arXiv:2210.13937v1 [cs.LG] 25 Oct 2022

Multi-Fidelity Bayesian Optimization with Unreliable Information Sources

Petrus Mikkola [1] Julien Martinelli [1] Louis Filstroff [2] *
Samuel Kaski [1,3]

[1] Department of Computer Science, Aalto University, Finland

[2] ENSAI, CREST

[3] Department of Computer Science, University of Manchester, UK

Bayesian optimization (BO) is a powerful framework for optimizing black-box, expensive-to-evaluate functions. Over the past decade, many algorithms have been proposed to integrate cheaper, lower-fidelity approximations of the objective function into the optimization process, with the goal of converging towards the global optimum at a reduced cost. This task is generally referred to as multi-fidelity Bayesian optimization (MFBO). However, MFBO algorithms can lead to higher optimization costs than their vanilla BO counterparts, especially when the low-fidelity sources are poor approximations of the objective function, therefore defeating their purpose. To address this issue, we propose rMFBO (robust MFBO), a methodology to make any GP-based MFBO scheme robust to the addition of unreliable information sources. rMFBO comes with a theoretical guarantee that its performance can be bound to its vanilla BO analog, with high controllable probability. We demonstrate the effectiveness of the proposed methodology on a number of numerical benchmarks, outperforming earlier MFBO methods on unreliable sources. We expect rMFBO to be particularly useful to reliably include human experts with varying knowledge within BO processes.

1. Introduction

Bayesian optimization (BO) has become a popular framework for global optimization of black-box functions, especially when they are expansive to evaluate (Jones et al., 1998;

*LF was with the Department of Computer Science, Aalto University, Finland at the time this research was conducted.

Brochu et al., 2010). Such functions have neither known functional form nor derivatives, and conventional optimization techniques such as gradient descent cannot be directly employed. BO rests upon two key elements. First, it constructs a probabilistic surrogate model of the objective function with built-in uncertainty estimates (typically, a Gaussian process, GP), based on evaluations of the function. The obtained surrogate is then used to select the next point to evaluate through the maximization of a so-called *acquisition function*, which quantifies the expected utility of evaluating a specific point with the purpose of optimizing the black-box function. Many off-the-shelf acquisition functions achieve this task while balancing exploration and exploitation. Iterating these two steps produces a sequence whose aim is to converge to the global optimum using a limited number of function queries. BO has proven effective for a variety of problems, including hyperparameter optimization (Snoek et al., 2012), materials science (Zhang et al., 2020), and drug discovery (Gómez-Bombarelli et al., 2018; Korovina et al., 2020).

In many scenarios, lower-fidelity approximations of the objective function are available, at a cheaper query cost. This occurs for instance when the evaluation of the objective function involves a numerical scheme, where computational cost and accuracy can be traded off. Another example is the knowledge of domain experts. Indeed, practitioners often have implicit knowledge of the objective function, such as good candidate regions about the location of the global optimum (Hvarfner et al., 2022). Such knowledge may naturally be considered as a low-fidelity version of the true objective function.

The problem of integrating these auxiliary information sources (ISs) to reduce the cost of BO has been tackled in the literature under the name multi-fidelity Bayesian optimization (MFBO) (Huang et al., 2006; Kandasamy et al., 2016; Zhang et al., 2017; Sen et al., 2018; Song et al., 2019; Takeno et al., 2020; Li et al., 2020; Moss et al., 2021) when the different sources can be ranked by their degree of fidelity; when this is not possible, the problem has been studied as multi-task BO (Swersky et al., 2013), non-hierarchical multi-fidelity BO (Lam et al., 2015), or multi-information source BO (Poloczek et al., 2017). However, as we will empirically demonstrate, state-of-the-art MFBO algorithms can fail when the auxiliary ISs are poor approximations of the primary IS. More precisely, for a fixed budget, these algorithms will lead to a higher regret w.r.t. their single-fidelity counterparts (i.e., vanilla BO, which uses the primary IS only), defeating their purpose. For instance, the guarantees of Kandasamy et al. (2016) require that the deviation between an auxiliary IS and the primary IS is bounded by a constant known beforehand, which hardly ever holds in practice, for example when working with a human expert, or when experimenting with simulations to find the optimal control parameters for a robotic system (Marco et al., 2017).

Surprisingly enough, this issue has not formally been addressed in the BO literature so far. We fix this gap, introducing rMFBO, a methodology to make any GP-based MFBO algorithm *robust* to the addition of unreliable information sources. More precisely, rMFBO comes with a theoretical guarantee that its performance can be bound to its vanilla BO analog, with high controllable probability. To the best of our knowledge, rMFBO is the first MFBO scheme providing such performance guarantees. We then proceed to

demonstrate the effectiveness of the proposed methodology on various numerical settings using different MFBO algorithms of the literature. Through its building block nature, rMFBO paves the way towards a more systematic usage of auxiliary ISs independently of their degree of fidelity, allowing human experts to join the optimization process in a reliable manner.

2. Preliminaries

Gaussian process regression

We begin by introducing the notation for single-output Gaussian process regression, the probabilistic surrogate upon which BO rests. Consider a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ with $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, for which we want to learn a model of the form $y_i = f(\mathbf{x}_i) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$ for all i. We may place a (zero-mean) GP prior on f:

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')). \tag{1}$$

This defines a distribution over functions f whose mean is $\mathbb{E}[f(\mathbf{x})] = 0$ and covariance $\text{cov}[f(\mathbf{x}), f(\mathbf{x}')] = k(\mathbf{x}, \mathbf{x}')$. Here, k is a kernel function measuring the similarity between inputs. Consequently, for any finite-dimensional collection of inputs $(\mathbf{x}_1, \dots, \mathbf{x}_n)$, the function values $\boldsymbol{f} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^{\mathrm{T}} \in \mathbb{R}^n$ follow a multivariate normal distribution $\boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, K)$, where $K \in \mathbb{R}^{n \times n} = (k(\mathbf{x}_i, \mathbf{x}_j))_{1 \le i,j \le n}$ is the kernel matrix.

Given $\mathcal{D}$, the posterior predictive distribution $p(f(\mathbf{x}) \mid \mathcal{D})$ is Gaussian for all $\mathbf{x}$ with mean $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$, such that

$$\begin{aligned} \mu(\mathbf{x}) &= \boldsymbol{k}_{\mathbf{x}} (K + \sigma^2_{\text{noise}} I)^{-1} y, \\ \sigma^2(\mathbf{x}) &= k(\mathbf{x}, \mathbf{x}) - \boldsymbol{k}_{\mathbf{x}} (K + \sigma^2_{\text{noise}} I)^{-1} \boldsymbol{k}_{\mathbf{x}}, \end{aligned}$$

where $\boldsymbol{k}_{\mathbf{x}} = [k(\mathbf{x}, \mathbf{x}_1), \cdots, k(\mathbf{x}, \mathbf{x}_n)]^{\mathrm{T}} \in \mathbb{R}^n$.

Multi-output Gaussian process regression

GPs can be extended to Multi-Output Gaussian processes (MOGP), modeling any collection of m-sized vector-valued outputs $(\mathbf{y}_1, ..., \mathbf{y}_n)$ based on inputs $(\mathbf{x}_1, ..., \mathbf{x}_n)$ as a multivariate normal distribution. One way to achieve this extension is through the addition of a $(d+1)^{\text{th}}$ dimension to the input space, representing the output index $1 \le l \le m$. This enables treating a MOGP as a single-output GP acting on the augmented space $\mathbb{R}^{d+1}$ through the kernel $k((\mathbf{x}, \ell), (\mathbf{x}', \ell'))$. The latter can, for instance, take the separable form $k((\mathbf{x}, \ell), (\mathbf{x}', \ell')) = k_{\text{input}}(\mathbf{x}, \mathbf{x}') \times k_{\text{IS}}(\ell, \ell')$. In particular, this setting allows the use of the readily-available analytical formulae for the posterior mean and variance of single-output GPs.

Problem setup

We consider the problem of optimizing a black-box function $f^{(m)} : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a subset of $\mathbb{R}^d$, i.e. solving

$$\arg\max_{\mathbf{x}\in\mathcal{X}} f^{(m)}(\mathbf{x}). \tag{2}$$

In addition to $f^{(m)}$ (the *primary* IS), we may also query $m-1$ other auxiliary functions (*auxiliary* ISs), $f^{(\ell)} : \mathcal{X} \to \mathbb{R}$, where $\ell \in [\![m-1]\!]$ denotes the IS index. The cost of evaluating $f^{(\ell)}(\mathbf{x})$ is λ_ℓ for any $\mathbf{x} \in \mathcal{X}$. We assume that $\lambda_\ell < \lambda_m$ for any auxiliary IS $\ell \in [\![m-1]\!]$. The objective is to solve (2) within the budget $\Lambda > 0$.

Bayesian optimization

At each round t, an input-IS pair $(\mathbf{x}, \ell) \in \mathcal{X} \times [\![m]\!]$ is selected by maximizing the acquisition function α, which depends on the GP surrogate model on f given all the data acquired up until round $t-1$:

$$\mathbf{x}_t, \ell_t = \arg\max_{(\mathbf{x},\ell)\in\mathcal{X}\times[\![m]\!]} \alpha(\mathbf{x}, \ell). \tag{3}$$

Querying for $f^{(\ell)}(\mathbf{x})$ returns a noisy observation $y_{\mathbf{x}}^{(\ell)} = f^{(\ell)}(\mathbf{x}) + \epsilon$, with i.i.d. noise $\epsilon \sim \mathcal{N}(0, \sigma^2_{noise})$. We refer to the sequence of queries $\{\mathbf{x}_t\}_{t=1}^T$ returned by a BO algorithm as an *acquisition trajectory*.

Note that vanilla BO (i.e., BO with the primary IS only) amounts to using the acquisition function $\mathbf{x} \mapsto \alpha(\mathbf{x}, m)$, which will be referred to as single-fidelity BO (SFBO) from now on.

Recall that we wish to optimize $f^{(m)}$ within the budget Λ. In this scenario, the performance metric of interest is the regret of the algorithm, whose definition is recalled below.

Definition 1 (BO regret)**.** *The regret of the BO algorithm that spends Λ_{cost} and returns the final choice $\mathbf{x}_{choice}$, is defined by*

$$R(\Lambda_{cost}, \mathbf{x}_{choice}) := \begin{cases} f^* - f^{(m)}(\mathbf{x}_{choice}) & \text{if } \Lambda_{cost} \leq \Lambda, \\ \infty & \text{otherwise} \end{cases}$$

where $f^* = \max_{\mathbf{x}\in\mathcal{X}} f^{(m)}$ is the global maximum of the primary IS. If $\Lambda_{cost} \leq \Lambda$, then the shorthand notation $R(\mathbf{x}_{choice})$ can be used.

Definition 2 (Number of queries)**.** *Let $T := \lfloor \Lambda/\lambda_m \rfloor$ be the available number of primary IS queries. Let $T^{(\ell)}$ be the random variable describing the number of ℓ^{th}-IS queries spent by the MFBO algorithm.*

There are two popular choices for $\mathbf{x}_{\text{choice}}$. First, the *Bayes-optimal choice*

$$\mathbf{x}_{\text{choice}} = \arg\max_{\mathbf{x}\in\mathcal{X}} \mu(\mathbf{x}, m),$$

where $\mu(\mathbf{x}, m)$ is the posterior mean of the GP model given all the data up to the final query $T_{\text{last}} = \sum_{\ell=1}^{m} T^{(\ell)}$. The regret in this case is called the *inference regret*. Second, the *simple choice*

$$\mathbf{x}_{\text{choice}} = \arg\max_{t\in[\![T^{(m)}]\!]} f^{(m)}(\mathbf{x}_t),$$

where $(\mathbf{x}_1, ..., \mathbf{x}_{T^{(m)}})$ is the primary IS acquisition trajectory returned by the MFBO algorithm. The regret in that case is called the *simple regret*.

3. Related work

As discussed in the introduction, many different multi-fidelity extensions of Bayesian optimization have been proposed in the literature; we refer the interested reader to Takeno et al. (2020, Section 5) for a review. The closest to our work are methods that do not assume a hierarchy between the sources (e.g., when the degree of fidelity cannot be assessed in advance), as by Lam et al. (2015), where the focus lies in designing a GP model that takes into account the non-hierarchical nature of the sources. The multi-fidelity kernel introduced by Poloczek et al. (2017) (Appendix E) is one example of such a design.

The problem of the potential performance degradation of MFBO algorithms has surprisingly been largely ignored by the literature, with the exception of Kandasamy et al. (2016), who noticed that their multi-fidelity method performed poorly compared to all single-fidelity variants in one experiment (Kandasamy et al., 2016, Appendix D.3).

Lastly, we mention that robustness has been studied for vanilla BO in the context of (sometimes adversarially) noisy inputs or outputs (Martinez-Cantin et al., 2018; Bogunovic et al., 2018; Fröhlich et al., 2020; Kirschner and Krause, 2021). This notion of robustness is fundamentally different from ours. Indeed, we wish to provide guarantees that the addition of an (or several) auxiliary IS will not lead to worse performance w.r.t. vanilla BO; we do not assume noisy inputs, or that the outputs of any of the sources are corrupted by outliers.

4. Pitfalls of MFBO methods

We now demonstrate, on a simple example, the influence of the auxiliary IS quality on the performance of MFBO algorithms. Let us consider the Hartmann6D function as the objective (i.e., the primary IS). We examine two scenarios: in the first one, the auxiliary IS is informative, consisting of a biased version of the primary IS, with a degree of fidelity $l = 0.2$. In the second scenario, the auxiliary IS is taken to be the 6-dimensional Rosenbrock function, an irrelevant source for this problem. Analytical forms for these

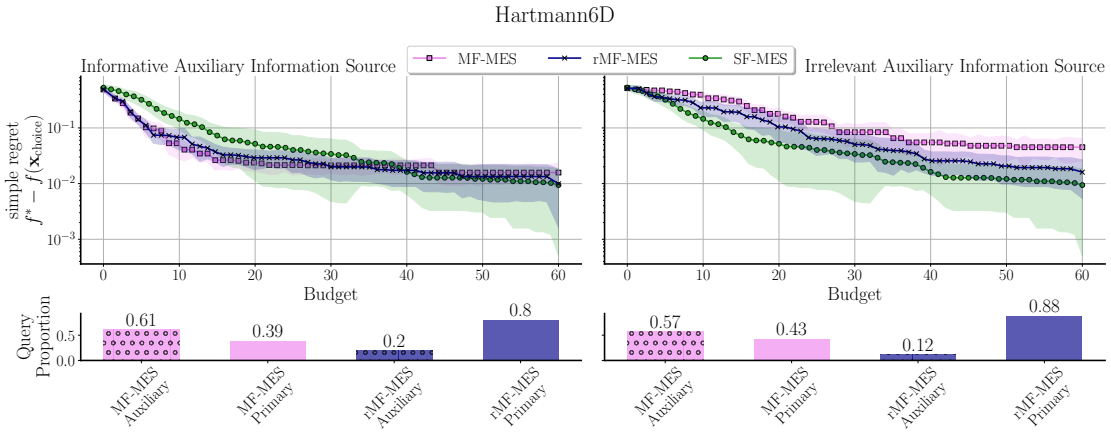

Figure 1: Simple regret as a function of budget spent in two multi-fidelty problems, averaged over 20 repetitions. The informative auxiliary IS helps reduce the cost of BO (left panel: MF-MES in purple converges faster than SF-MES in green), whereas the irrelevant IS catastrophically disrupts performance (right panel: MF-MES does not reach the low regret of SF-MES even in the long run). In both settings, the primary IS cost is set to 1 and the auxiliary IS cost to 0.2. From relevant to irrelevant IS, the proportion of auxiliary IS queries remains high for MF-MES, while rMF-MES is more consistent (lower panel).

examples can be found in Appendix F. We evaluate the multi-fidelity maximum-entropy search (MF-MES) method from Takeno et al. (2020) on these two scenarios as well as its single fidelity counterpart (SF-MES), and our proposed algorithm, rMFBO, built on top of these methods (rMF-MES). In both cases, the cost of the primary IS is set to 1, and that of the auxiliary IS to 0.2. The simple regret of the three algorithms averaged over 20 runs is displayed on Figure 1.

When the auxiliary IS is informative (left panel), MF-MES converges faster than SF-MES. This is the expected behavior from MFBO algorithms:they use cheap IS queries in the beginning to clear out unpromising regions of the space at a low cost, which eventually speeds up convergence. However, when the auxiliary IS is irrelevant (right panel), there is a clear gap between MF-MES and SF-MES, even in the long run. This demonstrates the inability of MF-MES to deal with an irrelevant IS. In that scenario, we hypothesize that the budget is wasted on uninformative queries, and thus too many rounds are spent on learning that the sources are not correlated (Figure 1, right bar plot), leading to a sub-optimal data acquisition trajectory.

There is therefore a need for a robust method to such a scenario. This is what the proposed rMF-MES, formally introduced in the next section, achieves, by taking the best of both worlds: sticking close to the single fidelity track in case of an irrelevant IS while using informative lower-fidelity queries to accelerate convergence.

5. Robust MFBO algorithm

In this section, we introduce rMFBO (robust MFBO), a methodology to make any GP-based MFBO scheme robust to the addition of unreliable ISs. The key idea is to control the quality of the acquisitions, to prevent the MFBO algorithm from behaving as described at the end of Section 4.

At round t, based on the acquisition function α, MFBO proposes the query $(\mathbf{x}_t^{\mathrm{MF}}, \ell_t)$ according to Eq. (3). The question is to decide whether to execute this query, or to go with a more conservative query from the primary IS. Indeed, we wish to curb the potential performance deterioration w.r.t. vanilla BO. To do so, we introduce a concurrent pseudo-SFBO algorithm, which constructs a GP surrogate based on data from the primary IS only, and so-called *pseudo-observations*, introduced later on. The pseudo-SFBO uses the acquisition function $\mathbf{x} \mapsto \alpha(\mathbf{x}, m)$ and a separate single-output GP, yielding the query $(\mathbf{x}_t^{\mathrm{pSF}}, m)$.

Let us denote the predictive mean and standard deviation of the MOGP model (used by the MFBO algorithm) by μ_{MF} and σ_{MF}, and those of the GP model (used by the pseudo-SFBO algorithm) by μ_{SF} and σ_{SF}. In a nutshell, the proposed rMFBO follows the conservative query $\mathbf{x}_t^{\mathrm{pSF}}$, unless the predictive variance of the MOGP model at $\mathbf{x}_t^{\mathrm{pSF}}$ is small enough:

$$\sigma_{\mathrm{MF}}(\mathbf{x}_t^{\mathrm{pSF}}, m) \leq c_1, \tag{4}$$

where $c_1 > 0$ is a user-specified parameter. The pseudo-SBFO learns from all the samples, even when the MFBO candidate $\mathbf{x}_t^{\mathrm{MF}}$ is queried, by adding the pseudo-obervation $\mu_{\mathrm{MF}}(\mathbf{x}_t^{\mathrm{pSF}}, m)$.

It can easily be seen that if $c_1 \to 0$, the described algorithm becomes the SFBO algorithm, since the MFBO proposals would be always ignored. Our main result, formally discussed in Section 4, is that we are able to derive a lower bound on the regret difference between robust MFBO and SFBO as a function of $c_1 > 0$.

While condition (4) ensures that we can achieve similar performance as SFBO when auxiliary IS is irrelevant, we also want to reap the benefits of the multi-fidelity approach when the auxiliary IS is relevant. To that end, we introduce a measure of IS relevance, s, and add this second condition for the acceptance of the MF query:

$$s(\mathbf{x}_t^{\mathrm{MF}}, \ell_t) \geq c_2, \tag{5}$$

with $c_2 > 0$ a user-specified parameter. We want to draw attention on the fact that condition (4) operates over SFBO proposal while condition (5) acts on the MFBO proposal, possibly based on an auxiliary IS. Condition (5) makes rMFBO revert more often to primary IS, and makes pseudo-observations more accurate overall. This allows the algorithm to consider less conservative values for c_1, opening the door for the exploitation of auxiliary ISs.

Algorithm 1: Robust MFBO algorithm

Input: Budget Λ, costs $(\lambda_1, ..., \lambda_m)$, acquisition function α, hyperparameters c_1 and c_2, relevance measure s
Initialize $\mathcal{D}^{\mathrm{pSF}}, \mathcal{D}^{\mathrm{MF}}$
Perform Bayesian updates $\mu_{\mathrm{SF}}, \sigma_{\mathrm{SF}}, \mu_{\mathrm{MF}}, \sigma_{\mathrm{MF}}$
$t \leftarrow 1$
while $\lfloor \Lambda/\lambda_m \rfloor \geq 2\lambda_m$ **do**
  $\mathbf{x}_t^{\mathrm{pSF}} \leftarrow \arg\max_{\mathbf{x}} \alpha(\mathbf{x}, m \mid \mu_{\mathrm{SF}}, \sigma_{\mathrm{SF}})$
  $(\mathbf{x}_t^{\mathrm{MF}}, \ell_t) \leftarrow \arg\max_{\mathbf{x},\ell} \alpha(\mathbf{x}, \ell \mid \mu_{\mathrm{MF}}, \sigma_{\mathrm{MF}})$
  if $\sigma_{\mathrm{MF}}(\mathbf{x}_t^{\mathrm{pSF}}, m) \leq c_1$ **and** $s(\mathbf{x}_t^{\mathrm{MF}}, \ell_t) \geq c_2$ **then**
    $y_t \leftarrow f(\mathbf{x}_t^{\mathrm{MF}}, \ell_t)$
    $\mathcal{D}^{\mathrm{MF}} \leftarrow \mathcal{D}^{\mathrm{MF}} \cup \{((\mathbf{x}_t^{\mathrm{MF}}, \ell_t), y_t)\}$
    Perform Bayesian updates $\mu_{\mathrm{MF}}, \sigma_{\mathrm{MF}}$
    $y_t \leftarrow \mu_{\mathrm{MF}}(\mathbf{x}_t^{\mathrm{pSF}}, m)$ # pseudo-observation
    $\mathcal{D}^{\mathrm{pSF}} \leftarrow \mathcal{D}^{\mathrm{pSF}} \cup \{(\mathbf{x}_t^{\mathrm{pSF}}, y_t)\}$
    $\Lambda \leftarrow \Lambda - \lambda_{\ell_t}$
  else
    $y_t \leftarrow f(\mathbf{x}_t^{\mathrm{pSF}}, m)$
    $\mathcal{D}^{\mathrm{pSF}} \leftarrow \mathcal{D}^{\mathrm{pSF}} \cup \{(\mathbf{x}_t^{\mathrm{pSF}}, y_t)\}$
    $\mathcal{D}^{\mathrm{MF}} \leftarrow \mathcal{D}^{\mathrm{MF}} \cup \{((\mathbf{x}_t^{\mathrm{pSF}}, m), y_t)\}$
    $\Lambda \leftarrow \Lambda - \lambda_m$
  end if
  Perform Bayesian updates $\mu_{\mathrm{SF}}, \sigma_{\mathrm{SF}}, \mu_{\mathrm{MF}}, \sigma_{\mathrm{MF}}$
  $t \leftarrow t + 1$
end while
$S \leftarrow \{\mathbf{x} \in \mathcal{X} \mid \sigma_{\mathrm{MF}}(\mathbf{x}, m) \leq c_1\}$
$\mathbf{x}_t^{\mathrm{pSF}} \leftarrow \arg\max_{\mathbf{x} \in S} \mu_{\mathrm{MF}}(\mathbf{x}, m)$
$y_t \leftarrow f(\mathbf{x}_t^{\mathrm{pSF}}, m)$

In this paper, we use a cost-weighted information gain of $(\mathbf{x}, \ell)$ (Takeno et al., 2020),

$$s(\mathbf{x}, \ell) = \frac{\mathrm{I}(f(\mathbf{x}, \ell), f_* \mid \mathcal{D}^{\mathrm{MF}})}{\lambda_\ell}, \tag{6}$$

where I is the mutual information between the observation $f(\mathbf{x}, \ell)$ and the maximal value of $f^{(m)}$, $f_* := \max_{\mathbf{x} \in \mathcal{X}} f^{(m)}(\mathbf{x})$; in other words the information gain on f_* brought by the observation $f(\mathbf{x}, \ell)$.

The whole procedure is summarized in Algorithm 1, and an extended version can be found and is discussed in Appendix C.

6. Theoretical results

In this section we tie the regret of rMFBO to that of its SFBO counterpart. The derivation holds for any relevance measure s.

Let us first define the function $f : \mathcal{X} \times [\![m]\!] \to \mathbb{R}$ such that $f(\mathbf{x}, \ell) = f^{(\ell)}(\mathbf{x})$ for all $(\mathbf{x}, \ell) \in \mathcal{X} \times [\![m]\!]$. We assume that $\mathcal{X}$ is a convex compact subset of $\mathbb{R}^d$, and we make the following assumptions about f:

Assumption 1 (f is drawn from a MOGP)**.** *Assume f is a draw from a MOGP with zero-mean and covariance function $\kappa((\mathbf{x}, \ell), (\mathbf{x}', \ell'))$. In other words, $\{f(\mathbf{x}_i, \ell_i)\}_i$ is multivariate normal for any finite set of input-IS pairs $\{(\mathbf{x}_i, \ell_i)\}_i$.*

Assumption 2. *κ is known.*

Assumption 3. *κ is at least twice differentiable.*

These assumptions are common in the Bayesian optimization literature. For instance, see Srinivas et al. (2012) and Kandasamy et al. (2016). We also follow these authors in the next assumption.

Assumption 4 (Bounded derivatives with high probability)**.**

$$\mathbb{P}\left(\sup_{\mathbf{x}\in\mathcal{X}} \left|\frac{\partial f}{\partial x_j}\right| > L\right) \leq a e^{-(L/b_j)^2}, \quad \forall\ j \in [\![d]\!]$$

for some constants $a, b_j > 0$.

Since a function with bounded partial derivatives (with an uniform bound L) is Lipschitz continuous (with a Lipschitz constant $\sqrt{d}L$), Assumption 4 implies by complementing and the union bound that

$$|f^{(m)}(\mathbf{x}) - f^{(m)}(\mathbf{x}')| \leq \sqrt{d} L \left\|\mathbf{x} - \mathbf{x}'\right\|_2 \quad \forall\ \mathbf{x}, \mathbf{x}' \in \mathcal{X},$$

with probability greater than $1 - dae^{-(L/b_j)^2}$. Further, Assumption 4 is satisfied for four times differentiable kernels (Ghosal and Roy, 2006, Theorem 5).

The next assumptions relate to the acquisition function.

Assumption 5. *For any round $t \in \mathbb{N}$, we assume that the mapping $(\mathbf{x}, \mathcal{D}_t) \mapsto \alpha(\mathbf{x}, m, \mathcal{D}_t)$ is C^2.*

Assumption 6. *The Hessian $\nabla^2_{\mathbf{x}} \alpha(\mathbf{x}, m)$ is a definite matrix at the optimum $\mathbf{x} = \mathbf{x}^*$.*

Running Algorithm 1 for T rounds returns the trajectory $\{\mathbf{x}_t^{\text{pSF}}\}_{t=1}^T$, which consists of primary IS queries or pseudo-primary IS queries. Moreover, we denote the acquisition

trajectory returned by the single-fidelity counterpart as $\{\mathbf{x}_t^{\mathrm{SF}}\}_{t=1}^T$. Our reasoning is as follows: we first control the closeness of the two acquisition trajectories, then derive a lower bound on the difference of their regret.

Let us consider the dataset as a $t(d+1)$-dimensional vector $\mathcal{D}_t = (x_1^{(1)}, ..., x_t^{(d)}, y_1, ..., y_t)$. Let $\mathbb{D}_t$ be the closed line segment joining two datasets $\mathcal{D}_t^A$ and $\mathcal{D}_t^B$. We introduce a concept, *the maximum rate of change of the next query with respect to* $\mathbb{D}_t$, defined as the random variable M_t

$$M_t = \max_{\mathcal{D}\in\mathbb{D}_t} \left\| \frac{\partial \mathbf{x}_{t+1}}{\partial \mathcal{D}}(\mathcal{D}) \right\|_{\mathrm{op}},$$

where $\|\cdot\|_{\mathrm{op}}$ is the operator norm. M_t measures the sensitivity of the next query when moving from a dataset $\mathcal{D}_t^A$ to a dataset $\mathcal{D}_t^B$. It depends on the smoothness of the objective function f, the kernel k, and the acquisition function α. The detailed formulas Eqs. (12)-(13) and the computation details for M_t can be found in Appendix B and D. Consider $\mathcal{D}_t^A = \mathcal{D}_t^{\mathrm{pSF}}$ and $\mathcal{D}_t^B = \mathcal{D}_t^{\mathrm{SF}}$, and let us denote by $\hat{M}_t$ the largest product of any combination of $M_0, ..., M_{t-1}$,

$$\hat{M}_t = \max_{S\in 2^{[\![t-1]\!]}} \prod_{k\in S} M_k. \tag{7}$$

Proposition 1. *Assume running Algorithm (1) with control parameter*

$$c_1(\varepsilon, q) = \frac{\varepsilon}{\sqrt{-2\log(1-q)}}, \tag{8}$$

for $\varepsilon > 0$ *and* $q \in]0,1[$. *For all* $t \in [\![T-1]\!]$,

$$\left\| \mathbf{x}_t^{SF} - \mathbf{x}_t^{pSF} \right\|_\infty \leq \varepsilon t \hat{M}_t d^t \tag{9}$$

holds with probability greater than $q\left(1 - da\exp(-1/b^2)\right)$.

Proof. The proof for the noiseless scenario ($\sigma_{\mathrm{noise}} = 0$) is given in Appendix B.1. For a noisy scenario ($\sigma_{\mathrm{noise}} > 0$), see the proof in Appendix B.2. □

Corollary 1. *The instant regret difference for all* $t \in [\![T-1]\!]$ *is bounded; we have:*

$$\left| f^{(m)}(\mathbf{x}_t^{SF}) - f^{(m)}(\mathbf{x}_t^{pSF}) \right| \leq \varepsilon t \hat{M}_t d^{t+1} \tag{10}$$

with probability greater than $q\left(1 - da\exp(-1/b^2)\right)$.

Proof. By Assumption 4 and the equivalence of the norms $\|\cdot\|_2$ and $\|\cdot\|_\infty$, it holds with probability greater than $1 - da\exp(-\frac{1}{b^2})$ that:

$$\left| f^{(m)}(\mathbf{x}) - f^{(m)}(\mathbf{x}') \right| \leq d \left\| \mathbf{x} - \mathbf{x}' \right\|_\infty,$$

for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Corollary 1 follows from Proposition 1. □

We can now present our main result, which states that with conservative control parameter $c_1(\varepsilon, q)$ (small ε and high q), the worst case remains close to the same regret of the SFBO algorithm, with high probability. This means that including an auxiliary IS in the robust MFBO algorithm will not cause any "harm", given conservative control parameters.

Theorem 1 ("No harm"). *Let's assume that both algorithms, the robust MFBO (Algorithm 1) and its SFBO variant, return their simple final choices. Then,*

$$R(\Lambda + \lambda_m, \mathbf{x}^{rMF}_{choice}) \leq R(\Lambda, \mathbf{x}^{SF}_{choice}) + \varepsilon \max\left\{T\hat{M}_T d^{T+1}, 2\right\}$$

with probability greater than $q\left(1 - da\exp(-\frac{1}{b^2})\right)$.

Proof. The proof is given in Appendix B.3. □

Among other things, the proof requires an amount λ_m of the budget Λ to be dedicated to a final target query (Algorithm 1, line 25).

Theorem 1 says that the magnitude of a possible regret loss compared to SFBO is proportional to ε with a probability proportional to q. For instance, if we tolerate 0.1 units of regret undershoot with 90% probability, then by Theorem 1 this is guaranteed with the control parameter value $c_1 = h(0.1, 0.9) \approx 0.05$ for early BO rounds.

7. Experimental results

We evaluate rMFBO on a benchmark of synthetic functions widely used in multi-fidelity studies. Our proposed method is used to make three state-of-the-art MFBO algorithms robust: Maximum Entropy Search (MF-MES) (Takeno et al., 2020), General-purpose Information-Based Bayesian OptimisatioN (GIBBON) (Moss et al., 2021) and Knowledge Gradient (MFKG) (Poloczek et al., 2017), the latter being benchmarked only on low-dimensional problems. The experiments are run within the BoTorch framework (Balandat et al., 2020). The probabilistic surrogate model uses the downsampling kernel from Wu et al. (2019): $k((\mathbf{x}, \ell), (\mathbf{x}', \ell')) = k_{\text{input}}(\mathbf{x}, \mathbf{x}') \times k_{\text{IS}}(\ell, \ell')$, where $k_{\text{input}}(\cdot, \cdot)$ is the RBF kernel, and $k_{\text{IS}}(\ell, \ell') := c + (1 - \ell)^{1+\delta}(1 - \ell')^{1+\delta}$. Here, $l \in [0, 1]$ represents the degree of fidelity of the IS, $l = 1$ referring to the target function, often denoted m. The hyperparameters c and δ are estimated marginal likelihood maximization, similarly as those of k_{input}. Each test function is rescaled between $[0, 1]$. Analytical forms can be found in Appendix F. The initial dataset consists of $5d$ evaluations of the primary IS and $4d$ for each auxiliary IS. For the remainder of the section and until the ablation study, we set rMFBO hyperparameters to $c_1 = c_2 = 0.1$. For each experiment, we report the average and standard deviation of the simple regrets computed over 20 repetitions with different initializations.

7.1. Synthetic functions with one auxiliary IS

The goal is to maximize a target function using noisy evaluations of the objective (primary IS), and an auxiliary IS. Figure 1 displays results obtained in such a setting, with an informative IS case, and an irrelevant IS case.

Negated Exponential Currin 2D: We reproduce the experiment performed by Kandasamy et al. (2016) and consider the Exponential Currin function as the primary IS while the auxiliary IS is the negated objective function itself, with $\lambda_l = 0.1, \lambda_m = 1$. rMFBO is able to find the global optimum using a lower budget than its MFBO counterpart (Figure 3), except in the case of the Knowledge Gradient acquisition function, where, surprisingly, the MFBO algorithm outperforms SFBO.

Sinus-perturbed Rosenbrock 2D: Next, we examine the Rosenbrock 2D function as the target. In this experiment, the auxiliary IS is equal to the objective corrupted with a sinusoidal signal whose magnitude is the target function mean. The costs are $\lambda_l = 0.2, \lambda_m = 1$. Here, rMFBO is able to rely on the corrupted version of the target (Figure 4), consistently outperforming SFBO in the early rounds regime for MES and GIBBON. For these methods, rMFBO and SFBO reach the global minimum at the same time, slightly sooner for rMF-KG and SF-KG. MF-MES, MF-GIBBON and MF-KG only attain the optimum 15, 7 and 15 points of budget later, respectively. It is worth noticing that a similar experiment was performed by Poloczek et al. (2017), however using a sinusoidal signal with a magnitude 200 times inferior to the mean of the objective function. Applying a more realistic perturbation still produces an informative IS while illustrating the increased robustness brought by rMF-KG with respect to MF-KG (Figure 4, right panel).

7.2. XGBoost hyperparameter tuning

We now assess the performance of rMFBO on a real-world hyperparameter tuning example. To that end, we follow the experiment introduced by Li et al. (2020) and train an XGBoost model (Chen and Guestrin, 2016) to predict a quantitative measure of the diabetes progression[1]. The dataset includes 442 examples, two-thirds are used for training and the remaining for evaluation. We employ the implementation from the scikit-learn library (Pedregosa et al., 2011) and optimize 5 continuous hyperparameters described in Section F. The primary IS m trains XGBoost with 100 weak learners trees, while the auxiliary IS l uses only 10, with $\lambda_1 = 0.1, \lambda_m = 1$. The optimization starts with 10 random queries at each IS. We used the nRMSE to evaluate the performance. rMF-MES is able to take advantage of the auxiliary IS and achieves faster convergence than SF-MES, while staying close to MF-MES (Figure 5). Note that the auxiliary IS is in this case of extremely good quality and comes at a 10-times cheaper cost (Figure 6),

[1] https://archive.ics.uci.edu/ml/datasets/diabetes

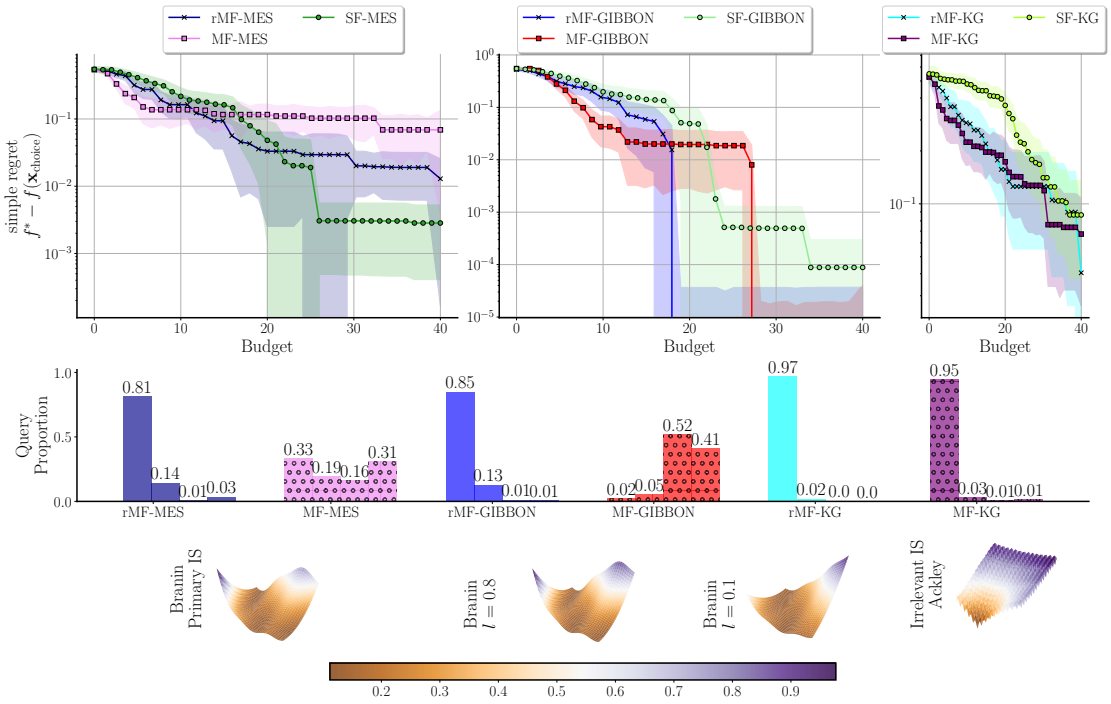

Figure 2: **Top:** simple regret depicted over budget spent in the Branin 2D multi-fidelity problem with 3 auxiliary ISs, averaged over 20 repetitions. **Middle:** distribution of IS queries. The value l refers to the degree of fidelity of the IS. The bars are sorted following the same order at the lower panel. **Bottom:** 2D plot of the available ISs. The primary IS cost is set to 1 and the auxiliary ISs cost to 0.2.

thus demonstrating the ability of rMF-MES to use these cheap queries even though the main purpose of the algorithm is to provide robustness against unreliable sources. On the other hand, rMF-GIBBON displays a behavior closer to SF-GIBBON. While a significant advantage over MF-GIBBON can be observed in the first half of the budget, rMF-GIBBON then gets distanced by its MFBO counterpart.

7.3. Synthetic functions with several auxiliary ISs of varying relevance

Next, we check whether rMFBO is able to distinguish between relevant and irrelevant auxiliary ISs in the presence of multiple auxiliary ISs on 2 examples. In the first problem, the 6D Hartmann function is selected as the primary IS, with 3 auxiliary ISs: Hartmann with degree of fidelity 0.8, 0.1 and the Rosenbrock function. In the second problem, we wish to optimize the 2D Branin function with 3 auxiliary ISs: Branin with degree of fidelity 0.8, 0.1 and the Ackley function. In both settings, the primary IS can be queried for a cost of 1, and all auxiliary ISs for a cost of 0.2. Figure 2 and 7 display the results,

showing that rMFBO provides a consistent decrease of regret compared to MFBO across methods. rMFBO also strongly improves over SFBO in one case (GIBBON - Branin2D), achieves slightly lower regret in three cases (MES and GIBBON - Hartmann2D, KG - Branin2D) and higher regret in the last case (MES - Branin2D). This again demonstrates the relevance of our method in the even more realistic setting where multiple auxiliary ISs are available, some being relevant and others not. Our claim is backed up by the query distribution bar plots, which show how rMFBO avoids querying uninformative sources while still taking advantage of the informative one, except in the case of rMFKG where almost all queries were addressed to the primary IS. Meanwhile, MF-MES queries are scattered across all ISs, independently of their relevance. MF-GIBBON delegates roughly a third of its queries to irrelevant IS, which leads to poor results in the Hartmann problem, but not in the Branin problem, where convergence to the global optimum is still achieved faster than SF-GIBBON.

7.4. Ablation study

We investigated the performances of rMFBO with respect to several changes such as auxiliary IS cost or algorithm hyperparameters variation. The benchmarks are only performed using the MES and GIBBON strategies, due to the high computational overhead of the KG method, more than ten times larger than MES and GIBBON (Moss et al., 2021). We study the same setting described at the end of Section 2.

Variation of the auxiliary IS cost λ_l: Figure 8 shows the evolution of the simple regret as the lower fidelity cost increases ($\lambda_l \in \{0.1, 0.2, 0.5, 0.8\}$). For the informative auxiliary IS case (first two columns), the increase in the cost naturally shifts the behavior of MFBO and rMFBO towards that of SFBO, losing the the acceleration of the convergence in the process. This illustrates the trade-off between informativeness of an auxiliary IS and its query cost. Interestingly, MFBO methods fail to respect that trade-off in an irrelevant auxiliary IS setting. Indeed, MFBO regret eventually flats out for both MES and GIBBON acquisition strategies at high auxiliary IS query cost. For $\lambda_l = 0.8$, MF-MES spent 82% of the budget on auxiliary IS queries on average, MF-GIBBON 42%, over the 20 repetitions. Instead, rMFBO is as expected only slightly positively affected by the cost increase, since the irrelevant auxiliary IS becomes less and less relevant, showing again robustness to uninformative IS.

Variation of rMFBO hyperparameters: Lastly, we investigate how the hyperparameters c_1 and c_2 affect the performance of rMFBO. c_1 constitutes a threshold on the primary IS MF model posterior variance evaluated at SFBO proposal $(\mathbf{x}_t^{\text{pSF}}, m)$, which leads to a primary IS query when exceeded. c_2 measures the information gain provided by the MFBO proposal $(\mathbf{x}_t^{\text{MF}}, l)$ for $l \in [\![m]\!]$, and leads to a primary IS query when not exceeded. We vary $c_1, c_2 \in \{0, 0.1, 0.2\}$ and display the results in Figure 9.

As expected, setting c_1 to 0 (first two rows) essentially reduces rMFBO to SFBO, no

matter the value of c_2. As c_1 increases and $c_2 = 0$ (first column, rows third to sixth), rMFBO quickly transitions to the MFBO dynamics regardless of the relevance of the IS. Then, increasing c_2 provides a reasonable trade-off between robustness to irrelevant auxiliary IS and exploitation of informative auxiliary IS.

8. Conclusions

In this paper, we introduced rMFBO, a building block to any MFBO method to make it robust to unreliable information sources, i.e., of unknown relevance. In particular, we showed that the regret bound of rMFBO can be tied to that of SFBO, with high probability. Upon extensive experiments, we further demonstrated that the current MFBO methods lack this notion of robustness, and how rMFBO was able to successfully fill this gap, while staying competitive when the auxiliary information sources are relevant.

The proposed rMFBO relies on two hyperparameters, c_1 and c_2. While c_1 is theoretically grounded, an heuristic value was picked for c_2, giving satisfactory results and whose soundness was empirically assessed through ablation studies. Nevertheless, gaining theoretical understanding on how this value should be selected is left for future research. Preliminary results suggest the threshold should adapt to the dimension of the problem and the number of BO rounds, as would be expected from an entropy-based measure. From a computational perspective, rMFBO keeps track of two acquisition trajectories, which leads to increased computation times, but negligible compared to the evaluation costs encountered in real-world settings.

Any safety-critical MFBO application can benefit from a methodology like rMFBO, as our algorithm gives guarantees against erroneous or even adversarial information sources. Lastly, rMFBO opens the door to a more systematic inclusion of human experts, with varying knowledge, within BO processes. Typically, these experts would have precise understanding on a specific region of the input domain, but would provide irrelevant feedback elsewhere. Our algorithm makes it possible to take into account these novel information sources with varying degree of fidelity across the input domain, opening exciting opportunities in Bayesian optimization.

References

Baggett, L. W. (1992). *Functional analysis: A primer*. Dekker.

Balandat, M., Karrer, B., Jiang, D., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. (2020). BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21524–21538.

Bogunovic, I., Scarlett, J., Jegelka, S., and Cevher, V. (2018). Adversarially robust optimization with gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Brochu, E., Cora, V. M., and De Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Chen, T. and Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Fröhlich, L., Klenske, E., Vinogradska, J., Daniel, C., and Zeilinger, M. (2020). Noisy-input entropy search for efficient robust Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2262–2272.

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31.

Ghosal, S. and Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 34(5):2413–2429.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276.

Huang, D., Allen, T. T., Notz, W. I., and Miller, R. A. (2006). Sequential kriging optimization using multiple-fidelity evaluations. *Structural and Multidisciplinary Optimization*, 32(5):369–382.

Hvarfner, C., Stoll, D., Souza, A., Lindauer, M., Hutter, F., and Nardi, L. (2022). πBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization. In *International Conference on Learning Representations (ICLR)*.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13:455–492.

Kandasamy, K., Dasarathy, G., Oliva, J. B., Schneider, J., and Poczos, B. (2016). Gaussian Process Bandit Optimisation with Multi-fidelity Evaluations. In *Advances in Neural Information Processing Systems (NIPS)*.

Kirschner, J. and Krause, A. (2021). Bias-Robust Bayesian Optimization via Dueling Bandits. In *International Conference on Machine Learning (ICML)*, pages 5595–5605.

Korovina, K., Xu, S., Kandasamy, K., Neiswanger, W., Poczos, B., Schneider, J., and Xing, E. (2020). ChemBO: Bayesian Optimization of Small Organic Molecules with Synthesizable Recommendations. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3393–3403.

Lam, R., Allaire, D. L., and Willcox, K. E. (2015). Multifidelity optimization using statistical surrogate modeling for non-hierarchical information sources. In *56th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*.

Li, S., Xing, W., Kirby, R., and Zhe, S. (2020). Multi-fidelity Bayesian optimization via deep neural networks. In *Advances in Neural Information Processing Systems*

(NeurIPS), pages 8521–8531.

Marco, A., Berkenkamp, F., Hennig, P., Schoellig, A. P., Krause, A., Schaal, S., and Trimpe, S. (2017). Virtual vs. real: Trading off simulations and physical experiments in reinforcement learning with Bayesian optimization. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1557–1563.

Martinez-Cantin, R., Tee, K., and McCourt, M. (2018). Practical Bayesian optimization in the presence of outliers. In *International conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1722–1731.

Moss, H. B., Leslie, D. S., González, J., and Rayson, P. (2021). GIBBON: General-purpose Information-Based Bayesian Optimisation. *Journal of Machine Learning Research*, 22(235):1–49.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Poloczek, M., Wang, J., and Frazier, P. (2017). Multi-information source optimization. In *Advances in Neural Information Processing Systems (NIPS)*.

Sen, R., Kandasamy, K., and Shakkottai, S. (2018). Multi-fidelity black-box optimization with hierarchical partitions. In *International conference on machine learning (ICML)*, pages 4538–4547.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing systems (NIPS)*.

Song, J., Chen, Y., and Yue, Y. (2019). A general framework for multi-fidelity Bayesian optimization with gaussian processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3158–3167.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. (2012). Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265.

Swersky, K., Snoek, J., and Adams, R. P. (2013). Multi-task Bayesian optimization. In *Advances in Neural Information Processing Systems (NIPS)*.

Takeno, S., Fukuoka, H., Tsukada, Y., Koyama, T., Shiga, M., Takeuchi, I., and Karasuyama, M. (2020). Multi-fidelity Bayesian optimization with max-value entropy search and its parallelization. In *International Conference on Machine Learning (ICML)*, pages 9334–9345.

Wu, J., Toscano-Palmerin, S., Frazier, P. I., and Wilson, A. G. (2019). Practical Multi-fidelity Bayesian Optimization for Hyperparameter Tuning. *arXiv preprint arXiv:1903.04703*.

Zhang, Y., Apley, D. W., and Chen, W. (2020). Bayesian optimization for materials

design with mixed quantitative and qualitative variables. *Scientific reports*, 10(1):1–13.

Zhang, Y., Hoang, T. N., Low, B. K. H., and Kankanhalli, M. (2017). Information-based multi-fidelity Bayesian optimization. In *NIPS Workshop on Bayesian Optimization*.

A. Additional Figures

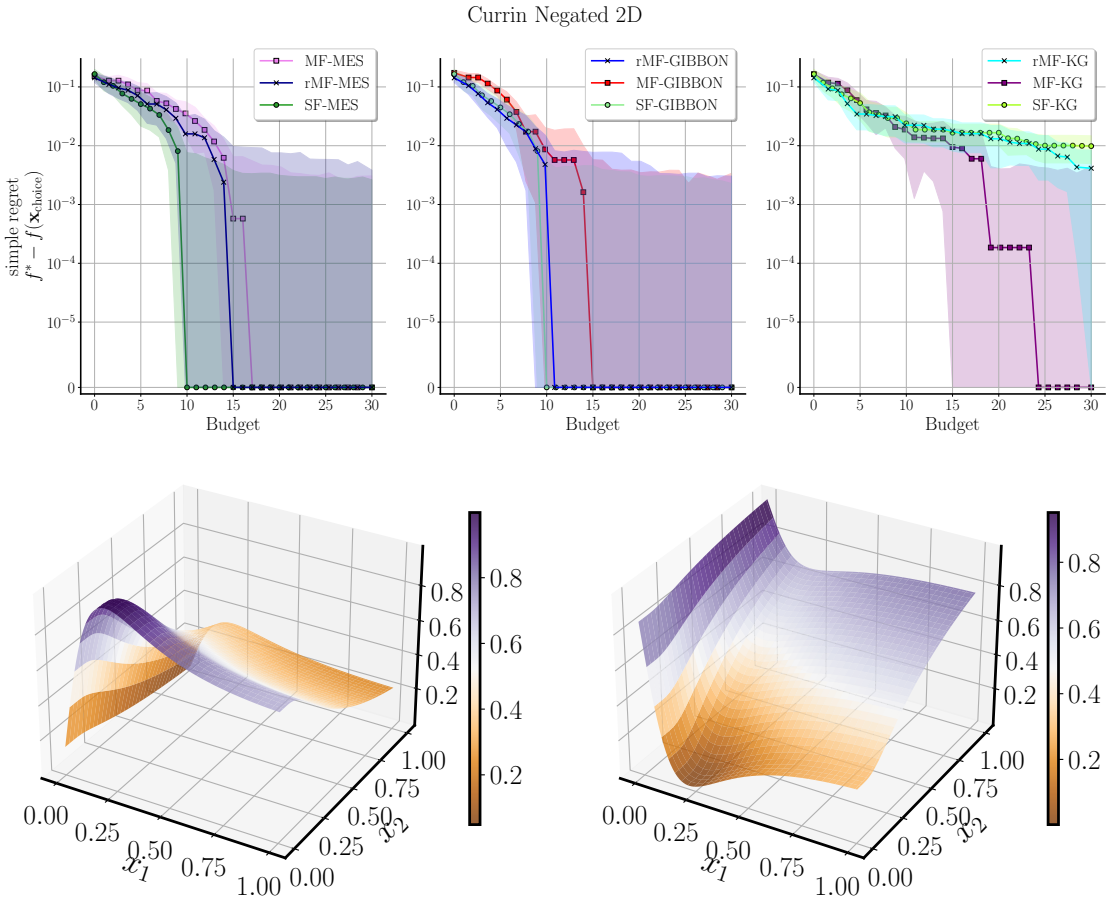

Figure 3: **Top:** simple regret depicted over budget spent for the Negated Currin2D multi-fidelity problem, averaged over 20 repetitions. The primary IS cost is set to 1 and the auxiliary IS cost to 0.1. In the computation of $k_{\text{IS}}(\ell, \ell')$, we used $\ell = 1,\ \ell' = 0.1$ for the primary IS and the auxiliary IS, respectively. **Bottom:** 2D plot of the primary IS (left) and the auxiliary IS (right).

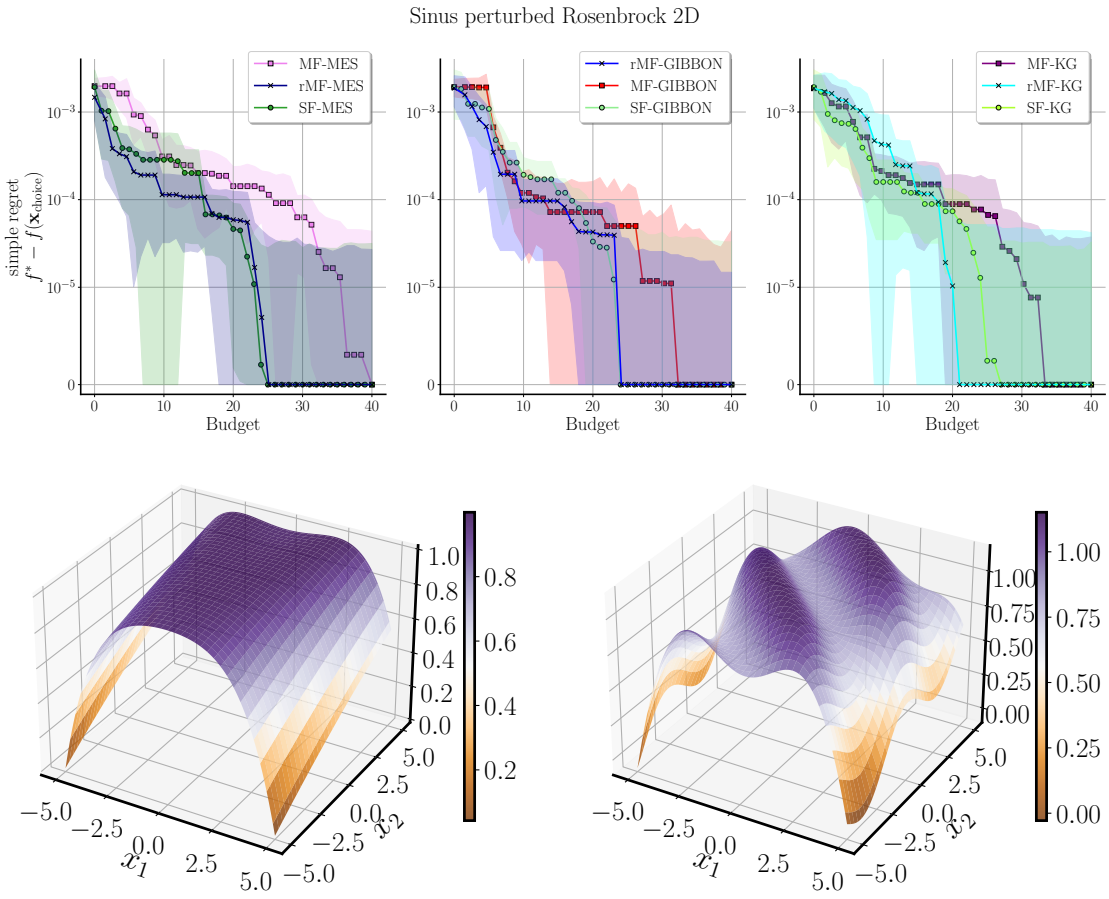

Figure 4: **Top:** simple regret depicted over budget spent for the sinus perturbed Rosenbrock2D multi-fidelity problem, averaged over 20 repetitions. The primary IS cost is set to 1 and the auxiliary IS cost to 0.2. In the computation of $k_{\text{IS}}(\ell, \ell')$, we used $\ell = 1$, $\ell' = 0.2$ for the primary IS and the auxiliary IS, respectively. **Bottom:** 2D plot of the primary IS (left), and its sinus perturbed version used as auxiliary IS (right).

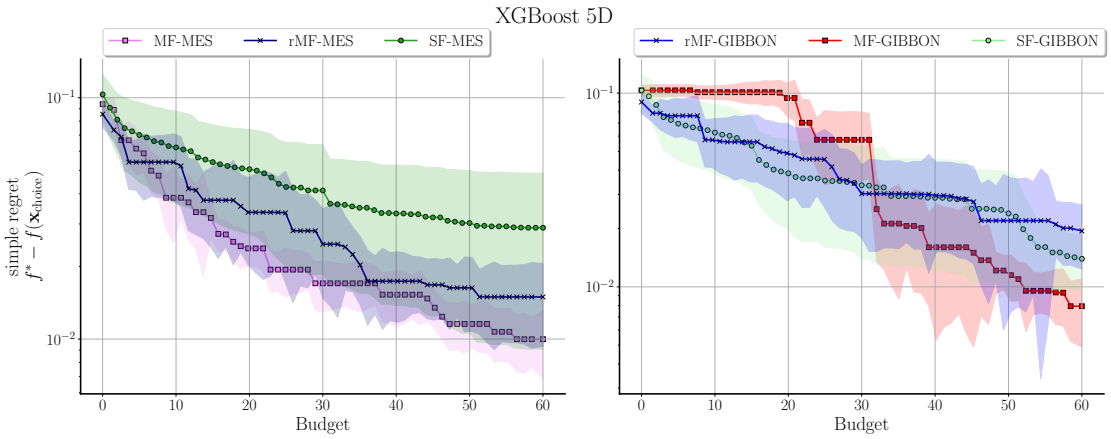

Figure 5: Simple regret depicted over budget spent for the XGBoost hyperparameter tuning multi-fidelity problem, averaged over 20 repetitions. The primary IS cost is set to 1 and the auxiliary IS cost to 0.1. In the computation of $k_{\text{IS}}(\ell, \ell')$, we used $\ell = 1,\ \ell' = 0.1$ for the primary IS and the auxiliary IS, respectively. For the simple regret computation, f^* has been obtained using 30000 evaluations of the primary IS at random points.

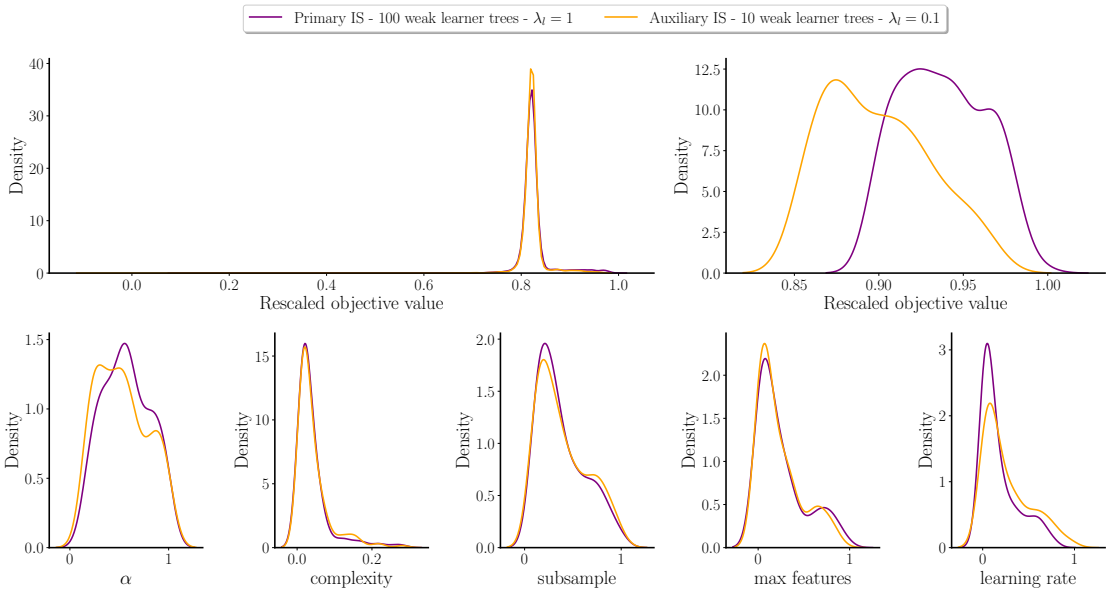

Figure 6: Distribution of rescaled objective values and best samples for the XGBoost 5D hyperparameter tuning problem. For each IS, distributions are computed using 3000 random uniform samples within hyperparameter bounds and a kernel density estimator. **Top left:** the whole rescaled objective values distribution. **Top right:** density plot of the 5% best values. **Bottom:** density plot of the hyperparameters samples associated with 5% best values, demonstrating the strong agreement between the primary and the auxiliary IS.

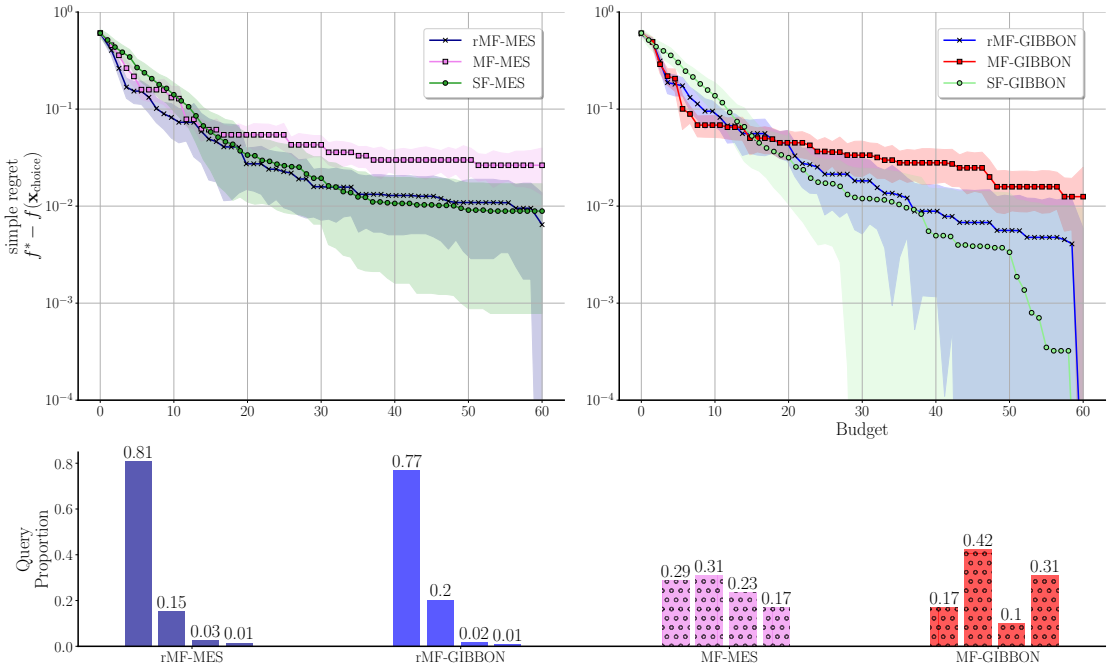

Figure 7: **Top:** simple regret depicted over budget spent in the Hartmann6D multi-fidelity problem with 3 auxiliary ISs, averaged over 20 repetitions. **Bottom:** distribution of IS queries. From left to right, the bars are sorted in the following order: Hartmann the primary IS, Hartmann with degree of fidelity $l = 0.8$, Hartmann with $l = 0.1$ and finally the Rosenbrock function (see Section F). In the computation of $k_{\text{IS}}(\ell, \ell')$, we used $\ell = 1,\ \ell' = 0.8,\ \ell'' = 0.1$ and $\ell''' = 0$ for the primary IS, the Hartmann with 0.8 bias IS, the Hartmann with 0.1 bias and the rosenbrock function, respectively. The primary IS cost is set to 1 and the auxiliary ISs cost to 0.2.

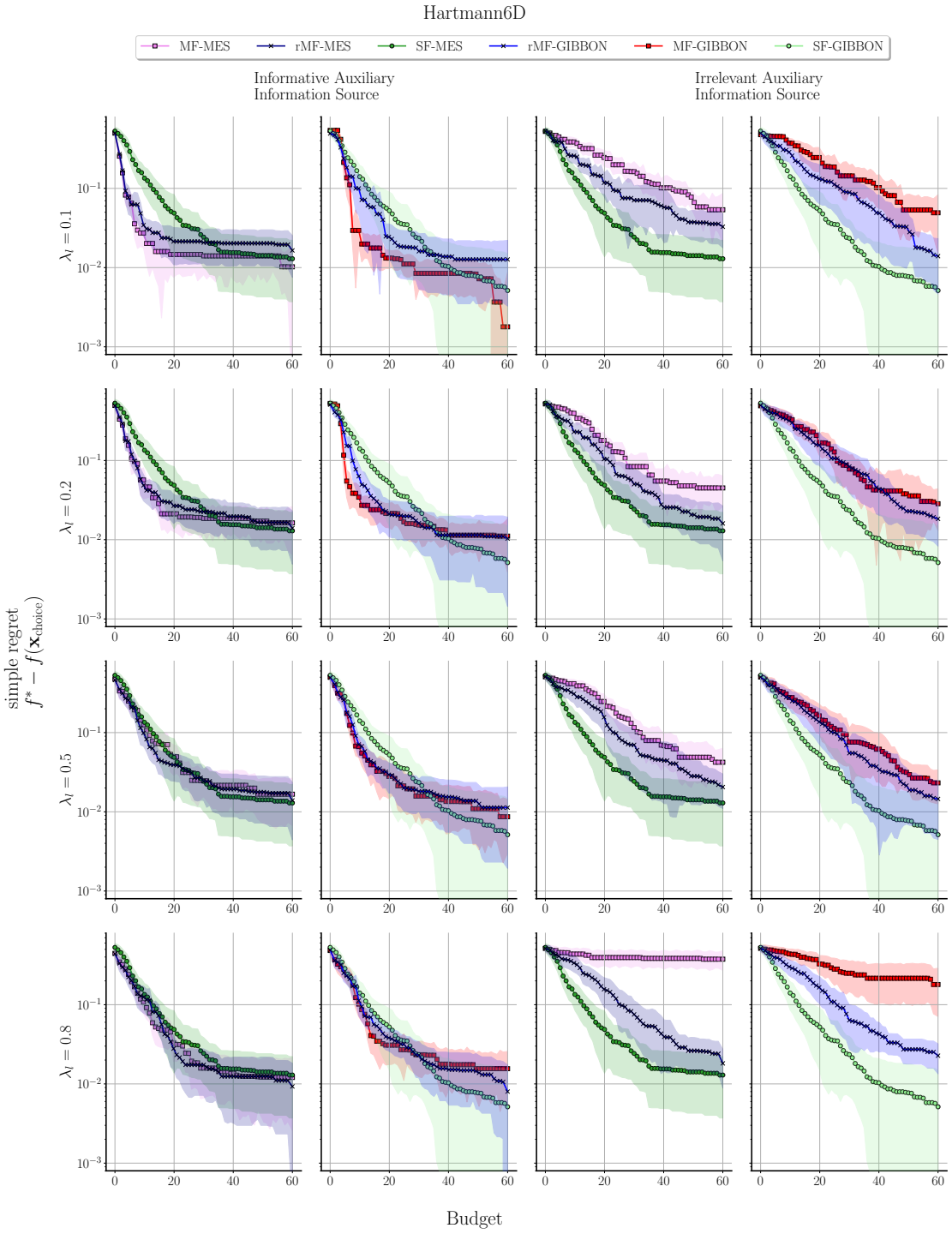

Figure 8: Simple regret depicted over budget spent for the Hartmann6D multi-fidelity problem, averaged over 20 repetitions. For each row, the auxiliary IS cost is varied. In the computation of $k_{\text{IS}}(\ell, \ell')$, we used $\ell = 1,\ \ell' = 0.2$ for the primary IS and the auxiliary IS, respectively.

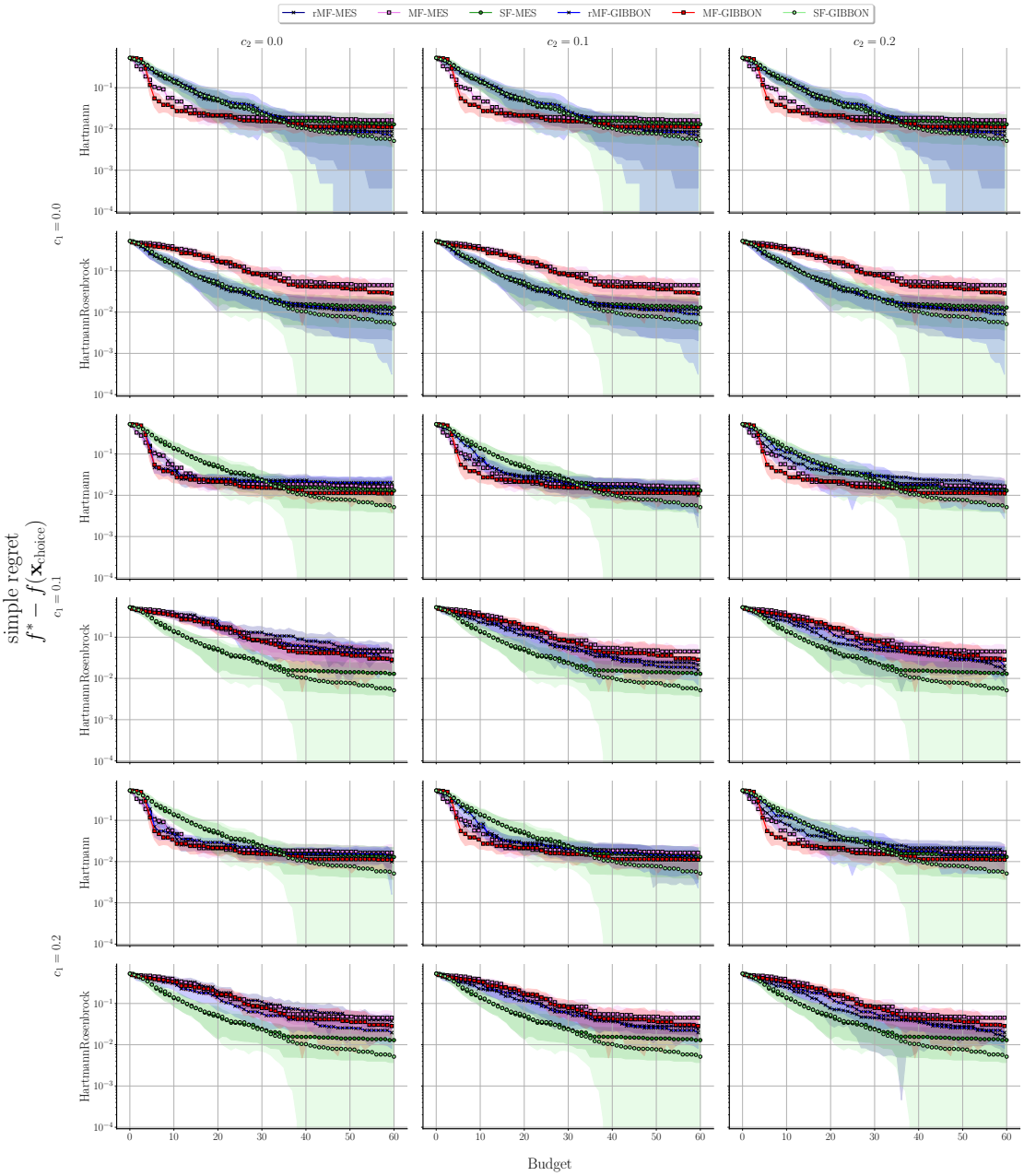

Figure 9: Simple regret depicted over budget spent for the Hartmann6D multi-fidelity problem, averaged over 20 repetitions. At the row level, the hyperparameter c_1 is varied, with even rows being the Hartmann/Hartmann0.2 (relevant IS) problem, while odd rows consider the Hartmann/HartmannRosenbrock (irrelevant IS) problem. At the column level, the hyperparameter c_2 is varied. In the computation of $k_{\text{IS}}(\ell, \ell')$, we used $\ell = 1,\ \ell' = 0.2$ for the primary IS and the auxiliary IS, respectively.

B. Additional proofs

B.1. Noiseless scenario: Proposition 1

Proof of Proposition 1. In the proof, we simplify the notations and denote the pseudo SFBO query at round t by $\mathbf{x}_t = \mathbf{x}_t^{\mathrm{pSF}}$, and $\mathcal{D}_t = \mathcal{D}_t^{\mathrm{pSF}}$ for the pseudo SFBO dataset. The SFBO query at round t is denoted by $\mathbf{x}_t^{\mathrm{SF}}$, as earlier. The acquisition function is treated as a function of an input-IS pair $(\mathbf{x}, \ell)$ and the dataset $\mathcal{D}$. Let us think the dataset as a $t(d+1)$-dimensional vector $\mathcal{D}_t = (x_1^{(1)}, \cdots, x_1^{(d)}, \cdots, x_t^{(1)}, \cdots, x_t^{(d)}, y_1, ..., y_t)$. Let us use a shorthand notation $\alpha(\mathbf{x}, m, \mathcal{D}) = \alpha(\mathbf{x}, \mathcal{D})$, and consider the next primary IS query

$$\mathbf{x}_{t+1} = \arg\max_{\mathbf{x} \in \mathcal{X}} \alpha(\mathbf{x}, \mathcal{D}_t) \tag{11}$$

which defines an implicit function $\mathbf{x}_{t+1}(\mathcal{D}_t)$ such that $\mathbf{x}_{t+1}$ solves $\nabla_{\mathbf{x}} \alpha(\mathbf{x}, \mathcal{D}_t) = \mathbf{0}_{\mathbb{R}^d}$. By applying the implicit function theorem to the continuously differentiable function $\nabla_{\mathbf{x}} \alpha : \mathcal{X} \times (\mathcal{X} \times \mathbb{R})^t \to \mathbb{R}^d$ (Assumption 5) with invertible Jacobian (Assumption 6), the rate of change of the next query with respect to the dataset can be defined as,

$$\left\| \frac{\partial \mathbf{x}_{t+1}}{\partial \mathcal{D}_t} (\mathcal{D}_t) \right\|_{\mathrm{op}} = \left\| \left[(\mathbf{H}_{\alpha,\mathbf{x}}(\mathbf{x}_{t+1}(\mathcal{D}_t), \mathcal{D}_t))^{-1} \frac{\partial \nabla_{\mathbf{x}} \alpha}{\partial \mathcal{D}_t^{(k)}} (\mathbf{x}_{t+1}(\mathcal{D}_t), \mathcal{D}_t) \right]_{k=1}^{t(d+1)} \right\|_{\mathrm{op}} \tag{12}$$

where $\|A\|_{\mathrm{op}} := \inf\{c > 0 : \|A\mathbf{x}\|_\infty \leq c \|\mathbf{x}\|_\infty \ \forall \mathbf{x} \in \mathcal{X}\}$ is the operator norm, and $\mathbf{H}_{\alpha,\mathbf{x}}$ denotes the Hessian of $\mathbf{x} \mapsto \alpha(\mathbf{x}, \mathcal{D}_t)$. Specifically, the $(i,j)^{\mathrm{th}}$-element of $\mathbf{H}_{\alpha,\mathbf{x}}$ reads as $\frac{\partial^2 \alpha}{\partial x_i \partial x_j}$. The i^{th}-element of $\frac{\partial \nabla_{\mathbf{x}} \alpha}{\partial \mathcal{D}_t^{(k)}}$ is $\frac{\partial^2 \alpha}{\partial \mathcal{D}_t^{(k)} \partial x_i}$, which denotes the partial derivative w.r.t. the first and the second variable of $\alpha(\cdot, \cdot)$.

We now show the proposition by induction.

1° Base case $t = 1$: The claim follows by the design of Algorithm, since $\mathbf{x}_1^{\mathrm{SF}} = \mathbf{x}_1$.

2° Induction step: Let us assume that $t \in \{1, ..., T-2\}$. The outcome vector of the SFBO algorithm is $\mathbf{y}_t^{SF} = (y_1^{\mathrm{SF}}, ..., y_t^{\mathrm{SF}})$. The outcome vector of the pseudo-SFBO algorithm, $\mathbf{y}_t = (y_1, ..., y_t)$, consists of observations $y_\tau = f^{(m)}(\mathbf{x}_\tau)$ and pseudo-observations $y_\tau = \mu_{\mathrm{MF}}(\mathbf{x}_\tau, m)$.

Consider the mapping $\mathbf{x}_{t+1} : (\mathcal{X} \times \mathbb{R})^t \to \mathcal{X}$ with the domain and codomain equipped with the sup-norms. By the above-mentioned implicit function theorem, there exists a neighbourhood such that $\mathbf{x}_{t+1}$ is continuously differentiable with bounded derivatives. Since $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, the spaces $((\mathcal{X} \times \mathbb{R})^t, \|\cdot\|_\infty)$ and $(\mathcal{X}, \|\cdot\|_\infty)$ are Banach spaces. Thus, by the mean value inequality on Banach spaces (Baggett, 1992, Theorem 12.6), if we consider the closed line segment joining $\mathcal{D}_t$ to $\mathcal{D}_t^{\mathrm{SF}}$ (using the convexity of $\mathcal{X}$), and define M_t to be the maximum rate of change over the line segment,

$$M_t := \max_{\mathcal{D} \in \mathbb{D}_t} \left\| \frac{\partial \mathbf{x}_{t+1}}{\partial \mathcal{D}} (\mathcal{D}) \right\|_{\mathrm{op}}, \quad \text{where } \mathbb{D}_t = \left\{ \mathcal{D} \ \middle| \ \mathcal{D} = (1-s)\mathcal{D}_t + s\mathcal{D}_t^{\mathrm{SF}}, \ s \in [0,1] \right\} \tag{13}$$

we can bound the closeness of the next queries at round t,

$$\left\|\mathbf{x}_{t+1}^{\mathrm{SF}} - \mathbf{x}_{t+1}\right\|_\infty = \left\|\mathbf{x}_{t+1}(\mathcal{D}_t^{\mathrm{SF}}) - \mathbf{x}_{t+1}(\mathcal{D}_t)\right\|_\infty \leq \left\|\mathcal{D}_t^{\mathrm{SF}} - \mathcal{D}_t\right\|_\infty M_t.$$

It remains to bound $\left\|\mathcal{D}_t^{\mathrm{SF}} - \mathcal{D}_t\right\|_\infty$. First, observe that,

$$\left\|\mathcal{D}_t^{\mathrm{SF}} - \mathcal{D}_t\right\|_\infty = \max\left\{\left\|\mathbf{x}_1^{\mathrm{SF}} - \mathbf{x}_1\right\|_\infty, ..., \left\|\mathbf{x}_t^{\mathrm{SF}} - \mathbf{x}_t\right\|_\infty, \left|y_1^{\mathrm{SF}} - y_1\right|, ..., \left|y_t^{\mathrm{SF}} - y_t\right|\right\}.$$

Let us consider $|y_\tau^{\mathrm{SF}} - y_\tau|$ for any $\tau \in \{1, ..., t\}$. It holds that,

$$|y_\tau^{\mathrm{SF}} - y_\tau| = \begin{cases} |f(\mathbf{x}_\tau^{\mathrm{SF}}) - f(\mathbf{x}_\tau)|, & \text{if Line 8 of Algorithm 1 false at query } \tau \\ |f(\mathbf{x}_\tau^{\mathrm{SF}}) - \mu_\tau(\mathbf{x}_\tau)|, & \text{if Line 8 true at query } \tau \end{cases}$$

where we use the shorthand notations $f(\mathbf{x}) := f(\mathbf{x}, m)$ and $\mu_{\tau,\mathrm{MF}}(\mathbf{x}) := \mu_{\mathrm{MF}}(\mathbf{x}, m \mid \mathcal{D}_\tau)$.

For the false case (real observation), we have

$$\begin{aligned} |y_\tau^{SF} - y_\tau| &= |f(\mathbf{x}_\tau^{\mathrm{SF}}) - f(\mathbf{x}_\tau)| \\ &\leq \sqrt{d}\left\|\mathbf{x}_\tau^{\mathrm{SF}} - \mathbf{x}_\tau\right\|_2 && (14) \\ &\leq d\left\|\mathbf{x}_\tau^{\mathrm{SF}} - \mathbf{x}_\tau\right\|_\infty && (15) \\ &\leq d\varepsilon\tau\hat{M}_\tau d^\tau && (16) \\ &= \varepsilon\tau\hat{M}_\tau d^{\tau+1} \end{aligned}$$

with probability greater than $1 - da\exp(-\frac{1}{b^2})$ by Assumption 4 exploited in (14). The inequalities (15) and (16) follow from the equivalence of the norms ($\|\mathbf{x}\|_2 \leq \sqrt{d}\|\mathbf{x}\|_\infty$) and the induction hypothesis, respectively. For the true case (pseudo-observations), as before we have

$$\begin{aligned} |y_\tau^{\mathrm{SF}} - y_\tau| &\leq |f(\mathbf{x}_\tau^{\mathrm{SF}}) - \mu_{\tau,\mathrm{MF}}(\mathbf{x}_\tau)| \\ &\leq |f(\mathbf{x}_\tau^{SF}) - f(\mathbf{x}_\tau)| + |f(\mathbf{x}_\tau) - \mu_{\tau,\mathrm{MF}}(\mathbf{x}_\tau)| \\ &\leq \varepsilon\tau\hat{M}_\tau d^{\tau+1} + |f(\mathbf{x}_\tau) - \mu_{\tau,\mathrm{MF}}(\mathbf{x}_\tau)| \end{aligned}$$

with probability greater than $1 - da\exp(-\frac{1}{b^2})$. The first term represents the error for being off from the single-fidelity algorithm acquisition track, and the second term is the prediction error of the MOGP surrogate model. Given that the objective f is drawn from a MOGP with same covariance kernel than that of the MOGP surrogate in the Algorithm (Assumption 1), the latter term can be bounded. For any constant $C > 0$, at round t,

$$\begin{aligned} \mathrm{P}\left(\frac{f(\mathbf{x}) - \mu_{t,\mathrm{MF}}(\mathbf{x})}{\sigma_{t,\mathrm{MF}}(\mathbf{x})} > C\right) &\leq \frac{1}{2}\exp\left(-\frac{C^2}{2}\right) \\ \mathrm{P}\left(|f(\mathbf{x}) - \mu_{t,\mathrm{MF}}(\mathbf{x})| > C\sigma_{t,\mathrm{MF}}(\mathbf{x})\right) &\leq \exp\left(-\frac{C^2}{2}\right) \\ \mathrm{P}\left(|f(\mathbf{x}) - \mu_{t,\mathrm{MF}}(\mathbf{x})| \leq C\sigma_{t,\mathrm{MF}}(\mathbf{x})\right) &\geq 1 - \exp\left(-\frac{C^2}{2}\right). \end{aligned}$$

Pick $C = \frac{\varepsilon}{\sigma_{t,\mathrm{MF}}(\mathbf{x})}$. Then, we know that $|f(\mathbf{x}) - \mu_{t,\mathrm{MF}}(\mathbf{x})| \leq \varepsilon$ holds at least with probability $1 - \exp\left(-\frac{\varepsilon^2}{2\sigma^2_{t,\mathrm{MF}}(\mathbf{x})}\right)$.

If $\sigma_t(\mathbf{x}) \leq \frac{\varepsilon}{\sqrt{-2\log(1-q)}}$, then $|f(\mathbf{x}) - \mu_{t,\mathrm{MF}}(\mathbf{x})| \leq \varepsilon$ holds with probability greater than q. Therefore, $|f(\mathbf{x}_\tau) - \mu_{\tau,\mathrm{MF}}(\mathbf{x}_\tau)| \leq \varepsilon$, and

$$|y_\tau^{\mathrm{SF}} - y_\tau| \leq \varepsilon\tau\hat{M}_\tau d^{\tau+1} + \varepsilon = \varepsilon(\tau\hat{M}_\tau d^{\tau+1} + 1).$$

By combining the results, we have

$$\begin{aligned}
\left\|\mathcal{D}_t^{\mathrm{SF}} - \mathcal{D}_t\right\|_\infty &= \max\left\{\left\|\mathbf{x}_1^{\mathrm{SF}} - \mathbf{x}_1\right\|_\infty, \cdots, \left\|\mathbf{x}_t^{\mathrm{SF}} - \mathbf{x}_t\right\|_\infty, \left|y_1^{\mathrm{SF}} - y_1\right|, \cdots, \left|y_t^{\mathrm{SF}} - y_t\right|\right\} \\
&\leq \max\left\{\varepsilon M_0 d, \cdots, \varepsilon t d^t \max_{S\in 2^{[\![t-1]\!]}} \prod_{k\in S} M_k, \varepsilon(M_0 d + 1), \cdots, \varepsilon(t d^{t+1} \max_{S\in 2^{[\![t-1]\!]}} \prod_{k\in S} M_k + 1)\right\} \\
&= \varepsilon\left(t d^{t+1} \max_{S\in 2^{[\![t-1]\!]}} \prod_{k\in S} M_k + 1\right)
\end{aligned}$$

with probability greater than $q\left(1 - da\exp(-\frac{1}{b^2})\right)$. Note that the event $|f(\mathbf{x}) - \mu_{t,\mathrm{MF}}(\mathbf{x})| \leq \varepsilon$ and the event in Assumption 4 are independent given the assumptions. For all $t \in \{1, ..., T-2\}$,

$$\begin{aligned}
\left\|\mathbf{x}_{t+1}^{\mathrm{SF}} - \mathbf{x}_{t+1}\right\|_\infty &\leq \varepsilon\left(t d^{t+1} \max_{S\in 2^{[\![t-1]\!]}} \prod_{k\in S} M_k + 1\right) M_t \\
&\leq \varepsilon\left(t M_t d^{t+1} \max_{S\in 2^{[\![t-1]\!]}} \prod_{k\in S} M_k + d^{t+1} \max_{S\in 2^{[\![t]\!]}} \prod_{k\in S} M_k\right) \\
&= \varepsilon d^{t+1}\left(t \max_{S\in 2^{[\![t]\!]}} \prod_{k\in S} M_k + \max_{S\in 2^{[\![t]\!]}} \prod_{k\in S} M_k\right) \\
&= \varepsilon(t+1)\hat{M}_{t+1} d^{t+1}
\end{aligned}$$

holds with probability greater than $q\left(1 - da\exp(-\frac{1}{b^2})\right)$. □

B.2. Noisy scenario : Proposition 1

We consider a noisy scenario, that is, $\sigma_{\mathrm{noise}} > 0$. It can be shown that Proposition 1 holds if $\frac{\sqrt{\sigma_{\mathrm{noise}}}}{\varepsilon} \leq (d^{t+1}\hat{M}_t - 1)/2$ (with a negligible lower probability, specifically a factor of $\mathrm{Erf}\left(\frac{1}{2\sqrt{\sigma_{\mathrm{noise}}}}\right)\mathrm{Erf}\left(\frac{1}{\sqrt{2\sigma_{\mathrm{noise}}}}\right)$ lower). Given empirical study on values $\hat{M}_t$ (Appendix D), it is highly unlikely that this condition does not hold with reasonable values for ε and σ_{noise}.

Proof. Proof B.1 should be modified as follows.

Let us consider $|y_\tau^{\mathrm{SF}} - y_\tau|$ for any $\tau \in \{1, ..., t\}$. It holds that,

$$|y_\tau^{\mathrm{SF}} - y_\tau| = \begin{cases} |f(\mathbf{x}_\tau^{\mathrm{SF}}) + \epsilon - f(\mathbf{x}_\tau) - \epsilon'|, & \text{if Line 8 of Algorithm 1 false at query } \tau \\ |f(\mathbf{x}_\tau^{\mathrm{SF}}) + \epsilon - \mu_{\tau,\mathrm{MF}}(\mathbf{x}_\tau, m)|, & \text{if Line 8 false at query } \tau. \end{cases}$$

Note that $|\epsilon - \epsilon'|$ follows a half-normal distribution with scale parameter $\sqrt{2}\sigma_{\mathrm{noise}}$, and $|\epsilon|$ follows a half-normal distribution with scale parameter σ_{noise}. This implies that $P(|\epsilon - \epsilon'| \leq \sqrt{\sigma_{\mathrm{noise}}}) = \mathrm{Erf}\left(\frac{1}{2\sqrt{\sigma_{\mathrm{noise}}}}\right)$ and $P(|\epsilon| \leq \sqrt{\sigma_{\mathrm{noise}}}) = \mathrm{Erf}\left(\frac{1}{\sqrt{2\sigma_{\mathrm{noise}}}}\right)$.

For the false case (real observation), we have

$$\begin{aligned} |y_\tau^{\mathrm{SF}} - y_\tau| &= |f(\mathbf{x}_\tau^{\mathrm{SF}}) + \epsilon - f(\mathbf{x}_\tau) - \epsilon'| \\ &\leq |f(\mathbf{x}_\tau^{\mathrm{SF}}) - f(\mathbf{x}_\tau)| + |\epsilon - \epsilon'| \\ &\leq \frac{1}{\sqrt{d}} \left\|\mathbf{x}_\tau^{\mathrm{SF}} - \mathbf{x}_\tau\right\| + \sqrt{\sigma_{\mathrm{noise}}} \\ &\leq \frac{1}{\sqrt{d}} \sqrt{d}\varepsilon\tau\hat{M}_\tau + \sqrt{\sigma_{\mathrm{noise}}} \\ &= \varepsilon\tau\hat{M}_\tau + \sqrt{\sigma_{\mathrm{noise}}} \end{aligned}$$

with probability greater than $\mathrm{Erf}\left(\frac{1}{2\sqrt{\sigma_{\mathrm{noise}}}}\right)\left(1 - da\exp(-\frac{1}{db^2})\right)$ by Assumption 4. The last two inequalities follow from the induction hypothesis and the equivalence of the norms, $\|\mathbf{x}\| \leq \sqrt{d}\,\|\mathbf{x}\|_\infty$.

For the true case (pseudo-observation), as before we have,

$$\begin{aligned} |y_\tau^{SF} - y_\tau| &\leq |f(\mathbf{x}_\tau^{SF}) - \mu(\mathbf{x}_\tau)| + |\epsilon| \\ &\leq |f(\mathbf{x}_\tau^{SF}) - f(\mathbf{x}_\tau)| + |f(\mathbf{x}_\tau) - \mu(\mathbf{x}_\tau)| + \sqrt{\sigma_{\mathrm{noise}}} \\ &\leq \varepsilon\tau\hat{M}_\tau + |f(\mathbf{x}_\tau) - \mu(\mathbf{x}_\tau)| + 2\sqrt{\sigma_{\mathrm{noise}}} \\ &\leq \varepsilon\tau\hat{M}_\tau + \varepsilon + 2\sqrt{\sigma_{\mathrm{noise}}} \\ &= \varepsilon(\tau\hat{M}_\tau + 1 + \frac{2\sqrt{\sigma_{\mathrm{noise}}}}{\varepsilon}) \end{aligned}$$

with probability greater than $\mathrm{Erf}\left(\frac{1}{2\sqrt{\sigma_{\mathrm{noise}}}}\right)\mathrm{Erf}\left(\frac{1}{\sqrt{2\sigma_{\mathrm{noise}}}}\right)\left(1 - da\exp(-\frac{1}{db^2})\right)$.

Hence, in this case we have

$$
\begin{aligned}
&\left\|\mathcal{D}_t^{\mathrm{SF}} - \mathcal{D}_t\right\|_\infty \\
&= \max\left\{\left\|\mathbf{x}_1^{\mathrm{SF}} - \mathbf{x}_1\right\|_\infty, ..., \left\|\mathbf{x}_t^{\mathrm{SF}} - \mathbf{x}_t\right\|_\infty, \left|y_1^{\mathrm{SF}} - y_1\right|, ..., \left|y_t^{\mathrm{SF}} - y_t\right|\right\} \\
&\leq \max\left\{\varepsilon M_0 d, \cdots, \varepsilon t d^t \max_{S\in 2^{[\![t-1]\!]}} \prod_{k\in S} M_k, \varepsilon(M_0 d + 1 + \frac{2\sqrt{\sigma_{\mathrm{noise}}}}{\varepsilon}), \cdots,\right. \\
&\qquad\left. \varepsilon\left(t d^{t+1} \max_{S\in 2^{[\![t-1]\!]}} \prod_{k\in S} M_k + 1 + \frac{2\sqrt{\sigma_{\mathrm{noise}}}}{\varepsilon}\right)\right\} \\
&= \varepsilon\left(t d^{t+1} \max_{S\in 2^{[\![t-1]\!]}} \prod_{k\in S} M_k + 1 + \frac{2\sqrt{\sigma_{\mathrm{noise}}}}{\varepsilon}\right)
\end{aligned}
$$

with probability greater than $\mathrm{Erf}\left(\frac{1}{2\sqrt{\sigma_{\mathrm{noise}}}}\right)\mathrm{Erf}\left(\frac{1}{\sqrt{2\sigma_{\mathrm{noise}}}}\right)\left(1 - da\exp(-\frac{1}{db^2})\right)q$.

For all $t \in \{1, ..., T-2\}$,

$$
\begin{aligned}
\left\|\mathbf{x}_{t+1}^{\mathrm{SF}} - \mathbf{x}_{t+1}\right\|_\infty &\leq \varepsilon\left(t d^{t+1} \max_{S\in 2^{[\![t-1]\!]}} \prod_{k\in S} M_k + 1 + \frac{2\sqrt{\sigma_{\mathrm{noise}}}}{\varepsilon}\right) M_t \\
&\leq \varepsilon\left(t M_t d^{t+1} \max_{S\in 2^{[\![t-1]\!]}} \prod_{k\in S} M_k + d^{t+1} \max_{S\in 2^{[\![t]\!]}} \prod_{k\in S} M_k\right) \\
&= \varepsilon d^{t+1}\left(t \max_{S\in 2^{[\![t]\!]}} \prod_{k\in S} M_k + \max_{S\in 2^{[\![t]\!]}} \prod_{k\in S} M_k\right) \\
&= \varepsilon(t+1)\hat{M}_{t+1} d^{t+1}
\end{aligned}
$$

holds with probability greater than $\mathrm{Erf}\left(\frac{1}{2\sqrt{\sigma_{\mathrm{noise}}}}\right)\mathrm{Erf}\left(\frac{1}{\sqrt{2\sigma_{\mathrm{noise}}}}\right)\left(1 - da\exp(-\frac{1}{db^2})\right)q$, when $1 + \frac{2\sqrt{\sigma_{\mathrm{noise}}}}{\varepsilon} \leq d^{t+1}\hat{M}_t$. Specifically, when $\frac{\sqrt{\sigma_{\mathrm{noise}}}}{\varepsilon} \leq (d^{t+1}\hat{M}_t - 1)/2$.

□

B.3. Theorem 1

Proof of Theorem 1. First, note that for any choice (simple or Bayes optimal) it holds,

$$R(\mathbf{x}_{\mathrm{choice}}^{\mathrm{SF}}) - R(\mathbf{x}_{\mathrm{choice}}^{\mathrm{rMF}}) = f(\mathbf{x}_{\mathrm{choice}}^{\mathrm{rMF}}) - f(\mathbf{x}_{\mathrm{choice}}^{\mathrm{SF}}).$$

For the simple choice, we have $\mathbf{x}_{\mathrm{choice}}^{\mathrm{SF}} = \arg\max_{t\in[\![T]\!]} f(\mathbf{x}_t^{\mathrm{SF}})$, and $\mathbf{x}_{\mathrm{choice}}^{\mathrm{rMF}} = \arg\max_{t\in[\![T^{(m)}]\!]} f(\mathbf{x}_t)$ where $\mathbf{x}_1, ..., \mathbf{x}_{T^{(m)}}$ is the primary IS acquisition sequence returned by Algorithm 1 (pseudo-queries removed from the output sequence). With a slight abuse of notation we write T and $T^{(m)}$ for both the number of queries (Definition 2) and the corresponding index sets (e.g. $t \in T^{(m)}$ means that y_t is not a pseudo-observation).

For all $t \in T$, it holds that $f(\mathbf{x}^{\text{rMF}}_{\text{choice}}) - f(\mathbf{x}^{\text{SF}}_{\text{choice}}) > -\varepsilon T \hat{M}_T d^{T+1}$ by Corollary 1 with probability greater than $q\left(1 - da\exp(-\frac{1}{b^2})\right)$. The problem is that the values y_t for $t \in T \setminus T^{(m)}$ are never observed, and we cannot take minimum over these "NaN values" (i.e. $\arg\max$ is not well-defined) in the computation of $\mathbf{x}^{\text{rMF}}_{\text{choice}}$. To solve this issue, a quantity λ_m was saved from the budget Λ (Algorithm 1, Lines 24-26), thus ensuring that if the true maximizer is one of the pseudo-observations, then it will be queried at primary IS, leading to an actual observation.

Specifically, for the last query at $T+1$. For any $\mathbf{x} \in S$, it holds that $P(|f(\mathbf{x}) - \mu_T(\mathbf{x})| \leq \varepsilon) \geq q$ (see Proof B.1). Thus,

$$|\max_{\mathbf{x}\in S} f(\mathbf{x}) - f(\arg\max_{x\in S} \mu_{T,\text{MF}}(\mathbf{x}))| \leq 2\varepsilon,$$

with probability greater than q, and $S = \{\mathbf{x} \in \mathcal{X} \mid \sigma_{\text{MF}}(\mathbf{x}, m) \leq c_1\}$. □

C. Full version of the algorithm

Some modifications can be done to improve the empirical performance of Algorithm 1 while all the theoretical results of Section 6 still hold.

Posterior mean update of the pseudo observations: Lines 12-13 of Algorithm 1 are for simplicity, the algorithm can be made more efficient by adjusting these. All the pseudo-observations in $\mathcal{D}_{sf}$ can be updated to correspond to the most recent predictive mean estimate of the joint surrogate model at the current round t. This does not break the condition $\sigma_{\text{MF}}(\mathbf{x}^{\text{pSF}}_t, m) \leq c_1$, since the posterior variance cannot increase as new data is added. We go further and also check whether the single-fidelity GP surrogate can provide a more accurate estimate of the pseudo-observation in the sense of the accuracy of a nearest neighbor. For the pseudo-observation, we choose the most recent predictive mean estimate of the single-fidelity surrogate model, if $f(\mathbf{x}^{nn}_t, m)$ is closer to $\mu_{\text{SF}}(\mathbf{x}^{\text{pSF}}_t)$ than $\mu_{\text{MF}}(\mathbf{x}^{\text{pSF}}_t, m)$, where $\mathbf{x}^{nn}_t$ is the nearest neighbor of $\mathbf{x}^{\text{pSF}}_t$ in the primary IS training data.

Multiple auxiliary IS relevance check: When the number of auxiliary ISs is more than two, the algorithm can give a chance also to other auxiliary IS, even if the first proposed query $\mathbf{x}^{\text{MF}}_t$ at IS ℓ_t is considered irrelevant by the algorithm. Looping over all IS, and checking their relevance, does not violate the conditions in Algorithm 1, so the theoretical results are preserved.

The pseudo code of the full algorithm is presented in Algorithm 2. Blue lines correspond

to addition w.r.t. the first improvement, red lines to the second.

Algorithm 2: Full version of robust MFBO algorithm

Input: Budget Λ, costs $(\lambda_1, ..., \lambda_m)$, acquisition function α, hyperparameters c_1 and c_2, relevance measure s
Initialize $\mathcal{D}^{\text{pSF}}, \mathcal{D}^{\text{MF}}$
Perform Bayesian updates $\mu_{\text{SF}}, \sigma_{\text{SF}}, \mu_{\text{MF}}, \sigma_{\text{MF}}$
$p_{\text{obs}} \leftarrow \{\}$
$t \leftarrow 1$
while $\lfloor \Lambda/\lambda_m \rfloor \geq 2\lambda_m$ **do**
  $\mathbf{x}_t^{\text{pSF}} \leftarrow \arg\max_{\mathbf{x}} \alpha(\mathbf{x}, m \mid \mu_{\text{SF}}, \sigma_{\text{SF}})$
  condition1 $\leftarrow \sigma_{\text{MF}}(\mathbf{x}_t^{\text{pSF}}, m) \leq c_1$
  condition2 $\leftarrow$ False
  if condition1 **then**
    $(\mathbf{x}_t^{\text{MF}}, \ell_t) \leftarrow \arg\max_{\mathbf{x},\ell} \alpha(\mathbf{x}, \ell \mid \mu_{\text{MF}}, \sigma_{\text{MF}})$
    if $\ell_t = m$ **then**
      condition2 $\leftarrow$ True
    else
      ISleft $\leftarrow [\![m-1]\!]$
      while |ISleft| > 0 **and not** condition2 **do**
        $(\mathbf{x}_t^{\text{MF}}, \ell_t) \leftarrow \arg\max_{\mathbf{x}\in\mathcal{X},\ell\in\text{ISleft}} \alpha(\mathbf{x}, \ell \mid \mu_{\text{MF}}, \sigma_{\text{MF}})$
        if $s(\mathbf{x}_t^{\text{MF}}, \ell_t) \geq c_2$ **then**
          condition2 $\leftarrow$ True
        else
          ISleft $\leftarrow$ ISleft $\setminus \{\ell_t\}$
        end if
      end while
    end if
  end if
  if condition1 **and** condition2 **then**
    $p_{\text{obs}} \leftarrow p_{\text{obs}} \cup \{t\}$
    $y_t \leftarrow f(\mathbf{x}_t^{\text{MF}}, \ell_t)$
    $\mathcal{D}^{\text{MF}} \leftarrow \mathcal{D}^{\text{MF}} \cup \{((\mathbf{x}_t^{\text{MF}}, \ell_t), y_t)\}$
    Perform Bayesian updates $\mu_{\text{MF}}, \sigma_{\text{MF}}$
    $y_t \leftarrow \mu_{\text{MF}}(\mathbf{x}_t^{\text{pSF}}, m)$ # pseudo-observation
    $\mathcal{D}^{\text{pSF}} \leftarrow \mathcal{D}^{\text{pSF}} \cup \{(\mathbf{x}_t^{\text{pSF}}, y_t)\}$
    $\Lambda \leftarrow \Lambda - \lambda_{\ell_t}$
  else
    $y_t \leftarrow f(\mathbf{x}_t^{\text{pSF}}, m)$
    $\mathcal{D}^{\text{pSF}} \leftarrow \mathcal{D}^{\text{pSF}} \cup \{(\mathbf{x}_t^{\text{pSF}}, y_t)\}$
    $\mathcal{D}^{\text{MF}} \leftarrow \mathcal{D}^{\text{MF}} \cup \{((\mathbf{x}_t^{\text{pSF}}, m), y_t)\}$
    $\Lambda \leftarrow \Lambda - \lambda_m$
  end if
  Perform Bayesian updates $\mu_{\text{SF}}, \sigma_{\text{SF}}, \mu_{\text{MF}}, \sigma_{\text{MF}}$
  $\mathcal{D}^{\text{pSF}} \leftarrow$ UPDATE-PSEUDO-OBS$(\mathcal{D}^{\text{pSF}}, \mathcal{D}^{\text{MF}}, \mu_{\text{MF}}, \mu_{\text{pSF}}, p_{\text{obs}})$
  Perform Bayesian updates $\mu_{\text{SF}}, \sigma_{\text{SF}}$
  $t \leftarrow t + 1$
end while
$S \leftarrow \{\mathbf{x} \in \mathcal{X} \mid \sigma_{\text{MF}}(\mathbf{x}, m) \leq c_1\}$
$\mathbf{x}_t^{\text{pSF}} \leftarrow \arg\max_{\mathbf{x}\in S} \mu_{\text{MF}}(\mathbf{x}, m)$
$y_t \leftarrow f(\mathbf{x}_t^{\text{pSF}}, m)$

Algorithm 3: UPDATE-PSEUDO-OBS

Input: $\mathcal{D}^{\text{pSF}}, \mathcal{D}^{\text{MF}}, \mu_{\text{MF}}, \mu_{\text{pSF}}, pobs$
for t in pobs **do**
  $\mathbf{x}_t^{nn} \leftarrow \text{NearestNeighbor}(\mathbf{x}_t^{\text{pSF}}, \mathcal{D}^{\text{MF}}[\ell = m])$
  $y \leftarrow f(\mathbf{x}_t^{nn}, m)$
  if $|\mu_{\text{MF}}(\mathbf{x}, m) - y| > |\mu_{\text{SF}}(\mathbf{x}) - y|$ **then**
    $\mathcal{D}^{\text{pSF}}[y_t] \leftarrow \mu_{\text{SF}}(\mathbf{x}_t^{\text{pSF}})$
  else
    $\mathcal{D}^{\text{pSF}}[y_t] \leftarrow \mu_{\text{MF}}(\mathbf{x}_t^{\text{pSF}}, m)$
  end if
end for
return $\mathcal{D}^{\text{pSF}}$

D. Computing constants M_t

Recall that the formula for M_t presented in Equations (12) and (13). The optimization over $\mathbb{D}_t$ makes the computation of M_t expensive. To avoid this, we consider a lower bound for M_t, defined as,

$$\underline{M}_t := \left\| \left[(\mathbf{H}_{\alpha,\mathbf{x}}(\mathbf{x}_{t+1}(\mathcal{D}_t), \mathcal{D}_t))^{-1} \frac{\partial \nabla_{\mathbf{x}} \alpha}{\partial \mathcal{D}_t^{(k)}} (\mathbf{x}_{t+1}(\mathcal{D}_t), \mathcal{D}_t) \right]_{k=1}^{t(d+1)} \right\|_{\text{op}}, \tag{17}$$

where $\|A\|_{\text{op}} := \inf\{c > 0 : \|A\mathbf{x}\|_\infty \leq c \|\mathbf{x}\|_\infty \ \forall \mathbf{x} \in \mathcal{X}\}$ is the operator norm, and $\mathbf{H}_{\alpha,\mathbf{x}}$ denotes the Hessian of $\mathbf{x} \mapsto \alpha(\mathbf{x}, \mathcal{D}_t)$. Specifically, the $(i,j)^{\text{th}}$-element of $\mathbf{H}_{\alpha,\mathbf{x}}$ reads as $\frac{\partial^2 \alpha}{\partial x_i \partial x_j}$. The i^{th}-element of $\frac{\partial \nabla_{\mathbf{x}} \alpha}{\partial \mathcal{D}_t^{(k)}}$ is $\frac{\partial^2 \alpha}{\partial \mathcal{D}_t^{(k)} \partial x_i}$, which denotes the partial derivative w.r.t. the first and the second variable of $\alpha(\cdot, \cdot)$. Note that $\mathbf{H}_{\alpha,\mathbf{x}}(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^{d \times d}$ and $\frac{\partial \nabla_{\mathbf{x}} \alpha}{\partial \mathcal{D}_t^{(k)}}(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^d$ for $\mathbf{a} \in \mathbb{R}^d, \mathbf{b} \in \mathbb{R}^{t(d+1)}$.

The gradient of $\nabla \alpha_{\mathbf{x}}$ can be obtained by exploiting the automatic differentiation tools available in different programming frameworks. We used the BoTorch-GPyTorch ecosystems (Balandat et al., 2020; Gardner et al., 2018). Matrices in Equation (17) need not to compute separately, but instead by taking the Jacobian of $(\mathbf{x}_{t+1}, \mathcal{D}_t) \mapsto \nabla_{\mathbf{x}} \alpha(\mathbf{x}_{t+1}, \mathcal{D}_t)$, and by considering its sub-matrices, all the terms can be obtained. However, the automatic differentiation comes at the cost of possible numerical instability. Especially, a reliable estimate of the Hessian $\mathbf{H}_{\alpha,\mathbf{x}}$ turned out to be difficult to obtain, resulting often a Hessian with complex eigen values and lacking symmetry. However, we run an experiment where the Hessian was forced to be symmetric and a large jitter term was added to the diagonal. The results on Rosenbrock 2D with rMF-GIBBON over 20 repetitions are depicted in Table 1.

Round	1	2	3	4	5	6	7	8	9
Mean	0.823668	1.081626	1.005563	1.199914	1.483624	1.454633	1.854412	2.625723	3.606273
Std	0.398510	0.475530	0.358448	0.610190	0.559461	0.511228	0.903042	1.302165	1.784013

Table 1: The mean and standard deviation of $\underline{M}_t, t \in [\![9]\!]$, over 20 repetitions.

E. Multi-fidelity kernels

We here give some insights about the different joint models that can be used in MFBO, as well as some additional numerical experiments using different kernels.

E.1. Kernels

The Downsampling kernel: Recall that the joint model employed in the experiments from the main text uses the following kernel:

$$k((\mathbf{x},\ell),(\mathbf{x}',\ell')) = k_{\text{input}}(\mathbf{x},\mathbf{x}') \times k_{\text{IS}}(\ell,\ell')$$
$$k_{\text{input}}(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\frac{(x_i-x_i')^2}{s_i}\right)$$
$$k_{\text{IS}}(\ell,\ell') = c + (1-\ell)^{1+\delta}(1-\ell')^{1+\delta}$$

The value $\ell \in [0,1]$ needs to be specified, and represents the confidence we have in the IS, with the primary IS m being associated to $\ell = 1$. The hyperparameters c, δ and $\{s_i\}_{1\leq i\leq d}$ are obtained through marginal likelihood maximization. When $\delta = 0$,

$$k((\mathbf{x},\ell),(\mathbf{x}',\ell')) = (c + (1-\ell)(1-\ell'))k_{\text{input}}(\mathbf{x},\mathbf{x}') \tag{18}$$

which can be written as

$$k((\mathbf{x},\ell),(\mathbf{x}',\ell')) = \begin{cases} ck_{\text{input}}(\mathbf{x},\mathbf{x}') + (1-\ell)(1-\ell')k_{\text{input}}(\mathbf{x},\mathbf{x}') & \ell \neq 1,\ \ell' \neq 1 \\ ck_{\text{input}}(\mathbf{x},\mathbf{x}') & \text{otherwise} \end{cases}$$

The Linear Truncated kernel: The Linear Truncated kernel implemented in BoTorch reads as

$$k_{\text{LT}}((\mathbf{x},\ell),(\mathbf{x}',\ell')) = k_{\text{input}}(\mathbf{x},\mathbf{x}') + c(\ell,\ell')k_{\text{IS}}(\mathbf{x},\mathbf{x}') \tag{19}$$
$$c(\ell,\ell') = (1-\ell)(1-\ell')(1+\ell\ell')^p \tag{20}$$

where k_{input} and k_{IS} are Matern kernels both with $\nu = 2.5$, but each with their own lengthscale. For $p = 0$, this leads to

$$k_{\text{LT}}((\mathbf{x},\ell),(\mathbf{x}',\ell')) = \begin{cases} k_{\text{input}}(\mathbf{x},\mathbf{x}') + (1-\ell)(1-\ell')k_{\text{IS}}(\mathbf{x},\mathbf{x}') & \ell \neq 1,\ \ell' \neq 1 \\ k_{\text{input}}(\mathbf{x},\mathbf{x}') & \text{otherwise} \end{cases}$$

This highlights a close correspondance with the Downsampling kernel when the hyperparameters of k_{IS} are close to that of k_{input}.

The MISO kernel: Following our notations, the Multi-Information Source Optimization (MISO) kernel introduced in (Poloczek et al., 2017, p.3) reads as

$$k_{\text{MISO}}((\mathbf{x}, \ell), (\mathbf{x}', \ell')) = k_{\text{input}}(\mathbf{x}, \mathbf{x}') + \mathbb{I}(\ell = \ell')k_{\ell}(\mathbf{x}, \mathbf{x}') \tag{21}$$

where k_{input} and k_{ℓ} are similar kernels, e.g. both Matern or RBF, but each with their own lengthscale. Here, ℓ and ℓ' take categorical values, corresponding to ISs indexes, with m being the primary IS index, equivalent to $\ell = 1$ for the Downsampling and Linear Truncated kernels. This can also be written as

$$k_{\text{MISO}}((\mathbf{x}, \ell), (\mathbf{x}', \ell')) = \begin{cases} k_{\text{input}}(\mathbf{x}, \mathbf{x}') + k_{\ell}(\mathbf{x}, \mathbf{x}') & \ell = \ell' \neq m \\ 2k_{\text{input}}(\mathbf{x}, \mathbf{x}') & \ell = \ell' = m \\ k_{\text{input}}(\mathbf{x}, \mathbf{x}') & \text{otherwise} \end{cases}$$

E.2. Additional numerical experiments

We evaluate the Linear Truncated kernel on the same setting described in Section 4, that we remind here, for completeness. Let us consider the Hartmann6D function as the objective (i.e., the primary IS). We examine two scenarios: in the first one, the auxiliary IS is informative, consisting of a biased version of the primary IS, with a degree of fidelity $l = 0.2$. In the second scenario, the auxiliary IS is taken to be the 6-dimensional Rosenbrock function, an irrelevant source for this problem. In both scenarios, the query costs are 1 for the primary IS and 0.2 for the auxiliary IS. The hyperparameter for rMFBO are set to $c_1 = c_2 = 0.1$ in all experiments of this section.

Figure 10 displays the results using the Linear Truncated kernel, with $\ell = 0.2$ for the auxiliary IS and $\ell = 1$ for the primary IS. Using this kernel, MF-MES exhibits a clear lack of robustness (third column, top row). Indeed, the bar plot (third column, middle row) indicates that the irrelevant IS is queried in majority. As the number of primary IS query is very low, we also provide the inference regret (bottom row), which shows a similar trend. MF-GIBBON is more robust in the irrelevant IS case, while still being outperformed by rMF-GIBBON.

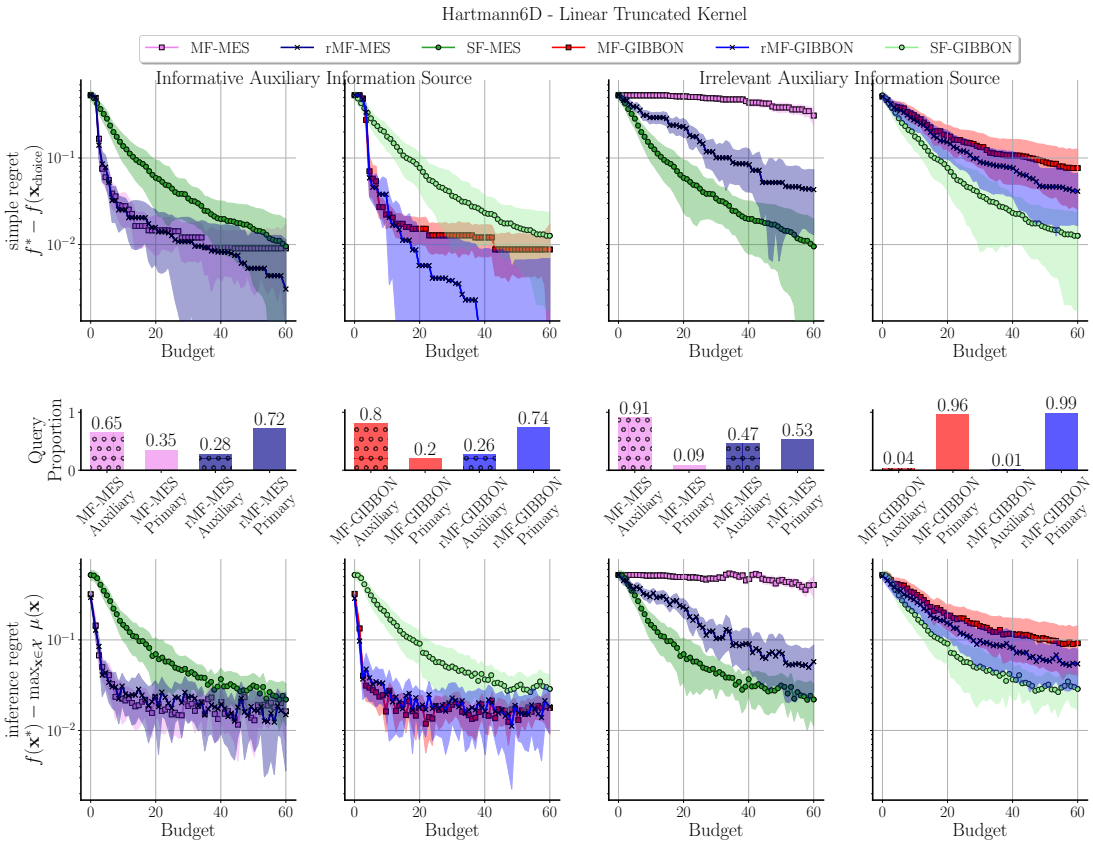

Figure 10: Simple regret depicted over budget spent for the Hartmann6D multi-fidelity problem, averaged over 20 repetitions. For the informative auxiliary IS, the Hartmann function with bias $l = 0.2$ is considered (see Section F). The irrelevant auxiliary IS is the 6D Rosenbrock function.

Varying the kernel confidence level: Both the Linear Truncated and Downsampling kernels (for $p > 0$) involve a contribution that depends on which value ℓ is associated to each IS (Equations (18) and (19)). While the primary IS m always correspond to the value $\ell = 1$, the value set for the auxiliary IS remains to be selected, and will determine whether or not that IS is perceived as relevant in the optimization of the objective. Figure 11 considers the same setting as described at the end of Section 4 using the Downsampling kernel, and for each row, the auxiliary IS value ℓ is varied in $\{0.1, 0.2, 0.5, 0.8\}$, thus simulating increased confidence in the auxiliary IS for the MOGP. As expected, overconfidence ($\ell = 0.8$) in the irrelevant IS leads to optimization failure for MF-MES and MF-GIBBON (right panel), while rMF-MES and rMF-GIBBON maintain consistent performances.

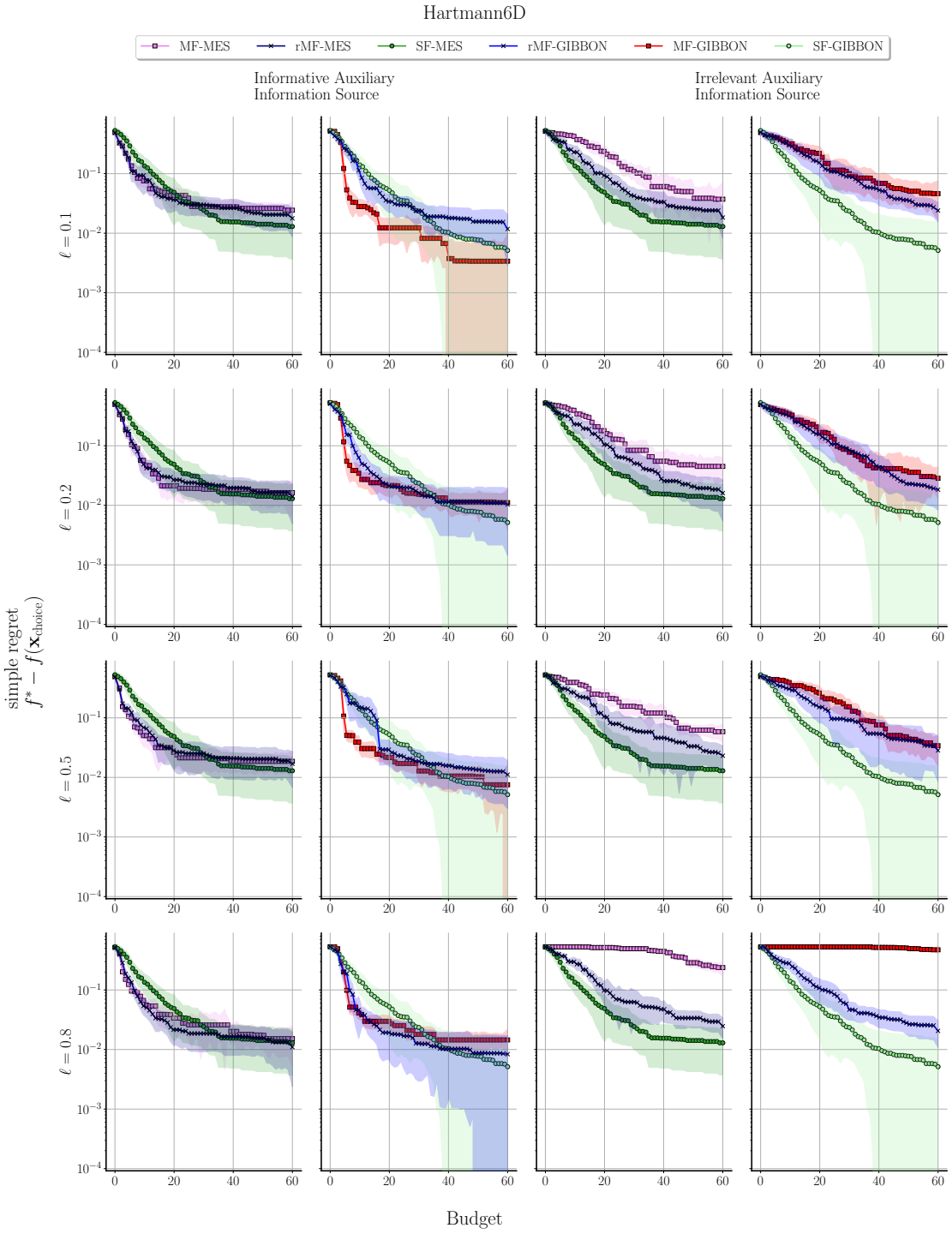

Figure 11: Simple regret depicted over budget spent for the Hartmann6D multi-fidelity problem, averaged over 20 repetitions. For the informative auxiliary IS, the Hartmann function with bias $l = 0.2$ is considered (see Section F). The irrelevant auxiliary IS is the 6D Rosenbrock function. For each row, the confidence level ℓ that the MOGP places in the auxiliary IS is varied

F. Test functions and experiment details

Hartmann-6D function:

$$f(\mathbf{x}, l) = -\sum_{i=1}^{4} \alpha_i \exp\left(-\sum_{j=1}^{6} A_{ij}(x_j - P_{ij})\right)$$
$$\alpha = (1.0 - 0.1(1-l), 1.2, 3.0, 3.2)^T$$
$$\mathbf{A} = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix}$$
$$\mathbf{P} = 10^{-4} \begin{pmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{pmatrix}$$

defined over $[0,1]^6$, and $l \in [0,1]$ is the degree of fidelity. The primary IS is then reached for $l = 1$.

Rosenbrock-dD function:

$$f(\mathbf{x}) = \sum_{i=1}^{d-1} \left(100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right)$$

defined over $[-5,5]^d$. The sinus-perturbed version used in the 2D case is defined as:

$$g(\mathbf{x}) = f(\mathbf{x}) + \mathbb{E}[f(X)] \times 0.8 \sin(x_1 + x_2)$$

The expectation is approximated by the empirical mean taken over a grid of 1000×1000 points linearly spaced across $[-5,5]^2$.

Exponential Currin 2D function:

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

defined over $[0,1]^2$.

Branin 2D function:

$$f(\mathbf{x}, l) = \left(x_2 - \left(\frac{5.1}{4\pi^2} - 0.1(1-l)\right)x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10$$

defined over $[-5,10] \times [0,15]$, and $l \in [0,1]$ is the degree of fidelity.

Ackley 2D function:

$$f(\mathbf{x}) = -20\exp\left(-0.2\sqrt{\frac{1}{2}(x_1+x_2)^2}\right) - \exp\left(\frac{1}{2}(\cos(2\pi x_1)+\cos(2\pi x_2)\right) + 20 + e^1$$

defined over $[-5, 10] \times [0, 15]$.

XGBoost hyperparameter tuning: The following hyperparameters are optimized: Huber loss parameter $\alpha \in [0.01, 0.1]$, complexity parameter used for minimal cost-complexity pruning ($[0.01, 100]$), fraction of samples used to fit individual base learners ($[0.1, 1]$), fraction of features considered when looking for the best tree split ($[0.01, 1]$) and learning rate ($[0.001, 1]$).