
Supplementary Material
Bayesian Optimization by Density Ratio Estimation

Anonymous Author(s)

Affiliation

Address

email

1 A Expected improvement

2 For completeness, we reproduce the derivations of [1]. Recall that expected improvement (EI) is
3 defined as the expectation of the improvement utility function $I_\gamma(\mathbf{x})$ under the posterior predictive
4 distribution $p(y | \mathbf{x}, \mathcal{D}_N)$. Expanding this out, we have

$$\alpha_\gamma(\mathbf{x}; \mathcal{D}_N) = \mathbb{E}_{p(y | \mathbf{x}, \mathcal{D}_N)}[I_\gamma(\mathbf{x})] = \int_{-\infty}^{\infty} I_\gamma(\mathbf{x}) p(y | \mathbf{x}, \mathcal{D}_N) dy \quad (1)$$

$$= \int_{-\infty}^{\tau} (\tau - y) p(y | \mathbf{x}, \mathcal{D}_N) dy \quad (2)$$

$$= \frac{1}{p(\mathbf{x} | \mathcal{D}_N)} \int_{-\infty}^{\tau} (\tau - y) p(\mathbf{x} | y, \mathcal{D}_N) p(y | \mathcal{D}_N) dy. \quad (3)$$

5 Note we have used Bayes' rule in the last step above. Next, the denominator evaluates to

$$p(\mathbf{x} | \mathcal{D}_N) = \int_{-\infty}^{\infty} p(\mathbf{x} | y, \mathcal{D}_N) p(y | \mathcal{D}_N) dy \quad (4)$$

$$= \ell(\mathbf{x}) \int_{-\infty}^{\tau} p(y | \mathcal{D}_N) dy + g(\mathbf{x}) \int_{\tau}^{\infty} p(y | \mathcal{D}_N) dy \quad (5)$$

$$= \gamma \ell(\mathbf{x}) + (1 - \gamma) g(\mathbf{x}), \quad (6)$$

6 since $\gamma = \Phi(\tau) = p(y < \tau | \mathcal{D}_N)$, by definition. Finally, we evaluate the numerator,

$$\int_{-\infty}^{\tau} (\tau - y) p(\mathbf{x} | y, \mathcal{D}_N) p(y | \mathcal{D}_N) dy = \ell(\mathbf{x}) \int_{-\infty}^{\tau} (\tau - y) p(y | \mathcal{D}_N) dy \quad (7)$$

$$= \ell(\mathbf{x}) \tau \int_{-\infty}^{\tau} p(y | \mathcal{D}_N) dy - \ell(\mathbf{x}) \int_{-\infty}^{\tau} y p(y | \mathcal{D}_N) dy \quad (8)$$

$$= \gamma \tau \ell(\mathbf{x}) - \ell(\mathbf{x}) \int_{-\infty}^{\tau} y p(y | \mathcal{D}_N) dy \quad (9)$$

$$= K \cdot \ell(\mathbf{x}), \quad (10)$$

7 where

$$K = \gamma \tau - \int_{-\infty}^{\tau} y p(y | \mathcal{D}_N) dy. \quad (11)$$

8 Hence, this shows that the EI function is equivalent to the γ -relative density ratio [3] up to a constant
9 factor K ,

$$\alpha_\gamma(\mathbf{x}; \mathcal{D}_N) \propto \frac{\ell(\mathbf{x})}{\gamma\ell(\mathbf{x}) + (1 - \gamma)g(\mathbf{x})} \quad (12)$$

$$= \left(\gamma + \frac{g(\mathbf{x})}{\ell(\mathbf{x})}(1 - \gamma) \right)^{-1}. \quad (13)$$

10 B Class-posterior probability

11 We provide an unabridged derivation of the identity in [eq. 13](#) of the main paper. First, the ordinary
12 density ratio is given by

$$r_0(\mathbf{x}) = \frac{\ell(\mathbf{x})}{g(\mathbf{x})} = \frac{p(\mathbf{x} | z = 1)}{p(\mathbf{x} | z = 0)} \quad (14)$$

$$= \left(\frac{p(z = 1 | \mathbf{x})p(\mathbf{x})}{p(z = 1)} \right) \left(\frac{p(z = 0 | \mathbf{x})p(\mathbf{x})}{p(z = 0)} \right)^{-1} \quad (15)$$

$$= \frac{p(z = 0)}{p(z = 1)} \cdot \frac{p(z = 1 | \mathbf{x})}{p(z = 0 | \mathbf{x})}. \quad (16)$$

13 By construction, we have,

$$\frac{p(z = 0)}{p(z = 1)} = \frac{p(y \geq \tau)}{p(y < \tau)} = \frac{1 - \gamma}{\gamma} = \left(\frac{\gamma}{1 - \gamma} \right)^{-1}. \quad (17)$$

14 Furthermore,

$$\frac{p(z = 1 | \mathbf{x})}{p(z = 0 | \mathbf{x})} = \frac{p(z = 1 | \mathbf{x})}{1 - p(z = 1 | \mathbf{x})} = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}. \quad (18)$$

15 Therefore,

$$r_0(\mathbf{x}) = \left(\frac{\gamma}{1 - \gamma} \right)^{-1} \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}. \quad (19)$$

16 Plugging this into the expression of [eq. 2](#) of the main paper, we have

$$r_\gamma(\mathbf{x}) = \left(\gamma + (1 - \gamma) \left(\frac{\gamma}{1 - \gamma} \right) \left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right)^{-1} \right)^{-1} \quad (20)$$

$$= \gamma^{-1} \left(1 + \left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right)^{-1} \right)^{-1} \quad (21)$$

$$= \gamma^{-1} \pi(\mathbf{x}), \quad (22)$$

17 as required.

18 C Binary cross-entropy

19 The binary cross-entropy (BCE) loss is given by

$$\mathcal{L}(\boldsymbol{\theta}) = -\beta \cdot \mathbb{E}_{\ell(\mathbf{x})}[\log \pi_{\boldsymbol{\theta}}(\mathbf{x})] - (1 - \beta) \cdot \mathbb{E}_{g(\mathbf{x})}[\log (1 - \pi_{\boldsymbol{\theta}}(\mathbf{x}))], \quad (23)$$

20 where β denotes the class balance rate. That is, let N_ℓ and N_g be the sizes of the support of $\ell(\mathbf{x})$ and
21 $g(\mathbf{x})$, respectively. Then, we have

$$\beta = \frac{N_\ell}{N}, \quad \text{and} \quad 1 - \beta = \frac{N_g}{N}, \quad (24)$$

22 where $N = N_\ell + N_g$.

23 **C.1 Empirical risk minimization**

24 We show that the BCE loss can be approximated by the empirical risk,

$$\mathcal{L}(\boldsymbol{\theta}) \simeq -\frac{1}{N} \left(\sum_{n=1}^N z_n \log \pi_{\boldsymbol{\theta}}(\mathbf{x}_n) + (1 - z_n) \log (1 - \pi_{\boldsymbol{\theta}}(\mathbf{x}_n)) \right). \quad (25)$$

25 Let ρ be the permutation of the set $\{1, \dots, N\}$, i.e. the bijection from $\{1, \dots, N\}$ to itself, such that
26 $y_{\rho(n)} < \tau$ if $0 < \rho(n) \leq N_{\ell}$, and $y_{\rho(n)} \geq \tau$ if $N_{\ell} < \rho(n) \leq N_g$. That is to say,

$$\mathbf{x}_{\rho(n)} \sim \begin{cases} \ell(\mathbf{x}) & \text{if } 0 < \rho(n) \leq N_{\ell}, \\ g(\mathbf{x}) & \text{if } N_{\ell} < \rho(n) \leq N_g. \end{cases} \quad \text{and} \quad z_{\rho(n)} = \begin{cases} 1 & \text{if } 0 < \rho(n) \leq N_{\ell}, \\ 0 & \text{if } N_{\ell} < \rho(n) \leq N_g. \end{cases} \quad (26)$$

27 Then, we have

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{N} (N_{\ell} \cdot \mathbb{E}_{\ell(\mathbf{x})}[\log \pi_{\boldsymbol{\theta}}(\mathbf{x})] + N_g \cdot \mathbb{E}_{g(\mathbf{x})}[\log (1 - \pi_{\boldsymbol{\theta}}(\mathbf{x}))]) \quad (27)$$

$$\simeq -\frac{1}{N} \left(\mathcal{N}_{\ell} \cdot \frac{1}{\mathcal{N}_{\ell}} \sum_{n=1}^{N_{\ell}} \log \pi_{\boldsymbol{\theta}}(\mathbf{x}_{\rho(n)}) + \mathcal{N}_g \cdot \frac{1}{\mathcal{N}_g} \sum_{n=N_{\ell}+1}^{N_g} \log (1 - \pi_{\boldsymbol{\theta}}(\mathbf{x}_{\rho(n)})) \right) \quad (28)$$

$$= -\frac{1}{N} \left(\sum_{n=1}^N z_{\rho(n)} \log \pi_{\boldsymbol{\theta}}(\mathbf{x}_{\rho(n)}) + (1 - z_{\rho(n)}) \log (1 - \pi_{\boldsymbol{\theta}}(\mathbf{x}_{\rho(n)})) \right) \quad (29)$$

$$= -\frac{1}{N} \left(\sum_{n=1}^N z_n \log \pi_{\boldsymbol{\theta}}(\mathbf{x}_n) + (1 - z_n) \log (1 - \pi_{\boldsymbol{\theta}}(\mathbf{x}_n)) \right). \quad (30)$$

28 **C.2 Optimum**

29 We show the identity of [eq. 15](#) of the main paper. Taking the functional derivative and setting it to
30 zero, we get

$$0 = \frac{\partial \mathcal{L}}{\partial \pi_{\boldsymbol{\theta}}} = -\mathbb{E}_{\ell(\mathbf{x})} \left[\frac{\beta}{\pi_{\boldsymbol{\theta}}(\mathbf{x})} \right] + \mathbb{E}_{g(\mathbf{x})} \left[\frac{1-\beta}{1-\pi_{\boldsymbol{\theta}}(\mathbf{x})} \right] \quad (31)$$

$$= \int -\frac{\beta \ell(\mathbf{x})}{\pi_{\boldsymbol{\theta}}(\mathbf{x})} + \frac{(1-\beta)g(\mathbf{x})}{1-\pi_{\boldsymbol{\theta}}(\mathbf{x})} d\mathbf{x} \quad (32)$$

31 This integral evaluates to zero iff the integrand itself evaluates to zero. Hence, we solve the following
32 for $\pi_{\boldsymbol{\theta}}(\mathbf{x})$,

$$\frac{\beta \ell(\mathbf{x})}{\pi_{\boldsymbol{\theta}}(\mathbf{x})} = \frac{(1-\beta)g(\mathbf{x})}{1-\pi_{\boldsymbol{\theta}}(\mathbf{x})}. \quad (33)$$

33 We re-arrange this expression to give

$$\frac{1-\pi_{\boldsymbol{\theta}}(\mathbf{x})}{\pi_{\boldsymbol{\theta}}(\mathbf{x})} = \left(\frac{1-\beta}{\beta} \right) \frac{g(\mathbf{x})}{\ell(\mathbf{x})} \Leftrightarrow \frac{1}{\pi_{\boldsymbol{\theta}}(\mathbf{x})} - 1 = \frac{\beta \ell(\mathbf{x}) + (1-\beta)g(\mathbf{x})}{\beta \ell(\mathbf{x})} - 1. \quad (34)$$

34 Finally, we add one to both sides and invert the result to give

$$\pi_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{\beta \ell(\mathbf{x})}{\beta \ell(\mathbf{x}) + (1-\beta)g(\mathbf{x})}, \quad (35)$$

35 as required.

36 **D Toy example**

37 Consider the following toy example where the densities $\ell(x)$ and $g(x)$ are *known* and given exactly
38 by the following (mixture of) Gaussians,

$$\ell(x) = 0.3\mathcal{N}(2, 1^2) + 0.7\mathcal{N}(-3, 0.5^2), \quad \text{and} \quad g(x) = \mathcal{N}(0, 2^2), \quad (36)$$

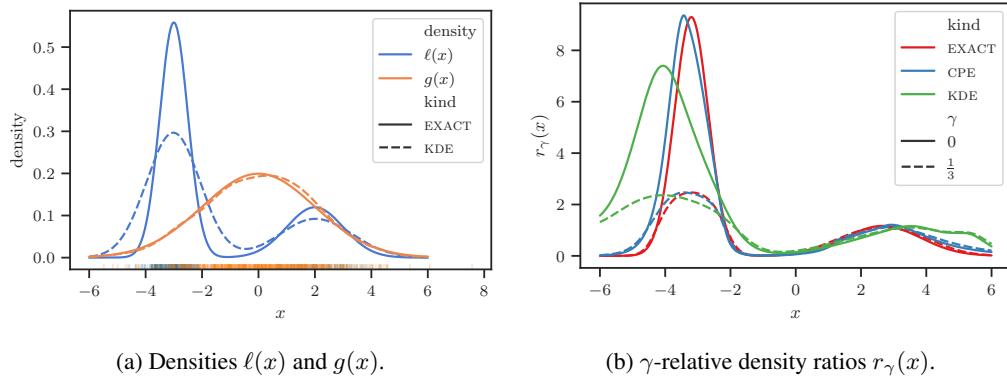


Figure 1: Synthetic toy example with (mixtures of) Gaussians.

as illustrated by the *solid blue* and *orange* lines in Figure 1a, respectively. We draw a total of $N = 1000$ samples from these distributions, with a fraction $\gamma = 1/3$ drawn from $\ell(x)$ and the remainder from $g(x)$. These are represented by the vertical markers along the bottom of the x -axis (a so-called “rug plot”). Then, two kernel density estimations (KDEs), shown with *dashed* lines, are fit on these respective sample sets, with kernel bandwidths selected according to the “normal reference” rule-of-thumb. We see that, for both densities, the modes are recovered well, while for $\ell(x)$, the variances are overestimated in both of its mixture components. As we shall see, this has deleterious effects on the resulting density ratio estimate.

In Figure 1b, we show the ordinary and γ -relative density ratios with the *solid* and *dashed red* lines, respectively. The density ratio estimates resulting from taking the ratio of the KDEs are shown in *green*. Those resulting from the class-probability estimation (CPE) method described in § 3 of the main paper are shown in *blue*. The probabilistic classifier consists of a simple multi-layer perceptron (MLP) with 3 hidden layers, each with 32 units and elu activations.

The CPE method appears, at least visually, to recover the exact density ratios well, while the KDE method does so quite poorly. Perhaps the more important quality to focus on, for the purposes of Bayesian optimization (BO), is the global maximum of the density ratio functions. In the case of the KDE method, we can see that this deviates significantly from that of the true density ratios. In this instance, even though KDE fit $g(x)$ well and recovered the mode of $\ell(x)$ accurately, a slight overestimation of the variance in the latter led to a dramatic shift in the maximum of the resulting density ratio functions.

59 E Implementation details

60 Software. Our method is implemented as a *configuration generator* plugin for the [HpBandSter](#) library of Falkner et al. [2].

62 Epochs per iteration. To ensure the training time on BO iteration N is nonincreasing as a function of N , instead of directly specifying the number of epochs (i.e. full passes over the data), we specify the number of (batchwise gradient) steps S to train for in each iteration. Since the number of steps per epoch is $M = \lceil N/B \rceil$, the effective number of epochs on the N -th BO iteration is then $E = \lfloor S/M \rfloor$. For example, if $S = 1024$ and $B = 64$, the number of epochs for iteration $N = 512$ would be $E = 128$. As another example, for all $0 < N \leq B$, we have $E = S = 1024$. We set $S = 100$.

68 F Qualitative analysis

69 In Figures 2 and 3, we show a scatterplot of the locations suggested by tree-structured Parzen estimator
70 (TPE) and BO, respectively, across 20 runs on the SIX-HUMP CAMEL problem.

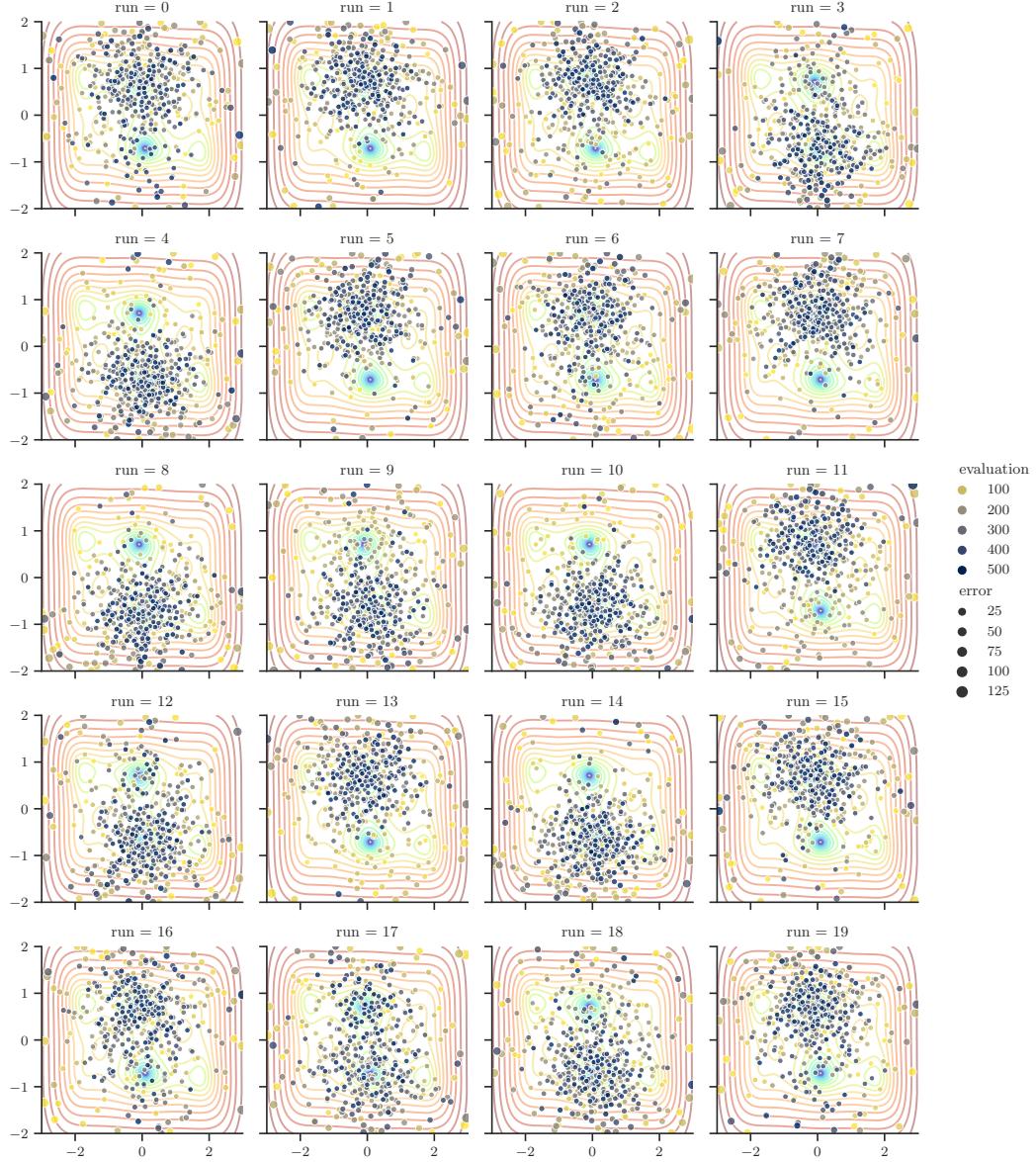


Figure 2: Candidates suggested by TPE on the SIX-HUMP CAMEL problem across 20 runs.

71 References

- 72 [1] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In
73 *Advances in neural information processing systems*, pages 2546–2554, 2011.
- 74 [2] S. Falkner, A. Klein, and F. Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. In
75 *International Conference on Machine Learning*, pages 1437–1446, 2018.
- 76 [3] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for
77 robust distribution comparison. In *Advances in neural information processing systems*, pages 594–602,
78 2011.

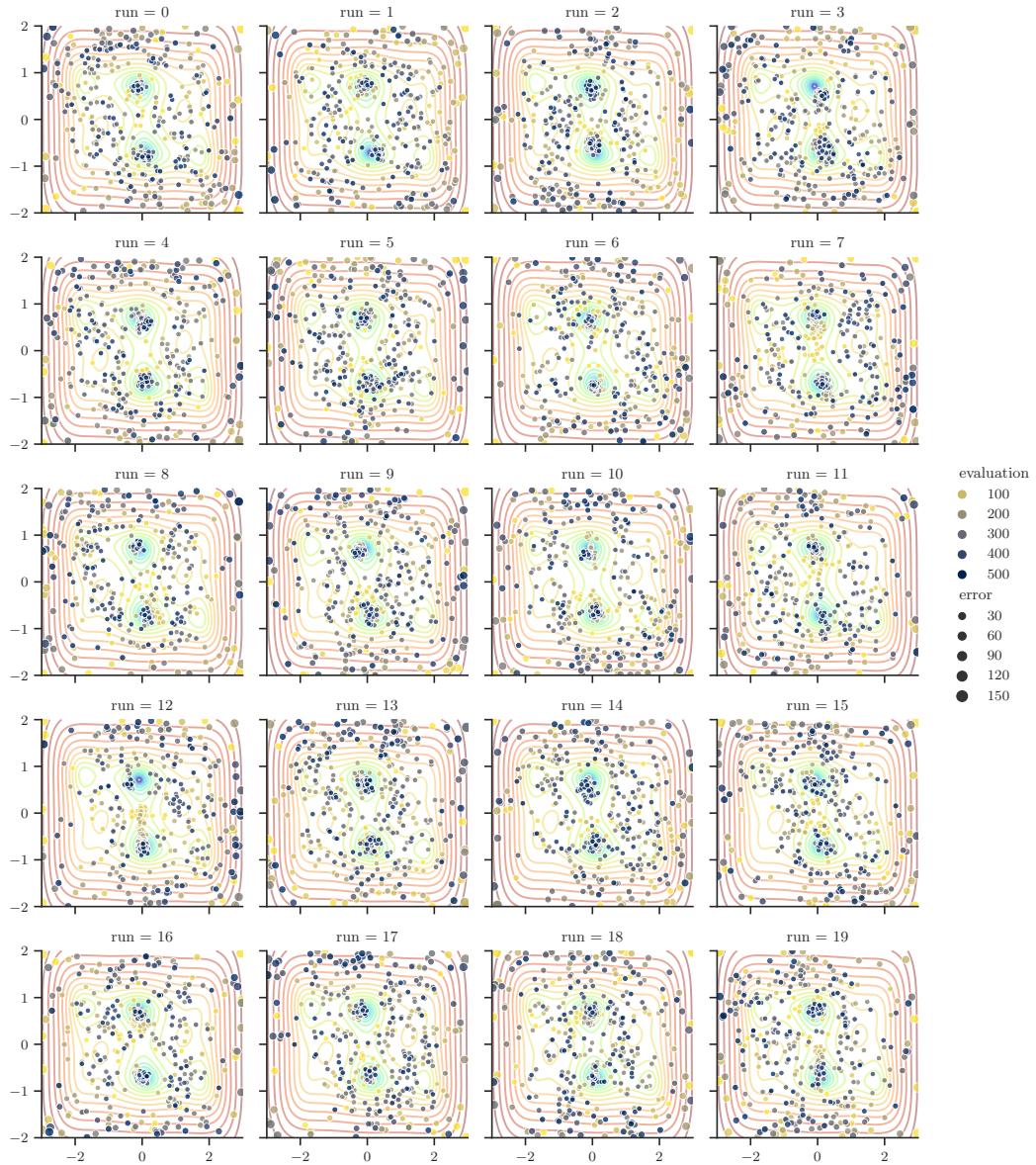


Figure 3: Candidates suggested by BORE on the SIX-HUMP CAMEL problem across 20 runs.