Bayesian Optimization with Conformal Prediction Sets

Samuel Stanton[1,2] **Wesley Maddox**[2] **Andrew Gordon Wilson**[2]
Prescient Design, Genentech[1] New York University[2]

Abstract

Bayesian optimization is a coherent, ubiquitous approach to decision-making under uncertainty, with applications including multi-arm bandits, active learning, and black-box optimization. Bayesian optimization selects decisions (i.e. objective function queries) with maximal expected utility with respect to the posterior distribution of a Bayesian model, which quantifies reducible, epistemic uncertainty about query outcomes. In practice, subjectively implausible outcomes can occur regularly for two reasons: 1) model misspecification and 2) covariate shift. Conformal prediction is an uncertainty quantification method with coverage guarantees even for misspecified models and a simple mechanism to correct for covariate shift. We propose conformal Bayesian optimization, which directs queries towards regions of search space where the model predictions have guaranteed validity, and investigate its behavior on a suite of black-box optimization tasks and tabular ranking tasks. In many cases we find that query coverage can be significantly improved without harming sample-efficiency.

1 INTRODUCTION

Bayesian optimization (BayesOpt) is a popular strategy to focus data collection towards improving a specific objective, such as discovering useful new materials or drugs (Terayama et al., 2021; Wang and Dowling, 2022). BayesOpt relies on a Bayesian model of the objective (a surrogate model) to select new observations (queries) that maximize the user's expected utility. If the surrogate does not fit the objective well, then the expected utility of new queries may not correspond well at all to their *actual* utility, leading to little or no improvement in the objective value after many rounds of data collection.

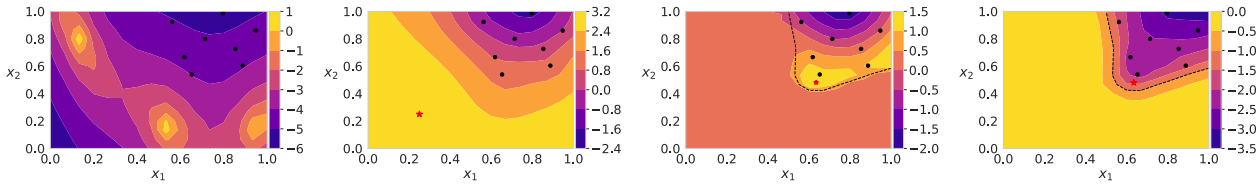
Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

The most practical way to check how well the surrogate fits the objective is to compute its accuracy on a random heldout subset of the training data. Unfortunately such a holdout set is not at all representative of points we are likely to query since the goal is to find queries that are *better* than the training data in some way. In other words there is *feedback covariate shift* between the likely query points and the existing training data which degrades the accuracy of the surrogate (Fannjiang et al., 2022). Even without covariate shift, we cannot guarantee the accuracy of the surrogate predictions at all, and instead can only hope that the predictions are accurate enough to provide a useful signal for data collection.

The crux of the issue is that the *coverage* (i.e., the frequency that a prediction set contains the true outcome over many repeated measurements) of Bayes credible prediction sets is directly tied to the correctness of our modeling assumptions, which we cannot entirely control (Datta et al., 2000; Duanmu et al., 2020). We would prefer the price of assumption error to be lost *efficiency* (i.e., wider prediction sets), rather than poor coverage.

Conformal prediction is a distribution-free uncertainty quantification method which provides prediction sets with reliable coverage under very mild assumptions (Vovk et al., 2005). In particular, conformal prediction can accomodate *post hoc* covariate shift (i.e., covariate shift that is only known after training the surrogate) and does not assume the surrogate is well-specified (e.g., we could use a linear model on data following a cubic trend). Unfortunately conformal prediction is challenging to use in a BayesOpt algorithm since it is non-differentiable, requires continuous outcomes to be discretized, and needs density ratio estimates for unknown densities. Furthermore, because conformal prediction sets are defined over observable outcomes, they cannot distinguish between epistemic and aleatoric uncertainty, a distinction that is important for effective exploration.

In this work we present conformal Bayesian optimization with a motivating example in Figure 1. Conformal BayesOpt adjusts how far new queries will move from the training data by choosing an acceptable miscoverage tolerance $\alpha \in (0, 1]$. If $\alpha = 1$ then we recover conventional BayesOpt, but if $\alpha < 1$ then the search will be directed to the region where conformal predictions are guaranteed coverage of at



(a) Objective fn. (b) UCB (c) CUCB (d) Normalized density ratio

Figure 1: A motivating example of feedback covariate shift. We want $\mathbf{x}^* \in [0,1]^2$ which maximizes the Branin objective **(a)**, starting from 8 examples in the upper right (the black dots). The upper-confidence bound (UCB) acquisition function **(b)** selects the next query (the red star) far from any training data, where we cannot guarantee reliable predictions. In higher dimensions, we will exhaust our query budget long before covering the whole search space with training data. Given a miscoverage tolerance $\alpha = 1/\sqrt{8}$, conformal UCB **(c)** directs the search to the region where conformal predictions are guaranteed coverage of at least $(1-\alpha)$. **(d)** The dashed line is the set $\mathbf{x}$ such that $w(\mathbf{x}) \propto p_{\text{query}}(\mathbf{x})/p_{\text{train}}(\mathbf{x})$ is exactly α.

least $(1-\alpha)$, keeping feedback covariate shift in check and accounting for potential error in modeling assumptions.

In summary, our contributions are as follows:

- We show how to integrate conformal prediction into BayesOpt through the conformal Bayes posterior, with corresponding generalizations of common BayesOpt acquisition functions, enabling the reliable coverage of conformal prediction while still distinguishing between epistemic and aleatoric uncertainty in a principled way.
- An efficient, differentiable implementation of full conformal Bayes for Gaussian process (GP) regression models, which is necessary for effective query optimization, and a practical procedure to estimate the density ratio for BayesOpt query proposal distributions.
- Demonstrations on synthetic black-box optimization tasks and real tabular ranking tasks that conformal BayesOpt has superior sample-efficiency when the surrogate is misspecified and is comparable otherwise, while improving query coverage significantly. Note that while conformal BayesOpt has promising performance, our goal is not primarily to "beat" classical alternatives; rather, we show how to introduce conformal prediction into BayesOpt, and explore the corresponding empirical behaviour and results. [1]

2 PRELIMINARIES

In this work we will focus on black-box optimization problems of the form $\max_{\mathbf{x}\in\mathcal{X}}(f_1^*(\mathbf{x}),\ldots,f_d^*(\mathbf{x}))$, where each $f_i^* : \mathcal{X} \to \mathbb{R}$ is an unknown function of decision variables $\mathbf{x} \in \mathcal{X}$, and d is the number of objectives. We do not observe f^* directly, but instead receive noisy outcomes (i.e. labels) $\mathbf{y} \in \mathcal{Y}$ according to some likelihood $p^*(\mathbf{y}|f)$.

[1] Code is available at `github.com/samuelstanton/conformal-bayesopt.git`

2.1 Bayesian optimization

BayesOpt alternates between *inference* and *selection*, inferring the expected utility of potential query points from available data, which then serves as a proxy objective to select the next batch of observations, which are fed back into the inference procedure, completing one iteration of a repeating loop (Brochu et al., 2010; Frazier, 2018). Inference consists of applying Bayes rule to a prior $p(f)$, a likelihood $p(\mathbf{y}|f)$ and dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=0}^{n-1}$ to obtain a Bayes posterior $p(f|\mathcal{D})$. The expected utility of $\mathbf{x}$ is given by an acquisition function $a : \mathcal{X} \to \mathbb{R}$ with the general form

$$a(\mathbf{x}, \mathcal{D}) = \int u(\mathbf{x}, f, \mathcal{D}) p(f|\mathcal{D}) df, \tag{1}$$

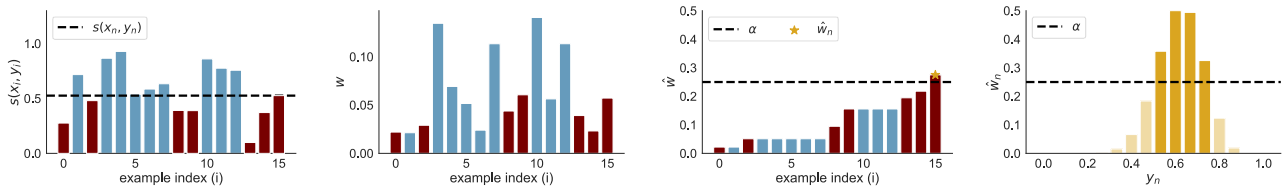
where u is a user-specified utility function. For example, taking $u(\mathbf{x}, f, \mathcal{D}) = [f(\mathbf{x}) - \max_{\mathbf{y}_i \in \mathcal{D}} \mathbf{y}_i]_+$, where $[\cdot]_+ = \max(\cdot, 0)$, yields the expected improvement (EI) acquisition function (Jones et al., 1998). Since a is the Bayes posterior expectation of u, maximizers $\mathbf{x}^* = \operatorname{argmax}_{\mathbf{x}\in\mathcal{X}} a(\mathbf{x}, \mathcal{D})$ are *Bayes-optimal* with respect to u. Bayes-optimality means a decision is *coherent* with our posterior beliefs about f^*. We think of $\operatorname{argmax}_{\mathbf{x}\in\mathcal{X}} a(\mathbf{x}, \mathcal{D})$ as inducing a distribution $p_{\text{query}}(\mathbf{x}) = p'(\mathbf{x}|\mathcal{D}) \propto \exp\{a(\mathbf{x}, \mathcal{D})\}$ (Levine, 2018).

2.2 Bayesian inference and model misspecification

One way to assess the quality of our posterior beliefs is to check the coverage of the corresponding Bayes β-credible prediction sets, which are subsets $\mathcal{K}_\beta(\mathbf{x}) \subseteq \mathcal{Y}$ satisfying

$$\beta = \int_{\mathbf{y}\in\mathcal{K}_\beta(\mathbf{x})} \int p(\mathbf{y}|f(\mathbf{x})) p(f|\mathcal{D}) df \, d\mathbf{y}, \tag{2}$$

where $\beta \in (0, 1]$ is the level of subjective credibility (Gneiting et al., 2007). β-credible sets may exhibit poor coverage, meaning the frequency of "implausible" events outside the set happening is much more than $1-\beta$ (Bachoc, 2013). Note that poor coverage does not necessarily imply that



(a) conformal scores $\mathbf{s}$ (b) imp. weights $\mathbf{w}$ (c) IW partial sums $\overline{\mathbf{w}}$ (d) $\mathcal{C}_\alpha(\mathbf{x}_n)$

Figure 2: Constructing a conformal prediction set $C_\alpha(\mathbf{x}_n)$ in the regression setting. First, **(a)** we choose some $\hat{\mathbf{y}}_n \in \mathcal{Y}$ and guess $\mathbf{y}_n = \hat{\mathbf{y}}_n$, computing conformal scores $\mathbf{s}$ of $\{(\mathbf{x}_0, \mathbf{y}_0), \ldots, (\mathbf{x}_{n-1}, \mathbf{y}_{n-1}), (\mathbf{x}_n, \hat{\mathbf{y}}_n)\}$. **(b)** we note which examples score better than our guess (shown in blue), and mask out the corresponding importance weights $\mathbf{w}$. **(c)** we compute the partial sums $\overline{\mathbf{w}}$ of the masked importance weights, adding $\hat{\mathbf{y}}_n$ to $C_\alpha(\mathbf{x}_n)$ if $\overline{w}_n > \alpha$, **(d)** repeat steps **(a - c)** for many guesses of $\mathbf{y}_n$. Rejected and accepted guesses are shaded light and dark, respectively.

$\mathcal{K}_\beta(\mathbf{x})$ was computed incorrectly, it may simply indicate a faulty assumption.

For example, in BayesOpt it is very common to assume $f^* \sim \mathcal{GP}(0, \kappa)$, where κ is a Matérn kernel. Matérn kernels support functions that are at least once differentiable, and can struggle to model objectives with discontinuities. As another example, we typically do not know the true likelihood $p^*(\mathbf{y}|f)$, and often choose a simple homoscedastic likelihood $p(\mathbf{y}|f) = \mathcal{N}(f, \sigma^2 I_d)$ for convenience. In reality the true noise process may be correlated with $\mathbf{x}$, across objectives, or may not be Gaussian at all (Assael et al., 2014, Griffiths et al., 2019, Makarova et al., 2021). These examples are common instances of *model misspecification*.

In practice faulty assumptions are nearly inevitable, and they are not always harmful, since simplifying assumptions can confer significant practical benefits. Indeed, theoretical convergence rates for acquisition functions like UCB (Srinivas et al., 2010) suggest that for BayesOpt we want to use the *smoothest possible* model, subject to the constraint that we can still model f^* *sufficiently* well. Similarly Gaussian likelihoods have significant computational advantages, and there is no guarantee that constructing a task-specific likelihood for every optimization problem would be worth the effort. We propose accepting that some assumption error will always be present, and instead focus on how alter BayesOpt to accomodate imperfect models.

2.3 Conformal prediction

See Shafer and Vovk (2008) for a complete tutorial on conformal prediction, or Angelopoulos and Bates (2021) for a modern, accessible introduction. Informally, a conformal prediction set $C_\alpha(\mathbf{x}_n) \subset \mathcal{Y}$ is a set of possible labels for a test point $\mathbf{x}_n$ given a sequence of observations $\mathcal{D}$. Candidate labels $\hat{\mathbf{y}}_n$ are included in $C_\alpha(\mathbf{x}_n)$ if the resulting pair $(\mathbf{x}_n, \hat{\mathbf{y}}_n)$ is sufficiently similar to the actual examples in $\mathcal{D}$. The degree of similarity is measured by a score function s and importance weights (IWs) $\mathbf{w}$, and the similarity threshold is determined by the miscoverage tolerance α. In Figure 2 we visualize the process of constructing $C_\alpha(\mathbf{x}_n)$.

Conformal prediction is a *distribution-free* uncertainty quantification method because it does not assume $\mathcal{D}$ is drawn from any particular distribution, nor does it assume s is derived from a well-specified model. In our context the critical assumption is that $\mathcal{D} \cup \{(\mathbf{x}_n, \mathbf{y}_n)\}$ is pseudo-exchangeable. Fannjiang et al. (2022) provide a formal statement of pseudo-exchangeability (Definition 2), and prove a coverage guarantee for conformal prediction sets in the special case when $\mathcal{D}$ is IID from $p(\mathbf{x})p^*(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x}|\mathcal{D})$ is invariant to shuffling of $\mathcal{D}$, which we restate below:

Definition 2.1. Let $\mathcal{D} \sim p(\mathbf{x})p^*(\mathbf{y}|\mathbf{x})$ and $(\mathbf{x}_n, \mathbf{y}_n) \sim p'(\mathbf{x}|\mathcal{D})p^*(\mathbf{y}|\mathbf{x})$, where $p'(\mathbf{x}|\mathcal{D})$ is chosen such that $\mathcal{D} \cup \{(\mathbf{x}_n, \mathbf{y}_n)\}$ is pseudo-exchangeable. Given $w_i \propto p'(\mathbf{x}_i|\hat{\mathcal{D}}_{-i})/p(\mathbf{x}_i)$ s.t. $\sum_i w_i = 1$, $\forall \alpha \in (0, 1]$, the conformal prediction set corresponding to score function s is

$$\mathcal{C}_\alpha(\mathbf{x}_n) := \left\{ \hat{\mathbf{y}}_n \in \mathcal{Y} \;\middle|\; \mathbf{h}^\top \mathbf{w} > \alpha \right\}, \tag{3}$$
$$\text{where } h_i := \mathbb{1}\left\{ s(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathcal{D}}) \leq s(\mathbf{x}_n, \hat{\mathbf{y}}_n, \hat{\mathcal{D}}) \right\},$$

$\hat{\mathcal{D}} = \mathcal{D} \cup \{(\mathbf{x}_n, \hat{\mathbf{y}}_n)\}$, and $p'(\mathbf{x}|\hat{\mathcal{D}}_{-i})$ is the query proposal density given training data $\hat{\mathcal{D}}_{-i} = \hat{\mathcal{D}} - \{(\mathbf{x}_i, \mathbf{y}_i)\}$. The importance weights $\mathbf{w}$ account for covariate shift (Tibshirani et al., 2019), and $w_i = 1/(n+1)$ $\forall i$ in the special case where $\mathcal{D} \cup \{(\mathbf{x}_n, \mathbf{y}_n)\}$ is fully exchangeable (e.g. IID).

Conformal prediction enjoys a frequentist marginal coverage guarantee on $\mathcal{C}_\alpha(\mathbf{x}_n)$ with respect to the joint distribution over $\mathcal{D} \cup \{(\mathbf{x}_n, \mathbf{y}_n)\}$,

$$\mathbb{P}[\mathbf{y}_n \in \mathcal{C}_\alpha(\mathbf{x}_n)] \geq 1 - \alpha, \tag{4}$$

meaning if we repeatedly draw $\mathcal{D} \sim p(\mathbf{x})p^*(\mathbf{y}|\mathbf{x})$, and $(\mathbf{x}_n, \mathbf{y}_n) \sim p'(\mathbf{x}|\mathcal{D})p^*(\mathbf{y}|\mathbf{x})$, $\mathcal{C}_\alpha(\mathbf{x}_n)$ will contain the observed label $\mathbf{y}_n$ with frequency at least $(1-\alpha)$. A prediction set with a coverage guarantee like Eq. (4) is *conservatively valid* at the $1-\alpha$ level. In Appendix B.1 we discuss *randomized* conformal prediction which is *exactly valid*, meaning the long run frequency of errors converges to exactly α.

Marginal coverage is distinct from *conditional* coverage, since it does not guarantee the coverage of C_α for any specific $\mathbf{x}_n$, only the average coverage over the whole domain $\mathcal{X}$.

Full conformal Bayes corresponds to the score function $s(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathcal{D}}) = \log p(\mathbf{y}_i|\mathbf{x}_i, \hat{\mathcal{D}})$. Conditioning an existing posterior $p(\mathbf{y}|\mathbf{x}_i, \mathcal{D})$ on the additional observation $(\mathbf{x}_n, \hat{\mathbf{y}}_n)$ is commonly referred to as "retraining" in the conformal prediction literature. *If* the surrogate just so happens to be correctly specified (e.g. $f^* \sim p(f)$), then $\log p(\mathbf{y}_i|\mathbf{x}_i, \hat{\mathcal{D}})$ is the optimal choice of score function, meaning it provides the most *efficient* prediction sets (i.e. smallest by expected volume w.r.t. the prior $p(f)$) among all prediction sets that are valid at the $1-\alpha$ level (Hoff, 2021). In the typical situation where we think our model assumptions are plausible but do not really believe them, full conformal Bayes rewards us if our assumptions turn out to be correct, yet it produces valid predictions as long as the true data generation process is some pseudo-exchangeable sequence.

BayesOpt and pseudo-exchangeability: unfortunately if $\mathcal{D}$ is collected by some active online selection strategy such as BayesOpt, then $\mathcal{D}$ is not IID and the coverage guarantee for Defn. 2.1 does not apply. Note that even large offline datasets are not guaranteed to be IID, so the same issue may arise even in single-round design setting considered by Fannjiang et al. (2022). Despite the gap in theory, in this work we investigate the technical challenges associated with incorporating conformal prediction sets into BayesOpt, and find that empirically they can still improve query coverage (Section 5.3). In Appendix A.1 we include further discussion of the assumptions and limitations of conformal prediction.

3 RELATED WORK

Conformal prediction: Our work is related to Fannjiang et al. (2022), who propose a black-box optimization method based on conformal prediction specifically to address feedback covariate shift. However, because they assume new queries are drawn from a closed-form proposal distribution, and because exact conformal prediction is not differentiable, their approach cannot be easily extended to most BayesOpt methods.[2] Bai et al. (2022) propose a differentiable approximation of conformal prediction, but it requires solving a minimax optimization subproblem. Stutz et al. (2021) independently proposed a continuous relaxation of conformal prediction, like our work, but only for fully exchangeable classification data. We propose a more general form that allows for covariate shift, and we also provide an efficient discretization procedure for regression and show how to estimate the importance weights when the queries are drawn from an implicit density.

[2]BayesOpt proposal distributions are usually implicit, obtained through gradient-based optimization of the acquisition function.

Robust BayesOpt: There is a substantial body of work on adaptation to model misspecification in the bandit setting (i.e. discrete actions), e.g. Lattimore et al. (2020) and Foster et al. (2020), however we are primarily focused on problems with continuous decisions. Since the seminal analysis of GP-UCB regret bounds by Srinivas et al. (2010), follow-up work has proposed UCB variants for misspecified likelihoods (Makarova et al., 2021), misspecified GP priors (Bogunovic and Krause, 2021), or to guarantee $f^*(\mathbf{x}_i) > c$ for some threshold $c \in \mathbb{R}$ (Sui et al., 2015). These approaches are not easy to extend to other acquisition functions, and tend to rely on fairly strong assumptions on the smoothness of f^* or fix a specific kind of model misspecification.[3] Wang et al. (2018) prove regret bounds for GP-UCB when f^* is drawn from a GP with unknown mean and kernel functions, but assume we know the right hypothesis class of mean and kernel functions and have a collection of offline datasets available for pretraining.

Finally, Eriksson et al. (2019) propose TuRBO, which is superficially similar to conformal BayesOpt since it constrains queries to a Latin hypercube trust region around the best known local optimum. While TuRBO can be very effective in practice, the size of the trust region is controlled by a heuristic with five hyperparameters in the single-objective case, and even more in the multi-objective case (Daulton et al., 2021b). Despite the additional complexity, the credible set coverage on queries in TuRBO trust regions can still vary wildly (see Section 5.3). In contrast, conformal prediction provides distribution-free coverage guarantees under very mild assumptions, and our approach can be applied to any reparameterizable acquisition function (Wilson et al., 2017). To our knowledge our approach is the first BayesOpt procedure to incorporate conformal prediction.

4 METHOD

We now describe the key ideas behind conformal Bayesian optimization. First in Section 4.1 we show how to efficiently compute $C_\alpha(\mathbf{x}_n)$, addressing differentiability and discretization of continuous outcomes. Our procedure is summarized in Algorithm 1. In Section 4.2 we introduce the conformal Bayes posterior $p_\alpha(f(\mathbf{x}_n)|\mathcal{D})$, allowing us to distinguish between aleatoric and epistemic uncertainty and combine to conformal prediction with many well-known BayesOpt utility functions. Finally in Section 4.3 we address feedback covariate shift without requiring closed-form expressions for $p(\mathbf{x})$ and $p'(\mathbf{x}_i|\hat{\mathcal{D}}_{-i})$. In Appendix D.1 we provide a detailed overview of the whole method, along with a discussion of the computational complexity in Appendix D.4.

[3]For example, it is commonly assumed that f^* has bounded RKHS norm w.r.t. the chosen GP kernel, that we know a good bound in order to set hyperparameters correctly, and that f^* is Lipschitz continuous.

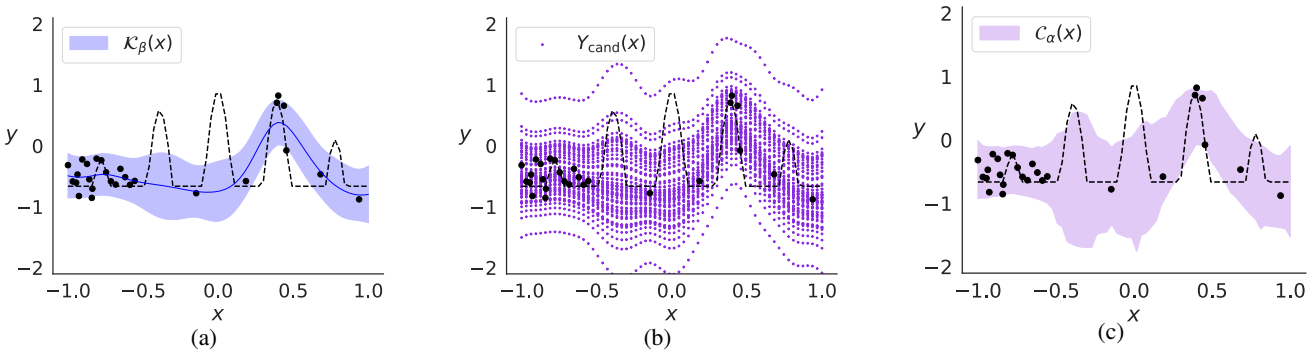


Figure 3: Constructing full conformal Bayes prediction sets starting with a Bayes posterior $p(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D})$. In this example $\mathcal{D}$ is composed of $n = 27$ noisy observations (shown as black dots) of the true objective (shown as a black dashed line) and $\alpha = 1 - \beta = 0.19$. In panel **(a)** we show $\mathcal{K}_\beta(\mathbf{x})$, the β-credible prediction set. In panel **(b)** we show Y_{cand} populated by samples from $p(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D})$. In panel **(c)** we show $\mathcal{C}_\alpha(\mathbf{x})$, the conformal prediction set. The coverage of $\mathcal{C}_\alpha(\mathbf{x})$ is noticeably better than $\mathcal{K}_\beta(\mathbf{x})$ in regions where there is little training data, though the nominal confidence level is the same.

Algorithm 1 Differentiable conformal prediction masks

Data: train data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=0}^{n-1}$, test point $\mathbf{x}_n$, imp. weights $\mathbf{w}$, label candidates Y_{cand}, score function s, miscoverage tolerance α, relaxation strength $\tau > 0$.

$m_j = 0, \forall j \in \{0, \dots, k-1\}$.

for $\hat{\mathbf{y}}_j \in Y_{\text{cand}}$ **do**

  $\hat{\mathcal{D}} \leftarrow \mathcal{D} \cup \{(\mathbf{x}_n, \hat{\mathbf{y}}_j)\}$

  $\mathbf{s} \leftarrow [s(\mathbf{x}_0, \mathbf{y}_0, \hat{\mathcal{D}}) \ \cdots \ s(\mathbf{x}_n, \hat{\mathbf{y}}_j, \hat{\mathcal{D}})]^\top$.

  $\mathbf{h} \leftarrow \texttt{sigmoid}(\tau^{-1}(\mathbf{s} - s_n))$.

  $\overline{w} \leftarrow \mathbf{h}^\top \mathbf{w}$.

  $m_j \leftarrow \texttt{sigmoid}(\tau^{-1}(\overline{w} - \alpha))$.

end

Result: $\mathbf{m}$

4.1 Full conformal Bayes with Gaussian processes

Efficient retraining: full conformal Bayes requires us to compute $\log p(\mathbf{y}_i|\mathbf{x}_i, \hat{\mathcal{D}})$ $\forall i \leq n$ and $\forall \hat{\mathbf{y}}_j \in Y_{\text{cand}}$, where Y_{cand} is some discretization of $\mathcal{Y}$. This incremental posterior update can be done very efficiently if the surrogate is a GP regression model (Gardner et al., 2018; Stanton et al., 2021; Maddox et al., 2021), and we will later reuse the conditioned posteriors to estimate expectations w.r.t. $p_\alpha(f(\mathbf{x})|\mathcal{D})$. Note that computing the GP posterior likelihood of training data can be numerically unstable, which we address in Appendix D. Other Bayesian predictive posteriors (e.g. from Bayesian neural networks) are conditioned on training data via iterative methods such as gradient descent, making full conformal Bayes very expensive (Fong and Holmes, 2021).

Differentiable prediction masks: the definition of $\mathcal{C}_\alpha(\mathbf{x}_n)$ in Eq. (3) can be broken down into a sequence of simple vector operations interspersed with Heaviside functions. The Heaviside function is piecewise constant, with ill-defined derivatives, so we replace it with its continuous relaxation, the `sigmoid` function (Algorithm 1). Informally, the output m_j of the final sigmoid can be interpreted as the *probability* of accepting some $\hat{\mathbf{y}}_j$ into $\mathcal{C}_\alpha(\mathbf{x}_n)$. The smoothness of the relaxation is controlled by a single hyperparameter $\tau \in (0, +\infty)$. As $\tau \to 0$ the relaxation becomes exact but the gradients become very poorly behaved.

Efficient discretization of $\mathcal{Y}$: now we need a good way to choose Y_{cand}. When $\mathbf{y}$ is low-dimensional (e.g. sequential, single-objective tasks), then Y_{cand} can be a dense grid, however dense grids are inefficient since they must be wide enough to capture all possible values of $\mathbf{y}$ and dense enough to pinpoint the boundary of $\mathcal{C}_\alpha(\mathbf{x}_n)$. Even if we do not fully believe $p(\mathbf{y}|\mathbf{x}_n, \mathcal{D})$, it is still our best guess of where $\mathbf{y}|\mathbf{x}_n$ should be, so instead of a dense grid we populate Y_{cand} with proposals $\hat{\mathbf{y}}_j \sim p(\mathbf{y}|\mathbf{x}_n, \mathcal{D})$. This approach not only reduces computational effort for low-dimensional $\mathbf{y}$, it also allows us to extend to tasks with multiple objectives and batched queries (Appendix B.6). In Figure 3 we visualize the computation of a conformal Bayes prediction set.

4.2 Conformal acquisition functions

For the sake of clarity in the following sections we will omit the subscript from $\mathbf{x}_n$. By the sum rule of probability, we can rewrite $p(f(\mathbf{x})|\mathcal{D})$ as an integral over all possible outcomes $\mathbf{y}|\mathbf{x}$,

$$p(f(\mathbf{x})|\mathcal{D}) = \int_{\hat{\mathbf{y}} \in \mathcal{Y}} p(f(\mathbf{x})|\hat{\mathcal{D}}) p(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D}) d\hat{\mathbf{y}}. \quad (5)$$

In other words, $p(f(\mathbf{x})|\mathcal{D})$ can be seen as a Bayesian model average, where we condition each component model on a different potential observation $(\mathbf{x}, \hat{\mathbf{y}})$, and weight the components by $p(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D})$.

We are free to change the component weights to any other valid distribution over $\hat{\mathbf{y}}$ we like. Now we introduce the conformal Bayes predictive posterior $p_\alpha(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D})$,

$$p_\alpha(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D}) := \begin{cases} (1-\alpha)/Z_1 & \text{if } \hat{\mathbf{y}} \in C_\alpha(\mathbf{x}), \\ \alpha p(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D})/Z_2 & \text{else,} \end{cases}$$

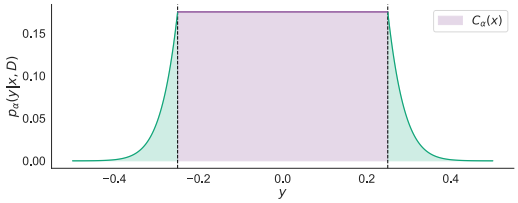


Figure 4: An illustration of $p_\alpha(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D})$. The total density of all outcomes in $\mathcal{C}_\alpha(\mathbf{x})$ is set to $1-\alpha$.

where Z_1, Z_2 are normalization constants. See Figure 4 for an illustration. We are partitioning the outcome space into two events, either $\hat{\mathbf{y}} \in C_\alpha(\mathbf{x})$ or it is not. Since $C_\alpha(\mathbf{x})$ is a valid prediction set, $\hat{\mathbf{y}} \in C_\alpha(\mathbf{x})$ with frequency $(1-\alpha)$, and we do not consider any particular $\hat{\mathbf{y}} \in C_\alpha(\mathbf{x})$ to be more likely than another, since the coverage guarantee holds for $C_\alpha(\mathbf{x})$ as a whole.[4] We also expect that $\hat{\mathbf{y}} \in \mathcal{Y} \setminus C_\alpha(\mathbf{x})$ with frequency α, and we weight each $\hat{\mathbf{y}} \notin C_\alpha(\mathbf{x})$ by $p(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D})$ to form a proper density (i.e. a density that integrates to 1).

If we had noiseless observations (i.e. $\mathbf{y}_i = f(\mathbf{x}_i)$), we could use $p_\alpha(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D})$ directly when computing the acquisition value of new queries. However managing the explore-exploit tradeoff with noisy outcomes requires us to distinguish between epistemic and aleatoric uncertainty. If we do not, optimistic acquisition functions like UCB may direct us towards queries whose outcomes are uncertain due to measurement error. Substituting $p_\alpha(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D})$ for $p(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D})$ in Eq. (5) results in the conformal Bayes posterior $p_\alpha(f(\mathbf{x})|\mathcal{D})$,

$$p_\alpha(f(\mathbf{x})|\mathcal{D}) := \frac{1-\alpha}{Z_1} \int_{\hat{\mathbf{y}} \in C_\alpha(\mathbf{x})} p(f|\hat{\mathcal{D}}) d\hat{\mathbf{y}} + \frac{\alpha}{Z_2} \int_{\hat{\mathbf{y}} \in \mathcal{Y} \setminus C_\alpha(\mathbf{x})} p(f(\mathbf{x})|\hat{\mathcal{D}}) p(\hat{\mathbf{y}}|\mathbf{x}, \mathcal{D}) d\hat{\mathbf{y}}. \tag{6}$$

Given $p_\alpha(f|\mathcal{D})$, we can "conformalize" any acquisition function written in the form of Eq. (1) by substituting $p_\alpha(f|\mathcal{D})$ for $p(f|\mathcal{D})$. In Appendix B.2 we show that $p_\alpha(f|\mathcal{D})$ converges pointwise to $p(f|\mathcal{D})$ as $\alpha \to 1$, and in Appendix B.4 we explicitly derive conformal variants of several popular BayesOpt acquisition functions.

Monte Carlo estimates of conformal acquisition values: in brief, given a query point $\mathbf{x}$ and utility function u, we first draw a candidate grid Y_{cand} and compute the corresponding prediction mask $\mathbf{m}$ according to Section 4.1. Then we estimate the conformal acquisition value as follows:

$$\begin{aligned} a_\alpha(\mathbf{x}, \mathcal{D}) &= \int u(\mathbf{x}, f, \mathcal{D}) p_\alpha(f|\mathcal{D}) df, \\ &\approx (1-\alpha)\mathbf{u}^\top \mathbf{v} + \alpha \mathbf{u}^\top \mathbf{v}', \\ \text{where } \mathbf{u} &= [u(\mathbf{x}, f^{(0)}, \mathcal{D}) \ \cdots \ u(\mathbf{x}, f^{(k-1)}, \mathcal{D})]^\top, \\ \mathbf{v}_i &= \frac{m_i}{p(\hat{\mathbf{y}}_i|\mathbf{x}, \mathcal{D})} \left(\sum_j \frac{m_j}{p(\hat{\mathbf{y}}_j|\mathbf{x}, \mathcal{D})} \right)^{-1}, \\ \mathbf{v}'_i &= (1-m_i)(\mathbf{1}^\top(\mathbf{1}-\mathbf{m}))^{-1}, \end{aligned} \tag{7}$$

and $f^{(j)} \sim p(f|\mathcal{D} \cup \{(\mathbf{x}, \hat{\mathbf{y}}_j)\}) \ \forall \hat{\mathbf{y}}_j \in Y_{\text{cand}}$, which is cheap since we already computed the conditioned posteriors when calculating $\mathbf{m}$. See Appendix B.3 for the full derivation.

4.3 Accounting for Feedback Covariate Shift

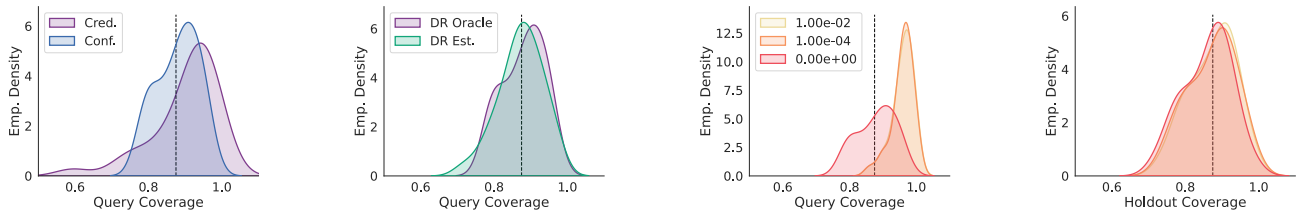
If we were merely ranking queries exchangeable with $\mathcal{D}$, then there would be no need to correct for covariate shift. However, our goal is to find queries with exceptional outcomes, and the more we optimize, the more severe we can expect the resulting feedback covariate shift to be.

Density ratio estimation: as we saw in Section 2.3, adapting $\mathcal{C}_\alpha(\mathbf{x})$ to covariate shift requires estimating importance weights $w_i \propto r(\mathbf{x}_i) = p'(\mathbf{x}_i|\hat{\mathcal{D}}_{-i})/p(\mathbf{x}_i)$, where $p'(\mathbf{x}_i|\hat{\mathcal{D}}_{-i})$ is the proposal distribution from which we *would* have drawn candidate query points if we had training data $\hat{\mathcal{D}}_{-i}$. If we have closed-form expressions for $p(\mathbf{x}_i)$ and $p'(\mathbf{x}_i|\hat{\mathcal{D}}_{-i})$ then we can compute $r(\mathbf{x}_i)$ easily, but in general we only have samples from $p(\mathbf{x})$. Furthermore if we wish to optimize queries with gradient based methods then $p'(\mathbf{x}_i|\hat{\mathcal{D}}_{-i})$ is implicitly defined as the distribution over iterates $\mathbf{x}_n^{(t)}$ induced by the gradient field $\nabla_{\mathbf{x}} a_\alpha$ and an initial distribution on $\mathbf{x}_n^{(0)}$. Fortunately we can still obtain samples from $p'(\mathbf{x}_i|\hat{\mathcal{D}}_{-i})$ by sampling from the energy distribution, $p'(\mathbf{x}_i|\hat{\mathcal{D}}_{-i}) \propto \exp\{a_\alpha(\mathbf{x}, \hat{\mathcal{D}}_{-i})\}$ via stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011). Note that this formulation requires us to run $(n+1) \times k$ SGLD chains, one for each density $p'(\mathbf{x}_i|\hat{\mathcal{D}}_{-i})$. Since we are already intending to use a sample-based empirical approximation of the density ratio, we make another approximation here, assuming $a_\alpha(\mathbf{x}, \hat{\mathcal{D}}_{-i}) \approx a_\alpha(\mathbf{x}, \mathcal{D})$, which allows us to rely on samples from a single SGLD chain.

Once we have samples from $p(\mathbf{x})$ (which are already in $\mathcal{D}$) and $p'(\mathbf{x}|\mathcal{D})$, we estimate $r(\mathbf{x})$ with a probabilistic classifier (Sugiyama et al., 2012). We assign labels z to the samples, corresponding to the conditional distributions $p(\mathbf{x}) = p(\mathbf{x}|z=0)$ and $p'(\mathbf{x}|\mathcal{D}) = p(\mathbf{x}|z=1)$. By Bayes theorem, we rewrite $r(\mathbf{x})$,

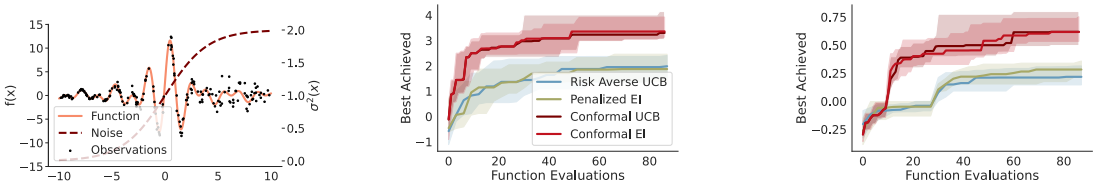
$$r(\mathbf{x}) = \frac{p(z=0)}{p(z=1)} \frac{p(z=1 \mid \mathbf{x})}{p(z=0 \mid \mathbf{x})}, \tag{8}$$

[4] We could also use a conformal predictive density here (Vovk et al., 2017; Marx et al., 2022), which we leave for future work.



(a) (b) (c) (d)

Figure 5: Here we evaluate the empirical coverage of $C_\alpha(\mathbf{x})$. The shaded regions in each panel depict a KDE estimate of the distribution of coverage when $n = 64$ and $\alpha = 0.125$, estimated from 32 independent trials. The black dashed line indicates $1 - \alpha$. In panel **(a)** we compare the coverage of Bayesian β-credible ($\beta = 1 - \alpha$) and randomized conformal prediction sets, if $\tau = 0$ and we have a density ratio oracle. Conformal prediction provides much more consistent coverage. Next in panel **(b)** replacing the density ratio oracle with learned density ratio estimates has fairly minimal effect on the coverage of the resulting conformal prediction sets. In panel **(c)** we investigate the effect of the sigmoid temperature τ when $p'(\mathbf{x}|\mathcal{D}) \neq p(\mathbf{x})$. In panel **(d)** we investigate the effect of τ when $p'(\mathbf{x}|\mathcal{D}) = p(\mathbf{x})$. Increasing τ makes the prediction sets more conservative.



(a) `sinc` objective fn. (b) `sinc` (c) `double-knot`

Figure 6: BayesOpt results on heteroscedastic, single-objective tasks `sinc` and `double-knot` (reporting median and its 95% conf. interval, estimated from 16 trials). **(a)** $\texttt{sinc}(\mathbf{x})$ (left y axis) and $\varepsilon(\mathbf{x})$ (right y axis). **(b)** the `sinc` task, best objective value found by conformal BayesOpt with homoscedastic likelihoods compared to baselines, risk-averse UCB and penalized EI, both with heteroscedastic likelihoods. **(c)** the `double-knot` task, same experiment as in panel **(b)**. Conformal BayesOpt with a misspecified likelihood outperforms the specialized baselines on both tasks.

such that we need only train a probabilistic classifier $\hat{p}(z \mid \mathbf{x})$ to discriminate the sample labels. We estimate the prior ratio $p(z = 0)/p(z = 1)$ empirically.

Which comes first, the acquisition function or the ratio estimator? To estimate r as just described we clearly must be able to compute $\nabla_{\mathbf{x}} a_\alpha$ to draw the required samples from $p'(\mathbf{x}|\mathcal{D}) \propto \exp\{a_\alpha(\mathbf{x}, \mathcal{D})\}$. Here we find a second and more serious issue, since a_α itself depends on r. We need an estimator $\hat{r}$ that simultaneously induces $p'(\mathbf{x}|\mathcal{D}) \propto \exp\{a_\alpha(\mathbf{x}, \mathcal{D})\}$ *and* accurately estimates $p'(\mathbf{x}|\mathcal{D})/p(\mathbf{x})$. For example, we could assume $\hat{r}(\mathbf{x}) = 1, \forall \mathbf{x}$, but the induced p' likely does not satisfy $p'(\mathbf{x}|\mathcal{D})/p(\mathbf{x}) = 1, \forall \mathbf{x}$.

To solve this issue, we begin with an initial estimator $\hat{r}_0(\mathbf{x}) = 1, \forall \mathbf{x}$, and for $t \geq 0$ we sample from $p'(\mathbf{x}|\mathcal{D}) \propto \exp\{a_\alpha(\mathbf{x}, \mathcal{D})\}$ via SGLD using the current estimator $\hat{r}_t$, then update the classifier on those new samples to produce an updated estimator $\hat{r}_{t+1}$ for the next iteration. To keep the acquisition surface from changing too rapidly (potentially destabilizing our SGLD chain), we compute an exponential moving average of the classifier weights, and the averaged weights are used when computing gradients of $a_\alpha(\mathbf{x}, \mathcal{D})$. Our approach is analogous to (and directly inspired by) bootstrapped deep Q-learning (Mnih et al., 2015).

5 EXPERIMENTS

In Section 5.1 we report the empirical coverage of credible and conformal prediction sets in a simplified setting. In Section 5.2 we show that conformal BayesOpt is robust to a misspecified likelihood. Finally in Section 5.3 we evaluate conformal BayesOpt on synthetic black-box optimization tasks, comparing the query coverage of credible and conformal prediction sets. See Appendix C for results on multi-objective synthetic tasks and real ranking tasks using drug and antibody design data, and see Appendix D for all experimental details.

5.1 Do Our Approximations Impact Coverage?

First we compare the empirical coverage of Bayes credible sets and randomized conformal prediction sets, and evaluate the sensitivity of conformal prediction to continuous relaxation and density ratio estimation. We consider a simplified offline regression setting where $p(\mathbf{x})$ and $p'(\mathbf{x}|\mathcal{D})$ are known 3D spherical Gaussian distributions with different means, f^*

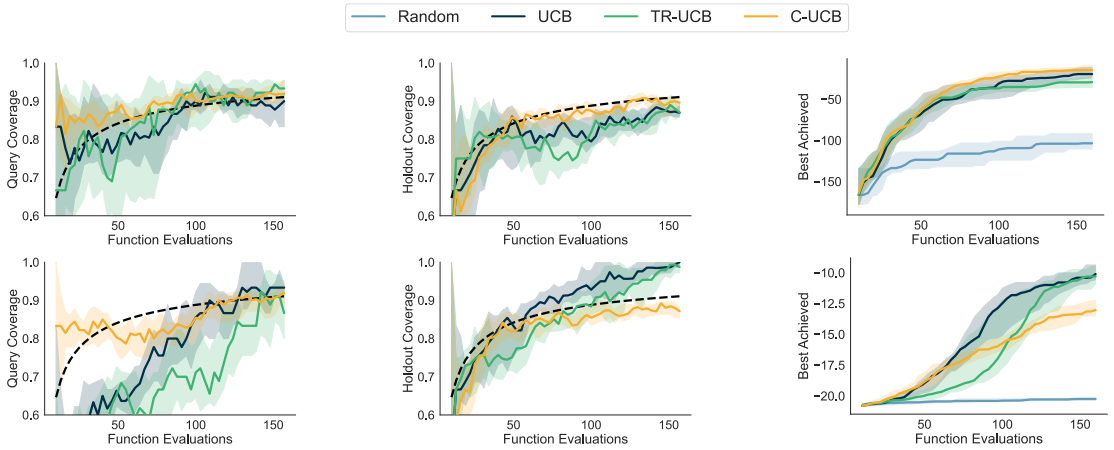


Figure 7: Using the same base acquisition function, we compare standard BayesOpt (UCB), TuRBO (TR-UCB), and conformal BayesOpt (C-UCB) optimizing `levy-20d` **(top row)** and `ackley-20d` **(bottom row)**. The midpoint, lower, and upper bounds of each curve depict the 50%, 20%, and 80% quantiles, estimated from 25 trials. In the **left column** we see credible set coverage varying significantly, despite reasonable coverage on a random holdout subset of the training data **(center column)**. The **right column** shows $\max_{0 \leq i \leq n} f^*(\mathbf{x}_i)$ as n increases, and we see the methods have comparable sample-efficiency. Conformal BayesOpt improves the objective value while predicting query outcomes much more reliably.

is the 3D Hartmann function, and $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(f^*(\mathbf{x}), \sqrt{.05})$. If the exact validity guarantee of randomized conformal prediction holds, then over many trials the coverage should concentrate around $(1-\alpha)$. Some deviation is to be expected due to sample variance and discretization error. In Figure 5a we see when we have the density ratio oracle and $\tau = 0$, that the distribution of conformal coverage is indeed concentrated around $(1-\alpha)$, especially relative to the distribution of credible coverage. In the other panels of Figure 5 we show that empirical density ratio estimates and the continuous relaxation do not compromise validity. In particular increasing τ makes the corresponding prediction sets more conservative, which is consistent with the limiting case $\lim_{\tau \to \infty} C_\alpha(\mathbf{x}) = \mathcal{Y}, \ \forall \mathbf{x}$.

5.2 Model Misspecification and Sample-Efficiency

Recall from Section 2.2 that BayesOpt surrogates often use a homoscedastic likelihood $p(\mathbf{y}|f) = \mathcal{N}(f, \sigma^2)$, where σ^2 is a learned constant. In Figure 6a we plot $f^*(\mathbf{x}) = \texttt{sinc}(\mathbf{x}) := (10\sin(\mathbf{x}) + 1)\sin(3\mathbf{x})/\mathbf{x}$ on $[-10, 10]$ with $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(f^*(\mathbf{x}), \varepsilon(\mathbf{x}))$ and $\varepsilon(\mathbf{x}) = 2/(1 + \exp\{\mathbf{x}/2\})$ In Figure 6b we compare conformal BayesOpt with $p(\mathbf{y}|f) = \mathcal{N}(f, \sigma^2)$ to two baselines specifically designed for tasks with heteroscedastic noise, risk-averse UCB (Makarova et al., 2021) and penalized EI (Griffiths et al., 2019), which both use heteroscedastic likelihoods. Both baselines require multiple replicates of each query to update their likelihoods, which significantly reduces sample efficiency. In Figure 6c, we repeat the same experiment on a second heteroscedastic task $f^*(\mathbf{x}) = \texttt{double-knot}(\mathbf{x}) := -\mathbf{x}_1 \exp\{-\mathbf{x}_1 - \mathbf{x}_2\}$ on $[-2, 6]^2$, with $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(f^*(\mathbf{x}), \varepsilon(\mathbf{x}))$ and $\varepsilon(\mathbf{x}) = ||\mathbf{x}||_2)$ (Gramacy, 2005). Despite having a simpler, misspecified noise model, conformal BayesOpt finds a better solution with fewer queries.

5.3 Good Query Coverage and Good Sample-Efficiency Are Not Mutually Exclusive

We use the batch UCB acquisition function ($q = 3$) to optimize two synthetic functions `levy` and `ackley`, taking $\mathcal{X} \subset \mathbb{R}^{20}$. For this experiment $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(f^*(\mathbf{x}), (\sigma^*)^2)$. To simulate the covariate shift that occurs in many applied problems, we sampled the initial training data from a random orthant of the input space. In Figure 7 we compare the sample efficiency and coverage of standard BayesOpt, TuRBO (Eriksson et al., 2019), and conformal BayesOpt. Each method is comparable in terms of sample efficiency, and the credible set coverage for standard BayesOpt and TuRBO looks reasonable on a random subset of the training data, if a bit unpredictable. However if we look at the *query coverage* we see that the credible set coverage varies wildly. The difference between coverage on a random holdout set and coverage on the query set is due to feedback covariate shift. In contrast, we see that the conformal set coverage for both random holdout points and query points very consistently tracks $(1-\alpha)$, where $\alpha = 1/\sqrt{n}$. In other words, of the methods considered conformal BayesOpt is the only approach that improves the objective while reliably predicting the query outcomes.

In Figure 8 we preview results showing we can also improve query coverage in tabular ranking tasks related to drug design. See Appendix C.3 for the full experiment.

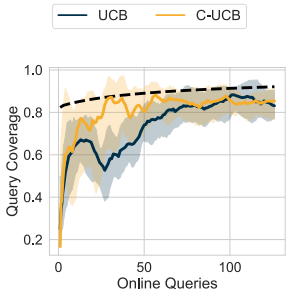


Figure 8: We used C-UCB to select small-molecule compounds for dopamine receptor binding affinity, showing that conformal acquisitions improve query coverage without harming sample efficiency.

6 DISCUSSION

We have shown that a combination of model misspecification and optimization-induced covariate shift can make Bayesian credible sets unreliable exactly where they are needed most — where the queries concentrate. Conformal prediction provides a principled solution to these issues, with distribution-free validity guarantees that help ensure robustness to model misspecification, and a natural mechanism to correct for covariate shift. To use conformal prediction inside of BayesOpt, we have developed differentiable and efficiently discretizable conformal prediction sets and coupled these with a practical density ratio estimation procedure, addressing key technical challenges in the conformal prediction literature. With the introduction of the conformal Bayes posterior, we have derived conformal generalizations of many popular acquisition functions, allowing us to accommodate features of practical tasks including batched queries, noisy observations, and multiple objectives. Empirically we find the combination of conformal prediction and BayesOpt to be very promising, since it improves query coverage and has sample-efficiency comparable to methods with no coverage guarantees at all.

Looking forward, although we focus on GP surrogates in the low-n regime, we expect many of these ideas to transfer to much larger models and datasets by replacing full conformal Bayes with split conformal Bayes, and either augmenting GPs with deep kernel learning, or by replacing GPs entirely with Bayesian linear models operating on pretrained representations learned by large self-supervised models. Extending conformal BayesOpt to discrete optimization, specifically biological sequence design, is a particularly exciting direction for future work. There are also intriguing theoretical directions, such as analyzing the effect of continuous relaxation and the effect of learned density ratio estimates on conformal coverage guarantees, and investigating whether the regret of conformal BayesOpt can be analyzed with milder assumptions than algorithms like GP-UCB (Srinivas et al., 2010).

As machine learning systems are deployed for increasingly impactful applications, we must confront the reality that machine learning models *will* be built on faulty assumptions, and those models *will* be asked to rank potential decisions without sufficient training data. The solution is not to blind ourselves to the error in our assumptions, nor is it to paralyze ourselves in pursuit of a perfect model. Instead we should develop methods that can gracefully accommodate imperfect models, balancing internal coherence with external validity.

Acknowledgments

The authors thank Sanyam Kapoor for his SGLD implementation, and Anastasios Angelopoulos, Matthias Seeger, Greg Benton, Andres Potapczynski, and Wanqian Yang for helpful discussions. This research is supported by NSF CAREER IIS-2145492, NSF I-DISRE 193471, NIH R01DA048764-01A1, NSF IIS-1910266, NSF 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science, Meta Core Data Science, Google AI Research, BigHat Biosciences, Capital One, and an Amazon Research Award.

References

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Assael, J.-A. M., Wang, Z., Shahriari, B., and de Freitas, N. (2014). Heteroscedastic treed bayesian optimisation. *arXiv preprint arXiv:1410.7172*.

Bachoc, F. (2013). Cross validation and maximum likelihood estimations of hyper-parameters of gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69.

Bai, Y., Mei, S., Wang, H., Zhou, Y., and Xiong, C. (2022). Efficient and differentiable conformal prediction with general function classes. *arXiv preprint arXiv:2202.11091*.

Balandat, M., Karrer, B., Jiang, D., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. (2020). BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems*, volume 33.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2021). Testing for outliers with conformal p-values. *arXiv preprint arXiv:2104.08279*.

Bekas, C., Kokiopoulou, E., and Saad, Y. (2007). An estimator for the diagonal of a matrix. *Applied numerical mathematics*, 57(11-12):1214–1229.

Beume, N., Fonseca, C. M., Lopez-Ibanez, M., Paquete, L., and Vahrenhold, J. (2009). On the complexity of computing the hypervolume indicator. *IEEE Transactions on Evolutionary Computation*, 13(5):1075–1082.

Bogunovic, I. and Krause, A. (2021). Misspecified gaussian process bandit optimization. *Advances in Neural Information Processing Systems*, 34.

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

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Datta, G. S., Ghosh, M., Mukerjee, R., and Sweeting, T. J. (2000). Bayesian prediction with approximate frequentist validity. *The Annals of Statistics*, 28(5):1414–1426.

Daulton, S., Balandat, M., and Bakshy, E. (2020). Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. *Advances in Neural Information Processing Systems*, 33:9851–9864.

Daulton, S., Balandat, M., and Bakshy, E. (2021a). Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *Advances in Neural Information Processing Systems*, 34.

Daulton, S., Cakmak, S., Balandat, M., Osborne, M. A., Zhou, E., and Bakshy, E. (2022). Robust multi-objective bayesian optimization under input noise. *arXiv preprint arXiv:2202.07549*.

Daulton, S., Eriksson, D., Balandat, M., and Bakshy, E. (2021b). Multi-objective bayesian optimization over high-dimensional search spaces. *arXiv preprint arXiv:2109.10964*.

Duanmu, H., Roy, D. M., and Smith, A. (2020). Existence of matching priors on compact spaces. *arXiv preprint arXiv:2011.03655*.

Emmerich, M. (2005). *Single-and multi-objective evolutionary design optimization assisted by gaussian random field metamodels*. PhD thesis, Dortmund University.

Emmerich, M. T., Deutz, A. H., and Klinkenberg, J. W. (2011). Hypervolume-based expected improvement: Monotonicity properties and exact computation. In *2011 IEEE Congress of Evolutionary Computation (CEC)*, pages 2147–2154. IEEE.

Eriksson, D., Pearce, M., Gardner, J., Turner, R. D., and Poloczek, M. (2019). Scalable global optimization via local bayesian optimization. *Advances in Neural Information Processing Systems*, 32.

Fannjiang, C., Bates, S., Angelopoulos, A., Listgarten, J., and Jordan, M. I. (2022). Conformal prediction for the design problem. *arXiv preprint arXiv:2202.03613*.

Fong, E. and Holmes, C. C. (2021). Conformal bayesian computation. *Advances in Neural Information Processing Systems*, 34.

Foster, D. J., Gentile, C., Mohri, M., and Zimmert, J. (2020). Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489.

Frazier, P. I. (2018). A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*.

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

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

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

Gramacy, R. B. (2005). *Bayesian treed Gaussian process models*. University of California, Santa Cruz.

Griffiths, R.-R., Aldrick, A. A., Garcia-Ortegon, M., Lalchand, V. R., and Lee, A. A. (2019). Achieving robustness to aleatoric uncertainty with heteroscedastic bayesian optimisation. *arXiv preprint arXiv:1910.07779*.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hoff, P. (2021). Bayes-optimal prediction with frequentist coverage control. *arXiv preprint arXiv:2105.14045*.

Hornung, V., Hartmann, R., Ablasser, A., and Hopfner, K.-P. (2014). Oas proteins and cgas: unifying concepts in sensing and responding to cytosolic nucleic acids. *Nature Reviews Immunology*, 14(8):521–528.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. (2021). Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*.

Ionescu, C., Vantzos, O., and Sminchisescu, C. (2015). Matrix backpropagation for deep networks with structured layers. In *Proceedings of the IEEE international conference on computer vision*, pages 2965–2973.

Johnstone, C. and Ndiaye, E. (2022). Exact and approximate conformal inference in multiple dimensions. *arXiv preprint arXiv:2210.17405*.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492.

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Landrum, G. (2016). Rdkit: Open-source cheminformatics software.

Lattimore, T., Szepesvari, C., and Weisz, G. (2020). Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR.

Letham, B., Karrer, B., Ottoni, G., and Bakshy, E. (2019). Constrained bayesian optimization with noisy experiments. *Bayesian Analysis*, 14(2):495–519.

Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*.

Liang, Q. and Lai, L. (2021). Scalable bayesian optimization accelerates process optimization of penicillin production. In *NeurIPS 2021 AI for Science Workshop*.

Maddox, W. J., Stanton, S., and Wilson, A. G. (2021). Conditioning sparse variational gaussian processes for online decision-making. *Advances in Neural Information Processing Systems*, 34:6365–6379.

Makarova, A., Usmanova, I., Bogunovic, I., and Krause, A. (2021). Risk-averse heteroscedastic bayesian optimization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17235–17245. Curran Associates, Inc.

Marx, C., Zhao, S., Neiswanger, W., and Ermon, S. (2022). Modular conformal calibration. In *International Conference on Machine Learning*, pages 15180–15195. PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Ndiaye, E. (2022). Stable conformal prediction sets. In *International Conference on Machine Learning*, pages 16462–16479. PMLR.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style,

high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Scholkopf, B. and Smola, A. J. (2018). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022.

Stanton, S., Maddox, W., Delbridge, I., and Wilson, A. G. (2021). Kernel interpolation for scalable online gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 3133–3141. PMLR.

Stanton, S., Maddox, W., Gruver, N., Maffettone, P., Delaney, E., Greenside, P., and Wilson, A. G. (2022). Accelerating bayesian optimization for biological sequence design with denoising autoencoders. *arXiv preprint arXiv:2203.12742*.

Stutz, D., Cemgil, A. T., Doucet, A., et al. (2021). Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Sui, Y., Gotovos, A., Burdick, J., and Krause, A. (2015). Safe exploration for optimization with gaussian processes. In *International conference on machine learning*, pages 997–1005. PMLR.

Terayama, K., Sumita, M., Tamura, R., and Tsuda, K. (2021). Black-box optimization for automated discovery. *Accounts of Chemical Research*, 54(6):1334–1346.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Vovk, V., Shen, J., Manokhin, V., and Xie, M.-g. (2017). Nonparametric predictive distributions based on conformal prediction. In *Conformal and Probabilistic Prediction and Applications*, pages 82–102. PMLR.

Wang, K. and Dowling, A. W. (2022). Bayesian optimization for chemical products and functional materials. *Current Opinion in Chemical Engineering*, 36:100728.

Wang, Z., Kim, B., and Kaelbling, L. P. (2018). Regret bounds for meta bayesian optimization with an unknown gaussian process prior. *Advances in Neural Information Processing Systems*, 31.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer.

Wilson, J. T., Moriconi, R., Hutter, F., and Deisenroth, M. P. (2017). The reparameterization trick for acquisition functions. *arXiv preprint arXiv:1712.00424*.

Appendices

The appendices are structured as follows:

- In Appendix A, we describe the assumptions, limitations, and broader impacts of this work.
- In Appendix B, we provide detailed derivations of the randomized differentiable conformal prediction, conformal Bayes posterior and of conformal acquisition functions.
- In Appendix C, we include more experimental results, in particular multi-objective black-box optimization and single-objective tabular ranking tasks with real data.
- In Appendix D, we give implementation details for all experiments.

A ASSUMPTIONS, LIMITATIONS, AND BROADER IMPACTS

A.1 Assumptions

The assumptions underlying the coverage guarantee for conformal prediction are strikingly mild. All else equal, any real-valued, measurable score function will produce a valid prediction set (Vovk et al., 2005). There are trivial examples that produce trivially valid prediction sets $\mathcal{C}_\alpha(\mathbf{x}) = \mathcal{Y},\ \forall \mathbf{x},\ \forall \alpha$. In general if we choose s poorly we pay a price in terms of *efficiency* (i.e. the volume of the prediction sets), but validity is still maintained.

The critical assumption is that $\{(\mathbf{x}_0, \mathbf{y}_0), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ are pseudo-exchangeable.[5] A sequence of random variables is pseudo-exchangeable if the joint density can be factored into terms that only depend on the values of the sequence, not the ordering (Fannjiang et al., 2022). Informally, we can see that BayesOpt satisfies pseudo-exchangeability because the likelihood of the training data is just the mixture of all the previous query likelihoods, and the query likelihoods do not depend on the order of the past observations. Because we make no assumptions about the data distribution beyond pseudo-exchangeability, conformal prediction belongs to a class of methods known as *distribution-free* uncertainty quantification.

A.2 Limitations

Marginal vs. conditional coverage guarantees: full and split conformal prediction sets have marginal coverage guarantee that is easy to confuse with conditional coverage guarantees (Angelopoulos and Bates, 2021). Marginal coverage guarantees must be interpreted with the same frequentist mindset as other frequentist measures of uncertainty, such as confidence intervals and p-values, with similar risks of misinterpretation by inexperienced users. We have attempted to make clear in the main text that the full conformal prediction coverage guarantee is only realized in the aggregate, as the average of coverages observed in many independent, parallel experiments. Coverage observed within any specific trial for any specific input can (and does) vary substantially from the aggregate tendency. There is very recent work which seeks to provide a stronger conditional validity guarantee that can be expected to hold for some $(1 - \delta)$ fraction of trials, which we hope to apply to conformal BayesOpt in future work (Bates et al., 2021).

Approximation error: we have introduced some necessary approximations in this work, notably the discretization of continuous labels and the continuous relaxation of conformal prediction sets. While we have given empirical evidence that the error introduced by these approximations does not appear to be too severe, practitioners should be aware that some deviation from the expected coverage level may occur, as we discuss in Section 5. This limitation is analogous to the numeric limitations of linear algebra implemented with floating point arithmetic. We may be able to make use of Ndiaye (2022) to avoid discretizing continuous outcomes entirely, which we leave for future work.

A.3 Broader Impacts:

Potential negative social impacts: black-box optimization algorithms are application-agnostic. The same algorithms that are being used to design new therapeutics could in theory be used to discover new toxins for bioterror or biowarfare. Similarly, the same algorithms used to design new materials for scientific discovery could be used to design new weapons or rocket fuels. Our work is not particularly vulnerable to misuse relative to the large body of existing work on black-box optimization algorithms.

Machine learning research: phenomena like model misspecification and covariate shift are often blamed on complexity in the external world, but they are also induced by our own behavior, such as choosing a convenient likelihood for a model (even when a more sophisticated option is available) or actively selecting new training data. We hope this work spurs more interest in understanding how to reliably interact with the models we have *today*, in addition to work on "better" models for tomorrow.

Experimental design: applications like materials science and drug discovery require the coordination of large, interdisciplinary teams of scientists and engineers. If machine learning systems are to play a central role in that coordination, they must be reliable, in the sense that the systems should have stable behavior and consistently valid predictions. That kind of reliability requires more than faith in an ad hoc collection of modeling assumptions with limited experimental validation. This work is a step towards machine learning systems with interpretable certificates of reliability that can serve as the foundation on which to build teams which push the boundaries of experimental science.

[5] Note that every IID sequence of random variables is exchangeable, but not every exchangeable sequence is IID. Similarly pseudo-exchangeability does not mean every element of the sequence except for the last is IID.

B PROOFS AND DERIVATIONS

B.1 Smoothed conformal prediction

Algorithm 2 Randomized differentiable conformal prediction masks

Data: train data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=0}^{n-1}$, test point $\mathbf{x}_n$, imp. weights $\mathbf{w}$, label candidates Y_{cand}, score function s, miscoverage tolerance α, relaxation strength τ.

$m_j = 0, \forall j \in \{0, \dots, k-1\}$.

for $\hat{\mathbf{y}}_j \in Y_{\text{cand}}$ **do**

$\hat{\mathcal{D}} \leftarrow \mathcal{D} \cup \{(\mathbf{x}_n, \hat{\mathbf{y}}_j)\}$

$\mathbf{s} \leftarrow [s(\mathbf{x}_0, \mathbf{y}_0, \hat{\mathcal{D}}) \ \cdots \ s(\mathbf{x}_n, \hat{\mathbf{y}}_j, \hat{\mathcal{D}})]^\top$.

$\mathbf{h} \leftarrow \texttt{sigmoid}(\tau^{-1}(\mathbf{s} - s_n))$.

$\overline{w} \leftarrow \mathbf{h}^\top \mathbf{w}$.

$\theta \leftarrow \texttt{clip}(w_n^{-1}(\overline{w} - \alpha), 0, 1)$.

$\eta \sim \text{Bernoulli}(\theta)$.

$\overline{w}' = \overline{w} - (1 - \eta) w_n$.

$m_j \leftarrow \texttt{sigmoid}(\tau^{-1}(\overline{w}' - \alpha))$.

end

Result: $\mathbf{m}$

As discussed in Section 2.3 of the main text, standard conformal prediction (Definition 2.1) is *conservatively valid*, meaning in the long run the coverage of conformal prediction sets is *at least* $1 - \alpha$. If the prediction sets are too conservative, they may be too wide to be helpful for decision-making. In the BayesOpt context we want prediction sets that are *exactly valid*, neither underconfident nor overconfident. Fortunately with a small change (i.e. randomization) conformal prediction sets can be made exactly valid.

Informally, exact validity only requires that we treat an edge case more carefully (see "smoothed conformal predictors" in Vovk et al. (2005) for more details). Specifically there will be some candidate labels $\hat{\mathbf{y}}_j$ that are right on the boundary of the prediction set, and we will introduce randomness to sometimes include such points, and sometimes not, depending on exactly how close to the boundary the points are.

More precisely, there are occasions when

$$\sum_{i=0}^{n-1} h_i w_i < \alpha < \sum_{i=0}^{n} h_i w_i = \overline{w}.$$

In standard conformal prediction the corresponding label $\hat{\mathbf{y}}_j$ would always be accepted into the prediction set. To make a smoothed conformal predictor, we flip a coin with bias $\theta = \texttt{clip}(w_n^{-1}(\overline{w} - \alpha)), 0, 1)$. We call the outcome of the flip η. If $\eta = 1$, then we accept $\hat{\mathbf{y}}_j$, similarly if $\eta = 0$ we reject $\hat{\mathbf{y}}_j$. Note if $\overline{w} - \alpha < 0$ then $\hat{\mathbf{y}}_j$ is always rejected, similarly if $\overline{w} - \alpha > w_n$ then $\hat{\mathbf{y}}_j$ is always accepted. Informally we are checking to see if the contribution of w_n is responsible for pushing $\overline{w}$ over the threshold α, and if so we are probabilistically accepting $\hat{\mathbf{y}}_j$ depending on how close $\overline{w} - w_n$ is to α. We give the continuous relaxation of smoothed conformal prediction in Algorithm 2.

B.2 Characterizing the conformal Bayes posterior

All conditional distributions are also conditioned on $\mathcal{D}$, which we omit from the notation for the sake of clarity. Recall that the conformal Bayes posterior is written as

$$\begin{aligned} p(f(\mathbf{x})|\mathbf{x}) &= \int_{\mathcal{Y}} p(f(\mathbf{x})|\mathbf{x}, \mathbf{y}) p(\mathbf{y}|\mathbf{x}) d\mathbf{y}, \\ &= \int_{C_\alpha(\mathbf{x})} p(f(\mathbf{x})|\mathbf{x}, \mathbf{y}) p(\mathbf{y}|\mathbf{x}) d\mathbf{y} + \int_{\mathcal{Y} \setminus C_\alpha(\mathbf{x})} p(f(\mathbf{x})|\mathbf{x}, \mathbf{y}) p(\mathbf{y}|\mathbf{x}) d\mathbf{y}. \end{aligned}$$

Now we define a new conformal Bayes posterior distribution as a mixture distribution over $\mathbf{y}$,

$$p_\alpha(\mathbf{y}|\mathbf{x}) = (1-\alpha)q_1(\mathbf{y}|\mathbf{x}) + \alpha q_2(\mathbf{y}|\mathbf{x}) \tag{9}$$

$$Z_1 = \int_{C_\alpha(\mathbf{x})} 1 d\mathbf{y}, \qquad q_1(\mathbf{y}|\mathbf{x}) = \begin{cases} 1/Z_1 & \text{if } \mathbf{y} \in C_\alpha(\mathbf{x}), \\ 0 & \text{else,} \end{cases}$$

$$Z_2 = \int_{\mathcal{Y}\setminus C_\alpha(\mathbf{x})} p(\mathbf{y}|\mathbf{x}) d\mathbf{y}, \qquad q_2(\mathbf{y}|\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{y} \in C_\alpha(\mathbf{x}), \\ p(\mathbf{y}|\mathbf{x})/Z_2 & \text{else,} \end{cases}$$

where the normalizing constants Z_1, Z_2 ensure that $\int p_\alpha(\mathbf{y}|\mathbf{x})d\mathbf{y} = 1$ (assuming $C_\alpha(\mathbf{x})$ is bounded and non-empty, so Z_1 is non-zero and finite). If $\mathcal{C}_\alpha(\mathbf{x}) = \mathcal{Y}$ and $\mathcal{Y}$ is unbounded then $p_\alpha(\mathbf{y}|\mathbf{x})$ is not a proper density.

The corresponding conformal Bayes posterior distribution over f is

$$\begin{aligned} p_\alpha(f(\mathbf{x})|\mathbf{x}) &= \int_{\mathcal{Y}} p(f(\mathbf{x})|\mathbf{x},\mathbf{y}) p_\alpha(\mathbf{y}|\mathbf{x}) d\mathbf{y} \\ &= \frac{1-\alpha}{Z_1} \int_{C_\alpha(\mathbf{x})} p(f(\mathbf{x})|\mathbf{x},\mathbf{y}) d\mathbf{y} + \frac{\alpha}{Z_2} \int_{\mathcal{Y}\setminus C_\alpha(\mathbf{x})} p(f(\mathbf{x})|\mathbf{x},\mathbf{y}) p(\mathbf{y}|\mathbf{x}) d\mathbf{y} \end{aligned}$$

Finally we can rewrite both integrals over all $\mathcal{Y}$ by introducing a binary mask,

Definition B.1.

$$\begin{aligned} p_\alpha(f(\mathbf{x})|\mathbf{x}) &:= \frac{1-\alpha}{Z_1} \int m_\alpha(\mathbf{x},\mathbf{y}) p(f(\mathbf{x})|\mathbf{x},\mathbf{y}) d\mathbf{y} \\ &\quad + \frac{\alpha}{Z_2} \int (1 - m_\alpha(\mathbf{x},\mathbf{y})) p(f(\mathbf{x})|\mathbf{x},\mathbf{y}) p(\mathbf{y}|\mathbf{x}) d\mathbf{y}, \\ m_\alpha(\mathbf{x},\mathbf{y}) &:= \begin{cases} 1 & \text{if } \mathbf{y} \in C_\alpha(\mathbf{x}), \\ 0 & \text{else.} \end{cases} \end{aligned}$$

Proposition B.1. Let $n > 1$ and $p_\alpha(f|D)$ be defined according to Definition B.1. Then $p_\alpha(f|D)$ converges pointwise in $\mathbf{x}$ to $p(f(\mathbf{x})|\mathbf{x}, D)$ as $\alpha \to 1$,

$$\lim_{\alpha \to 1} p_\alpha(f(\mathbf{x})|\mathbf{x}) = p(f|\mathbf{x}).$$

Proof:

Let $\varepsilon > 0$, $n > 2$, and define $\alpha_k = 1 - 1/(k+1)$ for $k \in \mathbb{N}$.

$$\begin{aligned} |p_{\alpha_k}(f(\mathbf{x})|\mathbf{x}) - p(f(\mathbf{x})|\mathbf{x})| &= |\Delta_1 + \Delta_2|, \\ &\leq |\Delta_1| + |\Delta_2|, \end{aligned}$$

where

$$\Delta_1 = \frac{1-\alpha_k}{Z_1} \int_{C_{\alpha_k}(\mathbf{x})} p(f(\mathbf{x})|\mathbf{x},\mathbf{y}) d\mathbf{y} - \int_{C_{\alpha_k}(\mathbf{x})} p(f(\mathbf{x})|\mathbf{x},\mathbf{y}) p(\mathbf{y}|\mathbf{x}) d\mathbf{y},$$

$$\Delta_2 = \frac{\alpha_k}{Z_2} \int_{\mathcal{Y}\setminus C_{\alpha_k}(\mathbf{x})} p(f(\mathbf{x})|\mathbf{x},\mathbf{y}) p(\mathbf{y}|\mathbf{x}) d\mathbf{y} - \int_{\mathcal{Y}\setminus C_{\alpha_k}(\mathbf{x})} p(f(\mathbf{x})|\mathbf{x},\mathbf{y}) p(\mathbf{y}|\mathbf{x}) d\mathbf{y}.$$

Recalling the definition of $C_\alpha(\mathbf{x})$ (Def. 2.1), we observe that $C_{\alpha_k}(\mathbf{x}) \supset C_{\alpha_{k+1}}(\mathbf{x}), \forall k \in \mathbb{N}$[6] Furthermore we see that since the importance weights $\mathbf{w}$ must sum to 1 that $\lim_{k\to\infty} C_{\alpha_k}(\mathbf{x}) = \emptyset$.

[6] $A \supset B$ indicates that A is a strict superset of B.

Bounding $|\Delta_1|$:

$$\begin{aligned} |\Delta_1| &\leq |\mathcal{O}(1-\alpha_k) - \mathcal{O}(1-\alpha_k)|, \\ \Rightarrow |\Delta_1| &\leq c_1(1-\alpha_k). \end{aligned}$$

Bounding $|\Delta_2|$:

$$\begin{aligned} |\Delta_2| &\leq |(\alpha_k - 1)\mathcal{O}(1)|, \\ \Rightarrow |\Delta_2| &\leq c_2(1-\alpha_k). \end{aligned}$$

Choose $k \in \mathbb{N}$ large enough that $(c_1 + c_2)(1-\alpha_k) < \varepsilon$. ■

B.3 Monte Carlo integration of conformal acquisition functions

We want to integrate acquisition functions of the form

$$\begin{aligned} a(\mathbf{x}, \mathcal{D}) &= \int u(\mathbf{x}, f, \mathcal{D}) p_\alpha(f|\mathcal{D}) df \\ &= \frac{1-\alpha}{Z_1} \int\int u(\mathbf{x}, f, \mathcal{D}) m_\alpha(\mathbf{x}, \mathbf{y}) p(f(\mathbf{x})|\mathbf{x}, \mathbf{y}) d\mathbf{y} df \\ &\quad + \frac{\alpha}{Z_2} \int\int u(\mathbf{x}, f, \mathcal{D})(1 - m_\alpha(\mathbf{x}, \mathbf{y})) p(f(\mathbf{x})|\mathbf{x}, \mathbf{y}) p(\mathbf{y}|\mathbf{x}) d\mathbf{y} df, \end{aligned}$$

Suppose we have sampled $Y_{\text{cand}} = \{\hat{\mathbf{y}}_j\}_{j=0}^{k-1}$, with $\hat{\mathbf{y}}_j \sim p(\mathbf{y}|\mathbf{x}, \mathcal{D})$, and $f^{(j)} \sim p(f|\mathcal{D} \cup \{(\mathbf{x}, \hat{\mathbf{y}}_j)\})$. Starting with the first term in the sum, we have

$$\frac{1-\alpha}{Z_1} \int\int u(\mathbf{x}, f, \mathcal{D}) m_\alpha(\mathbf{x}, \mathbf{y}) p(f(\mathbf{x})|\mathbf{x}, \mathbf{y}) d\mathbf{y} df \approx \frac{1-\alpha}{Z_1 k} \sum_{j=0}^{k-1} \frac{m_\alpha(\mathbf{x}, \hat{\mathbf{y}}_j)}{p(\hat{\mathbf{y}}_j|\mathbf{x})} u(\mathbf{x}, f^{(j)}, \mathcal{D}).$$

We estimate the normalization constant Z_1 as follows:

$$\begin{aligned} Z_1 = \int_{C_\alpha(\mathbf{x})} 1 d\mathbf{y} &= \int m_\alpha(\mathbf{x}, \mathbf{y}) d\mathbf{y} \\ &\approx \frac{1}{k} \sum_{j=0}^{k-1} \frac{m_\alpha(\mathbf{x}, \hat{\mathbf{y}}_j)}{p(\hat{\mathbf{y}}_j|\mathbf{x})} \end{aligned}$$

By similar logic the second term in the sum is estimated as follows:

$$\frac{\alpha}{Z_2} \int\int u(\mathbf{x}, f, \mathcal{D})(1 - m_\alpha(\mathbf{x}, \mathbf{y})) p(f(\mathbf{x})|\mathbf{x}, \mathbf{y}) p(\mathbf{y}|\mathbf{x}) d\mathbf{y} df \approx \frac{\alpha}{Z_2 k} \sum_{j=0}^{k-1} (1 - m_\alpha(\mathbf{x}, \hat{\mathbf{y}}_j)) u(f^{(j)}, \mathcal{D}),$$

where

$$\begin{aligned} Z_2 = \int_{\mathcal{Y} \setminus C_\alpha(\mathbf{x})} p(\mathbf{y}|\mathbf{x}) d\mathbf{y} &= \int (1 - m_\alpha(\mathbf{x}, \mathbf{y})) p(\mathbf{y}|\mathbf{x}) d\mathbf{y} \\ &\approx \frac{1}{k} \sum_{j=0}^{k-1} (1 - m_\alpha(\mathbf{x}, \hat{\mathbf{y}}_j)) \end{aligned}$$

Upon further inspection we see that k drops out of the equations, and in effect we are simply computing weighted sums, where the weights have been normalized to sum to 1, i.e.

$$\begin{aligned} a(\mathbf{x}, \mathcal{D}) \approx \hat{a}(\mathbf{x}, \mathcal{D}) &= (1-\alpha)\mathbf{u}^\top \mathbf{v} + \alpha \mathbf{u}^\top \mathbf{v}', \\ \mathbf{u} &= [u(\mathbf{x}, f^{(0)}, \mathcal{D}) \ \cdots \ u(\mathbf{x}, f^{(k-1)}, \mathcal{D})]^\top, \\ \mathbf{v}_j &= \frac{m_\alpha(\mathbf{x}, \hat{\mathbf{y}}_j)}{p(\hat{\mathbf{y}}_j|\mathbf{x})} \left(\sum_{i=0}^{k-1} \frac{m_\alpha(\mathbf{x}, \hat{\mathbf{y}}_i)}{p(\hat{\mathbf{y}}_i|\mathbf{x})} \right)^{-1}, \\ \mathbf{v}'_j &= (1 - m_\alpha(\mathbf{x}, \hat{\mathbf{y}}_j)) \left(\sum_{i=0}^{k-1} (1 - m_\alpha(\mathbf{x}, \hat{\mathbf{y}}_i)) \right)^{-1}. \end{aligned}$$

B.4 Conformalized Single Objective Acquisitions

Conformal NEI: rather than taking $u(\mathbf{x}, f, \mathcal{D}) = [f(\mathbf{x}) - \max_{\mathbf{y}_i \in \mathcal{D}} \mathbf{y}_i]_+$ (which corresponds to EI), take

$$u_{\text{NEI}}(\mathbf{x}, f, \mathcal{D}) = [f(\mathbf{x}) - \max_{\mathbf{x}'_i \in \mathcal{D}} f(\mathbf{x}'_i)]_+. \tag{10}$$

Note that u is now a function of the joint collection of function evaluations $\{f(\mathbf{x}), f(\mathbf{x}'_0), \ldots, f(\mathbf{x}'_{n-1})\}$ (Letham et al., 2019).

Conformal UCB: the reparameterized form of UCB was originally derived in Wilson et al. (2017) as follows:

$$\text{UCB}(\mathbf{x}) = \int u_{\text{UCB}}(\mathbf{x}, f, \mathcal{D}) \mathcal{N}\left(f \middle| \mu, \frac{\lambda\pi}{2}\Sigma \right) df,$$

where $\lambda > 0$ is a hyperparameter balancing the explore-exploit tradeoff, μ, Σ are the mean and covariance of $p(f|\mathcal{D})$, and $u_{\text{UCB}}(\mathbf{x}, f, \mathcal{D}) = \mu + \frac{\lambda\pi}{2}|\mu - f|$. Because UCB is optimistic, the conformalization procedure is a little different than the previous acquisition functions. When marginalizing out the outcomes $\mathbf{y}$ to obtain the conformal Bayes posterior, we integrate over the restricted outcome space $\mathcal{Y}_\mu = \{\mathbf{y} \in \mathcal{Y} | \mathbf{y} \geq \mu\}$. Hence we derive conformal UCB as

$$\text{CUCB}_\alpha(\mathbf{x}) = \int \int_{\mathcal{Y}_\mu} u_{\text{UCB}}(\mathbf{x}, f, \mathcal{D}) \mathcal{N}\left(f \middle| \mu(\mathbf{y}), \frac{\lambda\pi}{2}\Sigma(\mathbf{y}) \right) p_\alpha(\mathbf{y}|\mathbf{x}, \mathcal{D}) d\mathbf{y} df,$$

where $\mu(\mathbf{y}), \Sigma(\mathbf{y})$ are the predictive mean and covariance of $p(f|\mathcal{D} \cup \{(\mathbf{x}, \mathbf{y})\})$.

B.5 Conformalized Multi-Objective Acquisition Functions

When there are multiple objectives of interest, a single best design $\mathbf{x}^*$ may not exist. Suppose there are d objectives, $f^* : \mathcal{X} \to \mathbb{R}^d$. The goal of multi-objective optimization (MOO) is to identify the set of *Pareto-optimal* solutions such that improving one objective within the set leads to worsening another. We say that $\mathbf{x}$ dominates $\mathbf{x}'$, or $f^*(\mathbf{x}) \succ f^*(\mathbf{x}')$, if $f^*_k(\mathbf{x}) \geq f^*_k(\mathbf{x}')$ for all $k \in \{1, \ldots, d\}$ and $f^*_k(\mathbf{x}) > f^*_k(\mathbf{x}')$ for some k. The set of *non-dominated* solutions $\mathscr{X}^*$ is defined in terms of the Pareto frontier (PF) $\mathcal{P}^*$,

$$\mathscr{X}^\star = \{\mathbf{x} : f(\mathbf{x}) \in \mathcal{P}^\star\}, \quad \text{where } \mathcal{P}^\star = \{f(\mathbf{x}) \ : \ \mathbf{x} \in \mathcal{X}, \ \nexists \, \mathbf{x}' \in \mathcal{X} \text{ s.t. } f(\mathbf{x}') \succ f(\mathbf{x})\}. \tag{11}$$

MOO algorithms typically aim to identify a finite approximation to $\mathscr{X}^\star$, which may be infinite, within a reasonable number of iterations. One way to measure the quality of an approximate PF $\mathcal{P}$ is to compute the hypervolume $\text{HV}(\mathcal{P}|\mathbf{r}_{\text{ref}})$ of the polytope bounded by $\mathcal{P} \cup \{\mathbf{r}_{\text{ref}}\}$, where $\mathbf{r}_{\text{ref}} \in \mathbb{R}^d$ is a user-specified *reference point*.

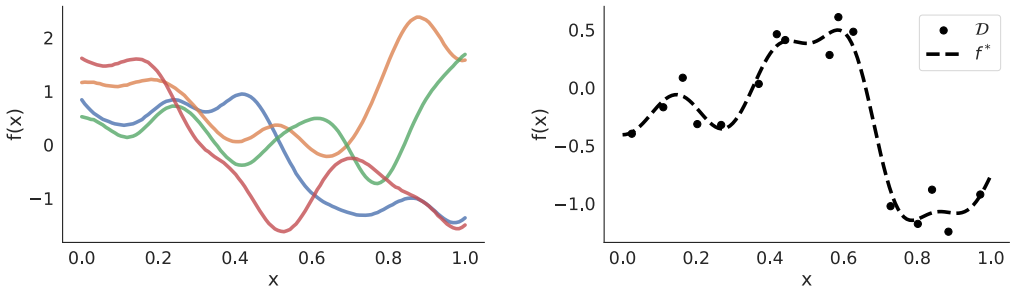
$$u_{\text{EHVI}}(\mathbf{x}, f, \mathcal{D}) = \text{HVI}(\mathcal{P}', \mathcal{P}|\mathbf{r}_{\text{ref}}) = [\text{HV}(\mathcal{P}'|\mathbf{r}_{\text{ref}}) - \text{HV}(\mathcal{P}|\mathbf{r}_{\text{ref}})]_+, \tag{12}$$

where $\mathcal{P}' = \mathcal{P} \cup \{\hat{f}(\mathbf{x})\}$ (Emmerich, 2005; Emmerich et al., 2011; Daulton et al., 2020). If our measurements of f are noisy we cannot compute HV exactly and instead must substitute $\hat{f} \sim p(f|\mathcal{D})$, i.e.

$$u_{\text{NEHVI}}(\mathbf{x}, f, \mathcal{D}) = \text{HVI}(\hat{\mathcal{P}}'_t, \hat{\mathcal{P}}_t|\mathbf{r}_{\text{ref}}), \tag{13}$$

where $\hat{\mathcal{P}}_t = \{\hat{f}(\mathbf{x}) \ : \ \mathbf{x} \in \mathcal{X}_t, \ \nexists \, \mathbf{x}' \in \mathcal{X}_t \text{ s.t. } \hat{f}(\mathbf{x}') \succ \hat{f}(\mathbf{x})\}$ and $\hat{\mathcal{P}}' = \hat{\mathcal{P}} \cup \{\hat{f}(\mathbf{x})\}$ (Daulton et al., 2021a).

Our derivations hold for so-called composite acquisitions as well, so we could also extend to qParEGO and qNParEGO variants for multi-objective optimization (Daulton et al., 2020, 2022).



(a) (b)

Figure 9: Here we demonstrate how to generate target functions with known RKHS norm w.r.t. a given kernel κ, in this case an RBF kernel with lengthscale $\ell = 0.1$ on $\mathcal{X} = [0, 1]$. In the **left** panel we show prior draws $f^{(i)} \sim \mathcal{GP}(0, \kappa)$. In the **right** panel we show a synthetic target function f^* with corresponding RKHS norm $||f^*||_\kappa = 2.0934$. We produced f^* according to Algorithm 3, using $n = 16$ basis points with dimension $d = 1$ and noise variance $\sigma^2 = 0.04$.

B.6 Conformalizing Batch Acquisitions

In general batch acquisitions have the form

$$a(\mathbf{x}_0, \dots, \mathbf{x}_{q-1}, \mathcal{D}) = \int \max_{i<q} u(\mathbf{x}_i, f, \mathcal{D}) p(f|\mathcal{D}) df. \tag{14}$$

Note that $f(\mathbf{x}_0), \dots, f(\mathbf{x}_{q-1})$ are sampled jointly when estimating Eq. (14) with Monte Carlo methods. Increasing the query batch size to q increases the dimensionality of the outcome to $q \times d$, where d is the number of objectives. Our importance-sampling MC integration procedure introduced in Section 4.1 scales gracefully with higher outcome dimensionality, we simply sample the elements of Y_{cand} from $p(y(\mathbf{x}_0), \dots, y(\mathbf{x}_{q-1})|\mathbf{x}_{0:q-1}, \mathcal{D})$.

The bigger challenge arises in computing the conformal masks for batched query outcomes. In our current implementation we compute the conformal scores (the Bayes posterior log-likelihood) pointwise for each query batch element, with corresponding pointwise conformal prediction masks. We apply the pointwise masks before computing $\max_{i<q} u$ across query batch elements. The alternative would be to compute a joint conformal score across all query batch elements (similarly computing joint scores for each of the previous query batches in the training data). Note that this second approach essentially reduces to replacing each datum $(\mathbf{x}_i, \mathbf{y}_i)$ in Eq. (3) with $(X_i, Y_i) = ([\mathbf{x}_0 \cdots \mathbf{x}_{q-1}]^\top, [\mathbf{y}_0 \cdots \mathbf{y}_q]^\top)$. We leave the implementation of this second approach for future work.

B.7 Out-of-distribution queries

If $p'(\mathbf{x}|\mathcal{D}) \neq p(\mathbf{x})$, and $w(\mathbf{x}, \mathcal{D}) > \alpha$, then every candidate label is automatically accepted to the prediction set and $\mathcal{C}_\alpha(\mathbf{x}) = \mathcal{Y}$, which makes $p_\alpha(\mathbf{x}|\mathcal{D})$ an improper density. Intuitively the issue is there are not enough points in $\mathcal{D}$ close enough to $\mathbf{x}$ to guarantee a miscoverage rate of at most α unless every possible label is included in the prediction set. Our solution is to set any conformal acquisition value $a_\alpha(\mathbf{x}, \mathcal{D})$ to 0 if $w(\mathbf{x}, \mathcal{D}) > \alpha$. In practice we achieve this effect by introducing a second mask $m' = \texttt{sigmoid}(\tau^{-1}(w_n - \alpha))$ which we apply to Eq. (7). For conformal EI and other similar acquisition functions, this mask simply means we will favor any point close enough to the dataset to guarantee coverage if it has positive expected utility over any out-of-distribution point. For conformal UCB, this mask means all out-of-distribution points are assigned the value $1/n \sum_{\mathbf{y}_i \in \mathcal{D}} \mathbf{y}_i$ (assuming the labels have been mean-centered during preprocessing).

B.8 Generating target functions with known RKHS norm

As we noted in Section 3, many formal regret bounds for BayesOpt rest on the assumption that 1) the RKHS norm of f^* corresponding to the choice of kernel κ is bounded (i.e. $||f^*||_\kappa < \infty$) and 2) that we know a reasonable upper bound on the RKHS norm. Although these assumptions almost certainly do *not* hold for every f^* encountered in practice, it is worthwhile to examine the behavior of conformal acquisition functions when the idealized assumptions do hold, to see what price we pay for robustness.

Although it is enough to know a bound on $||f^*||_\kappa$ in this section we will optimize target functions for which we know the

Algorithm 3 Generating target functions with known RKHS norm

Data: kernel κ, # basis points n, input dimension d, noise variance σ^2
Sample basis points $X = \{\mathbf{x}_0, \dots, \mathbf{x}_{n-1}\} \sim \texttt{SobolSequence}(n, d)$
Sample $\mathbf{u} = \{u_0, \dots, u_{n-1}\} \sim \mathcal{N}(\mathbf{0}, \kappa(X, X))$ (GP prior draw with kernel κ)
Sample $\mathbf{y} = \{y_0, \dots, y_{n-1}\} \sim \mathcal{N}(\mathbf{u}, \sigma^2 I_n)$
Compute $\mathbf{c} = (K_{XX} + \sigma^2 I)^{-1}\mathbf{y}$
$f^*(\cdot) = \kappa(\cdot, X)\mathbf{c}$ (GP posterior mean with kernel κ conditioned on $\mathcal{D}$)
$B = \sqrt{\mathbf{c}^\top K_{XX}\mathbf{c}}$ (RHKS norm $||f^*||_\kappa$)
Result: f^*, B

RKHS norm exactly, to remove slack in the bound as a potential experimental confounder. A simple approach, which we summarize in Algorithm 3, is to choose a random set of n basis points $X = \{\mathbf{x}_0, \dots \mathbf{x}_{n-1}\}$, draw random function values from the GP prior $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \kappa(X, X))$ (Figure 9a), draw noisy labels $\mathbf{y} \sim \mathcal{N}(\mathbf{u}, \sigma^2 I_n)$, and take f^* to be the GP posterior mean $\mu_{f|\mathcal{D}}$ conditioned on the synthetic dataset $\mathcal{D} = (X, \mathbf{y})$ (Figure 9b).

The proof that $\mu_{f|\mathcal{D}} \in \mathcal{H}_\kappa(X)$ and $||\mu_{f|\mathcal{D}}||_\kappa = \sqrt{\mathbf{c}^\top K_{XX}\mathbf{c}}$ follows directly from the definition of an RKHS (see Chapter 2.2 in Scholkopf and Smola (2018)). Although these are well-established results, we sketch the proof here for the reader's convenience.

Proof: Recall that by definition, the RKHS corresponding to κ with basis points X can be written as

$$\mathcal{H}_\kappa(X) = \overline{\{f \ : \ f(\cdot) = \sum_{i=0}^{n-1} c_i \kappa(\cdot, \mathbf{x}_i) \text{ for some } \mathbf{c} \in \mathbb{R}^n\}}, \tag{15}$$

where $\overline{\mathcal{S}}$ is the closure of $\mathcal{S}$.

Since the GP posterior mean conditioned on $\mathcal{D}$ is given by

$$\mu_{f|\mathcal{D}}(\cdot) = \kappa(\cdot, X)(K_{XX} + \sigma^2 I)^{-1}\mathbf{y},$$

and $\mathbf{c} = (K_{XX} + \sigma^2 I)^{-1}\mathbf{y}$ is in $\mathbb{R}^n$, we've verified that $\mu_{f|\mathcal{D}} \in \mathcal{H}_\kappa(X)$.

For any $f(\cdot) = \sum_{i=0}^{n-1} c_i \kappa(\cdot, \mathbf{x}_i)$ for some $\mathbf{c} \in \mathbb{R}^n$ we can derive the corresponding RKHS norm as follows:

$$\begin{aligned}
||f||_\kappa^2 &= \langle f, f \rangle_\kappa, \\
&= \left\langle \sum_{i=0}^{n-1} c_i \kappa(\cdot, \mathbf{x}_i), \sum_{j=0}^{n-1} c_j \kappa(\cdot, \mathbf{x}_j) \right\rangle_\kappa, \\
&= \sum_{i,j=0}^{n-1} c_i c_j \left\langle \kappa(\cdot, \mathbf{x}_i), \kappa(\cdot, \mathbf{x}_j) \right\rangle_\kappa, \\
&= \sum_{i,j=0}^{n-1} c_i c_j \kappa(\mathbf{x}_i, \mathbf{x}_j), \\
\Rightarrow ||f||_\kappa &= \sqrt{\mathbf{c}^\top K_{XX}\mathbf{c}}.
\end{aligned} \tag{16}$$

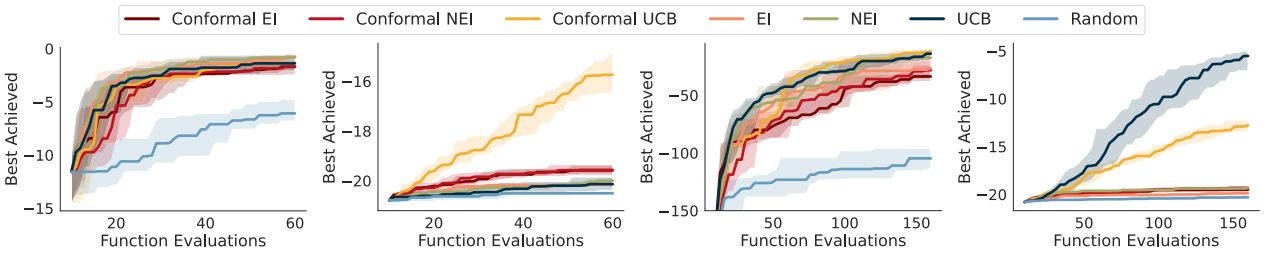
The second line follows from the linearity of $\langle \cdot, \cdot \rangle_\kappa$, and the third line follows from the reproducing property of $\mathcal{H}_\kappa(X)$ and symmetry of κ. ■

Remark: we are not required to use $\mu_{f|\mathcal{D}}$, once we have chosen κ and X we could in principle take any linear combination of $\{k(\cdot, \mathbf{x}_0), \dots, k(\cdot, \mathbf{x}_{n-1})\}$, e.g. by sampling $\mathbf{c} \sim \mathcal{N}(\mathbf{0}, I)$. However $\mu_{f|\mathcal{D}}$ is convenient and sufficient for our purposes.

When generating target functions in this way, it is important to account for the dimensionality of the input d when choosing the number of basis points n. Due to the curse of dimensionality, the number of points needed to produce "interesting" functions (i.e. functions that are not flat almost everywhere) grows exponentially with the dimension of the input when using default GP kernels (e.g. Matérn).

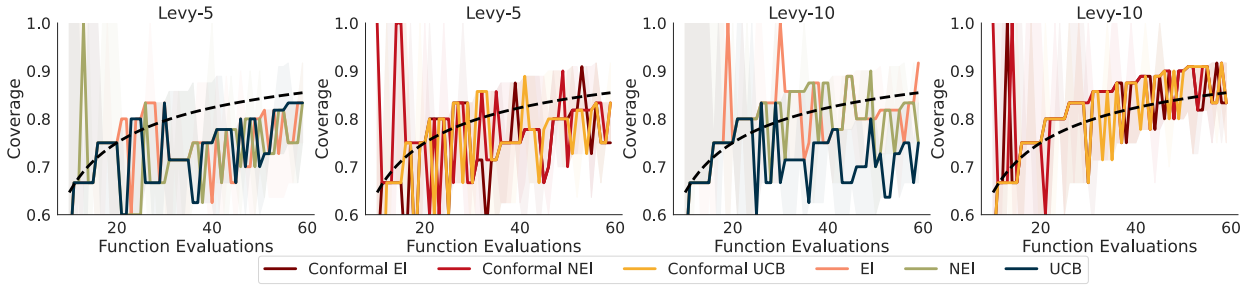
C ADDITIONAL EXPERIMENTAL RESULTS

C.1 Single-Objective Black-Box Optimization

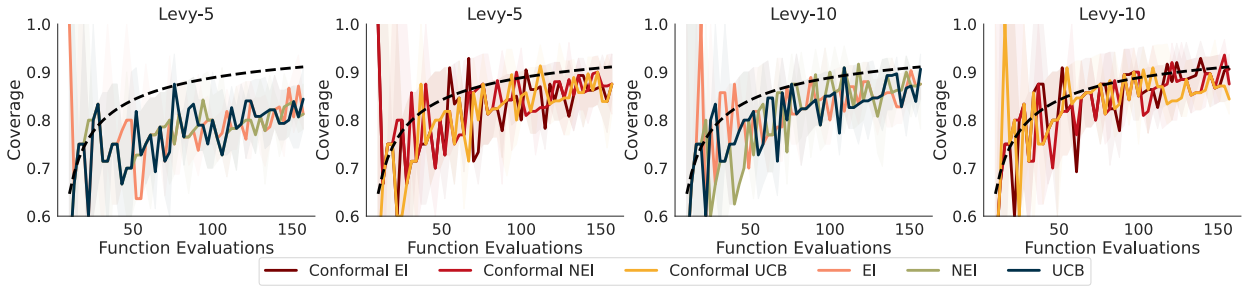


(a) Levy-5, $q = 1$. (b) Ackley-20, $q = 1$. (c) Levy-20, $q = 3$. (d) Ackley-20, $q = 3$.

Figure 10: BayesOpt best objective value found with conformal and standard acquisition functions on single-objective tasks Levy-d and Ackley-d (reporting median and its 95% conf. interval, estimated from 25 trials). qEI, qNEI, conformal qEI, and conformal qNEI all perform similarly, conformal qUCB is best everywhere except Ackley-20, where it comes second after qUCB.



(a) Levy 5 and 10, $q = 1$.



(b) Levy 5 and 10, $q = 3$.

Figure 11: BayesOpt empirical coverage of conformal and credible prediction sets evaluated on holdout data from single-objective task Levy-d (reporting median and its 95% conf. interval, estimated from 25 trials). The conformal coverage curves track the target $1 - \alpha$ (black dashed line) well, significantly better than the credible curves, which tend to be overconfident. Median w/ 95% confidence interval is shown.

In Figure 10 we investigate the effect of the choice of acquisition function on sample-efficiency, comparing conventional and conformal versions. In particular we consider expected improvement (EI), noisy expected improvement (NEI) and upper confidence bound (UCB) alongside their conformal counterparts. No clear ranking emerges here, however UCB and conformal UCB both perform well in general.

In Figure 11, we investigate the sensitivity of coverage on random holdout data to the query batch size q and the dimensionality of the inputs d. Here, we plot the median and its 95% confidence interval as shading, finding that the conformal sets are better calibrated in a frequentist sense than Bayesian credible sets.

C.2 Multi-Objective Black-Box Optimization

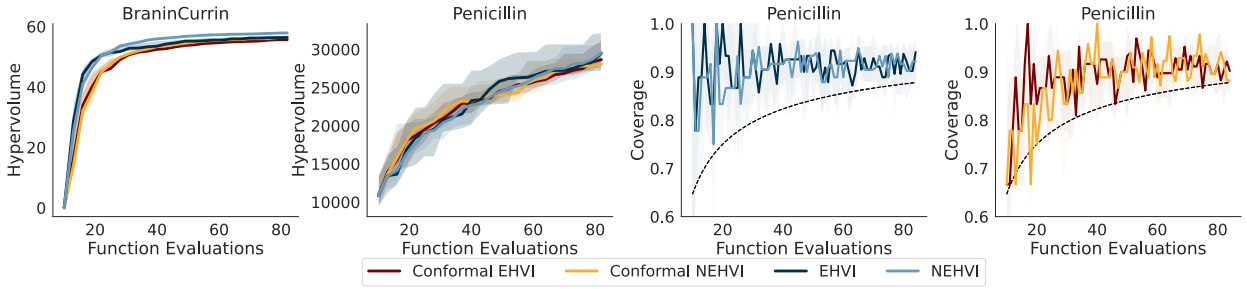
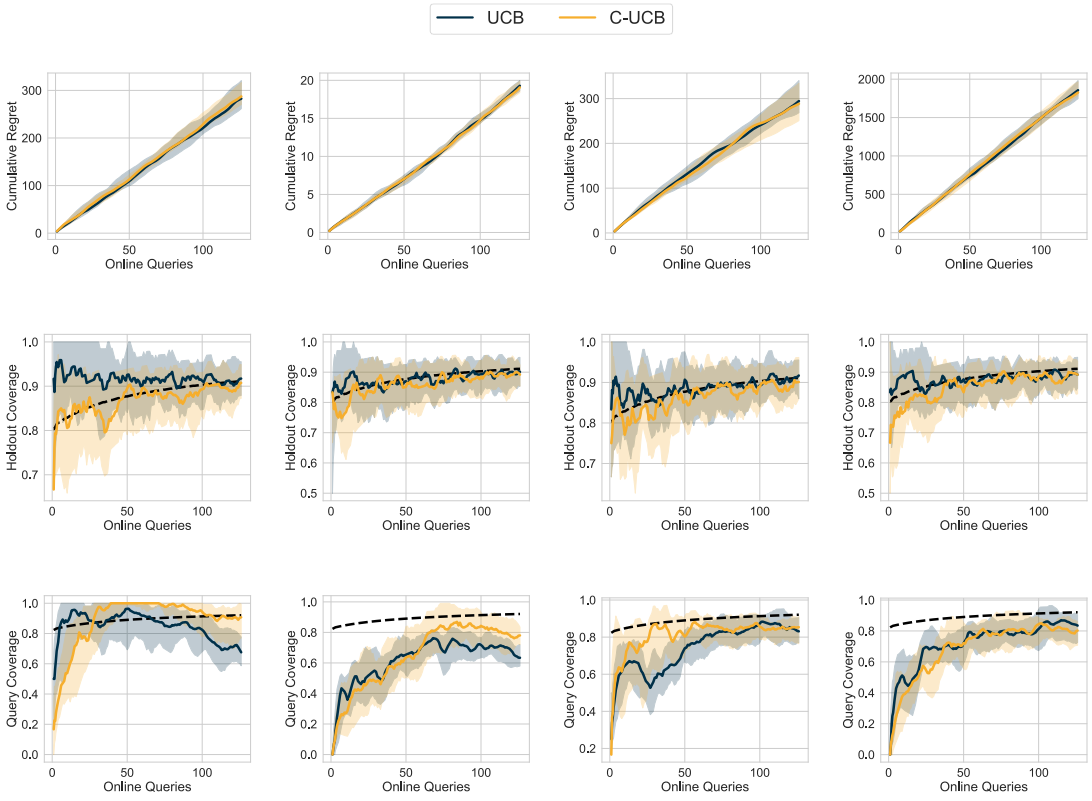


Figure 12: BayesOpt results on multi-objective tasks `branin-currin` and `penicillin` (reporting median and its 95% conf. interval, estimated from 25 trials). **Left two panels:** Both conformal and standard acquisitions find solution sets with similar hypervolumes. **Right two panels:** Credible and conformal empirical coverage curves. The conformal curves track the target $1 - \alpha$ (black dashed line) better than the credible curves, but both are underconfident.

To demonstrate that our approach scales to multi-objective tasks, we consider two tasks, `branin-currin` ($d = 2$) and `penicillin` ($d = 3$) (Liang and Lai, 2021). The goal is not to find a single $\mathbf{x}^*$, but rather to find the set of all non-dominated solutions, the Pareto front (Appendix B.5). By non-dominated, we mean the set of solutions with the property that the objective value cannot increase in one dimension without decreasing in another. We report results using the expected hypervolume improvement (EHVI) (Emmerich, 2005; Emmerich et al., 2011; Daulton et al., 2020) and noisy expected hypervolume improvement (NEHVI) (Daulton et al., 2021a) as the base acquisition functions in Figure 12. Like the single-objective case conformal BayesOpt is comparable in terms of sample-efficiency as quantified by the solution hypervolume relative to a common reference point (Beume et al., 2009), and conformal set coverage tracks $(1 - \alpha)$ more closely than credible set coverage. All black-box functions used in this paper are synthetic with implementations coming from BoTorch (Balandat et al., 2020). The Penicillin function was originally proposed by Liang and Lai (2021).

In this setting the performance of conventional BayesOpt and conformal BayesOpt is very similar in terms of solution quality, and the improvement to coverage is fairly small (Figure 12). The root issue appears to be that it is simply much more difficult to accurately characterize conformal prediction sets in multiple dimensions, since intervals become polyhedra (Johnstone and Ndiaye, 2022). Although our Bayesian discretization scheme avoids the exponential memory usage of dense grids, its ability to pinpoint the boundary of conformal prediction sets appears to degrade as the dimensionality of $\mathbf{y}$ increases.

C.3 Tabular Ranking with Real-World Drug and Antibody Data



(a) penalized logP (b) QED (c) DRD3 binding affinity (d) OAS Protein Stability

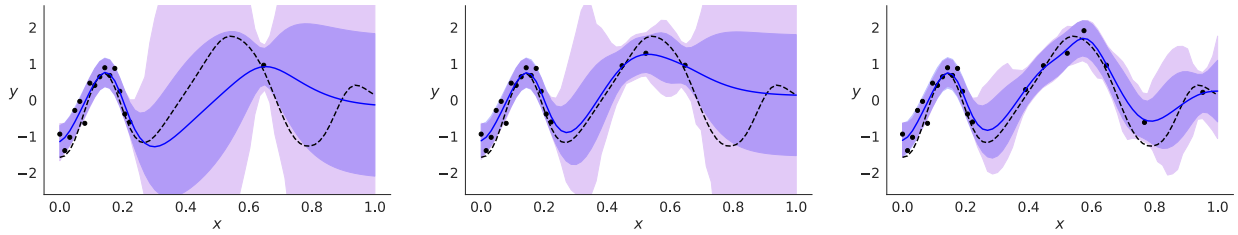
Figure 13: Result ranking tabular molecular datasets for drug-related properties such as solubility (logP) **(a)**, empirical drug-likeness score (QED) **(b)**, dopamine receptor (DRD3) binding affinity **(c)** and antibody stability **(d)**. Across datasets, CUCB selects queries with more consistent coverage than UCB **(bottom two rows)**, with identical sample efficiency **(top row)**. The midpoint, lower, and upper bounds of each curve depict the 50%, 20%, and 80% quantiles, estimated from 4 trials.

Sometimes instead of solving $\max_{\mathbf{x}\in\mathcal{X}} a(\mathbf{x}, \mathcal{D})$, the search space is restricted to a discrete subset of candidates $X_{\text{cand}} \subset \mathcal{X}$. This restriction is particularly common for tasks with discrete decision variables, such as biological sequence design (Stanton et al., 2022). This existence of a fixed candidate set simplifies the computation of conformal acquisition functions substantially, since we can use samples from X_{cand} directly when training the ratio estimator $\hat{r}$, rather than relying on bootstrapped SGLD as discussed in Section 4.3.

To emulate this kind of application, we compared standard and conformal UCB on a selection of small and large molecule ranking tasks. In particular, we ranked a subset of small molecules drawn from the ZINC dataset (Krenn et al., 2020) for three target properties, penalized logP (solubility), QED (drug-likeness), and DRD3 (dopamine receptor) binding affinity (Gómez-Bombarelli et al., 2018; Huang et al., 2021). We also ranked a subset of large antibody molecules drawn from the OAS dataset (Hornung et al., 2014) for stability. For simplicity we did not use sequence-based representations of the molecules, instead relying on RDKit chemical descriptors (Landrum, 2016) and BioPython sequence descriptors (Cock et al., 2009) to generate continuous feature representations of the small and large molecules, respectively.

Starting with the 32 worst entries in our labeled dataset, we selected 128 candidates sequentially ($q = 1$), revealing the corresponding label and retraining the surrogate after each new selection. We share our results in Figure 13. Because selection is restricted to a prespecified candidate set, the coverage is less consistent than the black-box optimization setting, however we find that conformal UCB still selects queries with better coverage overall, without sacrificing sample-efficiency (measured by cumulative regret, i.e. the difference between the selected candidate label and the best possible label of the remaining candidates).

C.4 Comparing Bayesian Credible Sets and Conformal Bayes Prediction Sets in the Well-Specified Regime



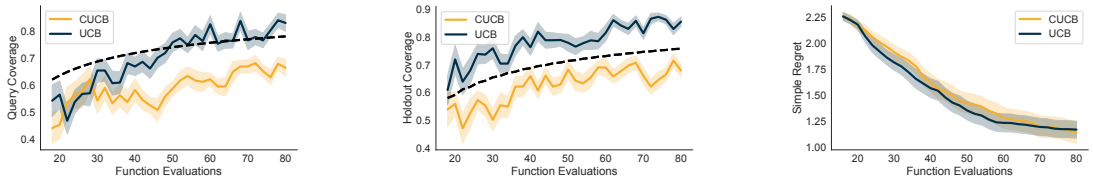
(a) $t = 2$ (b) $t = 4$ (c) $t = 8$

Figure 14: A qualitative example of the difference between conformal Bayes prediction sets and Bayes credible sets in the well-specified regime (f^* generated with a Matérn-$5/2$ kernel with lengthscale $\ell = 0.1$ and $n = 16$ basis points). In regions with plentiful training data conformal and credible predictions sets are essentially indistinguishable. In regions where training data is sparse, conformal prediction sets are underconfident and much wider than credible sets. The credible sets are well-calibrated across the full domain regardless of the amount of training data because the GP prior is highly concordant with the actual target function.

In Appendix B.8 we discussed a method for generating target functions from a pre-specified RKHS. Not only do target functions generated in this way have a computable RKHS norm, they also allow us to compare the behavior of conformal Bayes prediction sets and Bayes credible sets in the well-specified regime. Since we know the kernel we used to generate the target function, we can use the same kernel during inference, eliminating one of the possible causes of poor coverage.

As we noted in Section 2.3, conformal Bayes produces the most efficient (i.e. the smallest by volume in expectation under $p(f)$) prediction sets among all prediction rules which are guaranteed to be valid at the $1 - \alpha$ level. Nevertheless the validity guarantee does come with a price. In Figure 14 we show that in the well-specified regime conformal Bayes tends to be underconfident where training data is sparse, whereas Bayes credible sets are well-calibrated across the whole domain. This result is expected and typical of Bayesian methods. If we have very good knowledge of the nature of f^* we are better off fully exploiting that knowledge than relying solely on vague assumptions (e.g. pseudo-exchangeability). The choice between conformal Bayes and conventional Bayesian methods is inherently context-dependent, a function of the available data, the user's confidence in their prior and the cost of miscalibration if that confidence is misplaced.

In Figure 15 we repeat the experiment in Section 5.3 using the same procedure as above to generate f^*, increasing the input dimension to 5 and the number of basis points to 128. We report the query and holdout coverage, along with the simple regret, $f^*(\mathbf{x}^*) - \max_{\mathbf{x}_i \in \mathcal{D}} f^*(\mathbf{x}_i)$. Although there is no advantage to using conformal prediction in the well-specified regime, we find the performance to be comparable.



(a) Query Coverage (b) Holdout Coverage (c) Simple Regret

Figure 15: In this experiment we examine the coverage and simple regret of BayesOpt in the well-specified regime. We plot the mean and the standard error estimated from 25 trials. Here f^* is generated with a Matérn-$5/2$ kernel with lengthscale $\ell = 0.1$ and $n = 128$ basis points in $[0, 1]^5$. As expected, Bayesian credible sets (indicated by the **UCB** curve) are fairly well-calibrated. The coverage of conformal prediction sets is less satisfactory (the **CUCB** curve), due in part to the fact that the training data is not IID, and in part to density ratio approximation error.

D IMPLEMENTATION DETAILS

D.1 Bringing Everything Together

Algorithm 4 Pseudocode for the conformal BayesOpt inner loop

Input: train data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=0}^{n-1}$, initial solution $\mathbf{x}_n$, score function s, miscoverage tolerance α, sigmoid temperature τ_σ, SGLD learning rate $\eta_{\mathbf{x}}$, # SGLD steps $t_{\max}$, SGLD temperature τ_{SGLD}, classifier learning rate η_θ, EMA parameter γ.
Initialize classifier q_θ, set weight average $\bar{\theta} = \mathbf{0}$.
Initialize classifier dataset $\mathcal{D}' = \{(\mathbf{x}_i, 0)\}_{i=0}^{n-1}$
for $t = 0, \ldots, t_{\max} - 1$ **do**
  Estimate $\hat{r}_t(\mathbf{x}_i), \forall i \in \{0, \ldots, n\}$ with $q_{\bar{\theta}}$. (Eq. 8)
  $(\mathbf{w}_t)_i = \hat{r}_t(\mathbf{x}_i) / \sum_k \hat{r}_t(\mathbf{x}_k), \forall i \in \{0, \ldots, n\}$.
  $Y_{\text{cand}} \leftarrow \{\hat{\mathbf{y}}_j\}_{j=0}^{m-1}$ s.t. $\hat{\mathbf{y}}_j \sim \hat{p}(\mathbf{y}|\mathbf{x}'_t, \mathcal{D})$.
  $\mathbf{m} = \text{outcome\_mask}(\mathcal{D}, \mathbf{x}_n, \mathbf{w}_t, Y_{\text{cand}}, s, \alpha, \tau_\sigma)$. (Algorithm 1)
  Estimate acquisition value $a(\mathbf{x}_n)$. (Eq. (7))
  Update $\mathbf{x}_n \leftarrow \text{sgld\_step}(\mathbf{x}_n, a(\mathbf{x}_n), \eta_{\mathbf{x}}, \tau_{\text{SGLD}})$.
  Update $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(\mathbf{x}_n, 1)\}$.
  Update $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \ell(\theta, \mathcal{D}')$
  Update $\bar{\theta} \leftarrow (1 - \gamma)\bar{\theta} + \gamma\theta$.
end for
Return: $\mathbf{x}_n$

In Algorithm 4 we summarize the entire conformal BayesOpt inner loop used to select new queries.

D.2 Stable Predictions on the Training Set

We found that computing the GP posterior negative log-likelihood (and its gradients) on training data to be numerically unstable and so used stochastic diagonal estimation to estimate the posterior variances. Plugging in $K = \kappa(X, X)$ into that posterior mean and variance, we get that the posterior mean is $K(K + \sigma^2)^{-1}\mathbf{y}$ and the posterior covariance is $\Sigma = \sigma^2 I + K - K(K + \sigma^2 I)^{-1}K$. Unfortunately, the second term ends up being unstable as it requires solving (and then subtracting) a (batched) system of size $n \times n$. To see the reason for instability, note that as $\sigma^2 I \to 0$ then the entire covariance matrix tends to zero.

We originally tried backpropagating through an eigendecomposition; however, this produced ill-defined gradients, see the explanation in Ionescu et al. (2015). Instead we computed a stochastic diagonal estimate, using the identities

$$\Sigma = \sigma^2 I + \sigma^2 K(K + \sigma^2 I)^{-1}, \tag{17}$$

$$\text{diag}(\Sigma)_i \approx \sigma^2 \left(1 + \left(\sum_{j=1}^{J} \mathbf{z}^{(j)} \odot K(K + \sigma^2 I)^{-1}\mathbf{z}^{(j)}\right)_i \left(\sum_{j=1}^{J} \mathbf{z}^{(j)} \odot \mathbf{z}^{(j)}\right)_i^{-1}\right), \tag{18}$$

where the probe vector $\mathbf{z}^{(j)}$ has i.i.d Bernoulli entries and $\odot$ is the Hadamard product. This estimator comes from Bekas et al. (2007) and is in spirit quite similar to Hutchinson's trace estimator for the log determinant. We used $J = 10$ probe vectors.

D.3 Hyperparameters

For all GP models in this paper, we used the default single task GP (`SingleTaskGP`) model from BoTorch, which uses a scaled Matern-$5/2$ kernel with automatic relevance determination and a Gamma$(3, 6)$ prior over the lengthscales and a Gamma$(2, 0.15)$ prior over the outputscales. We used constant prior mean functions. For the likelihood, we used a softplus transformation to optimize the raw noise, constraining the noise to be between 5×10^{-4} and 0.5. To fit the GP kernel hyperparameters ϕ, we used BoTorch's default fitting utility, `fit_gpytorch_model`, which uses L-BFGS-B to maximize the log-marginal likelihood $\log p(\mathcal{D}|\phi)$.

For the miscoverage tolerance we used a simple schedule $\alpha = \max(0.05, 1/\sqrt{n})$. Note if $\alpha < 1/n$ then $\mathcal{C}_\alpha(\mathbf{x}) = \mathcal{Y},\ \forall \mathbf{x} \in \mathcal{X}$.

We initialized $\mathcal{D}_0$ with 10 Sobol points drawn from a random orthant of the normalized input space. The input normalization was computed from known bounds of the black-box functions.

Black-Box Optimization Hyperparameters	
Name	**Value**
# Optimization rounds	50
q (query batch size)	$\{1, 3\}$
$\lvert\mathcal{D}_0\rvert$	10
σ (normalized measurement noise scale)	0.1
τ_σ (sigmoid temp.)	1e-2
k (i.e. $\lvert Y_{\text{cand}}\rvert$)	256
# SGLD chains	5
$t_{\max}$ (# SGLD total steps)	100
t_{burn} (# SGLD burn-in steps)	25
$\eta_{\mathbf{x}}$ (SGLD learning rate)	1e-3
τ_{SGLD} (SGLD temp.)	1e-3
η_θ (classifier learning rate)	1e-3
γ (classifier EMA weight)	2e-2
λ (classifier weight decay)	1e-4
Random seeds	$\{0, \ldots, 24\}$

Tabular Ranking Hyperparameters	
Name	**Value**
# Optimization rounds	128
q (query batch size)	1
$\lvert\mathcal{D}_0\rvert$	32
σ (normalized measurement noise scale)	n/a
τ_σ (sigmoid temp.)	1e-6
k (i.e. $\lvert Y_{\text{cand}}\rvert$)	64
# number classifier gradient updates	256
η_θ (classifier learning rate)	1e-3
γ (classifier EMA weight)	1
λ (classifier weight decay)	1e-4
Random seeds	$\{0, \ldots, 3\}$

D.4 Computational Complexity

The cost of training the surrogate GP regression model is the same in our case as conventional BayesOpt, namely $\mathcal{O}(n^3)$ if exact GP inference is used without any approximations.

The cost of *retraining* the surrogate GP on a single new example $(\mathbf{x}_n, \hat{\mathbf{y}}_j)$ is $\mathcal{O}(n)$, since one can make use of efficient low-rank updates to the root decomposition of $(K_{XX} + \sigma^2 I)^{-1}$ (Gardner et al., 2018). The surrogate GP can be retrained on all k candidate labels in parallel on a GPU, which keeps the wall-clock cost of retraining to $\mathcal{O}(n)$ but increases the memory footprint by a factor of k. The increased memory footprint persists for the duration of the selection phase, during which the acquisition function is optimized to select the next query.

The cost of drawing a sample function $f^{(j)} \sim p_\alpha(f|\hat{\mathcal{D}}_j)$ is the same as drawing a sample from $p(f|\mathcal{D})$, and whereas conventional BayesOpt methods typically draw multiple samples from $p(f|\mathcal{D})$, we find that drawing a single sample from each $p_\alpha(f|\hat{\mathcal{D}}_j)$ is sufficient since $p_\alpha(f(\mathbf{x}_n)|\hat{\mathcal{D}}_j)$ tends to concentrate near $\hat{\mathbf{y}}_j$.

Therefore the complexity of each acquisition function gradient evaluation is dominated by the same $\mathcal{O}(qn^2)$ cost of exact GP test-time inference with q test examples as conventional BayesOpt, at the cost of a k-factor increase in memory usage.

Although this analysis would seem to indicate that conformal BayesOpt should run in similar wall-clock time to conventional BayesOpt, in practice it is considerably slower because the SGLD chains in conformal BayesOpt must be allowed to burn in, and then the bootstrapped ratio estimator must be given time to converge, which requires many more gradient evaluations

than optimizing a conventional BayesOpt acquisition function with L-BFGS. In our experiments we found conformal BayesOpt to be around an order of magnitude slower in terms of wall-clock time than conventional BayesOpt. This increase in runtime is a considerable drawback and merits further investigation in future work.

D.5 Compute Resources

Our experiments were conducted on a range of NVIDIA GPUs, including RTX 2080 Tis, Titan RTXs, V100s, and A100s in high-performance computing clusters. All experiments used a single GPU at a time. It would require approximately 250 GPU hours to reproduce the experiments in this paper by our estimate,

$$1 \text{ GPU hr/seed} \times 25 \text{ seeds per variant} \times 1 \text{ variant per experiment} \times 10 \text{ experiments} = 250 \text{ hrs}.$$

Other experimental runs, e.g. development and debugging, probably consumed an order of magnitude more GPU hours.

D.6 Software Packages

- Python 3, PSF License Agreement (Van Rossum and Drake, 2009).
- Matplotlib, Matplotlib License Agreement.
- Seaborn, BSD License.
- NumPy, BSD License (Harris et al., 2020).
- PyTorch, BSD License (Paszke et al., 2019).
- GPyTorch, MIT License (Gardner et al., 2018).
- BoTorch, MIT License (Balandat et al., 2020).