



Universiteit
Leiden
The Netherlands

Empirical analysis of the dynamic binary value problem with IOHprofiler

Vermetten, DL.; Lengler, J.; Rusin, D.; Bäck, T.H.W.; Doerr, C.; Affenzeller, M.; ... ; Machado, P.

Citation

Vermetten, D. L., Lengler, J., Rusin, D., Bäck, T. H. W., & Doerr, C. (2024). Empirical analysis of the dynamic binary value problem with IOHprofiler. *Lecture Notes In Computer Science*, 20-35. doi:10.1007/978-3-031-70068-2_2

Version: Publisher's Version

License: [Licensed under Article 25fa Copyright Act/Law \(Amendment Taverne\)](#)

Downloaded from: <https://hdl.handle.net/1887/4254858>

Note: To cite this publication please use the final published version (if applicable).

Empirical Analysis of the Dynamic Binary Value Problem with IOHprofiler

Diederick Vermetten¹, Johannes Lengler², Dimitri Rusin³, Thomas Bäck¹, Carola Doerr³

¹ LIACS, Leiden University, Leiden, The Netherlands

² ETH Zürich, Zürich, Switzerland

³ Sorbonne Université, CNRS, LIP6, Paris, France

Abstract. Optimization problems in dynamic environments have recently been the source of several theoretical studies. One of these problems is the monotonic Dynamic Binary Value problem, which theoretically has high discriminatory power between different Genetic Algorithms. Given this theoretical foundation, we integrate several versions of this problem into the IOHprofiler benchmarking framework. Using this integration, we perform several large-scale benchmarking experiments to both recreate theoretical results on moderate dimensional problems and investigate aspects of GA’s performance which have not yet been studied theoretically. Our results highlight some of the many synergies between theory and benchmarking and offer a platform through which further research into dynamic optimization problems can be performed.

Keywords: Evolutionary Algorithms · Benchmarks · Dynamic Environment · Dynamic Binary Value.

1 Introduction

Evolutionary and genetic algorithms (EAs, GAs) are an important family of randomized optimization heuristics. In order to better understand the behavior of these algorithms, we should take advantage of the wide range of perspectives from which they have been studied, including theoretical runtime analysis, systematic benchmarking studies, and practical experience. While each of these domains offers its own viewpoint, these should not stay isolated but rather benefit from and strengthen each other.

In the last years, there has been a substantial amount of theoretical work on two seemingly unrelated situations: optimization of monotonic functions [5], and optimization in dynamic linear environments [9,4] like dynamic binary value [7]. For both of these topics, theoretical analysis shows that even among simple EAs, performance can dramatically differ. Hence, these situations have high discriminative power between different EAs. To give a flavour of the theoretical results, we highlight two specific results here:

1. The $(1 + 1)$ -EA with mutation rate $1/N$ finds the optimum of the dynamic binary value environment in time $O(N \log N)$, while the same algorithm with mutation rate $2/N$ needs exponential time [9].
2. The $(1 + 1)$ -EA with mutation rate $1/N$ finds the optimum of every monotonic function in time $O(N \log^2 N)$ [6], while the $(\mu + 1)$ -EA with the same mutation rate and large population size μ needs exponential time on some hard instances [11].

These results illustrate the strengths, but also the limitations of theoretical analyses. They can identify interesting differences between similar algorithms, and may be able to prove results on whole classes of benchmarks like all monotonic functions. Moreover, they are able to provide a deep understanding of the given situation. On the other hand, they are often limited to rather simple algorithms, and to asymptotic statements like “ $O(N \log N)$ ” or “exponential”. Thus they leave important gaps that can be filled with systematic benchmarking studies.

In this paper we provide such a benchmarking study for two aforementioned situations of monotonic and of dynamic linear environments. To this end, we have several contributions:

- We develop practical variants of the theoretical benchmarks that have been used for runtime analyses. This requires non-trivial modifications, as the theoretical variants are not suited for efficient evaluations, see Section 1.2.

- We have integrated them into IOHexperimenter [16], a module of the open-source IOHprofiler [3] project that allows general access to the new benchmarks in a framework where they can be easily used to test other iterative optimization algorithms. Part of this integration is a generic extension of IOHexperimenter to support dynamic environments.
- We run a systematic evaluation of a large class of algorithms on the new benchmarks, identifying the parameters which are generally most crucial for the performance on these benchmarks.
- We investigate several of the theoretical results in a wider context, asking how big the effects are for small dimensionality N , how sensitive they are for other parameter settings and how well they transfer to modifications of the algorithms.

Since this paper brings together many different benchmarking aspects, we discuss them in more detail in the following sections.

1.1 Dynamic environments

As mentioned, we have extended the IOHprofiler framework [3] to support dynamic environments, i.e., environments in which the fitness function f can change over time. In our experiments, whenever the environment changes, we re-evaluate the population with respect to the new environment, to avoid comparing fitnesses from different environments with each other. However, the implementation also supports keeping the old fitness values.

Dynamic environments can occur in various ways through a range of applications [1]. For example, part of the problem description may be uncertain or become only available over time [15,14]. This work is based on dynamic linear functions [9,7,4] and the dynamic binary value function [8,7], which were motivated by shifting training data for a machine learning system, especially in the context of co-evolution [7].

1.2 Theory-inspired benchmarks

In recent years, there has been increasing interest and demand for composing a problem suite with theoretical benchmarks that are suitable for empirical testing. We provide a class of such benchmarks, based on two ideas from theory: dynamic binary value for dynamic environments and the HOTTOPIC function as an instance of a hard monotonic function.

The dynamic binary value function DBV is based on the static linear binary value function $BV : \{0,1\}^N \rightarrow \mathbb{R}; BV(x) = \sum_{i=1}^N 2^{i-1}x_i$, which computes for a bit string x the integer encoded by x in binary representation. In the dynamic setting, we use the same set of weights, but shuffle them via a random permutation. Note that the same permutation is applied for all strings that are evaluated in the same environment. Hence, every environment is given by a linear function with positive weights. In particular, all environments share the all-one string as their common global optimum, which gives the optimization process a clear objective: it should produce solutions which are as close as possible to the invariant global optimum. Moreover, the functions are never deceptive in any coordinate: whenever we flip a zero-bit into a one-bit, then the fitness increases in all possible environments. This ensures that the benchmark is feasible and can be solved efficiently by some algorithms and parameter configurations, but not by all. This yields the high discriminatory power of the benchmark.

Practically, the BV function can not easily be implemented because the weights are so large that they cause numerical problems. We do provide an implementation via a lexicographic comparison, but since this requires an approach that is not compatible with the usual framework of fitness functions, it is less well-suited as a module to be used by other researchers. Instead, we use the observation from [7] that DBV can be obtained in the limit from dynamic linear functions where we draw all weights independently from some heavy-tailed distribution. In this work, we consider the Uniform, Power-of-two (with a maximum power) and capped Pareto distributions, and compare them to the version based on lexicographic comparison.

Moreover, our implementation leaves the freedom to either change the environment every generation, or to change it less frequently. This also gives a way to test a function similar to the HOTTOPIC function introduced in [10] (named HOTTOPIC in [5]). This function has served an important role as theoretical benchmark [5,11], but it is forbiddingly slow to evaluate practically.

However, it has been pointed out in [8] that HOTTOPIC may be approximated by a dynamic linear function in which the environment changes every εn generations, for some constant ε .

Thus, we provide several practical implementations of theoretical benchmarks. We empirically validate our implementations in Section 3 by testing whether previous theoretical analyses can be recovered empirically, and find generally a good qualitative agreement.

1.3 Theoretical results on the benchmarks

Here we review very briefly the known theoretical results on the related benchmarks. For both DBV and HOTTOPIC, the $(1 + 1)$ -EA is very well understood. For this simple algorithm, the main parameter is the mutation rate. For both benchmarks, there is a sharp phase transition in the mutation rate χ/N . There is a constant χ_0 , which is $\chi_0 = 2.13\dots$ for DBV [8] and $\chi_0 = 1.59\dots$ for HOTTOPIC [10,5]. If $\chi < \chi_0$ then the $(1 + 1)$ -EA finds the optimum in time $O(N \log N)$, otherwise it needs exponential time in N . The same dichotomy also holds for other dynamic linear functions, where all weights are drawn independently and identically distributed. The threshold χ_0 depends on the distribution and is always larger than the threshold 1.59.. for DBV. For example, the threshold is $\chi_0 = 2$ for an exponential distribution and $\chi_0 = (2 - p)/(1 - p)$ for a geometric distribution with parameter p .

For HOTTOPIC, it is known that this result generalizes to a large number of functions, including the $(1 + \lambda)$ -EA, the $(\mu + 1)$ -EA, the $(1 + \lambda)$ -fEA, the $(\mu + 1)$ -fEA, all without crossover, and the $(1 + (\lambda, \lambda))$ -GA, where for the last three the parameter χ must be substituted appropriately [5]. However, all these results come with an important caveat: they hold only sufficiently close to the optimum. This may seem like a harmless condition as usually the hardest part of optimization is close to the optimum. However, this is not generally the case for HOTTOPIC. It was shown in [11] that for $(\mu + 1)$ -EA without crossover, the hardest part of HOTTOPIC is not around the optimum, but that the algorithm instead gets stuck in a region “in distance εn ” from the optimum, even though it would efficiently find the optimum if started in distance $\varepsilon n/2$ from the optimum [11]. This shows abstractly the surprising fact that progress from εn to $\varepsilon n/2$ is harder than from $\varepsilon n/2$ to 0. However, the ε could not be explicitly specified. One advantage of our benchmarking approach is that we are able to replace the abstract value of ε by some concrete numbers, see Section 3.

Another interesting result on HOTTOPIC is that crossover helps dramatically close to the optimum: for every mutation rate χ there is μ_0 such that the $(\mu + 1)$ -GA using either mutation or crossover, each with probability 0.5, is efficient close to the optimum. It remained an open question how the behavior is further away from the optimum, where there are the opposing effects of being (counter-intuitively) in a harder region of the search space, and having the benefits of crossover.

For DBV, results other than for the $(1 + 1)$ -EA are a bit more sparse. The benchmark has been introduced in [7]. Experimental results for a very limited set of parameters indicated that also here the hardest part of optimization is not always next to the optimum, at least not for the $(2 + 1)$ EA and $(3 + 1)$ EA without crossover. For the $(2 + 1)$ EA this could also be shown formally in [8]. However, it can be ruled out that the reason is the same as for the $(\mu + 1)$ -EA on HOTTOPIC, since the latter effect crucially depends on the fact that HOTTOPIC functions maintain the order of weights for some period of time [7]. So even though there is a similar surprising behavior, the reasons must be different.

1.4 IOHprofiler

IOHprofiler [3] is a modular toolbox for benchmarking iterative optimization heuristics. It provides access to a variety of problem suites through a common interface in its IOHexperimenter [16] module. This allows for a flexible benchmarking pipeline, which includes data logging. This data can be processed and visualized in the IOHanalyzer [19] module.

By providing access to a wide variety of problem types with common logging infrastructure, IOHprofiler has enabled a range of different research questions, both from a theoretical and practical perspective. For example, recent works have shown that the star-discrepancy optimization problem provides a challenging environment for a large set of optimization heuristics [2], or created environments for competitions on several sets of submodular optimization problems [13].

2 Experimental Setup

2.1 DBV in IOExperimenter

Our experimental setup is built upon the integration of several versions of the Dynamic BinVal problem class into the IOExperimenter environment. To handle the dynamic nature of the problem, a way of progressing its internal state (importance of each variable), a *step*-function has to be added to the problem itself. Note that this means that in practice, the algorithm determines when the problem’s state is updated, which is needed since the problem itself has no ability to detect when an evaluation belongs to a new generation.

The Dynamic BinVal problem is based on the Binary Value problem defined via BinVal: $\{0, 1\}^n \rightarrow \mathbb{R}; x \mapsto \sum_{i=1}^n 2^{i-1} x_i$. In the dynamic version, we draw a permutation π for each generation and compute all fitnesses with respect to the permuted weights $\text{DBV}(x) = \text{DBV}_\pi(x) = \sum_{i=1}^n 2^{\pi(i)-1} x_i$. This version of the problem can not be implemented directly, as the weights would become too large for the standard datatypes used within IOExperimenter. As such, we follow the observation from [8] and draw weights independently from different distributions instead. In this context, we consider three distinct distributions:

- **Uniform:** Uniform Weights from $\mathcal{U}(0, 1)$
- **PowersOfTwo:** Uniformly from the set $\{2^i : 1 \leq i \leq 31 - \log_2(N)\}$ (to avoid overflow of the summed value)
- **Pareto:** Pareto distribution with a limited upper bound $((1 - \mathcal{U}(0, 0.75))^{-10})$, where 0.75 is chosen to avoid overflow of the summed value)

In addition to these 3 problem variants, we can also consider a ‘true’ version of the DBV (**Ranking**) by not implementing an evaluation, but a ranking function instead. This ranking function sorts the given individuals lexicographically according to the order of the weights, and thus does not require using the large powers of 2 directly. While this version is true to the problem, it requires any used optimization algorithm to be modified from objective-value evaluations to this ranking scheme, and thus does not give good modularity.

The integration of the 4 versions of DBV into IOExperimenter has the additional benefit that we can utilize several transformation methods to validate algorithmic invariances. These transformations, originally proposed in the context of the ‘PBO’ suite, include changing the target string with a random bitstring, swapping the order of variables and translating/scaling the returned objective value. In our experiments, when we refer to a problem instance, this corresponds to a function combined with a setting of these transformations.

2.2 Used Algorithm

In this paper, we make use of a standard GA which is heavily parameterized to allow for a variety of experimental setups. An outline of this algorithm is given in Algorithm 1, while the available parameters are indicated in Table 1. For our termination criterion, we use a combination of function evaluation budget ($100 \cdot N$ by default) and optimality (to save computation time).

2.3 Experimental Setup and Reproducibility

Throughout our experiments, we make use of $N = 1000$, and perform independent runs on 15 different instances for each configuration. Each run is given a budget of $1000 \cdot N$ evaluations. Our full experimental setup, including the data collection, processing and visualization, is available on our Zenodo repository [18]. This repository also includes the full datasets from our experiments in an IOHanalyzer-compatible format.

3 Results

3.1 Exploration of Used Algorithms

For our first set of experiments, we perform a coarse grid sampling of our parameter space to identify which factors most impact the algorithm’s performance. For this setting, performance is

Parameter	Range	Default	Notes
Mutation factor χ	$[0, N]$	1	Mutation rate is $\frac{\chi}{N}$
Number of offspring λ	$[1, \infty)$	1	
Number of parents μ	$[1, \infty)$	1	
Selection	plus, comma	plus	Number of generations
Crossover rate p_c	$[0, 1]$	0	
Dynamic frequency	$[0, \infty)$	1	
Minimum bits flipped n_{min}	$[0, N)$	0	Mutate crossover-result
Mutation after crossover	yes, no	yes	
Function Version	Rank, Uniform Power2, Pareto	Rank	Rank=lexicographic
Instance	$[0, \infty]$	$[1, 25]$	Determines transformations
Dimensionality N	$[1, \infty]$	1000	

Table 1: Parameters available in the used GA version (top) and parameterization of the DBV functions (bottom). In our experiments, we use the default value unless stated otherwise.

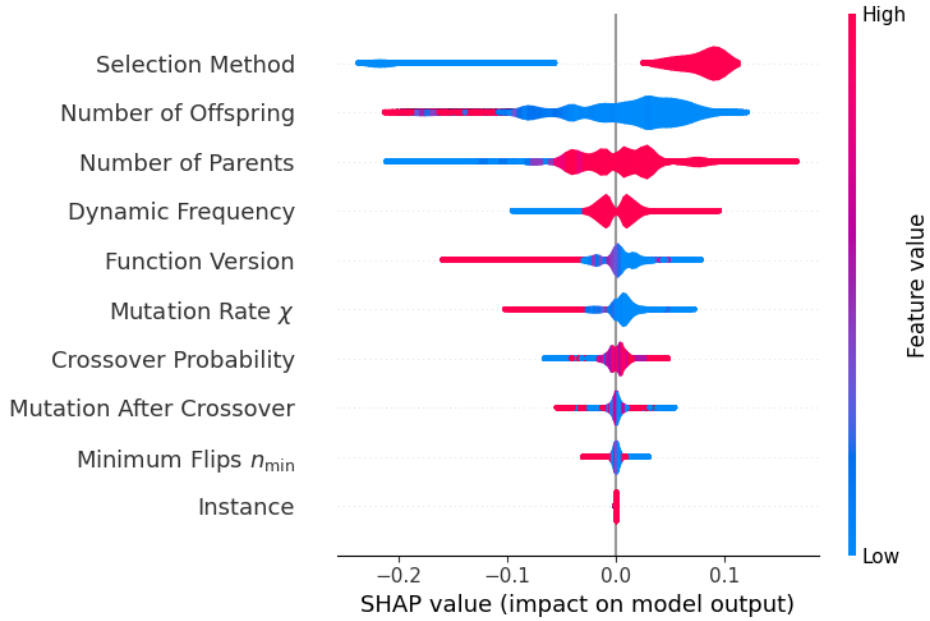


Fig. 1: Dimensionality 1000, SHAP-values for each of the varied parameter settings. For selection, red corresponds to plus-selection. For the function version, red corresponds to the rank-based version.

Algorithm 1 Outline of the used Genetic Algorithm.

```

1: Inputs: Parameters from Table 1, function  $f$  to maximize
2: procedure GA
3:   Initialize Population  $\mathcal{P}$  ▷ At random or with fixed distance to optimum
4:   while termination criterion not met do
5:     if Update required then ▷ Based on update frequency
6:        $f \rightarrow \text{step}()$ 
7:     end if
8:     for  $i$  in number of offspring do
9:       Perform crossover with probability  $p_c$ , creating  $\mathcal{O}_i$  ▷ Uniformly
10:      if No crossover or Mutation after cross disabled then
11:         $n_{\text{flip}} \leftarrow \max(n_{\min}, \text{Binom}(\frac{X}{N}, N))$ 
12:        Flip  $n_{\text{flip}}$  bits to create  $\mathcal{O}_i$ 
13:      end if
14:    end for
15:    Evaluate  $\mathcal{O}$  (and  $\mathcal{P}$  if required) ▷ Rank or evaluate based on function
16:     $\mathcal{P} \leftarrow (\mathcal{P}; \mathcal{O})$  ▷ Plus or comma selection
17:  end while
18: end procedure

```

measured as the percentage of correct bits after $100 \cdot N$ evaluations. To analyze the resulting data, we make use of the SHAP approach [12], which is a popular approach in the context of global explainability [17]. The resulting SHAP-values are shown in Figure 1, sorted from most impactful (top) to least (bottom). In this figure, the color of each dot corresponds to the chosen option, e.g. for selection blue corresponds to a comma while red corresponds to a plus-strategy.

While the grid used to create Figure 1 is rather coarse, it does provide an initial overview of the components of the algorithm which influence the selected performance measure in the most drastic ways. In particular, we notice that using a comma selection greatly deteriorates the ability to get close to the optimum, and that using a low number of parents and a large number of offspring is beneficial in this experimental setting. This generalizes the theoretical result in [11], where it was identified as a problem of the $(\mu + 1)$ -EA that the population is exchanged only slowly, and the family trees within the population become deep. Those properties are generally mitigated if the ratio of offspring versus parents is large. Also of note is the fact that the version of the function has quite some impact on the algorithm’s performance, while the instantiation of the function is negligible. This suggests that the used GA is invariant to the used instance transformations, but not to the used weight-distribution.

3.2 Mutation in 1+1

For our second experiment, we focus on the (1+1) GA and study the impact of mutation and the used function version on its performance. To achieve this, we vary the mutation rate between 0 and 6 (with increments of 0.1), as well as the minimum number of mutated bits between 0 and 1. This is repeated for each version of the DBV, with 25 independent runs of budget $1000 \cdot N$.

In Figure 2, we show the number of function evaluations required to reach the optimum for each of the selected GA versions. In Figure 3 we plot the fraction of correct bits after the evaluation budget has been exhausted. A dotted black line highlights the mutation rate of 1.6, which is the theoretical threshold for the ranking-based version, the solid blue line, in the case where it is allowed that mutation flips zero bits (the parent is duplicated). Also, theory predicts that there are no other versions with smaller thresholds [9]. Indeed, both theoretical predictions are confirmed in Figures 2 and 3.

When comparing the different versions of DBV in Figure 2, we observe quite a stark difference in algorithm performance between the Uniform-based DBV and the other settings. While for low mutation rates all considered versions result in the same behavior, which results from flipping 1 bit, which is always accepted if it was a zero-bit and rejected otherwise, with larger mutation rates the uniform problem becomes visibly easier than the rest. This is visible both in the number of evaluations required to reach optimality, as well as in the number of correct bits after the budget threshold is reached (Figure 3). We also see that both the PowersOfTwo and Pareto are very similar

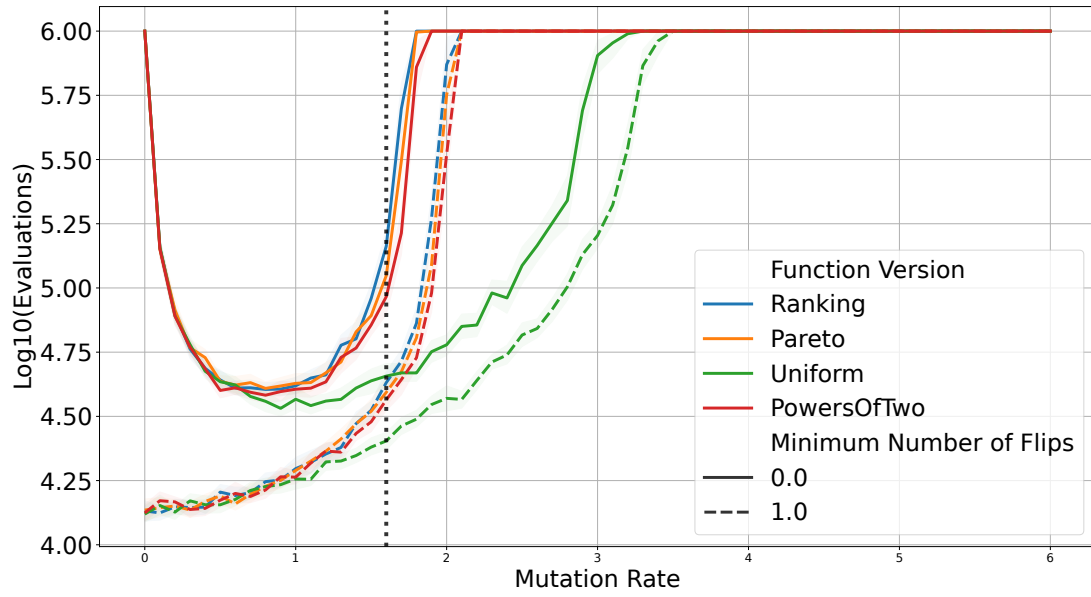


Fig. 2: Dimensionality 1000, number of evaluations (log-scaled) needed to reach optimum given different mutation rates.

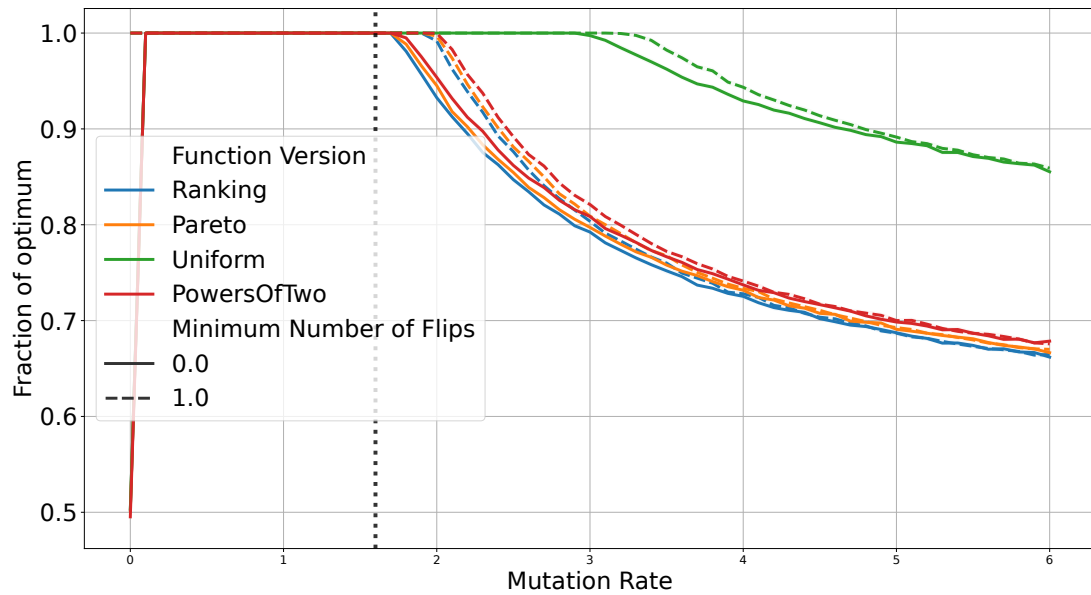


Fig. 3: Dimensionality 1000, relative number of correct bits after the budget (1000 times N) is used, given different mutation rates.

to the Rank-based DBV (with a slight preference for Pareto), so they seem to both be suitable to model the problem when function-evaluation-based approaches are required.

To understand the dashed lines, we observe that in the $(1+1)$ case, generations have no effect if the parent is duplicated (except for counting as an idle step). Hence, we should consider the number n_{flip} of flipped bits *conditioned* on flipping at least one bit. A bit counter-intuitively, the expectation $\mathbb{E}[n_{\text{flip}} \mid n_{\text{flip}} > 0]$ goes *down* by overwriting the case $n_{\text{flip}} = 0$ with $n_{\text{flip}} = 1$, which means that this change *decreases* the number of flipped bits in the conditional space. Effectively this change decreases the mutation rate, which is why the threshold is shifted to the right.

Figure 3 shows that the effect of a large mutation rate is quite dramatic. For example, for a mutation rate of $3/n$ the algorithm is only able to set less than 80% of the bits correctly after a very generous budget of $1000N$. This is still a large distance from the optimum. Even with mutation rate of $3/n$, the algorithm has a chance of roughly 15% (for $n_{\text{min}} = 0$) to flip exactly one bit in a mutation, compared to 37% for mutation rate $1/n$. Since those are the mutations that bring most progress in this range, the poor performance can not be explained by the moderate slowdown (factor 2-3) that larger mutation rates entail on other benchmarks like ONEMAX. The large effect shown in Figure 3 has not been quantified by theory before.

While the heavy aggregation of Figures 2 and 3 shows the final performance of the considered algorithm configurations, it does not show the full behavior of the underlying runs. Since we recorded the full performance trajectory, we can load this data into IOAnalyzer to generate more fine-grained visualization, such as ERT curves [19] illustrated in Figure 4. The ERT gives some estimates for the expected running time under some pessimistic assumptions on the running time in which the optimum is not reached [19]. It tends to overestimate the actual running times [7]. Again, this shows how suddenly the performance drops at a given level, and how little is to be gained by increasing the computation budget.

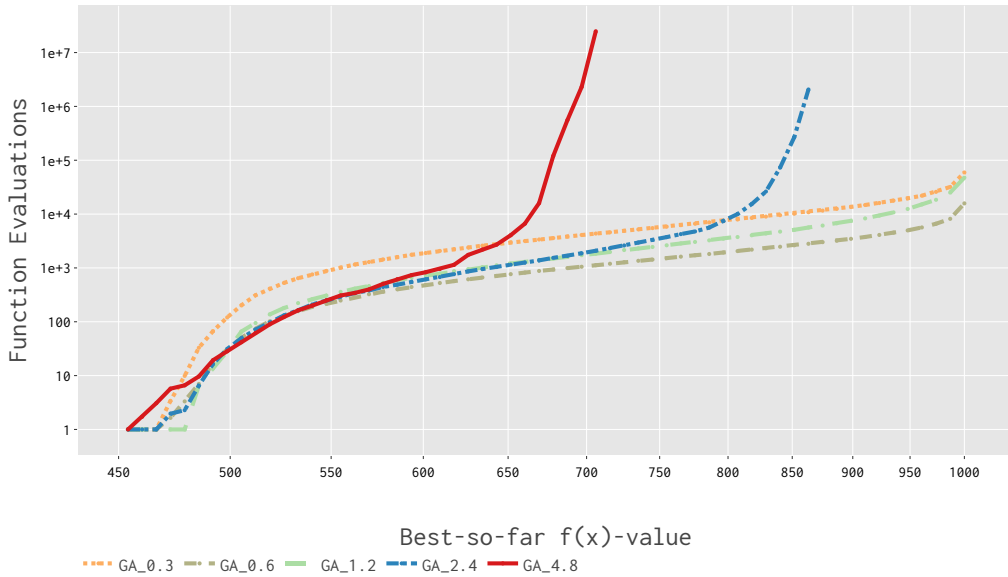


Fig. 4: Dimensionality 1000, some selected ERT curves for different mutation rates (rank-based function).

3.3 Adding multiple parents: Crossover

For our next set of experiments, we investigate the effects of crossover (uniform) on the performance of the $(\mu+1)$ -GA. We vary μ in $\{2, 3, 4, 8, 16\}$, and crossover rate in $\{0, 0.5, 0.9\}$. In addition to the crossover rate, we change whether an individual resulting from crossover still undergoes mutation or not. Mutation is not applied after crossover in the previous theory results.

To illustrate the impact of the different versions of crossover, we plot the number of evaluations until optimality for a given crossover rate 0.5 in Figure 5. In this figure, we can see a transition between these two mechanisms, where for lower mutation rates it is beneficial to still perform mutation after crossover, while disabling mutation once crossover is performed allows for more stable behavior at higher mutation rates. One perspective on this is that there is something like an optimal amount of mutation. If the mutation rate is smaller, then it may be beneficial to add mutation also after crossover. If the mutation rate is larger, then it is beneficial to not apply it in every generation.

A similar trade-off can be observed in Figure 5 for population sizes. While an increase in population size requires more function evaluations per iteration (each individual needs to be re-evaluated to perform selection), it also improves the ability of the algorithm to work with much higher mutation rates.

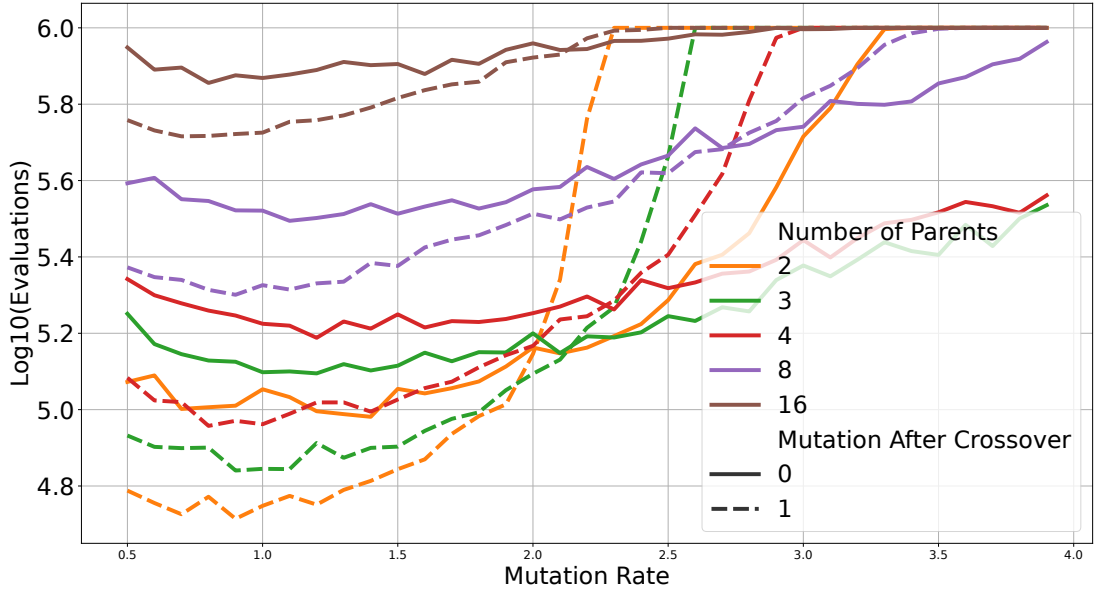


Fig. 5: Dimensionality 1000, number of evaluations (log-scaled) needed to reach optimum, on the rank-based functions, given different number of parents and mutation after crossover setting. Crossover probability is 0.5.

3.4 Difficulty of the Search Stages

A remarkable theoretical result is that the hardest region for optimization is not always around the optimum. This has been shown theoretically for the $(2+1)$ EA on DBV [8] and for the $(\mu+1)$ EA with large constant μ on HOTTOPIC [11]. Notably, the underlying cause is very different in both settings. In this section we want to investigate how relevant this phenomenon for DBV in practice, which has also been shown in some narrow-focused experiments in [7], for example for the $(3+1)$ EA for mutation rate $2.7/N$. To this end, we seeded the algorithm with bit strings of a given Hamming distance from the optimum (fraction of correct bits). By seeding the algorithm with strings ranging from 0.5 to 0.95 and measuring the time needed to improve this fraction by 0.05, we get an overview of the relative hardness of each stage of the optimization procedure. The results of this experiment are visualized in Figure 6.

In our experiments, we could not reproduce the finding that the hardest region was in some intermediate range. Comparing with the experiments in [7], we suspect that this effect only shows for very large budgets. Since we want to systematically explore different parameter setting in this paper, the required budget becomes forbiddingly large. We believe our results do show that the

effect may not be very common in practical setting, and we leave it to future work to nail down the difference further. However, in Figure 6 we see another interesting effect. As for the mutation rate, the number of parents has a trade-off between different regions of the search space. We generally see that larger population sizes are beneficial in “easy” regions of the search space, but are detrimental in “harder regions”. This suggests that in the setting of dynamic optimization and DBV, there may not be a fixed population size that is optimal throughout all phases of the optimization process.

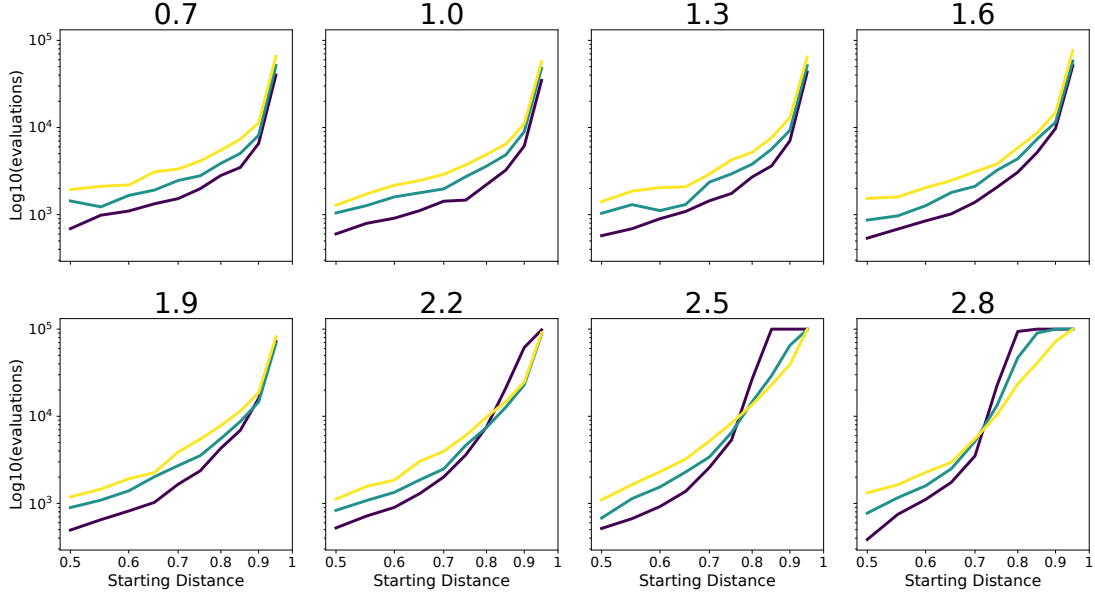


Fig. 6: Dimensionality 1000, number of evaluations (log-scaled) needed to improve the hamming distance to the optimum solution by 50, on the rank-based functions, given different number of parents (blue = 2, green = 3, yellow = 4) and mutation rate (plot titles). Crossover probability is 0.

3.5 HotTopic: impact of update frequency

While the DBV problem is defined as a fully dynamic problem where the environment changes every iteration of the algorithm, we may also change it less frequently. This has been conjectured to be related to the well-studied HOTTOPIC problem, which is a *static* monotonic problem, but where different regions of the search space behave like different linear functions.

To illustrate the impact of the function update frequency, we again show the number of evaluations required by our $(\mu, 1)$ -GA with different settings of μ . The results of this experiment, where the crossover rate is 0.9 and mutation after crossover is enabled, can be seen in Figure 7. From this figure, we observe that for $\mu > 1$ the difficulty of the problem increases with the update frequency up until some point, after which the problem difficulty drops again. With a large enough update frequency, the problem becomes essentially static, at which point the problem collapses to a static binary value problem. While this function is by folklore conjectured to be the hardest static linear function, the class of static linear functions is rather easy, and this includes the static binary value problem.

Interestingly, the intermediate range where the problem is hardest may correspond to HOTTOPIC functions, where informally the “environment” changes after εN generations, where ε is a small constant. There is a remarkable analogy to the theoretical results in [11]. There it was also shown that an intermediate regime is the hardest, but not for the update frequency, but rather for the distance from the optimum. However, the two quantities may be related in this context. In the HOTTOPIC analysis, the crucial observation was that if the weight structure is stable, then some

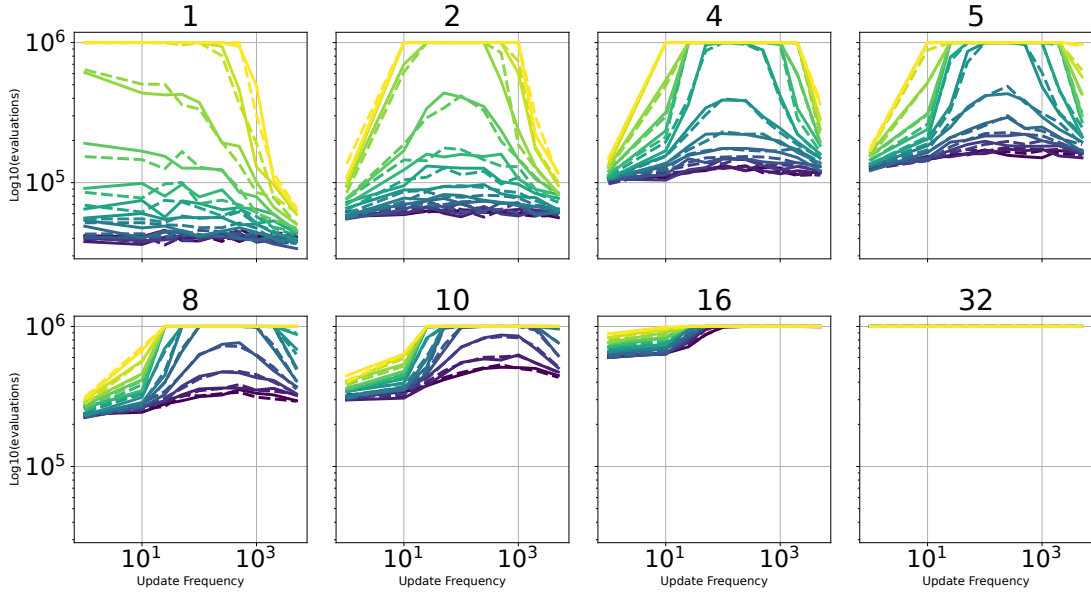


Fig. 7: Dimensionality 1000, number of evaluations (log-scaled) needed to reach optimum, on the rank-based functions, given different number of parents, mutation rates and mutation after crossover setting ($p_c = 0.9$). The x-axis corresponds to the frequency at which the weights are updated (in number of generations)

bad family tree structure may evolve temporarily. However, in some distance from the optimum, improvements are found while this bad family tree structures dominate. Similarly, here the same bad family tree structures might evolve while the environment is constant, and new offspring may evolve while these bad structures are around. We hope that this hypothesis gives inspiration for future work, be it empirically or theoretically.

4 Conclusions and Future Work

Through a varied set of empirical experiments, we have highlighted several ways in which benchmarking can be used to build upon results from theoretical analysis in sometimes surprising ways. By integrating different versions of the Dynamic Binary Value problem into the IOHprofiler platform, we were able to compare several versions of DBV, highlighting that sampling-based versions, specifically based on the Pareto distribution, lead to similar performance as the original DBV formulation, without requiring impractically large weights.

To highlight the potential of the integration of DBV into different benchmarking pipelines, we analyzed the impact of a variety of algorithmic modifications to a standard GA, which allows for the exploration of research questions related to mutation rates, crossover methods and population sizes within the same setup. Perhaps most intriguing from a theory perspective is the non-monotone dependence on the update frequency in Figure 7. For symmetry reasons, the update frequency makes no difference for population size $\mu = 1$. Hence, the observation that a lower update frequency makes the problem harder must be due to the structure of the population. It would be very interesting to understand how the structure of the evolving populations depends on the update frequency, and why a lower update frequency leads to more susceptible populations.

The experimental environment described in this paper can be further extended to look at different kinds of dynamic environments or different algorithmic ideas. Such new empirical studies might be used to determine interesting aspects of these problems to further analyze theoretically, leading to a positive feedback loop between theory and practice.

Acknowledgements This work was initiated as part of the Dagstuhl Seminar 23332 "Synergizing Theory and Practice of Automated Algorithm Design for Optimization". Our work was supported by CNRS Sciences informatiques via the AAP project IOHprofiler.

References

1. Branke, J.: Evolutionary optimization in dynamic environments, vol. 3. Springer Science & Business Media (2012)
2. Clément, F., Vermetten, D., De Nobel, J., Jesus, A.D., Paquete, L., Doerr, C.: Computing star discrepancies with numerical black-box optimization algorithms. In: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 1330–1338 (2023)
3. Doerr, C., Ye, F., Horesh, N., Wang, H., Shir, O.M., Bäck, T.: Benchmarking discrete optimization heuristics with iohprofiler. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. pp. 1798–1806 (2019)
4. Janett, D., Lengler, J.: Two-dimensional drift analysis: Optimizing two functions simultaneously can be hard. Theoretical Computer Science **971**, 114072 (2023)
5. Lengler, J.: A general dichotomy of evolutionary algorithms on monotone functions. IEEE Transactions on Evolutionary Computation **24**(6), 995–1009 (2019)
6. Lengler, J., Martinsson, A., Steger, A.: When does hillclimbing fail on monotone functions: an entropy compression argument. In: 2019 Proceedings of the Sixteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO). pp. 94–102. SIAM (2019)
7. Lengler, J., Meier, J.: Large population sizes and crossover help in dynamic environments. Natural Computing pp. 1–15 (2022)
8. Lengler, J., Riedi, S.: Runtime analysis of the $(\mu + 1)$ -EA on the dynamic binval function. Evolutionary Computation in Combinatorial Optimization **12692**, 84–99 (2021)
9. Lengler, J., Schaller, U.: The $(1+1)$ -EA on noisy linear functions with random positive weights. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 712–719. IEEE (2018)
10. Lengler, J., Steger, A.: Drift analysis and evolutionary algorithms revisited. Combinatorics, Probability and Computing **27**(4), 643–666 (2018)
11. Lengler, J., Zou, X.: Exponential slowdown for larger populations: the $(\mu + 1)$ -ea on monotone functions. In: Proceedings of the 15th ACM/SIGEVO Conference on Foundations of Genetic Algorithms. pp. 87–101 (2019)
12. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
13. Neumann, F., Neumann, A., Qian, C., Do, A.V., de Nobel, J., Vermetten, D., Ahouei, S.S., Ye, F., Wang, H., Bäck, T.: Benchmarking algorithms for submodular optimization problems using iohprofiler. CoRR **abs/2302.01464** (2023). <https://doi.org/10.48550/arXiv.2302.01464>, <https://doi.org/10.48550/arXiv.2302.01464>
14. Neumann, F., Pourhassan, M., Roostapour, V.: Analysis of evolutionary algorithms in dynamic and stochastic environments. Theory of evolutionary computation: recent developments in discrete optimization pp. 323–357 (2020)
15. Nguyen, T.T., Yang, S., Branke, J.: Evolutionary dynamic optimization: A survey of the state of the art. Swarm and Evolutionary Computation **6**, 1–24 (2012)
16. de Nobel, J., Ye, F., Vermetten, D., Wang, H., Doerr, C., Bäck, T.: IOHexperimenter: Benchmarking platform for iterative optimization heuristics. CoRR **abs/2111.04077** (2021), <https://arxiv.org/abs/2111.04077>
17. van Stein, N., Vermetten, D., Kononova, A.V., Bäck, T.: Explainable benchmarking for iterative optimization heuristics. arXiv preprint arXiv:2401.17842 (2024)
18. Vermetten, D., Lengler, J., Rusin, D., Bäck, T., Doerr, C.: Reproducibility files and additional figures (2024), code and data repository (Zenodo): doi.org/10.5281/zenodo.10964455 Figure repository (Figshare): doi.org/10.6084/m9.figshare.25592904
19. Wang, H., Vermetten, D., Ye, F., Doerr, C., Bäck, T.: Iohanalyzer: Detailed performance analysis for iterative optimization heuristic. ACM Trans. Evol. Learn. Optim. **2**(1), 3:1–3:29 (2022). <https://doi.org/10.1145/3510426>, <https://doi.org/10.1145/3510426>, iOAnalyzer is available at CRAN, on GitHub, and as web-based GUI, see <https://iohprofiler.github.io/IOHAnalyzer/> for links