HPO × ELA: Investigating Hyperparameter Optimization Landscapes by Means of Exploratory Landscape Analysis

Lennart Schneider[1(✉)] , Lennart Schäpermeier[2] , Raphael Patrick Prager[3] , Bernd Bischl[1] , Heike Trautmann[3,4] , and Pascal Kerschke[2]

[1] Chair of Statistical Learning and Data Science, LMU Munich, Munich, Germany
{lennart.schneider,bernd.bischl}@stat.uni-muenchen.de
[2] Big Data Analytics in Transportation, TU Dresden, Dresden, Germany
{lennart.schaepermeier,pascal.kerschke}@tu-dresden.de
[3] Data Science: Statistics and Optimization, University of Münster,
Münster, Germany
{raphael.prager,heike.trautmann}@wi.uni-muenster.de
[4] Data Management and Biometrics Group, University of Twente,
Enschede, Netherlands

Abstract. Hyperparameter optimization (HPO) is a key component of machine learning models for achieving peak predictive performance. While numerous methods and algorithms for HPO have been proposed over the last years, little progress has been made in illuminating and examining the actual structure of these black-box optimization problems. Exploratory landscape analysis (ELA) subsumes a set of techniques that can be used to gain knowledge about properties of unknown optimization problems. In this paper, we evaluate the performance of five different black-box optimizers on 30 HPO problems, which consist of two-, three- and five-dimensional continuous search spaces of the XGBoost learner trained on 10 different data sets. This is contrasted with the performance of the same optimizers evaluated on 360 problem instances from the black-box optimization benchmark (BBOB). We then compute ELA features on the HPO and BBOB problems and examine similarities and differences. A cluster analysis of the HPO and BBOB problems in ELA feature space allows us to identify how the HPO problems compare to the BBOB problems on a structural meta-level. We identify a subset of BBOB problems that are close to the HPO problems in ELA feature space and show that optimizer performance is comparably similar on these two sets of benchmark problems. We highlight open challenges of ELA for HPO and discuss potential directions of future research and applications.

Keywords: Hyperparameter optimization · Exploratory landscape analysis · Machine learning · Black-box optimization · Benchmarking

1   Introduction

In machine learning (ML), hyperparameter optimization (HPO) constitutes one of the most frequently used tools for improving the predictive performance of a model [3]. The goal of classical single-objective HPO is to find a hyperparameter configuration that minimizes the estimated generalization error. Generally, neither a closed-form mathematical representation nor analytic gradient information is available, making HPO a *black-box* optimization problem and evolutionary algorithms (EAs) and model-based optimizers good candidate algorithms. As a consequence, no prior information about the optimization landscape – which could allow comparisons of HPO and other black-box problems, or provide guidance regarding the choice of optimizer – is available. This also extends to automated ML (AutoML) [14], which builds upon HPO.

In contrast, in the domain of *continuous* black-box optimization, a sophisticated toolbox for landscape analysis and the characterization of their properties has been developed over the years. In exploratory landscape analysis (ELA), optimization landscape features are calculated from small samples of evaluated points from the original black-box problem. It has been shown in numerous studies that ELA feature sets capture relevant landscape characteristics and that they can be used for automated algorithm selection, improving upon the state-of-the-art selector [5,17]. Particularly well-studied are the functions from the black-box optimization benchmark (BBOB) [12].

Empirical studies [30,31] in the closely related area of algorithm configuration hint that performance landscapes often are rather benign, i.e., unimodal and convex, although this only holds for an aggregation over larger instance sets and their analysis does not allow further characterization of individual problem landscapes. There exists some work to circumvent HPO altogether, by automatically configuring an algorithm for a given problem instance [1,28]. However, these are limited to configuring optimization algorithms rather than ML models. In addition, they are often restricted in the number and type of variables they are able to configure. [26] apply fitness landscape analysis on AutoML landscapes, computing fitness distance correlations and neutrality ratios on various AutoML problems. They utilize these features only in an exploratory manner, characterizing the landscapes, without a link to optimizer performance, and cannot compare the analyzed landscapes to other black-box problems in a natural way. Similar work on fitness landscape analysis exists but focuses mostly on neural networks [6,35]. Some preliminary work [9] on the hyperparameters of a $(1+1)$-EA on a OneMax problem suggests that the ELA feature distribution of a HPO problem can be significantly different from other benchmark problems. Recently, [32] developed statistical tests for the deviation of loss landscapes from uni-modality and convexity and showed that loss landscapes of AutoML problems are highly structured and often uni-modal.

In this work, we characterize continuous HPO problems using ELA features, enabling comparisons between different black-box optimization problems and optimizers. Our main contributions are as follows:

1. We examine similarities and differences of HPO and BBOB problems by investigating the performance of different black-box optimizers.
2. We compute ELA features for all HPO and BBOB problems and demonstrate their usefulness in distinguishing between HPO and BBOB.
3. We demonstrate how HPO problems position themselves in ELA feature space on a meta-level by performing a cluster analysis on principle components derived from ELA features of HPO and BBOB problems and investigate performance differences of optimizers on HPO problems and BBOB problems that are close to the HPO problems in ELA feature space.
4. We discuss how ELA can be used for HPO in future work and highlight open challenges of ELA in the context of HPO.
5. We release code and data of all our benchmark experiments hoping to facilitate future research (which currently may be hindered due to the computationally expensive HPO black-box evaluations).

The remainder of this paper is structured as follows: Fundamentals for HPO and ELA are introduced in Sect. 2. The experimental setup is presented in Sect. 3, with the results regarding the algorithm performance and ELA feature space analysis in Sect. 4 and 5, respectively. Section 6 concludes this paper and offers future research directions.

2   Background

Hyperparameter Optimization. Hyperparameter optimization (HPO) methods aim to identify a well-performing hyperparameter configuration $\boldsymbol{\lambda} \in \tilde{\Lambda}$ for an ML algorithm $\mathcal{I}_{\boldsymbol{\lambda}}$ [3]. An ML *learner* or *inducer* $\mathcal{I}$ configured by hyperparameters $\boldsymbol{\lambda} \in \Lambda$ maps a data set $\mathcal{D} \in \mathbb{D}$ to a model $\hat{f}$, i.e., $\mathcal{I} : \mathbb{D} \times \Lambda \to \mathcal{H}, (\mathcal{D}, \boldsymbol{\lambda}) \mapsto \hat{f}$. $\mathcal{H}$ denotes the so-called hypothesis space, i.e., the function space to which a model belongs [3]. The considered search space $\tilde{\Lambda} \subset \Lambda$ is typically a subspace of the set of all possible hyperparameter configurations: $\tilde{\Lambda} = \tilde{\Lambda}_1 \times \tilde{\Lambda}_2 \times \cdots \times \tilde{\Lambda}_d$, where $\tilde{\Lambda}_i$ is a bounded subset of the domain of the i-th hyperparameter Λ_i. This $\tilde{\Lambda}_i$ can be either real, integer, or category valued, and the search space can contain dependent hyperparameters, leading to a possibly hierarchical search space. The classical (single-objective) HPO problem is defined as:

$$\boldsymbol{\lambda}^* \in \underset{\boldsymbol{\lambda} \in \tilde{\Lambda}}{\arg\min} \, \widehat{\mathrm{GE}}(\boldsymbol{\lambda}), \tag{1}$$

i.e., the goal is to minimize the estimated generalization error. This typically involves a costly resampling procedure that can take a significant amount of time, see [3] for further details. $\widehat{\mathrm{GE}}(\boldsymbol{\lambda})$ is a black-box function, as it generally has no closed-form mathematical representation, and analytic gradient information is generally not available. Therefore, the minimization of $\widehat{\mathrm{GE}}(\boldsymbol{\lambda})$ forms an *expensive black-box* optimization problem. In general, $\widehat{\mathrm{GE}}(\boldsymbol{\lambda})$ is only a stochastic estimate of the true unknown generalization error. Formally, $\widehat{\mathrm{GE}}(\boldsymbol{\lambda})$ depends on

the concrete inducer, a resampling strategy (e.g., cross-validation) and a performance metric, for more details see [3]. In the following, we use the logloss as performance metric:

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(-\sum_{k=1}^{g} \sigma_k \left(y^{(i)} \right) \log \left(\hat{\pi}_k \left(\mathbf{x}^{(i)} \right) \right) \right). \tag{2}$$

Here, g is the total number of classes, $\sigma_k \left(y^{(i)} \right)$ is 1 if y is class k, and 0 otherwise (multi-class one-hot encoding), and $\hat{\pi}_k \left(\mathbf{x}^{(i)} \right)$ is the estimated probability for observation $\mathbf{x}^{(i)}$ belonging to class k.

Exploratory Landscape Analysis. The optimization landscapes of black-box functions, by design, carry no prior problem information, beyond the definition of their search parameters, which can be used for their characterization. In the continuous domain, ELA [23] addresses this problem by computing features on a small sample of evaluated points, which can be used for better understanding optimizer performance [24], algorithm selection [17] and even algorithm configuration [28].

The original ELA features consist, e.g., of meta model features (`ela_meta`) such as adjusted R^2 values for quadratic and linear models and y-distribution features (`ela_distr`) such as the skewness and kurtosis of the objective values. Over time, researchers continued to propose further feature sets, including nearest better clustering (`nbc`) [16] and dispersion (`disp`) [22] features to measure multi-modality, and information content (`ic`) features [25], which extract features from random walks across the problem landscape. The R package `flacco` [18] and Python package `pflacco` [27] implement a collection of the most widely used ELA feature sets.

ELA studies often focus on the noiseless BBOB functions, as they offer diverse, well-understood challenges (such as conditioning and multimodality) and a wide range of algorithm performance data is readily available. BBOB consists of 24 minimization problems, which are identified by their function ID (FID) and scalable with respect to their dimensionality, which ranges from 2 to 40. Furthermore, different instances, identified by instance IDs (IIDs), are defined for each function, creating slightly different optimization problems with the same fundamental characteristics by means of randomized transformations in the decision and objective space. All D-dimensional BBOB problems share a decision space of $[-5, 5]^D$, which is guaranteed to contain the (known) optimum.

3   Experimental Setup

We compare the following optimizers: `CMAES` (a simple CMA-ES with $\sigma_0 = 0.5$ and no restarts), `GENSA` (a generalized simulated annealing approach as described in [37]), `Grid` (a grid search performed by generating a uniform sized grid over the search space and evaluating configurations of the grid in random order),

Random (random search performed by sampling configurations uniformly at random), and MBO (Bayesian optimization using a Gaussian process as surrogate model and expected improvement as acquisition function [15], similarly configured as in [20]). All optimizers were given a budget of $50D$ function evaluations in total (where D is the dimensionality of the problem). All optimizer runs were replicated 10 times. We choose these optimizers for the following reasons: (1) they cover a wide range of optimizers that can be used for a black-box problem, (2) Grid and especially Random are frequently used for HPO and Random often can be considered a strong baseline [2].

As HPO problems, we tune XGBoost[1] [8] on ten different OpenML [36] data sets (classification tasks) chosen from the OpenML-CC18 benchmarking suite [4]. The specific data sets were chosen to cover a variety of the number of classes, instances, and features (cf. Table 1). To reduce noise as much as possible, performance (logloss) is estimated via 10-fold cross-validation with a fixed instantiating per data set. On each data set, we create $2, 3$ and 5 dimensional XGBoost problems by tuning nrounds, eta ($2D$), lambda ($3D$), gamma and alpha ($5D$), resulting in 30 problems in total. We selected these hyperparameters because (1) they can be incorporated in a purely continuous search space which is generally required for the computation of ELA features, (2) they have been shown to be influential on performance [29] and (3) have a straightforward interpretation, i.e., nrounds controls the number of boosting iterations (typically increasing performance but also the tendency to overfit) while the other hyperparameters counteract overfitting and control various aspects of regularization. The full search space is described in Table 2. Note that nrounds is tuned on a logarithmic scale and therefore all parameters are treated as continuous during optimization. Missing values of numeric features were imputed using Histogram imputation (values are drawn uniformly at random between lower and upper histogram breakpoints with cells being sampled according to the relative frequency of points contained in a cell). Missing values of factor variables were imputed by adding a new factor level and factor variables were encoded using one-hot-encoding. While XGBoost is a practically relevant learner we do have to note that only considering a single learner is somewhat restrictive. We discuss this limitation in Sect. 6. In the following, individual HPO problems are abbreviated by <name>_<d>, i.e., wilt_2 for the $2D$ wilt problem.

As BBOB problems we select FIDs 1–24 with IIDs 1–5 with a dimensionality of $\{2, 3, 5\}$, resulting in 360 problems in total. We abbreviate individual BBOB problems by <fid>_<iid>_<dim>, i.e., 24_1_5 for FID 24 with IID 1 in the $5D$ setting. Experiments have been conducted in R [33], where the individual implementation of an optimizer is referenced in the mlr3 ecosystem [19]. The package smoof [7] provides the aforementioned BBOB problems. We release all data and code for running the benchmarks and analyzing results via the following GitHub repository: https://github.com/slds-lmu/hpo_ela. HPO benchmarks took around 2.2 CPU years on Intel Xeon E5-2670 instances, with optimizer overhead ranging from 10% (MBO for $5D$) to less than 1% (Random or Grid).

---

[1] Using a gbtree booster.

Table 1. OpenML data sets.

ID	Name	Number of		
		Cl.	Inst.	Feat.
40983	wilt	2	4839	5
469	analcatdata_dmft	6	797	4
41156	ada	2	4147	48
6332	cylinder-bands	2	540	37
23381	dresses-sales	2	500	12
1590	adult	2	48842	14
1461	Bank-marketing	2	45211	16
40975	car	4	1728	6
41146	sylvine	2	5124	20
40685	shuttle	7	58000	9

IDs correspond to OpenML data set IDs, which enable to query data set properties via https://www.openml.org/d/<id>.

Table 2. XGBoost search space.

Hyper-param.	Type	Range	Trafo
nrounds	int.	$[3, 2000]$	log
eta	cont.	$[\exp(-7), \exp(0)]$	log
lambda	cont.	$[\exp(-7), \exp(7)]$	log
gamma	cont.	$[\exp(-10), \exp(2)]$	log
alpha	cont.	$[\exp(-7), \exp(7)]$	log

"log" in the Trafo column indicates that this parameter is optimized on a (continuous) logarithmic scale, i.e., the range is given by [log(lower), log(upper)], and values are re-transformed via the exponential function prior to their evaluation. Parameters part of the full XGBoost search space that are not shown are set to their default.

4   Optimizer Performance

For each BBOB problem, we computed optimizer rankings based on the average final performance (best target value of an optimizer run averaged over replications). Figures 1a to 1c visualize the differences in rankings on the BBOB problems split for the dimensionality. Friedman tests indicated overall significant differences in rankings ($2D$: $\chi^2(4) = 154.55, p < 0.001$, $3D$: $\chi^2(4) = 219.16, p < 0.001$, $5D$: $\chi^2(4) = 258.69, p < 0.001$). We observe that `MBO` and `CMAES` perform well throughout all three dimensionalities, whereas `GENSA` only is significantly better than `Grid` or `Random` for dimensionalities 3 and 5. Moreover, `Grid` only falls behind `Random` for the $5D$ problems.

Figures 1d to 1f analogously visualize differences in rankings on the HPO problems split for the dimensionality. Friedman tests indicated overall significant differences in rankings ($2D$: $\chi^2(4) = 36.32, p < 0.001$, $3D$: $\chi^2(4) = 34.32, p < 0.001$, $5D$: $\chi^2(4) = 34.80, p < 0.001$). Again, `MBO` and `CMAES` perform well throughout all three dimensionalities. Notably, `GENSA` shows lacklustre performance regardless of the dimensionality, failing to outperform `Grid` or `Random`. Similarly as on the BBOB problems, `Grid` tends to fall behind `Random` for the higher-dimensional problems. We do want to note that critical difference plots for the HPO problems are somewhat underpowered when compared to the BBOB problems due to the difference in the number of benchmark problem which results in larger critical distances, as seen in the figures.

In Fig. 2, we visualize the anytime performance of optimizers by the mean normalized regret averaged over replications split for the dimensionality of problems. The normalized regret is defined for an optimizer trace on a benchmark problem as the distance of the current best solution to the overall best solution found across all optimizers and replications, scaled by the overall range of empirical solution values for this benchmark problem. We choose this metric due to the theoretical optimal solutions being unknown for HPO problems, and apply it to both BBOB and HPO problems to enable performance comparisons. We
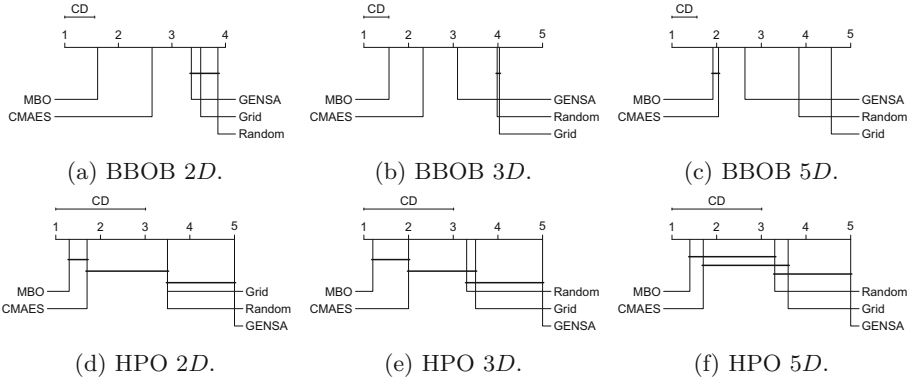
Fig. 1. Critical differences plots for mean ranks of optimizers on BBOB and HPO problems split with respect to the dimensionality.
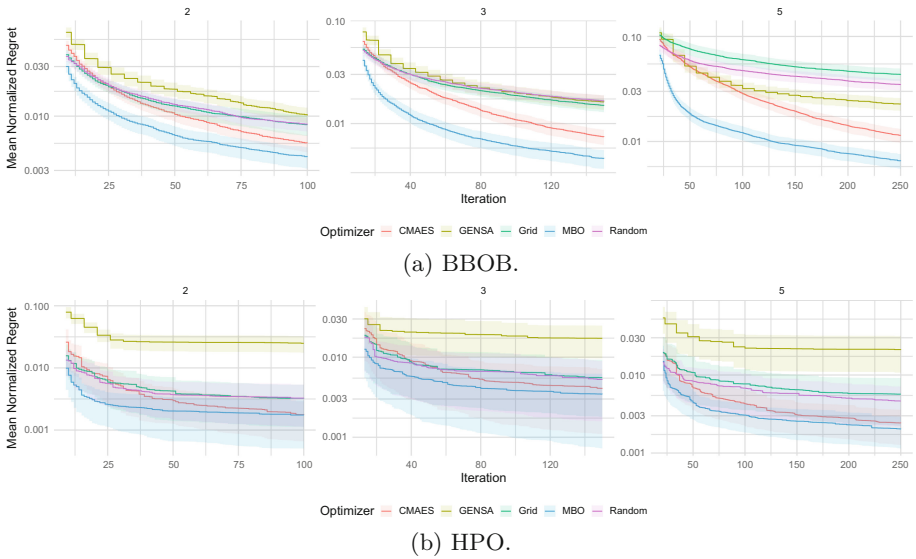

Fig. 2. Anytime mean normalized regret of optimizers on BBOB and HPO problems averaged over replications split for the dimensionality of problems. Ribbons represent standard errors. The x-axis starts after 8% of the optimization budget has been used (initial `MBO` design).

observe strong anytime performance of `MBO` and `CMAES` on both BBOB and HPO problems regardless their dimensionality. `GENSA` shows good performance on the 5D BBOB problems but shows poor anytime performance on HPO problems in general. Differences in anytime performance are less pronounced on the HPO problems, although we do want to note that the width of the standard error ribbons is strongly influenced by the number of benchmark problems.

(a) BBOB.

(b) HPO.
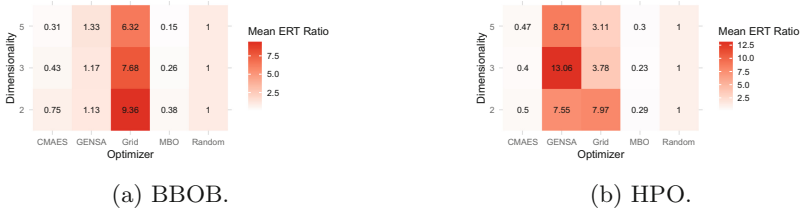
Fig. 3. Average ERT ratios (optimizers to `Random`) for HPO and BBOB problems.

As an additional performance evaluation, we calculated the Expected Running Time (ERT) [11]. In essence, for a given algorithm and problem, the ERT is defined as $\text{ERT} = \frac{1}{n}\sum_{i=1}^{10} \text{FE}_i$, where n is the number of repetitions which are able to reach a specific target, i refers to an individual repetition, and FE_i denotes the number of function evaluations used. We investigated the ERT of optimizers with the target given as the median of the best `Random` solutions (using $50D$ evaluations) over the ten replications per benchmark problem. We choose this (for BBOB unusual) target due to (1) the theoretical optimum of HPO problems being unknown and (2) `Random` being considered a strong baseline in HPO [2]. To bring all ERTs on the same scale, we computed the ERT ratios between optimizers and `Random` per benchmark problem which further allows us to aggregate these ratios over benchmark problems[2]. We visualize these aggregated ERT ratios separately for the dimensionality of benchmark problems in Fig. 3. We observe that average ERT ratios of `MBO` and `CMAES` are comparably similar for BBOB and HPO problems although the tendency that these optimizers become even more efficient with increasing dimensionality is less pronounced on the HPO problems. `Grid` generally falls behind and `GENSA` shows lacklustre performance on HPO.

5   ELA Feature Space Analysis

For each HPO and BBOB problem, we use $50D$ points sampled by LHS (Min-Max) as an initial design for computing ELA features. We normalize the search space to the unit cube and standardize objective function values per benchmark problem $((y - \hat{\mu})/\hat{\sigma})$ prior to calculating ELA features. This is done to counter potential artefacts that could be seen in ELA features solely due to different value ranges in decision and, in particular, in objective space. We calculate the feature sets `ela_meta`, `ic`, `ela_distr`, `nbc` and `disp`, which were introduced in Sect. 2, using the `flacco` R package [18].

To answer the question whether ELA can be used to distinguish HPO from BBOB problems, we construct a binary classification task using ELA features to predict the label "HPO" vs. "BBOB". We use a decision tree and estimate

---

[2] Following [17], optimizers that did not meet the target in any run were assigned an ERT of the worst ERT on a benchmark problem multiplied by a factor of 10.
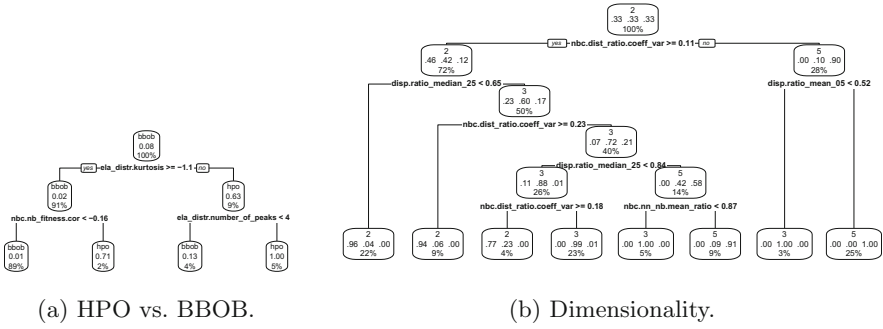
(a) HPO vs. BBOB.   (b) Dimensionality.

Fig. 4. Decision trees for classifying benchmark problems into HPO or BBOB problems (left) and classifying the dimensionality of BBOB problems (right).

the generalization error via 10 times repeated 10-fold cross-validation (stratified for the target). We obtain an estimated classification error of 3.54%. Figure 4a illustrates the decision tree obtained after training on all data. We observe that only few ELA features are needed to correctly classify problems: HPO problems tend to exhibit a lower `ela_distr.kurtosis` combined with more `ela_distr.number_of_peaks` or show a higher `nbc.nb_fitness.cor` than BBOB problems if the first split with respect to the kurtosis has not been affirmed. This finding is supported by visualizations of the 2D HPO problems, which we present in our online appendix, i.e., most 2D HPO problems have large plateaus resulting in negative kurtosis.

To answer the question whether dimensionality is a different concept for HPO compared to BBOB problems[3] we perform the following analysis: We construct a classification task using ELA features to predict the dimensionality of the problem but only use the BBOB subset for the training of a decision tree. We estimate the generalization error via 10 times repeated 10-fold cross-validation (stratified for the target) and obtain an estimated classification error of 7.39%. We then train the decision tree on all BBOB problems (illustrated in Fig. 4b) and determine the holdout performance on the HPO problems and obtain a classification error of 10%. Only few ELA features of the `disp` and `nbc` group are needed to predict the dimensionality of problems with high accuracy. Intuitively, this is sensible, due to `nbc` features involving the calculation of distance metrics (which themselves should be affected by the dimensionality) and both `nbc` and `disp` features being sensible to the multimodality of problems [16,22] which should also be affected by the dimensionality. Based on the reasonable good hold-out performance of the classifier on the HPO problems, we conclude that "dimensionality" is a similar concept for BBOB and HPO problems.

---

[3] For HPO problems, it is a priori often unclear whether a change in a parameter value also results in relevant objective function changes, i.e., the intrinsic dimensionality of a HPO problem may be lower than the number of hyperparameter suggests.
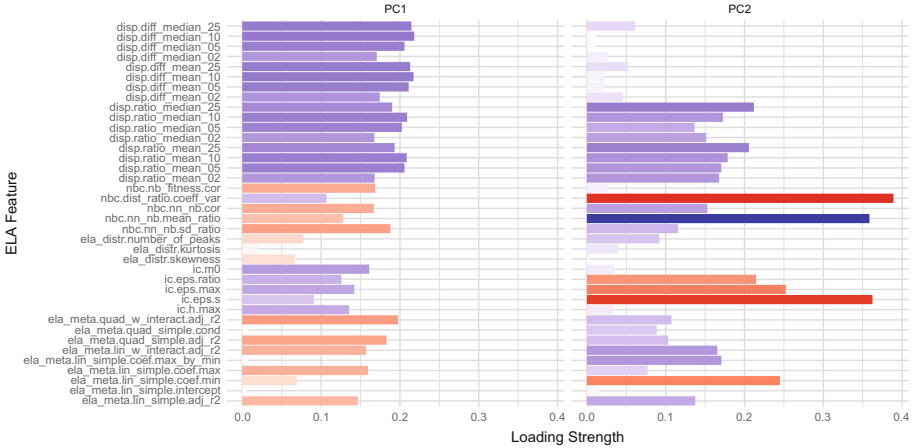
Fig. 5. Factor loadings of ELA features on the first two principle components. Blue indicates a positive loading, whereas red indicates a negative loading.

To gain insight on a meta-level, we performed a PCA on the scaled and centered ELA features of both the HPO and BBOB problems. To ease further interpretation, we select a two component solution that explains roughly 60% of the variance. Figure 5 summarizes factor loadings of ELA features on the first two principle components. Most `disp` features show a medium positive loading on PC1, whereas some `nbc` show medium negative loadings. `ela_meta` features, including R^2 measures of linear and quadratic models, also exhibit medium negative loadings on PC1. We therefore summarize PC1 as a latent dimension that mostly reflects multimodality of problems. Regarding PC2, three features stand out with strong loadings: `nbc.dist_ratio.coeff_var`, `nbc.nn_nb.mean_ratio` and `ic.eps.s`. Moreover, `disp.ratio_*` features generally have a medium negative loading. We observe that all features used by the decision tree in Fig. 4b also have comparably large loadings on PC2. Therefore, we summarize PC2 as an indicator of the dimensionality of problems.

We then performed k-means clustering on the two scaled and centered principal component scores. A silhouette analysis suggested the selection of three clusters. In Fig. 6, we visualize the assignment of HPO and BBOB problems to these three clusters. Labels represent IDs of BBOB and HPO problems. We observe that the dimensionality of problems is almost perfectly reflected in the PC2 alignment. Cluster 2 and 3 can be mostly distinguished along PC2 (cluster 3 contains low dimensional problems and cluster 2 contains higher dimensional problems) whereas cluster 1 contains problems with large PC1 values. HPO problems are exclusively assigned to cluster 2 or 3, exhibiting low variance with respect to their PC1 score, with the PC1 values indicating low multimodality.

As a final analysis we determined the nearest BBOB neighbors of the HPO problems (in ELA feature space based on the cluster analysis, i.e., minimizing the Euclidean distance over the first two principal component scores). For a
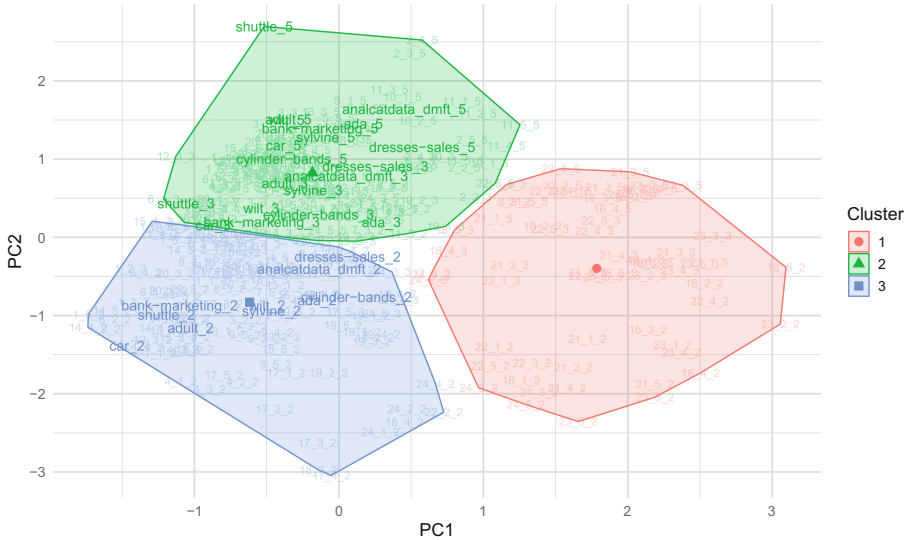
Fig. 6. Cluster analysis of BBOB and HPO problems on the first two principle component scores in ELA feature space.

complete list, see our online appendix. We again computed optimizer rankings based on the average final performance of the optimizers (over the replications), but this time for all HPO problems (regardless their dimensionality) and the subset of BBOB problems that are closest to the HPO problems in ELA feature space (see Fig. 7). Friedman tests indicated overall significant differences in rankings for both HPO ($\chi^2(4) = 104.99, p < 0.001$) and nearest BBOB ($\chi^2(4) = 61.01, p < 0.001$) problems. We observe similar optimizer rankings, with MBO and CMAES outperforming Random or Grid, indicating that closeness in ELA feature space somewhat translates to optimizer performance. Nevertheless, we do have to note that GENSA exhibits poor performance on the HPO problems compared to the nearest BBOB problems. We hypothesize that this may be caused by the performance of GENSA being strongly influenced by its hyperparameter configuration itself and provide an initial investigation in our online appendix.
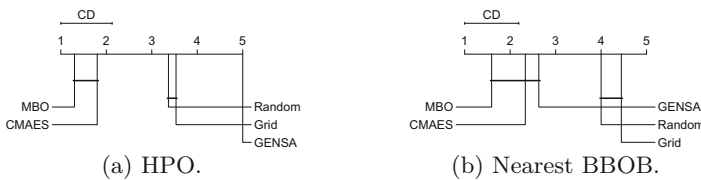

(a) HPO.   (b) Nearest BBOB.

Fig. 7. Critical differences plots for mean ranks of optimizers on all HPO problems (left) and the subset of nearest BBOB problems.

6 Conclusion

In this paper, we characterized the landscapes of continuous hyperparameter optimization problems using ELA. We have shown that ELA features can be used to (1) accurately distinguish HPO from BBOB problems and (2) classify the dimensionality of problems. By performing a cluster analysis in ELA feature space, we have shown that our HPO problems mostly position themselves with BBOB problems of little multimodality, mirroring the results of [30,32]. Determining the nearest BBOB neighbor of HPO problems in ELA feature space allowed us to investigate performance differences of optimizers with respect to HPO problems and their nearest BBOB problems and we observed comparably similar performance. We believe that this work is an important first step in identifying BBOB problems that can be used in lieu of real HPO problems when, for example, configuring or developing novel HPO methods.

Our work still has several limitations. A major one is that traditional ELA is only applicable to continuous HPO problems, which constitute a minority of real-world problems. In many practical applications, search spaces include categorical and conditionally active hyperparameters – so-called hierarchical, mixed search spaces [34]. In such scenarios, measures such as the number of local optima, fitness-distance correlation or auto-correlation of fitness along a path of a random walk [10,13] can be used to gain insight into the fitness landscape. Another limitation is that our studied HPO problems all stem from tuning XGBoost, with little variety of comparably low dimensional search spaces, which limits the generalizability of our results.

In future work, we would like to extend our experiments to cover a broader range of HPO settings, in particular different learners and search spaces, but also data sets. We also want to reiterate that HPO is generally noisy and expensive. In our benchmark experiments, costly 10-fold cross-validation with a fixed instantiating per data set was employed to reduce noise to a minimal level. Future work should explore the effect of the variance of the estimated generalization error on the calculation and usage of ELA features which poses a serious challenge for ELA applied to HPO in practice. Besides, we used logloss as a performance metric which by definition is rather "smooth" compared to other metrics such as the classification accuracy (but the concrete choice of performance metric typically depends on the concrete application at hand). Moreover, ELA requires the evaluation of an initial design, which is very costly in the context of HPO. In general, HPO often can be performed with evaluations on multiple fidelity levels, i.e., by reducing the size of training data, and plenty of HPO methods make use of this resulting in significant speed-up [21]. Future work could explore the possibility of using low fidelity evaluations for the initial design required by ELA and how multiple fidelity levels of HPO affect ELA features.

We consider our work as pioneer work and hope to ignite the research interest in studying the landscape properties of HPO problems going beyond fitness measures. We envision that, by improved understanding of HPO landscapes and identifying relevant landscape properties, better optimizers may be designed, and eventually instance-specific algorithm selection and configuration for HPO may be enabled.

References

1. Belkhir, N., Dréo, J., Savéant, P., Schoenauer, M.: Per instance algorithm configuration of CMA-ES with limited budget. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 681–688 (2017)
2. Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems, pp. 2546–2554 (2011)
3. Bischl, B., et al.: Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. arXiv:2107.05847 [cs, stat] (2021)
4. Bischl, B., et al.: OpenML benchmarking suites. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, vol. 1 (2021)
5. Bischl, B., Mersmann, O., Trautmann, H., Preuß, M.: Algorithm selection based on exploratory landscape analysis and cost-sensitive learning. In: Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation, pp. 313–320 (2012)
6. Bosman, A.S., Engelbrecht, A.P., Helbig, M.: Progressive gradient walk for neural network fitness landscape analysis. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion, pp. 1473–1480 (2018)
7. Bossek, J.: smoof: Single- and multi-objective optimization test functions. R J. **9**(1), 103–113 (2017). https://journal.r-project.org/archive/2017/RJ-2017-004/index.html
8. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
9. Doerr, C., Dreo, J., Kerschke, P.: Making a case for (hyper-)parameter tuning as benchmark problems. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion, pp. 1755–1764 (2019)

10. Hains, D.R., Whitley, L.D., Howe, A.E.: Revisiting the big valley search space structure in the TSP. J. Oper. Res. Soc. **62**(2), 305–312 (2011). https://doi.org/10.1057/jors.2010.116

11. Hansen, N., Auger, A., Finck, S., Ros, R.: Real-parameter black-box optimization benchmarking 2010: Experimental setup. Research Report RR-7215, Inria (2010). https://hal.inria.fr/inria-00462481

12. Hansen, N., Finck, S., Ros, R., Auger, A.: Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Technical report RR-6829, Inria (2009). https://hal.inria.fr/inria-00362633/document

13. Hernando, L., Mendiburu, A., Lozano, J.A.: An evaluation of methods for estimating the number of local optima in combinatorial optimization problems. Evol. Comput. **21**(4), 625–658 (2013)

14. Hutter, F., Kotthoff, L., Vanschoren, J. (eds.): Automated Machine Learning. TSSCML, Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05318-5

15. Jones, D.R., Schonlau, M., Welch, W.J.: Efficient global optimization of expensive black-box functions. J. Global Optim. **13**(4), 455–492 (1998). https://doi.org/10.1023/A:1008306431147

16. Kerschke, P., Preuss, M., Wessing, S., Trautmann, H.: Detecting funnel structures by means of exploratory landscape analysis. In: Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation, pp. 265–272 (2015)

17. Kerschke, P., Trautmann, H.: Automated algorithm selection on continuous blackbox problems by combining exploratory landscape analysis and machine learning. Evol. Comput. **27**(1), 99–127 (2019)

18. Kerschke, P., Trautmann, H.: Comprehensive feature-based landscape analysis of continuous and constrained optimization problems using the R-package Flacco. In: Bauer, N., Ickstadt, K., Lübke, K., Szepannek, G., Trautmann, H., Vichi, M. (eds.) Applications in Statistical Computing. SCDAKO, pp. 93–123. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-25147-5_7

19. Lang, M., et al.: mlr3: a modern object-oriented machine learning framework in R. J. Open Source Softw. **4**(44), 1903 (2019)

20. Le Riche, R., Picheny, V.: Revisiting Bayesian optimization in the light of the COCO benchmark. Struct. Multidiscip. Optim. **64**(5), 3063–3087 (2021). https://doi.org/10.1007/s00158-021-02977-1

21. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A.: Hyperband: A novel bandit-based approach to hyperparameter optimization. J. Mach. Learn. Res. **18**(1), 6765–6816 (2017)

22. Lunacek, M., Whitley, D.: The dispersion metric and the CMA evolution strategy. In: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, pp. 477–484 (2006)

23. Mersmann, O., Bischl, B., Trautmann, H., Preuss, M., Weihs, C., Rudolph, G.: Exploratory landscape analysis. In: Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation, pp. 829–836 (2011)

24. Mersmann, O., Preuss, M., Trautmann, H., Bischl, B., Weihs, C.: Analyzing the BBOB results by means of benchmarking concepts. Evol. Comput. **23**(1), 161–185 (2015)

25. Muñoz, M.A., Kirley, M., Halgamuge, S.K.: Exploratory landscape analysis of continuous space optimization problems using information content. IEEE Trans. Evol. Comput. **19**(1), 74–87 (2014)

26. Pimenta, C.G., de Sá, A.G.C., Ochoa, G., Pappa, G.L.: Fitness landscape analysis of automated machine learning search spaces. In: Paquete, L., Zarges, C. (eds.) EvoCOP 2020. LNCS, vol. 12102, pp. 114–130. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-43680-3_8

27. Prager, R.P.: pflacco: A Python Interface of the R Package Flacco, April 2022. https://github.com/Reiyan/pflacco

28. Prager, R.P., Trautmann, H., Wang, H., Bäck, T.H.W., Kerschke, P.: Per-instance configuration of the modularized CMA-ES by means of classifier chains and exploratory landscape analysis. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 996–1003. IEEE (2020)

29. Probst, P., Boulesteix, A.L., Bischl, B.: Tunability: Importance of hyperparameters of machine learning algorithms. J. Mach. Learn. Res. **20**(53), 1–32 (2019)

30. Pushak, Y., Hoos, H.: Algorithm configuration landscapes: In: Auger, A., Fonseca, C.M., Lourenço, N., Machado, P., Paquete, L., Whitley, D. (eds.) PPSN 2018, Part II. LNCS, vol. 11102, pp. 271–283. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99259-4_22

31. Pushak, Y., Hoos, H.H.: Golden parameter search: Exploiting structure to quickly configure parameters in parallel. In: Proceedings of the 2020 Genetic and Evolutionary Computation Conference, pp. 245–253 (2020)

32. Pushak, Y., Hoos, H.H.: AutoML landscapes. ACM Trans. Evol. Learn. Optim. (TELO) (2022, in print). https://www.cs.ubc.ca/labs/algorithms/Projects/ACLandscapes/PusHoo22a.pdf

33. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021). https://www.R-project.org/

34. Thornton, C., Hutter, F., Hoos, H.H., Leyton-Brown, K.: Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 847–855 (2013)

35. Traoré, K.R., Camero, A., Zhu, X.X.: Fitness landscape footprint: A framework to compare neural architecture search problems. arXiv:2111.01584 [cs] (2021)

36. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: OpenML: Networked science in machine learning. SIGKDD Explor. **15**(2), 49–60 (2013)

37. Xiang, Y., Gubian, S., Suomela, B., Hoeng, J.: Generalized simulated annealing for global optimization: The GenSA package. R J. **5**(1), 13–28 (2013)