Non-parametric model of the space of continuous black-box optimization problems

Mario A. Muñoz
School of Mathematical Sciences, Monash University
Clayton, Victoria 3800, Australia
mario.munoz@monash.edu

Kate Smith-Miles
School of Mathematical Sciences, Monash University
Clayton, Victoria 3800, Australia
kate.smith-miles@monash.edu

ABSTRACT

Exploratory Landscape Analysis are data driven methods used for automated algorithm selection in continuous black-box optimization. Most of these methods follow strong assumptions that limit their characterization power, or loose information by compressing the data into a few scalar features. A more flexible approach is to avoid explicit measuring and comparing of specific structures. In this paper we present a proof-of-concept for a more general method, which produces non-parametric models of the space of problems. Using non-metric multidimensional scaling, we generate synthetic features for each problem, which could replace or complement the existing ones. We demonstrate approaches to produce algorithm recommendations and visual representations of the space. To validate the model, we compare our results with those obtained through existing methods, which show that our models have competitive performance.

CCS CONCEPTS

• **Mathematics of computing** → **Nonparametric representations**; **Continuous optimization**;

KEYWORDS

Black-box optimization, Continuous optimization, Exploratory landscape analysis

ACM Reference format:
Mario A. Muñoz and Kate Smith-Miles. 2017. Non-parametric model of the space of continuous black-box optimization problems. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3075971

1 METHOD

The goal in a continuous black-box optimization (BBO) problem is to minimize a cost function $f : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{X} \subset \mathbb{R}^D$ is the input space, $\mathcal{Y} \subset \mathbb{R}$ is the output space, and $D \in \mathbb{N}^*$ is the dimension of the problem. A candidate solution $\mathbf{x} \in \mathcal{X}$ is a D-dimensional vector, and $y \in \mathcal{Y}$ is the candidate's cost. Let $\mathbf{x}_{\min}$ and $\mathbf{x}_{\max}$ be vectors composed of the lower and upper bounds of $\mathcal{X}$ respectively, and $\varrho_{\max} = \|\mathbf{x}_{\max} - \mathbf{x}_{\min}\|$. Let $\varphi = y_i - y_j$, and $\varrho = \|\mathbf{x}_i - \mathbf{x}_j\|$. A length scale r is defined as $|\varphi| / \varrho$ [6].

GECCO '17 Companion, Berlin, Germany

DOI: http://dx.doi.org/10.1145/3067695.3075971

Assume an uninformed random search algorithm which considers the current candidate, $\mathbf{x}_i$, to be the origin of $\mathcal{X}$. Therefore, the coordinates of any other candidate, $\mathbf{x}_j$, can be expressed in D-spherical coordinates as $[\varrho, \theta_1, \ldots, \theta_{D-1}]^\top$, where $\theta_i \in [0, 2\pi)$, $i = 1 \ldots, D-1$. Let $R \in [0, \varrho_{\max})$, $\Theta_i \in [0, 2\pi)$ and Φ be random variables such that R defines the size of a step, Θ_i its direction, and Φ the change of cost due to the step. Assume that the probability of taking a step of size R in any direction is constant. As such, we ignore the values of θ_i. Let $p(\varphi, \varrho)$ be the probability that a step of size ϱ produces a change in cost of magnitude φ for the function f.

Consider a second cost function $g : \mathcal{X} \to \mathcal{Y}$ with $q(\varphi, \varrho)$ being the probability that a step of size ϱ produces a change in cost of magnitude φ. Then, the amount of information lost when $q(\varphi, \varrho)$ is used to approximate $p(\varphi, \varrho)$ is given by the Kullback-Leiber (KL) divergence [2]. The KL-divergence is not a true metric: it is non-symmetric, zero iff $p = q$ almost everywhere, and undefined if $q = 0$ for any φ and ϱ. If $p = 0$, then $0 \ln 0 \equiv 0$. A symmetric alternative, the J-divergence, is defined as $D_J(p\|q) = D_{KL}(p\|q) + D_{KL}(q\|p)$ [6]. However, it disregards the possibility that p or q are equal to zero for some values of φ or ϱ, resulting in an undefined J-divergence. Therefore, we define a symmetric divergence, δ, as $\delta = D_{KL}(p\|q)$ if $p \neq q \wedge D_{KL}(q\|p) \leq 0$, $\delta = D_{KL}(q\|p)$ if $p \neq q \wedge D_{KL}(p\|q) \leq 0$, and $\delta = 0.5 D_J(p\|q)$ otherwise. To reduce the risk of undefined divergences due to $q = 0$ for some values, ϱ is scaled with $1/\varrho_{\max}$ and φ is normalized to variance one. The KL-divergence is estimated using the nearest neighbor with variable neighborhood method (`KL_kNN_kiTi`) [9].

Let $\{f_1, \ldots, f_m\}$ be a set of cost functions represented by their joint pdfs $\{p_1, \ldots, p_m\}$. Let $\mathbf{\Delta}$ be the matrix of cross divergences between $\{p_1, \ldots, p_m\}$, which defines the relationships between functions as a manifold embedded in a M-dimensional space. However $\mathbf{\Delta}$ cannot be used alongside other ELA methods to fit a prediction model. Therefore, we generate a $m \times M$ matrix of synthetic features, $\mathbf{\Psi}$. We use non-metric multidimensional scaling (NMDS) minimizing the Kruskal's $Stress$ (1) to generate $\mathbf{\Psi}$. To find a value of M, ten randomly seeded iterations of NMDS are carried out for $M \in [1, 30]$, until the change in $Stress$ (1) stabilizes. We follow a similar procedure to generate synthetic features from the distribution of r as defined by [6], which we call Y.

2 EXPERIMENTAL VALIDATION

As a representative subset of the space of continuous BBO problems, we use the first 30 instances from the noiseless BBOB/COCO benchmark set [1] at $D = 2$. We define a binary performance measure in which '0' represents BFGS having lower expected running time than BIPOP-CMA-ES and '1' represents the opposite. To calculate $\mathbf{\Delta}$, we take 2×10^2 sample points from $\mathcal{X}$ using Latin hypercube design

and evaluate it on each instance from the COCO benchmark. We define the matrix $\mathbf{\Lambda}$ as the comparison features [3–5, 7, 8]: (a) the dispersion at 1%, $D_{1\%}$, (b) the adjusted coefficient of determination of a purely quadratic model $\bar{R}^2_Q$, (c) the ratio between the minimum and the maximum absolute values of the quadratic term coefficients in the purely quadratic model, CN, (d) the significance of first order, $\xi^{(1)}$, (e) the skewness of the cost distribution, γ (Y), (f) the entropy of the cost distribution, H (Y), (g) the number of peaks, PKS, (h) the maximum information content, $H_{\max}$, (i) the mean cross-validation error (MCVE) of a LDA at 10%, EL_{10}, (j) the ratio between the MCVE of a LDA and a QDA at 10%, LQ_{10}, (k) the MCVE of a LDA at 25%, EL_{25}, (l) the ratio between the MCVE of a LDA and a QDA at 25%, LQ_{25}, and (m) the ratio between the MCVE of a LDA and a QDA at 50%, LQ_{50}. All levelset features and PKS were scaled using $\log_{10}$, and all features were standardized. These features are uncorrelated between each other, and correlated to other features proposed in the same references[1].

Minimizing the Kruskal's *Stress* (1) results in seven features for $\mathbf{\Psi}$, and four features for Y. Testing the correlations between features, we find that ψ_1 is highly correlated with γ (Y) and PKS, and ψ_2 is highly correlated with $\bar{R}^2_Q$. On the other hand, υ_1 is highly correlated with H (Y). No feature from $\mathbf{\Psi}$ is highly correlated to a feature from Y, which indicates that both approaches could complement each other.

We fitted a SVM to predict whether BFGS is preferred over BIPOP-CMA-ES, with seven feature sets: $\mathbf{\Lambda}$, $\mathbf{\Psi}$, Y, $\{\mathbf{\Psi}, \mathrm{Y}\}$, $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$, $\{\mathbf{\Lambda}, \mathrm{Y}\}$ and $\{\mathbf{\Lambda}, \mathbf{\Psi}, \mathrm{Y}\}$. For the last three feature sets, if two features have absolute correlation higher than 0.7, we remove the feature from $\mathbf{\Lambda}$. We fine tune the SVM and select the best features from each set using 10-fold cross-validation, whose accuracy is 96.0% for $\mathbf{\Lambda}$, 85.6% for $\mathbf{\Psi}$, 92.6% for Y, 95.7% for $\{\mathbf{\Psi}, \mathrm{Y}\}$, 96.5% for $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$, 97.4% for $\{\mathbf{\Lambda}, \mathrm{Y}\}$, and 97.6% for $\{\mathbf{\Lambda}, \mathbf{\Psi}, \mathrm{Y}\}$. Therefore, excepting the model using only $\mathbf{\Psi}$, all other models achieve a cross-validation accuracy above 90%. This can be explained by the lack of representation of H (Y) in $\mathbf{\Psi}$, which by itself achieves an accuracy of 78.9%. By including this information through Y, the accuracy increases to 95.7%. This is comparable to $\mathbf{\Lambda}$, which achieves an accuracy of 96.0%. If all features can be calculated, it is possible to obtain the maximum accuracy of 97.6%.

We employed t-SNE [10] to project the selected features from $\mathbf{\Lambda}$, $\{\mathbf{\Psi}, \mathrm{Y}\}$, and $\{\mathbf{\Lambda}, \mathbf{\Psi}, \mathrm{Y}\}$ into two dimensions, such that the relationships between problem instances can be visualized. From a group of 30 iterations we selected the one with the lowest error. Figure 1 show the results from the selected features of $\{\mathbf{\Lambda}, \mathbf{\Psi}, \mathrm{Y}\}$. This figure demonstrate how the non-parametric features produce clearer clusters than $\mathbf{\Lambda}$. For example, $\{f_{16}, f_{21}, f_{22}, f_{23}, f_{24}\}$ form clusters in their own right at the top of the figure.

3 LIMITATIONS AND WAYS FORWARD

Our next steps are: (a) to test the method with problems of higher dimensionalities; (b) to evaluate the effects of the sample size used to calculate the divergences; (c) to test other algorithms to verify the predictive power of the new features; and (d) to include some amount of information about the direction of the step, which was discarded by assuming that each candidate was the origin of the input space. For the later issue, we could assume assuming the existence of a hyper-plane described by the vectors representing two candidates, effectively reducing the space to two dimensions. A possible advantage of our method is its potential to generate an statistical manifold. Through information geometry techniques we could obtain a theoretically robust characterization method, which could provide clues regarding how to transform the structure of a problem so it can be quickly solved by a specific search algorithm.

[1]Complete results are presented in an extended version of this paper, available at http://users.monash.edu.au/~andresm/

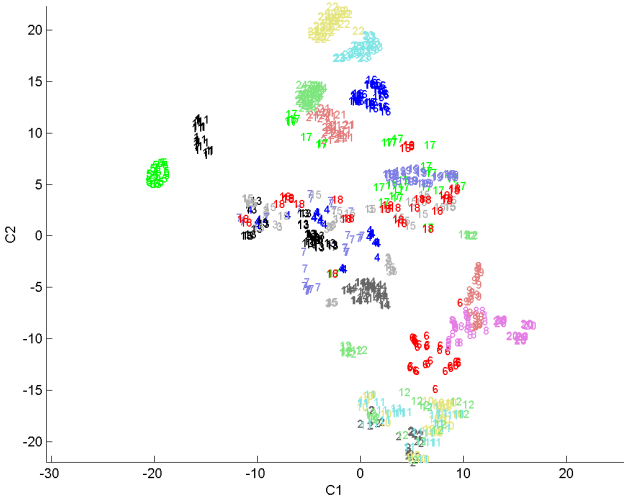

Figure 1: Two dimensional projections of the selected features from $\{\mathbf{\Lambda}, \mathbf{\Psi}, \mathrm{Y}\}$ obtained through t-SNE.

ACKNOWLEDGMENTS

This work is funded by the ARC through the Australian Laureate Fellowship FL140100012.

REFERENCES

[1] N. Hansen, A. Auger, S. Finck, and R. Ros. 2014. *Real-Parameter Black-Box Optimization Benchmarking BBOB-2010: Experimental Setup.* Technical Report RR-7215. INRIA.
[2] S. Kullback and R.A. Leibler. 1951. On Information and Sufficiency. *Ann. Math. Stat.* 22, 1 (1951), 79–86.
[3] M. Lunacek and D. Whitley. 2006. The dispersion metric and the CMA evolution strategy. In *GECCO '06*. ACM, New York, NY, USA, 477–484.
[4] J. Marin. 2012. How landscape ruggedness influences the performance of real-coded algorithms: a comparative study. *Soft Comput.* 16, 4 (2012), 683–698.
[5] O. Mersmann, B. Bischl, H. Trautmann, M. Preuß, C. Weihs, and G. Rudolph. 2011. Exploratory landscape analysis. In *GECCO '11*. ACM, New York, NY, USA, 829–836.
[6] R. Morgan and M. Gallagher. 2017. Analysing and characterising optimization problems using length scale. *Soft Comput.* 21, 7 (2017), 1735–1752.
[7] M.A. Mu noz, M. Kirley, and S.K. Halgamuge. 2015. Exploratory landscape analysis of continuous space optimization problems using information content. *IEEE Trans. Evol. Comput.* 19, 1 (2015), 74–87.
[8] D.I. Seo and B.R. Moon. 2007. An Information-Theoretic Analysis on the Interactions of Variables in Combinatorial Optimization Problems. *Evol. Comput.* 15, 2 (2007), 169–198.
[9] Z. Szabó. 2014. Information Theoretical Estimators Toolbox. *J. Mach. Learn. Res.* 15 (2014), 283–287.
[10] L. van der Maaten. 2014. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.* 15 (2014), 3221–3245.