Journal of Artificial Intelligence Research 66 (2019) 151-196 Submitted 03/2019; published 09/2019

Multi-fidelity Gaussian Process Bandit Optimisation

Kirthevasan Kandasamy KANDASAMY@EECS.BERKELEY.EDU
University of California, Berkeley, CA, USA

Gautam Dasarathy GAUTAMD@ASU.EDU
Arizona State University, AZ, USA

Junier Oliva JOLIVA@CS.UNC.EDU
University of North Carolina at Chapel Hill, NC, USA

Jeff Schneider SCHNEIDE@CS.CMU.EDU
Barnabás Póczos BAPOCZOS@CS.CMU.EDU
Carnegie Mellon University, PA, USA

Abstract

In many scientific and engineering applications, we are tasked with the maximisation of an expensive to evaluate black box function f. Traditional settings for this problem assume just the availability of this single function. However, in many cases, cheap approximations to f may be obtainable. For example, the expensive real world behaviour of a robot can be approximated by a cheap computer simulation. We can use these approximations to eliminate low function value regions cheaply and use the expensive evaluations of f in a small but promising region and speedily identify the optimum. We formalise this task as a *multi-fidelity* bandit problem where the target function and its approximations are sampled from a Gaussian process. We develop MF-GP-UCB, a novel method based on upper confidence bound techniques. In our theoretical analysis we demonstrate that it exhibits precisely the above behaviour and achieves better bounds on the regret than strategies which ignore multi-fidelity information. Empirically, MF-GP-UCB outperforms such naive strategies and other multi-fidelity methods on several synthetic and real experiments.

1. Introduction

In stochastic bandit optimisation, we wish to optimise a function $f : \mathcal{X} \to \mathbb{R}$ by sequentially querying it and obtaining *bandit feedback*, i.e. when we query at any $x \in \mathcal{X}$, we observe a possibly noisy evaluation of $f(x)$. f is typically expensive and the goal is to identify its maximum while keeping the number of queries as low as possible. Some applications are hyper-parameter tuning in expensive machine learning algorithms (Snoek, Larochelle, & Adams, 2012), optimal policy search in complex systems (Martinez-Cantin, de Freitas, Doucet, & Castellanos, 2007), online advertising (Kar, Li, Narasimhan, Chawla, & Sebastiani, 2016), scientific experiments (Parkinson, Mukherjee, & Liddle, 2006), and statistical tasks such as collaborative filtering (S. Li, Karatzoglou, & Gentile, 2016) and clustering (Gentile et al., 2017). Historically, bandit problems were studied in settings where the goal is to maximise the cumulative reward of all queries to the payoff instead of just finding the maximum. Applications in this setting include clinical trials and online advertising.

Conventional methods in these settings assume access to only this single expensive function of interest f. We will collectively refer to them as *single fidelity* methods. In many practical problems however, cheap approximations to f might be available. For instance, when tuning hyper-parameters of learning algorithms, the goal is to maximise a cross validation score on a training set, which can be expensive if the training set is large. However validation curves tend to vary smoothly with training

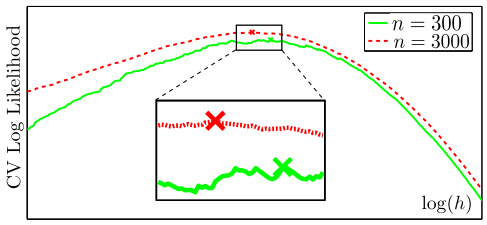


Figure 1: Average 5-fold CV log likelihood on datasets of size $300, 3000$ on a synthetic kernel density estimation task. The crosses are the maxima.

set size; therefore, we can train and cross validate on small subsets to approximate the validation accuracies of the entire dataset. For a concrete example, consider kernel density estimation (KDE), where we need to tune the bandwidth h of a kernel when using a dataset of size 3000. Figure 1 shows the average cross validation likelihood against h for a dataset of size $n = 3000$ and a smaller subset of size $n = 300$. Since the cross validation performance of a hyper-parameter depends on the training set size (Vapnik & Vapnik, 1998), we can obtain only a biased estimate of the cross validation performance with 3000 points using a subset of size 300. Consequently, the two maximisers are also different. That said, the curve for $n = 300$ approximates the $n = 3000$ curve quite well. Since training and cross validation on small n is cheap, we can use it to eliminate bad values of the hyper-parameters and reserve the expensive experiments with the entire dataset for the promising hyper-parameter values (for example, boxed region in Figure 1).

In the conventional treatment for online advertising, each query to f is, say, the public display of an ad on the internet for a certain time period. However, we could also choose smaller experiments by, say, confining the display to a small geographic region and/or for shorter periods. The estimate is biased, since users in different geographies are likely to have different preferences, but will nonetheless be useful in gauging the all round performance of an ad. In optimal policy search in robotics and autonomous driving, vastly cheaper computer simulations are used to approximate the expensive real world performance of the system (Cutler, Walsh, & How, 2014; Urmson et al., 2008). Scientific experiments can be approximated to varying degrees using less expensive data collection, analysis, and computational techniques (Parkinson et al., 2006).

In this paper, we cast these tasks as *multi-fidelity bandit optimisation* problems assuming the availability of cheap approximate functions (fidelities) to the payoff f. **Our contributions** are:

1. We present a formalism for multi-fidelity bandit optimisation using Gaussian process (GP) assumptions on f and its approximations. We develop a novel algorithm, Multi-Fidelity Gaussian Process Upper Confidence Bound (MF-GP-UCB) for this setting.

2. Our theoretical analysis proves that MF-GP-UCB explores the space $\mathcal{X}$ at lower fidelities and uses the high fidelities in successively smaller regions to converge on the optimum. As lower fidelity queries are cheaper, MF-GP-UCB has better upper bounds on the regret than single fidelity strategies which have to rely on the expensive function to explore the entire space.

3. We demonstrate that MF-GP-UCB outperforms single fidelity methods and other alternatives empirically, via a series of synthetic examples, three hyper-parameter tuning tasks and one inference problem in astrophysics. Our matlab implementation and experiments are available at `github.com/kirthevasank/mf-gp-ucb`.

Related Work

Since the seminal work by Robbins (1952), the multi-armed bandit problem has been studied extensively in the K-armed setting. Recently, there has been a surge of interest in the optimism under uncertainty principle for K-armed bandits, typified by upper confidence bound (UCB) methods (Auer, 2003; Bubeck & Cesa-Bianchi, 2012). UCB strategies have also been used in bandit tasks with linear (Dani, P. Hayes, & Kakade, 2008) and GP (Srinivas, Krause, Kakade, & Seeger, 2010) payoffs. There is a plethora of work on single fidelity methods for global optimisation both with noisy and noiseless evaluations. Some examples are branch and bound techniques such as dividing rectangles (DiRect), simulated annealing, genetic algorithms and more (Jones, Perttunen, & Stuckman, 1993; Kawaguchi, Kaelbling, & Lozano-Pérez, 2015; Kirkpatrick, Gelatt, & Vecchi, 1983; Munos, 2011). A suite of single fidelity methods in the GP framework closely related to our work is Bayesian Optimisation (BO). While there are several techniques for BO (Hernández-Lobato, Hoffman, & Ghahramani, 2014; Jones, Schonlau, & Welch, 1998; Mockus, 1994; Thompson, 1933), of particular interest to us is the Gaussian process upper confidence bound (GP-UCB) algorithm of Srinivas et al. (2010).

Many applied domains of research such as aerodynamics, industrial design and hyper-parameter tuning have studied multi-fidelity methods (Forrester, Sóbester, & Keane, 2007; Huang, Allen, Notz, & Miller, 2006; Klein, Bartels, Falkner, Hennig, & Hutter, 2015; L. Li, Jamieson, DeSalvo, Rostamizadeh, & Talwalkar, 2017; Swersky, Snoek, & Adams, 2013, 2014); a plurality of them use BO techniques. However these treatments neither formalise nor analyse any notion of *regret* in the multi-fidelity setting. In contrast, MF-GP-UCB is an intuitive UCB idea with good theoretical properties. Bogunovic, Scarlett, Krause, and Cevher (2016) study a version of BO where an algorithm might use cheap, noisy, yet unbiased approximations to a function f; but as we will explain in Section 2, this is different to the multi-fidelity problem. Agarwal, Duchi, Bartlett, and Levrard (2011) derive oracle inequalities for hyper-parameter tuning with ERM under computational budgets. Our setting is more general as it applies to any bandit optimisation task. Sabharwal, Samulowitz, and Tesauro (2015) present a UCB based idea for tuning hyper-parameters with incremental data allocation. However, their theoretical results are for an idealised non-realisable algorithm. Cutler et al. (2014) study reinforcement learning with multi-fidelity simulators by treating each fidelity as a Markov Decision Process. Finally, Zhang and Chaudhuri (2015) study active learning when there is access to a cheap weak labeler and an expensive strong labeler. These works study problems different to optimisation.

Recently, in Kandasamy, Dasarathy, Poczos, and Schneider (2016) we studied the classical K-armed bandit in multi-fidelity settings. Here, we build on this work to study multi-fidelity Bayesian optimisation; as such, we share similarities in the assumptions, algorithm, and some analysis techniques. A preliminary version of this paper appeared in Kandasamy, Dasarathy, Oliva, Schenider, and Póczos (2016) where we provided theoretical results in continuous domains and with two fidelities (one approximation). In this paper, we expand on the above and provide results both in discrete domains and for a general number of fidelities. Furthermore, we eliminate some technical assumptions from our previous work and present cleaner and more interpretable versions of our theorems. In follow up work (Kandasamy, Dasarathy, Schneider, & Poczos, 2017), we extend multi-fidelity optimisation to settings with continuous approximations. While the assumptions there are considerably different, it builds on the main intuitions from this work. To the best of our knowledge,

this is the first line of work to formalise a notion of regret and provide a theoretical analysis for multi-fidelity optimisation.

Subsequent to our work, there has been a line of research on multi-fidelity optimisation in various settings. Sen, Kandasamy, and Shakkottai (2018, 2019) develop an algorithm in frequentist settings which builds on the key intuitions here, i.e. query at low fidelities and proceed higher only when the uncertainty has shrunk. In addition, Song, Chen, and Yue (2018) develop a Bayesian algorithm which chooses fidelities based on the mutual information. Poloczek, Wang, and Frazier (2017); Wu, Toscano-Palmerin, Frazier, and Wilson (2019) use knowledge gradient methods for multi-fidelity Bayesian optimisation while Hoag and Doppa (2018) use techniques from search based optimisation for this problem.

The remainder of this manuscript is organised as follows. Section 2 presents our formalism including a notion of simple regret for multi-fidelity GP optimisation. Section 4 presents our algorithm. We present our theoretical results in Section 5 beginning with an informal discussion of results for $M = 2$ fidelities in Section 5.1 to elucidate the main ideas. The proofs are given in Section 8. Section 7 presents our experiments with some details deferred to Appendix A. Appendix B collects some ancillary material including a table of notations and abbreviations in Appendix B.2.

2. Problem Set Up

We wish to maximise a function $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is a finite discrete or compact subset of $[0, r]^d$, where $r > 0$ and d is the dimension of $\mathcal{X}$. We can interact with f only by querying it at some $x \in \mathcal{X}$ and obtaining a noisy evaluation $y = f(x) + \epsilon$ of f, where the noise satisfies $\mathbb{E}[\epsilon] = 0$. Let $x_\star \in \operatorname{argmax}_{x\in\mathcal{X}} f(x)$ be a maximiser of f and $f_\star = f(x_\star)$ be the maximum value. Let $x_t \in \mathcal{X}$ be the point queried at time t by a sequential procedure. The goal in bandit optimisation is to achieve small *simple regret* S_n, defined below, after n queries to f.

$$S_n = \min_{t=1,\ldots,n} f_\star - f(x_t). \tag{1}$$

Our primary distinction from the usual setting is that we have access to $M - 1$ successively accurate approximations $f^{(1)}, f^{(2)}, \ldots, f^{(M-1)}$ to the function of interest $f = f^{(M)}$. We refer to these approximations as fidelities. The multi-fidelity framework is attractive when the following two conditions are true about the problem.

1. *The approximations $f^{(1)}, \ldots, f^{(M-1)}$ approximate $f^{(M)}$.* To this end, we will assume a uniform bound for the fidelities, $\|f^{(M)} - f^{(m)}\|_\infty \leq \zeta^{(m)}$ for $m = 1, \ldots, M$, where the bounds $\zeta^{(1)} > \zeta^{(2)} > \cdots > \zeta^{(M)} = 0$ are known.

2. *The approximations are cheaper than evaluating at $f^{(M)}$.* We will assume that a query at fidelity m expends a cost $\lambda^{(m)}$ of a resource, such as computational effort or money. The costs are known and satisfy $0 < \lambda^{(1)} < \lambda^{(2)} < \cdots < \lambda^{(M)}$.

Above, and throughout this manuscript, for any $h : \mathcal{X} \to \mathbb{R}$, we define $\|h\|_\infty = \sup_{x\in\mathcal{X}} |h(x)|$. As the fidelity m increases, the approximations become better but are also more costly. An algorithm for multi-fidelity bandits is a sequence of query-fidelity pairs $\{(x_t, m_t)\}_{t\geq 0}$, where at time n, the algorithm chooses (x_n, m_n) using information from previous query-observation-fidelity triples $\{(x_t, y_t, m_t)\}_{t=1}^{n-1}$. Here $y_t = f^{(m_t)}(x_t) + \epsilon_t$ where, the ϵ_t values are independent noise at each time step t and $\mathbb{E}[\epsilon_t] = 0$.

Some smoothness assumptions on $f^{(m)}$'s are needed to make the problem tractable. A standard in the Bayesian nonparametric literature is to use a Gaussian process (GP) prior (Rasmussen & Williams, 2006) with covariance kernel κ. Two popular kernels of choice are the squared exponential (SE) kernel $\kappa_{\sigma,h}$ and the Matérn kernel $\kappa_{\nu,h}$. Writing $z = \|x - x'\|_2$, they are defined as

$$\kappa_{\sigma,h}(x,x') = \sigma \exp\left(-\frac{z^2}{2h^2}\right), \qquad \kappa_{\nu,\rho}(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}z}{\rho}\right)^{\nu} B_\nu\left(\frac{\sqrt{2\nu}z}{\rho}\right),$$

respectively. Here $\sigma, h, \nu, \rho > 0$ are parameters of the kernels and Γ, B_ν are the Gamma and modified Bessel functions. A convenience the GP framework offers is that posterior distributions are analytically tractable. If $f \sim \mathcal{GP}(0,\kappa)$ is a sample from a GP, and we have observations $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^t$, where $y_i = f(x_i) + \epsilon$ and $\epsilon \sim \mathcal{N}(0,\eta^2)$ is Gaussian noise, then the posterior distribution for $f(x)|\mathcal{D}_t$ is also Gaussian $\mathcal{N}(\mu_t(x), \sigma_t^2(x))$ with (Rasmussen & Williams, 2006),

$$\mu_t(x) = k^\top (K + \eta^2 I_t)^{-1} Y, \qquad \sigma_t^2(x) = \kappa(x,x) - k^\top (K + \eta^2 I_t)^{-1} k. \tag{2}$$

Here, $Y \in \mathbb{R}^t$ is a vector with $Y_i = y_i$, $k \in \mathbb{R}^t$ is a vector with $k_i = \kappa(x, x_i)$. The matrix $K \in \mathbb{R}^{t\times t}$ is given by $K_{i,j} = \kappa(x_i, x_j)$. $I_t \in \mathbb{R}^{t\times t}$ is the $t \times t$ identity matrix.

2.1 The Generative Process for Multi-fidelity Optimisation

In keeping with the above framework, we assume the following generative model for the functions $f^{(1)}, \ldots, f^{(M)}$. A generative mechanism is given constants $\zeta^{(1)}, \ldots, \zeta^{(M-1)}$. It then generates the functions as follows.

Step 1. Sample $f^{(m)} \sim \mathcal{GP}(0,\kappa)$ for $m = 1, \ldots, M$. **(A1)**

Step 2. Check if $\|f^{(M)} - f^{(m)}\|_\infty \leq \zeta^{(m)}$ for all $m = 1, \ldots, M-1$. If true, then deliver $f^{(1)}, \ldots, f^{(M)}$. If false, go back to Step 1. **(A2)**

In addition to this, we will also assume that upon querying $f^{(m)}$ at x_t we observe $f^{(m)}(x_t) + \epsilon$ where $\epsilon \sim \mathcal{N}(0,\eta^2)$ is Gaussian noise with variance η^2. Note that for well behaved kernels, such as the SE and Matérn kernels, GP sample paths are continuous with probabiliy 1 (Adler, 1990).

Condition **A2** characterises the approximation conditions for the lower fidelities. Lemma 2 shows that **A2** is satisfied with positive probability when $f^{(1)}, \ldots, f^{(M)}$ are sampled from a GP. Hence this is a valid generative process since **A2** will eventually be satisfied. Moreover, in Section 4 we argue that while **A2** renders the computation of the true posterior of all GPs inefficient via closed form equations such as in (2), it is still possible to derive an efficient algorithm that uses (2) to determine future points for evaluation.

We note that other natural approximation conditions can be used to characterise the cheaper fidelities. We choose a uniform bound condition because it provides a simple way to reason about one fidelity from the others, hence keeping the analysis tractable while ensuring the model is interesting enough so that empirical performance is not compromised. Our theoretical analysis assumes that the algorithm needs to know the uniform bounds $\zeta^{(1)}, \ldots, \zeta^{(M-1)}$ which can be unrealistic in practical settings. In Section 6 we describe a heuristic for choosing these values in a data dependent manner. That said, we believe that the intuitions in this work can be used to develop other upper confidence based multi-fidelity BO algorithms for other approximation conditions. In fact, the approximation

conditions in our follow up work in Kandasamy et al. (2017), are of a Bayesian flavour via a kernel on the fidelities. The algorithm, BOCA, builds on the key insights developed here.

It is worth mentioning that while our theoretical results are valid for arbitrary In this work, we will assume that M is a small fixed value and that $\lambda^{(1)}$ is comparable to $\lambda^{(M)}$. For instance, in many practical applications of multi-fidelity optimisation, while an approximation may be cheaper than the real experiment, it could itself be quite expensive and hence require an intelligence procedure, such as Bayesian optimisation, to choose the next point. This is the regime the current paper focuses on, as opposed to asymptotic regimes where $M \to \infty$ and/or $\lambda^{(1)} \to 0$. Moreover, very large values of M are better handled by the formalism in our follow up work in Kandasamy et al. (2017).

Finally, we note that Assumption **A1** can be relaxed to hold for different kernels and noise variances for each fidelity, i.e. different $\kappa^{(m)}, \eta^{(m)}$ for $m = 1, \ldots, M$, with minimal modifications to our analysis but we use the above form to simplify the presentation of the results. In fact, our practical implementation uses different kernels.

2.2 Simple Regret for Multi-fidelity Optimisation

Our goal is to achieve small simple regret $S(\Lambda)$ after spending capital Λ of a resource. We will aim to provide *any-capital* bounds, meaning that we will assume that the game is played indefinitely and will try to bound the regret for all (sufficiently large) values of Λ. This is similar in spirit to any-time analyses in single fidelity bandit methods as opposed to fixed time horizon analyses. Let $\{m_t\}_{t\geq 0}$ be the fidelities queried by a multi-fidelity method at each time step. Let N be the *random* quantity such that $N = \max\{n \geq 1 : \sum_{t=1}^{n} \lambda^{(m_t)} \leq \Lambda\}$, i.e. it is the number of queries the strategy makes across all fidelities until capital Λ. Only the optimum of $f = f^{(M)}$ is of interest to us. The lower fidelities are useful to the extent that they help us optimise $f^{(M)}$ with less cost, but there is no reward for optimising a cheaper approximation. Accordingly, we set the instantaneous reward q_t at time t to be $-\infty$ if $m_t \neq M$ and $f^{(M)}(x_t)$ if $m_t = M$. If we let $r_t = f_\star - q_t$ denote the instantaneous regret, we have $r_t = +\infty$ if $m_t \neq M$ and $f_\star - f^{(M)}(x_t)$ if $m_t = M$. For optimisation, the simple regret is simply the best instantaneous regret, $S(\Lambda) = \min_{t=1,\ldots,N} r_t$. Equivalently,

$$S(\Lambda) = \min_{t=1,\ldots,N} r_t = \begin{cases} \min\limits_{\substack{t=1,\ldots,N \\ t\,:\,m_t=M}} f_\star - f^{(M)}(x_t) & \text{if we have queried at the } M^{\text{th}} \text{ fidelity at least once,} \\ +\infty & \text{otherwise.} \end{cases} \tag{3}$$

Note that the above reduces to S_n in (1) when we only have access to $f^{(M)}$ with $n = N = \lfloor \Lambda/\lambda^{(M)} \rfloor$.

Before we proceed, we note that it is customary in the bandit literature to analyse *cumulative regret*. The definition of cumulative regret depends on the application at hand (Kandasamy, Dasarathy, Poczos, & Schneider, 2016) and our results can be extended to many sensible notions of cumulative regret. However, both to simplify exposition and since our focus in this paper is optimisation, we stick to simple regret.

Challenges: We conclude this subsection with a commentary on some of the challenges in multi-fidelity optimisation using Figure 2 for illustration. For simplicity, we will focus on 2 fidelities when we have one approximation $f^{(1)}$ to an expensive function $f^{(2)}$. For now assume that (unrealistically) $f^{(1)}$ and its optimum $x_\star^{(1)}$ are known. Typically $x_\star^{(1)}$ is suboptimal for $f^{(2)}$. A seemingly straightforward solution might be to search for $x_\star$ in an appropriate subset, such as a neighborhood

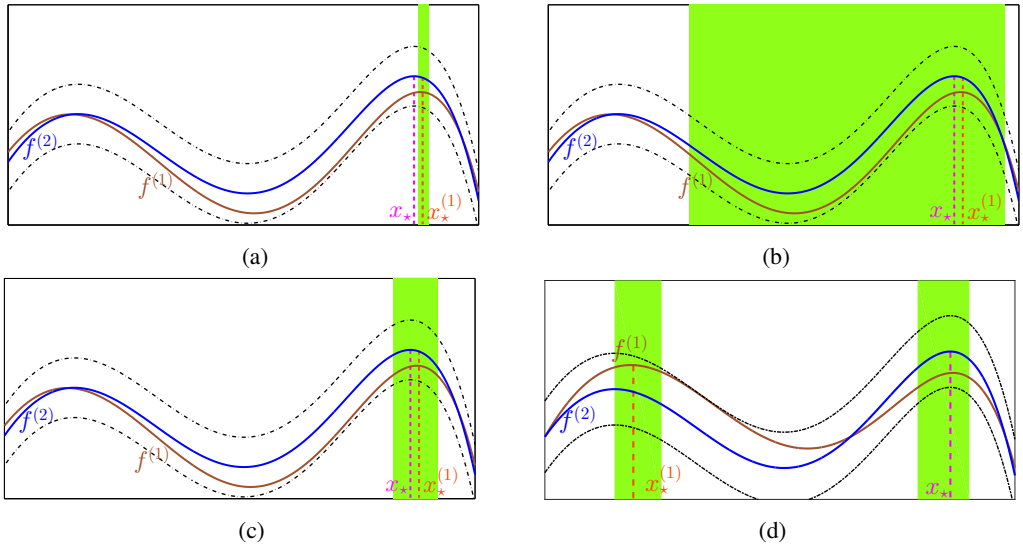


Figure 2: An illustration of the challenges in multi-fidelty optimisation. See Section 2.

of $x_\star^{(1)}$. However, if this neighborhood is too small, we might miss the optimum $x_\star$ (green region in Figure 2(a)). A crucial challenge for multi-fidelity methods is to not get stuck at the optimum of a lower fidelity. While exploiting information from lower fidelities, it is also important to *explore* sufficiently at higher fidelities. In our experiments, we demonstrate that naive strategies which do not do so could get stuck at the optimum of a lower fidelity. Alternatively, if we pick a very large subset (Figure 2(b)) we might not miss $x_\star$; however, it defeats the objectives of the multi-fidelity set up where the goal is to use the approximation to be prudent about where we query $f^{(2)}$. Figure 2(c) displays a seemingly sensible subset, but it remains to be seen how it is chosen. Further, this subset might not even be a neighborhood as illustrated in Figure 2(d), where $f^{(1)}, f^{(2)}$ are multi-modal and the optima are in different modes. In such cases, an appropriate algorithm should explore all such modes. On top of the above, an algorithm does not actually know $f^{(1)}$. A sensible algorithm should explore $f^{(1)}$ and simultaneously identify the above subset, either implicitly or explicitly, for exploration at the second fidelity $f^{(2)}$. Finally, it is also important to note that $f^{(1)}$ is not simply a noisy version of $f^{(2)}$; this setting is more challenging as an algorithm needs to explicitly account for the bias in the approximations.

2.3 Some Useful Properties of GPs

For what follows, we present some useful properties and concepts related to GPs with well behaved kernels. We we will denote probabilities when $f^{(1)}, \ldots, f^{(M)} \sim \mathcal{GP}(0, \kappa)$ independently, by $\mathbb{P}_{\mathcal{GP}}$. $\mathbb{P}$ will denote probabilities under the prior in the multi-fidelity setting which includes **A2** after sampling the functions; i.e. for any event E, $\mathbb{P}(E) = \mathbb{P}_{\mathcal{GP}}(E|\mathbf{A2})$. First, 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 $\nu > 2$; see Ghosal and Roy (2006), Theorem 5.

Assumption 1. *(Theorem 5 in Ghosal and Roy, 2006) Let $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$, where $\kappa : [0, r]^d \times [0, r]^d \to \mathbb{R}$ is a stationary kernel (Rasmussen & Williams, 2006). The partial derivatives of f satisfies the*

following condition. There exist constants $a, b > 0$ such that,

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

Observe that we have used notation $\mathbb{P}_{\mathcal{GP}}$ to indicate the prior probability when $f \sim \mathcal{GP}(0, \kappa)$ for consistency. Next, the following assumption supposes that there is a positive probability to the event that the supremum of a GP in a bounded domain is smaller than any given $\epsilon > 0$.

Assumption 2. *Let $\mathcal{X} = [0, r]^d$ and $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$. Let $\kappa : \mathcal{X} \times \mathcal{X}^d \to \mathbb{R}$ be such that for all $\epsilon > 0$, there exists $\mathrm{Q}(\epsilon) > 0$ such that,*

$$\mathbb{P}_{\mathcal{GP}}\Big(\sup_{x \in \mathcal{X}} |f(x)| < \epsilon\Big) > \mathrm{Q}(\epsilon).$$

As shown by Theorem 4 in Ghosal and Roy (2006), this is satisfied for the SE and Matérn kernels. Finally, following Srinivas et al. (2010), our theoretical results will be given in terms of the *Maximum Information Gain* (MIG), defined below.

Definition 1. *(Maximum Information Gain Srinivas et al., 2010) Let $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$. Consider any $A \subset \mathbb{R}^d$ any let $\widetilde{A} = \{x_1, \ldots, x_n\} \subset A$ be a finite subset. Let $f_{\widetilde{A}}, \epsilon_{\widetilde{A}} \in \mathbb{R}^n$ be such that $(f_{\widetilde{A}})_i = f(x_i)$, $(\epsilon_{\widetilde{A}})_i \sim \mathcal{N}(0, \eta^2)$ for $i = 1, \ldots, n$, and $y_{\widetilde{A}} = f_{\widetilde{A}} + \epsilon_{\widetilde{A}}$. Let I denote the Shannon mutual information. The Maximum Information Gain $\Psi_n(A)$ of set A after n evaluations is the maximum mutual information between the function values and observations among all choices of n points in A. Precisely,*

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

The MIG, which depends on the kernel and the set A, will be an important quantity in our analysis as it characterises the statistical difficulty of GP Bandits. For a given kernel it typically scales with the volume of A (Srinivas et al., 2010)[1]. For example, if $A = [0, r]^d$ then $\Psi_n(A) \in \mathcal{O}(r^d \Psi_n([0, 1]^d))$. It is known that for the SE kernel, $\Psi_n([0, 1]^d) \in \mathcal{O}((\log(n))^{d+1})$ and for the Matérn kernel, $\Psi_n([0, 1]^d) \in \mathcal{O}(n^{\frac{d(d+1)}{2\nu + d(d+1)}} \log(n))$ (Seeger, Kakade, & Foster, 2008; Srinivas et al., 2010).

3. A Review of GP-UCB

Sequential optimisation methods adopting UCB principles maintain a high probability upper bound $\varphi_t : \mathcal{X} \to \mathbb{R}$ for $f(x)$ for all $x \in \mathcal{X}$ (Auer, 2003). At time t we query at the maximiser of this upper bound $x_t = \mathrm{argmax}_{x \in \mathcal{X}} \varphi_t(x)$. Our work builds on GP-UCB (Srinivas et al., 2010), where φ_t takes the form $\varphi_t(x) = \mu_{t-1}(x) + \beta_t^{1/2} \sigma_{t-1}(x)$. Here μ_{t-1}, σ_{t-1} are the posterior mean and standard deviation of the GP conditioned on the previous $t-1$ queries $\{(x_i, y_i)\}_{i=1}^{t-1}$ and $\beta_t > 0$. The key intuition here is that the mean μ_{t-1} encourages an exploitative strategy – in that we want to query where we know the function is high – and the standard deviation σ_{t-1} encourages an explorative strategy – in that we want to query at regions we are uncertain about f lest we miss out on high

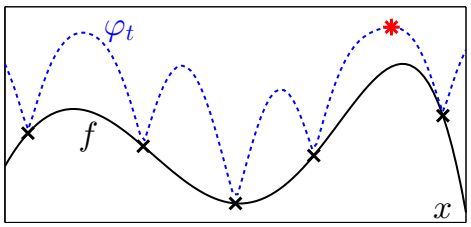


Figure 3: Illustration of GP-UCB. The solid black line is $f(x)$ and the dashed blue line is $\varphi_t(x)$. The observations until $t-1$ are shown as black crosses. At time t, we query at the maximiser $x_t = \text{argmax}_{x\in\mathcal{X}} \varphi_t(x)$ shown via the red star.

Algorithm 1 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 (2).

valued regions. β_t will control the trade-off between exploration and exploitation. We have presented GP-UCB in Algorithm 1 and illustrated it in Figure 3.

The following theorem from Srinivas et al. (2010) bounds the simple regret S_n (1) for GP-UCB. They give their bounds in terms of the cumulative regret, but converting it to simple regret is straightforward.

Theorem 1. *(Theorems 1 and 2 in Srinivas et al., 2010) Let $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$, $f : \mathcal{X} \to \mathbb{R}$ and the kernel κ satisfies Assumption 1). 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 $\delta \in (0,1)$. The following bounds on the simple regret S_n hold with $\mathbb{P}_{\mathcal{GP}}$-probability $> 1-\delta$ for all $n \geq 1$.*

- *If $\mathcal{X}$ is a finite discrete set, run* GP-UCB *with $\beta_t = 2\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)$. Then,*

$$\text{for all } n \geq 1, \qquad S_n \leq \sqrt{\frac{C_1\beta_n\Psi_n(\mathcal{X})}{n}}$$

- *If $\mathcal{X} = [0,r]^d$, run* GP-UCB *with $\beta_t = 2\log\left(\frac{2\pi^2t^2}{3\delta}\right) + 2d\log\left(t^2bdr\sqrt{\frac{4ad}{\delta}}\right)$. Then,*

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

1. In section C.2 of Srinivas et al. (2010), the kernel's eigenspectrum is defined with respect to the uniform measure on the domain $\mathcal{X}$. When we consider any subset $A \subset \mathcal{X}$ with the same measure and eigenspectrum, a multiplicative $\text{vol}(A)$ term appears.

4. Multi-fidelity Gaussian Process Upper Confidence Bound (MF-GP-UCB)

We now propose MF-GP-UCB, which extends GP-UCB to the multi-fidelity setting. Like GP-UCB, MF-GP-UCB will also maintain a UCB for $f^{(M)}$ obtained via the previous queries at *all* fidelities. Denote the posterior GP mean and standard deviation of $f^{(m)}$ conditioned *only* on the previous queries at fidelity m by $\mu_t^{(m)}, \sigma_t^{(m)}$ respectively (See (2)). Then define,

$$\varphi_t^{(m)}(x) = \mu_{t-1}^{(m)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(m)}(x) + \zeta^{(m)}, \quad \forall\, m, \qquad\qquad \varphi_t(x) = \min_{m=1,\ldots,M} \varphi_t^{(m)}(x). \tag{4}$$

For appropriately chosen β_t, $\mu_{t-1}^{(m)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(m)}(x)$ will upper bound $f^{(m)}(x)$ with high probability. By **A2** and (4), $\varphi_t^{(m)}(x)$ upper bounds $f^{(M)}(x)$ for all m. We have M such bounds, and their minimum $\varphi_t(x)$ gives the best upper bound for $f^{(M)}$. Following UCB strategies such as GP-UCB, our next query is at the maximiser of this UCB, $x_t = \text{argmax}_{x\in\mathcal{X}}\, \varphi_t(x)$.

Next we need to decide which fidelity to query at. Consider any $m < M$. The $\zeta^{(m)}$ constraints on $f^{(m)}$ restrict the value of $f^{(M)}$ – the confidence band $\beta_t^{1/2}\sigma_{t-1}^{(m)}$ for $f^{(m)}$ is lengthened by $\zeta^{(m)}$ to obtain confidence on $f^{(M)}$. If $\beta_t^{1/2}\sigma_{t-1}^{(m)}(x_t)$ for $f^{(m)}$ is large, it means that we have not constrained $f^{(m)}$ sufficiently well at x_t and should query at the m^{th} fidelity. On the other hand, querying indefinitely in the same region to reduce the uncertainty $\beta_t^{1/2}\sigma_{t-1}^{(m)}$ at the m^{th} fidelity in that region will not help us much as the $\zeta^{(m)}$ elongation caps off how much we can learn about $f^{(M)}$ from $f^{(m)}$; i.e. even if we knew $f^{(m)}$ perfectly, we will only have constrained $f^{(M)}$ to within a $\pm\zeta^{(m)}$ band. Our algorithm captures this simple intuition. Having selected x_t, we begin by checking at the first fidelity. If $\beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t)$ is smaller than a threshold $\gamma^{(1)}$, we proceed to the second fidelity. If at any stage $\beta_t^{1/2}\sigma_{t-1}^{(m)}(x_t) \geq \gamma^{(m)}$ we query at fidelity $m_t = m$. If we proceed all the way to fidelity M, we query at $m_t = M$. We will discuss choices for $\gamma^{(m)}$ in Sections 5.1 and 6. We summarise the resulting procedure in Algorithm 2.

Algorithm 2 MF-GP-UCB

Inputs: kernel κ, bounds $\{\zeta^{(m)}\}_{m=1}^M$, thresholds $\{\gamma^{(m)}\}_{m=1}^M$.

- For $m = 1, \ldots, M$: $\quad \mathcal{D}_0^{(m)} \leftarrow \varnothing, \quad (\mu_0^{(m)}, \sigma_0^{(m)}) \leftarrow (\mathbf{0}, \kappa^{1/2})$.
- for $t = 1, 2, \ldots$
  1. $x_t \leftarrow \text{argmax}_{x\in\mathcal{X}}\ \varphi_t(x)$. See (4) for φ_t and Sections 5, 6 for β_t.
  2. $m_t = \min\{\, m \,|\beta_t^{1/2}\sigma_{t-1}^{(m)}(x_t) \geq \gamma^{(m)} \ \text{ or } \ m = M\}$.
  3. $y_t \leftarrow$ Query $f^{(m_t)}$ at x_t.
  4. Update $\mathcal{D}_t^{(m_t)} \leftarrow \mathcal{D}_{t-1}^{(m_t)} \cup \{(x_t, y_t)\}$. Obtain $\mu_t^{(m_t)}, \sigma_t^{(m_t)}$ conditioned on $\mathcal{D}_t^{(m_t)}$ (2). Set $\mathcal{D}_t^{(m)} \leftarrow \mathcal{D}_{t-1}^{(m)}$, $\mu_t^{(m)} \leftarrow \mu_{t-1}^{(m)}$, $\sigma_t^{(m)} \leftarrow \sigma_{t-1}^{(m)}$ for $m \neq m_t$.

Before we proceed, we make an essential observation. The posterior for any $f^{(m)}(x)$ conditioned on previous queries at *all* fidelities $\bigcup_{\ell=1}^M \mathcal{D}_t^{(\ell)}$ is not Gaussian due to the $\zeta^{(m)}$ constraints (**A2**). However, $|f^{(m)}(x) - \mu_{t-1}^{(m)}(x)| < \beta_t^{1/2}\sigma_{t-1}^{(m)}(x)$ holds with high probability, since, by conditioning only on queries at the m^{th} fidelity we have Gaussianity for $f^{(m)}(x)$. (See Lemma 9, Section 8.1).

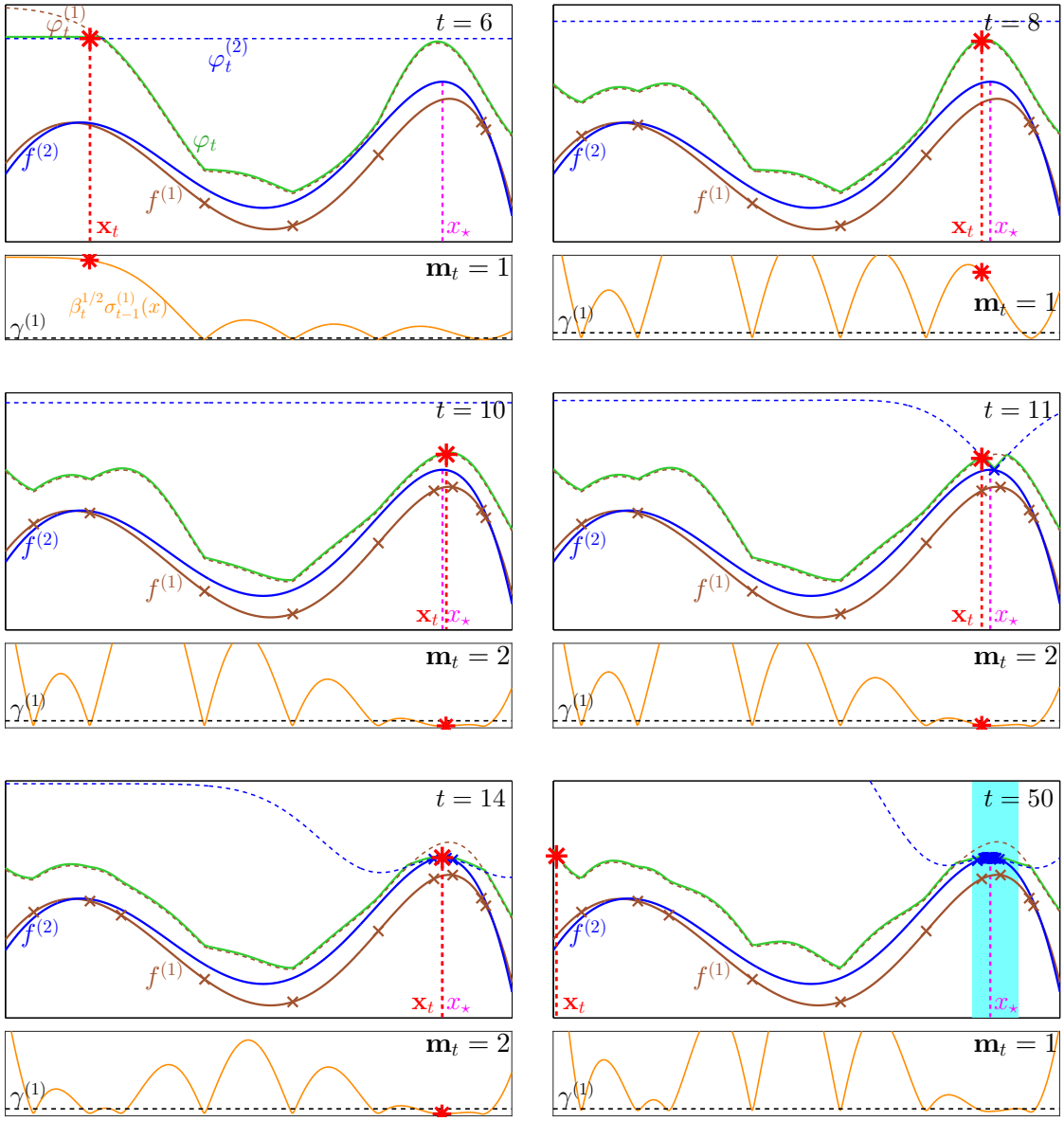


Figure 4: The 6 panels illustrate an execution of MF-GP-UCB in 2 fidelities at times $t = 6, 8, 10, 11, 14, 50$. In each panel, the top figure illustrates the upper bounds and selection of x_t while the bottom figure illustrates the selection of m_t. We have initialised MF-GP-UCB with 5 random points at the first fidelity. In the top figures, the solid lines in brown and blue are $f^{(1)}, f^{(2)}$ respectively, and the dashed lines are $\varphi_t^{(1)}, \varphi_t^{(2)}$. The solid green line is $\varphi_t = \min(\varphi_t^{(1)}, \varphi_t^{(2)})$. The small crosses are queries from 1 to $t-1$ and the red star is the maximiser of φ_t, i.e. the next query x_t. $x_\star$, the optimum of $f^{(2)}$ is shown in magenta. In the bottom figures, the solid orange line is $\beta_t^{1/2}\sigma_{t-1}^{(1)}$ and the dashed black line is $\gamma^{(1)}$. When $\beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t) \leq \gamma^{(1)}$ we play at fidelity $m_t = 2$ and otherwise at $m_t = 1$. The cyan region in the last panel is the good set $\mathcal{X}_g^\star$ described in Section 5.1.

An Illustration of MF-GP-UCB: Figure 4 illustrates MF-GP-UCB via a simulation on a 2–fidelity problem. At the initial stages, MF-GP-UCB is mostly exploring $\mathcal{X}$ in the first fidelity. $\beta_t^{1/2}\sigma_{t-1}^{(1)}$ is large and we are yet to constrain $f^{(1)}$ well to proceed to $m = 2$. At $t = 10$, we have constrained $f^{(1)}$ sufficiently well at a region around the optimum. $\beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t)$ falls below $\gamma^{(1)}$ and we query at $m_t = 2$. Notice that once we do this (at $t = 11$), $\varphi_t^{(2)}$ dips to change φ_t in that region. At $t = 14$, MF-GP-UCB has identified the maximum $x_\star$ with just 4 queries to $f^{(2)}$. The region shaded in cyan in the last figure is the "good set" $\mathcal{X}_g^\star$, which we alluded to in Section 2. We will define it formally and explain its significance in the multi-fidelity set up shortly. Our analysis predicts that most second fidelity queries in MF-GP-UCB will be be confined to this set (roughly) and the simulation corroborates this claim. For example, in the last figure, at $t = 50$, the algorithm decides to explore at a point far away from the optimum. However, this query occurs in the first fidelity since we have not sufficiently constrained $f^{(1)}(x_t)$ in this region and $\beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t)$ is large. The key idea is that it is *not necessary* to query such regions at the second fidelity as the first fidelity alone is enough to conclude that it is suboptimal. In addition, observe that in a large portion of $\mathcal{X}$, φ_t is given by $\varphi_t^{(1)}$ except in a small neighborhood around $x_\star$, where it is given by $\varphi_t^{(2)}$.

Next we present our main theoretical results. We wish to remind the reader that a table of notations is available in Appendix B.2.

5. Theoretical Results

First and foremost, we will show that condition **A2** occurs with positive probability when we sample the functions from a GP. The following lemma shows that $\mathbb{P}_{\mathcal{GP}}(\mathbf{A2}) = \xi_{\mathbf{A2}} > 0$ which establishes that the generative mechanism is valid. The proof is given in Section 8.

Lemma 2. *Let $f^{(1)}, \ldots, f^{(M)}$ be sampled from $\mathcal{GP}(0, \kappa)$ and* **A2** *denote the event $\{\|f^{(M)} - f^{(m)}\|_\infty \leq \zeta^{(m)}, \forall m \leq M-1\}$. Then,*

$$\mathbb{P}_{\mathcal{GP}}(\mathbf{A2}) = \xi_{\mathbf{A2}} \;\geq\; \mathrm{Q}\left(\frac{\zeta^{(M-1)}}{2}\right) \cdot \prod_{m=1}^{M-1} \mathrm{Q}\left(\frac{\zeta^{(m)}}{2}\right) \tag{5}$$

Here Q *is from Assumption 2. $\xi_{\mathbf{A2}} > 0$ since each of the terms in the product are positive.*

We are now ready to present our theoretical results. We begin with an informal yet intuitive introuduction to our theorems in $M = 2$ fidelities.

5.1 A Preview of our Theorems

In this subsection, we will ignore constants and polylog terms when they are dominated by other terms. $\lesssim, \gtrsim, \asymp$ denote inequality and equality ignoring constants. When $A \subset \mathcal{X}$, we will denote its complement by $\overline{A}$.

Fundamental to the 2-fidelity problem is the good set $\mathcal{X}_g^\star = \{x \in \mathcal{X}; f_\star - f^{(1)}(x) \leq \zeta^{(1)}\}$. $\mathcal{X}_g^\star$ is a high-valued region for $f^{(2)}(x)$: for all $x \in \mathcal{X}_g^\star$, $f^{(2)}(x)$ is at most $2\zeta^{(1)}$ away from the optimum. If a multi-fidelity strategy were to use *all* its second fidelity queries only in $\mathcal{X}_g^\star$, then, by Theorem 1, the regret will only have $\Psi_n(\mathcal{X}_g^\star)$ dependence after n high fidelity queries. In contrast, a strategy that only operates at the highest fidelity, such as GP-UCB, will have $\Psi_n(\mathcal{X})$ dependence. When $\zeta^{(1)}$ is small, i.e. when $f^{(1)}$ is a good approximation to $f^{(2)}$, $\mathcal{X}_g^\star$ will be much smaller than $\mathcal{X}$.

Then, $\Psi_n(\mathcal{X}_g^\star) \ll \Psi_n(\mathcal{X})$, and the multi-fidelity strategy will have better bounds on the regret than a single fidelity strategy. Alas, achieving this somewhat ideal goal is not possible without perfect knowledge of the approximation. However, with MF-GP-UCB we can come quite close. As we will show shortly, *most* second fidelity queries will be confined to the slightly inflated good set $\mathcal{X}_g = \{x \in \mathcal{X}; f_\star - f^{(1)}(x) \leq \zeta^{(1)} + 3\gamma^{(1)}\}$. The following lemma bounds the number of first and second fidelity evaluations in $\mathcal{X}_g$ and its complement $\overline{\mathcal{X}_g}$. We denote the number of queries at the m^{th} fidelity in a set $A \subset \mathcal{X}$ within the first n time steps by $T_n^{(m)}(A)$.

Lemma 3 (Informal, Bounding the number of evaluations for $M = 2$)**.** *Let $\mathcal{X} \subset [0, r]^d$. Consider* MF-GP-UCB *after n total evaluations at either fidelity. Let $T_n^{(m)}(A)$ denote the number of fidelity m queries in some set $A \subset \mathcal{X}$ in n steps. Then,*

$$T_n^{(1)}\left(\overline{\mathcal{X}}_g\right) \lesssim \text{polylog}(n) \cdot \Pi(\mathcal{X}_g), \qquad T_n^{(1)}(\mathcal{X}_g) \lesssim \frac{\text{polylog}(n)}{{\gamma^{(1)}}^2} \cdot \Pi(\mathcal{X}_g),$$

$$T_n^{(2)}\left(\overline{\mathcal{X}}_g\right) \lesssim \tau_n \cdot \Pi\left(\overline{\mathcal{X}}_g\right), \qquad T_n^{(2)}(\mathcal{X}_g) \asymp n.$$

Here $\Pi(A) = |A|$ for discrete A and $\Pi(A) = \text{vol}(A)$ for continuous A. The bound for $T_n^{(2)}\left(\overline{\mathcal{X}}_g\right)$ holds for any sublinear increasing sequence $\{\tau_n\}_{n\geq 1}$.

The above lemma will be useful for two reasons. First, the bounds on $T_n^{(2)}(\cdot)$ show that most second fidelity queries are inside $\mathcal{X}_g$; the number of such expensive queries outside $\mathcal{X}_g$ is small. This *strong* result is only possible in the multi-fidelity setting. From the results of Srinivas et al. (2010), we can infer that the best achievable bound on the number of plays for GP-UCB inside a suboptimal set is $\asymp n^{1/2}$ for the SE kernel and even worse for the Matérn kernel. For example, in the simulation of Figure 4, all queries to $f^{(2)}$ are in fact confined to $\mathcal{X}_g^\star$ which is a subset of $\mathcal{X}_g$. This allows us to obtain regret that scales with $\Psi_n(\mathcal{X}_g)$ as explained above. Second, we will use Lemma 3 to control N, the (random) number of queries by MF-GP-UCB within capital Λ. Let $\underline{n}_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$ be the (non-random) number of queries by a single fidelity method operating only at the second fidelity. As $\lambda^{(1)} < \lambda^{(2)}$, N could be large for an arbitrary multi-fidelity method. However, using the bounds on $T_n^{(1)}(\cdot)$ we can show that N is $\asymp \underline{n}_\Lambda$ when Λ is larger than some value Λ_0. Below, we detail the main ingredients in the proof of Lemma 3.

- $T_n^{(1)}(\mathcal{X}_g)$: By the design of our algorithm, MF-GP-UCB will begin querying $f^{(1)}$. To achieve finite regret we need to show that we will eventually query $f^{(2)}$. For any region in $\mathcal{X}_g$ the switching condition of step 2 in Algorithm 2 ensures that we do not query that region indefinitely. That is, if we keep querying a certain region, the first fidelity GP uncertainty $\beta_t^{1/2}\sigma_{t-1}^{(m)}$ will reduce below $\gamma^{(1)}$ in that region. We will discuss the implications of the choice of $\gamma^{(1)}$ at the end of this subsection and in Section 6.
- $T_n^{(1)}\left(\overline{\mathcal{X}}_g\right)$: For queries to $f^{(1)}$ outside $\mathcal{X}_g$, we use the following reasoning: as $f^{(1)}$ is small outside $\mathcal{X}_g$, it is unlikely to contain the UCB maximiser and be selected in step 1 of Algorithm 2 several times.
- $T_n^{(2)}\left(\overline{\mathcal{X}}_g\right)$: We appeal to previous first fidelity queries. If we are querying at the second fidelity at a certain region, it can only be because the first fidelity confidence band is small. This implies

that there must be several first fidelity queries in that region which in turn implies that we can learn about $f^{(1)}$ with high confidence. As $f^{(1)}$ alone would tell us that any point in $\mathcal{X}_g$ is suboptimal for $f^{(2)}$, the maximiser of the UCB is unlikely to lie in this region frequently. Hence, we will not query outside $\mathcal{X}_g$ often.

It follows from the above that the number of second fidelity queries in $\mathcal{X}_g$ scales $T_n^{(2)}(\mathcal{X}_g) \asymp n$. Finally, we invoke techniques from Srinivas et al. (2010) to control the regret using the MIG. However, unlike them, we can use the MIG of $\mathcal{X}_g$ since an overwhelming amount of evaluations at the second fidelity are in $\mathcal{X}_g$. This allows us to obtain a tighter bound on $S(\Lambda)$ of the following form.

Theorem 4 (Informal, Regret of MF-GP-UCB for $M = 2$)**.** *Let $\mathcal{X} \subset [0, r]^d$. Then there exists Λ_0 depending only on $\gamma^{(1)}$, $\lambda^{(1)}$ and the approximation $f^{(1)}$ such that, for all $\Lambda > \Lambda_0$ the following holds with high probability.*

$$S(\Lambda) \lesssim \sqrt{\frac{\beta_{\underline{n}_\Lambda} \Psi_{\underline{n}_\Lambda}(\mathcal{X}_g)}{\underline{n}_\Lambda}}$$

It is instructive to compare the above rates against that for GP-UCB in Theorem 1. By dropping the common and sub-dominant terms, the rate for GP-UCB is $\Psi^{1/2}_{\underline{n}_\Lambda}(\mathcal{X})$ whereas for MF-GP-UCB it is $\Psi^{1/2}_{\underline{n}_\Lambda}(\mathcal{X}_g)$. Therefore, whenever the approximation is very good ($\mathrm{vol}(\mathcal{X}_g) \ll \mathrm{vol}(\mathcal{X})$) the rates for MF-GP-UCB are very appealing. When the approximation worsens and $\mathcal{X}_g^\star, \mathcal{X}_g$ become larger, the bound decays gracefully. In the worst case, MF-GP-UCB is never worse than GP-UCB up to constant terms for $\Lambda \geq \Lambda_0$. The Λ_0 term is required since at the initial stages, MF-GP-UCB will be exploring $f^{(1)}$ before proceeding to $f^{(2)}$, at which stage its regret will still be $+\infty$. The costs $\lambda^{(1)}, \lambda^{(2)}$ get factored into the result via the $\Lambda > \Lambda_0$ condition. If $\lambda^{(1)}$ is large, for fixed $\gamma^{(1)}$, a larger amount of capital is spent at the first fidelity, so Λ_0 will be large. We will make the dependence on Λ_0 on the lower fidelities explicit in the formal theorem statements.

Now let us analyse the effect of the parameter $\gamma^{(1)}$ on the result. At first sight, large $\gamma^{(1)}$ seems to increase the size of $\mathcal{X}_g$ which would suggest that we should keep it as small as possible. However, smaller $\gamma^{(1)}$ also increases Λ_0; intuitively, if $\gamma^{(1)}$ is too small, then one will wait for a long time in step 2 of Algorithm 2 for $\beta_t^{1/2}\sigma_{t-1}^{(1)}$ to decrease without proceeding to $f^{(2)}$. As one might expect, an "optimal" choice of $\gamma^{(1)}$ depends on how large a Λ_0 we are willing to tolerate; i.e. how long we are willing to wait investigating the cheap approximation. Moreover, if the approximation is extremely cheap, it makes sense to use very small $\gamma^{(1)}$ and learn as much as possible about $f^{(2)}$ from $f^{(1)}$. However, it also depends on other problem dependent quantities such as $\mathcal{X}_g^\star$. In Section 5.2 we describe a choice for $\gamma^{(1)}$ based on $\lambda^{(1)}, \lambda^{(2)}$ and $\zeta^{(1)}$ that aims to balance the cost spent at each fidelity. In our experiments however, we found that more aggressive choices for these threshold values $\gamma^{(m)}$ perform better in practice. We describe one such technique in Section 6.

For general M, we will define a hierarchy of good sets, the complement of which will be eliminated when we proceed from one fidelity to the next. At the highest fidelity, we will be querying mostly inside a small subset of $\mathcal{X}$ informed by the approximations $f^{(1)}, \ldots, f^{(M-1)}$. We will formalise these intuitions in the next two subsections.

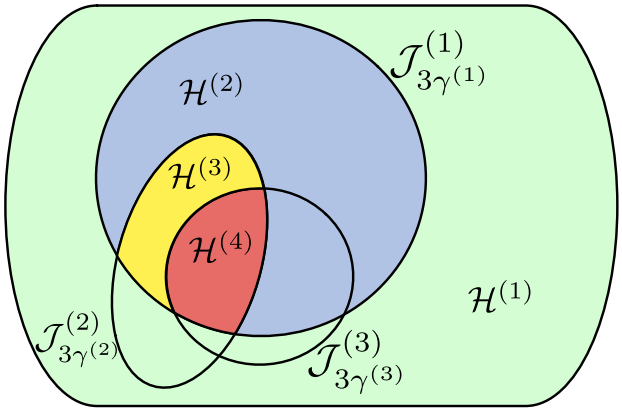


Figure 5: Illustration of the partition $\mathcal{H}^{(m)}$'s for a $M = 4$ fidelity problem. The sets $\mathcal{J}_0^{(m)}$ are indicated next to their boundaries. The sets $\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \mathcal{H}^{(3)}, \mathcal{H}^{(4)}$ are shown in green, blue, yellow and red respectively. Most of the capital invested at points in $\mathcal{H}^{(m)}$ will be due to queries to the m^{th} fidelity function $f^{(m)}$.

5.2 Discrete $\mathcal{X}$

We first analyse the case when $\mathcal{X}$ is a discrete subset of $[0, r]^d$. Denote $\Delta^{(m)}(x) = f_\star - f^{(m)}(x) - \zeta^{(m)}$ and $\mathcal{J}_\eta^{(m)} = \{x \in \mathcal{X}; \Delta^{(m)}(x) \leq \eta\}$. Note that $\Delta^{(m)} > 0$ for all m by our assumptions. Central to our analysis will be the partitioning $(\mathcal{H}^{(m)})_{m=1}^M$ of $\mathcal{X}$. First define $\mathcal{H}^{(1)} = \overline{\mathcal{J}}_{3\gamma}^{(1)} = \{x : f^{(1)}(x) < f_\star - \zeta^{(1)} - 3\gamma^{(1)}\}$ to be the arms whose $f^{(1)}$ value is at least $\zeta^{(1)} + 3\gamma^{(1)}$ below the optimum $f_\star$. Then recursively define,

$$\mathcal{H}^{(m)} = \overline{\mathcal{J}}_{3\gamma}^{(m)} \cap \left(\bigcap_{\ell=1}^{m-1} \mathcal{J}_{3\gamma}^{(\ell)} \right) \quad \text{for } 2 \leq m \leq M-1, \qquad \mathcal{H}^{(M)} = \bigcap_{\ell=1}^{M-1} \mathcal{J}_{3\gamma}^{(\ell)}. \tag{6}$$

In addition to the above, we will also find it useful to define the sets "above" $\mathcal{H}^{(m)}$ as $\widehat{\mathcal{H}}^{(m)} = \bigcup_{\ell=m+1}^{M} \mathcal{H}^{(\ell)}$ and the sets "below" $\mathcal{H}^{(m)}$ as $\breve{\mathcal{H}}^{(m)} = \bigcup_{\ell=1}^{m-1} \mathcal{H}^{(\ell)}$. Our analysis reveals that most of the capital invested at points in $\mathcal{H}^{(m)}$ will be due to queries to the m^{th} fidelity function $f^{(m)}$. $\breve{\mathcal{H}}^{(m)}$ is the set of points that can be excluded from queries at fidelities m and beyond due to information from lower fidelities. $\widehat{\mathcal{H}}^{(m)}$ are points that will be queried at fidelities higher than m several times. In the 2 fidelity setting described in Section 5, $\mathcal{X}_g = \mathcal{H}^{(2)}$ and $\overline{\mathcal{X}_g} = \mathcal{H}^{(1)} = \breve{\mathcal{H}}^{(2)}$. We have illustrated these sets in Figure 5.

Recall that $\underline{n}_\Lambda = \lfloor \Lambda / \lambda^{(M)} \rfloor$ is the number of queries by a single-fidelity method; it is a lower bound on N, the number of queries by a multi-fidelity method. Similarly, $\overline{n}_\Lambda = \lfloor \Lambda / \lambda^{(1)} \rfloor$ will be an upper bound on N. We will now define two quantities Λ_1, Λ_2 where $\Lambda_1 < \Lambda_2$. We will show improved simple regret over GP-UCB when the capital Λ is larger than these quantities, with the $\Lambda > \Lambda_2$ regime being better by an additive $\log(\lambda^{(M)}/\lambda^{(1)})$ factor over the $\Lambda > \Lambda_1$ case. Formally, we define Λ_1 to be the smallest Λ satisfying the following condition,

$$\sum_{m=2}^{M} \lambda^{(m)} |\mathcal{H}^{(m-1)}| + \sum_{m=1}^{M-1} \lambda^{(m)} |\mathcal{H}^{(m)} \cup \widehat{\mathcal{H}}^{(m)}| \left\lceil \frac{\eta^2}{{\gamma^{(m)}}^2} \beta_{\overline{n}_\Lambda} \right\rceil \leq \frac{\Lambda}{2}, \tag{7}$$

and Λ_2 to be the smallest Λ satisfying the following condition,

$$\lambda^{(M)}|\mathcal{X}| \;+\; \lambda^{(M)} \sum_{m=1}^{M-1} |\mathcal{H}^{(m)} \cup \widehat{\mathcal{H}}^{(m)}| \left\lceil \frac{\eta^2}{{\gamma^{(m)}}^2} \beta_{\overline{n}_\Lambda} \right\rceil \;\leq\; \frac{\Lambda}{2}. \tag{8}$$

We can find such Λ_1, Λ_2, since for fixed $\gamma^{(m)}$'s, in both cases, the right side is linear in Λ and the left is logarithmic since $\beta_n \asymp \mathcal{O}(\log(n))$ and $\overline{n}_\Lambda \asymp \Lambda$. Since $\{\mathcal{H}^{(m)}\}_{m=1}^M$ form a partition of $\mathcal{X}$ and $\lambda^{(1)} < \cdots < \lambda^{(M)}$, we see that $\Lambda_1 < \Lambda_2$. Recall that at the initial stages, MF-GP-UCB has infinite simple regret since the evaluations are at lower fidelities. $\Lambda > \Lambda_1$ indicates the phase where $\Theta(\underline{n}_\Lambda)$ evaluations have been made inside $\mathcal{H}^{(M)}$, but the total number of evaluations N could be much larger. When $\Lambda > \Lambda_2$, we have reached a phase where N is also in $\Theta(\underline{n}_\Lambda)$.

Moreover, note that when the approximations are good, i.e. the sets $\mathcal{H}^{(m)}$ are small, both Λ_1 and Λ_2 are small. Λ_1 is also small when the approximations are cheap, i.e. $\lambda^{(m)}$'s are small. Therefore, the cheaper and better the approximations, we have to wait less time (for fixed $\gamma^{(m)}$) before MF-GP-UCB starts querying at the M^{th} fidelity and achieves good regret.

We now state our main theorem for discrete $\mathcal{X}$. To simplify the analysis, we will introduce an additional condition in the fidelity selection criterion in step 2 of Algorithm 2. We will always evaluate $f^{(m)}$ at x_t only if x_t has been evaluated at all lower fidelities, $1, \ldots, m-1$; precisely, that $m_t = \min_m\{\, m \,|\beta_t^{1/2}\sigma_{t-1}^{(m)}(x_t) \geq \gamma^{(m)} \;\text{ or }\; m = M \;\text{ or }\; T_n^{(m)}(x_t) = 0\}$. Both this condition, and the dependence of Λ_2 on $|\mathcal{X}|$ in (8) are an artefact of our analysis. They arise only because we do not account for the correlations between the arms in our discrete analysis; doing so requires us to make assumptions about the locations of the arms in $[0,r]^d$. We will not need this condition or have Λ_2 depend on $|\mathcal{X}|$ for the continuous case.

Theorem 5. *Let $\mathcal{X}$ be a discrete subset of $[0,r]^d$. Let $f^{(m)} \sim \mathcal{GP}(\mathbf{0}, \kappa)$ for all m. Assume that $f^{(m)}$'s satisfy assumptions* **A1**, **A2** *and κ satisfies Assumption 1. Pick $\delta \in (0,1)$ and run* MF-GP-UCB *(Algorithm 2) with $\beta_t = 2\log\left(M|\mathcal{X}|\pi^2 t^2/(3\xi_{\mathbf{A2}}\delta)\right)$. Then, we have the following bounds on $S(\Lambda)$ with $\mathbb{P}$-probability greater than $1-\delta$.*

$$\text{for all } \Lambda > \Lambda_1, \quad S(\Lambda) \;\leq\; \sqrt{\frac{2C_1\beta_{\overline{n}_\Lambda}\Psi_{\underline{n}_\Lambda}(\mathcal{H}^{(M)})}{\underline{n}_\Lambda}}$$

$$\text{for all } \Lambda > \Lambda_2, \quad S(\Lambda) \;\leq\; \sqrt{\frac{2C_1\beta_{2\underline{n}_\Lambda}\Psi_{\underline{n}_\Lambda}(\mathcal{H}^{(M)})}{\underline{n}_\Lambda}}$$

Here $C_1 = 8/\log(1+\eta^2)$ is a constant, $\underline{n}_\Lambda = \lfloor \Lambda/\lambda^{(M)} \rfloor$, $\overline{n}_\Lambda = \lfloor \Lambda/\lambda^{(1)} \rfloor$, and $\xi_{\mathbf{A2}}$ is from (5).

The difference between the two results is the $\beta_{\overline{n}_\Lambda}$ dependence in the former setting and $\beta_{\underline{n}_\Lambda}$ in the latter; the latter bound is better by an additive $\log(\lambda^{(M)}/\lambda^{(1)})$ term, but we have to wait for longer. Dropping constant and polylog terms and comparing to the result in Theorem 1 reveals that we outperform GP-UCB by a factor of $\sqrt{\Psi_{\underline{n}_\Lambda}(\mathcal{H}^{(M)})/\Psi_{\underline{n}_\Lambda}(\mathcal{X})} \asymp \sqrt{\text{vol}(\mathcal{H}^{(M)})/\text{vol}(\mathcal{X})}$ asymptotically. The set $\mathcal{H}^{(M)}$ from (6) is determined by the $\zeta^{(1)}, \ldots . \zeta^{(M-1)}$ values, the approximations $f^{(1)}, \ldots, f^{(M-1)}$ and the parameters $\gamma^{(1)}, \ldots, \gamma^{(M-1)}$. The better the approximations, the smaller the set $\mathcal{H}^{(M)}$ and there is more advantage over single fidelity strategies. In Figure 6, we have shown the ratio $\text{vol}(\mathcal{H}^{(2)})/\text{vol}(\mathcal{X})$ for a two fidelity problem as $\zeta^{(1)}$ decreases—the figure corroborates our

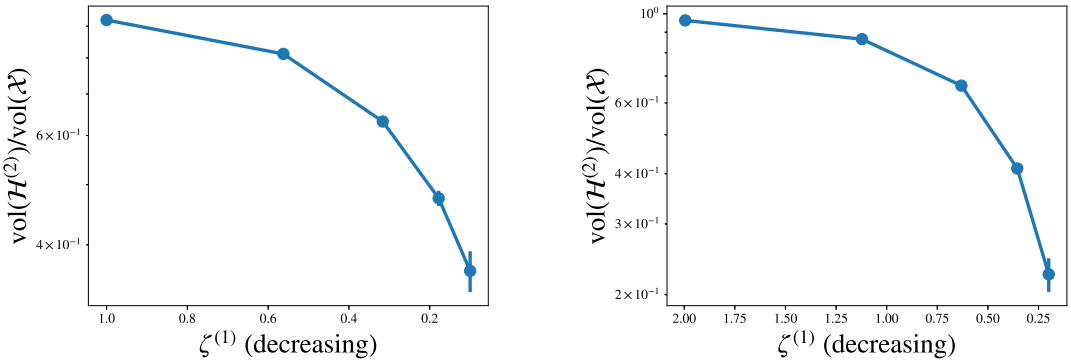


Figure 6: Empirically computed values for the ratio $\mathrm{vol}(\mathcal{H}^{(2)})/\mathrm{vol}(\mathcal{X})$ for a one dimensional (left) and two dimensional (right) 2-fidelity problem. For this, the samples $f^{(1)}, f^{(2)}$ were generated using the generative mechanism of Section 2.1, under the stipulated value for $\zeta^{(1)}$. In both cases, we used an SE kernel with bandwidth 1 and scale parameter 1. The y-axis is the mean value for the ratio over several samples and the x-axis is $\zeta^{(1)}$. In both cases, we used $\gamma^{(1)} = \zeta^{(1)}/3$, and approximated the continuous domain with a uniform grid of size 10^4. The figure indicates that as the approximation improves, i.e. $\zeta^{(1)}$ decreases, the ratio decreases and consequently, we get better bounds.

claim that the rates improve as the $\zeta^{(m)}$ values decrease. As the approximations worsen, the advantage to multi-fidelity optimisation diminishes as expected, but we are never worse than GP-UCB up to constant factors.

A few remarks are in order. First, note that the dependence on $\underline{n}_\Lambda$ (or equivalently Λ) is the same for both GP-UCB and MF-GP-UCB. In fact, one should not expect multi-fidelity optimisation to yield "rate" improvements since such $\sqrt{1/n}$ dependencies are typical in the bandit literature (Bubeck, Munos, Stoltz, & Szepesvári, 2011; Shang, Kaufmann, & Valko, 2017). The multi-fidelity framework allows us to find a good region, i.e. $\mathcal{H}^{(M)}$, where the optimum exists, and as such, we should expect the improvements to be in terms of the size of this set, relative to $\mathcal{X}$. Second, even when the kernels for each GP are different, the MIG dependence in Theorem 5 will be that of the highest fidelity GP $f^{(M)}$. The dependence of the other kernels will factor in via the $\xi_{\mathbf{A2}}$ bound; precisely, the more $\kappa^{(m)}$ is different from $\kappa^{(M)}$, the corresponding $\mathrm{Q}(\zeta^{(m)}/2)$ term will be smaller, leading to a smaller $\xi_{\mathbf{A2}}$ value. Finally, the bound is given in terms of $\mathcal{H}^{(M)}$ which, as illustrated by Figure 6, gives us insight into the types of gains we can expect from multi-fidelity optimisation. However, $\mathcal{H}^{(M)}$ is a random quantity and obtaining high probability bounds on its volume could shed more light on the gains of our multi-fidelity optimisation framework; this is an interesting avenue for future work.

Choice of $\gamma^{(m)}$. It should be noted that an "optimal" choice of $\gamma^{(m)}$ depends on the available budget, i.e. how long we are willing to wait before achieving non-trivial regret. If we are willing to wait long, we can afford to choose small $\gamma^{(m)}$ and consequently have better guarantees on the regret. This optimal choice also depends on several unknown problem dependent factors – such as the sizes of the sets $\mathcal{H}^{(m)}$. In Kandasamy, Dasarathy, Poczos, and Schneider (2016), the choice $\gamma^{(m)} = \zeta^{(m)}\sqrt{\lambda^{(m)}/\lambda^{(m+1)}}$ was used which ensures that for an arm $x \in \mathcal{H}^{(m)}$, the cost spent at lower fidelities $1, \ldots, m-1$ is not more than the cost spent at fidelity m. Beyond this intuitive

property, this choice further achieves a lower bound on the K-armed multi-fidelity problem. The same choice for $\gamma^{(m)}$ here ensures that the cost spent at the lower fidelities is not more than an upper bound on the cost spent at fidelity m – we have elaborated more in Remark 2 after our proofs. We have empirically demonstrated the effect of different choice of $\gamma^{(m)}$ values via an experiment in Figure 8(b). Building on these ideas, an explicit prescription for the choice of $\gamma^{(m)}$ is bound to be a fruitful avenue of research, and we leave this to future work. In the meanwhile, in Section 6, we describe a heuristic for adaptively choosing $\gamma^{(m)}$ adaptively which worked well in our experiments.

5.3 Continuous and Compact $\mathcal{X}$

We define the sets $\mathcal{H}^{(m)}, \widehat{\mathcal{H}}^{(m)}$ for $m = 1, \ldots, M$ as in the discrete case. Let $\{\nu_n\}_{n\geq 0}$ be any sublinear sequence such that $\nu_n \to \infty$. Let

$$\mathcal{H}_n^{(m)} = \left\{ x \in \mathcal{X} : \ \mathrm{B}_2\big(x, r\sqrt{d}/\nu_n^{\frac{1}{2d}}\big) \cap \mathcal{H}^{(m)} \neq \varnothing \quad \wedge \quad x \notin \widehat{\mathcal{H}}^{(m)} \right\}$$

to be a ν_n-dependent L_2 dilation of $\mathcal{H}_n^{(m)}$ by $r\sqrt{d}/\nu_n^{\frac{1}{2d}}$. Here, $\mathrm{B}_2(x, \epsilon)$ is an L_2 ball of radius ϵ centred at x. Notice that as $n \to \infty$, $\mathcal{H}_n^{(m)} \to \mathcal{H}^{(m)}$. Similar to the discrete case, we define Λ_1 to be the smallest Λ satisfying the following the condition,

$$\lambda^{(M)}\nu_{\overline{n}_\Lambda} \ + \ C_\kappa \eta^2 \beta_{\overline{n}_\Lambda}^{p+1} \sum_{m=1}^{M-1} \lambda^{(m)} \frac{\mathrm{vol}(\mathcal{H}_{\underline{n}_\Lambda}^{(m)} \cup \widehat{\mathcal{H}}^{(m)})}{{\gamma^{(m)}}^{2p}} \ \leq \ \frac{\Lambda}{2}, \tag{9}$$

and Λ_2 to be the smallest Λ satisfying the following condition,

$$\lambda^{(M)}\nu_{\overline{n}_\Lambda} \ + \ C_\kappa \eta^2 \beta_{\overline{n}_\Lambda}^{p+1} \lambda^{(M)} \sum_{m=1}^{M-1} \frac{\mathrm{vol}(\mathcal{H}_{\underline{n}_\Lambda}^{(m)} \cup \widehat{\mathcal{H}}^{(m)})}{{\gamma^{(m)}}^{2p}} \ \leq \ \frac{\Lambda}{2}. \tag{10}$$

Here $p = 1/2$ for the SE kernel and $p = 1$ for the Matérn kernel. C_κ is a kernel dependent constant elucidated in our proofs; for the SE kernel, $C_\kappa = 2^{2+d/2}(d\kappa_0/h^2)^{d/2}$ where κ_0, h are parameters of the kernel. Via a reasoning similar to the discrete case we see that $\Lambda_1 < \Lambda_2$. Our main theorem is as follows.

Theorem 6. *Let $\mathcal{X} \subset [0, r]^d$ be compact and convex. Let $f^{(m)} \sim \mathcal{GP}(\mathbf{0}, \kappa) \ \forall\, m$, and satisfy assumptions* **A1**, **A2**. *Let κ satisfy Assumption 1 with some constants a, b. Pick $\delta \in (0, 1)$ and run* MF-GP-UCB *(Algorithm 2) with*

$$\beta_t = 2\log\left(\frac{M\pi^2 t^2}{2\xi_{\mathbf{A2}}\delta}\right) + 4d\log(t) + \max\left\{0\,,\ 2d\log\left(brd\sqrt{\log\left(\frac{6Mad}{\xi_{\mathbf{A2}}\delta}\right)}\right)\right\}.$$

Then, we have the following bounds on $S(\Lambda)$ with $\mathbb{P}$-probability greater than $1 - \delta$.

$$\text{for all } \Lambda > \Lambda_1, \quad S(\Lambda) \ \leq \ \sqrt{\frac{2C_1\beta_{\overline{n}_\Lambda}\Psi_{\underline{n}_\Lambda}(\mathcal{H}_{\underline{n}_\Lambda}^{(M)})}{\underline{n}_\Lambda}} + \frac{\pi^2}{3\underline{n}_\Lambda}$$

$$\text{for all } \Lambda > \Lambda_2, \quad S(\Lambda) \ \leq \ \sqrt{\frac{2C_1\beta_{2\underline{n}_\Lambda}\Psi_{\underline{n}_\Lambda}(\mathcal{H}_{\underline{n}_\Lambda}^{(M)})}{\underline{n}_\Lambda}} + \frac{\pi^2}{3\underline{n}_\Lambda}$$

Here $C_1 = 8/\log(1 + \eta^2)$ is a constant, $\underline{n}_\Lambda = \lfloor \Lambda/\lambda^{(M)} \rfloor$, and $\overline{n}_\Lambda = \lfloor \Lambda/\lambda^{(1)} \rfloor$.

Note that the sets $\mathcal{H}^{(M)}_{\underline{n}_\Lambda}$ depend on the sublinear increasing sequence $\{\nu_n\}_{n\geq 0}$ – the theorem is valid for any such choice of ν_n. The comparison of the above bound against GP-UCB is similar to the discrete case. The main difference is that we have an additional dilation of $\mathcal{H}^{(M)}$ to $\mathcal{H}^{(M)}_{\underline{n}_\Lambda}$ which occurs due to a covering argument in our analysis. Recall that $\mathcal{H}^{(m)}_{\underline{n}_\Lambda} \to \mathcal{H}^{(m)}$ as $\Lambda \to \infty$. The bound is determined by the MIG of the set $\mathcal{H}^{(M)}_{\underline{n}_\Lambda}$, which is small when the approximations are good.

6. Some Implementation Details of MF-GP-UCB and other Baselines

Our implementation uses some standard techniques in the Bayesian optimisation literature given below. In addition, we describe the heuristics used to set the $\gamma^{(m)}, \zeta^{(m)}$ parameters of our method.

Initialisation: Following recommendations in Brochu, Cora, and de Freitas (2010), all GP methods were initialised with uniform random queries using an initialisation capital Λ_0. For single fidelity methods, we used it at the M^{th} fidelity, whereas for multi-fidelity methods we used $\Lambda_0/2$ at the first fidelity and $\Lambda_0/2$ at the second fidelity.

Kernel: In all our experiments, we used the SE kernel. We initialise the kernel by maximising the GP marginal likelihood (Rasmussen & Williams, 2006) on the initial sample and then update the kernel every 25 iterations using marginal likelihood.

Choice of β_t: β_t, as specified in Theorems 1, 6 has unknown constants and tends to be too conservative in practice. Following Kandasamy, Schenider, and Póczos (2015) we use $\beta_t = 0.2d\log(2t)$ which captures the dominant dependencies on d and t.

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

Choice of $\zeta^{(m)}$'s: Algorithm 2 assumes that the $\zeta^{(m)}$'s are given with the problem description, which is hardly the case in practice. In our implementation, instead of having to deal with $M-1$, $\zeta^{(m)}$ values we will assume $\|f^{(m)} - f^{(m-1)}\|_\infty \leq \zeta$. This satisfies assumption **A2** with $(\zeta^{(1)}, \zeta^{(2)}, \ldots, \zeta^{(M-1)}) = ((M-1)\zeta, (M-2)\zeta, \ldots, \zeta)$. This allows us to work with only one value of ζ. We initialise ζ to a small value, 1% of the range of initial queries. Whenever we query at any fidelity $m > 1$ we also check the posterior mean of the $(m-1)^{\text{th}}$ fidelity. If $|f^{(m)}(x_t) - \mu^{(m-1)}_{t-1}(x_t)| > \zeta$, we query again at x_t, but at the $(m-1)^{\text{th}}$ fidelity. If $|f^{(m)}(x_t) - f^{(m-1)}(x_t)| > \zeta$, we update ζ to twice the violation.

Choice of $\gamma^{(m)}$'s: The role of the $\gamma^{(m)}$ values at each fidelity is to ensure that we do not spend too much effort at the lower fidelities, where if $\gamma^{(m)}$ is too small, MF-GP-UCB spends a large number of queries at fidelity m to reduce the variance below $\gamma^{(m)}$. This might cause MF-GP-UCB to spend an unnecessarily large number of evaluations at fidelity m. Hence, we start with small values for all $\gamma^{(m)}$. However, if the algorithm does not query above the m^{th} fidelity for more than $\lambda^{(m+1)}/\lambda^{(m)}$ iterations, we double $\gamma^{(m)}$. All $\gamma^{(m)}$ values were initialised to 1% of the range of initial queries.

Whilst the first four choices are standard in the BO literature (Brochu et al., 2010; Snoek et al., 2012), our methods for selecting the $\zeta^{(m)}$ and $\gamma^{(m)}$ parameters are heuristic in nature. We obtained robust implementations of MF-GP-UCB with little effort in tweaking these choices. In fact, we found our implementation was able to recover even from fairly bad approximations at the lower fidelities (see experiment in Figure 9). We believe that other reasonable heuristics can also be used in place of our choices here, and a systematic investigation into protocols for the same will be a fruitful avenue for future research.

7. Experiments

We present experiments for compact and continuous $\mathcal{X}$ since it is the more practically relevant setting. We compare MF-GP-UCB to the following baselines. **Single fidelity methods:** GP-UCB; EI: the expected improvement criterion for BO (Jones et al., 1998); DiRect: the dividing rectangles method (Jones et al., 1993). **Multi-fidelity methods:** MF-NAIVE: a naive baseline where we use GP-UCB to query at the *first* fidelity a large number of times and then query at the last fidelity at the points queried at $f^{(1)}$ in decreasing order of $f^{(1)}$-value; MF-SKO: the multi-fidelity sequential kriging method from Huang et al. (2006). Previous works on multi-fidelity methods (including MF-SKO) had not made their code available and were not straightforward to implement. We discuss this more in Appendix A.1 along with some other single and multi-fidelity baselines we tried but excluded in the comparison to avoid clutter in the figures. We also detail some design choices and hyper-parameters for the baselines in Appendix A.1.

7.1 Synthetic Examples

We begin with a series of synthetic experiments, designed to demonstrate the applicability and limitations of MF-GP-UCB. We use the Currin exponential ($d = 2$), Park ($d = 4$) and Borehole ($d = 8$) functions in $M = 2$ fidelity experiments and the Hartmann functions in $d = 3$ and 6 with $M = 3$ and 4 fidelities respectively. The first three functions are taken from previous multi-fidelity literature (Xiong, Qian, & Wu, 2013) while we tweaked the Hartmann functions to obtain the lower fidelities for the latter two cases. In Appendix A we give the formulae for these functions and the approximations used for the lower fidelities. We show the simple regret $S(\Lambda)$ against capital Λ in Figure 7. The number of fidelities and the costs used for each fidelity are also given in Figure 7. MF-GP-UCB outperforms other baselines on all problems.

The last panel of Figure 7 shows a histogram of the number of queries at each fidelity after 184 queries of MF-GP-UCB, for different ranges of $f^{(3)}(x)$ for the Hartmann-3D function. Many of the queries at the low $f^{(3)}$ values are at fidelity 1, but as we progress they decrease and the second fidelity queries increase. The third fidelity dominates very close to the optimum but is used sparingly elsewhere. This corroborates the prediction in our analysis that MF-GP-UCB uses low fidelities to explore and successively higher fidelities at promising regions to zero in on $x_\star$. (Also see Figure 4.)

A common occurrence with MF-NAIVE was that once we started querying at fidelity M, the regret barely decreased. The diagnosis in all cases was the same: it was stuck around the maximum of $f^{(1)}$ which is suboptimal for $f^{(M)}$. This suggests that while we have cheap approximations, the problem is by no means trivial. As explained previously, it is also important to *explore* at higher fidelities to achieve good regret. The efficacy of MF-GP-UCB when compared to single fidelity methods is that it confines this exploration to a small set containing the optimum. In our experiments we found that MF-SKO did not consistently beat other single fidelity methods. Despite our best efforts to reproduce MF-SKO, we found it to be quite brittle. In fact, we also tried another multi-fidelity method and found that it did not perform as desired (See Appendix A.1 for details).

Effect of the cost of the approximations: We now test the effect the cost of the approximation on performance. Figure 8(a) shows the results when MF-GP-UCB was run on the 2-fidelity Borehole experiment for different costs for the approximation $f^{(1)}$. We fixed $\lambda^{(2)} = 1$ and varied $\lambda^{(1)}$ between 0.01 to 0.5. As $\lambda^{(1)}$ increases, the performance worsens as expected. At $\lambda^{(1)} = 0.5$ it is indistinguishable from GP-UCB as the overhead of managing 2 fidelities becomes significant when compared to the improvements of using the approximation.

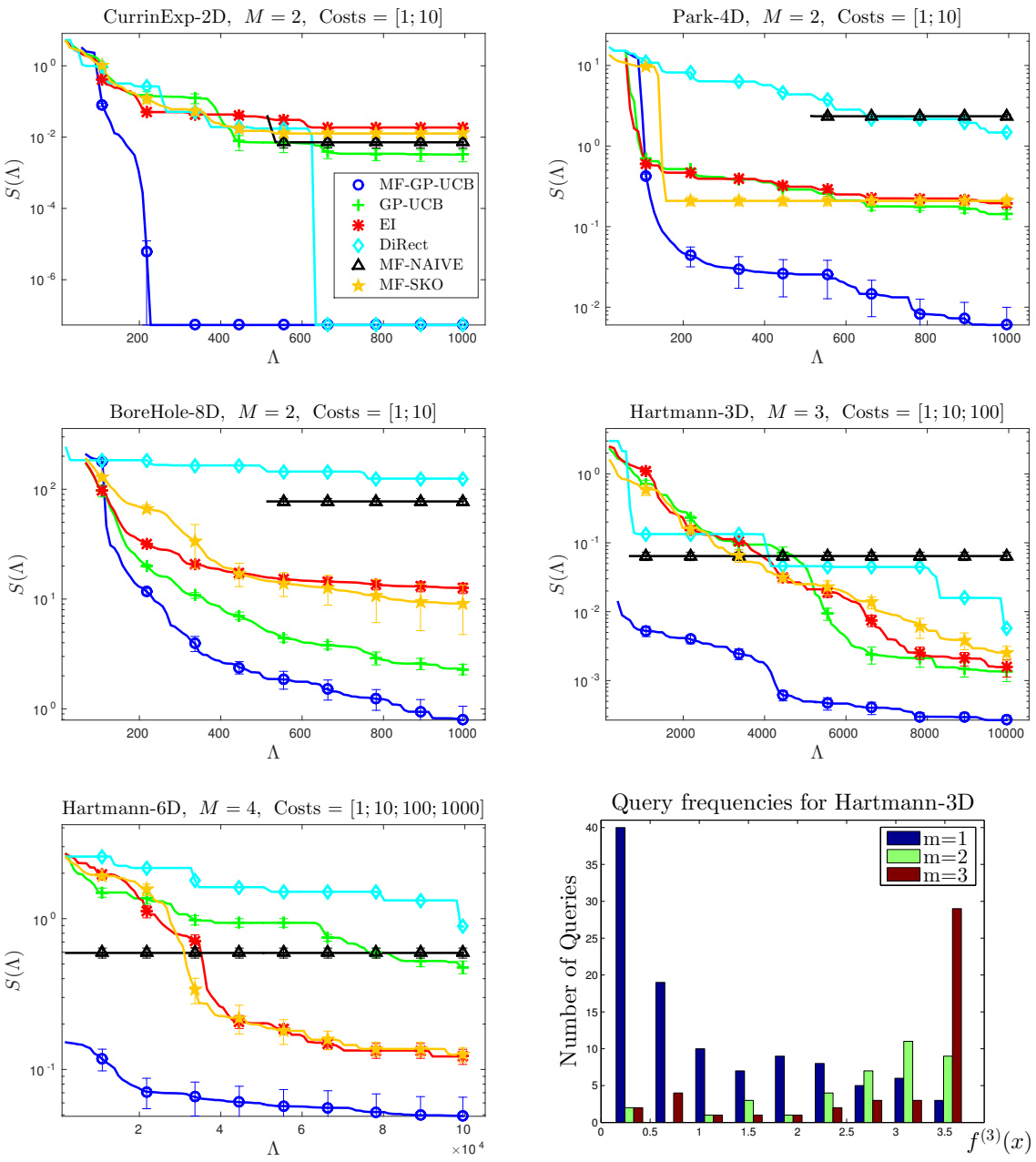
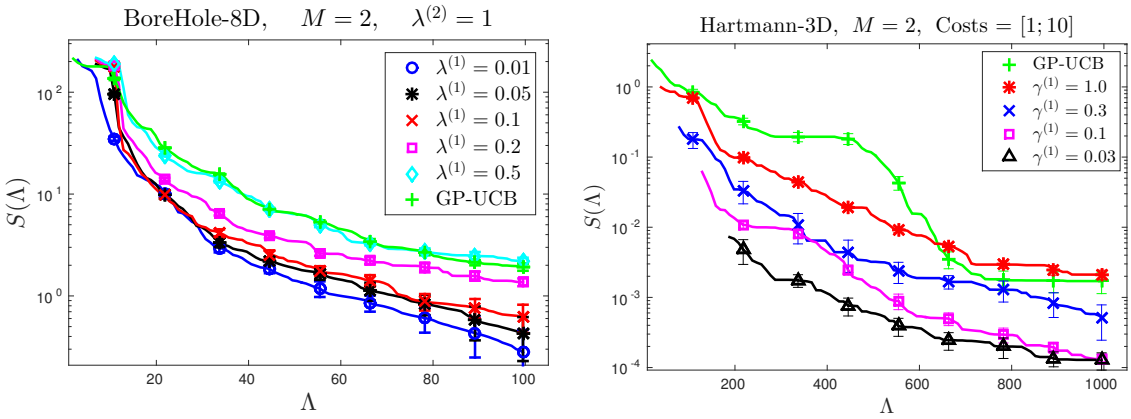


Figure 7: The simple regret $S(\Lambda)$ (3) against the spent capital Λ on the synthetic functions. The title states the function, its dimensionality, the number of fidelities and the costs we used for each fidelity in the experiment; for example, in the fourth panel, we used $M = 3$ fidelities, with costs $\lambda^{(1)} = 1, \lambda^{(2)} = 10, \lambda^{(3)} = 100$ on the 3 dimensional Hartmann function. All curves barring DiRect (which is a deterministic), were produced by averaging over 20 experiments. The error bars indicate one standard error. All figures follow the legend in the first figure for the Currin exponential function. The last panel shows the number of queries at different function values at each fidelity for the Hartmann-3D example.



(a) (b)

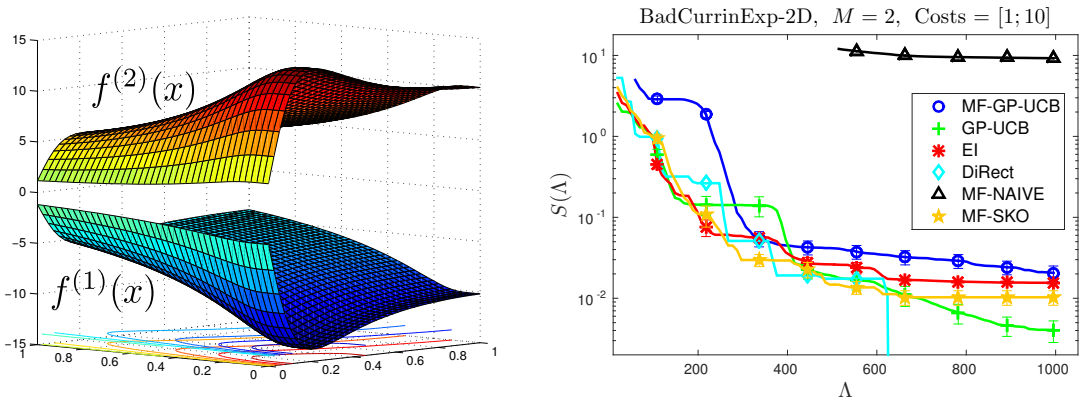
Figure 8: (a): The performance of our implementation of MF-GP-UCB for different values of $\lambda^{(1)}$ in the 2 fidelity Borehole experiment. Our implementation uses the techniques and heuristics described in Section 6. In all experiments we used $\lambda^{(2)} = 1$. We have also shown the curve for GP-UCB for reference. (b): The performance of MF-GP-UCB for different choices of *fixed* threshold values $\gamma^{(1)}$. The curves were averaged over 20 independent runs, and, in this figure, they start when at least 10 of the 20 runs have queried at least once at the top (second) fidelity. This experiment was run on the $3-$dimensional Hartmann function in the two fidelity set up where $\zeta^{(1)} \approx 0.112$.

Effect of threshold values on MF-GP-UCB: We now demonstrate the effect of different choices for $\gamma^{(1)}$ on MF-GP-UCB as described in Algorithm 2. We use the 3 dimensional Hartmann function in a 2 fidelity set up where $\zeta^{(1)} \approx 0.112$, $\lambda^{(1)} = 1$ and $\lambda^{(2)} = 10$. The implementation follows the description in Section 6, except that the true $\zeta^{(1)}$ value is made known to MF-GP-UCB and the threshold value $\gamma^{(1)}$ is kept fixed at values $0.03, 0.1, 0.3, 1.0$. The result is shown in Figure 8(b). We see that as $\gamma^{(1)}$ decreases the curves start later in the figure indicating that MF-GP-UCB spends more time at the approximation $f^{(1)}$ before proceeding to $f^{(2)}$; however, the simple regret is also generally better for smaller $\gamma^{(1)}$. Therefore, if we have a large computational budget and are willing to wait longer, we can choose small $\gamma^{(m)}$ values and achieve better simple regret.

Bad Approximations: It is natural to ask how MF-GP-UCB performs with bad approximations at lower fidelities. We found that our implementation with the heuristics suggested in Section 6 to be quite robust. We demonstrate this using the Currin exponential function, but using the negative of $f^{(2)}$ as the first fidelity approximation, i.e. $f^{(1)}(x) = -f^{(2)}(x)$. Figure 9 illustrates $f^{(1)}$, $f^{(2)}$ and gives the simple regret $S(\Lambda)$. Understandably, it loses to the single fidelity methods since the first fidelity queries are wasted and it spends some time at the second fidelity recovering from the bad approximation. However, it eventually is able to achieve low regret.

7.2 Model Selection and Astrophysics Experiments

We now present results on three hyper-parameter tuning tasks and a maximum likelihood inference task in Astrophysics. We compare methods on computation time since that is the "cost" in all



(a) (b)

Figure 9: (a): The functions used in the Bad Currin Exponential experiment where $f^{(1)} = -f^{(2)}$. (b): The simple regret for this experiment. See caption under Figure 7 for more details.

experiments. We include the processing time for each method in the comparison (i.e. the cost of determining the next query). The results are given in Figure 10, where, as we see MF-GP-UCB outperforms other baselines on all tasks. The experimental set up for each optimisation problem is described below.

Classification using SVMs (SVM): We trained a Support vector classifier on the magic gamma dataset using the sequential minimal optimisation algorithm to an accuracy of 10^{-12}. The goal is to tune the kernel bandwidth and the soft margin coefficient in the ranges $(10^{-3}, 10^{1})$ and $(10^{-1}, 10^{5})$ respectively on a dataset of size 2000. We set this up as a $M = 2$ fidelity experiment with the entire training set at the second fidelity and 500 points at the first. Each query to $f^{(m)}$ required 5-fold cross validation on the respective training sets.

Regression using additive kernels (SALSA): We used the SALSA method for additive kernel ridge regression (Kandasamy & Yu, 2016) on the 4-dimensional coal power plant dataset. We tuned the 6 hyper-parameters –the regularisation penalty, the kernel scale and the kernel bandwidth for each dimension– each in the range $(10^{-3}, 10^{4})$ using 5-fold cross validation. This experiment used $M = 3$ and $2000, 4000, 8000$ points at each fidelity respectively.

Viola & Jones face detection (V&J): The Viola & Jones cascade face classifier (Viola & Jones, 2001), which uses a cascade of weak classifiers, is a popular method for face detection. To classify an image, we pass it through each classifier. If at any point the classifier score falls below a threshold, the image is classified as negative. If it passes through the cascade, then it is classified as positive. One of the more popular implementations comes with OpenCV and uses a cascade of 22 weak classifiers. The threshold values in the OpenCV implementation are pre-set based on some heuristics and there is no reason to think they are optimal for a given face detection problem. The goal is to tune these 22 thresholds by optimising them over a training set. We modified the OpenCV implementation to take in the thresholds as parameters. As our domain $\mathcal{X}$ we chose a neighbourhood around the configuration used in OpenCV. We set this up as an $M = 2$ fidelity experiment where

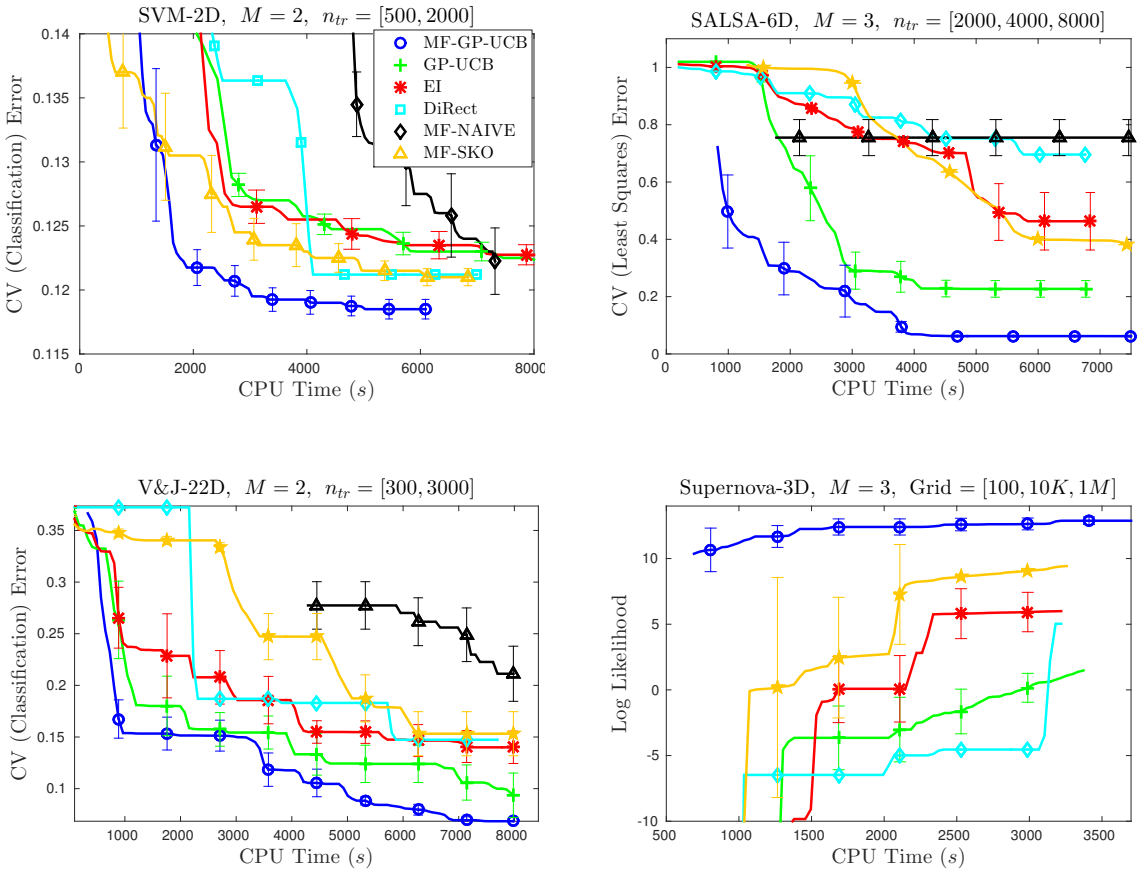


Figure 10: Results on the real experiments. The first three figures are hyper-parameter tuning tasks while the last is an astrophysical maximum likelihood problem. The title states the experiment, dimensionality (number of hyper-parameters or cosmological parameters) and the number of fidelities. For the three hyper-parameter tuning tasks we plot the best cross validation error (lower is better) and for the astrophysics task we plot the highest log likelihood (higher is better). For the hyper-parameter tuning tasks we obtained the lower fidelities by using smaller training sets, indicated by n_{tr} in the figures and for the astrophysical problem we used a coarser grid for numerical integration, indicated by "Grid". MF-NAIVE is not visible in the last experiment because it performed very poorly. All curves were produced by averaging over 10 experiments. The error bars indicate one standard error. The lengths of the curves are different in time as we ran each method for a pre-specified number of iterations and they concluded at different times.

the second fidelity used 3000 images from the Viola and Jones face database and the first used just 300. Interestingly, on an independent test set, the configurations found by MF-GP-UCB consistently achieved over 90% accuracy while the OpenCV configuration achieved only 87.4% accuracy.

Type Ia Supernovae (Supernova): We use Type Ia supernovae data from Davis et al (2007) for maximum likelihood inference on 3 cosmological parameters, the Hubble constant $H_0 \in (60, 80)$, the

dark matter fraction $\Omega_M \in (0,1)$ and the dark energy fraction $\Omega_\Lambda \in (0,1)$. Unlike typical parametric maximum likelihood problems we see in machine learning, the likelihood is only available as a black-box. It is computed using the Robertson–Walker metric Davis et al (2007), which requires a (one dimensional) numerical integration for each sample in the dataset. We set this up as a $M = 3$ fidelity task. At the third fidelity, the integration was performed using the trapezoidal rule on a grid of size 10^6. For the first and second fidelities, we used grids of size $10^2, 10^4$ respectively. The goal is to maximise the likelihood at the third fidelity.

8. Proofs

In this section we present the proofs of our main theorems. While it is self contained, the reader will benefit from first reading the more intuitive discussion in Section 5. The goal in this section is to bound the simple regret $S(\Lambda)$ given in (3). Recall that N is the random number of plays within capital Λ. While $N \leq \lfloor \Lambda/\lambda^{(1)} \rfloor$ is a trivial upper bound for N, this will be too loose for our purposes. In fact, we will show that after a sufficiently large number of queries at any fidelity, the number of queries at fidelities smaller than M will be sublinear in N. Hence $N \in \mathcal{O}(\underline{n}_\Lambda)$ where $\underline{n}_\Lambda = \lfloor \Lambda/\lambda^{(M)} \rfloor$ is the number of plays by any algorithm that operates only at the highest fidelity.

We introduce some notation to keep track of the evaluations at each fidelity in MF-GP-UCB. After n steps, we will have queried multiple times at any of the M fidelities. $T_n^{(m)}(x)$ denotes the number of queries at $x \in \mathcal{X}$ at fidelity m after n steps. $T_n^{(m)}(A)$ denotes the same for a subset $A \subset \mathcal{X}$. $\mathcal{D}_n^{(m)} = \{(x_t, y_t)\}_{t:m_t=m}$ is the set of query-value pairs at the m^{th} fidelity until time n.

Roadmap: To bound $S(\Lambda)$ in both the discrete and continuous settings, we will begin by studying the algorithm after n evaluations at any fidelity and analyse the following quantity,

$$\tilde{R}_n = \sum_{\substack{t:m_t=M \\ x_t \in \mathcal{Z}}} \left(f_\star - f^{(M)}(x_t) \right) \tag{11}$$

Readers familiar with the bandit literature will see that this is similar to the notion of *cumulative regret*, except we only consider queries at the M^{th} fidelity and inside a set $\mathcal{Z} \subset \mathcal{X}$. $\mathcal{Z}$ contains the optimum and generally has high value for the payoff function $f^{(M)}(x)$; it will be determined by the approximations provided via the lower fidelity evaluations. We will show that most of the M^{th} fidelity evaluations will be inside $\mathcal{Z}$ in the multi-fidelity setting, and hence, the regret for MF-GP-UCB will scale with $\Psi_n(\mathcal{Z})$ instead of $\Psi_n(\mathcal{X})$ as is the case for GP-UCB. Finally, to convert this bound in terms of n to one that depends on Λ, we show that both the total number of evaluations N and the number of highest fidelity evaluations $T_N^{(M)}(\mathcal{X})$ are on the order of $\underline{n}_\Lambda$ when Λ is sufficiently large. For this, we bound the number of plays at the lower fidelities (see Lemma 3). Then $S(\Lambda)$ can be bounded by,

$$S(\Lambda) \;\leq\; \frac{1}{T_N^{(M)}(\mathcal{X})} \tilde{R}_N \;\lesssim\; \frac{1}{\underline{n}_\Lambda} \tilde{R}_{\underline{n}_\Lambda}. \tag{12}$$

Before we proceed, we will prove a series of results that will be necessary in our proofs of Theorems 5 and 6. We first prove Lemma 2.

Proof of Lemma 2. Let $\mathbf{A2}' = \left\{\|f^{(M)}\|_\infty \leq \zeta^{(M-1)}/2 \ \cap \ \bigcap_{m=1}^{M-1} \|f^{(m)}\|_\infty \leq \zeta^{(m)}/2\right\}$. It is straightforward to see that $\mathbf{A2}' \subset \mathbf{A2}$ since for any $m \leq M-1$,

$$\|f^{(M)} - f^{(m)}\|_\infty \leq \|f^{(M)}\|_\infty + \|f^{(m)}\|_\infty \leq \zeta^{(M-1)}/2 + \zeta^{(m)}/2 \leq \zeta^{(m)}.$$

Hence, $\mathbb{P}_{\mathcal{GP}}(\mathbf{A2}) \geq \mathbb{P}_{\mathcal{GP}}(\mathbf{A2}')$. We can now bound,

$$\mathbb{P}_{\mathcal{GP}}(\mathbf{A2}') \ = \ \mathbb{P}_{\mathcal{GP}}\big(\|f^{(M)}\|_\infty \leq \zeta^{(M-1)}/2\big) \cdot \prod_{m=1}^{M-1} \mathbb{P}_{\mathcal{GP}}\big(\|f^{(m)}\|_\infty \leq \zeta^{(m)}/2\big) \ \geq \ \xi_{\mathbf{A2}}.$$

Here the equality in the first step comes from the observation that the $f^{(m)}$'s are independent under the $\mathbb{P}_{\mathcal{GP}}$ probability. The last inequality comes from Assumption 2. ■

Remark 1. It is worth noting that the above bound is a fairly conservative lower bound on $\xi_{\mathbf{A2}}$ since $\mathbf{A2}'$ essentially requires that all samples $f^{(m)}$ be small so as to make the differences $f^{(M)} - f^{(m)}$ small. We can obtain a more refined bound on $\xi_{\mathbf{A2}}$ by noting that $f^{(M)} - f^{(m)} \sim \mathcal{GP}(\mathbf{0}, 2\kappa)$ and following proofs for bounding the supremum of a GP (e.g. Theorem 5.4 in Adler, 1990, or Theorem 4 in Ghosal and Roy, 2006). This leads to smaller values for β_t in Theorems 5 and 6 and consequently better constants in our bounds. However, this analysis will require accounting for correlations when analysing multiple GPs which is beyond the scope and tangential to the goals of this paper. Moreover, from a practical perspective it would not result in anything actionable since many quantities in the expression for β_t are already unknown in practice, even for GP-UCB. It is also worth noting that the dependence of $\xi_{\mathbf{A2}}$ on our regret bounds is mild since it appears as a $\sqrt{\log(1/\xi_{\mathbf{A2}})}$ term.

Next, Lemma 7 provides a way to bound the probability of an event under our prior (**A1** and **A2**) using the probability of the event when the functions are sampled from a GP (**A1** only).

Lemma 7. *Let E be a $\mathbb{P}_{\mathcal{GP}}$-measurable event. Then, $\mathbb{P}(E) \leq \xi_{\mathbf{A2}}^{-1} \mathbb{P}_{\mathcal{GP}}(E)$.*

Proof This follows via a straightforward application of Bayes' rule, shown below. The last step uses Lemma 2 and that the intersection of two sets is at most as large as either set.

$$\mathbb{P}(E) = \mathbb{P}_{\mathcal{GP}}(E|\mathbf{A2}) = \frac{\mathbb{P}_{\mathcal{GP}}(E \cap \mathbf{A2})}{\mathbb{P}_{\mathcal{GP}}(\mathbf{A2})} \leq \frac{1}{\xi_{\mathbf{A2}}} \mathbb{P}_{\mathcal{GP}}(E).$$

■

For our analysis, we will also need to control the sum of conditional standard deviations for queries in a subset $A \subset \mathcal{X}$. We provide the lemma below, whose proof is based of a similar result in Srinivas et al. (2010).

Lemma 8. *Let $f \sim \mathcal{GP}(0, \kappa)$, $f : \mathcal{X} \to \mathbb{R}$ and each time we query at any $x \in \mathcal{X}$ we observe $y = f(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \eta^2)$. Let $A \subset \mathcal{X}$. Assume that we have queried f at n points, $(x_t)_{t=1}^n$ of which s points are in A. Let σ_{t-1} denote the posterior variance at time t, i.e. after $t-1$ queries. Then, $\sum_{x_t \in A} \sigma_{t-1}^2(x_t) \leq \frac{2}{\log(1+\eta^{-2})} \Psi_s(A)$.*

Proof Let $A_s = \{z_1, z_2, \ldots, z_s\}$ be the queries inside A in the order they were queried. Now, assuming that we have only queried inside A at A_s, denote by $\tilde{\sigma}_{t-1}(\cdot)$, the posterior standard deviation after $t-1$ such queries. Then,

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

Queries outside 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 A. The third step uses the inequality $u^2/v^2 \leq \log(1+u^2)/\log(1+v^2)$ with $u = \tilde{\sigma}_{t-1}(z_t)/\eta$ and $v = 1/\eta$ and the last step uses Lemma 15 in Appendix B.1. The result follows from the fact that $\Psi_s(A)$ maximises the mutual information among all subsets of size s. ■

8.1 Discrete $\mathcal{X}$

Proof of Theorem 5. Without loss of generality, we can assume that MF-GP-UCB is run indefinitely. Let N denote the (random) number of queries within Λ, i.e. the quantity satisfying $N = \max\{n \geq 1; \sum_{t=1}^{n} \lambda^{(m_t)} \leq \Lambda\}$. Note that $\mathbf{supp}(N) \subset \{n \in \mathbb{N} : \underline{n}_\Lambda \leq n \leq \overline{n}_\Lambda\}$. In our analysis, we will first analyse MF-GP-UCB after n steps and control the regret and the number of lower fidelity evaluations.

Bounding the regret after n evaluations: We will need the following lemma to establish that $\varphi_t(x)$ upper bounds $f^{(M)}(x)$. The proof is given in Section 8.1.1.

Lemma 9. *Pick $\delta \in (0,1)$ and choose $\beta_t \geq 2\log\left(\frac{M|\mathcal{X}|\pi^2 t^2}{3\xi_{\mathbf{A2}}\delta}\right)$. Then, with probability at least $1-\delta/2$, for all $t \geq 1$, for all $x \in \mathcal{X}$ and for all $m \in \{1,\ldots,M\}$, we have*

$$\left|f^{(m)}(x) - \mu^{(m)}_{t-1}(x)\right| \;\leq\; \beta_t^{1/2}\sigma^{(m)}_{t-1}(x).$$

First note the following bound on the instantaneous regret when $m_t = M$,

$$f_\star - f^{(M)}(x_t) \;\leq\; \varphi_t(x_\star) - (\mu^{(M)}_{t-1}(x_t) - \beta_t^{1/2}\sigma^{(M)}_{t-1}(x_t)) \tag{13}$$
$$\leq\; \varphi_t(x_t) - (\mu^{(M)}_{t-1}(x_t) - \beta_t^{1/2}\sigma^{(M)}_{t-1}(x_t)) \;\leq\; 2\beta_t^{1/2}\sigma^{(M)}_{t-1}(x_t).$$

The first step uses that $\varphi^{(m)}_t(x)$ is an upper bound for $f^{(M)}(x)$ by Lemma 9 and the assumption **A2**, and hence so is the minimum $\varphi_t(x)$. The second step uses that x_t was the maximiser of $\varphi_t(x)$ and the third step that $\varphi^{(M)}_t(x) \geq \varphi_t(x)$. To control $\tilde{R}_n$, we will use $\mathcal{Z} = \mathcal{H}^{(M)}$ in (11) and invoke Lemma 8. Applying the Cauchy Schwarz inequality yields,

$$\tilde{R}^2_n \;\leq\; T^{(m)}_n(\mathcal{H}^{(M)}) \sum_{\substack{t:m_t=M \\ x_t\in\mathcal{H}^{(M)}}} \left(f_\star - f^{(M)}(x_t)\right)^2 \;\leq\; T^{(m)}_n(\mathcal{H}^{(M)}) \sum_{\substack{t:m_t=M \\ x_t\in\mathcal{H}^{(M)}}} 4\beta_t(\sigma^{(m)}_{t-1}(x_t))^2$$
$$\leq\; C_1 T^{(m)}_n(\mathcal{H}^{(M)})\beta_n \Psi_{T^{(m)}_n(\mathcal{H}^{(M)})}(\mathcal{H}^{(M)}). \tag{14}$$

Here $C_1 = 8/\log(1+\eta^{-2})$.

Bounding the number of evaluations: Lemma 10, given below, bounds the number of evaluations at different fidelities in different regions of $\mathcal{X}$. This will allow us to bound, among other things, the total number of plays N and the number of M^{th} fidelity evaluations outside $\mathcal{Z}$. The proof of Lemma 10 is given in Section 8.1.2. Recall that $T_n^{(m)}(x)$ denotes the number of queries at point $x \in \mathcal{X}$ at fidelity m. Similarly, we will denote $T_n^{(>m)}(x)$ to denote the number of queries at point x at fidelities larger than m.

Lemma 10. *Pick $\delta \in (0,1)$ and set $\beta_t = 2\log\left(\frac{M|\mathcal{X}|\pi^2 t^2}{3\xi_{\mathbf{A2}}\delta}\right)$. Further assume $\varphi_t(x_\star) \geq f_\star$. Consider any $x \in \mathcal{H}^{(m)}\backslash\{x_\star\}$ for $m < M$. We then have the following bounds on the number of queries at any given time step n,*

$$\begin{aligned} T_n^{(\ell)}(x) &\leq \frac{\eta^2}{{\gamma^{(m)}}^2}\beta_n + 1, \quad \text{for } \ell < m, \\ \mathbb{P}\left(T_n^{(m)}(x) > \left\lceil 5\Big(\frac{\eta}{\Delta^{(m)}(x)}\Big)^2\beta_n\right\rceil\right) &\leq \frac{3\delta}{2\pi^2}\frac{1}{|\mathcal{X}|n^2}, \\ \mathbb{P}\left(T_n^{(>m)}(x) > u\right) &\leq \frac{3\delta}{2M\pi^2}\frac{1}{|\mathcal{X}|u}. \end{aligned}$$

First whenever $\varphi_t(x_\star) \geq f_\star$, by using the union bound on the second result of Lemma 10,

$$\mathbb{P}\left(\exists\, n \geq 1,\ \exists\, m \in \{1,\ldots,M\},\ \exists\, x \in \mathcal{H}^{(m)}\backslash\{x_\star\},\ T_n^{(m)}(x) > \left\lceil 5\Big(\frac{\eta}{\Delta^{(m)}(x)}\Big)^2\beta_n\right\rceil\right) \leq \frac{\delta}{4}.$$

Here we have used $\sum n^{-2} = \pi^2/6$. The last two quantifiers just enumerates over all $x \in \mathcal{X}\backslash\{x_\star\}$. Similarly, applying the union bound for $u = 1$ on the third result, we have, for any given n,

$$\mathbb{P}\left(\exists\, m \in \{1,\ldots,M\},\ \exists\, x \in \mathcal{H}^{(m)},\ T_n^{(>m)}(x) > 1\right) \leq \frac{3\delta}{2\pi^2} < \frac{\delta}{4}.$$

We will apply the above result for $n = \lfloor \Lambda/\lambda^{(1)} \rfloor$ and observe that $T_n^{(>m)}(x)$ is non-decreasing in n. Hence,

$$\mathbb{P}\left(\forall\, n \leq \Lambda/\lambda^{(1)},\ \forall\, m \in \{1,\ldots,M\},\ \forall\, x \in \mathcal{H}^{(m)},\ T_n^{(>m)}(x) \leq 1\right) > 1 - \frac{\delta}{4}. \tag{15}$$

The condition for Lemma 10 holds with probability at least $1 - \delta/2$ (by Lemma 9), and therefore the above bounds hold together with probability $> 1 - \delta$. We have tabulated these bounds in Table 1. We therefore have the following bound on the number of fidelity m ($< M$) plays $T_n^{(m)}(\mathcal{X})$,

$$\begin{aligned} T_n^{(m)}(\mathcal{X}) &\leq T_n^{(m)}(\breve{\mathcal{H}}^{(m)}) + \sum_{x \in \mathcal{H}^{(m)}} \left\lceil \frac{5\eta^2}{{\Delta^{(m)}(x)}^2}\beta_n \right\rceil + |\widehat{\mathcal{H}}^{(m)}| \left\lceil \frac{\eta^2}{{\gamma^{(m)}}^2}\beta_n \right\rceil \\ &\leq T_n^{(m)}(\breve{\mathcal{H}}^{(m)}) + |\mathcal{H}^{(m)} \cup \widehat{\mathcal{H}}^{(m)}| \left\lceil \frac{\eta^2}{{\gamma^{(m)}}^2}\beta_n \right\rceil \qquad (16) \\ &\leq |\mathcal{H}^{(m-1)}| + |\mathcal{H}^{(m)} \cup \widehat{\mathcal{H}}^{(m)}| \left\lceil \frac{\eta^2}{{\gamma^{(m)}}^2}\beta_n \right\rceil \qquad (17) \end{aligned}$$

	$\mathcal{H}^{(1)}$	$\mathcal{H}^{(2)}$	$\mathcal{H}^{(m)}$	$\mathcal{H}^{(M)}\backslash\{x_\star\}$
$T_n^{(1)}(x)$	$\frac{5\eta^2}{{\Delta^{(1)}(x)}^2}\beta_n + 1$	$\frac{\eta^2}{{\gamma^{(1)}}^2}\beta_n + 1$	$\dots$ $\frac{\eta^2}{{\gamma^{(1)}}^2}\beta_n + 1$ $\dots$	$\frac{\eta^2}{{\gamma^{(1)}}^2}\beta_n + 1$
$T_n^{(2)}(x)$		$\frac{5\eta^2}{{\Delta^{(2)}(x)}^2}\beta_n + 1$	$\dots$ $\frac{\eta^2}{{\gamma^{(2)}}^2}\beta_n + 1$ $\dots$	$\frac{\eta^2}{{\gamma^{(2)}}^2}\beta_n + 1$
$\vdots$ $T_n^{(m)}(x)$ $\vdots$	1	1	$\vdots$ $\dots$ $\frac{5\eta^2}{{\Delta^{(m)}(x)}^2}\beta_n + 1$ $\dots$ $\vdots$	$\vdots$ $\frac{\eta^2}{{\gamma^{(m)}}^2}\beta_n + 1$ $\vdots$
$T_n^{(M)}(x)$			1	$\frac{5\eta^2}{{\Delta^{(M)}(x)}^2}\beta_n + 1$

Table 1: Bounds on the number of queries for each $x \in \mathcal{H}^{(m)}$ (columns) at each fidelity (rows). The bound for $T_n^{(M)}(x)$ in $\mathcal{H}^{(M)}$ holds for all arms except the optimal arm $x_\star$ (note $\Delta^{(M)}(x_\star) = 0$).

The second step uses that $\Delta^{(m)}(x) \geq 3\gamma^{(m)}$ for $x \in \mathcal{H}^{(m)}$ and the last step uses the modification to the discrete algorithm which ensures that we will always play an arm at a lower fidelity before we play it at a higher fidelity. Hence, for an arm in $\mathcal{H}^{(m)}$, the 1 play at fidelities larger than m will be played at fidelity $m+1$.

Proof of first result: First consider the total cost $\Lambda'(n)$ expended at fidelities $1, \dots, M-1$ and at the M^{th} fidelity outside of $\mathcal{H}^{(M)}$ after n evaluations. Using (17), we have,

$$\begin{aligned}\Lambda'(n) &= \sum_{m=1}^{M-1} \lambda^{(m)} T_n^{(m)}(\mathcal{X}) + \lambda^{(M)} T_n^{(M)}(\breve{\mathcal{H}}^{(M)}) \\ &\leq \sum_{m=2}^{M} \lambda^{(m)} |\mathcal{H}^{(m-1)}| + \sum_{m=1}^{M-1} \lambda^{(m)} |\mathcal{H}^{(m)} \cup \widehat{\mathcal{H}}^{(m)}| \left\lceil \frac{\eta^2}{{\gamma^{(m)}}^2}\beta_n \right\rceil.\end{aligned}$$

Since $N \leq \overline{n}_\Lambda$, we have for all $n \in \mathbf{supp}\,(N)$, $\Lambda'(n)$ is less than the LHS of (7) and hence less than $\Lambda/2$. Therefore, the amount of cost spent at the M^{th} fidelity inside $\mathcal{H}^{(M)}$ is at least $\Lambda/2$ and since each such evaluation expends $\lambda^{(M)}$, we have $T_N^{(M)}(\mathcal{H}^{(M)}) \geq \underline{n}_\Lambda/2$. Therefore using (14) we have,

$$S(\Lambda) \leq \frac{1}{T_N^{(M)}(\mathcal{H}^{(M)})}\tilde{R}_N \leq \sqrt{\frac{C_1\beta_N \Psi_{T_N^{(M)}(\mathcal{H}^{(M)})}(\mathcal{H}^{(M)})}{T_N^{(M)}(\mathcal{H}^{(M)})}} \leq \sqrt{\frac{2C_1\beta_{\overline{n}_\Lambda}\Psi_{\underline{n}_\Lambda}(\mathcal{H}^{(M)})}{\underline{n}_\Lambda}}.$$

Here, we have used $N \leq \overline{n}_\Lambda$ and that $\underline{n}_\Lambda \geq T_N^{(M)}(\mathcal{H}^{(M)}) \geq \underline{n}_\Lambda/2$.

Proof of second result: Using (16), the total number of queries at fidelities less than M and the number of M^{th} fidelity queries outside of $\mathcal{H}^{(M)}$ can be bounded as follows,

$$\sum_{m=1}^{M-1}\sum_{x\in\mathcal{X}} T_n^{(m)}(x) + T_n^{(M)}(\breve{\mathcal{H}}^{(M)}) \leq |\mathcal{X}| + \sum_{m=1}^{M-1} |\mathcal{H}^{(m)} \cup \widehat{\mathcal{H}}^{(m)}| \left\lceil \frac{\eta^2}{{\gamma^{(m)}}^2}\beta_n \right\rceil. \tag{18}$$

The first term of the RHS above follows via (15) and the following argument. In particular, this does not use the additional condition on the discrete algorithm – we will use a similar argument in the continuous domain setting.

$$\sum_{m=1}^{M}\sum_{x\in\widetilde{\mathcal{H}}^{(m)}} T_n^{(m)}(x) = \sum_{m=1}^{M}\sum_{\ell=1}^{m-1}\sum_{x\in\mathcal{H}^{(\ell)}} T_n^{(m)}(x) \leq \sum_{m=1}^{M-1}\sum_{x\in\mathcal{H}^{(m)}} T_n^{(>m)}(x) \;\leq\; |\mathcal{X}|. \tag{19}$$

Let the LHS of (18) be A and the RHS be B when $n = N$. When $\Lambda > \Lambda_2$, by (8) and using the fact that $N \leq \overline{n}_\Lambda$, we have $B < \underline{n}_\Lambda/2 < N/2$. Since $N = A + T_N^{(M)}(\mathcal{H}^{(M)})$, we have $T_N^{(M)}(\mathcal{H}^{(M)}) > N/2 > \underline{n}_\Lambda/2$. Further, since the total expended budget after N rounds $\Lambda(N)$ satisfies $\Lambda(N) \geq T_N^{(M)}(\mathcal{H}^{(M)})\lambda^{(M)} > \lambda^{(M)}N/2$, we also have $N < 2\underline{n}_\Lambda$. Putting these results together we have for all $\Lambda > \Lambda_2$,

$$S(\Lambda) \leq \sqrt{\frac{C_1\beta_N \Psi_{T_N^{(M)}(\mathcal{H}^{(M)})}(\mathcal{H}^{(M)})}{T_N^{(M)}(\mathcal{H}^{(M)})}} \leq \sqrt{\frac{2C_1\beta_{2\underline{n}_\Lambda}\Psi_{\underline{n}_\Lambda}(\mathcal{H}^{(M)})}{\underline{n}_\Lambda}}.$$

■

Remark 2. ***Choice of*** $\gamma^{(m)}$***:*** As described in the main text, the optimal choice for $\gamma^{(m)}$ depends on the available budget and unknown problem dependent quantities. However the choice $\gamma^{(m)} = \sqrt{\lambda^{(m)}/\lambda^{(m+1)}}\zeta^{(m)}$ ensures that for any $x \in \mathcal{H}^{(m)}$, the bounds on the number of plays in Table 1 are on the same order for fidelities m and below. To see this, consider any $\ell < m$. Then,

$$\Delta^{(m)}(x) = \Delta^{(\ell)}(x)+\zeta^{(\ell)}-\zeta^{(m)}+f^{(\ell)}(x)-f^{(M)}(x)+f^{(M)}(x)-f^{(m)}(x) \leq 3\gamma^{(\ell)}+2\zeta^{(\ell)} \leq 5\zeta^{(\ell)}.$$

We therefore have,

$$\lambda^{(\ell)}\cdot\frac{\eta^2}{{\gamma^{(\ell)}}^2} = \lambda^{(\ell+1)}\frac{\eta^2}{{\zeta^{(\ell)}}^2} \leq 5\left(\lambda^{(m)}\cdot\frac{5\eta^2}{{\Delta^{(m)}(x)}^2}\right)$$

Above, by Table 1, the left most expression is an upper bound on the cost spent at fidelity ℓ and the term inside the parantheses is an upper bound on the cost spent at fidelity m. Hence, the capital spent at the lower fidelities is within a constant factor of this bound. In the K-armed setting (Kandasamy, Dasarathy, Poczos, & Schneider, 2016), we showed a $\mathcal{O}(\eta^2/{\Delta^{(m)}(x)}^2)$ lower bound on the number of plays at the m^{th} fidelity as well; such a result is not straightforward in the GP setting due to correlations between arms.

8.1.1 PROOF OF LEMMA 9

This is a straightforward argument using Gaussian concentration and the union bound. Consider any given m, t, x.

$$\begin{aligned}
&\mathbb{P}\left(|f^{(m)}(x) - \mu_{t-1}^{(m)}(x)| > \beta_t^{1/2}\sigma_{t-1}^{(m)}(x)\right)\\
&\quad= \frac{1}{\xi_{\mathbf{A2}}}\mathbb{P}_{\mathcal{GP}}\left(|f^{(m)}(x) - \mu_{t-1}^{(m)}(x)| > \beta_t^{1/2}\sigma_{t-1}^{(m)}(x)\right)\\
&\quad= \frac{1}{\xi_{\mathbf{A2}}}\mathbb{E}_{\mathcal{GP}}\left[\mathbb{E}_{\mathcal{GP}}\left[\mathbb{1}\left\{|f^{(m)}(x) - \mu_{t-1}^{(m)}(x)| > \beta_t^{1/2}\sigma_{t-1}^{(m)}(x)\right\}\;\middle|\;\mathcal{D}_{t-1}^{(m)}\right]\right]
\end{aligned}$$

$$= \frac{1}{\xi_{\mathbf{A2}}}\mathbb{E}_{\mathcal{GP}}\left[\mathbb{E}_{\mathcal{GP}}\left[\mathbb{1}\left\{|f^{(m)}(x) - \mu^{(m)}_{t-1}(x)| > \beta_t^{1/2}\sigma^{(m)}_{t-1}(x)\right\} \;\middle|\; \mathcal{D}^{(m)}_{t-1}\right]\right]$$
$$= \frac{1}{\xi_{\mathbf{A2}}}\mathbb{E}_{\mathcal{GP}}\left[\mathbb{P}_{Z\sim\mathcal{N}(0,1)}\left(|Z| > \beta_t^{1/2}\right)\right] \;\leq\; \frac{1}{\xi_{\mathbf{A2}}}\exp\left(\frac{\beta_t}{2}\right) \;=\; \frac{3\delta}{M|\mathcal{X}|\pi^2 t^2}.$$

The first step uses Lemma 7. In the second step we have conditioned w.r.t $\mathcal{D}^{(m)}_{t-1}$ which allows us to use Lemma 14. Recall that conditioning on all queries will not be a Gaussian due to the $\zeta^{(m)}$ constraints. The statement follows via a union bound over all $m \in \{1, \ldots, M\}$, $x \in \mathcal{X}$ and all t and noting that $\sum_t t^{-2} = \pi^2/6$. ■

8.1.2 PROOF OF LEMMA 10

First consider any $\ell < m$. Assume that we have already queried $\lceil \eta^2\beta_n/{\gamma^{(m)}}^2\rceil$ times at any $t \leq n$. Since the Gaussian variance after s observations is η^2/s and that queries elsewhere will only decrease the conditional variance we have, $\kappa^{(\ell)}_{t-1}(x,x) \leq \eta^2/T^{(\ell)}_{t-1}(x) < {\gamma^{(m)}}^2/\beta_n$. Therefore, $\beta_t^{1/2}\sigma^{(\ell)}_{t-1}(x) < \beta_n^{1/2}\sigma^{(\ell)}_{t-1}(x) < \gamma^{(m)}$ and by the design of our algorithm we will not play at the ℓ^{th} fidelity at time t for all t until n. This establishes the first result.

To bound $T^{(m)}_n(x)$ we first observe,

$$\begin{aligned}
\mathbb{1}\{T^{(m)}_n(x) > u\} \;\leq\; & \mathbb{1}\{\exists t : u+1 \leq t \leq n :\; \varphi_t(x) \text{ was maximum} \;\wedge \\
& \qquad \beta_t^{1/2}\sigma^{(\ell)}_{t-1}(x) < \gamma^{(m)},\, \forall\, \ell < m \;\wedge\; \beta_t^{1/2}\sigma^{(m)}_{t-1}(x) \geq \gamma^{(m)} \wedge \\
& \qquad T^{(m)}_{t-1}(x) \geq u\} \\
\leq\; & \mathbb{1}\{\exists t : u+1 \leq t \leq n :\; \varphi_t(x) > \varphi_t(x_\star) \quad\wedge\quad T^{(m)}_{t-1}(x) \geq u\} \\
\leq\; & \mathbb{1}\{\exists t : u+1 \leq t \leq n :\; \varphi^{(m)}_t(x) > f_\star \quad\wedge\quad T^{(m)}_{t-1}(x) \geq u\}. \qquad (20)
\end{aligned}$$

The first line just enumerates the conditions in our algorithm for it to have played x at time t at fidelity m. In the second step we have relaxed some of those conditions, noting in particular that if $\varphi_t(\cdot)$ was maximised at x then it must be larger than $\varphi_t(x_\star)$. The last step uses the fact that $\varphi^{(m)}_t(x) \geq \varphi_t(x)$ and the assumption on $\varphi_t(x_\star)$. Consider the event $\{\varphi^{(m)}_t(x) > f_\star \wedge T^{(m)}_{t-1}(x) \geq u\}$. We will choose $u = \lceil 5\eta^2\beta_n/{\Delta^{(m)}(x)}^2\rceil$ and bound its probability via,

$$\begin{aligned}
& \mathbb{P}\Big(\varphi^{(m)}_t(x) > f_\star \wedge T^{(m)}_{t-1}(x) \geq u\Big) \\
& \quad = \frac{1}{\xi_{\mathbf{A2}}}\mathbb{P}_{\mathcal{GP}}\Big(\mu^{(m)}_{t-1}(x) + \beta_t^{1/2}\sigma^{(m)}_{t-1}(x) + \zeta^{(m)} > f_\star \quad\wedge\quad T^{(m)}_{t-1}(x) \geq u\Big) \\
& \quad = \frac{1}{\xi_{\mathbf{A2}}}\mathbb{P}_{\mathcal{GP}}\Bigg(\mu^{(m)}_{t-1}(x) - f^{(m)}(x) > \underbrace{f_\star - f^{(m)}(x) - \zeta^{(m)}}_{\Delta^{(m)}(x)} - \beta_t^{1/2}\sigma^{(m)}_{t-1}(x) \wedge \\
& \qquad\qquad T^{(m)}_{t-1}(x) > u\Bigg) \\
& \quad \leq \frac{1}{\xi_{\mathbf{A2}}}\mathbb{P}_{\mathcal{GP}}\Big(\mu^{(m)}_{t-1}(x) - f^{(m)}(x) > (\sqrt{5}-1)\beta_n^{1/2}\sigma^{(m)}_{t-1}(x)\Big) \\
& \quad \leq \frac{1}{\xi_{\mathbf{A2}}}\mathbb{P}_{Z\sim\mathcal{N}(0,1)}\Big(Z > \frac{(\sqrt{5}-1)^2}{2}\beta_n^{1/2}\Big)
\end{aligned}$$

$$\leq \frac{1}{\xi_{\mathbf{A2}}}\frac{1}{2}\exp\left(-\frac{3}{4}\beta_n\right) = \frac{1}{\xi_{\mathbf{A2}}}\frac{1}{2}\left(\frac{3\xi_{\mathbf{A2}}\delta}{M|\mathcal{X}|\pi^2}\right)^{\frac{3}{2}} n^{-3} \leq \frac{1}{2}\frac{3\delta}{M|\mathcal{X}|\pi^2}n^{-3}$$

Above in the third step we have used, if $u \geq 5\eta^2\beta_n/\Delta^{(m)}(x)^2$, then $\Delta^{(m)}(x) \geq \sqrt{5}\beta_n^{1/2}\sigma_{t-1}^{(m)}(x)$ and that $\beta_n \geq \beta_t$. The fourth step uses Lemma 14 after conditioning on $\mathcal{D}_{t-1}^{(m)}$, the fifth step uses $(\sqrt{5}-1)^2 > 3/2$ and the last step uses $3\delta/|\mathcal{X}|\pi^2 < 1$. Using the union bound on (20), we get $\mathbb{P}(T_n^{(m)}(x) > u) \leq \sum_{t=u+1}^n \mathbb{P}(\varphi_t^{(m)}(x) > f_\star \wedge T_{t-1}^{(m)}(x) \geq u)$. Now (20) implies that $\mathbb{P}(T_n^{(m)}(x) > u) \leq \sum_{t=u+1}^n \mathbb{P}(\varphi_t^{(m)}(x) > f_\star \wedge T_{t-1}^{(m)}(x) \geq u)$. The second inequality of the lemma follows by noting that there are at most n terms in the summation.

Finally, for the third inequality we observe

$$\mathbb{P}(T_n^{(>m)}(x) > u) \leq \mathbb{P}\big(\exists t : u+1 \leq t \leq n\,;\, \varphi_t^{(m)}(x) > f_\star \quad \wedge \quad \beta_t^{1/2}\sigma_{t-1}^{(m)}(x) < \gamma^{(m)}\big). \quad (21)$$

As before, we have used that if x is to be queried at time t, then $\varphi_t(x)$ should be at least larger than $\varphi_t(x_\star)$ which is larger than $f_\star$ due to the assumption in the theorem. The second condition is necessary to ensure that the switching procedure proceeds beyond the m^{th} fidelity. It is also necessary to have $\beta_t^{1/2}\sigma_{t-1}^{(\ell)}(x) < \gamma^{(\ell)}$ for $\ell < m$, but we have relaxed them. We first bound the probability of the event $\{\varphi_t^{(m)}(x) > f_\star \quad \wedge \quad \beta_t^{1/2}\sigma_{t-1}^{(m)}(x) < \gamma^{(m)}\}$.

$$\begin{aligned}
&\mathbb{P}\big(\varphi_t^{(m)}(x) > f_\star \wedge \beta_t^{1/2}\sigma_{t-1}^{(m)}(x) < \gamma^{(m)}\big) \\
&\quad= \frac{1}{\xi_{\mathbf{A2}}}\mathbb{P}_{\mathcal{GP}}\big(\varphi_t^{(m)}(x) > f_\star \wedge \beta_t^{1/2}\sigma_{t-1}^{(m)}(x) < \gamma^{(m)}\big) \\
&\quad= \frac{1}{\xi_{\mathbf{A2}}}\mathbb{P}_{\mathcal{GP}}\big(\mu_{t-1}^{(m)}(x) - f^{(m)}(x) > \Delta^{(m)}(x) - \beta_t^{1/2}\sigma_{t-1}^{(m)}(x) \wedge \beta_t^{1/2}\sigma_{t-1}^{(m)}(x) < \gamma^{(m)}\big) \\
&\quad\leq \frac{1}{\xi_{\mathbf{A2}}}\mathbb{P}_{\mathcal{GP}}\big(\mu_{t-1}^{(m)}(x) - f^{(m)}(x) > 2\gamma^{(m)} - \beta_t^{1/2}\sigma_{t-1}^{(m)}(x) \quad \wedge \quad \beta_t^{1/2}\sigma_{t-1}^{(m)}(x) < \gamma^{(m)}\big) \\
&\quad\leq \frac{1}{\xi_{\mathbf{A2}}}\mathbb{P}_{\mathcal{GP}}\big(\mu_{t-1}^{(m)}(x) - f^{(m)}(x) > \beta_t^{1/2}\sigma_{t-1}^{(m)}(x)\big) \\
&\quad\leq \frac{1}{\xi_{\mathbf{A2}}}\mathbb{P}_{Z\sim\mathcal{N}(0,1)}\Big(Z > \beta_t^{1/2}\Big) \leq \frac{1}{\xi_{\mathbf{A2}}}\frac{1}{2}\exp\left(-\frac{1}{2}\beta_t\right) \\
&\quad= \frac{1}{\xi_{\mathbf{A2}}}\frac{1}{2}\left(\frac{3\xi_{\mathbf{A2}}\delta}{M|\mathcal{X}|\pi^2}\right)t^{-2} \leq \frac{1}{2}\frac{3\delta}{M|\mathcal{X}|\pi^2}t^{-2}
\end{aligned}$$

Here, the second step uses that for all $x \in \mathcal{H}^{(m)}$, $\Delta^{(m)}(x) > 3\gamma^{(m)} > 2\gamma^{(m)}$ and the third step uses the second condition. Using the union bound on (21) and bounding the sum by an integral gives us,

$$\begin{aligned}
\mathbb{P}(T_n^{(>m)}(x) > u) &\leq \sum_{t=u+1}^n \frac{1}{2}\frac{3\delta}{M|\mathcal{X}|\pi^2}t^{-2} \leq \frac{1}{2}\frac{3\delta}{M|\mathcal{X}|\pi^2}\int_u^\infty t^{-2}\mathrm{d}t \\
&\leq \frac{1}{2}\,\frac{3\delta}{M|\mathcal{X}|\pi^2}\,\frac{1}{u}.
\end{aligned}$$

■

8.2 Compact and Convex $\mathcal{X}$

To prove theorem 6 we will require a fairly delicate set up for the continuous setting. Given a sequence $\{\nu_n\}_{n\geq 0}$, at time n we will consider a $r\sqrt{d}/(2\nu_n^{1/2d})$-covering of the space $\mathcal{X}$ of size $\nu_n^{1/2}$.

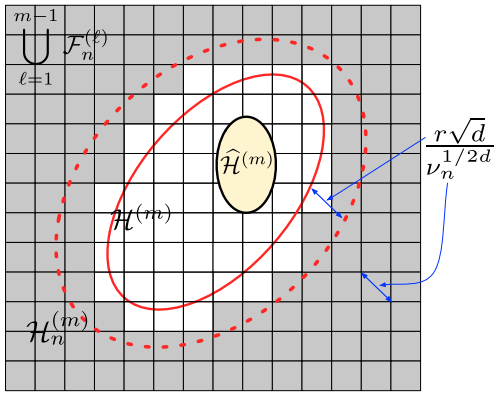


Figure 11: Illustration of the sets $\{\mathcal{F}_n^{(\ell)}\}_{\ell=1}^{m-1}$ with respect to $\mathcal{H}^{(m)}$. The grid represents a $r\sqrt{d}/n^{1/(2d)}$ covering of $\mathcal{X}$. The yellow region is $\widehat{\mathcal{H}}^{(m)}$. The area enclosed by the solid red line (excluding $\widehat{\mathcal{H}}^{(m)}$) is $\mathcal{H}^{(m)}$. $\mathcal{H}_n^{(m)}$, shown by a dashed red line, is obtained by dilating $\mathcal{H}^{(m)}$ by $r\sqrt{d}/n^{\alpha/2d}$. The grey shaded region represents $\bigcup_{\ell=1}^{m-1} \mathcal{F}_n^{(\ell)}$. By our definition, $\bigcup_{\ell=1}^{m-1} \mathcal{F}_n^{(\ell)}$ contains the cells which are entirely outside $\mathcal{H}^{(m)}$. However, the inflation $\mathcal{H}_n^{(m)}$ is such that $\widehat{\mathcal{H}}^{(m)} \cup \mathcal{H}_n^{(m)} \cup \bigcup_{\ell=1}^{m-1} \mathcal{F}_n^{(\ell)} = \mathcal{X}$. We further note that as $n \to \infty$, $\mathcal{H}_n^{(m)} \to \mathcal{H}^{(m)}$.

For instance, if $\mathcal{X} = [0, r]^d$ a sufficient discretisation would be an equally spaced grid having $\nu_n^{1/2d}$ points per side. Let $\{a_{i,n}\}_{i=1}^{n^{\frac{\alpha}{2}}}$ be the points in the covering, $F_n = \{A_{i,n}\}_{i=1}^{n^{\frac{\alpha}{2}}}$ be the "cells" in the covering, i.e. $A_{i,n}$ is the set of points which are closest to $a_{i,n}$ in $\mathcal{X}$ and the union of all sets $A_{i,n}$ in F_n is $\mathcal{X}$. Next we will define another partitioning of the space similar using this covering. First let $F_n^{(1)} = \{A_{i,n} \in F_n : \ A_{i,n} \subset \mathcal{J}^{(1)}_{\max(\tau,\rho\gamma)}\}$. Next,

$$F_n^{(m)} = \left\{ A_{i,n} \in F_n : \ A_{i,n} \subset \overline{\mathcal{J}}^{(m)}_{\max(\tau,\rho\gamma)} \ \wedge \ A_{i,n} \notin \bigcup_{\ell=1}^{m-1} F_n^{(\ell)} \right\} \ \text{ for } \ 2 \leq m \leq M-1. \tag{22}$$

Note that $F_n^{(m)} \subset F_n^{(m)}$. We define the following *disjoint* subsets $\{\mathcal{F}_n^{(m)}\}_{m=1}^{M-1}$ of $\mathcal{X}$ via $\mathcal{F}_n^{(m)} = \bigcup_{A_{i,n} \in F_n^{(m)}} A_{i,n}$. We have illustrated $\bigcup_{\ell=1}^{m-1} \mathcal{F}_n^{(\ell)}$ with respect to $\mathcal{H}^{(m)}$ and $\mathcal{H}_n^{(m)}$ in Figure 11. By observing that $\mathcal{H}_n^{(1)} = \mathcal{H}^{(1)}$ and that $\overline{\mathcal{H}_n^{(m)} \cup \widehat{\mathcal{H}}^{(m)}} \subset \bigcup_{\ell=1}^{m-1} \mathcal{F}_n^{(\ell)}$ (see Figure 11) we have the following,

$$\forall m \in \{1, \ldots, M\}, \quad T_n^{(m)}(\mathcal{X}) \leq \left(\sum_{\ell=1}^{m-1} T_n^{(m)}(\mathcal{F}_n^{(\ell)}) \right) + T_n^{(m)}(\mathcal{H}_n^{(m)}) + T_n^{(m)}(\widehat{\mathcal{H}}^{(m)}). \tag{23}$$

We are now ready to prove Theorem 6. We will denote the ε covering number of a set $A \subset \mathcal{X}$ in the $\|\cdot\|_2$ metric by $\Omega_\varepsilon(A)$.

Proof of Theorem 6. As in the discrete case, we will first control the regret and the number of lower fidelity evaluations by controlling each term in (23).

Bounding the regret after n evaluations: We will need the following lemma whose proof is given in Section 8.1.1.

Lemma 11. *For β_t as given in Theorem 6, the following holds with probability $> 1 - 5\delta/6$.*

$$\forall\, m \in \{1, \ldots, M\},\ \ \forall\, t \geq 1, \qquad \Delta^{(m)}(x_t) = f_\star - f^{(m)}(x_t) \ \leq\ 2\beta_t \sigma_{t-1}^{(m)}(x_t) \ +\ 1/t^2.$$

As in the discrete setting, we set $\mathcal{Z} = \mathcal{H}_n^{(M)}$ in (11) to bound $\tilde{R}_n$. Using $m = M$ in Lemma 11 and using calculations similar to the discrete case yields,

$$\tilde{R}_n \leq \sum_{\substack{m_t = M \\ x_t \in \mathcal{Z}}} \left(2\beta_t^{1/2} \sigma_{t-1}^{(M)}(x_t) + \frac{1}{t^2} \right) \leq \sqrt{C_1 T_n^{(m)}(\mathcal{H}_n^{(M)}) \beta_n \Psi_{T_n^{(m)}(\mathcal{H}_n^{(m)})}(\mathcal{H}_n^{(m)})} + \frac{\pi^2}{6}. \quad (24)$$

Here $C_1 = 8/\log(1 + \eta^{-2})$. We have also used the fact $\sum_{t>0} t^{-2} = \frac{\pi^2}{6}$.

Bounding the number of evaluations: The following lemma will be used to bound the number of plays in $\mathcal{H}_n^{(m)} \cup \widehat{\mathcal{H}}^{(m)}$. The proof is given in Section 8.2.2.

Lemma 12. *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 $A \subset \mathcal{X}$ such that its L_2 diameter $\mathrm{diam}(A) \leq D$. Say we have n queries $(x_t)_{t=1}^n$ of which s points are in A. Then the posterior variance of the GP, $\kappa'(x, x)$ at any $x \in A$ satisfies*

$$\kappa'(x, x) \leq \begin{cases} C_{SE} D^2 + \frac{\eta^2}{s} & \text{if } \kappa \text{ is the SE kernel,} \\ C_{Mat} D + \frac{\eta^2}{s} & \text{if } \kappa \text{ is the Matérn kernel,} \end{cases}$$

for appropriate kernel dependent constants C_{SE}, C_{Mat}.

First consider the SE kernel. At time t consider any $\varepsilon_n = \frac{\gamma^{(m)}}{\sqrt{8 C_{SE} \beta_n}}$ covering $(B_i)_{i=1}^{\varepsilon_n}$ of $\mathcal{H}_n^{(m)} \cup \widehat{\mathcal{H}}^{(m)}$. The number of queries inside any B_i of this covering at time n will be at most $\left\lceil \frac{2\eta^2}{{\gamma^{(m)}}^2} \beta_n \right\rceil$. To see this, assume we have already queried this many times inside B_i at time $t \leq n$. By Lemma 12 the maximum variance in A_i can be bounded by

$$\max_{x \in A_i} \kappa_{t-1}^{(m)}(x, x) \leq C_{SE}(2\varepsilon_n)^2 \ +\ \frac{\eta^2}{T_t^{(m)}(A_i)} \leq \frac{{\gamma^{(m)}}^2}{\beta_n}.$$

Therefore, $\beta_t^{1/2} \sigma_{t-1}^{(m)}(x) \leq \beta_n^{1/2} \sigma_{t-1}^{(m)}(x) < \gamma^{(m)}$ and we will not query inside A_i until time n. A similar result is obtained for the Matérn kernel by setting $\varepsilon_n = \frac{{\gamma^{(m)}}^2}{4 C_{Mat} \beta_n}$. Therefore we have,

$$\begin{aligned} T_n^{(m)}(\mathcal{H}_n^{(m)} \cup \widehat{\mathcal{H}}^{(m)}) \ &\leq\ \Omega_{\varepsilon_n}(\widehat{\mathcal{H}}^{(m)} \cup \widehat{\mathcal{H}}^{(m)}) \left\lceil \frac{2\eta^2}{{\gamma^{(m)}}^2} \beta_n \right\rceil \\ &\leq\ C_\kappa \eta^2 \beta_n^{p+1} \frac{\mathrm{vol}(\mathcal{H}_n^{(m)} \cup \widehat{\mathcal{H}}^{(m)})}{{\gamma^{(m)}}^{2p}}. \end{aligned} \quad (25)$$

Here $C_\kappa = 2^{2+d/2}(dC_{SE})^{\frac{d}{2}}$ and $p = 1/2$ for the SE kernel while $C_\kappa = 2^{2+d}(C_{Mat})^d d^{d/2}$ and $p = 1$ for the Matérn kernel. We have also used the fact that $\lceil k \rceil \leq 2k$ for large enough k and the following bound for a δ-packing in the Euclidean metric $\Omega_\delta(A) \leq \mathrm{vol}(A) d^{d/2} / (2^{d/2} \delta^d)$.

Next, we will bound $T_n^{(m)}(\overline{\mathcal{H}_n^{(m)} \cup \widehat{\mathcal{H}}^{(m)}})$ by controlling. $T_n^{(>m)}(\mathcal{F}_n^{(m)})$. To that end we provide the following Lemma whose proof is given in Section 8.2.3.

Lemma 13. *Consider any $A_{i,n} \in F_n^{(m)}$ where $F_n^{(m)}$ is as defined in (22). Let β_t be as given in Theorem 6. Then for all $n' \geq u \geq (3\eta)^{-2/3}$ we have,*

$$\mathbb{P}(T_{n'}^{(>m)}(A_{i,n}) > u) \;\leq\; \frac{\delta}{\pi^2} \cdot \frac{1}{u}$$

Using the above result with $n' = \overline{n}_\Lambda$ gives us the result for all $n' \leq \overline{n}_\Lambda$ since $T_{n'}^{(>m)}(A_{i,n})$ is nondecreasing with n. Setting $u = \max\{(3\eta)^{-2/3}, \nu_n^{1/2}\}$, and applying the union bound over all $m \in \{1, \ldots, M\}$ and $A_{i,n} \in F_n^{(m)}$, yields the following bound for all $n' \leq \overline{n}_\Lambda$,

$$\begin{aligned}
&\mathbb{P}\left(\exists\, m \in \{1,\ldots,M\},\, T_{n'}^{(>m)}(\mathcal{F}_n^{(m)}) > |F_n^{(m)}|\nu_n^{1/2}\right) \;\leq\; \sum_{m=1}^{M} \mathbb{P}\left(T_{n'}^{(>m)}(\mathcal{F}_n^{(m)}) > |F_n^{(m)}|\nu_n^{1/2}\right) \\
&\leq \sum_{m=1}^{M} \sum_{A_{i,n} \in F_n^{(m)}} \mathbb{P}\left(T_{n'}^{(>m)}(A_{i,n}) > \nu_n^{1/2}\right) \;\leq\; \sum_{m=1}^{M} |F_n^{(m)}| \frac{\delta}{\pi^2} \frac{1}{\nu_n^{1/2}} \\
&\leq\; |F_n| \frac{\delta}{\pi^2} \frac{1}{\nu_n^{1/2}} \;=\; \frac{\delta}{\pi^2} \leq \frac{\delta}{6}. \qquad (26)
\end{aligned}$$

Henceforth, all statements we make will make use of the bounds above and will hold with probability $> 1 - \delta$ for all $n \in \mathbf{supp}\,(N)$.

Proof of first result: Consider the cost $\Lambda'(n)$ spent at fidelities $1, \ldots, M-1$ and at the M^{th} fidelity outside of $\mathcal{H}_n^{(M)}$ after n evaluations.

$$\begin{aligned}
\Lambda'(n) \;&=\; \sum_{m=1}^{M-1} \lambda^{(m)} T_n^{(m)}(\mathcal{X}) \;+\; \lambda^{(M)} T_n^{(M)}(\mathcal{H}_n^{(M)}) \\
&= \sum_{m=1}^{M} \lambda^{(m)} \left(\sum_{\ell=1}^{m-1} T_n^{(m)}(\mathcal{F}_n^{(\ell)}) \right) + \sum_{m=1}^{M-1} \lambda^{(m)} T_n^{(m)}(\mathcal{H}_n^{(m)} \cup \widehat{\mathcal{H}}^{(m)}) \\
&\leq\; \lambda^{(M)} \nu_n + C_\kappa \eta^2 \beta_n^{p+1} \sum_{m=1}^{M-1} \lambda^{(m)} \frac{\mathrm{vol}(\mathcal{H}_n^{(m)} \cup \widehat{\mathcal{H}}^{(m)})}{{\gamma^{(m)}}^{2p}}
\end{aligned}$$

The second step uses (23). The third step uses (25), (26), and the following argument,

$$\sum_{m=1}^{M} \left(\sum_{\ell=1}^{m-1} T_n^{(m)}(\mathcal{F}_n^{(\ell)}) \right) \leq \sum_{m=1}^{M-1} T_n^{(>m)}(\mathcal{F}_n^{(\ell)}) \;\leq\; \sum_{m=1}^{M-1} |F_n^{(m)}| \nu_n^{1/2} \;\leq\; \nu_n^{1/2} |F_n| \;\leq\; \nu_n. \qquad (27)$$

The remainder of the proof follows similar to the discrete case. Noting that $\underline{n}_\Lambda \leq n \leq \overline{n}_\Lambda$ and that $\mathcal{H}_n^{(m)}$ is shrinking with n, we can conclude that $\Lambda'(n)$ is less than the LHS of (10). Therefore, $T_N^{(M)}(\mathcal{H}_n^{(M)}) \geq \underline{n}_\Lambda/2$ and hence,

$$S(\Lambda) \leq \sqrt{\frac{C_1 \beta_N \Psi_{T_N^{(M)}(\mathcal{H}_n^{(m)})}(\mathcal{H}_n^{(m)})}{T_N^{(M)}(\mathcal{H}_n^{(M)})} + \frac{\pi^2}{6 T_N^{(M)}(\mathcal{H}_n^{(M)})}} \;\leq\; \sqrt{\frac{2C_1 \beta_{\overline{n}_\Lambda} \Psi_{\underline{n}_\Lambda}(\mathcal{H}_{\underline{n}_\Lambda}^{(M)})}{\underline{n}_\Lambda} + \frac{\pi^2}{3\underline{n}_\Lambda}}.$$

Proof of second result: As in the discrete case, we bound the number of queries at fidelity $m < M$ and the M^{th} fidelity queries outside $\mathcal{H}_n^{(M)} \cup \mathcal{H}^{(M)}$ as follows.

$$\sum_{m=1}^{M-1} T_n^{(m)}(\mathcal{X}) \ + \ T_n^{(M)}(\overline{\mathcal{H}_n^{(M)}}) \ \leq \ \sum_{m=1}^{M} \left(\sum_{\ell=1}^{m-1} T_n^{(m)}(\mathcal{F}_n^{(\ell)}) \right) \ + \ \sum_{m=1}^{M-1} T_n^{(m)}(\mathcal{H}_n^{(m)} \cup \widehat{\mathcal{H}}^{(m)})$$
$$\leq \ \nu_n \ + \ C_\kappa \beta_n^{p+1} \sum_{m=1}^{M-1} \frac{\text{vol}(\mathcal{H}_n^{(M)} \cup \mathcal{H}^{(M)})}{{\gamma^{(m)}}^{2p}} \tag{28}$$

The first step uses (23) while the second step uses (25) and (27). Once again, similar to the discrete case we can argue that for all $\Lambda > \Lambda_2$, the RHS B of (28) satisfies $B < \underline{n}_\Lambda/2 < N/2$, the M^{th} fidelity plays in $\mathcal{H}_n^{(M)}$ satisfies $T_N^{(M)}(\mathcal{H}_n^{(M)}) > N/2 > \underline{n}_\Lambda/2$, and the number of plays satisfies $N \leq 2\underline{n}_\Lambda$. Combining this with (24) gives us the following for all $n \leq \underline{n}_\Lambda$,

$$S(\Lambda) \leq \sqrt{\frac{C_1 \beta_N \Psi_{T_N^{(M)}(\mathcal{H}_n^{(m)})}(\mathcal{H}_n^{(m)})}{T_N^{(M)}(\mathcal{H}_n^{(M)})} + \frac{\pi^2}{6 T_N^{(M)}(\mathcal{H}_n^{(M)})}} \ \leq \ \sqrt{\frac{2C_1 \beta_{2\underline{n}_\Lambda} \Psi_{\underline{n}_\Lambda}(\mathcal{H}_{\underline{n}_\Lambda}^{(M)})}{\underline{n}_\Lambda} + \frac{\pi^2}{3\underline{n}_\Lambda}}.$$

■

8.2.1 PROOF OF LEMMA 11

The first part of the proof mimics the arguments in Lemmas 5.6, 5.7 of Srinivas et al. (2010). By Assumption 1 for any given $m \in \{1, \ldots, M\}$ and $i \in \{1, \ldots, d\}$ we have,

$$\mathbb{P}_{\mathcal{GP}}\left(\left| \frac{\partial f^{(m)}(x)}{\partial x_i} \right| \ > \ b\sqrt{\log\left(\frac{6Mad}{\xi_{\mathbf{A2}}\delta}\right)} \right) \leq \frac{\xi_{\mathbf{A2}}\delta}{6Md}$$

Then, by the union bound and Lemma 7 we have,

$$\mathbb{P}\Bigg(\forall\, m \in \{1, \ldots, M\},\, \forall\, i \in \{1, \ldots, d\},\, \forall x \in \mathcal{X}, \ \left| \frac{\partial f^{(m)}(x)}{\partial x_i} \right| \ < \ b\sqrt{\log\left(\frac{6Mad}{\xi_{\mathbf{A2}}\delta}\right)} \Bigg)$$
$$\geq \ 1 - \frac{\delta}{6}.$$

Now we construct a discretisation F_t of $\mathcal{X}$ of size $(\nu_t)^d$ such that we have for all $x \in \mathcal{X}$, $\|x - [x]_t\|_1 \leq rd/\nu_t$. Here $[x]_t$ is the closest point to x in the discretisation. (Note that this is different from the discretisation appearing in Theorem 6 even though we have used the same notation). By choosing $\nu_t = t^2 brd\sqrt{\log(6Mad/(\xi_{\mathbf{A2}}\delta))}$ and using the above we have

$$\forall\, x \in \mathcal{X}, \quad |f^{(m)}(x) - f^{(m)}([x]_t)| \ \leq \ b \log(6Mad/\delta) \|x - [x]_t\|_1 \ \leq \ 1/t^2 \tag{29}$$

for all $f^{(m)}$'s with probability $> 1 - \delta/6$.

Noting that $\beta_t \geq 2\log(M|F_t|\pi^2 t^2/2\delta)$ for the given choice of ν_t we have the following with probability $> 1 - \delta/3$.

$$\forall\, t \geq 1, \ \forall\, m \in \{1, \ldots, M\}, \ \forall\, a \in F_t, \quad |f^{(m)}(a) - \mu_{t-1}^{(m)}(a)| \ \leq \ \beta_t^{1/2} \sigma_{t-1}^{(m)}(a). \tag{30}$$

The proof mimics that of Lemma 9 using the same conditioning argument. However, instead of a fixed set over all t, we change the set at which we have confidence based on the discretisation. Similarly we can show that with probability $> 1 - \delta/3$ we also have confidence on the decisions x_t at all time steps. Precisely,

$$\forall\, t \geq 1,\ \forall\, m \in \{1, \ldots, M\}, \quad |f^{(m)}(x_t) - \mu^{(m)}_{t-1}(x_t)| \ \leq\ \beta_t^{1/2} \sigma^{(m)}_{t-1}(x_t). \tag{31}$$

Using (29),(30) and (31) the following statements hold with probability $> 1 - 5\delta/6$. First we can upper bound $f_\star$ by,

$$f_\star \ \leq\ f^{(m)}(x_\star) + \zeta^{(m)} \ \leq\ f^{(m)}([x_\star]_t) + \zeta^{(m)} + \frac{1}{t^2} \ \leq\ \varphi^{(m)}_t([x_\star]_t) + \frac{1}{t^2}. \tag{32}$$

Since the above holds for all m, we have $f_\star \leq \varphi_t([x_\star]_t) + 1/t^2$. Now, using similar calculations as (13) we bound $\Delta^{(m)}(x_t)$.

$$\begin{aligned}
\Delta^{(m)}(x_t) =&\ f_\star - f^{(m)}(x_t) - \zeta^{(m)} \\
\leq&\ \varphi_t([x_\star]_t) + \frac{1}{t^2} - f^{(m)}(x_t) - \zeta^{(m)} \ \leq\ \varphi_t(x_t) - f^{(m)}(x_t) - \zeta^{(m)} + \frac{1}{t^2} \\
\leq&\ \varphi^{(m)}_t(x_t) - \mu^{(m)}_{t-1}(x_t) + \beta_t^{1/2}\sigma^{(m)}_{t-1}(x_t) - \zeta^{(m)} + \frac{1}{t^2} \ \leq\ 2\beta_t^{1/2}\sigma^{(m)}_{t-1}(x_t) + \frac{1}{t^2}.
\end{aligned}$$

■

8.2.2 PROOF OF LEMMA 12

Since the posterior variance only decreases with more observations, we can upper bound $\kappa'(x, x)$ for any $x \in A$ by considering its posterior variance with only the s observations in A. Further the maximum variance within A occurs if we pick 2 points x_1, x_2 that are distance D apart and have all observations at x_1; then x_2 has the highest posterior variance. Therefore, we will bound $\kappa'(x, x)$ for any $x \in A$ with $\kappa(x_2, x_2)$ in the above scenario. Let $\kappa_0 = \kappa(x, x)$ and $\kappa(x, x') = \kappa_0\phi(\|x - x'\|_2)$, where $\phi(\cdot) \leq 1$ depends on the kernel. Denote the gram matrix in the scenario described above by $\Delta = \kappa_0 \mathbf{1}\mathbf{1}^\top + \eta^2 I$. Then using the Sherman-Morrison formula on the posterior variance (2),

$$\begin{aligned}
\kappa'(x, x) \ \leq\ \kappa'(x_2, x_2) \ =&\ \kappa(x_2, x_2) - [\kappa(x_1, x_2)\mathbf{1}]^\top \Delta^{-1} [\kappa(x_1, x_2)\mathbf{1}] \\
=&\ \kappa_0 - \kappa_0^2\phi^2(D)\mathbf{1}^\top \left[\kappa_0 \mathbf{1}\mathbf{1}^\top + \eta^2 I\right]^{-1} \mathbf{1} \\
=&\ \kappa_0 - \kappa_0\phi^2(D)\mathbf{1}^\top \left[\frac{\kappa_0}{\eta^2} I - \frac{\left(\frac{\kappa_0}{\eta^2}\right)^2 \mathbf{1}\mathbf{1}^\top}{1 + \frac{\kappa_0}{\eta^2} s}\right] \mathbf{1} \\
=&\ \kappa_0 - \kappa_0\phi^2(D) \left(\frac{\kappa_0}{\eta^2} s - \frac{\left(\frac{\kappa_0}{\eta^2}\right)^2 s^2}{1 + \frac{\kappa_0}{\eta^2} s}\right) \\
=&\ \kappa_0 - \kappa_0\phi^2(D) \frac{s}{\frac{\eta^2}{\kappa_0} + s} \ =\ \frac{1}{1 + \frac{\eta^2}{\kappa_0 s}} \left(\kappa_0 - \kappa_0\phi^2(D) + \frac{\eta^2}{s}\right) \\
\leq&\ \kappa_0(1 - \phi^2(D)) + \frac{\eta^2}{s}.
\end{aligned}$$

For the SE kernel $\phi^2(D) = \exp\left(\frac{-D^2}{2h^2}\right)^2 = \exp\left(\frac{-D^2}{h^2}\right) \leq 1 - \frac{D^2}{h^2}$. Plugging this into the bound above retrieves the first result with $C_{SE} = \kappa_0/h^2$. For the Matérn kernel we use a Lipschtiz constant L_{Mat} of ϕ. Then $1 - \phi^2(D) = (1 - \phi(D))(1 + \phi(D)) \leq 2(\phi(0) - \phi(D)) \leq 2L_{Mat}D$. We get the second result with $C_{Mat} = 2\kappa_0 L_{Mat}$. Since the SE kernel decays fast, we get a stronger result on its posterior variance which translates to a better bound in our theorems. ■

8.2.3 PROOF OF LEMMA 13

First, we will invoke the same discretisation used in the proof of Lemma 11 via which we have $\varphi_t([x_\star]_t) \geq f_\star - 1/t^2$ (32). (Therefore, Lemma 13 holds only with probability $> 1 - \delta/6$, but this event has already been accounted for in Lemma 11.) Let $b_{i,n,t} = \operatorname{argmax}_{x \in A_{i,n}} \varphi_t(x)$ be the maximiser of the upper confidence bound in $A_{i,n}$ at time t. Note that the discretisation is fixed ahead of time and $b_{i,n,t}$ is deterministic given the data $\{(x_t, m_t, y_t)\}_{i=1}^{t-1}$ at time t. Now using the relaxation $x_t \in A_{i,n} \implies \varphi_t(b_{i,n,t}) > \varphi_t([x_\star]_t) \implies \varphi_t^{(m)}(b_{i,n,t}) > f_\star - 1/t^2$ and proceeding,

$$
\begin{aligned}
\mathbb{P}(T_{n'}^{(>m)}(A_{i,n}) > u) \;&\leq\; \frac{1}{\xi_{\mathbf{A2}}} \mathbb{P}_{\mathcal{GP}}\big(\exists t : u+1 \leq t \leq n,\;\; \varphi_t^{(m)}(b_{i,n,t}) > f_\star - 1/t^2 \;\;\wedge \\
&\qquad\qquad \beta_t^{1/2}\sigma_{t-1}^{(m)}(b_{i,n,t}) < \gamma^{(m)}\big) \\
&\leq\; \frac{1}{\xi_{\mathbf{A2}}} \sum_{t=u+1}^{n'} \mathbb{P}_{\mathcal{GP}}\big(\mu_{t-1}^{(m)}(b_{i,n,t}) - f^{(m)}(b_{i,n,t}) > \Delta^{(m)}(b_{i,n,t}) - \beta_t^{1/2}\sigma_{t-1}^{(m)}(b_{i,n,t}) - 1/t^2 \;\;\wedge \\
&\qquad\qquad \beta_t^{1/2}\sigma_{t-1}^{(m)}(b_{i,n,t}) < \gamma^{(m)}\big) \\
&\leq\; \frac{1}{\xi_{\mathbf{A2}}} \sum_{t=u+1}^{n'} \mathbb{P}_{\mathcal{GP}}\big(\mu_{t-1}^{(m)}(b_{i,n,t}) - f^{(m)}(b_{i,n,t}) > 2\beta_t^{1/2}\sigma_{t-1}^{(m)}(b_{i,n,t}) - 1/t^2\big) \\
&\leq\; \frac{1}{\xi_{\mathbf{A2}}} \sum_{t=u+1}^{n'} \mathbb{P}_{Z\sim\mathcal{N}(0,1)}\left(Z > \beta_t^{1/2}\right) \;\leq\; \sum_{t=u+1}^{n'} \frac{1}{\xi_{\mathbf{A2}}}\frac{1}{2}\exp\left(\frac{-\beta_t}{2}\right) \\
&\leq\; \frac{1}{\xi_{\mathbf{A2}}}\frac{1}{2}\left(\frac{2\xi_{\mathbf{A2}}\delta}{M\pi^2}\right) \sum_{t=u+1}^{n'} t^{-2} \;\leq\; \frac{\delta}{M\pi^2}\frac{1}{u}
\end{aligned}
\tag{33}
$$

In the second step we have rearranged the terms and used the definition of $\Delta^{(m)}(x)$. In the third step, as $A_{i,n} \subset \overline{\mathcal{J}}^{(m)}_{\max(\tau,\rho\gamma)}$, we have $\Delta^{(m)}(b_{i,n,t}) > 3\gamma^{(m)} > 3\beta_t^{1/2}\sigma_{t-1}^{(m)}(b_{i,n,t})$. The last step bounds the sum by an integral. For the fourth step, we have used, $t > u \geq 1/(3\eta)^{2/3}$, $\beta_t > 2\log(M\pi^2 t^2/2\delta) > (3/2)^2$, and $\sigma_{t-1}^{(m)}(b_{i,n,t}) > \eta/\sqrt{t}$ to conclude,

$$
t > \frac{1}{(3\eta)^{2/3}} \implies \frac{3t^{3/2}}{2} > \frac{1}{2\eta} \implies t^{3/2}\beta_t^{1/2} > \frac{1}{2\eta} \implies 2\beta_t^{1/2}\sigma_{t-1}^{(m)} > \frac{1}{t^2}.
$$

■

9. Conclusion

We introduced and studied the multi-fidelity bandit problem under Gaussian Process assumptions. Our theorems demonstrate that MF-GP-UCB explores the space using the cheap lower fidelities, and uses the higher fidelity queries on successively smaller regions, hence performing better than

single fidelity strategies. Via experiments on synthetic functions, three hyper-parameter tuning tasks, and an astrophysical maximum likelihood estimation problem, we demonstrate the efficacy of our method and more generally, the utility of the multi-fidelity framework. Our Matlab implementation and experiments can be downloaded from github.com/kirthevasank/mf-gp-ucb.

Going forward we wish to study multi-fidelity optimisation under different model assumptions, and extend the algorithm when we have to deal with approximations from structured fidelity spaces.

Acknowledgments

We wish to thank Bharath Sriperumbudur for the helpful email discussions. This research is partly funded by DOE grant DESC0011114, NSF grant IIS1563887, and the Darpa D3M program. KK was supported by a Facebook fellowship and a Siebel scholarship. This work was done when KK, GD, and JO were at Carnegie Mellon University.

Appendix A. Addendum to Experiments

A.1 Some Implementation Details of other Baselines

For MF-NAIVE we limited the number of first fidelity evalutions to $\max\left(\frac{1}{2}\frac{\Lambda}{\lambda^{(1)}}, 500\right)$ where Λ was the total budget used in the experiment. The 500 limit was set to avoid unnecessary computation – for all of these problems, 500 queries are not required to find the maximum. While there are other methods for multi-fidelity optimisation (discussed under Related Work) none of them had made their code available nor were their methods straightforward to implement - this includes MF-SKO.

A straightforward way to incorporate lower fidelity information to GP-UCB and EI is to share the same kernel parameters. This way, the kernel κ can be learned by by jointly maximising the marginal likelihood. While the idea seems natural, we got mixed results in practice. On some problems this improved the performance of all GP methods (including MF-GP-UCB), but on others all performed poorly. One explanation is that while lower fidelities approximate function values, they are not always best described by the same kernel. The results presented do not use lower fidelities to learn κ as it was more robust. For MF-GP-UCB, each $\kappa^{(m)}$ was learned independently using only the queries at fidelity m.

In addition to the baselines presented in the figures, we also compared our method to the following methods. The first two are single fidelity and the last two are mutlti-fidelity methods.

- The probability of improvement (PI) criterion for BO (Brochu et al., 2010). We found that in general either GP-UCB or EI performed better.
- Querying uniformly at random at the highest fidelity and taking the maximum. On all problems this performed worse than other methods.
- A variant of MF-NAIVE where instead of GP-UCB we queried at the first fidelity uniformly at random. On some problems this did better than querying with GP-UCB, probably since unlike GP-UCB it was not stuck at the maximum of $f^{(1)}$. However, generally it performed worse.
- The multi-fidelity method from Forrester et al. (2007) also based on GPs. We found that this method did not perform as desired: in particular, it barely queried beyond the first fidelity.

A.2 Description of Synthetic Experiments

The following are the descriptions of the synthetic functions used. The first three functions and their approximations were taken from Xiong et al. (2013).

Currin exponential function: The domain is the two dimensional unit cube $\mathcal{X} = [0,1]^2$. The second and first fidelity functions are,

$$f^{(2)}(x) = \left(1 - \exp\left(\frac{-1}{2x_2}\right)\right)\left(\frac{2300x_1^3 + 1900x_1^2 + 2092x_1 + 60}{100x_1^3 + 500x_1^2 + 4x_1 + 20}\right),$$
$$f^{(1)}(x) = \frac{1}{4}f^{(2)}(x_1 + 0.05, x_2 + 0.05) + \frac{1}{4}f^{(2)}(x_1 + 0.05, \max(0, x_2 - 0.05)) +$$
$$\frac{1}{4}f^{(2)}(x_1 - 0.05, x_2 + 0.05) + \frac{1}{4}f^{(2)}(x_1 - 0.05, \max(0, x_2 - 0.05)).$$

Park function: The domain is $\mathcal{X} = [0,1]^4$. The second and first fidelity functions are,

$$f^{(2)}(x) = \frac{x_1}{2}\left(\sqrt{1 + (x_2 + x_3^2)\frac{x_4}{x_1^2}} - 1\right) + (x_1 + 3x_4)\exp(1 + \sin(x_3)),$$
$$f^{(1)}(x) = \left(1 + \frac{\sin(x_1)}{10}\right)f^{(2)}(x) - 2x_1^2 + x_2^2 + x_3^2 + 0.5.$$

Borehole function: The second and first fidelity functions are,

$$f^{(2)}(x) = \frac{2\pi x_3(x_4 - x_6)}{\log(x_2/x_1)\left(1 + \frac{2x_7x_3}{\log(x_2/x_1)x_1^2x_8} + \frac{x_3}{x_5}\right)},$$
$$f^{(1)}(x) = \frac{5x_3(x_4 - x_6)}{\log(x_2/x_1)\left(1.5 + \frac{2x_7x_3}{\log(x_2/x_1)x_1^2x_8} + \frac{x_3}{x_5}\right)}.$$

The domain of the function is $[0.05, 0.15; 100, 50K; 63.07K, 115.6K; 990, 1110; 63.1, 116; 700, 820; 1120, 1680; 9855, 12045]$. We first linearly transform the variables to lie in $[0,1]^8$.

Hartmann-3D function: The M^{th} fidelity function is $f^{(M)}(x) = \sum_{i=1}^4 \alpha_i \exp\big(-\sum_{j=1}^3 A_{ij}(x_j - P_{ij})^2\big)$ where $A, P \in \mathbb{R}^{4\times 3}$ are fixed matrices given below and $\alpha = [1.0, 1.2, 3.0, 3.2]$. For the lower fidelities we use the same form except changing α to $\alpha^{(m)} = \alpha + (M - m)\delta$ where $\delta = [0.01, -0.01, -0.1, 0.1]$ and $M = 3$. The domain is $\mathcal{X} = [0,1]^3$.

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

Hartmann-6D function: The 6-D Hartmann function takes the same form as the 3-D case except $A, P \in \mathbb{R}^{4\times 6}$ are as given below. We use the same modifications as above to obtain the lower fidelities using $M = 4$.

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

Appendix B. Other Material

B.1 Some Ancillary Results

The following results were used in our analysis. The first is a standard Gaussian concentration result and the second is an expression for the Information Gain in a GP from Srinivas et al. (2010).

Lemma 14 (Gaussian Concentration). *Let $Z \sim \mathcal{N}(0,1)$. Then $\mathbb{P}(Z > \epsilon) \leq \frac{1}{2}\exp(-\epsilon^2/2)$.*

Lemma 15 (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. Using the basic Gaussian properties it can be shown that the 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_{t-1}^2 is the posterior GP variance after observing the first $t-1$ points.

B.2 A Table of Notations and Abbreviations

The following table summarises the notation and abbreviations used in the manuscript. The table continues to multiple pages.

Notation	Description
$\mathbb{E}_{\mathcal{GP}}, \mathbb{P}_{\mathcal{GP}}$	Expectations and probabilities when $f^{(1)}, \ldots, f^{(M)}$ are sampled from $\mathcal{GP}(0, \kappa)$.
$\mathbb{E}, \mathbb{P}$	Expectations and probabilities under the prior, which includes condition **A2** after $f^{(1)}, \ldots, f^{(M)}$ are sampled from $\mathcal{GP}(0, \kappa)$.
$\xi_{\mathbf{A2}}$	A lower bound on the probability that condition **A2** holds when $f^{(1)}, \ldots, f^{(M)}$ are sampled, see (5).
Q	The function which controls the probability on the supremum of a GP, see Assumption 2.
M	The number of fidelities.
$f, f^{(m)}$	The payoff function and its m^{th} fidelity approximation. $f^{(M)} = f$.
Λ	Λ typically denotes the capital of some resource which is expended upon each evaluation of at any fidelity.
$\lambda^{(m)}$	The cost, i.e. amount of capital expended, of querying at fidelity m.
N	The random number of queries at any fidelity within capital Λ. $N = \max\{n \geq 1 : \sum_{t=1}^{n} \lambda^{(m_t)} \leq \Lambda\}$
$\mathcal{X}$	The domain over which we are optimising f.
$x_\star, f_\star$	The optimum point and value of the M^{th} fidelity function.
$\overline{A}$	The complement of a set $A \subset \mathcal{X}$. $\overline{A} = \mathcal{X} \backslash A$.
$\lvert A \rvert$	The cardinality of a set $A \subset \mathcal{X}$ if it is countable.
$\vee, \wedge$	Logical *Or* and *And* respectively.
$\lesssim, \gtrsim, \asymp$	Inequalities and equality ignoring constant terms.
q_t, r_t	The instantaneous reward and regret respectively. $q_t = f^{(M)}(x_t)$ if $m_t = M$ and $-\infty$ if $m_t \neq M$. $r_t = f_\star - q_t$.

$S(\Lambda)$	The simple regret after spending capital Λ. $S(\Lambda) = f_\star - \min_{t=1,\ldots,N} f(x_t)$.
$\zeta^{(m)}$	The bound on the maximum difference between $f^{(m)}$ and $f^{(M)}$, $\|f^{(M)} - f^{(m)}\|_\infty \leq \zeta^{(m)}$.
$\mu_t^{(m)}$	The mean of the m^{th} fidelity GP $f^{(m)}$ conditioned on $\mathcal{D}_t^{(m)}$ at time t.
$\kappa_t^{(m)}$	The covariance of the m^{th} fidelity GP $f^{(m)}$ conditioned on $\mathcal{D}_t^{(m)}$ at time t.
$\sigma_t^{(m)}$	The standard deviatiation of the m^{th} fidelity GP $f^{(m)}$ conditioned on $\mathcal{D}_t^{(m)}$ at time t.
x_t, y_t	The queried point and observation at time t.
m_t	The queried fidelity at time t.
$\mathcal{D}_n^{(m)}$	The set of queries at the m^{th} fidelity until time n $\{(x_t, y_t)\}_{t:m_t=m}$.
β_t	The coefficient trading off exploration and exploitation in the UCB. See Theorems 5 and 6.
$\varphi_t^{(m)}(x)$	The upper confidence bound (UCB) provided by the m^{th} fidelity on $f^{(M)}(x)$. $\varphi_t^{(m)}(x) = \mu_{t-1}^{(m)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(m)}(x) + \zeta^{(m)}$.
$\varphi_t(x)$	The combined UCB provided by all fidelities on $f^{(M)}(x)$. $\varphi_t(x) = \min_m \varphi_t^{(m)}(x)$.
$\gamma^{(m)}$	The parameter in MF-GP-UCB for switching from the m^{th} fidelity to the $(m+1)^{\text{th}}$.
$\tilde{R}_n$	The M^{th} fidelity cumulative regret after n rounds. See (11)
$T_n^{(m)}(A)$	The number of queries at fidelity m in subset $A \subset \mathcal{X}$ until time n.
$T_n^{(>m)}(A)$	Number of queries at fidelities greater than m in any subset $A \subset \mathcal{X}$ until time n.
$\underline{n}_\Lambda$	$\underline{n}_\Lambda = \lfloor \Lambda/\lambda^{(M)} \rfloor$. Number of plays by a strategy querying only at fidelity M within capital Λ; also a lower bound on N, the number of plays by a multi-fidelity strategy.
$\overline{n}_\Lambda$	An upper bound on N, the number of plays by a multi-fidelity strategy within capital Λ. $\overline{n}_\Lambda = \lfloor \Lambda/\lambda^{(1)} \rfloor$.
$\Psi_n(A)$	The maximum information gain of a set $A \subset \mathcal{X}$ after n queries in A. See Definition 1.
$\Delta^{(m)}(x)$	$\Delta^{(m)}(x) = f_\star - f^{(m)} - \zeta^{(m)}$.
$\mathcal{J}_\eta^{(m)}$	The points in $\mathcal{X}$ whose $f^{(m)}$ value is within $\zeta^{(m)} + \eta$ of the optimum $f_\star$. $\mathcal{J}_\eta^{(m)} = \{x \in \mathcal{X}; \Delta^{(m)}(x) \leq \eta\}$.
$\mathcal{H}^{(m)}$	$(\mathcal{H}^{(m)})_{m=1}^M$ is a partitioning of $\mathcal{X}$. See Equation (6). The analysis of MF-GP-UCB hinges on these partitioning.
$\widehat{\mathcal{H}}^{(m)}, \breve{\mathcal{H}}^{(m)}$	The arms "above"/"below" $\mathcal{H}^{(m)}$. $\widehat{\mathcal{H}}^{(m)} = \bigcup_{\ell=m+1}^M \mathcal{H}^{(\ell)}$, $\breve{\mathcal{H}}^{(m)} = \bigcup_{\ell=1}^{m-1} \mathcal{H}^{(\ell)}$.
$\mathcal{H}_n^{(m)}$	An n-dependent dilation of $\mathcal{H}^{(m)}$ in the continuous setting. See Section 5.3.
$\widehat{\mathcal{H}}^{(m)}, \breve{\mathcal{H}}^{(m)}$	The arms "above"/"below" $\mathcal{H}^{(m)}$. $\widehat{\mathcal{H}}^{(m)} = \bigcup_{\ell=m+1}^M \mathcal{H}^{(\ell)}$, $\breve{\mathcal{H}}^{(m)} = \bigcup_{\ell=1}^{m-1} \mathcal{H}^{(\ell)}$.
$\mathcal{X}_g^\star$	The good set for $M = 2$ fidelity problems. $\mathcal{X}_g^\star = \{x \in \mathcal{X}; f_\star - f^{(1)}(x) \leq \zeta^{(1)}\}$.
$\mathcal{X}_g$	The inflated good set for MF-GP-UCB. $\mathcal{X}_g = \{x; f_\star - f^{(1)}(x) \leq \zeta^{(1)} + 3\gamma\}$.
$\Omega_\varepsilon(A)$	The ε–covering number of a subset $A \subset \mathcal{X}$ in the $\|\cdot\|_2$ metric.
Λ_1, Λ_2	The minimum capitals that need to be expended before the bound on $S(\Lambda)$ hold in Theorems 5 and 6.

Abbreviation	Description
UCB	Upper Confidence Bound
BO	Bayesian Optimisation
GP-UCB	Gaussian Process Upper Confidence Bound (Srinivas et al., 2010)
MF-GP-UCB	Multi-fidelity Gaussian Process Upper Confidence Bound
EI	(Gaussian Process) Expected Improvement (Jones et al., 1998)
MF-SKO	Multi-fidelity Sequential Kriging Optimisation (Huang et al., 2006)
MF-NAIVE	Naive multi-fidelity method described in Section 7.
DiRect	DIviding RECTangles (Jones et al., 1993)
SE	Squared Exponential (in reference to the kernel)

References

Adler, R. J. (1990). *An introduction to continuity, extrema, and related topics for general gaussian processes*. IMS.

Agarwal, A., Duchi, J. C., Bartlett, P. L., & Levrard, C. (2011). Oracle inequalities for computationally budgeted model selection. In *Colt*.

Auer, P. (2003). Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*.

Bogunovic, I., Scarlett, J., Krause, A., & Cevher, V. (2016). Truncated variance reduction: A unified approach to bayesian optimization and level-set estimation. In *Advances in neural information processing systems* (pp. 1507–1515).

Brochu, E., Cora, V. M., & de Freitas, N. (2010). A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *CoRR*.

Bubeck, S., & Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*.

Bubeck, S., Munos, R., Stoltz, G., & Szepesvári, C. (2011). X-armed bandits. *Journal of Machine Learning Research*, *12*(May), 1655–1695.

Cutler, M., Walsh, T. J., & How, J. P. (2014). Reinforcement Learning with Multi-Fidelity Simulators. In *International conference on robotics and automation*.

Dani, V., P. Hayes, T. P., & Kakade, S. M. (2008). Stochastic Linear Optimization under Bandit Feedback. In *Colt*.

Davis et al, T. M. (2007). Scrutinizing Exotic Cosmological Models Using ESSENCE Supernova Data Combined with Other Cosmological Probes. *Astrophysical Journal*.

Forrester, A. I. J., Sóbester, A., & Keane, A. J. (2007). Multi-fidelity optimization via surrogate modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, *463*.

Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., & Etrue, E. (2017). On context-dependent

clustering of bandits. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 1253–1262).

Ghosal, S., & Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression". *Annals of Statistics*.

Hernández-Lobato, J. M., Hoffman, M. W., & Ghahramani, Z. (2014). Predictive Entropy Search for Efficient Global Optimization of Black-box Functions. In *Neural information processing systems.*

Hoag, E., & Doppa, J. R. (2018). Bayesian optimization meets search based optimization: A hybrid approach for multi-fidelity optimization. In *Thirty-second aaai conference on artificial intelligence.*

Huang, D., Allen, T., Notz, W., & Miller, R. (2006). Sequential Kriging Optimization Using Multiple-fidelity Evaluations. *Structural and Multidisciplinary Optimization*.

Jones, D. R., Perttunen, C. D., & Stuckman, B. E. (1993). Lipschitzian Optimization Without the Lipschitz Constant. *Journal of Optimization Theory and Applications*.

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*.

Kandasamy, K., Dasarathy, G., Oliva, J., Schenider, J., & Póczos, B. (2016). Gaussian Process Bandit Optimisation with Multi-fidelity Evaluations. In *Advances in neural information processing systems.*

Kandasamy, K., Dasarathy, G., Poczos, B., & Schneider, J. (2016). The Multi-fidelity Multi-armed Bandit. In *Advances in neural information processing systems* (pp. 1777–1785).

Kandasamy, K., Dasarathy, G., Schneider, J., & Poczos, B. (2017). Multi-fidelity Bayesian Optimisation with Continuous Approximations. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 1799–1808).

Kandasamy, K., Schenider, J., & Póczos, B. (2015). High Dimensional Bayesian Optimisation and Bandits via Additive Models. In *International conference on machine learning.*

Kandasamy, K., & Yu, Y. (2016). Additive Approximations in High Dimensional Nonparametric Regression via the SALSA. In *International conference on machine learning.*

Kar, P., Li, S., Narasimhan, H., Chawla, S., & Sebastiani, F. (2016). Online optimization methods for the quantification problem. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1625–1634).

Kawaguchi, K., Kaelbling, L. P., & Lozano-Pérez, T. (2015). Bayesian Optimization with Exponential Convergence. In *Nips.*

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *SCIENCE*, *220*.

Klein, A., Bartels, S., Falkner, S., Hennig, P., & Hutter, F. (2015). Towards efficient Bayesian Optimization for Big Data. In *Bayesopt.*

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel

bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, *18*(1), 6765–6816.

Li, S., Karatzoglou, A., & Gentile, C. (2016). Collaborative filtering bandits. In *Proceedings of the 39th international acm sigir conference on research and development in information retrieval* (pp. 539–548).

Martinez-Cantin, R., de Freitas, N., Doucet, A., & Castellanos, J. (2007). Active Policy Learning for Robot Planning and Exploration under Uncertainty. In *Proceedings of robotics: Science and systems.*

Mockus, J. (1994). Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*.

Munos, R. (2011). Optimistic Optimization of Deterministic Functions without the Knowledge of its Smoothness. In *Nips.*

Parkinson, D., Mukherjee, P., & Liddle, A. R. (2006). A Bayesian Model Selection Analysis of WMAP3. *Physical Review*.

Poloczek, M., Wang, J., & Frazier, P. (2017). Multi-information source optimization. In *Advances in neural information processing systems* (pp. 4288–4298).

Rasmussen, C., & Williams, C. (2006). *Gaussian Processes for Machine Learning*. University Press Group Limited.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*.

Sabharwal, A., Samulowitz, H., & Tesauro, G. (2015). Selecting near-optimal learners via incremental data allocation. In *Aaai.*

Seeger, M., Kakade, S., & Foster, D. (2008). Information Consistency of Nonparametric Gaussian Process Methods. *IEEE Transactions on Information Theory*.

Sen, R., Kandasamy, K., & Shakkottai, S. (2018). Multi-Fidelity Black-Box Optimization with Hierarchical Partitions. In *International conference on machine learning.*

Sen, R., Kandasamy, K., & Shakkottai, S. (2019). Noisy blackbox optimization using multi-fidelity queries: A tree search approach. In *The 22nd international conference on artificial intelligence and statistics* (pp. 2096–2105).

Shang, X., Kaufmann, E., & Valko, M. (2017). Adaptive black-box optimization got easier: Hct only needs local smoothness. In *European workshop on reinforcement learning.*

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In *Neural information processing systems.*

Song, J., Chen, Y., & Yue, Y. (2018). A general framework for multi-fidelity bayesian optimization with gaussian processes. *ArXiv*, *abs/1811.00755*.

Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2010). Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *International conference on machine learning.*

Swersky, K., Snoek, J., & Adams, R. P. (2013). Multi-task bayesian optimization. In *Advances in neural information processing systems* (pp. 2004–2012).

Swersky, K., Snoek, J., & Adams, R. P. (2014). Freeze-thaw bayesian optimization. *ArXiv*, *abs/1406.3896*.

Thompson, W. R. (1933). On the Likelihood that one Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*.

Urmson, C., Anhalt, J., Bagnell, D., Baker, C., Bittner, R., Clark, M., ... others (2008). Autonomous Driving in Urban Environments: Boss and the Urban Challenge. *Journal of Field Robotics*, *25*.

Vapnik, V. N., & Vapnik, V. (1998). *Statistical learning theory*. Wiley New York.

Viola, P. A., & Jones, M. J. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. In *Computer vision and pattern recognition.*

Wu, J., Toscano-Palmerin, S., Frazier, P. I., & Wilson, A. G. (2019). Practical multi-fidelity bayesian optimization for hyperparameter tuning. In *Uai.*

Xiong, S., Qian, P. Z. G., & Wu, C. F. J. (2013). Sequential design and analysis of high-accuracy and low-accuracy computer codes. *Technometrics*, *55*.

Zhang, C., & Chaudhuri, K. (2015). Active Learning from Weak and Strong Labelers. In *Nips.*