BORE: Bayesian Optimization by Density Ratio Estimation

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

[1]University of Sydney, [2]Amazon Web Services, [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 scalability, increased flexibility, and greater representational capacity.

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)

- Let **threshold** $\tau = \Phi^{-1}(\gamma)$ where constant γ denotes some quantile of the observed y values, i.e. $\gamma = \Phi(\tau) = p(y < \tau \,|\, \mathcal{D}_N)$.
- The **improvement** utility function quantifies the improvement over τ

$$I_\gamma(\mathbf{x}) = \max(\tau - y, 0).$$

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

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

- Under *Gaussian* $p(y \,|\, \mathbf{x}, \mathcal{D}_N)$ we get closed-form expression for $\alpha_\gamma(\mathbf{x}; \mathcal{D}_N)$
- But this also 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* $\alpha_\gamma(\mathbf{x}; \mathcal{D}_N)$.

- So why not just find an alternative formulation of $\alpha_\gamma(\mathbf{x}; \mathcal{D}_N)$ that doesn't depend 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 $0 \leq \gamma < 1$ [4].

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

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

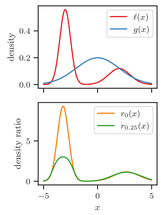


Figure: Example densities

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 its variants, such as the tree-structured Parzen estimator (TPE) [1].
- This approach has major flaws, and has long since been superseded by *direct* DRE methods such as CPE, KMM, KLIEP, ULSIF, RULSIF, etc [3].
- 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*.

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

EI vs. Density Ratio

- Define τ as a function of γ as before. Let $\ell(\mathbf{x})$ and $g(\mathbf{x})$ be distributions such that $\mathbf{x} \sim \ell(\mathbf{x})$ if $y < \tau$, and $\mathbf{x} \sim g(\mathbf{x})$ if $y \geq \tau$.
- Define conditional $p(\mathbf{x} \,|\, y, \mathcal{D}_N)$ in terms of $\ell(\mathbf{x})$ and $g(\mathbf{x})$

$$p(\mathbf{x} \,|\, y, \mathcal{D}_N) = \begin{cases} \ell(\mathbf{x}) & \text{if } y < \tau \\ g(\mathbf{x}) & \text{if } y \geq \tau \end{cases}$$

- Remarkably, it can be shown EI can be expressed as the γ-relative density ratio, up to some constant factor [1]

$$\boxed{\alpha_\gamma(\mathbf{x}; \mathcal{D}_N) \propto r_\gamma(\mathbf{x})}$$

- See example in Figure 2 below with $\gamma = 1/3$

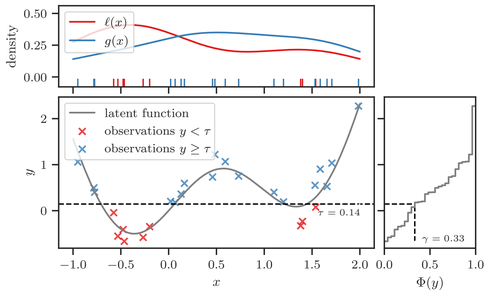


Figure: Synthetic test function $f(x) = \sin(3x) + x^2 - 0.7x$ with observation noise $\varepsilon \sim \mathcal{N}(0, 0.2^2)$.

Density Ratio vs. Class-posterior Probability

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

$$z = \begin{cases} 1 & \text{if } y < \tau \\ 0 & \text{if } y \geq \tau \end{cases}$$

- Let the **class-posterior probability** be denoted 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

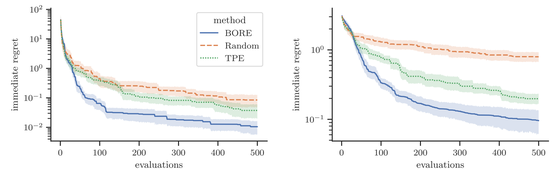
- Estimate $\pi(\mathbf{x})$ by training a probabilistic classifier $\pi_{\boldsymbol{\theta}}(\mathbf{x})$ parameterized by $\boldsymbol{\theta}$
- An obvious candidate—a *feed-forward neural network*
  - expressive (universal approximation), flexible, composable
  - easily scalable with stochastic optimization
  - differentiable end-to-end wrt inputs $\mathbf{x}$
- The so-called BO loop is summarized in Algorithm 1 below.

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

while *under budget* **do**
- $\boldsymbol{\theta}^* \leftarrow \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ // update classifier by optimizing parameters $\boldsymbol{\theta}$ wrt binary cross-entropy (BCE) loss
- $\mathbf{x}_N \leftarrow \arg\max_{\mathbf{x} \in \mathcal{X}} \pi_{\boldsymbol{\theta}^*}(\mathbf{x})$ // suggest new candidate by optimizing input x wrt classifier output
- $y_N \leftarrow f(\mathbf{x}_N)$ // obtain y_N by evaluating blackbox function at $\mathbf{x}_N$
- $\mathcal{D}_N \leftarrow \mathcal{D}_{N-1} \cup \{(\mathbf{x}_N, y_N)\}$ // update dataset

end

Results – Synthetic Test Benchmarks



(a) BRANIN ($D = 2$) (b) HARTMANN6D ($D = 6$)

Figure: Immediate regret over function evaluations on synthetic test benchmarks

Results – Meta-surrogate Benchmarks

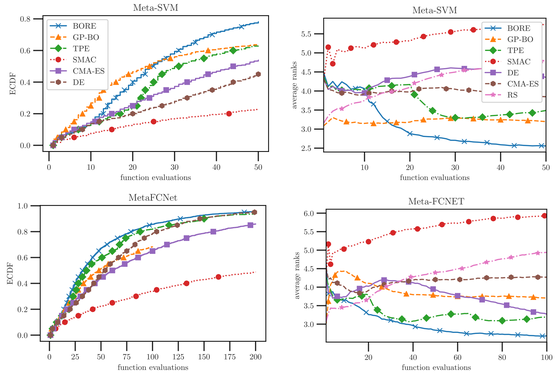


Figure: Empirical cdfs (ECDFS) (left) and ranks (right) across the SVM and FCNET problem classes of *Profet* [2]

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] A. Klein, Z. Dai, F. Hutter, N. Lawrence, and J. Gonzalez.
Meta-surrogate benchmarking for hyperparameter optimization.
In *Advances in Neural Information Processing Systems*, pages 6270–6280, 2019.

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

[4] 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.