

---

# Bayesian Optimization with Binary Auxiliary Information

---

Yehong Zhang, Zhongxiang Dai, and Bryan Kian Hsiang Low

Department of Computer Science, National University of Singapore, Republic of Singapore  
{yehong, daizhongxiang, lowkh}@comp.nus.edu.sg

## Abstract

This paper presents novel mixed-type *Bayesian optimization* (BO) algorithms to accelerate the optimization of a target objective function by exploiting correlated auxiliary information of binary type that can be more cheaply obtained, such as in policy search for reinforcement learning and hyperparameter tuning of machine learning models with early stopping. To achieve this, we first propose a mixed-type *multi-output Gaussian process* (MOGP) to jointly model the continuous target function and binary auxiliary functions. Then, we propose information-based acquisition functions such as *mixed-type entropy search* (MT-ES) and *mixed-type predictive ES* (MT-PES) for mixed-type BO based on the MOGP predictive belief of the target and auxiliary functions. The exact acquisition functions of MT-ES and MT-PES cannot be computed in closed form and need to be approximated. We derive an efficient approximation of MT-PES via a novel mixed-type random features approximation of the MOGP model whose cross-correlation structure between the target and auxiliary functions can be exploited for improving the belief of the global target maximizer using observations from evaluating these functions. We propose new practical constraints to relate the global target maximizer to the binary auxiliary functions. We empirically evaluate the performance of MT-ES and MT-PES with synthetic and real-world experiments.

## 1 INTRODUCTION

*Bayesian optimization* (BO) has recently demonstrated with notable success to be highly effective in optimizing an unknown (possibly noisy, non-convex, and/or with

no closed-form expression/derivative) target function using a finite budget of often expensive function evaluations (Shahriari *et al.*, 2016). As an example, BO is used by Snoek *et al.* (2012) to determine the setting of input hyperparameters (e.g., learning rate, batch size of data) of a *machine learning* (ML) model that maximize its validation accuracy (i.e., output of the unknown target function). Conventionally, a BO algorithm relies on some choice of acquisition function (e.g., improvement-based (Shahriari *et al.*, 2016) such as probability of improvement or *expected improvement* (EI) over currently found maximum, information-based (Villemonteix *et al.*, 2009) such as *entropy search* (ES) (Hennig and Schuler, 2012) and *predictive entropy search* (PES) (Hernández-Lobato *et al.*, 2014), or *upper confidence bound* (UCB) (Srinivas *et al.*, 2010)) as a heuristic to guide its search for the global target maximizer. To do this, the BO algorithm exploits the chosen acquisition function to repeatedly select an input for evaluating the unknown target function that trades off between sampling at or near to a likely target maximizer based on a *Gaussian process* (GP) belief of the unknown target function (exploitation) vs. improving the GP belief (exploration) until the budget is expended.

In practice, the expensive-to-evaluate target function often correlates well with some cheaper-to-evaluate *binary* auxiliary function(s) that delineate the input regions potentially containing the global target maximizer and can thus be exploited to boost the BO performance. For example, automatically tuning the hyperparameters of a sophisticated ML model (e.g., deep neural network) with BO is usually time-consuming as it may incur several hours to days to evaluate the validation accuracy of the ML model under each selected hyperparameter setting when training with a massive dataset. To accelerate this process, consider an auxiliary function whose output is a binary decision of whether the validation accuracy of the ML model under the selected input hyperparameter setting will exceed a pre-specified threshold, which is recommended by some early/optimal stopping mecha-

nism (Müller *et al.*, 2007) after a small number of training epochs. Such auxiliary information of binary type is cheaper to obtain and can quickly delineate the input regions containing the best hyperparameter setting, hence incurring less time for exploration. Similarly, to find the best reinforcement learning policy for an AI agent in a game or a real robot in a task with binary outcomes (e.g., success or failure) (Tesch *et al.*, 2013), maximizing the success rate (i.e., the unknown target function with a continuous output type) averaged over multiple random environments can be accelerated by deciding whether the selected setting of input policy parameters is promising in a single environment (i.e., the auxiliary function with a binary output type). To search for the optimal setting of a system via user interaction (Shahriari *et al.*, 2016), gathering implicit/binary user feedback (e.g., click or not, like or dislike) is often easier than asking for an explicit rating/ranking of a shown example. The above practical examples motivate the need to design and develop a *mixed-type* BO algorithm that can naturally trade off between exploitation vs. exploration over the target function with a *continuous* output type and the cheaper-to-evaluate auxiliary function(s) with a *binary* output type for finding or improving the belief of the global target maximizer, which is the focus of our work here.

In this paper, we generalize information-based acquisition functions like ES and PES to *mixed-type ES* (MT-ES) and *mixed-type PES* (MT-PES) for mixed-type BO (Section 4). To the best of our knowledge, these are the first BO algorithms that exploit correlated *binary* auxiliary information for accelerating the optimization of a continuous target objective function. Different from *continuous* auxiliary functions which have been exploited by a number of multi-fidelity BO algorithms (Huang *et al.*, 2006; Swersky *et al.*, 2013; Kandasamy *et al.*, 2016, 2017; Poloczek *et al.*, 2017; Sen *et al.*, 2018), the binary auxiliary functions in our problem make the widely used Gaussian likelihood inappropriate and prevent a direct application of existing multi-fidelity BO algorithms.<sup>1</sup>

To resolve this, we first propose a mixed-type multi-output GP to jointly model the unknown continuous target function and binary auxiliary functions. Although the exact acquisition function of MT-PES cannot be computed in closed form, the main contribution of our work here is to show that it is in fact possible to derive an efficient approximation of MT-PES via (a) a novel *mixed-type random features* (MT-RF) approximation of the MOGP model whose cross-correlation structure between the target and auxiliary functions can be exploited for improving the belief of the global target maximizer using the observations from evaluating these functions (Section 5.1), and (b) new

practical constraints relating the global target maximizer to the binary auxiliary functions (Section 5.2). We empirically evaluate the performance of MT-ES and MT-PES with synthetic and real-world experiments (Section 6).

## 2 PROBLEM SETUP

In this work, we have access to an unknown target objective function  $f_1$  and  $M - 1$  auxiliary functions  $f_2, \dots, f_M$  defined over a bounded input domain  $D \subset \mathbb{R}^d$  such that each input  $x \in D$  is associated with a noisy output  $y_i(x)$  for  $i = 1, \dots, M$ . As mentioned in Section 1, a cost  $\lambda_i(x)$  is incurred to evaluate function  $f_i$  at each input  $x \in D$  and the target function is more costly to evaluate than the auxiliary functions, i.e.,  $\lambda_1(x) > \lambda_i(x)$  for  $i = 2, \dots, M$ . Then, the objective is to find the global target maximizer  $x_* \triangleq \arg \max_{x \in D} f_1(x)$  with a lower cost by exploiting the cheaper auxiliary function evaluations, as compared to evaluating only the target function. Our problem differs from that of the conventional multi-fidelity BO in that only the target function returns continuous outputs (i.e.,  $y_1(x) \in \mathbb{R}$ ) while the auxiliary functions return binary outputs (i.e.,  $y_i(x) \in \{1, -1\}$  for  $i = 2, \dots, M$ ).

## 3 MIXED-TYPE MULTI-OUTPUT GP

Various types of multi-output GP models (Cressie, 1993; Wackernagel, 1998; Webster and Oliver, 2007; Skolidis, 2012; Bonilla *et al.*, 2007; Teh and Seeger, 2005; Álvarez and Lawrence, 2011) have been used to jointly model target and auxiliary functions with continuous outputs. However, none of them can be used straightforwardly in our problem to model the mixed output types due to the non-Gaussian likelihood  $p(y_i(x)|f_i(x))$  of the auxiliary functions. To resolve this issue, we generalize the *convolved multi-output Gaussian process* (CMOGP) to model the correlated functions with mixed continuous and binary output types by approximating the non-Gaussian likelihood using *expectation propagation* (EP), as discussed later. The CMOGP model is chosen for generalization due to its convolutional structure which can be exploited for deriving an efficient approximation of our acquisition function, as described in Section 5.

Let the target and auxiliary functions  $f_1, \dots, f_M$  be jointly modeled as a CMOGP which defines each function  $f_i$  as a convolution between a smoothing kernel  $K_i$  and a latent function<sup>2</sup>  $L$  with an additive bias  $m_i$ :

$$f_i(x) \triangleq m_i + \int_{x' \in D} K_i(x - x') L(x') dx'. \quad (1)$$

<sup>2</sup>To ease exposition, we consider a single latent function. Note, however, multiple latent functions can be used to improve the modeling (Álvarez and Lawrence, 2011). More importantly, our proposed MT-RF approximation and MT-PES algorithm can be easily generalized to handle multiple latent functions, as shown in Appendix G.

<sup>1</sup>We discuss other related works in Appendix A.

Let  $D_i^+ \triangleq \{\langle x, i \rangle\}_{x \in D}$  and  $D^+ \triangleq \bigcup_{i=1}^M D_i^+$ . As shown by Álvarez and Lawrence (2011), if  $\{L(x)\}_{x \in D}$  is a GP, then  $\{f_i(x)\}_{\langle x, i \rangle \in D^+}$  is also a GP, that is, every finite subset of  $\{f_i(x)\}_{\langle x, i \rangle \in D^+}$  follows a multivariate Gaussian distribution. Such a GP is fully specified by its *prior* mean  $\mu_i(x) \triangleq \mathbb{E}[f_i(x)]$  and covariance  $\sigma_{ij}(x, x') \triangleq \text{cov}[f_i(x), f_j(x')]$  for all  $\langle x, i \rangle, \langle x', j \rangle \in D^+$ , the latter of which characterizes both the correlation structure within each function (i.e.,  $i = j$ ) and the cross-correlation between different functions (i.e.,  $i \neq j$ ). Specifically, let  $\{L(x)\}_{x \in D}$  be a GP with zero mean, prior covariance  $\sigma_{xx'} \triangleq \mathcal{N}(x - x' | \mathbf{0}, \Gamma^{-1})$ , and  $K_i(x) \triangleq \sigma_{s_i} \mathcal{N}(x | \mathbf{0}, P_i^{-1})$  where  $\sigma_{s_i}^2$  is the signal variance controlling the intensity of the outputs of  $f_i(x)$ ,  $\Gamma$  and  $P_i$  are diagonal precision matrices controlling, respectively, the degrees of correlation between outputs of latent function  $L(x)$  and cross-correlation between outputs of  $L(x)$  and  $f_i(x)$ . Then,  $\mu_i(x) = m_i$  and

$$\sigma_{ij}(x, x') = \sigma_{s_i} \sigma_{s_j} \mathcal{N}(x - x' | \mathbf{0}, \Gamma^{-1} + P_i^{-1} + P_j^{-1}). \quad (2)$$

In this work, we assume the Gaussian and probit likelihoods for the target and auxiliary functions, respectively:

$$\begin{aligned} p(y_1(x) | f_1(x)) &\triangleq \mathcal{N}(f_1(x), \sigma_{n_1}^2), \\ p(y_i(x) | f_i(x)) &\triangleq \Phi_{\text{cdf}}(y_i(x) | f_i(x)) \end{aligned} \quad (3)$$

for  $i = 2, \dots, M$ . Supposing a column vector  $y_X \triangleq (y_i(x))_{\langle x, i \rangle \in X}^\top$  of outputs are observed by evaluating each  $i$ -th function  $f_i$  at a set  $X_i \subset D_i^+$  of input tuples where  $X \triangleq \bigcup_{i=1}^M X_i$ , the predictive belief/distribution of  $f_Z \triangleq (f_i(x))_{\langle x, i \rangle \in Z}^\top$  for any set  $Z \subseteq D^+ \setminus X$  of input tuples can be computed by

$$p(f_Z | y_X) = \int p(f_Z | f_X) p(f_X | y_X) df_X. \quad (4)$$

For conventional CMOGP with only continuous output types, (4) can be computed analytically since both  $p(f_Z | f_X)$  and  $p(f_X | y_X)$  are Gaussians (Álvarez and Lawrence, 2011). Unfortunately, the non-Gaussian likelihood in (3) makes the integral in (4) intractable. To resolve this issue, the work of Pourmohamad and Lee (2016) has proposed a sampling strategy based on a sequential Monte Carlo algorithm which, however, is computationally inefficient and makes the approximation of our proposed acquisition function (Section 5) prohibitively expensive. In contrast, we approximate the non-Gaussian likelihood using EP to derive an analytical approximation of (4), as detailed later. EP will be further exploited in Section 5 for approximating our proposed acquisition function efficiently.

### 3.1 MIXED-TYPE CMOGP PREDICTIVE INFERENCE

Let  $X_B \triangleq \bigcup_{i=2}^M X_i$  be a set of input tuples of the auxiliary functions. The posterior distribution  $p(f_X | y_X)$  in (4)

can be computed by

$$\begin{aligned} &p(f_{X_1}, f_{X_B} | y_{X_1}, y_{X_B}) \\ &\propto p(f_{X_1}, f_{X_B}) p(y_{X_1} | f_{X_1}) p(y_{X_B} | f_{X_B}) \\ &= p(f_{X_1} | f_{X_B}) p(f_{X_B}) p(y_{X_1} | f_{X_1}) \prod_{\langle x, i \rangle \in X_B} p(y_i(x) | f_i(x)) \\ &= p(f_{X_1} | f_{X_B}) p(y_{X_1} | f_{X_1}) q(f_{X_B}) \end{aligned} \quad (5)$$

where  $q(f_{X_B}) \triangleq p(f_{X_B}) \prod_{\langle x, i \rangle \in X_B} p(y_i(x) | f_i(x))$  can be approximated with a multivariate Gaussian  $\mathcal{N}(f_{X_B} | \tilde{\mu}_B, \tilde{\Sigma}_B)$  using EP by approximating each non-Gaussian likelihood as a Gaussian. Let

$$\begin{aligned} p(y_i(x) | f_i(x)) &= \Phi_{\text{cdf}}(y_i(x) | f_i(x)) \\ &\approx \tilde{Z}_i(x) \mathcal{N}(f_i(x) | \tilde{\mu}_i(x), \tilde{\sigma}_i^2(x)) \end{aligned} \quad (6)$$

for all  $\langle x, i \rangle \in X_B$ . Following the EP procedure in Section 3.6 of Rasmussen and Williams (2006), the parameters  $\tilde{\mu}_i(x)$  and  $\tilde{\sigma}_i^2(x)$  can be computed analytically and

$$\begin{aligned} \tilde{\mu}_B &= \Sigma_{X_B X_B} (\tilde{\Sigma}^{-1} \tilde{\mu} + \Sigma_{X_B X_B}^{-1} \mu_{X_B}) \\ \tilde{\Sigma}_B &= (\tilde{\Sigma}^{-1} + \Sigma_{X_B X_B}^{-1})^{-1} \end{aligned} \quad (7)$$

where  $\tilde{\mu} \triangleq (\tilde{\mu}_i(x))_{\langle x, i \rangle \in X_B}^\top$ ,  $\tilde{\Sigma}$  is a diagonal matrix with diagonal components  $\tilde{\sigma}_i^2(x)$  for  $\langle x, i \rangle \in X_B$ ,  $\Sigma_{AA'} \triangleq (\sigma_{ij}(x, x'))_{\langle x, i \rangle \in A, \langle x', j \rangle \in A'}$ , and  $\mu_A \triangleq (\mu_i(x))_{\langle x, i \rangle \in A}^\top$  for any  $A, A' \subseteq D^+$ .

By combining (7), (5), and (3) with (4) (Appendix B), the predictive belief  $p(f_Z | y_X)$  can be approximated by a multivariate Gaussian  $\mathcal{N}(\mu_{Z|X}, \Sigma_{Z|X})$  with the following *posterior* mean vector and covariance matrix:

$$\begin{aligned} \mu_{Z|X} &\triangleq \mu_Z + \Sigma_{ZX} \Lambda^{-1} (\tilde{y}_X - \mu_X) \\ \Sigma_{Z|X} &\triangleq \Sigma_{ZZ} - \Sigma_{ZX} \Lambda^{-1} \Sigma_{XZ} \end{aligned} \quad (8)$$

where  $\Lambda \triangleq \begin{bmatrix} \Sigma_{X_1 X_1} + \Sigma_n & \Sigma_{X_1 X_B} \\ \Sigma_{X_B X_1} & \Sigma_{X_B X_B} + \tilde{\Sigma} \end{bmatrix}$ ,  $\tilde{y}_X \triangleq [y_{X_1}; \tilde{\mu}]$ , and  $\Sigma_n$  is a  $|X_1| \times |X_1|$  diagonal matrix with diagonal components  $\sigma_{n_1}^2$ . Consequently, the approximated predictive belief of  $y_i(x)$  for any input tuple  $\langle x, i \rangle \in D^+$  can be computed using  $p(y_i(x) | y_X) = \int p(y_i(x) | f_i(x)) p(f_i(x) | y_X) df_i(x)$ . Due to (3) and (8),

$$\begin{aligned} p(y_1(x) | y_X) &\approx \mathcal{N}(y_1(x) | \mu_{\{1\}|X}, \sigma_{\{1\}|X}^2 + \sigma_{n_1}^2) \\ p(y_i(x) = 1 | y_X) &\approx \Phi_{\text{cdf}}\left(\mu_{\{i\}|X} / \sqrt{1 + \sigma_{\{i\}|X}^2}\right) \end{aligned} \quad (9)$$

for  $i = 2, \dots, M$  where  $\sigma_{\{i\}|X}^2 \triangleq \Sigma_{\{\langle x, i \rangle\} \{\langle x, i \rangle\} | X}$  for  $i = 1, \dots, M$ .

## 4 BO WITH BINARY AUXILIARY INFORMATION

To achieve the objective described in Section 2, our BO algorithm repeatedly selects the next input tuple  $\langle x, i \rangle$

for evaluating the  $i$ -th function  $f_i$  at  $x$  that maximizes a choice of acquisition function  $\alpha(y_X, \langle x, i \rangle)$  per unit cost given the past observations  $(X, y_X)$ :

$$\langle x, i \rangle^+ \triangleq \arg \max_{\langle x, i \rangle \in D^+ \setminus X} \alpha(y_X, \langle x, i \rangle) / \lambda_i(x)$$

and updates  $X \leftarrow X \cup \{\langle x, i \rangle^+\}$  until the budget is expended. Since the costs of evaluating the target vs. auxiliary functions differ, we use the above *cost-sensitive* acquisition function such that the cheaper auxiliary function evaluations can be exploited. We will focus on designing the acquisition function  $\alpha$  first and the estimation of  $\lambda_i(x)$  in real-world applications will be discussed later in Section 6.

Intuitively,  $\alpha$  should be designed to enable its BO algorithm to jointly and naturally optimize the non-trivial trade-off between exploitation vs. exploration over the target and auxiliary functions for finding or improving the belief of the global target maximizer  $x_*$  by utilizing information from the mixed-type CMOGP predictive belief of these functions (8). To do this, one may be tempted to directly use the conventional EI (Mockus *et al.*, 1978) and  $\text{EI}_\pi$  (Tesch *et al.*, 2013) acquisition functions for selecting inputs to evaluate the target and auxiliary functions, respectively.  $\text{EI}_\pi$  is a variation of EI and, to the best of our knowledge, the only acquisition function designed for optimizing an unknown function with a binary output type. However, this does not satisfy our objective since  $\text{EI}_\pi$  aims to find the global maximizer of the auxiliary function which can differ from the global target maximizer if the target and auxiliary functions are not perfectly correlated. To resolve this issue, we propose to exploit information-based acquisition functions and generalize them to our mixed-type BO problem such that input tuples for evaluating the target and auxiliary functions are selected to directly maximize *only* the unknown target objective function, as detailed later.

#### 4.1 INFORMATION-BASED ACQUISITION FUNCTIONS FOR MIXED-TYPE BO

Information-based acquisition functions like ES and PES have been designed to enable their BO algorithms to improve the belief of the global target maximizer. In mixed-type BO, we can similarly define a belief of the maximizer  $x_{*i}$  of each  $i$ -th function  $f_i$  as  $p(x_{*i} | y_X) \triangleq p(f_i(x_{*i}) = \max_{x \in D} f_i(x) | y_X)$  for  $i = 1, \dots, M$ . To achieve the objective of maximizing *only* the target function in mixed-type BO, ES can be used to measure the information gain of *only* the global target maximizer  $x_*$  (i.e.,  $x_{*1}$ ) from selecting the next input tuple  $\langle x, i \rangle$  for evaluating the  $i$ -th (possibly binary auxiliary) function  $f_i$  at  $x$  given the past observations  $(X, y_X)$ :

$$\alpha(y_X, \langle x, i \rangle) \triangleq H(x_* | y_X) - \mathbb{E}_{p(y_i(x) | y_X)} [H(x_* | y_X \cup \{\langle x, i \rangle\})]. \quad (10)$$

Similar to the *multi-task ES* algorithm (Swersky *et al.*, 2013) which is designed for BO with *continuous* auxiliary information, we can use Monte Carlo sampling to approximate (10) by utilizing information from the mixed-type CMOGP predictive belief (i.e., (8) and (9)) of the target and auxiliary functions. To make the Monte Carlo approximation tractable and efficient, we need to discretize the input domain and assume that the search space for evaluating (10) is pruned to a small set of input candidates which, following the work of Swersky *et al.* (2013), can be selected by applying EI to *only* the target function. Such a form of approximation, however, faces two critical limitations: (a) Computing (10) incurs cubic time in the size of the discretized input domain and is thus expensive to evaluate with a large input domain (or risks being approximated poorly), and (b) the pruning of the search space artificially constrains the exploration of auxiliary functions and requires a parameter in EI (i.e., to control the exploration-exploitation trade-off) to be manually tuned to fit different real-world applications.

To circumvent the above-mentioned issues, we can exploit the symmetric property of conditional mutual information and rewrite (10) as

$$\alpha(y_X, \langle x, i \rangle) = H(y_i(x) | y_X) - \mathbb{E}_{p(x_* | y_X)} [H(y_i(x) | y_X, x_*)] \quad (11)$$

which we call *mixed-type PES* (MT-PES). Intuitively, the selection of an input tuple  $\langle x, i \rangle$  to maximize (11) has to trade off between exploration of every target and auxiliary function (hence inducing a large Gaussian predictive entropy  $H(y_i(x) | y_X)$ ) vs. exploitation of the current belief  $p(x_* | y_X)$  of the global target maximizer  $x_*$  to choose a nearby input  $x$  of function  $f_i$  (i.e., convolutional structures and maximizers of the target and auxiliary functions are similar or close (Section 3)) to be evaluated (hence inducing a small expected predictive entropy  $\mathbb{E}_{p(x_* | y_X)} [H(y_i(x) | y_X, x_*)]$ ) to yield a highly informative observation that in turn improves the belief of  $x_*$ . Note that the entropy of continuous random variables (i.e., differential entropy) and discrete/binary random variables (i.e., Shannon entropy) are not comparable<sup>3</sup>. So, the differential entropy terms in (11) for  $i = 1$  are not comparable to the Shannon entropy terms in (11) for  $i = 2, \dots, M$ . Fortunately, the difference of the two entropy terms in (11) is exactly the information gain of the global target maximizer  $x_*$  in (10) which is comparable between  $i = 1$  vs.  $i = 2, \dots, M$  regardless of whether the output  $y_i(x)$  is continuous or binary. Next, we will describe how to evaluate (11) efficiently.

<sup>3</sup>For example, the Shannon entropy is always non-negative while the differential entropy can be negative. A detailed discussion of their difference and connection is available in Chapter 8 of Cover and Thomas (2006).

## 5 APPROXIMATION OF MIXED-TYPE PREDICTIVE ENTROPY SEARCH

Due to (9), the first Gaussian predictive/posterior entropy term in (11) can be computed analytically:

$$H(y_1(x)|y_X) \triangleq 0.5 \log(2\pi e(\sigma_{(x,1)|X}^2 + \sigma_{n_1}^2))$$

$$H(y_i(x)|y_X) \triangleq - \sum_{y_i(x) \in \{1, -1\}} p(y_i(x)|y_X) \log p(y_i(x)|y_X) \quad (12)$$

for  $i = 2, \dots, M$ . Unfortunately, the second term in (11) cannot be evaluated in closed form. Although this second term appears to resemble that in PES (Hernández-Lobato *et al.*, 2014), their approximation method, however, cannot be applied straightforwardly here since it cannot account for either the binary auxiliary information or the complex cross-correlation structure between the target and auxiliary functions. To achieve this, we will first propose a novel mixed-type random features approximation of the CMOGP model whose cross-correlation structure between the target and auxiliary functions can be exploited for sampling the global target maximizer  $x_*$  more accurately using the past observations  $(X, y_X)$  from evaluating these functions (especially when the target function is sparsely evaluated due to its higher cost), which is in turn used to approximate the expectation in (11). Then, we will formalize some practical constraints relating the global target maximizer to the binary auxiliary functions, which are used to approximate the second entropy term within the expectation in (11).

### 5.1 MIXED-TYPE RANDOM FEATURES

To approximate the expectation in (11) efficiently by averaging over samples of the target maximizer from  $p(x_*|y_X)$  in a continuous input domain, we will derive an analytic sample of the unknown function  $f_i$  given the past observations  $(X, y_X)$ , which is differentiable and can be optimized by any existing gradient-based optimization method to search for its maximizer. Unlike the work of Hernández-Lobato *et al.* (2014) that achieves this in PES using the *single-output random features* (SRF) method for handling a single continuous output type (Lázaro-Gredilla *et al.*, 2010; Rahimi and Recht, 2007), we have to additionally consider how the binary auxiliary functions and their complex cross-correlation structure with the target function can be exploited for sampling the target maximizer  $x_*$  more accurately. To address this, we will now present a novel *mixed-type random features* (MT-RF) approximation of the CMOGP model by first deriving an analytic form of the latent function  $L$  with SRF and then an analytic approximation of  $f_i$  using the convolutional structure of the CMOGP model. The results of EP (6) can be reused here to ap-

proximate the non-Gaussian likelihood  $p(y_i(x)|f_i(x))$  for  $i = 2, \dots, M$ .

Using SRF (Rahimi and Recht, 2007), the latent function  $L$  modeled using GP can be approximated by a linear model  $L(x) \approx \phi(x)^\top \theta$  where  $\phi(x)$  is a random vector of an  $m$ -dimensional feature mapping of the input  $x$  for  $L(x)$  and  $\theta \sim \mathcal{N}(\underline{0}, I)$  is an  $m$ -dimensional vector of weights. Then, interestingly, by exploiting the convolutional structure of the CMOGP model in (1),  $f_i(x)$  can also be approximated analytically by a linear model:  $f_i(x) \approx m_i + \phi_i(x)^\top \theta$  where the random vector  $\phi_i(x) \triangleq \sigma_{s_i} \text{diag}(\exp(-0.5W^\top P_i^{-1}W)) \phi(x)$  can be interpreted as input features of  $f_i(x)$ ,  $W$  is a  $d \times m$  random matrix which is used to map  $x \rightarrow \phi(x)$  in SRF, and function  $\text{diag}(A)$  returns a diagonal matrix with the same diagonal components as  $A$ . The exact definition of  $\phi(x)$  and the derivation of  $\phi_i(x)$  are in Appendix C.

Then, a sample of  $f_i$  can be constructed using  $f_i^{(s)}(x) \triangleq m_i + \phi_i^{(s)}(x)^\top \theta^{(s)}$  where  $\phi_i^{(s)}(x)$  and  $\theta^{(s)}$  are vectors of features and weights sampled, respectively, from the random vector  $\phi_i(x)$  and the posterior distribution of weights  $\theta$  given the past observations  $(X, y_X)$ , the latter of which is approximated to be Gaussian by exploiting the conditional independence property of MT-RF and the results of EP (6) from the mixed-type CMOGP model:

$$p(\theta|y_X) = \mathcal{N}(\theta|A^{-1}\Phi(\Lambda - \Sigma_{XX})^{-1}(\tilde{y}_X - \mu_X), A^{-1})$$

where  $A \triangleq \Phi(\Lambda - \Sigma_{XX})^{-1}\Phi^\top + I$  and  $\Phi \triangleq (\phi_j(x))_{(x,j) \in X}$ , as detailed in Appendix C.2.

Consequently, the expectation in (11) can be approximated by averaging over  $S$  samples of the target maximizer  $x_*^{(s)}$  of  $f_1^{(s)}$  to yield an approximation of MT-PES:

$$\alpha(y_X, \langle x, i \rangle) \approx H(y_i(x)|y_X) - \frac{1}{S} \sum_{s=1}^S H(y_i(x)|y_X, x_*^{(s)}) \quad (13)$$

where  $x_*^{(s)} \triangleq x_{*1}^{(s)}$  and  $x_{*i}^{(s)} \triangleq \arg \max_{x \in D} f_i^{(s)}(x)$  for  $i = 1, \dots, M$ . Drawing a sample of  $x_*^{(s)}$  incurs  $\mathcal{O}(m^3 + m^2|X|)$  time if  $m \leq |X|$  and  $\mathcal{O}(|X|^3 + |X|^2m)$  time if  $m > |X|$ , which is more efficient than using Thompson sampling to sample  $f_i$  over a discretized input domain that incurs cubic time in its size since a sufficiently fine discretization of the entire input domain is typically larger in size than the no.  $|X|$  of observations.

### 5.2 APPROXIMATING THE PREDICTIVE ENTROPY CONDITIONED ON THE TARGET MAXIMIZER

We will now discuss how the second entropy term in (13) is approximated. Firstly, the posterior distribution of  $y_i(x)$  given the past observations and target maximizer is

computed by

$$p(y_i(x)|y_X, x_*) = \int p(y_i(x)|f_i(x)) p(f_i(x)|y_X, x_*) df_i(x) \quad (14)$$

where  $p(y_i(x)|f_i(x))$  is defined in (3) and  $p(f_i(x)|y_X, x_*)$  will be approximated by EP, as detailed later. As shown in Section 3, the Gaussian predictive belief  $p(f_i(x)|y_X)$  (8) can be computed analytically. Then,  $p(f_i(x)|y_X, x_*)$  can be considered as a constrained version of  $p(f_i(x)|y_X)$  by further conditioning on the target maximizer  $x_*$ . It is intuitive that the posterior distribution of  $f_i(x)$  is constrained by  $f_i(x) \leq f_i(x_{*i}), \forall \langle x, i \rangle \in D^+$ . However, since only the target maximizer  $x_*$  is of interest, how should the value of  $f_i(x)$  be constrained by  $x_*$  instead of  $x_{*i}$  if  $i = 2, \dots, M$ ? To resolve this, we introduce a slack variable  $c_i$  to formalize the relationship between maximizers of the target and auxiliary functions:

$$f_i(x) \leq f_i(x_*) + c_i \quad \forall x \in D, i \neq 1 \quad (15)$$

where  $c_i \triangleq \mathbb{E}_{p(x_{*i}|y_X)}[f_i(x_{*i})] - \mathbb{E}_{p(x_*|y_X)}[f_i(x_*)]$  measures the gap between the expected maximum of  $f_i$  and the expected output of  $f_i$  evaluated at  $x_*$  and can be approximated efficiently using our MT-RF method even though  $f_i$  is unknown, as detailed later. Consequently, the following simplified constraints instead of (15) will be used to approximate  $p(f_i(x)|y_X, x_*)$ :

- C1.  $f_i(x) \leq f_i(x_*) + \delta_i c_i$  for a given  $\langle x, i \rangle \in D^+$  where  $\delta_i$  equals to 0 if  $i = 1$ , and 1 otherwise.
- C2.  $f_1(x_*) \geq y_{\max} + \epsilon_1$  where  $\epsilon_1 \sim \mathcal{N}(0, \sigma_{n_1}^2)$  and  $y_{\max} \triangleq \max_{(x,1) \in X_1} y_1(x)$  is the largest among the noisy outputs observed by evaluating the target function  $f_1$  at  $X_1$ .
- C3.  $\Phi_{\text{cdf}}(f_j(x_*) + c_j) \geq 0.5$  for  $j = 2, \dots, M$ .<sup>4</sup>

The first constraint C1 keeps the influence of  $x_*$  to the next input tuple  $\langle x, i \rangle$  to be selected by MT-PES. Instead of constraining all unknown functions over the entire input domain, C2 and C3 relax (15) to be valid only for the outputs observed from evaluating these functions. When the target and auxiliary functions are highly correlated (i.e., small  $c_j$ ), C3 means that a positive label can be observed with high probability by evaluating an auxiliary function at the target maximizer  $x_*$ . Using these constraints,  $p(f_i(x)|y_X, x_*) \approx p(f_i(x)|y_X, C1, C2, C3)$  which can be approximated analytically using EP. To

<sup>4</sup>Like the work of Swersky *et al.* (2013) (Section 2.2), we assume the cross-correlation between the target and auxiliary functions to be positive. An auxiliary function that is negatively correlated with the target function can be easily transformed to be positively correlated by negating all its outputs.

achieve this, we will first derive a tractable approximation of the posterior distribution  $p(f_i(x_*)|y_X, C2, C3)$  which does not depend on the next selected input  $x$ . Note that such terms can be computed once and reused in the approximation of  $p(f_i(x)|y_X, x_*)$  in (14) which depends on  $x$ , as detailed later.

**Approximating terms independent of  $x$ .** Let  $f_j^* \triangleq f_j(x_*)$  and  $f^* \triangleq (f_j^*)_{j=1, \dots, M}^\top$ . We can use the cdf of a standard Gaussian distribution and an indicator function to represent the probability of C2 and C3, respectively. Then, the posterior distribution  $p(f^*|y_X)$  can be constrained with C2 and C3 by

$$p(f^*|y_X, C2, C3) \propto p(f^*|y_X) \Phi_{\text{cdf}}\left(\frac{f_1^* - y_{\max}}{\sigma_{n_1}}\right) \prod_{j=2}^M \mathbb{I}(f_j^* + c_j \geq 0). \quad (16)$$

Interestingly, by sampling the target and auxiliary maximizers  $x_*$  and  $x_{*j}$  using our MT-RF method proposed in Section 5.1, the value of  $c_j$  in (16) can be approximated by Monte Carlo sampling<sup>5</sup>:

$$c_j \approx \mathbb{E}_{p(x_{*j}|y_X)}[f_j(x_{*j})] - \mathbb{E}_{p(x_*|y_X)}[f_j(x_*)] \approx S^{-1} \sum_{s=1}^S \left( f_j^{(s)}(x_{*j}^{(s)}) - f_j^{(s)}(x_*^{(s)}) \right).$$

With the multiplicative form of (16),  $p(f^*|y_X, C2, C3)$  can be approximated to be a multivariate Gaussian  $\mathcal{N}(f^*|\mu, \Sigma)$  using EP by approximating each non-Gaussian factor (i.e.,  $\Phi_{\text{cdf}}$  and  $\mathbb{I}$ ) in (16) to be a Gaussian, as detailed in Appendix D. Consequently, the posterior distribution  $p(f_i^*|y_X, C2, C3)$  can be approximated by a Gaussian  $\mathcal{N}(f_i^*|\mu_i, \tau_i)$  where  $\mu_i$  is the  $i$ -th component of  $\mu$  and  $\tau_i$  is the  $i$ -th diagonal component of  $\Sigma$ .

**Approximating terms that depend on  $x$ .** In C2 and C3,  $f_i^*$  is the only term that is related to C1. It follows that  $f_i(x)$  is conditionally independent of C2 and C3 given  $f_i^*$ . Let  $f^+ \triangleq [f_i(x_*); f_i(x)]$ . So,  $p(f^+|y_X, C2, C3) = p(f_i(x)|y_X, f_i^*) p(f_i^*|y_X, C2, C3) = \mathcal{N}(f^+|\mu^+, \Sigma^+)$  where  $\mu^+$  and  $\Sigma^+$  can be computed analytically using  $\mu_i, \tau_i$ , and (8), as detailed in Appendix E.

To involve C1, an indicator function  $\mathbb{I}(f_i(x) \leq f_i(x_*) + \delta_i c_i)$  is used to represent the probability that C1 holds. Then,  $p(f_i(x)|y_X, x_*) \approx \int p(f^+|y_X, C1, C2, C3) df_i^*$  where

$$p(f^+|y_X, C1, C2, C3) \approx Z'^{-1} p(f^+|y_X, C2) \mathbb{I}(f_i(x) \leq f_i(x_*) + \delta_i c_i). \quad (17)$$

Since the posterior of  $f_i(x_*)$  has been updated according to C2 and C3 (16),  $c_i$  in (17) is updated likewise:

$$c_i \approx S^{-1} \sum_{s=1}^S \left( f_i^{(s)}(x_{*i}^{(s)}) - \mu_i^{(s)} \right)$$

<sup>5</sup>When  $j = 1$ ,  $c_j$  is equal to 0 since  $x_{*j} = x_*$ .

where  $\mu_i^{(s)}$  is computed in (16) using a sampled  $x_*^{(s)}$ . Similar to that in (Hernández-Lobato *et al.*, 2014), a one-step EP can be used to approximate (17) as a multivariate Gaussian with the following posterior mean vector and covariance matrix:

$$\begin{aligned}\mu_{f^+} &\triangleq \mu^+ - (\gamma/\sqrt{v})\Sigma^+a \\ \Sigma_{f^+} &\triangleq \Sigma^+ - v^{-1}\gamma(\gamma - (\eta - \delta_i c_i)/\sqrt{v})\Sigma^+aa^\top\Sigma^+\end{aligned}\quad (18)$$

where  $\gamma \triangleq \phi((\delta_i c_i - \eta)/\sqrt{v})/\Phi_{\text{cdf}}((\delta_i c_i - \eta)/\sqrt{v})$ ,  $\eta \triangleq a^\top\mu^+$ ,  $v \triangleq a^\top\Sigma^+a$  and  $a = [-1; 1]$ . The derivation of (18) is in Appendix F. So, the posterior mean and variance of  $p(f_i(x)|y_X, x_*)$  can be approximated, respectively, using the 2-th component of  $\mu_{f^+}$  and (2, 2)-th component of  $\Sigma_{f^+}$  denoted by  $\mu_{f_i}$  and  $v_{f_i}$ . As a result, the posterior entropy  $H(y_i(x)|y_X, x_*^{(s)})$  in (13) can be approximated using (12) by replacing  $\mu_{\{(x,i)\}|X}$  and  $\sigma_{\{(x,i)\}|X}^2$  in (12) with, respectively,  $\mu_{f_i}^{(s)}$  and  $v_{f_i}^{(s)}$  where  $\mu_{f_i}^{(s)}$  and  $v_{f_i}^{(s)}$  are computed in (18) using a sampled  $x_*^{(s)}$ .

## 6 EXPERIMENTS AND DISCUSSION

This section empirically evaluates the performance of our MT-PES algorithm against that of (a) the state-of-the-art PES (Hernández-Lobato *et al.*, 2014) without utilizing the binary auxiliary information and (b) MT-ES performing Monte Carlo approximation of (10). In all experiments, we use  $m \triangleq 200$  random features and  $S \triangleq 50$  samples of the target maximizer in MT-PES. The input candidates with top 30 EI values are selected for evaluating MT-ES. The *mixed-type MOGP* (MT-MOGP) hyperparameters are learned via maximum likelihood estimation. The performance of the tested algorithms are evaluated using *immediate regret* (IR)  $|f_1(x_*) - f_1(\tilde{x}_*)|$  where  $\tilde{x}_* \triangleq \arg \max_{x \in D} \mu_{\{(x,1)\}|X}$  is their recommended target maximizer. In each experiment, one observation of the target function is randomly selected as the initialization.

### 6.1 SYNTHETIC EXPERIMENTS

The performance of the tested algorithms are firstly evaluated using synthetic and benchmark functions.

**Synthetic functions.** The synthetic functions are generated using  $M \triangleq 2$  and  $D \triangleq [0, 1]^2$ . To do this, the CMOGP hyperparameters with one latent function are firstly fixed as the values shown in Appendix H.1, which are also used in the tested algorithms as optimal hyperparameters. Then, a set  $X$  of 450 input tuples are uniformly sampled from  $D^+$  and their corresponding outputs are sampled from the CMOGP prior. The target function is set to be the predictive mean  $\mu_{\{(x,1)\}|X}$  of the CMOGP model. The outputs of the auxiliary function are set to be 1 if  $\mu_{\{(x,2)\}|X} \geq 0$ , and  $-1$  otherwise. An example of the synthetic functions can be found in Figs. 1a to 1c. As

can be seen in Figs. 1b and 1c, we can generate multiple auxiliary functions with different proportions of positive outputs from a target function (Fig. 1a) by varying the bias  $m_2$ . All these auxiliary functions correlate well with the target function but delineate the input regions containing the target maximizer differently and thus result in different MT-PES performance, as will be shown later.

**Empirical analysis of MT-MOGP and MT-RF.** Firstly, we verify that the MT-MOGP model and MT-RF can outperform the conventional GP model and single-output RF by exploiting cross-correlation structure between the target and auxiliary function aux1 (i.e., Figs. 1a and 1b). Figs. 1d and 1e show the predictive mean and the sampled maximizers of the target function using randomly sampled observations. By comparing Figs. 1d and 1e with Fig. 1a, it can be observed that the MT-MOGP model and MT-RF can predict the target function and sample the target maximizer more accurately than the conventional GP model and single-output RF using an additional 50 observations from evaluating aux1.

**Empirical analysis of mixed-type BO.** Next, the performance of the tested BO algorithms are evaluated using ten groups (i.e., one target function, two auxiliary functions aux1 and aux2 with different  $m_2$ ) of synthetic functions generated using the above procedure. We adjust  $m_2$  such that around 20% of auxiliary outputs are positive for each aux1 and set  $m_2 = 0$  for each aux2. An averaged IR is obtained by optimizing the target function in each of them with 10 different initializations for each tested algorithm.

Fig. 2 shows the results of all tested algorithms for synthetic functions with a cost budget of 2500. From Fig. 2a, MT-PES can achieve a similar averaged IR with a much lower cost than PES, which implies that the BO performance can be accelerated by exploiting the binary auxiliary information of lower evaluation cost. MT-ES achieves lower averaged IR than PES with a cost less than 1000 but unfortunately performs less well in the remaining BO iterations. Even though the cheap auxiliary outputs provide additional information for finding the target maximizer at the beginning of BO, the multimodal nature of the synthetic function (see Fig. 1a) causes MT-ES to be trapped easily in some local maximum since its search space has been pruned using EI for time efficiency.

To investigate how the performance of MT-PES will be affected by the proportion of positive outputs in different auxiliary functions, we vary the number and bias  $m_2$  of the auxiliary function(s) and show the results in Fig. 2b. It can be observed that MT-PES using aux2 as the auxiliary function does not converge as fast as MT-PES using aux1, which is expected since aux2 with a larger proportion of positive outputs is less informative in delineating the input regions containing the target maximizer than aux1. Also,

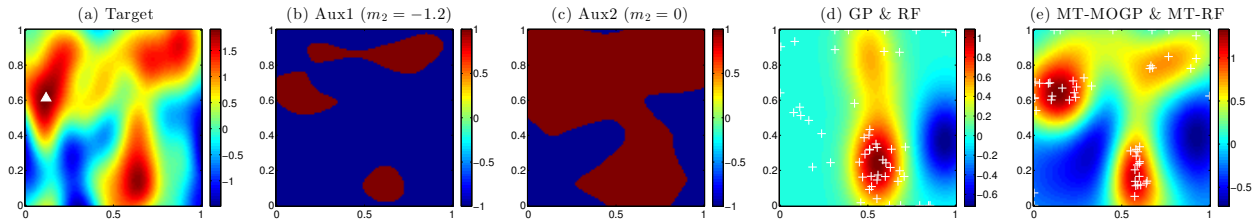


Figure 1: (a-c) Example of the synthetic functions where ‘ $\triangle$ ’ is the global target maximizer, (d) target function predicted by conventional GP model and the target maximizers (‘+’) sampled by RF with 5 observations from evaluating the target function, and (e) target function predicted by MT-MOGP model and the target maximizers (‘+’) sampled by MT-RF with 5 and 50 observations from evaluating the target and aux1 functions, respectively.

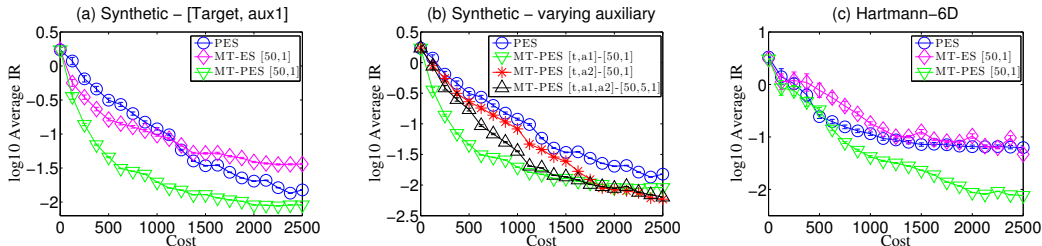


Figure 2: Graphs of  $\log_{10}$ (averaged IR) vs. cost incurred by tested algorithms for (a-b) synthetic functions and (c) Hartmann-6D function. The type and cost of functions used in each experiment are shown in the title and legend of each graph where ‘t’ denotes target function and ‘a1’ and ‘a2’ denote aux1 and aux2 functions, respectively. The error bars are computed in the form of standard error.

Fig. 2b shows that MT-PES is able to exploit multiple auxiliary functions with different costs to achieve a lower averaged IR than PES with a much lower cost.

*Remark.* From the results in Fig. 2b, one may expect MT-PES to converge faster using an auxiliary function with a smaller proportion of positive outputs, which is not always the case. If the auxiliary function has sparse positive outputs, MT-PES will face difficulty finding a positive output when exploring the auxiliary function and start to evaluate the target function after only several negative outputs are observed from evaluating the cheap auxiliary function. These negative outputs may not be informative enough to guide the algorithm to directly evaluate the target function near to the likely target maximizer. To reduce the negative effect of such an unexpected behavior in real-world applications with an unknown auxiliary function, we can set MT-PES to evaluate only the auxiliary function using a small amount (e.g., 10%) of the budget at the beginning of BO so that positive auxiliary outputs are highly likely to be observed before MT-PES chooses to evaluate the expensive target function.

To provide more insight into the approximations of MT-PES, we follow the PES paper (Hernández-Lobato *et al.*, 2014) and show the accuracy of the EP approximations (Section 5.2) compared to that of the ground truth constructed using the rejection sampling method. To verify how sensitive the performance of MT-PES is to different settings, we have also evaluated the performance of the

tested algorithms using synthetic functions with varying costs  $\lambda_i$ , random features dimension  $m$ , and sampling size  $S$ . The results are reported in Appendix H.1.

**Hartmann-6D function.** The original Hartmann-6D function is used as the target function and to construct the binary auxiliary function, as detailed in Appendix H.2. Fig. 2c shows results of the tested algorithms with 10 different initializations. It can be observed that MT-PES converges faster to a lower averaged IR than PES. However, MT-ES does not perform well for Hartmann-6D function which is difficult to optimize due to their multimodal nature (i.e., 1 global maximum and 6 local maxima) and large input domain. The former causes MT-ES to be trapped easily in some local maximum while the latter prohibits MT-ES from finely discretizing the input domain to remain computationally tractable.

## 6.2 REAL-WORLD EXPERIMENTS

The tested algorithms are next used in hyperparameter tuning of a ML model in an image classification task and policy search for reinforcement learning.

**Convolutional neural network (CNN) with CIFAR-10 dataset.** The six CNN<sup>6</sup> hyperparameters to be tuned in our experiments are the learning rate of SGD in the range of  $[10^{-5}, 1]$ , three dropout rates in the range of

<sup>6</sup>We use the example code of keras (i.e., cifar10\_cnn.py) and switch the optimizer in their code to SGD.



$[0, 1]$ , batch size in the range of  $[100, 1000]$ , and number of learning epochs in the range of  $[100, 1000]$ . We use training and validation data of size 50000 and 10000, respectively. The unknown target function to be maximized is the validation accuracy evaluated by training the CNN with all the training data. The auxiliary function is the decision made using the *Bayesian optimal stopping* (BOS) mechanism in (Dai *et al.*, 2019; Müller *et al.*, 2007) by setting 0.5 as a threshold of the validation accuracy. In particular, we train the same CNN model with a smaller fixed dataset of size 10000 randomly selected from the original training data and apply the BOS after 20 training epochs. The BOS will early-stop the training and return 1 if it predicts that a final validation accuracy of 0.5 can be achieved with a high probability, and  $-1$  otherwise.<sup>7</sup> The real training time is not known and varies with different settings of hyperparameters. To simplify the setting of the evaluation costs, we use  $\lambda_1(x) = 1$  and  $\lambda_2(x) = 0.2 \times (20/x_{\text{epochs}})$  where  $x_{\text{epochs}}$  is the number of learning epochs in each selected hyperparameter setting.<sup>8</sup> For this experiment, we additionally compare the tested algorithms with *multi-fidelity GP-UCB* (MF-GP-UCB) (Kandasamy *et al.*, 2016) that can only exploit *continuous* auxiliary functions. The auxiliary function of MF-GP-UCB is the validation accuracy evaluated by training the same CNN with the same data used for the auxiliary function of MT-PES.<sup>9</sup> The actual wall-clock time shown in the results includes the time of both CNN training and BO. The validation accuracy  $f_1(\tilde{x}_*)$  is evaluated by training the CNN with  $\tilde{x}_*$  for the tested algorithms.

**Policy search for reinforcement learning (RL).** We apply the tested algorithms to the CartPole task from OpenAI Gym and use a linear policy consisting of 8 parameters in the range of  $[0, 1]$ . This task is defined to be a success (i.e., reward of 1) if the episode length reaches 200, and a failure (reward of  $-1$ ) otherwise. The target function to be maximized is the success rate averaged over 100 episodes with random starting states. The auxiliary function is the reward of one episode with a fixed starting state  $(0, 0, 0.02, 0.02)$ .  $\lambda_1(x) = 100$  and  $\lambda_2(x) = 1$  are used in the experiments. The success rate  $f_1(\tilde{x}_*)$  is evaluated by running the CartPole task with  $\tilde{x}_*$  as the policy parameters over 100 episodes for the tested algorithms.

Fig. 3 shows results of the tested algorithms with 5 dif-

<sup>7</sup>A description of BOS is provided in Appendix H.3.

<sup>8</sup>We use 20% of the training data for evaluating the auxiliary function and early-stop the training after around 20 epochs.

<sup>9</sup>One may consider constructing the auxiliary function of MF-GP-UCB with an even smaller training dataset such that its cost is similar to that of the binary auxiliary function. However, for any smaller training dataset, we can always early-stop the training and achieve a much cheaper binary auxiliary function, as compared to the continuous auxiliary function of MF-GP-UCB constructed using the same dataset.

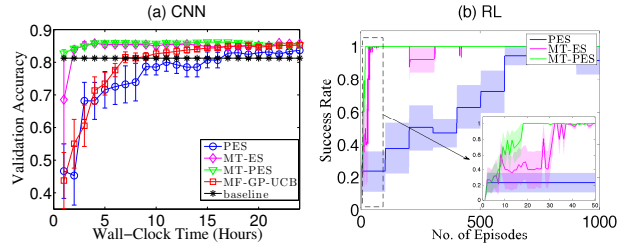


Figure 3: Graphs of (a) validation accuracy vs. wall-clock time incurred by tested algorithms for CNN and (b) success rate vs. no. of episodes incurred by tested algorithms for RL. The results for the first 50 episodes are zoomed in for a clearer comparison.

ferent initializations for the CNN hyperparameter tuning and RL policy search tasks. It can be observed that both MT-ES and MT-PES converge faster to a smaller IR than other tested algorithms. MT-PES also converges faster than MT-ES in both experiments. MT-ES and MT-PES outperform MF-GP-UCB since evaluating the *binary* auxiliary function by early-stopping the CNN training incurs much less time than evaluating the true validation accuracy for MF-GP-UCB. Using only 1 hour, MT-PES can improve the performance of CNN over that of the baseline achieved using the default hyperparameters in the existing code, which shows that MT-PES is promising in quickly finding more competitive hyperparameters of complex ML models.

## 7 CONCLUSION

This paper describes novel MT-ES and MT-PES algorithms for mixed-type BO that can exploit cheap binary auxiliary information for accelerating the optimization of a target objective function. A novel mixed-type CMOGP model and its MT-RF approximation are proposed for improving the belief of the unknown target function and the global target maximizer using observations from evaluating the target and binary auxiliary functions. New practical constraints are proposed to relate the global target maximizer to the binary auxiliary functions such that MT-PES can be approximated efficiently. Empirical evaluation on synthetic functions and real-world applications shows that MT-PES outperforms the state-of-the-art BO algorithms. For future work, our proposed mixed-type BO algorithms can be easily extended to handle both binary and continuous auxiliary information, hence generalizing multi-fidelity PES (Zhang *et al.*, 2017).<sup>10</sup>

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

<sup>10</sup>A closely related counterpart is multi-fidelity active learning (Zhang *et al.*, 2016).

## References

- Álvarez, M. A. and Lawrence, N. D. (2011). Computationally efficient convolved multiple output Gaussian processes. *JMLR*, **12**, 1459–1500.
- Bonilla, E. V., Chai, K. M. A., and Williams, C. K. I. (2007). Multi-task Gaussian process prediction. In *Proc. NIPS*, pages 153–160.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons, Inc., second edition.
- Dai, Z., Yu, H., Low, K. H., and Jaillet, P. (2019). Bayesian optimization meets Bayesian optimal stopping. In *Proc. ICML*, pages 1496–1506.
- Falkner, S., Klein, A., and Hutter, F. (2018). BOHB: Robust and efficient hyperparameter optimization at scale. In *Proc. ICML*, pages 1436–1445.
- González, J., Dai, Z., Damianou, A., and Lawrence, N. D. (2017). Preferential Bayesian optimization. In *Proc. ICML*, pages 1282–1291.
- 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.
- Hernández-Lobato, J. M., Gelbart, M. A., Adams, R. P., Hoffman, M. W., and Ghahramani, Z. (2016). A general framework for constrained Bayesian optimization using information-based search. *JMLR*, **17**(1), 5549–5601.
- Huang, D., Allen, T. T., Notz, W. I., and Miller, R. A. (2006). Sequential kriging optimization using multiple-fidelity evaluations. *Struct. Multidisc. Optim.*, **32**(5), 369–382.
- Kandasamy, K., Dasarthy, G., Oliva, J. B., Schneider, J., and Póczos, B. (2016). Gaussian process bandit optimisation with multi-fidelity evaluations. In *Proc. NIPS*, pages 992–1000.
- Kandasamy, K., Dasarthy, G., Schneider, J., and Póczos, B. (2017). Multi-fidelity Bayesian optimisation with continuous approximations. In *Proc. ICML*, pages 1799–1808.
- Lázaro-Gredilla, M., Quiñero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. (2010). Sparse spectrum Gaussian process regression. *JMLR*, **11**, 1865–1881.
- Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. *JMLR*, **18**, 1–52.
- Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference*. Ph.D. thesis, Massachusetts Institute of Technology.
- Mockus, J., Tiešis, V., and Žilinskas, A. (1978). The application of Bayesian methods for seeking the extremum. In L. C. W. Dixon and G. P. Szegő, editors, *Towards Global Optimization 2*, pages 117–129. North-Holland Publishing Company.
- Müller, P., Berry, D. A., Grieve, A. P., Smith, M., and Krams, M. (2007). Simulation-based sequential Bayesian design. *J. Statistical Planning and Inference*, **137**(10), 3140–3150.
- Poloczek, M., Wang, J., and Frazier, P. I. (2017). Multi-information source optimization. In *Proc. NIPS*, pages 4288–4298.
- Pourmohamad, T. and Lee, H. K. H. (2016). Multivariate stochastic process models for correlated responses of mixed type. *Bayesian Anal.*, **11**(3), 797–820.
- Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Proc. NIPS*, pages 1177–1184.
- Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*. MIT Press.
- Russo, D. J., van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, **11**(1), 1–96.
- Schön, T. B. and Lindsten, F. (2011). Manipulating the multivariate Gaussian density. Technical report, Division of Automatic Control, Linköping University, Sweden.
- Sen, R., Kandasamy, K., and Shakkottai, S. (2018). Multi-fidelity black-box optimization with hierarchical partitions. In *Proc. ICML*, pages 4538–4547.
- 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.
- Skolidis, G. (2012). *Transfer Learning with Gaussian Processes*. Ph.D. thesis, University of Edinburgh.
- Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Proc. NIPS*, pages 2951–2959.
- 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.
- Swersky, K., Snoek, J., and Adams, R. P. (2013). Multi-task Bayesian optimization. In *Proc. NIPS*, pages 2004–2012.
- Teh, Y. W. and Seeger, M. (2005). Semiparametric latent factor models. In *Proc. AISTATS*, pages 333–340.
- Tesch, M., Schneider, J., and Choset, H. (2013). Expensive function optimization with stochastic binary outcomes. In *Proc. ICML*, pages 1283–1291.
- Villemonteix, J., Vazquez, E., and Walter, E. (2009). An informational approach to the global optimization of expensive-to-evaluate functions. *J. Glob. Optim.*, **44**(4), 509–534.
- Wackernagel, H. (1998). *Multivariate Geostatistics: An Introduction with Applications*. Springer, second edition.
- Webster, R. and Oliver, M. (2007). *Geostatistics for Environmental Scientists*. John Wiley & Sons, Inc., second edition.
- 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*.