A GENERAL FRAMEWORK FOR USER-GUIDED BAYESIAN OPTIMIZATION

Carl Hvarfner
Lund University
carl.hvarfner@cs.lth.se

Frank Hutter
University of Freiburg
Prior Labs
fh@cs.uni-freiburg.de

Luigi Nardi
Lund University
Stanford University
DBTune
luigi.nardi@cs.lth.se

ABSTRACT

The optimization of expensive-to-evaluate black-box functions is prevalent in various scientific disciplines. Bayesian optimization is an automatic, general and sample-efficient method to solve these problems with minimal knowledge of the underlying function dynamics. However, the ability of Bayesian optimization to incorporate prior knowledge or beliefs about the function at hand in order to accelerate the optimization is limited, which reduces its appeal for knowledgeable practitioners with tight budgets. To allow domain experts to customize the optimization routine, we propose `ColaBO`, the first Bayesian-principled framework for incorporating prior beliefs beyond the typical kernel structure, such as the likely location of the optimizer or the optimal value. The generality of `ColaBO` makes it applicable across different Monte Carlo acquisition functions and types of user beliefs. We empirically demonstrate `ColaBO`'s ability to substantially accelerate optimization when the prior information is accurate, and to retain approximately default performance when it is misleading.

1 INTRODUCTION

Bayesian Optimization (BO) (Mockus et al., 1978; Jones et al., 1998; Snoek et al., 2012) is a well-established methodology for the optimization of expensive-to-evaluate black-box functions. Known for its sample efficiency, BO has been successfully applied to a variety of domains where laborious system tuning is prominent, such as hyperparameter optimization (Snoek et al., 2012; Bergstra et al., 2011b; Lindauer et al., 2022), neural architecture search (Ru et al., 2021; White et al., 2021), robotics (Calandra et al., 2014; Mayr et al., 2022), hardware design (Nardi et al., 2019; Ejjeh et al., 2022), and chemistry (Griffiths & Hernández-Lobato, 2020).

Typically employing a Gaussian Process (Rasmussen & Williams, 2006) (GP) surrogate model, BO allows the user to specify a prior over functions $p(f)$ via the Gaussian Process kernel, as well as an optional prior over its hyperparameters. Within the framework of the prior, the user can specify expected smoothness, output range and possible noise level of the function at hand, with the hopes of accelerating the optimization if accurate. However, the prior beliefs that can be specified within the framework of the kernel hyperparameters do not span the full range of beliefs that practitioners may possess. For example, practitioners may know which *parts of the input space* tend to work best (Oh et al., 2018; Perrone et al., 2019; Smith, 2018; Wang et al., 2019), know a range or upper bound on the optimal output (Jeong & Kim, 2021; Nguyen & Osborne, 2020) such as a maximal achievable accuracy of 100%, or other properties of the objective, such as preference relations between configurations (Huang et al., 2022). The limited ability of practitioners to interact and

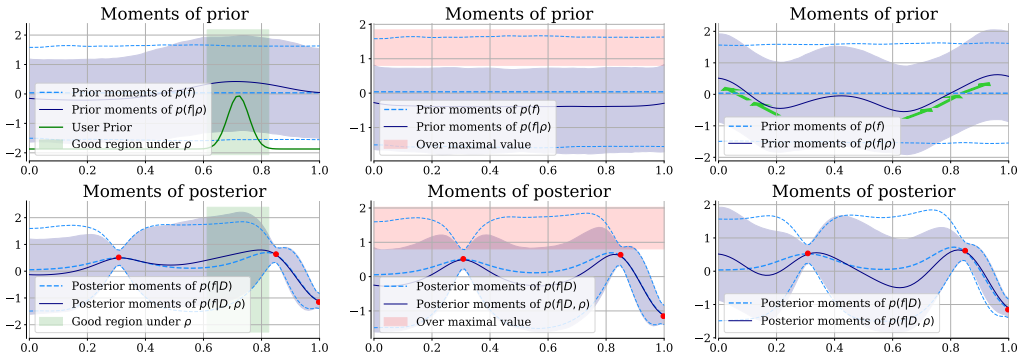

Figure 1: Three different `ColaBO` priors: (left) Prior over the optimum $\boldsymbol{x}_*$, and the induced changed in the GP for an optimum located in the green region. (middle) Prior over optimal value, $f^* < 0.8$. (right) Prior over preference relations $f(\boldsymbol{x})_1 \geqslant f(\boldsymbol{x}_2)$ for five preferences (green arrows, e.g. $f(0.0) \geqslant f(0.1) \geqslant f(0.2)$.

collaborate with the BO machinery (Kumar et al., 2022) thus runs the risk of failing to use valuable domain expertise, or alienating knowledgeable practitioners altogether. While knowledge injection beyond what is natively supported by the GP kernel is crucial to further increase the efficiency of Bayesian optimization, so far no current approach allows for the integration of arbitrary types of user knowledge. To address this gap, we propose an intuitive framework that effectively allows the user to reshape the Gaussian process at will to mimic their held beliefs.

Figure 1 illustrates how, for the three aforementioned priors, the GP is reshaped to *faithfully represent* the belief held by the user - whether it be a prior over the optimum (left), optimal value (middle), or preference relations between points (right). Our novel framework for *Collaborative Bayesian Optimization* (`ColaBO`) diverges from the typical assumption of Gaussian posteriors, and is applicable to any Monte Carlo acquisition function (Wilson et al., 2017; 2018; Balandat et al., 2020). Formally, we make the following contributions:

1. We introduce `ColaBO`, a framework for incorporating user knowledge over the optimizer, optimal value and preference relations into Bayesian optimization in the form of an additional prior on the surrogate, orthogonal to the conventional prior on the kernel hyperparameters,
2. We demonstrate that the proposed framework is generally applicable to Monte Carlo acquisition functions, inheriting MC acquisiiton function utility,
3. We empirically show that `ColaBO` accelerates optimization when injected with for priors over optimal location and optimal value.

2 Background

We outline Bayesian optimization, Gaussian Processes and Monte Carlo (MC) acquisition functions, as well as the concept of a prior over the optimum.

2.1 Bayesian optimization

We consider the problem of optimizing a black-box function f across a set of feasible inputs $\mathcal{X} \subset \mathbb{R}^d$:

$$\boldsymbol{x}^* \in \arg\max_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}). \tag{1}$$

We assume that $f(\boldsymbol{x})$ is expensive to evaluate and can potentially only be observed through a noise-corrupted estimate, $y_{\boldsymbol{x}}$, where $y_{\boldsymbol{x}} = f(\boldsymbol{x}) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ for some noise level σ_ϵ^2. In this setting, we wish to maximize f in an efficient manner. Bayesian optimization (BO) aims to globally maximize f by an initial design and thereafter sequentially choosing new points $\boldsymbol{x}_n$ for some iteration n, creating the data $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{(\boldsymbol{x}_n, y_n)\}$ (Brochu et al., 2010; Shahriari et al., 2016; Garnett, 2022). After each new observation, BO constructs a probabilistic surrogate model $p(f|\mathcal{D}_n)$ (Snoek et al., 2012; Hutter et al., 2011; Bergstra et al., 2011a; Müller et al., 2023) and uses that surrogate to build an acquisition function $\alpha(\boldsymbol{x}; \mathcal{D}_n)$ which selects the next query.

2.2 GAUSSIAN PROCESSES

When constructing the surrogate, the most common choice is a *Gaussian process* (GP) (Rasmussen & Williams, 2006). The GP utilizes a covariance function k, which encodes a prior belief for the smoothness of f, and determines how previous observations influence prediction. Given observations $\mathcal{D}_n$ at iteration n, the Gaussian posterior $p(f|\mathcal{D}_n)$ over the objective is characterized by the posterior mean $\mu_n(\boldsymbol{x}, \boldsymbol{x}')$ and (co-)variance $\Sigma_n(\boldsymbol{x}, \boldsymbol{x}')$ of the GP:

$$\mu_n(\boldsymbol{x}) = \mathbf{k}_n(\boldsymbol{x})^\top(\mathbf{K}_n + \sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{y}, \quad \Sigma_n(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}') - \mathbf{k}_n(\boldsymbol{x})^\top(\mathbf{K} + \sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{k}_n(\boldsymbol{x}'),$$

where $(\mathbf{K}_n)_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\mathbf{k}_n(\boldsymbol{x}) = [k(\boldsymbol{x}, \boldsymbol{x}_1), \dots, k(\boldsymbol{x}, \boldsymbol{x}_n)]^\top$ and σ_ϵ^2 is the noise variance. For applications in BO and beyond, samples from the posterior are required either directly for optimization (Eriksson et al., 2019) through Thompson sampling (Thompson, 1933), or to estimate auxiliary quantities of interest (Hernández-Lobato et al., 2015; Neiswanger et al., 2021; Hvarfner et al., 2023). For a finite set of k query locations ($\boldsymbol{X} = \boldsymbol{x}_1, \dots, \boldsymbol{x}_k$), the classical method of generating samples is via a location-scale transform of Gaussian random variables, $f(\boldsymbol{X}) = \mu_n(\boldsymbol{X}) + \boldsymbol{L}\boldsymbol{\epsilon}$, where $\boldsymbol{L}$ is the Cholesky decomposition of $\boldsymbol{K}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$. Unfortunately, the classic approach is intrinsically non-scalable, incurring a $\mathcal{O}(k^3)$ cost due to the aforementioned matrix decomposition.

2.3 DECOUPLED POSTERIOR SAMPLING

To remedy the issue of scalability in posterior sampling, $\mathcal{O}(k)$ weight-space approximations based on Random Fourier Features (RFF) (Rahimi & Recht, 2007) obtain approximate (continuous) function draws $\tilde{f}(\boldsymbol{x}) = \sum_{i=1}^m w_i\phi_i(\boldsymbol{x})$, where $\phi_i(\boldsymbol{x}) = \frac{2}{\ell}(\boldsymbol{\psi}_i^\top\boldsymbol{x} + b_i)$. The random variables $w_i \sim \mathcal{N}(0, 1)$, $b_i \sim \mathcal{U}(0, 2\pi)$, and $\boldsymbol{\psi}_i$ are sampled proportional to the spectral density of k.

While achieving scalability, the seminal RFF approach by Rahimi & Recht (2007) suffers from the issue of variance starvation (Mutny & Krause, 2018; Wang et al., 2018; Wilson et al., 2020). As a remedy, Wilson et al. (2020) decouple the draw of functions from the approximate posterior $p(\tilde{f}|\mathcal{D})$ into a more accurate draw from the prior $p(\tilde{f})$, followed by a deterministic data-dependent update:

$$(\tilde{f}|\mathcal{D})(\boldsymbol{x}) \overset{d}{=} \underbrace{\tilde{f}(\boldsymbol{x})}_{\text{draw from prior}} + \underbrace{\mathbf{k}_n(\boldsymbol{x})^\top(\mathbf{K}_n + \sigma_\epsilon^2\mathbf{I})^{-1}(\mathbf{y} - \tilde{f}(\boldsymbol{x}) - \boldsymbol{\epsilon})}_{\text{deterministic update}} \tag{2}$$

Eq. 2 deviates from the distribution-first approach that is typically prevalent in GPs in favor of a variable-first approach utilizing Matheron's rule (Journel & Huijbregts, 1976).

2.4 MONTE CARLO ACQUISITION FUNCTIONS

Acquisition functions act on the surrogate model to quantify the utility of a point in the search space. They encode a trade-off between exploration and exploitation, typically using a greedy heuristic to do so. A simple and computationally cheap heuristic is Expected Improvement (`EI`) (Jones et al., 1998; Bull, 2011). For a noiseless function and a current best observation y_n^*, the `EI` acquisition function is $\alpha_{EI}(\boldsymbol{x}) = \mathbb{E}_{y_{\boldsymbol{x}}}\left[(y_n^* - y_{\boldsymbol{x}})^+\right]$. For noisy problem settings, a noise-adapted variant of `EI` (Letham et al., 2018) is frequently considered, where both the incumbent y_n^* and the upcoming query $y_{\boldsymbol{x}}$ are substituted for the non-observable noiseless incumbent f_n^* and noiseless upcoming query $f_{\boldsymbol{x}}$. Other frequently used acquisition functions are the Upper Confidence Bound (`UCB`) (Srinivas et al., 2012), Probability of Improvement (PI) (Kushner, 1964) and Knowledge Gradient (KG) (Frazier et al., 2009). Information-theoretic acquisition functions consider the mutual information to select the next query $\alpha_{\texttt{MI}}(\boldsymbol{x}) = I((\boldsymbol{x}, y_{\boldsymbol{x}}); *|\mathcal{D}_n)$, where $*$ can entail either the optimum $\boldsymbol{x}_*$ as in (Predictive) Entropy Search (`ES`/`PES`) (Hennig & Schuler, 2012; Hernández-Lobato et al., 2014), the optimal value f_* as in Max-value Entropy Search (`MES`) (Wang & Jegelka, 2017; Moss et al., 2021) or the tuple $(\boldsymbol{x}_*, f_*)$ for Joint Entropy Search (`JES`) (Hvarfner et al., 2022a; Tu et al., 2022).

All the aforementioned acquisition functions compute expectations $\mathbb{E}_{f_{\boldsymbol{x}}}$ (or alternatively $\mathbb{E}_{y_{\boldsymbol{x}}}$) over some utility $u(f_{\boldsymbol{x}})$ of the output (Wilson et al., 2017; 2018), which typically have simple, or even closed-form, solutions for Gaussian posteriors. However, approximating the expectation through Monte Carlo integration has proven useful in the context of batch optimization (Wilson et al., 2018), efficient acquisition function approximation (Balandat et al., 2020), and non-Gaussian posteriors (Astudillo & Frazier, 2021). By sampling over possible outputs $f_{\boldsymbol{x}}$ and utilizing the reparametrization

trick (Kingma & Welling, 2014; Rezende et al., 2014), utilities u can be easily computed across a larger set of applications and be optimized to greater accuracy.

2.5 Prior over the Optimum

A prior over the optimum (Souza et al., 2021; Hvarfner et al., 2022b; Mallik et al., 2023) is a user-specified belief $\pi : \mathcal{X} \to \mathbb{R}$ of the subjective likelihood that a given $\boldsymbol{x}$ is optimal. Formally,

$$\pi(\boldsymbol{x}) = \mathbb{P}\left(\boldsymbol{x} = \arg\max_{\boldsymbol{x}'} f(\boldsymbol{x}')\right). \tag{3}$$

This prior is generally considered to be independent of observed data, but rather a result of previous experimentation or anecdotal evidence. Regions that the user expects to contain the optimum will typically have a high value, but this does not exclude the chance of the user belief $\pi(\boldsymbol{x})$ to be inaccurate, or even misleading. Lastly, we require π to be strictly positive in all of $\mathcal{X}$, which suggests that any point included in the search space may be optimal.

3 Methodology

We now introduce `ColaBO`, the first Bayesian-principled BO framework that flexibly allows users to *collaborate* with the optimizer by injecting prior knowledge about the objective that substantially exceeds the type of prior knowledge natively supported by GPs. In Sec. 3.1, we introduce and derive a novel prior over function properties, which yields a surrogate model conditioned on the user belief. Thereafter, in Sec. 3.2, we demonstrate how the hierarchical prior integrates with MC acquisition functions. Lastly, in Sec. 3.3, we state practical considerations to assure the performance of `ColaBO`.

3.1 Prior over Function Properties

We consider the typical GP prior over functions $p(f) = \mathcal{GP}(\mu, \Sigma)$, where the characteristics of f, such as smoothness and output magnitude, are fully defined by the kernel k (and its associated hyper-parameters $\boldsymbol{\theta}$, which are omitted for brevity). We seek to inject an additional, user-defined prior belief over f into the GP, such as the prior over the optimum in Sec. 2.5, $\pi(\boldsymbol{x}) = \mathbb{P}\left(\boldsymbol{x} = \arg\max_{\boldsymbol{x}'} f(\boldsymbol{x}')\right)$. By postulating that π is accurate, we wish to form a belief-weighted prior - a prior over *functions* where the distribution over the optimum is exactly $\pi(\boldsymbol{x})$. We start by considering the user belief $\pi : \mathcal{X} \to \mathbb{R}$ from Eq. (3), and extend the definition to involve the integration over f, similarly to the Thompson sampling definition of Kandasamy et al. (2018). Formally,

$$\pi(\boldsymbol{x}) = \mathbb{P}\left(\boldsymbol{x} = \arg\max_{\boldsymbol{x}'} f(\boldsymbol{x}')\right) = \int_f \pi(\delta_*(\boldsymbol{x}|f))p(f)df \tag{4}$$

where $\delta_*(\boldsymbol{x}|f) = 1$, if $\boldsymbol{x} = \arg\max_{\boldsymbol{x}' \in \mathcal{X}} f(\boldsymbol{x}')$, and zero otherwise. As such, $\delta_*(\boldsymbol{x}|f)$ maps a function $f_i \sim p(f)$ to its $\arg\max$, and evaluates whether this $\arg\max$ is equal to $\boldsymbol{x}$.

However, a belief over the optimum, or any other property, of a function f is implicitly a belief over the function f itself. As such, a non-uniform $\pi(\boldsymbol{x})$ should reasonably induce a change in the prior $p(f)$ to reflect the non-uniform optimum. To this end, we introduce an augmented user belief over the optimum $\rho^*_{\boldsymbol{x}} \sim \mathcal{P}^*_{\boldsymbol{x}}$, where $\mathcal{P}^*_{\boldsymbol{x}}$ is the prior over possible user beliefs, and draws are random functions $\rho^*_{\boldsymbol{x}} : \mathcal{X} \to \mathbb{R}^+$ which themselves take a function f as input, and output a positive real number quantifying the likelihood of a sample f_i under $\pi(\boldsymbol{x})$. Formally, we define $\rho^*_{\boldsymbol{x}}$ as

$$\rho^*_{\boldsymbol{x}}(f) = \mathbb{P}\left(\boldsymbol{x} = \arg\max_{\boldsymbol{x}'} f(\boldsymbol{x}')\right) = \frac{1}{Z_{\rho^*_{\boldsymbol{x}}}} \int_{\mathcal{X}} \delta_*(\boldsymbol{x}|f)\pi(\boldsymbol{x})d\boldsymbol{x} \tag{5}$$

where the intractible normalizing constant $Z_{\rho^*_{\boldsymbol{x}}}$ arises from the fact that the integrated density $\pi(\boldsymbol{x})$ acts on a finite-dimensional *property* of f, and not f itself. Under $\rho^*_{\boldsymbol{x}}(f)$, functions whose $\arg\max$ lies in a high-density region under π will be assigned a higher probability. Notably, the definition in 5 can extend to other properties of f as well: a user belief p_{f*} over the optimal value f_* analogously yields a belief over functions $\rho^*_{f_{\boldsymbol{x}}}(f)$:

$$\rho^*_{f_{\boldsymbol{x}}}(f) = \mathbb{P}\left(\boldsymbol{x} = \max_{\boldsymbol{x}'} f(\boldsymbol{x}')\right) = \frac{1}{Z_{\rho^*_{f_{\boldsymbol{x}}}}} \int_{f_{\boldsymbol{x}}} \delta_*(\boldsymbol{x}|f)p_{f*}(y)df_{\boldsymbol{x}}. \tag{6}$$

It is worthwhile to reflect on the meaning of $\rho(f)$, and how beliefs over function properties propagate to $p(f)$. Concretely, if the user belief $\rho^*_{f_{\boldsymbol{x}}}(f)$ asserts that the maximal value lies within $C_1 < \max f < C_2$, the resulting distribution over f should only contain functions whose $\max$ falls within this range. Using rejection sampling, functions which disobey this criterion are filtered out, which yields the posterior $p(f|\rho)$. Having defined and exemplified how user beliefs impact the prior over functions $p(f)$, the role of ρ as a likelihood should be apparent: given a prior over functions $p(f)$ and a user belief over functions $\rho(f)$ which places a probability on all possible draws $f_i\ p(f)$, we can form a belief-weighted prior $p(f|\rho) \propto p(f)\rho(f)$. Thus, we introduce the formal definition of a user belief over a function property:

Definition 3.1 (User Belief over Functions)**.** *The user belief over functions* $\rho(f) \propto \frac{p(f|\rho)}{p(f)}$.

As the subsequent derived methodology applies independently of the specific property of f that a prior is placed on, we will henceforth consider a belief over a general function property ρ. Having defined the role of ρ and the posterior over functions it produces, a natural question arises: *How is* $p(f|\rho)$ *updated once observations* $\mathcal{D}$ *are obtained?*

Since the data $\mathcal{D}$ is independent of the prior (the data generation process is intrinsically unaffected by the belief held by the user), application of Bayes' rule yields the following posterior $p(f|\mathcal{D}, \rho)$,

$$p(f|\mathcal{D}, \rho) = \frac{p(\mathcal{D}, \rho|f)p(f)}{p(\mathcal{D}, \rho)} = \frac{p(\mathcal{D}|f)p(\rho|f)p(f)}{p(\mathcal{D})p(\rho)} = \frac{p(f|\rho)}{p(f)} p(f|\mathcal{D}) \propto \rho(f)p(f|\mathcal{D}), \tag{7}$$

where the right side of the proportionality in Eq. 7 suggests an intuitive generation process for samples $(f|\mathcal{D}, \rho)$ to approximate the density $p(f|\mathcal{D}, \rho)$. Utilizing the pathwise update from Eq. 2, we note that given an approximate draw $\tilde{f}$ from the prior, the subsequent data-dependent update is deterministic. Recalling Eq. 2 and assuming independence between ρ and $\mathcal{D}$, ρ only affects the draw from the prior, whereas $\mathcal{D}$ only affects the update. Consequently, we obtain

$$(\tilde{f}|\mathcal{D}, \rho)(\boldsymbol{x}) \overset{d}{=} \underbrace{(\tilde{f}|\rho)(\boldsymbol{x})}_{\text{draw from prior}} + \underbrace{\mathbf{k}_n(\boldsymbol{x})^\top(\mathbf{K}_n + \sigma_\epsilon^2\mathbf{I})^{-1}(\mathbf{y} - (\tilde{f}|\rho)(\boldsymbol{x}) - \boldsymbol{\epsilon})}_{\text{deterministic update}}, \tag{8}$$

where $(\tilde{f}|\rho) \sim p(f)\rho(\tilde{f})$ are once again obtained using rejection sampling on draws from $p(\tilde{f})$. Figure 2 displays this in detail: given the typical GP prior over functions *and* a user belief over the optimum, we obtain a distribution over functions $p(\tilde{f}|\rho^*_{\boldsymbol{x}})$ before having observed any data (top right). Samples from the approximate prior $p(\tilde{f})$ (light blue) are re-sampled proportionally to their probability of occurring under the prior $\rho^*_{\boldsymbol{x}}(\tilde{f})$ in green, leaving samples $(\tilde{f}|\rho^*_{\boldsymbol{x}})$ in navy blue, which are highly probable under $\rho^*_{\boldsymbol{x}}$. Once data is obtained, these samples are updates according to Eq. 8, which preserves the shape of the samples far away from observed data and yields the desired posterior.

3.2 Prior-weighted Monte Carlo Acquisition Functions

Naturally, neither the belief-weighted prior $p(f|\rho)$ nor the belief-weighted posterior $p(f|\mathcal{D}, \rho)$ have a closed-form expression. Both are inherently non-Gaussian for non-uniform beliefs. As such, we resort to MC acquisition functions to compute utilities that are amenable to BO. In the subsequent section, we focus on the prevalent acquisition functions `EI`, and `MES`.

Expected Improvement The computation of the MC-`EI` within the `ColaBO` framework requires only minor adaptations of the original MC acquisition function. By definition, MC-`EI` assigns utility u as $u_{\texttt{EI}}(f(\boldsymbol{x})) = \max(f_n^* - f(\boldsymbol{x}), 0)$, which yields

$$\alpha_{\texttt{EI}}(\boldsymbol{x}; \mathcal{D}) = \mathbb{E}_{f_{\boldsymbol{x}}|\mathcal{D}}[u_{\texttt{EI}}(f_{\boldsymbol{x}})] \approx \tag{9}$$

$$\sum_\ell \max(f_n^* - f_{\boldsymbol{x}}^{(\ell)}, 0),\ f_{\boldsymbol{x}}^{(\ell)} \sim p(f(\boldsymbol{x})|\mathcal{D}). \tag{10}$$

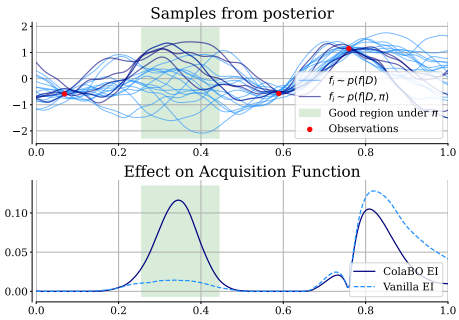

Figure 3: (Top) Draws from $p(f|\mathcal{D})$ (light blue) and $p(f|\rho, \mathcal{D})$ with a prior ρ located in the green region. (Bottom) Vanilla MC-`EI` and `ColaBO` MC-`EI`, resulting from computing the acquisition function from sample draws from $p(f|\rho, \mathcal{D})$.

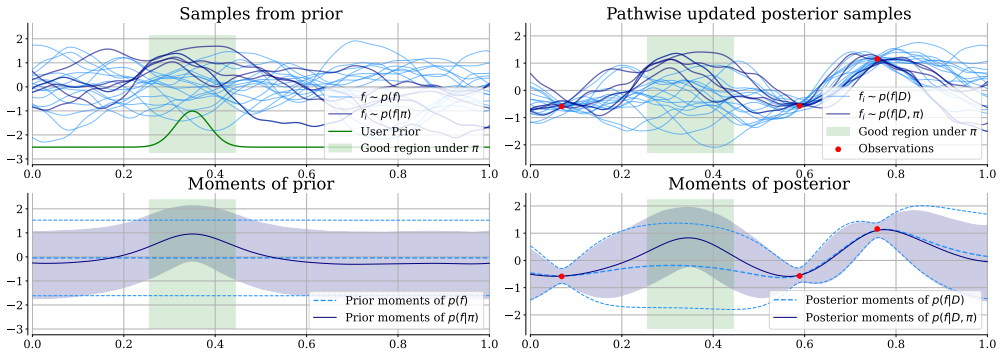

Figure 2: (Top left) Draws from the prior $p(f)$ (light blue) and the belief-weighted prior $p(f|\rho)$ whose members are likely to have their optimum within the green region. (Top right) Pathwise updated draws based on observed data. As the green region is distant from the observed data, samples are almost unaffected by the data in this region. (Bottom left) Exact mean and standard deviation $(\mu_{\boldsymbol{x}}, \sigma_{\boldsymbol{x}})$ of $p(f)$ and estimated mean and standard deviation of $p(f|\rho)$. (Bottom right) Exact $p(f|\mathcal{D})$ and estimated $p(f|\rho, \mathcal{D})$. As $p(f|\rho)$ constitutes of functions whose optimum is located within the green region the resulting model has a higher mean and lower variance within this region. Moreover, $p(f|\rho)$ globally displays lower upside variance compared to the vanilla GP.

Utilizing rejection sampling, we can compute the MC-`EI` under the `ColaBO` posterior accordingly,

$$\alpha_{\texttt{EI}}(\boldsymbol{x}; \mathcal{D}, \rho) = \mathbb{E}_{f_{\boldsymbol{x}}|\mathcal{D},\rho}[u_{\texttt{EI}}(f_{\boldsymbol{x}})] \propto \tag{11}$$

$$\int_f u_{\texttt{EI}}(f_{\boldsymbol{x}})\rho(f)p(f|\mathcal{D})df \approx \sum_{\ell} \rho(f^{(\ell)}) \max(f_n^* - f_{\boldsymbol{x}}^{(\ell)}, 0), \quad f_{\boldsymbol{x}}^{(\ell)} \sim p(f(\boldsymbol{x})|\mathcal{D}), \tag{12}$$

wherein samples in Eq. 12 are drawn from the prior, retained with probability $\rho(f^{(\ell)})/\max \rho$, and pathwise updated. In Figure 3, we demonstrate how `ColaBO-EI` differs from MC-`EI` for an identical posterior as in Figure 2. By computing $\alpha_{\texttt{EI}}$ from samples biased by ρ, `ColaBO` substantially directs the search towards good regions under ρ. Derivations for `PI` and `KG` are analogous to that of `EI`.

Max-Value Entropy Search We derive a `ColaBO-MES` acquisition function by first considering the definition of the entropy, $\mathrm{H}[p(y_{\boldsymbol{x}}|\mathcal{D})] = \mathbb{E}_{y_{\boldsymbol{x}}|\mathcal{D}}[-\log p(y_{\boldsymbol{x}}|\mathcal{D})]$. When considering the belief-weighted posterior, we further condition the posterior on ρ and obtain

$$\alpha_{\texttt{MES}}(\boldsymbol{x}) = \mathbb{E}_{f_*|\mathcal{D},\rho}\left[\mathbb{E}_{y_{\boldsymbol{x}}|\mathcal{D},\rho,f_*}[\log p(y_{\boldsymbol{x}}|\mathcal{D},\rho,f_*)]\right] - \mathbb{E}_{y_{\boldsymbol{x}}|\mathcal{D},\rho}[\log p(y_{\boldsymbol{x}}|\mathcal{D},\rho)] \tag{13}$$

$$\propto \mathbb{E}_{f_*|\mathcal{D},\rho}\left[\mathbb{E}_{f_{\boldsymbol{x}}|\mathcal{D},\rho}[\mathbb{E}_{y_{\boldsymbol{x}}|f_{\boldsymbol{x}}}[\log p(y_{\boldsymbol{x}}|f_{\boldsymbol{x}},\rho,f_*)]]\right] - \mathbb{E}_{f_{\boldsymbol{x}}|\mathcal{D},\rho}[\mathbb{E}_{y_{\boldsymbol{x}}|f_{\boldsymbol{x}}}[\log p(y_{\boldsymbol{x}}|f_{\boldsymbol{x}},\rho)]] \tag{14}$$

$$\approx \frac{1}{Z_J}\sum_{j=1}^{J}\sum_{\ell=1}^{L}\sum_{k=1}^{K} \log p(y_{\boldsymbol{x}}^{(k)}|f_{\boldsymbol{x}}^{(\ell)}, f_*^{(j)})\rho(f^{(\ell)})\rho(f^{(j)}) - \sum_{\ell=1}^{L}\sum_{k=1}^{K} \log p(y_{\boldsymbol{x}}^{(k)}|f_{\boldsymbol{x}}^{(\ell)})\rho(f^{(\ell)}), \tag{15}$$

where Z_J is a normalizing constant $\sum_J \rho(f^{(j)})$ brought on by sampling optimal values, $y_{\boldsymbol{x}}|f_{\boldsymbol{x}}$ can trivially be obtained by sampling Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ to a noiseless observation $f_{\boldsymbol{x}}|\mathcal{D}$ in the innermost expectation, and $f_{\boldsymbol{x}}$ and f_* are obtained through the pathwise sampling procedure outlined in Eq. 8. The samples are evaluated on $p((y_{\boldsymbol{x}}|f_{\boldsymbol{x}}),(y_{\boldsymbol{x}}|f_{\boldsymbol{x}}, f_*))$. As evident by Eq. 15, ρ affects the posterior distribution of both the observations $y_{\boldsymbol{x}}$ and the optimal values f_*. `PES` and `JES` are derived analogously. However, these acquisition function require conditioning on additional, simulated data and consequently, additional pathwise updates, to compute.

3.3 Practical Considerations

`ColaBO` introduces additional flexibility to MC-based BO acquisition functions. The `ColaBO` framework deviates from vanilla (q-)MC acquisition functions (Wilson et al., 2017; Balandat et al., 2020) by utilizing approximate sample functions from the posterior, as opposed to pointwise draws from the posterior predictive and the reparametrization trick (Kingma & Welling, 2014; Rezende et al., 2014). `ColaBO` holds three shortcomings not prevalent in vanilla MC acquisition functions: (1) it cannot utilize Quasi-MC in the draws from the predictive posterior (only in the RFF weights),

Algorithm 1 `ColaBO` iteration

Input: User prior ρ, number of function samples L, current data $\mathcal{D}$
Output: Next query location $\boldsymbol{x}'$.
for $\ell \in \{1, \dots, L\}$ **do**
  $\rho^{(\ell)} = \rho(\tilde{f}^{(\ell)}; n), \tilde{f}^{(\ell)} \sim p(\tilde{f})$ ▷ `Sample functions and evaluate on` π
  $(\tilde{f}^{(\ell)}|\mathcal{D}) = \texttt{PathwiseUpdate}(\tilde{f}^{(\ell)}, \mathcal{D})$ ▷ `Per-sample update as in Eq. 8`
end for
$p(\tilde{f}|\mathcal{D}, \rho) \approx \sum_\ell \rho^{(\ell)}(\tilde{f}^{(\ell)}|\mathcal{D})$ ▷ `Form MC estimate of posterior`
$\boldsymbol{x}' = \arg\max_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}_{p(\tilde{f}|\mathcal{D},\rho)}[u(\tilde{f}_{\boldsymbol{x}})]$ ▷ `Maximize MC acquisition`

(2) it cannot fix the base samples (Balandat et al., 2020) drawn from the posterior for acquisition function consistency across the search space, and (3) the RFF approximation of $p(f)$ introduces bias. This approximation error is substantially more pronounced for the Matérn 5/2-kernel than the squared exponential, leaving ColaBO best suited for the latter. In Sec. 4.1, we empirically display the impact of these shortcomings. While acquisition function optimization no longer enjoys improved accuracy resulting from reparametrization, the acquisition function can still benefit from the fact that `ColaBO` backpropagates through quantities computed as sums of smooth functions.

4 Results

We evaluate the performance of `ColaBO` on various tasks, using priors over the optimum $\pi_{\boldsymbol{x}_*}$ obtained from known optima on synthetic tasks, as well as from prior work (Mallik et al., 2023) on realistic tasks. We consider two variants of `ColaBO`: one using `LogEI` (Ament et al., 2023), a numerically stable, smoothed `logsumexp` transformation of `EI` with analogous derivation, and one variant using `MES`. We benchmark against the vanilla variants of each acquisition function, as well as π`BO` (Hvarfner et al., 2022b) and decoupled Thompson sampling Thompson (1933); Wilson et al. (2020). All acquisition functions are implemented in BoTorch (Balandat et al., 2020) using a squared exponential kernel and MAP hyperparameter estimation. We present experiments with a Matérn-5/2 (Matérn, 1960) kernel in App. C.1. Unless stated otherwise, all methods are initialized with the mode of the prior followed by 2 Sobol samples. The experimental setup is outlined in Appendix B, and our code is publicly available at `https://github.com/hvarfner/colabo`.

4.1 Approximation Quality of the ColaBO Framework

Firstly, we demonstrate the approximation quality of `ColaBO` *without* user priors to assert its accuracy compared to a vanilla MC acquisition function. To facilitate comparison, we randomly sample 10 points on the Hartmann (3D) function, and optimize `LogEI` with a large budget. We subsequently optimize `ColaBO-LogEI` on the same set of points and compare the $\arg\max$ to the solution found by the gold standard. Figure 4 displays the (log10) Euclidian distance between the $\arg\max$ of `LogEI` and its `ColaBO` variant. We note that, for small amounts ($\leqslant 256$) of posterior samples, the error induced by RFF bias is relatively low, which is evidenced by all RFF variants being roughly equal in distance to the true acquisition function optimizer.

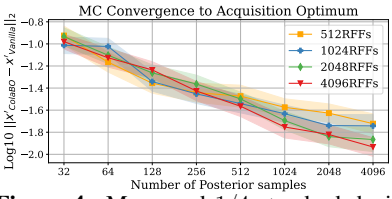

Figure 4: Mean and $1/4$ standard deviation of MC-induced errors of `ColaBO-LogEI` relative vanilla `LogEI` as measured by the distance to the $\arg\max$ of the acquisition function on Hartmann (3D) on 10 randomly sampled points for 40 seeds.

4.2 Synthetic Functions with Known Priors

We adapt a similar evaluation protocol to Hvarfner et al. (2022b), and evaluate `ColaBO` for two types of user beliefs for synthetic tasks: well-located and poorly located priors over the optimal location, designed to emulate a well-informed and poorly-informed practitioner, respectively. The well-located prior is offset by a small (10%) amount from the optimum, and the poorly located prior is maximally offset, while retaining its mode inside the search space. Complete details on the priors can be found

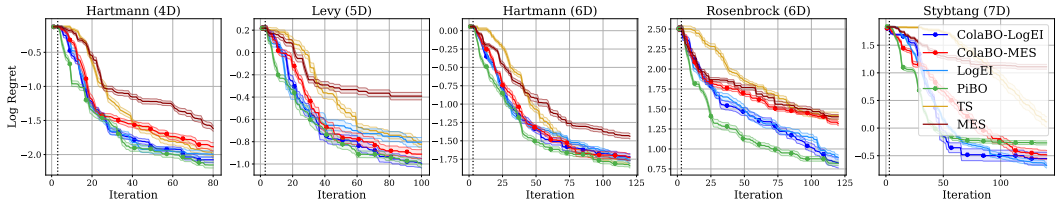

Figure 5: Performance on synthetic functions with well-located priors. Both `ColaBO-LogEI` and `ColaBO-MES` offer drastic speed-ups over their vanilla variants, and offer similar performance to π`BO`. The ranking of `ColaBO` acquisition functions are generally consistent with their respective vanilla variants. This is most prominent on Rosenbrock (6D), where `ColaBO-MES` struggles similarly to vanilla `MES`.

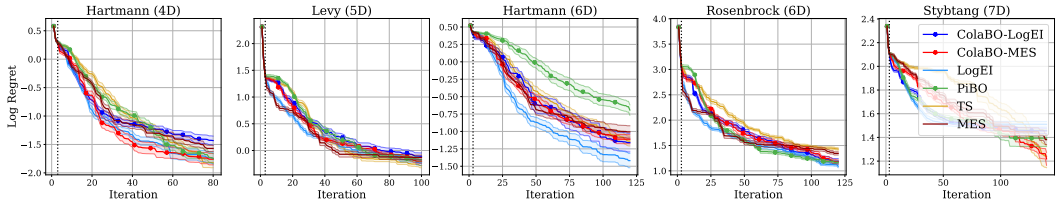

Figure 6: Performance on poorly located priors. `ColaBO` acquisition functions are more robust than π`BO`, as it frequently recovers the performance of the vanilla acquisition function before the total budget is depleted. `ColaBO-LogEI` struggles marginally on Hartmann (6D). `ColaBO-MES` recovers the baseline on all tasks.

in Appendix B.3. On well-located priors, both `ColaBO-LogEI` and `ColaBO-MES` demonstrate substantially improved performance relative to their vanilla counterparts, comparable to π`BO` on all benchmarks. On poorly located priors, `ColaBO` demonstrates superior robustness, recovering the performance of the vanilla acquisition function within the maximal budget of $20D$ iterations and clearly outperforming π`BO`, which more frequently misled by the poor prior. In Appendix C.2, we also demonstrate `ColaBO` utilizing (accurate) beliefs over the optimal value: similarly to Figure 5, `ColaBO` yields increased efficiency relative to baselines, albeit not as substantial. Moreover, we demonstrate its usage with batch evaluations on well-located priors in Sec. C.3, showing that the drop in performance from batching evaluations is marginal at worst.

4.3 Hyperparameter Tuning tasks

Lastly, we evaluate `ColaBO` on three $4D$ deep learning HPO tasks from the PD1 (Wang et al., 2023) benchmarking suite. While the optima for these tasks are ultimately unknown, we utilize the priors provided in MF-Prior-Bench [1] (Mallik et al., 2023), which are intended to provide a good starting point for further optimization. To emulate a realistic HPO setting, we consider a smaller optimization budget of $10D$ iterations, and initalize all methods that utilize user beliefs with only one initial sample, that being the mode of the prior. The two `ColaBO` variants perform best in this evaluation, producing the best terminal performance on two tasks (CIFAR, LM1B), with all methods being tied on the third (CIFAR). `ColaBO` demonstrates consistent speed-ups compared to its vanilla counterparts, surpassing the terminal performance of the baseline within a third of the budget on CIFAR and LM1B.

5 Related Work

In BO, auxiliary prior information can be conveyed in multiple ways. We outline meta learning/transfer learning for BO based on data from previous experiments, and data-less approaches.

Learning from Previous Experiments Transfer learning and meta learning for BO aims to automatically extract and use knowledge from prior executions of BO by pre-training the model on data acquired from previous executions (Swersky et al., 2013; Wistuba et al., 2015; Perrone et al.,

[1] `https://github.com/automl/mf-prior-bench`

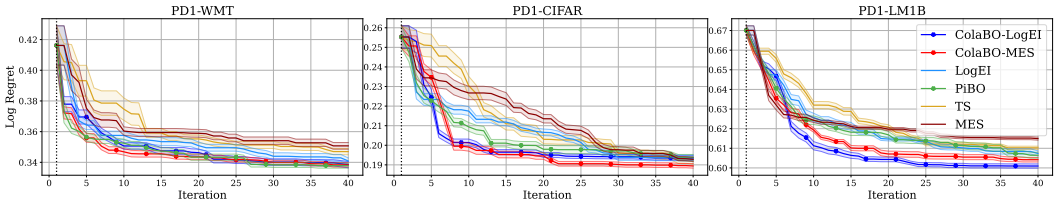

Figure 7: Performance on the 4D PD1 hyperparameter tuning tasks of various deep learning pipelines. `ColaBO` drastically accelerates optimization initially, finding configurations with close to terminal performance quickly. π`BO` offers competitive performance, but lacks the rapid initial progress of `ColaBO` on CIFAR and LM1B.

2019; Feurer et al., 2015; 2018; Rothfuss et al., 2021a;b; Wistuba & Grabocka, 2021; Feurer et al., 2022). Typically, meta- and transfer learning exploit relevant previous data for training the GP for the current task while retaining predictive uncertainty to account for imperfect task correlation.

Expert Priors over Function Optimum Few previous works have proposed to inject explicit prior distributions over the location of an optimum into BO. In these cases, users explicitly define a prior that encodes their beliefs on where the optimum is more likely to be located. Bergstra et al. (2011a) suggest an approach that supports prior beliefs from a fixed set of distributions, which affects the very initial stage of optimization. However, this approach cannot be combined with standard acquisition functions. BOPrO (Souza et al., 2021) employs a similar structure that combines the user-provided prior distribution with a data-driven model into a pseudo-posterior. From the pseudo-posterior, configurations are selected using the EI acquisition function, using the formulation in Bergstra et al. (2011a). πBO (Hvarfner et al., 2022b) suggests a general-purpose prior-weighted acquisition function, where the influence of the prior decreases over time. They provide convergence guarantees for when the framework is applied to the EI acquisition function. While effective, none of these approaches act on the surrogate model in a Bayesian-principled fashion, but strictly as heuristics. Moreover, they solely focus on priors over optimal inputs, thus offering less utility than `ColaBO`.

Priors over Optimal Value Similarly few works have addressed the issue of auxilliary knowledge of the optimal value. Both Jeong & Kim (2021) and Nguyen & Osborne (2020) propose altering the GP and accompanying it with tailored acquisition functions. Jeong & Kim (2021) employ variational inference, proposing distinct variational families depending on the type of knowledge pertaining to the optimal value. Nguyen & Osborne (2020) use a parabolic transformation of the output space to ensure an upper bound is preserved. Unlike `ColaBO`, neither of these methods is general enough to accompany arbitrary user priors to guide the optimization.

6 Conclusion, Limitations and Future Work

We presented `ColaBO`, a flexible BO framework that allows practitioners to inject beliefs over function properties in a Bayesian-principled manner, allowing for increased efficiency in the BO procedure. `ColaBO` works across a collection of MC acquisition functions, inheriting their flexibility in batch optimization and ability to work with non-Gaussian posteriors. It demonstrates competitive performance for well-located priors, using them to substantially accelerate optimization. Moreover, it retains approximately baseline performance when applied to detrimental priors, demonstrating greater robustness than π`BO`. `ColaBO` crucially relies on multiple steps of MC. While flexible, this approach drives computational expense in order to assert sufficient accuracy, requiring tens of seconds per evaluation to achieve desired accuracy, depending on the size of the benchmark. Moreover, obtaining draws from $\rho^*_{\boldsymbol{x}}$ scales exponentially in the dimensionality of the prior. While practitioners are unlikely to specify priors over more than a handful of variables, `ColaBO` may become impractical when priors of higher dimensionality are employed. Paths for future work could involve more accurate and efficient sampling procedures (Lin et al., 2023) from the belief-weighted prior, as well as variational (Titsias, 2009) or pre-trained Müller et al. (2022); Müller et al. (2023) approaches to obtain a representative belief-biased model with an analytical posterior. This would likely bring down the runtime of `ColaBO` and broaden its potential use. Lastly, applying `ColaBO` to multi-fidelity optimization (Kandasamy et al., 2016; Mallik et al., 2023) offers an additional avenue for increased efficiency which would further increase its viability on costly deep learning pipelines.

REFERENCES

Sebastian Ament, Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected improvements to expected improvement for bayesian optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1vyAG6j9PE.

Raul Astudillo and Peter Frazier. Bayesian optimization of function networks. *Advances in neural information processing systems*, 34:14463–14475, 2021.

M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization. In *Advances in Neural Information Processing Systems*, 2020. URL http://arxiv.org/abs/1910.06403.

J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Proceedings of the 25th International Conference on Advances in Neural Information Processing Systems (NeurIPS'11)*, pp. 2546–2554, 2011a.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24. Curran Associates, Inc., 2011b.

E. Brochu, V. Cora, and N. de Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv:1012.2599v1 [cs.LG]*, 2010.

Adam D. Bull. Convergence rates of efficient global optimization algorithms. 12:2879–2904, 2011.

R. Calandra, N. Gopalan, A. Seyfarth, J. Peters, and M. Deisenroth. Bayesian gait optimization for bipedal locomotion. In P. Pardalos and M. Resende (eds.), *Proceedings of the Eighth International Conference on Learning and Intelligent Optimization (LION'14)*, 2014.

Adel Ejjeh, Leon Medvinsky, Aaron Councilman, Hemang Nehra, Suraj Sharma, Vikram Adve, Luigi Nardi, Eriko Nurvitadhi, and Rob A Rutenbar. Hpvm2fpga: Enabling true hardware-agnostic fpga programming. In *Proceedings of the 33rd IEEE International Conference on Application-specific Systems, Architectures, and Processors*, 2022.

David Eriksson, Michael Pearce, Jacob Gardner, Ryan D Turner, and Matthias Poloczek. Scalable global optimization via local Bayesian optimization. In *Advances in Neural Information Processing Systems*, pp. 5496–5507, 2019. URL http://papers.nips.cc/paper/8788-scalable-global-optimization-via-local-bayesian-optimization.pdf.

M. Feurer, Jost Tobias Springenberg, and F. Hutter. Initializing bayesian hyperparameter optimization via meta-learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1128–1135, 2015.

M. Feurer, B. Letham, F. Hutter, and E. Bakshy. Practical transfer learning for bayesian optimization. *ArXiv abs/1802.02219*, 2018.

Matthias Feurer, Benjamin Letham, Frank Hutter, and Eytan Bakshy. Practical transfer learning for Bayesian optimization. *arXiv preprint 1802.02219*, 2022.

Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613, 2009.

R. Garnett. *Bayesian Optimization*. Cambridge University Press, 2022. Available for free at https://bayesoptbook.com/.

Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science*, 2020.

P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(1):1809–1837, June 2012. ISSN 1532-4435.

J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, 2014. URL https://proceedings.neurips.cc/paper/2014/file/069d3bb002acd8d7dd095917f9efe4cb-Paper.pdf.

José Miguel Hernández-Lobato, Michael Gelbart, Matthew Hoffman, Ryan Adams, and Zoubin Ghahramani. Predictive entropy search for bayesian optimization with unknown constraints. In *International conference on machine learning*, pp. 1699–1707. PMLR, 2015.

Daolang Huang, Louis Filstroff, Petrus Mikkola, Runkai Zheng, and Samuel Kaski. Bayesian optimization augmented with actively elicited expert knowledge, 2022.

F. Hutter, H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In C. Coello (ed.), *Proceedings of the Fifth International Conference on Learning and Intelligent Optimization (LION'11)*, volume 6683, pp. 507–523, 2011.

Carl Hvarfner, Frank Hutter, and Luigi Nardi. Joint entropy search for maximally-informed bayesian optimization. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022a.

Carl Hvarfner, Danny Stoll, Artur Souza, Marius Lindauer, Frank Hutter, and Luigi Nardi. PiBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization. In *International Conference on Learning Representations*, 2022b.

Carl Hvarfner, Erik Hellsten, Frank Hutter, and Luigi Nardi. Self-correcting bayesian optimization through bayesian active learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=dX9MjUtP1A.

Taewon Jeong and Heeyoung Kim. Objective bound conditional gaussian process for bayesian optimization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4819–4828. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jeong21a.html.

D. Jones, M. Schonlau, and W. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 12 1998. doi: 10.1023/A:1008306431147.

A G Journel and C J Huijbregts. Mining geostatistics, Jan 1976.

K. Kandasamy, G. Dasarathy, J. Oliva, J. Schneider, and B. Póczos. Gaussian Process Bandit Optimisation with Multi-fidelity Evaluations. In D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett (eds.), *Proceedings of the 30th International Conference on Advances in Neural Information Processing Systems (NeurIPS'16)*, pp. 992–1000, 2016.

K. Kandasamy, A. Krishnamurthy, J. Schneider, and B. Póczos. Parallelised Bayesian optimisation via Thompson sampling. In A. Storkey and F Perez-Cruz (eds.), *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, pp. 133–142. Proceedings of Machine Learning Research, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014. URL https://arxiv.org/abs/1312.6114.

Arun Kumar, Santu Rana, Alistair Shilton, and Svetha Venkatesh. Human-ai collaborative bayesian optimisation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 16233–16245. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6751611b394a3464cea53eed91cf163c-Paper-Conference.pdf.

H. J. Kushner. A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. *Journal of Basic Engineering*, 86(1):97–106, 03 1964. ISSN 0021-9223. doi: 10.1115/1.3653121. URL https://doi.org/10.1115/1.3653121.

B. Letham, K. Brian, G. Ottoni, and E. Bakshy. Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, 2018.

Jihao Andreas Lin, Javier Antorán, Shreyas Padhy, David Janz, José Miguel Hernández-Lobato, and Alexander Terenin. Sampling from gaussian process posteriors using stochastic gradient descent. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Sf9goJtTCE.

Marius Lindauer, Katharina Eggensperger, Matthias Feurer, André Biedenkapp, Difan Deng, Carolin Benjamins, Tim Ruhkopf, René Sass, and Frank Hutter. Smac3: A versatile bayesian optimization package for hyperparameter optimization. *Journal of Machine Learning Research*, 23(54):1–9, 2022. URL http://jmlr.org/papers/v23/21-0888.html.

Neeratyoy Mallik, Edward Bergman, Carl Hvarfner, Danny Stoll, Maciej Janowski, Marius Lindauer, Luigi Nardi, and Frank Hutter. Priorband: Practical hyperparameter optimization in the age of deep learning. *arXiv preprint 2306.12370*, 2023.

B. Matérn. Spatial variation. *Meddelanden fran Statens Skogsforskningsinstitut*, 1960.

Matthias Mayr, Carl Hvarfner, Konstantinos Chatzilygeroudis, Luigi Nardi, and Volker Krueger. Learning skill-based industrial robot tasks with user priors. *IEEE 18th International Conference on Automation Science and Engineering*, 2022. URL https://arxiv.org/abs/2208.01605.

J. Mockus, V. Tiesis, and A. Zilinskas. The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2(117-129):2, 1978.

Henry B. Moss, David S. Leslie, Javier Gonzalez, and Paul Rayson. Gibbon: General-purpose information-based bayesian optimisation. *Journal of Machine Learning Research*, 22(235):1–49, 2021. URL http://jmlr.org/papers/v22/21-0120.html.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Samuel Müller, Matthias Feurer, Noah Hollmann, and Frank Hutter. PFNs4BO: In-context learning for Bayesian optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25444–25470. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/muller23a.html.

Mojmir Mutny and Andreas Krause. Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/4e5046fc8d6a97d18a5f54beaed54dea-Paper.pdf.

L. Nardi, D. Koeplinger, and K. Olukotun. Practical design space exploration. In *2019 IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pp. 347–358. IEEE, 2019.

Willie Neiswanger, Ke Alexander Wang, and Stefano Ermon. Bayesian algorithm execution: Estimating computable properties of black-box functions using mutual information. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8005–8015. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/neiswanger21a.html.

Vu Nguyen and Michael A. Osborne. Knowing the what but not the where in Bayesian optimization. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7317–7326. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/nguyen20d.html.

C. Oh, E. Gavves, and M. Welling. BOCK : Bayesian optimization with cylindrical kernels. In *International Conference on Machine Learning*, pp. 3865–3874, 2018.

V. Perrone, H. Shen, M. Seeger, C. Archambeau, and R. Jenatton. Learning search spaces for bayesian optimization: Another view of hyperparameter transfer learning. In *Advances in Neural Information Processing Systems*, 2019.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/rezende14.html.

Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9116–9126, 2021a.

Jonas Rothfuss, Dominique Heyn, Jinfan Chen, and Andreas Krause. Meta-learning reliable priors in the function space. In *Advances in Neural Information Processing Systems*, volume 34, 2021b.

Binxin Ru, Xingchen Wan, Xiaowen Dong, and Michael Osborne. Interpretable neural architecture search via bayesian optimisation with weisfeiler-lehman kernels. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=j9Rv7qdXjd.

B. Shahriari, K. Swersky, Z. Wang, R. Adams, and N. de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

L. Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

J. Snoek, H. Larochelle, and R. Adams. Practical Bayesian optimization of machine learning algorithms. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger (eds.), *Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems (NeurIPS'12)*, pp. 2960–2968, 2012.

A. Souza, L. Nardi, L. Oliveira, K. Olukotun, M. Lindauer, and F. Hutter. Bayesian optimization with a prior for the optimum. In *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17, 2021, Proceedings, Part III*, volume 12977 of *Lecture Notes in Computer Science*, pp. 265–296. Springer, 2021.

N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, May 2012. ISSN 1557-9654. doi: 10.1109/tit.2011.2182033. URL http://dx.doi.org/10.1109/TIT.2011.2182033.

K. Swersky, J. Snoek, and R. Adams. Multi-task Bayesian optimization. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Proceedings of the 27th International Conference on Advances in Neural Information Processing Systems (NeurIPS'13)*, pp. 2004–2012, 2013.

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In David van Dyk and Max Welling (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL https://proceedings.mlr.press/v5/titsias09a.html.

Ben Tu, Axel Gandy, Nikolas Kantas, and Behrang Shafei. Joint entropy search for multi-objective bayesian optimization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ZChgD8OoGds.

Q. Wang, Y. Ming, Z. Jin, Q. Shen, D. Liu, M. J. Smith, K. Veeramachaneni, and H. Qu. Atmseer: Increasing transparency and controllability in automated machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–12. Association for Computing Machinery, 2019.

Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient bayesian optimization. In *International Conference on Machine Learning (ICML)*, 2017.

Zi Wang, Clement Gehring, Pushmeet Kohli, and Stefanie Jegelka. Batched large-scale bayesian optimization in high-dimensional spaces. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 745–754. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/wang18c.html.

Zi Wang, George E. Dahl, Kevin Swersky, Chansoo Lee, Zachary Nado, Justin Gilmer, Jasper Snoek, and Zoubin Ghahramani. Pre-trained Gaussian processes for Bayesian optimization. *arXiv preprint arXiv:2109.08215*, 2023.

C. White, W. Neiswanger, and Y. Savani. BANANAS: Bayesian optimization with neural architectures for neural architecture search. In Q. Yang, K. Leyton-Brown, and Mausam (eds.), *Proceedings of the Thirty-Fifth Conference on Artificial Intelligence (AAAI'21)*, pp. 10293–10301. Association for the Advancement of Artificial Intelligence, AAAI Press, 2021.

James Wilson, Frank Hutter, and Marc Deisenroth. Maximizing acquisition functions for bayesian optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/498f2c21688f6451d9f5fd09d53edda7-Paper.pdf.

James T. Wilson, Riccardo Moriconi, Frank Hutter, and Marc Peter Deisenroth. The reparameterization trick for acquisition functions, 2017. URL https://arxiv.org/abs/1712.00424.

James T. Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Peter Deisenroth. Efficiently sampling functions from gaussian process posteriors. In *International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/2002.09309.

M. Wistuba, N. Schilling, and L. Schmidt-Thieme. Hyperparameter search space pruning - A new component for sequential model-based hyperparameter optimization. In A. Appice, P. Rodrigues, V. Costa, J. Gama, A. Jorge, and C. Soares (eds.), *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD'15)*, volume 9285, pp. 104–119, 2015.

Martin Wistuba and Josif Grabocka. Few-shot bayesian optimization with deep kernel surrogates. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=bJxgv5C3sYc.

Lucas Zimmer, Marius Thomas Lindauer, and Frank Hutter. Auto-pytorch tabular: Multi-fidelity metalearning for efficient and robust autodl. *ArXiv*, abs/2006.13799, 2020. URL https://api.semanticscholar.org/CorpusID:220041844.

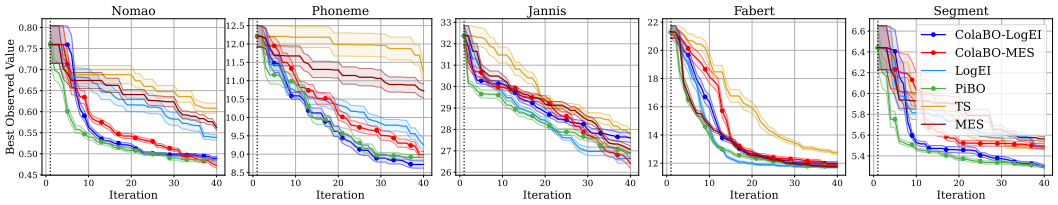

Figure 8: Performance on the 6D LCBench hyperparameter tuning tasks of various deep learning pipelines. `ColaBO` substantially improves on the non-prior baselines for 3 out of five tasks. π`BO` performs best on aggregate, and achieves the best acceleration in performance at early iterations.

A LCBENCH DL EXPERIMENTS

We evaluate all methods on five deep learning tasks (6D) from the LCBench (Zimmer et al., 2020) suite, once again utilizing priors from MF-Prior-Bench. The chosen tasks were the five tasks with available priors of the best (good) strength, as per the benchmark suite. Figure 8 shows the performance of all methods on the LCBench tasks. `ColaBO` improves substantially on the baseline approaches for 3 out of 5 tasks. π`BO` is the overall best-performing method, followed by `ColaBO-LogEI`.

B EXPERIMENTAL SETUP

B.1 MODEL

We outline the model used and the budget allocated to the various MC approximations involved with `ColaBO`. For all experiments, we utilize MAP estimation of the hyperparameters, and update the hyperparameters at every iteration of BO. All hyperparameters - lengthscale, outputscale and observation noise ($\boldsymbol{\theta} = \{\boldsymbol{\ell}, \sigma_\varepsilon^2, \sigma_f^2\}$) are given conventional $\mathcal{LN}(0,1)$ prior, applied on normalized inputs and standardized outputs. Furthermore, we fit the constant c of the mean function, assigning it a $\mathcal{N}(0,1)$ prior as well. In Tab. 1, we display the parameters of the MC approximations for various tasks. *No.* f is the maximal number of functions used in the MC computation of the acquisition function. *No. Reamples* is the number of initial posterior draws maximally used for the re-sampling of functions from the posterior $p(f|\rho)$. Lastly, . *No.* f_* is the number of optimal values used in the computation of `ColaBO-MES`.

Task	No. f	No. RFFs	No. Resamples	No. f_*
Synthetic Good	768	2048	$1.5 * 10^5$	32
Synthetic Bad	768	2048	$1.5 * 10^5$	32
PD1	512	4096	$2 * 10^5$	32
Appendix	512	1024	10^5	32

Table 1: Budget-related parameters of the Monte Carlo approximations for all tasks.

B.2 BENCHMARKS

We outline the benchmarks used, their search spaces and the amount of synthetic noise added. When adding noise, we intend for the ratio of noise variance to total output range to be approximately equal across benchmarks.

B.3 PRIORS

For synthetic benchmarks, the approximate optima of all included functions can be obtained in advance. Thus, the correctness of the prior is ultimately known in advance. For a function of

Task	Dimensionality	σ_ϵ	Search space
Hartmann (4D)	4	0.25	$[0, 1]^D$
Levy (5D)	5	0.5	$[-5, 5]^D$
Hartmann (6D)	6	0.25	$[0, 1]^D$
Rosenbrock (6D)	6	5	$[-2.048, 2.048]^D$
Stybtang (7D)	7	1	$[-4, 4]^D$

Table 2: Benchmarks used for the Bayesian optimization experiments.

dimensionality d with optimum at $\boldsymbol{x}_*$, the well-located prior is constructed by sampling an offset direction $\boldsymbol{\epsilon}$ and scaling the offset by a dimensionality- and quality-specific term $c(d, q) = q\sqrt{d}$. For the well-located prior on synthetic tasks, we use $q = 0.1$, which implies that the optimum is located 10% of the distance across the search space away from the optimum, and construct a Gaussian prior as

$$\pi_{\boldsymbol{x}_*}(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{x}_* + c_d \boldsymbol{\epsilon}/||\boldsymbol{\epsilon}||, \sigma_s), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}). \tag{16}$$

with $\sigma_s = 25\%$ for all tasks and prior qualities. For our 20 runs of the well-located prior, this procedure yields us 20 unique priors per quality type, with identical offsets from the true optimum. No priors with a mode outside the search space were allowed, such priors were simply replaced. For the misinformed priors, we set $q = 1$, guaranteeing that the mode of the prior will be outside of the search space, and subsequently relocating to the edge of the search space by its shortest path. Priors for all tasks are displayed in Tab. 3. For the PD1 tasks, the location for the priors were obtained from MF-Prior-Bench(`https://github.com/automl/mf-prior-bench`). However, these priors require offsetting in order to not be too strong, thus making subsequent BO obsolete. PD1 priors are provided in $[0, 1]$-normalized space for simplicity.

Task	Location	Offset, good	Offset, bad	σ_s
Hartmann (4D)	$[0.19, 0.19, 0.56, 0.26]$	$0.1\sqrt{D}$	max	0.25
Levy (5D)	$[1]^D$	$1\sqrt{D}$	max	2.5
Hartmann (6D)	$[0.20, 0.15, 0.48, 0.28, 0.31, 0.66]$	$0.1\sqrt{D}$	max	0.25
Rosenbrock (6D)	$[1]^D$	$0.4096\sqrt{D}$	max	1.024
Stybtang (7D)	$[-2.9]^D$	$0.8\sqrt{D}$	max	2
PD1-WMT	$[0.90, 0.69, 0.02, 0.97]$	$0.05\sqrt{D}$	N/A	0.25
PD1-CIFAR	$[1, 0.80, 0.0, 0.0]$	$0.05\sqrt{D}$	N/A	0.25
PD1-LM1B	$[0.91, 0.67, 0.36, 0.85]$	$0.05\sqrt{D}$	N/A	0.25

Table 3: $\pi_{\boldsymbol{x}_*}$ for synthetic BO tasks of both prior qualities and PD1.

C Additional Experiments

We provide complementary experiments to those introduced in the main paper. Firstly, we display results when `ColaBO` is used with a prior π_{f_*} over the optimal value in Sec. C.2. In Sec. C.3, we demonstrate `ColaBO`:s extensibility to batch evaluations, seamlessly extending the work of (Wilson et al., 2017).

C.1 Synthetic Matern Kernel Experiments

We evaluate `ColaBO`and all baselines on the synthetic tasks with a Matern-5/2 kernel and the good user belief over the optimum. We note that roughly half of all π`BO` runs struggle with numerical instability from iteration 60 onwards, which produces stagnation in performance and infrequent gains.

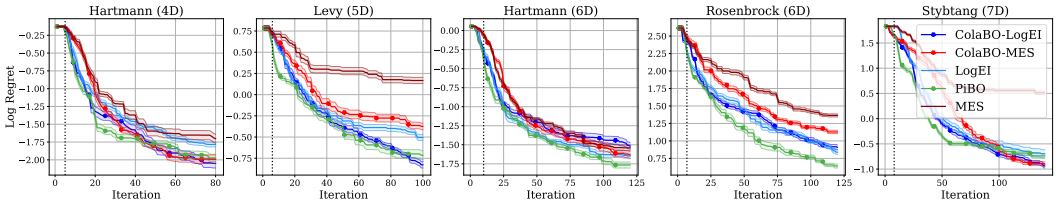

Figure 9: ColaBO on the synthetic tasks with a Matern kernel. Due to the difficulty of the RFF approximation, ColaBO-LogEI struggles on Hartmann (6D), and ColaBO performance is marginally worse on aggregate.

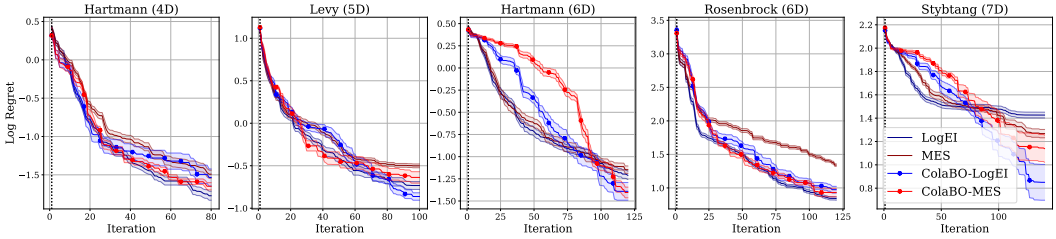

Figure 10: ColaBO with priors over the optimal value. Terminal performance substantially increases on 3 out of 5 benchmarks (Levy, Hartmann (6D), Stybtang), and is approximately preserved on the final two. ColaBO-MES improves marginally more than ColaBO-LogEI when utilizing a prior π_{f_*} over the optimal value.

C.2 Max-value Priors

We evaluate ColaBO with priors over the optimal value π_{f_*} in Figure 10. For each task, we place a Gaussian prior over the optimal value, centering it exactly at the optimal value. Notably, such a prior substantially influences the exploration-exploitiation trade-off; if the prior suggests that the incumbent has a value close to the optimal one, we are encouraged to exploit as samples with well-above-optimal values in exploratory will be discarded. Conversely, we are heavily encouraged to explore if the current best observation holds a value that we believe is far from optimal. On Hartmann (6D), we can see this behavior at play. Initial performance is poorer for ColaBO than their respective baselines, presumably due to above-average exploration, but terminal performance is better.

C.3 Batch Evaluations

We evaluate ColaBO on batch evaluations, utilizing the sequential greedy technique for MC acquisition functions from Wilson et al. (2018). Drop-off from sequential to batch evaluations is not evident from the plots, as ordering between sequential and batch varies with the benchmark. While unpredictable, we speculate that the altered exploration-exploitation trade-off provided by the batched acquisition function is occasionally beneficial in the presence of auxilliary user beliefs $\pi_{\boldsymbol{x}_*}$.

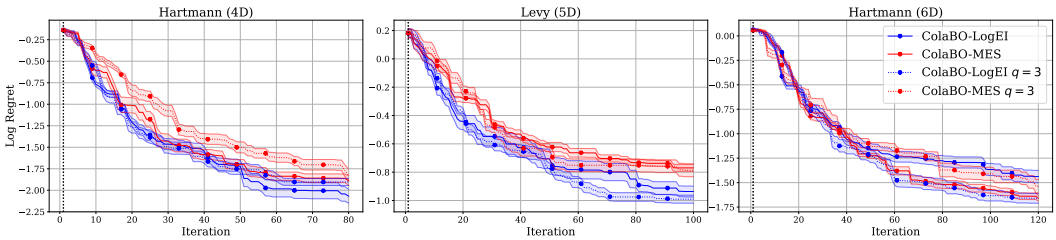

Figure 11: $q = 1$ (sequential) and $q = 3$ batch evaluation on a subset of synthetic functions with well-located priors for ColaBO-LogEI and ColaBO-MES. Total function evaluations are plotted for both sequential and batched variants, leaving them with the same number of total function evaluations.