Imprecise Acquisitions in Bayesian Optimization

Valentin Margraf
MCML, LMU Munich

Jonas Hanselle
MCML, LMU Munich

Julian Rodemann
CISPA, LMU Munich

Marcel Wever
L3S Research Center, Leibniz University Hannover

Sebastian Vollmer
DFKI (DSA)

Eyke Hüllermeier
LMU Munich, MCML, DFKI (DSA)

Abstract

Gaussian processes (GPs) are widely used as surrogate models in Bayesian optimization (BO). However, their predictive performance is highly sensitive to the choice of hyperparameters, often leading to markedly different posterior predictions. Hierarchical BO addresses this issue by marginalizing over hyperparameters to produce an aggregated posterior, which is then evaluated using an acquisition function (AF). Yet, this aggregation can obscure the disagreement among individual GP posteriors, an informative source of uncertainty that could be exploited for more robust decision-making. To overcome this limitation, we propose **I**mprecise **A**cquisitions in **B**ayesian **O**ptimization (IABO), which maintains a set of GP models and evaluates the AF separately under each one. This results in an imprecise, set-valued AF whose spread naturally captures model disagreement. We investigate two aggregation strategies applied at different stages: (i) acquisition-level aggregation, where AF values are combined into a single scalar via an aggregated AF, and (ii) decision-level aggregation, where each AF is optimized independently and the resulting maximizers are compared using stochastic dominance criteria. Our approach is applicable to arbitrary AFs, and experiments show that our decision-level strategies are highly competitive, often outperforming standard BO baselines across a range of benchmarks problems.

1 Introduction

Bayesian optimization (BO) is a widely used framework for optimizing expensive *black-box* functions [18]. It has been successfully applied across diverse domains, including materials design [49], drug discovery [36], agent design [14], and hyperparameter optimization [42, 11]. At its core, BO employs a probabilistic surrogate model to approximate the unknown objective function. This surrogate model provides posterior mean and variance estimates, which are used by an acquisition function (AF) to trade off exploration and exploitation. By maximizing the AF, BO selects the next evaluation point, incorporates the new observation, and refits the surrogate model, repeating this cycle until the evaluated budget is exhausted.

Gaussian processes (GPs) are the most commonly used surrogate models in BO [38], but their performance is highly sensitive to the choice of hyperparameters [33, 23, 39]. While hyperparameters are typically learned by maximizing the marginal likelihood, this procedure is often unstable and subject to high variability [28]. A more robust alternative, hierarchical BO, adopts a fully Bayesian treatment by sampling hyperparameters from a hyperprior and aggregating the corresponding GP posteriors into a single, averaged posterior.

1st Workshop on Epistemic Intelligence in Machine Learning (EIML 2025) .

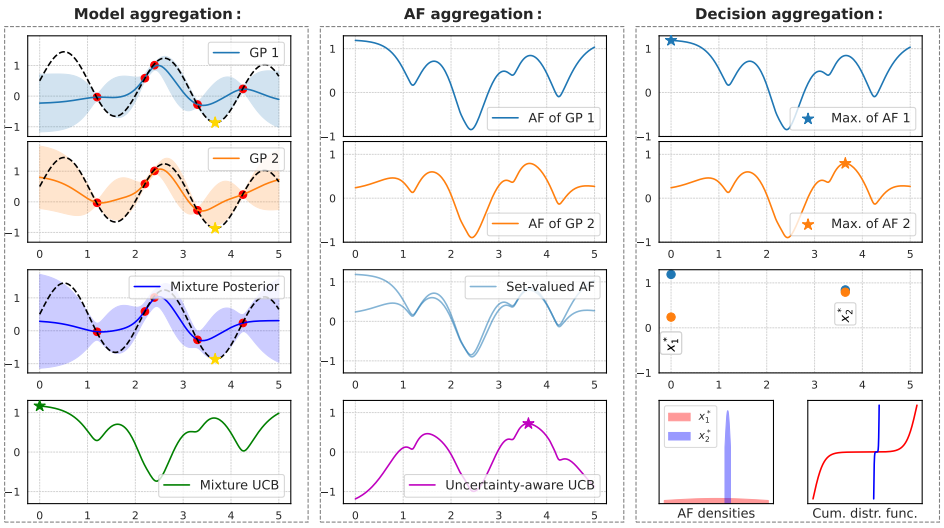

Figure 1: Aggregation can occur at the model-level (standard hierarchical BO aggregates posterior predictions [28]), at the acquisition-level, or at the decision-level: **Left:** Two GPs with distinct hyperparameters are combined into a mixture posterior, which is evaluated under the AF and maximized, yielding $\mathbf{x} \approx 0$ as the next candidate. **Middle:** The AF is evaluated under each GP separately, forming a set-valued AF whose spread reflects model disagreement. This spread can be leveraged (e.g., via $\mu - \mathrm{Var}$) to obtain an uncertainty-aware AF, whose maximizer is $\mathbf{x} \approx 3.7$. **Right:** Each AF is optimized independently, producing two maximizers, $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$. Their AF values form a set from which we select the candidate whose empirical cumulative distribution function (CDF) stochastically dominates; if none dominates, we choose the one with the higher mean AF value, here $\mathbf{x}_2^*$.

However, such aggregation can obscure disagreement among models. Two GPs may fit the data equally well, yet highlight different regions of uncertainty or potential optima. Averaging their predictions smooths over these conflicting beliefs, often leading to overconfident posteriors that no single model fully supports. In the context of BO, where the goal is to decide where to evaluate next, this inter-model disagreement encodes valuable information about epistemic uncertainty.

Motivated by ideas from imprecise probability theory [4], we propose to retain this information [47] by explicitly maintaining a set of GP models rather than collapsing them into one. Each GP posterior is evaluated separately under the AF, yielding an imprecise, set-valued AF whose spread quantifies model disagreement. Figure 1 illustrates this concept: instead of aggregating model predictions early (as in hierarchical BO), we postpone aggregation to later stages. This can occur either at the acquisition level, by combining AF values, or later at the decision level, where each AF is optimized independently and the next candidate is chosen from the resulting maximizers.

Contributions. (1) We propose the imprecise BO framework IABO, which preserves model disagreement by maintaining a set of plausible surrogate models rather than collapsing them into a single aggregated one. Aggregation is deferred and can occur either (i) at the acquisition-level, by combining AF values through an aggregated AF, or (ii) at the decision-level, by independently maximizing each AF and selecting the next candidate from the resulting maximizers. (2) Our framework is general and acquisition-function agnostic, seamlessly extending to standard choices such as Expected Improvement and Upper Confidence Bound. (3) Empirically, we demonstrate that deferring aggregation to later stages yields consistent performance gains across a range of benchmark problems, highlighting the value of explicitly modeling inter-surrogate disagreement in BO.

2 Hierarchical Bayesian Optimization

We consider the problem of sequentially optimizing a *black-box* function $f : \mathcal{X} \to \mathbb{R}$. We assume homoscedastic noise, observing $y_t = f(\mathbf{x}_t) + \xi_t$ with $\xi_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Our goal is to identify the global minimizer of f, i.e. $\mathbf{x}^\star = \arg\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$ while minimizing the number of evaluations In

sequential decision-making, we typically begin by specifying a prior over the unknown function, which encodes our initial belief. At time step t, we have a dataset of observations $\mathcal{D}_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^t$ to fit a surrogate model such as a Gaussian process (GP). A GP is fully specified by a prior mean function $m(\mathbf{x})$ and a kernel $k_\theta(\mathbf{x}, \mathbf{x}')$, parameterized by hyperparameters $\theta = \{l, \sigma_f, \sigma_\epsilon\}$, where $l \in \mathbb{R}^d$ are lengthscales, $\sigma_f \in \mathbb{R}$ is the output scale, and $\sigma_\epsilon \in \mathbb{R}$ is the noise variance. The GP posterior predictive mean and variance at a test point $\mathbf{x}$ at timestep t can then be derived analytically [38].

$$\mu_t(\mathbf{x}) = m(\mathbf{x}) + \mathbf{k}_t(\mathbf{x})^\top (\mathbf{K}_t + \Sigma_t)^{-1}(\mathbf{y}_t - m_t), \tag{1}$$

$$k_t(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_t(\mathbf{x})^\top (\mathbf{K}_t + \Sigma_t)^{-1}\mathbf{k}_t(\mathbf{x}'), \tag{2}$$

$$\sigma_t^2(\mathbf{x}) = k_t(\mathbf{x}, \mathbf{x}), \tag{3}$$

where $\mathbf{y}_t = [y_1, \ldots, y_t]^\top$, $m_t = [m(x_1), \ldots, m(x_t)]^\top$, $\mathbf{K}_t = [k_\theta(x_i, x_j)]_{i,j}$ is the kernel matrix over observed points, $\Sigma_t = \mathrm{diag}(\sigma_\epsilon^2, \ldots, \sigma_\epsilon^2)$, and $\mathbf{k}_t(\mathbf{x}) = [k_\theta(x_1, \mathbf{x}), \ldots, k_\theta(x_t, \mathbf{x})]^\top$.

Hierarchical Bayesian optimization [34, 42] offers a principled approach to estimating suitable hyperparameters θ. In this setting, hyperparameters are drawn from a hyperprior $\theta \sim P(\theta)$, and the predictive posterior over the latent function at a test point $\mathbf{x}$ is obtained by marginalizing over these hyperparameters: $p(f(\mathbf{x}) \mid \mathcal{D}_t) = \int p(f(\mathbf{x}) \mid \mathcal{D}_t, \theta)\, P(\theta \mid \mathcal{D}_t)\, d\theta$. The resulting hierarchical predictive posterior is a mixture of Gaussians [21], which is typically approximated by sampling hyperparameters $\theta^{(j)} \sim P(\theta \mid \mathcal{D}_t)$ and averaging the corresponding predictive posterior distributions: $p(f(\mathbf{x}) \mid \mathcal{D}_t) \approx \frac{1}{M} \sum_{j=1}^M p(f(\mathbf{x}) \mid \mathcal{D}_t, \theta^{(j)})$.

$$\mu_{\text{mix}}(\mathbf{x}) = \frac{1}{M}\sum_{j=1}^M \mu_j(\mathbf{x}), \quad \sigma^2_{\text{mix}}(\mathbf{x}) = \frac{1}{M}\sum_{j=1}^M \left(\sigma_j^2(\mathbf{x}) + \mu_j^2(\mathbf{x})\right) - \mu^2_{\text{mix}}(\mathbf{x}), \tag{4}$$

where $\mu_j(\mathbf{x})$ and $\sigma_j^2(\mathbf{x})$ are the predictive mean and variance of the j-th GP.

3 Related Work

Several works have proposed robust BO methods designed to reduce their sensitivity to GP hyperparameter settings. For instance, Berkenkamp et al. [10] adapt kernel hyperparameters iteratively during optimization. Bogunovic and Krause [12] introduce the enlarged-confidence GP-UCB algorithm, which augments the standard GP-UCB [43] with an additional exploration term. Wynne et al. [48] study the impact of misspecified smoothness assumptions and likelihoods in GPs and provide strategies for adjusting kernels and hyperparameters accordingly. More recently, Rodemann and Augustin [39] proposed imprecise GPs as surrogate models to mitigate prior mean misspecification. However, their approach is limited to constant means and does not account for uncertainty in kernel parameters.

Another line of work enhances robustness by maintaining multiple priors and selecting among them during optimization. Pautrat et al. [35] choose the prior that maximizes the product of its likelihood and AF value. Ziomek et al. [50] focus on time-varying domains and assume a set of expert-defined priors. They iteratively discard implausible ones and select at each step the prior whose posteriors yields the highest UCB value. Adachi et al. [1] frame the aggregation through the lens of social choice theory, treating each surrogate or AF as an agent with potentially biased preferences. Their dual voting mechanism combines noisy public votes with accurate private votes to correct for social influence and promote consensus among agents.

The methods most conceptually related to ours are ScoreBO and SAASBO. ScoreBO [23] defines an AF based on the discrepancy between the hierarchical posterior predictive distribution and the current posterior, averaging this measure over samples. SAASBO [17], in turn, samples hyperparameters from a sparsity-inducing prior over lengthscales and averages the resulting expected improvement values. Both approaches, therefore, perform aggregation at the acquisition-function level (cf. Figure 1), but only through simple averaging. Consequently, they neglect potential disagreement among models, since the variance of the AF values is not taken into account.

Table 1: Aggregated AFs with corresponding uncertainty preferences. Let $\mathbb{E}_q := \mathbb{E}_{q\in\mathcal{Q}(\mathbf{x})}[q]$ and $\mathrm{Var}_q := \mathrm{Var}_{q\in\mathcal{Q}(\mathbf{x})}[q]$ denote the expectation and variance over AF values at $\mathbf{x}$, with $\lambda > 0$.

Aggregated AF	$\mathbb{E}_q - \lambda\,\mathrm{Var}_q$	$\mathbb{E}_q$	$\mathbb{E}_q + \lambda\,\mathrm{Var}_q$	$\min\mathcal{Q}(x)$	$\mathrm{median}\,\mathcal{Q}(x)$	$\max\mathcal{Q}(x)$
Uncertainty preference	averse	neutral	affine	averse	neutral	affine

4 Imprecise Acquisitions

At the core of our framework, IABO, is the concept of imprecise AFs. Rather than relying on a single AF from one GP or a mixture posterior, we consider a set of plausible AFs, each computed from a different GP posterior. This captures both the risk inherent in a probabilistic model (e.g., a GP's predictive variance) and ambiguity, i.e., uncertainty about which model to trust [16]. Thus, proposing candidate points in BO becomes a problem of decision-making under imprecise probabilities (IP) [46], allowing us to leverage decades of research on decision criteria in this setting [9, 24, 2, 3, 25, 27, 45, 44, 22] tracing back to [16], see [44] for an overview. Formally, each hyperparameter sample $\theta^{(j)}$ produces a Gaussian process posterior $p(f(\mathbf{x}) \mid \mathcal{D}_t, \theta^{(j)}) := p(f^{(j)})$ with mean $\mu_{\theta^{(j)}}$ and variance $\sigma^2_{\theta^{(j)}}$. The AF evaluated under this posterior is $\alpha_{\theta^{(j)}} := \alpha\big(\mu_{\theta^{(j)}}, \sigma^2_{\theta^{(j)}}\big)$. Evaluating the AF across all sampled hyperparameters $\{\theta^{(j)}\}_{j=1}^M$ produces a set of plausible AF values for every $\mathbf{x}$. To decide on the next candidate for evaluation, i.e., which set is "the best", we propose two aggregation strategies operating at different levels, which can incorporate various risk preferences.

4.1 Acquisition-level aggregation

Definition 1 (Set of plausible acquisition values)**.** Given the set of AFs $\{\alpha_{\theta^{(j)}}\}_{j=1}^M$ we define for each candidate point $\mathbf{x} \in \mathcal{X}$

$$\mathcal{Q}(\mathbf{x}) := \{a_{\theta^{(1)}}(\mathbf{x}), \ldots, a_{\theta^{(M)}}(\mathbf{x})\} \tag{5}$$

the set of plausible AF values at $\mathbf{x}$ under posterior hyperparameter uncertainty.

Given such a set, we aggregate information by transforming each set into a single scalar value.

Definition 2 (Aggregated acquisition function)**.** Given a set of plausible acquisition values $\mathcal{Q}(\mathbf{x})$ at a candidate point $\mathbf{x} \in \mathcal{X}$, an aggregated acquisition function (AAF) is a mapping

$$\rho : \mathcal{Q}(\mathbf{x}) \;\to\; \rho(\mathcal{Q}(\mathbf{x})) \in \mathbb{R}, \tag{6}$$

which assigns a single real value to the set of plausible acquisition values of the given candidate.

We propose several variants in Table 1 that reflect different preferences: uncertainty-neutral, uncertainty-affine, or uncertainty-averse. The first AAF, $\mathbb{E}_q + \lambda\,\mathrm{Var}_q$, is inspired by UCB. Here, $\mathbb{E}_q$ denotes the average AF value across models (risk), while Var_q quantifies inter-model disagreement (ambiguity). In particular, setting $\lambda = 0$ recovers the SAASBO acquisition strategy [17] with UCB as the AF. The $\min\mathcal{Q}(x)$ criterion corresponds to the Γ-maximin criterion [41], representing a worst-case, uncertainty-averse strategy: it assumes the prior producing the lowest acquisition value is most suitable.

Lemma 3 (Robust max–min dominance)**.** *Let $\{\theta^{(j)}\}_{j=1}^M$ denote the set of hyperparameters with corresponding GPs and $\mathcal{Q}(\mathbf{x}) = \{a_{\theta^{(1)}}(\mathbf{x}), \ldots, a_{\theta^{(M)}}(\mathbf{x})\}$ the set of plausible AF values at a candidate $\mathbf{x}$ as defined above. Define the robust (max–min) choice as*

$$\mathbf{x}_{\mathrm{robust}} \;=\; \arg\max_{\mathbf{x}\in\mathcal{X}} \min_{j=1,\ldots,M} a_{\theta^{(j)}}(\mathbf{x}) \;=\; \arg\max_{\mathbf{x}\in\mathcal{X}} \min_{q\in\mathcal{Q}(\mathbf{x})} q.$$

Then, for any $x' \in \mathcal{X}$,

$$\min_{j=1,\ldots,M} a_{\theta^{(j)}}(\mathbf{x}') \;\le\; \min_{j=1,\ldots,M} a_{\theta^{(j)}}(\mathbf{x}_{\mathrm{robust}}). \tag{7}$$

Proof. Define the worst-case acquisition value $\gamma(\mathbf{x}) := \min_{j=1,\ldots,M} a_{\theta^{(j)}}(\mathbf{x})$. By definition of $\mathbf{x}_{\mathrm{robust}}$,

$$\gamma(x_{\mathrm{robust}}) = \max_{x\in\mathcal{X}} \gamma(x).$$

Hence, for any $\mathbf{x}' \in \mathcal{X}$, $\gamma(x') \le \gamma(x_{\mathrm{robust}})$, which directly implies (7). □

This lemma formalizes that the robust (max–min) selection guarantees a worst-case acquisition value that is at least as high as that of any other candidate, including the point obtained by maximizing the acquisition function under the aggregated posterior (as in standard hierarchical Bayesian optimization). Intuitively, hierarchical BO optimizes the expected acquisition value under the hyperparameter distribution $p(\theta)$, while the robust formulation optimizes its worst-case value across all plausible models. This ensures resilience to model misspecification, favoring solutions that remain reliable even when some surrogate models poorly reflect the true data-generating process.

Conversely, $\max \mathcal{Q}(x)$ implements an uncertainty-affine or optimistic approach, while the median rule provides a risk-neutral alternative that is robust to outliers among the plausible AF values. The AF-level procedure is detailed in Algorithm 1 in the appendix. To select the next candidate, one can maximize the AAF using standard optimization methods, since the expectation–variance mapping is fully continuous, whereas the min, max, and median mappings are piecewise-continuous. However, all of these risk mappings reduce the full set of AF values to a single scalar, inevitably compressing information that could otherwise inform more robust decision-making. One way to retain this information is to postpone aggregation to a later stage, as follows.

4.2 Decision-level aggregation

Our second strategy first maximizes each AF independently, producing a set of candidate maximizers.

Definition 4 (Set of maximizers)**.** Let $\mathbf{x}_j^* \in \arg\max_{\mathbf{x}} a_{\theta^{(j)}}(\mathbf{x})$, then we define the set of AF maximizers as

$$\mathcal{X}^* := \{\arg\max_{\mathbf{x}} a_{\theta^{(j)}}(\mathbf{x})\}_{j=1}^M = \{\mathbf{x}_j^*\}_{j=1}^M. \tag{8}$$

If the AF is interpreted as expected utility, the set $\mathcal{X}^*$ consists of points that maximize Walley's maximality criterion [46]: each point maximizes expected utility for at least[1] one probability measure (here: one $p(f^{(j)})$) from the ones under consideration (here: all $\{p(f^{(j)})\}_{j=1}^M$). A simple strategy is to select the candidate closest to the median of $\mathcal{X}^*$, but this only considers the candidates' positions in the $\mathcal{X}$ domain, ignoring the AF values themselves. To incorporate this additional information, we define a mapping that directly compares two sets of AF values when selecting among candidates.

Definition 5 (Dominance mapping)**.** Given two sets of plausible AF values (5), $\mathcal{Q}(\mathbf{x}_i^*)$ and $\mathcal{Q}(\mathbf{x}_k^*)$, corresponding to two candidate maximizers $\mathbf{x}_i^*, \mathbf{x}_k^* \in \mathcal{X}^*$, a dominance mapping

$$\varphi : \mathcal{Q}(\mathbf{x}_i^*) \times \mathcal{Q}(\mathbf{x}_k^*) \;\to\; \varphi(\mathcal{Q}(\mathbf{x}_i^*), \mathcal{Q}(\mathbf{x}_k^*)) \in \{0, 1\}, \tag{9}$$

assigns a binary value that quantifies the relative dominance between these two sets. Specifically, $\varphi(\mathcal{Q}(\mathbf{x}_i^*), \mathcal{Q}(\mathbf{x}_k^*)) = 1$ if $\mathbf{x}_i^*$ is dominated by $\mathbf{x}_k^*$, and 0 otherwise.

Note that since $\mathcal{X}^*$ contains at most M candidates, we only need to compute the dominance mapping for $M \times (M-1)$ pairwise comparisons, far fewer than would be required over the full domain $\mathcal{X}$. While pairwise comparisons can be implemented in various ways, we primarily rely on first- and second-order stochastic dominance criteria [20], which induce a partial order over random variables. Here, the acquisition values obtained under different priors are interpreted as samples from a random variable. To apply these criteria, we approximate the distribution of $\mathcal{Q}(\mathbf{x}_i^*)$ using its empirical cumulative distribution function $\hat{F}_i(q) = \frac{1}{M}\sum_{j=1}^M \mathbb{I}\{q_i^{(j)} \le q\}$, where $q_i^{(j)} \in \mathcal{Q}(\mathbf{x}^*)$.

First-order stochastic dominance (FSD): Candidate $\mathbf{x}_i^*$ first-order stochastically dominates $\mathbf{x}_k^*$ if $\hat{F}_i(q) \le \hat{F}_k(q) \quad \forall q \in \mathbb{R}$, with strict inequality for at least one q. Since we want to maximize the AF, $\mathbf{x}_i^*$ dominates $\mathbf{x}_k^*$ because it places more probability mass on higher AF values.

Second-order stochastic dominance (SSD): Candidate $\mathbf{x}_i^*$ second-order stochastically dominates $\mathbf{x}_k^*$ if $\int_{-\infty}^q \hat{F}_i(t)\,dt \le \int_{-\infty}^q \hat{F}_k(t)\,dt \quad \forall q \in \mathbb{R}$, with strict inequality for at least one q. For each pair of candidates, we first check for first-order stochastic dominance. If neither candidate exhibits FSD, we evaluate second-order stochastic dominance. For each candidate, we count how many times it is dominated by other candidates, under either FSD or SSD. The final selection rule is straightforward and interpretable: we choose the candidate $\mathbf{x}_i^* \in \mathcal{X}^*$ that is dominated the fewest times. This procedure naturally favors candidates whose acquisition values are robustly high across the set of plausible models, effectively balancing risk and ambiguity.

[1] It may happen that $\mathbf{x}_i^* = \mathbf{x}_k^*$ for $i \neq k$. Hence, $\mathcal{X}^*$ contains at most M elements

Definition 6 (Dominance score)**.** Given a set of maximizers $\mathcal{X}^* = \{\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*\}$ and a dominance mapping φ as defined in 9, the overall dominance score of a candidate $\mathbf{x}_i^*$ is defined as

$$\Phi(\mathbf{x}_i^*) = \sum_{k \neq i} \varphi(\mathcal{Q}(\mathbf{x}_i^*), \mathcal{Q}(\mathbf{x}_k^*)). \tag{10}$$

The candidate with the lowest dominance score is then selected, i.e., the one dominated the fewest times. In decision-theoretic terms, the set of undominated candidates $\{\mathbf{x}_i^* : \Phi(\mathbf{x}_i^*) = 0\}$ is called admissible [8]. If multiple candidates are tied, ties can be resolved using one of the following strategies: (1) select randomly among them, (2) choose the candidate with the lowest empirical mean, or (3) choose the candidate with the lowest empirical variance. The complete procedure for decision-level aggregation is presented in Algorithm 2 in the appendix.

4.3 Refinement of plausible hyperparameter samples

The aggregated acquisition functions Γ-maximin and $\max \mathcal{Q}(x)$, discussed in Section 4.1, can be overly pessimistic or overly optimistic, respectively. To mitigate this, Cattaneo [13] proposed an attenuated approach known as the α-cut or soft revision [19, 5], which has since been applied in classification and pseudolabeling contexts [40, 15, 30]. We adopt a similar strategy for selecting plausible hyperparameters. The core idea is to retain only those samples whose marginal likelihood (evidence) is sufficiently close to the best observed. Concretely, for each sampled $\theta^{(j)}$, we compute its marginal likelihood and normalize it relative to the maximum across all samples. We then retain only those $\theta^{(j)}$ whose relative likelihood exceeds a threshold γ:

$$p(\mathcal{D} \mid \theta^{(j)}) \geq \gamma \max_k p(\mathcal{D} \mid \theta^{(k)}). \tag{11}$$

5 Extensions

Risk-averse objective. In portfolio optimization, investors seek to avoid outcomes below a minimal acceptable threshold $f_{\min}$, which are considered disproportionately undesirable. Following [32], we employ a prospect-theory-inspired reward to formalize this risk-averse preference.

$$R_T^{\text{risk-averse}} = \sum_{t=1}^{T} \Big[(f(\mathbf{x}_t) - f_{\min})^{\alpha} \, \mathbf{1}_{\{f(\mathbf{x}_t) \geq f_{\min}\}} - \lambda \, (f_{\min} - f(\mathbf{x}_t))^{\beta} \, \mathbf{1}_{\{f(\mathbf{x}_t) < f_{\min}\}} \Big],$$

where $\alpha, \beta \in (0, 1]$ control the curvature of gains and losses, and $\lambda > 1$ represents loss aversion, amplifying the penalty for outcomes below $f_{\min}$.

Extending on Table 1, we propose the following AAF:

Risk-averse acquisition (averaged). To emphasize robustness to worst-case outcomes while normalizing across models, we define a PT-inspired acquisition function by averaging over the set of plausible acquisition values. Let $\mathcal{Q}(\mathbf{x}) = \{a_{\theta^{(1)}}(\mathbf{x}), \ldots, a_{\theta^{(M)}}(\mathbf{x})\}$ and let $f_{\min}$ be the minimal acceptable threshold. **to write: that fmin has to be on the same scale as the q, so maybe some normalization?** The averaged risk-averse acquisition is

$$a^{\text{risk}}(\mathbf{x}) = \frac{1}{M} \sum_{q \in \mathcal{Q}(\mathbf{x})} \Big[(q - f_{\min})^{\alpha} \, \mathbf{1}_{\{q \geq f_{\min}\}} - \lambda \, (f_{\min} - q)^{\beta} \, \mathbf{1}_{\{q < f_{\min}\}} \Big],$$

where $\alpha, \beta \in (0, 1]$ control the curvature of gains and losses, and $\lambda > 1$ amplifies penalties for outcomes below $f_{\min}$.

The next candidate is chosen as

$$\mathbf{x}_{\text{next}} = \arg\max_{\mathbf{x} \in \mathcal{X}} a^{\text{risk}}(\mathbf{x}).$$

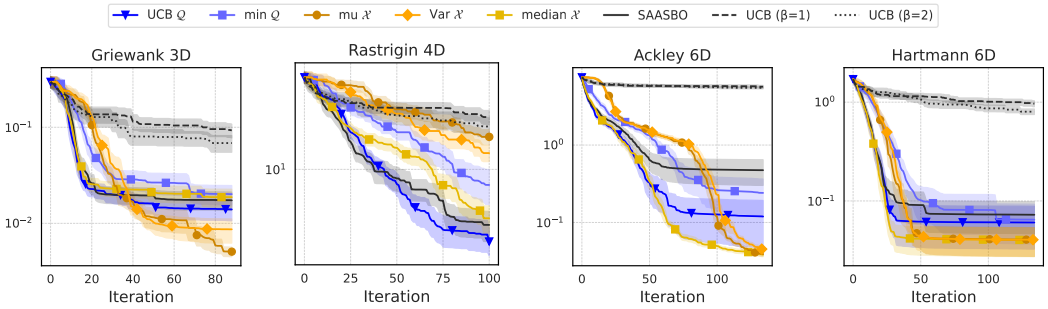

Figure 2: Log-regret averaged over 20 repetitions, using UCB as AF. Our approaches (UCB $\mathcal{Q}$, $\min \mathcal{Q}$, mu $\mathcal{X}$, Var $\mathcal{X}$, median $\mathcal{X}$) use the full set of GPs. The shaded area represents one standard error.

Remarks.

- Averaging normalizes the acquisition across the number of models, ensuring scale consistency.
- The extreme losses are still emphasized: outcomes below $f_{\min}$ are amplified by λ and the curvature β, enforcing robustness to worst-case scenarios.
- Setting $\alpha = \beta = 1$ and $\lambda = 1$ recovers the standard robust max–min acquisition:

$$\mathbf{x}_{\text{next}} = \arg\max_{\mathbf{x}} \min_{q \in \mathcal{Q}(\mathbf{x})} q.$$

6 Experiments

We conduct an experimental study on several standard benchmark functions: Griewank (3D), Rastrigin (4D), Ackley (6D), and Hartmann (6D). For clarity, we focus on two acquisition-level strategies: $\mathbb{E}_q + \mathrm{Var}_q$ (denoted UCB $\mathcal{Q}$ in the plots) and $\min \mathcal{Q}(\mathbf{x})$ (min $\mathcal{Q}$). We also present results for decision-level strategies based on stochastic dominance: selecting the candidate with the highest mean (mu $\mathcal{X}$) or lowest variance (Var $\mathcal{X}$), as well as the median rule, which chooses the candidate closest to the median maximizer (median $\mathcal{X}$).

In Figure 2, we use UCB as the AF and compare against standard UCB (single GP with hyperparameters learned via marginal likelihood maximization) with exploration parameters 1 and 2, as well as SAASBO with UCB [17]. In Figure 3, we use EI as the AF and compare against standard EI (single GP with hyperparameters learned via marginal likelihood maximization) and SAASBO with EI [17]. Additional experimental details, results for other methods and test functions, and analyses of the α-cut are provided in Appendix A.

Overall, our decision-level strategies perform strongly. Particularly, the median $\mathcal{X}$ strategy combined with EI consistently ranks among the top approaches across benchmarks, except for Griewank under both AFs. Notably, for Rastrigin 4D and Ackley 6D, this combination substantially outperforms all other methods. The acquisition-level strategy (UCB $\mathcal{Q}$) is also highly competitive to SAASBO, matching or surpassing it on Griewank 3D, Rastrigin 4D, and Ackley 6D with UCB as the AF. Standard EI and UCB are generally outperformed by our approaches, except for Rastrigin (4D) under EI. We note that decision-level strategies introduce additional computation overhead, as they require optimizing M separate acquisition functions, but this cost is offset by the consistent performance gains across benchmarks.

7 Conclusion and Future Work

We introduced the imprecise BO framework (IABO), which accounts for hyperparameter uncertainty by maintaining a set of plausible GP models. By evaluating the AF under each posterior separately, we obtain a set-valued AF whose spread naturally reflects model disagreement. We proposed acquisition-level and decision-level strategies to guide candidate selection under different uncertainty preferences. Empirical results on standard benchmark functions show that our methods, particularly

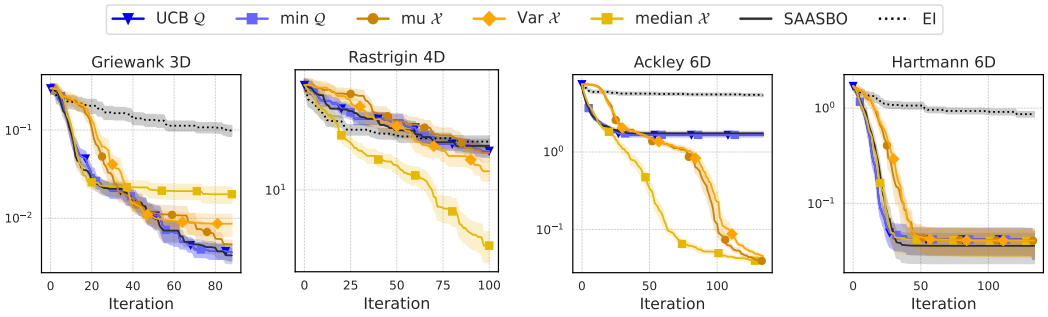

Figure 3: Log-regret averaged over 20 repetitions, using EI as AF. Our approaches (UCB $\mathcal{Q}$, min $\mathcal{Q}$, mu $\mathcal{X}$, Var $\mathcal{X}$, median $\mathcal{X}$) use the full set of GPs. The shaded area represents one standard error.

the decision-level strategies, consistently outperform standard UCB and are competitive with or superior to SAASBO across most tasks.

Several promising directions remain for future research. First, the framework could be extended to more advanced AFs that, e.g., handle heteroscedastic noise such as RAHBO [31]. Second, alternative decision criteria such as Levi's E-admissibility [29, 26] could be explored to capture different aspects of model uncertainty or accommodate user-specific risk preferences. Finally, investigating the impact of different hyperpriors—including poorly chosen ones as considered in ScoreBO [23] and implementing ScoreBO as an additional baseline could provide further insights into the robustness of our approach.

Broader Impact This work advances robust and adaptive Bayesian optimization methods, with potential applications in AutoML, materials science, and healthcare. As with any optimization tool, care must be taken to avoid biased objectives or unintended outcomes in sensitive domains. Beyond that, the authors foresee no significant societal or environmental risks.

References

[1] Adachi, M., Chau, S. L., Xu, W., Singh, A., Osborne, M. A., and Muandet, K. (2025). Bayesian optimization for building social-influence-free consensus. *CoRR*, abs/2502.07166.

[2] Augustin, T. (2001). On decision making under ambiguous prior and sampling information. In *ISIPTA*, volume 1, pages 9–16.

[3] Augustin, T. (2004). Optimal decisions under complex uncertainty–basic notions and a general algorithm for data-based decision making with partial prior knowledge described by interval probability. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik: Applied Mathematics and Mechanics*, 84(10-11):678–687.

[4] Augustin, T., Coolen, F. P. A., de Cooman, G., and Troffaes, M. C. M. (2014). *Introduction to Imprecise Probabilities*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd.

[5] Augustin, T. and Schollmeyer, G. (2021). Comment: on focusing, soft and strong revision of Choquet capacities and their role in statistics. *Statistical Science*, 36(2):205–209.

[6] Balandat, M., Karrer, B., Jiang, D., Daulton, S., Letham, B., Wilson, A., and Bakshy, E. (2020). Botorch: A framework for efficient monte-carlo Bayesian optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H., editors, *Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems (NeurIPS'20)*. Curran Associates.

[7] Benjamins, C., Graf, H., Segel, S., Deng, D., Ruhkopf, T., Hennig, L., Basu, S., Mallik, N., Bergman, E., Chen, D., Clément, F., Feurer, M., Eggensperger, K., Hutter, F., Doerr, C., and Lindauer, M. (2025). carps: A framework for comparing N hyperparameter optimizers on M benchmarks. *CoRR*.

[8] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis, Second Edition*. Springer.

[9] Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., et al. (1994). An overview of robust bayesian analysis. *Test*, 3(1):5–124.

[10] Berkenkamp, F., Schoellig, A., and Krause, A. (2019). No-regret Bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research*, 20:50:1–50:24.

[11] Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A., Deng, D., and Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1484.

[12] Bogunovic, I. and Krause, A. (2021). Misspecified gaussian process bandit optimization. In [37], pages 3004–3015.

[13] Cattaneo, M. E. G. V. (2014). A continuous updating rule for imprecise probabilities. In Laurent, A., Strauss, O., Bouchon-Meunier, B., and Yager, R. R., editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems - 15th International Conference, IPMU 2014, Montpellier, France, July 15-19, 2014, Proceedings, Part III*.

[14] Chen, Y., Huang, A., Wang, Z., Antonoglou, I., Schrittwieser, J., Silver, D., and de Freitas, N. (2018). Bayesian optimization in alphago. *arXiv:1812.06855 [cs.LG]*.

[15] Dietrich, S., Rodemann, J., and Jansen, C. (2024). Semi-supervised learning guided by the generalized bayes rule under soft revision. In *International Conference on Soft Methods in Probability and Statistics*, pages 110–117. Springer.

[16] Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 75(4):643–669.

[17] Eriksson, D. and Jankowiak, M. (2021). High-dimensional bayesian optimization with sparse axis-aligned subspaces. In Peters, J. and Sontag, D., editors, *Proceedings of The 37th Uncertainty in Artificial Intelligence Conference (UAI'21)*, pages 493–503. PMLR.

[18] Garnett, R. (2023). *Bayesian Optimization*. Cambridge University Press. Available for free at `https://bayesoptbook.com/`.

[19] Gong, R. and Meng, X.-L. (2021). Judicious judgment meets unsettling updating. *Statistical Science*, 36(2):169–190.

[20] Hadar, J. and Russell, W. R. (1969). Rules for ordering uncertain prospects. *The American Economic Review*.

[21] Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*

[22] Huntley, N., Hable, R., and Troffaes, M. C. M. (2014). Decision making. In Augustin, T., Coolen, F. P. A., de Cooman, G., and Troffaes, M. C. M., editors, *Introduction to Imprecise Probabilities*, pages 190–206. Wiley.

[23] Hvarfner, C., Hellsten, E., Hutter, F., and Nardi, L. (2023). Self-correcting bayesian optimization through bayesian active learning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Proceedings of the 37th International Conference on Advances in Neural Information Processing Systems (NeurIPS'23)*. Curran Associates.

[24] Jaffray, J.-Y. (1999). Rational decision making with imprecise probabilities. In *ISIPTA*, volume 99, pages 324–332.

[25] Jansen, C., Schollmeyer, G., and Augustin, T. (2018). Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences. *International Journal of Approximate Reasoning*, 98:112–131.

[26] Jansen, C., Schollmeyer, G., and Augustin, T. (2022). Quantifying degrees of e-admissibility in decision making with imprecise probabilities. In *Reflections on the Foundations of Probability and Statistics: Essays in Honor of Teddy Seidenfeld*, pages 319–346. Springer.

[27] Jansen, C., Schollmeyer, G., Rodemann, J., and Augustin, T. (2025). Empirical decision theory. *Poster presented at the International Symposium on Imprecise Probabilities (ISIPTA)*.

[28] Lalchand, V. and Rasmussen, C. E. (2019). Approximate inference for fully bayesian gaussian process regression. volume 118 of *Proceedings of Machine Learning Research*, pages 1–12. PMLR.

[29] Levi, I. (1974). On indeterminate probabilities. *The Journal of Philosophy*, 71(13):391–418.

[30] Löhr, T., Hofman, P., Mohr, F., and Hüllermeier, E. (2025). Credal prediction based on relative likelihood. *CoRR*.

[31] Makarova, A., Usmanova, I., Bogunovic, I., and Krause, A. (2021). Risk-averse heteroscedastic bayesian optimization. In [37], pages 17235–17245.

[32] Maringer, D. (2008). *Risk Preferences and Loss Aversion in Portfolio Optimization*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[33] Neiswanger, W. and Ramdas, A. (2021). Uncertainty quantification using martingales for misspecified gaussian processes. volume 132 of *Proceedings of Machine Learning Research*, pages 963–982. PMLR.

[34] Osborne, M. A. (2010). *Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature*. Phd thesis, University of Oxford, Oxford, UK.

[35] Pautrat, R., Chatzilygeroudis, K., and Mouret, J.-B. (2018). Bayesian optimization with automatic prior selection for data-efficient direct policy search.

[36] Pyzer-Knapp, E. O. (2018). Bayesian optimization for accelerated drug discovery. *IBM J. Res. Dev.*, 62(6):2:1–2:7.

[37] Ranzato, M., Beygelzimer, A., Nguyen, K., Liang, P., Vaughan, J., and Dauphin, Y., editors (2021). *Proceedings of the 35th International Conference on Advances in Neural Information Processing Systems (NeurIPS'21)*. Curran Associates.

[38] Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

[39] Rodemann, J. and Augustin, T. (2024). Imprecise bayesian optimization. *Knowledge-Based Systems*, pages 112–186.

[40] Rodemann, J., Jansen, C., Schollmeyer, G., and Augustin, T. (2023). In all likelihoods: robust selection of pseudo-labeled data. In Miranda, E., Montes, I., Quaeghebeur, E., and Vantaggi, B., editors, *International Symposium on Imprecise Probability: Theories and Applications*, volume 215, pages 412–425. PMLR.

[41] Seidenfeld, T. (2004). A contrast between two decision rules for use with (convex) sets of probabilities: Γ-maximin versus E-admissibility. *Synthese*, 140.

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

[43] Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In Fürnkranz, J. and Joachims, T., editors, *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*, pages 1015–1022. Omnipress.

[44] Troffaes, M. C. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29.

[45] Utkin, L. V. and Augustin, T. (2005). Powerful algorithms for decision making under partial prior information and general ambiguity attitudes. In *ISIPTA*, volume 5, pages 349–358.

[46] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall.

[47] Walter, G. and Coolen, F. P. A. (2016). Sets of priors reflecting prior-data conflict and agreement. volume 610 of *Communications in Computer and Information Science*, pages 153–164. Springer.

[48] Wynne, G., Briol, F.-X., and Girolami, M. (2021). Convergence guarantees for gaussian process means with misspecified likelihoods and smoothness. *Journal of Machine Learning Research*, 22(123):1–40.

[49] Zhang, Y., Apley, D. W., and Chen, W. (2020). Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Scientific Reports*, 10(4924).

[50] Ziomek, J., Adachi, M., and Osborne, M. A. (2025). Time-varying gaussian process bandits with unknown prior. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2025, Mai Khao, Thailand, 3-5 May 2025*.

A Appendix

We provide pseudocode for IABO for the two aggregation variants, acquisition-level and decision-level, in Algorithms 1 and 2, respectively.

Algorithm 1 IABO: acquisition-level

Require: AF α, uncertainty mapping ρ
for $t = 0, \dots, T$ **do**
  for $j = 1, \dots, M$ **do**
    $\theta^{(j)} \sim P(\theta \mid \mathcal{D}_t)$
    Compute $a_{\theta^{(j)}}$ from $p(f^{(j)})$
  end for
  Compute $\mathcal{Q}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ (5)
  $\mathbf{x}_t \in \arg\max_{\mathbf{x} \in \mathcal{X}} \rho(\mathcal{Q}(\mathbf{x}))$ (6)
  Observe $y_t = f(\mathbf{x}_t) + \xi_t$
  $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(\mathbf{x}_t, y_t)\}$
end for
Return: $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathcal{D}_T} f(\mathbf{x})$

Algorithm 2 IABO: decision-level

Require: AF α, dominance mapping φ
for $t = 0, \dots, T$ **do**
  for $j = 1, \dots, M$ **do**
    $\theta^{(j)} \sim P(\theta \mid \mathcal{D}_t)$
    Compute $a_{\theta^{(j)}}$ from $p(f^{(j)})$
  end for
  Compute $\mathcal{X}^*$ (8) and φ (9)
  $\mathbf{x}_t \in \arg\min_{\mathbf{x} \in \mathcal{X}^*} \Phi(\mathbf{x})$ (10)
  Observe $y_t = f(\mathbf{x}_t) + \xi_t$
  $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(\mathbf{x}_t, y_t)\}$
end for
Return: $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathcal{D}_T} f(\mathbf{x})$

All experiments were conducted using 2 CPU cores and 8 GiB of RAM. The HPC nodes utilized for the computations are equipped with two AMD Milan 7763 processors and a total of 256 GiB of main memory. The total computational time for the experiments, as tracked in the database, is approximately 1.13 CPU years. We provide code here `https://anonymous.4open.science/r/imprecise_acquisitions-5E61/README.md`.

We present additional results for the Branin (2D) and Rastrigin (8D) functions using EI and UCB in Figures 4 and 5, respectively. For UCB, we also include results with the α-cut variant in Figure 6.

All surrogate models employ a Matérn 2.5 kernel. For our methods, we adopt the same hyperprior as SAASBO to ensure a fair comparison. All experiments are implemented in BoTorch [6] (v0.14.0), a widely used Python framework for Bayesian optimization.

Following the SAASBO setup, we use the default parameters for hyperparameter sampling: a warm-up phase of 512 iterations, generation of 256 samples, and thinning by a factor of 16 due to sample correlation, resulting in 16 GP models. Our methods likewise operate with 16 GPs.

We set the observation noise to a small value, $\sigma_\epsilon^2 = 1 \times 10^{-3}$, primarily to avoid numerical instabilities. Each experiment is averaged over 20 random seeds (0–19), with optimization budgets of 88, 100, and 134 for problem dimensions 3, 4, and 6, respectively, following the setup from CARP-S [7]. The initial design consists of 20% of the optimization budget, generated using Sobol sequences.

With EI as the AF, our decision-level strategies, especially the median aggregation over $\mathcal{X}$ performs particularly well, as seen for the Rastrigin (4D, 8D), Ackley (6D), and Rosenbrock (4D) functions. The AF-level approaches perform comparably to SAASBO, except for Branin 2D. This is like due to the fact that SAASBO effectively corresponds to the mean of $\mathcal{Q}$, hence differing only in the choice of risk mapping. Standard EI is consistently outperformed by our methods, except on the Branin (2D) and Rastrigin (4D) problems.

With UCB as the AF, the results show a somewhat different trend, a shown in Figure 6. The decision-level approaches perform less favorably overall. The $\max \mathcal{Q}$ strategy performs best on the Ackley (6D) function, corresponding to a uncertainty-affine behavior.

In Figure 6, we show results for UCB with an α-cut at $\alpha = 0.05$. Interestingly, the differences are relatively minor, except that the $\min(\mathcal{Q})$ strategy leads to improvements for some functions, such as Branin (2D) and Griewank (3D). One possible reason is that, with only 16 GPs, the $\alpha = 0.05$ threshold filters out only a small number of models, limiting its overall effect.

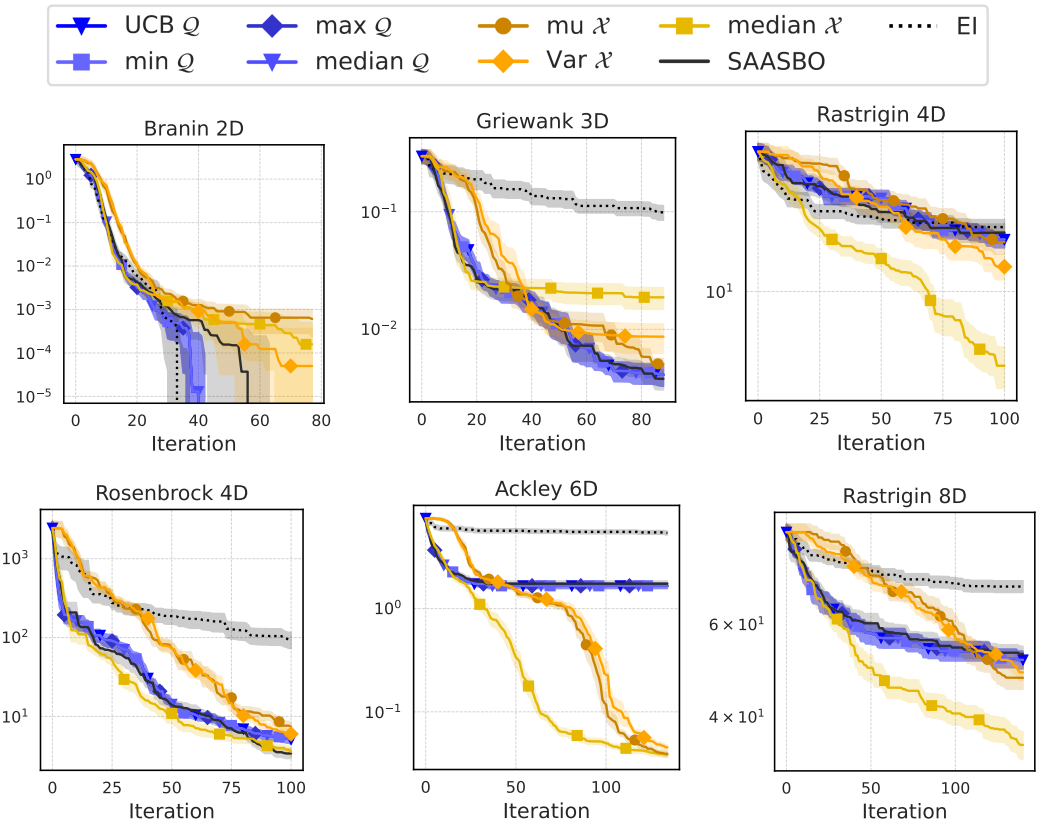

Figure 4: Log-regret averaged over 20 repetitions, using EI as AF. Our approaches (UCB $\mathcal{Q}$, $\min \mathcal{Q}$, mu $\mathcal{X}$, Var $\mathcal{X}$, median $\mathcal{X}$) use the full set of GPs. The shaded area represents one standard error.

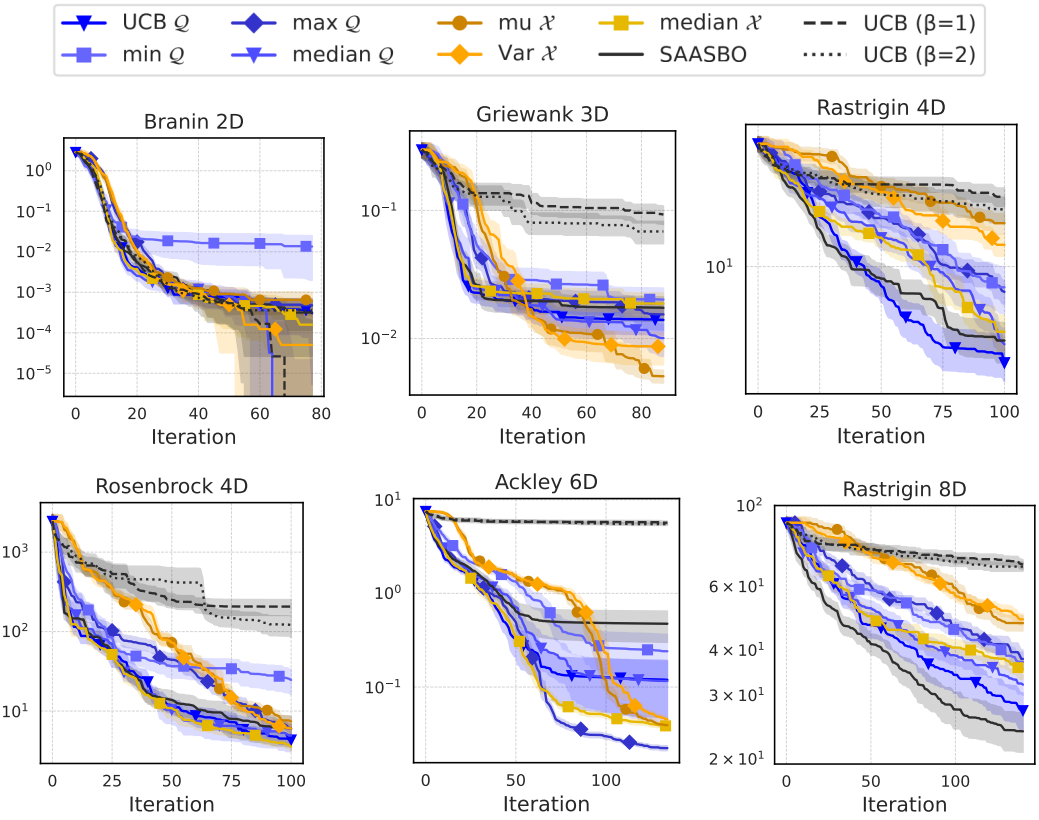

Figure 5: Log-regret averaged over 20 repetitions, using UCB as AF. Our approaches (UCB $\mathcal{Q}$, min $\mathcal{Q}$, mu $\mathcal{X}$, Var $\mathcal{X}$, median $\mathcal{X}$) use the full set of GPs. The shaded area represents one standard error.

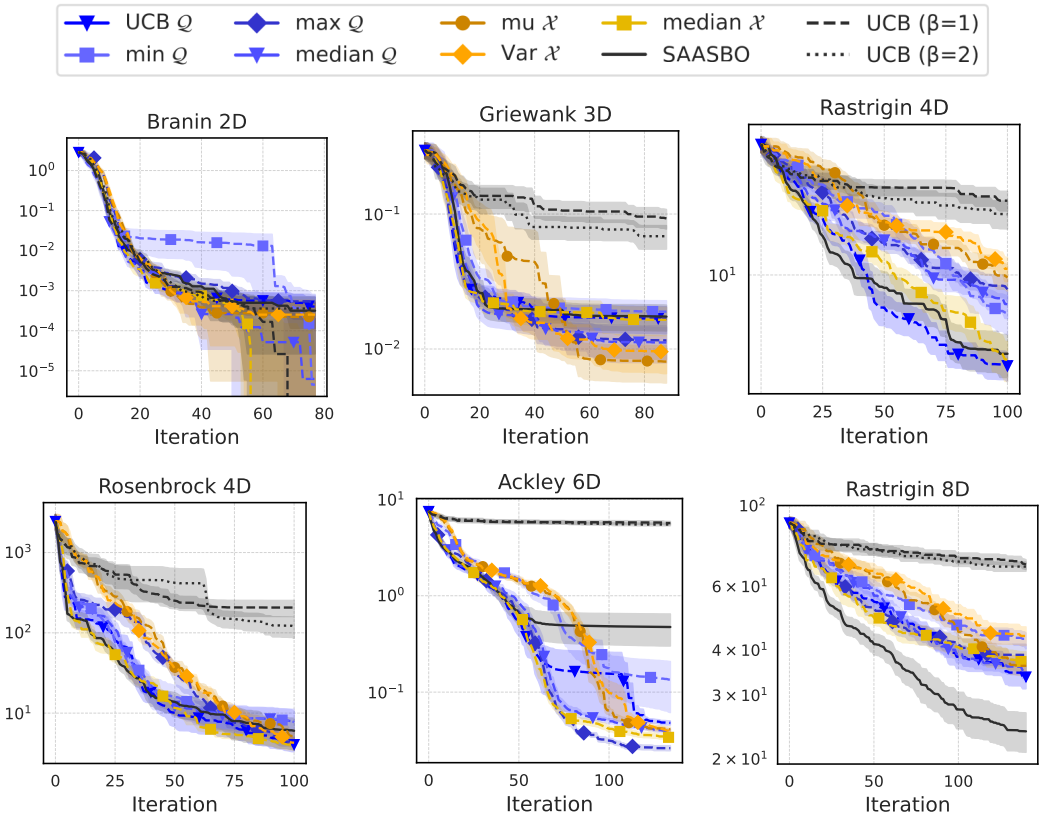

Figure 6: Log-regret averaged over 20 repetitions, using UCB as AF. Our approaches (UCB $\mathcal{Q}$, min $\mathcal{Q}$, max $\mathcal{Q}$, median $\mathcal{Q}$, mu $\mathcal{X}$, Var $\mathcal{X}$, median $\mathcal{X}$) use the set of GPs after the α-cut from Section 4.3. The shaded area represents one standard error.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state that the paper's main contribution is a new framework for using imprecise acquisition functions in Bayesian optimization, i.e. decision-making under imprecise probabilities. The claims made are consistent with the experimental results, and the paper does not overstate its contributions.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 7.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In the Appendix A.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide a link to an anonymized GitHub repository in Appendix A, which includes the code, as well as detailed instructions on how to set up the Python environment and execute the experiments.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In the Appendix A.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Standard errors are plotted.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Specified in the Appendix A.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presented in this paper fully adheres to the NeurIPS Code of Ethics. All experiments use publicly available, synthetic datasets, there are no human subjects or sensitive personal data involved, and the methods and results are reported transparently.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Section 7.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Botorch is cited.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]