Density Ratio Estimation-based Bayesian Optimization with Semi-Supervised Learning

Jungtaek Kim

University of Wisconsin–Madison

Bayesian Optimization

- Bayesian optimization has attracted immense attention from various research areas:
  - hyperparameter optimization,
  - chemical reaction optimization,
  - language model fine-tuning.
- It is capable of efficiently finding a global optimum of a costly black-box function.
- Generally, a probabilistic regression model such as Gaussian processes is widely used as a surrogate.

Density Ratio Estimation-based Bayesian Optimization

- This line of research utilizes $p(\mathbf{x} \mid y \leq y^\dagger, \mathcal{D})$ and $p(\mathbf{x} \mid y > y^\dagger, \mathcal{D})$, where $y^\dagger$ is a threshold for dividing inputs to two groups that are relatively close and relatively far to a global solution.
- Instead of modeling two densities separately, it allows us to solve Bayesian optimization using binary classification.
- Its acquisition function is defined by the following:

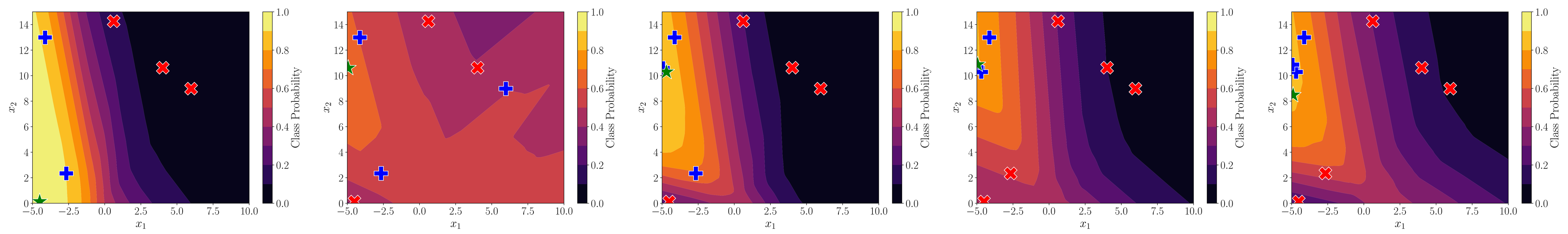
$$A(\mathbf{x} \mid \zeta, \mathcal{D}_t) = \frac{p(\mathbf{x} \mid z=1)}{\zeta p(\mathbf{x} \mid z=1) + (1-\zeta) p(\mathbf{x} \mid z=0)}, \quad (1)$$

  where $\zeta = p(y \leq y^\dagger) \in [0, 1)$ is a threshold ratio.
- Therefore, a class probability over $\mathbf{x}$ for Class 1 is considered as an acquisition function:

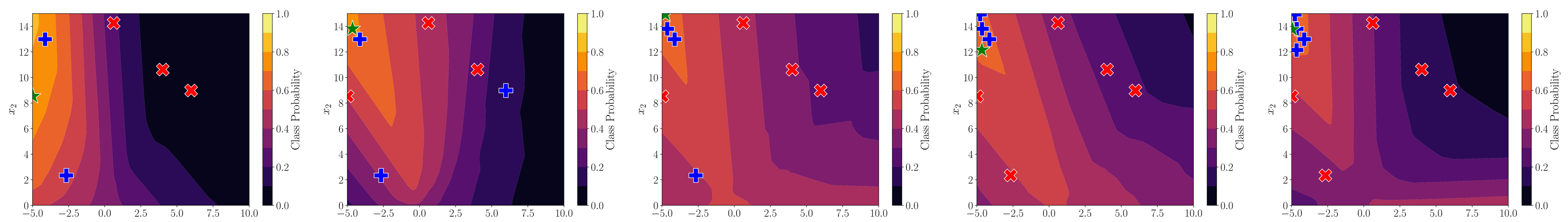
$$A(\mathbf{x} \mid \zeta, \mathcal{D}_t) = \zeta^{-1} \pi(\mathbf{x}). \quad (2)$$

Over-Exploitation Problem

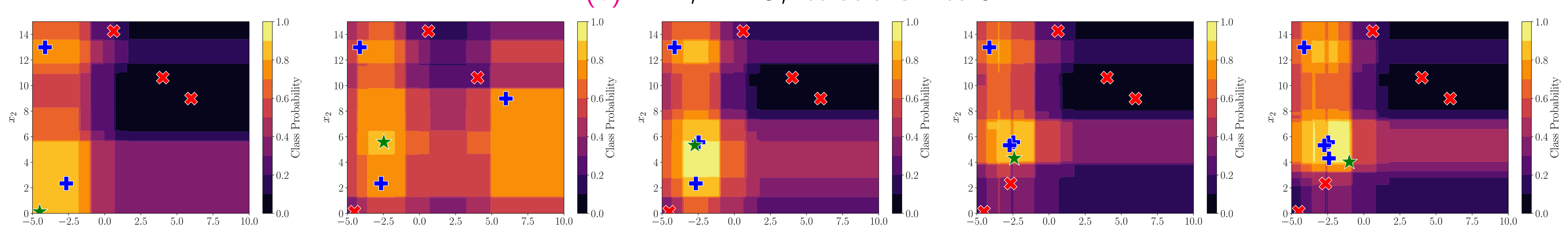
- The supervised classifiers used in density ratio estimation-based Bayesian optimization suffer from the over-exploitation problem.
- It indicates the problem of overconfidence over known knowledge on global solution candidates.
- At early iterations, a supervised classifier tends to overfit to a small size of $\mathcal{D}_t$ due to a relatively large model capacity.
- This consequence makes a Bayesian optimization algorithm highly focus on exploitation.



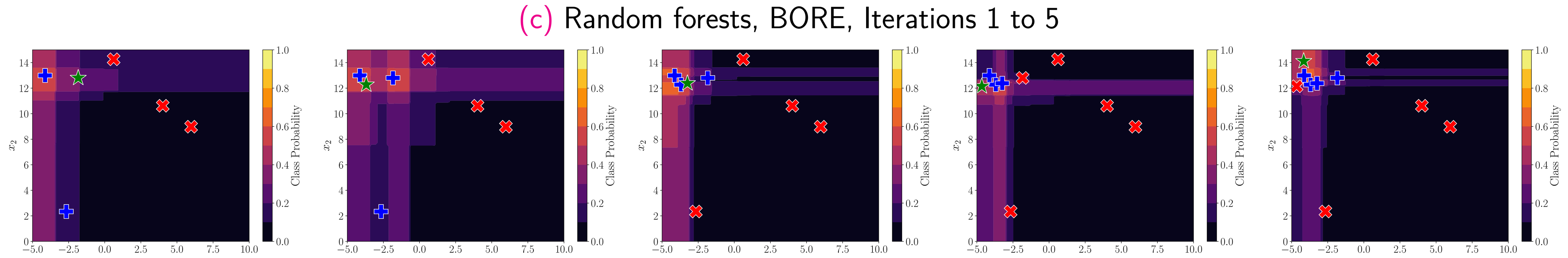
(a) MLP, BORE, Iterations 1 to 5



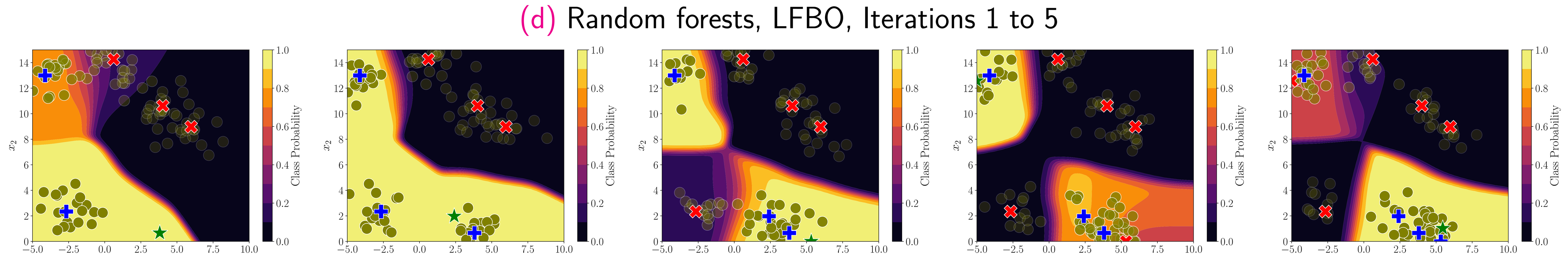
(b) MLP, LFBO, Iterations 1 to 5



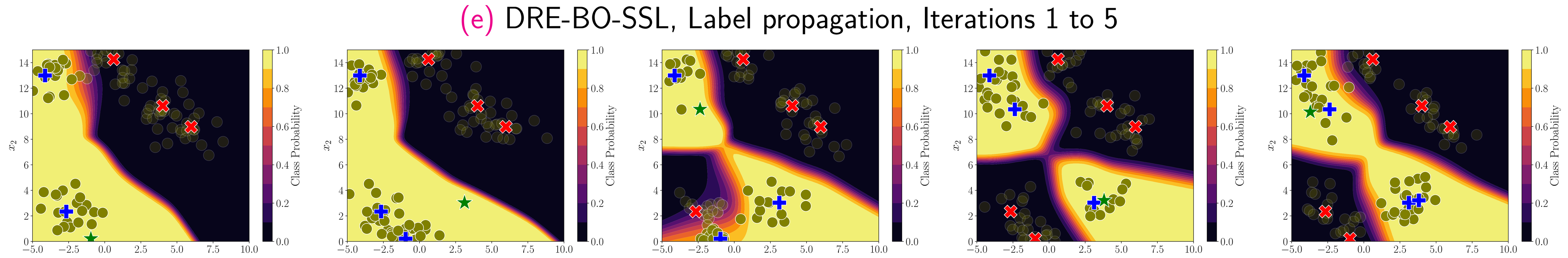
(c) Random forests, BORE, Iterations 1 to 5



(d) Random forests, LFBO, Iterations 1 to 5



(e) DRE-BO-SSL, Label propagation, Iterations 1 to 5



(f) DRE-BO-SSL, Label spreading, Iterations 1 to 5

Figure: Comparisons of BORE, LFBO, and DRE-BO-SSL with label propagation and label spreading for the Branin function.

DRE-BO-SSL

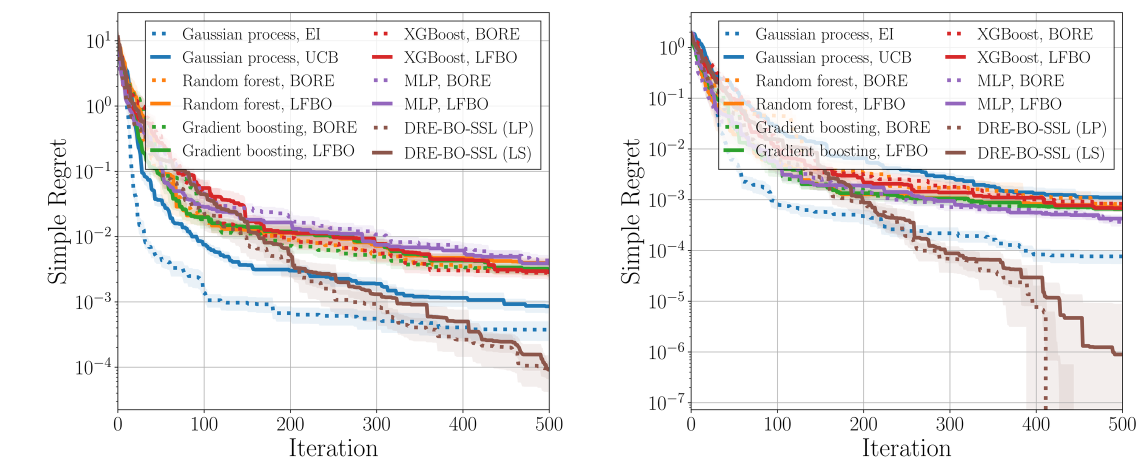
- We introduce DRE-BO-SSL defined with semi-supervised learning.
- Using the pseudo-labels $\widehat{\mathbf{C}}_t$ of a semi-supervised model, it chooses the next query point $\mathbf{x}_{t+1}$:

$$\mathbf{x}_{t+1} = \underset{\mathbf{x} \in \mathcal{X}}{\arg\max}\ \pi_{\widehat{\mathbf{C}}_t}(\mathbf{x}; \zeta, \mathcal{D}_t, \mathbf{X}_u), \quad (3)$$

  where $\pi_{\widehat{\mathbf{C}}_t}(\mathbf{x}; \zeta, \mathcal{D}_t, \mathbf{X}_u)$ predicts a class probability over $\mathbf{x}$ for Class 1.
- We adopt a multi-started local optimization technique, e.g., L-BFGS-B, to solve (3).
- Two semi-supervised learning methods are used:
  - label propagation,
  - label spreading.
- Two scenarios are tackled:
  - a scenario with unlabeled point sampling, which assumes that unlabeled data points are unavailable,
  - a scenario with fixed-size pools, which assumes that the pools are provided as sets of possible query points.

Experimental Results

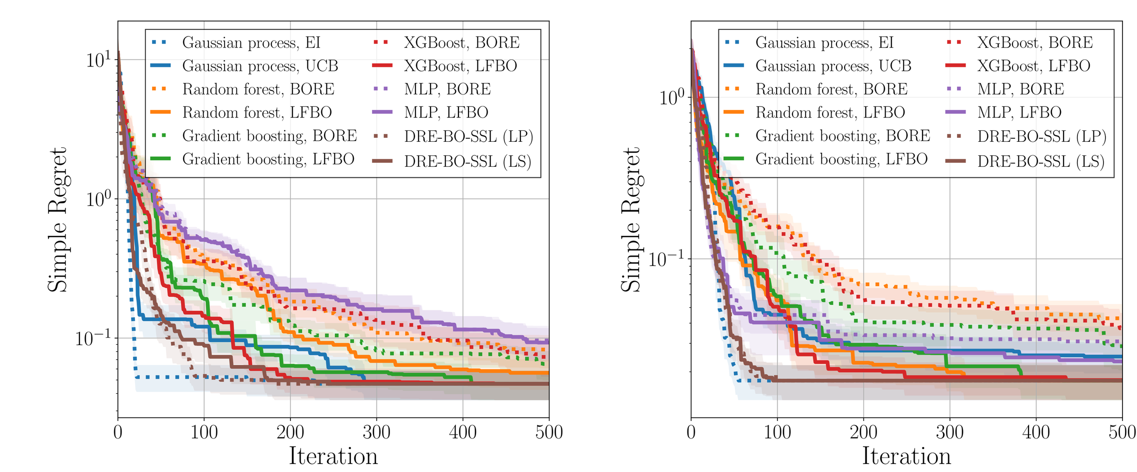
- A scenario with unlabeled point sampling



(a) Branin (b) Six-hump camel

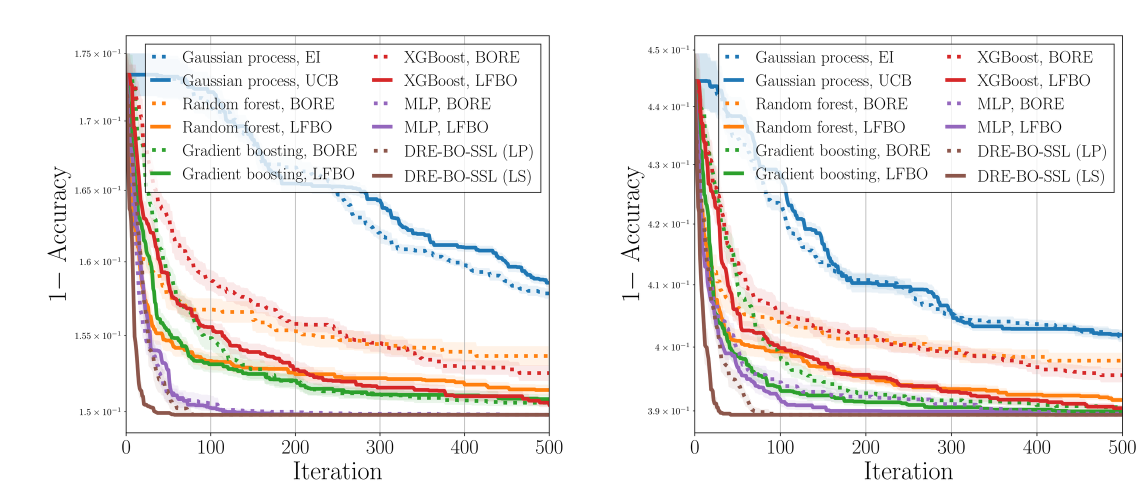
Figure: Synthetic benchmark functions.

- Scenarios with fixed-size pools



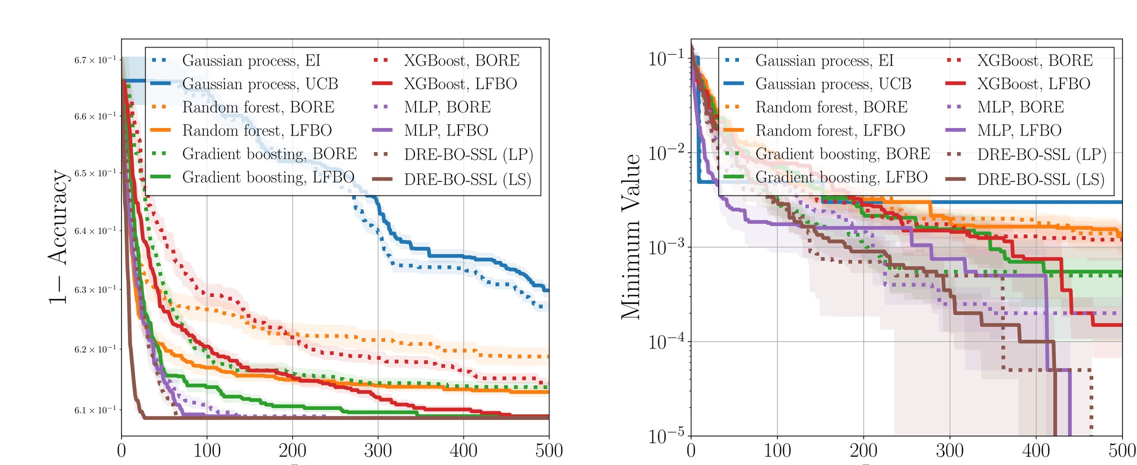
(a) Branin (b) Six-hump camel

Figure: Synthetic benchmark functions.



(a) CIFAR-10 (b) CIFAR-100

Figure: NATS-Bench.



(a) ImageNet-16-120 (b) Multi-digit MNIST search

Figure: NATS-Bench and minimum multi-digit MNIST search.

arXiv GitHub My Profile