SUPPLEMENTARY MATERIALS

A TECHNICAL PRELIMINARIES AND PROOFS

A.1 PROOFS OF THEORETICAL RESULTS

We begin with the preliminary Lemmas 1 and 2.

Lemma 1. *(Theorem 2 of [Chowdhury and Gopalan, 2017]). Let f be a function lying in the RKHS $\mathcal{H}_{k_0}$ of kernel k_0 such that $\|f\|_{\mathcal{H}_k} \leq 1$ with input dimension D. Assume the process of observation noise $\{\epsilon_t\}$ is σ-sub-gaussian. Then setting*

$$\beta_t = 1 + \sigma\sqrt{2\left(\gamma_{t-1} + 1 + \ln(1/\delta)\right)},$$

we have the following holds with probability at least $1-\delta$,

$$|\mu_t(x) - f(x)| \leq \beta_{t+1}\sigma_t(x), \quad \forall x \in \mathcal{X}, \quad \forall t \geq 1,$$

where μ_t, σ_t are given by the formula

$$\mu_t(\mathbf{x}) = \mathbf{k}_{t-1}^T(\mathbf{x})(\mathbf{K}_{t-1} + \sigma\mathbf{I})^{-1}\mathbf{y}_{t-1},$$
$$\sigma_t^2(\mathbf{x}) = k_0(\mathbf{x}, \mathbf{x}) - \mathbf{k}_{t-1}^T(\mathbf{x})(\mathbf{K}_{t-1} + \sigma\mathbf{I})^{-1}\mathbf{k}_{t-1}(\mathbf{x}).$$

Lemma 2. *(Lemma 4 of [Chowdhury and Gopalan, 2017]). Suppose we sample the objective function f at $\{\mathbf{x}^1, \ldots, \mathbf{x}^{T-1}\}$ then the sum of standard deviations is bounded by,*

$$\sum_{t=1}^{T} \sigma_{t-1}\left(\mathbf{x}^t\right) \leq \sqrt{4(T+2)\gamma_T}.$$

Our first goal is to prove Lemma 1.

Lemma 1. *Suppose the learning rate of the LaMBO is set to be $\eta = \sqrt{2^{-H}T^{-1}\log|\mathcal{K}|}$, where H is the depth of the MSET, then the expected cumulative regret of LaMBO is:*

$$\mathbb{E}[R(T)] = \mathcal{O}\left(\sqrt{2^H T \log|\mathcal{K}|}\right).$$

Remark 1. *We can compare the result with SMB in [Koren et al., 2017b] where $\mathbb{E}[R_T] = \mathcal{O}\left(\sqrt{kT\log|\mathcal{K}|}\right)$. Note that ours is a lower bound of it as $2^H \leq k$. They could potentially have a large gap between them in terms of order. This performance improvement is due to our loss estimator adapted to arm correlation, whereas [Koren et al., 2017b] considers the pure bandit information.*

The following Lemma 3 is a key to prove Lemma 1 in the main text.

Lemma 3. *For any sequence of $\tilde{\ell}_1, \ldots, \tilde{\ell}_T$, denote i^* to be the solution of $\max_i \sum_{t=1}^T \tilde{\ell}_t(i)$ and assume $2^H \leq c\frac{T}{\log|K|}$ for some constant $c > 0$, then there exists an $\eta = \Theta(\sqrt{2^{-H}T^{-1}\log|\mathcal{K}|})$ such that LaMBO has the property*

$$\mathbb{E}[\sum_{t=1}^{T}(p_t \cdot \tilde{\ell}_t - \tilde{\ell}_t(i^*))] \leq \frac{\log|K|}{\eta} + \eta T 2^{H+1}.$$

For the proof of 3 we follow the path in [Koren et al., 2017a]. Before we start the proof of Lemma 3, we will need Lemma 4, 5, 6, and 7.

Lemma 4.

$$0 \leq \bar{\ell}_{t,h}(i) \leq \prod_{j=0}^{h-1}(1 + \sigma_{t,j}), \quad \forall i \in K. \tag{1}$$

In particular, if $\sigma_{t,h} = -1$ then $\bar{\ell}_{t,j} = 0$ for all $j > h$.

Proof. The last statement is trivial from definition. We will prove Eq. (4) by induction on h. Since $0 \leq \bar{\ell}_{t,0}(i) \leq 1$ by Eq. (8) and the UCB is upper bounded by 1, the statement holds for $h = 0$. Now assume it holds for $h - 1$. Then firstly,

$$\bar{\ell}_{t,h}(i) \geq -\frac{1}{\eta} \log \left(\sum_{j \in A_h(i)} \frac{p_t(j)}{p_t(A_h(i))} \right) = 0.$$

Secondly, applying Jensen's inequality, we have

$$\begin{aligned}
\bar{\ell}_{t,h}(i) &\leq -\frac{1}{\eta} \sum_{k \in A_h(i)} \frac{p_t(k)}{p_t(A_h(i))} \log(\exp(-\eta(1 + \sigma_{t,h-1})\bar{\ell}_{t,h-1}(k))) \\
&= (1 + \sigma_{t,h-1}) \sum_{k \in A_h(i)} \frac{p_t(k)}{p_t(A_h(i))} \bar{\ell}_{t,h-1}(k) \\
&\leq \sum_{k \in A_h(i)} \frac{p_t(k)}{p_t(A_h(i))} \prod_{j=0}^{h-1} (1 + \sigma_{t,j}) \\
&= \prod_{j=0}^{h-1} (1 + \sigma_{t,j}),
\end{aligned}$$

where the second inequality is followed by the induction assumption. Therefore the proof is complete by mathematical induction. □

Lemma 5. *For all t and $0 \leq h \leq H$ the followings hold:*

- *For all t we have*

$$\mathbb{E}\left[p_t \cdot \tilde{\ell}_t\right] = \mathbb{E}\left[\tilde{\ell}_t(i_t)\right]. \tag{2}$$

- *With probability at least $1 - 2^{-(h+1)}$, we have that $A_h(i_t) = A_h(i_{t-1})$.*

Proof. The proof of the second property is identical to Lemma 8 in [Koren et al., 2017a] and is thus omitted . Now we prove the first property. Note that we only need to prove

$$\mathbb{E}[\mathbf{1}(i_t = i)] = \mathbb{E}[p_t(i)], \quad \forall t > 0, \quad \forall i \in \mathcal{K}. \tag{3}$$

We will again use the mathematical induction to prove the above statement. The initial case $t = 1$ holds trivially. Now assume the statement is true for $t = k$. Then for $t = k + 1$,

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}(i_{k+1} = i)|h_k = 0] = \mathbb{E}[\mathbf{1}(i_k = i)|h_k = 0] \\
=&\mathbb{E}[\mathbf{1}(i_k = i)] = \mathbb{E}[p_k(i)],
\end{aligned}$$

where the last equality follows from the induction assumption. On the other hand,

$$\mathbb{E}[p_{k+1}(i)|h_k = 0] = \mathbb{E}[p_k(i)|h_k = 0] = \mathbb{E}[p_k(i)],$$

where the last equality follows from the independence between p_k and h_k. Hence we have

$$\mathbb{E}[\mathbf{1}(i_{k+1} = i)|h_k = 0] = \mathbb{E}[p_{k+1}(i)|h_k = 0]. \tag{4}$$

Now if $h_k = h' > 0$. Let $A' \in \mathcal{A}_{h'}$ be the subtree such that $\{i\} \subset A'$ then by the tower rule for expectation we have

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}(i_{k+1} = i)|h_k = h', p_{k+1}] \\
=&\mathbb{E}[\mathbf{1}_{A'}(i_{k+1})\mathbb{E}[\mathbf{1}(i_{k+1} = i)|h_k, p_{k+1}, i_k \in A']|h_k = h', p_{k+1}] \\
=&\mathbb{E}[p_{k+1}(A')p_{k+1}(i|A')|h_k = h', p_{k+1}] \\
=&\mathbb{E}[p_{k+1}(i)|h_k = h', p_{k+1}].
\end{aligned}$$

Therefore,

$$\mathbb{E}[\mathbf{1}(i_{k+1} = i | h_k = h')] = \mathbb{E}[p_{k+1}(i) | h_k = h']. \tag{5}$$

By (4), (5) now holds for every possible value of h_k, so we must have

$$\mathbb{E}[\mathbf{1}(i_{k+1} = i)] = \mathbb{E}[p_{k+1}(i)],$$

which completes the proof by induction. □

Lemma 6. *For all t, we have* $\mathbb{E}[p_t \cdot \tilde{\ell}_t^2] \leq 2^{H+1}$.

Proof. Observe

$$\tilde{\ell}_t^2(i) \leq \left(\bar{\ell}_{t,0}(i) + \sum_{h=0}^{H-1} \sigma_{t,h} \bar{\ell}_{t,h}(i) \right)^2.$$

Since $\mathbb{E}[\sigma_{t,h}] = 0$ and $\mathbb{E}[\sigma_{t,h}\sigma_{t,h'}] = 0$ for $h \neq h'$, we have

$$\mathbb{E}[\tilde{\ell}_t^2(i)] \leq 2 \sum_{h=0}^{H-1} \mathbb{E}[\bar{\ell}_{t,h}^2(i)]. \tag{6}$$

Now by Lemma 4 we have

$$p_t \cdot \bar{\ell}_{t,h}^2 \leq \sum_{i \in \mathcal{K}} p_t(i) \prod_{j=0}^{h-1} (1 + \sigma_{t,h})^2.$$

Then taking expectation on both sides leads to

$$\mathbb{E}[p_t \cdot \bar{\ell}_{t,h}^2] \leq \sum_{i \in K} p_t(i) 2^h = 2^h. \tag{7}$$

Finally, combining Eq. (6) with Eq. (7), we get

$$\mathbb{E}[p_t \cdot \tilde{\ell}_t^2] \leq 2 \sum_{h=0}^{H-1} \mathbb{E}[p_t \cdot \bar{\ell}_{t,h}^2] \leq 2^{H+1}.$$

□

Lemma 7. *[Alon et al., 2015]. Let* $\eta > 0$ *and* $\mathbf{z}_1, \ldots, \mathbf{z}_T \in \mathbb{R}^{|\mathcal{K}|}$ *be real vectors such that* $\mathbf{z}_t(i) \geq -\frac{1}{\eta}$ *then a sequence of probability vectors* $p_1, \ldots, p_T$ *defined by* $p_1 = (1/|\mathcal{K}|, \ldots, 1/|\mathcal{K}|)$ *and for all* $t > 1$*,*

$$p_t(i) = \frac{p_{t-1}(i) \exp(-\eta \mathbf{z}_t(i))}{\sum_{j \in \mathcal{K}} q_{t-1}(j) \exp(-\eta \mathbf{z}_t(j))},$$

have the property that

$$\sum_{t=1}^{T} p_t \cdot \mathbf{z}_t \leq \sum_{t=1}^{T} \mathbf{z}_t(i^*) + \frac{\log |\mathcal{K}|}{\eta} + \eta \sum_{t=1}^{T} p_t \cdot \mathbf{z}_t^2,$$

for any $i^* \in \mathcal{K}$*.*

Now we are ready to prove Lemma 3.

Proof. By the assumption that $2^H \le c\frac{T}{\log|K|}$ for some constant $c > 0$, if we set $\eta = \sqrt{c^{-1}2^{-H}T^{-1}\log|K|}$ then we have $2^H \le \frac{1}{\eta}$. Also observe that $\tilde{\ell}_t = \bar{\ell}_{t,0} + \sum_{j=0}^{h_t-1}\bar{\ell}_{t,j} - \bar{\ell}_{t,h_t}$, so Lemma 4 implies that $\bar{\ell}_t \ge -\frac{1}{\eta}$. Now we apply Lemma 7 to the sequence $\{\tilde{\ell}_t\}_t$ to obtain

$$\sum_{t=1}^{T} p_t \cdot \tilde{\ell}_t - \sum_{t=1}^{T} \tilde{\ell}_t(i^*) \le \frac{\log|\mathcal{K}|}{\eta} + \eta \sum_{t=1}^{T} p_t \cdot \tilde{\ell}_t^2. \tag{8}$$

Finally, we take expectation on both sides of Eq. (8) together with Lemma 5 and 6, then

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T} \tilde{\ell}_t(i_t) - \sum_{t=1}^{T} \tilde{\ell}_t(i^*)] \\
=&\mathbb{E}[\sum_{t=1}^{T} p_t \cdot \tilde{\ell}_t - \sum_{t=1}^{T} \tilde{\ell}_t(i^*)] \\
\le&\frac{\log|K|}{\eta} + \eta \sum_{t=1}^{T} \mathbb{E}[p_t \cdot \tilde{\ell}_t^2] \\
\le&\frac{\log|K|}{\eta} + \eta T 2^{H+1},
\end{aligned}$$

which completes the proof. □

Next we prove Lemma 2 in the main text.

Lemma 2. *For sufficient large T, suppose for $m = 1, \dots, N-1$ that the parameters d_m of an MSET are chosen recursively,*

$$\begin{aligned}
d_1 &= \left\lfloor -\log\left(\frac{1}{\sqrt{\lambda}}\sum_{j=2}^{N-1} c_j / \sum_{j=1}^{N-1} c_j\right)\right\rfloor, \\
d_m &= \left\lfloor -\log\left(\frac{1}{\sqrt{\lambda}}\sum_{j=m+1}^{N-1} c_j / \sum_{j=1}^{N-1} c_j\right)\right\rfloor - \sum_{n=1}^{m-1} d_n, \\
i &= 2, \dots, N-2, \\
d_{N-1} &= \log(T^{1/3}/\log|\mathcal{K}|) - \sum_{m=1}^{N-2} d_m.
\end{aligned} \tag{9}$$

Then LaMBO results in cumulative costs

$$\mathbb{E}[\sum_{t=1}^{T} \Gamma^t] = \mathcal{O}\left(\sum_{m=1}^{N-1} \sqrt{\lambda} c_m T^{2/3} \log|\mathcal{K}| \log\frac{T^{1/3}}{\log|\mathcal{K}|}\right). \tag{10}$$

Proof. The proof follows by showing firstly that the movement cost is dominated by a HST metric, and secondly that under the tree metric the cumulative cost is bounded by the quantity in the lemma. To define the HST metric formally, let us introduce the following terminology in accordance to [Koren et al., 2017b]. Given u, v be nodes in the MSET $\mathcal{T}$, let LCA(u, v) be their least common ancestor node. Then the scaled HST metric is defined as follows:

$$\Delta_{\mathcal{T}}(u, v) = (\sqrt{\lambda}\sum_{j=1}^{N-1} c_j)\frac{2^{\text{level}(\text{LCA}(u,v))}}{2^{\text{depth}(\mathcal{T})}}, \ \forall u, v \in \mathcal{K}. \tag{11}$$

Under this metric, the cost incurred from changing variables in the i^{th} module is

$$(\sqrt{\lambda}\sum_{j=1}^{N-1} c_j)\frac{2^{d_i+\dots+d_{N-1}}}{2^{d_1+\dots+d_{N-1}}} = \frac{\sum_{j=1}^{N-1} c_j}{2^{d1+\dots+d_{i-1}}}.$$

Then the condition of dominance over the original cost is, for $i = 1, \dots, N-2$,

$$\frac{1}{\sqrt{\lambda}} \frac{\sum_{j=1}^{N-1} c_j}{2^{d1+\cdots+d_{i-1}}} \geq \sum_{j=i}^{N-1} c_i,$$

$$\Rightarrow d_1 + \cdots + d_{i-1} \leq -\log\left(\frac{\sum_{j=i}^{N-1} c_j}{\sum_{j=1}^{N-1} c_j}\right) - \frac{\log\lambda}{2}.$$

Rearrangements of these linear inequalities yield the solution for d_1 to d_{N-2} as

$$\begin{aligned} d_1 &= \left\lfloor -\log\left(\frac{\sum_{j=2}^{N-1} c_j}{\sum_{j=1}^{N-1} c_j}\right) - \frac{\log\lambda}{2}\right\rfloor, \\ d_i &= \left\lfloor -\log\left(\frac{\sum_{j=i+1}^{N-1} c_j}{\sum_{j=1}^{N-1} c_j}\right) - \frac{\log\lambda}{2}\right\rfloor - \sum_{n=1}^{i-1} d_n, \\ i &= 2, \dots, N-2. \end{aligned} \tag{12}$$

Under the condition in Eq. (12), the cost incurred from the HST metric Eq. (11) is larger than our original cost. Hence, an upper bound for the cost incurred from the metric will also bound our cumulative cost.

Now we bound the cumulative cost under this HST metric. Observe i_t and i_{t-1} belongs to the same subtree on level h of the tree with probability at least $1 - 2^{h-H}$, therefore we have

$$\begin{aligned} \mathbb{E}[\Delta_{\mathcal{T}}(i_t, i_{t-1})] &\leq \sum_{j=1}^{N-1} \sqrt{\lambda} c_j \sum_{h=0}^{H-1} 2^{h-H} \cdot 2^{h-1} \\ &\leq \sum_{j=1}^{N-1} \sqrt{\lambda} c_j \frac{H}{2^{H+1}}. \end{aligned} \tag{13}$$

On the other hand, the condition of $d_{N-1} = \mathcal{O}(T^{1/3}/log|K|) - d_1 - \cdots - d_{N-2}$ admits a non-negative solution of d_{N-1} for sufficient large T. This condition implies an upper bound on $H = d_1 + \cdots + d_{N-1} = \mathcal{O}(\log(T^{1/3}\log|K|))$. Finally, combining this upper bound of H with Eq. (13) completes the proof. □

Now we are in the last stage of proving Theorem 1.

Theorem 1. *For $1 \leq m \leq N-1$, let D_m denote the dimension of $\mathcal{X}_m$ and suppose for all $t > 0$, we set $\beta_t = \Theta(\sqrt{\gamma_{t-1} + \ln T})$, $\eta = \Theta(T^{-2/3}\sum_{m=1}^{N-1} D_m \log(\frac{LT^{1/3}}{D_m \log T}))$, and we have an MSET with a uniform partition of each $\mathcal{X}_m$ with diameters $r_m = \frac{D_m}{L}T^{-\frac{1}{3}}\log T$, where the depth parameters d_m follows from Lemma 2. Then LaMBO achieves the expected movement regret*

$$\mathbb{E}[R^+] = \mathcal{O}(\lambda(\sum_{j=1}^{N-1} c_j \sum_{m=1}^{N-1} D_m T^{\frac{2}{3}}(\log T)^2) + \gamma_T\sqrt{T}).$$

Proof. We first bound the ordinary regret. Choose $\beta_t = 1 + \sigma\sqrt{2\left(\gamma_{t-1} + 1 + \ln T\right)}$. Then, with probability $1 - 1/T$, we have

$$\begin{aligned} R(T) &= \sum_{t=1}^{T} f(x_t) - f^* \\ &\overset{(a)}{\leq} \sum_{t=1}^{T} \alpha^t(x_t) - \min_{x\in\mathcal{X}} \alpha^t(x) + 2\beta_t\sigma_{t-1}(x_t) \\ &\overset{(b)}{\leq} \sum_{t=1}^{T} \bar{\ell}_{t,0}(i_t) - \bar{\ell}_{t,0}(i^*) + L\sum_{i=1}^{N-1} r_i T + \sum_{t=1}^{T} 2\beta_t\sigma_{t-1}(x_t), \end{aligned} \tag{14}$$

where (a) follows from Lemma 1 and (b) from the fact that $\bar{\ell}_{t,0}(i^*) = \min_{\mathbf{z}\in\mathcal{Z}} \alpha_t(\mathbf{x}^{t-1}_{1:j-1}, \mathbf{z})$ for some $\mathcal{Z}$ and that f is L-Lipschitz.

Note that when the above inequality fails it only contributes to cumulative regret in expectation by $1/T \times \mathcal{O}(T) = \mathcal{O}(1)$, so we can ignore this term in later calculation.

Now, taking expectation on both sides of Eq. (14) yields

$$
\begin{aligned}
\mathbb{E}[R] &\overset{(c)}{\leq} \mathbb{E}[\sum_{t=1}^{T}(\bar{\ell}_{t,0}(i_t) - \bar{\ell}_{t,0}(i^*))] + L\sum_{i=1}^{N-1} r_i T + \mathcal{O}(\gamma_T\sqrt{T}) \\
&\overset{(d)}{=} \mathbb{E}[\sum_{t=1}^{T}(\tilde{\ell}_{t}(i_t) - \tilde{\ell}_{t}(i^*))] + L\sum_{i=1}^{N-1} r_i T + \mathcal{O}(\gamma_T\sqrt{T}) \\
&\overset{(e)}{=} \mathcal{O}(\sqrt{2^H T \log|\mathcal{K}|} + L\sum_{i=1}^{N-1} r_i T + \gamma_T\sqrt{T}),
\end{aligned}
$$

where (c) follows from Lemma 2, (d) from that $\mathbb{E}[\ell_t] = \mathbb{E}[\bar{\ell}_{t,0} + \sum_{j=0}^{H-1}\bar{\ell}_{t,j}] = \mathbb{E}[\bar{\ell}_{t,0}] + \sum_{j=0}^{H-1}\mathbb{E}[\sigma_{t,j}]\mathbb{E}[\bar{\ell}_{t,j}] = \mathbb{E}[\bar{\ell}_{t,0}]$, and (e) from Lemma 3 where $\frac{\log|K|}{\eta} + \eta T 2^{H+1} = \mathcal{O}(\sqrt{2^{H+1}T\log|\mathcal{K}|})$ for $\eta = \sqrt{2^{-H}T^{-1}\log|\mathcal{K}|}$.

On the other hand, the cumulative movement cost by Lemma 2 is

$$
\sum_{t=1}^{T}\Gamma^t = \mathcal{O}(\sum_{j=1}^{N-1}\sqrt{\lambda}c_j\frac{H}{2^H}T). \tag{15}
$$

From Eq. (15), we plug in $H = \log(T^{1/3}/\log|\mathcal{K}|)$, $r_i = \Theta(\frac{D_i}{L}T^{-1/3}\log T)$ and $|\mathcal{K}| = \Theta(\prod_{i=1}^{N-1} 1/r_i^{D_i})$.

Then, we have

$$
\begin{aligned}
\eta =& \sqrt{2^{-H}T^{-1}\log|\mathcal{K}|} \\
=& \Theta(T^{-2/3}\sum_{i=1}^{N-1} D_i \log(LT^{1/3}/D_i\log T)),
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}[R^+] = \mathbb{E}[R] + \mathbb{E}[\sum_{t=1}^{T}\sqrt{\lambda}\Gamma^t] \\
\leq& \mathcal{O}(\sum_{j=1}^{N-1} c_j T^{\frac{2}{3}}\log T\log|K| + L\sum_{i=1}^{N-1} r_i T + \gamma_T\sqrt{T}) \\
\leq& \mathcal{O}(\sum_{j=1}^{N-1} c_j \sum_{i=1}^{N-1} D_i T^{\frac{2}{3}}(\log T)^2 + \gamma_T\sqrt{T}),
\end{aligned}
$$

which completes the proof. □

A.2 BOUNDS ON THE MAXIMUM MUTUAL INFORMATION FOR COMMON KERNELS

The following lists known bounds of the maximum mutual information for common kernels [Srinivas et al., 2010]:

- $\gamma_t = O(D\log t)$, for linear kernel.
- $\gamma_t = O((\log t)^{D+1})$, for Squared Exponential kernel.
- $\gamma_t = O(t^{\frac{D(D+1)}{2\nu + D(D+1)}}\log t)$ for Matérn kernels with $\nu > 1$,

where D is the dimension of input space.

B PRACTICAL CONSIDERATIONS AND IMPLEMENTATION DETAILS

B.1 DETAILS ON MODEL SELECTION

Below we detail the extensions we use in the experiments to improve the algorithm's performance.

Restart with Epochs: A plausible strategy is to refresh the arm-selection probability every τ iterations to escape from local optimum. In our implementation we choose $\tau = 25$ as the default value.

Adaptive Resolution Increase: In experiments, a simple extension allows LaMBO to discard the arms that have probability of selection being less than a threshold ($\tau = 0.9$ in our implementation), and partition each remaining subset into 2 subsets. We found that combining this with restart can accelerate the optimization in many cases.

Update of Kernel: We choose RBF kernel and Mátern class. As commonly found in practice, we update our kernel hyperparameters every 25 iterations based on the maximum likelihood estimation.

Aggressive learning rate: Our experiments show that constant learning rate $\eta = 1$ usually outperforms the rate $\Theta(T^{-2/3})$ suggested by the theory.

B.2 FURTHER DESIGN OF MSET AND PARTITION STRATEGIES

Construction of MSET: A crucial part of algorithm is in the construction of the MSET, which involves partitioning the variables in each module, and setting the depth parameters (d_i's). For a MSET with $|\mathcal{K}|$ leaves to choose from, LaMBO requires solving $|\mathcal{K}|$ local BO optimization problems per iteration. Hence initially, we partition each variable space of module to two subsets only, and abandon subsets when their arm selection probability p_t is below some threshold. In our experiments, we always set it to be $0.2/|\mathcal{K}|$, where $|\mathcal{K}|$ denotes the number of leaves of MSET. After that, we further divide the remaining subsets again to increase the resolution. This procedure could be iterated upon further although we typically do not go beyond two stages of refinement.

In our implementation, we set the depth parameter to be $d_i = 1$ or $d_i \propto \log \lambda c_i$ when c_i could be estimated in prior. Empirically, we found that the performance is quite robust when $d_i \leq 5$ for the different cost ratios in both synthetic and real experiments we tested. To avoid accumulating cost too fast in early stages of LaMBO, we record the number of times that variable changes in the first module and dynamically increase the first depth parameters d_1 by 1 every 20 iterations when the the number has increased beyond 5 ($1/4$ of the cycle) during the period. In all experiements, we have found this simple add-on perform on par or better than fixing depth parameters through an entire run.

Partitioning method: Although LaMBO achieves theoretical guarantees with a uniform partition, such a partition does not fully leverage the structure of the function. For this reason, the computational complexity can be very large for high-dimensional problems. On the other hand, we observe from our experiments that simple bisection aligned with coordinates yields good performance on many synthetic data and on our neural data. To further improve the performance, we adopt the multi-scale optimization strategy [Wang et al., 2014, Azar et al., 2014], which adaptively increases the partition resolution through iterations. Practically it has often leads to more computational savings, which involves partitioning more finely in regions that have high rewards. A simple version of this strategy is also employed in our experiments where regions are discarded with probability of selection being below some threshold (typically 0.1) and the remaining regions are further partitioned with increasing resolution. Another remedy is to use domain specific information to help restrict the search space or define the hierarchy in the MSET. The generality of the MSET makes it possible to use expert or prior knowledge to constrain switching between specific sets of variables that may be implausible. For instance, in our study of application in neuroscience, there are certain combinations of parameters that would violate certain size constraints related to the underlying biology that could be incorporated into the design of the MSET.

C FURTHER DETAILS ON THE NEUROIMAGING EXPERIMENTS

Figure S1 illustrates the scheme of our brain-imaging experiment without pre-processing. In this brain mapping pipeline, we varied the U-Net training hyperparameters and the 3D reconstruction post-processing hyperparameters. In the first module (U-Net training), we optimized the learning rate $\in [1 \times 10^{-7}, 1 \times 10^{-1}]$ and batch size $\in [4, 12]$ for the U-Net. In the second module, we applied post-processing operations to the U-Net output 3D reconstructions, including label purity $\in [0.51, 0.8]$,

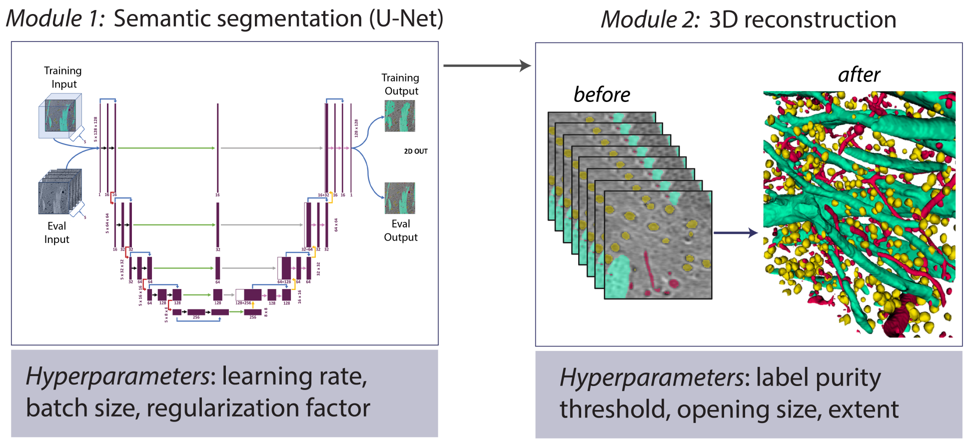

Figure S1: *Pipeline for neuroimage segmentation.* From from left to right, we show the training of U-Net which outputs segmentation of 2D images, and a post-processed 3D reconstruction.

cell opening size $\in [0, 2]$, and a shape parameter (extent) to determine whether uncertain components are either cells or blood vessels $\in [0.3, 0.8]$.

We also performed a 3-module experiment by adding a pre-processing before U-Net training. We varied the pre-processing hyperparameters, U-Net training hyperparameters, and 3D reconstruction post-processing hyperparameters. In the pre-processing, we used a contrast parameter $\in [1, 2]$ and denoising parameter $[1, 15]$ (regularization strength in Non-Local Means [Buades et al., 2011]), and in the second module (U-Net training), we varied the learning rate $\in [1 \times 10^{-5}, 8.192 \times 10^{-2}]$ and batch size $\in [2, 14]$. During the third module (post-processing of 3D reconstructions), we varied label purity $\in [0.51, 0.8]$, cell opening size $\in [0, 2]$, and extent $\in [0.3, 0.8]$. In our experiments, we define the cost to be the aggregate recorded clock time for generating an output after changing a variable in a specific module (see Figure S2, right). To test LaMBO on the problem, we gathered a data set consisting of $606,000$ combinations of hyperparameters by exhaustive search.

D PSEUDO-CODE FOR THE SLOWING MOVING BANDIT ALGORITHM

For completeness, we include a pseudo-code for slowly moving bandit algorithm below.

Algorithm 1 Slowly Moving Bandit (SMB)

Input: A tree $\mathcal{T}$ with a set of finite leaves $K, \eta > 0$.
Initialize: $p_1 = \text{Unif}(K), h_0 = H$ and $i_0 \sim p_1$
for $t = 1$ to T **do**
  Select arm $i_t \sim p_t(\cdot | A_{h_{t-1}}(i_{t-1}))$.
  Let $\sigma_{t,h}$, $h = 1, \dots, H-1$, be i.i.d. Unif$(\{-1, 1\})$.
  let $h_t = \min\{0 \le h \le H : \sigma_{t,h} = -1\}$ where $\sigma_{t,H} = -1$.
  Compute vectors $\bar{\ell}_{t,0}, \dots, \bar{\ell}_{t,H-1}$ recursively via $\bar{\ell}_{t,0}(i) = \frac{\mathbf{1}(i_t = i)}{p_t(i)} \ell_t(t)$, and for all $h \ge 1$:

$$\bar{\ell}_{t,h}(i) = -\frac{1}{\eta} \log \left(\sum_{j \in A_h(i)} \frac{p_t(j) \zeta_{t,h}(j)}{p_t(A_h(i))} \right), \quad \zeta_{t,h}(j) = e^{-\eta(1+\sigma_{t,h-1}) \bar{\ell}_{t,h-1}(j)}.$$

  $\tilde{\ell}_t = \bar{\ell}_{t,h} + \sum_{h=0}^{H-1} \sigma_{t,h} \bar{\ell}_{t,h}$.
  $p_{t+1} = \frac{p_t(i) e^{-\eta \ell_t(i)}}{\sum_{j=1}^{|\mathcal{K}|} p_t(j) e^{-\eta \ell_t(j)}}, \ \forall i \in \mathcal{K}$.
end for

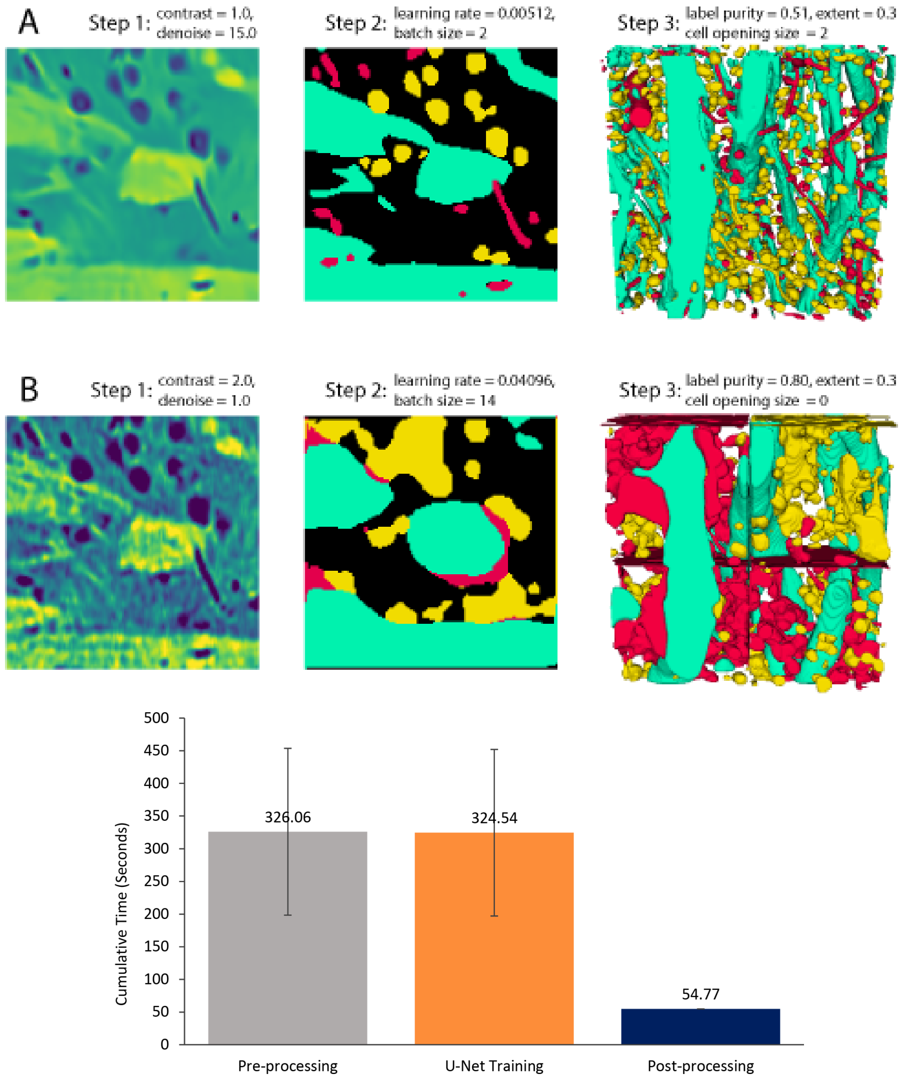

Figure S2: *Results from the three-stage pipeline for the optimal set of parameters and a suboptimal set.* Along the top row (A), we show the results obtained for an optimal set of hyperparameters selected by our approach (as measured by the f1-score). Along the bottom row (B), we show the same results for a suboptimal hyperparameter combination with poor performance. Below, we show the statistics of the timing costs for each stage of the pipeline.

E FURTHER EXPERIMENTS ON SYNTHETIC FUNCTIONS

The synthetic functions used in the experiment are taken from [Surjanovic and Bingham] and [Kandasamy et al., 2016]. We use linear transformation to normalize all the function to the range $[0, 1]$.

We compare LaMBO with other BO algorithms on four synthetic functions, (A) Hartmann 6D, (B) Rastrigin 6D, (C) Ackley 8D, and (D) Griewank 6D. The plots on the top shows the regret performance, the plots on the button show their surface. We observe when objective have multiple local optimum comparable with the global one, LaMBO has comparable performance with the alternative. However, LaMBO performs significantly better than the baselines when the objective has a sharper global optimum. Unlike deterministic decision rule proposed in the alternatives, LaMBO has randomized decision rule and does not rely on the GP regression alone, which allows it to have more incentive for exploration.

Hartmann 6D function:

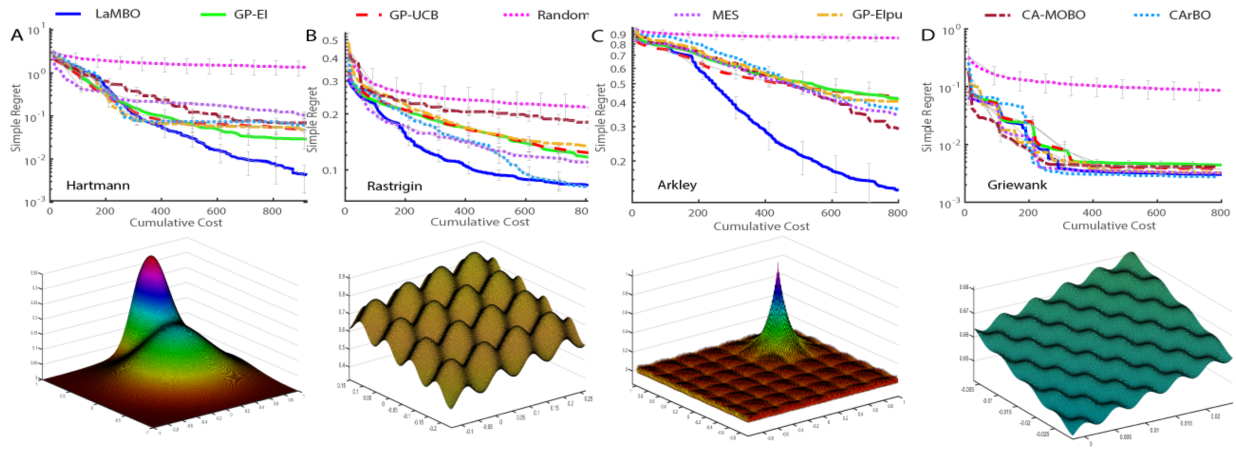

Figure S3: *Synthetic functions.* This section contains details and further experiment results on the synthetic functions. We compare LaMBO with other BO algorithms on four synthetic functions, (A) Hartmann 6D, (B) Rastrigin 6D, (C) Ackley 8D, and (D) Griewank 6D. The plots on the top shows the regret performance and the 3D plots on the bottom show their surface.

The function is $f(x) = \sum_{i=1}^{4} \alpha_i \exp(-\sum_{j=1}^{6} A_{ij}(x_j - P_{ij}))$, where

$$\alpha = [1, 1, 2, 3, 3.2],$$

$$A = \begin{bmatrix} 10 & 3 & 17 & 3.5 & 1.7 \\ 0.05 & 10 & 17 & 0.1 & 8 \\ 3 & 3.5 & 1.7 & 10 & 17 \\ 17 & 8 & 0.05 & 10 & 0.1 \end{bmatrix},$$

$$P = 10^{-4} \times \begin{bmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 0.05 & 10 & 17 & 8 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{bmatrix},$$

and the domain is $[0, 1]^6$.

Ackley 8D function:

$$f(\mathbf{x}) - 20\exp(-0.2\sqrt{\frac{1}{8}\sum_{i=1}^{8} x_i^2}) - \exp\left(\frac{1}{8}\sum_{i=1}^{8}\cos(2\pi x_i)\right)$$
$$+20 + \exp(1),$$

where the domain is $[-32.768, -32.768]^8$.

Rastrigin 6D function:

$$f(\mathbf{x}) = 60 + \sum_{i=1}^{6}\left[x_i^2 - 10\cos(2\pi x_i)\right],$$

where the domain is $[-5.12, 5.12]^6$.

Griewank 6D function:

$$f(\mathbf{x}) = \sum_{i=1}^{6}\frac{x_i^2}{4000} - \prod_{i=1}^{6}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1,$$

where the domain is $[-600, 600]^6$.

$$\alpha_c(\mathbf{x}) = \alpha(\mathbf{x})/c(\mathbf{x})$$

References

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Annual Conference on Learning Theory*, volume 40. Microtome Publishing, 2015.

Mohammad Gheshlaghi Azar, Alessandro Lazaric, and Emma Brunskill. Online stochastic optimization under correlated bandit feedback. In *ICML*, pages 1557–1565, 2014.

Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-local means denoising. *Image Processing On Line*, 1:208–212, 2011.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 844–853. JMLR.org, 2017. URL `http://dl.acm.org/citation.cfm?id=3305381.3305469`.

Kirthevasan Kandasamy, Gautam Dasarathy, Junier B Oliva, Jeff Schneider, and Barnabás Póczos. Gaussian process bandit optimisation with multi-fidelity evaluations. In *Advances in Neural Information Processing Systems*, pages 992–1000, 2016.

Tomer Koren, Roi Livni, and Yishay Mansour. Multi-armed bandits with metric movement costs. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4119–4128. Curran Associates, Inc., 2017a.

Tomer Koren, Roi Livni, and Yishay Mansour. Bandits with movement costs and adaptive pricing. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1242–1268, Amsterdam, Netherlands, 07–10 Jul 2017b. PMLR. URL `http://proceedings.mlr.press/v65/koren17a.html`.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 1015–1022, USA, 2010. Omnipress. ISBN 978-1-60558-907-7. URL `http://dl.acm.org/citation.cfm?id=3104322.3104451`.

S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets. Retrieved February 3, 2020, from `http://www.sfu.ca/~ssurjano`.

Ziyu Wang, Babak Shakibi, Lin Jin, and Nando Freitas. Bayesian multi-scale optimistic optimization. In *Artificial Intelligence and Statistics*, pages 1005–1014. PMLR, 2014.