Trusted-Maximizers Entropy Search for Efficient Bayesian Optimization (Supplementary material)

Quoc Phong Nguyen[*1] **Zhaoxuan Wu**[*3,4] **Bryan Kian Hsiang Low**[1] **Patrick Jaillet**[2]

[1]Department of Computer Science, National University of Singapore, Republic of Singapore
[2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA
[3]Institute of Data Science, National University of Singapore, Republic of Singapore
[4]NUSGS Integrative Sciences and Engineering Programme, National University of Singapore, Republic of Singapore

A EVALUATION OF $p(\mathbf{x}^\star|\mathbf{y}_\mathcal{D})$

The probability $p(\mathbf{x}^\star|\mathbf{y}_\mathcal{D})$ can be expressed as:

$$\begin{aligned}p(\mathbf{x}^\star|\mathbf{y}_\mathcal{D}) &= p(f(\mathbf{x}^\star) \geq f(\mathbf{x})\ \forall \mathbf{x} \in \mathcal{X}^\star|\mathbf{y}_\mathcal{D})\\ &= p(\mathbf{J}_{\mathbf{x}^\star}\mathbf{f}_{\mathcal{X}^\star} \leq 0|\mathbf{y}_\mathcal{D})\end{aligned}$$

where $\mathbf{J}_{\mathbf{x}^\star}$ is a matrix of size $|\mathcal{X}^\star| \times |\mathcal{X}^\star|$ with 0 entries except for $\mathbf{J}_{ii} = 1, \mathbf{J}_{ij} = -1$ if $i \neq j$ where j is the index of $\mathbf{x}^\star$ in $\mathcal{X}^\star$. As $p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D})$ is the p.d.f. of a multivariate Gaussian (1), so is $p(\mathbf{J}_{\mathbf{x}^\star}\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D})$. Hence, $p(\mathbf{J}_{\mathbf{x}^\star}\mathbf{f}_{\mathcal{X}^\star} \leq 0|\mathbf{y}_\mathcal{D})$ can be computed efficiently.

B IMPORTANCE SAMPLING FROM $p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star)$

Let $\mathbf{f}_{\setminus\star} \triangleq (f(\mathbf{x}'))^\top_{\mathbf{x}' \in \mathcal{X}^\star \setminus \{\mathbf{x}^\star\}}$ denote the function values at $\mathcal{X}^\star \setminus \{\mathbf{x}^\star\}$. We have $p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star) \propto p(\mathbf{f}_{\mathcal{X}^\star}, \mathbf{x}^\star|\mathbf{y}_\mathcal{D})$ which equals

$$p(\mathbf{x}^\star|\mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})p(\mathbf{f}_{\setminus\star}|\mathbf{y}_\mathcal{D})p(f(\mathbf{x}^\star)|\mathbf{x}^\star, \mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})\ .$$

We can first draw samples of $\mathbf{f}_{\mathcal{X}^\star}$ from the p.d.f. $p(\mathbf{f}_{\setminus\star}|\mathbf{y}_\mathcal{D})p(f(\mathbf{x}^\star)|\mathbf{x}^\star, \mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})$, then weight these samples with $p(\mathbf{x}^\star|\mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})$. Sampling from $p(\mathbf{f}_{\setminus\star}|\mathbf{y}_\mathcal{D})p(f(\mathbf{x}^\star)|\mathbf{x}^\star, \mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})$ is a 2-step process:

1. Drawing a sample of $\mathbf{f}_{\setminus\star}$ from $p(\mathbf{f}_{\setminus\star}|\mathbf{y}_\mathcal{D})$ which is the p.d.f. of a multivariate Gaussian distribution.
2. Given a sample of $\mathbf{f}_{\setminus\star}$, drawing a sample of $f(\mathbf{x}^\star)$ from $p(f(\mathbf{x}^\star)|\mathbf{x}^\star, \mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})$ which is a lower-truncated Gaussian distribution (truncation of lower tail at $f^+ \triangleq \max_{\mathbf{x}' \in \mathcal{X}^\star \setminus \{\mathbf{x}^\star\}} f(\mathbf{x}')$).

The weight of a sample $\mathbf{f}_{\mathcal{X}^\star}$ is $p(\mathbf{x}^\star|\mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})$ which can be computed efficiently.

$$\begin{aligned}&p(\mathbf{x}^\star|\mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})\\ &= \int p(\mathbf{x}^\star, f(\mathbf{x}^\star)|\mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})\ \mathrm{d}f(\mathbf{x}^\star)\\ &= \int p(\mathbf{x}^\star|\mathbf{f}_{\mathcal{X}^\star}, \mathbf{y}_\mathcal{D})p(f(\mathbf{x}^\star)|\mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})\ \mathrm{d}f(\mathbf{x}^\star)\\ &= \int \mathbf{I}_{f(\mathbf{x}^\star) \geq f^+} p(f(\mathbf{x}^\star)|\mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})\ \mathrm{d}f(\mathbf{x}^\star)\\ &= 1 - \Phi_{p(f(\mathbf{x}^\star)|\mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})}\left(f^+\right)\end{aligned}$$

where $\Phi_{p(f(\mathbf{x}^\star)|\mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})}\left(f^+\right)$ is the c.d.f. of the GP predictive belief $p(f(\mathbf{x}^\star)|\mathbf{f}_{\setminus\star}, \mathbf{y}_\mathcal{D})$ evaluated at f^+, and $\mathbf{I}_{f(\mathbf{x}^\star) \geq f^+}$ is an indicator function such that it is 1 if $f(\mathbf{x}^\star) \geq f^+$ and 0 otherwise.

C CLOSED-FORM EXPRESSION OF $p(y_\mathbf{x}|\mathbf{f}_{\mathcal{X}^\star}, \mathbf{y}_\mathcal{D})$

At a BO iteration, let $\mathbf{z} \triangleq [\mathbf{f}_{\mathcal{X}^\star}; \mathbf{y}_\mathcal{D}]^\top$ be a column vector where the first $|\mathcal{X}^\star|$ elements are $\mathbf{f}_{\mathcal{X}^\star}$ and the last $|\mathcal{D}|$ elements are observations $\mathbf{y}_\mathcal{D}$. Let $\mathbf{t}$ be a column vector where the first $|\mathcal{X}^\star|$ inputs are $\mathcal{X}^\star$ and the last $|\mathcal{D}|$ inputs are observed inputs $\mathcal{D}$. We have the expression: $p(y_\mathbf{x}|\mathbf{f}_{\mathcal{X}^\star}, \mathbf{y}_\mathcal{D}) = \mathcal{N}(y_\mathbf{x}; \mu^+_\mathbf{x}, \sigma^2_{\mathbf{x}|\mathcal{X}^\star} + \sigma^2_n)$ specified by

$$\mu^+_\mathbf{x} \triangleq \mathbf{k}_{\mathbf{x}\mathbf{t}}\left(\mathbf{K}_{\mathbf{t}\mathbf{t}} + \sigma^2_n\tilde{\mathbf{I}}\right)^{-1}\mathbf{z} \tag{1}$$

$$\sigma^2_{\mathbf{x}|\mathcal{X}^\star} \triangleq k_{\mathbf{x}\mathbf{x}} - \mathbf{k}_{\mathbf{x}\mathbf{t}}\left(\mathbf{K}_{\mathbf{t}\mathbf{t}} + \sigma^2_n\tilde{\mathbf{I}}\right)^{-1}\mathbf{k}_{\mathbf{t}\mathbf{x}} \tag{2}$$

where $\tilde{\mathbf{I}}$ is a matrix of size $(|\mathcal{X}^\star| + |\mathcal{D}|) \times (|\mathcal{X}^\star| + |\mathcal{D}|)$ such that $\tilde{\mathbf{I}}_{ij} = 1$ if $i = j > |\mathcal{X}^\star|$ and 0 otherwise. Note that $\mu^+_\mathbf{x}$ and $\sigma_{\mathbf{x}|\mathcal{X}^\star}$ are the mean and standard deviation of $p(f(\mathbf{x})|\mathbf{f}_{\mathcal{X}^\star}, \mathbf{y}_\mathcal{D})$, which are used in Section E.

[*] Equal contribution

D EXPECTATION PROPAGATION APPROXIMATION FOR $p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star)$

To approximate $p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star)$, it is expressed as

$$\begin{aligned}
p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star) &\propto p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}) p(\mathbf{x}^\star|\mathbf{f}_{\mathcal{X}^\star}, \mathbf{y}_\mathcal{D}) \\
&= p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}) \prod_{\mathbf{x}' \in \mathcal{X}^\star \setminus \{\mathbf{x}^\star\}} \mathbf{I}_{f(\mathbf{x}^\star) \geq f(\mathbf{x}')} \\
&= p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}) \prod_{\mathbf{x}' \in \mathcal{X}^\star \setminus \{\mathbf{x}^\star\}} \mathbf{I}_{\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star} \geq 0}
\end{aligned}$$

where $\mathbf{c}'_\mathbf{x}$ is a column vector of length $|\mathcal{X}^\star|$ with 0 entries except for the i-th entry of value -1, the j-th entry of value 1 where i and j are the indices of $\mathbf{x}'$ and $\mathbf{x}^\star$ in $\mathcal{X}^\star$, respectively; $\mathbf{I}$ is the indicator function.

The p.d.f. $p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star)$ is approximated with a Gaussian distribution by EP, i.e., $p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star) \approx \mathcal{N}(\mathbf{f}_{\mathcal{X}^\star}; \boldsymbol{\mu}_{\text{ep}}, \boldsymbol{\Sigma}_{\text{ep}}) \triangleq q_{\text{ep}}(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star)$. As in PPES Shah and Ghahramani [2015], to construct this Gaussian approximation, each indicator term (involved $\mathbf{I}$) is approximated with a univariate scaled Gaussian p.d.f.:

$$\begin{aligned}
q_{\text{ep}}(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star) &\triangleq \mathcal{N}(\mathbf{f}_{\mathcal{X}^\star}; \boldsymbol{\mu}_{\text{ep}}, \boldsymbol{\Sigma}_{\text{ep}}) \\
&= \mathcal{N}(\mathbf{f}_{\mathcal{X}^\star}; \tilde{\mathbf{m}}, \tilde{\mathbf{K}}) \prod_{\mathbf{x}' \in \mathcal{X}^\star \setminus \{\mathbf{x}^\star\}} \tilde{Z}_{\mathbf{x}'} \mathcal{N}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}; \tilde{\mu}_{\mathbf{x}'}, \tilde{\tau}_{\mathbf{x}'})
\end{aligned}$$

where $\mathcal{N}(\mathbf{f}_{\mathcal{X}^\star}; \tilde{\mathbf{m}}, \tilde{\mathbf{K}})$ denotes the GP predictive belief of $\mathbf{f}_{\mathcal{X}^\star}$ given $\mathbf{y}_\mathcal{D}$, the scale factor $\tilde{Z}_{\mathbf{x}'}$ and the variance $\tilde{\tau}_{\mathbf{x}'}$ are positive, $\tilde{\mu}_{\mathbf{x}'} \in \mathbb{R}$. The parameters $\{\tilde{Z}_{\mathbf{x}'}, \tilde{\mu}_{\mathbf{x}'}, \tilde{\tau}_{\mathbf{x}'}\}_{\mathbf{x}' \in \mathcal{X}^\star \setminus \{\mathbf{x}^\star\}}$ are call the *site parameters*, which are optimized such that the Kullback-Leibler divergence of $q_{\text{ep}}(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star) / \int \tilde{p}(\mathbf{f}_{\mathcal{X}^\star})\ \mathrm{d}\mathbf{f}_{\mathcal{X}^\star}$ from $p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star)$ is minimized, i.e., their means and covariance matrices match. As $\{\tilde{Z}_{\mathbf{x}'}\}_{\mathbf{x}' \in \mathcal{X}^\star \setminus \{\mathbf{x}^\star\}}$ are independent from $\mathbf{f}_{\mathcal{X}^\star}$, we only optimize $\{\tilde{\mu}_{\mathbf{x}'}, \tilde{\tau}_{\mathbf{x}'}\}_{\mathbf{x}' \in \mathcal{X}^\star \setminus \{\mathbf{x}^\star\}}$.

As the product of Gaussian p.d.f. leads to a Gaussian p.d.f., we have

$$\boldsymbol{\Sigma}_{\text{ep}} = \left(\tilde{\mathbf{K}}^{-1} + \sum_{\mathbf{x}' \in \mathcal{X}^\star \setminus \{\mathbf{x}^\star\}} \frac{1}{\tilde{\tau}_{\mathbf{x}'}} \mathbf{c}_{\mathbf{x}'} \mathbf{c}_{\mathbf{x}'}^\top \right)^{-1} \quad (3)$$

$$\boldsymbol{\mu}_{\text{ep}} = \boldsymbol{\Sigma}_{\text{ep}} \left(\tilde{\mathbf{K}}^{-1} \tilde{\mathbf{m}} + \sum_{\mathbf{x}' \in \mathcal{X}^\star \setminus \{\mathbf{x}^\star\}} \frac{\tilde{\mu}_{\mathbf{x}'}}{\tilde{\tau}_{\mathbf{x}'}} \mathbf{c}_{\mathbf{x}'} \right) . \quad (4)$$

To update the site parameters, we first compute the *cavity distributions*

$$p_{\setminus \mathbf{x}'}(\mathbf{f}_{\mathcal{X}^\star}) \triangleq \frac{q_{\text{ep}}(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_\mathcal{D}, \mathbf{x}^\star)}{\tilde{Z}_{\mathbf{x}'} \mathcal{N}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}; \tilde{\mu}_{\mathbf{x}'}, \tilde{\tau}_{\mathbf{x}'})} .$$

Note that in the next step, we would like to approximate the distribution of $\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}$, so we only require the cavity distribution of the r.v. $\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}$ which is denoted as:

$$p_{\setminus \mathbf{x}'}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}) \triangleq \frac{\tilde{p}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star})}{\tilde{Z}_{\mathbf{x}'} \mathcal{N}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}; \tilde{\mu}_{\mathbf{x}'}, \tilde{\tau}_{\mathbf{x}'})}$$

where $\tilde{p}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}) = \mathcal{N}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}; \mathbf{c}_{\mathbf{x}'}^\top \boldsymbol{\mu}_{\text{ep}}, \mathbf{c}_{\mathbf{x}'}^\top \boldsymbol{\Sigma}_{\text{ep}} \mathbf{c}_{\mathbf{x}'})$. Let the Gaussian mean and variance of $p_{\setminus \mathbf{x}'}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star})$ are $\mu_{\setminus \mathbf{x}'}$ and $\tau_{\setminus \mathbf{x}'}$, respectively, we have

$$\begin{aligned}
\tau_{\setminus \mathbf{x}'} &= \left(\left(\mathbf{c}_{\mathbf{x}'}^\top \boldsymbol{\Sigma}_{\text{ep}} \mathbf{c}_{\mathbf{x}'} \right)^{-1} - \tilde{\tau}_{\mathbf{x}'}^{-1} \right)^{-1} \\
\mu_{\setminus \mathbf{x}'} &= \tau_{\setminus \mathbf{x}'} \left(\frac{\mathbf{c}_{\mathbf{x}'}^\top \boldsymbol{\mu}_{\text{ep}}}{\mathbf{c}_{\mathbf{x}'}^\top \boldsymbol{\Sigma}_{\text{ep}} \mathbf{c}_{\mathbf{x}'}} - \frac{\tilde{\mu}_{\mathbf{x}'}}{\tilde{\tau}_{\mathbf{x}'}} \right) .
\end{aligned}$$

Next, the *projection step* of EP is to do moment matching of $\tilde{Z}_{\mathbf{x}'} \mathcal{N}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}; \tilde{\mu}_{\mathbf{x}'}, \tilde{\tau}_{\mathbf{x}'}) p_{\setminus \mathbf{x}'}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star})$ with $\mathbf{I}_{\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star} \geq 0}\, p_{\setminus \mathbf{x}'}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star})$. We use the derivatives of the zeroth moment to compute moments of $\mathbf{I}_{\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star} \geq 0}\, p_{\setminus \mathbf{x}'}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star})$, denoted as $\hat{\mu}_\mathbf{x}$ (mean) and $\hat{\tau}_\mathbf{x}$ (variance). The zeroth moment is computed as:

$$\begin{aligned}
\hat{Z}_{\mathbf{x}'} &\triangleq \int \mathbf{I}_{\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star} \geq 0}\, p_{\setminus \mathbf{x}'}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star})\ \mathrm{d}\left(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}\right) \\
&= \int \mathbf{I}_{\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star} \geq 0}\, \mathcal{N}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}; \mu_{\setminus \mathbf{x}'}, \tau_{\setminus \mathbf{x}'})\ \mathrm{d}\left(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}\right) \\
&= \Phi(\beta_{\mathbf{x}'}) \text{ where } \beta_{\mathbf{x}'} \triangleq \frac{\mu_{\setminus \mathbf{x}'}}{\sqrt{\tau_{\setminus \mathbf{x}'}}}
\end{aligned}$$

$$\frac{\partial \hat{Z}_{\mathbf{x}'}}{\partial \mu_{\setminus \mathbf{x}'}} = \frac{\Phi(\beta_{\mathbf{x}'})}{\partial \mu_{\setminus \mathbf{x}'}} = \frac{\phi(\beta_{\mathbf{x}'})}{\sqrt{\tau_{\setminus \mathbf{x}'}}} \quad (5)$$

$$\frac{\partial \hat{Z}_{\mathbf{x}'}}{\partial \tau_{\setminus \mathbf{x}'}} = \frac{\Phi(\beta_{\mathbf{x}'})}{\partial \tau_{\setminus \mathbf{x}'}} = -\frac{1}{2} \frac{\mu_{\setminus \mathbf{x}'}}{\sqrt{\tau_{\setminus \mathbf{x}'}^3}} \phi(\beta_{\mathbf{x}'}) \quad (6)$$

where ϕ and Φ are the p.d.f. and c.d.f. of the standard Gaussian distribution, respectively. On the other hand, we can express the derivatives of $\hat{Z}_{\mathbf{x}'}$ as:

$$\begin{aligned}
\frac{\partial \hat{Z}_{\mathbf{x}'}}{\partial \mu_{\setminus \mathbf{x}'}} &= \frac{\partial}{\partial \mu_{\setminus \mathbf{x}'}} \int \mathbf{I}_{\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star} \geq 0}\, \mathcal{N}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}; \mu_{\setminus \mathbf{x}'}, \tau_{\setminus \mathbf{x}'})\ \mathrm{d}\left(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}\right) \\
&= \hat{Z}_{\mathbf{x}'} \frac{\hat{\mu}_{\mathbf{x}'}}{\tau_{\setminus \mathbf{x}'}} - \hat{Z}_{\mathbf{x}'} \frac{\mu_{\setminus \mathbf{x}'}}{\tau_{\setminus \mathbf{x}'}} \\
&= \Phi(\beta_{\mathbf{x}'}) \frac{\hat{\mu}_{\mathbf{x}'}}{\tau_{\setminus \mathbf{x}'}} - \Phi(\beta_{\mathbf{x}'}) \frac{\mu_{\setminus \mathbf{x}'}}{\tau_{\setminus \mathbf{x}'}} \quad (7)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \hat{Z}_{\mathbf{x}'}}{\partial \tau_{\setminus \mathbf{x}'}} &= \frac{\partial}{\partial \tau_{\setminus \mathbf{x}'}} \int \mathbf{I}_{\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star} \geq 0}\, \mathcal{N}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}; \mu_{\setminus \mathbf{x}'}, \tau_{\setminus \mathbf{x}'})\ \mathrm{d}\left(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}\right) \\
&= \frac{1}{2} \hat{Z}_{\mathbf{x}'} \frac{\hat{\mu}_{\mathbf{x}'}^2 + \hat{\tau}_{\mathbf{x}'}}{\tau_{\setminus \mathbf{x}'}^2} - \hat{Z}_{\mathbf{x}'} \frac{\mu_{\setminus \mathbf{x}'}}{\tau_{\setminus \mathbf{x}'}^2} \hat{\mu}_{\mathbf{x}'} + \frac{1}{2} \hat{Z}_{\mathbf{x}'} \frac{\mu_{\setminus \mathbf{x}'}^2}{\tau_{\setminus \mathbf{x}'}^2} - \frac{1}{2} \frac{\hat{Z}_{\mathbf{x}'}}{\tau_{\setminus \mathbf{x}'}} \\
&= \frac{1}{2} \Phi(\beta_{\mathbf{x}'}) \frac{\hat{\mu}_{\mathbf{x}'}^2 + \hat{\tau}_{\mathbf{x}'}}{\tau_{\setminus \mathbf{x}'}^2} - \Phi(\beta_{\mathbf{x}'}) \frac{\mu_{\setminus \mathbf{x}'}}{\tau_{\setminus \mathbf{x}'}^2} \hat{\mu}_{\mathbf{x}'} \\
&\quad + \frac{1}{2} \Phi(\beta_{\mathbf{x}'}) \frac{\mu_{\setminus \mathbf{x}'}^2}{\tau_{\setminus \mathbf{x}'}^2} - \frac{1}{2} \frac{\Phi(\beta_{\mathbf{x}'})}{\tau_{\setminus \mathbf{x}'}} . \quad (8)
\end{aligned}$$

Equating (5) with (7), and (6) with (8), we have

$$\begin{aligned}
\hat{\mu}_{\mathbf{x}'} &= \sqrt{\tau_{\setminus \mathbf{x}'}} \frac{\phi(\beta_{\mathbf{x}'})}{\Phi(\beta_{\mathbf{x}'})} + \mu_{\setminus \mathbf{x}'} \\
\hat{\tau}_{\mathbf{x}'} &= -\mu_{\setminus \mathbf{x}'} \sqrt{\tau_{\setminus \mathbf{x}'}} \frac{\phi(\beta_{\mathbf{x}'})}{\Phi(\beta_{\mathbf{x}'})} - \hat{\mu}_{\mathbf{x}'}^2 + 2\mu_{\setminus \mathbf{x}'} \hat{\mu}_{\mathbf{x}'} - \mu_{\setminus \mathbf{x}'}^2 + \tau_{\setminus \mathbf{x}'} .
\end{aligned}$$

Then, we update the site parameters to get the moments of $\tilde{Z}_{\mathbf{x}'}\mathcal{N}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}; \tilde{\mu}_{\mathbf{x}'}, \tilde{\tau}_{\mathbf{x}'}) = \mathbf{I}_{\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star} \geq 0}\, p_{\setminus \mathbf{x}'}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star}) / p_{\setminus \mathbf{x}'}(\mathbf{c}_{\mathbf{x}'}^\top \mathbf{f}_{\mathcal{X}^\star})$ as

$$\begin{aligned}
\tilde{\tau}_{\mathbf{x}'} &= \left(\hat{\tau}_{\mathbf{x}'}^{-1} - \tau_{\setminus \mathbf{x}'}^{-1}\right)^{-1} \\
\tilde{\mu}_{\mathbf{x}'} &= \tilde{\tau}_{\mathbf{x}'} \left(\hat{\tau}_{\mathbf{x}'}^{-1}\hat{\mu}_{\mathbf{x}'} - \tau_{\setminus \mathbf{x}'}^{-1}\mu_{\setminus \mathbf{x}'}\right) .
\end{aligned}$$

Finally, we update the parameters $\boldsymbol{\mu}_{\text{ep}}$ and $\boldsymbol{\Sigma}_{\text{ep}}$ by (3) and (4). The process is repeated for all $\mathbf{x}' \in \mathcal{X}^\star \setminus \{\mathbf{x}^\star\}$ until convergence.

E CLOSED-FORM EXPRESSION OF $p(f(\mathbf{x})|\mathbf{y}_{\mathcal{D}}, \mathbf{x}^\star)$

We can express the predictive posterior distribution $p(f(\mathbf{x})|\mathbf{y}_{\mathcal{D}}, \mathbf{x}^\star)$ as:

$$\begin{aligned}
&p(f(\mathbf{x})|\mathbf{y}_{\mathcal{D}}, \mathbf{x}^\star) \\
&= \int p(f(\mathbf{x})|\mathbf{f}_{\mathcal{X}^\star}, \mathbf{y}_{\mathcal{D}}, \mathbf{x}^\star) p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_{\mathcal{D}}, \mathbf{x}^\star)\, \mathrm{d}\mathbf{f}_{\mathcal{X}^\star} \\
&= \int p(f(\mathbf{x})|\mathbf{f}_{\mathcal{X}^\star}, \mathbf{y}_{\mathcal{D}}) p(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_{\mathcal{D}}, \mathbf{x}^\star)\, \mathrm{d}\mathbf{f}_{\mathcal{X}^\star} \\
&\approx \int p(f(\mathbf{x})|\mathbf{f}_{\mathcal{X}^\star}, \mathbf{y}_{\mathcal{D}}) q_{\text{ep}}(\mathbf{f}_{\mathcal{X}^\star}|\mathbf{y}_{\mathcal{D}}, \mathbf{x}^\star)\, \mathrm{d}\mathbf{f}_{\mathcal{X}^\star} \\
&= \int \mathcal{N}(f(\mathbf{x}); \mu_{\mathbf{x}}^+, \sigma_{\mathbf{x}|\mathcal{X}^\star}^2) \mathcal{N}(\mathbf{f}_{\mathcal{X}^\star}; \boldsymbol{\mu}_{\text{ep}}, \boldsymbol{\Sigma}_{\text{ep}})\, \mathrm{d}\mathbf{f}_{\mathcal{X}^\star} .
\end{aligned}$$

where $p(f(\mathbf{x})|\mathbf{f}_{\mathcal{X}^\star}, \mathbf{y}_{\mathcal{D}}) \triangleq \mathcal{N}(f(\mathbf{x}); \mu_{\mathbf{x}}^+, \sigma_{\mathbf{x}|\mathcal{X}^\star}^2)$; $\mu_{\mathbf{x}}^+$ and $\sigma_{\mathbf{x}|\mathcal{X}^\star}^2$ are defined in Eq. 1 and Eq. 2, respectively.

Let $\mathbf{r} \triangleq \mathbf{K}_+^{-1}\mathbf{k}_{\mathbf{t}\mathbf{x}}$, $\mathbf{a} \triangleq \mathbf{r}_{\mathcal{X}^\star}$, and $b \triangleq \mathbf{r}_{\mathcal{D}}^\top \mathbf{y}_{\mathcal{D}}$ where $\mathbf{r}_{\mathcal{X}^\star}$ and $\mathbf{r}_{\mathcal{D}}$ are the first $|\mathcal{X}^\star|$ and the last $|\mathcal{D}|$ elements of the column vector $\mathbf{r}$. Then, we have

$$\begin{aligned}
\mu_{\mathbf{x}}^+ &= \mathbf{a}^\top \mathbf{f}_{\mathcal{X}^\star} + b \\
p(f(\mathbf{x})|\mathbf{y}_{\mathcal{D}}, \mathbf{x}^\star) &\approx \int \mathcal{N}(f(\mathbf{x}); \mathbf{a}^\top \mathbf{f}_{\mathcal{X}^\star} + b, \sigma_{\mathbf{x}|\mathcal{X}^\star}^2) \\
&\qquad \times \mathcal{N}(\mathbf{f}_{\mathcal{X}^\star}; \boldsymbol{\mu}_{\text{ep}}, \boldsymbol{\Sigma}_{\text{ep}})\, \mathrm{d}\mathbf{f}_{\mathcal{X}^\star} \\
&= \mathcal{N}(f(\mathbf{x}); \mathbf{a}^\top \boldsymbol{\mu}_{\text{ep}} + b, \sigma_{\mathbf{x}|\mathcal{X}^\star}^2 + \mathbf{a}^\top \boldsymbol{\Sigma}_{\text{ep}}\mathbf{a}) .
\end{aligned}$$

Hence,

$$\begin{aligned}
p(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, \mathbf{x}^\star) &\approx \mathcal{N}(f(\mathbf{x}); \mathbf{a}^\top \boldsymbol{\mu}_{\text{ep}} + b, \sigma_{\mathbf{x}|\mathcal{X}^\star}^2 + \mathbf{a}^\top \boldsymbol{\Sigma}_{\text{ep}}\mathbf{a} + \sigma_n^2) \\
&\triangleq q_{\text{ep}}(y_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}, \mathbf{x}^\star) .
\end{aligned}$$

F AN EXAMPLE ON EXPLOITATION VS. EXPLORATION OF DIFFERENT TES APPROXIMATION METHODS

Fig. 1 shows an example where TES_{sp} and TES_{ep} select different inputs. In particular, TES_{ep} selects an input where the GP posterior mean of its function value is higher than that of the input TES_{sp} selects. Scrutinizing more carefully on the acquisition function values $f(0)$ and $f(2)$, we observe that the difference in values of TES_{sp} is smaller than that of TES_{ep}. In these cases, it means that TES_{ep} has a stronger preference to a larger mean function value than TES_{sp}. Hence, it is observed in these cases that TES_{ep} exploits more than TES_{sp}.

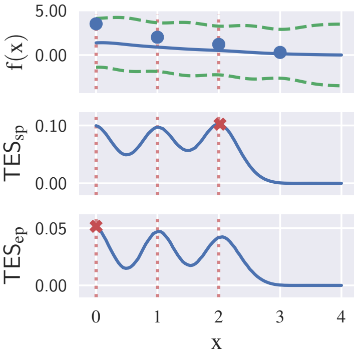


Figure 1: An example: TES_{ep} exploits more than TES_{sp}. The top plot shows the GP posterior mean as a solid blue line, uncertainty (variance) as dashed green lines, and data points as blue points. The dotted red lines show the positions of $\mathcal{X}^\star$. The red crosses indicate the maximizers of acquisition functions.

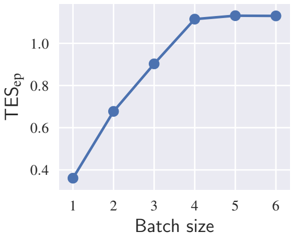
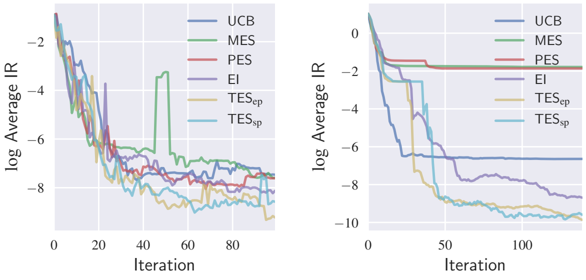


Figure 2: Plot of TES_{ep} against the batch size for $n = 5$.

G TES_{EP} OF DIFFERENT BATCH SIZES

Note that $\alpha^\star(\mathbf{y}_{\mathcal{D}}, \mathcal{B})$ is the information gain about $\mathbf{x}^\star$ through observing $\mathbf{y}_{\mathcal{B}}$. Thus, it is comparable between batches of different sizes. Given the size of $\mathcal{X}^\star$ as 5, Fig. 2 shows the maximum TES_{ep} value (i.e., maximum information gain) for different batch sizes. We observe that increasing the batch size to be larger than $|\mathcal{X}^\star|$ only yields an insignificant amount of TES_{ep}. Therefore, the size of $\mathcal{X}^\star$ should be at least the batch size to make the most out of the observations $\mathbf{y}_{\mathcal{B}}$. Besides, we can select the batch size adaptively at each BO iteration based on a trade-off between the increase in the information gain and the cost of an observation.



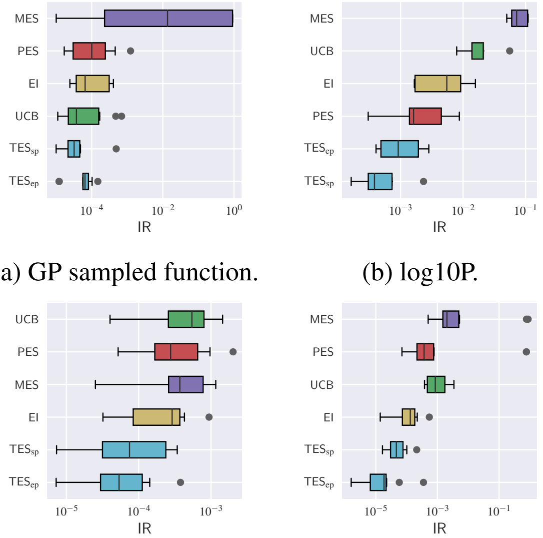
(a) Branin. (b) Hartmann-3D.

Figure 3: Additional BO results.

H ADDITIONAL EXPERIMENT RESULTS

Fig. 3 shows the average of the results over 10 random runs for 2 synthetic functions: Branin defined in $[0,1]^2$, and Hartmann-3D defined in $[0,1]^3$ [Lizotte, 2008]. As our optimization problem is maximization and these synthetic functions are often minimized, we take the negative values of these functions as objective functions. In these experiments, the noise variance is fixed to $\sigma_n^2 = 10^{-4}$. Note that unlike the synthetic function we sample from a GP posterior, the assumption of using an *isotropic* SE kernel[1] might be violated in these synthetic functions. However, the empirical results in Fig. 3 still show a reasonable performance. Overall, TES acquisition functions outperform others, but the difference is less obvious in the Branin function. It could be because Branin is a simple function to optimize (i.e., having 3 local maxima in a 2-dimensional input space). In the Hartmann-3D experiment, EI outperforms PES, which could be because PES explores more than EI as explained by Hernández-Lobato et al. [2014].

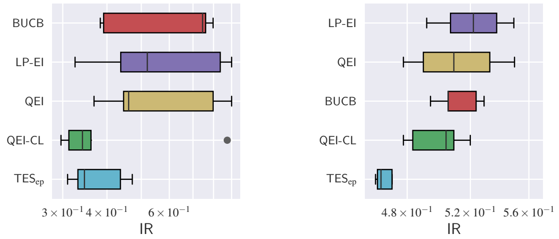
Additionally, the box plots of IR in the last iteration of random experimental runs are shown in Figs. 4, 5, and 6.



(a) GP sampled function. (b) log10P.

(c) Branin. (d) Hartmann-3D.

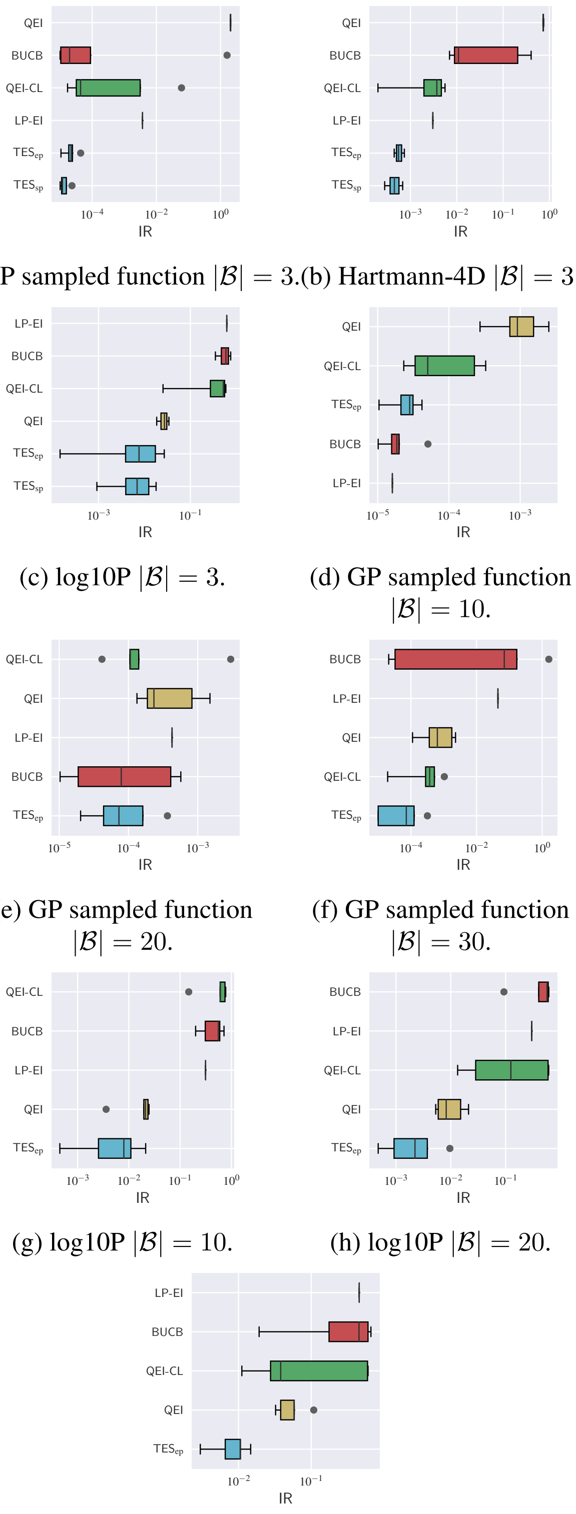
Figure 4: Box plots of BO with $|\mathcal{B}| = 1$.



(a) The CIFAR-10. (b) The face attack function.

Figure 5: Box plots of batch BO results for CIFAR-10 and the face attack function. $|\mathcal{B}| = 20$.

[1] A kernel $k_{\mathbf{x}\mathbf{x}'}$ is *isotropic* if it depends on $|\mathbf{x} - \mathbf{x}'|$ only.



(a) GP sampled function $|\mathcal{B}| = 3$.(b) Hartmann-4D $|\mathcal{B}| = 3$.

(c) log10P $|\mathcal{B}| = 3$.

(d) GP sampled function $|\mathcal{B}| = 10$.

(e) GP sampled function $|\mathcal{B}| = 20$.

(f) GP sampled function $|\mathcal{B}| = 30$.

(g) log10P $|\mathcal{B}| = 10$.

(h) log10P $|\mathcal{B}| = 20$.

(i) log10P $|\mathcal{B}| = 40$.

Figure 6: Box plots of batch BO (i.e., $|\mathcal{B}| > 1$) results for the GP sampled function, Hartmann-4D, and log10P.