Appendix

A. Some Ancillary Material

A.1. Review of GP-UCB

We present a review of the GP-UCB algorithm of Srinivas et al. (2010) which we build on in this work. Here we will assume $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$ where $\kappa : \mathcal{X}^2 \to \mathbb{R}$ is a radial kernel defined on the domain $\mathcal{X}$. The algorithm is given below.

Algorithm 2 GP-UCB (Srinivas et al., 2010)

Input: kernel κ.

- $\mathcal{D}_0 \leftarrow \varnothing$, $(\mu_0, \sigma_0) \leftarrow (\mathbf{0}, \kappa^{1/2})$.
- **for** $t = 1, 2, \ldots$
  1. $x_t \leftarrow \text{argmax}_{x \in \mathcal{X}} \mu_{t-1}(x) + \beta_t^{1/2} \sigma_{t-1}(x)$
  2. $y_t \leftarrow$ Query f at x_t.
  3. Perform Bayesian posterior updates to obtain μ_t, σ_t See (1).

To present the theoretical results for GP-UCB, we begin by defining the *Maximum Information Gain* (MIG) which characterises the statistical difficulty of GP bandits.

Definition 2. *(Maximum Information Gain (Srinivas et al., 2010)) Let $f \sim \mathcal{GP}(\mathbf{0}, \phi_{\mathcal{X}})$. Consider any $A \subset \mathbb{R}^d$ and let $A' = \{x_1, \ldots, x_n\} \subset A$ be a finite subset. Let $f_{A'}, \epsilon_{A'} \in \mathbb{R}^n$ such that $(f_{A'})_i = f(x_i)$ and $(\epsilon_{A'})_i \sim \mathcal{N}(0, \eta^2)$. Let $y_{A'} = f_{A'} + \epsilon_{A'}$. Denote the Shannon Mutual Information by I. The Maximum Information Gain of A is*

$$\Psi_n(A) = \max_{A' \subset A, |A'| = n} I(y_{A'}; f_{A'}).$$

Next, we will need the following regularity conditions on the kernel. It is satisfied for four times differentiable kernels such as the SE kernel and Matérn kernel when $\nu > 2$ (Ghosal & Roy, 2006).

Assumption 3. *Let $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$, where $\kappa : \mathcal{X}^2 \to \mathbb{R}$ is a stationary kernel. The partial derivatives of f satisfies the following condition. There exist constants $a, b > 0$ such that,*

$$\text{for all } J > 0, \text{ and for all } i \in \{1, \ldots, d\}, \quad \mathbb{P}\left(\sup_x \left|\frac{\partial f(x)}{\partial x_i}\right| > J\right) \leq a e^{-(J/b)^2}.$$

The following theorem is a bound on the simple regret S_n (2) for GP-UCB.

Theorem 4. *((Srinivas et al., 2010)) Let $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$, where $\mathcal{X} = [0,1]^d$, $f : \mathcal{X} \to \mathbb{R}$ and the kernel κ satisfies Assumption 3). At each query, we have noisy observations $y = f(x) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \eta^2)$. Denote $C_1 = 8/\log(1+\eta^{-2})$. Pick a failure probability $\delta \in (0,1)$ and run GP-UCB with $\beta_t = 2\log\left(\frac{2\pi^2 t^2}{3\delta}\right) + 2d\log\left(t^2 bdr\sqrt{\frac{4ad}{\delta}}\right)$. The following holds with probability $> 1 - \delta$,*

$$\text{for all } n \geq 1, \qquad S_n \leq \sqrt{\frac{C_1 \beta_n \Psi_n(\mathcal{X})}{n}} + \frac{\pi^2}{6}.$$

A.2. Some Technical Results

Here we present some technical lemmas we will need for our analysis.

Lemma 5 (Gaussian Concentration)**.** *Let $Z \sim \mathcal{N}(0,1)$. Then $\mathbb{P}(Z > \epsilon) \leq \frac{1}{2}\exp(-\epsilon^2/2)$.*

Lemma 6 (Mutual Information in GP, (Srinivas et al., 2010) Lemma 5.3)**.** *Let $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$, $f : \mathcal{X} \to \mathbb{R}$ and we observe $y = f(x) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \eta^2)$. Let A be a finite subset of $\mathcal{X}$ and f_A, y_A be the function values and observations on this set respectively. Then the Shannon Mutual Information $I(y_A; f_A)$ is,*

$$I(y_A; f_A) = \frac{1}{2}\sum_{t=1}^{n} \log(1 + \eta^{-2}\sigma_{t-1}^2(x_t)).$$

where σ^2_{t-1} is the posterior GP variance after observing the first $t-1$ points.

Our next result is a technical lemma taken from Kandasamy et al. (2016a). It will be used in controlling the posterior variance of our f and g GPs.

Lemma 7 (Posterior Variance Bound (Kandasamy et al., 2016a)). *Let $f \sim (\mathbf{0}, \kappa)$, $f : \mathcal{U} \to \mathbb{R}$ where $\kappa(u, u') = \kappa_0 \phi(\|u - u'\|)$ and ϕ is a radial kernel. Upon evaluating f at u we observe $y = f(u) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \eta^2)$. Let $u_1 \in \mathcal{U}$ and suppose we have s observations at u_1 and no observations elsewhere. Then the posterior variance κ' (see (1)) at all $u \in \mathcal{U}$ satisfies,*

$$\kappa'(u, u) \leq \kappa_0(1 - \phi^2(\|u - u_1\|)) + \frac{\eta^2/s}{1 + \frac{\eta^2}{\kappa_0 s}}.$$

Proof: The proof is in Section C.0.1 of Kandasamy et al. (2016a) who prove this result as part of a larger proof. ■

B. Analysis

We will first state a formal version of Theorem 1. Recall from the main text where we stated that most evaluations at $z_\bullet$ are inside the following set $\mathcal{X}_\rho$.

$$\mathcal{X}_\rho = \{x \in \mathcal{X} : f_\star - f(x) \leq 2\rho\sqrt{\kappa_0}\|\xi\|_\infty\}.$$

This is not entirely accurate as it hides a dilation that arises due to a covering argument in our proofs. Precisely, we will show that after n queries at any fidelity, BOCA will use most of the $z_\bullet$ evaluations in $\mathcal{X}_{\rho,n}$ defined below using $\mathcal{X}_\rho$.

$$\mathcal{X}_{\rho,n} = \left\{ x \in \mathcal{X} : \mathrm{B}_2\big(x, \sqrt{d}/n^{\alpha/2d}\big) \cap \mathcal{X}_{\rho,n} \neq \varnothing \right\} \tag{10}$$

Here $\mathrm{B}_2(x, \epsilon)$ is an L_2 ball of radius ϵ centred at x. $\mathcal{X}_{\rho,n}$ is a dilation of $\mathcal{X}_\rho$ by $\sqrt{d}/n^{\alpha/2d}$. Notice that for all $\alpha > 0$, as $n \to \infty$, $\mathcal{X}_{\rho,n}$ approaches $\mathcal{X}_\rho$ at a polynomial rate. We now state our main theorem below.

Theorem 8. *Let $\mathcal{Z} = [0,1]^p$ and $\mathcal{X} = [0,1]^d$. Let $g \sim \mathcal{GP}(\mathbf{0}, \kappa)$ where κ is of the form (3). Let $\phi_\mathcal{X}$ satisfy Assumption 3 with some constants $a, b > 0$. Pick $\delta \in (0,1)$ and run* BOCA *with*

$$\beta_t = 2\log\left(\frac{\pi^2 t^2}{2\delta}\right) + 4d\log(t) + \max\left\{0\,,\, 2d\log\left(brd\log\left(\frac{6ad}{\delta}\right)\right)\right\}.$$

Then, for all $\alpha \in (0,1)$ there exists ρ, Λ_0 such that with probability at least $1-\delta$ we have for all $\Lambda \geq \Lambda_0$,

$$S(\Lambda) \;\leq\; \sqrt{\frac{2C_1\beta_{2n_\Lambda}\Psi_{2n_\Lambda}(\mathcal{X}_{\rho,n})}{n_\Lambda}} \;+\; \sqrt{\frac{2C_1\beta_{2n_\Lambda}\Psi_{2n_\Lambda^\alpha}(\mathcal{X})}{n_\Lambda^{2-\alpha}}} \;+\; \frac{\pi^2}{6n_\Lambda}.$$

Here $C_1 = 8/\log(1+\eta^2)$ is a constant and $n_\Lambda = \lfloor \Lambda/\lambda(z_\bullet) \rfloor$. ρ satisfies $\rho > \rho_0 = \max\{2, 1 + \sqrt{(1+2/\alpha)/(1+d)}\}$.

In addition to the dilation, Theorem 1 in the main text also suppresses the constants and polylog terms. The next three subsections are devoted to proving the above theorem. In Section B.1 we describe some discretisations for $\mathcal{Z}$ and $\mathcal{X}$ which we will use in our proofs. Section B.2 gives some lemmas we will need and Section B.3 gives the proof.

B.1. Set Up & Notation

Notation: Let $U \subset \mathcal{Z} \times \mathcal{X}$. $T_n(U)$ will denote the number of queries by BOCA at points $(z, x) \in U$ within n time steps. When $A \subset \mathcal{Z}$ and $B \subset \mathcal{X}$, we will overload notation to denote $T_n(A, B) = T_n(A \times B)$. For $z \in \mathcal{Z}$, $[> z]$ will denote the fidelities which are more expensive than z, i.e. $[> z] = \{z' \in \mathcal{Z} : \lambda(z') > \lambda(z)\}$.

We will require a fairly delicate set up before we can prove Theorem 8. Let $\alpha > 0$. All sets described in the rest of this subsection are defined with respect to α. First define

$$\tilde{\mathcal{H}}_n = \{(z, x) \in \mathcal{Z} \times \mathcal{X} : f_\star - f(x) < 2\rho\beta_n^{1/2}\sqrt{\kappa_0}\xi(z)\},$$

where recall from (4), $\xi(z) = \sqrt{1 - \phi_{\mathcal{Z}}^2(\|z - z_\bullet\|)}$ is the information gap function. We next define $\mathcal{H}'_n$ to be an L_2 dilation of $\tilde{\mathcal{H}}_n$ in the $\mathcal{X}$ space, i.e.

$$\mathcal{H}'_n = \{(z, x) \in \mathcal{Z} \times \mathcal{X} : \mathrm{B}_2\big(x, \sqrt{d}/n^{\alpha/2d}\big) \cup \tilde{\mathcal{H}}_n \neq \varnothing\}.$$

Finally, we define $\mathcal{H}_n$ to be the intersection of $\mathcal{H}'_n$ with all fidelities satisfying the third condition in (7). That is,

$$\mathcal{H}_n = \mathcal{H}'_n \cap \Big\{(z, x) \in \mathcal{Z} \times \mathcal{X} : \xi(z) > \|\xi\|_\infty / \beta_n^{1/2}\Big\}. \tag{11}$$

In our proof we will use the second condition in (7) to control the number of queries in $\mathcal{H}_n$.

To control the number of queries outside $\mathcal{H}_n$ we first introduce a $\frac{\sqrt{d}}{2n^{\frac{\alpha}{2d}}}$-covering of the space $\mathcal{X}$ of size $n^{\alpha/2}$. If $\mathcal{X} = [0, 1]^d$, a sufficient covering would be an equally spaced grid having $n^{\frac{\alpha}{2d}}$ points per side. Let $\{a_{i,n}\}_{i=1}^{n^{\frac{\alpha}{2}}}$ be the points in the covering. $A_{i,n} \subset \mathcal{X}$ to be the points in $\mathcal{X}$ which are closest to $a_{i,n}$ in $\mathcal{X}$. Therefore $F_n = \{A_{i,n}\}_{i=1}^{n^{\frac{\alpha}{2}}}$ is a partition of $\mathcal{X}$.

Now define $Q_t : 2^{\mathcal{X}} \to 2^{\mathcal{Z}}$ to be the following function which maps subsets of $\mathcal{X}$ to subsets of $\mathcal{Z}$.

$$Q_t(A) = \Big\{z \in \mathcal{Z} : \ \forall\, x \in A, \quad f_\star - f(x) \geq 2\rho\beta_t^{1/2}\sqrt{\kappa_0}\xi(z)\Big\}. \tag{12}$$

That is, Q_t maps $A \subset \mathcal{X}$ to fidelities where the information gap ξ is smaller than $(f_\star - f(x))/(2\rho\beta_t^{1/2})$ for all $x \in A$. Next we define $\theta_t : 2^{\mathcal{X}} \to \mathcal{Z}$, to be the cheapest fidelity in $Q_t(A)$ for a subset $A \in \mathcal{X}$.

$$\theta_t(A) \ = \ \operatorname*{arginf}_{z \in Q_t(A)} \lambda(z). \tag{13}$$

We will see that BOCA will not query inside an $A_{i,n} \in F_n$ at fidelities larger than $\theta_t(A_{i,n})$ too many times (see Lemma 12). That is, $T_n([> \theta_n(A_{i,n})], A_{i,n})$ will be small. We now define $\mathcal{F}_n$ as follows,

$$\mathcal{F}_n = \bigcup_{A_{i,n} \subset \mathcal{X} \setminus \mathcal{X}_{\rho,n}} [> \theta_n(A_{i,n})] \times A_{i,n}. \tag{14}$$

That is, we first choose $A_{i,n}$'s that are completely outside $\mathcal{X}_{\rho,n}$ and take their cross product with fidelities more expensive than $\theta_t(A_{i,n})$. By design of the above sets, and using the third condition in (7) we can bound the total number of queries as follows,

$$n = T_n(\mathcal{Z}, \mathcal{X}) \ \leq \ T_n(\{z_\bullet\}, \mathcal{X}_{\rho,n}) + T_n(\mathcal{F}_n) + T_n(\mathcal{H}_n)$$

We will show that the last two terms on the right hand side are small for BOCA and consequently, the first term will be large. But first, we establish a series of technical results which will be useful in proving theorem 8.

B.2. Some Technical Lemmas

The first lemma proves that the UCB φ_t in (6) upper bounds $f(x_t)$ on all the domain points $\{x_t\}_{t \geq 1}$ chosen for evaluation.

Lemma 9. *Let $\beta_t > 2\log(\pi^2 t^2/2\delta)$. Then, with probability $> 1 - \delta/3$, we have*

$$\forall\, t \geq 1, \quad |f(x_t) - \mu_{t-1}(x_t)| \ \leq \ \beta_t^{1/2}\sigma_{t-1}(x_t).$$

Proof: This is a straightforward argument using Lemma 5 and the union bound. At $t \geq 1$,

$$\begin{aligned}\mathbb{P}\Big(|f(x) - \mu_{t-1}(x)| > \beta_t^{1/2}\sigma_{t-1}(x)\Big) &= \mathbb{E}\left[\mathbb{E}\left[|f(x) - \mu_{t-1}(x)| > \beta_t^{1/2}\sigma_{t-1}(x) \;\middle|\; \mathcal{D}_{t-1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{P}_{Z \sim \mathcal{N}(0,1)}\left(|Z| > \beta_t^{1/2}\right)\right] \leq \ \exp\left(\frac{-\beta_t}{2}\right) \ = \ \frac{2\delta}{\pi^2 t^2}.\end{aligned}$$

In the first step we have conditioned w.r.t $\mathcal{D}_{t-1} = \{(z_i, x_i, y_i)\}_{i=1}^{t-1}$ which allows us to use Lemma 5 as $f(x)|\mathcal{D}_{t-1} \sim \mathcal{N}(\mu_{t-1}(x), \sigma_{t-1}^2(x))$. The statement follows via a union bound over all $t \geq 0$ and the fact that $\sum_t t^{-2} = \pi^2/6$. ■

Next we show that the GP sample paths are well behaved and that $\varphi_t(x)$ upper bounds $f(x)$ on a sufficiently dense subset at each time step. For this we use the following lemma.

Lemma 10. *Let β_t be as given in Theorem 8. Then for all t, there exists a discretisation G_t of $\mathcal{X}$ of size $(t^2brd\sqrt{6ad/\delta})^d$ such that the following hold.*

- *Let $[x]$ be the closest point to $x \in \mathcal{X}$ in the discretisation. With probability $> 1-\delta/6$, we have*

$$\forall\, t \geq 1, \quad \forall\, x \in \mathcal{X}, \quad |f(x) - f([x]_t)| \leq 1/t^2.$$

- *With probability $> 1-\delta/3$, for all $t \geq 1$ and for all $a \in G_t$, $|f(a) - \mu_{t-1}(a)| \leq \beta_t^{1/2}\sigma_{t-1}(a)$.*

Proof: The first part of the proof, which we skip here, uses the regularity condition for $\phi_{\mathcal{X}}$ in Assumption 3 and mimics the argument in Lemmas 5.6, 5.7 of Srinivas et al. (2010). The second part mimics the proof of Lemma 9 and uses the fact that $\beta_t > 2\log(|G_t|\pi^2t^2/2\delta)$. ■

The discretisation in the above lemma is different to the coverings introduced in Section B.1. The next lemma is about the information gap function in (4).

Lemma 11. *Let $g \sim \mathcal{GP}(0,\kappa)$, $g : \mathcal{Z}\times\mathcal{X} \to \mathbb{R}$ and κ is of the form (3). Suppose we have s observations from g. Let $z \in \mathcal{Z}$ and $x \in \mathcal{X}$. Then $\tau_{t-1}(z,x) < \alpha$ implies $\sigma_{t-1}(x) < \alpha + \sqrt{\kappa_0}\xi(z)$.*

Proof: The proof uses the observation that for radial kernels, the maximum difference between the variances at two points u_1 and u_2 occurs when all s observations are at u_2 or vice versa. Now we use $u_1 = (z,x)$ and $u_2 = (z_\bullet, x)$ and apply Lemma 7 to obtain $\tau^2_{t-1}(z_\bullet,x) \leq \kappa_0(1-\phi_{\mathcal{Z}}(\|z_\bullet - z\|))^2 + \frac{\eta^2/s}{1+\frac{\eta^2}{s\kappa_0}}$. However, As $\tau^2_{t-1}(z,x) = \frac{\eta^2/s}{1+\frac{\eta^2}{s\kappa_0}}$ when all observations are at (z,x) and noting that $\sigma^2_{t-1}(x) = \tau^2_{t-1}(z_\bullet,x)$, we have $\sigma^2_{t-1}(x) \leq \kappa_0(1-\phi_{\mathcal{Z}}(\|z_\bullet - z\|))^2 + \tau^2_{t-1}(z,x)$. Since the above situation characterised the maximum difference between $\sigma^2_{t-1}(x)$ and $\tau^2_{t-1}(z,x)$, this inequality is valid for any general observation set. The proof is completed using the elementary inequality $a^2+b^2 \leq (a+b)^2$ for $a,b>0$. ■

We are now ready to prove Theorem 8. The plan of attack is as follows. We will analyse BOCA after n time steps and bound the number of plays at fidelities $z \neq z_\bullet$ and outside $\mathcal{X}_{\rho,n}$ at $z_\bullet$. Then we will show that for sufficiently large Λ, the number of *random* plays N is bounded by $2n_\Lambda$ with high probability. Finally we use techniques from Srinivas et al. (2010), specifically the maximum information gain, to control the simple regret. However, unlike them we will obtain a tighter bound as we can control the regret due to the sets $\mathcal{X}_{\rho,n}$ and $\mathcal{X}\setminus\mathcal{X}_{\rho,n}$ separately.

B.3. Proof of Theorem 8

Let $\alpha > 0$ be given. We invoke the sets $\mathcal{X}_{\rho,n}, \mathcal{H}_n, \mathcal{F}_n$ in equations (10), (11), (14) for the given α. The following lemma establishes that for any $A \subset \mathcal{X}$, we will not query inside A at fidelities larger than $\theta_t(A)$ (13) too many times. The proof is given in Section B.3.1.

Lemma 12. *Let $A \subset \mathcal{X}$ which does not contain the optimum. Let ρ, β_t be as given in Theorem 8. Then for all $u > \max\{3, (2(\rho-\rho_0)\eta)^{-2/3}\}$, we have*

$$\mathbb{P}\Big(T_n([>\theta_t(A)], A) \;>\; u\Big) \;\leq\; \frac{\delta}{\pi^2}\frac{1}{u^{1+4/\alpha}}$$

To bound $T(\mathcal{F}_n)$, we will apply Lemma 12 with $u = n^{\alpha/2}$ on all $A_{i,n} \in F_n$ satisfying $A_{i,n} \subset \mathcal{X}\setminus\mathcal{X}_{\rho,n}$. Since $\mathcal{X}_\rho \subset \mathcal{X}_{\rho,n}$, $A_{i,n}$ does not contain the optimum. As $\mathcal{F}_n$ is the union of such sets (14), we have for all n (larger than a constant),

$$\begin{aligned}
\mathbb{P}(T(\mathcal{F}_n) > n^\alpha) \;&\leq\; \mathbb{P}\Big(\exists A_{i,n} \subset \mathcal{X}\setminus\mathcal{X}_{\rho,n},\;\; T_n([>\theta_t(A_{i,n})], A_{i,n}) \;>\; n^{\alpha/2}\Big) \\
&\leq \sum_{\substack{A_{i,n}\in F_n \\ A_{i,n}\subset\mathcal{X}\setminus\mathcal{X}_{\rho,n}}} \mathbb{P}\Big(T_n([>\theta_t(A_{i,n})], A_{i,n}) \;>\; n^{\alpha/2}\Big) \leq |F_n|\frac{\delta}{\pi^2}\frac{1}{n^{\alpha/2+2}} \leq \frac{\delta}{\pi^2}\frac{1}{n^2}
\end{aligned}$$

Now applying the union bound over all n, we get $\mathbb{P}(\forall\, n \geq 1,\; T(\mathcal{F}_n) > n^\alpha) \;\leq\; \delta/6$.

Now we will bound the number of plays in $\mathcal{H}_n$ using the second condition in (7). We begin with the following Lemma. The proof mimics the argument in Lemma 11 of Kandasamy et al. (2016a) who prove a similar result for GPs defined on just the domain, i.e. $f \sim \mathcal{GP}(\mathbf{0},\kappa)$ where $f : \mathcal{X} \to \mathbb{R}$.

Lemma 13. *Let $A \subset \mathcal{Z} \times \mathcal{X}$ and the L_2 diameter of A in $\mathcal{X}$ be $D_\mathcal{X}$ and that in $\mathcal{Z}$ be $D_\mathcal{Z}$. Suppose we have n evaluations of g of which s are in A. Then for any $(z, x) \in A$, the posterior variance τ'^2 satisfies,*

$$\tau'^2(z,x) \leq \kappa_0(1 - \phi^2_\mathcal{Z}(D_\mathcal{Z})\phi^2_\mathcal{X}(D_\mathcal{X})) + \frac{\eta^2}{s}.$$

Let $\lambda_r = \lambda_{min}/\lambda(z_\bullet)$ where $\lambda_{min} = \min_{z\in\mathcal{Z}} \lambda(z)$. If the maximum posterior variance in a certain region is smaller than $\gamma(z)$, then we will not query within that region by the second condition in (7). Further by the third condition, since we will only query at fidelities satisfying $\xi(z) > \|\xi\|_\infty/\beta_n^{1/2}$, it is sufficient to show that the posterior variance is bounded by $\kappa_0\|\xi\|^2_\infty\lambda_r^{2q}/\beta_n$ at time n to prove that we will not query again in that region. For this we can construct a covering of $\mathcal{H}_n$ such that $1 - \phi^2_\mathcal{Z}(D_\mathcal{Z})\phi^2_\mathcal{X}(D_\mathcal{X}) < \frac{1}{2}\|\xi\|^2_\infty\lambda_r^{2q}/\beta_n$. For any $A \subset \mathcal{Z} \times \mathcal{X}$, the covering number, which we denote $\Omega_n(A)$ of this construction will typically be poly-logarithmic in n (See Remark 15 below). Now if there are $\frac{2\beta_n\eta^2}{\lambda_r^{2q}\|\xi\|^2_\infty\kappa_0} + 1$ queries inside a ball in this covering, the posterior variance, by Lemma 13 will be smaller than $\kappa_0\|\xi\|^2_\infty\lambda_r^{2q}/\beta_n$. Therefore, we will not query any further inside this ball. Hence, the total number of queries in $\mathcal{H}_n$ is $T_n(\mathcal{H}_n) \leq C_2\Omega_n(\mathcal{H}_n)\frac{\beta_n}{\lambda_r^{2q}} \leq C_3\text{vol}(\mathcal{H}_n)\frac{\text{polylog}(n)}{\text{poly}(\lambda_r)}$ for appropriate constants C_2, C_3. (Also see Remark 16).

Next, we will argue that the number of queries for sufficiently large Λ, is bounded by $n_\Lambda/2$ where, recall $n_\Lambda = \lfloor\Lambda/\lambda(z_\bullet)\rfloor$. This simply follows from the bounds we have for $T_n(\mathcal{F}_n)$ and $T_n(\mathcal{H}_n)$.

$$T_n(\mathcal{Z} \setminus \{z_\bullet\}, \mathcal{X}) \leq T_n(\mathcal{F}_n) + T_n(\mathcal{H}_n) \leq n^\alpha + \mathcal{O}(\text{polylog}(n)).$$

Since the right hand side is sub-linear in n, we can find n_0 such that for all n_0, $n/2$ is larger than the right hand side. Therefore for all $n \geq n_0$, $T_n(\{z_\bullet\}, \mathcal{X}) > n/2$. Since our bounds hold with probability $> 1 - \delta$ for all n we can invert the above inequality to bound N, the random number of queries after capital Λ. We have $N \leq 2\Lambda/\lambda(z_\bullet)$. We only need to make sure that $N \geq n_0$ which can be guaranteed if $\Lambda > \Lambda_0 = n_0\lambda(z_\bullet)$.

The final step of the proof is to bound the simple regret after n time steps in BOCA. This uses techniques that are now standard in GP bandit optimisation, so we only provide an outline. We will need the following Lemma, whose proof is given in Section B.3.2.

Lemma 14. *Assume that we have queried g at n points, $(z_t, x_t)_{t=1}^n$ of which s points are in $\{z_\bullet\} \times A$ for any $A \subset \mathcal{X}$. Let σ_{t-1} denote the posterior variance of f at time t, i.e. after $t-1$ queries. Then, $\sum_{x_t\in A, z_t=z_\bullet} \sigma^2_{t-1}(x_t) \leq \frac{2}{\log(1+\eta^{-2})}\Psi_s(A)$. Here $\Psi_s(A)$ is the MIG of $\phi_\mathcal{X}$ after s queries to A as given in Definition 2.*

We now define the quantity R_n below. Readers familiar with the GP bandit literature might see that it is similar to the notion of cumulative regret, but we only consider queries at $z_\bullet$.

$$R_n = \sum_{\substack{t=1\\ z_t=z_\bullet}}^{n} f_\star - f(x_t) \quad = \sum_{\substack{z_t=z_\bullet\\ x_t\in\mathcal{X}_{\rho,n}}} f_\star - f(x_t) \quad + \sum_{\substack{z_t=z_\bullet\\ x_t\notin\mathcal{X}_{\rho,n}}} f_\star - f(x_t). \tag{15}$$

For any $A \subset \mathcal{X}$ we can use Lemmas 9, 10, and 14 and the Cauchy Schwartz inequality to obtain,

$$\sum_{\substack{z_t=z_\bullet\\ x_t\in A}} f_\star - f(x_t) \leq \sqrt{C_1 T_n(z_\bullet, A)\beta_n\Psi_{T_n(z_\bullet,A)}(A)} + \sum_{\substack{z_t=z_\bullet\\ x_t\in A}} \frac{1}{t^2}. \tag{16}$$

For the first term in (15), we set $A = \mathcal{X}_{\rho,n}$ in (16) and use the trivial bound $T_n(z_\bullet, \mathcal{X}_{\rho,n}) \leq n$. For the second term we note that $\{z_\bullet\} \times (\mathcal{X} \setminus \mathcal{X}_{\rho,n}) \subset \mathcal{F}_n$ and hence, $T_n(z_\bullet, \mathcal{X} \setminus \mathcal{X}_{\rho,n}) \leq T_n(\mathcal{F}_n) \leq n^\alpha$. As $A \subset B \implies \Psi_n(A) \leq \Psi_n(B)$, we have $R_n \leq \sqrt{C_1 n\beta_n\Psi_n(\mathcal{X}_{\rho,n})} + \sqrt{C_1 n^\alpha\beta_n\Psi_{n^\alpha}(\mathcal{X})} + \pi^2/6$. Now, using the fact that $N \leq 2n_\Lambda$ for large enough N we have,

$$R_N \leq \sqrt{2C_1 n_\Lambda\beta_{2n_\Lambda}\Psi_{2n_\Lambda}(\mathcal{X}_{\rho,n})} \; + \; \sqrt{2^\alpha C_1 n_\Lambda^\alpha\beta_{2n_\Lambda}\Psi_{2n_\Lambda^\alpha}(\mathcal{X})} \; + \; \frac{\pi^2}{6}.$$

The theorem now follows from the fact that $S(\Lambda) \leq \frac{1}{N}R_N$ by definition and that $N \geq n_\Lambda$. The failure instances arise out of Lemmas 9, 10 and the bound on $T_n(\mathcal{F}_n)$, the summation of whose probabilities are bounded by δ. ■

Remark 15 (Construction of covering for the SE kernel)**.** We demonstrate that such a construction is always possible using the SE kernel. Using the inequality $e^{-x} \geq 1-x$ for $x > 0$ we have,

$$1 - \phi_{\mathcal{X}}^2(D_{\mathcal{X}})\phi_{\mathcal{Z}}^2(D_{\mathcal{Z}}) < \frac{D_{\mathcal{X}}^2}{h_{\mathcal{X}}^2} + \frac{D_{\mathcal{Z}}^2}{h_{\mathcal{Z}}^2}$$

where $D_{\mathcal{Z}}, D_{\mathcal{X}}$ will be the L_2 diameters of the balls in the covering. Now let $h = \min\{h_{\mathcal{Z}}, h_{\mathcal{X}}\}$ and choose

$$D_{\mathcal{X}} = D_{\mathcal{Z}} = \frac{h}{2}\frac{\|\xi\|_\infty}{\beta_n^{1/2}}\lambda_r^q,$$

via which we have $1 - \phi_{\mathcal{Z}}^2(z)\phi_{\mathcal{X}}^2(x) < \frac{1}{2}\xi(\sqrt{p})^2\lambda_r^{2q}/\beta_n$ as stated in the proof. Noting that $\beta_n \asymp \log(n)$, using standard results on covering numbers, we can show that the size of this covering will be $\log(n)^{\frac{d+p}{2}}/\lambda_r^{q(d+p)}$. A similar argument is possible for Matérn kernels, but the exponent on $\log(n)$ will be worse.

Remark 16 (Choice of q for SE kernel)**.** From the arguments in our proof and Remark 15, we have that the number of plays in a set $S \subset (\mathcal{Z} \times \mathcal{X})$ is $T(S) \leq \text{vol}(S)\log(n)^{\frac{d+p+2}{2}}\left(\frac{\lambda(z_\bullet)}{\lambda_{min}}\right)^{q(p+d+2)}$. However, we chose to work work λ_{min} mostly to simplify the proof. It is not hard to see that for $A \subset \mathcal{X}$ and $B \subset \mathcal{Z}$ if $\lambda(z) \approx \lambda'$ for all $z \in B$, then $T_n(B, A) \approx \text{vol}(B \times A)\log(n)^{\frac{d+p+2}{2}}\left(\frac{\lambda(z_\bullet)}{\lambda'}\right)^{q(p+d+2)}$. As the capital spent in this region is $\lambda' T_n(A, B)$, by picking $q = 1/(p+d+2)$ we ensure that the capital expended for a certain $A \subset \mathcal{X}$ at all fidelities is roughly the same, i.e. for any A, the capital density in fidelities z such that $\lambda(z) < \lambda(\theta_t(A))$ will be roughly the same. Kandasamy et al. (2016c) showed that doing so achieved a nearly minimax optimal strategy for cumulative regret in K-armed bandits. While it is not clear that this is the best strategy for optimisation under GP assumptions, it did reasonably well in our experiments. We leave it to future work to resolve this.

B.3.1. Proof of Lemma 12

For brevity, we will denote $\theta = \theta_t(A)$. We will invoke the discretisation G_t used in Lemma 10 via which we have $\varphi_t([x_\star]_t) \geq f_\star - 1/t^2$ for all $t \geq 1$. Let $b = \text{argmax}_{x \in A}\varphi_t(x)$ be the maximiser of the upper confidence bound φ_t in A at time t. Now note that, $x_t \in A \implies \varphi_t(b) > \varphi_t([x_\star]_t) \implies \varphi_t(b) > f_\star - 1/t^2$. We therefore have,

$$\begin{aligned}\mathbb{P}\big(T_n([> \theta], A) > u\big) &\leq \mathbb{P}\big(\exists t : u+1 \leq t \leq n, \ \ \varphi_t(b) > f_\star - 1/t^2 \ \ \wedge \ \ \tau_{t-1}(\theta, b) < \gamma(\theta)\big) \\ &\leq \sum_{t=u+1}^{n} \mathbb{P}\big(\mu_{t-1}(b) - f(b) > f_\star - f(b) - \beta_t^{1/2}\sigma_{t-1}(b) - 1/t^2 \ \ \wedge \ \ \tau_{t-1}(\theta, b) < \gamma(\theta)\big) \quad (17)\end{aligned}$$

We now note that

$$\tau_{t-1}(\theta, b) < \gamma(\theta) \implies \sigma_{t-1}(b) < \gamma(\theta) + \sqrt{\kappa_0}\xi(\theta) \leq 2\sqrt{\kappa_0}\xi(\theta) \leq \frac{1}{\beta_t^{1/2}\rho}(f_\star - f(b)).$$

The first step uses Lemma 11. The second step uses the fact that $\gamma(\theta) = \sqrt{\kappa_0}\xi(\theta)(\lambda(z)/\lambda(z_\bullet))^{1/(p+d+2)} \leq \sqrt{\kappa_0}\xi(\theta)$ and the last step uses the definition of $Q_t(A)$ in (12) whereby we have $f_\star - f(x) \geq 2\rho\beta_t^{1/2}\sqrt{\kappa_0}\xi(\theta)$. Now plugging this back into (17), we can bound each term in the summation by,

$$\begin{aligned}\mathbb{P}\big(\mu_{t-1}(b) - f(b) > (\rho-1)\beta_t^{1/2}\sigma_{t-1}(b) - 1/t^2\big) &\leq \mathbb{P}_{Z\sim\mathcal{N}(0,1)}\left(Z > (\rho_0 - 1)\beta_t^{1/2}\right) \\ \leq \frac{1}{2}\exp\left(\frac{(\rho_0-1)^2}{2}\beta_t\right) &\leq \frac{1}{2}\left(\frac{2\delta}{\pi^2}\right)^{(\rho_0-1)^2} t^{-(\rho_0-1)^2(2+2d)} \leq \frac{\delta}{\pi^2}t^{-(\rho_0-1)^2(2+2d)}. \quad (18)\end{aligned}$$

In the first step we have used the following facts, $t > u \geq \max\{3, (2(\rho-\rho_0)\eta)^{-2/3}\}$, $\pi^2/2\delta > 1$ and $\sigma_{t-1}(b) > \eta/\sqrt{t}$ to conclude,

$$(\rho-\rho_0)\frac{\eta\sqrt{4\log(t)}}{\sqrt{t}} > \frac{1}{t^2} \implies (\rho-\rho_0)\cdot\sqrt{2\log\left(\frac{\pi^2t^2}{2\delta}\right)}\cdot\frac{\eta}{\sqrt{t}} > \frac{1}{t^2} \implies (\rho-\rho_0)\beta_t^{1/2}\sigma_{t-1}(b) > \frac{1}{t^2}.$$

The second step of (18) uses Lemma 5, the third step uses the conditions on $\beta_t^{1/2}$ as given in theorem 8 and the last step uses the fact that $\pi^2/2\delta > 1$. Now plug (18) back into (17). The result follows by bounding the sum by an integral and noting that $\rho_0 > 2$ and $\rho_0 \geq 1 + \sqrt{(1+2/\alpha)/(1+d)}$. ■

B.3.2. PROOF OF LEMMA 14

Let $A_s = \{u_1, u_2, \ldots, u_s\}$ be the queries in $\{z_\bullet\} \times A$ in the order they were queried. Now, assuming that we have queried g only inside $\{z_\bullet\} \times A$, denote by $\tilde{\sigma}_{t-1}(\cdot)$, the posterior standard deviation after $t-1$ such queries. Then,

$$\sum_{t:x_t \in A, z_t = z_\bullet} \sigma^2_{t-1}(x_t) \leq \sum_{t=1}^{s} \tilde{\sigma}^2_{t-1}(u_t) \leq \sum_{t=1}^{s} \eta^2 \frac{\tilde{\sigma}^2_{t-1}(u_t)}{\eta^2} \leq \sum_{t=1}^{s} \frac{\log(1+\eta^{-2}\tilde{\sigma}^2_{t-1}(u_t))}{\log(1+\eta^{-2})} \leq \frac{2}{\log(1+\eta^{-2})} I(y_{A_s}; f_{A_s}).$$

Queries outside $\{z_\bullet\} \times A$ will only decrease the variance of the GP so we can upper bound the first sum by the posterior variances of the GP with only the queries in $\{z_\bullet\} \times A$. The third step uses the inequality $u^2/v^2 \leq \log(1+u^2)/\log(1+v^2)$. The result follows from the fact that $\Psi_s(A)$ maximises the mutual information among all subsets of size s. ■

C. Addendum to Experiments

C.1. Implementation Details

We describe some of our implementation details below.

Domain and Fidelity space: Given a problem with arbitrary domain $\mathcal{X}$ and $\mathcal{Z}$, we mapped them to $[0,1]^d$ and $[0,1]^p$ by appropriately linear transforming the coordinates.

Initialisation: Following recommendations in Brochu et al. (2010) all GP methods were initialised with uniform random queries with $\Lambda/10$ capital, where Λ is the total capital used in the experiment. For GP-UCB and GP-EI all queries were initialised at $z_\bullet$ whereas for the multi-fidelity methods, the fidelities were picked at random from the available fidelities.

GP Hyper-parameters: Except in the first two experiments of Fig. 3, the GP hyper-parameters were learned after initialisation by maximising the GP marginal likelihood (Rasmussen & Williams, 2006) and then updated every 25 iterations. We use an SE kernel for both $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Z}}$ and instead of using one bandwidth for the entire fidelity space and domain, we learn a bandwidth for each dimension separately. We learn the kernel scale, bandwidths and noise variance using marginal likelihood. The mean of the GP is set to be the median of the observations.

Choice of β_t: β_t, as specified in Theorem 8 has unknown constants and tends to be too conservative in practice (Srinivas et al., 2010). Following the recommendations in Kandasamy et al. (2015) we set it to be of the correct "order"; precisely, $\beta_t = 0.5d\log(2\ell t + 1)$. Here, ℓ is the effective L_1 diameter of $\mathcal{X}$ and is computed by scaling each dimension by the inverse of the bandwidth of the SE kernel for that dimension.

Maximising φ_t: We used the DiRect algorithm (Jones et al., 1993).

Fidelity selection: Since we only worked in low dimensional fidelity spaces, the set $\mathcal{Z}_t$ was constructed in practice by obtaining a finely sampled grid of $\mathcal{Z}$ and then filtering out those which satisfied the 3 conditions in (7). In the second condition of (7), the threshold $\gamma(z)$ can be multiplied up to a constant factor, i.e $c\gamma(z)$ without affecting our theoretical results. In practice, we started with $c = 1$ but we updated it every 20 iterations via the following rule: if the algorithm has queried $z_\bullet$ more than 75% of the time in the last 20 iterations, we decrease it to $c/2$ and if it queried less than 25% of the time we increase it to $2c$. But the c value is always clipped inbetween 0.1 and 20. In practice we observed that the value for c usually stabilised around 1 and 8 although in some experiments it shot up to 20. Changing c this way resulted in slightly better performance in practice.

C.2. Description of Synthetic Functions

The following are the synthetic functions used in the paper.

GP Samples: For the GP samples in the first two experiments of Figure 3 we used an SE kernel with bandwidth 0.1 for $\phi_{\mathcal{X}}$. For $\phi_{\mathcal{Z}}$ we used bandwidths 1 and 0.01 for the first and second experiments respectively. The function was constructed by obtaining the GP function values on a 50×50 grid in the two dimensional $\mathcal{Z} \times \mathcal{X}$ space and then interpolating for evaluations in between via bivariate splines. For both experiments we used $\eta^2 = 0.05$ and the cost function $\lambda(z) = 0.2 + 6z^2$.

Currin exponential function: The domain is the two dimensional unit cube $\mathcal{X} = [0,1]^2$ and the fidelity was $\mathcal{Z} = [0,1]$ with $z_\bullet = 1$. We used $\lambda(z) = 0.1 + z^2$, $\eta^2 = 0.5$ and,

$$g(z,x) = \left(1 - 0.1(1-z)\exp\left(\frac{-1}{2x_2}\right)\right)\left(\frac{2300x_1^3 + 1900x_1^2 + 2092x_1 + 60}{100x_1^3 + 500x_1^2 + 4x_1 + 20}\right).$$

Hartmann functions: We used $g(z,x) = \sum_{i=1}^4 (\alpha_i - \alpha_i'(z)) \exp\big(-\sum_{j=1}^3 A_{ij}(x_j - P_{ij})^2\big)$. Here A, P are given below for the 3 and 6 dimensional cases and $\alpha = [1.0, 1.2, 3.0, 3.2]$. Then α_i' was set as $\alpha_i'(z) = 0.1(1 - z_i)$ if $i \leq p$ for $i = 1, 2, 3, 4$. We constructed the $p = 4$ and $p = 2$ Hartmann functions for the 3 and 6 dimensional cases respectively this way. When $z = z_\bullet = \mathbf{1}_p$, this reduces to the usual Hartmann function commonly used as a benchmark in global optimisation.

For the 3 dimensional case we used $\lambda(z) = 0.05 + (1 - 0.05)z_1^3 z_2^2$, $\eta^2 = 0.01$ and,

$$A = \begin{bmatrix} 3 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3 & 10 & 30 \\ 0.1 & 10 & 35 \end{bmatrix}, \quad P = 10^{-4} \times \begin{bmatrix} 3689 & 1170 & 2673 \\ 4699 & 4387 & 7470 \\ 1091 & 8732 & 5547 \\ 381 & 5743 & 8828 \end{bmatrix}.$$

For the 3 dimensional case we used $\lambda(z) = 0.05 + (1 - 0.05)z_1^3 z_2^2 z_3^{1.5} z_4^1$, $\eta^2 = 0.05$ and,

$$A = \begin{bmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{bmatrix}, \quad P = 10^{-4} \times \begin{bmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{bmatrix}.$$

Borehole function: This function was taken from (Xiong et al., 2013). We first let,

$$f_2(x) = \frac{2\pi x_3(x_4 - x_6)}{\log(x_2/x_1)\left(1 + \frac{2x_7 x_3}{\log(x_2/x_1)x_1^2 x_8} + \frac{x_3}{x_5}\right)},$$
$$f_1(x) = \frac{5x_3(x_4 - x_6)}{\log(x_2/x_1)\left(1.5 + \frac{2x_7 x_3}{\log(x_2/x_1)x_1^2 x_8} + \frac{x_3}{x_5}\right)}.$$

Then we define $g(z,x) = z f_2(x) + (1 - z) f_1(x)$. The domain of the function is $\mathcal{X} = [0.05, 0.15; 100, 50K; 63.07K, 115.6K;\ 990, 1110; 63.1, 116; 700, 820;\ 1120, 1680; 9855, 12045]$ and $\mathcal{Z} = [0,1]$ with $z_\bullet = 1$. We used $\lambda(z) = 0.1 + z^{1.5}$ for the cost function and $\eta^2 = 5$ for the noise variance.

Branin function: We use the following function where $\mathcal{X} = [[-5, 10], [0, 15]]^2$ and $\mathcal{Z} = [0,1]^3$.

$$g(z,x) = a(x_2 - b(z_1)x_1^2 + c(z_2)x_1 - r)^2 + s(1 - t(z))\cos(x_1) + s,$$

where $a = 1$, $b(z_1) = 5.1/(4\pi^2) - 0.01(1 - z_1)$ $c(z_2) = 5/\pi - 0.1(1 - z_2)$, $r = 6$, $s = 10$ and $t(z_3) = 1/(8\pi) + 0.05(1 - z_3)$. At $z = z_\bullet = \mathbf{1}_p$, this becomes the standard Branin function used as a benchmark in global optimisation. We used $\lambda(z) = 0.05 + z_1^3 z_2^2 z_3^{1.5}$ for the cost function and $\eta^2 = 0.05$ for the noise variance.