

---

# Rectified Max-Value Entropy Search for Bayesian Optimization

---

Quoc Phong Nguyen<sup>1</sup>, Bryan Kian Hsiang Low<sup>1</sup>, Patrick Jaillet<sup>2</sup>

<sup>1</sup>Dept. of Computer Science, National University of Singapore, Republic of Singapore  
{qphong, lowkh}@comp.nus.edu.sg

<sup>2</sup>Dept. of Electrical Engineering and Computer Science, MIT, USA  
jaillet@mit.edu

## Abstract

Although the existing *max-value entropy search* (MES) is based on the widely celebrated notion of mutual information, its empirical performance can suffer due to two misconceptions whose implications on the exploration-exploitation trade-off are investigated in this paper. These issues are essential in the development of future acquisition functions and the improvement of the existing ones as they encourage an accurate measure of the mutual information such as the *rectified MES* (RMES) acquisition function we develop in this work. Unlike the evaluation of MES, we derive a closed-form probability density for the observation conditioned on the max-value and employ stochastic gradient ascent with reparameterization to efficiently optimize RMES. As a result of a more principled acquisition function, RMES shows a consistent improvement over MES in several synthetic function benchmarks and real-world optimization problems.

## 1 Introduction

*Bayesian optimization* (BO) has demonstrated to be highly effective in optimizing an unknown complex objective function (i.e., possibly noisy, non-convex, without a closed-form expression nor derivative) with a finite budget of expensive function evaluations [Brochu et al., 2010, Shahriari et al., 2015, Snoek et al., 2012]. A BO algorithm depends on a choice of acquisition function (e.g., improvement-based such as probability of improvement [Kushner, 1964] and *expected improvement* (EI) [Mockus et al., 1978], information-based such as those described below, or *upper confidence bound* (UCB) [Srinivas et al., 2010]) as a heuristic to guide its search for the global maxi-

mizer. To do this, the BO algorithm utilizes the chosen acquisition function to iteratively select an input query for evaluating the unknown objective function that trades off between observing at or near to a likely maximizer based on a *Gaussian process* (GP) belief of the objective function (exploitation) vs. improving the GP belief (exploration) until the budget is expended.

In this paper, we consider information-based acquisition functions based on the widely celebrated notion of mutual information, which include *entropy search* (ES) [Hennig and Schuler, 2012], *predictive ES* (PES) [Hernández-Lobato et al., 2014], *output-space PES* (OPES) [Hoffman and Ghahramani, 2015], *max-value ES* (MES) [Wang and Jegelka, 2017], and fast information-theoretic BO [Ru et al., 2018]. In general, they either maximize the information gain on the *max-value* (i.e., global maximum of objective function) or its corresponding global *maximizer*. Though ES and PES perform the latter and hence directly achieve the goal of BO, they require a series of approximations. On the other hand, MES and OPES perform the former and enjoy the advantage of requiring sampling of only a 1-dimensional random variable representing the max-value (instead of a multi-dimensional random vector representing the maximizer). In particular, MES can be expressed in closed form and thus optimized easily. Unfortunately, its BO performance is compromised by a number of misconceptions surrounding its design, as discussed below. Since there may be subsequent works building on MES (e.g., [Knudde et al., 2018, Takeno et al., 2019]), it is imperative that we rectify these pressing misconceptions.

So, our first contribution of this paper is to review the principle of information gain underlying MES, which will shed light on two (perhaps) surprising misconceptions (Section 3) and their negative implications on its interpretation as a mutual information measure and its BO performance. We give an intuitive illustration in Section 4 using simple synthetic experiments how they can cause its search to behave suboptimally in

trading off between exploitation vs. exploration.

Our second contribution is the development of a *rectified max-value entropy search* (RMES) acquisition function that resolves the misconceptions present in MES and hence provides a more precise measure of mutual information. In contrast to the straightforward implementation of MES, the evaluation of RMES may seem challenging at first glance: Unlike the noiseless observations assumed by MES that follow the well-known truncated Gaussian distribution when conditioned on the max-value, the true noisy observations obtained from evaluating the objective function do not. However, by deriving a closed-form expression for the conditional probability density of the noisy observation, we make the evaluation of RMES possible. Furthermore, the optimization is made efficient by using stochastic gradient ascent with a reparameterization trick [Kingma and Welling, 2013]. As a result of a more principled acquisition function, RMES gives a rewarding performance that is improved over the existing MES in several synthetic function benchmarks and real-world optimization problems.

## 2 Background

### 2.1 Gaussian Processes in BO

Consider the problem of sequentially optimizing an unknown objective function  $f : \mathcal{X} \rightarrow \mathbb{R}$  over a bounded input domain  $\mathcal{X} \subset \mathbb{R}^d$ . BO algorithms repeatedly select an input query  $\mathbf{x} \in \mathcal{X}$  for evaluating  $f$  to obtain a noisy observed output  $y_{\mathbf{x}} \triangleq f_{\mathbf{x}} + \epsilon$  with  $f_{\mathbf{x}} \triangleq f(\mathbf{x})$ , i.i.d. Gaussian noise  $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ , and noise variance  $\sigma_n^2$ . Since it is costly to evaluate  $f$ , our goal is to strategically select input queries for finding the maximizer  $\mathbf{x}_* \triangleq \operatorname{argmax}_{\mathbf{x}' \in \mathcal{X}} f_{\mathbf{x}'}$  as rapidly as possible. To achieve this, the belief of  $f$  is modeled by a GP. Let  $\{f_{\mathbf{x}'}\}_{\mathbf{x}' \in \mathcal{X}}$  denote a GP, that is, every finite subset of  $\{f_{\mathbf{x}'}\}_{\mathbf{x}' \in \mathcal{X}}$  follows a multivariate Gaussian distribution [Rasmussen and Williams, 2006]. Then, the GP is fully specified by its *prior* mean  $\mathbb{E}[f_{\mathbf{x}'}]$  and covariance  $k_{\mathbf{x}'\mathbf{x}''} \triangleq \operatorname{cov}[f_{\mathbf{x}'}, f_{\mathbf{x}''}]$  for all  $\mathbf{x}', \mathbf{x}'' \in \mathcal{X}$ , the latter of which can be defined, for example, by the widely-used *squared exponential* (SE) kernel  $k_{\mathbf{x}'\mathbf{x}''} \triangleq \sigma_s^2 \exp(-0.5(\mathbf{x}' - \mathbf{x}'')^\top \Lambda^{-2}(\mathbf{x}' - \mathbf{x}''))$  where  $\Lambda \triangleq \operatorname{diag}[\ell_1, \dots, \ell_d]$  and  $\sigma_s^2$  are its length-scale and signal variance hyperparameters, respectively. For notational simplicity (and w.l.o.g.), the prior mean is assumed to be zero. Given a column vector  $\mathbf{y}_{\mathcal{D}} \triangleq (y_{\mathbf{x}'})_{\mathbf{x}' \in \mathcal{D}}^\top$  of noisy outputs observed from evaluating  $f$  at a set  $\mathcal{D}$  of input queries selected in previous BO iterations, the GP predictive belief of  $f$  at any input query  $\mathbf{x}$  is a Gaussian  $f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}} \sim \mathcal{N}(\mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2)$  with the

following *posterior* mean  $\mu_{\mathbf{x}}$  and variance  $\sigma_{\mathbf{x}}^2$ :

$$\begin{aligned} \mu_{\mathbf{x}} &\triangleq K_{\mathbf{x}\mathcal{D}}(K_{\mathcal{D}\mathcal{D}} + \sigma_n^2 I)^{-1} \mathbf{y}_{\mathcal{D}} \\ \sigma_{\mathbf{x}}^2 &\triangleq k_{\mathbf{x}\mathbf{x}} - K_{\mathbf{x}\mathcal{D}}(K_{\mathcal{D}\mathcal{D}} + \sigma_n^2 I)^{-1} K_{\mathcal{D}\mathbf{x}} \end{aligned} \quad (1)$$

where  $K_{\mathbf{x}\mathcal{D}} \triangleq (k_{\mathbf{x}\mathbf{x}'}')_{\mathbf{x}' \in \mathcal{D}}$ ,  $K_{\mathcal{D}\mathcal{D}} \triangleq (k_{\mathbf{x}'\mathbf{x}''}')_{\mathbf{x}', \mathbf{x}'' \in \mathcal{D}}$ , and  $K_{\mathcal{D}\mathbf{x}} \triangleq K_{\mathbf{x}\mathcal{D}}^\top$ . Then,  $y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}} \sim \mathcal{N}(\mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2 + \sigma_n^2)$ .

### 2.2 Max-value information gain

Mutual information between 2 random variables has been used to quantify the amount of information gain about one random variable through observing the other. In the context of BO where the observations are the noisy function output  $y_{\mathbf{x}}$  at the input query  $\mathbf{x}$ , the information gain is about the maximizer (e.g., *entropy search* (ES) [Hennig and Schuler, 2012] and *predictive entropy search* (PES) [Hernández-Lobato et al., 2014]), i.e.,  $\mathbf{x}_*$ , or the max-value (e.g., *max-value entropy search* (MES) [Wang and Jegelka, 2017]), denoted as  $f_* \triangleq f_{\mathbf{x}_*}$ . In this paper, we consider the information gain on the max-value. The information gain on the max-value can be interpreted as the reduction in the uncertainty of the max-value  $f_*$  by observing  $y_{\mathbf{x}}$  where the uncertainty is measured by the entropy, i.e., the mutual information between  $f_*$  and  $y_{\mathbf{x}}$ :

$$I(f_*; y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}) = H(p(f_*|\mathbf{y}_{\mathcal{D}})) - \mathbb{E}_{p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})}[H(p(f_*|\mathbf{y}_{\mathcal{D}}, y_{\mathbf{x}}))] \quad (2)$$

where, given a random variable  $z$  following a distribution specified by the probability density  $p(z)$ ,  $H(p(z))$  denotes the entropy of  $z$  and  $\mathbb{E}_{p(z)} f'(z)$  denotes the expectation of a function  $f'$  over  $z$ . As (2) requires  $p(f_*|\mathbf{y}_{\mathcal{D}}, y_{\mathbf{x}})$  which is computationally expensive to evaluate for different values of  $y_{\mathbf{x}}$ , the symmetry property of the mutual information is often exploited to express the acquisition function as:

$$I(f_*; y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}) = H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})) - \mathbb{E}_{p(f_*|\mathbf{y}_{\mathcal{D}})}[H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*))] \quad (3)$$

To evaluate the acquisition function in (3), it requires the evaluation of  $p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*)$ . This probability is difficult to evaluate as it requires imposing the condition that  $f_{\mathbf{x}'} \leq f_* \forall \mathbf{x}' \in \mathcal{X}$ , so MES only imposes the condition of the max-value at the input query:<sup>1</sup>

$$p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*) \approx \int p(y_{\mathbf{x}}|f_{\mathbf{x}})p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})\mathbb{I}_{f_{\mathbf{x}} \leq f_*} df_{\mathbf{x}} \quad (4)$$

where  $\mathbb{I}_{f_{\mathbf{x}} \leq f_*}$  is the indicator function such that it is 1 if  $f_{\mathbf{x}} \leq f_*$  and 0 otherwise. We adopt this assumption throughout the paper.

Furthermore, to make (4) a truncated Gaussian density whose entropy has a closed-form expression, MES

<sup>1</sup>It is noted that OPES uses a more relaxed assumption than MES, but it requires a difficult approximation.

replaces  $y_{\mathbf{x}}$  with  $f_{\mathbf{x}}$  on the right hand side of (4), which results in  $\alpha_{\text{MES}}(\mathbf{x}, \mathcal{Y}_{\mathcal{D}}) \triangleq I(f_*; f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}}) =$

$$H(p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}})) - \mathbb{E}_{p(f_* | \mathcal{Y}_{\mathcal{D}})} [H(p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}}, f_*))]. \quad (5)$$

This formulation of MES has been applied in quite a few works, e.g., [Knudde et al., 2018, Takeno et al., 2019]. In (5),  $f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}}, f_*$  follows a truncated Gaussian distribution whose entropy has a closed-form expression. Hence, the MES expression can be reduced to [Wang and Jegelka, 2017]:

$$\frac{1}{|\mathcal{F}|} \sum_{f_* \in \mathcal{F}} \left[ \frac{h_{f_*}(\mathbf{x}) \psi(h_{f_*}(\mathbf{x}))}{2\Psi(h_{f_*}(\mathbf{x}))} - \log \Psi(h_{f_*}(\mathbf{x})) \right] \quad (6)$$

where  $\mathcal{F}$  is a finite set of samples of  $f_*$  drawn from  $p(f_* | \mathcal{Y}_{\mathcal{D}})$ ,  $h_{f_*}(\mathbf{x}) \triangleq \frac{f_* - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}}$ ,  $\psi(h_{f_*}(\mathbf{x})) \triangleq \mathcal{N}(h_{f_*}(\mathbf{x}); 0, 1)$  denotes the probability density function at  $h_{f_*}(\mathbf{x})$  of the standard Gaussian distribution, and  $\Psi(h_{f_*}(\mathbf{x}))$  denotes the cumulative density function value at  $h_{f_*}(\mathbf{x})$  of the standard Gaussian distribution. This set  $\mathcal{F}$  is obtained by either optimizing sampled functions from the GP posterior [Hernández-Lobato et al., 2014, Wang and Jegelka, 2017] or approximating with a Gumbel distribution [Wang and Jegelka, 2017].

### 3 Misconceptions in MES

This section investigates two main issues with the existing MES [Wang and Jegelka, 2017], which paves the way to an improved variant of MES in the next section.

#### 3.1 Noiseless Observations

Since BO observations are often the noisy function output  $y_{\mathbf{x}}$  due to uncontrolled factors such as random noise in environmental sensing, stochastic optimization, and random mini-batches of data in machine learning, replacing  $y_{\mathbf{x}}$  with  $f_{\mathbf{x}}$  in (5) fundamentally changes the principle behind the information gain on the max-value. In other words, MES measures the amount of information gain about  $f_*$  through observing  $f_{\mathbf{x}}$  which is often not observed in practice. In fact, the noisy observation  $y_{\mathbf{x}}$  contains less information about the latent function than its noiseless counterpart  $f_{\mathbf{x}}$ . Thus, replacing  $y_{\mathbf{x}}$  with  $f_{\mathbf{x}}$  potentially overestimates the amount of information gain as shown in Fig. 4 in Section 4.

Fig. 1 illustrates the difference between the distribution of the noisy function output  $y_{\mathbf{x}}$  (i.e., the observation of BO) and that of the noiseless function output  $f_{\mathbf{x}}$ . It is noted that conditioned on the max-value  $f_*$ , the distribution of the noiseless function output  $f_{\mathbf{x}} | f_*$

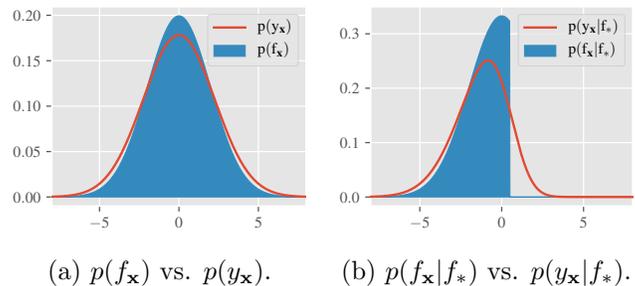


Figure 1: A comparison between the distribution of the noiseless function output  $f_{\mathbf{x}}$  and that of the noisy function output  $y_{\mathbf{x}}$  where  $f_{\mathbf{x}} \sim \mathcal{N}(0, 4)$ ,  $y_{\mathbf{x}} = f_{\mathbf{x}} + \epsilon$ ,  $\epsilon \sim \mathcal{N}(0, 1)$ , and  $f_* = 0.5$ .

is a truncated Gaussian distribution having a closed-form expression for its entropy. On the contrary, it is challenging to evaluate the probability density of the noisy function output conditioned on the max-value, i.e.,  $y_{\mathbf{x}} | f_*$  in Fig. 1b, not to mention its entropy. We will resolve this issue in Section 4.

#### 3.2 Discrepancy in Evaluation

The mutual information can be interpreted as the mutual dependence between two random variables via the *Kullback-Leibler* (KL) divergence. In (5), these random variables are  $f_{\mathbf{x}}$  and  $f_*$ , and the mutual information can be expressed as  $I(f_*; f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}}) = \text{D}_{\text{KL}} [p(f_*, f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}}) \parallel p(f_* | \mathcal{Y}_{\mathcal{D}}) p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}})]$  which denotes the KL divergence between  $p(f_*, f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}})$  and  $p(f_* | \mathcal{Y}_{\mathcal{D}}) p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}})$ . If the KL divergence between  $p(f_*, f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}})$  and the fully factorized distribution  $p(f_* | \mathcal{Y}_{\mathcal{D}}) p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}})$  is large,  $f_*$  and  $f_{\mathbf{x}}$  are dependent on each other. Hence, the distributions of  $f_*$  and  $f_{\mathbf{x}}$  should be defined consistently throughout the evaluation of the mutual information. On the contrary, if there exist discrepancies in the evaluation of the mutual information such that  $p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}})$  or  $p(f_* | \mathcal{Y}_{\mathcal{D}})$  is not uniquely evaluated, the mutual information is no longer properly defined. Therefore, for MES to have an interpretation as a measure of the mutual information,  $p(f_* | \mathcal{Y}_{\mathcal{D}})$  and  $p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}})$  should be consistently defined throughout the MES evaluation. For example, in (5),  $p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}})$  should have the same probability density under assumptions in  $H(p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}}))$  and those in  $\mathbb{E}_{p(f_* | \mathcal{Y}_{\mathcal{D}})} [H(p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}}, f_*))]$ .

Let us consider the conditional probability density  $p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}})$  under assumptions imposed when evaluating  $H(p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}}))$  and  $\mathbb{E}_{p(f_* | \mathcal{Y}_{\mathcal{D}})} [H(p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}}, f_*))]$  of MES in order to identify the discrepancy in the evaluation of these two terms.

- Evaluating  $H(p(f_{\mathbf{x}} | \mathcal{Y}_{\mathcal{D}}))$ : In MES, there is no ap-

proximation when evaluating  $H(p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}))$  since  $p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})$  is the density of a Gaussian distribution with a closed-form expression for its entropy, i.e.,

$$f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}} \sim \mathcal{N}(\mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2). \quad (7)$$

- Evaluating  $\mathbb{E}_{p(f_*|\mathbf{y}_{\mathcal{D}})}[H(p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*))]$ : As the expectation over  $f_*|\mathbf{y}_{\mathcal{D}}$  is intractable, MES approximates the expectation with an average over a finite set  $\mathcal{F}$  of max-value samples, i.e.,  $\mathbb{E}_{p(f_*|\mathbf{y}_{\mathcal{D}})}[H(p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*))] \approx \frac{1}{|\mathcal{F}|} \sum_{f_* \in \mathcal{F}} H(p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*))$ . This set  $\mathcal{F}$  is obtained in the same approach in (6). Under this assumption, it follows that

$$f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}} \sim \frac{1}{|\mathcal{F}|} \sum_{f_* \in \mathcal{F}} \mathcal{N}_{f_*}(\mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2) \quad (8)$$

where  $\mathcal{N}_{f_*}(\mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2)$  denotes a distribution obtained from restricting a Gaussian distribution  $\mathcal{N}(\mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2)$  to the interval of  $(-\infty, f_*]$ , i.e., an upper truncated Gaussian distribution.

At first glance, the use of the Gaussian distribution in (7) is convenient and fairly straightforward for the closed-form expression of its entropy. However, putting the probability densities inferred from the evaluation of the two terms (i.e., (7) and (8)) together, it becomes obvious that there exists a discrepancy in the evaluation of MES as  $p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})$  is different under assumptions imposed on  $H(p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}))$  and those imposed on  $\mathbb{E}_{p(f_*|\mathbf{y}_{\mathcal{D}})}[H(p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*))]$ . This is because (7) is the result of marginalizing out  $f_*$  for all of its possible values in  $\mathbb{R}$ , while (8) is the result of marginalizing out  $f_*$  over a finite set  $\mathcal{F} \subset \mathbb{R}$ . The discrepancy between a Gaussian distribution and the average over truncated Gaussian distributions is illustrated in Fig. 2. This difference violates the definition of a random variable as  $f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}$  does not have a unique distribution throughout the evaluation of MES. As a result, MES cannot be interpreted as the mutual dependence (mutual information) between 2 random variables since there is not a consistent description for the distribution of  $f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}$  in its evaluation. An undesired consequence is that MES might over-explore as shown in Fig. 3 in Section 4. It is noted that a similar issue also exists in PES [Hernández-Lobato et al., 2014].

## 4 Rectified Max-Value Entropy Search

In this section, we propose a *rectified max-value entropy search* (RMES) measuring the information gain  $I(f_*; y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})$  without the above misconceptions,

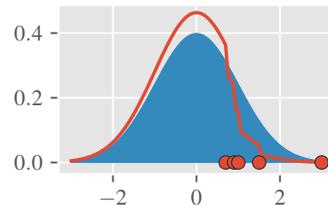


Figure 2: Difference between  $\mathcal{N}(0, 1)$  (shade blue area) and  $\sum_u \mathcal{N}_u(0, 1)/4$  (red line) for  $u \in \{0.7, 0.9, 1, 1.5, 3\}$  (red circles).

i.e.,  $\alpha_{\text{RMES}}(\mathbf{x}, \mathbf{y}_{\mathcal{D}}) \triangleq I(f_*; y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}) = H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})) - \mathbb{E}_{p(f_*|\mathbf{y}_{\mathcal{D}})}[H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*))]$ . The expectation over  $f_*|\mathbf{y}_{\mathcal{D}}$  is approximated with a finite set  $\mathcal{F}$  of samples obtained by optimizing sampled functions from the GP posterior [Hernández-Lobato et al., 2014, Wang and Jegelka, 2017].

In MES,  $f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  follows a truncated Gaussian distribution whose entropy has a closed-form expression. On the contrary, in RMES,  $y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  is the sum of a Gaussian random variable  $\epsilon$  and a truncated Gaussian random variable  $f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  ( $p(y_{\mathbf{x}}|f_*)$ ) in Fig. 1b). Optimizing RMES which involves the entropy of  $y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  is challenging as no analytical expression of the entropy is available. To resolve this challenge, we first derive a closed-form expression for the probability density of  $y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  in the following theorem.

**Theorem 1** *The probability density function of  $y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  is expressed as:*

$$p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*) = \mathcal{N}(y_{\mathbf{x}}; \mu_{\mathbf{x}}, \sigma_+^2) \frac{\Psi(g_{f_*}(y_{\mathbf{x}}))}{\Psi(h_{f_*}(\mathbf{x}))} \quad (9)$$

where  $g_{f_*}(y_{\mathbf{x}}) \triangleq (\sigma_+^2 f_* - \sigma_n^2 \mu_{\mathbf{x}} - \sigma_{\mathbf{x}}^2 y_{\mathbf{x}}) / (\sigma_{\mathbf{x}} \sigma_n \sigma_+)$ ;  $\sigma_+^2 \triangleq \sigma_{\mathbf{x}}^2 + \sigma_n^2$ ;  $h_{f_*}(\mathbf{x})$  and  $\Psi$  are specified in (6).

The derivation of (9) is in Appendix A. Although we cannot evaluate the entropy of  $y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  analytically, we can optimize it using the stochastic gradient ascent. Given the closed-form expression in (9), a straightforward solution is to express  $H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*))$  as  $\mathbb{E}_{p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*)}[-\log p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*)]$ . Then, one can sample batches of  $y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  by first sampling  $f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  from a truncated Gaussian distribution, then, sampling the Gaussian noise  $\epsilon$ , and summing them up. Stochastic gradient optimization has been made feasible for an expectation over a random sample following a truncated Gaussian distribution by a clever trick recently, namely, implicit reparameterization gradients [Figurnov et al., 2018]. However, as samples of  $y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  depend on  $f_*$ , we need to draw different samples of  $y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  for different values of  $f_* \in \mathcal{F}$ , which is inefficient.

Fortunately, we can design a more efficient approach to stochastically optimize  $H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*))$  where samples are drawn independently from  $f_*$  and the optimization only requires the simple reparameterization trick: the Gaussian standardization. By factoring  $p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*)$  into  $w_{f_*}(y_{\mathbf{x}}) \triangleq \Psi(g_{f_*}(y_{\mathbf{x}}))/\Psi(h_{f_*}(\mathbf{x}))$  dependent on  $f_*$  and  $p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}) = \mathcal{N}(y_{\mathbf{x}}; \mu_{\mathbf{x}}, \sigma_{+}^2)$  independent from  $f_*$ . We can express  $H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*))$  in an importance sampling approach with the weight  $w_{f_*}(y_{\mathbf{x}})$  as

$$H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*)) = \mathbb{E}_{p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})} [-w_{f_*}(y_{\mathbf{x}}) \log p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*)]$$

where  $p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*)$  is computed in (9). The expectation over the Gaussian distribution  $p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})$  which is independent from  $f_*$  can be reparameterized using the Gaussian standardization in [Kingma and Welling, 2013]:

$$H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*)) = \mathbb{E}_{p(\nu)} [-w_{f_*}(t(\nu)) \log p(t(\nu)|\mathbf{y}_{\mathcal{D}}, f_*)]$$

where  $\nu \sim \mathcal{N}(0, 1)$ ,  $t(\nu) \triangleq \nu\sigma_{+} + \mu_{\mathbf{x}}$ , and  $p(t(\nu)|\mathbf{y}_{\mathcal{D}}, f_*) \triangleq p(y_{\mathbf{x}} = t(\nu)|\mathbf{y}_{\mathcal{D}}, f_*)$ . The expectation over  $\nu$  does not depend on  $\mathbf{x}$  nor  $f_*$ , so samples of  $\nu$  can be shared between different samples of  $f_*$ . Furthermore, the gradient of the acquisition function is not propagated through this sampling procedure. Hence, the term  $\mathbb{E}_{p(f_*|\mathbf{y}_{\mathcal{D}})} [H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*))]$   $\approx \frac{1}{|\mathcal{F}|} \sum_{f_* \in \mathcal{F}} H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*))$  in RMES can be expressed as:

$$\mathbb{E}_{p(\nu)} \left[ -\frac{1}{|\mathcal{F}|} \sum_{f_* \in \mathcal{F}} w_{f_*}(t(\nu)) \log p(t(\nu)|\mathbf{y}_{\mathcal{D}}, f_*) \right]. \quad (10)$$

Regarding the term  $H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}))$  in RMES, to avoid the discrepancy in the evaluation of MES (Section 3.2), we apply the same restriction of  $f_*$  to the finite set  $\mathcal{F}$  onto  $H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}))$  to obtain

$$H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})) = \mathbb{E}_{p(\nu)} \left[ -\frac{1}{|\mathcal{F}|} \sum_{f_* \in \mathcal{F}} w_{f_*}(t(\nu)) \times \log \left( \frac{1}{|\mathcal{F}|} \sum_{f'_* \in \mathcal{F}} p(t(\nu)|\mathbf{y}_{\mathcal{D}}, f'_*) \right) \right].$$

where the Gaussian standardization is used to reparameterize  $y_{\mathbf{x}}$ . Hence, from the above equation and (10), we have  $\alpha_{\text{RMES}}(\mathbf{x}, \mathbf{y}_{\mathcal{D}}) \triangleq I(f_*; y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}) =$

$$\mathbb{E}_{p(\nu)} \left[ \frac{1}{|\mathcal{F}|} \sum_{f_* \in \mathcal{F}} w_{f_*}(t(\nu)) \log \frac{|\mathcal{F}| p(t(\nu)|\mathbf{y}_{\mathcal{D}}, f_*)}{\sum_{f'_* \in \mathcal{F}} p(t(\nu)|\mathbf{y}_{\mathcal{D}}, f'_*)} \right] \quad (11)$$

which can be optimized using a stochastic optimization algorithm such as Adam [Kingma and Ba, 2015]. It is

noted that in (11), samples of  $\nu$  are shared for different samples of the max-value  $f_*$  and both terms in (3). Hence, the number of samples of  $\nu$  can be reduced in comparison with the approach of directly sampling  $y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, f_*$  mentioned above.

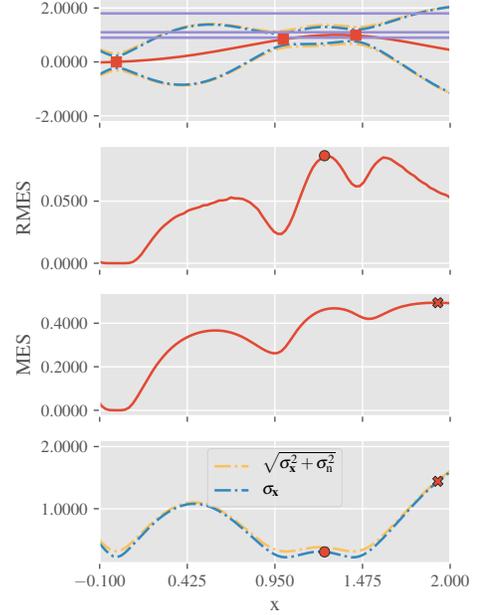


Figure 3: Example where MES tends towards an exploration strategy. The plots share the x-axis. The top plot shows the data samples as red squares; the GP posterior mean as a red line; the GP posterior standard deviations of  $y_{\mathbf{x}}$  and  $f_{\mathbf{x}}$  as dashed yellow lines and dashed blue lines, respectively; and the max-value samples as purple lines. The middle two plots show the values of RMES and MES acquisition functions with the maximum values of RMES and MES (the input queries) as a circle and a cross, respectively. The inputs that maximize RMES and MES are also shown in the bottom plot as a circle and a cross, respectively.

As mentioned in Section 3, the misconceptions in MES cause an imbalance in the exploration-exploitation trade-off. To observe the effects of correcting these misconceptions in RMES, let us investigate two simple examples where RMES and MES select different input queries such that MES over-explores and over-exploits in Fig. 3 and Fig. 4, respectively.

Fig. 3 illustrates an example where MES tends towards an exploration strategy. In this figure, the noise variance is small to minimize the effect of the noise in the observation, i.e., the noiseless observation issue in Section 3.1. Hence, the dashed blue and dashed yellow lines in Fig. 3 are mostly overlapped since the standard deviation values of  $y_{\mathbf{x}}$  and  $f_{\mathbf{x}}$  are almost the same. We observe that MES selects an input query (the cross)

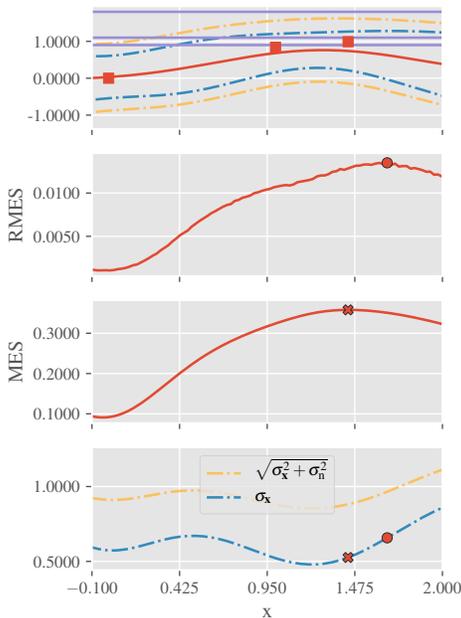


Figure 4: Example where MES tends towards an exploitation strategy. The notations are adopted from Fig. 3.

whose function output has a higher posterior variance and a smaller posterior mean than RMES. In other words, MES tends towards exploration. However, MES explores at a cost of not exploiting an uncertain input with a high posterior mean which is selected by RMES (the circle). This could be because the discrepancy in the evaluation issue in Section 3.2 causes the uncertainty in the GP posterior (i.e., the term  $H(p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}))$ ) to have a high influence on MES. As a result, MES selects an input query far away from the data samples in this example. Therefore, it is noted that a small noise variance does not guarantee a good selection strategy of MES due to the discrepancy in the evaluation. An undesired consequence of this over-exploration is that MES does not query inputs close to the maximizer compared with RMES as shown in the experiments in Section 5.

Fig. 4 illustrates an example where MES tends towards an exploitation strategy. In this example, the length-scale is set such that function values in Fig. 4 are more correlated with one another in comparison with those in Fig. 3. The noise variance is set to a larger value such that the variance of  $y_{\mathbf{x}}$  is significantly larger than that of  $f_{\mathbf{x}}$ . Hence, there is a gap between the dashed blue and dashed yellow lines in the top and bottom plots of Fig. 4. MES selects an input query (the cross) whose function value has a smaller posterior variance and a larger posterior mean than RMES. In other words, MES tends towards exploitation. How-

ever, MES exploits at an input query where the uncertainty of the noise overwhelms the uncertainty of the unknown objective function since the yellow line is much larger than the blue line at the cross in the bottom plot. This could be because MES overestimates the amount of information gain by not taking into account the noise, i.e., replacing the noisy observation  $y_{\mathbf{x}}$  with a noiseless function output  $f_{\mathbf{x}}$  in the acquisition function (Section 3.1). On the other hand, RMES selects an input query (the circle) that balances between the noise in  $y_{\mathbf{x}}$  and information about  $f_{\mathbf{x}}$ . Over-exploitation could prevent MES from leaving a suboptimal maximizer, which causes its poor performance in the experiments in Section 5.

## 5 Experiments

In this section, we empirically evaluate the performance of our RMES and existing acquisition functions such as EI [Mockus et al., 1978], UCB [Srinivas et al., 2010], PES [Hernández-Lobato et al., 2014], and MES [Wang and Jegelka, 2017]. Similar to the MES work [Wang and Jegelka, 2017], we use 2 evaluation criteria: *simple regret* (SR) and *inference regret* (IR). The simple regret measures the regret of the best input query so far, i.e.,  $f_{\mathbf{x}^*} - \max_{\mathbf{x}' \in \mathcal{D}} f_{\mathbf{x}'}$ . The inference regret measures the regret of the inferred maximizer which is often defined as the maximizer of the GP posterior mean function [Hennig and Schuler, 2012, Hernández-Lobato et al., 2014, Wang and Jegelka, 2017]. In other words, the inference regret is defined as  $f_{\mathbf{x}^*} - f_{\arg\max_{\mathbf{x}' \in \mathcal{X}} \mu_{\mathbf{x}'}}$  where  $\mu_{\mathbf{x}}$  is defined in (1).

The experiments include (1) synthetic function benchmarks such as a function sample drawn a GP, the Branin-Hoo function, the 2-dimensional Michaelwicz function, and the eggholder function which is a difficult function to optimize due to its many local maxima; and (2) real-world optimization problems. As an environmental sensing problem, we use the pH field of Broom’s Barn farm [Webster and Oliver, 2007] which is spatially distributed over a 1200m by 680m region discretized into  $31 \times 18$  grid of sampling locations. To generate a continuous objective function from the dataset, we fit a GP model to the dataset and use its mean function as the objective function which is unknown to all BO algorithms. For tuning hyperparameters of a machine learning model, we compare the performance of different acquisition functions to tune the hyperparameters of a *support vector machine* (SVM) to fit the Wisconsin breast cancer dataset from the UCI repository [Dua and Graff, 2017]. There are two hyperparameters of the SVM: the penalty parameter of the error term in the range  $[0.5, 2]$  and the natural

logarithm of the kernel coefficient for the radial basis function kernel of the SVM in the range  $[-5, -3]$ . Given the SVM’s hyperparameters, the unknown objective function is the 100-fold cross-validation accuracy and the noisy observation is the 20-fold cross-validation accuracy. The latter is provided to BO as observations at input queries.

To account for the randomness in the observations, synthetic experiments are repeated 15 times and the real-word experiments are repeated 10 times with random initialization of 2 training data samples. The logarithm to the base 10 of the performance measure average is reported. The objective functions are each modeled as a sample of a GP whose kernel hyperparameters are learned using maximum likelihood estimation [Rasmussen and Williams, 2006]. The zero mean and the SE kernel are used. The objective functions are shifted to have a zero mean.

For MES and RMES, there are 5 samples of the max-value at each BO iteration. For PES, there are 5 samples of the maximizer at each BO iteration. The max-value and maximizer samples are drawn by optimizing function samples from the GP posterior [Hernández-Lobato et al., 2014, Wang and Jegelka, 2017]. The approximated sampling of the max-value via a Gumbel distribution [Wang and Jegelka, 2017] is not used since we do not want the approximation quality to tamper with the performance of the acquisition functions.

### 5.1 Synthetic Function Benchmarks

In this section, we consider synthetic function benchmarks: the Branin-Hoo function (Fig. 5a) and the eggholder function (Fig. 5b). The experiments are conducted with both small and large noise standard deviations:  $\sigma_n = 0.01$  and  $\sigma_n = 0.3$ . As the Branin-Hoo and eggholder functions are often used in minimization problems, the negative values of these functions are used as the objective function. Experiments with a function sample drawn from a GP and the 2-dimensional Michaelwicz function are in the Appendix B.

The logarithm to the base 10 of the average of the simple regret (SR) and that of the inference regret (IR) at each BO iteration are shown in Figs. 6 and 8. In general, RMES outperforms MES in all these experiments by converging to both small simple and inference regrets. This empirically illustrates the beneficial effects of correcting misconceptions in MES. In particular, when the noise is small, i.e.,  $\sigma_n = 0.01$ , it is likely that the over-exploration of MES prevents it from properly exploiting the location of the maximizer as shown in Fig. 3. On the other hand, when the noise

is large, i.e.,  $\sigma_n = 0.3$ , the over-exploitation traps MES at suboptimal maxima.

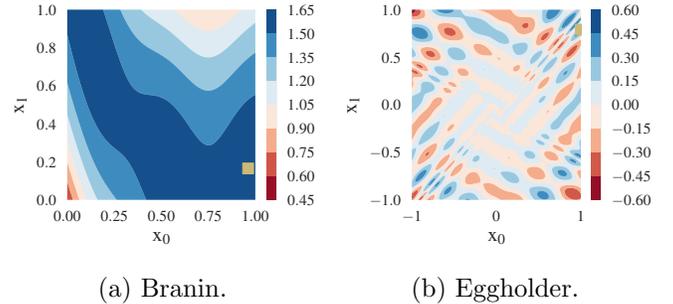


Figure 5: Synthetic functions whose maximizers are denoted as yellow squares.

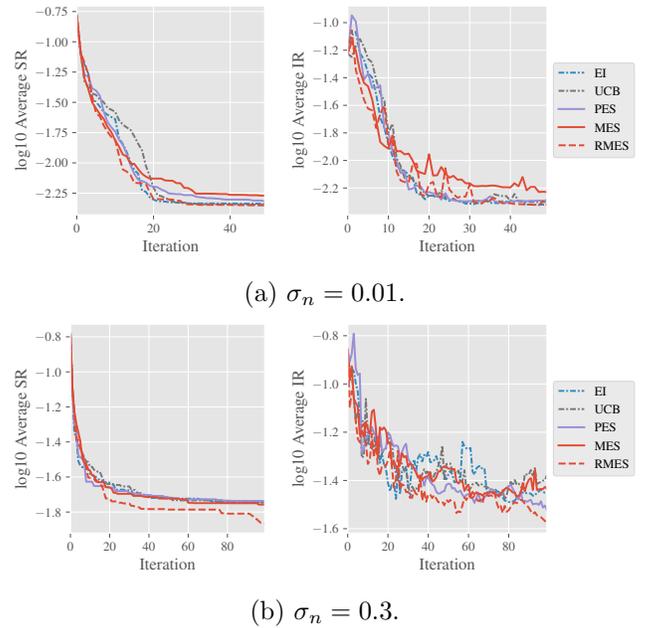


Figure 6: Branin-Hoo function.

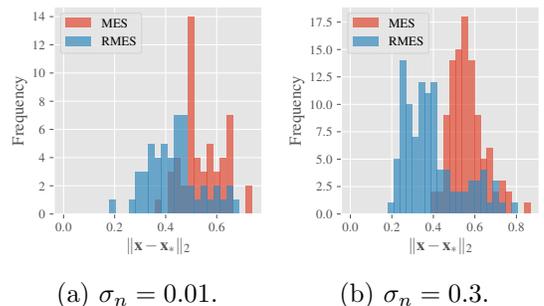
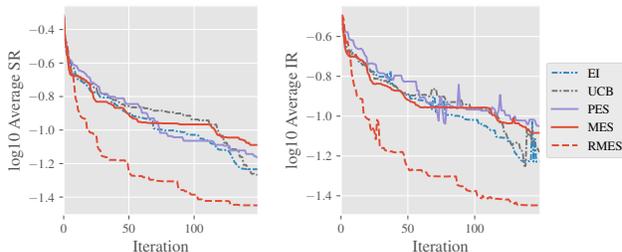


Figure 7: Distance of input queries to the maximizer of the Branin-Hoo function.

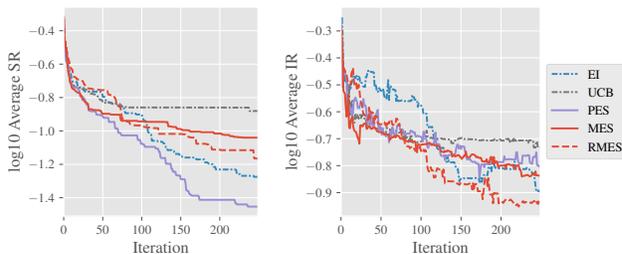
The Branin-Hoo function is a simple function with a high correlation between function values in the input

domain (Fig. 5a), so all the acquisition functions perform similarly in Fig. 6. Nonetheless, MES still underperforms slightly in comparison with RMES. Fig. 7 shows the histogram of the average distance from input queries to the maximizer over 15 repetitions of the experiment, i.e.,  $\|\mathbf{x} - \mathbf{x}_*\|_2 \triangleq \sqrt{(\mathbf{x} - \mathbf{x}_*)^\top (\mathbf{x} - \mathbf{x}_*)}$ , for RMES and MES. Due to the high correlation between function values, it does not require either RMES or MES to query for inputs close to the maximizer to obtain a good performance. However, we still observe that MES queries for inputs farther from the maximizer than RMES, which explains its poorer performance. This phenomenon is most likely due to the imbalance in the exploration-exploitation trade-off as illustrated in Section 4. The same observation is noted in other experiments: the function sample drawn from a GP and the Michaelwicz function in Fig. 13 and Fig. 15 in the Appendix B, respectively.

The eggholder function is a difficult function to optimize as it has many local maxima (Fig. 5b). Hence, it requires an acquisition function to balance between exploration and exploitation to be query efficient. In this case, RMES outperforms other acquisition functions by converging to a smaller regret except for the simple regret when  $\sigma_n = 0.3$  in Fig. 8.



(a)  $\sigma_n = 0.01$ .



(b)  $\sigma_n = 0.3$ .

Figure 8: Eggholder function.

## 5.2 Real-world Optimization Problems

This section presents the results of BO algorithms for real-world optimization problems including an environment sensing problem of finding the location with the maximum pH value (Fig. 11b in Appendix B),

and a machine learning training problem of tuning an SVM model for training the Wisconsin breast cancer dataset. The noise standard deviation values of the pH field and tuning the SVM model problems are  $\sigma_n = 0.25$  and  $\sigma_n = 0.02$ , respectively. Although both MES and RMES show reasonable performance in these experiments, RMES has an advantage in the simple regret of the tuning SVM experiment. In comparison with EI and UCB, the difference in performance is insignificant. Regarding PES, it does not perform as well as the other acquisition functions, especially in the tuning SVM experiment.

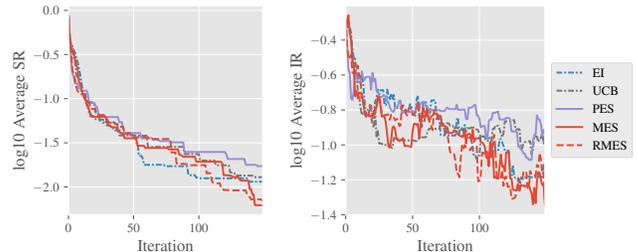


Figure 9: The pH field experiment.

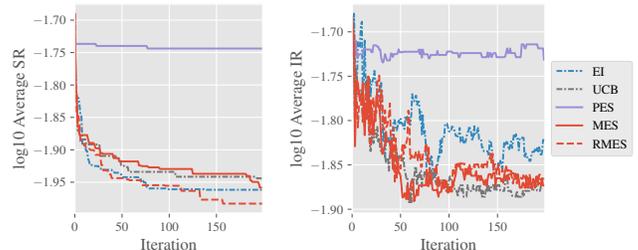


Figure 10: Tuning SVM model experiment.

## 6 Conclusion

In this paper, we illustrate two misconceptions in the existing MES acquisition function: noiseless observations and discrepancy in the evaluation, which have negative implications on its interpretation as the mutual information as well as on its empirical performance. Based on the insights of these issues, we develop the RMES acquisition function that produces a more accurate measure of the information gain about the max-value through observing noisy function outputs. As a result, it has a superior performance compared with MES thanks to the correction of these issues. Nonetheless, optimizing RMES is more challenging than the existing MES as it does not have any closed-form expression. To overcome this hurdle, we derive a closed-form expression for the probability density of the noisy observation given the max-value, and

design an efficient sampling approach to do stochastic gradient ascent with a reparameterization trick. It is empirically shown in several synthetic function benchmarks and real-world optimization problems that the performance of RMES is preferable over MES and competitive with existing acquisition functions.

## References

- [Brochu et al., 2010] Brochu, E., Cora, V. M., and de Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv:1012.2599*.
- [Dua and Graff, 2017] Dua, D. and Graff, C. (2017). UCI machine learning repository.
- [Figurnov et al., 2018] Figurnov, M., Mohamed, S., and Mnih, A. (2018). Implicit reparameterization gradients. In *Proc. NeurIPS*, pages 441–452.
- [Hennig and Schuler, 2012] Hennig, P. and Schuler, C. J. (2012). Entropy search for information-efficient global optimization. *JMLR*, pages 1809–1837.
- [Hernández-Lobato et al., 2014] 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.
- [Hoffman and Ghahramani, 2015] Hoffman, M. W. and Ghahramani, Z. (2015). Output-space predictive entropy search for flexible global optimization. In *NIPS workshop on Bayesian Optimization*.
- [Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proc. ICLR*.
- [Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv:1312.6114*.
- [Knudde et al., 2018] Knudde, N., Couckuyt, I., Spina, D., Łukasik, K., Barmuta, P., Schreurs, D., and Dhaene, T. (2018). Data-efficient Bayesian optimization with constraints for power amplifier design. In *2018 IEEE MTT-S International conference on numerical electromagnetic and multiphysics modeling and optimization (NEMO)*, pages 1–3. IEEE.
- [Kushner, 1964] Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multi-peak curve in the presence of noise. *Journal of basic engineering*, 86(1):97–106.
- [Mockus et al., 1978] Mockus, J., Tiesis, V., and Žilinskas, A. (1978). The application of Bayesian methods for seeking the extremum. In Dixon, L. C. W. and Szegö, G. P., editors, *Towards Global Optimization 2*, pages 117–129. North-Holland Publishing Company.
- [Rasmussen and Williams, 2006] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.
- [Ru et al., 2018] Ru, B., McLeod, M., Granzio, D., and Osborne, M. A. (2018). Fast information-theoretic Bayesian optimisation. In *Proc. ICML*, pages 4381–4389.
- [Shahriari et al., 2015] Shahriari, B., Swersky, K., Wang, Z., Adams, R., and de Freitas, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175.
- [Snoek et al., 2012] Snoek, J., Larochelle, H., and Adams, R. (2012). Practical Bayesian optimization of machine learning algorithms. In *Proc. NIPS*, pages 2951–2959.
- [Srinivas et al., 2010] 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.
- [Takeno et al., 2019] Takeno, S., Fukuoka, H., Tsukada, Y., Koyama, T., Shiga, M., Takeuchi, I., and Karasuyama, M. (2019). Multi-fidelity Bayesian optimization with max-value entropy search. *arXiv:1901.08275*.
- [Wang and Jegelka, 2017] Wang, Z. and Jegelka, S. (2017). Max-value entropy search for efficient Bayesian optimization. In *Proc. ICML*, pages 3627–3635.
- [Webster and Oliver, 2007] Webster, R. and Oliver, M. (2007). *Geostatistics for environmental scientists*. John Wiley & Sons.

## A Probability Density of $y_{\mathbf{x}}|y_{\mathcal{D}}, f_*$

This section derives a closed-form expression of  $p(y_{\mathbf{x}}|y_{\mathcal{D}}, f_*)$ . Different from the relatively straightforward expression in (4), we can express  $p(y_{\mathbf{x}}|y_{\mathcal{D}}, f_*)$  in an importance sampling manner. That is, samples of  $y_{\mathbf{x}}|y_{\mathcal{D}}, f_*$  can be obtained by first drawing a sample  $y$  of  $y_{\mathbf{x}}|y_{\mathcal{D}}$  and then, weighting the sample with  $p(f_{\mathbf{x}} \leq f_*|y_{\mathcal{D}}, y_{\mathbf{x}} = y)$ . Let  $y$  denote a realization of the random variable  $y_{\mathbf{x}}$ , we have

$$\begin{aligned} & p(y_{\mathbf{x}} = y|y_{\mathcal{D}}, f_*) \\ & \propto p(y_{\mathbf{x}} = y|y_{\mathcal{D}})p(f_{\mathbf{x}} \leq f_*|y_{\mathcal{D}}, y_{\mathbf{x}} = y) \end{aligned} \quad (12)$$

where  $p(f_{\mathbf{x}} \leq f_*|y_{\mathcal{D}}, y_{\mathbf{x}} = y)$  can be considered as the cumulative density function of the distribution specified by  $p(f_{\mathbf{x}}|y_{\mathcal{D}}, y_{\mathbf{x}} = y)$ . Let  $f$  denote a realization of the random variable  $f_{\mathbf{x}}$ . Recall that  $y_{\mathbf{x}} = f_{\mathbf{x}} + \epsilon$  where the noise  $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ , we can evaluate the probability density  $p(f_{\mathbf{x}} = f|y_{\mathcal{D}}, y_{\mathbf{x}} = y)$  as below:

$$\begin{aligned} & p(f_{\mathbf{x}} = f|y_{\mathcal{D}}, y_{\mathbf{x}} = y) \\ & = p(f_{\mathbf{x}} = f|y_{\mathcal{D}})p(\epsilon = y - f) \\ & = \mathcal{N}(f; \mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2)\mathcal{N}(y - f; 0, \sigma_n^2) \\ & = \mathcal{N}\left(f; \frac{\sigma_n^2\mu_{\mathbf{x}} + \sigma_{\mathbf{x}}^2y}{\sigma_+^2}, \frac{\sigma_{\mathbf{x}}^2\sigma_n^2}{\sigma_+^2}\right) \\ & = \psi\left(\frac{\sigma_+^2f - \sigma_n^2\mu_{\mathbf{x}} - \sigma_{\mathbf{x}}^2y}{\sigma_{\mathbf{x}}\sigma_n\sigma_+}\right) \end{aligned}$$

where  $p(f_{\mathbf{x}} = f|y_{\mathcal{D}}) = \mathcal{N}(f; \mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2)$  is the GP posterior distribution (1);  $\psi$  denotes the probability density of the standard Gaussian distribution, and  $\sigma_+^2 \triangleq \sigma_{\mathbf{x}}^2 + \sigma_n^2$ . Hence, we have the cumulative density function  $p(f_{\mathbf{x}} \leq f_*|y_{\mathcal{D}}, y_{\mathbf{x}} = y)$  expressed as:

$$p(f_{\mathbf{x}} \leq f_*|y_{\mathcal{D}}, y_{\mathbf{x}} = y) = \Psi\left(\frac{\sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}} - \sigma_{\mathbf{x}}^2y}{\sigma_{\mathbf{x}}\sigma_n\sigma_+}\right) \quad (13)$$

where  $\Psi$  denotes the cumulative density function of the standard Gaussian distribution. By substituting (13) into (12), we obtain

$$\begin{aligned} & p(y_{\mathbf{x}} = y|y_{\mathcal{D}}, f_*) \\ & \propto p(y_{\mathbf{x}} = y|y_{\mathcal{D}})\Psi\left(\frac{\sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}} - \sigma_{\mathbf{x}}^2y}{\sigma_{\mathbf{x}}\sigma_n\sigma_+}\right) \\ & = \mathcal{N}(y; \mu_{\mathbf{x}}, \sigma_+^2)\Psi\left(\frac{\sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}} - \sigma_{\mathbf{x}}^2y}{\sigma_{\mathbf{x}}\sigma_n\sigma_+}\right) \\ & = \mathcal{N}(y; \mu_{\mathbf{x}}, \sigma_+^2)\Psi(g_{f_*}(y)) \end{aligned} \quad (14)$$

where  $g_{f_*}(y) \triangleq \frac{\sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}} - \sigma_{\mathbf{x}}^2y}{\sigma_{\mathbf{x}}\sigma_n\sigma_+}$ . To obtain the expression for  $p(y_{\mathbf{x}} = y|y_{\mathcal{D}}, f_*)$ , we need to evaluate

the integral of (14):

$$\begin{aligned} & \int \mathcal{N}(y; \mu_{\mathbf{x}}, \sigma_+^2)\Psi(g_{f_*}(y)) \, dy \\ & = \int p(y_{\mathbf{x}} = y|y_{\mathcal{D}})\Psi\left(\frac{\sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}} - \sigma_{\mathbf{x}}^2y}{\sigma_{\mathbf{x}}\sigma_n\sigma_+}\right) \, dy \\ & = \int p(y_{\mathbf{x}} = y|y_{\mathcal{D}})p\left(\nu \leq \frac{\sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}} - \sigma_{\mathbf{x}}^2y}{\sigma_{\mathbf{x}}\sigma_n\sigma_+}\right) \, dy \\ & = p\left(\nu \leq \frac{\sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}} - \sigma_{\mathbf{x}}^2y_{\mathbf{x}}}{\sigma_{\mathbf{x}}\sigma_n\sigma_+} \middle| y_{\mathcal{D}}\right) \\ & = p(\nu\sigma_{\mathbf{x}}\sigma_n\sigma_+ + \sigma_{\mathbf{x}}^2y_{\mathbf{x}} \leq \sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}}|y_{\mathcal{D}}) \end{aligned}$$

where  $\nu \sim \mathcal{N}(0, 1)$ . Recall that  $y_{\mathbf{x}}|y_{\mathcal{D}} \sim \mathcal{N}(\mu_{\mathbf{x}}, \sigma_+^2)$ , it implies  $\nu\sigma_{\mathbf{x}}\sigma_n\sigma_+ + \sigma_{\mathbf{x}}^2y_{\mathbf{x}}$  follows a Gaussian distribution  $\mathcal{N}(\sigma_{\mathbf{x}}^2\mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2\sigma_+^4)$ . Therefore,  $p(\nu\sigma_{\mathbf{x}}\sigma_n\sigma_+ + \sigma_{\mathbf{x}}^2y_{\mathbf{x}} \leq \sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}}|y_{\mathcal{D}})$  is the cumulative density function at  $\sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}}$  of a Gaussian distribution  $\mathcal{N}(\sigma_{\mathbf{x}}^2\mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2\sigma_+^4)$ , i.e.,

$$\begin{aligned} & \int \mathcal{N}(y; \mu_{\mathbf{x}}, \sigma_+^2)\Psi(g_{f_*}(y)) \, dy \\ & = \Psi(\sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}}; \sigma_{\mathbf{x}}^2\mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2\sigma_+^4) \\ & = \Psi\left(\frac{f_* - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}}\right). \end{aligned} \quad (15)$$

Hence, from (14) and (15), we obtain the exact probability density function of  $y_{\mathbf{x}}|y_{\mathcal{D}}, f_*$  as below:

$$p(y_{\mathbf{x}}|y_{\mathcal{D}}, f_*) = \mathcal{N}(y_{\mathbf{x}}; \mu_{\mathbf{x}}, \sigma_+^2) \frac{\Psi(g_{f_*}(y_{\mathbf{x}}))}{\Psi(h_{f_*}(\mathbf{x}))} \quad (16)$$

where  $g_{f_*}(y_{\mathbf{x}}) \triangleq \frac{\sigma_+^2f_* - \sigma_n^2\mu_{\mathbf{x}} - \sigma_{\mathbf{x}}^2y_{\mathbf{x}}}{\sigma_{\mathbf{x}}\sigma_n\sigma_+}$  and  $h_{f_*}(\mathbf{x}) \triangleq \frac{f_* - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}}$ .

## B Other Synthetic Function Benchmarks

In this section, we describe experiments with other synthetic function benchmarks including: a function sample drawn from a GP with hyperparameters:  $l = 0.33$ ,  $\sigma_s^2 = 1$  (Fig. 11a); and the 2-dimensional Michaelwicz function.

In the function sample drawn from a GP experiment (Fig. 12), MES converges to a much larger regret in comparison to other acquisition functions. We plot the distance from the input queries to the maximizer, i.e.,  $\|\mathbf{x} - \mathbf{x}_*\|_2 \triangleq \sqrt{(\mathbf{x} - \mathbf{x}_*)^\top (\mathbf{x} - \mathbf{x}_*)}$ , for RMES and MES in Fig. 13 which shows that MES does not query inputs close the maximizer for both values of  $\sigma_n$  due to the imbalance in the exploration-exploitation trade-off. On the other hand, RMES spends a large proportion of queries for inputs close to the maximizer.

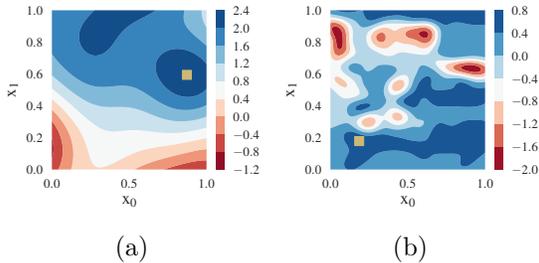


Figure 11: Function sample drawn from GP (a) and the GP posterior mean of the pH field (b).

RMES explores more than EI, UCB, and PES in this experiment as RMES converges slower in Fig. 12. However, RMES converges to a better simple regret when  $\sigma_n = 0.01$ . As this function is relatively easy to optimize (Fig. 11a), EI can quickly exploit to get to the maximizer, so it outperforms the other acquisition functions in the inference regret when  $\sigma_n = 0.01$  and in the simple regret when  $\sigma_n = 0.3$ . On the other hand, PES achieves the best inference regret when  $\sigma_n = 0.3$ .

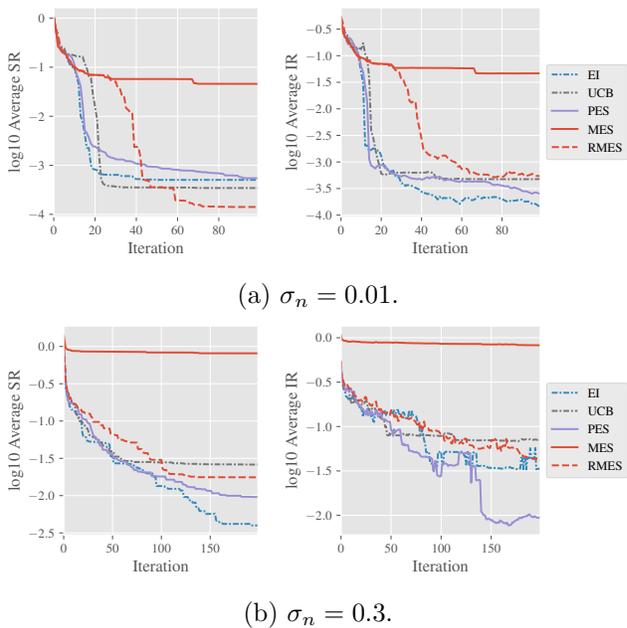
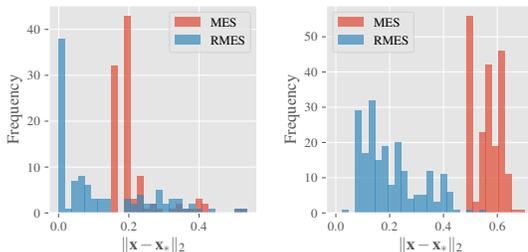


Figure 12: A function sample drawn from a GP.

Fig. 14 shows the results of the 2-dimensional Michaelwicz function. We can observe that MES does not perform as well as the other acquisition functions. Fig. 15 of the distance of input queries to the maximizer also shows that MES does not query inputs close to the maximizer in comparison to RMES, which means MES cannot properly search for the maximizer. Among EI, UCB, PES, and RMES, we observe that in terms of the simple regret, RMES is on par with EI and they outperform the other acquisition functions. Regard-



(a)  $\sigma_n = 0.01$ . (b)  $\sigma_n = 0.3$ .

Figure 13: Distance of input queries to the maximizer of a function sample drawn from a GP.

ing the inference regret, PES and EI have the best performance though RMES matches the performance of UCB.

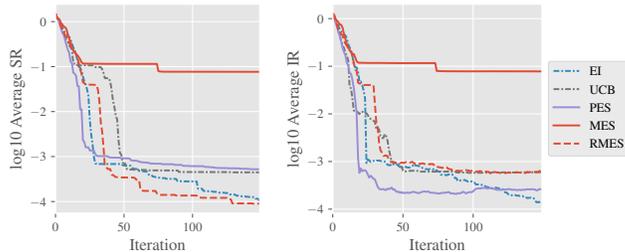


Figure 14: Michaelwicz function with  $\sigma_n = 0.01$ .

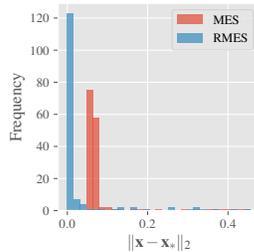


Figure 15: Distance of input queries to the maximizer of the Michaelwicz function.