Supplementary Documents for paper: Batch Bayesian Optimization via Simulation Matching

Javad Azimi, Alan Fern, Xiaoli Z. Fern
School of EECS, Oregon State University
{azimi, afern, xfern}@eecs.oregonstate.edu

Proposition 1. *Let* $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *then*

$$p(\mathbf{x}_i \geq \mathbf{x}_1, \cdots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \cdots, \mathbf{x}_n) = \prod_{j=1}^{n-1} \left(1 - \Phi(-\mu_z^j)\right), \tag{1}$$

such that $\mathbf{z} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{z}} = \left(A\boldsymbol{\Sigma}A'\right)^{-\frac{1}{2}} A\boldsymbol{\mu},\ \boldsymbol{\Sigma}_{\mathbf{z}} = I\right)$*, and* $A_{n-1\times n}$ *is a sparse matrix such that* $\forall j\ a_{ji} = 1$ *,and* $\forall k,\ 0 \leq k < i,\ a_{kk} = -1$ *,and* $\forall k,\ i < k \leq n,\ a_{k-1,k} = -1$.

Proof. Let $\mathbf{x} = (x_1, x_2, \cdots, x_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\begin{aligned} p(x_1 > x_2, \cdots, x_n) &= p(x_1 > x_2, x_1 > x_3, \cdots, x_1 > x_n) \\ &= p(x_1 - x_2 > 0, x_1 - x_3 > 0, \cdots, x_1 - x_n > 0) \end{aligned} \tag{2}$$

Let $\mathbf{y} = (y_1, y_2, \cdots, y_{n-1})$ such that $y_i = x_1 - x_{i+1}$. Therefore, $\mathbf{y} \sim \mathcal{N}\left(\mathrm{A}\boldsymbol{\mu}, \mathrm{A}\boldsymbol{\Sigma}\mathrm{A}'\right)$ and A is defined as follows

$$\mathrm{A} = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots \\ 1 & 0 & -1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 & -1 \end{bmatrix}_{n-1\times n} \tag{3}$$

Let define $\boldsymbol{\mu}_y = \mathrm{A}\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_y = \mathrm{A}\boldsymbol{\Sigma}\mathrm{A}'$. Now, we need to obtain $p(\mathbf{y} > 0),\ \mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right)$. Let define $\mathbf{z} = \boldsymbol{\Sigma}_y^{-\frac{1}{2}}\mathbf{y}$. Therefore, $\mathbf{z} \sim \mathcal{N}\left(\boldsymbol{\Sigma}_y^{-\frac{1}{2}}\boldsymbol{\mu}_y,\ \mathrm{I}\right)$. Therefore we can derive $p(\mathbf{y} > 0)$ as follows:

$$\begin{aligned} p(\mathbf{y} > 0) &= p\left(\boldsymbol{\Sigma}_y^{\frac{1}{2}}\mathbf{z} > 0\right) \\ &\equiv p(\mathbf{z} > 0) \qquad \text{since } \boldsymbol{\Sigma}_y^{\frac{1}{2}} \text{ is P.D} \\ &= p(z_1 > 0)p(z_2 > 0)\cdots p(z_{n-1} > 0) \\ &= \left(1 - \Phi(-\mu_z^1)\right)\left(1 - \Phi(-\mu_z^2)\right)\cdots\left(1 - \Phi(-\mu_z^{n-1})\right) \end{aligned} \tag{4}$$

Therefore,

$$p(x_1 > x_2, \cdots, x_n) = \left(1 - \Phi(-\mu_z^1)\right)\left(1 - \Phi(-\mu_z^2)\right)\cdots\left(1 - \Phi(-\mu_z^{n-1})\right) \tag{5}$$

such that $\mathbf{z} \sim \mathcal{N}\left(\left(\mathrm{A}\boldsymbol{\Sigma}\mathrm{A}'\right)^{-\frac{1}{2}}\mathrm{A}\boldsymbol{\mu},\ \mathrm{I}\right)$.

By changing the matrix A we can Compute the above probability for any x_i. □

Lemma 1. *The proposed objective function is a nonincreasing supermodular function.*

Proof. We prove it for the case when we have only one set $\mathbf{S}$ and then we can simply extend it to multiple sets $\mathbf{S}_i,\ i = \{1, 2, \cdots, n_s\}$ by taking a summation.

First we prove the nonincreasing property of our objective function. We have to prove $f(\mathbf{S}^b \cup \mathbf{x}^*) \leq f(\mathbf{S}^b)$ which can be simply proved as follow:

$$
\begin{aligned}
f(\mathbf{S}^b \cup \mathbf{x}^*) &= \sum_{j=1}^{k} \min_{\mathbf{x}_b \in \{\mathbf{S}^b \cup \mathbf{x}^*\}} w_j \parallel \mathbf{x}_j - \mathbf{x}_b \parallel \\
&= \sum_{j=1}^{k} \min \left(\min_{\mathbf{x}_b \in \{\mathbf{S}\}} w_j \parallel \mathbf{x}_j - \mathbf{x}_b \parallel,\ w_j \parallel \mathbf{x}_j - \mathbf{x}^* \parallel \right) \\
&\leq \sum_{j=1}^{k} \min_{\mathbf{x}_b \in \mathbf{S}^b} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel \\
&= f(\mathbf{S}^b)
\end{aligned} \tag{6}
$$

Next we will prove the supermodularity property of our objective function.

Let $\mathbf{S}_1^b \subset \mathbf{S}_2^b$ such that $\mathbf{S}_1^b = \{\mathbf{x}_1, \mathbf{x}_2, \cdots \cdot \mathbf{x}_{n_1}\}$, $\mathbf{S}_2^b = \{\mathbf{x}_1, \mathbf{x}_2, \cdots \cdot \mathbf{x}_{n_1}, \cdots, \mathbf{x}_{n_2}\}$. We need to prove that

$$
\begin{aligned}
&\underbrace{\sum_{j=1}^{k} \min_{\mathbf{x}_b \in \mathbf{S}_1^b} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel}_{p_1} - \underbrace{\sum_{j=1}^{k} \min_{\mathbf{x}_b \in \left\{\mathbf{S}_1^b \cup \mathbf{x}^*\right\}} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel}_{p_2} \\
&\geq \underbrace{\sum_{j=1}^{k} \min_{\mathbf{x}_b \in \mathbf{S}_2^b} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel}_{p_3} - \underbrace{\sum_{j=1}^{k} \min_{\mathbf{x}_b \in \left\{\mathbf{S}_2^b \cup \mathbf{x}^*\right\}} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel}_{p_4}
\end{aligned} \tag{7}
$$

Let rewrite the equation 7 as follows:

$$
\begin{aligned}
&\sum_{j=1}^{k} \underbrace{\min_{\mathbf{x}_b \in \mathbf{S}_1^b} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel}_{p_1} - \underbrace{\min \left(\min_{\mathbf{x}_b \in \left\{\mathbf{S}_1^b\right\}} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel,\ \parallel \mathbf{x}_j - \mathbf{x}^* \parallel \right)}_{p_2} \\
&\geq \sum_{j=1}^{k} \underbrace{\min \left(\min_{\mathbf{x}_b \in \mathbf{S}_1^b} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel,\ \min_{\mathbf{x}_b \in \mathbf{S}_2^b \setminus \mathbf{S}_1^b} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel \right)}_{p_3} \\
&\quad - \underbrace{\min \left(\min_{\mathbf{x}_b \in \mathbf{S}_1^b} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel,\ \min_{\mathbf{x}_b \in \mathbf{S}_2^b \setminus \mathbf{S}_1^b} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel,\ \parallel \mathbf{x}_j - \mathbf{x}^* \parallel \right)}_{p_4}
\end{aligned} \tag{8}
$$

Since we already proved the objective function in non-increasing then it is clear that $p_1 \geq p_3$. For each $\mathbf{x}_j$ we might have an improvement $I_l^j = p1 - p2 \geq 0,\ I_r^j = p3 - p4 \geq 0$ in both sides of the above equation. We need to prove that $\forall j, I_l^j \geq I_r^j$.

If for a particular $\mathbf{x}_j$ we have $p_1 = p_2$ then it can be easily seen that $p_3 = p_4$ and then $I_l^j = I_r^j$. Suppose for a particular $\mathbf{x}_j$ we have $p_1 > p_2$ and consequently we have the improvement $I_l^j = p1 - p2 > 0$. Therefor, we can have 2 different cases in the right hand side:

1. $p_4 = p_3$ which means

$$\min_{\mathbf{x}_b \in \mathbf{S}_2^b \setminus \mathbf{S}_1^b} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel < \min \left(\min_{\mathbf{x}_b \in \mathbf{S}_1^b} \parallel \mathbf{x}_j - \mathbf{x}_b \parallel,\ \parallel \mathbf{x}_j - \mathbf{x}^* \parallel \right)$$

   Therefor $I_r^j = 0$ and we show that $I_l^j > I_r^j$

2. $p_4 < p_3$ which can be easily shown that it means $p_4 = p_2 = \| \mathbf{x}_j - \mathbf{x}^* \|$, and since $p_1 \geq p_2$ we can conclude that $p_1 - p_2 \geq p_3 - p_4$ which means $I_l^j \geq I_r^j$.

□

Corollary 1. *Let* $t = 1$*, then we have the following bound for the greedy algorithm.*

$$f(\mathbf{S}^b) \leq (e-1) f\left(Opt\right). \tag{9}$$

Proof.

$$\begin{aligned} f(\mathbf{S}^b) &\leq \frac{1}{t}\left[\left(\frac{q+t}{q}\right)^q - 1\right] f(Opt) \\ &= \left(\left(\frac{q+1}{q}\right)^q - 1\right) f(Opt) \qquad t = 1 \\ &\leq \left(\lim_{q\to\infty}\left(\frac{q+1}{q}\right)^q - 1\right) f(Opt) \\ &= (e-1) f(Opt) \end{aligned} \tag{10}$$

□

Lemma 2. *The expected maximum,* $E[max(.)]$*, over a set of normal random variables is a submodular monotonic function.*

Proof. First, let proof that the $E[max(.)]$ is a monotonically increasing function. Suppose $p < k$ and $(f_1, f_2, \cdots, f_p)$ and $(f_1, f_2, \cdots, f_p, \cdots, f_k)$ have joint normal distributions. We need to prove that

$$E\left[\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right) \Big| \mathcal{D}\right] \geq E\left[\max\left(f_1, f_2, \cdots, f_p\right) \Big| \mathcal{D}\right]. \tag{11}$$

We use the definition of the expectation to prove the result.

$$\begin{aligned} & E\left[\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right) \Big| \mathcal{D}\right] \\ &\quad = \int \cdots \int \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right) p_{f_1, f_2, \cdots, f_p, \cdots, f_k | \mathcal{D}} df_1 df_2 \cdots df_p \cdots df_k \\ &\quad \geq \int \cdots \int \max\left(f_1, f_2, \cdots, f_p\right) p_{f_1, f_2, \cdots, f_p, \cdots, f_k | \mathcal{D}} df_1 df_2 \cdots df_p \cdots df_k \\ &\quad = \int \cdots \int \max\left(f_1, f_2, \cdots, f_p\right) \left(\int \cdots \int p_{f_1, f_2, \cdots, f_p, \cdots, f_k | \mathcal{D}} df_{p+1} \cdots df_k\right) df_1 df_2 \cdots df_p \\ &\quad = \int \cdots \int \max\left(f_1, f_2, \cdots, f_p\right) p_{f_1, f_2, \cdots, f_p | \mathcal{D}} df_1 df_2 \cdots df_p \\ &\quad = E\left[\max\left(f_1, f_2, \cdots, f_p\right) \Big| \mathcal{D}\right]. \end{aligned} \tag{12}$$

A set function F is called *submodular*, if for any 2 sets of samples $\mathbf{S}_1^b, \mathbf{S}_2^b \subseteq \mathcal{S}$ and $\mathbf{x}^* \in \mathcal{S} \setminus \mathbf{S}_2^b$ with corresponding realization f^*, it holds that $F(\mathbf{S}_1^b \cup \mathbf{x}) - F(\mathbf{S}_1^b) \geq F(\mathbf{S}_2^b \cup \mathbf{x}) - F(\mathbf{S}_2^b)$. Again, in our application F is $E[max(\cdot)]$. Let $\mathbf{S}_1^b = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p\}$ and $\mathbf{S}_2^b = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$ where $p \leq k$. We need to prove that

$$\begin{aligned} & E\left[\max\left(f_1, f_2, \cdots, f_p, f^*\right) \Big| \mathcal{D}\right] - E\left[\max\left(f_1, f_2, \cdots, f_p\right) \Big| \mathcal{D}\right] \\ &\quad \geq E\left[\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k, f^*\right) \Big| \mathcal{D}\right] - E\left[\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right) \Big| \mathcal{D}\right]. \end{aligned} \tag{13}$$

To prove this, we start from the right hand side of the inequality and the basic definition of the expectation.

$$
\begin{aligned}
&E\left[\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k, f^*\right) \Big| \mathcal{D}\right] - E\left[\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right) \Big| \mathcal{D}\right] \\
&\quad = \int \cdots \int \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k, f^*\right) p_{f_1, f_2, \cdots, f_p, \cdots, f_k, f^* | \mathcal{D}} df_1 df_2 \cdots df_p \cdots df_k df^* \\
&\qquad - \int \cdots \int \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right) p_{f_1, f_2, \cdots, f_p, \cdots, f_k | \mathcal{D}} df_1 df_2 \cdots df_p \cdots df_k \\
&\quad = \int \cdots \int \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k, f^*\right) p_{f_1, f_2, \cdots, f_p, \cdots, f_k, f^* | \mathcal{D}} df_1 df_2 \cdots df_p \cdots df_k df^* \\
&\qquad - \int \cdots \int \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right) p_{f_1, f_2, \cdots, f_p, \cdots, f_k, f^* | \mathcal{D}} df_1 df_2 \cdots df_p \cdots df_k df^* \\
&\quad = \int \cdots \int \left[\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k, f^*\right) - \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right)\right] \\
&\qquad\qquad p_{f_1, f_2, \cdots, f_p, \cdots, f_k, f^* | \mathcal{D}} df_1 df_2 \cdots df_p \cdots df_k df^* \\
&\quad \leq \int \cdots \int \left[\max\left(f_1, f_2, \cdots, f_p, f^*\right) - \max\left(f_1, f_2, \cdots, f_p\right)\right] \\
&\qquad\qquad p_{f_1, f_2, \cdots, f_p, \cdots, f_k, f^* | \mathcal{D}} df_1 df_2 \cdots df_p \cdots df_k df^* \\
&\quad = \int \cdots \int \left[\max\left(f_1, f_2, \cdots, f_p, f^*\right) - \max\left(f_1, f_2, \cdots, f_p\right)\right] p_{f_1, f_2, \cdots, f_p, f^* | \mathcal{D}} df_1 df_2 \cdots df_p df^* \\
&\quad = \int \cdots \int \max\left(f_1, f_2, \cdots, f_p, f^*\right) p_{f_1, f_2, \cdots, f_p, f^* | \mathcal{D}} df_1 df_2 \cdots df_p df^* \\
&\qquad\qquad - \int \cdots \int \max\left(f_1, f_2, \cdots, f_p\right) p_{f_1, f_2, \cdots, f_p | \mathcal{D}} df_1 df_2 \cdots df_p \\
&\quad = E\left[\max\left(f_1, f_2, \cdots, f_p, f^*\right) \Big| \mathcal{D}\right] - E\left[\max\left(f_1, f_2, \cdots, f_p\right) \Big| \mathcal{D}\right]
\end{aligned} \tag{14}
$$

Notice that the inequality holds if we can prove:

$$
\begin{aligned}
&\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k, f^*\right) - \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right) \\
&\leq \max\left(f_1, f_2, \cdots, f_p, f^*\right) - \max\left(f_1, f_2, \cdots, f_p\right)
\end{aligned} \tag{15}
$$

There are two possible cases as follows:

$$
\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k, f^*\right) = \begin{cases} f^* \\ \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right). \end{cases} \tag{16}
$$

1. In the first case, if $\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k, f^*\right) = f^*$, then we also have $\max\left(f_1, f_2, \cdots, f_p, f^*\right) = f^*$. Hence,

$$
\begin{aligned}
&\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k, f^*\right) - \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right) \\
&\quad = f^* - \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right) \\
&\quad \leq f^* - \max\left(f_1, f_2, \cdots, f_p\right) \\
&\quad = \max\left(f_1, f_2, \cdots, f_p, f^*\right) - \max\left(f_1, f_2, \cdots, f_p\right)
\end{aligned} \tag{17}
$$

2. In the second case, if $\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k, f^*\right) = \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right)$, then we have

$$
\begin{aligned}
&\max\left(f_1, f_2, \cdots, f_p, \cdots, f_k, f^*\right) - \max\left(f_1, f_2, \cdots, f_p, \cdots, f_k\right) \\
&\quad = 0 \\
&\quad \leq \max\left(f_1, f_2, \cdots, f_p, f^*\right) - \max\left(f_1, f_2, \cdots, f_p\right) \\
&\quad = \max\left(f_1, f_2, \cdots, f_p, f^*\right) - \max\left(f_1, f_2, \cdots, f_p\right)
\end{aligned} \tag{18}
$$

Notice that $\max\left(f_1, f_2, \cdots, f_p, f^*\right) - \max\left(f_1, f_2, \cdots, f_p\right)$ is always non-negative.

□