Likelihood-free Density Ratio Acquisition Functions are not Equivalent to Expected Improvements

Jiaming Song
Department of Computer Science
Stanford University
tsong@cs.stanford.edu

Stefano Ermon
Department of Computer Science
Stanford University
ermon@cs.stanford.edu

Abstract

Bayesian Optimization (BO) is one of the most effective black-box optimization methods, yet the need to ensure analytical tractability in the posterior predictive makes it challenging to apply BO to large-scale problems with high-dimensional observations. For these problems, likelihood-free methods present a promising avenue since they can work with more expressive models and are often more efficient. Previous papers (Bergstra et al., 2011; Tiao et al., 2021) have claimed that density ratios acquired from the likelihood-free inference are equivalent to the widely popular expected improvement acquisition function, allowing us to perform BO without expensive exact posterior inference. Unfortunately, we show in this paper that the claim is false; we identify errors in their reasoning and illustrate a counter-example where density ratios are inversely correlated to expected improvements. Our results suggest that additional care is needed when interpreting and applying density ratio acquisition functions from likelihood-free inference.

1 Introduction

Bayesian Optimization (BO) is a popular class of global optimization methods over a design space aimed to use as few queries as possible (Mockus et al., 1978). Crucial to BO is the acquisition function, *i.e.*, a function that selects where to query next given current observations. Various acquisition functions have been proposed, but they are often difficult to compute, let alone optimize in high-dimension inputs (Wilson et al., 2018). Even ones that are relatively easy to compute, such as expected improvement (EI, (Mockus et al., 1978)), may require expensive posterior inference via Gaussian processes (Rasmussen, 2003). In contrast, approximate inference methods can work with more expressive models and are often more efficient (Rezende et al., 2014); integrating effective acquisition functions with approximate inference could scale to problems with large amounts of high-dimensional observations, at the cost of analytical tractability (Cranmer et al., 2020).

For example, one can use density ratios acquired from likelihood-free inference (Mohamed & Lakshminarayanan, 2016; Gutmann et al., 2016) as acquisition functions (Bergstra et al., 2011; Tiao et al., 2021). Intuitively, if we learn a probabilistic classifier to separate samples with high fitness inputs apart from the others, then this classifier should tell us if a new input has high fitness or not. If the classifier is Bayes-optimal, then it can recover density ratios (Sugiyama et al., 2012). In one of the first papers that performs Bayesian optimization for neural network hyperparameter optimization, Bergstra et al. (2011) claimed that these density ratios are proportional to expected improvement, so one can bypass expensive Gaussian process entirely and use density ratios as acquisition functions.

Unfortunately, we show in this paper that the above claim is false. First, we identify an error in their reasoning. Next, we illustrate a counter-example supporting our claims; not only are density ratios not proportional to expected improvements, but they can be inversely correlated to them. Finally, we discuss the ramifications of the result.

Bayesian Deep Learning workshop, NeurIPS 2021.

2 Findings

Background Given a black-box function $f : \mathcal{X} \to \mathbb{R}$, Bayesian Optimization (BO) aims to find an input $\mathbf{x}$ where $f(\mathbf{x})$ is maximized, given n observations $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Denote $p(y|\mathbf{x}, \mathcal{D}_n)$ as the posterior predictive distribution (usually a Gaussian process), and the expected improvement function for a point $\mathbf{x}$ is as follows:

$$\mathrm{EI}(\mathbf{x}, \tau) = \mathbb{E}_{y \sim p(y|\mathbf{x}, \mathcal{D}_n)}[\max(y - \tau, 0)] \tag{1}$$

where τ is a threshold (often set as the currently observed maximum value). Following notations in Tiao et al. (2021), let τ be the γ-th percentile of observed y values (*i.e.*, $\gamma = p(y \leq \tau)$), and let two densities $\ell(\mathbf{x}) := p(\mathbf{x}|y \leq \tau; \mathcal{D}_n)$ and $g(\mathbf{x}) := p(\mathbf{x}|y > \tau; \mathcal{D}_n)$ be the densities for $\mathbf{x}$ conditioned on y being less or greater than the threshold, respectively. The following claim restates the results in Bergstra et al. (2011); Tiao et al. (2021), which shows an equivalence between density (or likelihood) ratio acquisition functions and expected improvements[1]:

Claim 1 ((Bergstra et al., 2011; Tiao et al., 2021)). *Define τ, γ, $\ell(\mathbf{x})$, $g(\mathbf{x})$ as above. Then:*

$$\mathrm{EI}(\mathbf{x}, \tau) \propto \frac{g(\mathbf{x})}{\gamma \ell(\mathbf{x}) + (1 - \gamma) g(\mathbf{x})}, \tag{2}$$

in other words, EI is proportional to the ratio between $g(\mathbf{x})$ and $\gamma \ell(\mathbf{x}) + (1 - \gamma) g(\mathbf{x})$.

The ratio can then be estimated with a Bayes optimal classifier that learns from binary class labels that assigns label 1 to points with $y > \tau$ and 0 to points with $y \leq \tau$ (Grover et al., 2019).

Counter claim We will now go through the "proof" of the claim and identify where the argument does not hold. First we have that:

$$\mathrm{EI}(\mathbf{x}, \tau) = \mathbb{E}_{p(y|\mathbf{x}, D_n)}[\max(y - \tau, 0)] = \int_\tau^\infty \max(y - \tau, 0) p(y|\mathbf{x}, D_n) \mathrm{d}y \tag{3}$$

$$= \frac{\int_\tau^\infty \max(y - \tau, 0) p(y|D_n) p(\mathbf{x}|y, D_n) \mathrm{d}y}{p(\mathbf{x}|D_n)}, \tag{4}$$

where we applied Bayes' rule in Eq (4). Then, the denominator evaluates as:

$$p(\mathbf{x}|D_n) = p(\mathbf{x}|y \leq \tau, D_n) p(y \leq \tau|D_n) + p(\mathbf{x}|y > \tau, D_n) p(y > \tau|D_n) \tag{5}$$

$$= \ell(\mathbf{x}) p(y \leq \tau|D_n) + g(\mathbf{x}) p(y > \tau|D_n) = \gamma \ell(\mathbf{x}) + (1 - \gamma) g(\mathbf{x}). \tag{6}$$

where we use the definition that $\gamma := p(y \leq \tau|D_n)$. Finally, Bergstra et al. (2011); Tiao et al. (2021) claims that the numerator evaluates to:

$$\int_\tau^\infty \max(y - \tau) p(y|D_n) p(\mathbf{x}|y, D_n) \mathrm{d}y = g(\mathbf{x}) \int_\tau^\infty \max(y - \tau) p(y|D_n) \mathrm{d}y \tag{7}$$

which is $g(\mathbf{x})$ times a value independent of $\mathbf{x}$. Dividing Eq. (7) with Eq. (6) and using Eq. (4) should recover the result. However, in Eq. (7), $g(\mathbf{x}) := p(\mathbf{x}|y > \tau)$ is directly taken out of the integral, as Bergstra et al. (2011) assumed it is independent of y as long as $y > \tau$. From the definition of conditional probability:

$$g(\mathbf{x}) := p(\mathbf{x}|y > \tau, D_n) = \frac{\int_\tau^\infty p(\mathbf{x}, y|D_n) \mathrm{d}y}{\int_\tau^\infty p(y|D_n) \mathrm{d}y} \neq p(\mathbf{x}|y, D_n), \tag{8}$$

so Eq. (7) does not hold immediately. Thus, we have illustrated that the above density ratio may not be proportional to expected improvement.

3 Empirical Evaluations

In this section, we raise a counter-example where density ratio is mostly inversely proportional to expected improvement. Consider the following joint distribution as our posterior:

$$p(x) = \mathrm{Uniform}[-1, 1], \quad p(y|x) = \mathcal{N}(-x^2, (0.01 + 10x^2)^2). \tag{9}$$

[1] We note that Bergstra et al. (2011); Tiao et al. (2021) minimizes the BO objective in their formulation, whereas we paraphrase the claim to adapt to maximization.

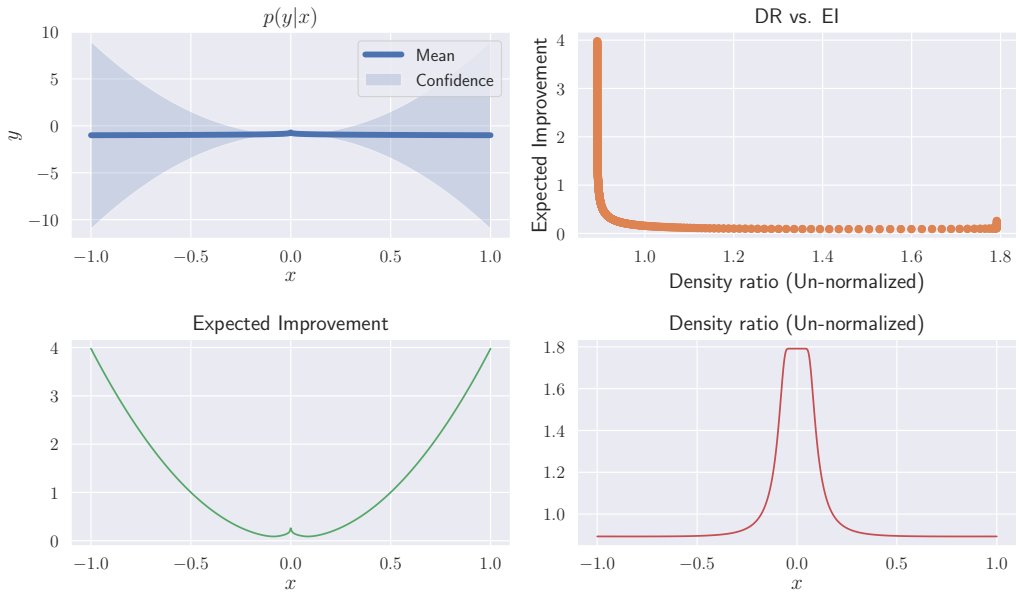

Figure 1: *Top left*: the conditional distribution $p(y|x)$, while the the posterior mean is maximized at $x = 0$, it has large variance when x is large absolute values. *Top right*: the un-normalized density ratio (DR) versus expected improvement (EI), computed analytically. *Bottom*: exact values of EI (*left*) and DR (*right*) as a function of x.

The intuition behind such a design is to make predictive with sub-optimal mean exhibit large variance; for a certain threshold, a lower sub-optimal mean indicates lower probability of improvement but does not indicate lower expected improvement due to the large variances. Let us take $\tau = -0.5^2$, which is the median of the the posterior mean (any threshold between the extreme values should work in principle). In the density ratio, the denominator is simply $p(x)$, and the numerator is $p(x|y > \tau) \propto p(y > \tau|\mathbf{x})p(x)$, and we can compute it analytically; so is the EI function.

Figure 1 shows $p(y|x)$, the corresponding expected improvement, un-normalized density ratios, and a comparison of the two. Here, the points with high EI is at $x = \pm 4$, whereas the points with high density ratios are at $x \approx 0$. Clearly, a higher density ratio does not indicate a higher expected improvement. In this example, they are mostly inversely correlated.

Admittedly, the example is an extreme case that is unlikely to happen in realistic scenarios (*e.g.*, noise level depends on input); but since the equivalence claim should be agnostic to the posterior predictive and the selected threshold, our example suffices as a counter-example to the claim.

4 Discussions

In this paper, we revisit the claim about the equivalence between density ratio (DR) acquisition functions and expected improvements (EI). We identify a potential error in the proof of the claim and illustrate an example where DR and EI are mostly inversely correlated.

Meanwhile, the fact that these two are not equivalent does not imply density ratio acquisition functions cannot be useful in practice. Actually, Tiao et al. (2021) show that they can be quite successful empirically. In our example, density ratio acquisition functions selects around $x = 0$, the point with maximum posterior mean, showing some promise as an acquisition function as well. Our results merely suggest that extra care is needed when interpreting and applying acquisition functions with likelihood-free inference. It would be interesting to investigate when density ratio acquisition function positively correlate with expected improvement and how we can apply likelihood-free inference to general Bayesian optimal experimental design problems.

References

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Aditya Grover, Jiaming Song, Alekh Agarwal, Kenneth Tran, Ashish Kapoor, Eric Horvitz, and Stefano Ermon. Bias correction of learned generative models using Likelihood-Free importance weighting. *arXiv preprint arXiv:1906.09531*, June 2019.

Michael U Gutmann, Jukka Corander, et al. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 2016.

Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2, 1978.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, October 2016.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pp. 63–71. Springer, 2003.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Louis C Tiao, Aaron Klein, Matthias Seeger, Edwin V Bonilla, Cedric Archambeau, and Fabio Ramos. Bore: Bayesian optimization by density-ratio estimation. *arXiv preprint arXiv:2102.09009*, 2021.

James T Wilson, Frank Hutter, and Marc Peter Deisenroth. Maximizing acquisition functions for bayesian optimization. *arXiv preprint arXiv:1805.10196*, 2018.