A Proofs

We begin by introducing some notation. Let $x^*_h$ denote the optimal node at level $h$. That is the cell of $x^*_h$ contains the optimizer $x^*$. Also let $f^+$ and $x^+$ represent the best function value observed thus far and the associated node respectively.

A.1 Technical Lemmas

Lemma 1 (Lemma 5 of de Freitas et al. (2012)). Given a set of points $x_{1:T} := \{x_1, \ldots, x_T\} \in \mathcal{D}$ and a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ with kernel $\kappa$ the following bounds hold:

1. Any $f \in \mathcal{H}$ is Lipschitz continuous with constant $\|f\|_{\mathcal{H}} L$, where $\| \cdot \|_{\mathcal{H}}$ is the Hilbert space norm and $L$ satisfies the following:

$$L^2 \leq \sup_{x \in \mathcal{D}} \partial_x \partial_{x'} \kappa(x, x')|_{x=x'}$$

and for $\kappa(x, x') = \kappa(x - x')$ we have

$$L^2 \leq \partial_x^2 \kappa(x)|_{x=0}.$$  

2. The projection operator $P_{1:T}$ on the subspace $\text{span}\{\kappa(x_t, \cdot)\}_{t=1:T}$ is given by $P_{1:T} = k(\cdot)K^{-1}(\kappa(\cdot), f)$ where $k(\cdot) = k_{1:T}(\cdot) := [\kappa(x_1, \cdot) \cdots \kappa(x_T, \cdot)]^\top$ and $K := [\kappa(x_i, x_j)]_{i,j=1:T}$; moreover, we have that

$$\langle k(\cdot), f \rangle := \begin{bmatrix} \langle \kappa(x_1, \cdot), f \rangle \\ \vdots \\ \langle \kappa(x_T, \cdot), f \rangle \end{bmatrix} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_T) \end{bmatrix}.$$  

Here $P_{1:T}P_{1:T} = P_{1:T}$ and $\|P_{1:T}\| \leq 1$ and $\|1 - P_{1:T}\| \leq 1$.

3. Given tuples $(x_i, f_i)$ with $f_i = f(x_i)$, the minimum norm interpolation $\bar{f}$ with $\bar{f}(x_i) = f(x_i)$ is given by $\bar{f} = P_{1:T}f$. Consequently its residual $g := (1 - P_{1:T})f$ satisfies $g(x_i) = 0$ for all $x_i \in x_{1:T}$.

Lemma 2 (Lemma 6 of de Freitas et al. (2012)). Under the assumptions of Lemma 1 it follows that

$$|f(x) - P_{1:T}f(x)| \leq \|f\|_{\mathcal{H}} \sigma_T(x),$$

where $\sigma_T^2(x) = \kappa(x, x) - k_{1:T}(x)K^{-1}k_{1:T}(x)$ and this bound is tight. Moreover, $\sigma_T^2(x)$ is the posterior predictive variance of a Gaussian process with the same kernel.

Lemma 3 (Adapted from Proposition 1 of de Freitas et al. (2012)). Let $\kappa : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ be a kernel that is twice differentiable along the diagonal $\{(x, x) \mid x \in \mathbb{R}^D\}$, with $L$ defined as in Lemma 1.1, and $f$ be an element of the RKHS with kernel $\kappa$. If $f$ is evaluated at point $x$, then for any other point $y$ we have $\sigma_T(y) \leq L\|x - y\|$.

Proof. Let $\mathcal{H}$ be the RKHS corresponding to $\kappa$ and $f \in \mathcal{H}$ an arbitrary element with $g := (1 - P_{1:T})f$; the residual defined in lemma 1.3. Since $g \in \mathcal{H}$, we have by Lemma 1.1, $g$ is Lipschitz. Thus we have that for any point $y$:

$$|g(y)| \leq L\|g\|_{\mathcal{H}}\|y - x\| \leq L\|f\|_{\mathcal{H}}\|y - x\|,$$

where the second inequality is guaranteed by Lemma 1.2. On the other hand, by Lemma 2, we know that for all $y$ we have the following tight bound:

$$|g(y)| \leq \|f\|_{\mathcal{H}} \sigma_T(y)$$

(3)

Now, given the fact that both inequalities (2) and (3) are bounding the same quantity and that the latter is a tight estimate, we necessarily have that:

$$\|f\|_{\mathcal{H}} \sigma_T(y) \leq L\|f\|_{\mathcal{H}}\|y - x\|.$$  

Canceling $\|f\|_{\mathcal{H}}$ gives us the result. □
Lemma 4 (Adapted from Lemma 5.1 of Srinivas et al. (2010)). Let \( f \) be a sample from a GP. Consider \( \eta \in (0, 1) \) and set \( B_T = 2 \log (\pi_T / \eta) \) where \( \sum_{i=1}^{\infty} \pi_i = 1 \), \( \pi_T > 0 \). Then,

\[
|f(x_T) - \mu_T(x_T)| \leq B_T^2 \sigma_T(x_T) \forall T \geq 1
\]

holds with probability at least \( 1 - \eta \).

Proof. For \( x_T \) we have that \( f(x) \sim \mathcal{N}(\mu_T(x_T), \sigma_T(x_T)) \) since \( f \) is a sample from the GP. Now, if \( r \sim \mathcal{N}(0, 1) \), then

\[
P(r > c) = e^{-c^2/2(2\pi)^{-1/2}} \int e^{-(r-c)^2/2-c(r-c)} dr < e^{-c^2/2} \mathbb{P}(r > 0) = \frac{1}{2} e^{-c^2/2}.
\]

Thus we have that

\[
P \left( f(x) - \mu_T(x) > B_T^{1/2} \sigma_T(x) \right) = P(r > B_T^{1/2}) < \frac{1}{2} e^{-B_T^{1/2}}.
\]

By symmetry and the union bound, we have that \( P \left( |f(x) - \mu_T(x)| > B_T^{1/2} \sigma_T(x) \right) < e^{-B_T^{1/2}} \). By applying the union bound again, we derive

\[
P \left( |f(x) - \mu_T(x)| > B_T^{1/2} \sigma_T(x) \forall T \geq 1 \right) < \sum_{T=1}^{\infty} e^{-B_T^{1/2}}.
\]

By substituting \( B_T = 2 \log (\pi_T / \eta) \), we obtain the result. As in Srinivas et al. (2010), we can set \( \pi_T = \pi^2 T^2 / 6 \). \( \square \)

Since each node’s UCB and LCB are only evaluated at most once, we give the following shorthands in notation. Let \( N(x) \) be the number of evaluations of confidence bounds by the time the UCB of \( x \) is evaluated (line 12 of Algorithm 3) and let \( T(x) = |D_t| \) be the time the UCB of \( x \) is evaluated. Define \( U(x) = U_N(x) |D_T(x)| = \mu(x) |D_T(x)| + B_N(x) \sigma(x) |D_T(x)| \) and \( L(x) = L_N(x) |D_T(x)| = \mu(x) |D_T(x)| - B_N(x) \sigma(x) |D_T(x)| \).

Lemma 5. Consider \( B(x^*, \rho) \) and \( \gamma \in (0, 1) \) as in Assumptions 2 and 3. Suppose \( \mathcal{L}(x^*_h) \leq f(x^*_h) \leq U(x^*_h) \). If \( x^*_h \in B(x^*, \rho) and \delta(h) < c_0 \) then there exists a constant \( c \) such that \( \mathcal{L}(x^*_h) \geq f^* - \gamma \mathcal{L}(x^*_h) \).

Proof. If \( x^*_h \) is not evaluated then \( f(x^*) \geq U_T(x^*_h) \geq f^* - \delta(h) \geq f^* - c_0 \) which implies that \( x^* \in B(x^*, \rho) \). Therefore, \( f^* - c_2 \| x^* - x^* \| \geq f(x^*) \geq U_T(x^*_h) \geq f^* \) which in turn implies that \( \| x^* - x^* \| \leq \gamma \mathcal{L}(x^*_h) \). Similarly \( f^* - c_2 \| x^*_h - x^* \| \geq f(x^*_h) \geq f^* - \delta(h) \). Therefore \( \| x^*_h - x^* \| \leq \gamma \mathcal{L}(x^*_h) \). By the triangle inequality, we have

\[
\| x^* - x^*_h \| \leq \| x^* - x^* \| + \| x^*_h - x^* \| \leq 2 \gamma \mathcal{L}(x^*_h).
\]

By Lemma 3, we have that \( \sigma_T(x^*_h) \leq 2L \sqrt{\frac{\delta(h)}{c_2}} \). By the definition of \( \mathcal{L}_T \), we can argue that

\[
\mathcal{L}(x^*_h) \geq U(x^*_h) - 4B_N(x^*_h) L \sqrt{\frac{\delta(h)}{c_2}} \geq f^* - \delta(h) - 4B_N(x^*_h) L \sqrt{\frac{\delta(h)}{c_2}} = f^* - c_2 \gamma^h - 4B_N(x^*_h) L \sqrt{\frac{\delta(h)}{c_2}}.
\]

Note that since \( \gamma \in (0, 1), \gamma < \gamma \). Assume that \( B_1 = b \). Let \( \bar{c} = c/b + 4L \sqrt{\frac{\delta(h)}{c_2}} \). Since \( B_N > B_1 \forall N > 1 \), we have the statement.

If \( x^*_h \) is evaluated then the statement is trivially true. \( \square \)
Lemma 6. Assume that $h_{\text{max}} = n$. For a node $x_{h,i}$ at level $h$, $B_{N(x_{h,i})} = O(\sqrt{h})$.

Proof. Assume that there are $n_i$ nodes expanded at the end of iteration $i$ of the outer loop. In the $i + 1$th iteration of the outer loop, there can be at most $h_{\text{max}}(n_i)$ additional expansions added. Thus the total number of expansions at the end of iteration $i$ is at most $n_{i-1} + h_{\text{max}}(n_{i-1})$. We can prove by induction that $n_i \leq i^{2/3}$. Since it is clear that BaMSOO would expand every node after a finite number of node expansions, we only have to prove that $f^* - \hat{\delta}(h) \leq f^* - e_0$, which implies $n = \sum_{i=0}^{n_{\text{max}}} (2^i)^{2/3}$ where $n$ is the total number of expansions. Thus, there would be at most $2 \times 2^{i^{2/3}}$ evaluations. Hence,

$$B_{N(x_{h,i})} \leq \sqrt{2 \log(2^{2\sqrt{\log h}}/6\eta)} \leq 2 \log(2^{2\sqrt{\log h}}) + 2 \log(\pi^2/6\eta) = O(\sqrt{h}).$$

Lemma 7. After a finite number of node expansions, an optimal node $x_{h_0}^* \in B(x_0, \rho)$ is expanded such that $\hat{\delta}(h) \leq \epsilon_0$ and $x^*_h \in B(x_0, \rho)$.

Proof. Since it is clear that BaMSOO would expand every node after a finite number of node expansions, we only have to show that there exists an $h_0$ that satisfies the conditions. By Lemma 6, we have that $\forall h > h_0$, we have $\hat{\delta}(h) \leq \epsilon_0$. Thus, there exists an $h_0$ such that $\hat{\delta}(h) < \epsilon_0$. Since $f(x^*_h) > f^* - \hat{\delta}(h) > f^* - e_0$, we have by Assumption 2 that, $x^*_h \in B(x_0, \rho)$.

Lemma 8. $\sum_{h=0}^{H} |I_{\hat{\delta}(H)}| \leq C \left( B_{N(x_{h_0}^*)} \right)^{D/2} \gamma^{(D/4-hD/\alpha)H}$ for some constant $C$ for all $H > h_0$.

Proof. By Lemma 7, we know that $\hat{\delta}(H) = \hat{\delta}(H) < \epsilon_0$ if $H > h_0$. Therefore, by Assumption 2, we have that $\chi_{\hat{\delta}(H)} = \{ x \in \chi : f(x) \geq f^* - \hat{\delta}(H) \} \subseteq B(x_0, \rho)$. Again by Assumption 2, we have that

$$f^* - \hat{\delta}(H) \leq f(x) \leq f^* - c_2 \| x - x^* \|^2 \forall x \in \chi_{\hat{\delta}(H)}.$$

Thus $\chi_{\hat{\delta}(H)} \subseteq B \left( x_0, \sqrt{\hat{\delta}(H)/c_2} \right) = B \left( x_0, \sqrt{\hat{\delta}(H)/c_2} \right)$.

Since each cell $(h,i)$ contains a $\ell$-ball of radius $\nu \delta(h)$ centered at $x_{h,i}$ we have that each cell contains a ball $B(x_{h,i}, (\nu \delta(h))^{1/\alpha}) = B(x_{h,i}, (\nu \delta(h))^{1/\alpha})$. By the argument of volume, we have that $|I_{\hat{\delta}(H)}| \leq C_1 \left( B_{N(x_{h_0}^*)} \right)^{D/2} \gamma^{HD/4 - hD/\alpha}$ for some constant $C_1$. Finally,

$$\sum_{h=0}^{H} |I_{\hat{\delta}(H)}| \leq C_1 \sum_{h=0}^{H} \left( B_{N(x_{h_0}^*)} \right)^{D/2} \gamma^{HD/4 - hD/\alpha}$$

$$= C_1 \left( B_{N(x_{h_0}^*)} \right)^{D/2} \gamma^{HD/4} \sum_{h=0}^{H} \gamma^{-hD/\alpha}$$

$$= C_1 \left( B_{N(x_{h_0}^*)} \right)^{D/2} \gamma^{HD/4} \sum_{h=0}^{H} \left( \gamma^{D/\alpha} \right)^{h-H}$$

$$\leq C_1 \left( B_{N(x_{h_0}^*)} \right)^{D/2} \gamma^{HD/4} \sum_{h=0}^{\infty} \left( \gamma^{D/\alpha} \right)^{h-H}$$

$$= C_1 \left( B_{N(x_{h_0}^*)} \right)^{D/2} \gamma^{HD/4} \frac{\gamma^{-DH/\alpha}}{1 - \gamma^{D/\alpha}}$$

$$= C_1 \frac{1 - \gamma^{D/\alpha}}{1 - \gamma^{D/\alpha}} \left( B_{N(x_{h_0}^*)} \right)^{D/2} \gamma^{(D/4-D/\alpha)H}. $$
Bayesian Multi-Scale Optimistic Optimization

Setting $C = \frac{c_1}{1 - \gamma D/\alpha}$ gives us the desired result. \hfill \Box

**Lemma 9.** Suppose $L(x^*_h) \leq f(x^*_h) \leq U(x^*_h)$. If $x^*_h$ is not evaluated (that is $U(x^*_h) < f^+$) then $f^+$ is $\delta(h)$-optimal.

**Proof.** $f^+ > U(x^*_h) \geq f(x^*_h) > f^+ - \delta(h)$. \hfill \Box

### A.2 Main Results

#### A.2.1 Simple Regret

Let $h^*_n$ be the deepest level of an expanded optimal node with $n$ node expansions. This following lemma is adapted from Lemma 2 of Munos (2011).

**Lemma 10.** Suppose $L(x) \leq f(x) \leq U(x)$ for all $x$ whose confidence region are evaluated. Whenever $h \leq h_{\text{max}}(n)$ and $n \geq C h_{\text{max}}(n) \sum_{i=0}^{h} \left( B_N(x^*_i) \right)^{D/2} \gamma(D/4-D/\alpha)i + n_0$ for some constant $C$, we have $h^*_n \geq h$.

**Proof.** We prove the statement by induction. By Lemma 7, we have that after $n_0$ node expansions, a node $x^*_{h_0} \in B(x^*, \rho)$ is expanded. Also $\forall h > h_0$, we have that $cB_N(x^*_h) \leq \epsilon_0$ and $x^*_h \in B(x^*, \rho)$. For $h = h_0$, the statement is trivially satisfied. Thus assume that the statement is true for $h$. Let $n$ be such that $n \geq C h_{\text{max}}(n) \sum_{i=0}^{h+1} \left( B_N(x^*_i) \right)^{D/2} \gamma(D/4-D/\alpha)i + n_0$. By the inductive hypothesis we have that $h^*_n \geq h$. Assume $h^*_n = h$ since otherwise the proof is finished. As long as the optimal node at level $h+1$ is not expanded, all nodes expanded at the level are $\delta(h+1)$-optimal by Lemma 5. By Lemma 8, we know that after $C h_{\text{max}}(n) \left( B_N(x^*_{h+1}) \right)^{D/2} \gamma(D/4-D/\alpha)(h+1)$ node expansions, the optimal node at level $h+1$ will be expanded since there are at most $\sum_{i=0}^{h+1} \left| f_i \right| \delta(h+1)$-optimal nodes at or beneath level $h+1$. Thus $h^*_n \geq h+1$. \hfill \Box

**Theorem 1.** Suppose $L(x) \leq f(x) \leq U(x)$ for all $x$ whose confidence region is evaluated. Let us write $h(n)$ to be the smallest integer $h \geq h_0$ such that

$$Ch_{\text{max}}(n) \sum_{i=0}^{h} \left( B_N(x^*_i) \right)^{\frac{D}{2}} \gamma(D/4-D/\alpha)i + n_0 \geq n.$$ 

Then the loss is bounded as

$$r_n \leq \delta(\min\{h(n), h_{\text{max}}(n) + 1\})$$

and $h^*_n \geq \min\{h(n) - 1, h_{\text{max}}(n)\}$.

**Proof.** From Lemma 8, and the definition of $h(n)$ we have that

$$Ch_{\text{max}}(n) \sum_{i=0}^{h(n)-1} \left( B_N(x^*_i) \right)^{\frac{D}{2}} \gamma(D/4-D/\alpha)i + n_0 < n.$$ 

By Lemma 10, we have that $h^*_n \geq h(n) - 1$ if $h(n) - 1 \leq h_{\text{max}}(n)$ and $h^*_n \geq h_{\text{max}}(n)$ otherwise. Therefore $h^*_n \geq \min\{h(n) - 1, h_{\text{max}}(n)\}$.

By Lemma 9, we know that if $x^*_{h^*_n + 1}$ is not evaluated then $f^+$ is $\delta(h^*_n + 1)$-optimal. If $x^*_{h^*_n + 1}$ is evaluated, then $f \left( x^*_{h^*_n + 1} \right)$ is $\delta(h^*_n + 1)$-optimal. Thus $r_n \leq \delta(\min\{h(n), h_{\text{max}}(n) + 1\})$. \hfill \Box

**Proof of Corollary 1.** Suppose $L(x) \leq f(x) \leq U(x)$ for all $x$ whose confidence region is evaluated. By Lemma 4, we know that this holds with probability at least $1 - \eta$. 


By the definition of \( h(n) \) we have that
\[
 n \leq \text{Ch}_{\max}(n) \sum_{i=\text{h}_0}^{h(n)} \left(B_N\left(x_i^*\right)\right)^{D/2} \gamma^{(D/4-D/n)i} + n_0
\]
\[
 \leq \text{Ch}_{\max}(n) \left(B_N\left(x_{\text{h}(n)}^*\right)\right)^{D/2} \sum_{i=\text{h}_0}^{h(n)} \gamma^{-di} + n_0
\]
\[
 \leq \text{Ch}_{\max}(n) \left(B_N\left(x_{\text{h}(n)}^*\right)\right)^{D/2} \gamma^{-dh_0} \gamma^{-dh(n) - 1} \frac{\gamma^{-d} - 1}{\gamma^{-d} - 1} + n_0
\]
(4)

If \( h(n) \leq h_{\max}(n) + 1 \), then by Theorem 1, we have that \( h^*_n \geq h(n) - 1 \). After \( n \) expansions, the optimal node \( x_{h(n)-1}^* \) has been expanded which suggests that its children’s confidence bounds have been evaluated. Hence, \( N\left(x_{h(n)}^*\right) < 2n \) since there have only been \( n \) expansions. Therefore,
\[
(4) \leq Kn^e \left(B_{2n}\right)^{D/2} \gamma^{-dh(n)}
\]
for some constant \( K \) which implies that
\[
\gamma^{h(n)} \leq K^{1/d} B_{2n}^{\frac{2\alpha}{d\gamma}} n^{-\frac{1+\varepsilon}{d}} = K^{1/d} \left[2 \log(4\pi^2 n^2/6\eta)\right]^{\frac{\alpha}{\gamma^2}} n^{-\frac{1+\varepsilon}{d}}.
\]

By Theorem 1, we have that
\[
r_n \leq c \min \left\{K^{1/d} \left[2 \log(4\pi^2 n^2/6\eta)\right]^{\frac{\alpha}{\gamma^2}} n^{-\frac{1+\varepsilon}{d}}, \gamma^{(n+1)^e}\right\} = O\left(n^{-\frac{1+\varepsilon}{d}} \log \frac{n^2}{\eta}\right).
\]

\( \square \)