BORE: Bayesian Optimization by Density-Ratio Estimation

Louis C. Tiao[1,3] Aaron Klein[2] Matthias Seeger[2] Edwin V. Bonilla[3,1] Cédric Archambeau[2] Fabio Ramos[1,4]

[1]University of Sydney [2]Amazon [3]CSIRO's Data61 [4]NVIDIA

Summary

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.
- Most acquisition functions are derived from the posterior predictive of a probabilistic surrogate model. Prevalent among these is the expected improvement (EI).
- The need to ensure analytical tractability in the model poses limitations that can hinder the efficiency and applicability of BO.
- We cast the computation of EI as a probabilistic 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 in terms of expressiveness, versatility, and scalability.

Bayesian Optimization (BO)

- Find input $\mathbf{x} \in \mathcal{X}$ that maximizes blackbox function $f : \mathcal{X} \to \mathbb{R}$

$$\mathbf{x}_\star = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$$

given noisy observations $y \sim \mathcal{N}(f(\mathbf{x}), \sigma^2)$ with noise variance σ^2.

- Build *probabilistic surrogate model* upon observations $\mathcal{D}_N = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$.

Expected Improvement (EI)

- The **improvement** utility function quantifies the improvement over some τ

$$U(\mathbf{x}, y, \tau) \coloneqq \max(\tau - y, 0).$$

- Then, the **expected improvement** acquisition function is the expected value of $U(\mathbf{x}, y, \tau)$ under the **posterior predictive** $p(y \,|\, \mathbf{x}, \mathcal{D}_N)$

$$\alpha(\mathbf{x}; \mathcal{D}_N, \tau) = \mathbb{E}_{p(y \,|\, \mathbf{x}, \mathcal{D}_N)}[U(\mathbf{x}, y, \tau)].$$

- If $p(y \,|\, \mathbf{x}, \mathcal{D}_N)$ is *Gaussian*, $\alpha(\mathbf{x}; \mathcal{D}_N, \tau)$ has analytic form (easy to evaluate)
- But this comes at a price—*guaranteeing analytical tractability of the posterior often requires placing strong and oversimplifying assumptions at the expense of expressiveness.*

Strategy

Observation—*we only care about* $p(y \,|\, \mathbf{x}, \mathcal{D}_N)$ *to the extent that we can compute* $\alpha(\mathbf{x}; \mathcal{D}_N, \tau)$.

- Why not instead find an alternative formulation of $\alpha(\mathbf{x}; \mathcal{D}_N, \tau)$ that doesn't depend explicitly on $p(y \,|\, \mathbf{x}, \mathcal{D}_N)$?

Relative Density-Ratio

- Let $\ell(\mathbf{x})$ and $g(\mathbf{x})$ be a pair of distributions.
- The γ-**relative density-ratio** of $\ell(\mathbf{x})$ and $g(\mathbf{x})$ is defined as

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

where $\gamma \ell(\mathbf{x}) + (1-\gamma) g(\mathbf{x})$ is the γ-mixture density with $0 \leq \gamma < 1$ [3].

- For $\gamma = 0$, we recover the **ordinary density-ratio**

$$r_0(\mathbf{x}) = \frac{\ell(\mathbf{x})}{g(\mathbf{x})}.$$

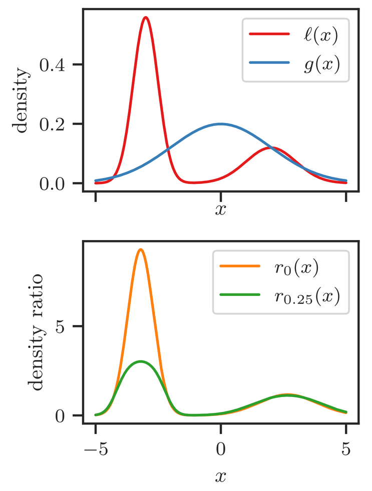


Figure: Example 1D densities

Density-Ratio Estimation (DRE)

- Since $r_\gamma(\mathbf{x}) = h(r_0(\mathbf{x}))$ where h is *monotonically non-decreasing*, it is justifiable to maximize $r_\gamma(\mathbf{x})$ by instead maximizing $r_0(\mathbf{x})$.
- An obvious way to estimate $r_0(\mathbf{x})$ is to separately estimate $\ell(\mathbf{x})$ and $g(\mathbf{x})$ using kernel density estimation (KDE) or some variant thereof, such as the tree-structured Parzen estimator (TPE) [1].
- This simplistic approach has major flaws, and has long since been superseded by *direct* DRE methods such as CPE, KMM, KLIEP, ULSIF, RULSIF, etc [2].
- Conceptually, the simplest of these is class-probability estimation (CPE), i.e. **probabilistic classification**—*something we know how to do well!*

Key Connections

The *expected improvement* function is proportional to a *class-posterior probability*. Hence, it can be readily estimated through *probabilistic classification*.

$$\underbrace{\alpha\left(\mathbf{x}; \mathcal{D}_N, \Phi^{-1}(\gamma)\right)}_{\text{expected improvement}} \propto \underbrace{r_\gamma(\mathbf{x})}_{\text{relative density-ratio}} \propto \underbrace{\pi(\mathbf{x})}_{\text{class-posterior probability}} \qquad \therefore \quad \mathbf{x}_\star = \arg\max_{\mathbf{x}\in\mathcal{X}} \alpha\left(\mathbf{x}; \mathcal{D}_N, \Phi^{-1}(\gamma)\right) = \arg\max_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x})$$

EI vs. Density-Ratio

- Let **threshold** τ be γ-th quantile of observed y values $\tau \coloneqq \Phi^{-1}(\gamma)$ where

$$\gamma = \Phi(\tau) \coloneqq p(y \leq \tau; \mathcal{D}_N).$$

- Define $\ell(\mathbf{x}) \coloneqq p(\mathbf{x} \,|\, y \leq \tau; \mathcal{D}_N)$ and $g(\mathbf{x}) \coloneqq p(\mathbf{x} \,|\, y > \tau; \mathcal{D}_N)$.
- Remarkably, it can be shown that EI can be expressed as the γ-relative density-ratio, up to some constant factor [1]

$$\boxed{\alpha\left(\mathbf{x}; \mathcal{D}_N, \Phi^{-1}(\gamma)\right) \propto r_\gamma(\mathbf{x})}$$

- *Example.* See 1D example in Figure 2 below with $\gamma = 1/3$

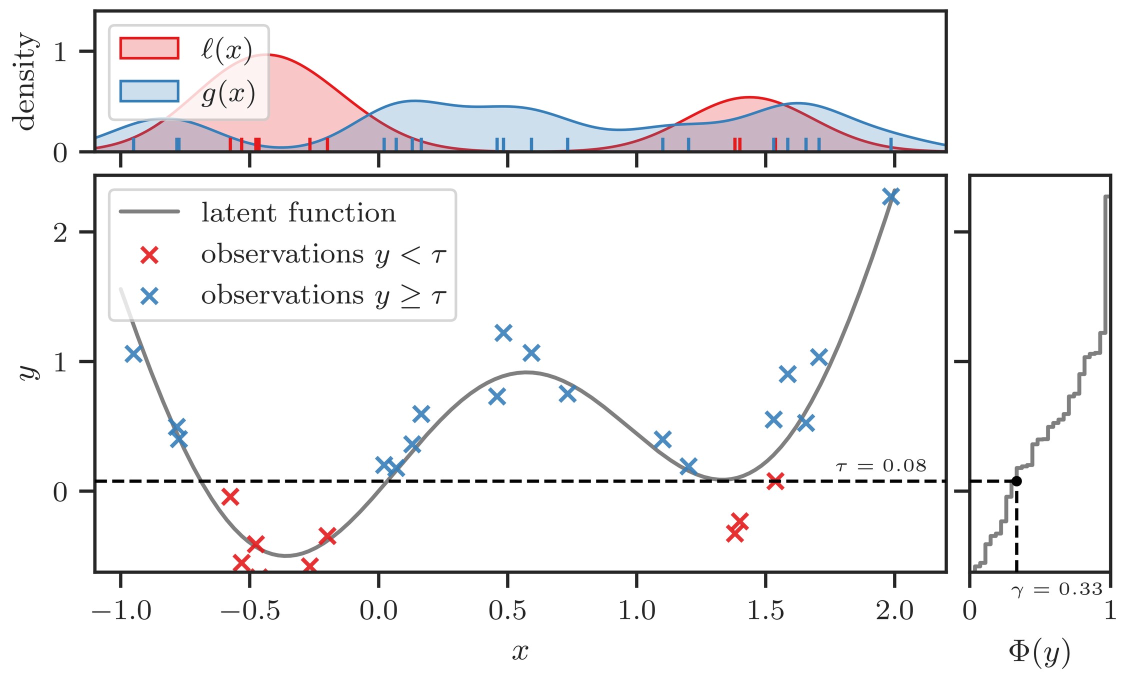


Figure: Synthetic test function $f(x) = \sin(3x) + x^2 - 0.7x$ with observation noise $\varepsilon \sim \mathcal{N}(0, 0.2^2)$.

Density-Ratio vs. Class-posterior Probability

- Construct a **binary classification problem** by introducing labels

$$z = \begin{cases} 1 & \text{if } y \leq \tau, \\ 0 & \text{if } y > \tau. \end{cases}$$

- Denote the **class-posterior probability** by $\pi(\mathbf{x}) = p(z = 1 \,|\, \mathbf{x})$.
- The γ-relative density-ratio is equivalent to the class-posterior probability, up to a constant factor

$$\boxed{r_\gamma(\mathbf{x}) = \gamma^{-1}\pi(\mathbf{x})}$$

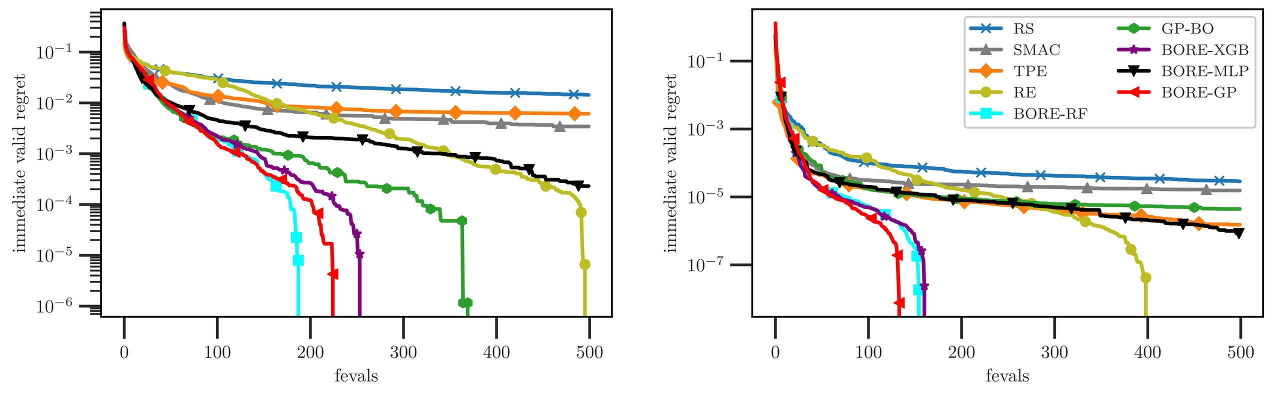
BO by Probabilistic Classification

- Estimate $\pi(\mathbf{x})$ by training a probabilistic classifier $\pi_{\boldsymbol{\theta}}(\mathbf{x})$ parameterized by $\boldsymbol{\theta}$
- Different families of classifiers have complementary strengths, e.g.,
  - feed-forward neural networks: multi-layer perceptrons (MLPs)
  - ensembles of decision trees: random forests (RFs), gradient-boosted trees (XGBOOST)
  - GP classifiers (GPCs)
- The so-called BO loop is summarized in Algorithm 1 below.

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

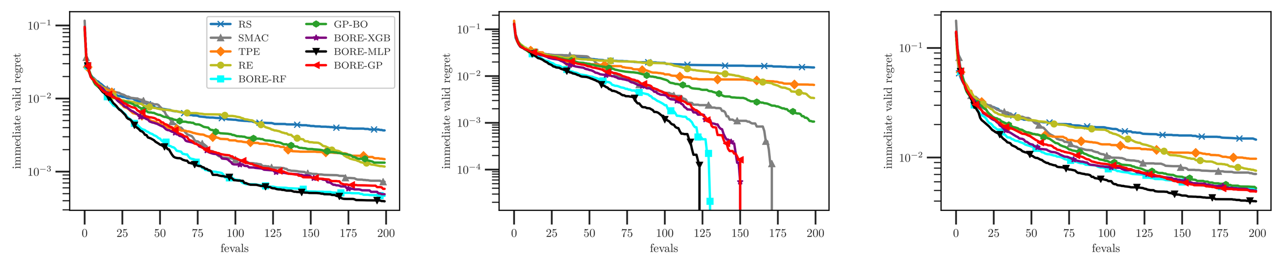
```
while under budget do
    θ* ← arg min_θ L(θ)          // update classifier by optimizing parameters θ wrt binary cross-entropy (BCE) loss
    x_N ← arg max_{x∈X} π_θ*(x)   // suggest new candidate by optimizing input x wrt classifier output
    y_N ← f(x_N)                 // obtain y_N by evaluating blackbox function at x_N
    D_N ← D_{N−1} ∪ {(x_N, y_N)}  // update dataset
end
```

Experimental Results



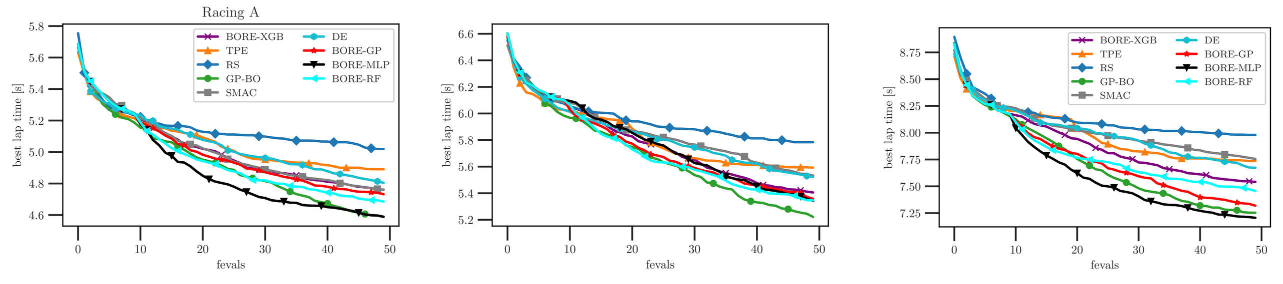
(a) PROTEIN (b) NAVAL (c) PARKINSONS (d) SLICE

Figure: Results on the HPOBench **neural network tuning** problems ($D = 9$).



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet-16

Figure: Results on the NASBench201 **neural architecture search** problems ($D = 6$).



(a) UC BERKELEY ($D = 12$) (b) ETH ZÜRICH I ($D = 20$) (c) ETH ZÜRICH II ($D = 21$)

Figure: Results on the **racing line optimization** problems.

References

[1] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pages 2546–2554, 2011.

[2] M. Sugiyama, T. Suzuki, T. Kanamori. *Density Ratio Estimation in Machine Learning.* Cambridge University Press, 2012.

[3] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems*, pages 594–602, 2011.