Ordinal versus Metric Gaussian Process Regression in Surrogate Modelling for CMA Evolution Strategy

Zbyněk Pitra
National Institute of Mental Health, Czech Republic
Czech Technical University
Prague 115 19, Czech Republic
z.pitra@gmail.com

Lukáš Bajer
Charles University
Malostranské náměstí 25
Prague 118 00, Czech Republic
bajeluk@gmail.com

Jakub Repický
Charles University
Malostranské náměstí 25
Prague 118 00, Czech Republic
j.repicky@gmail.com

Martin Holeňa
Czech Academy of Sciences
Pod Vodárenskou věží 2
Prague 182 07, Czech Republic
martin@cs.cas.cz

ABSTRACT

This work presents an ordinal-based Gaussian process surrogate model for the state-of-the-art continuous black-box optimizer CMA-ES in scenarios where the objective evaluations are very expensive. Such model is motivated by the CMA-ES' invariance with respect to order preserving transformations. Alongside with the model's description, comparison with the standard (metric) Gaussian process surrogate for the CMA-ES is given.

CCS CONCEPTS

•**Computing methodologies → Continuous space search;** *Model development and analysis; Uncertainty quantification;*

KEYWORDS

black-box optimization, evolutionary optimization, surrogate modelling, Gaussian-process regression

ACM Reference format:
Zbyněk Pitra, Lukáš Bajer, Jakub Repický, and Martin Holeňa. 2017. Ordinal versus Metric Gaussian Process Regression
in Surrogate Modelling for CMA Evolution Strategy. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3076086

1 INTRODUCTION

Surrogate regression models replacing the original expensive fitness in some of the evaluated points have been in use since the early 2000s. In this paper, surrogate modelling is studied in connection with the state-of-the-art method for continuous black-box optimization – the CMA-ES (Covariance Matrix Adaptation Evolution Strategy). The considered models are Gaussian processes (GP), which differ from other common surrogate models through estimating the whole probability distribution of fitness values. To combine them with the CMA-ES is challenging because CMA-ES invariance with respect to order preserving transformations suggests ordinal regression, whereas Gaussian process continuity suggests metric regression.

In history, one of the most promising surrogate models combined with the CMA-ES, according to the results on the COCO platform bechmarks, uses ranking SVM, i. e., ordinal support vector regression [4]. On the other hand, two other surrogate models for the CMA-ES, which are similarly successful on the COCO platform, use metric regression: the response-surface model in [1] and the GP in [5]. This balance between ordinal and metric surrogates was a starting point for our investigation.

DOI: http://dx.doi.org/10.1145/3067695.3076086

2 SURROGATE MODEL BASED ON ORDINAL GP REGRESSION

2.1 Probabilistic Least Squares Ordinal Regression

The Probabilistic least squares approach of ordinal GP [6], which we have chosen for this work, consists in a linear mapping of latent GP $f(\mathbf{x})$ as $\alpha_0 - \alpha f(\mathbf{x})$ into r intervals $I_1 = (-\infty, b_1], I_2 = (b_1, b_2], \ldots, I_r = (b_{r-1}, \infty)$, where $-\infty = b_0 < b_1 < \cdots < b_{r-1} < b_r = \infty$. The probability that a random variable $f(\mathbf{x})$ with probability distribution $\mathcal{N}(\mu, \sigma)$ is mapped to a particular interval I_k, $k = 1, \ldots, r$, is

$$P(f(\mathbf{x}) \in I_k) = \Phi\left(\frac{b_k - (\alpha_0 - \alpha\mu)}{\sqrt{1 + \alpha^2\sigma^2}}\right) - \Phi\left(\frac{b_{k-1} - (\alpha_0 - \alpha\mu)}{\sqrt{1 + \alpha^2\sigma^2}}\right) = \\ = \Phi\left(\frac{\alpha\mu + \beta_k}{\sqrt{1 + \alpha^2\sigma^2}}\right) - \Phi\left(\frac{\alpha\mu + \beta_{k-1}}{\sqrt{1 + \alpha^2\sigma^2}}\right), \quad (1)$$

where Φ is the distribution function of the standard normal distribution $\mathcal{N}(0, 1)$ and $\beta_k = b_k - \alpha_0, k = 0, \ldots, r$.

Taking into account (1), the PLSOR approach estimates the likelihood that the prediction of $f(\mathbf{x}_i)$ based on the remaining training data without $(\mathbf{x}_i, y_i)$ is mapped to the same interval $I_{y_i} = (\beta_{y_i - 1} + \alpha_0, \beta_{y_i} + \alpha_0)$ to which y_i is mapped. Denoting the mean of that prediction μ_{-i} and its variance σ^2_{-i} with hyperparameters of the GP estimated only from the remaining training data, this leads to the final estimated likelihood of the observed assignment of the

Algorithm 1 Ordinal GP model training

Input: $(\mathbf{x}_i, y_i)_{i=1}^n$ (training points),
r (the number of bins for clustering),
$\boldsymbol{\theta}^0$ (initial values of latent GP hyperparameters $\boldsymbol{\theta}$),
α^0, $\{\beta_j^0\}_{j=1}^{r-1}$ (initial values of PLSOR hyperparameters α, $\{\beta_j\}_{j=1}^{r-1}$)

$\{y_i^{\text{ord}}\}_{i=1}^n \leftarrow \text{cluster}(\{y_i\}_{i=1}^n, r)$

$(\alpha, \{\beta_j\}_{j=1}^{r-1}, \boldsymbol{\theta})^* \leftarrow \underset{\alpha, \{\beta_j\}_{j=1}^{r-1}, \boldsymbol{\theta}}{\arg\max} \log \hat{\mathcal{L}}(\{y_i^{\text{ord}}\}_{i=1}^n | \{\mathbf{x}_i\}_{i=1}^n, \alpha, \{\beta_j\}_{j=1}^{r-1}, \boldsymbol{\theta})$ (see Eq. (2))

Output: $(\alpha, \{\beta_j\}_{j=1}^{r-1}, \boldsymbol{\theta})^*$ (trained model hyperparameters)

Algorithm 2 Ordinal GP model prediction

Input: $\{\mathbf{x}_i\}_{i=1}^\lambda$ (population of points),
$\boldsymbol{\theta}$ (trained latent GP hyperparameters),
α, $\{\beta_j\}_{j=1}^{r-1}$ (trained PLSOR hyperparameters)

$p_i^k \leftarrow P(f(\mathbf{x}_i) \in I_k | \mathbf{x}_i, \alpha, \{\beta_j\}_{j=1}^{r-1}, \boldsymbol{\theta}), \forall k = 1, \dots, r, \forall i = 1, \dots, \lambda$ (see Eq. (1))

$q_i \leftarrow \sum_{k=1}^r p_i^k k \quad \forall i = 1, \dots, \lambda$

$\{\mathbf{x}_{i:\lambda}\}_{i=1}^\lambda \leftarrow$ order $\{\mathbf{x}_i\}_{i=1}^\lambda$ according to $q_{1:\lambda} \leq q_{2:\lambda} \leq \dots \leq q_{\lambda:\lambda}$

Output: $\{\mathbf{x}_{i:\lambda}\}_{i=1}^\lambda$ (ordered population)

training data to the intervals $I_1, \dots, I_r$:

$$\hat{\mathcal{L}}(y_i \in I_{y_i}, i = 1, \dots, n | \{\mathbf{x}_l\}_{l=1}^n, \alpha, \beta_1, \dots, \beta_{r-1}, \boldsymbol{\theta}) = \prod_{i=1}^n \left(\Phi\left(\frac{\alpha \mu_{-i} + \beta_{y_i}}{\sqrt{1 + \alpha^2 \sigma_{-i}^2}} \right) - \Phi\left(\frac{\alpha \mu_{-i} + \beta_{y_i - 1}}{\sqrt{1 + \alpha^2 \sigma_{-i}^2}} \right) \right). \tag{2}$$

2.2 Ordinal GP in DTS-CMA-ES

We present implementation details of ordinal GP model for the DTS-CMA-ES [5].

The ordinal GP model-building phase, depicted in Algorithm 1, starts with clustering the input data $(\mathbf{x}_i, y_i)_{i=1}^n$ to intervals $I_1, \dots, I_r$. After that, the hyperparameters are selected to maximize the likelihood (2).

The ordinal GP model prediction procedure is depicted in Algorithm 2. The prediction of the ordinal class q_i of a point $\mathbf{x}_i$ is calculated as the expectation of the ordinal class values of $\mathbf{x}_i$ with respect to the probability distribution defined for $\mathbf{x} = \mathbf{x}_i$ according to (1). The output of the GP model is the ordered set of CMA-ES generated population $\{\mathbf{x}_{i:\lambda}\}_{i=1}^\lambda$, where the index $i{:}\lambda$ denotes the index of the i-th ranked point, that is $q_{1:\lambda} \leq q_{2:\lambda} \leq \dots \leq q_{\lambda:\lambda}$.

3 EXPERIMENTS ON COCO PLATFORM

We performed the experiments[1] on the noiseless part of the COCO framework to compare the proposed implementation of ordinal GP regression model (Ord-DTS) with the metric regression model from DTS-CMA-ES [5], BIPOP-^{s*}ACM-ES-k [3], the original CMA-ES [2], and lmm-CMA-ES [1].

First, PLSOR parameters resulting in the best regression performance were identified. These parameters are the kernel of the latent GP process (K_{SE}, $\mathbf{K}_{\text{Matérn}}^{\nu=5/2}$), the type of clustering (no clustering, **quantile clustering**, agglomerative hierarchical clustering), and the number of ordinal classes for clustering (μ, $\boldsymbol{\lambda}$, 2λ), where one of the best-observed values are typeset in bold and used in the experiments on the COCO benchmark.

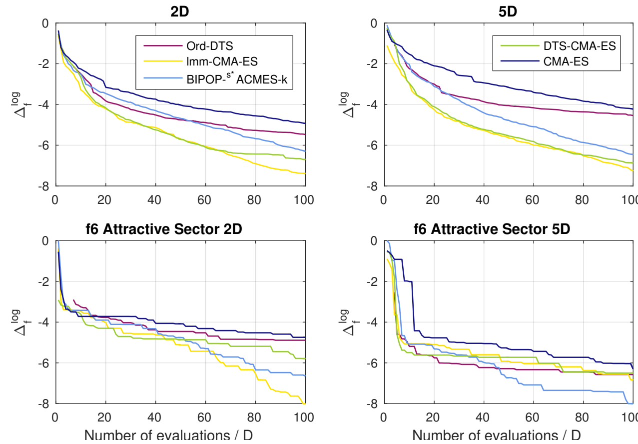

Figure 1: Comparison of optimization algorithms on 24 COCO noiseless functions and on Attractive sector function (f_6) in 2 and 5D

The summary in Figure 1 show the effect of usage of the PLSOR model instead of the metric GP in the DTS-CMA-ES optimizer on the COCO benchmark. The graphs show the scaled logarithm $\Delta_f^{\log}$ of the median of minimal distance from the function optimum over runs on 15 independent instances dependent on function evaluations divided by dimension (see [5] for details). The values are scaled to the $[-8, 0]$ interval, where -8 corresponds to the minimal and 0 to the maximal distance.

Its easy to see that the performance of PLSOR model in DTS-CMA-ES is lower than the performance of metric model with the exception of the *attractive sector* function f_6.

ACKNOWLEDGMENTS

The reported research was supported by the CSF grant No. 17-01251, the GA of the CTU No. SGS17/193/OHK4/3T/14, and by the project Nr. LO1611 with a financial support from the MEYS under the NPU I program.

REFERENCES

[1] A. Auger, D. Brockhoff, and N. Hansen. 2013. Benchmarking the Local Metamodel CMA-ES on the Noiseless BBOB'2013 Test Bed. In *GECCO 2013*. 1225–1232.
[2] N. Hansen. 2006. The CMA Evolution Strategy: A Comparing Review. In *Towards a New Evolutionary Computation*. Springer, 75–102.
[3] I. Loshchilov, M. Schoenauer, and M. Sebag. 2013. BI-population CMA-ES Algorithms with Surrogate Models and Line Searches. In *Genetic and Evolutionary Computation Conference (GECCO Companion)*. ACM Press, 1177–1184.
[4] I. Loshchilov, M. Schoenauer, and M. Sebag. 2013. Intensive Surrogate Model Exploitation in Self-Adaptive Surrogate-Assisted CMA-ES (saACM-ES). In *GECCO 2013*. 439–446.
[5] Z. Pitra, L. Bajer, and M. Holeňa. 2016. *Doubly Trained Evolution Control for the Surrogate CMA-ES.* Springer International Publishing, Cham, 59–68.
[6] P. K. Srijith, S. Shevade, and S. Sundararajan. 2012. *A Probabilistic Least Squares Approach to Ordinal Regression.* Springer Berlin Heidelberg, 683–694.

[1] the source code is available at https://github.com/repjak/surrogate-cmaes/tree/ordgp