proc. NeurIPS*.

Yu, H., Hoang, T. N., Low, K. H., and Jaillet, P. (2019b). Stochastic variational inference for Bayesian sparse Gaussian process regression. In *Proc. IJCNN*.

Zhang, Y., Hoang, T. N., Low, K. H., and Kankanhalli, M. (2016). Near-optimal active learning of multi-output Gaussian processes. In *Proc. AAAI*, pages 2351–2357.

Zhang, Y., Hoang, T. N., Low, K. H., and Kankanhalli, M. (2017). Information-based multi-fidelity Bayesian optimization. In *Proc. NIPS Workshop on Bayesian Optimization*.

Zhang, Y., Dai, Z., and Low, K. H. (2019). Bayesian optimization with binary auxiliary information. In *Proc. UAI*.