Fitness Landscape Analysis Metrics based on Sobol Indices and Fitness- and State-Distributions

Christoph Waibel
Architecture and Building Systems
Inst. of Technology in Architecture
ETH Zurich
Zurich, Switzerland
waibel@arch.ethz.ch

Georgios Mavromatidis
Group for Sustainability and Technology
Dept. of Management, Technology, and Ec.
ETH Zurich
Zurich, Switzerland
gmavroma@ethz.ch

Yong-Wei Zhang, *Member, IEEE*
College of Electronics and Information
Jiangsu University of Science and Technology
Zhenjiang, China
ywzhang@just.edu.cn

***Abstract*—This paper proposes new Fitness Landscape Analysis (FLA) metrics as measures for problem difficulty in heuristic search. The sensitivity and variable interaction metrics are based on Sobol indices, a common technique in sensitivity analysis. The fitness- and state-variance and the fitness- and state-skewness metrics are based on the second and third central statistical moments. We compute metric values for around 550 continuous test functions in 2, 5 and 10 dimensions and compare it to well-established metrics (Fitness Distance Correlation, Autocorrelation, Information Content, Density-Basin Information and Partial Information Content). By conducting two-sample statistical tests (T-Test, F-Test, Kolmogorov-Smirnov Test, and Rank-Sum Test) for all combinations of FLA metrics, we demonstrate that our proposed metrics result in significantly different distributions. Thus, we can conclude that they reveal fitness landscape characteristics not captured by the existing metrics that were considered.**

***Index Terms*—Fitness Landscape Analysis, Sensitivity Analysis, Central Moments, Black-Box Optimization, Problem Difficulty**

I. Introduction

In contrast to dynamic (exploratory) Fitness Landscape Analysis (FLA), static FLA aims to examine characteristics of a function $f(x)$ independently of the optimization algorithm that solves it [3]. By conducting an FLA, features of an (black-box) optimization problem related to algorithm performance may be identified; thus, they help with the selection of the ideal algorithm and/or calibrating its hyper-parameters. At first sight, dynamic FLA appears to be more useful in this respect, as it provides information specific to a solver [4]–[7]. Especially hyper-parameter optimization methods are therefore often online, i.e. only when a solver is initialized to a new problem are its hyper-parameters tuned to this particular problem instance [8]–[10].

Static FLA provides extensive information on problem features that may also be used to improve algorithm performance. When encountering a new unknown problem, solver behavior may be observed and cross-correlated to previously studied, known test function problems. Therefore, pre-tuning methods such as in [11], [12], and generally hyper-parameter optimization and algorithm selection research can utilize information from static FLA.

This paper is based on results from a PhD thesis [1] and findings have been partially published in [2].

It has been shown that several FLA metrics are required to properly capture the diverse characteristics of real-world problems [13], [14]. When facing a new problem with no prior information about its structure, it remains unclear which FLA metric yields the highest descriptive content. Therefore, this paper introduces the following (static) metrics to add to the existing FLA literature:

- Degree of Variable Interaction, v_{inter}
- Coefficient of Variation in Variable Sensitivity, v_{cv}
- Fitness Variance, $\mu_2(y)$
- State Variance, $\mu_2(\|d\|)$
- Fitness Skewness, $s(y)$
- State Skewness, $s(\|d\|)$

v_{inter} and v_{cv} are based on the Sobol indices, a common technique in Global Sensitivity Analysis [15]. v_{cv} describes how sensitive the cost value of a function $f(x)$ reacts to individual decision variables x_i. v_{inter} describes the degree of variable interaction, i.e. how much the cost value is varied due to interaction effects between multiple x_i. The authors in [16] have also developed a metric to quantify variable interactions; however, the main difference to v_{cv} is that our metric is strictly static, as it relies on a Monte-Carlo based sequence as sampling input. The metric by [16], on the other hand, is for dynamic FLA.

$\mu_2(y)$ and $\mu_2(\|d\|)$ are the second central moment—i.e. the variance—of fitness and state distributions. Larger variations have a negative influence on the effectiveness of probabilistic sampling or optimization methods.

$s(y)$ and $s(\|d\|)$ are based on the third central moment—i.e. the skewness—however not centered around the mean of a distribution. Thus, they inform on whether probabilistic methods are more likely to find low- or high-cost solutions, as well as how dispersed solutions are in variable space.

For computing $\mu_2\|d\|$ and $s(\|d\|)$, we introduce a new distance measure, state dispersion, that informs on the dispersion in variable space for specific cost-value ranges.

Our proposed metrics are described in detail in Section II. In Section III we compare metric values of our metrics with other established popular FLA metrics (Fitness Distance Correlation [17], Autocorrelation [18], Information Content, Density-Basin Information and Partial Information Content

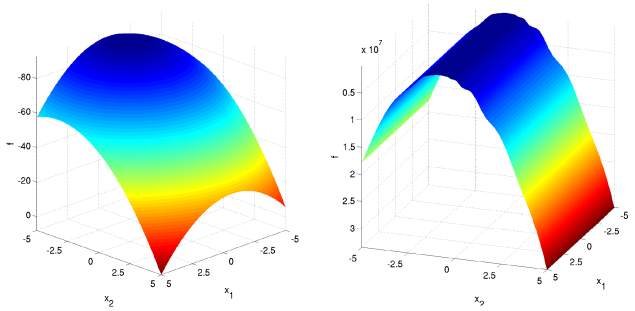
978-1-7281-6929-3/20/$31.00 ©2020 IEEE

[19]), using a large set of over 550 continuous optimization problems as compiled by [7].

Although we only consider continuous functions in this study, the proposed metrics can also be applied to discrete / combinatorial problems with minor modifications, e.g. by using the Hamming distance instead of Euclidean as a state-distance measure. Furthermore, $\mu_2(y)$, $\mu_2(\|d\|)$, $s(y)$ and $s(\|d\|)$ could also be applied to dynamic FLA, but we compute them as static measures in this study.

II. Proposed FLA Metrics

The proposed FLA metrics require an input sequence X. More specifically, v_{inter} and v_{cv} require a Monte-Carlo based sequence, such as the Sobol-sequence, which is a low discrepancy quasi-random sampling technique. In [20] a sampling size of $T = m \times (n+2)$ is suggested, with m ranging at least between 500 and 1000, and n being the problem dimension[1]. The other metrics $\mu_2(y)$, $\mu_2(\|d\|)$, $s(y)$ and $s(\|d\|)$ work with any kind of sequence, but naturally benefit from an extensive sampling. The individual FLA metrics are described in detail in the following sections and exemplified with four test functions in $n = 10$ from the COCO set [21], shown in $n = 2$ in Fig. 1.



(a) Sphere function (b) Ellipsoidal function

(c) Weierstrass function (d) Gallagher's Gaussian 21-hi Peaks function

Fig. 1. Test functions used in $n = 10$ to exemplify proposed Fitness Landscape metrics (visualized in $n = 2$ here). Images from [21]

A. Metrics based on Sobol Indices

The Sobol method [15] is a variance-based Global Sensitivity Analysis technique that quantifies the variation in cost value $y = f(x)$ caused by changing decision variable values

[1]In order to reduce numerical issues, we even use $m = 10,000$ in our later numerical experiments in Section III.

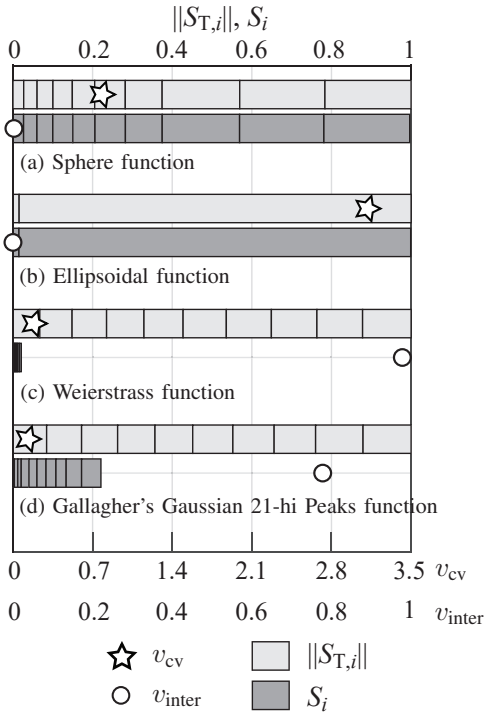


Fig. 2. First- and total-order effects as computed by the Sobol method for four test functions in $n = 10$. Light grey barcharts show $\|S_{T,i}\|$ (normalized between minimal and maximal $S_{T,i}$ per problem), dark grey barcharts show S_i. Star markers show v_{cv}, circle markers show v_{inter}. The bar segments indicate individual variable values of x_i for $\|S_{T,i}\|$ and S_i respectively.

x_i, $i \in n$. The Sobol method provides two indices: (i) The first-order effect S_i that indicates the amount of variation of y caused solely by an individual variable x_i (i.e. without considering variable interactions), and (ii) the total-order Sobol index $S_{T,i}$ that indicates the amount of variation in y caused by x_i including its interaction effects with other variables. We use the `sobol2007` function of the `R` package `sensitivity` that implements the equations by [22], [23].

S_i and $S_{T,i}$ are shown in Fig. 2 as stacked barplots for four test functions in $n = 10$. Large S_i values translate to a high influence on the output variance by the variable itself without interactions. Large $S_{T,i}$ mean a high influence on the output variance by a variable including its interaction effects. Fig. 1a and 1b are examples of separable problems (small v_{inter}), Fig. 1c and 1d are examples for inseparable problems (large v_{inter}). Also, Fig. 1b shows that this instance of the Ellipsoidal function from the COCO set has a single variable x_i that contributes almost entirely to the output variance. We aim to quantify such function characteristics with the FLA metrics as described in the following.

1) Degree of Variable Interaction: Our first metric v_{inter} measures the degree of variable interactions of a function $f(x)$. Only the first-order effect, S_i, is used for its computation:

$$v_{\text{inter}} := 1 - \sum_{i=1}^{n} S_i. \tag{1}$$

Since S_i indicates the contribution that a single variable x_i has on the variance of y (ignoring possible interactions with other x_j, $j \neq i$), v_{inter} describes how much of the variation in y can be described by first-order effects only. Theoretically, $\sum_{i=1}^{n} S_i = 1$ for completely separable problems and generally $\sum_{i=1}^{n} S_i \leq 1$, meaning that when v_{inter} approaches 1, interactions between variables have a higher contribution to the output variance. On the contrary, a small value for v_{inter} indicates variable separability (see circle markers in Fig. 2).

2) Coefficient of Variation in Variable Sensitivities: The second proposed metric based on the Sobol method measures the variation in variable sensitivity by computing the coefficient of variation, i.e. the ratio of the standard deviation to the mean of all $S_{\text{T},i}$ of a problem:

$$v_{\text{cv}} \coloneqq \frac{\sigma_{S_\text{T}}}{\bar{S}_\text{T}}, \tag{2}$$

where σ_{S_T} is the standard deviation and $\bar{S}_\text{T}$ is the mean of all $S_{\text{T},i}$, $i \in n$. Since $S_{\text{T},i}$ indicates the impact of x_i on the output variance y including its interactions with other variables, a large value for v_{cv} indicates that only few x_i are important in a problem. On the contrary, small values for v_{cv} mean that most x_i are equally important (see star markers in Fig. 2).

B. Fitness- and State-Variance

The fitness- and state-variance metrics $\mu_2(y)$ and $\mu_2(\|d\|)$ are derived from histograms as a way to visualize the distribution of a sampled problem [14], [24]. We propose using the second central statistical moment—i.e. the variance—as a summarizing FLA metric to such histograms.

1) Fitness Variance: The fitness variance is computed as:

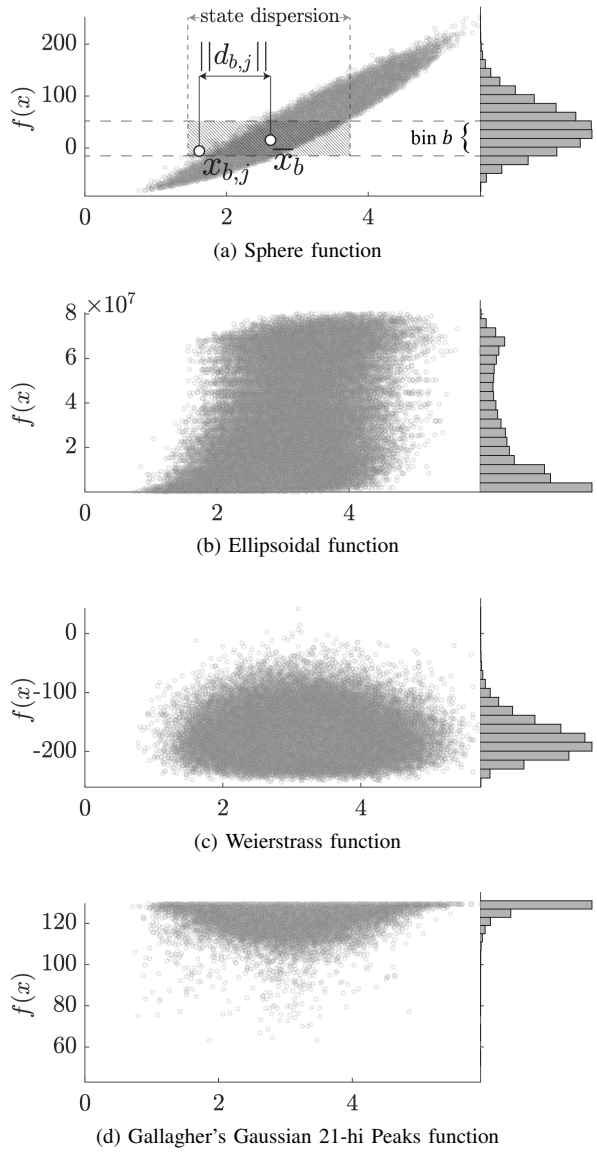
$$\mu_2(y) \coloneqq \mathbb{E}(y' - \mu_{y'})^2, \tag{3}$$

where y' is $y = f(X)$ after mean normalization, $\mu_{y'}$ is the mean of y' and $\mathbb{E}$ denotes the expected value. A large fitness variance means that the probability of sampling a wider range of different cost values increases, whereas a small fitness variance indicates that probabilistic methods will more likely find solutions with similar cost values. Therefore, this metric can be used as an indicator of how likely randomized operators in a solver may find low- or high-cost solutions (Fig. 3 and 4).

2) State Variance: For computing the state variance, we first express the state distribution X as the normalized distance values per bin b of the cost-values histograms (Fig. 4). Per b, we compute the mean normalized distance of all states $x_{b,j}$—j being the index of states in bin b—to the mean solution $\overline{x}_b$ within that bin. Hence, we define the mean normalized distance of all states within a bin b—also understood as state "dispersion"—by:

$$\left\|\overline{d}_b\right\| \coloneqq \frac{1}{\#b} \sum_{j=1}^{\#b} \left(\frac{1}{n} \sum_{i=1}^{n} \sqrt{(\|x_{b,j,i}\| - \|\overline{x}_{b,i}\|)^2} \right) , \tag{4}$$

where $i \in n$, n being the problem dimension, and $\#b$ being the number of elements in bin b. The number of bins B is a parameter that will influence the resolution of the histogram. We found $B = 20$ to be a reasonable value.



(a) Sphere function

(b) Ellipsoidal function

(c) Weierstrass function

(d) Gallagher's Gaussian 21-hi Peaks function

Fig. 3. Input sequences X plotted as Fitness Distance Correlation plots (Euclidean distance). Histograms show the frequency at which certain cost value intervals occur. For each bin b (in Fig. 3a four bins are grouped for illustrative purposes) Eq. 4 is computed.

The computation of a single distance $\|d_{b,j}\|$ is illustrated in Fig. 3a. Large values for $\left\|\overline{d}_b\right\|$ mean that, at the given cost-value range, solutions are more spread in state space and thus possibly deceiving an optimization algorithm, as the solution is more ambiguous. In other words, two solution vectors $x^{(1)}$ and $x^{(2)}$ might have similar variables but very different cost values. Small $\left\|\overline{d}_b\right\|$ indicate easier problems because the close solutions (in variable space) are also associated by similar cost values, i.e. a non-deceptive landscape.

Fig. 4 shows $\left\|\overline{d}_b\right\|$ plotted as solid black curves over the cost-value histograms of four different test functions; the dashed curves indicate minimal and maximal $\|d_{b,j}\|$ per bin b.

As shown in Fig. 4d at $f(x) \approx 50$ highlighted in orange, the curve for $\left\|\overline{d}_b\right\|$ might unexpectedly spike. This is caused by too few samples in a bin, causing $\|d_{b,j}\|$ to behave seemingly random. Since it would create a bias in the state variance towards non-representative domains, we introduce following condition for such cases:

$$\exists\{b \in B \mid \#b < b_{\text{lb}} n\} \;\Rightarrow \left\|\overline{d}_b\right\| \coloneqq 0, \tag{5}$$

which assigns a mean distance of zero to those underpopulated bins. b_{lb} is a parameter to define the smallest cardinality of b to be considered for the state variance metric, and B is a parameter that defines the total number of bins. We use $b_{\text{lb}} = 1.5$ and $B = 20$ in our later numeric experiments in Section III.

Finally, we can compute the state variance of $\|d\|$ with:

$$\mu_2(\|d\|) = \mathbb{E}(d_\sim - \mu_d)^2 \;, \tag{6}$$

where $d_\sim$ denotes the distribution of $\|d\|$ and μ_d is its mean. We obtain $d_\sim$ by re-interpreting the black solid curves of $\left\|\overline{d}_b\right\|$ in Fig. 4 as probability density functions, i.e. $\|d\| \mapsto P$ (left Y-axis in Fig.4), with $\sum_{b=1}^{B} P(b) = 1.0$. Hence, $d_\sim \in \{1/B, 2/B, \ldots, 1\}$ is a sequence with normalized values, where the occurrence of each element is proportional to $P(b)$.

C. Fitness- and State Skewness

We also propose a modified skewness metric that is computed from the distribution of fitness and state values. It maintains bias information, i.e. whether a function is more likely to contain high- or low-cost solutions (fitness skewness), or whether the state dispersion is higher at high- or low-cost solutions (state skewness) (Fig. 4).

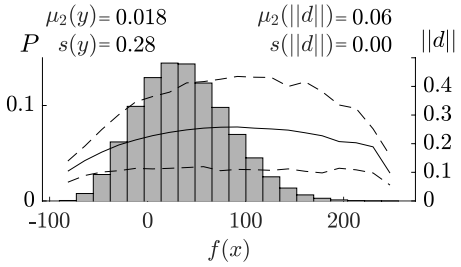
Unlike the third statistical moment, we do not center our measure around the mean of a distribution. The difference to Pearson's moment coefficient of skewness $\tilde{\mu}_3$ is that the range of our measure is $[-1, 1]$, hence allowing a comparison between functions with different cost value- and state dispersion ranges. However, our measure is proportional to $\tilde{\mu}_3$. E.g. in Fig. 4a to 4d, we have $s(y) = \{0.28, 0.39, 0.48, -0.93\}$, while the regular skewness $\tilde{\mu}_3(y)$ would yield $\{0.35, 0.82, 0.75, -3.88\}$.

1) Fitness Skewness: We define the fitness skewness as:

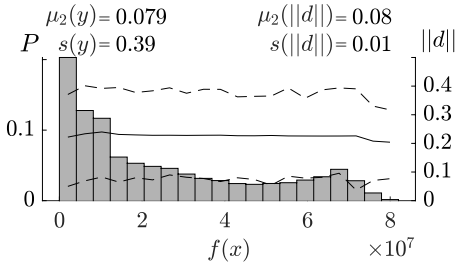
$$s(y) \coloneqq \frac{1}{T} \sum_{j=1}^{T} \Big(\|y\|_s \left(\check{y} - y_j\right)\Big) \;, \tag{7}$$

where $\|y\|_s = |((\max(y) - \min(y))/2)^{-1}|$ is a normalization factor ($|\cdot|$ denotes the absolute value), such that $s(y) \in [-1, 1]$, $\check{y} = ((\max(y) - \min(y))/2) + \min(y)$ is the central value of y, and T being the population size of the input sequence X.

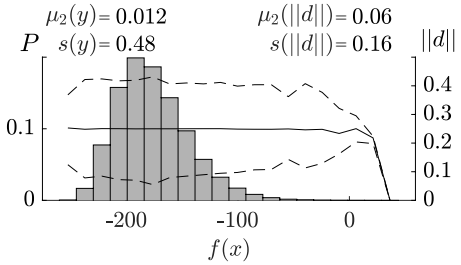
The fitness skewness is shown in Fig. 4 for different test functions. A large positive skewness $s(y) = 1$ means that a function has a high probability of sampling a low cost solution, whereas a negative skewness $s(y) = -1$ means that it is more likely to sample a high cost solution.



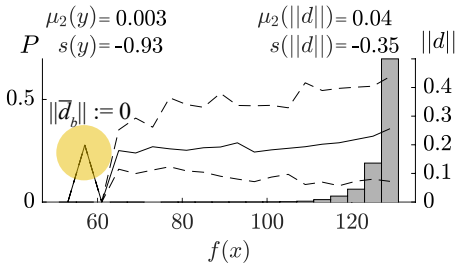
(a) Sphere function



(b) Ellipsoidal function



(c) Weierstrass function



(d) Gallagher's Gaussian 21-hi Peaks function

Fig. 4. Cost histograms and state dispersion curves (black solid) for metrics $\mu_2(y)$, $s(y)$, $\mu_2(\|d\|)$ and $s(\|d\|)$. Dashed curves indicate the max and min $\|d_b\|$ per bin. Image modified after [2].

2) State Skewness: The fitness skewness only addressed the distribution in function value space. With the state skewness metric, we intend to relate cost value probabilities with how dispersed solutions are in state space x. In accordance to Eq. 7, we define the state skewness as:

$$s(\|d\|) \coloneqq \frac{1}{T} \sum_{j=1}^{T} \Big(\check{d}^{-1}\left(\check{d} - d_{\sim,j}\right)\Big) \;, \tag{8}$$

where $\check{d} = 0.5 - 0.5/B$ is the centre of the histogram.

A positive state skewness can be interpreted as a higher dispersion of states at low cost values, whereas a negative

state skewness means a higher dispersion at high cost values. Gallagher's Gaussian 21-hi Peaks function in Fig. 4d is an example of a large negative state skewness, which means that solutions x with high cost values can be substantially different in state space. The Sphere function in Fig. 4a is an example with no state skewness, as at both cost extremes states seem to converge, meaning that solutions become less ambiguous in such domains.

III. Comparison to alternative Static FLA Metrics

In this section, we compare our proposed FLA metrics with established ones from the literature (Section III-B). We use an extensive set of test functions (described in Section III-A) and conduct statistical significance tests between all considered metrics to assess whether new perspectives and problem characteristics are obtained with our newly proposed metrics (Section III-C).

A. Test Functions Investigated

We apply our developed FLA metrics to an extensive problem set that has been compiled in [7]. It contains 272 mostly continuous benchmark problems from various sources, including 219 problems collected from [25]–[27], the 29 problems from the CEC 2017 optimization competition [28], and the 24 problems from BBOB 2009 (also known as COCO test set) [21], [29]. In [7] it has been shown that this set covers a very high diversity and thus can be considered representative for the domain of continuous optimization. We compute FLA metrics for $n = 2, 5, 10$ for the entire set, even though in [2], [7] it has been shown that n does not have a significant impact on FLA metrics for most of the test functions investigated. However, [7] showed that n can still have an impact on algorithm performance on a given problem, i.e. an algorithm that dominates a benchmark at low dimensions may not perform as well as other algorithms on higher dimensions.

Some functions in the set are not entirely scalable in n, which is why we have 558 problems / problem instances (we use 15 different instances of the COCO test set per n) in $n = 10$, 556 in $n = 5$, and 547 in $n = 2$.

B. FLA Metrics Considered

Apart from our proposed metrics from Section II, we compute the following established metrics from the literature: Fitness Distance Correlation, FDC [17]; Autocorrelation, $\tilde{\rho}$ [18]; Information Content, $\tilde{H}(\varepsilon)$; Density-Basin Information, $\tilde{h}(\varepsilon)$; and Partial Information Content, $\tilde{M}(\varepsilon)$ [19].

We compute FDC using the same Sobol sequence with $T = (2+n) \times 10,000$ as for our proposed metrics. The tilde accent in $\tilde{\rho}$, $\tilde{H}(\varepsilon)$, $\tilde{h}(\varepsilon)$ and $\tilde{M}(\varepsilon)$ denote the medians of 20 random walks with different starting points and length $T = n \times 100$. Their sequence is sampled as:

$$x_t = x_{t-1} + \sigma_{\text{rw}} \hat{v}, \tag{9}$$

where σ_{rw} is the step size—we use 2% of the search domain—and $\hat{v} = v|v|^{-1}$ is the unit vector, with $v \in \mathbb{R}^n$ being a random vector drawn from a normal distribution, i.e. $\forall i \in n$, $v_i \sim \mathcal{N}(0, 1)$.

The autocorrelation measure may fluctuate across changing lag l. Therefore, for one random walk, we propose measuring the median from $l = 1$ to $n + 1$ as:

$$\rho \coloneqq \tilde{r}_{l=1:\, n+1}, \tag{10}$$

which means that $\tilde{\rho}$ is the median of 20 ρ.

$\tilde{H}(\varepsilon)$, $\tilde{h}(\varepsilon)$ and $\tilde{M}(\varepsilon)$ are functions of a tolerance parameter ε, for which we use 0.2% throughout our experiments. In initial experiments [2] for one instance of the 24 COCO test functions, we have found that this value results in a reasonable degree of discrimination of the problems, while with a too large $\varepsilon = 1\%$ all measures converged towards the same value. With a too small $\varepsilon = 0.001\%$, measures were almost uniformly distributed, with no grouping observed among the test functions. Related to this, [6] have addressed the issue of an appropriate step size. The authors propose variable step sizes to improve sampling quality, but we leave such improvements for future work.

C. Statistical Difference between FLA Metrics

In this section, we perform statistical tests for all 11 considered FLA metrics v_{inter}, v_{cv}, $\mu_2(y)$, $\mu_2(\|d\|)$, $s(y)$, $s(\|d\|)$, FDC, $\tilde{\rho}$, $\tilde{H}(\varepsilon)$, $\tilde{h}(\varepsilon)$, and $\tilde{M}(\varepsilon)$. From the test results, we can assess whether our proposed metrics inform us on different fitness landscape characteristics that are not covered by the alternative metrics.

We perform the two sample T-Test for equal means (TT), the F-Test for equal variances (FT), the Rank-Sum Test for equal medians (RS), and the Kolmogorov-Smirnov Test for equal continuous distributions (KS) for all $\mathcal{M}(\mathcal{M}-1)/2$ combinations of FLA metrics, $\mathcal{M}$ being our 11 considered metrics. The null hypothesis for one combination states that two FLA metrics m_i and m_j have the same distributions, $H_0 \colon m_i = m_j$, with $j, i \in \mathcal{M}$ and $i \neq j$. The rejection of H_0 consequently means that the two tested metrics are statistically different.

All FLA metric values are normalized beforehand between the maximal and minimal value found for the entire test function set, such that they can be compared in the statistical tests (not necessary for KS). This procedure is required, since most FLA metrics have inherently different value ranges (see Fig. 2 and 4). It assumes that the minimal and maximal values represent the true upper and lower bounds for a theoretical complete set of functions, otherwise there is a risk of losing bias information of distributions. Naturally, we cannot know the true bounds. However, as has been shown in [7], the test set has a very large coverage and thus we rely on the assumption of this test set being representative enough for the class of continuous optimization problems. It should also be noted that by normalizing our data, we skew the test towards a more conservative direction. Thus, it would only become more likely to not find any statistical difference.

Fig. 5 shows a scatterplot matrix and p-values for all tests and metrics at $n = 10$, Table I shows the p-values for $n = 2$

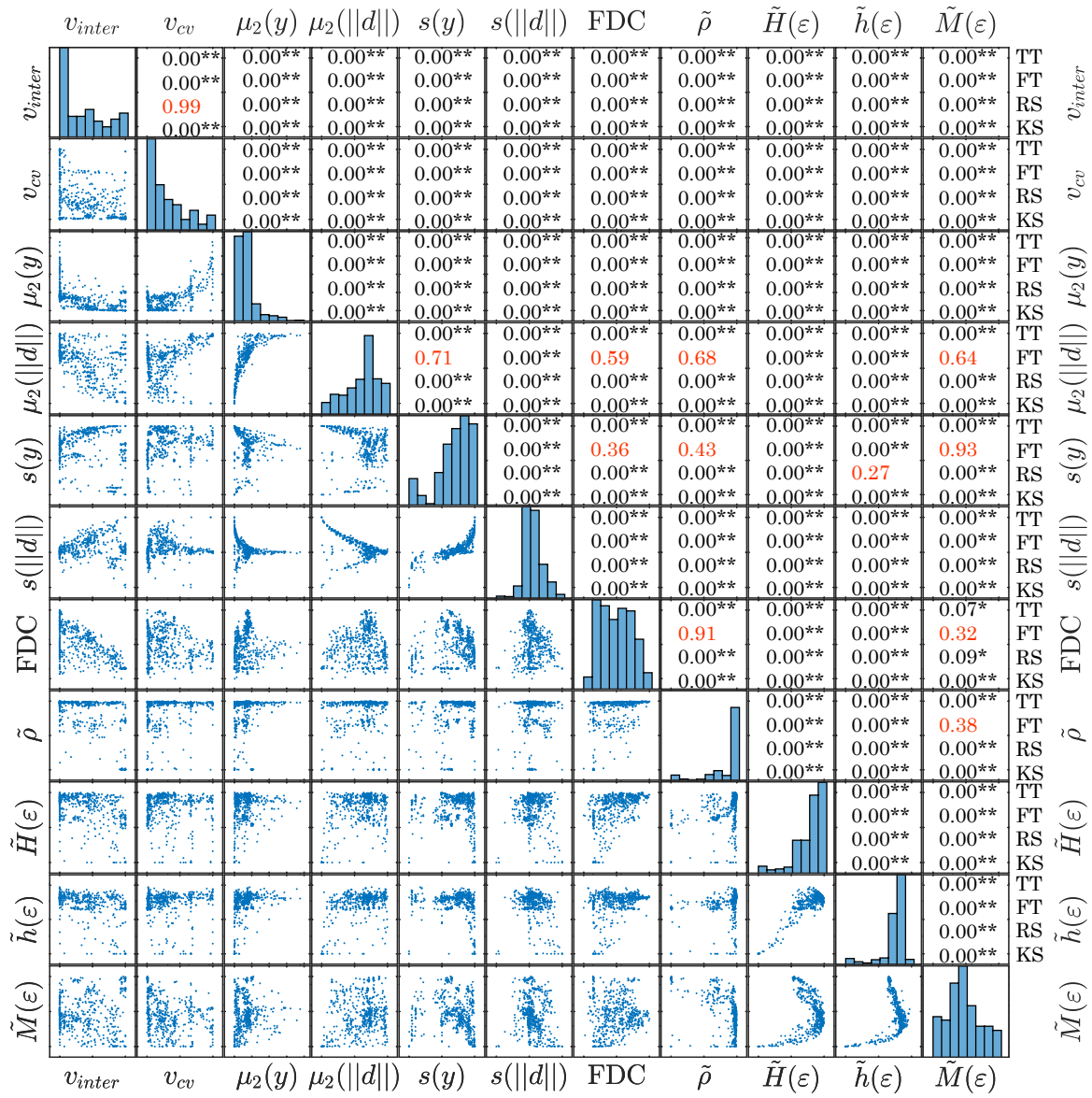


Fig. 5. Scatterplot matrix for $n = 10$. p-values marked with ** are significant at $\alpha = 0.05$ and marked with * at $\alpha = 0.1$. Values in red font indicate cases for which the Null-hypothesis is accepted (i.e. no significant difference). TT = T-Test; FT = F-Test; RS = Rank-Sum Test; KS = Kolmogorov-Smirnov Test. Values are rounded to 2 decimal places.

and $n = 5$. As already shown by [7], problem dimensionality does not noticeably change FLA metrics, therefore we omit scatterplots for $n = 2$ and 5.

The results clearly show that there is a significant difference between FLA metrics for the majority of statistical tests conducted at a high confidence, i.e. $p = 0.00$. Only in few cases is the null hypothesis accepted (p-values marked in red font in Fig. 5 and Table I). However, accepted H_0 are not consistent throughout problem dimensions $n = 2, 5, 10$. While we do not show it here, we did find that the scatterplots for the different n exhibit similar patterns. The reason for partly accepting H_0 might be because our input sequence also increases with problem dimension, thus possibly changing distributions slightly. Since most tests do show a statistical difference over n, we can still be confident in concluding that all considered FLA metrics describe different characteristics of a function.

Another way of comparing similarity is by studying the scatterplots in Fig. 5. It is striking that for the metrics based on information analysis—$\tilde{H}(\varepsilon)$, $\tilde{h}(\varepsilon)$, and $\tilde{M}(\varepsilon)$—distinct patterns emerge, albeit no linear correlation. The same is true for the state- and fitness-distribution metrics, which appear to correlate. For most of the other FLA metric combinations, points are rather scattered over the entire domain. This can be interpreted as a high diversity of the considered test function set. Also, assuming all considered FLA metrics produce meaningful measures, this result can be interpreted as evidence for the new metrics describing significantly different fitness

landscape characteristics.

IV. Discussion and Conclusion

In this paper, we have introduced new FLA metrics based on Sobol indices and state- and fitness-distributions. They inform on variable sensitivities, variable interaction, state dispersion and fitness probabilities. Such information may be used for selecting an appropriate optimization algorithm or calibrating its hyper-parameters for a specific problem.

We have compared our metrics to other established ones (Fitness Distance Correlation, Autocorrelation, Information Content, Density-Basin Information and Partial Information Content) using an extensive set of around 550 continuous test functions, tested in $n = 2, 5, 10$.

Summarizing the results of this study, the majority of statistical tests conducted showed a significant difference of FLA metric distributions. Thus, it is evident that the proposed metrics result in different distributions as compared to the existing well-established metrics. A sensible interpretation is that all metrics reveal different characteristics and features of fitness landscapes, thus should all be mutually considered in FLA.

This assumes our proposed metrics result in coherent measures for the test functions, which we have demonstrated in Section II at the example of the Sphere-, the Ellipsoidal-, the Weierstrass-, and Gallagher's 21-hi Peaks-function.

It should be critically noted that both empirical and theoretical evidence on the usefulness of the proposed metrics for FLA tasks, such as hyper-parameter optimization or algorithm selection, is still missing. As a first step, they should be correlated to algorithm performance when solving benchmark test functions or real engineering problems.

Future work should also adapt the proposed metrics for discrete / combinatorial functions, e.g. by using the Hamming distance. Also, the impact of soft and hard constraints on metric values needs to be investigated, as well as the impact of high dimensional problems $n > 100$. These are essential topics especially in real-world optimization problems. Unlike the metrics based on the Sobol indices, the state- and fitness-distribution metrics can also be applied to exploratory / dynamic FLA.

Finally, future work should find a means to integrate our (and other) FLA metrics to the algorithm selection problem and to hyper-parameter optimization. A possible approach could be comprised of an initial screening phase, in which approximate information of an unknown problem is obtained. Using cross-correlation, previously conducted static FLA with test functions could be brought into relation to the unknown problem, and the hyper-parameters of an optimization algorithm could be calibrated accordingly.

References

[1] C. Waibel, "Simulation-Based Optimization of Buildings and Multi-Energy Systems," PhD Dissertation, ETH Zurich, 2018.

[2] C. Waibel, G. Mavromatidis, R. Evins, and J. Carmeliet, "A Comparison of Building Energy Optimization Problems and Mathematical Test Functions using Static Fitness Landscape Analysis," *Journal of Building Performance Simulation*, vol. 12, no. 6, pp. 789–811, 2019.

[3] J.-P. Watson, "An Introduction to Fitness Landscape Analysis and Cost Models for Local Search," in *Handbook of Metaheuristics*, M. Gendreau and J.-Y. Potvin, Eds. Boston, MA: Springer US, 2010, pp. 599–623.

[4] M. Gallagher and B. Yuan, "A general-purpose tunable landscape generator," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 5, pp. 590–603, 2006.

[5] M. a. Muñoz, Y. Sun, M. Kirley, and S. K. Halgamuge, "Algorithm selection for black-box continuous optimization problems: A survey on methods and challenges," *Information Sciences*, vol. 317, pp. 224–245, 2015.

[6] M. A. Muñoz, M. Kirley, and S. K. Halgamuge, "Exploratory landscape analysis of continuous space optimization problems using information content," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 1, pp. 74–87, 2015.

[7] Y.-W. Zhang and S. K. Halgamuge, "Similarity of Continuous Optimization Problems from the Algorithm Performance Perspective," in *IEEE Congress on Evolutionary Computation*, Wellington, New Zealand, 2019, pp. 2950–2958.

[8] F. Hutter, H. H. Hoos, K. Leyton-Brown, and T. Stützle, "ParamILS: An automatic algorithm configuration framework," *Journal of Artificial Intelligence Research*, vol. 36, pp. 267–306, 2009.

[9] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential Model-Based Optimization for General Algorithm Configuration," in *Learning and Intelligent Optimization: 5th International Conference, LION 5, Rome, Italy, January 17-21, 2011. Selected Papers*, C. Coello Coello, Ed. Springer Berlin Heidelberg, 2011, pp. 507–523.

[10] M. López-Ibáñez, J. Dubois-Lacoste, L. Pérez Cáceres, M. Birattari, and T. Stützle, "The irace package: Iterated racing for automatic algorithm configuration," *Operations Research Perspectives*, vol. 3, pp. 43–58, 2016.

[11] F. Kursawe, "Grundlegende empirische Untersuchungen der Parameter von Evolutionsstrategien - Metastrategien," PhD Thesis, Department of Computer Science, University of Dortmund, 1999.

[12] C. Waibel, R. Evins, and J. Carmeliet, "Clustering and Ranking Based Methods for Selecting Tuned Search Heuristic Parameters," in *IEEE Congress on Evolutionary Computation*, Wellington, New Zealand, 2019, pp. 2932–2941.

[13] E. Pitzer and M. Affenzeller, "A Comprehensive Survey on Fitness Landscape Analysis," *Recent Advances in Intelligent Engineering Systems*, pp. 161–191, 2012.

[14] K. M. Malan and A. P. Engelbrecht, "A survey of techniques for characterising fitness landscapes and some possible ways forward," *Information Sciences*, vol. 241, pp. 148–163, 2013.

[15] I. M. Sobol', "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates," *Mathematics and Computers in Simulation*, vol. 55, no. 1-3, pp. 271–280, 2001.

[16] Y. Sun, M. Kirley, and S. K. Halgamuge, "Quantifying Variable Interactions in Continuous Optimization Problems," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 2, pp. 249–264, 2016.

[17] T. Jones and S. Forrest, "Fitness Distance Correlation as a Measure of Problem Difficulty for Genetic Algorithms," in *Proceedings of the 6th International Conference on Genetic Algorithms*, vol. 95, 1995, pp. 184–192.

[18] E. Weinberger, "Correlated and uncorrelated fitness landscapes and how to tell the difference," *Biological Cybernetics*, vol. 63, no. 5, pp. 325–336, 1990.

[19] V. K. Vassilev, T. C. Fogarty, and J. F. Miller, "Information characteristics and the structure of landscapes." *Evolutionary computation*, vol. 8, no. 1, pp. 31–60, 2000.

[20] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons, 2004.

[21] S. Finck, N. Hansen, R. Ros, and A. Auger, "Real-Parameter Black-Box Optimization Benchmarking 2010: Presentation of the Noiseless Functions," Research Center PPE, Technical Report 2009/20, Tech. Rep., 2010.

[22] S. Tarantola, D. Gatelli, S. S. Kucherenko, W. Mauntz, and Others, "Estimating the approximation error when fixing unessential factors in global sensitivity analysis," *Reliability Engineering & System Safety*, vol. 92, no. 7, pp. 957–960, 2007.

[23] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola, "Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index," *Computer Physics Communications*, vol. 181, no. 2, pp. 259–270, 2010.

TABLE I

STATISTICAL TESTS FOR $n = 2$ AND $n = 5$. P-VALUES MARKED WITH ** ARE SIGNIFICANT AT $\alpha = 0.05$ AND MARKED WITH * AT $\alpha = 0.1$. VALUES IN RED FONT INDICATE CASES FOR WHICH THE NULL-HYPOTHESIS IS ACCEPTED (I.E. NO SIGNIFICANT DIFFERENCE). TT = T-TEST; FT = F-TEST; RS = RANK-SUM TEST; KS = KOLMOGOROV-SMIRNOV TEST. VALUES ROUNDED TO 2 DECIMAL PLACES.

			v_{cv}	$\mu_2(y)$	$\mu_2(\|d\|)$	$s(y)$	$s(\|d\|)$	FDC	$\tilde{\rho}$	$\tilde{H}(\varepsilon)$	$\tilde{h}(\varepsilon)$	$\tilde{M}(\varepsilon)$		$n = 5$
			0.56	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	TT	
			0.01**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	FT	v_{inter}
			0.01**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	RS	
			0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	KS	
v_{cv}	TT	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	TT	
	FT	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	FT	v_{cv}
	RS	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	RS	
	KS	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	KS	
$\mu_2(y)$	TT	0.00**	010*		0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	TT	
	FT	0.00**	0.00**		0.00**	0.00**	0.32	0.00**	0.00**	0.00**	0.00**	0.00**	FT	$\mu_2(y)$
	RS	0.00**	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	RS	
	KS	0.00**	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	KS	
$\mu_2(\|d\|)$	TT	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	TT	
	FT	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	0.00**	0.34	0.00**	0.00**	FT	$\mu_2(\|d\|)$
	RS	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	RS	
	KS	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	KS	
$s(y)$	TT	0.00**	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	0.02**	0.00**	0.00**	TT	
	FT	0.00**	0.00**	0.91	0.00**		0.00**	0.16	0.01**	0.00**	0.00**	0.24	FT	$s(y)$
	RS	0.00**	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	0.60	0.57	0.00**	RS	
	KS	0.00**	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	KS	
$s(\|d\|)$	TT	0.00**	0.00**	0.00**	0.94	0.00**		0.03**	0.00**	0.00**	0.00**	0.05**	TT	
	FT	0.00**	0.00**	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	FT	$s(\|d\|)$
	RS	0.00**	0.00**	0.00**	0.55	0.00**		0.10**	0.00**	0.00**	0.00**	0.01**	RS	
	KS	0.00**	0.00**	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	0.00**	0.00**	KS	
FDC	TT	0.00**	0.00**	0.00**	0.03**	0.00**	0.03**		0.00**	0.00**	0.00**	0.91	TT	
	FT	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**		0.23	0.00**	0.00**	0.82	FT	FDC
	RS	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	0.82	RS	
	KS	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	0.04**	KS	
$\tilde{\rho}$	TT	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	TT	
	FT	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.02**		0.00**	0.00**	0.15	FT	$\tilde{\rho}$
	RS	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	RS	
	KS	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**		0.00**	0.00**	0.00**	KS	
$\tilde{H}(\varepsilon)$	TT	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**		0.23	0.00**	TT	
	FT	0.00**	0.00**	0.00**	0.00**	0.00**	0.04**	0.00**	0.00**		0.01**	0.00**	FT	$\tilde{H}(\varepsilon)$
	RS	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**		0.83	0.00**	RS	
	KS	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**		0.00**	0.00**	KS	
$\tilde{h}(\varepsilon)$	TT	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.07*	0.00**		0.00**	TT	
	FT	0.00**	0.00**	0.00**	0.00**	0.00**	0.29	0.00**	0.00**	0.31		0.00**	FT	$\tilde{h}(\varepsilon)$
	RS	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**		0.00**	RS	
	KS	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**		0.00**	KS	
$\tilde{M}(\varepsilon)$	TT	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.51	0.00**	0.00**	0.00**			
	FT	0.00**	0.00**	0.04**	0.00**	0.03**	0.00**	0.00**	0.00**	0.00**	0.00**			
	RS	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.94	0.00**	0.00**	0.00**			
	KS	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**			
n=2		v_{inter}	v_{cv}	$\mu_2(y)$	$\mu_2(\|d\|)$	$s(y)$	$s(\|d\|)$	FDC	$\tilde{\rho}$	$\tilde{H}(\varepsilon)$	$\tilde{h}(\varepsilon)$			

[24] S. Reid, "Fitness Landscape Analysis for Computational Finance," 2015. [Online]. Available: http://www.turingfinance.com/fitness-landscape-analysis-for-computational-finance/

[25] M. Molga and C. Smutnicki, "Test Functions for optimization algorithm," 2005.

[26] E. P. Adorio and U. P. Diliman, "MVFmultivariate test functions library in C for unconstrained global optimization." [Online]. Available: http://www.geocities.ws/eadorio/mvf.pdf

[27] M. Jamil and X.-S. Yang, "A literature survey of benchmark functions for global optimization problems," *International Journal of Mathematical Modelling and Numerical Optimisation*, vol. 4, no. 2, pp. 150–194, 2013.

[28] P. N. Suganthan, N. Hansen, J. J. Liang, K. Deb, Y. P. Chen, A. Auger, and S. Tiwari, "Problem definitions and evaluation criteria for the CEC 2017 special session and competition on single objective real-parameter numerical optimization," Nanyang Technological University, Jordan University of Science and Technology, Zhengzhou University, Singapore, Jordan, Zhengzhou, Tech. Rep., 2017.

[29] N. Hansen, A. Auger, S. Finck, and R. Ros, "Real-Parameter Black-Box Optimization Benchmarking 2009: Experimental Setup," RR-6828, INRIA, Tech. Rep., 2009.