

Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance

Shuhei Watanabe *

Preferred Networks Inc., Japan

SHUHEIWATANABE@PREFERRED.JP

Abstract

Recent scientific advances require complex experiment design, necessitating the meticulous tuning of many experiment parameters. Tree-structured Parzen estimator (TPE) is a widely used Bayesian optimization method in recent parameter tuning frameworks such as Hyperopt and Optuna. Despite its popularity, the roles of each control parameter in TPE and the algorithm intuition have not been discussed so far. The goal of this paper is to identify the roles of each control parameter and their impacts on parameter tuning based on the ablation studies using diverse benchmark datasets. The recommended setting concluded from the ablation studies is demonstrated to improve the performance of TPE. Our TPE implementation used in this paper is available at <https://github.com/nabenabe0928/tpe/tree/single-opt>¹.

1. Introduction

Recent scientific advances have seen the possibility of black-box optimization (BBO) in research fields such as drug discovery (Schneider et al., 2020), material discovery (Xue et al., 2016; Li et al., 2017a; Vahid et al., 2018), financial applications (Gonzalvez et al., 2019), and hyperparameter optimization (HPO) of machine learning algorithms (Loshchilov & Hutter, 2016; Chen et al., 2018; Feurer & Hutter, 2019). This trend has spurred the development of sample-efficient parameter-tuning frameworks such as Optuna (Akiba et al., 2019), Ray (Liaw et al., 2018), BoTorch (Balandat et al., 2020), and Hyperopt (Bergstra et al., 2011, 2013a, 2013b, 2015), enabling researchers to make significant strides in these domains.

Tree-structured Parzen estimator (TPE) is a widely used Bayesian optimization (BO) method in these frameworks and it has achieved various outstanding performances so far. For example, TPE played a pivotal role for HPO of deep learning models in winning Kaggle competitions (Alina et al., 2019; Addison et al., 2022) and Watanabe et al. (2023a) won the AutoML 2022 competition on “Multiobjective Hyperparameter Optimization for Transformers” using TPE. Furthermore, TPE has been extended to multi-fidelity (Falkner et al., 2018), multi-objective (Ozaki et al., 2020, 2022b), meta-learning (Watanabe et al., 2022, 2023a), and constrained (Watanabe & Hutter, 2022, 2023) settings to tackle diverse scenarios. Despite its versatility, its algorithm intuition and the roles of each control parameter have not been

*. This work was done at the University of Freiburg.

1. As our experiments have complex package dependencies, we provide the stable reproduced implementation of our TPE in OptunaHub: <https://hub.optuna.org/samplers/tpeTutorial/>.

discussed so far. Therefore, we describe the algorithm intuition and empirically present the roles of each control parameter. The rest of this paper is structured as follows:

1. **Background:** explains the knowledge required for this paper,
2. **Algorithm Details of TPE:** describes the TPE algorithm and empirically presents the roles of each control parameter, and
3. **Ablation Study:** performs the ablation study of the control parameters in the original TPE and investigates the enhancement in the bandwidth selection on diverse benchmark datasets.

The recommended setting drawn from the analysis is compared against recent baseline methods. This paper narrows down our scope solely to single-objective optimization problems for simplicity. We defer the details of the extensions and the applications of TPE to Appendix B and some general tips for HPO to Appendix F.

2. Background

This section describes the knowledge required to read through this paper.

2.1 Notations

We first define notations in this paper.

- $\mathcal{X}_d \subseteq \mathbb{R}$ (for $d = 1, \dots, D$), a domain of the d -th (transformed) hyperparameter,
- $\mathbf{x} \in \mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_D \subseteq \mathbb{R}^D$, a (transformed) hyperparameter configuration,
- $y = f(\mathbf{x}) + \varepsilon$, an observation of the objective function $f : \mathcal{X} \rightarrow \mathbb{R}$ with a noise ε ,
- $\mathcal{D} := \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, a set of observations (the size $N := |\mathcal{D}|$),
- $\mathcal{D}^{(l)}, \mathcal{D}^{(g)}$, a better group and a worse group in \mathcal{D} (the sizes $N^{(l)} := |\mathcal{D}^{(l)}|, N^{(g)} := |\mathcal{D}^{(g)}|$),
- $\gamma \in (0, 1]$, a top quantile used for the better group $\mathcal{D}^{(l)}$,
- $y^\gamma \in \mathbb{R}$, the top- γ quantile objective value in \mathcal{D} ,
- $p(\mathbf{x}|\mathcal{D}^{(l)}), p(\mathbf{x}|\mathcal{D}^{(g)})$, the probability density functions (PDFs) of the better group and the worse group built by kernel density estimators (KDEs),
- $r(\mathbf{x}|\mathcal{D}) := p(\mathbf{x}|\mathcal{D}^{(l)})/p(\mathbf{x}|\mathcal{D}^{(g)})$, the density ratio (equivalent to acquisition function) used to judge the promise of a hyperparameter configuration,
- $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$, a kernel function with bandwidth $b \in \mathbb{R}_+$ that changes based on a provided dataset,
- $b^{(l)}, b^{(g)} \in \mathbb{R}_+$, bandwidth (a control parameter) for the kernel function based on $\mathcal{D}^{(l)}$ and $\mathcal{D}^{(g)}$,
- $w_n \in [0, 1]$ (for $n = 1, \dots, N$), a weight for each basis in KDEs,
- $\stackrel{\text{rank}}{\simeq}$, the order isomorphic between left hand side and right hand side and $\phi(\mathbf{x}) \stackrel{\text{rank}}{\simeq} \psi(\mathbf{x})$ means $\phi(\mathbf{x}_1) < \phi(\mathbf{x}_2) \Leftrightarrow \psi(\mathbf{x}_1) < \psi(\mathbf{x}_2)$.

Note that “transformed” implies that some parameters might be preprocessed by such as log transformation or logit transformation, and the notations l (*lower* or better) and g (*greater* or worse) come from the original paper (Bergstra et al., 2011).

2.2 Bayesian Optimization

Bayesian optimization (BO)² aims to minimize the objective function $f(\mathbf{x})$ as follows:

$$\mathbf{x}_{\text{opt}} \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}). \quad (1)$$

Note that this paper consistently considers **minimization** problems. For example, hyperparameter optimization (HPO) of machine learning algorithms aims to find an optimal hyperparameter configuration \mathbf{x}_{opt} (e.g., learning rate, dropout rate, and the number of layers) that exhibits the best performance (e.g., the error rate in classification tasks, and mean squared error in regression tasks). BO iteratively searches for \mathbf{x}_{opt} using the so-called acquisition function to trade off the degree of exploration and exploitation. Roughly speaking, while exploitation searches near promising observations, exploration searches unseen regions. A common choice for the acquisition function is the following expected improvement (EI) (Jones et al., 1998):

$$\text{EI}_{y^*}[\mathbf{x}|\mathcal{D}] := \int_{-\infty}^{y^*} (y^* - y)p(y|\mathbf{x}, \mathcal{D})dy. \quad (2)$$

Another choice is the probability of improvement (PI) (Kushner, 1964):

$$\mathbb{P}(y \leq y^*|\mathbf{x}, \mathcal{D}) := \int_{-\infty}^{y^*} p(y|\mathbf{x}, \mathcal{D})dy. \quad (3)$$

Note that y^* is a control parameter specified by algorithms or users. In principle, PI is exploitative and EI is explorative. TPE is an exploitative BO method because the acquisition function of TPE is equivalent to PI as shown by Watanabe and Hutter (2022, 2023), Song et al. (2022)³. Owing to the exploitative nature, TPE is inclined to search locally. The posterior $p(y|\mathbf{x}, \mathcal{D})$ computation varies depending on the BO methods. Although a typical choice is Gaussian process regression (Williams & Rasmussen, 2006), practical methods such as SMAC (Hutter et al., 2011) and TPE use random forests and KDEs, respectively. The next section describes the posterior modeling $p(y|\mathbf{x}, D)$ for TPE by KDEs.

2.3 Tree-Structured Parzen Estimator

TPE is a variant of BO methods first invented by Bergstra et al. (2011). The name comes from the method being able to handle a tree-structured search space, and using Parzen estimators, aka kernel density estimators (KDEs). Notice that a tree-structured search space is a search space that includes some conditional parameters⁴. TPE models the posterior $p(y|\mathbf{x}, \mathcal{D})$ using the following assumption:

$$p(\mathbf{x}|y, \mathcal{D}) := \begin{cases} p(\mathbf{x}|\mathcal{D}^{(l)}) & (y \leq y^\gamma) \\ p(\mathbf{x}|\mathcal{D}^{(g)}) & (y > y^\gamma) \end{cases}, \quad (4)$$

-
- 2. We encourage readers to check recent surveys (Brochu et al., 2010; Shahriari et al., 2016; Garnett, 2022).
 - 3. Bergstra et al. (2011) originally state that the acquisition function of TPE is EI, but EI is equivalent to PI in the TPE formulation.
 - 4. For example, when we optimize the dropout rates at each layer in an L -layered neural network where $L \in \{2, 3\}$, the dropout rate at the third layer does not exist for $L = 2$. Therefore, we call the dropout rate at the third layer a *conditional parameter*. Note that this paper tests TPE only on non tree-structured spaces.

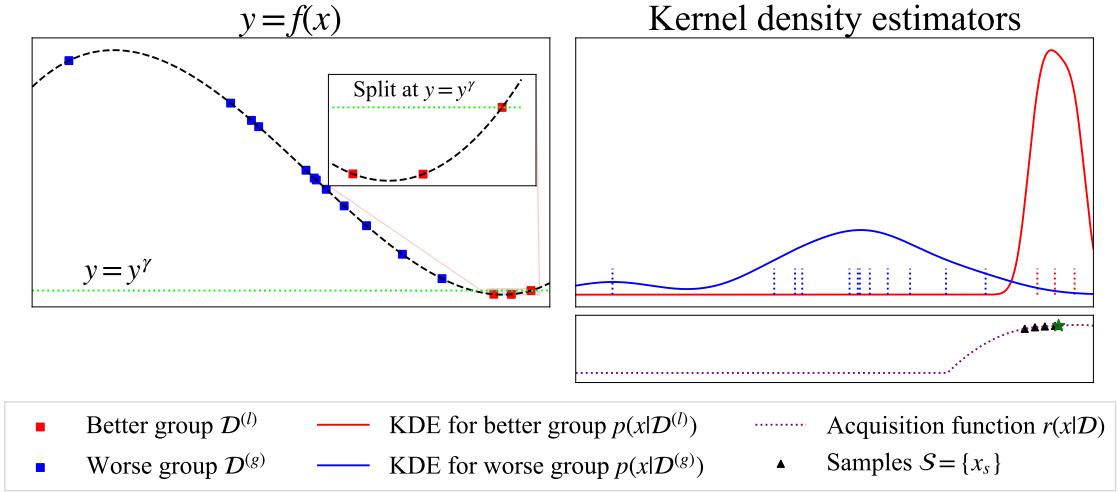


Figure 1: The conceptual visualization of TPE. **Left:** the objective function $y = f(\mathbf{x})$ (black dashed line) and its observations \mathcal{D} . The magnified figure shows the boundary $y = y^\gamma$ (green dotted line) of $\mathcal{D}^{(l)}$ (red squares) and $\mathcal{D}^{(g)}$ (blue squares). **Top right:** the KDEs built by $\mathcal{D}^{(l)}$ (red solid line) and $\mathcal{D}^{(g)}$ (blue solid line). **Bottom right:** the density ratio $p(\mathbf{x}|\mathcal{D}^{(l)})/p(\mathbf{x}|\mathcal{D}^{(g)})$ (purple dotted line) used for the acquisition function. We pick the configuration with the best acquisition function value (green star) in the samples (black triangles) from $p(\mathbf{x}|\mathcal{D}^{(l)})$.

where the top-quantile γ is computed at each iteration based on the number of observations $N (= |\mathcal{D}|)$ (see Section 3.1), and y^γ is the top- γ -quantile objective value in the set of observations \mathcal{D} ; see Figure 1 for the intuition. For simplicity, we assume that \mathcal{D} is already sorted by y_n such that $y_1 \leq y_2 \leq \dots \leq y_N$. Then the better group and the worse group are obtained as $\mathcal{D}^{(l)} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N^{(l)}}$ and $\mathcal{D}^{(g)} = \{(\mathbf{x}_n, y_n)\}_{n=N^{(l)}+1}^N$ where $N^{(l)} = \lceil \gamma N \rceil$. The KDEs in Eq. (4) are estimated via:

$$\begin{aligned} p(\mathbf{x}|\mathcal{D}^{(l)}) &= w_0^{(l)} p_0(\mathbf{x}) + \sum_{n=1}^{N^{(l)}} w_n k(\mathbf{x}, \mathbf{x}_n | b^{(l)}), \\ p(\mathbf{x}|\mathcal{D}^{(g)}) &= w_0^{(g)} p_0(\mathbf{x}) + \sum_{n=N^{(l)}+1}^N w_n k(\mathbf{x}, \mathbf{x}_n | b^{(g)}) \end{aligned} \quad (5)$$

where the weights $\{w_n\}_{n=1}^N$ are determined at each iteration (see Section 3.2), k is a kernel function (see Section 3.3), $b^{(l)}, b^{(g)} \in \mathbb{R}_+$ are the bandwidth (see Section 3.3.4) and p_0 is non-informative prior (see Section 3.3.5). Note that the summations of weights are 1, meaning that $w_0^{(l)} + \sum_{n=1}^{N^{(l)}} w_n = 1$ and $w_0^{(g)} + \sum_{n=N^{(l)}+1}^N w_n = 1$ hold. Using the assumption in Eq. (4), we obtain the following acquisition function:

$$\mathbb{P}(y \leq y^\gamma | \mathbf{x}, \mathcal{D}) \stackrel{\text{rank}}{\simeq} r(\mathbf{x}|\mathcal{D}) := \frac{p(\mathbf{x}|\mathcal{D}^{(l)})}{p(\mathbf{x}|\mathcal{D}^{(g)})}. \quad (6)$$

Algorithm 1 Tree-structured Parzen estimator (TPE)

N_{init} (The number of initial configurations, `n_startup_trials` in Optuna), N_s (The number of candidates to consider in the optimization of the acquisition function, `n_ei_candidates` in Optuna), Γ (A function to compute the top quantile γ , `gamma` in Optuna), W (A function to compute weights $\{w_n\}_{n=0}^{N+1}$, `weights` in Optuna), k (A kernel function), B (A function to compute a bandwidth b for k).

-
- ```

1: $\mathcal{D} \leftarrow \emptyset$
2: for $n = 1, 2, \dots, N_{\text{init}}$ do ▷ Initialization
3: Randomly pick \boldsymbol{x}_n
4: $y_n := f(\boldsymbol{x}_n) + \epsilon_n$ ▷ Evaluate the (expensive) objective function
5: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\boldsymbol{x}_n, y_n)\}$
6: while Budget is left do
7: Compute $\gamma \leftarrow \Gamma(N)$ with $N := |\mathcal{D}|$ ▷ Section 3.1 (Splitting algorithm)
8: Split \mathcal{D} into $\mathcal{D}^{(l)}$ and $\mathcal{D}^{(g)}$
9: Compute $\{w_n\}_{n=0}^{N+1} \leftarrow W(\mathcal{D})$ ▷ See Section 3.2 (Weighting algorithm)
10: Compute $b^{(l)} \leftarrow B(\mathcal{D}^{(l)})$, $b^{(g)} \leftarrow B(\mathcal{D}^{(g)})$ ▷ Section 3.3.4 (Bandwidth selection)
11: Build $p(\boldsymbol{x}|\mathcal{D}^{(l)})$, $p(\boldsymbol{x}|\mathcal{D}^{(g)})$ based on Eq. (5) ▷ Use $\{w_n\}_{n=0}^{N+1}$ and $b^{(l)}, b^{(g)}$
12: Sample $\mathcal{S} := \{\boldsymbol{x}_s\}_{s=1}^{N_s} \sim p(\boldsymbol{x}|\mathcal{D}^{(l)})$
13: Pick $\boldsymbol{x}_{N+1} := \boldsymbol{x}^* \in \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{S}} r(\boldsymbol{x}|\mathcal{D})$ ▷ The evaluations by the acquisition function
14: $y_{N+1} := f(\boldsymbol{x}_{N+1}) + \epsilon_{N+1}$ ▷ Evaluate the (expensive) objective function
15: $\mathcal{D} \leftarrow \mathcal{D} \cup \{\boldsymbol{x}_{N+1}, y_{N+1}\}$

```
- 

The intermediate process is provided in Appendix A. The main routine of the TPE algorithm lies in Lines 6–15 of Algorithm 1. Each iteration repeats the TPE algorithm parameter calculations and the selection of the next configuration based on the acquisition function. The selection candidates are sampled from the KDE built by the better group  $p(\boldsymbol{x}|\mathcal{D}^{(l)})$ . Although the global convergence of TPE is guaranteed when the  $\epsilon$ -greedy algorithm is used in Line 13 as shown by Watanabe et al. (2023a), this paper focuses on the greedy algorithm due to our assumption of a restrictive budget ( $\sim 200$  evaluations).

### 3. Algorithm Details of Tree-Structured Parzen Estimator

This section describes each component of TPE and elucidates the roles of each control parameter. More specifically, we highlight the effects of each control parameter on the trade-off between exploitation and exploration. Roughly speaking, exploitation searches near promising observations and exploration searches unseen regions. Each subsection title accompanies (`arg_name`), which refers to the corresponding argument's name of `TPESampler` in Optuna. Each visualization uses the default setting except `multivariate=True` of Optuna v4.0.0 if not specified. Table 2 presents the implementational differences in TPE variants.

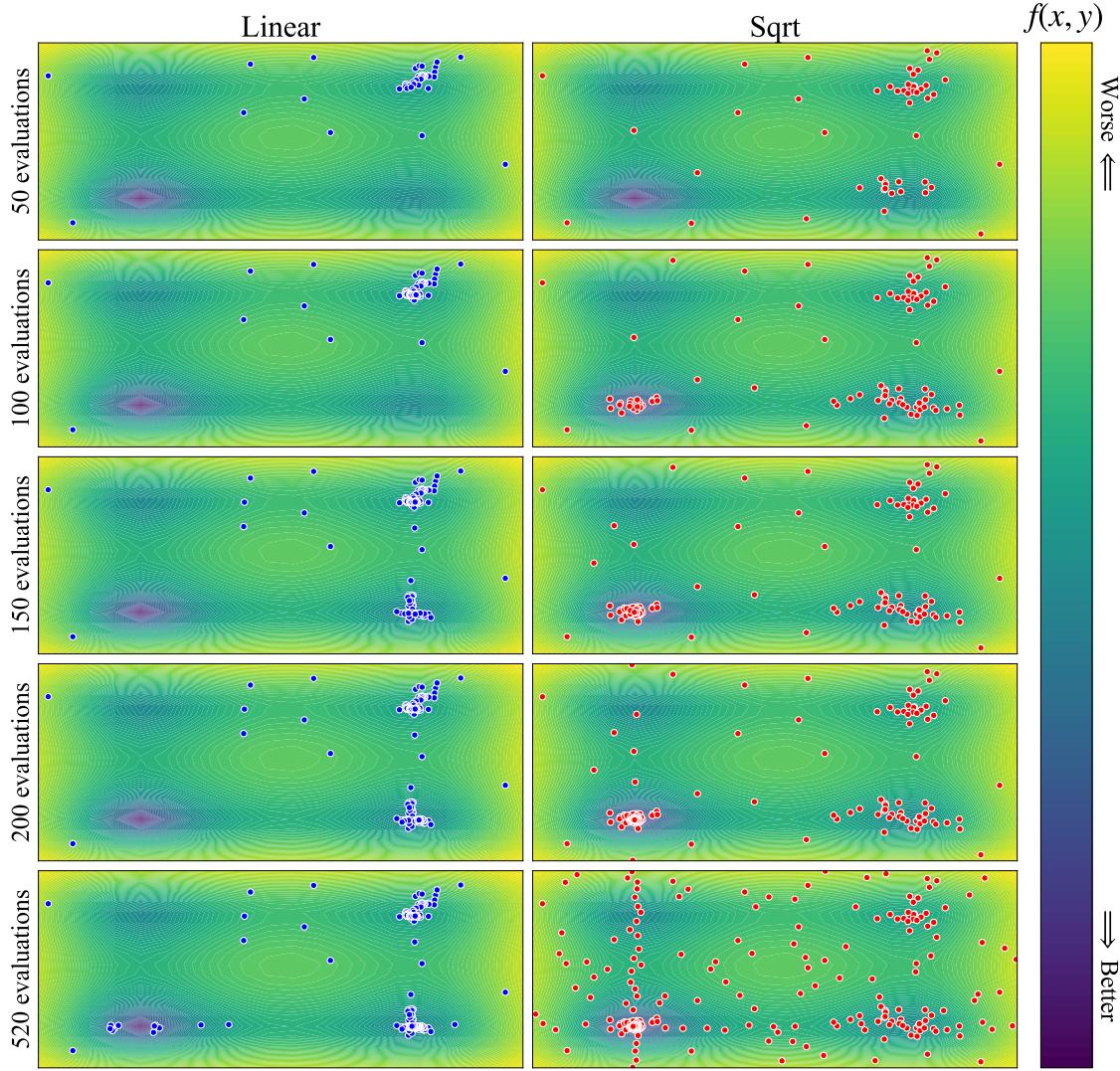


Figure 2: The optimizations of the Styblinski function using the splitting algorithm `linear` and `sqrt`. The red and blue dots show the observations till each “X evaluations”. The lower left blue shade in each figure is the optimal point and this area should be found with as few observations as possible. **Left column:** the optimization using `linear`. The optimal area is found with around 500 evaluations owing to strong exploitation. **Right column:** the optimization using `sqrt`. The optimal area is found with around 100 evaluations thanks to exploration. Although there is no observation near the optimal area for both methods at 50 evaluations, `sqrt` finds the optimal area thanks to its exploration nature.

### 3.1 Splitting Algorithm ( $\gamma$ )

The splitting algorithm of  $\mathcal{D}$  into the better group  $\mathcal{D}^{(l)}$  and the worse group  $\mathcal{D}^{(g)}$  is controlled by the quantile  $\gamma$  and a function  $\Gamma(N)$ . The following two variants of  $\Gamma$  are proposed separately by Bergstra et al. (2011) and Bergstra et al. (2013a), respectively:

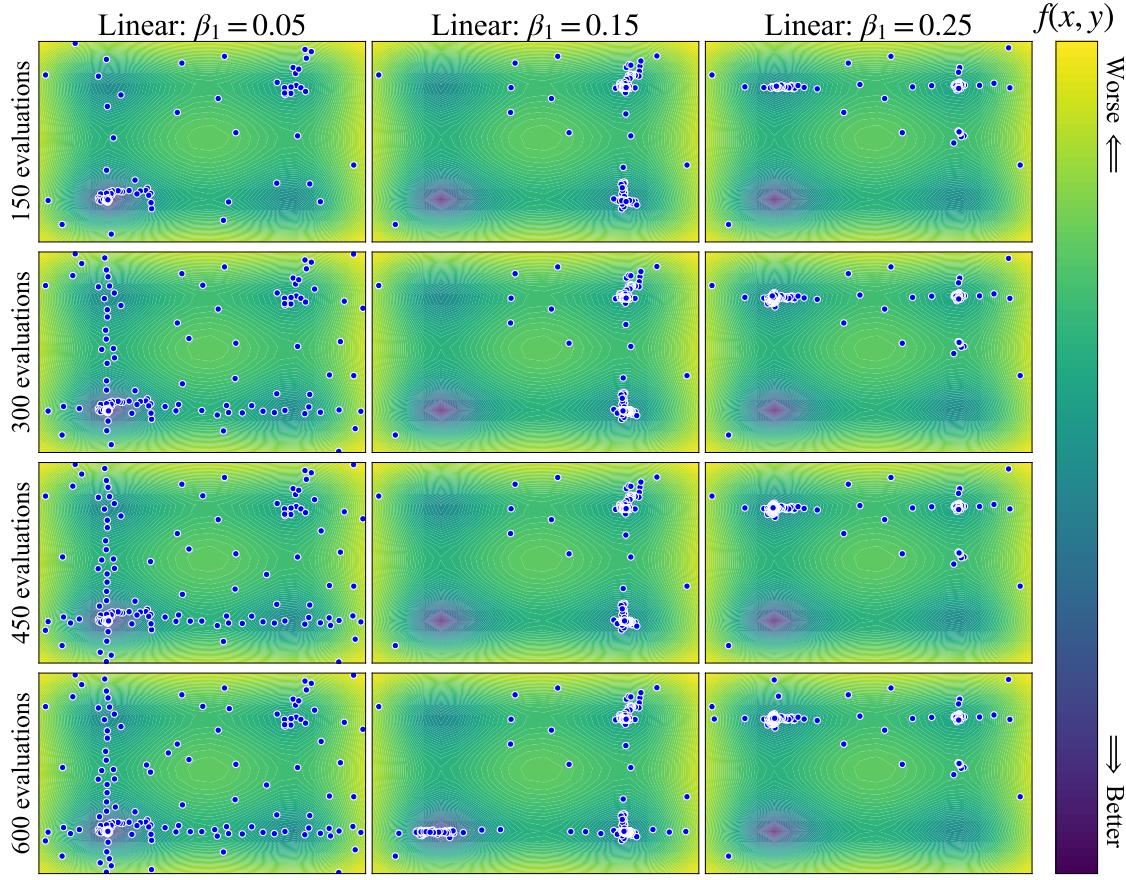


Figure 3: The optimizations of the Styblinski function using the splitting algorithm `linear` with different  $\beta_1$  (**Left column**:  $\beta_1 = 0.05$ , **Center column**:  $\beta_1 = 0.15$ , **Right column**:  $\beta_1 = 0.25$ ). The **blue dots** show the observations till each “X evaluations”. The lower left blue shade in each figure is the optimal point and this area should be found with as few observations as possible. We see scattered dots (more explorative) for a small  $\beta_1$  and concentrated dots (more exploitative) for a large  $\beta_1$ .

- (`linear`)  $\gamma := \Gamma(N) = \beta_1 \in (0, 1]$ ,
- (`Square root (sqrt)`)  $\gamma := \Gamma(N) = \beta_2/\sqrt{N}, \beta_2 \in (0, \sqrt{N}]$ ,

where  $\beta_1 = 0.15$  and  $\beta_2 = 0.25$  are used in the original papers and the number of observations in the better group  $\mathcal{D}^{(l)}$  is limited to 25 at maximum. Figure 2 shows that `sqrt` promotes more exploration and suppresses exploitation, and Figures 3, 4 demonstrate smaller  $\beta_1$  and  $\beta_2$  lead to more exploration and less exploitation. Small  $\beta_1$  and  $\beta_2$  are explorative because:

1.  $p(\mathbf{x}|\mathcal{D}^{(l)})$  would have a narrower modal that requires few observations for  $p(\mathbf{x}|\mathcal{D}^{(g)})$  to cancel out the contribution from  $p(\mathbf{x}|\mathcal{D}^{(l)})$  in the density ratio  $r(\mathbf{x}|\mathcal{D})$ , and thus it takes less time to switch to exploration, and

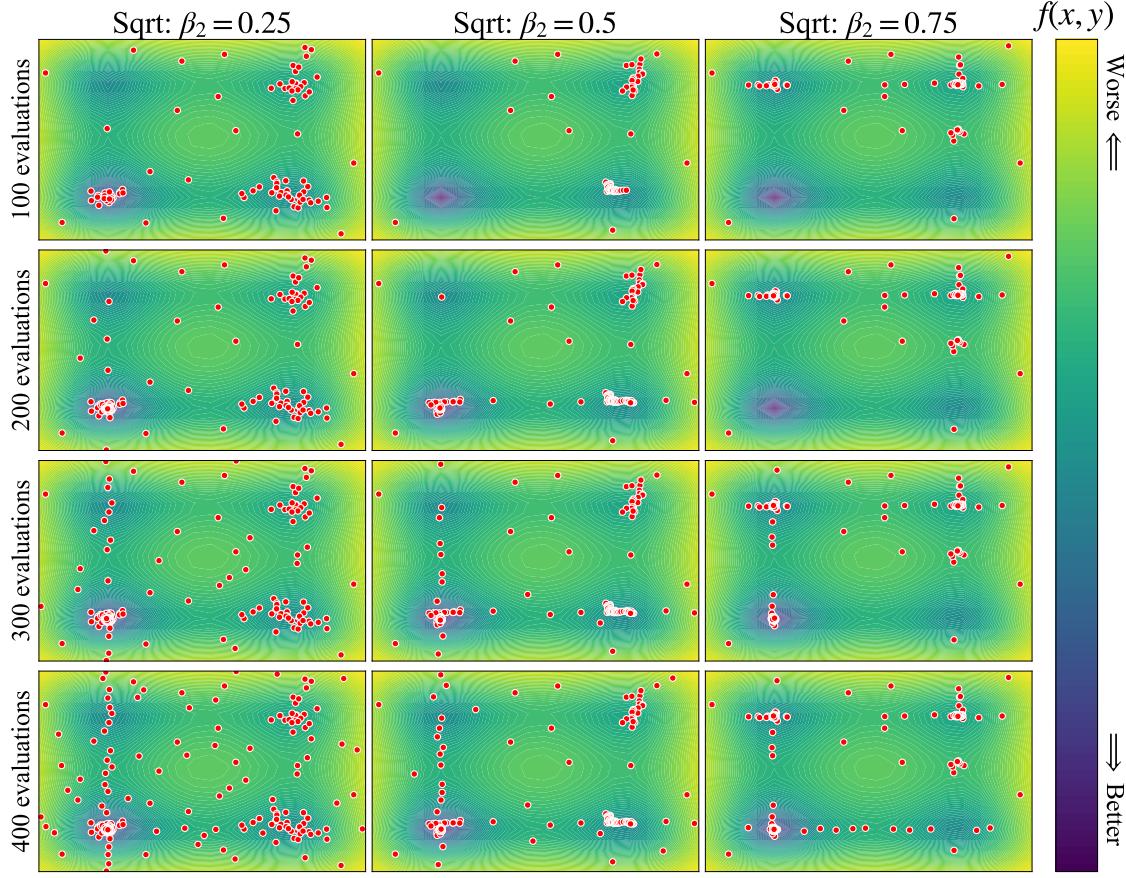


Figure 4: The optimizations of the Styblinski function using the splitting algorithm `sqrt` with different  $\beta_2$  (**Left column**:  $\beta_2 = 0.25$ , **Center column**:  $\beta_2 = 0.5$ , **Right column**:  $\beta_2 = 0.75$ ). The **red dots** show the observations till each “X evaluations”. The lower left blue shade in each figure is the optimal point and this area should be found with as few observations as possible We see scattered dots (more explorative) for a small  $\beta_2$  and concentrated dots (more exploitative) for a large  $\beta_2$ .

2. Since the prior weight  $w_0^{(l)}$  becomes larger due to a smaller number of observations in the better group  $\mathcal{D}^{(l)}$ , the prior effect becomes more dominant in  $p(\mathbf{x}|\mathcal{D}^{(l)})$  of Eq. (5) and it promotes exploration.

On the other hand, when the objective function has multiple modals, the multi-modality in  $p(\mathbf{x}|\mathcal{D}^{(l)})$  allows to explore all modalities, and large  $\beta_1$  or  $\beta_2$  does not necessarily lead to poor performance. It explains why Ozaki et al. (2020, 2022b) use `linear`, which gives a larger  $\gamma$ , for multi-objective settings.

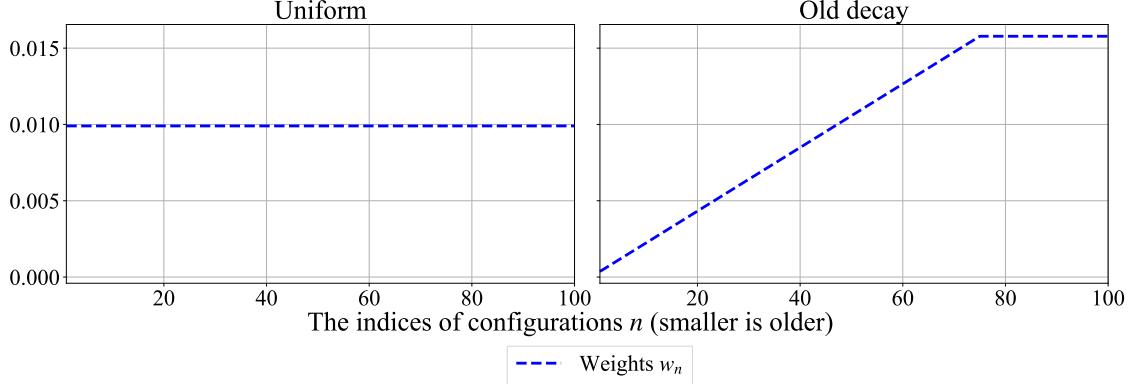


Figure 5: The distributions of each weighting algorithm when using  $N^{(g)} = 100$ . **Left:** the weight distribution for the uniform. **Right:** the weight distribution for the old decay. Older observations get lower weights and the latest 25 observations get the uniform weight.

Table 1: The advantages and disadvantages of each weighting algorithm.

| Weighting algorithm  | Advantages                                                                          | Disadvantages                                                                              |
|----------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|
| Uniform              | - Use all observations equally<br>- Need no careful preprocessing of $y$            | - Take time to account the recent observations<br>- Not consider the ranking in each group |
| Old decay            | - Take less time to switch to exploration<br>- Need no careful preprocessing of $y$ | - Might waste the knowledge from the past<br>- Not consider the ranking in each group      |
| Expected improvement | - Consider the ranking in the better group                                          | - Need careful preprocessing of $y$                                                        |

### 3.2 Weighting Algorithm (weights, prior\_weight)

The weighting algorithm  $W$  is used to determine the weights  $\{w_n\}_{n=1}^N$  for KDEs. For simplicity, we denote the prior weights as  $w_0^{(l)} := w_0$  and  $w_0^{(g)} := w_{N+1}$ , and we consider only `prior_weight=1.0`, which is the default value; see Section 3.3.5 for more details about `prior_weight`. Bergstra et al. (2011) use the following *uniform* weighting algorithm:

$$w_n := \begin{cases} \frac{1}{N^{(l)}+1} & (n = 0, \dots, N^{(l)}) \\ \frac{1}{N^{(g)}+1} & (n = N^{(l)} + 1, \dots, N + 1) \end{cases}, \quad (7)$$

In contrast, Bergstra et al. (2013a) use the following *old decay* weighting algorithm:

$$w_n := \begin{cases} \frac{1}{N^{(l)}+1} & (n = 0, \dots, N^{(l)}) \\ \frac{w'_n}{\sum_{n=N^{(l)}+1}^{N+1} w'_n} & (n = N^{(l)} + 1, \dots, N + 1) \end{cases}. \quad (8)$$

where  $w'_n$  (for  $n = N^{(l)} + 1, \dots, N + 1$ ) is defined as:

$$w'_n := \begin{cases} 1 & (t_n > N^{(g)} + 1 - T_{\text{old}}) \\ \tau(t_n) + \frac{1-\tau(t_n)}{N^{(g)}+1} & (\text{otherwise}) \end{cases}. \quad (9)$$

Note that  $t_n \in \{1, \dots, N^{(g)} + 1\}$  for  $n = N^{(l)} + 1, \dots, N + 1$  is the query order of the  $n$ -th observation in  $\mathcal{D}^{(g)}$ , which means  $t_n = 1$  is the oldest and  $t_n = N^{(g)} + 1$  is the youngest, and the decay rate is  $\tau(t_n) := (t_n - 1)/(N^{(g)} - T_{\text{old}})$  where  $T_{\text{old}} = 25$  is used by Bergstra et al. (2013a). We take  $t_{N+1} = 1$  to view the prior as the oldest information. We visualize the weight distribution in Figure 5. The old decay assigns smaller weights to older observations as the search should focus on the current region of interest but not on the old regions of interest. Furthermore, the following computation makes the acquisition function *expected improvement* (EI) more strictly (Song et al., 2022):

$$w_n := \begin{cases} \frac{y^\gamma - y_n}{\sum_{n'=1}^{N^{(l)}} (1 + 1/N^{(l)}) (y^\gamma - y_{n'})} & (n = 1, \dots, N^{(l)}) \\ \frac{1}{N^{(g)} + 1} & (n = N^{(l)} + 1, \dots, N + 1) \end{cases} \quad (10)$$

where  $w'_0$  is the mean of  $y^\gamma - y_n$  for  $n = 1, \dots, N^{(l)}$  and

$$w_0 := \frac{\sum_{n=1}^{N^{(l)}} (y^\gamma - y_n) / N^{(l)}}{\sum_{n=1}^{N^{(l)}} (1 + 1/N^{(l)}) (y^\gamma - y_n)}. \quad (11)$$

Note that the weighting algorithm used in MOTPE (Ozaki et al., 2022b) is proven to be EI by Song et al. (2022) although this fact is not explicitly mentioned by Ozaki et al. (2022b).

Table 1 lists the advantages and disadvantages of each weighting algorithm. While EI can consider the scale of  $y$ ,  $y$  must be carefully preprocessed such as by standardization and log transformation. Although uniform and old decay should be used when an abundant budget is available, multiple independent runs of TPE are preferred in such cases.

### 3.3 Kernel Functions

This section explains the control parameters in the kernel functions. Notice that this section consistently denotes the kernel function for the  $d$ -th dimension as  $k_d : \mathcal{X}_d \times \mathcal{X}_d \rightarrow \mathbb{R}_{\geq 0}$  and the uniform weight is used for simplicity.

#### 3.3.1 Kernel for Numerical Parameters

The Gaussian kernel is used for numerical parameters:

$$g(x, x_n | b) := \frac{1}{\sqrt{2\pi b^2}} \exp \left[ -\frac{1}{2} \left( \frac{x - x_n}{b} \right)^2 \right]. \quad (12)$$

Bergstra et al. (2011) employ the truncated Gaussian kernel  $k_d(x, x_n) := g(x, x_n | b) / Z(x_n)$  for  $x$  defined on  $\mathcal{X}_d := [L, R]$  where  $Z(x_n) := \int_L^R g(x, x_n | b) dx$  is a normalization constant. The parameter  $b \in \mathbb{R}_+$  in the Gaussian kernel is called *bandwidth*, and Falkner et al. (2018) use Scott's rule (Scott, 2015) (see Appendix C.3.2) and Bergstra et al. (2011) use a heuristic to determine the bandwidth  $b$  as described in Appendix C.3.1. This paper uses Scott's rule as a main algorithm to be consistent with the classical KDE basis. The kernel function for a numerical discrete parameter  $x \in \{L, L + q, \dots, R\}$  is computed as:

$$k_d(x, x_n) := \frac{1}{Z(x_n)} \int_{x-q/2}^{x+q/2} g(x', x_n | b) dx' \quad (13)$$

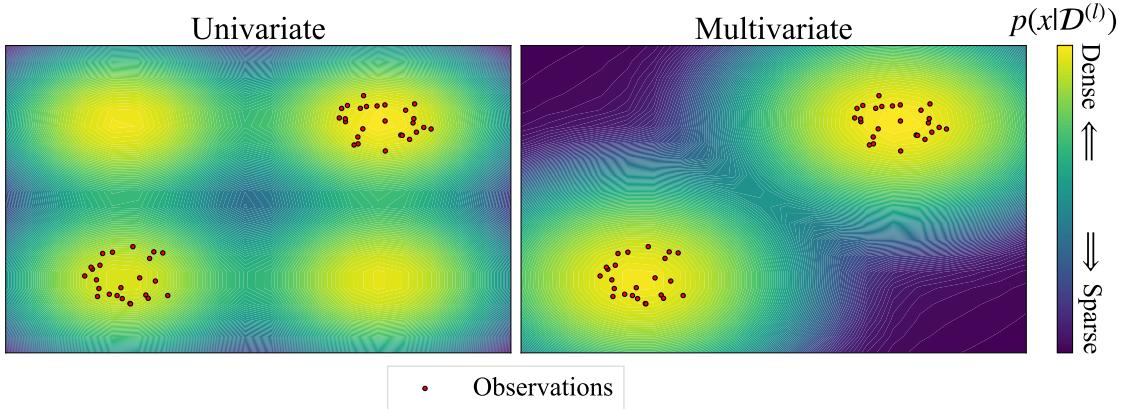


Figure 6: The difference between KDEs by univariate and multivariate kernels. The brighter color shows high density. **Left:** the contour plot of the KDE by the univariate kernel. As the univariate kernel cannot capture the interaction effects, we see bright colors even at the top-left and bottom-right regions. **Right:** the contour plot of the KDE by the multivariate kernel. As the multivariate kernel can capture the interaction effects, we see bright colors only near the observations.

where we defined  $R := L + (K - 1)q$ ,  $q \in \mathbb{R}$ , and  $K \in \mathbb{Z}_+$ . The normalization constant for the discrete kernel is computed as  $Z(x_n) := \int_{L-q/2}^{R+q/2} g(x', x_n | b) dx'$ . In general, large and small bandwidths are explorative and exploitative, respectively, as discussed in Section 3.3.4.

### 3.3.2 Kernel for Categorical Parameters

The Aitchison-Aitken kernel (Aitchison & Aitken, 1976) is used for categorical parameters:

$$k_d(x, x_n | b) = \begin{cases} 1 - b & (x = x_n) \\ \frac{b}{C-1} & (\text{otherwise}) \end{cases} \quad (14)$$

where  $C$  is the number of choices in the categorical parameter and  $b \in [0, 1]$  is the *bandwidth* for this kernel. Optuna v4.0.0 uses a heuristic for bandwidth computation shown in Appendix C.3, which we refer to `optuna` in the ablation study. The bandwidth in this kernel also controls the degree of exploration and a large bandwidth is explorative.

### 3.3.3 Univariate Kernel vs Multivariate Kernel (multivariate)

Bergstra et al. (2011, 2013a) use the so-called *univariate* KDEs:

$$p(\mathbf{x} | \{\mathbf{x}_n\}_{n=1}^N) := \frac{1}{N} \prod_{d=1}^D \sum_{n=1}^N k_d(x_d, x_{n,d} | b), \quad (15)$$

where  $x_{n,d} \in \mathcal{X}_d$  is the  $d$ -th dimension of the  $n$ -th observation  $\mathbf{x}_n$ . Although the independence assumption of each dimension allows the univariate kernel to handle conditional parameters, this assumption limits the performance if no conditional parameters exist. Falkner et al.

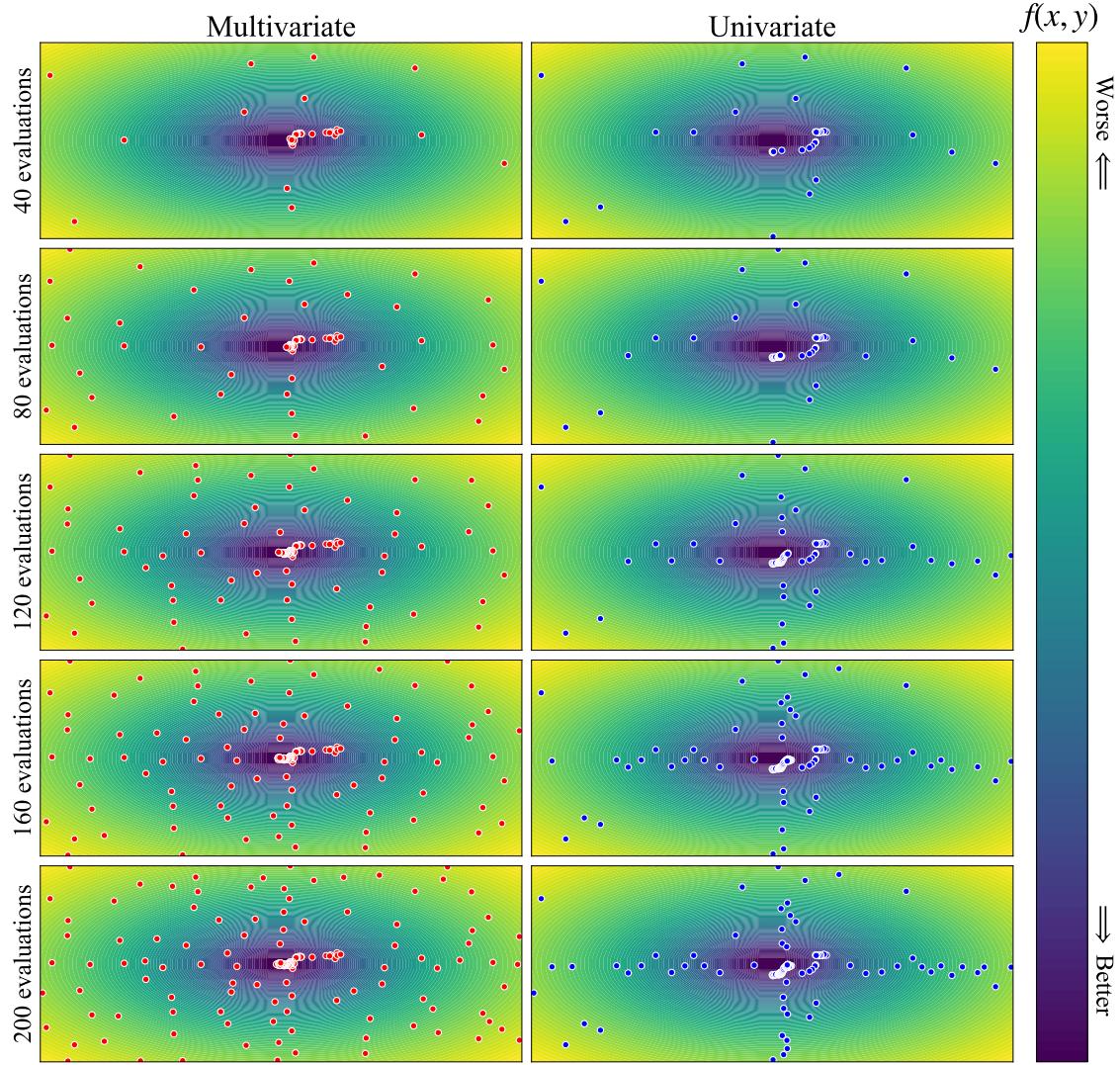


Figure 7: The optimizations of the Sphere function using the multivariate and univariate kernels. The red and blue dots show the observations till each “X evaluations”. The blue shade in each figure is the optimal point and this area should be found with as few observations as possible. **Left column:** the optimization using the multivariate kernel. Exploration starts after the center part is covered. **Right column:** the optimization using the univariate kernel. The observations gather close to the two lines  $x_1 = 0$  and  $x_2 = 0$  because the univariate kernel cannot account for interaction effects.

(2018) tackle this issue using the following *multivariate* KDEs:

$$p(\mathbf{x}|\{\mathbf{x}_n\}_{n=1}^N) := \frac{1}{N} \sum_{n=1}^N \prod_{d=1}^D k_d(x_d, x_{n,d}|b). \quad (16)$$

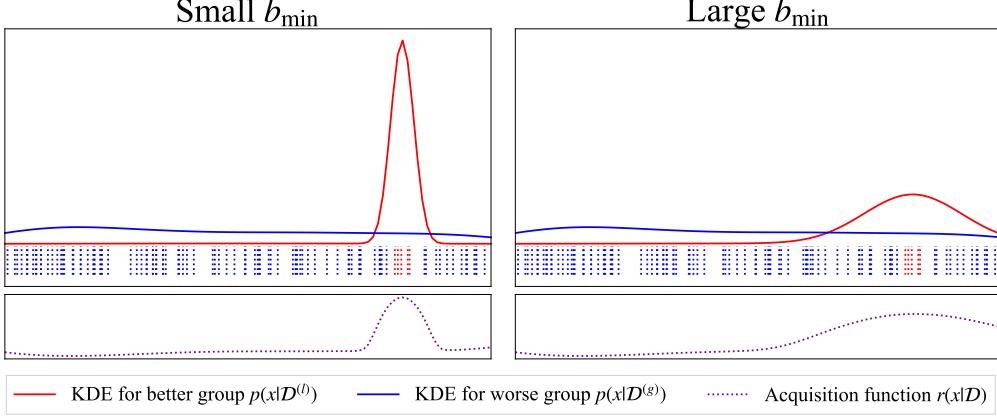


Figure 8: The change in the density ratio  $r(\mathbf{x}|\mathcal{D})$  (purple dotted lines in **Bottom row**) with respect to the minimum bandwidth  $b_{\min}$ . Both settings use the same observations (red and blue dotted lines in **Top row**). **Left:** a small  $b_{\min}$ . Since the KDE for the better group (red line) has a sharp peak, the density ratio is sharply peaked. **Right:** a large  $b_{\min}$ . The density ratio is horizontally distributed.

Although the multivariate kernel cannot be applied to the tree-structured search space as is, `group=True` in Optuna overcomes this limitation as explained in Appendix C.2. As mentioned earlier, the multivariate kernel is important to improve the performance. According to Eq. (15), since  $\sum_{n=1}^N k_d(x_d, x_{n,d}|b)$  is independent<sup>5</sup> of that of another dimension  $d' (\neq d)$ , the optimization of each dimension can be separately performed. However, as Figure 6 visualizes, the univariate kernel cannot capture the interaction effects. Meanwhile, the multivariate kernel considers interaction effects, making it unlikely to be misguided. Figure 7 shows the multivariate kernel recognizes the exact location of the modal while the univariate kernel ends up searching the axes  $x_1 = 0$  and  $x_2 = 0$  separately due to the incapability to recognize the exact modal location. Interestingly, however, the separate search of each dimension is effective for objective functions with many modals such as the Xin-She-Yang function and the Rastrigin function.

### 3.3.4 Bandwidth Modification (consider\_magic\_clip)

The bandwidth  $b$  of the numerical kernel defined on  $[L, R]$  is first computed by a heuristic, e.g., Scott's rule (Scott, 2015), and then is clipped by the so-called *magic clipping* invented by Bergstra et al. (2011):

$$b_{\text{new}} := \max(\{b, b_{\min}\}) \text{ where } b_{\min} := \frac{R - L}{\min(\{100, N\})} \quad (17)$$

where  $N = N^{(l)} + 1$  is used for  $p(\mathbf{x}|\mathcal{D}^{(l)})$  and  $N^{(g)} + 1$  is used for  $p(\mathbf{x}|\mathcal{D}^{(g)})$ . Note that if  $p(\mathbf{x}|\mathcal{D}^{(\cdot)})$  does not include the prior  $p_0$ ,  $N = N^{(\cdot)}$  is used instead. Appendix C.3 discusses

---

5. If  $f(\mathbf{x}) = \prod_{d=1}^D f_d(x_d)$  where  $f_d : \mathcal{X}_d \rightarrow \mathbb{R}_{\geq 0}$ , it is obvious that  $\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) = \prod_{d=1}^D \min_{x_d \in \mathcal{X}_d} f_d(x_d)$ . Therefore, the individual optimization of each dimension leads to the optimality. Notice that if  $f_d$  can map to a negative number, the statement is not necessarily true.

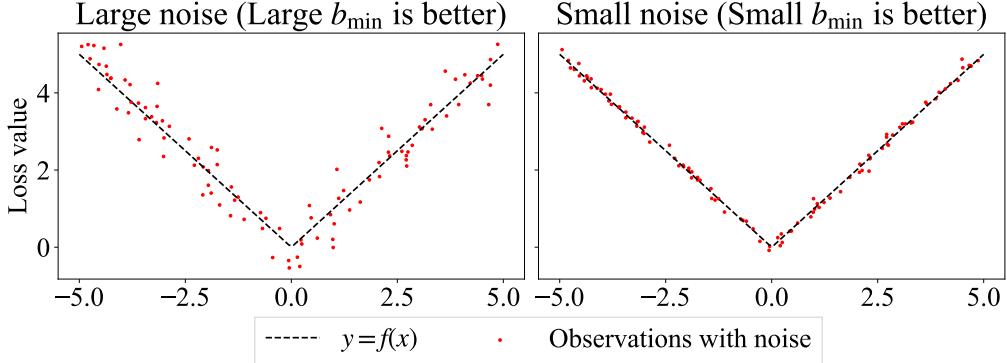


Figure 9: A concrete example where the minimum bandwidth matters. The black dashed lines are the true objective function without noise and the red dots are the observations with noise. **Left:** a function that has a large noise compared to the variation with respect to  $x$ . As can be seen, some observations near  $x = 1$  are better than some observations near  $x = 1$  due to the noise. In such cases, smaller  $b_{\min}$ , which leads to more precise optimizations, does not make much difference. **Right:** a function that has a small noise compared to the variation with respect to  $x$ . Since the noise is small, smaller  $b_{\min}$  helps to yield precise solutions.

more details of the bandwidth selection algorithms used in TPE. In principle,  $b_{\min}$  becomes exploitative when it is small and becomes explorative when it is large as shown in Figure 8. The magic clipping mostly expands the bandwidth, changing the behavior of TPE from exploitative to explorative. For example, when we search for the best dropout rate of neural networks from the range of  $[0, 1]$ , the difference between 0.4 and 0.5 is important but that between 0.40 and 0.41 is not so important because the noise is likely to be dominant in the performance variation. Figure 9 intuitively illustrates this point. Since the performance variation is explained by the noise rather than the parameter (**Left**), more precise optimization by a small bandwidth is not necessary. In contrast, when the noise is negligible (**Right**), a small bandwidth is essential. The noisy example may be optimized sufficiently by picking a value from  $\{-5, -4, \dots, 4, 5\}$ , which we call intrinsic set. We denote the size of an intrinsic set as *intrinsic cardinality*, which is 11 in this example. The scale of  $b_{\min}$  should be inversely proportional to the *intrinsic cardinality* of each parameter. The bandwidth modification is individually analyzed in the ablation study for better performance.

### 3.3.5 Non-Informative Prior (consider\_prior, prior\_weight)

Prior  $p_0$  in Eq. (5) essentially controls the regularization effect in KDEs. The PDF of Gaussian distribution  $\mathcal{N}((R+L)/2, (R-L)^2)$  is used for a numerical parameter defined on  $[L, R]$ , and the probability mass function of uniform categorical distribution  $\mathcal{U}(\{1, \dots, C\})$  is used for a categorical parameter  $\{1, \dots, C\}$ . Prior is especially important for  $p(\mathbf{x}|\mathcal{D}^{(l)})$  to prevent strong exploitation, which is seen in Figure 10, due to a small  $N^{(l)}$  throughout an optimization. Prior is mostly indispensable to TPE as discussed later. The regularization effect of prior can be controlled by `prior_weight` as well. For example, `prior_weight=2.0` doubles  $w_0^{(l)}, w_0^{(g)}$ , promoting exploration.

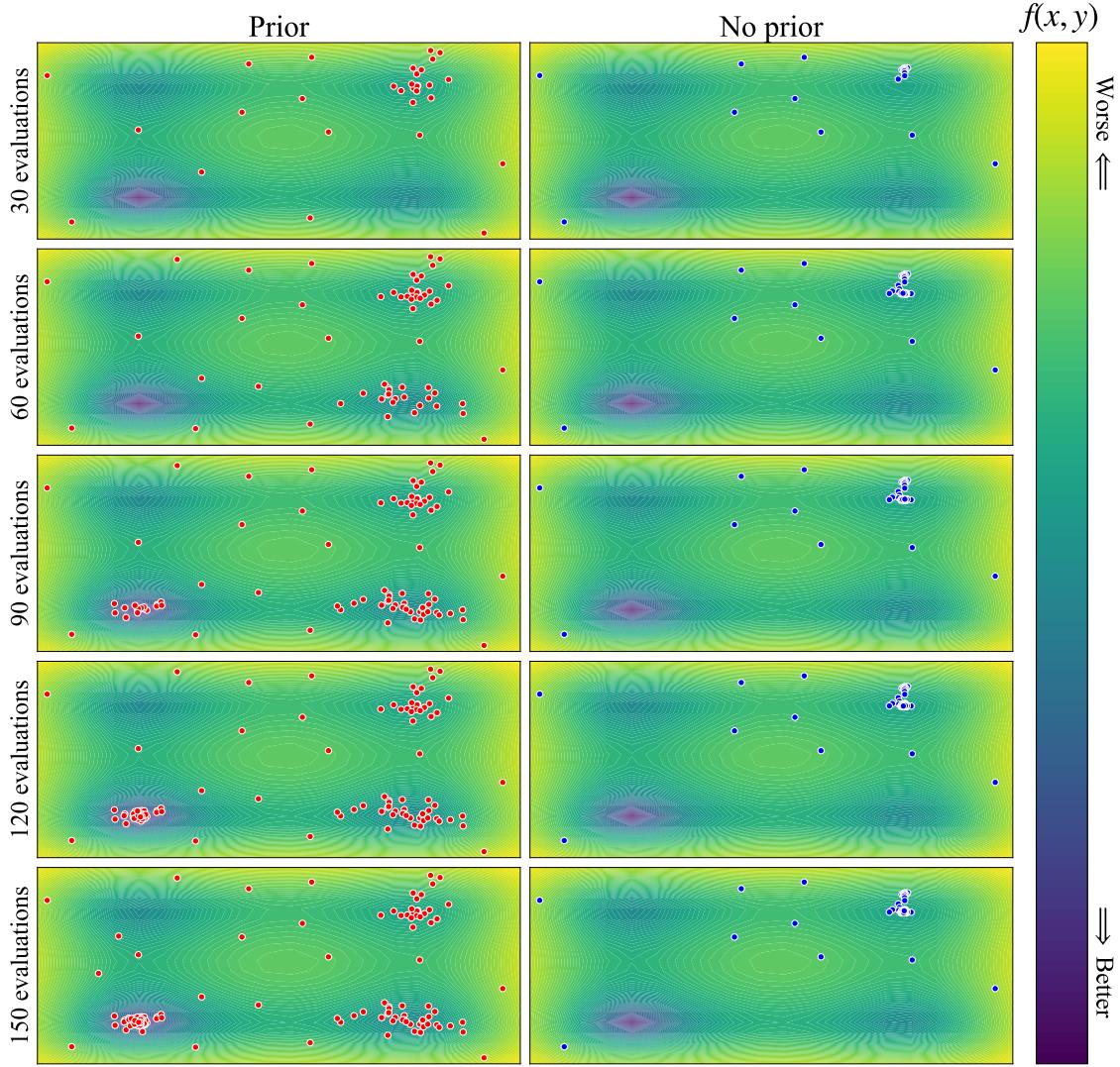


Figure 10: The optimizations of the Styblinski function with and without prior  $p_0$ . The red and blue dots show the observations till each “X evaluations”. The lower left blue shade in each figure is the optimal point and this area should be found with as few observations as possible. **Left column:** the optimization with prior. Unseen regions are explored every time each modal is covered by many observations. **Right column:** the optimization without prior. No exploration happens after one of the modals are found.

#### 4. Experiments

This section first provides the ablation study of each control parameter and presents the recommended default values. Then, we further investigate the enhancement for bandwidth selection. Finally, we compare the enhanced TPE with various baseline methods.

Table 2: The implementational differences in each component of TPE variants. Other components such as prior and the number of initial configurations are shared across all the implementations. TPE (2011), TPE (2013), BOHB, MOTPE, c-TPE, and Optuna refer to TPE by Bergstra et al. (2011), TPE (Hyperopt) by Bergstra et al. (2013a), TPE in BOHB (Falkner et al., 2018), TPE in MOTPE (Ozaki et al., 2022b), TPE in c-TPE (Watanabe & Hutter, 2023), and TPE in Optuna v4.0.0, respectively. Note that `uniform` of BOHB is computed by Eq. (21), and c-TPE uses a fixed  $b_{\min}$  instead of magic clipping. For the bandwidth selection, `hyperopt`, `scott`, and `optuna` refer to Eqs. (22), (23), and (24) in Appendix C.3, respectively. BOHB calculates the bandwidth of categorical parameters using `scott` as if they are numerical parameters.

| Version    | Splitting<br>(gamma)     | Weighting<br>(weights) | Bandwidth |             | Multivariate<br>(multivariate) | Magic clipping<br>(consider_magic_clip) |
|------------|--------------------------|------------------------|-----------|-------------|--------------------------------|-----------------------------------------|
|            |                          |                        | Numerical | Categorical |                                |                                         |
| TPE (2011) | linear, $\beta_1 = 0.15$ | uniform                | hyperopt  | $b = 0$     | False                          | True                                    |
| TPE (2013) | sqrt, $\beta_2 = 0.25$   | old decay              | hyperopt  | $b = 0$     | False                          | True                                    |
| BOHB       | linear, $\beta_1 = 0.15$ | uniform*               | scott     | scott       | True                           | False                                   |
| MOTPE      | linear, $\beta_1 = 0.10$ | EI                     | hyperopt  | $b = 0$     | False                          | True                                    |
| c-TPE      | sqrt, $\beta_2 = 0.25$   | uniform                | scott     | $b = 0.2$   | True                           | False*                                  |
| Optuna     | linear, $\beta_1 = 0.10$ | old decay              | optuna    | Eq. (25)    | True                           | True                                    |

Table 3: The search space of the control parameters used in the ablation study. Note that  $\beta_1, \beta_2$  are conditional parameters and we have  $2 \times 2 \times 2 \times (4 + 4) \times 4 \times 4 = 1024$  possible combinations in total.

| Component                                | Choices                            |
|------------------------------------------|------------------------------------|
| Multivariate (multivariate)              | {True, False}                      |
| Use prior $p_0$ (consider_prior)         | {True, False}                      |
| Use magic clipping (consider_magic_clip) | {True, False}                      |
| Splitting algorithm $\Gamma$ (gamma)     | {linear, sqrt}                     |
| 1. $\beta_1$ in linear                   | {0.05, 0.10, 0.15, 0.20}           |
| 2. $\beta_2$ in sqrt                     | {0.25, 0.50, 0.75, 1.0}            |
| Weighting algorithm $W$ (weights)        | {uniform, old-decay, old-drop, EI} |
| Categorical bandwidth $b$ in Eq. (14)    | {0.0, 0.1, 0.2, optuna}            |

## 4.1 Ablation Study

This section aims to identify the importance of each control parameter discussed in the previous section via the ablation study and provides the recommended default setting.

### 4.1.1 Setup

We exhaustively evaluate all the possible combinations of the control parameters specified in Table 3 to find performant default values. Note that `old-drop` is added to the choices of the weighting algorithm to verify whether old information is necessary. `old-drop` gives uniform weights to the recent  $T_{\text{old}} = 25$  observations and drops the weights (i.e., to give

Table 4: The hyperparameter importance (HPI) of each control parameter on the benchmark functions to achieve top 50% and to improve to top 5% from top 50%. HPI is measured at  $\{50, 100, 150, 200\}$  evaluations. We bold the top-2 HPI. Note that while the HPI in  $\beta$  quantifies whether variating  $\beta$  given either `linear` or `sqrt` matters, that in “Splitting algorithm” quantifies whether switching between `linear` and `sqrt` matters.

| Dimension | Control parameter                 | The number of function evaluations |               |                 |               |                 |               |                 |               |
|-----------|-----------------------------------|------------------------------------|---------------|-----------------|---------------|-----------------|---------------|-----------------|---------------|
|           |                                   | 50 evaluations                     |               | 100 evaluations |               | 150 evaluations |               | 200 evaluations |               |
|           |                                   | Top 50%                            | Top 5%        | Top 50%         | Top 5%        | Top 50%         | Top 5%        | Top 50%         | Top 5%        |
| 5D        | Multivariate                      | <b>23.67%</b>                      | 12.77%        | <b>41.39%</b>   | 14.71%        | <b>48.67%</b>   | 10.90%        | <b>50.81%</b>   | 11.28%        |
|           | Prior                             | 6.21%                              | 9.45%         | 4.65%           | 10.47%        | 6.57%           | 8.87%         | 8.45%           | 10.19%        |
|           | Magic clipping                    | <b>36.11%</b>                      | 10.35%        | <b>21.65%</b>   | 9.33%         | 15.19%          | 8.77%         | 10.45%          | 12.60%        |
|           | Splitting algorithm               | 1.09%                              | 0.97%         | 1.78%           | 3.33%         | 1.95%           | 2.95%         | 3.05%           | 3.09%         |
|           | $\beta_1$ ( <code>linear</code> ) | 9.80%                              | <b>17.95%</b> | 7.10%           | 11.04%        | 5.80%           | 8.37%         | 7.24%           | 16.86%        |
|           | $\beta_2$ ( <code>sqrt</code> )   | 19.24%                             | <b>39.12%</b> | 20.80%          | <b>27.89%</b> | <b>17.64%</b>   | <b>39.32%</b> | <b>16.08%</b>   | <b>22.36%</b> |
|           | Weighting algorithm               | 3.87%                              | 9.39%         | 2.62%           | <b>23.22%</b> | 4.18%           | <b>20.82%</b> | 3.92%           | <b>23.62%</b> |
| 10D       | Multivariate                      | 8.39%                              | 16.64%        | <b>26.23%</b>   | 14.87%        | <b>44.48%</b>   | 14.38%        | <b>48.94%</b>   | 14.24%        |
|           | Prior                             | <b>25.69%</b>                      | 11.48%        | 12.96%          | 14.76%        | 6.31%           | <b>18.81%</b> | 4.50%           | <b>19.72%</b> |
|           | Magic clipping                    | <b>38.79%</b>                      | 20.09%        | <b>31.63%</b>   | 14.98%        | <b>25.45%</b>   | 14.60%        | <b>21.81%</b>   | 11.44%        |
|           | Splitting algorithm               | 0.87%                              | 0.69%         | 2.34%           | 2.23%         | 3.32%           | 2.48%         | 3.76%           | 3.23%         |
|           | $\beta_1$ ( <code>linear</code> ) | 7.83%                              | <b>20.42%</b> | 8.49%           | <b>17.92%</b> | 5.77%           | 15.04%        | 5.68%           | 13.81%        |
|           | $\beta_2$ ( <code>sqrt</code> )   | 14.70%                             | <b>24.06%</b> | 15.79%          | <b>22.84%</b> | 11.83%          | <b>21.78%</b> | 11.51%          | <b>25.76%</b> |
|           | Weighting algorithm               | 3.74%                              | 6.61%         | 2.55%           | 12.41%        | 2.84%           | 12.91%        | 3.78%           | 11.81%        |
| 30D       | Multivariate                      | <b>32.83%</b>                      | 17.15%        | 18.62%          | 18.37%        | <b>31.01%</b>   | 16.89%        | <b>36.80%</b>   | 15.23%        |
|           | Prior                             | <b>29.53%</b>                      | 14.52%        | <b>24.91%</b>   | 15.55%        | 12.38%          | <b>19.74%</b> | 9.52%           | <b>20.66%</b> |
|           | Magic clipping                    | 16.31%                             | <b>26.46%</b> | <b>31.82%</b>   | <b>18.99%</b> | <b>28.84%</b>   | 14.69%        | <b>27.98%</b>   | 12.38%        |
|           | Splitting algorithm               | 1.06%                              | 1.87%         | 2.12%           | 3.17%         | 1.58%           | 5.57%         | 2.45%           | 5.69%         |
|           | $\beta_1$ ( <code>linear</code> ) | 4.82%                              | <b>17.80%</b> | 4.54%           | <b>23.22%</b> | 8.66%           | <b>22.77%</b> | 6.26%           | <b>22.58%</b> |
|           | $\beta_2$ ( <code>sqrt</code> )   | 12.17%                             | 8.56%         | 11.24%          | 8.94%         | 8.56%           | 7.93%         | 9.86%           | 8.71%         |
|           | Weighting algorithm               | 3.28%                              | 13.65%        | 6.75%           | 11.75%        | 8.96%           | 12.40%        | 7.13%           | 14.75%        |

zero weights) for the rest. The other parameters not specified in the table are fixed to the default values of Optuna v4.0.0 except Scott’s rule in Eq. (23) is used for the numerical bandwidth selection. The experiments are conducted over 51 objective functions listed in Appendix D to strengthen the reliability of the analysis. The objective functions include 36 benchmark functions (12 different functions  $\times$  3 different dimensionalities), 8 tasks in HPOBench (Eggensperger et al., 2021), 4 tasks in HPOlib (Klein & Hutter, 2019), and 3 tasks in JAHS-Bench-201 (Bansal et al., 2022). The default scale of  $y$  is used in EI except on HPOlib where the log scale of the validation MSE is used. Each optimization observes 200 configurations and is repeated 10 times each with a different random seed. The initialization follows the default setting of Optuna v4.0.0, meaning that 10 random configurations are evaluated before each optimization (i.e., `n_startup_trials=10`). The analysis for Figures 11 – 14 is performed based on Watanabe et al. (2023b). Roughly speaking, parameter choices above the black-dotted lines are likely to achieve good performance. More specifically, the red and blue shades above the black-dotted lines contribute to the top-5% and the top-50% among all the possible configurations, respectively. For more details, see Appendix D.



Figure 11: The probability mass values of each control parameter in the top-5% and top-50% observations at  $\{50, 100, 150, 200\}$  evaluations on the benchmark functions. High probability mass in a specific value implies that we are likely to yield top-5% or top-50% performance with the specific value.

#### 4.1.2 Results & Discussion

Figures 11,12 present the probability mass of each choice in the top-50% and -5% observations and Tables 4,5 present the HPI of each control parameter. Appendix E provides the individual results to supplement the information loss by the summarization in the figures.

## A TUTORIAL OF TREE-STRUCTURED PARZEN ESTIMATOR

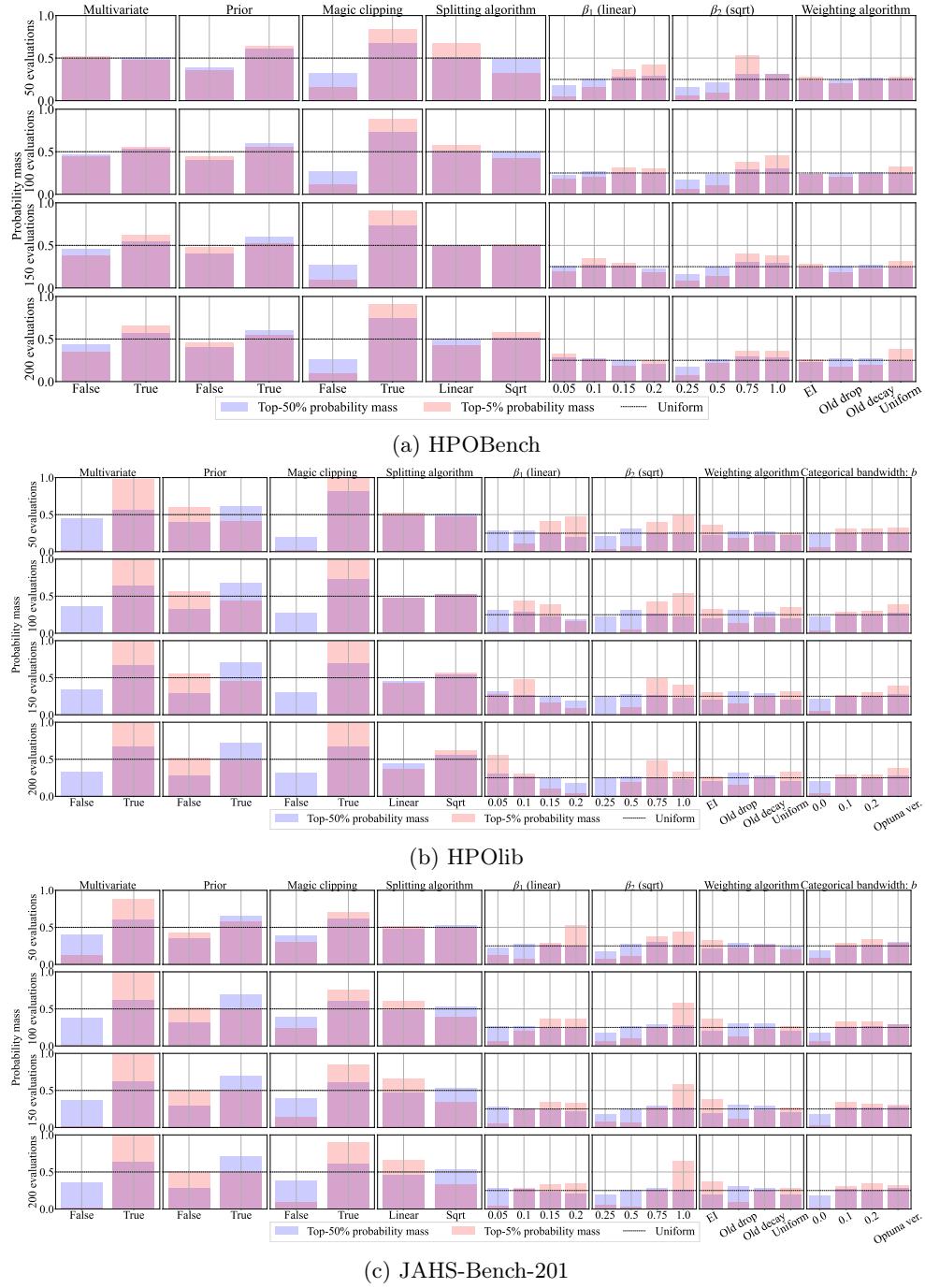


Figure 12: The probability mass values of each control parameter in the top-5% and top-50% observations at  $\{50, 100, 150, 200\}$  evaluations on the HPO benchmarks. High probability mass in a specific value implies that we are likely to yield top-5% or top-50% performance with the specific value.

**Multivariate (multivariate):** the settings with the multivariate kernel yield a strong peak in the top-5% probability mass for almost all settings and the mean performance in the individual results shows that the multivariate kernel outperforms the univariate kernel except on the 10- and 30-dimensional Xin-She-Yang function. For the benchmark functions,

Table 5: The hyperparameter importance (HPI) of each control parameter on the HPO benchmarks to achieve top 50% and to improve to top 5% from top 50%. HPI is measured at  $\{50, 100, 150, 200\}$  evaluations. We bold the top-2 HPI. Note that while the HPI in  $\beta$  quantifies whether variating  $\beta$  given either `linear` or `sqrt` matters, that in “Splitting algorithm” quantifies whether switching between `linear` and `sqrt` matters.

| Benchmark      | Control parameter                 | The number of function evaluations |               |                 |               |                 |               |                 |               |
|----------------|-----------------------------------|------------------------------------|---------------|-----------------|---------------|-----------------|---------------|-----------------|---------------|
|                |                                   | 50 evaluations                     |               | 100 evaluations |               | 150 evaluations |               | 200 evaluations |               |
|                |                                   | Top 50%                            | Top 5%        | Top 50%         | Top 5%        | Top 50%         | Top 5%        | Top 50%         | Top 5%        |
| HPOBench       | Multivariate                      | 1.41%                              | 6.79%         | 2.99%           | 8.35%         | 3.29%           | 4.78%         | 5.11%           | 6.36%         |
|                | Prior                             | 15.43%                             | 0.60%         | 11.11%          | 1.95%         | 11.78%          | 2.90%         | <b>10.93%</b>   | 3.30%         |
|                | Magic clipping                    | <b>38.76%</b>                      | 10.16%        | <b>63.38%</b>   | 11.95%        | <b>60.96%</b>   | 12.77%        | <b>62.16%</b>   | 12.43%        |
|                | Splitting algorithm               | 2.00%                              | 7.61%         | 1.29%           | 12.26%        | 1.11%           | 4.44%         | 1.02%           | 2.98%         |
|                | $\beta_1$ ( <code>linear</code> ) | 12.61%                             | <b>37.74%</b> | 3.12%           | <b>29.52%</b> | 6.00%           | <b>36.62%</b> | 7.94%           | <b>42.21%</b> |
|                | $\beta_2$ ( <code>sqrt</code> )   | <b>27.87%</b>                      | <b>29.51%</b> | <b>17.18%</b>   | <b>28.03%</b> | <b>15.49%</b>   | <b>23.06%</b> | 10.87%          | 15.18%        |
|                | Weighting algorithm               | 1.93%                              | 7.59%         | 0.93%           | 7.94%         | 1.38%           | 15.43%        | 1.97%           | <b>17.53%</b> |
| HPOlib         | Multivariate                      | 3.06%                              | 20.88%        | 15.51%          | 16.10%        | 19.25%          | <b>17.85%</b> | 21.00%          | 17.09%        |
|                | Prior                             | <b>8.88%</b>                       | 6.11%         | <b>23.51%</b>   | 7.56%         | <b>31.62%</b>   | 10.95%        | <b>35.41%</b>   | 10.44%        |
|                | Magic clipping                    | <b>72.53%</b>                      | 5.89%         | <b>38.75%</b>   | 10.62%        | <b>28.31%</b>   | 13.90%        | <b>22.96%</b>   | 16.06%        |
|                | Splitting algorithm               | 0.30%                              | 0.51%         | 0.53%           | 0.20%         | 1.51%           | 0.30%         | 2.52%           | 0.80%         |
|                | $\beta_1$ ( <code>linear</code> ) | 6.32%                              | <b>28.93%</b> | 7.62%           | <b>17.43%</b> | 6.33%           | 15.05%        | 6.58%           | <b>19.45%</b> |
|                | $\beta_2$ ( <code>sqrt</code> )   | 6.70%                              | <b>26.03%</b> | 4.92%           | <b>29.81%</b> | 2.34%           | <b>26.18%</b> | 0.97%           | <b>19.84%</b> |
|                | Weighting algorithm               | 1.98%                              | 4.91%         | 7.99%           | 9.79%         | 8.92%           | 8.40%         | 8.36%           | 8.23%         |
| JAHS-Bench-201 | Categorical bandwidth: $b$        | 0.23%                              | 6.73%         | 1.16%           | 8.50%         | 1.72%           | 7.37%         | 2.20%           | 8.08%         |
|                | Multivariate                      | 11.47%                             | <b>20.27%</b> | <b>16.16%</b>   | <b>23.69%</b> | <b>17.05%</b>   | <b>20.26%</b> | <b>19.79%</b>   | <b>16.73%</b> |
|                | Prior                             | <b>33.05%</b>                      | 5.63%         | <b>38.34%</b>   | 8.17%         | <b>40.41%</b>   | 8.09%         | <b>41.88%</b>   | 7.62%         |
|                | Magic clipping                    | <b>19.53%</b>                      | 3.75%         | 13.59%          | 5.26%         | 12.29%          | 9.45%         | 11.90%          | 11.42%        |
|                | Splitting algorithm               | 1.16%                              | 1.22%         | 0.94%           | 3.98%         | 1.45%           | 6.03%         | 1.52%           | 5.28%         |
|                | $\beta_1$ ( <code>linear</code> ) | 3.24%                              | <b>34.47%</b> | 1.34%           | 12.62%        | 2.68%           | 10.23%        | 2.94%           | 11.18%        |
|                | $\beta_2$ ( <code>sqrt</code> )   | 12.84%                             | 20.15%        | 10.78%          | <b>28.97%</b> | 8.81%           | <b>25.91%</b> | 5.29%           | <b>28.18%</b> |
| HPOlib         | Weighting algorithm               | 5.80%                              | 5.28%         | 10.22%          | 11.52%        | 9.89%           | 12.87%        | 10.38%          | 12.66%        |
|                | Categorical bandwidth: $b$        | 12.91%                             | 9.22%         | 8.62%           | 5.79%         | 7.41%           | 7.16%         | 6.30%           | 6.93%         |

the multivariate kernel is one of the most dominant factors and the HPI goes up as the number of evaluations increases in the low-dimensional problems. It matches the intuition that the multivariate kernel needs more observations to be able to exploit useful information. Although the multivariate kernel is not essential for HPOBench, it is recommended to use the multivariate kernel.

**Prior (consider\_prior):** while the settings with the prior  $p_0$  yield a strong peak in the top-50% probability mass for all the settings, it is not the case in the top-5% probability mass for the HPO benchmarks. For HPOBench and JAHS-Bench-201, although the settings with the prior are more likely to achieve the top-5% performance in the early stage of optimizations, the likelihood decreases over time. It implies that the prior (more exploration) is more effective in the beginning. On the other hand, for HPOlib, the likelihood of achieving the top-5% performance is higher in the settings without the prior. It implies that the prior (more exploitation) should be reduced in the beginning for HPOlib. According to the individual results, while the performance distributions of the settings without the prior are close to uniform, those with the prior have a stronger modal. It means that the prior is primarily important for the top-50% as can be seen in Tables 4,5 as well and the settings without the prior require more careful tuning. Although some settings on HPOlib without the

prior outperform those with the prior, we recommend using the prior because the likelihood difference is not striking.

**Magic Clipping (consider\_magic\_clip):** according to the probability mass, the magic clipping has a negative impact on the benchmark functions and a positive impact on the HPO benchmarks. Almost no settings achieve the top-5% performance with the magic clipping for the high-dimensional benchmark functions. The results relate to the noise level and the intrinsic cardinality discussed in Section 3.3.4. Since the magic clipping affects the performance strongly, we investigate it further in the next section.

**Splitting Algorithm and  $\beta$  (gamma):** the choice of either `linear` or `sqrt` does not strongly affect the likelihood of achieving the top-50%. Although the HPI of the splitting algorithm is dominated by the other HPs, the splitting algorithm choice slightly affects the results. For example, while `linear` is more effective for the 5D benchmark functions, `sqrt` is more effective for the 30D benchmark functions. Since `linear` and `sqrt` promotes exploitation and exploration, respectively, as discussed in Section 3.1, it might be useful to change the splitting algorithm depending on the dimensionality of the search space. However, the choice of  $\beta$  is much more important according to the tables and we, unfortunately, cannot see similar patterns in each figure although the following findings are observed:

- peaks of  $\beta_1$  in `linear` largely change over time,
- peaks of  $\beta_2$  in `sqrt` do not change drastically,
- `linear` with a small  $\beta_1$  is effective for the high-dimensional benchmark functions,
- `sqrt` with a large  $\beta_2$  is effective for the HPO benchmarks, and
- the peaks change from larger  $\beta$  to small  $\beta$  over time (exploitation to exploration).

Another finding is that while the magic clipping needs to be adjusted to promote exploitation for the benchmark functions and exploration for the HPO benchmarks,  $\beta$  needs to be adjusted to promote exploration (small  $\beta$ ) for the benchmark functions and exploitation (large  $\beta$ ) for the HPO benchmarks. It implies that each component controls the trade-off between exploration and exploitation differently, necessitating a careful tuning of these control parameters. However, `linear` with  $\beta_1 = 0.1$  and `sqrt` with  $\beta_2 = 0.75$  exhibit relatively stable performance for each problem.

**Weighting Algorithm (weights):** although the weighting algorithm is not an important factor to attain the top-50% performance, the results show that `EI` is effective to attain the top-5% performance, and `uniform` comes next. Note that since the behavior of `EI` heavily depends on the distribution of the objective value  $p(y)$ , `EI` might require special treatment on  $y$  as in the scale of HPOlib. For example, if the objective returns infinity, the weights in Eq. (10) cannot be defined anymore. According to the top-50% probability mass of the HPO benchmarks, `old-decay` and `old-drop` are the most frequent choices, implying that they are relatively robust to the choice of other control parameters. However, the regularization effect caused by dropping past observations limits the performance, leading to less probability mass at the top-5%.

**Categorical Bandwidth  $b$ :** the categorical bandwidth with  $b = 0$  achieves the top-5% performance in nearly no cases due to the overfitting to one category. Otherwise, any choices exhibit more or less similar performance while `optuna` slightly outperforms the other choices.

**Ablation Study Summary:** to sum up the discussion, our recommendation is to use:

- `multivariate=True`,
- `consider_prior=True`,
- `consider_magic_clip=True` for the HPO benchmarks and `False` for the benchmark functions,
- `gamma=linear` with  $\beta_1 = 0.1$  or `gamma=sqrt` with  $\beta_2 = 0.75$ ,
- `weights=EI` with some processing on  $y$  or `weights=uniform`, and
- `optuna` of the categorical bandwidth selection.

The recommended setting allows TPE to outperform Optuna v4.0.0 except for some tasks of HPOBench and HPOlib. The next section further discusses enhancements to the bandwidth selection.

## 4.2 Analysis of Bandwidth Selection

This section investigates the effect of various bandwidth selection algorithms on the performance and provides a recommended default setting.

### 4.2.1 Setup

The experiments investigate the effect of modifications on Eq. (17). Recall that the bandwidth  $b_{\text{new}} = \max(\{b, b_{\min}\})$  is used and the modified minimum bandwidth  $b_{\min}$  is computed as:

$$b_{\min} := \max(\{\Delta(R - L), b_{\text{magic}}\}), \quad (18)$$

In the experiments, we will modify:

1. the bandwidth selection heuristic, which computes  $b$ , discussed in Appendix C.3,
2. the minimum bandwidth factor  $\Delta$ ,
3. the algorithm to compute  $b_{\text{magic}}$ , and
4. whether to use the magic clipping (we use  $b_{\text{magic}} = 0$  for the non magic-clipping setting).

The algorithm of  $b_{\text{magic}}$  uses  $b_{\text{magic}} = (R - L)/N^\alpha$  where  $\alpha = 1$  is used in Optuna v4.0.0 by default. Table 6 shows the search space of the control parameters. Note that  $b_{\text{magic}} = 0$  included in the category of Modification 4 corresponds to  $\alpha = \infty$ . As the magic clipping uses a function of the observation size determined by the splitting algorithm, all the eight choices in Section 4 regarding the splitting algorithm is included in the search space along with the weighting algorithms `uniform`, `EI`, and `old-decay`. The other control parameters are fixed to the recommended setting in Section 4.1.2.

### 4.2.2 Results & Discussion

Figures 13,14 present the probability mass of each choice in the top-50% and -5% observations and Tables 7,8 present the HPI of each control parameter for the bandwidth selection. Appendix E provides the individual results to supplement the information loss by the summarization in the figures.

Table 6: The search space of the control parameters used in the investigation of better bandwidth selection. We provide the details of the bandwidth selection heuristic in Appendix C.3. We have  $6 \times 4 \times 3 = 72$  possible combinations for the bandwidth selection and  $8 \times 3 = 24$  for the splitting and the weighting algorithms. Therefore, we evaluate  $72 \times 24 = 1728$  possible combinations.

| Component                                    | Choices                                                    |
|----------------------------------------------|------------------------------------------------------------|
| The bandwidth selection heuristic            | {hyperopt, optuna, scott}                                  |
| The minimum bandwidth factor $\Delta$        | {0.01, 0.03, 0.1, 0.3}                                     |
| The exponent $\alpha$ for $b_{\text{magic}}$ | { $2^{-2}$ , $2^{-1}$ , $2^0$ , $2^1$ , $2^2$ , $\infty$ } |

Table 7: The hyperparameter importance (HPI) of each control parameter for the bandwidth selection on the benchmark functions to achieve top 50% and to improve to top 5% from top 50% for the bandwidth selection. HPI is measured at {50, 100, 150, 200} evaluations. We bold the top-2 HPI. Note that while the HPI in  $\beta$  quantifies whether variating  $\beta$  given either `linear` or `sqrt` matters, that in “Splitting algorithm” quantifies whether switching between `linear` and `sqrt` matters.

| Dimension | Control parameter                 | The number of function evaluations |               |                 |               |                 |               |                 |               |
|-----------|-----------------------------------|------------------------------------|---------------|-----------------|---------------|-----------------|---------------|-----------------|---------------|
|           |                                   | 50 evaluations                     |               | 100 evaluations |               | 150 evaluations |               | 200 evaluations |               |
|           |                                   | Top 50%                            | Top 5%        | Top 50%         | Top 5%        | Top 50%         | Top 5%        | Top 50%         | Top 5%        |
| 5D        | Bandwidth heuristic               | 0.95%                              | <b>32.15%</b> | 1.02%           | <b>33.50%</b> | 1.31%           | <b>30.52%</b> | 1.54%           | <b>29.79%</b> |
|           | $\Delta$ (factor of $b_{\min}$ )  | 7.27%                              | <b>23.56%</b> | 11.43%          | <b>31.34%</b> | <b>17.73%</b>   | <b>36.59%</b> | <b>21.60%</b>   | <b>39.45%</b> |
|           | Exponent for $b_{\text{magic}}$   | <b>67.05%</b>                      | 13.83%        | <b>66.02%</b>   | 10.76%        | <b>59.48%</b>   | 10.37%        | <b>58.83%</b>   | 10.99%        |
|           | Splitting algorithm               | 0.78%                              | 0.56%         | 1.48%           | 0.38%         | 2.26%           | 0.71%         | 2.68%           | 1.63%         |
|           | $\beta_1$ ( <code>linear</code> ) | 9.21%                              | 5.83%         | 5.93%           | 5.41%         | 3.55%           | 3.25%         | 2.56%           | 2.75%         |
|           | $\beta_2$ ( <code>sqrt</code> )   | <b>13.94%</b>                      | 11.57%        | <b>13.22%</b>   | 8.09%         | 15.18%          | 7.44%         | 12.45%          | 6.89%         |
|           | Weighting algorithm               | 0.81%                              | 12.49%        | 0.90%           | 10.50%        | 0.47%           | 11.13%        | 0.34%           | 8.50%         |
| 10D       | Bandwidth heuristic               | 1.23%                              | <b>41.90%</b> | 2.65%           | <b>35.62%</b> | 3.08%           | <b>33.22%</b> | 3.14%           | <b>34.17%</b> |
|           | $\Delta$ (factor of $b_{\min}$ )  | <b>20.61%</b>                      | <b>18.65%</b> | <b>18.04%</b>   | <b>33.10%</b> | <b>18.67%</b>   | <b>32.97%</b> | <b>22.85%</b>   | <b>33.27%</b> |
|           | Exponent for $b_{\text{magic}}$   | <b>67.00%</b>                      | 17.16%        | <b>63.13%</b>   | 14.19%        | <b>59.39%</b>   | 14.24%        | <b>57.82%</b>   | 13.11%        |
|           | Splitting algorithm               | 0.58%                              | 0.83%         | 1.09%           | 0.56%         | 1.77%           | 1.40%         | 1.68%           | 1.20%         |
|           | $\beta_1$ ( <code>linear</code> ) | 3.43%                              | 11.52%        | 3.62%           | 8.10%         | 2.71%           | 9.35%         | 2.39%           | 7.75%         |
|           | $\beta_2$ ( <code>sqrt</code> )   | 5.75%                              | 7.89%         | 8.34%           | 5.39%         | 11.34%          | 6.23%         | 10.03%          | 6.35%         |
|           | Weighting algorithm               | 1.40%                              | 2.04%         | 3.13%           | 3.04%         | 3.04%           | 2.60%         | 2.09%           | 4.16%         |
| 30D       | Bandwidth heuristic               | 1.05%                              | <b>54.42%</b> | 3.10%           | <b>45.27%</b> | 4.67%           | <b>40.70%</b> | 7.02%           | <b>37.59%</b> |
|           | $\Delta$ (factor of $b_{\min}$ )  | <b>39.11%</b>                      | 12.94%        | <b>44.11%</b>   | <b>20.67%</b> | <b>44.83%</b>   | <b>25.66%</b> | <b>45.65%</b>   | <b>27.45%</b> |
|           | Exponent for $b_{\text{magic}}$   | <b>47.92%</b>                      | <b>16.94%</b> | <b>41.19%</b>   | 16.66%        | <b>38.49%</b>   | 15.65%        | <b>37.06%</b>   | 14.72%        |
|           | Splitting algorithm               | 0.17%                              | 1.24%         | 0.76%           | 2.05%         | 1.04%           | 2.31%         | 1.21%           | 2.13%         |
|           | $\beta_1$ ( <code>linear</code> ) | 4.70%                              | 4.83%         | 3.48%           | 7.71%         | 2.62%           | 7.83%         | 1.82%           | 7.33%         |
|           | $\beta_2$ ( <code>sqrt</code> )   | 6.76%                              | 7.07%         | 6.88%           | 5.01%         | 7.66%           | 4.05%         | 6.11%           | 5.63%         |
|           | Weighting algorithm               | 0.28%                              | 2.56%         | 0.48%           | 2.63%         | 0.69%           | 3.79%         | 1.13%           | 5.15%         |

**Minimum Bandwidth Factor  $\Delta$ :** this factor is the second most important parameter for the benchmark functions and small factors  $\Delta = 0.01, 0.03$  are effective due to the intrinsic cardinality as discussed in Section 3.3.4. Although the factor  $\Delta$  is not a primarily important parameter for the HPO benchmarks, discrete search spaces may require a large  $\Delta$ . Note, however, that  $\Delta = 0.3$  is too large in our setups. Although the varying noise level depending on tasks makes it impossible to generalize the optimal  $\Delta$ , we recommend using  $\Delta = 0.03$ .

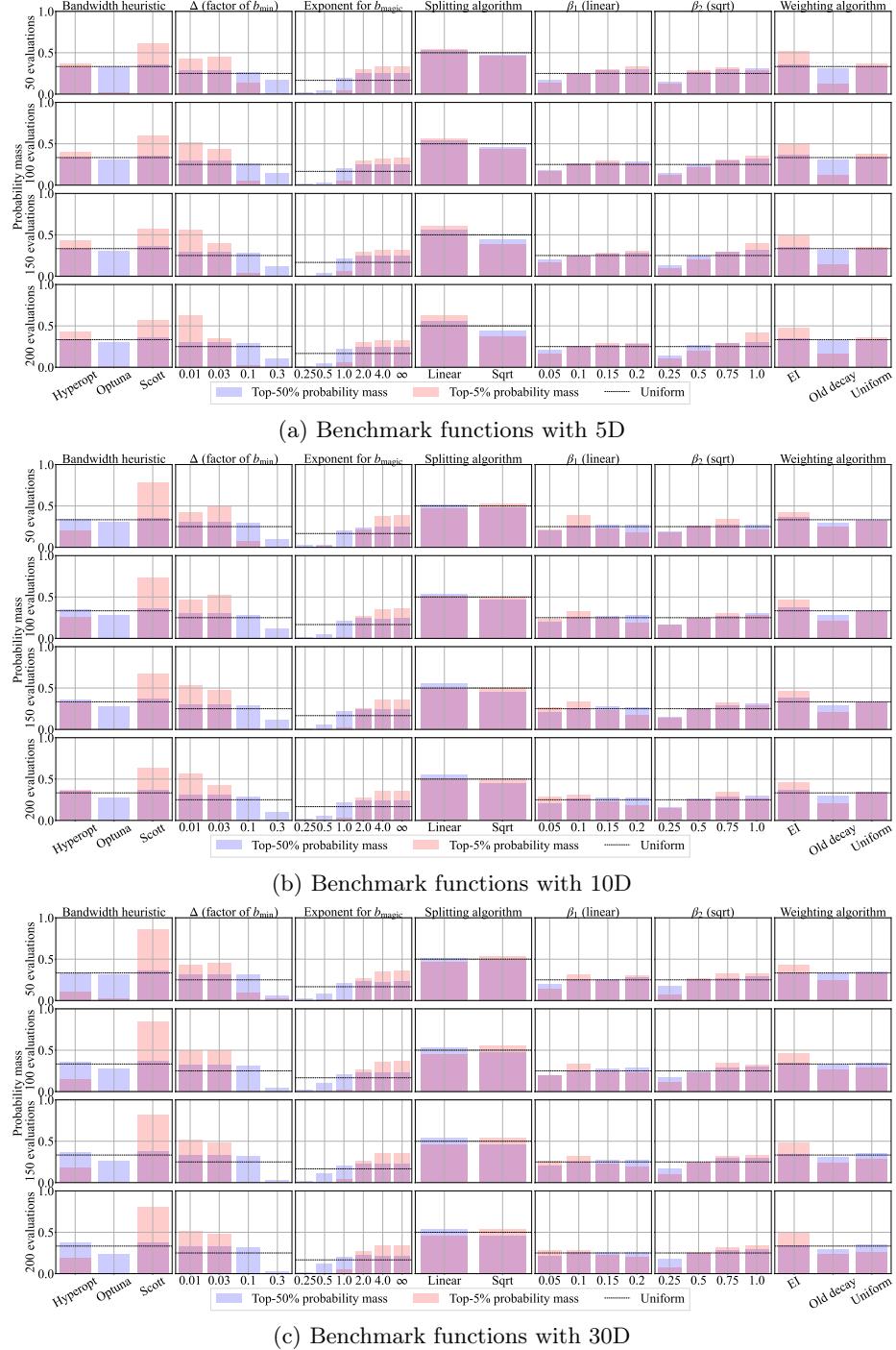


Figure 13: The probability mass values of each control parameter for the bandwidth selection in the top-5% and top-50% observations at  $\{50, 100, 150, 200\}$  evaluations on the benchmark functions. High probability mass in a specific value implies that we are likely to yield top-5% or top-50% performance with the specific value.

**Bandwidth Selection Heuristic:** according to the top-5% probability mass, an appropriate bandwidth selection heuristic varies across tasks. The best choice is `scott` for the benchmark functions, `optuna` for HPOBench and HPOlib, and `hyperopt` for JAHS-Bench-201,

## A TUTORIAL OF TREE-STRUCTURED PARZEN ESTIMATOR

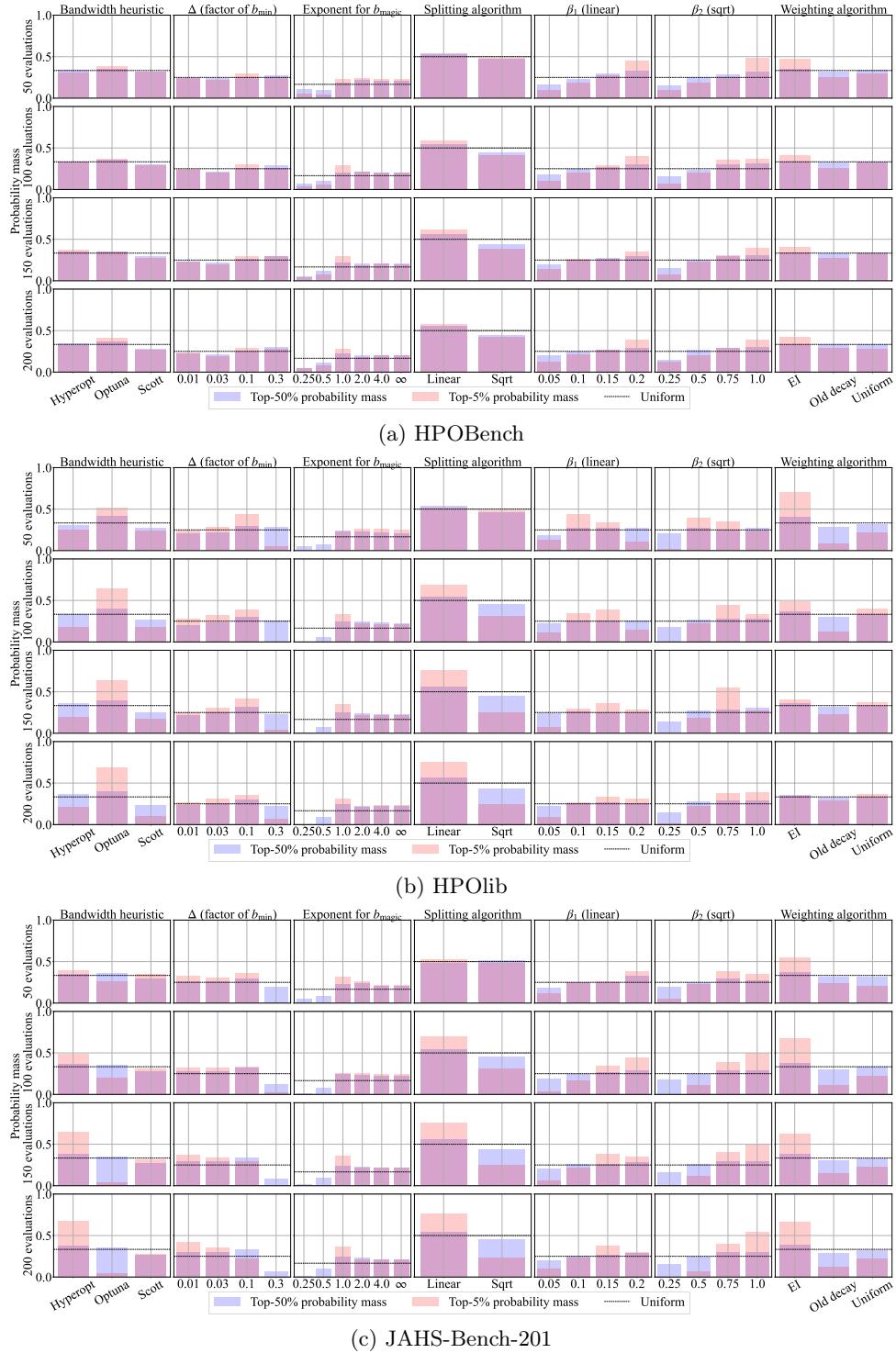


Figure 14: The probability mass values of each control parameter for the bandwidth selection in the top-5% and top-50% observations at  $\{50, 100, 150, 200\}$  evaluations on the HPO benchmarks. High probability mass in a specific value implies that we are likely to yield top-5% or top-50% performance with the specific value.

respectively. Interestingly, `optuna` rarely achieves the top-5% performance on search spaces with continuous parameters. Indeed, this finding relates to the discussion about the minimum

Table 8: The hyperparameter importance (HPI) of each control parameter for the bandwidth selection on the HPO benchmarks to achieve top 50% and to improve to top 5% from top 50%. HPI is measured at  $\{50, 100, 150, 200\}$  evaluations. We bold the top-2 HPI. Note that while the HPI in  $\beta$  quantifies whether variating  $\beta$  given either `linear` or `sqrt` matters, that in “Splitting algorithm” quantifies whether switching between `linear` and `sqrt` matters.

| Benchmark      | Control parameter                 | The number of function evaluations |               |                 |               |                 |               |                 |               |
|----------------|-----------------------------------|------------------------------------|---------------|-----------------|---------------|-----------------|---------------|-----------------|---------------|
|                |                                   | 50 evaluations                     |               | 100 evaluations |               | 150 evaluations |               | 200 evaluations |               |
|                |                                   | Top 50%                            | Top 5%        | Top 50%         | Top 5%        | Top 50%         | Top 5%        | Top 50%         | Top 5%        |
| HPOBench       | Bandwidth heuristic               | 1.35%                              | 7.00%         | 2.05%           | 7.67%         | 2.94%           | 5.98%         | 5.05%           | 6.42%         |
|                | $\Delta$ (factor of $b_{\min}$ )  | 4.16%                              | 8.54%         | 7.25%           | 8.82%         | 7.29%           | 7.57%         | 8.28%           | 8.07%         |
|                | Exponent for $b_{\text{magic}}$   | <b>36.61%</b>                      | <b>21.51%</b> | <b>45.66%</b>   | <b>22.88%</b> | <b>48.82%</b>   | <b>23.75%</b> | <b>50.08%</b>   | <b>25.10%</b> |
|                | Splitting algorithm               | 1.98%                              | 4.60%         | 4.10%           | 7.42%         | 6.01%           | 3.99%         | 5.14%           | 1.39%         |
|                | $\beta_1$ ( <code>linear</code> ) | <b>28.42%</b>                      | 14.47%        | 14.35%          | 19.82%        | 9.49%           | 17.06%        | 7.49%           | 18.82%        |
|                | $\beta_2$ ( <code>sqrt</code> )   | 26.51%                             | <b>33.87%</b> | <b>25.72%</b>   | <b>25.85%</b> | <b>25.00%</b>   | <b>33.13%</b> | <b>23.51%</b>   | <b>26.16%</b> |
|                | Weighting algorithm               | 0.97%                              | 10.01%        | 0.87%           | 7.55%         | 0.45%           | 8.53%         | 0.45%           | 14.04%        |
| HPOlib         | Bandwidth heuristic               | <b>9.53%</b>                       | 3.97%         | 5.83%           | 13.15%        | 7.09%           | <b>16.80%</b> | 9.83%           | <b>27.77%</b> |
|                | $\Delta$ (factor of $b_{\min}$ )  | 8.56%                              | 17.98%        | 5.52%           | <b>16.57%</b> | 5.51%           | 13.77%        | 6.27%           | 14.28%        |
|                | Exponent for $b_{\text{magic}}$   | <b>58.61%</b>                      | 8.98%         | <b>73.31%</b>   | 5.68%         | <b>65.36%</b>   | 7.53%         | <b>62.96%</b>   | 8.18%         |
|                | Splitting algorithm               | 2.24%                              | 0.29%         | 1.73%           | 4.78%         | 3.44%           | 10.45%        | 3.95%           | 12.01%        |
|                | $\beta_1$ ( <code>linear</code> ) | 7.80%                              | 20.76%        | 3.10%           | 13.24%        | 1.02%           | 13.25%        | 1.17%           | 12.28%        |
|                | $\beta_2$ ( <code>sqrt</code> )   | 6.86%                              | <b>22.93%</b> | <b>8.68%</b>    | <b>37.13%</b> | <b>16.79%</b>   | <b>28.33%</b> | <b>15.37%</b>   | <b>18.11%</b> |
|                | Weighting algorithm               | 6.39%                              | <b>25.10%</b> | 1.83%           | 9.46%         | 0.78%           | 9.86%         | 0.45%           | 7.37%         |
| JAHS-Bench-201 | Bandwidth heuristic               | 5.82%                              | 7.53%         | 5.43%           | 8.66%         | 5.68%           | <b>26.17%</b> | 4.58%           | <b>23.71%</b> |
|                | $\Delta$ (factor of $b_{\min}$ )  | 7.40%                              | <b>21.60%</b> | <b>17.52%</b>   | 7.27%         | <b>28.02%</b>   | 6.48%         | <b>33.17%</b>   | 7.55%         |
|                | Exponent for $b_{\text{magic}}$   | <b>61.77%</b>                      | <b>22.37%</b> | <b>58.50%</b>   | 6.91%         | <b>47.60%</b>   | 9.09%         | <b>44.72%</b>   | 7.32%         |
|                | Splitting algorithm               | 0.32%                              | 2.36%         | 1.64%           | 4.82%         | 2.93%           | 8.21%         | 1.63%           | 8.05%         |
|                | $\beta_1$ ( <code>linear</code> ) | <b>14.95%</b>                      | 13.60%        | 6.21%           | 17.50%        | 3.31%           | 10.23%        | 3.01%           | 7.99%         |
|                | $\beta_2$ ( <code>sqrt</code> )   | 7.59%                              | 17.91%        | 8.77%           | <b>33.75%</b> | 10.17%          | <b>25.37%</b> | 10.64%          | <b>29.95%</b> |
|                | Weighting algorithm               | 2.15%                              | 14.63%        | 1.93%           | <b>21.10%</b> | 2.30%           | 14.46%        | 2.25%           | 15.42%        |

bandwidth. For example,  $b$  calculated by `optuna` takes about  $(R - L)/10 \sim (R - L)/5$  based on Eq. (24), matching the poor performance with the minimum bandwidth factor  $\Delta \geq 0.1$ . On the other hand, `optuna`, which enforces  $b \geq (R - L)/10$ , outperforms the other heuristics on discrete search spaces in HPOBench and HPOlib, coinciding with the observations at the top-5% probability mass for the factor  $\Delta$  peaking at  $\Delta = 0.1$ . For `scott`, zero bandwidths are often observed for discrete parameters, suggesting tuning the factor  $\Delta$  carefully for noisy problems when using `scott`. While both `scott` and `optuna` exhibit clear drawbacks, `hyperopt` shows the most stable performance thanks to a flexible handling of the intrinsic cardinality, making our recommendation `hyperopt` by default.

**Exponent  $\alpha$  for  $b_{\text{magic}}$ :** As discussed in Section 4.1.2, a large  $\alpha$ , which leads to a small  $b_{\text{magic}}$ , and a small  $\alpha$ , which leads to a large  $b_{\text{magic}}$ , are effective for the benchmark functions and the HPO benchmarks, respectively. We recommend  $\alpha = 2.0$ , which is the compromise for both the benchmark functions and the HPO benchmarks based on the results. However, the exponent for  $b_{\text{magic}}$  is the most important parameter, so careful tuning is still necessary.

**Splitting and Weighting Algorithms (`gamma`, `weights`):** the conclusion for these parameters does not change largely from Section 4.1.2 except `linear` with  $\beta_1 = 0.15$  generalizes the most.

**Ablation Study Summary:** to sum up the results of the ablation study for the bandwidth selection, our recommendation is to use:

- `hyperopt` by default, `scott` for noiseless objectives, and `optuna` for discrete spaces,
- $\Delta = 0.03$  by default, small  $\Delta$  (0.01 or 0.03) for noiseless objectives, and large  $\Delta$  (0.03 or 0.1) for noisy objectives,
- $\alpha = 2$  by default,  $\alpha = \infty$  and  $\alpha = 1$  for noiseless and noisy objective, respectively, and
- `linear` with  $\beta_1 = 0.15$  (or `sqrt` with  $\beta_2 = 0.75$ ) and `weights=EI`.

The next section validates the performance of our recommended setting.

### 4.3 Comparison with Baseline Methods

This section discusses the performance improvement by our recommended setting against various BO methods. To generalize the assessment, our recommended setting is evaluated not only on the problem set used in the earlier sections but also on an unused problem set.

#### 4.3.1 Setup

The following baseline methods are used:

- `BORE` (Tiao et al., 2021): a classifier-based BO method inspired by TPE,
- `HEBO`<sup>6</sup> (Cowen-Rivers et al., 2022): the winner solution of the black-box optimization challenge 2020 (Turner et al., 2021),
- `Random search` (Bergstra & Bengio, 2012): the most basic baseline method in HPO,
- `SMAC`<sup>7</sup> (Hutter et al., 2011; Lindauer et al., 2022): a random forest-based BO method widely used in practice, and
- `TurBO`<sup>8</sup> (Eriksson et al., 2019): a recent strong baseline method used in the black-box optimization challenge 2020.

Each package is used by its default setting. Note that since the default setting of `BORE` is not specified in the original paper (Tiao et al., 2021), we use the default setting of `BORE` in Syne Tune (Salinas et al., 2022). More specifically, XGBoost is used as a classifier model in `BORE` and the best point is picked from 500 random points. `TurBO` follows the default setting in `SMAC3`, which uses `TurBO-1` where `TurBO-M` refers to `TurBO` with  $M$  trust regions. One-hot encoding is applied to the categorical parameters in each benchmark for `TurBO` and `HEBO`, which use the Gaussian process. TPE uses the following control parameters:

- `multivariate=True`,
- `consider_prior=True`,
- `consider_magic_clip=True` with the exponent  $\alpha = 2.0$  for  $b_{\text{magic}}$ ,
- `gamma=linear` with  $\beta = 0.15$ ,

---

6. We used v0.3.2 in <https://github.com/huawei-noah/HEBO>.

7. We used v1.4.0 in <https://github.com/automl/SMAC3>.

8. We used the implementation by <https://github.com/uber-research/TuRBO> (Accessed on Jan 2023).

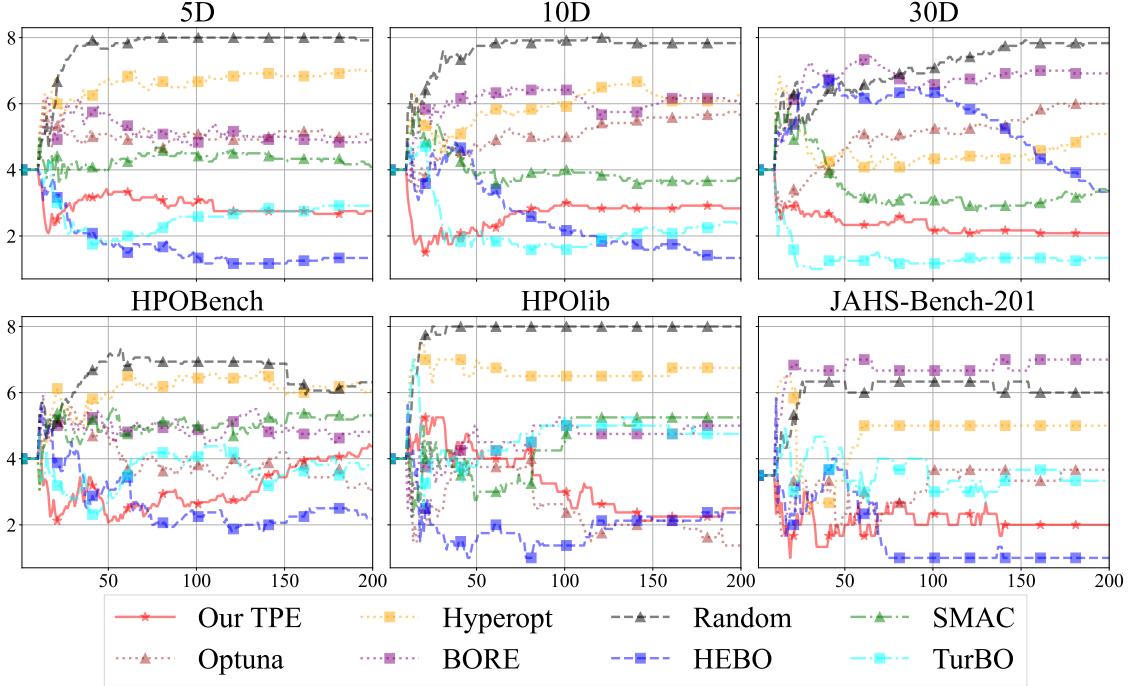


Figure 15: The average rank of each optimization method on the benchmarked task set. The initial average ranks are constant over all samplers due to the shared initial configurations. Note that SMAC are omitted for JAHS-Bench-201 due to the package dependency issue. The individual results are provided in Appendix E.

- `weights=EI`,
- `optuna` for the categorical bandwidth selection,
- `hyperopt` for the bandwidth selection heuristic, and
- the minimum bandwidth factor  $\Delta = 0.03$ .

Furthermore, we run Optuna v4.0.0 and Hyperopt v0.2.7, both of which use TPE internally, to see the improvements. The visualization uses the average rank of the median performance over 10 random seeds. To avoid the overfitting to the previously used benchmark problems, we conduct experiments using some additional benchmark functions detailed in Appendix D, LCBench in YAHPO Gym (Pfisterer et al., 2022), and the surrogate version<sup>9</sup> of Olympus (Hickman et al., 2023). Note that we call the benchmarks used in the previous sections *benchmarked task set* and the additional benchmarks *validation task set* hereafter.

#### 4.3.2 Results & Discussion

Figure 15, 16 present the average rank of each optimization method on the benchmarked and validation task sets, respectively. Most importantly, the benchmarking results show very similar performance on both benchmarked and validation task sets, implying that

9. <https://github.com/nabenabe0928/olympus-surrogate-bench/tree/main>

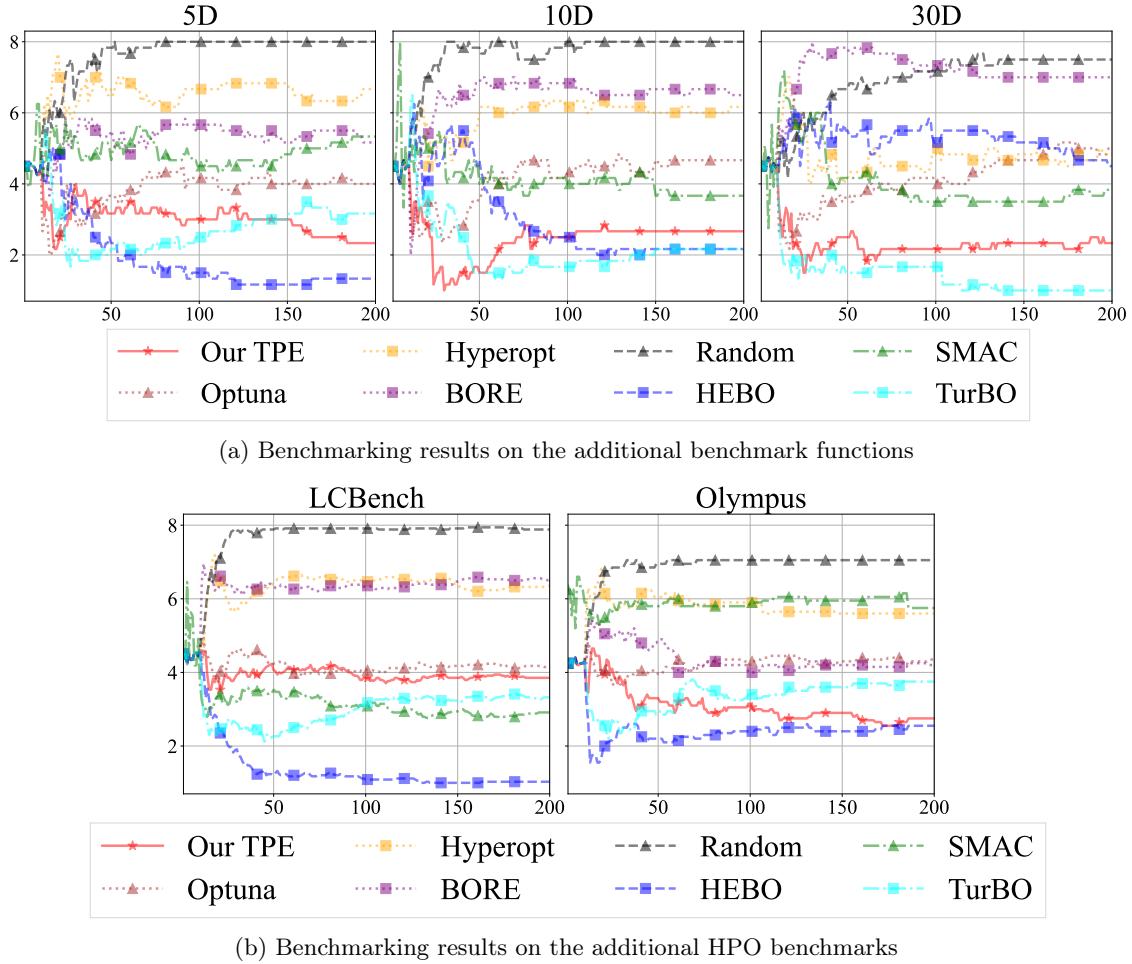


Figure 16: The average rank of each optimization method on the benchmarked task set. The initial average ranks are constant over all samplers due to the shared initial configurations.

Table 9: The qualitative summary of the results obtained from the comparison. Only HEBO, TurBO, and our TPE are listed as they exhibit the best performance on at least one of the benchmark sets. Each method is qualitatively rated for each column by  $\bigcirc$  (relatively good),  $\triangle$  (medium), and  $\times$  (relatively bad).

|         | Runtime     | Continuous (Low/High dimensional)<br>Low ( $\sim 15D$ ) | Continuous (Low/High dimensional)<br>High ( $15D \sim$ ) | Discrete (W/O categorical)<br>With | Discrete (W/O categorical)<br>Without |
|---------|-------------|---------------------------------------------------------|----------------------------------------------------------|------------------------------------|---------------------------------------|
| Our TPE | $\bigcirc$  | $\triangle$                                             | $\triangle$                                              | $\bigcirc$                         | $\triangle$                           |
| HEBO    | $\times$    | $\bigcirc$                                              | $\times$                                                 | $\bigcirc$                         | $\bigcirc$                            |
| TurBO   | $\triangle$ | $\triangle$                                             | $\bigcirc$                                               | $\times$                           | $\triangle$                           |

our recommendation setting is robust to diverse tasks. Furthermore, our TPE consistently outperforms the Hyperopt TPE, which is used by many research papers, and the Optuna TPE both on the benchmarked and validation task sets. Notice, however, that the Optuna TPE

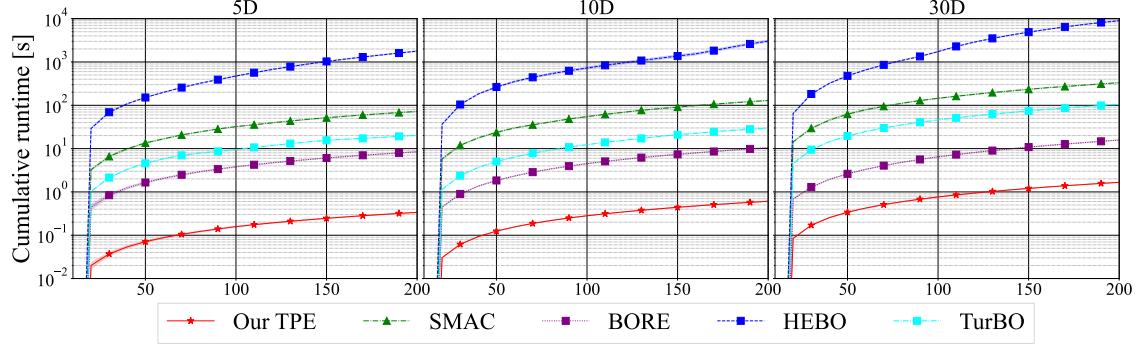


Figure 17: The cumulative runtime of each optimization method on the sphere function with different dimensionalities (5D for **Left**, 10D for **Center**, and 30D for **Right**). The horizontal axis is the number of configuration evaluations and the weak-color bands show the standard error over 10 independent runs. Note that we used 8 cores of Intel Core i7-10700.

outperforms our TPE on discrete search spaces. The high performance of the Optuna TPE on discrete search spaces is already discussed in ‘‘Bandwidth Selection Heuristic’’ of Section 4.2.2. Although the Optuna TPE is better than our TPE on discrete search spaces, the performance of our TPE still remains comparable, suggesting that our recommendation setting is preferred. With that being said, our TPE is not a silver bullet. Indeed, HEBO consistently outperforms our TPE except for the 30D problems and TurBO exhibits very strong performance on the benchmark functions. Meanwhile, our TPE is much quicker compared to HEBO, which takes 2.5 hours to sample 200 configurations on the 30D problems, as shown in Figure 17. As the computational overhead of an objective varies, higher performance does not necessarily mean one method is better than the others. Hence, practitioners must consider which optimization method to use based on their specific use cases. For example, if the evaluation of an objective takes only a few seconds, HEBO is probably not an appropriate choice as the sampling overhead dominates the evaluation overhead. The qualitative evaluations of HEBO, TurBO, and our TPE are summarized in Table 9.

## 5. Conclusion

This paper provided detailed explanations of each component in the TPE algorithm and showed how we should determine each control parameter. Roughly speaking, the control parameter settings should be changed depending on how much noise objective functions have. Accounting for the noise level is especially important for the bandwidth selection algorithm owing to the intrinsic cardinality. Despite that, we provided a recommended default setting as a conclusion and compared our recommended version of TPE with various baseline methods on diverse problems. The results demonstrated that our TPE performs much better than the existing TPE-related packages such as Hyperopt and Optuna, and potentially outperforms the state-of-the-art BO methods with much shorter computational overhead. As this paper focused on the single-objective optimization setting, we did not discuss settings such as multi-objective, constrained, multi-fidelity, and batch optimization and the investigation of these settings would be our future work.

## Appendix A. The Derivation of the Acquisition Function

Watanabe and Hutter (2022, 2023), Song et al. (2022) independently prove that EI and PI are equivalent under the assumption in Eq. (4). For this reason, we only discuss PI for simplicity. We show the detail of the transformations to obtain Eq. (6). We first plug in Eq. (4) to Eq. (3) (the formulation of PI) as follows:

$$\begin{aligned}
\int_{-\infty}^{y^\gamma} p(y|\mathbf{x}, \mathcal{D}) dy &= \int_{-\infty}^{y^\gamma} \frac{p(\mathbf{x}|y, \mathcal{D})p(y|\mathcal{D})}{p(\mathbf{x}|\mathcal{D})} dy \quad (\because \text{Bayes' theorem}) \\
&= \frac{p(\mathbf{x}|\mathcal{D}^{(l)})}{p(\mathbf{x}|\mathcal{D})} \underbrace{\int_{-\infty}^{y^\gamma} p(y|\mathcal{D}) dy}_{=\gamma} \quad (\because y \in (-\infty, y^\gamma] \Rightarrow p(\mathbf{x}|y, \mathcal{D}) = p(\mathbf{x}|\mathcal{D}^{(l)})) \\
&= \frac{\gamma p(\mathbf{x}|\mathcal{D}^{(l)})}{\gamma p(\mathbf{x}|\mathcal{D}^{(l)}) + (1 - \gamma)p(\mathbf{x}|\mathcal{D}^{(g)})} \tag{19}
\end{aligned}$$

where the last transformation used the following marginalization:

$$\begin{aligned}
p(\mathbf{x}|\mathcal{D}) &= \int_{-\infty}^{\infty} p(\mathbf{x}|y, \mathcal{D})p(y|\mathcal{D}) dy \\
&= p(\mathbf{x}|\mathcal{D}^{(l)}) \int_{-\infty}^{y^\gamma} p(y|\mathcal{D}) dy + p(\mathbf{x}|\mathcal{D}^{(g)}) \int_{y^\gamma}^{\infty} p(y|\mathcal{D}) dy \tag{20} \\
&= \gamma p(\mathbf{x}|\mathcal{D}^{(l)}) + (1 - \gamma)p(\mathbf{x}|\mathcal{D}^{(g)}).
\end{aligned}$$

## Appendix B. Related Work of Tree-Structured Parzen Estimator

While this paper focuses solely on the single-objective setting, strict generalizations with multi-objective (MOTPE) (Ozaki et al., 2020, 2022b) and constrained optimization (c-TPE) (Watanabe & Hutter, 2023) settings are available. Namely, MOTPE and c-TPE are identical to the original TPE when the number of objectives is 1 and the violation probability is almost everywhere zero, i.e.,  $\mathbb{P}[\bigcap_{i=1}^K c_i \leq c_i^* | \mathbf{x}] = 0$  for almost all <sup>10</sup>  $\mathbf{x} \in \mathcal{X}$ , respectively. TPE has been adapted to the meta-learning, multi-fidelity, and combinatorial optimization settings as well. Falkner et al. (2018) extend TPE to the multi-fidelity setting by combining TPE and Hyperband (Li et al., 2017b). Watanabe et al. (2023a) introduce meta-learning by considering task similarity via the overlap of promising domains. Abe et al. (2025) introduce a categorical kernel for TPE to handle categorical parameters more efficiently.

Furthermore, BORE (Tiao et al., 2021) is inspired by TPE. BORE replaces the density ratio with a classifier model based on the fact that TPE evaluates the promise of configurations by binary classification. Song et al. (2022) and Oliveira et al. (2022) build some theories on top of BORE. Song et al. (2022) formally derive the expected improvement for a classifier-based surrogate. While TPE and BORE train a binary classifier with uniform weights for each sample, the expected improvement requires a binary classifier with weights proportional to the improvement from a given threshold. Oliveira et al. (2022) provide a theoretical

---

10. More formally, almost everywhere zero is defined by  $\mu(\{\mathbf{x} \in \mathcal{X} | \mathbb{P}[\bigcap_{i=1}^K c_i \leq c_i^* | \mathbf{x}] = 0\}) = 0$  where  $\mu : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$  is the Lebesgue measure.

analysis of regret using the Gaussian process classifier. In contrast to BORE, TPE has almost no theories due to the multimodality nature of KDE and the explicit handling of density ratio. Watanabe et al. (2023a) show that  $\mathcal{D}^{(l)}$  asymptotically converges to a set of the  $\gamma'$ -quantile ( $\gamma' < \gamma$ ) configurations if the objective function  $f$  does not have noise, we use the  $\epsilon$ -greedy algorithm in Line 13, and we use a fixed  $\gamma$ . To the best of our knowledge, this is the only theory available for TPE. This analysis suggests picking a random configuration once in a while, which is, in principle, the  $\epsilon$ -greedy algorithm. However, it is not realistic to use the  $\epsilon$ -greedy algorithm in practice due to the severely limited computational budget.

## Appendix C. More Details of Kernel Density Estimator

This section describes the implementation details of kernel density estimator in each package.

### C.1 Uniform Weighting Algorithm in BOHB

Suppose we have  $\{x_n\}_{n=1}^N$  where  $x_n \in [L, R]$ , the KDE in BOHB is computed as follows:

$$\sum_{n=1}^N \frac{z_n}{\sum_{i=1}^N z_i} g(x, x_n | b). \quad (21)$$

Recall that  $g$  is defined in Eq. (12) and  $z_n = \int_L^R g(x, x_n) dx$ . In principle, this formulation allows to flatten PDFs even when the observations concentrate near a boundary. Note that this formulation is not implemented in our experiments.

### C.2 group Parameter in Optuna

In this section, we use  $x_{\text{null}}$  as a placeholder for an undefined value and  $\mathbb{R}_{\text{null}} := \mathbb{R} \cup \{x_{\text{null}}\}$  for convenience. We first formally define the tree-structured search space. Suppose  $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_D$  is a  $D$ -dimensional search space with  $\mathcal{X}_d \subseteq \mathbb{R}_{\text{null}}$  for each  $d \in [D] := \{1, \dots, D\}$  and the  $d_i$ -th ( $d_i \in [D]$ ) dimension is conditional for each  $i \in [D_c]$  where  $D_c (< D)$  is the number of conditional parameters. Furthermore, assume binary functions  $c_{d_i} : \prod_{d \neq d_i} \mathcal{X}_d \rightarrow \{0, 1\}$  for each  $i \in [D_c]$  are given. For example, consider the search space with (1) the number of layers  $x_1 := L \in [3] = \mathcal{X}_1$ , and (2) the dropout rates at the  $l$ -th layer  $x_{l+1} := p_l \in [0, 1] = \mathcal{X}_{l+1}$  for  $l \in [3]$ , then  $x_3$  and  $x_4$  are the conditional parameters in this search space. The binary functions for this search space are  $c_3(x_1, x_2, x_4) = \mathbb{I}[x_1 \geq 2] = \mathbb{I}[L \geq 2]$  and  $c_4(x_1, x_2, x_3) = \mathbb{I}[x_1 \geq 3] = \mathbb{I}[L \geq 3]$ . Note that the dropout rate at the  $l$ -th layer will not be defined if there are no more than  $l$  layers. Another example is the following search space:

- The number of layers  $x_1 := L \in [2] = \mathcal{X}_1$ ,
- The optimizer at the 1st layer  $x_2 \in \{\text{adam}, \text{sgd}\} = \mathcal{X}_2$ <sup>11</sup>,
- The coefficient  $\beta_1 \in (0, 1) = \mathcal{X}_3$  for `adam` in the 1st layer,
- The coefficient  $\beta_2 \in (0, 1) = \mathcal{X}_4$  for `adam` in the 1st layer,
- The momentum  $m \in (0, 1) = \mathcal{X}_5$  for `sgd` in the 1st layer,
- The optimizer at the 2nd layer  $x_6 \in \{\text{adam}, \text{sgd}\} = \mathcal{X}_6$ ,

11. See <https://pytorch.org/docs/stable/generated/torch.optim.Adam.html> for `adam` and <https://pytorch.org/docs/stable/generated/torch.optim.SGD.html> for `sgd`.

- The coefficient  $\beta_1 \in (0, 1) = \mathcal{X}_7$  for `adam` in the 2nd layer,
- The coefficient  $\beta_2 \in (0, 1) = \mathcal{X}_8$  for `adam` in the 2nd layer, and
- The momentum  $m \in (0, 1) = \mathcal{X}_9$  for `sgd` in the 2nd layer.

All the parameters except  $x_1, x_2$  are conditional in this example, making the binary functions for each dimension,  $c_3 = c_4 = \mathbb{I}[x_2 = \text{adam}]$ ,  $c_5 = \mathbb{I}[x_2 = \text{sgd}]$ ,  $c_6 = \mathbb{I}[x_1 \geq 2]$ ,  $c_7 = c_8 = \mathbb{I}[x_1 \geq 2]\mathbb{I}[x_6 = \text{adam}]$ , and  $c_9 = \mathbb{I}[x_1 \geq 2]\mathbb{I}[x_6 = \text{sgd}]$ . As can be seen,  $x_7, x_8, x_9$  require  $x_6$  to be defined and  $x_6$  requires  $x_1$  to be no less than 2. This hierarchical structure, i.e.,  $x_7, x_8, x_9$  require  $x_6$  and, in turn,  $x_1$  as well to be defined, is the exact reason why we call search spaces with conditional parameters tree-structured. Strictly speaking,  $x_7, x_8, x_9$  cannot be provided to  $c_6$ , but they are unnecessary for  $c_6$  thanks to the tree structure. Hence, we simply pad  $x_7, x_8, x_9$  with a placeholder  $x_{\text{null}}$ .

Using the second example above, we will explain the `group` parameter in Optuna. First, we define a set of dimensions as  $s \subseteq [D]$  and a subspace of  $\mathcal{X}$  that takes the dimensions specified in  $s$  as  $\mathcal{X}_s := \prod_{d \in s} \mathcal{X}_d$ . Then `group` enumerates all possible  $\mathcal{X}_s$  based on a set of observations  $\mathcal{D}$  and optimizes the acquisition function separately in each subspace. For example, the second example above could take  $\mathcal{X}_{\{1, 2, 3, 4\}}$ ,  $\mathcal{X}_{\{1, 2, 5\}}$ ,  $\mathcal{X}_{\{1, 2, 3, 4, 6, 7, 8\}}$ ,  $\mathcal{X}_{\{1, 2, 3, 4, 6, 9\}}$ ,  $\mathcal{X}_{\{1, 2, 5, 6, 7, 8\}}$ , and  $\mathcal{X}_{\{1, 2, 5, 6, 9\}}$ , as subspaces when we ignore dimensions filled with a placeholder  $x_{\text{null}}$ . The enumeration of the subspaces can be easily reproduced by checking missing values in each observation from  $\mathcal{D}$  with the time complexity of  $O(DN)$ . Then we perform a sampling by the TPE algorithm in each subspace using the observations that belong exactly to  $\mathcal{X}_s$ . In other words, when we have an observation  $\{2, \text{sgd}, x_{\text{null}}, x_{\text{null}}, 0.5, \text{sgd}, x_{\text{null}}, x_{\text{null}}, 0.5\}$ , we use this observation only for the sampling in  $\mathcal{X}_{\{1, 2, 5, 6, 9\}}$ , but not for that in  $\mathcal{X}_{\{1, 2, 5\}}$ .

### C.3 Bandwidth Selection

Throughout this section, the bandwidth of KDE is assumed to be computed based on a set of observations  $\{x_n\}_{n=1}^N \in [L, R]^N$  where  $x_n$  is sorted such that  $x_1 \leq x_2 \leq \dots \leq x_N$  holds. If the prior is included, we simply include  $x = (L + R)/2$  in the observations. Note that all methods select the bandwidth for each dimension independently.

#### C.3.1 Hyperopt Implementation

When `consider_endpoints=True`, we first augment the observations as  $\{x_n\}_{n=0}^{N+1}$  where  $x_0 = L, x_{N+1} = R$ ; otherwise, we just use  $\{x_n\}_{n=1}^N$ . Then the bandwidth  $b_n$  for the  $n$ -th kernel function  $k(\cdot, x_n | b_n)$  ( $n = 1, 2, \dots, N$ ) is computed as follows:

$$b_n := \begin{cases} x_n - x_{n-1} & (\text{if } x_{n+1} \text{ not exist}) \\ x_{n+1} - x_n & (\text{if } x_{n-1} \text{ not exist}) \\ \max(\{x_{n+1} - x_n, x_n - x_{n-1}\}) & (\text{otherwise}) \end{cases}. \quad (22)$$

This heuristic adapts the bandwidth depending on the concentration of observations. Namely, the bandwidth will be wider at sparse regions and narrower at dense regions, respectively.

#### C.3.2 BOHB Implementation (Scott's Rule)

Scott's rule calculates bandwidth for the univariate Gaussian kernel as follows:

$$b = \left( \frac{4}{3N} \right)^{1/5} \min \left( \sigma, \frac{\text{IQR}}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)} \right) \simeq 1.059 N^{-1/5} \min \left( \sigma, \frac{\text{IQR}}{1.34} \right) \quad (23)$$

Table 10: The list of the benchmark functions. Each function can take an arbitrary dimension  $D \in \mathbb{Z}_+$ . The right column shows the domain of each function. One of the dimensions in  $D \in \{5, 10, 30\}$  is used in the experiments.

| Name            | $f(\mathbf{x})$ ( $\mathbf{x} := [x_1, x_2, \dots, x_D]$ )                                                                                                            | $ x_d  \leq R$ |
|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
| Ackley          | $\exp(1) + 20 \left( 1 - \exp \left( -\frac{1}{5} \sqrt{\frac{1}{D} \sum_{d=1}^D x_d^2} \right) \right) - \exp \left( \frac{1}{D} \sum_{d=1}^D \cos 2\pi x_d \right)$ | 32.768         |
| Griewank        | $1 + \frac{1}{4000} \sum_{d=1}^D x_d^2 - \prod_{d=1}^D \cos \frac{x_d}{\sqrt{d}}$                                                                                     | 600            |
| K-Tablet        | $\sum_{d=1}^K x_d^2 + \sum_{d=K+1}^D (100x_d)^2$ where $K = \lceil D/4 \rceil$                                                                                        | 5.12           |
| Levy            | $\sin^2 \pi w_1 + \sum_{d=1}^{D-1} (w_d - 1)^2 (1 + 10 \sin^2(\pi w_d + 1)) + (w_D - 1)^2 (1 + \sin^2 2\pi w_D)$ where $w_d = 1 + \frac{x_d - 1}{4}$                  | 10             |
| Perm            | $\sum_{d_1=1}^D \left( \sum_{d_2=1}^D (d_2 + 1) (x_{d_2}^{d_1} - \frac{1}{d_2^{d_1}}) \right)^2$                                                                      | 1              |
| Rastrigin       | $10D + \sum_{d=1}^D (x_d^2 - 10 \cos 2\pi x_d)$                                                                                                                       | 5.12           |
| Rosenbrock      | $\sum_{d=1}^{D-1} \left( 100(x_{d+1} - x_d^2)^2 + (x_d - 1)^2 \right)$                                                                                                | 5              |
| Schwefel        | $-\sum_{d=1}^D x_d \sin \sqrt{ x_d }$                                                                                                                                 | 500            |
| Sphere          | $\sum_{d=1}^D x_d^2$                                                                                                                                                  | 5              |
| Styblinski      | $\frac{1}{2} \sum_{d=1}^D (x_d^4 - 16x_d^2 + 5x_d)$                                                                                                                   | 5              |
| Weighted sphere | $\sum_{d=1}^D dx_d^2$                                                                                                                                                 | 5              |
| Xin-She-Yang    | $\sum_{d_1=1}^D  x_{d_1}  \exp \left( -\sum_{d_2=1}^D \sin x_{d_2}^2 \right)$                                                                                         | $2\pi$         |

where  $(4/3N)^{1/5}$  comes from Eq. (3.28) by Silverman (2018), IQR is the interquartile range of the observations,  $\sigma$  is the standard deviation of the observations, and  $\Phi : \mathbb{R} \rightarrow [0, 1]$  is the cumulative distribution of  $\mathcal{N}(0, 1)$ . BOHB calculates the bandwidth for each dimension separately and calculates the bandwidth for categorical parameters as if they are numerical parameters.

### C.3.3 Optuna v4.0.0 Implementation

The bandwidth is computed in Optuna v4.0.0 as follows:

$$b = \frac{R - L}{5} N^{-1/(D+4)} \quad (24)$$

where  $D$  is the dimension of search space,  $N$  is the number of observations, and the target parameter is defined on  $[L, R]$ . In contrast to the other methods, the bandwidth depends only on the number of observations and the dimension of search space. The bandwidth of a categorical parameter with  $C \in \mathbb{Z}_+$  categories is determined in Optuna v4.0.0 as follows:

$$b = \frac{C - 1}{N + C} \quad (25)$$

where  $N$  is the number of observations, and the Aitchison-Aitken kernel is used.

## Appendix D. The Details of Benchmarks

Table 10 lists the 12 different benchmark functions used in the experiments with 3 different dimensionalities. Their characteristics are available in Table 11. For the HPO benchmarks,

Table 11: The characteristics of the benchmark functions.

| Name            | Characteristics            |
|-----------------|----------------------------|
| Ackley          | Multimodal                 |
| Griewank        | Multimodal                 |
| K-Tablet        | Monomodal, ill-conditioned |
| Levy            | Multimodal                 |
| Perm            | Monomodal                  |
| Rastrigin       | Multimodal                 |
| Rosenbrock      | Monomodal                  |
| Schwefel        | Multimodal                 |
| Sphere          | Convex, monomodal          |
| Styblinski      | Multimodal                 |
| Weighted sphere | Convex, monomodal          |
| Xin-She-Yang    | Multimodal                 |

Table 12: The search space of HPOBench (5 discrete parameters). HPOBench is a tabular benchmark and we can query the performance of a specified configuration from the tabular. HPOBench stores all possible configurations of an MLP in this table for 8 different OpenML datasets (`Vehicle`, `Segmentation`, `Car evaluation`, `Australian credit approval`, `German credit`, `Blood transfusion service center`, `KC1 software detect prediction`, `Phoneme`). Each parameter except “Depth” has 10 grids and the grids are taken so that each grid is equally distributed in the log-scaled range.

| Hyperparameter        | Parameter type | Range          |
|-----------------------|----------------|----------------|
| L2 regularization     | Discrete       | $[10^{-8}, 1]$ |
| Batch size            | Discrete       | $[4, 256]$     |
| Depth                 | Discrete       | $[1, 3]$       |
| Initial learning rate | Discrete       | $[10^{-5}, 1]$ |
| Width                 | Discrete       | $[16, 1024]$   |

HPOBench (Eggensperger et al., 2021) in Table 12, HPOlib (Klein & Hutter, 2019) in Table 13, and JAHS-Bench-201 (Bansal et al., 2022) in Table 14 are used. The validation task set includes LCBench in YAHPO Gym, the surrogate version of Olympus, and 6 benchmark functions. LCBench provides the seven-dimensional continuous search space for 33 different datasets and the Olympus surrogate benchmark provides 10 chemistry continuous BBO problems each with a different dimensionality. The benchmark functions used in the validation task set are different power, Dixon-Price, Langermann, Michalewicz, Powell, and Trid. HPO benchmarks have two types:

1. **tabular benchmark**, which queries pre-recorded performance metric values from a static table which is why it cannot handle continuous parameters, and

Table 13: The search space of HPOlib (6 discrete + 3 categorical parameters). HPOlib is a tabular benchmark and we can query the performance of a specified configuration from the tabular. HPOlib stores all possible configurations of an MLP in this table for 4 different datasets (**Parkinsons telemonitoring**, **Protein structure**, **Naval propulsion**, **Slice localization**).

| Hyperparameter             | Parameter type | Range                                                                                 |
|----------------------------|----------------|---------------------------------------------------------------------------------------|
| Initial learning rate      | Discrete       | $\{5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}\}$ |
| Dropout rate {1, 2}        | Discrete       | {0.0, 0.3, 0.6}                                                                       |
| Number of units {1, 2}     | Discrete       | $\{2^4, 2^5, 2^6, 2^7, 2^8, 2^9\}$                                                    |
| Batch size                 | Discrete       | $\{2^3, 2^4, 2^5, 2^6\}$                                                              |
| Learning rate scheduling   | Categorical    | {cosine, constant}                                                                    |
| Activation function {1, 2} | Categorical    | {ReLU, tanh}                                                                          |

Table 14: The search space of JAHS-Bench-201 (2 continuous + 2 discrete + 8 categorical parameters). JAHS-Bench-201 is a surrogate benchmark and it uses an XGBoost surrogate model trained on pre-evaluated configurations for 3 different datasets (**CIFAR10**, **Fashion MNIST**, **Colorectal histology**). We use the output of the XGBoost surrogate given a specified configuration as a query.

| Hyperparameter                                  | Parameter type | Range                                                           |
|-------------------------------------------------|----------------|-----------------------------------------------------------------|
| Learning rate                                   | Continuous     | $[10^{-3}, 1]$                                                  |
| L2 regularization                               | Continuous     | $[10^{-5}, 10^{-2}]$                                            |
| Depth multiplier                                | Discrete       | {1, 2, 3}                                                       |
| Width multiplier                                | Discrete       | $\{2^2, 2^3, 2^4\}$                                             |
| Cell search space                               | Categorical    | {none, avg-pool-3x3, bn-conv-1x1, bn-conv-3x3, skip-connection} |
| (NAS-Bench-201 (Dong & Yang, 2020), Edge 1 – 6) | Categorical    | {ReLU, Hardswish, Mish}                                         |
| Activation function                             | Categorical    | {True, False}                                                   |
| Trivial augment (Müller & Hutter, 2021)         | Categorical    | {True, False}                                                   |

2. **surrogate benchmark**, which returns the predicted performance metric values by a surrogate model for the corresponding HP configurations.

HPOlib and HPOBench are tabular benchmarks and JAHS-Bench-201 is a surrogate benchmark. The visualization for Figures 11 – 14 uses the methodologies invented by Watanabe et al. (2023b). Assume there are  $K$  possible combinations of control parameters and a set of observations  $\{(\mathbf{x}_{k,n}^{(m)}, y_{k,n}^{(m)})\}_{n=1}^N$  are obtained on the  $m$ -th benchmark ( $m = 1, \dots, M$ ) with the  $k$ -th possible set of the control parameters  $\boldsymbol{\theta}_k$ .  $\boldsymbol{\theta}_k$  ( $k = 1, \dots, K$ ) is one of the possible sets of the control parameters specified in Table 3 and  $N = 200$  is used in the experiments. Then we collect a set of results  $\mathcal{R}_n^{(m)} := \{(\boldsymbol{\theta}_k, \zeta_{k,n}^{(m)})\}_{k=1}^K$  with the budget of  $n$  where  $\zeta_{k,n}^{(m)} := \min_{n' \leq n} y_{k,n'}^{(m)}$  and  $n \in \{50, 100, 150, 200\}$  are used in the experiments. Furthermore, we define  $i_{m,k}$  ( $k = 1, \dots, K$ ) such that  $\zeta_{i_{m,1},n}^{(m)} \leq \zeta_{i_{m,2},n}^{(m)} \leq \dots \leq \zeta_{i_{m,K},n}^{(m)}$ . The visualization of the probability mass functions performs the following:

1. Pick a top-performance quantile  $\alpha$  (in our case,  $\alpha = 0.05, 0.5$ ),

2. Extract the top- $\alpha$  quantile observations  $\{(\boldsymbol{\theta}_{i_{m,k}}, \zeta_{i_{m,k},n}^{(m)})\}_{k=1}^{\lceil \alpha K \rceil}$ ,
3. Build 1D KDEs  $p_d^{(m)}(\theta_d) := \sum_{k=1}^{\lceil \alpha N \rceil} k(\theta_d, \theta_{i_{m,k},d})$  for each control parameter,
4. Compute the mean of the KDEs from all the tasks  $\bar{p}_d := 1/M \sum_{m=1}^M p_d^{(m)}(\theta_d)$ ,
5. Plot the probability mass function of the mean  $\bar{p}_d$  of the KDEs.

The hyperparameter importance (HPI) is computed by PED-ANOVA (Watanabe et al., 2023b). The top-50% HPI captures the importance of each control parameter to achieve the top-50% performance (global HPI) while the top-5% HPI captures the importance of each control parameter to achieve the top-5% performance from the top-50% performance (local HPI).

## Appendix E. Additional Results

This section provides additional results. The visualization of figures performs the following operations (see the notations invented in Section 4.1.1):

1. Fix a control parameter (e.g., `multivariate=True`),
2. Gather all the cumulative minimum performance curves  $\{\{\zeta_{k,n}^{(m)}\}_{k=1}^K\}_{n=1}^N$  where  $N = 200$  was used in the experiments,
3. Plot the mean value  $1/K \sum_{k=1}^K \zeta_{k,n}^{(m)}$  over the gathered results at each “ $n$  evaluations” ( $n = 1, \dots, N$ ) as a solid line,
4. Plot vertically the (violin-plot-like) distribution of the gathered results  $\{\zeta_{k,n}^{(m)}\}_{k=1}^K$  at  $n \in \{50, 100, 150, 200\}$  evaluations as a transparent shadow, and
5. Repeat Operations 1.–4. for all settings.

The violin-plot-like distributions are used in the visualization instead of the standard error to compensate for the lack of distributional information. Each result includes optimizations by Optuna v4.0.0 as a baseline. The following results are presented:

1. **Ablation Study:** Figures 18–21 present the results on the benchmark functions and Figures 22,23 present the results on the HPO benchmarks.
2. **Analysis of Bandwidth Selection:** Figures 24–35 present the results on the benchmark functions and Figures 36–39 present the results on the HPO benchmarks.
3. **Comparison with Baseline Methods:** Figures 40–43 present the results on the benchmark functions and Figure 44 presents the results on the HPO benchmarks.

## Appendix F. General Advice for Hyperparameter Optimization

This section briefly discusses simple strategies to effectively design search spaces for HPO. There are several tips to design search spaces well:

1. reduce or bundle hyperparameters as much as possible,
2. include strong baseline settings in initial configurations,

3. use ordinal parameters instead of continuous parameters,
4. consider other possible optimization algorithms,
5. restart optimization after a certain number of evaluations.

The first point is essential due to the curse of dimensionality in high dimensions. For example, when hyperparameter configurations of neural networks are considered, it is simpler to bundle hyperparameters for each layer such as dropout rate and activation functions. Such design significantly reduces the dimensionality. The second point is to include strong baselines if available. This follows the basic principle of warm-starting methods (Feurer et al., 2015; Nomura et al., 2021; Hvarfner et al., 2022), which start near known strong baselines. The third point is to define hyperparameters as ordinal parameters rather than continuous parameters according to intrinsic cardinality. Recall that intrinsic cardinality is defined in Section 3.3.4. We illustrate an example below:

---

```

low, high = 0.0, 1.0
Definition as a continuous parameter
dropout = trial.suggest_float("dropout", low, high)

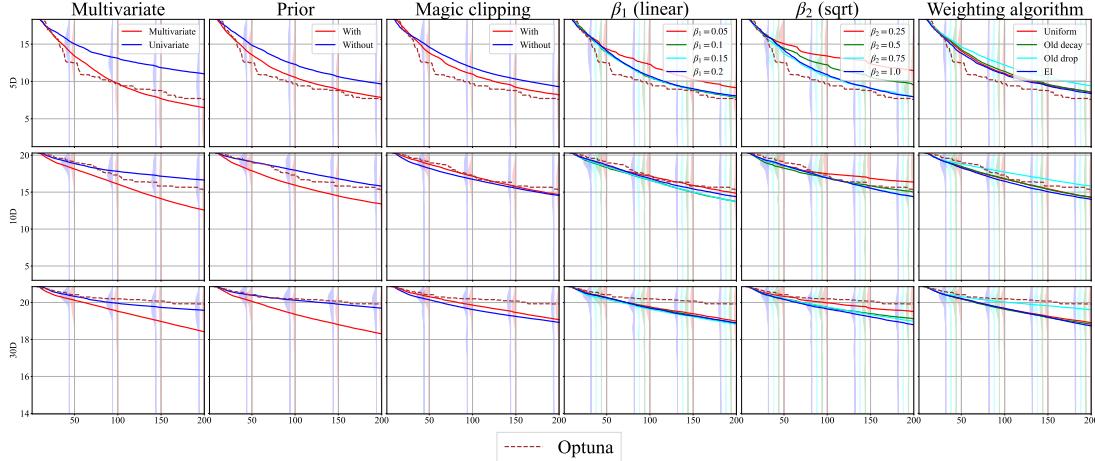
Definition as an ordinal parameter
intrinsic_cardinality = 11
dropout_choices = np.linspace(
 low, high, intrinsic_cardinality
)
index = trial.suggest_int(
 "dropout-index", 0, intrinsic_cardinality - 1
)
dropout = dropout_choices[index]

```

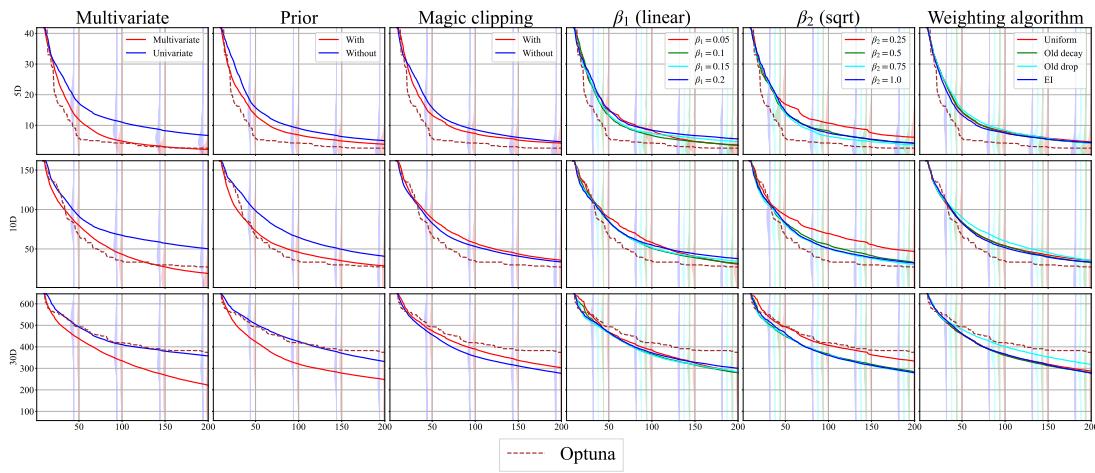
---

Notice that the discretization is not preferred in some methods such as Gaussian process-based BO because they often optimize the acquisition function by gradient-based approaches. The fourth point is algorithm selection. If the search space contains only numerical parameters, a promising candidate would be CMA-ES (Hansen, 2016) for large-budget settings and the Nelder-Mead method (Nelder & Mead, 1965) for small-budget settings. On the other hand, if the search space contains categorical or conditional parameters, random search or BO becomes a strong candidate. Note that Ozaki et al. (2022a) report that local search methods such as Nelder-Mead and CMA-ES consistently outperform global search methods. Although Ozaki et al. (2022a) do not test TPE, TPE is a promising option considering its local search nature especially when the search space contains categorical or conditional parameters. The fifth point is to restart optimizations. Restarting is especially important for non-global methods such as TPE and Nelder-Mead method because the optimization often gets stuck in a local optimum, missing promising regions.

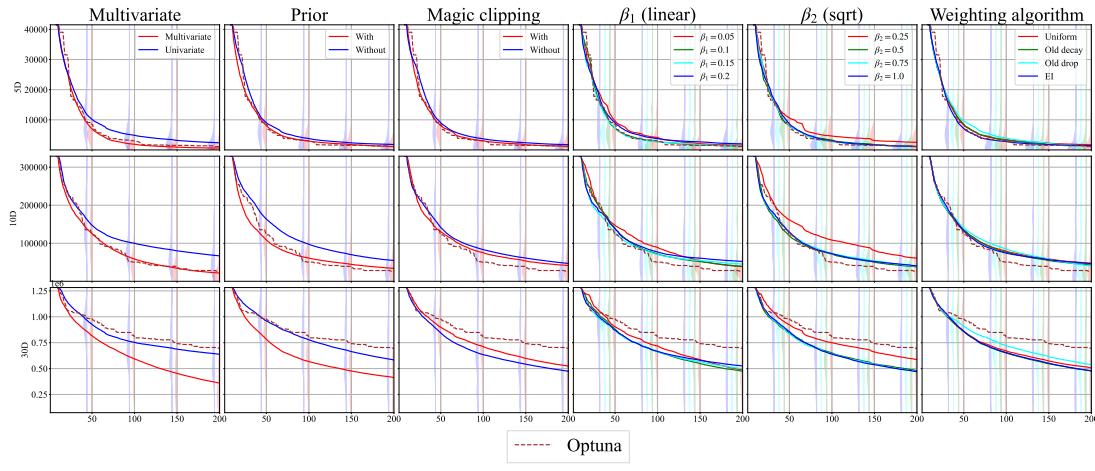
## A TUTORIAL OF TREE-STRUCTURED PARZEN ESTIMATOR



(a) Ackley with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)



(b) Griewank with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)



(c) K-Tablet with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)

Figure 18: The ablation study on benchmark functions. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

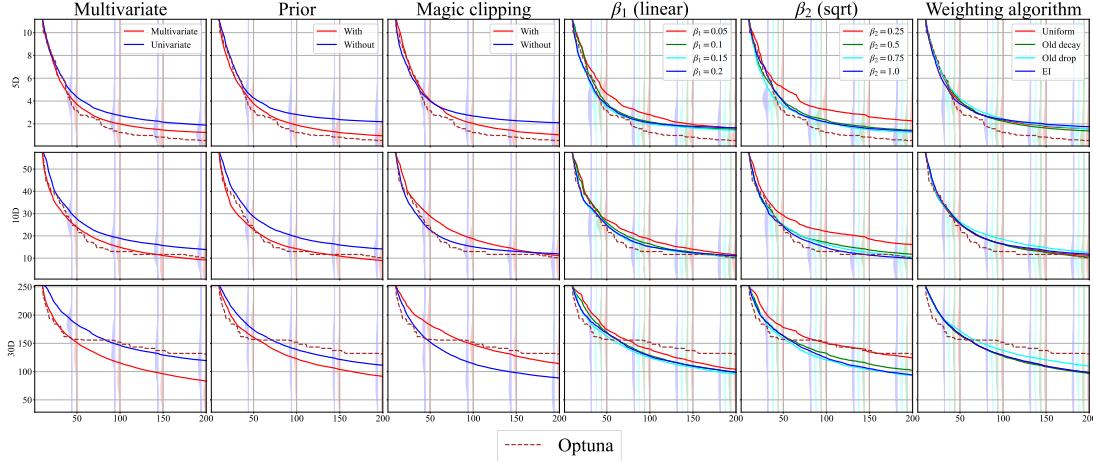
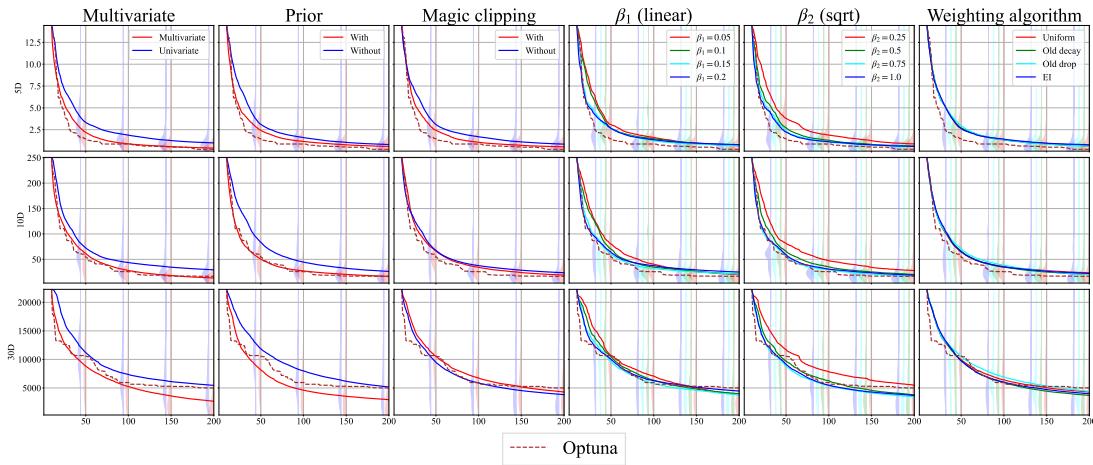
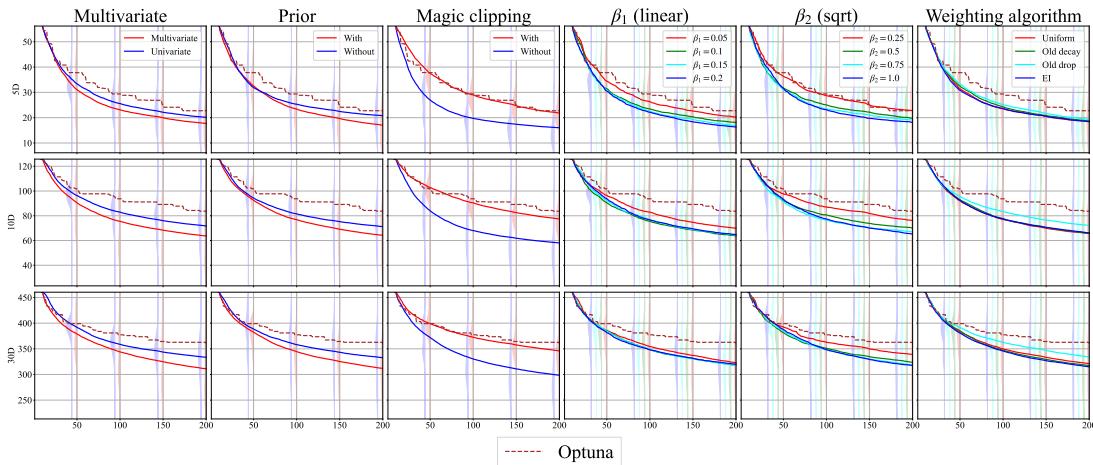
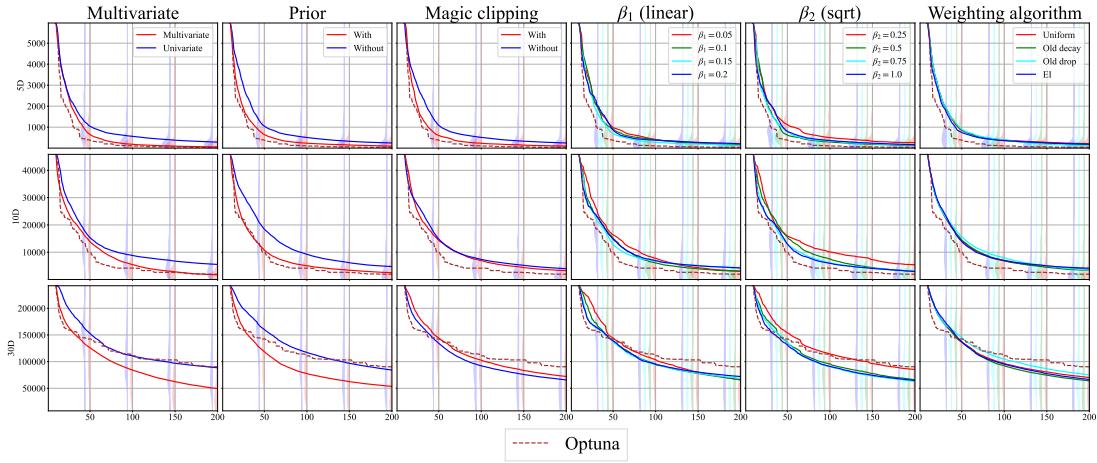
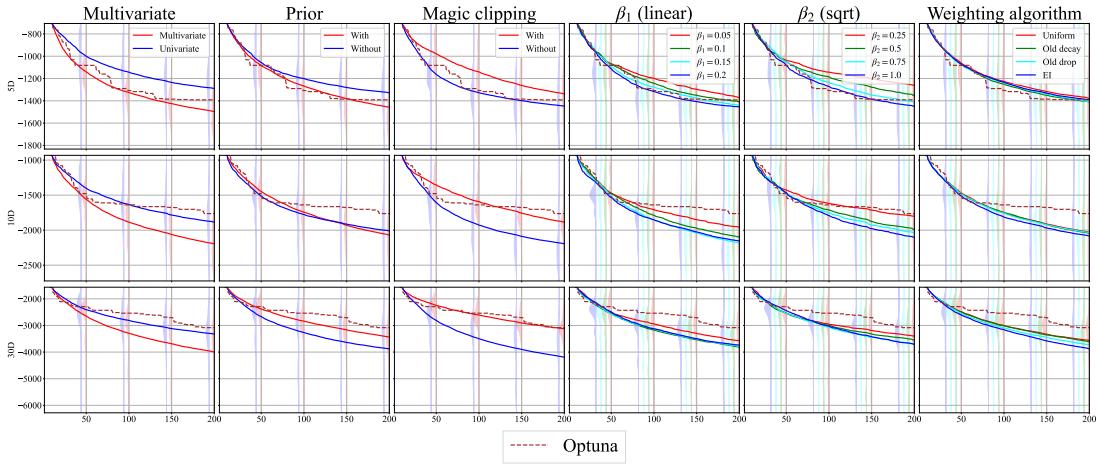
(a) Levy with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)(b) Perm with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)(c) Rastrigin with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)

Figure 19: The ablation study on benchmark functions. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

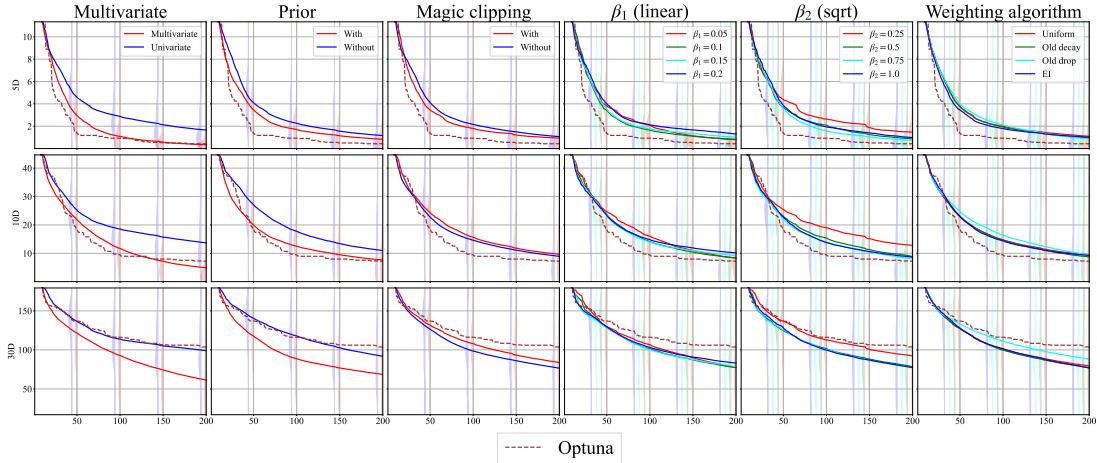
## A TUTORIAL OF TREE-STRUCTURED PARZEN ESTIMATOR



(a) Rosenbrock with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)



(b) Schwefel with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)



(c) Sphere with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)

Figure 20: The ablation study on benchmark functions. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

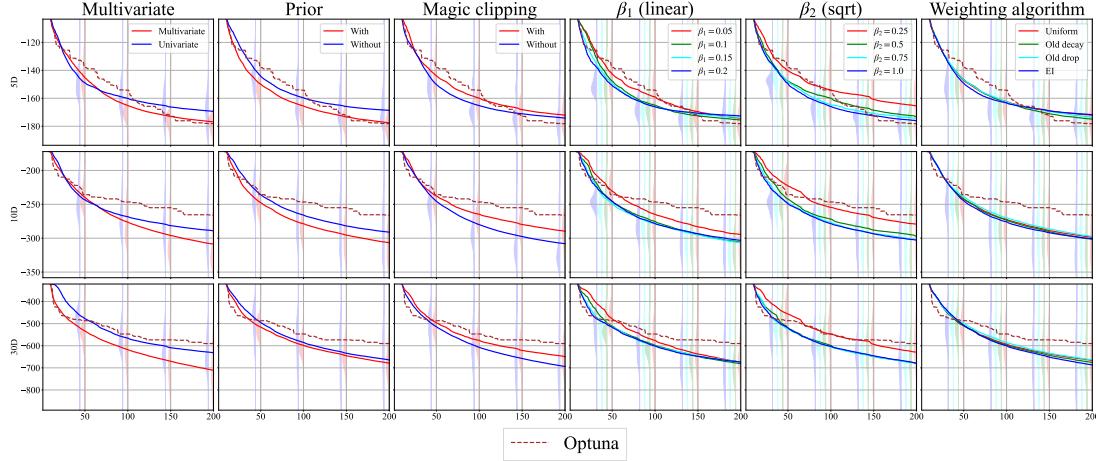
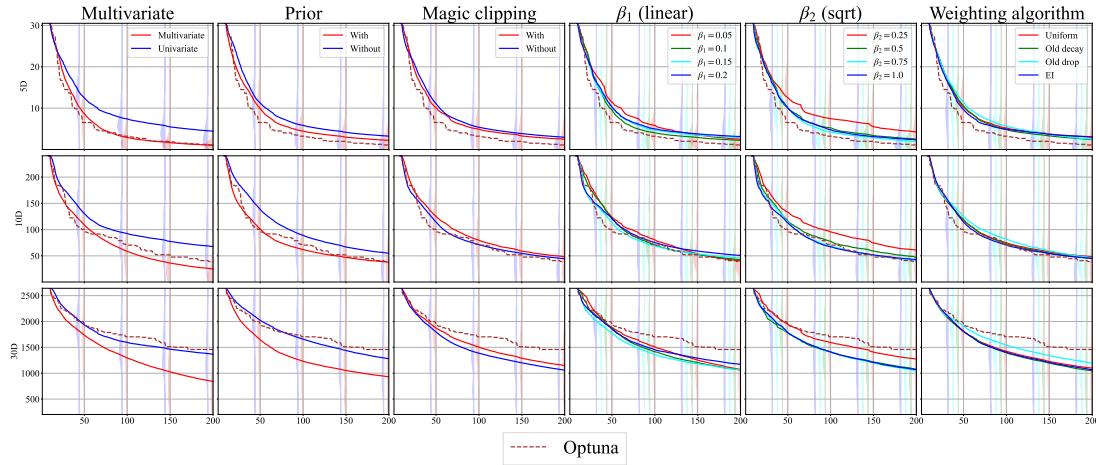
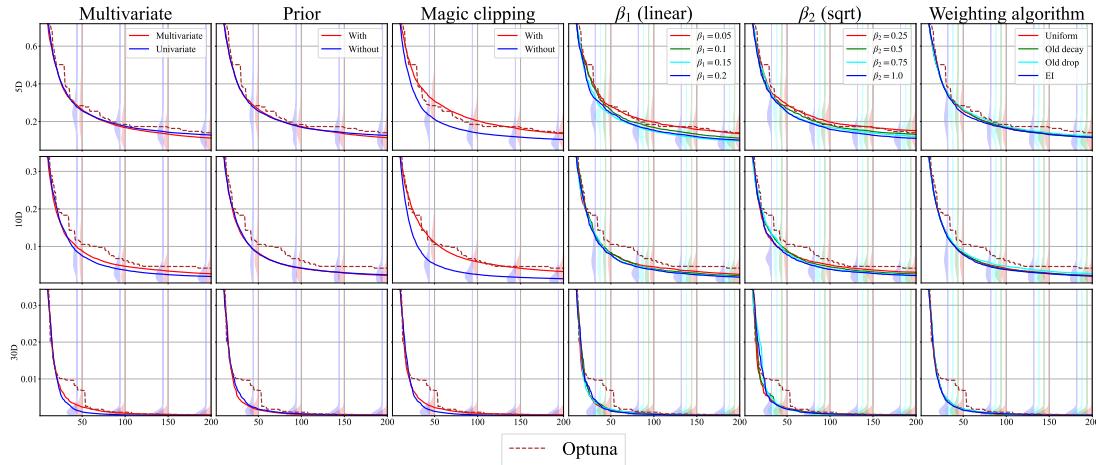
(a) Styblinski with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)(b) Weighted sphere with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)(c) Xin-She-Yang with 5D (**Top row**), 10D (**Middle row**), and 30D (**Bottom row**)

Figure 21: The ablation study on benchmark functions. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

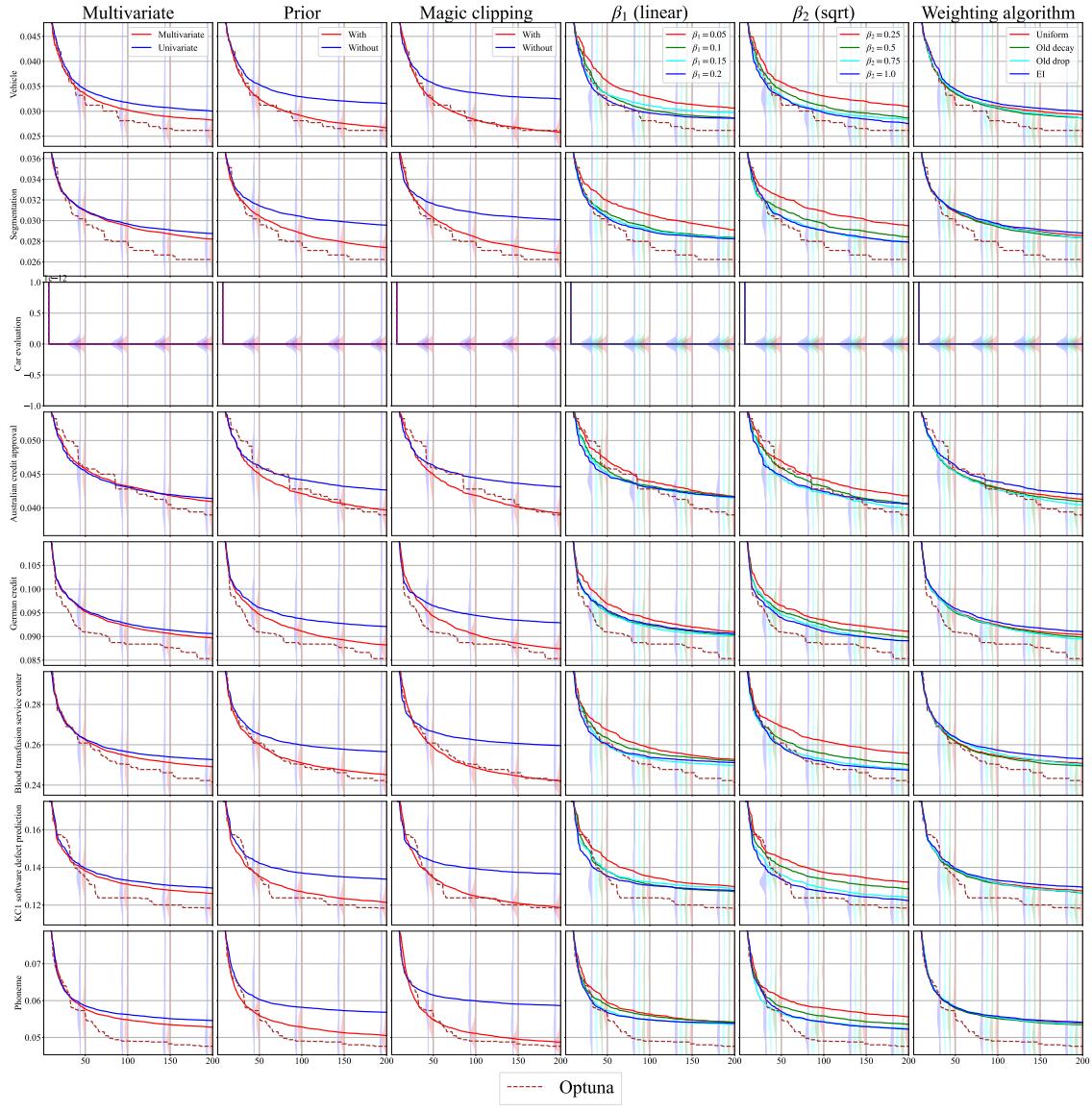


Figure 22: The ablation study on HPOBench (8 tasks). The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

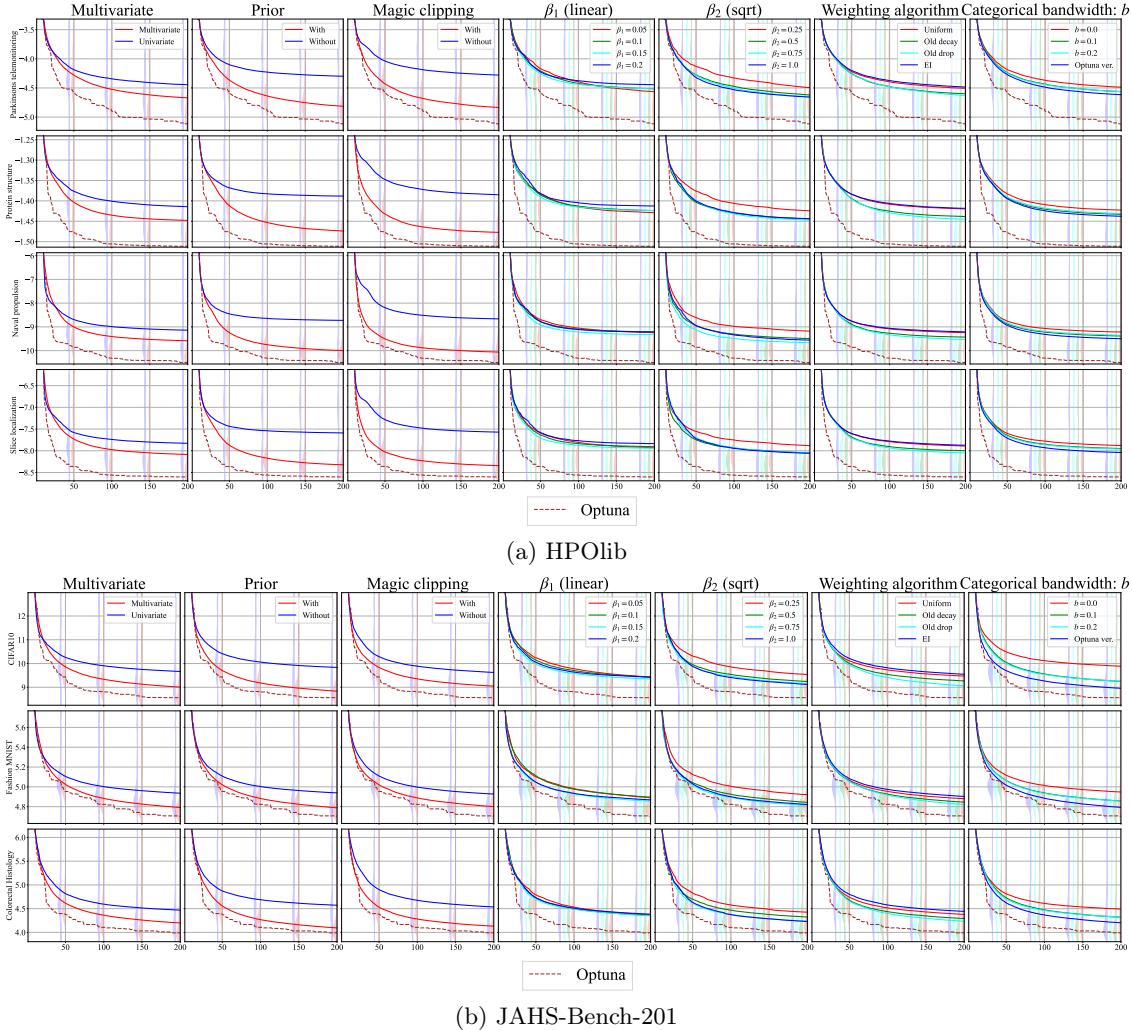


Figure 23: The ablation study on HPOlib (4 tasks) and JAHS-Bench-201 (3 tasks), which have categorical parameters. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. Note that the objective of HPOlib is the log of validation mean squared error. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

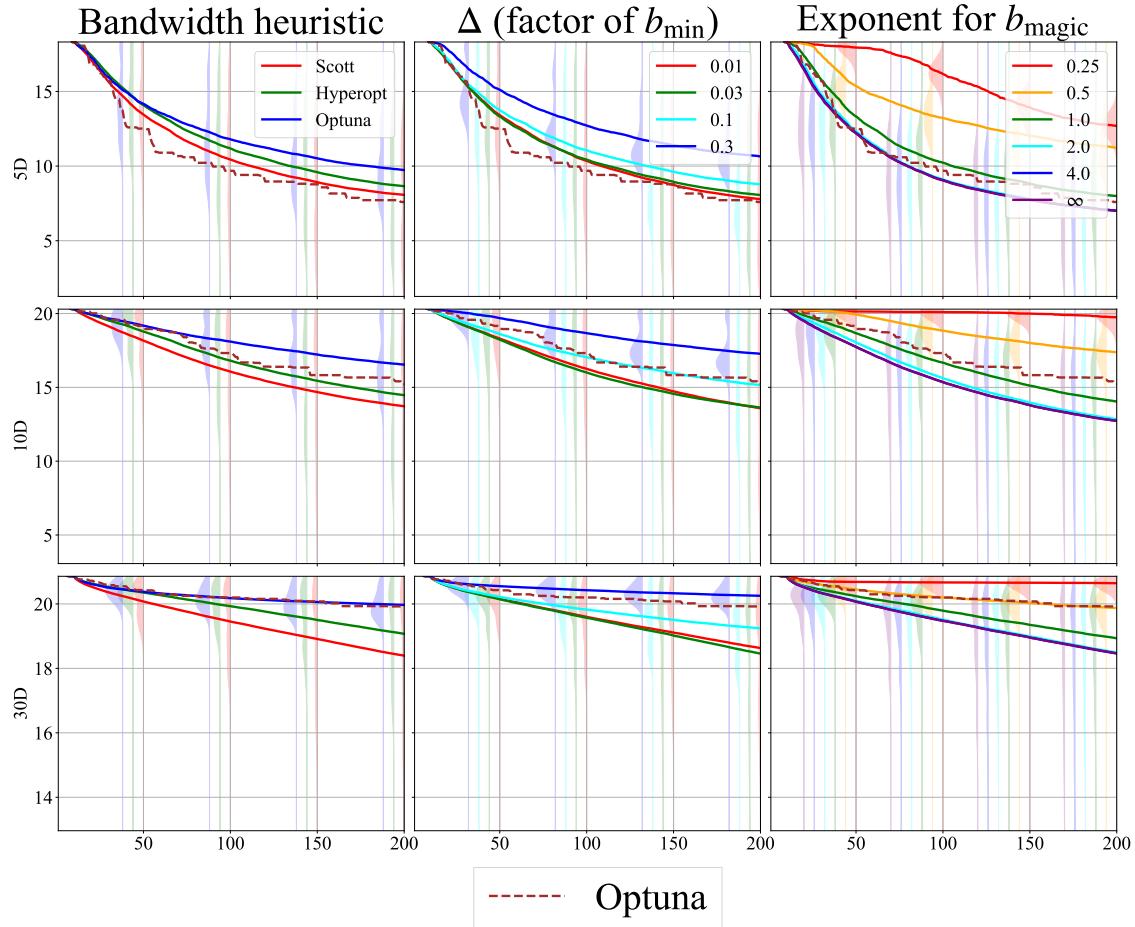


Figure 24: The ablation study of bandwidth related algorithms on the Ackley function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

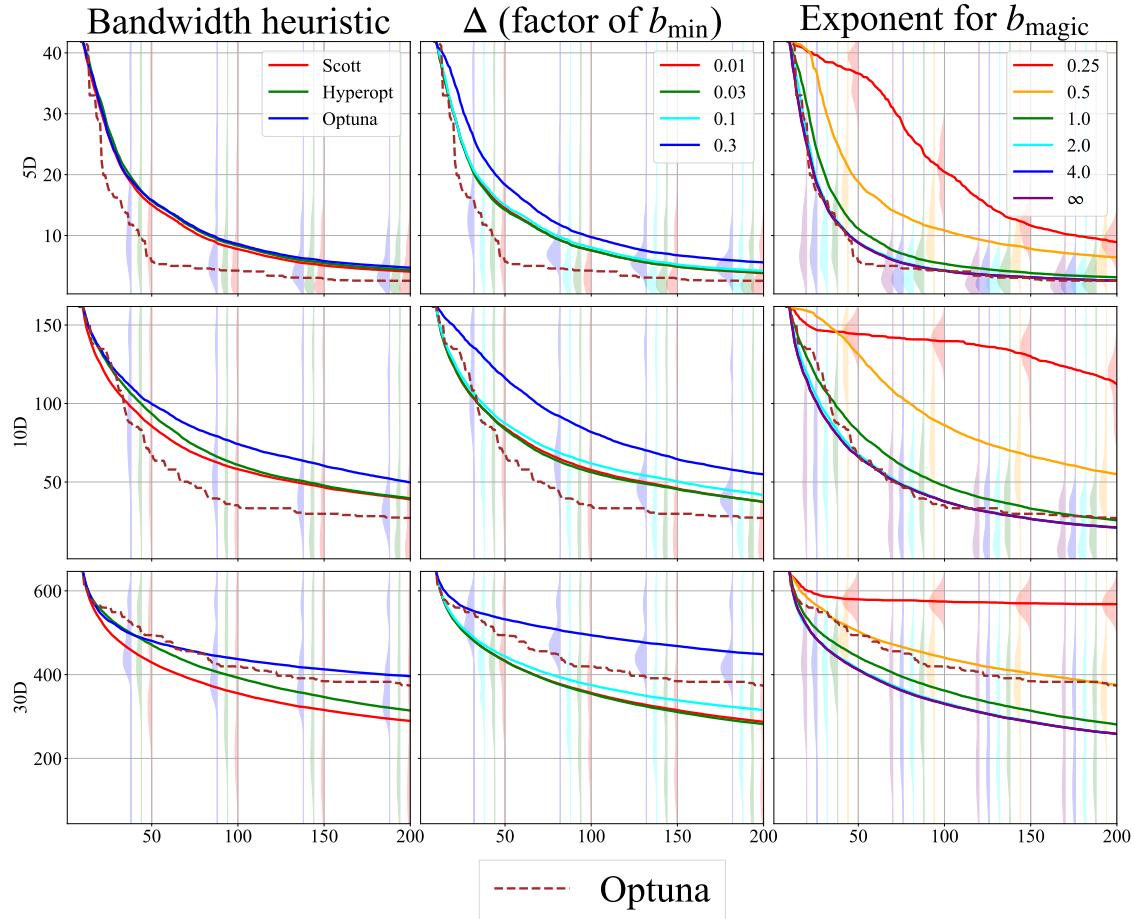


Figure 25: The ablation study of bandwidth related algorithms on the Griewank function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

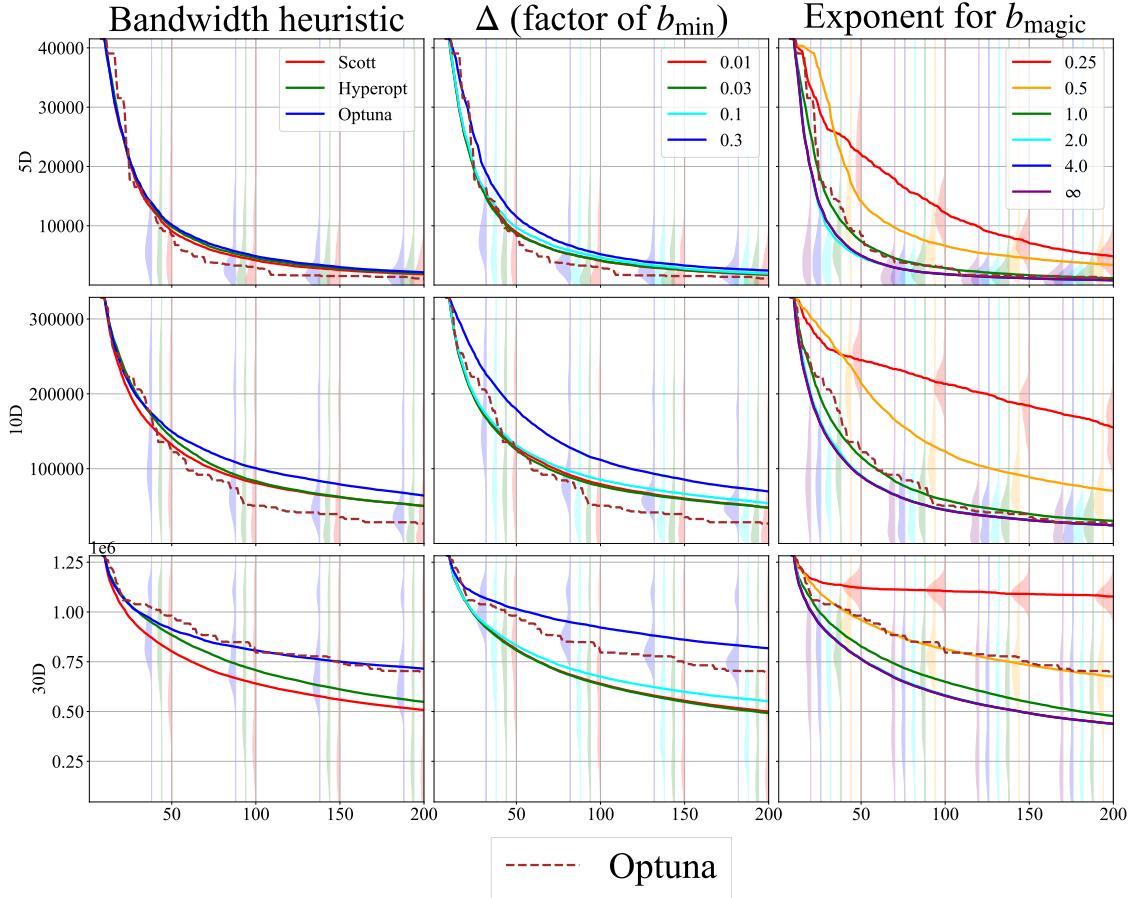


Figure 26: The ablation study of bandwidth related algorithms on the K-Tablet function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

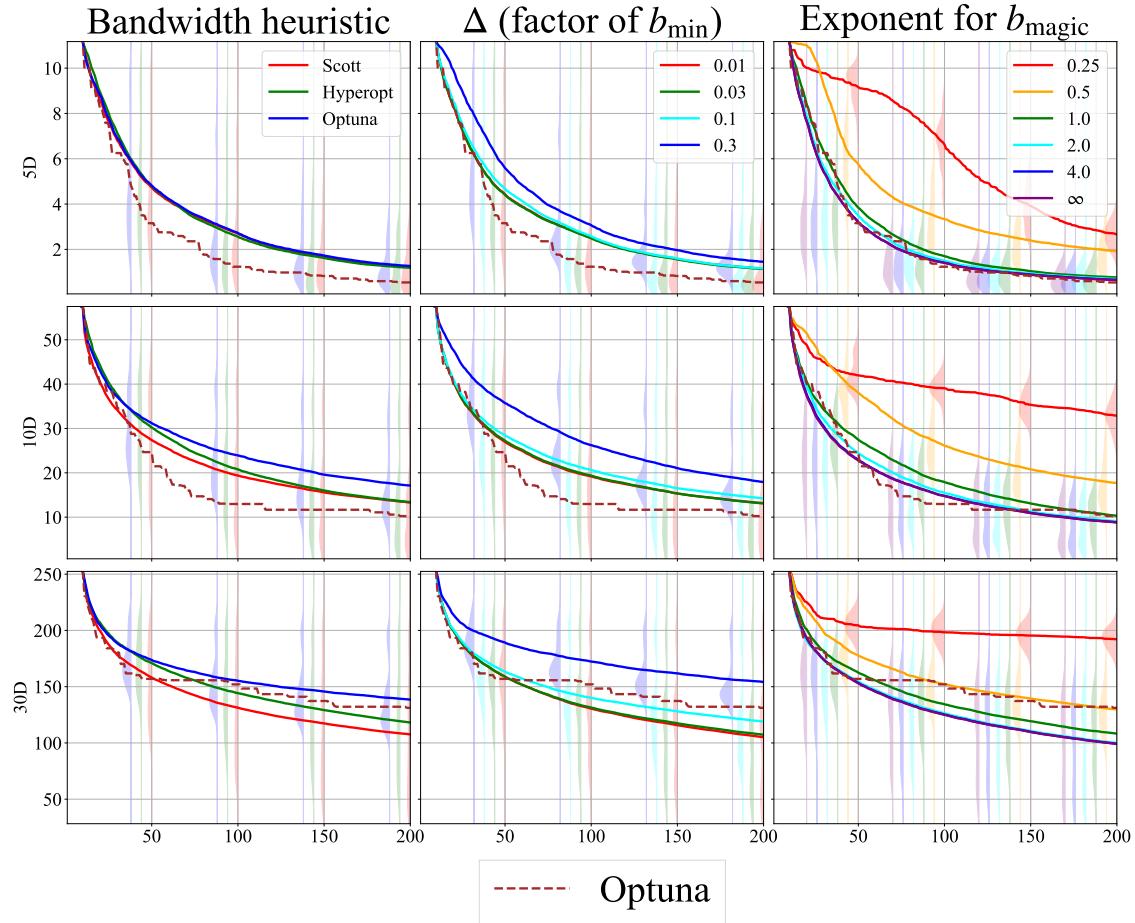


Figure 27: The ablation study of bandwidth related algorithms on the Levy function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

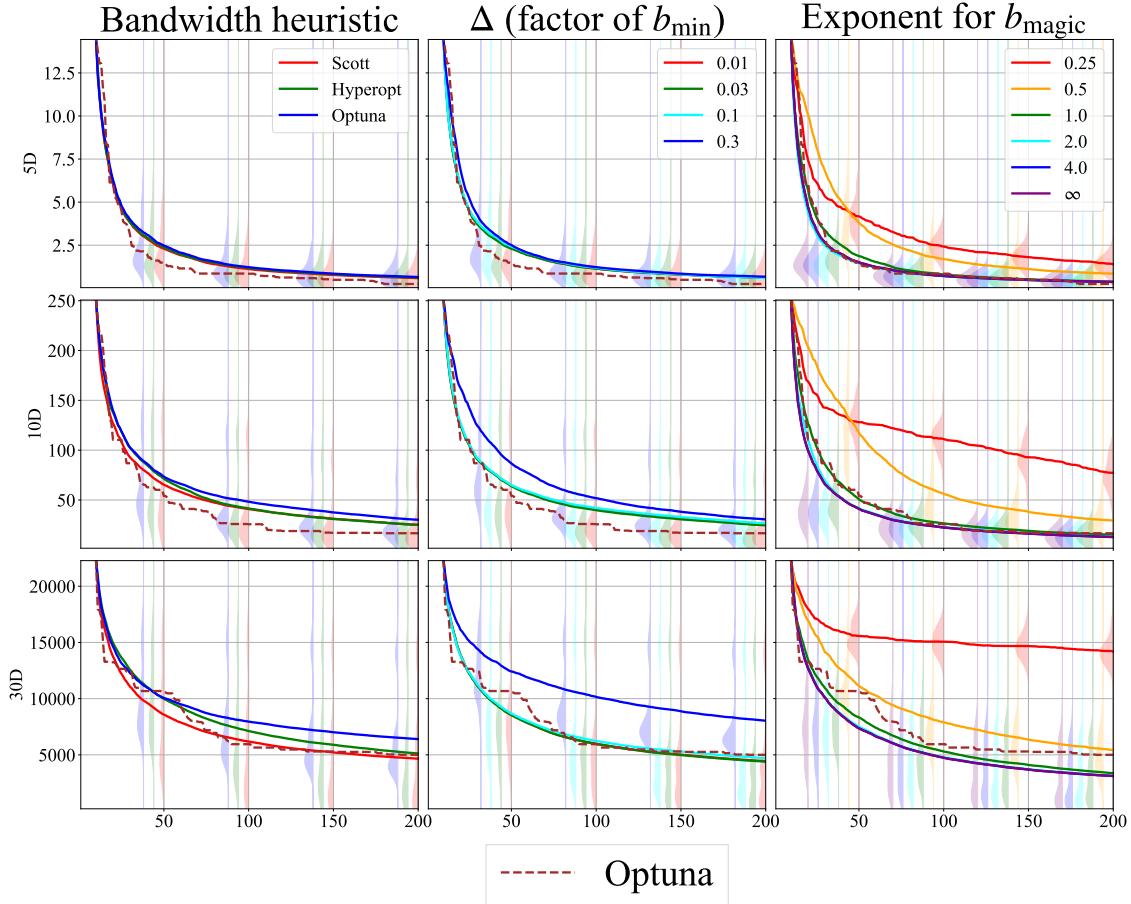


Figure 28: The ablation study of bandwidth related algorithms on the Perm function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

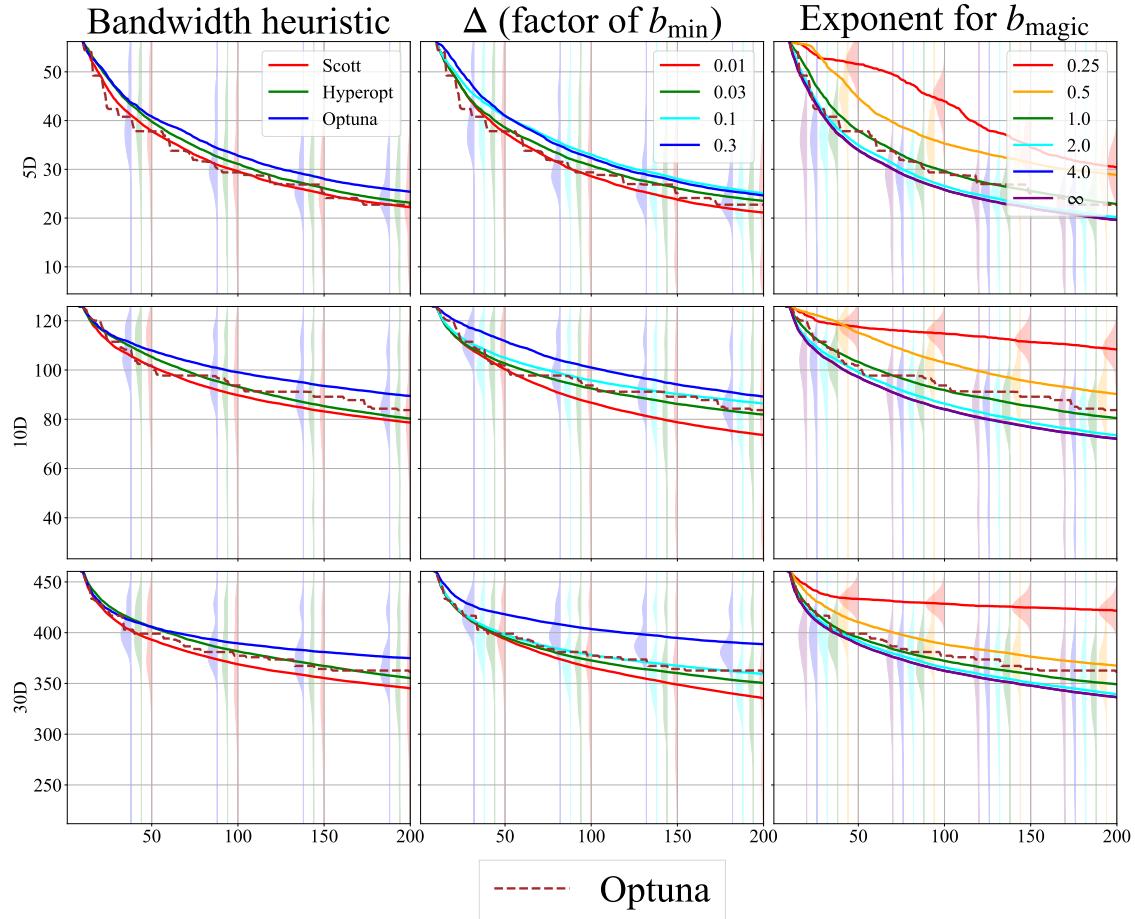


Figure 29: The ablation study of bandwidth related algorithms on the Rastrigin function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

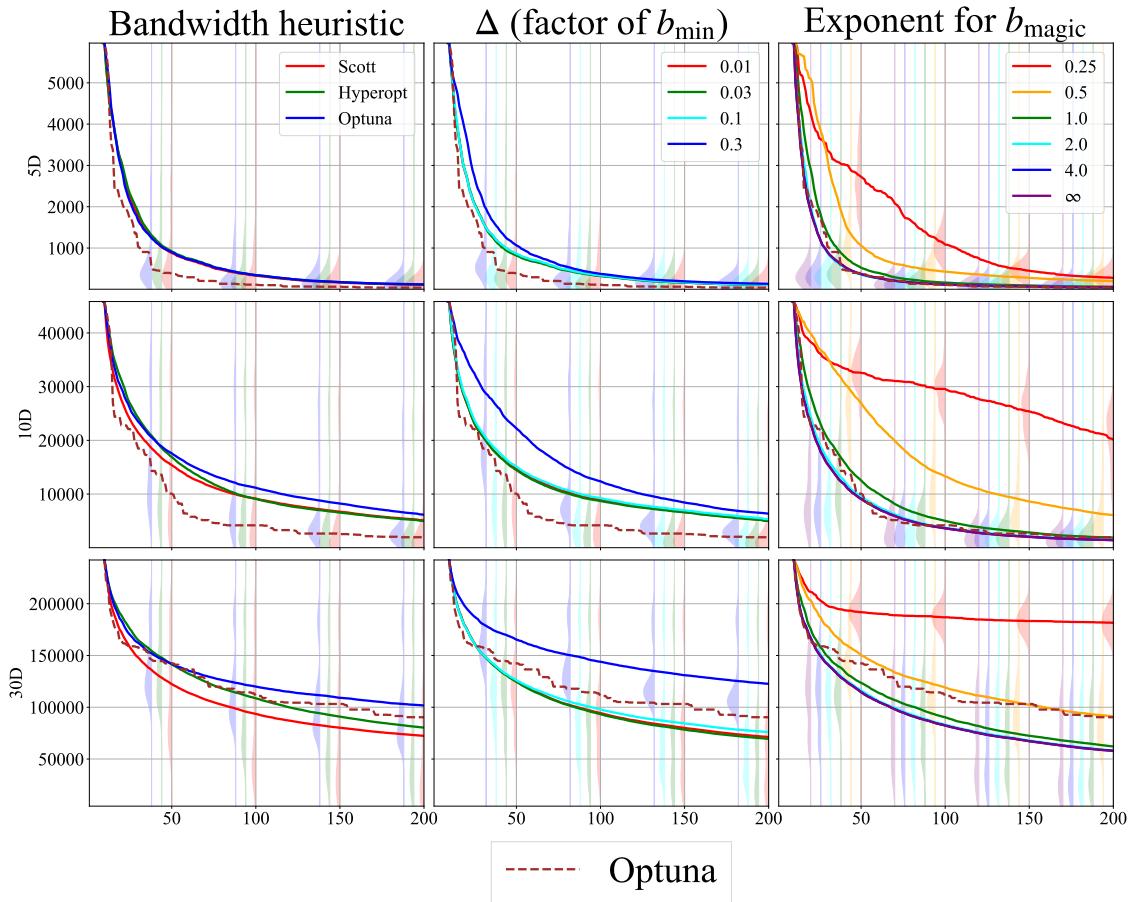


Figure 30: The ablation study of bandwidth related algorithms on the Rosenbrock function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

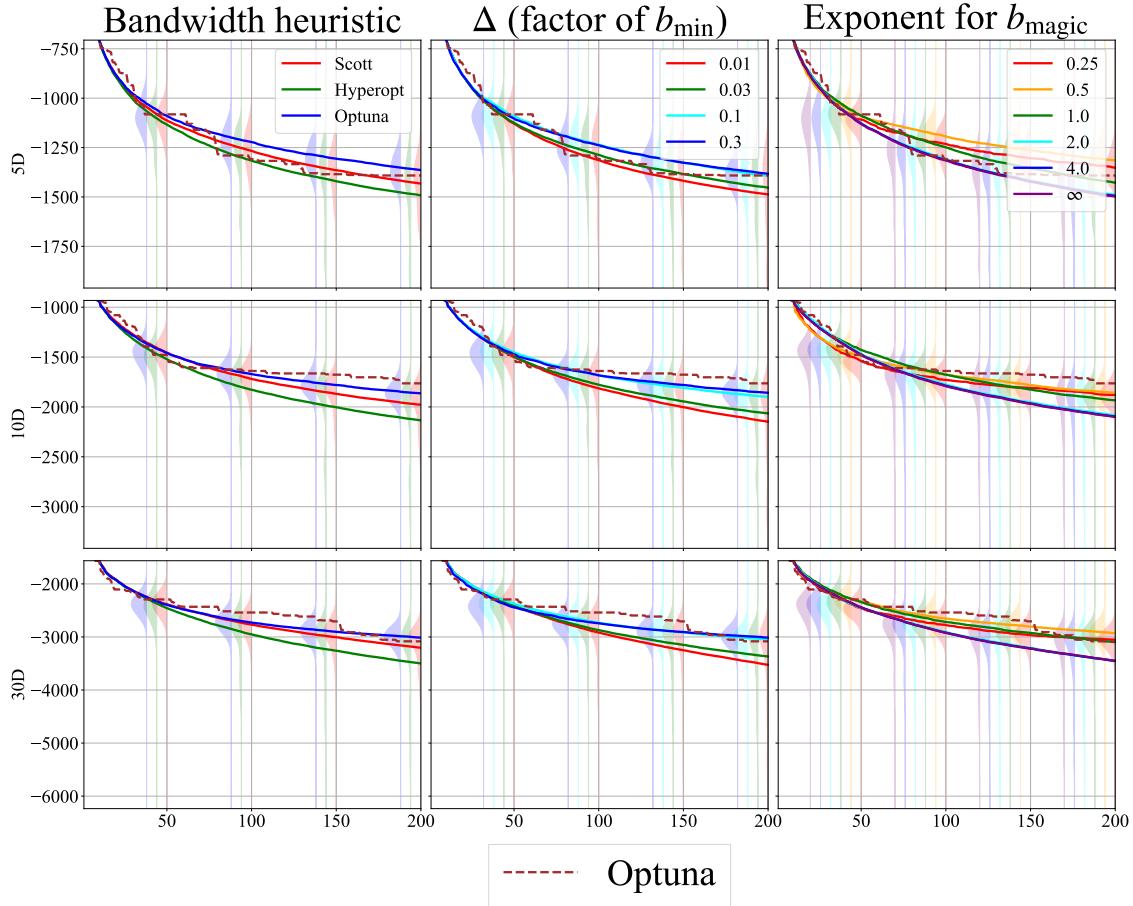


Figure 31: The ablation study of bandwidth related algorithms on the Schwefel function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

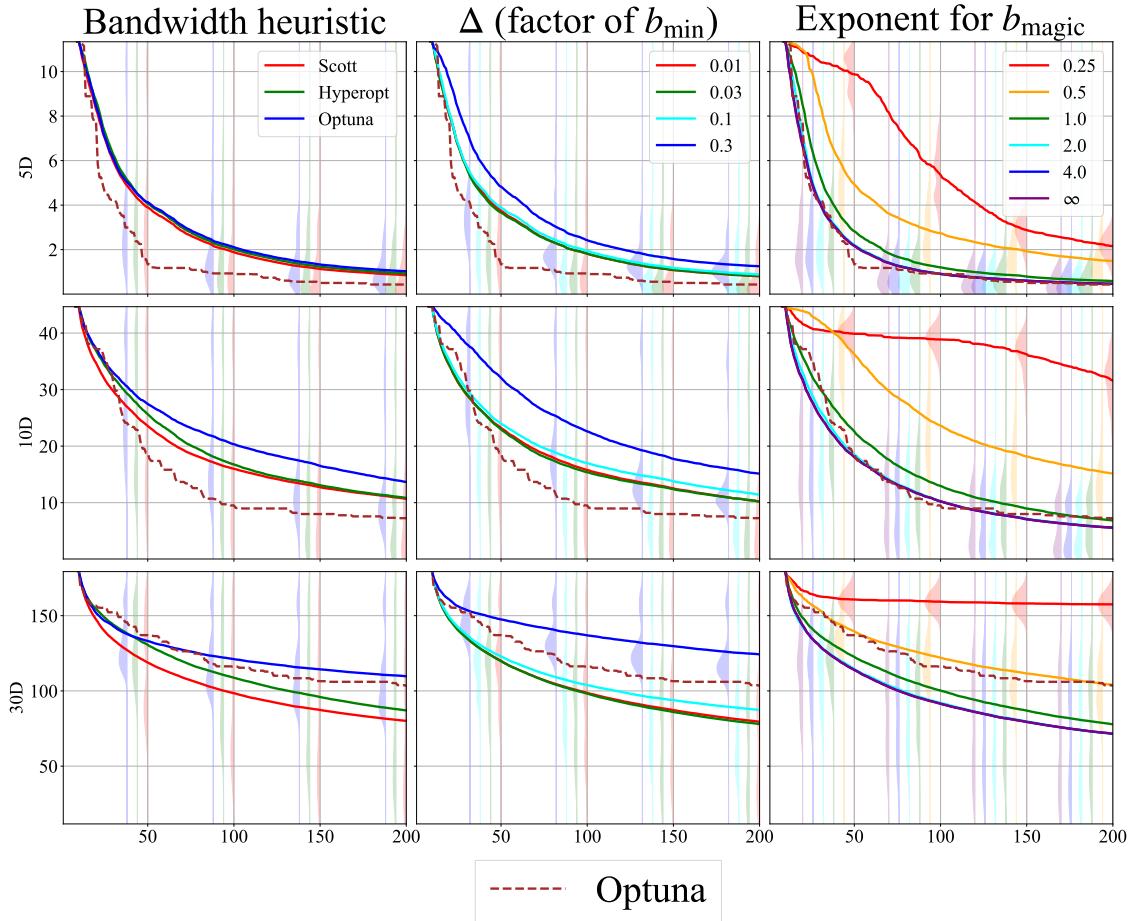


Figure 32: The ablation study of bandwidth related algorithms on the Sphere function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

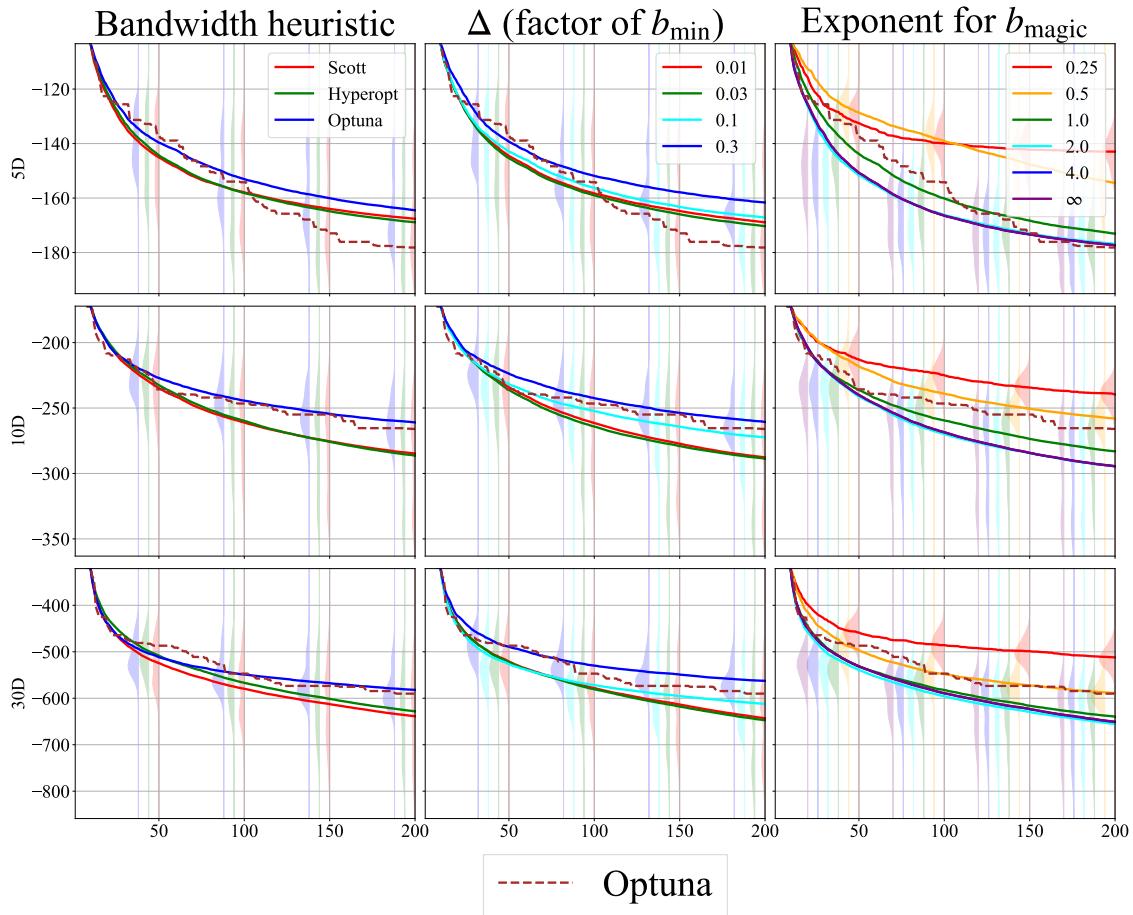


Figure 33: The ablation study of bandwidth related algorithms on the Styblinski function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

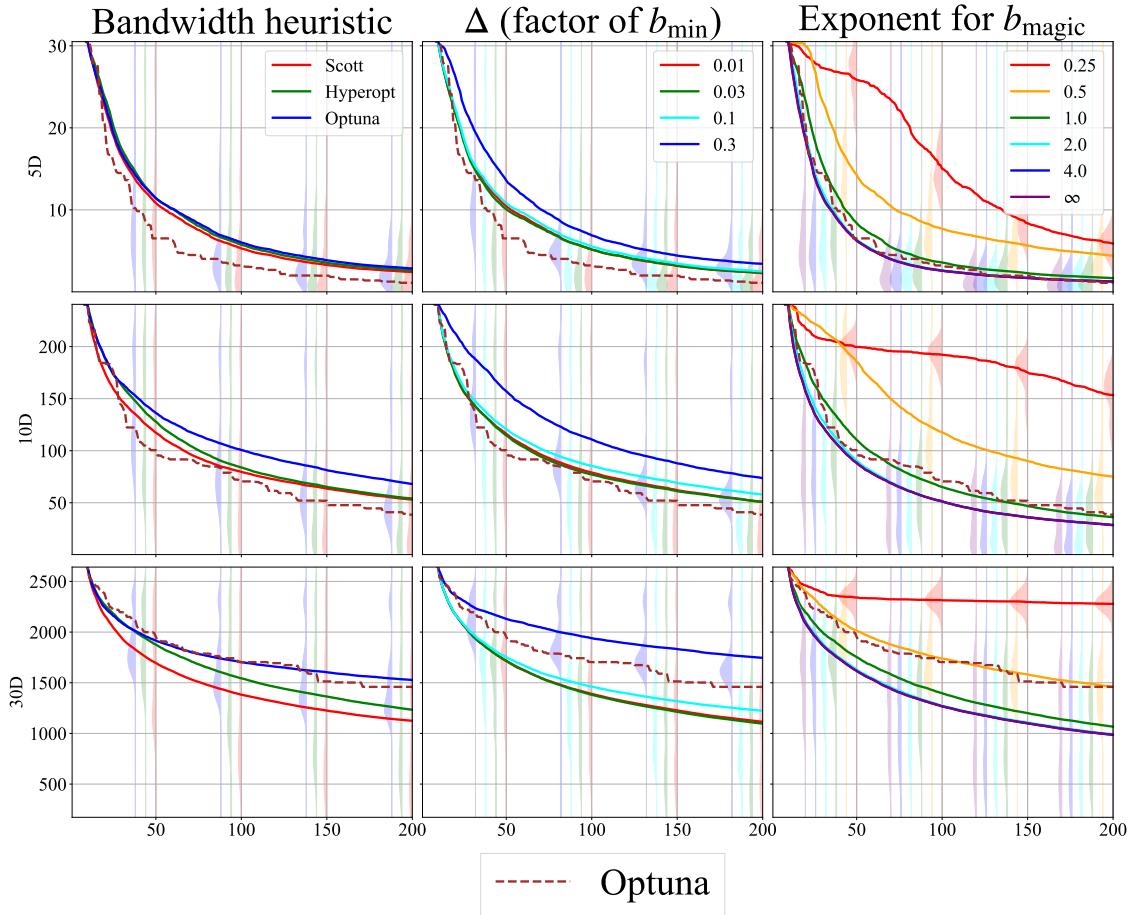


Figure 34: The ablation study of bandwidth related algorithms on the weighted sphere function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

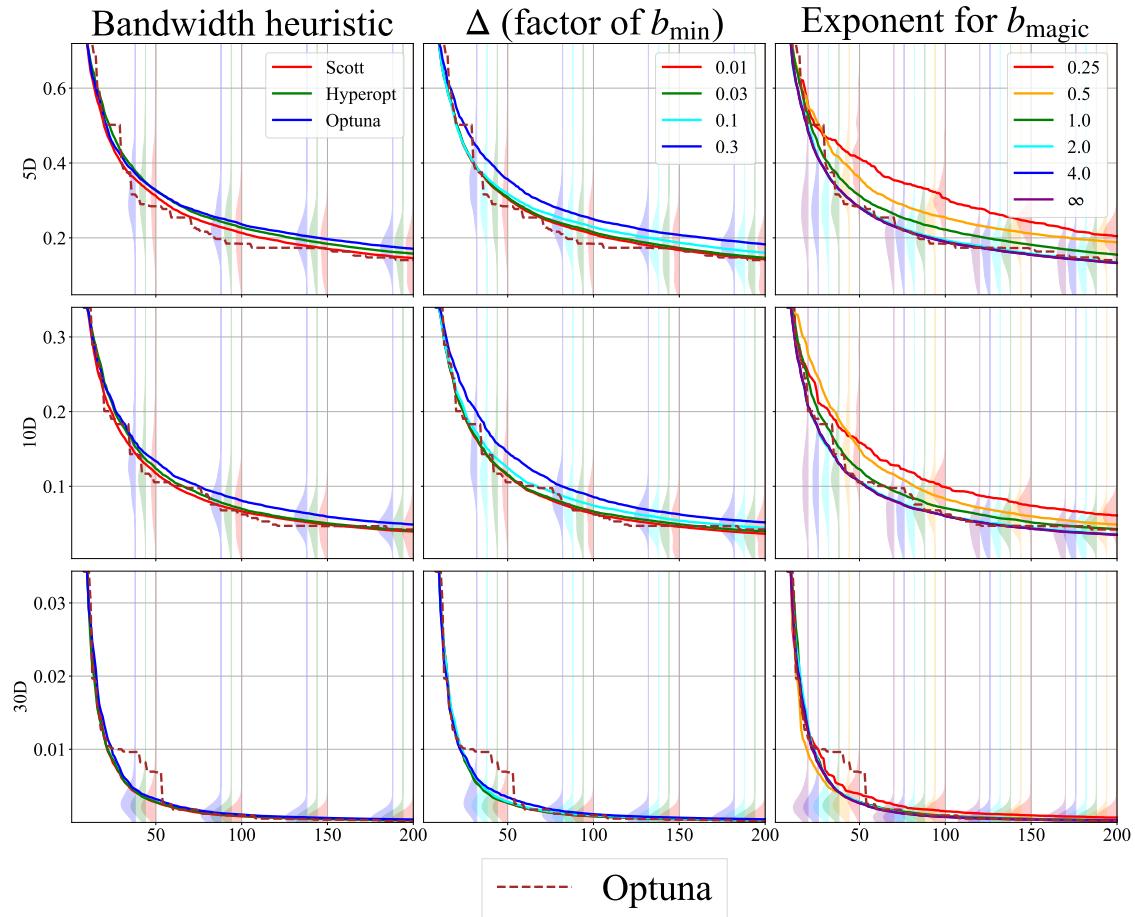


Figure 35: The ablation study of bandwidth related algorithms on the Xin-She-Yang function. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

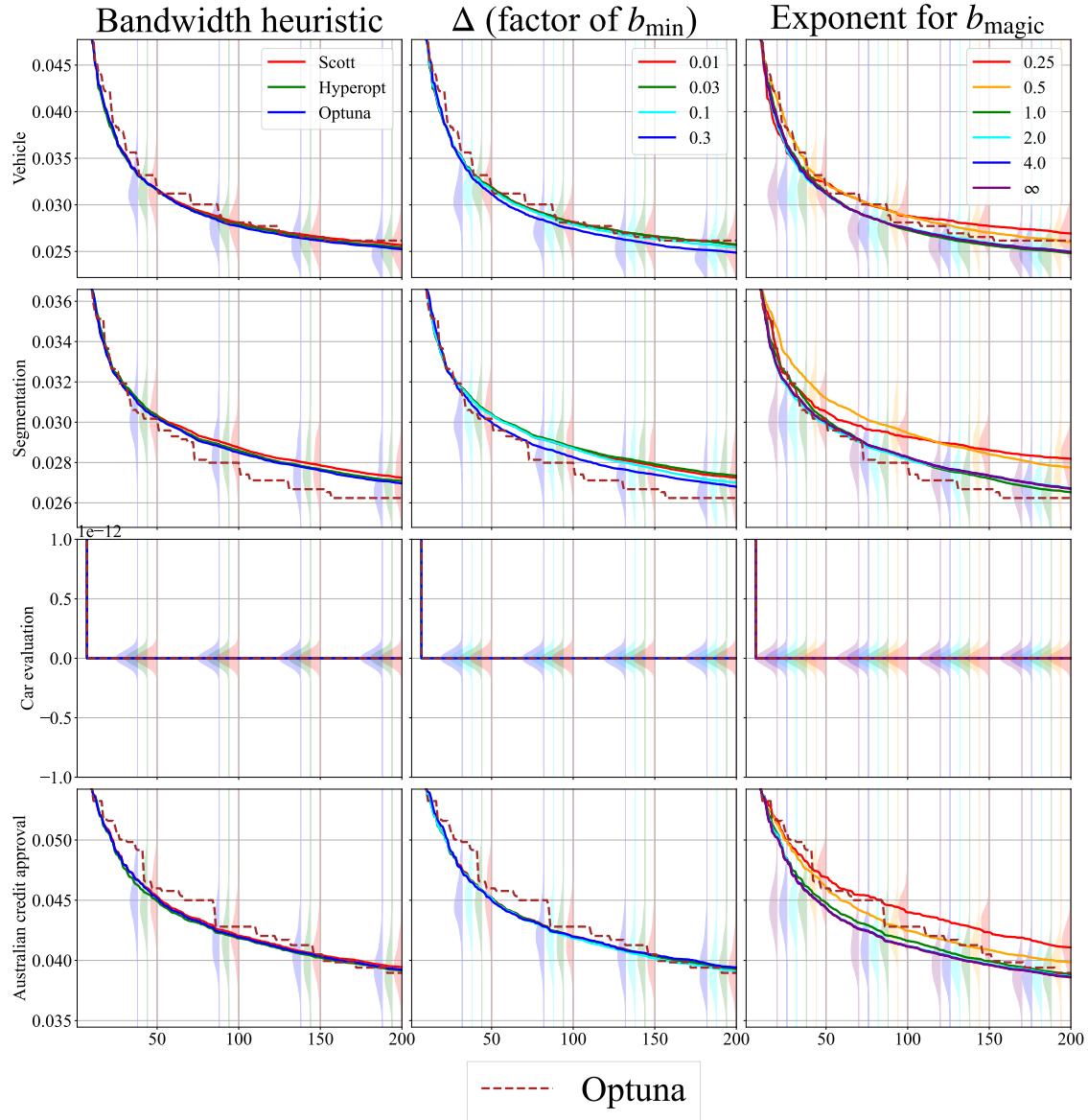


Figure 36: The ablation study of bandwidth algorithms on HPOBench (**Vehicle**, **Segmentation**, **Car evaluation**, **Australian credit approval**). The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

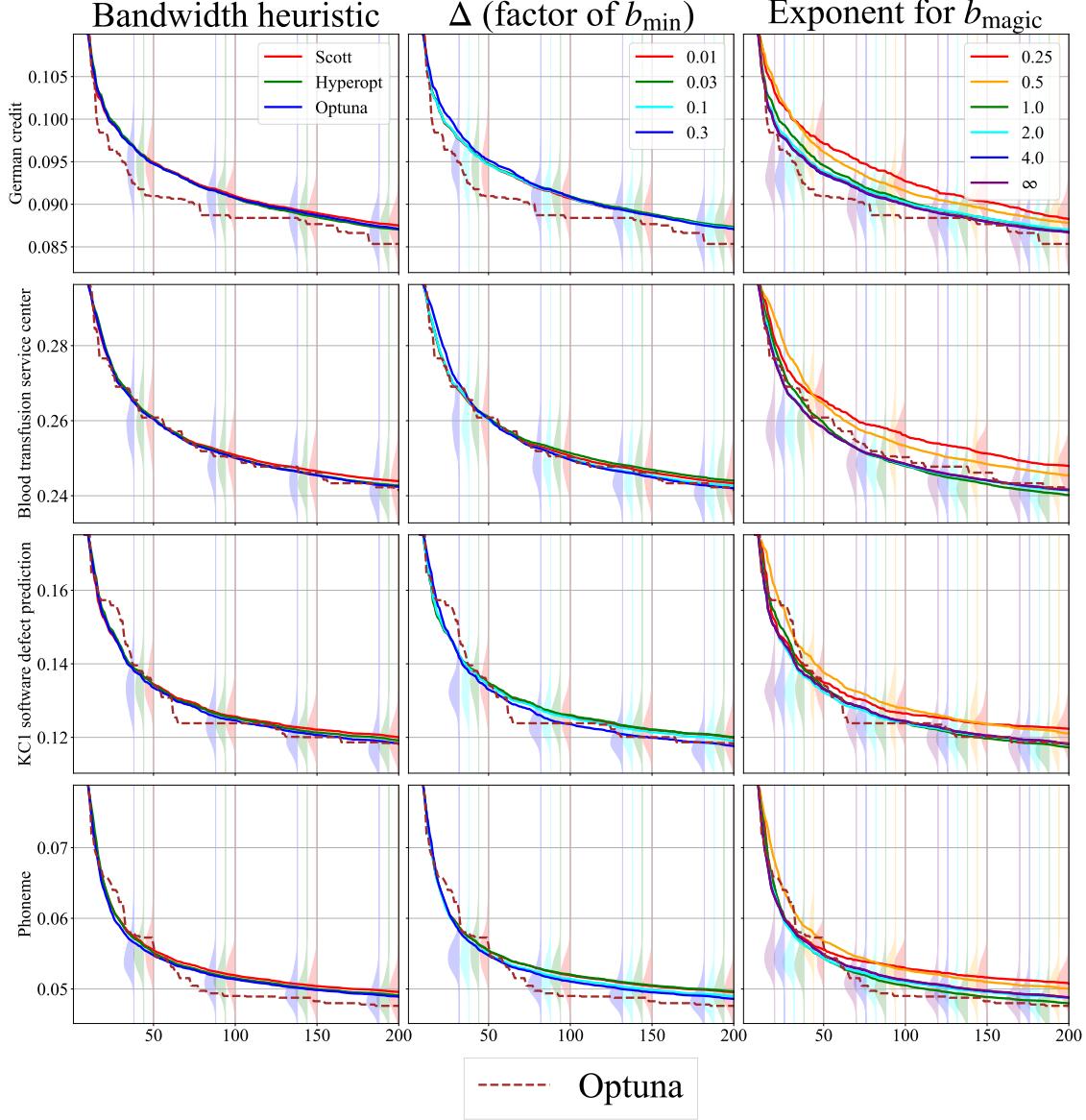


Figure 37: The ablation study of bandwidth algorithms on HPOBench (**German credit**, **Blood transfusion service center**, **KC1 software defect prediction**, **Phoneme**). The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

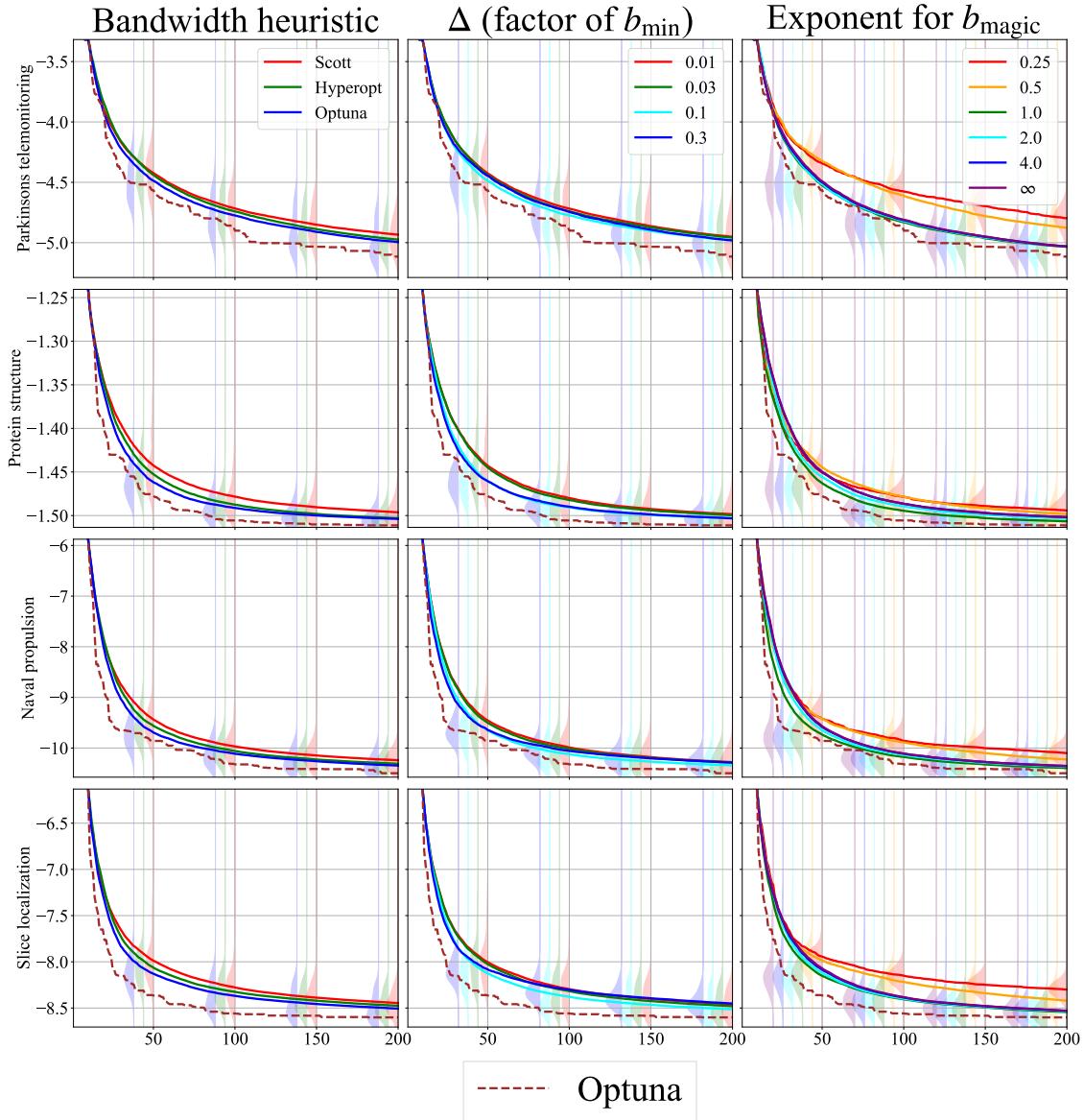


Figure 38: The ablation study of bandwidth algorithms on HPOlib. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. Note that the objective of HPOlib is the log of validation mean squared error. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

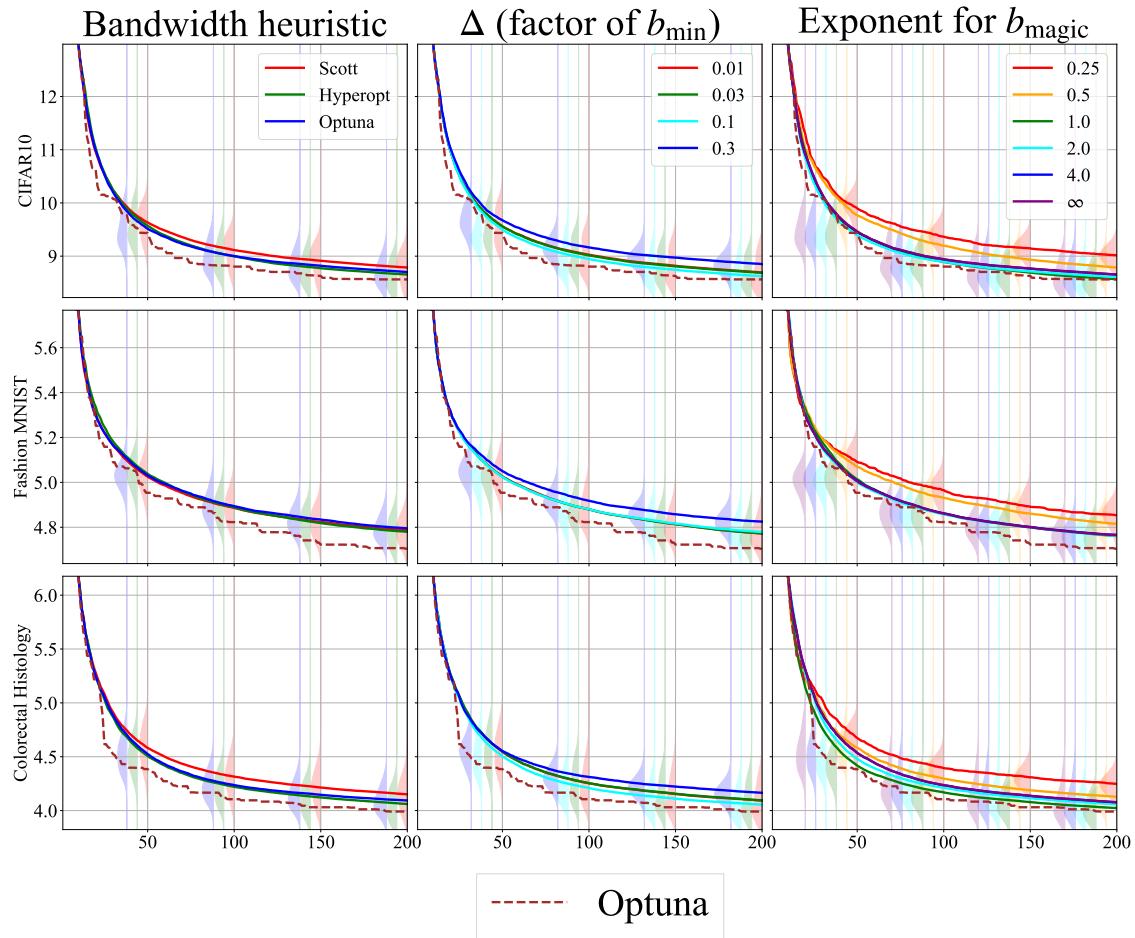


Figure 39: The ablation study of bandwidth algorithms on JAHS-Bench-201. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective. The solid lines in each figure show the mean of the cumulative minimum objective over all control parameter configurations. The transparent shades represent the distributions of the cumulative minimum objective at  $\{50, 100, 150, 200\}$  evaluations. The performance of Optuna v4.0.0 (brown dotted lines) is provided as a baseline.

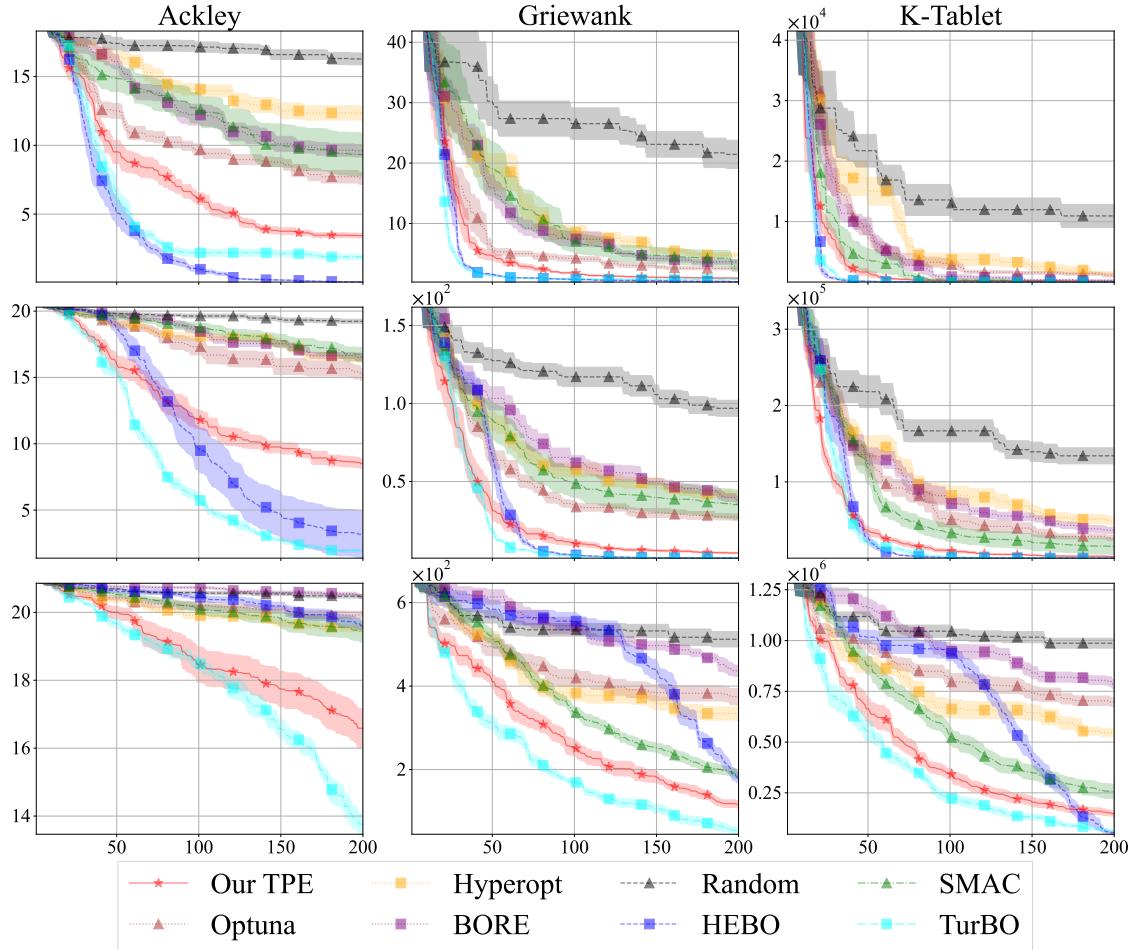


Figure 40: The comparison of optimization methods on benchmark functions. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective value. Each optimization method was run with 10 different random seeds and the weak-color bands represent the standard error.

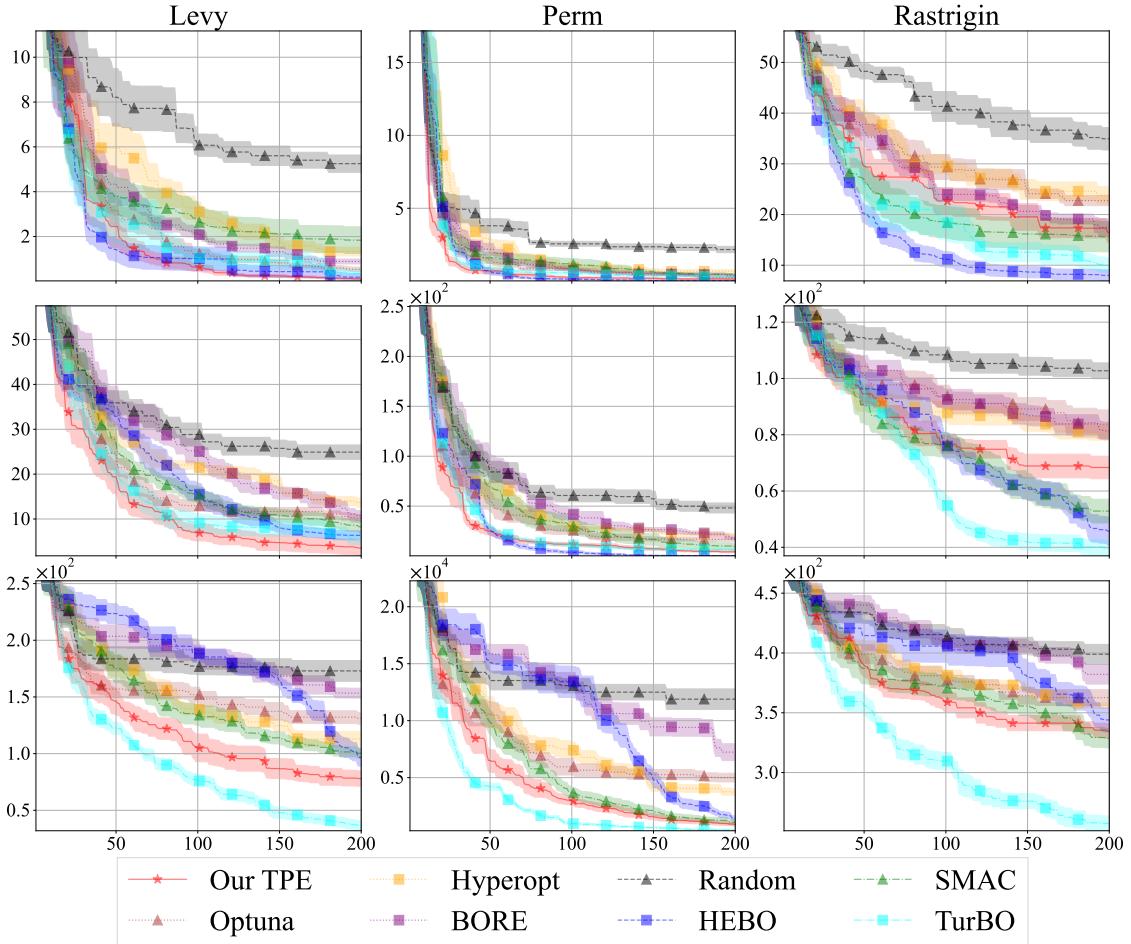


Figure 41: The comparison of optimization methods on benchmark functions. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective value. Each optimization method was run with 10 different random seeds and the weak-color bands represent the standard error.

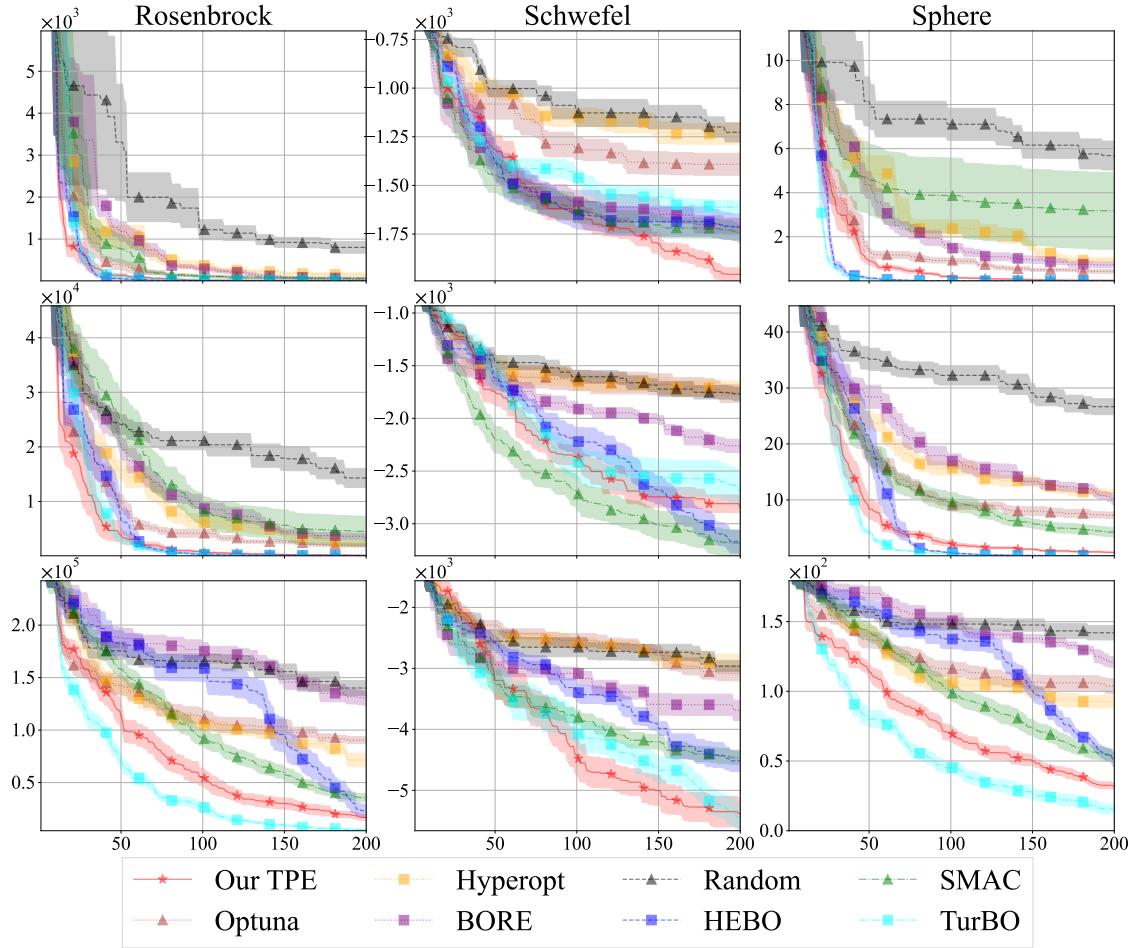


Figure 42: The comparison of optimization methods on benchmark functions. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective value. Each optimization method was run with 10 different random seeds and the weak-color bands represent the standard error.

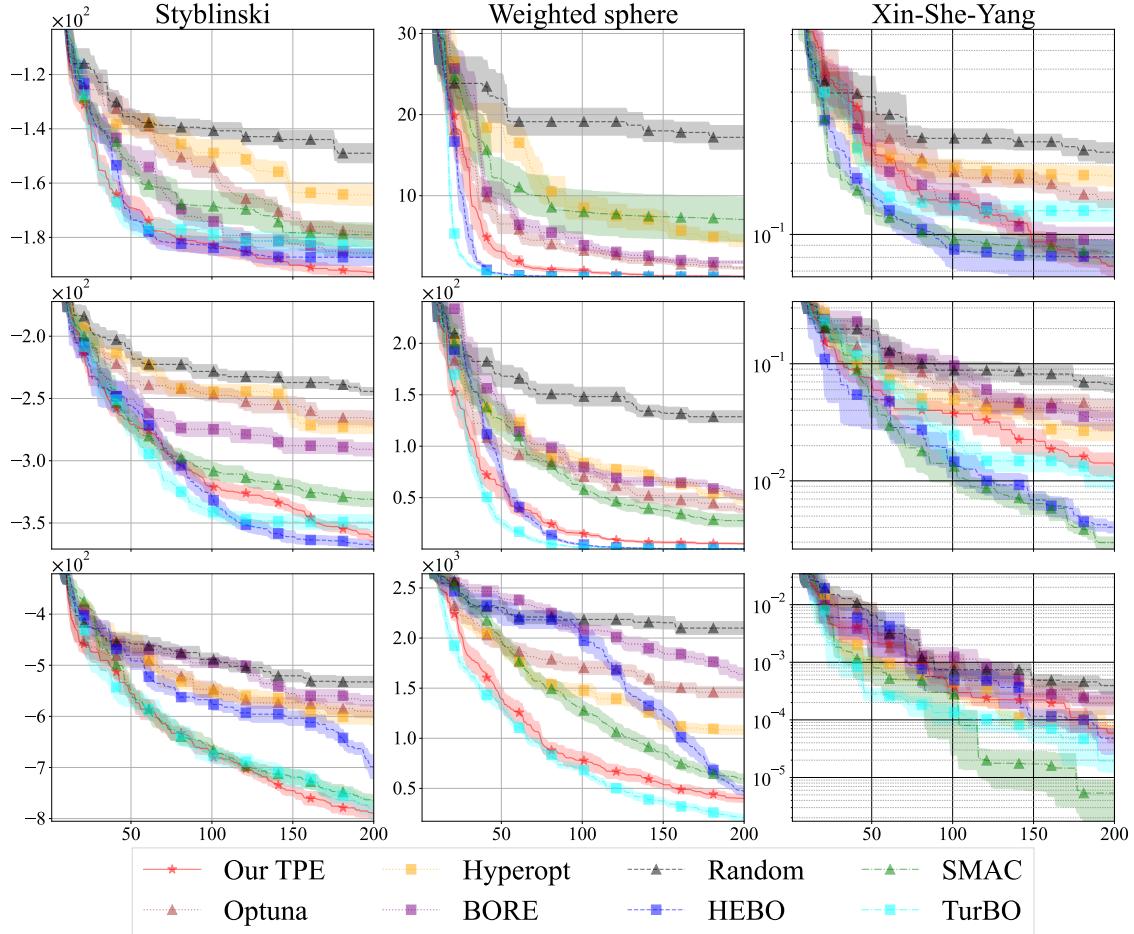


Figure 43: The comparison of optimization methods on benchmark functions. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective value. Each optimization method was run with 10 different random seeds and the weak-color bands represent the standard error.

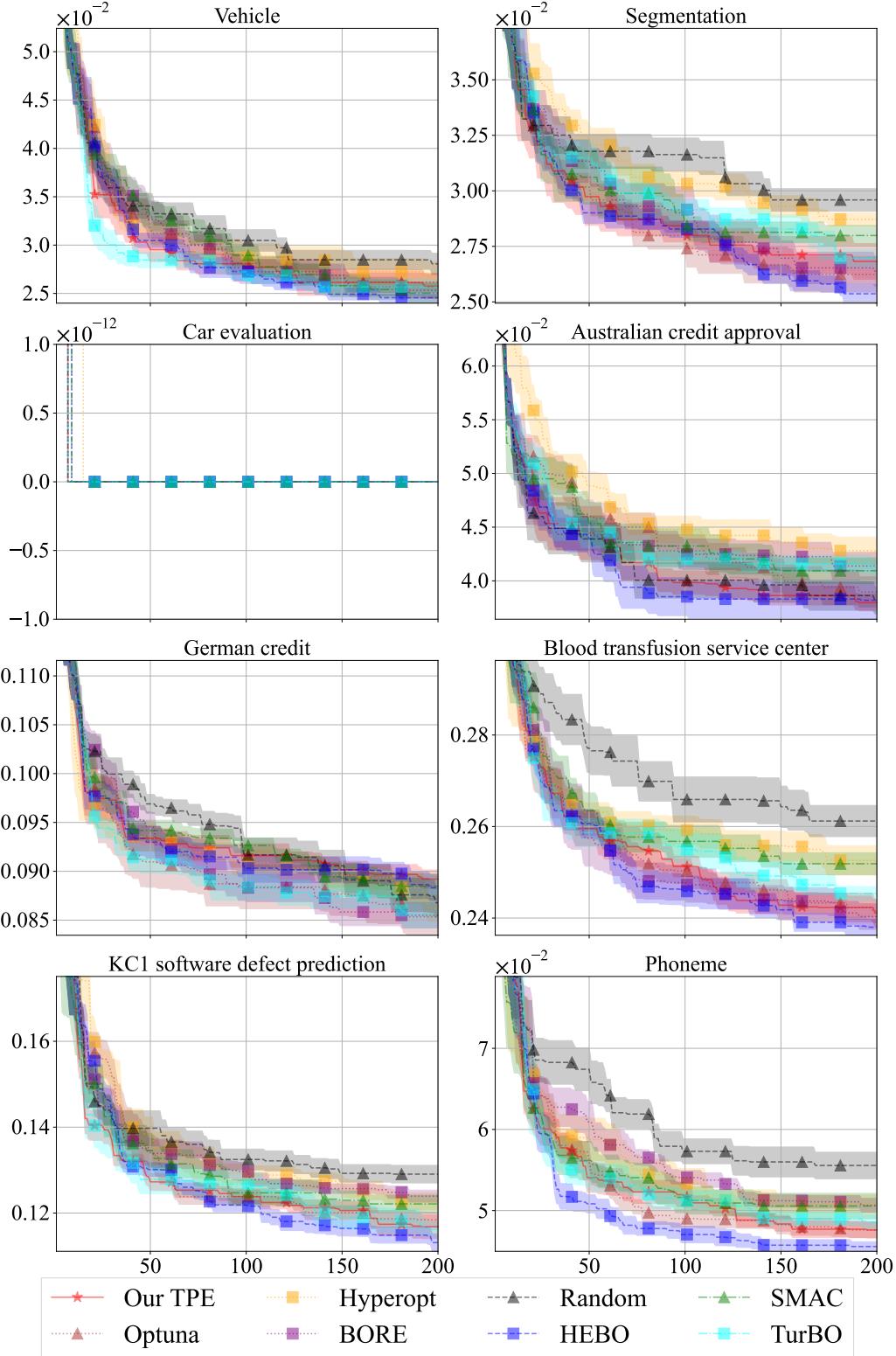


Figure 44: The comparison of optimization methods on HPOBench. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective value. Each optimization method was run with 10 different random seeds and the weak-color bands represent the standard error.

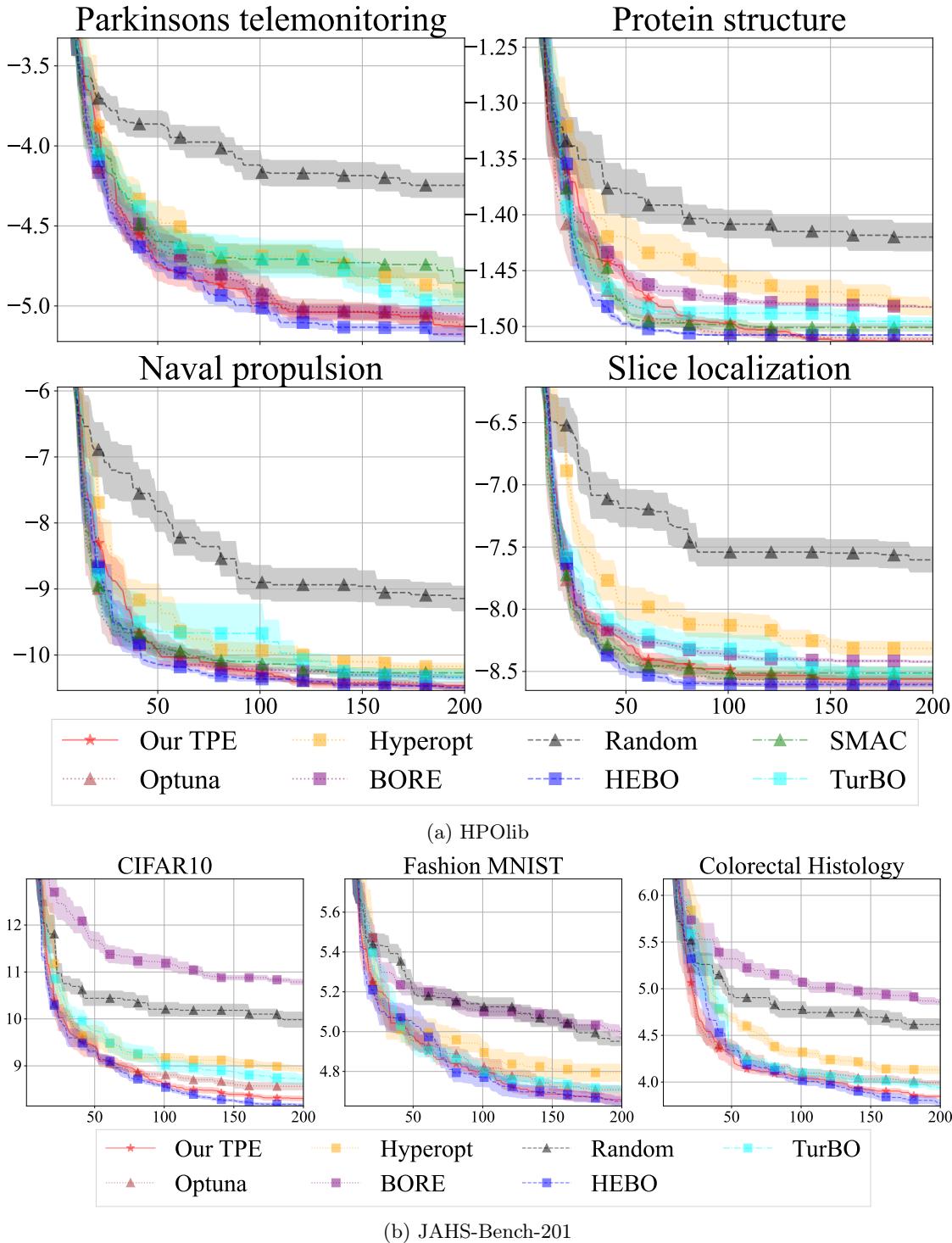


Figure 45: The comparison of optimization methods on the HPO benchmarks. The  $x$ -axis is the number of evaluations and the  $y$ -axis is the cumulative minimum objective value (for HPOlib, we took the log-scale of validation MSE). Each optimization method was run with 10 different random seeds and the weak-color bands represent the standard error. Note that SMAC are omitted for JAHS-Bench-201 due to the package dependency issue.

## References

- Abe, K., Wang, Y., & Watanabe, S. (2025). Tree-structured Parzen estimator can solve black-box combinatorial optimization more efficiently. *arXiv:2507.08053*.
- Addison, H., Inversion, K., Ryan, H., & Ted, C. (2022). Happywhale - whale and dolphin identification..
- Aitchison, J., & Aitken, C. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63.
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *International Conference on Knowledge Discovery & Data Mining*.
- Alina, J., Phil, C., Rodrigo, B., & Victor, G. (2019). Open images 2019 - object detection..
- Balandat, M., Karrer, B., Jiang, D., Daulton, S., Letham, B., Wilson, A., & Bakshy, E. (2020). BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. In *Advances in Neural Information Processing Systems*.
- Bansal, A., Stoll, D., Janowski, M., Zela, A., & Hutter, F. (2022). JAHS-Bench-201: A foundation for research on joint architecture and hyperparameter search. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*.
- Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2).
- Bergstra, J., Komera, B., Eliasmith, C., Yamins, D., & Cox, D. (2015). Hyperopt: a Python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8.
- Bergstra, J., Yamins, D., & Cox, D. (2013a). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International Conference on Machine Learning*.
- Bergstra, J., Yamins, D., Cox, D., et al. (2013b). Hyperopt: A Python library for optimizing the hyperparameters of machine learning algorithms. In *Python in Science Conference*, Vol. 13.
- Brochu, E., Cora, V., & de Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv:1012.2599*.
- Chen, Y., Huang, A., Wang, Z., Antonoglou, I., Schrittwieser, J., Silver, D., & de Freitas, N. (2018). Bayesian optimization in AlphaGo. *arXiv:1812.06855*.
- Cowen-Rivers, A., Lyu, W., Tutunov, R., Wang, Z., Grosnit, A., Griffiths, R., Maraval, A., Jianye, H., Wang, J., Peters, J., et al. (2022). HEBO: pushing the limits of sample-efficient hyper-parameter optimisation. *Journal of Artificial Intelligence Research*, 74.

- Dong, X., & Yang, Y. (2020). NAS-Bench-201: Extending the scope of reproducible neural architecture search. *arXiv:2001.00326*.
- Eggensperger, K., Müller, P., Mallik, N., Feurer, M., Sass, R., Klein, A., Awad, N., Lindauer, M., & Hutter, F. (2021). HPOBench: A collection of reproducible multi-fidelity benchmark problems for HPO. *arXiv:2109.06716*.
- Eriksson, D., Pearce, M., Gardner, J., Turner, R., & Poloczek, M. (2019). Scalable global optimization via local Bayesian optimization. *Advances in Neural Information Processing Systems*.
- Falkner, S., Klein, A., & Hutter, F. (2018). BOHB: Robust and efficient hyperparameter optimization at scale. In *International Conference on Machine Learning*.
- Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated Machine Learning*, pp. 3–33. Springer.
- Feurer, M., Springenberg, J., & Hutter, F. (2015). Initializing Bayesian hyperparameter optimization via meta-learning. In *Association for the Advancement of Artificial Intelligence*.
- Garnett, R. (2022). *Bayesian Optimization*. Cambridge University Press.
- Gonzalvez, J., Lezmi, E., Roncalli, T., & Xu, J. (2019). Financial applications of Gaussian processes and Bayesian optimization. *arXiv:1903.04841*.
- Hansen, N. (2016). The CMA evolution strategy: A tutorial. *arXiv:1604.00772*.
- Hickman, R., Parakh, P., Cheng, A., Ai, Q., Schrier, J., Aldeghi, M., & Aspuru-Guzik, A. (2023). Olympus, enhanced: benchmarking mixed-parameter and multi-objective optimization in chemistry and materials science..
- Hutter, F., Hoos, H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*.
- Hvarfner, C., Stoll, D., Souza, A., Lindauer, M., Hutter, F., & Nardi, L. (2022).  $\pi$ BO: Augmenting acquisition functions with user beliefs for Bayesian optimization. *arXiv:2204.11051*.
- Jones, D., Schonlau, M., & Welch, W. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13.
- Klein, A., & Hutter, F. (2019). Tabular benchmarks for joint architecture and hyperparameter optimization. *arXiv:1905.04970*.
- Kushner, H. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise..
- Li, C., de Celis Leal, D. R., Rana, S., Gupta, S., Sutti, A., Greenhill, S., Slezak, T., Height, M., & Venkatesh, S. (2017a). Rapid Bayesian optimisation for synthesis of short polymer fiber materials. *Scientific Reports*, 7.
- Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017b). Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18.

- Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J., & Stoica, I. (2018). Tune: A research platform for distributed model selection and training. *arXiv:1807.05118*.
- Lindauer, M., Eggensperger, K., Feurer, M., Biedenkapp, A., Deng, D., Benjamins, C., Ruhkopf, T., Sass, R., & Hutter, F. (2022). SMAC3: A versatile Bayesian optimization package for hyperparameter optimization. *Journal of Machine Learning Research*, 23.
- Loshchilov, I., & Hutter, F. (2016). CMA-ES for hyperparameter optimization of deep neural networks. *arXiv:1604.07269*.
- Müller, S., & Hutter, F. (2021). TrivialAugment: Tuning-free yet state-of-the-art data augmentation. In *International Conference on Computer Vision*.
- Nelder, J., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7.
- Nomura, M., Watanabe, S., Akimoto, Y., Ozaki, Y., & Onishi, M. (2021). Warm starting CMA-ES for hyperparameter optimization. In *Association for the Advancement of Artificial Intelligence*.
- Oliveira, R., Tiao, L., & Ramos, F. (2022). Batch Bayesian optimisation via density-ratio estimation with guarantees. *arXiv:2209.10715*.
- Ozaki, Y., Takenaga, S., & Onishi, M. (2022a). Global search versus local search in hyperparameter optimization. In *Congress on Evolutionary Computation*.
- Ozaki, Y., Tanigaki, Y., Watanabe, S., Nomura, M., & Onishi, M. (2022b). Multiobjective tree-structured Parzen estimator. *Journal of Artificial Intelligence Research*, 73.
- Ozaki, Y., Tanigaki, Y., Watanabe, S., & Onishi, M. (2020). Multiobjective tree-structured Parzen estimator for computationally expensive optimization problems. In *Genetic and Evolutionary Computation Conference*.
- Pfisterer, F., Schneider, L., Moosbauer, J., Binder, M., & Bischl, B. (2022). Yahpo gym – an efficient multi-objective multi-fidelity benchmark for hyperparameter optimization. In *International Conference on Automated Machine Learning*.
- Salinas, D., Seeger, M., Klein, A., Perrone, V., Wistuba, M., & Archambeau, C. (2022). Syne Tune: A library for large scale hyperparameter tuning and reproducible research. In *International Conference on Automated Machine Learning*.
- Schneider, P., Walters, W., Plowright, A., Sieroka, N., Listgarten, J., Goodnow, R., Fisher, J., Jansen, J., Duca, J., Rush, T., et al. (2020). Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery*, 19.
- Scott, D. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.
- Shahriari, B., Swersky, K., Wang, Z., Adams, R., & de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104.
- Silverman, B. (2018). *Density estimation for statistics and data analysis*. Routledge.
- Song, J., Yu, L., Neiswanger, W., & Ermon, S. (2022). A general recipe for likelihood-free Bayesian optimization. In *International Conference on Machine Learning*.

- Tiao, L., Klein, A., Seeger, M., Bonilla, E., Archambeau, C., & Ramos, F. (2021). BORE: Bayesian optimization by density-ratio estimation. In *International Conference on Machine Learning*.
- Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z., & Guyon, I. (2021). Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In *Advances in Neural Information Processing Systems Competition and Demonstration Track*.
- Vahid, A., Rana, S., Gupta, S., Vellanki, P., Venkatesh, S., & Dorin, T. (2018). New Bayesian-optimization-based design of high-strength 7xxx-series alloys from recycled aluminum. *Jom, 70*.
- Watanabe, S., Awad, N., Onishi, M., & Hutter, F. (2022). Multi-objective tree-structured Parzen estimator meets meta-learning. *Meta-learning Workshop at Advances in Neural Information Processing Systems*.
- Watanabe, S., Awad, N., Onishi, M., & Hutter, F. (2023a). Speeding up multi-objective hyperparameter optimization by task similarity-based meta-learning for the tree-structured Parzen estimator. *arXiv:2212.06751*.
- Watanabe, S., Bansal, A., & Hutter, F. (2023b). PED-ANOVA: Efficiently quantifying hyperparameter importance in arbitrary subspaces. In *arXiv:2304.10255*.
- Watanabe, S., & Hutter, F. (2022). c-TPE: Generalizing tree-structured Parzen estimator with inequality constraints for continuous and categorical hyperparameter optimization. *Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems Workshop at Advances in Neural Information Processing Systems*.
- Watanabe, S., & Hutter, F. (2023). c-TPE: Tree-structured Parzen estimator with inequality constraints for expensive hyperparameter optimization. *arXiv:2211.14411*.
- Williams, C., & Rasmussen, C. (2006). *Gaussian processes for machine learning*, Vol. 2. MIT press.
- Xue, D., Balachandran, P., Hogden, J., Theiler, J., Xue, D., & Lookman, T. (2016). Accelerated search for materials with targeted properties by adaptive design. *Nature Communications, 7*.