Supplementary Material

Bayesian Optimistic Optimisation with Exponentially Decaying Regret

1. Review of SOO, BaMSOO, IMGPO algorithms

In the first section of the Supplementary Material, we provide the details of SOO (Munos, 2011) and BamSOO (Wang et al., 2014). The main difference between our proposed BOO algorithm and these algorithms are in the following blue color lines. As we can see, SOO and BaMSOO select a node to be expanded at line 4 in each algorithm. At depth h, among the leaf

---

Algorithm 1 The SOO Algorithm (Munos, 2011)

Input: Parameter m

Initialisation: Set $\mathcal{T}_0 = \{(0, 0)\}$ (root node). Set $p = 1$. Sample initial points to build $\mathcal{D}_0$.

```
1: while True do
2:    Set v_max = −∞
3:    for h = 0 to min(depth(T_p), h_max(p)) do
4:       Among all leaves (h, j) of depth h, select (h, i) ∈ argmax_{(h,j)∈L} f(c_{h,j})
5:       if f(c_{h,i}) ≥ v_max then
6:          Expand node (h, i) by adding m children (h + 1, i_j) to tree T_p
7:          Evaluate all m functional values f(c_{h+1,i_j}), where (h + 1, i_j) are children of (h, i)
8:          Update v_max = f(c_{h,i})
9:          Update p = p + 1
10:      end if
11:   end for
12: end while
```

---

nodes, SOO selects the node with the maximum functional value, BaMSOO selects the node with the maximum value of function g. The function g is defined at line 9 and line 13 in Algorithm 3. Otherwise, the proposed BOO selects the node with the maximum GP-UCB value.

Once a node is selected to be expanded, SOO needs to sample the function at all m children nodes (at line 7 in Algorithm 2), BamSOO needs to sample the function at m' children nodes (at line 9 in Algorithm 3), where $0 \leq m' \leq m$ depending on the condition at line 9 in Algorithm 3. In the worst case, $m' = m$, BamSOO spends m evaluations like SOO. Otherwise, our sampling strategy samples the function only at the parent node. As a result, our strategy requires only one function evaluation irrespective of the value of m. IMGPO (Kawaguchi et al., 2016) is quite similar to BaMSOO except two differences. Frist, IMGPO do not force the tree to a maximum depth of $h_{max}(p)$ like SOO, BamSOO. Second, IMGPO add a strategy to reduce the computation when searching in the tree is inefficient. Please see their paper (Kawaguchi et al., 2016) for details.

1.1. Strict Negative Correlation

As we discussed in section 4.1 of the main paper. Most of tree-based optimistic optimisation algorithms like SOO, StoSOO (Valko et al., 2013), BaMSOO and IMGPO face a *strict negative correlation* between the branch factor m and the number of tree expansions given a fixed function evaluation budget N. In this part, we provide a summary table showing the simple regret (in the worst case) of these algorithms given a fixed function evaluation budget N.

Algorithm	Simple Regret
SOO	$\mathcal{O}(e^{\sqrt{\frac{N}{m}}})$
BaMSOO	$\mathcal{O}((\frac{N}{m})^{-\frac{2\alpha}{D(4-\alpha)}})$
IMGPO	$\mathcal{O}(e^{\sqrt{\frac{N}{m}}})$

Table 1. The simple regret of SOO depends on the near-optimality dimension d. If $d > 0$ then the simple regret is sublinear, if $d = 0$ then the simple regret is exponential as we show in this table. BaMSOO has a sublinear rate because it uses $d = D/\alpha - D/4$ where $\alpha = 1$ or 2. IMGPO uses a fixed $m = 3$. We here generalize their proof to any m.

Algorithm 2 The BaMSOO Algorithm (Wang et al., 2014)

Input: Parameter m

Initialisation: Set $g_{0,0} = f(c_{0,0})$, $f^+ = g_{0,0}$, $t = 1$, $p = 1$, $\mathcal{T}_0 = \{(0, 0)\}$ (root node). Sample initial points to build $\mathcal{D}_0$.

while True **do**
  Set $v_{max} = -\infty$
  for $h = 0$ to min(depth($\mathcal{T}_p$), $h_{max}(p)$) **do**
    Among all leaves (h, j') of depth h, select $(h, j) \in \text{argmax}_{(h,j')\in\mathcal{L}} g(c_{h,j'})$
    if $g(c_{h,j}) \geq v_{max}$ **then**
      for $i = 0$ to $k - 1$ **do**
        Update $p = p + 1$
        if $\mathcal{U}_p(c_{h+1,mj+i}) \geq f^+$ **then**
          Set $g(c_{h+1,mj+i}) = f(c_{h+1,mj+i})$
          Set $t = t + 1$
          $\mathcal{D}_t = \{\mathcal{D}_{t-1}, (c_{h+1,mj+i}, g(c_{h+1,mj+i}))\}$
        else
          Set $g(c_{h+1,mj+i}) = \mathcal{L}_p(c_{h+1,mj+i})$
        end if
        if $g(c_{h+1,mj+i}) > f^+$ **then**
          Set $f^+ = g(c_{h+1,mj+i})$
        end if
      end for
      Add the children of (h, j) to $\mathcal{T}_p$
      Set $v_{max} = g(c_{h,j})$
    end if
  end for
end while

The Table 3 shows the strict negative correlation of tree-based optimistic optimisation algorithms like SOO, BamSOO, IMGPO. The larger m is, the higher the simple regret is. This explains why most of tree-based optimistic optimisation algorithms often use a small value of m like $m = 2$, $m = 3$. In contrast, our algorithm leverages the large value of m to improve the regret bound.

2. Proof of Lemma 1

Lemma 1 (Lemma 1 in the main paper)**.** *Given any $(a, b) \in M(m)$ and a partitioning procedure $P(m; a, b)$, then*

1. *the longest side of a cell at depth h is at most $a^{-\lfloor \frac{bh}{D} \rfloor}$, and*
2. *the smallest side of a cell at depth h is at least $a^{-\lceil \frac{bh}{D} \rceil}$.*

Proof. We prove the statement by induction. At depth $h = 1$, we partition the search space $\mathcal{X}$ into $m = a^b$ cells using the partitioning procedure $P(m; a, b)$. There are two cases on b.

- $b = D$. Then the longest side of a cell at depth $h = 1$ is $1/a = a^{-\lfloor \frac{b}{D} \rfloor}$. Also, the smallest side of a cell at depth $h = 1$ is $1/a = a^{-\lceil \frac{b}{D} \rceil}$.
- $b < D$. Then by the partitioning procedure, the longest side of a cell at depth $h = 1$ is still 1. $a^{-\lfloor \frac{b}{D} \rfloor} = a^0 = 1$. Hence, the longest side of a cell at depth 1 is $a^{-\lfloor \frac{b}{D} \rfloor}$. Also, the smallest side of a cell at depth $h = 1$ is $1/a = a^{-\lceil \frac{b}{D} \rceil}$.

For both cases, the statement is true for $h = 1$. We assume that the statement is true for $h \geq 1$. We consider any cell at depth $h + 1$. By our algorithm, this cell is divided from a cell at depth h. Similar to the case $h = 1$, we also consider two cases on b.

- $b = D$. By the inductive hypothesis, the longest side of a cell at depth h is at most a^{-h}. Then the longest side of a child cell of this cell is $a^{-(h+1)} = a^{-\lfloor \frac{b(h+1)}{D} \rfloor}$. Also, the smallest side of a child cell of this cell is $a^{-(h+1)} = a^{-\lceil \frac{b(h+1)}{D} \rceil}$.
- $b < D$. By the inductive hypothesis, the longest side of a cell at depth h is at most $a^{-\lfloor \frac{bh}{D} \rfloor}$. If we divide a cell at depth h by the partitioning procedure, then the longest side of the sub-cell is at most $a^{-\lfloor \frac{bh}{D} \rfloor}/a = a^{-1-\lfloor \frac{bh}{D} \rfloor}$. However, since $b < D$, $\lfloor \frac{b(h+1)}{D} \rfloor \leq 1 + \lfloor \frac{bh}{D} \rfloor$. It follows that $a^{-1-\lfloor \frac{bh}{D} \rfloor} \leq a^{-\lfloor \frac{b(h+1)}{D} \rfloor}$. Thus, the longest side of a cell at depth $h+1$ is at most $a^{-\lfloor \frac{b(h+1)}{D} \rfloor}$.

  Also, by the inductive hypothesis, the smallest side of a cell at depth h is at least $a^{-\lceil \frac{bh}{D} \rceil}$. If we divide a cell at depth h then the smallest side of the sub-cell is at least $a^{-\lceil \frac{bh}{D} \rceil}/a = a^{-\lceil \frac{bh}{D} \rceil - 1}$. However, since $b < D$, $\lceil \frac{bh}{D} \rceil + 1 \geq \lceil \frac{b(h+1)}{D} \rceil$. As a result, $a^{-\lceil \frac{bh}{D} \rceil - 1} \geq a^{-\lceil \frac{b(h+1)}{D} \rceil}$. Thus, the smallest side of a cell at depth $h+1$ is at least $a^{-\lfloor \frac{b(h+1)}{D} \rfloor}$.

Thus, the statement holds for every $h \geq 1$. □

3. Proof of Lemma 4

To derive an upper bound on variance function σ_p as in Lemma 4, we use a concept, called the *fill distance*. Given a set of points $\mathcal{D}_{p-1}$, we define the fill distance $\text{FD}(\mathcal{D}_{p-1}, \mathcal{X})$ as the largest distance from any point in $\mathcal{X}$ to the points in $\mathcal{D}_{p-1}$, as

$$\text{FD}(\mathcal{D}_{p-1}, \mathcal{X}) = \text{sup}_{x \in \mathcal{X}} \text{inf}_{c_i \in D_{p-1}} ||x - c_i||.$$

The following result, which is proven by Wu & Schaback (1992) [Theorem 5.14], after is reviewed by Kanagawa et al. (2018) [Theorem 5.4], provides an upper bound for the posterior variance in terms of the fill distance. It applies the cases where the kernel whose RKHS is norm-equivalent to the Sobolev space.

Lemma 2 ((Wu & Schaback, 1992; Kanagawa et al., 2018)). *Let k be a kernel on $\mathbb{R}^d$ whose RKHS is norm equivalent to the Sobolev space. There exist constants $h_0 > 0$ and $C' > 0$ satisfying the following: for any $x \in \mathcal{X}$ and any set of observations $\mathcal{D}_{p-1} = \{c_1, c_2, ..., c_{p-1}\} \in \mathcal{X}$ satisfying $FD(\mathcal{D}_{p-1}, \mathcal{X}) \leq h_0$, we have*

$$\sigma_p(x) \leq C' FD(\mathcal{D}_{p-1}, \mathcal{X}))^{\nu - D/2}.$$

It was shown in (Bull, 2011) [Lemma 3] and in (Kanagawa et al., 2018) that the Matérm kernels's RKHS is norm-equivalent to the Sobolev space. Therefore, Lemma 2 is correct all functions satisfying our Assumption 1 and 2 (in Baysian setting).

Based on Lemma 2, we obtain the following result which is similar to Lemma 4 of Vakili et al. (2020) but for the Bayesian setting.

Lemma 3 (Based on Lemma 4 of Vakili et al. (2020)). *There exist constants $h_0 > 0$ and $C' > 0$ satisfying the following: for any $x \in \mathcal{X}$ and any set of observations $\mathcal{D}_{p-1} = \{x_1, x_2, ..., x_{p-1}\} \in \mathcal{X}$ satisfying $FD(\mathcal{D}_{p-1}, \mathcal{X}) \leq h_0$, we have*

$$\sigma_p(x) \leq min_{c_i \in \mathcal{D}_{p-1}} C' ||x - c_i||^{\nu - D/2}$$

Proof. The proof is very similar to their proof. We include it for the purpose of being self-contained. For $x \in \mathcal{X}$, let $c_i \in \mathcal{D}_{p-1}$ be the closet point to x: $||x - c_j|| = \min_{x_i \in \mathcal{D}_{p-1}} ||x - c_i||$. Define $\mathcal{X}' = \mathcal{B}_D(c_j, ||x - c_j||)$, the D-dimensional hyper-ball centered at c_j with radius $||x - c_j||$. Let $X" = \mathcal{D}_{p-1} \cap X'$. The fill distance of the points $X"$ in X' satisfies:

$$\text{FD}(X", X') = \text{sup}_{x' \in X'} \text{inf}_{c_i \in X"} ||x' - c_i|| \leq \text{sup}_{x' \in X'} ||x' - c_j|| = ||x - c_j||.$$

Define $\mu'(x) = \mathbb{E}[f(x)|X"]$ and $k'(x, x') = \mathbb{E}[(f(x) - \mu'(x))(f(x') - \mu'(x'))|X"]$. Let $\sigma'(x) = \sqrt{k'(x, x')}$ be the predictive standard deviation conditioned on observations $X"$. Applying Lemma 2 to $\sigma'(x)$, we have

$$\sigma'(x) \leq C' \text{FD}(X", X')^{\nu - D/2} \leq C' ||x - c_j||^{\nu - D/2}.$$

The lemma holds because $\sigma_p(x) \leq \sigma'(x)$. This is the decreasing monotonicity of the variance function. $\sigma'(x)$ is constructed from set $X'' \subset \mathcal{D}_{p-1}$. A more formal proof that $\sigma_p(x) \leq \sigma'(x)$ can be found in (Chevalier et al., 2014). □

Next, we apply this result to our context in which the set of the sampling points c_i, $\mathcal{D}_{p-1} = \{c_1, ..., c_{p-1}\}$, contains the centers of cells $A_{h,i}$ of a tree structured search space.

Now we prove Lemma 4 in the main paper.

Lemma 4. *Assuming that node $c_{h,i}$ at the depth $h \geq 1$ was sampled at the p-th expansion, where $p \geq h$, then we have that*

$$\sigma_p(c_{h,i}) \leq C_1(\delta(h-1;a,b))^{\nu/2-D/4},$$

where C_1 is a constant.

Proof. By Lemma 3, for every $x \in \mathcal{X}$, we have that

$$\sigma_p(x) \leq \min_{c_i \in \mathcal{D}_{p-1}} C'||x - c_i||^{\nu - D/2}$$

where C' is a constant.

By assumption, node $c_{h,i}$ at depth h is sampled at the p-th expansion, where $p \geq h$. By hierarchical structure of the sampled points, node $c_{h,i}$ is sampled only if its parent node was sampled. We denote this node by $c_{h-1,j}$ which is at depth $h-1$ with some index j. It follows that

$$\begin{aligned} \sigma_p(c_{h,i}) &\leq \min_{c_i \in \mathcal{D}_{p-1}} C'||c_{h,i} - c_i||^{\nu - D/2} \\ &\leq C'||c_{h,i} - c_{h-1,j}||^{\nu - D/2} \\ &\leq C'(L_1 D)^{D/4-\nu/2}(\delta(h-1;a,b))^{\nu/2-D/4}, \end{aligned}$$

where in the first inequality, we apply Lemma 3. In the second inequality, we use the property of $c_{h-1,j} \in \mathcal{D}_{p-1}$, hence $\min_{c_i \in \mathcal{D}_{p-1}} ||c_{h,i} - c_i||^{\upsilon - D/2} \leq ||c_{h,i} - c_{h-1,j}||^{\upsilon - D/2}$. In the last inequality, we have that $c_{h,i}$ belongs to the cell $A_{h-1,j}$ with center $c_{h-1,j}$. Hence, distance $||c_{h,i} - c_{h-1,j}||$ must be shorter than the diameter of that cell. By Lemma and the definition of $\delta(h-1;a,b) = L_1 D a^{-2\lfloor \frac{b(h-1)}{D} \rfloor}$, the last inequality is proven.

Finally, by setting $C_1 = C'(L_1 D)^{D/4-\nu/2}$, the lemma holds. □

4. Proof of Theorem 1

To prove Theorem 1, we will involve two stages:

- **Stage 1**: we first prove that if N is large enough, then under some assumptions, all the centers of nodes of expandable nodes will fall into the ball $\mathcal{B}(x^*, \theta)$ which is centered at x^* with radius θ as defined in Property 1. We prove this in the following Lemma 5.

- **Stage 2**: when a set of expandable nodes fallen into the ball $\mathcal{B}(x^*, \theta)$, the quadratic behaviours of the objective function surrounding the global optimum x^* will occur. We exploit this property to prove that $|I_h| \leq C$, where C is some constant.

Lemma 5. *Assume Algorithm 1 uses partitioning procedure $P(m;a,b)$ where $a = \mathcal{O}(N^{1/D})$ and $b = D$. Thus there exists a constant N_0 such that for every $N \geq N_0$, if (1) node $(h,i) \in I_h$, where $h \geq 2$ and (2) $\mathcal{L}_p(c_{h,i}) \leq f(c_{h,i})$ for every $h \leq p \leq N$, then*

$$c_{h,i} \in \mathcal{B}(x^*, \theta),$$

where $\mathcal{B}(x^*, \theta)$ is the ball centered at x^* with radius θ, which is defined in Property 1.

Proof. By definition, the expansion set $I_h = \{(h,i) | \exists h \leq p \leq N : \mathcal{U}_p(c_{h,i}) \geq f(x^*) - \delta(h;a,b)\}$. Therefore, if node $(h,i) \in I_h$ then there must exist some $h \leq p \leq N$ such that

$$\mathcal{U}_p(c_{h,i}) \geq f(x^*) - \delta(h;a,b) \tag{1}$$

On the other hand, for the same upper confidence bound $\mathcal{U}_p(c_{h,i})$ of $f(c_{h,i})$ as above, we have that

$$\begin{aligned}
\mathcal{U}_p(c_{h,i}) &= \mu_p(c_{h,i}) + \beta_p^{1/2}\sigma_p(c_{h,i}) && (2)\\
&= \mu_p(c_{h,i}) - \beta_p^{1/2}\sigma_p(c_{h,i}) + 2\beta_p^{1/2}\sigma_p(c_{h,i}) && (3)\\
&= \mathcal{L}_p(c_{h,i}) + 2\beta_p^{1/2}\sigma_p(c_{h,i}) && (4)\\
&\leq f(c_{h,i}) + 2\beta_p^{1/2}\sigma_p(c_{h,i}) && (5)\\
&\leq f(c_{h,i}) + 2\beta_p^{1/2}C_1(\delta(p-1;a,b))^{\nu/2-D/4}, && (6)
\end{aligned}$$

where in Eq (5), we use the assumption that $\mathcal{L}_p(c_{h,i}) \leq f(c_{h,i})$. In Eq (6), we use Lemma 4.

Combining Eq (1) and Eq (7), we obtain

$$\begin{aligned}
f(x^*) - f(c_{h,i}) &\leq f(c_{h,i}) + 2\beta_p^{1/2}C_1(\delta(p-1;a,b))^{\nu/2-D/4} && (7)\\
&= L_1Da^{-2\lfloor\frac{bh}{D}\rfloor} + 2\beta_p^{1/2}C_1(L_1Da^{-2\lfloor\frac{b(h-1)}{D}\rfloor})^{\nu/2-D/4} && (8)\\
&= L_1Da^{-2h} + 2\beta_p^{1/2}C_1(L_1D)^{\nu/2-D/4}a^{-(\nu-D/2)(h-1)}, && (9)
\end{aligned}$$

where Eq (8) uses the definition of $\delta(h;a,b)$ and Eq (9) uses the assumption that $b = D$.

We continue to go further with Eq (9) by using the assumptions that $a = \mathcal{O}(N^{1/D})$, $h \geq 2$ (from assumptions of Lemma 5), and $\nu - D/2 > 4$ (from Assumption 1):

$$\begin{aligned}
f(x^*) - f(c_{h,i}) &\leq L_1Da^{-2h} + 2\beta_p^{1/2}C_1(L_1D)^{\nu/2-D/4}a^{-(\nu-D/2)(h-1)} && (10)\\
&\leq L_1Da^{-4} + 2\beta_N^{1/2}C_1(L_1D)^{\nu/2-D/4}a^{-4} && (11)\\
&= \frac{L_1D + 2C_1(L_1D)^{\nu/2-D/4}\sqrt{2log(\pi^2N^3/3\eta)}}{a^4} && (12)\\
&= \mathcal{O}(\frac{\sqrt{log(N/3\eta)}}{N^{4/D}}), && (13)
\end{aligned}$$

where, in Eq (11), we use $h \geq 2$ and the increasing monotonicity of function β_p. We recall that β_p is the trade-off parameter used on our BOO proposed. Formally, $\beta_p = 2log(\frac{\pi^2p^3}{3\eta})$, where $\eta \in (0,1)$. In Eq (13), we use $a = \mathcal{O}(N^{1/d})$.

We have that $\frac{\sqrt{log(N/3\eta)}}{N^{4/D}} \to 0$ as $N \to \infty$. Therefore, for any $\epsilon_0 > 0$, there exists a constant $N_0 > 0$ such that for every $N \geq N_0$, $f(x^*) - f(c_{h,i}) \leq \epsilon_0$. Thus, by definition of $\mathcal{B}(x^*,\theta)$ in Property 1, $c_{h,i} \in \mathcal{B}(x^*,\theta)$. □

We now start to prove Theorem 1.

Theorem 1. *Assume that the proposed BOO algorithm uses partitioning procedure $P(m;a,b)$ where $a = \mathcal{O}(N^{1/D})$ and $b = D$. We consider set I_h, where $h \geq 2$ and assume that $\mathcal{L}_p(c_{h,i}) \leq f(c_{h,i}) \leq \mathcal{U}_p(c_{h,i})$ for all node $(h,i) \in I_h$ and for all $h \leq p \leq N$. Then there exist constants $N_1 > 0$ and $C > 0$ such that for every for $N \geq N_1$,*

$$|I_h| \leq C.$$

Proof. The proof involves three steps.

Step 1: **for each node** $(h,i) \in I_h$**, we seek to bound gap** $||x^* - c_{h,i}||$.

By Lemma 5 and the assumptions of Lemma 1, there exists a constant N_0 such that for every $N \geq N_0$, for any $(h,i) \in I_h$ then $c_{h,i} \in \mathcal{B}(x^*,\theta)$. Hence following Property 1, for any $(h,i) \in I_h$, the following result is guaranteed:

$$L_2||x^* - c_{h,i}||^2 \leq f(x^*) - f(c_{h,i}). \quad (14)$$

On the other hand, by definition of I_h, there exists $h \leq p \leq N$ such that

$$\mathcal{U}_p(c_{h,i}) \geq f(x^*) - \delta(h;a,b). \quad (15)$$

Combining Eq (14) and Eq (15), we have that

$$L_2||x^* - c_{h,i}||^2 \leq \mathcal{U}_p(c_{h,i}) + \delta(h; a, b) - f(c_{h,i}). \tag{16}$$

Similar to Lemma 5, we continue to analyze the right hand side of Eq (16) as follows:

$$\begin{aligned}
L_2||x^* - c_{h,i}||^2 &\leq \mathcal{U}_p(c_{h,i}) + \delta(h; a, b) - f(c_{h,i}) & (17)\\
&= \mu_p(c_{h,i}) + \beta_p^{1/2}\sigma_p(c_{h,i}) + \delta(h; a, b) - f(c_{h,i}) & (18)\\
&= \mu_p(c_{h,i}) - \beta_p^{1/2}\sigma_p(c_{h,i}) + 2\beta_p^{1/2}\sigma_p(c_{h,i}) + \delta(h; a, b) - f(c_{h,i}) & (19)\\
&\leq \mathcal{L}_p(c_{h,i}) + 2\beta_p^{1/2}\sigma_p(c_{h,i}) + \delta(h; a, b) - f(c_{h,i}) & (20)\\
&\leq 2\beta_p^{1/2}\sigma_p(c_{h,i}) + \delta(h; a, b) & (21)\\
&\leq 2\beta_p^{1/2}C_1(\delta(p-1; a, b))^{\nu/2-D/4} + \delta(h; a, b) & (22)\\
&\leq 2\beta_N^{1/2}C_1(\delta(h-1; a, b))^{\nu/2-D/4} + \delta(h; a, b), & (23)
\end{aligned}$$

where in Eq (18), we use the definition of $\mathcal{U}_p(c_{h,i})$, in Eq (20), we use the definition of $\mathcal{L}_p(c_{h,i})$. In Eq (21), we use the assumption that $\mathcal{L}_p(c_{h,i}) \leq f(c_{h,i})$. In Eq (22), we use Lemma 4. Finally, in the last inequality at Eq (23), we use the decreasing monotonicity of function $\delta(h; a, b)$ and the increasing monotonicity of function β_p. By assumption that $h \leq p \leq N$, hence $\delta(p-1; a, b) \leq \delta(h-1; a, b)$ and $\beta_p^{1/2} \leq \beta_N^{1/2}$. We recall that $\delta(h; a, b) = L_1 D a^{-2\lfloor \frac{bh}{D} \rfloor}$ as in Definition 1.

Thus, for any $(h, i) \in I_h$, where $h \geq 2$, we have that

$$L_2||x^* - c_{h,i}||^2 \leq 2\beta_N^{1/2}C_1(\delta(h-1; a, b))^{\nu/2-D/4} + \delta(h; a, b).$$

Step 2: **Bounding $|I_h|$ using covering balls**.

We let Ω_h be the set of nodes (h, i) at depth h generated by partitioning procedure $P(m; a, b)$. From Ω_h we define set $\overline{I}_h$ as

$$\overline{I}_h = \{(h, i) \in \Omega_h \text{ such that } L_2||x^* - c_{h,i}||^2 \leq 2\beta_N^{1/2}C_1(\delta(h-1; a, b))^{\nu/2-D/4} + \delta(h; a, b)\}.$$

By this definition, $I_h \subseteq \overline{I}_h$ which implies directly that $|I|_h \leq |\overline{I}_h|$. Now we consider the set of points $c_{h,i}$ of these nodes. This set is defined as

$$\overline{P}_h = \{c_{h,i} \in \mathcal{X} | (h, i) \in \overline{I}_h\}.$$

We can see that all the points of $\overline{P}_h$ are covered by a hypersphere centered at x^* with radius $\sqrt{\frac{2\beta_N^{1/2}C_1(\delta(h-1;a,b))^{\nu/2-D/4}+\delta(h;a,b)}{L_2}}$. We call this hypersphere $\mathcal{S}_h$.

On the other hand, by Lemma 3, the smallest side of a cell $A_{h,i}$ at depth h is at least $a^{\lceil -\frac{bh}{D} \rceil}$. Therefore, if we bound a point $c_{h,i} \in \overline{I}_h$ by a D-ball centered $c_{h,i}$ with radius $a^{\lceil -\frac{bh}{D} \rceil}/2$ then all these balls are disjoint. Further, even if there are several centers $c_{h,i}$ of these balls lying on the boundary of $\mathcal{S}_h$ then all these balls must be within the hypersphere centered at x^* with radius

$$\sqrt{\frac{2\beta_N^{1/2}C_1(\delta(h-1; a, b))^{\nu/2-D/4} + \delta(h; a, b)}{L_2}} + a^{\lceil -\frac{bh}{D} \rceil}/2.$$

Thus, $|\overline{P}_h|$ cannot exceed the number of disjoint balls which fit in the hypersphere centered at x^* with radius $\sqrt{\frac{2\beta_N^{1/2}C_1(\delta(h-1;a,b))^{\nu/2-D/4}+\delta(h;a,b)}{L_2}} + a^{\lceil -\frac{bh}{D} \rceil}/2$.

The number of these disjoint balls cannot exceed the proportion of the volume of the hypersphere of radius $\sqrt{\frac{\delta(h)+2\beta_N^{1/2}C_2L^{-\nu+D/2}(\delta(h-1))^{\nu-D/2}}{C_1}}$ and the volume of small balls of radius $a^{\lceil -\frac{bh}{D} \rceil}/2$. This proportion is measured by

$$(\frac{\sqrt{\frac{2\beta_N^{1/2}C_1(\delta(h-1;a,b))^{\nu/2-D/4}+\delta(h;a,b)}{L_2}}+a^{\lceil -\frac{bh}{D}\rceil}/2}{a^{\lceil -\frac{bh}{D}\rceil}/2})^D.$$

Thus, we have that

$$\begin{aligned}|\overline{P}_h| &\leq (\frac{\sqrt{\frac{2\beta_N^{1/2}C_1(\delta(h-1;a,b))^{\nu/2-D/4}+\delta(h;a,b)}{L_2}}+a^{\lceil -\frac{bh}{D}\rceil}/2}{a^{\lceil -\frac{bh}{D}\rceil}/2})^D &(24)\\ &= (\sqrt{\frac{4\beta_N^{1/2}C_1(\delta(h-1;a,b))^{\nu/2-D/4}+2\delta(h;a,b)}{L_2a^{2\lceil -\frac{bh}{D}\rceil}}}+1)^D &(25)\end{aligned}$$

However, by definition of $\overline{I}_h$ and $\overline{P}_h$, $|I_h| \leq |\overline{I}_h| = |\overline{P}_h|$. Therefore, we have

$$|I_h| \leq (\sqrt{\frac{4\beta_N^{1/2}C_1(\delta(h-1;a,b))^{\nu/2-D/4}+2\delta(h;a,b)}{L_2a^{2\lceil -\frac{bh}{D}\rceil}}}+1)^D.$$

Step 3: proving that there exists a constant C such that $|I_h| \leq C$.

Using the assumption that $b = D$, we have $a^{\lceil -2\frac{bh}{D}\rceil} = a^{-2h}$, $\delta(h-1;a,b) = L_1Da^{-2(h-1)}$, and $\delta(h;a,b) = L_1Da^{-2h}$. Replacing these results to Eq (25), we get

$$\begin{aligned}|I_h| &\leq (\sqrt{\frac{4\beta_N^{1/2}C_1(\delta(h-1;a,b))^{\nu/2-D/4}+2\delta(h;a,b)}{L_2a^{2\lceil -\frac{bh}{D}\rceil}}}+1)^D &(26)\\ &= (\sqrt{\frac{4\beta_N^{1/2}C_1(L_1D)^{\nu/2-D/4}a^{-(\nu-D/2)(h-1)}+2L_1Da^{-2h}}{L_2a^{-2h}}}+1)^D &(27)\\ &= (\sqrt{\frac{4C_1(L_1D)^{\nu/2-D/4}}{L_2}\times\beta_N^{1/2}\times a^{2h-(\nu-D/2)(h-1)}+\frac{2L_1}{L_2}}+1)^D &(28)\\ &\leq (\sqrt{\frac{4C_1(L_1D)^{\nu/2-D/4}}{L_2}\times\beta_N^{1/2}\times a^{4+D/2-\nu}+\frac{2L_1}{L_2}}+1)^D &(29)\\ &= C'(\sqrt{log(N/3\eta)}\times N^{(4+D/2-\nu)/D})^{D/2} &(30)\end{aligned}$$

where, Eq (29) holds because $a^{2h-(\nu-D/2)(h-1)} \leq a^{4+D/2-\nu}$. Indeed, by using the assumption that $\nu > 4 + D/2$ and $h \geq 2$, we have that

$$\begin{aligned}a^{2h-(\nu-D/2)(h-1)} &= a^{h(2+D/2-\nu)+(\nu-D/2)}\\ &\leq a^{2(2+D/2-\nu)+(\nu-D/2)}\\ &= a^{4+D/2-\nu}.\end{aligned}$$

For the last inequality at Eq (30), we use the assumption $a = \mathcal{O}(N^{1/D})$, $\beta_N = 2log(\pi^2N^3/3\eta)$, where $\eta \in (0,1)$, and the fact that $\frac{4C_1(L_1D)^{\nu/2-D/4}}{L_2}$ and $\frac{2L_1}{L_2}$ are constants independent of N. Thus, such a constant C' at Eq (30) exists.

Since $\nu > 4+D/2$, we have that $\frac{\sqrt{log(N/3\eta)}}{N^{\frac{\nu-D/2-4}{D}}} \to 0$ as $N \to \infty$. Therefore $\sqrt{log(N/3\eta)}\times N^{(4+D/2-\nu)/D} \to 0$ as $N \to \infty$. Thus, there exists constant $N_1 > 0$ and $C > 0$ such that for every $N \geq N_1$, $|I_h| \leq C$ for every $h \geq 2$. □

5. Proof of Lemma 5

Let (h_p^*+1, i^*) be an optimal node of depth h_p^*+1 (i.e., $x^* \in A_{h_p^*+1,i^*}$). We define a node (h,i) at depth h as $\delta(h;a,b)$-optimal if $\mathcal{U}(c_{h,i}) \geq f(c_{h,i}) - \delta(h;a,b)$. We obtains the following result.

Lemma 6. *Assume that* $f(c_{h_p^*+1,i^*}) \leq \mathcal{U}(c_{h_p^*+1,i^*})$. *Then any node* (h_p^*+1, i) *of depth* h_p^*+1 *before* (h_p^*+1, i^*) *is expanded, is* $\delta(h_p^*+1; a, b)$*-optimal.*

Proof. If the node (h_p^*+1, i^*) has not been expanded yet, then by Algorithm 1 (line 4) we have that $\mathcal{U}(c_{h_p^*+1,i}) \geq \mathcal{U}(c_{h_p^*+1,i^*})$. Combining with the assumptions, we get

$$\begin{aligned} \mathcal{U}(c_{h_p^*+1,i}) &\geq \mathcal{U}(c_{h_p^*+1,i^*}) & (31)\\ &\geq f(c_{h_p^*+1,i^*}) & (32)\\ &\geq f^* - \delta(h_p^*+1; a, b), & (33) \end{aligned}$$

where Eq (32) use the assumption that $f(c_{h_p^*+1,i^*}) \leq \mathcal{U}(c_{h_p^*+1,i^*})$, and Eq (33) use Lemma 3. Thus, the lemma holds. □

From Lemma 12, we deduce that once an optimal node of depth h is expanded, it takes at most $|I_{h+1}|$ node expansions at depth $h+1$ before the optimal node of depth $h+1$ is expanded. From that observation, we deduce the following lemma (corresponding to Lemma 5 in the main paper.)

Lemma 7. *Assume that* $f(c_{h,i}) \leq \mathcal{U}(c_{h,i})$ *for all optimal node* (h, i) *at each depth* $0 \leq h \leq h_{max}(n)$. *Then for any depth* $0 \leq h \leq h_{max}(n)$*, whenever* $n \geq h_{max}(n) \sum_{i=0}^{h} |I_i|$*, we have* $h_n^* \geq h$.

Proof. We prove it by induction. For $h = 0$, we have $h_n^* \geq 0$.

Assume that the proposition is true for all $o \leq h \leq h_0$ with $h_0 < h_{max}(n)$. Let us prove that it is also true for $h_0 + 1$. Let $n \geq h_{max}(n)(|I_0| + |I_1| + ... + |I_{h_0+1}|)$. Since $n \geq h_{max}(n)(|I_0| + |I_1| + ... + |I_{h_0}|)$, we have $h_n^* \geq h_0$. If $h_n^* \geq h_0 + 1$ then the proof is finished. If $h_n^* = h_0$, we consider the nodes of depth $h_0 + 1$ that are expanded. We have seen that as long as the optimal node of depth $h_0 + 1$ is not expanded, any node of depth $h_0 + 1$ that is expanded must be $\delta(h_0 + 1; a, b)$-optimal, i.e., belongs to I_{h_0+1}. Since there are $|I_{h_0+1}|$ of them, after $h_{max}(n)|I_{h_0+1}|$ node expansions, the optimal one must be expanded, thus $h_n^* \geq h_0 + 1$. □

6. Proof of Lemma 6

We use A_p to denote the set of all points evaluated by the algorithm and all centers of optimal nodes of the tree $\mathcal{T}_p$ after p evaluations.

Lemma 8. *Pick a* $\eta \in (0, 1)$. *Set* $\beta_p = 2log(\pi^2 p^3/3\eta)$ *and* $\mathcal{L}_p(c) = \mu_p(c) - \beta_p^{1/2}\sigma_p(c)$. *With probability* $1 - \eta$*, we have*

$$\mathcal{L}_p(c) \leq f(c) \leq \mathcal{U}_p(c),$$

for every $p \geq 1$ *and for every* $c \in A_p$.

Proof. After p evaluations, there are at most p evaluated points by the algorithm. On the other hand, after p evaluations, the deepest depth of the tree $\mathcal{T}_p$ is p. In addition, at each depth, there is only one optimal node which contains x^*. Therefore, there are at most p centers of optimal nodes which belong to tree $\mathcal{T}_p$. Thus, $|A_p| \leq 2p$.

The proof is similar to Lemma 5.1 in (Srinivas et al., 2012) and Lemma 4 in (Wang et al., 2014) with the set A_p (here we use the fact that f is a sample from the GP). If we let $\beta_p = 2log(\pi^2 p^2 |A_p|/6\eta)$, then with probability $1 - \eta$, we have

$$\mathcal{L}_p(c) \leq f(c) \leq \mathcal{U}_p(c),$$

for every $p \geq 1$ and for every $c \in A_p$. Since $|A_p| \leq 2p$, we will use $\beta_p = 2log(\pi^2 p^3/3\eta)$ instead and the lemma also holds with this $\beta_p = 2log(\pi^2 p^3/3\eta)$. □

Lemma 6 implies that with probability $1 - \eta$, all conditions $\mathcal{L}_p(c) \leq f(c) \leq \mathcal{U}_p(c)$ in Lemma 4, Theorem 1, and Lemma 5 in the main paper hold for every $1 \leq p \leq N$.

7. Proof of Theorem 2

Theorem 2 (Regret Bound)**.** *Assume that there is a partitioning procedure $P(m;a,b)$ where $a=\mathcal{O}(N^{1/D})$, $b=D$ and $2\le m<\sqrt{N}-1$. Let the depth function $h_{max}(p)=\sqrt{p}$. We consider $m^2<p\le N$, and define $h(p)$ as the smallest integer h such that*

$$h\ge\frac{\sqrt{p}-m-1}{C}+2,$$

where C is the constant defined by Theorem 1. Pick a $\eta\in(0,1)$. Then for every $N\ge N_1$, the loss is bounded as

$$r_p \quad\le\quad \delta(min\{h(p),\sqrt{p}+1\};a,b)+4C_1\beta_p^{1/2}(\delta(min\{h(p)-1,\sqrt{p}\};a,b))^{\nu/2-D/4},$$

with probability $1-\eta$, where N_1 is the constant defined in Theorem 1, C_1 is the constant defined in lemma 4 and $\beta_N=\sqrt{2log(\pi^2N^3/3\eta)}$.

Proof. By Theorem 1, the definition of $h(p)$ and the facts that $|I_0|=1$ and $|I_1|\le m$, we have

$$\begin{aligned}\sum_{l=0}^{h(p)-1}|I_l| &= |I_0|+|I_1|+(|I_2|+...+|I|_{h(p)-1})\\ &\le 1+m+C(h(p)-2)\le\sqrt{p}\end{aligned}$$

Therefore, $\sum_{l=0}^{h(p)-1}|I_l|\le\sqrt{p}$. By Lemma 5 when $h(p)-1\le h_{max}(p)=\sqrt{p}$, we have $h_p^*\ge h(p)-1$. If $h(p)-1>\sqrt{p}$ then $h_p^*=h_{max}(p)=\sqrt{p}$ since the BOO algorithm does not expand nodes beyond depth $h_{max}(p)$. Thus, in all cases, $h_p^*\ge\min\{h(p)-1,\sqrt{p}\}$.

Let (h,j) be the deepest node in $\mathcal{T}_p$ that has been expanded by the algorithm up to p expansions. Thus $h\ge h_p^*$. By Algorithm 1, we only expand a node when its GP-UCB value is larger than v_{max} which is updated at Line 10 of Algorithm 1. Thus, since the node (h,j) has been expanded, its GP-UCB value is at least as high as that of the some node (h_p^*+1,o) at depth h_p^*+1, such that

- (1) node (h_p^*+1,o) has been evaluated at some p'-th expansion before node (h,j) and
- (2) $(h_p^*+1,o)\in\text{argmax}_{(h_p^*+1,i)\in\mathcal{L}}\mathcal{U}_{p'}(c_{h_p^*+1,i})$ (see Line 4 of Algorithm 1).

We let node (h_p^*+1,o^*) be the optimal node at depth h_p^*+1. With probability $1-\eta$,

$$\begin{aligned}f(x^*)-\delta(h_p^*+1;a,b) &\le f(c_{h_p^*+1,o^*}) &(34)\\ &\le \mathcal{U}_{p'}(c_{h_p^*+1,o^*}) &(35)\\ &\le \mathcal{U}_{p'}(c_{h_p^*+1,o}) &(36)\\ &\le \mu_{p'}(c_{h_p^*+1,o})+\beta_{p'}^{1/2}\sigma_{p'}(c_{h_p^*+1,o}) &(37)\\ &\le \mu_{p'}(c_{h_p^*+1,o})-\beta_{p'}^{1/2}\sigma_{p'}(c_{h_p^*+1,o})+2\beta_{p'}^{1/2}\sigma_{p'}(c_{h_p^*+1,o}) &(38)\\ &\le \mathcal{L}_{p'}(c_{h_p^*+1,o})+2\beta_{p'}^{1/2}\sigma_{p'}(c_{h_p^*+1,o}) &(39)\\ &\le f(c_{h_p^*+1,o})+2\beta_{p'}^{1/2}\sigma_{p'}(c_{h_p^*+1,o}) &(40)\\ &\le \mathcal{U}_p(c_{h,j})+2\beta_{p'}^{1/2}\sigma_{p'}(c_{h_p^*+1,o}), &(41)\end{aligned}$$

where in Eq (34), we use Lemma 3. Eq (35) holds with probability $1-\eta$ by using Lemma 6. In Eq (36), we use the above condition (2). Eq (37) uses the definition of $\mathcal{U}_{p'}$. Eq (39) uses the definition of $\mathcal{L}_{p'}$. Eq (40) holds with probability $1-\eta$ by using Lemma 6. Finally, Eq (41) uses the updating condition at Line 5 and Line 10 of Algorithm 1.

Eq (41) implies that with probability $1-\eta$,

$$f(x^*)-\mathcal{U}_p(c_{h,j})\le\delta(h_p^*+1;a,b)+2\beta_{p'}^{1/2}\sigma_{p'}(c_{h_p^*+1,o}).$$

On the other hand, by Lemma 6, with probability $1-\eta$, we have

$$\begin{aligned} U_p(c_{h,j}) &= \mu_p(c_{h,j}) + \beta_p^{1/2}\sigma_p(c_{h,j}) \\ &= \mathcal{L}_p(c_{h,j}) + 2\beta_p^{1/2}\sigma_p(c_{h,j}) \\ &\leq f(c_{h,j}) + 2\beta_p^{1/2}\sigma_p(c_{h,j}) \end{aligned}$$

Combining these two results, we have

$$f(x^*) - f(c_{h,j}) \leq \delta(h_p^* + 1; a, b) + 2\beta_{p'}^{1/2}\sigma_{p'}(c_{h_p^*+1,o}) + 2\beta_p^{1/2}\sigma_p(c_{h,j}),$$

with a probability $1-\eta$.

Finally, by using Lemma 4 to bound $\sigma_{p'}(c_{h_p^*+1,o})$ and $\sigma_p(c_{h,j})$ and using the fact that the function $\delta(*; a, b)$ decreases with their depths, we achieve

$$\begin{aligned} r_p &\leq f(x^*) - f(c_{h,j}) \\ &\leq \delta(\min\{h(p), \sqrt{p}+1\}; a, b) + 4C_1\beta_p^{1/2}(\delta(\min\{h(p)-1, \sqrt{p}\}; a, b))^{\nu/2-D/4} \end{aligned}$$

with a probability $1-\eta$. □

8. Proof of Corollary 1

Corollary 1. *Pick a* $\eta \in (0,1)$. *There exists a constant* $N_2 > 0$ *such that for every* $N \geq N_2$ *we have that the simple regret of the proposed BOO with the partitioning procedure* $P(m; a, b)$ *where* $a = \lfloor(\frac{\sqrt{N}}{2})^{\frac{1}{D}}\rfloor$, $b = D$, *is bounded as*

$$r_N \leq \mathcal{O}(N^{-\sqrt{N}}),$$

with probability $1-\eta$.

Proof. With $a = \lfloor(\frac{\sqrt{N}}{2})^{\frac{1}{D}}\rfloor$ and $b = D$, $m = a^b \leq \sqrt{N}/2$. These conditions satisfy the assumptions of Theorem 2, therefore following Theorem 2 with probability $1-\eta$, we have that

$$r_N \leq \underbrace{\delta(\min\{h(N), \sqrt{N}+1\}; a, b)}_{\text{Term 1}} + \underbrace{4C_1\beta_N^{1/2}(\delta(\min\{h(N)-1, \sqrt{N}\}; a, b))^{\nu/2-D/4}}_{\text{Term 2}}.$$

We consider Term 1. There are two cases:

(1) If $\min\{h(N), \sqrt{N}+1\} = \sqrt{N}+1$ then $\delta(\min\{h(N), \sqrt{N}+1\}; a, b) = \delta(\sqrt{N}+1; a, b) = L_1 D a^{-2(\sqrt{N}+1)} \leq \mathcal{O}(N^{-\sqrt{N}})$ by replacing $a = \lfloor(\frac{\sqrt{N}}{2})^{\frac{1}{D}}\rfloor$.

(2) If $\min\{h(N), \sqrt{N}+1\} = h(N)$. By definition of $h(N)$ in Theorem 2, $h(N) \geq \frac{\sqrt{N}-m-1}{C} + 2 \geq \frac{\sqrt{N}}{2C} - \frac{1}{C} + 2$. Therefore, $\delta(\min\{h(N), \sqrt{N}+1\}; a, b) = \delta(h(N); a, b) = L_1 D a^{-2h(N)} \leq \mathcal{O}(N^{-\sqrt{N}})$.

Thus, for both cases, Term 1 is bounded by $O(N^{-\sqrt{N}})$. We now consider Term 2. There are also two cases:

(1) If $\min\{h(N)-1, \sqrt{N}\} = \sqrt{N}$ then $4C_1\beta_N^{1/2}(\delta(\min\{h(N)-1, \sqrt{N}\}; a, b))^{\nu/2-D/4} = 4C_1\beta_N^{1/2}(\delta(\sqrt{N}; a, b))^{\nu/2-D/4} = 4C_1L_1D\beta_N^{1/2}a^{-2(\nu/2-D/4)\sqrt{N}} \leq 4C_1L_1D\beta_N^{1/2}a^{-4\sqrt{N}}$. In the last inequality, we use the assumption that $\nu > 4 + D/2$. The component $a^{-4\sqrt{N}}$ with $a = \lfloor(\frac{\sqrt{N}}{2})^{\frac{1}{D}}\rfloor$ dominates β_N which is $\mathcal{O}(\sqrt{N})$. Therefore Term 2 is bounded by $O(N^{-\sqrt{N}})$.

(2) If $\min\{h(N)-1, \sqrt{N}\} = h(N)-1$. By definition of $h(N)$ in Theorem 2, $h(N)-1 \geq \frac{\sqrt{N}-m-1}{C} + 1 \geq \frac{\sqrt{N}}{2C} - \frac{1}{C} + 1$. Then $4C_1\beta_N^{1/2}(\delta(\min\{h(N)-1, \sqrt{N}\}; a, b))^{\nu/2-D/4} = 4C_1\beta_N^{1/2}(\delta(h(N)-1; a, b))^{\nu/2-D/4} = 4C_1L_1D\beta_N^{1/2}a^{-2(\nu/2-D/4)(h(N)-1)} \leq 4C_1L_1D\beta_N^{1/2}a^{-4(h(N)-1)}$. By the argument similar as above, we have that Term 2 is bounded by $O(N^{-\sqrt{N}})$.

Finally, for all cases, we get that $r_N \leq O(N^{-\sqrt{N}})$ with probability $1-\eta$. □

References

Bull, A. D. Convergence rates of efficient global optimization algorithms. *J. Mach. Learn. Res.*, 12:2879–2904, November 2011. ISSN 1532-4435.

Chevalier, C., Ginsbourger, D., and Emery, X. Corrected kriging update formulae for batch-sequential data assimilation. In Pardo-Igúzquiza, E., Guardiola-Albert, C., Heredia, J., Moreno-Merino, L., Durán, J. J., and Vargas-Guzmán, J. A. (eds.), *Mathematics of Planet Earth*, pp. 119–122, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences, 2018.

Kawaguchi, K., Kaelbling, L. P., and Lozano-Pérez, T. Bayesian optimization with exponential convergence, 2016. URL https://arxiv.org/abs/1604.01348.

Munos, R. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24*, pp. 783–791. Curran Associates, Inc., 2011.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Trans. Inf. Theor.*, 58(5):3250–3265, May 2012. ISSN 0018-9448. doi: 10.1109/TIT.2011.2182033. URL http://dx.doi.org/10.1109/TIT.2011.2182033.

Vakili, S., Picheny, V., and Durrande, N. Regret bounds for noise-free bayesian optimization, 2020.

Valko, M., Carpentier, A., and Munos, R. Stochastic simultaneous optimistic optimization. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 19–27, 2013.

Wang, Z., Shakibi, B., Jin, L., and de Freitas, N. Bayesian multi-scale optimistic optimization. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pp. 1005–1014, 2014.

Wu, Z. and Schaback, R. Local error estimates for radial basis function interpolation of scattered data. *IMA J. Numer. Anal*, 13:13–27, 1992.