Model Guided Sampling Optimization with Gaussian Processes for Expensive Black-Box Optimization

Lukáš Bajer
Faculty of Mathematics and Physics, Charles University
Malostranske nam. 25
118 00 Prague 1
and
Institute of Computer Science
Academy of Sciences
Czech Republic
bajer@cs.cas.cz

Viktor Charypar
Faculty of Nuclear Sciences and Physical Engineering
Czech Technical University in Prague
Trojanova 13
120 00 Prague 2
Czech Republic
charyvik@fjfi.cvut.cz

Martin Holeňa
Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodarenskou vezi 2
182 07 Prague 8
Czech Republic
martin@cs.cas.cz

ABSTRACT

Model Guided Sampling Optimization (MGSO) is a novel expensive black-box optimization method based on a combination of ideas from Estimation of Distribution Algorithms and global optimization methods using Gaussian Processes. The algorithm is described and its implementation tested on three benchmark functions as a proof of concept.

Categories and Subject Descriptors

G.1.6 [**Numerical Analysis**]: Optimization—*global optimization, unconstrained optimization*; F.2.1 [**Analysis of Algorithms and Problem Complexity**]: Numerical Algorithms and Problems

General Terms

Algorithms

Keywords

Benchmarking; Black-box optimization; Modeling; Gaussian Processes

1. INTRODUCTION

When optimizing expensive objective functions with evolutionary algorithms, the cost of the function evaluation dominates the cost of the whole optimization. Our goal is therefore to exploit the knowledge of all the previous evaluations. A typical solution in this case is to employ a model based on the evaluated candidate solutions. We can then either replace the objective function with this model for some evaluations (a practice known as surrogate modeling) or use the model to guide the selection of candidate solutions. We are concerned with the latter.

Estimation of Distribution Algorithms (EDA) use such approach of generating solutions based on a probabilistic model. An EDA evolves a population of candidate solutions in generations, each of which starts with solutions evaluation and selection of the promising ones. A model of the distribution of the better solutions is then built and new solutions are sampled from it. A comprehensive overview of EDAs and related methods can be found in [5].

Another approach is using Gaussian Process (GP) model for optimization fitting a GP regression model to the data and selecting next candidate solution based on the predictive distribution given by the GP. The most promising candidate is added to the dataset and a new model is fitted on the resulting data. A number of candidate selection criteria based on predictive distribution were examined by Jones et al. [4] and recently by Hennig and Schuler [3]; the Probability of Improvement (PoI) is of interest for this work.

We propose a novel optimization method – Model Guided Sampling Optimization (MGSO) – based on a combination of the GP optimization and EDAs. One can either view the novelty as replacing the distribution model in EDA with a probability of improvement of a GP model fitted to all known data, or, from the latter point of view, as, instead of maximizing the PoI to find a single candidate solution, sampling a set of several candidate solutions proportionally to the PoI.

Copyright is held by the author/owner(s).
GECCO'13, July 6-10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

2. GAUSSIAN PROCESS REGRESSION

A Gaussian process [6] is an infinite-dimensional probability density such that each finite-dimensional marginal is a multivariate Gaussian. The infinite-dimensional realizations are functions, and the realizations of the marginals are values of those functions at locations $\{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^N$. A prediction in a testing point then gives a one-dimensional Gaussian distribution.

A Gaussian process is completely specified by its mean and covariance functions

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right) \\ m(\mathbf{x}) &= \mathbb{E}\left[f(\mathbf{x})\right] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right]. \end{aligned}$$

We use a constant mean function and a squared exponential covariance function with automatic relevance determination

$$k_{\mathrm{SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^\top M(\mathbf{x} - \mathbf{x}')\right)$$

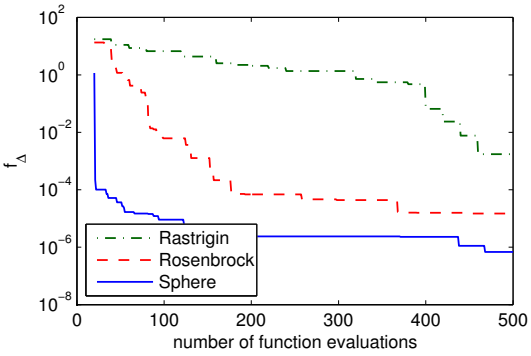

Figure 1: Medians (of 15 trials) of distances to optimum (f_Δ) for three benchmark functions

where $M = \boldsymbol{\ell}^{-2} I$, $\boldsymbol{\ell}$ is a vector of characteristic lengths of the process in each coordinate dimension and σ_f^2 is a noise level. The hyperparameters $\theta = \{\boldsymbol{\ell}, \sigma_f^2\}$ are fitted by maximizing the likelihood $p(\mathbf{y}|X, \theta)$ ($\mathbf{y}$ are the observations, X is the input points matrix).

Since the GP gives a one-dimensional Gaussian distribution as prediction, we can think of the distribution's variance as confidence of the model in its prediction and use it to guide optimization. Based on both mean and standard error of the prediction, we can easily get a probability of improvement [4] over a certain target T as

$$\mathrm{PoI} = \Phi\left(\frac{T - \mu_*}{\sigma_*}\right)$$

where μ_* is mean and σ_* standard error of the model prediction (both returned by the GP) and Φ is a standard $\mathcal{N}(0, 1)$ cumulative distribution function.

MGSO samples this probability of improvement to get a next set of candidate solutions.

3. THE MGSO

The basic idea of MGSO is similar to methods proposed by Jones [4]: Fit a GP model to an initial sample of data and let the model predictive distribution guide the optimization. The key difference is that MGSO does not maximize the PoI used as a criterion to find a single solution candidate to evaluate. It rather samples the PoI distribution to get a new population of candidate solutions. Those are then evaluated, the model is fitted to this augmented dataset and the process is repeated until a stopping condition is met.

To abstract from the search space scaling, MGSO works in an internal, linearly transformed coordinate system mapped to $[-1, 1]^D$ and the transformation can be updated during the optimization. Updating it restricts the learning and sampling space to a neighborhood of the optimum, which enables sampling in the situation when the PoI is non-zero only in a very small region. Finally, when the neighborhood of the optimum is sufficiently sampled, a local search of the GP is performed to fine tune the optimum.

Formally, the algorithm proceeds as follows:

1. Generate N_i initial samples $\{\mathbf{x}_0\}$ and evaluate them to get observed values $\{y_0\}$ forming a dataset $\mathcal{S}_0 = \{(\mathbf{x}_0, y_0)\}$
2. Until a stopping condition is met, for $i = 1, 2, \ldots$ repeat steps 3–8
3. Build a GP model $\mathcal{M}_i$ and fit its hyperparameters θ to the dataset $\mathcal{S}_{i-1}$
4. Sample N new candidate solutions $\{\mathbf{x}_i\}$ based on the PoI distribution of $\mathcal{M}_i$, optionally with different targets T as proposed in [4]
5. Evaluate $\{\mathbf{x}_i\}$ to get $\{y_i\}$
6. Augment the dataset obtaining $\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{(\mathbf{x}_i, y_i)\}$
7. Store the best $(\tilde{\mathbf{x}}, \tilde{y}) \in \mathcal{S}_i : \tilde{y} \leq y_k \, \forall (\mathbf{x}_k, y_k) \in \mathcal{S}_i$
8. If rescale conditions are met, restrict the dataset to $\mathbf{x} \in \times_{i=1}^{D} [l_i, u_i]$ and transform it to $[-1, 1]^D$
9. Return the best $(\tilde{\mathbf{x}}, \tilde{y})$

The bounds l_i and u_i in each coordinate dimension are found as a bounding box of ten nearest samples from the current optimum expanded by 10%. The initial number of samples N_i and population size N are input parameters.

Sampling is performed using the Gibbs method [1] enabling us to sample multivariate distributions. In our case, we sample candidate solutions from the empirical distribution proportional to the PoI. Samples resulting in ill-conditioned covariance matrices of the GP are rejected.

4. PRELIMINARY RESULTS

As a proof of concept, the method was tested on three benchmark functions from the BBOB toolbox [2] in 2D. For each function, 15 optimization runs were performed and best function values f_{best} in each generation were recorded. Figure 1 shows the median of distance to optimum $f_\Delta = f_{\mathrm{best}} - f_{\mathrm{opt}}$ reached using a given number of function evaluations (limited to 500) for each benchmark.

5. ACKNOWLEDGMENTS

This work was supported by the Czech Science Foundation (GAČR) grants P202/11/1368 and 13-17187S, the Grant Agency of the Charles Univ. (GAUK) 278511/2011 grant, and the Grant Agency of the Czech Technical University in Prague, grant No. SGS12/196/OHK3/3T/14.

6. REFERENCES

[1] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[2] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Technical Report RR-6829, INRIA, 2009. Updated February 2010.

[3] P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *J. Mach. Learn. Res.*, 13:1809–1837, 2012.

[4] D. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21:345–383, 2001.

[5] M. Pelikan, K. Sastry, and E. Cantú-Paz. *Scalable Optimization via Probabilistic Modeling*, volume 33 of *Studies in Computational Intelligence*. Springer, 2006.

[6] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.