A Benchmark with Artificial Functions

We evaluate existing MBO techniques on some popular artificial functions which were also used in the original ParEGO paper by Knowles [31]. KNO1 [39], OKA1 [39] and VLMOP2 [49] are popular benchmark functions with two input and two output dimensions. VLMOP3 [49] is a function with two input and three output dimensions. For each function we let each algorithm iteratively evaluate 100 points and compute the dominated hyper-volume of the Pareto front at each step with respect to a reference point which is determined as described in [31].

Figure 3 visualizes the average dominated hyper-volume and variance for 5 random seeds. There are rather clear differences in performance between the individual methods. The more advanced MBO algorithms seem to outperform the simpler methods although RS and RW are competitive on the artificial functions KNO1 and VLMOP2. The average per iteration time between the individual methods varies largely and is provided in Table 2. EHI scales very poorly with the number of objectives and we aborted the runs on VLMOP3 after 24h. We note that the two-dimensional input of these artificial functions has a much smaller dimension compared to the 7- and 10-dimensional hyperparameter spaces used in our HPO experiments.

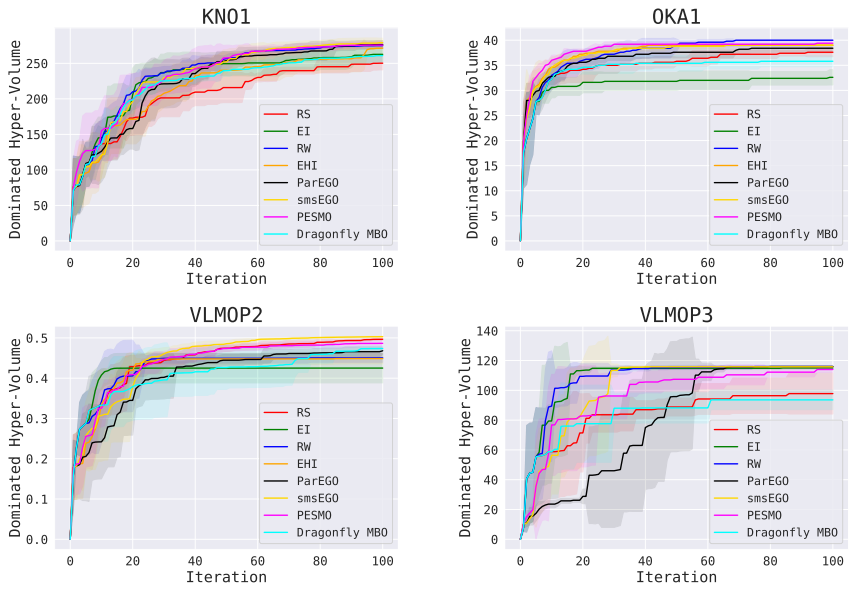

Figure 3: Dominated hyper-volume values for various multi-objective Bayesian optimization algorithms achieved on common artificial functions over 100 iterations. The average and standard deviation for 5 random seeds is shown. The more advanced methods have an advantage over the simpler ones. RS and RW are surprisingly competitive on KNO1 and VLMOP2.

Table 2: Average per iteration time over 100 steps in seconds. The time is solely dominated by the time it takes to determine the next evaluation candidate. The per iteration time varies largely between the individual methods. The EHI runs for 3-objective problem VLMOP3 were aborted after 24h.

Function	RS	EI	RW	ParEGO	PESMO	smsEGO	EHI
KNO1	0.01	2.50	2.86	18.52	31.46	276.36	327.79
OKA1	0.01	2.78	2.74	19.37	39.50	278.02	372.88
VLMOP2	0.01	1.97	2.48	23.17	42.94	305.19	802.11
VLMOP3	0.01	2.35	2.14	25.02	60.40	414.38	-

B Benchmark with FairML Tasks

Adding to Section 6, we provide more details about the experimental setup and visualize additional results. A detailed overview of the hyperparameter spaces used for the MLP and XGBoost classifiers is given by Table 3 and 4 respectively. Results for an experiment comparing various MBO methods on two fairness related datasets are visualized in Figure 4. Figure 5 visualizes dominated hyper-volume over time for XGB classifiers which are optimized for error and DSP. Figure 6 and 7 visualize experimental results for a 3 objective setting with error, DSP and DEO objective and a 4 objective setting with error, DSP, DEO and DFP objective respectively.

Table 3: Sklearn MLP search space

Parameter	Type	Domain	Scaling
n_layers	integer	$\{1, 2, 3, 4\}$	linear
layer_1	integer	$\{2, \ldots, 32\}$	logarithmic
layer_2	integer	$\{2, \ldots, 32\}$	logarithmic
layer_3	integer	$\{2, \ldots, 32\}$	logarithmic
layer_4	integer	$\{2, \ldots, 32\}$	logarithmic
alpha	real	$[10^{-6}, \ldots, 10^{-1}]$	logarithmic
learning_rate_init	real	$[10^{-6}, \ldots, 10^{-2}]$	logarithmic
beta_1	real	$[0.001, 0.99]$	logarithmic
beta_2	real	$[0.001, 0.99]$	logarithmic
tol	real	$[10^{-5}, 10^{-2}]$	logarithmic

Table 4: XGBoost search space

Parameter	Type	Domain	Scaling
n_estimators	integer	$\{1, 2, ..., 256\}$	logarithmic
learning_rate	real	$[0.01, 1.0]$	logarithmic
gamma	real	$[0.0, 0.1]$	linear
reg_alpha	real	$[10^{-3}, 10^{3}]$	logarithmic
reg_lambda	real	$[10^{-3}, 10^{3}]$	logarithmic
subsample	real	$[0.01, 1.0]$	linear
max_depth	integer	$\{1, 2, \ldots, 16\}$	linear

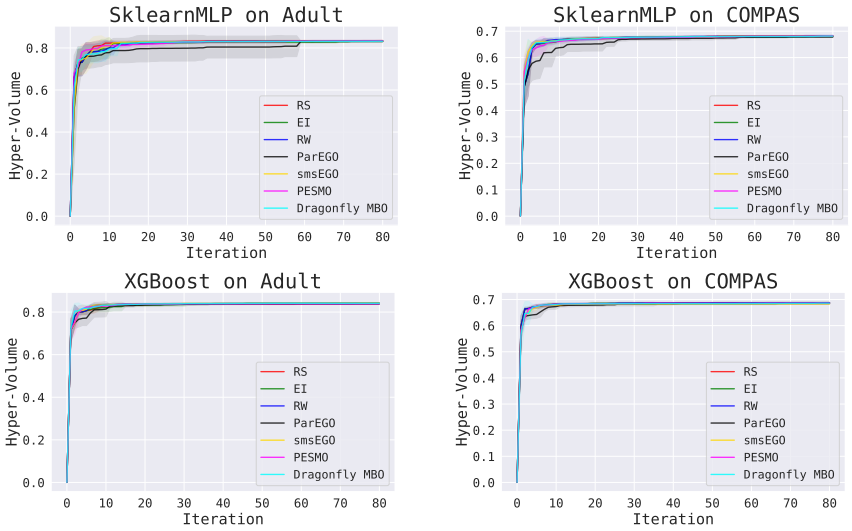

Figure 4: Dominated hyper-volume for MLP and XGBoost classifiers under error and DSP objectives on Adult and COMPAS dataset. The average and standard deviation for 5 random seeds is shown. Among the different approaches, the quality of the generated approximations is very similar, with RS and RW being surprisingly competitive

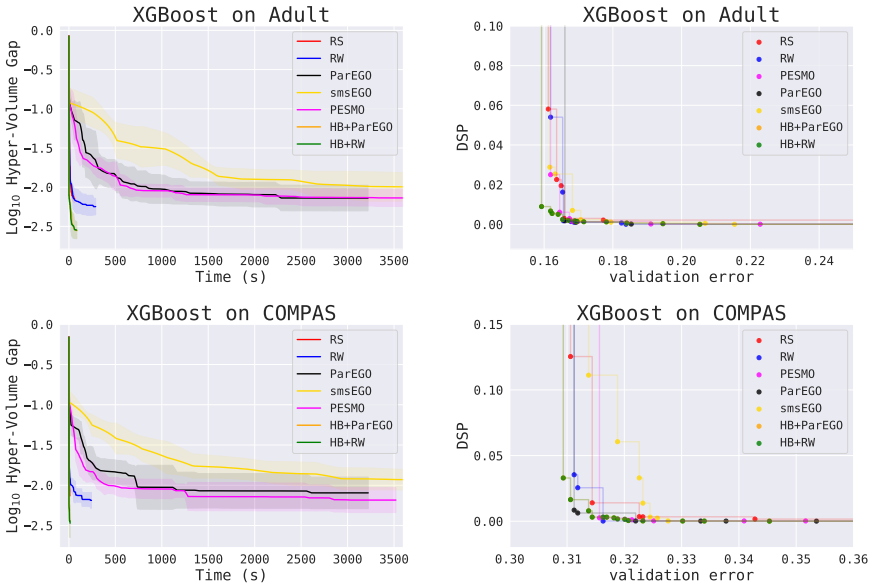

Figure 5: [Left] Dominated hyper-volume of the Pareto front approximations of XGBoost classifiers over time under error and DSP objective on Adult and COMPAS dataset. The average and standard deviation for 5 random seeds is shown. [Right] Corresponding Pareto front approximations. For both scalarization schemes, our method obtains Pareto front approximations with larger hyper-volume in a significantly shorter wall-clock time.

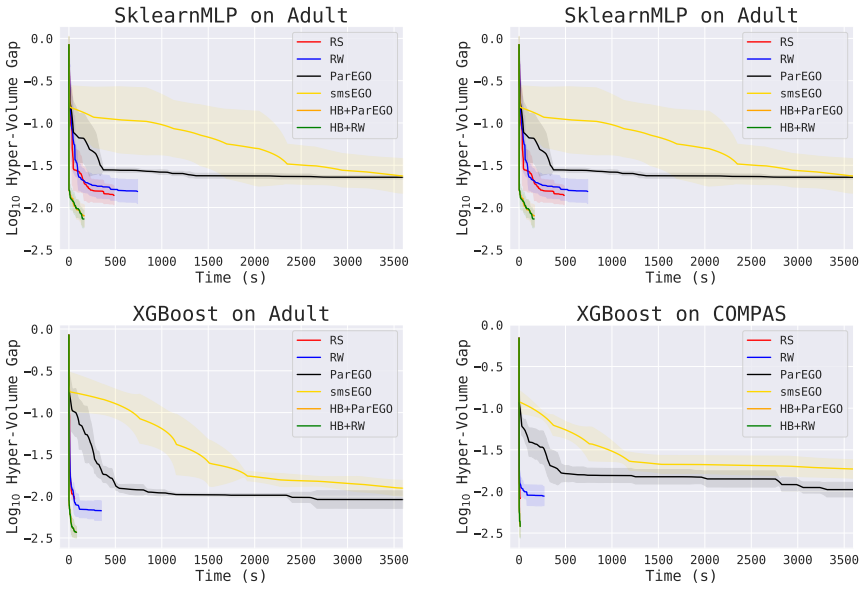

Figure 6: Dominated hyper-volume of the Pareto front approximations of MLP and XGBoost classifiers over time under error, DSP and DEO objective on Adult and COMPAS dataset. The average and standard deviation for 5 random seeds is shown. For both scalarization schemes, our method obtains Pareto front approximations with larger hyper-volume in a significantly shorter wall-clock time. With increasing problem dimension smsEGO requires a larger computational budget to determine the next candidate configuration at each step.

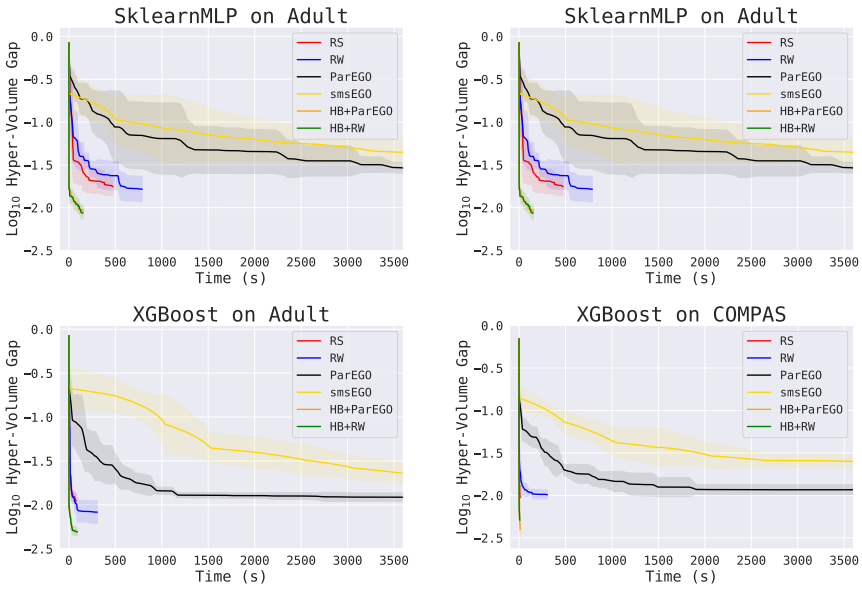

Figure 7: Dominated hyper-volume of the Pareto front approximations of MLP and XGBoost classifiers over time under error, DSP, DEO and DFP objective on Adult and COMPAS dataset. The average and standard deviation for 5 random seeds is shown. For both scalarization schemes, our method obtains Pareto front approximations with larger hyper-volume in a significantly shorter wall-clock time. With increasing problem dimension smsEGO requires a larger computational budget to determine the next candidate configuration at each step.