Supplementary Materials for the Submission: "Multi-fidelity Bayesian Optimization with Max-value Entropy Search and its Parallelization"

A. Semiparametric Latent Factor Model and its RFM approximation

A.1. Model Definition

Semiparametric Latent Factor Model (SLFM) is a Gaussian process based multiple response model (Teh et al., 2005). SLFM represents each output as a sum of C functions having different kernel functions $k_1, \dots, k_C$, where $k_c : \boldsymbol{x} \times \boldsymbol{x} \to \mathbb{R}$ is a kernel function. Let $w_{mc} \in \mathbb{R}$ be a weight that the m-th output (fidelity) assigns to the c-th function. By introducing an independent term $\kappa_{cm} > 0$, the kernel function is written as

$$k((\boldsymbol{x}, m), (\boldsymbol{x}', m')) = \sum_{c=1}^{C} (w_{cm} w_{cm'} + \kappa_{cm} \delta_{m=m'}) k_c(\boldsymbol{x}, \boldsymbol{x}'),$$

where $\delta_{m=m'} = 1$ if $m = m'$, and 0 otherwise. The parameters w_{cm} and κ_{cm} which control dependence between multiple outputs are regarded as hyper-parameters, and standard approaches such as marginal likelihood optimization are often used to set them.

A.2. RFM for SLFM

Let $\boldsymbol{f}_{\boldsymbol{x}} \coloneqq (f_{\boldsymbol{x}}^{(1)}, \dots, f_{\boldsymbol{x}}^{(M)})^\top$ be the M-dimensional output vector, and

$$\mathrm{cov}(\boldsymbol{f}_{\boldsymbol{x}}, \boldsymbol{f}_{\boldsymbol{x}'}) \coloneqq \begin{bmatrix} k((\boldsymbol{x}, 1), (\boldsymbol{x}', 1)) & \cdots & k((\boldsymbol{x}, 1), (\boldsymbol{x}', M)) \\ \vdots & & \vdots \\ k((\boldsymbol{x}, M), (\boldsymbol{x}', 1)) & \cdots & k((\boldsymbol{x}, M), (\boldsymbol{x}', M)) \end{bmatrix}$$

be the $M \times M$ covariance matrix of $\boldsymbol{x}$ and $\boldsymbol{x}'$. By defining $\boldsymbol{w}_c \coloneqq (w_{c1}, \dots, w_{cM})$ and $\boldsymbol{\kappa}_c \coloneqq (\kappa_{c1}, \dots, \kappa_{cM})$, this covariance is written as

$$\mathrm{cov}(\boldsymbol{f}_{\boldsymbol{x}}, \boldsymbol{f}_{\boldsymbol{x}'}) = \sum_{c=1}^{C} (\boldsymbol{w}_c \boldsymbol{w}_c^\top + \mathrm{diag}(\boldsymbol{\kappa}_c)) k_c(\boldsymbol{x}, \boldsymbol{x}').$$

Since $k_c(\boldsymbol{x}, \boldsymbol{x}')$ is assumed to be one of stationary kernel functions (e.g., Gaussian kernel), RFM can produce a feature vector representation $\boldsymbol{\phi}_c$ which approximates the kernel function as $k_c(\boldsymbol{x}, \boldsymbol{x}') \approx \boldsymbol{\phi}_c^\top(\boldsymbol{x}) \boldsymbol{\phi}_c(\boldsymbol{x})$. To transform $\boldsymbol{w}_c \boldsymbol{w}_c^\top + \mathrm{diag}(\boldsymbol{\kappa}_c)$ into a form of inner product, we use the Cholesky decomposition

$$\boldsymbol{w}_c \boldsymbol{w}_c^\top + \mathrm{diag}(\boldsymbol{\kappa}_c) = \boldsymbol{L}_c \boldsymbol{L}_c^\top,$$

where $\boldsymbol{L}_c \in \mathbb{R}^{M \times M}$ is a lower triangular matrix. Then, we obtain

$$\begin{aligned} \mathrm{cov}(\boldsymbol{f}_{\boldsymbol{x}}, \boldsymbol{f}_{\boldsymbol{x}'}) &\approx \sum_{c=1}^{C} \boldsymbol{L}_c \boldsymbol{L}_c^\top \left(\boldsymbol{\phi}_c^\top(\boldsymbol{x}) \boldsymbol{\phi}_c(\boldsymbol{x}') \right) \\ &= \sum_{c=1}^{C} \boldsymbol{\Psi}_c^\top(\boldsymbol{x}) \boldsymbol{\Psi}_c(\boldsymbol{x}') \end{aligned}$$

where $\boldsymbol{\Psi}_c(\boldsymbol{x}) \coloneqq \boldsymbol{L}_c^\top \otimes \boldsymbol{\phi}_c(\boldsymbol{x})$. Here, in the last line, we use the mixed-product property of Kronecker product. Then, the m-th column of $\boldsymbol{\Psi}_c(\boldsymbol{x})$ is defined as the feature of $\boldsymbol{x}$ for the m-th fidelity $\boldsymbol{\phi}(\boldsymbol{x}, m)$.

B. Proof of Lemma 3.1

Using Bayes' theorem, we obtain

$$\begin{aligned} & p(f_{\boldsymbol{x}}^{(m)} \mid f_{\boldsymbol{x}}^{(M)} \leq f_*, \mathcal{D}_t) \\ & = \frac{p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid f_{\boldsymbol{x}}^{(m)}, \mathcal{D}_t) p(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t)}{p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid \mathcal{D}_t)}. \end{aligned} \tag{15}$$

The densities $p(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t)$ and $p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid \mathcal{D}_t)$ are directly obtained from the predictive distribution:

$$\begin{aligned} p(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t) &= \phi(\gamma_{f_{\boldsymbol{x}}^{(m)}}^{(m)}(\boldsymbol{x}))/\sigma_{\boldsymbol{x}}^{(m)}, \\ p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid \mathcal{D}_t) &= \Phi(\gamma_{f_*}^{(M)}(\boldsymbol{x})). \end{aligned} \tag{16}$$

In addition, from (5), $p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid f_{\boldsymbol{x}}^{(m)}, \boldsymbol{x}, \mathcal{D}_t)$ is written as the cumulative distribution of this Gaussian:

$$p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid f_{\boldsymbol{x}}^{(m)}, \mathcal{D}_t) = \Phi((f_* - u(\boldsymbol{x}))/s(\boldsymbol{x})). \tag{17}$$

Substituting (16) and (17) into (15), the entropy is obtained.

C. Information Gain with Noisy Observation

Here, we describe calculation of the mutual information between f_* and noisy observation $y_{\boldsymbol{x}}^{(m)}$, where $y_{\boldsymbol{x}}^{(m)} := y^{(m)}(\boldsymbol{x})$ in this section. The mutual information can be written as the difference of the entropy:

$$I(f_*; y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, \mathcal{D}_t) = H(y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, \mathcal{D}_t) - \mathbb{E}_{p(f_* \mid \boldsymbol{x}, \mathcal{D}_t)}\big[H(y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, f_*, \mathcal{D}_t)\big]. \tag{18}$$

The first term in the right hand side is

$$H(y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, \mathcal{D}_t) = \log \left(\sqrt{2\pi e (\sigma_{\boldsymbol{x}}^{2(m)} + \sigma_{\mathrm{noise}}^2)} \right). \tag{19}$$

Using the sampling approximation of f_*, the second term in (18) is

$$\mathbb{E}_{p(f_* \mid \boldsymbol{x}, \mathcal{D}_t)}\big[H(y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, f_*, \mathcal{D}_t)\big] \approx \sum_{f_* \in \mathcal{F}_*} \frac{1}{|\mathcal{F}_*|} H(y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, f_*, \mathcal{D}_t). \tag{20}$$

For any $\zeta \in \mathbb{R}$, define

$$\gamma_\zeta^{(m)}(\boldsymbol{x}) := (\zeta - \mu_{\boldsymbol{x}}^{(m)})/\sigma_{\boldsymbol{x}}^{(m)},$$

and

$$\rho_\zeta^{(m)}(\boldsymbol{x}) := (\zeta - \mu_{\boldsymbol{x}}^{(m)})/\sqrt{\sigma_{\boldsymbol{x}}^{2(m)} + \sigma_{\mathrm{noise}}^2}.$$

In this case, even for the highest fidelity M, the density $p(y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, f_{\boldsymbol{x}}^{(M)} \leq f_*, \mathcal{D}_t)$ is not the truncated normal because of the noise term. Using Bayes' theorem, we decompose this density as

$$p(y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, f_{\boldsymbol{x}}^{(M)} \leq f_*, \mathcal{D}_t) = \frac{p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid y_{\boldsymbol{x}}^{(m)}, \boldsymbol{x}, \mathcal{D}_t) p(y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, \mathcal{D}_t)}{p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid \boldsymbol{x}, \mathcal{D}_t)}. \tag{21}$$

The densities $p(y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, \mathcal{D}_t)$ and $p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid \boldsymbol{x}, \mathcal{D}_t)$ are directly obtained from the predictive distribution:

$$\begin{aligned} p(y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, \mathcal{D}_t) &= \frac{1}{\sqrt{\sigma_{\boldsymbol{x}}^{2(m)} + \sigma_{\mathrm{noise}}^2}} \phi(\rho_{y_{\boldsymbol{x}}^{(m)}}^{(m)}(\boldsymbol{x})), \\ p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid \boldsymbol{x}, \mathcal{D}_t) &= \Phi(\gamma_{f_*}^{(M)}(\boldsymbol{x})). \end{aligned} \tag{22}$$

The joint marginal distribution $p(f_{\boldsymbol{x}}^{(M)}, y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, \mathcal{D}_t)$ is written as

$$\begin{bmatrix} y_{\boldsymbol{x}}^{(m)} \\ f_{\boldsymbol{x}}^{(M)} \end{bmatrix} \mid \boldsymbol{x}, \mathcal{D}_t \sim \mathcal{N}\left(\begin{bmatrix} \mu_{\boldsymbol{x}}^{(m)} \\ \mu_{\boldsymbol{x}}^{(M)} \end{bmatrix}, \begin{bmatrix} \sigma^{2(m)}_{\boldsymbol{x}} + \sigma^2_{\text{noise}} & \sigma^{2(mM)}_{\boldsymbol{x}} \\ \sigma^{2(mM)}_{\boldsymbol{x}} & \sigma^{2(M)}_{\boldsymbol{x}} \end{bmatrix} \right),$$

From this distribution, we obtain $p(f_{\boldsymbol{x}}^{(M)} \mid y_{\boldsymbol{x}}^{(m)}, \boldsymbol{x}, \mathcal{D}_t)$ as

$$f_{\boldsymbol{x}}^{(M)} \mid y_{\boldsymbol{x}}^{(m)}, \boldsymbol{x}, \mathcal{D}_t \sim \mathcal{N}(u_{\text{noise}}(\boldsymbol{x}), s^2_{\text{noise}}(\boldsymbol{x})),$$

where

$$u_{\text{noise}}(\boldsymbol{x}) = \frac{\sigma^{2(mM)}_{\boldsymbol{x}} \left(y_{\boldsymbol{x}}^{(m)} - \mu_{\boldsymbol{x}}^{(m)} \right)}{\sigma^{2(m)}_{\boldsymbol{x}} + \sigma^2_{\text{noise}}} + \mu_{\boldsymbol{x}}^{(M)},$$
$$s^2_{\text{noise}}(\boldsymbol{x}) = \sigma^{2(M)}_{\boldsymbol{x}} - \frac{\left(\sigma^{2(mM)}_{\boldsymbol{x}} \right)^2}{\sigma^{2(m)}_{\boldsymbol{x}} + \sigma^2_{\text{noise}}}.$$

Thus, $p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid y_{\boldsymbol{x}}^{(m)}, \boldsymbol{x}, \mathcal{D}_t)$ is written as the cumulative distribution of this Gaussian:

$$p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid y_{\boldsymbol{x}}^{(m)}, \boldsymbol{x}, \mathcal{D}_t) = \Phi(\gamma'_{f_*}(\boldsymbol{x})), \tag{23}$$

where, $\gamma'_{f_*}(\boldsymbol{x}) \coloneqq (f_* - u_{\text{noise}}(\boldsymbol{x}))/s_{\text{noise}}(\boldsymbol{x})$. Using (15), (16), and (17) in the proof of Lemma 3.1, the entropy is obtained as

$$\begin{aligned} & H(y_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, f_{\boldsymbol{x}}^{(M)} \leq f_*, \mathcal{D}_t) \\ & \quad = - \int Z \Phi\big(\gamma'_{f_*}(\boldsymbol{x})\big) \phi\big(\rho^{(m)}_{y_{\boldsymbol{x}}^{(m)}}(\boldsymbol{x})\big) \cdot \log \Big(Z \Phi\big(\gamma'_{f_*}(\boldsymbol{x})\big) \phi\big(\rho^{(m)}_{y_{\boldsymbol{x}}^{(m)}}(\boldsymbol{x})\big) \Big) \, \mathrm{d} y_{\boldsymbol{x}}^{(m)}, \end{aligned} \tag{24}$$

where $Z \coloneqq 1/\sqrt{\sigma_{\boldsymbol{x}}^{2(m)} + \sigma^2_{\text{noise}}} \Phi(\gamma^{(M)}_{f_*}(\boldsymbol{x}))$. The integral in (24) can be calculated by using numerical integration in the same way as (6).

Using $I(f_*; y_{\boldsymbol{x}}^{(m)})$ instead of $I(f_*; f_{\boldsymbol{x}}^{(m)})$ would be more natural when the observations are assumed to contain the observation noise with large variance σ^2_{noise}, but in practice, difference of these two formulations would not largely effect on performance of BO when σ^2_{noise} is small.

Note that the mutual information of parallel querying $I(f_*; f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}})$ can be replaced with the noisy observation $I(f_*; y_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}})$ by using same procedure.

D. Additional Information for Parallel Querying

D.1. Proof of Lemma 3.2

The first term of (8) is

$$\begin{aligned} \mathbb{E}_{\boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}_t} \left[H(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}) \right] &= \mathbb{E}_{\boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}_t} \left[\log \left(\sigma^{(m)}_{\boldsymbol{x} \mid \boldsymbol{f}_{\mathcal{Q}}} \sqrt{2\pi e} \right) \right] \\ &= \log \left(\sigma^{(m)}_{\boldsymbol{x} \mid \boldsymbol{f}_{\mathcal{Q}}} \sqrt{2\pi e} \right). \end{aligned}$$

The last equation holds since $\sigma^{(m)}_{\boldsymbol{x} \mid \boldsymbol{f}_{\mathcal{Q}}}$ does not depend on $\boldsymbol{f}_{\mathcal{Q}}$.

The second term of (8) is written as

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{f}_{\mathcal{Q}}, f_* \mid \mathcal{D}_t} \left[H(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}, f_{\boldsymbol{x}}^{(M)} \leq f_*) \right] \\ & = - \int \int p(\boldsymbol{f}_{\mathcal{Q}}, f_* \mid \mathcal{D}_t) \int p(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}, f_{\boldsymbol{x}}^{(M)} \leq f_*) \log p(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}, f_{\boldsymbol{x}}^{(M)} \leq f_*) \mathrm{d} f_{\boldsymbol{x}}^{(m)} \mathrm{d} \boldsymbol{f}_{\mathcal{Q}} \mathrm{d} f_*. \end{aligned} \tag{25}$$

For the conditional distribution

$$f_{\boldsymbol{x}}^{(M)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}, f_{\boldsymbol{x}}^{(m)} \sim \mathcal{N}(u_p(\boldsymbol{x}), s_p^2(\boldsymbol{x})),$$

the mean and the variance function can be written as

$$u_p(\boldsymbol{x}) = \frac{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(mM)} \left(f_{\boldsymbol{x}}^{(m)} - \mu_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)} \right)}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(m)}} + \mu_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)},$$

$$s_p^2(\boldsymbol{x}) = \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(M)} - \left(\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(mM)} \right)^2 \Big/ \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(m)}.$$

Then, from Bayes' theorem, we see

$$\begin{aligned} p(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}, f_{\boldsymbol{x}}^{(M)} \leq f_*) &= \frac{p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid \mathcal{D}_t, f_{\boldsymbol{x}}^{(m)}, \boldsymbol{f}_{\mathcal{Q}}) p(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}})}{p(f_{\boldsymbol{x}}^{(M)} \leq f_* \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}})} \\ &= \frac{\Phi\left(\frac{f_* - u_p(\boldsymbol{x})}{s_p(\boldsymbol{x})}\right) \phi\left(\frac{f_{\boldsymbol{x}}^{(m)} - \mu_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)}}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)}}\right)}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)} \Phi\left(\frac{f_* - \mu_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)}}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)}}\right)}. \end{aligned} \tag{26}$$

By defining

$$A \coloneqq \frac{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(mM)}}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(m)}},$$

we can re-write

$$\tilde{f}_* - u_p(\boldsymbol{x}) = \tilde{f}_* - A\tilde{f}_{\boldsymbol{x}}^{(m)},$$

and then, (26) is transformed into

$$\frac{\Phi\left(\frac{\tilde{f}_* - A\tilde{f}_{\boldsymbol{x}}^{(m)}}{s_p(\boldsymbol{x})}\right) \phi\left(\frac{\tilde{f}_{\boldsymbol{x}}^{(m)}}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)}}\right)}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)} \Phi\left(\frac{\tilde{f}_*}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)}}\right)} =: \eta(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}).$$

By further defining $h(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) \coloneqq \eta(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) \log \eta(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)})$, we simplify (25) as follows

$$-\int \int p(\boldsymbol{f}_{\mathcal{Q}}, f_* \mid \mathcal{D}_t) \int h(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) \mathrm{d}f_{\boldsymbol{x}}^{(m)} \mathrm{d}\boldsymbol{f}_{\mathcal{Q}} \mathrm{d}f_*. \tag{27}$$

This indicates that the most inner integrand can be shown as a function which only depends two random variables $\tilde{f}_*$ and $\tilde{f}_{\boldsymbol{x}}^{(m)}$. We change the variables of integration from $(f_*, f_{\boldsymbol{x}}^{(m)}, \boldsymbol{f}_{\mathcal{Q}}^\top)^\top$ to $(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}, \boldsymbol{f}_{\mathcal{Q}}^\top)^\top$.

$$\begin{aligned} \mathcal{J} &\coloneqq \begin{bmatrix} \frac{\partial \tilde{f}_*}{\partial f_*} & \frac{\partial \tilde{f}_*}{\partial f_{\boldsymbol{x}}^{(m)}} & \frac{\partial \tilde{f}_*}{\partial \boldsymbol{f}_{\mathcal{Q}}^\top} \\ \frac{\partial \tilde{f}_{\boldsymbol{x}}^{(m)}}{\partial f_*} & \frac{\partial \tilde{f}_{\boldsymbol{x}}^{(m)}}{\partial f_{\boldsymbol{x}}^{(m)}} & \frac{\partial \tilde{f}_{\boldsymbol{x}}^{(m)}}{\partial \boldsymbol{f}_{\mathcal{Q}}^\top} \\ \frac{\partial \boldsymbol{f}_{\mathcal{Q}}}{\partial f_*} & \frac{\partial \boldsymbol{f}_{\mathcal{Q}}}{\partial f_{\boldsymbol{x}}^{(m)}} & \frac{\partial \boldsymbol{f}_{\mathcal{Q}}}{\partial \boldsymbol{f}_{\mathcal{Q}}^\top} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{I}_2 & \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{Q}} \boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_{|\mathcal{Q}|} \end{bmatrix} \end{aligned}$$

where $\boldsymbol{I}_2$ and $\boldsymbol{I}_{|\mathcal{Q}|}$ are the identity matrices with size 2 and $|\mathcal{Q}|$, respectively. Note that determinant of $\mathcal{J}$ is $|\mathcal{J}| = 1$. Thus, by changing variables of integration and variables of the densities, (27) can be transformed into

$$\begin{aligned}
-\int\int p(\boldsymbol{f}_{\mathcal{Q}}, f_* \mid \mathcal{D}_t) \int h(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) \mathrm{d} f_{\boldsymbol{x}}^{(m)} \mathrm{d}\boldsymbol{f}_{\mathcal{Q}} \mathrm{d} f_* &= -\int\int p(\boldsymbol{f}_{\mathcal{Q}}, \tilde{f}_* \mid \mathcal{D}_t) \int h(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) \mathrm{d}\tilde{f}_{\boldsymbol{x}}^{(m)} \mathrm{d}\boldsymbol{f}_{\mathcal{Q}} \mathrm{d}\tilde{f}_* \\
&= -\int p(\tilde{f}_* \mid \mathcal{D}_t) \int h(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) \mathrm{d}\tilde{f}_{\boldsymbol{x}}^{(m)} \mathrm{d}\tilde{f}_* \\
&= -\mathbb{E}_{\tilde{f}_* \mid \mathcal{D}_t} \left[\int h(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) \mathrm{d}\tilde{f}_{\boldsymbol{x}}^{(m)} \right]. \qquad (28)
\end{aligned}$$

D.2. Analytical Calculation of Entropy for $m = M$

When $m = M$, the most inner integral in (25) can be further simplified because it is equal to the entropy of the truncated normal $p(f_{\boldsymbol{x}}^{(M)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}, f_{\boldsymbol{x}}^{(M)} \leq f_*)$, which is written as

$$\begin{aligned}
&-\int p(f_{\boldsymbol{x}}^{(M)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}, f_{\boldsymbol{x}}^{(M)} \leq f_*) \log p(f_{\boldsymbol{x}}^{(M)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}, f_{\boldsymbol{x}}^{(M)} \leq f_*) \mathrm{d} f_{\boldsymbol{x}}^{(M)} \\
&= \log \left(\sqrt{2\pi e} \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)} \Phi \left(\frac{\tilde{f}_*}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)}} \right) \right) - \frac{\tilde{f}_*}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)}} \frac{\phi\left(\frac{\tilde{f}_*}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)}} \right)}{2\Phi\left(\frac{\tilde{f}_*}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)}} \right)} \\
&=: \omega(\tilde{f}_*),
\end{aligned}$$

By using the same change of variables as (28), we obtain

$$-\int\int p(\boldsymbol{f}_{\mathcal{Q}}, f_* \mid \mathcal{D}_t) \omega(\tilde{f}_*) \mathrm{d}\boldsymbol{f}_{\mathcal{Q}} \mathrm{d} f_* = -\mathbb{E}_{\tilde{f}_* \mid \mathcal{D}_t} \left[\omega(\tilde{f}_*) \right].$$

D.3. Algorithm

As shown in Algorithm 2, the acquisition function maximization is performed when a worker becomes available. The sampling of $\tilde{f}_* \in \widetilde{\mathcal{F}}_*$ is performed through an RFM approximation of MF-GPR: $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, m)$. For the entropy calculation in line 19, one dimensional numerical integration is necessary for the integral in (14) when $m \neq M$, while the analytical formula is available when $m = M$ as shown in Appendix D.2.

D.4. Synchronous Parallelization

D.4.1. Single-fidelity Setting

In the main text, we focus on the asynchronous setting because of the diversity of sampling costs in MFBO. On the other hand, many parallel BO studies on the single-fidelity setting consider the synchronous setting (Figure 4). To our knowledge, a parallel extension of MES has not been studies even in the single-fidelity setting. Our approach is actually applicable to defining the single fidelity acquisition function. Although our main focus is in MFBO, we here show a counterpart of our multi-fidelity acquisition function in the single fidelity setting.

Suppose that we need to select q points written as $\mathcal{Q} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q\}$ for the single fidelity parallel BO. Unlike the asynchronous setting, q points is needed to be selected simultaneously. By setting $\boldsymbol{f}_{\mathcal{Q}} := (f_{\boldsymbol{x}_1}, \ldots, f_{\boldsymbol{x}_q})^\top$, a natural extension of MES for synchronous single-fidelity setting is written as

$$I(f_*; \boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}_t) := H(\boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}_t) - \mathbb{E}_{\boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}_t} \left[H(\boldsymbol{f}_{\mathcal{Q}} \mid \boldsymbol{f}_{\mathcal{Q}} \leq f_*, \mathcal{D}_t) \right]. \qquad (29)$$

Note that we impose the condition $\boldsymbol{f}_{\mathcal{Q}} \leq f_*$, indicating that all the elements of $\boldsymbol{f}_{\mathcal{Q}}$ is less than or equal to f_*, instead of $f_{\boldsymbol{x}} \leq f_*$ in the usual MES. The first term is the entropy of the q-dimensional Gaussian distribution which can be analytically calculated. The second term is the entropy of the multi-variate truncated normal distribution, for which we show analytical and approximate approaches to the computation.

Algorithm 2 Parallel MF-MES

```
function PARALLEL MF-MES($\mathcal{D}_0, M, \mathcal{X}, \{\lambda^{(m)}\}_{m=1}^{M}$)
    for $t = 0, \ldots, T$ do
        Wait for a worker to be available
        Generate $\widetilde{\mathcal{F}}_*$ from RFM
        $(\boldsymbol{x}_{t+1}, m_{t+1}) \leftarrow$ argmax$_{\boldsymbol{x} \in \mathcal{X}, m}$
          INFOGAIN($\boldsymbol{x}$, $m$, $\widetilde{\mathcal{F}}_*$, $\mathcal{D}_t$) / $\lambda^{(m)}$
        $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup (\boldsymbol{x}_{t+1}, y^{(m_{t+1})}(\boldsymbol{x}_{t+1}), m_{t+1})$
    end for
end function
function INFOGAIN($\boldsymbol{x}$, $m$, $\mathcal{F}_*$, $\mathcal{D}_t$)
    Calculate $\mu^{(m)}_{\boldsymbol{x}|\boldsymbol{f}_Q}$ and $\sigma^{(m)}_{\boldsymbol{x}|\boldsymbol{f}_Q}$
    Set $H_0 \leftarrow \log\left(\sigma_{\boldsymbol{x}|\boldsymbol{f}_Q}\sqrt{2\pi e}\right)$
    if $m \neq M$ then
        Calculate $\mu^{(M)}_{\boldsymbol{x}|\boldsymbol{f}_Q}, \sigma^{(m)}_{\boldsymbol{x}|\boldsymbol{f}_Q}$, and $\sigma^{2(mM)}_{\boldsymbol{x}|\boldsymbol{f}_Q}$
    end if
    Set $H_1 \leftarrow$ (14)
    Return $H_0 - H_1$
end function
```

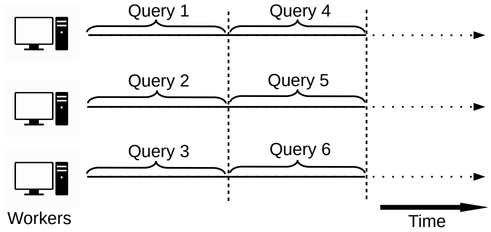


Figure 4: Synchronous setting in parallel BO.

First, we consider the analytical approach. The density $p(\boldsymbol{f}_Q \mid \mathcal{D}_t)$ is the predictive distribution of GPR, and we define $\boldsymbol{\mu}_Q$ and $\boldsymbol{\Sigma}_Q$ as the mean and covariance matrix, respectively. The truncated normal in the second term is defined through this density as follows

$$p(\boldsymbol{f}_Q \mid \boldsymbol{f}_Q \leq f_*, \mathcal{D}_t) = \begin{cases} p(\boldsymbol{f}_Q \mid \mathcal{D}_t)/Z, & \text{if } \boldsymbol{f}_Q \leq f_*, \\ 0, & \text{otherwise}, \end{cases} \tag{30}$$

where

$$Z \coloneqq \int_{\boldsymbol{f}_Q \leq f_*} p(\boldsymbol{f}_Q \mid \mathcal{D}_t) \mathrm{d}\boldsymbol{f}_Q.$$

We refer to the truncated normal (30) as $\mathrm{TN}(\boldsymbol{\mu}_Q^{\mathrm{TN}}, \boldsymbol{\Sigma}_Q^{\mathrm{TN}})$, where $\boldsymbol{\mu}_Q^{\mathrm{TN}}$ and $\boldsymbol{\Sigma}_Q^{\mathrm{TN}}$ are the mean and covariance matrix, respectively. Let $\mathbb{E}_{\rm TN}$ be the expectation by the density (30). Then, the entropy in the second term of (29) is re-written as

$$
\begin{aligned}
H[\boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}, \boldsymbol{f}_{\mathcal{Q}} \leq f_*] &= - \int_{\boldsymbol{f}_{\mathcal{Q}} \leq f_*} \frac{p(\boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}_t)}{Z} \log \frac{p(\boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}_t)}{Z} \mathrm{d}\boldsymbol{f}_{\mathcal{Q}} \\
&= -\mathbb{E}_{\rm TN} \left[\log \frac{p(\boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}_t)}{Z} \right] \\
&= -\mathbb{E}_{\rm TN} \left[\log p(\boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}_t) - \log Z \right] \\
&= -\mathbb{E}_{\rm TN} \left[\log p(\boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}_t) \right] + \log Z \\
&= -\mathbb{E}_{\rm TN} \left[-\frac{1}{2} \log |2\pi \boldsymbol{\Sigma}_{\mathcal{Q}}| - \frac{1}{2} (\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}})^\top \boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} (\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}}) \right] + \log Z \\
&= \frac{1}{2} \log |2\pi \boldsymbol{\Sigma}_{\mathcal{Q}}| + \frac{1}{2} \underbrace{\mathbb{E}_{\rm TN} \left[(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}})^\top \boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} (\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}}) \right]}_{=:B} + \log Z.
\end{aligned}
$$

By defining $\boldsymbol{d} = \boldsymbol{\mu}_{\mathcal{Q}}^{\rm TN} - \boldsymbol{\mu}_{\mathcal{Q}}$, we see

$$
\begin{aligned}
B &= \mathbb{E}_{\rm TN} \left[{\rm Tr}\left(\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} (\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}})(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}})^\top \right) \right] \\
&= {\rm Tr}\left(\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} \mathbb{E}_{\rm TN} \left[(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}})(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}})^\top \right] \right) \\
&= {\rm Tr}\left(\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} \mathbb{E}_{\rm TN} \left[(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}}^{\rm TN} + \boldsymbol{d})(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}}^{\rm TN} + \boldsymbol{d})^\top \right] \right) \\
&= {\rm Tr}\left(\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} \mathbb{E}_{\rm TN} \left[(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}}^{\rm TN})(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}}^{\rm TN})^\top + \boldsymbol{d}(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}}^{\rm TN})^\top + (\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}}^{\rm TN})\boldsymbol{d}^\top + \boldsymbol{d}\boldsymbol{d}^\top \right] \right).
\end{aligned}
$$

Since $\mathbb{E}_{\rm TN}[(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}})] = \boldsymbol{0}$, we further obtain

$$
\begin{aligned}
B &= {\rm Tr}\left(\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} \mathbb{E}_{\rm TN} \left[(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}}^{\rm TN})(\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}}^{\rm TN})^\top + \boldsymbol{d}\boldsymbol{d}^\top \right] \right) \\
&= {\rm Tr}\left(\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} (\boldsymbol{\Sigma}_{\mathcal{Q}}^{\rm TN} + \boldsymbol{d}\boldsymbol{d}^\top) \right)
\end{aligned}
$$

Therefore, we obtain

$$
H[\boldsymbol{f}_Q \mid \mathcal{D}, \boldsymbol{f}_Q \leq f_*] = \frac{1}{2} \Big(\log |2\pi \boldsymbol{\Sigma}_{\mathcal{Q}}| + {\rm Tr}\left(\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} (\boldsymbol{\Sigma}_{\mathcal{Q}}^{\rm TN} + \boldsymbol{d}\boldsymbol{d}^\top) \right) \Big) + \log Z.
$$

If Z, $\boldsymbol{\mu}_{\mathcal{Q}}^{\rm TN}$, and $\boldsymbol{\Sigma}_{\mathcal{Q}}^{\rm TN}$ are available, the above equation is easily calculated. The normalization term Z is the q-dimensional Gaussian CDF, for which a lot of fast computation algorithms have been proposed (e.g., Genz, 1992; Genton et al., 2017). A method proposed by (Genz, 1992) has been widely used, which requires $O(q^2)$ computations. For $\boldsymbol{\mu}_{\mathcal{Q}}^{\rm TN}$, and $\boldsymbol{\Sigma}_{\mathcal{Q}}^{\rm TN}$, G & Wilhelm (2012) shows analytical formulas which also depend on the multivariate Gaussian CDF. This needs q times computations of the $q-1$ dimensional CDF, and $q(q-1)$ times computations of the $q-2$ dimensional CDF.

To avoid many computations of $q-1$ dimensional CDF, we can introduce approximation of the entropy calculation or greedy selection of $\mathcal{Q}$. As a fast approximation, expectation propagation (EP) can be used to replace the truncated normal distribution with a Gaussian distribution, which makes the entropy calculation analytical. The similar technique is also used in (Hernández-Lobato et al., 2014). For the greedy strategy, we can choose a next point to add $\mathcal{Q}$ by maximizing $I(f_*; \boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}_t, \boldsymbol{f}_{\tilde{\mathcal{Q}}})$, where $\tilde{\mathcal{Q}}$ is a set of $(\boldsymbol{x}, m)$ already determined to be included in $\mathcal{Q}$. This information can be evaluated by the same way as we saw in the asynchronous setting (8) because the equation has the same form of conditional mutual information.

D.4.2. Multi-fidelity Setting

Combining the synchronous setting with multi-fidelity functions $m = 1, \ldots, M$ results in a combinatorial selection of $\mathcal{Q} = \{(\boldsymbol{x}_1, m_1), \ldots, (\boldsymbol{x}_q, m_q)\}$ because of the discreteness of the fidelity level m. When a simple greedy strategy is employed to select $\mathcal{Q}$, the procedure is reduced to the almost the same procedure as the synchronous single fidelity case described above. This indicates that we can avoid the q dimensional integral by using the technique shown in Section 3.2.

Table 1: Summary of possible settings. "FF-based" indicates the setting that the fidelity feature z is available, while "FF-free" does not assume it. Synchronous querying is denoted as 'sync', and asynchronous querying is denoted as 'asyn'.

	Fidelity	(S)equential/ (P)arallel	Our description	Note
Parallel BO	Single	P (sync)	Appendix D.4.1	-
	Single	P (asyn)	Special case of Parallel MF-MES	-
MFBO	Multiple (FF-based)	S	Appendix E	-
	Multiple (FF-free)	S	Section 3.1	-
Parallel MFBO	Multiple (FF-based)	P (sync)	Appendix E	(Wu & Frazier, 2017)
	Multiple (FF-based)	P (asyn)	Appendix E	No prior work
	Multiple (FF-free)	P (sync)	Appendix D.4.2	No prior work
	Multiple (FF-free)	P (asyn)	Section 3.2	No prior work

E. Incorporating Fidelity Feature

Our proposed method is applicable to the case that the fidelity is defined as a point of a fidelity feature (FF) space $\mathcal{Z}$ instead of the discrete fidelity level $1, \ldots, M$ (Kandasamy et al., 2017). Let $f_{\boldsymbol{x}}^{(\boldsymbol{z})}$ be the predictive distribution for the fidelity $\boldsymbol{z} \in \mathcal{Z}$. The goal is to solve $\max_{\boldsymbol{x} \in \mathcal{X}} f_{\boldsymbol{x}}^{(\boldsymbol{z}_*)}$, where $\boldsymbol{z}_* \in \mathcal{Z}$ is the highest fidelity to be optimized. For example, in the neural network hyper-parameter optimization, $\mathcal{Z}$ can be a two dimensional space defined by the number of training data and the number of training iterations.

In this case, our acquisition function (1) is extended to

$$a(\boldsymbol{x}, \boldsymbol{z}) \coloneqq I(f_*; f_{\boldsymbol{x}}^{(\boldsymbol{z})}) \; / \; \lambda^{(\boldsymbol{z})}, \tag{31}$$

where $f_* \coloneqq \max_{\boldsymbol{x} \in \mathcal{X}} f_{\boldsymbol{x}}^{(\boldsymbol{z}_*)}$ in this case, and $\lambda^{(\boldsymbol{z})}$ is known cost for $\boldsymbol{z} \in \mathcal{Z}$. As with (Kandasamy et al., 2017), we represent the output $f_{\boldsymbol{x}}^{(\boldsymbol{z})}$ as a Gaussian process on the direct product space $\mathcal{X} \times \mathcal{Z}$. Suppose that the observed training data set is written as $\mathcal{D}_n = \{(\boldsymbol{x}_i, y^{(\boldsymbol{z}_i)}(\boldsymbol{x}_i), \boldsymbol{z}_i)\}_{i=1}^n$, where $y^{(\boldsymbol{z}_i)}(\boldsymbol{x}_i)$ is an observation of $\boldsymbol{x}_i$ at the fidelity $\boldsymbol{z}_i$. A standard approach to defining a kernel on the joint space $\mathcal{X} \times \mathcal{Z}$ is to use the product form $k((\boldsymbol{x}_i, \boldsymbol{z}_i), (\boldsymbol{x}_j, \boldsymbol{z}_j)) = k_x(\boldsymbol{x}_i, \boldsymbol{x}_j) \; k_z(\boldsymbol{z}_i, \boldsymbol{z}_j)$, where $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel for the input space $\mathcal{X}$, and $k_z : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is a kernel for the fidelity space $\mathcal{Z}$. Based on this kernel, predictive distribution of GPR can be defined for any pair of $(\boldsymbol{x}, \boldsymbol{z})$, and thus the numerator of (31) can be calculated by using the same approach as $I(f_*; f_{\boldsymbol{x}}^{(m)})$ which we describe in Section 3.1.

Parallelization can also be considered in this FF-based case. For the asynchronous setting, the acquisition function is

$$a_{\mathrm{para}}(\boldsymbol{x}, \boldsymbol{z}) = I(f_*; f_{\boldsymbol{x}}^{(\boldsymbol{z})} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}) / \lambda^{(\boldsymbol{z})},$$

in which information gain is conditioned on the set of points currently under evaluation $\mathcal{Q} = \{(\boldsymbol{x}_1, m_1), \ldots, (\boldsymbol{x}_{q-1}, m_{q-1})\}$. As in the sequential case above, the calculation of this acquisition function is almost same as the discrete case in Section 3.2. For the synchronous case, the same discussion as Appendix D.4 also holds.

F. Summary of Settings in Sequential/Parallel MFBO

A possible combination of the single/multiple fidelity and sequential/parallel querying are summarized in Table 1. Our main focus is in FF-free MFBO, and FF-free parallel MFBO with asynchronous querying. In particular, for parallel MFBO, except for the FF-based synchronous querying, no prior works exist to our knowledge.

G. Additional Information of Empirical Evaluation

G.1. Other Experimental Settings

G.1.1. Settings of Methods

We trained the GPR model using normalized training observations (mean 0, and standard deviation 1), other than the GP-based synthetic function. Model hyper-parameters were optimized by marginal-likelihood at every 5 iterations. For the

GP-based synthetic function, we set the GPR hyper-parameters as parameters used for sampling the function. For the initial observations, we employed the Latin hypercube approach shown by (Huang et al., 2006). The number of initial training points $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ were set as follows:

- $5d$ and $4d$ for $m = 1$ and 2, respectively, if $M = 2$
- $6d$, $3d$ and $2d$ for $m = 1, 2$ and 3, respectively, if $M = 3$
- $10d$, $7d$ and $3d$ for $m = 1, 2$ and 3, respectively, in the material dataset

We used the Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sum_{i=1}^{d} (\boldsymbol{x}_i - \boldsymbol{x}_i')^2/(2\ell_i^2))$ for all kernels. The length scale parameter ℓ_d was optimized through marginal-likelihood in the following interval:

- $\ell_d \in [\mathrm{Domain\ size}/10, \mathrm{Domain\ size} \times 10]$ for the GP-based synthetic function and the benchmark functions, here $\mathrm{Domain\ size}$ is the difference between the maximum and the minimum of the input domain in each dimension. The input domain of each function is shown in Appendix G.1.2.
- $\ell_d \in [10^{-3}, 10^{-1}]$ for the material dateset
- The task kernel in BOCA: $\ell_d \in [2, (M-1) \times 10]$ for benchmark functions, and $\ell_d \in [10, 10^3]$ for the material dataset

The noise parameter of GPR was fixed as $\sigma_{\mathrm{noise}}^2 = 10^{-6}$. The number of kernels in SLFM was $C = 2$. The hyper-parameters in covariance among different output dimension were also optimized through marginal-likelihood in the following interval:

- $w_{c1} \in [\sqrt{0.75}, 1]$ for $c = 1, 2$
- $w_{c2} \in [-\sqrt{0.25}, \sqrt{0.25}]$ for $c = 1, 2$
- $\kappa_{cm} \in [10^{-3}, 10^{-1}]$ for $c = 1, 2$ and $m = 1, \dots, M$

The number of basis D in RFM was 1000, which was used by MF-MES, MF-PES, MES-LP, and AsyTS. The number of samplings for f^* in MES and PES was 10.

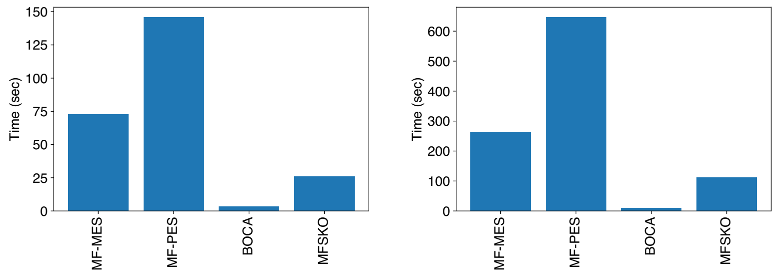
For all compared methods, including BOCA, MFSKO, local penalization in MES-LP, GP-UCB-PE, and AsyTS, we followed the settings of hyper-parameters in their original papers.

G.1.2. Details of Benchmark Datasets

GP-based Synthetic functions We used RFM for SLFM described in Appendix A.2. The input dimension is $d = 3$ and the domain is $x_i \in [0, 1]$. The parameters are $C = 1, \boldsymbol{w} = (0.9, 0.9)^\top, \boldsymbol{\kappa} = (0.1, 0.1)^\top$, and $\ell_i = 0.1$ for $i = 1, 2, 3$.

Styblinski-Tang function

$$
\begin{aligned}
f^{(1)} &= \frac{1}{2} \sum_{i=1}^{2} (0.9x_i^4 - 15x_i^2 + 6x_i), \\
f^{(2)} &= \frac{1}{2} \sum_{i=1}^{2} (x_i^4 - 16x_i^2 + 5x_i), \\
x_i &\in [-5, 5], i = 1, 2
\end{aligned}
$$



(a) Styblinski-Tang

(b) HartMann6

Figure 5: Computational time for acquisition function maximization.

HartMann6 function

$$
\begin{aligned}
f^{(1)} &= -\sum_{i=1}^{4} (\alpha_i - 0.2) \exp\left(-\sum_{j=1}^{6} A_{ij} (x_j - P_{ij})^2 \right), \\
f^{(2)} &= -\sum_{i=1}^{4} (\alpha_i - 0.1) \exp\left(-\sum_{j=1}^{6} A_{ij} (x_j - P_{ij})^2 \right) \\
f^{(3)} &= -\sum_{i=1}^{4} \alpha_i \exp\left(-\sum_{j=1}^{6} A_{ij} (x_j - P_{ij})^2 \right) \\
\boldsymbol{\alpha} &= [1.0, 1.2, 3.0, 3.2]^\top \\
\boldsymbol{A} &= \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix} \\
\boldsymbol{P} &= 10^{-4} \begin{pmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{pmatrix} \\
\boldsymbol{x}_j &\in [0,1], j = 1, \ldots, 6
\end{aligned}
$$

Materials Data As an example of practical application, we applied our method to the parameter optimization of computational simulation model in materials science. There is a computational model (Tsukada et al., 2014) that predicts equilibrium shape of precipitates in the α-Mg phase when material parameters are given. We estimate two material parameters (lattice mismatch and interface energy between the α-Mg and precipitate phases) from experimental data on precipitate shape measured by transmission electron microscopy (TEM) (Bhattacharjee et al., 2013). The objective function is the discrepancy between precipitate shape predicted by the computational model and one measured by TEM.

G.2. Measuring Computational Time of Acquisition Functions

We measured the computational time for the maximization of the acquisition functions. We assume that the predictive distribution of the GPR model is already obtained, because it is almost common for all the methods. The training dataset is created by the initialization process in our experiment described in Appendix G.1.

Figure 5 shows the results on three benchmark dataset, used in the main text. BOCA and MFSKO are relatively easy to compute because they are based on UCB and EI, respectively. Their acquisition function is simple, but difficult to incorporate global utility of the candidate without tuning parameters as we discuss in the main text. MF-MES was much faster than MF-PES. We emphasize that MF-PES employs the approximation based on EP to accelerate the computation, unlike our MF-MES which is almost analytical. This indicates that MF-MES provides more reliable entropy computation with smaller amount of computations than MF-PES.