Bayesian Optimization by Density Ratio Estimation

Louis C. Tiao, Aaron Klein, Cédric Archambeau, Edwin V. Bonilla, Matthias Seeger, Fabio Ramos
1University of Sydney 2Amazon Web Services 3CSIRO’s Data61 4NVIDIA

Abstract

Bayesian optimization (BO) is among the most effective and widely-used blackbox optimization methods. BO proposes solutions according to an explore-exploit trade-off criterion encoded in an acquisition function, many of which are derived from the posterior predictive of a probabilistic surrogate model. Prevalent among these is the expected improvement (EI). Naturally, the need to ensure analytical tractability in the model poses limitations that can ultimately hinder the efficiency and applicability of BO. In this paper, we cast the computation of EI as a binary classification problem, building on the well-known link between class-probability estimation (CPE) and density ratio estimation (DRE), and the lesser-known link between density ratios and EI. By circumventing the tractability constraints imposed on the model, this reformulation provides numerous natural advantages, not least in scalability, increased flexibility, and greater representational capacity.

1 Introduction

Bayesian optimization (BO) is a sample-efficient methodology for the optimization of expensive black-box functions [4, 23]. In brief, BO proposes candidate solutions according to an acquisition function that encodes the explore-exploit trade-off. At the core of BO is a probabilistic surrogate model based on which the acquisition function can be computed. The probabilistic model of choice in BO is commonly the Gaussian process (GP), owing to its flexibility and ability to yield uncertainty estimates.

However, GP-based BO can also be hampered by the limitations of GPs. Notably, they a) scale cubically with the number of observations [33] and b) assume stationarity, i.e. that the covariances between outputs are translation-invariant with respect to their inputs [26]. Further, they are not inherently equipped to deal with c) discrete variables, ordered or otherwise (i.e. categorical), and d) variables with conditional dependency structures [13]. Naturally, to address these issues, much of the focus has been directed toward extending the surrogate model itself. This has often led to ad hoc extensions that, by necessity of ensuring analytical tractability, place strong and oversimplifying assumptions at the expense of expressiveness.

Recognizing that the surrogate model is only a means to an end (i.e. of constructing an acquisition function), we seek to express acquisition functions in an alternate form that does not impose analytical tractability constraints on the surrogate model. Of particular interest is the expected improvement (EI) function [20], which has seen widespread adoption. Remarkably, Bergstra et al. [2] demonstrate that the EI function can be expressed as the relative ratio between two densities [34]. To estimate this density ratio, they propose a method, known as the tree-structured Parzen estimator (TPE), which can naturally deal with tree-structured inputs, discrete inputs, and scales linearly with the number of observations.

In this paper, we underscore the potential shortcomings of the TPE approach for tackling the general density ratio estimation (DRE) problem. In § 2, we highlight, among other issues that may lead to numerical instability, its tendency to scale poorly to higher dimensions [28]. In § 3, we explore
more powerful alternatives to fully exploit the link between DRE and the EI function, namely, DRE by class-probability estimation (CPE). This approach retains the strengths of TPE while scaling better with the dimensionality, and enables one to build arbitrarily expressive classifiers. Depending on the choice of classifier, it is possible to capture not only non-linear, but also non-stationary and heteroscedastic behaviours. Our experiments in § 4 demonstrate that our BO-by-DRE approach, termed BORE, competes favorably with state-of-the-art blackbox optimization algorithms on a variety of challenging synthetic test problems and meta-surrogate benchmarks for automated machine learning (AUTOML) [16].

2 Background

Let \( x \in \mathcal{X} \) denote an input to the blackbox function \( f : \mathcal{X} \to \mathbb{R} \) and \( y \sim \mathcal{N}(f(x), \sigma^2) \) the noisy observation of the corresponding output with noise variance \( \sigma^2 \). Further, let \( \mathcal{D}_N = \{(x_n, y_n)\}_{n=1}^N \) denote the set of observations.

**Expected improvement.** First we define a threshold \( \tau = \Phi^{-1}(\gamma) \) where constant \( \gamma \) denotes some quantile of the observed \( y \) values, i.e. \( \gamma = \Phi(\tau) = p(y < \tau \mid \mathcal{D}_N) \). Let \( I_\gamma(x) \) be the utility function that quantifies the non-negative improvement over \( \tau \)

\[
I_\gamma(x) = \max(\tau - y, 0).
\]

Then, the EI function \( \alpha_\gamma(x; \mathcal{D}_N) \) is defined as the expected value of \( I_\gamma(x) \) under the posterior predictive distribution \( p(y \mid x, \mathcal{D}_N) \)

\[
\alpha_\gamma(x; \mathcal{D}_N) = \mathbb{E}_{p(y \mid x, \mathcal{D}_N)}[I_\gamma(x)].
\]

Note that the dependence of \( \alpha_\gamma(x; \mathcal{D}_N) \) on \( \gamma \) occurs in \( \tau \) (which is implicitly a function of \( \gamma \)). For example, \( \gamma = 0 \) leads to the conventional setting of \( \tau = \min_n y_n \). In what follows, we will consider a relaxation of \( \tau \) with settings where \( \gamma > 0 \).

**Relative density ratio.** Let \( \ell(x) \) and \( g(x) \) be a pair of densities. The \( \gamma \)-relative density ratio of \( \ell(x) \) and \( g(x) \) is defined as

\[
r_\gamma(x) = \frac{\ell(x)}{\gamma \ell(x) + (1 - \gamma) g(x)},
\]

where \( \gamma \ell(x) + (1 - \gamma) g(x) \) denotes the \( \gamma \)-mixture density with mixing proportion \( 0 \leq \gamma < 1 \) [34]. Note that for \( \gamma = 0 \), we recover the ordinary density ratio \( r_0(x) = \ell(x)/g(x) \). Further, observe that we can directly express \( r_\gamma(x) \) as a function of \( r_0(x) \),

\[
r_\gamma(x) = (\gamma + r_0(x)^{-1}(1 - \gamma))^{-1}.
\]
Now, suppose $\ell(x)$ and $g(x)$ are distributions such that $x \sim \ell(x)$ if $y < \tau$, and $x \sim g(x)$ if $y \geq \tau$. Then, let us express the conditional $p(x \mid y, D_N)$ in terms of $\ell(x)$ and $g(x)$

$$p(x \mid y, D_N) = \begin{cases} \ell(x) & \text{if } y < \tau, \\ g(x) & \text{if } y \geq \tau. \end{cases}$$

(5)

Remarkably, under this construction (illustrated in Figure 1), Bergstra et al. [2] demonstrate that the EI function can be expressed as the $\gamma$-relative density ratio, up to some constant multiplicative factor

$$\alpha_{\gamma}(x; D_N) \propto r_{\gamma}(x).$$

(6)

For completeness, we provide the self-contained derivation in Appendix A. Hence, this result shows that the problem of maximizing EI reduces to that of maximizing the $\gamma$-relative density ratio,

$$x^* = \arg \max_{x \in \mathcal{X}} \alpha_{\gamma}(x; D_N) = \arg \max_{x \in \mathcal{X}} r_{\gamma}(x),$$

(7)

for which a wide variety of approaches are available [29].

### 2.1 Tree-structured Parzen estimator

In practice, to solve the optimization problem of eq. 7, Bergstra et al. [2] propose taking the following approach:

1. Since the relative density ratio $r_{\gamma}(x)$ is a monotonically non-decreasing function of the ordinary density ratio $r_0(x)$, they restrict their attention to maximizing the latter,

$$x^* = \arg \max_{x \in \mathcal{X}} r_0(x),$$

(8)

thus, effectively ignoring the mixing proportion $\gamma$ altogether.

2. Then, they estimate the ordinary density ratio $r_0(x)$ by separately estimating its constituent numerator $\ell(x)$ and denominator $g(x)$ using a tree-based variant of kernel density estimation (KDE) [25].

It is easy to see why this approach might be favorable compared to methods based on GP regression: one now incurs an $O(N)$ computational cost as opposed to the $O(N^3)$ cost of GP posterior inference. Furthermore, it is equipped to deal with tree-structured, mixed continuous, ordered and unordered discrete inputs. In spite of its advantages, TPE is not without potential pitfalls as discussed next.

### 2.2 Potential pitfalls

The first major drawback of TPE lies within step 1:

**Singularities.** Relying on the ordinary density ratio can result in numerical instabilities since it is unbounded, and often diverges to infinity even in simple toy scenarios. In contrast, the $\gamma$-relative density ratio is always bounded above by $\gamma^{-1}$ when $\gamma > 0$ [34].

The other potential problems of the TPE lie within step 2:

**Vapnik’s principle.** Conceptually, independently estimating the densities is actually a more cumbersome approach that violates Vapnik’s principle—namely, that when solving a problem of interest, one should refrain from resorting to solve a more general problem as an intermediate step [32]. In this instance, density estimation is a more general problem that is arguably more difficult than density ratio estimation.

**Kernel bandwidth.** KDE depends crucially on the selection of an appropriate kernel bandwidth, which is notoriously difficult [21, 24]. Furthermore, even with an optimal selection of a single fixed bandwidth, it cannot simultaneously adapt to low- and high-density regions [31].

**Error sensitivity.** These difficulties are exacerbated by the fact that one is required to select two bandwidths, whereby the optimal bandwidth for an individual density may not necessarily be appropriate for approximating the density ratio. Indeed, it may even have deleterious effects. This makes this approach unforgiving to any error in the individual approximations, particularly in that of the denominator $g(x)$, which has an outsized influence on the resulting density ratio.
We refer to Appendix B for derivations. Equations (12) and (13) establish the precise link between the density ratio and the class-posterior probability. Hence, in the so-called BO loop (summarized in Algorithm 1), we alternately optimize the classifier parameters \( \theta \) wrt to the BCE loss (to improve its approximation to the true class-posterior probability; Line 2) and the classifier input \( x \) wrt to its output (to suggest the next candidate to evaluate; Line 3).

In traditional GP-based methods, Line 3 typically consists of maximizing the EI function, explicitly expressed using the statistics of the GP posterior predictive (its mean, variance, pdf and cdf), while
We describe the experiments conducted to empirically evaluate our method. The classifier with our approach, the parameterized function $\pi$, is known to have a tendency of over-exploitation. This is especially manifest in the SIX-HUMP CAMEL problem, which is designed to have multiple local minima. In visualizations included in Appendix F, we show that TPE consistently gets bogged down in a local minimum, while BORE is able to balance the exploitation of the various local minima with the exploration of other parts of the space.

4 Experiments

We describe the experiments conducted to empirically evaluate our method. The classifier $\pi_\theta(x)$ is a multi-layer perceptron (MLP), with 2 hidden layers, each with 32 units. We consistently found $\text{elu}$ activations [7] to be particularly effective for low-dimensional problems, with $\text{relu}$ remaining otherwise the best choice. We optimize the weights with ADAM [14] using batch size of $B = 64$. For candidate suggestion, we optimize the input of the classifier wrt to its output using multi-started L-BFGS with three random restarts. To encourage exploration, we suggest random candidates at a proportion $\epsilon = 0.1$ of the time. Further details concerning the implementation and setting of additional hyperparameters are provided in Appendix E.

Synthetic test functions. We first consider a number of challenging synthetic test functions for optimization [30], namely, the BRANIN, SIX-HUMP CAMEL, MICHALEWICZ5D, and HART-MANN6D functions. To quantitatively assess performance, we report the immediate regret, defined as the absolute error between the global minimum and the lowest function value attained thus far. To assess the sample efficiency, we compare the immediate regret over the number of function evaluations against baselines. The baselines shown here are Random Search and TPE, as implemented in the HyperOpt library [1]. For each method, we show the mean and 95% confidence interval (CI) across results obtained from 20 repeated runs.

The results are shown in Figure 2. Across these problems, we find BORE to be competitive against TPE. In the BO folklore, TPE is known to have a tendency of over-exploitation. This is especially manifest in the SIX-HUMP CAMEL problem, which is designed to have multiple local minima. In visualizations included in Appendix F, we show that TPE consistently gets bogged down in a local minimum, while BORE is able to balance the exploitation of the various local minima with the exploration of other parts of the space.

Meta-surrogate benchmarks. Finally, we compare against a range of state-of-the-art optimization methods on the meta-surrogate benchmarks, PROFET [16]. First, BO methods with different kinds of probabilistic models: GP-based BO (GP-BO) [15], the random forest-based SMAC [12] and TPE [2], and second, two evolutionary algorithms: CMA-ES [10] and differential evolution (DE) [27]. PROFET emulates the hyperparameter tuning of common machine learning algorithms such as support vector machines (SVMs) [8], gradient boosting (XGBoost) [5], and others, on classification and regression problems, by sampling tasks (in the form of objective functions) from a generative

---

**Algorithm 1**: Bayesian optimization by density ratio estimation (BORE).

1. **while** under budget do
2. \[ \theta^* \leftarrow \arg \min_{\theta} \mathcal{L}(\theta) \] // update classifier by optimizing parameters $\theta$ wrt BCE loss
3. \[ x_N \leftarrow \arg \max_{x \in X} \pi_{\theta^*}(x) \] // suggest new candidate by optimizing input $x$ wrt classifier output
4. \[ y_N \leftarrow f(x_N) \] // obtain $y_N$ by evaluating blackbox function at $x_N$
5. \[ D_N \leftarrow D_{N-1} \cup \{(x_N, y_N)\} \] // update dataset
6. **end**

Line 2 would be the optimization of the GP hyperparameters wrt the marginal likelihood. By analogy with our approach, the parameterized function $\pi_\theta(x)$ is itself an approximation to the EI function to be maximized directly, while the approximation is tightened by improving its fit to the true class-posterior probabilities, in turn through optimization wrt the BCE loss.

In short, we have reduced the problem of computing EI to that of training a probabilistic classifier, thus unlocking a broad range of possible alternatives to GPs. This enables one to employ virtually any state-of-the-art classification method available to parameterize the classifier with arbitrarily expressive approximators that have the capacity to deal with non-linear, non-stationary, and heteroscedastic phenomena commonly encountered in BO.

In this work, we parameterize $\pi_\theta(x)$ by a feed-forward neural network (NN). This is an attractive choice not only for its universal approximation guarantees [11] but because a) one can easily adopt stochastic gradient descent (SGD) methods to scale up its parameter learning [17], and b) it is differentiable end-to-end, which enables the use of quasi-Newton methods such as L-BFGS [18] for candidate suggestion.
meta-model. We sampled 50 tasks for each of the three benchmark classes: META-SVM, META-FCNET, and META-XGBoost. Each optimizer was evaluated on each task with 20 independent runs using different random seeds. To aggregate the performance across tasks, we follow the protocol of Klein et al. [16] and report the average ranks and the empirical cdf (ECDF) of the runtime, using a single run of Random Search with 200 iterations as targets.

Figure 3 shows the results on all benchmark classes, from which we see that BORE consistently performs better than all other baselines. In particular, observe that despite GP-BO approaching the optimum faster in the early stages, it is eventually outperformed by BORE after having observed
a sufficient amount of data. Lastly, note that we were only able to run GP-BO for 100 function evaluations on the META-FCNET and META-XGBoost benchmarks due to its prohibitively-high computational overhead.

5 Conclusion

We have presented a novel methodology for BO based on the observation that the problem of computing EI can be reduced to that of probabilistic classification. This observation is made through the well-known link between CPE and DRE, and the lesser-known insight that EI can be expressed as a relative density ratio between two unknown distributions. We discussed important ways in which TPE, an early attempt to exploit the latter, falls short. Further, we demonstrated that a prototype implementation of this methodology, based on a simple feed-forward NN, can outperform TPE and be competitive with state-of-the-art derivative-free optimization methods.

A key appeal of this methodology lies in the room it allows for variations. Indeed, any other state-of-the-art classification method can readily be applied. In particular, SVMs, random forests, and XGBoost may prove to be strong contenders against NNs. Another axis of variation worth exploring is the potential benefits that other direct DRE methods may have to offer. In future work, we will also explore the effects of a fully-Bayesian treatment of the classifier parameters.

Overall, the simplicity and effectiveness of BORE make it a promising approach for blackbox optimization, and its high degree of extensibility provides many exciting avenues for future work.

Acknowledgments and Disclosure of Funding

L. Tiao is supported by an Australian Government Research Training Program (RTP) Scholarship and a CSIRO Data61 Postgraduate Scholarship.

References


