Continuous black-box optimization in linearly constrained domains using efficient Gibbs sampling

Christian L. Müller[a]

[a]Department of Computer Science, ETH Zürich
Institute of Theoretical Computer Science and Swiss Institute of Bioinformatics
Universitätstrasse 6, 8092 Zürich
christian.mueller@inf.ethz.ch

Abstract

We propose to combine state-of-art continuous black-box optimization heuristics that use the multivariate normal distribution as search operator with an efficient Gibbs sampler for the truncated normal distribution when the search domain is subject to linear inequality constraints. This synthesis provides a generic way for constrained continuous black-box optimization because the optimizer is guaranteed to only operate on the feasible domain. No problem- or domain-specific constraint-handling techniques have thus to be developed for the given optimization task. The proposed sampler works for normal distributions with arbitrary mean and covariance structure for any number of linear constraints that form a non-empty domain. Using the Gibbs sampling methodology, the computational complexity of generating constrained samples is poly(n). As a proof of concept we couple the sampler with two state-of-the-art search heuristics: (i) the Evolution Strategy with Covariance Matrix Adaptation and (ii) Gaussian Adaptation. We present numerical examples that show the efficacy and efficiency of our approach on selected test problems from the field of evolutionary computation and mathematical programming.

Introduction

Many of today's state-of-the-art continuous black-box optimization heuristics use the multivariate normal distribution as a means to iteratively generate new candidate solutions in the search space. Prominent examples are Evolution Strategies (ES) [1], Evolution Strategies with Covariance Matrix Adaptation (CMA-ES) [2], several Estimation of Distribution (EDA) algorithms [3], Gaussian Adaptation (GaA) [4, 5], and the Cross-Entropy method in continuous domains [6]. These methods are usually designed to solve black-box minimization problems over the unconstrained search space $\mathbb{R}^n$:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})\,. \tag{1}$$

In general, only zeroth-order information can be extracted from the black-box objective function $f(\mathbf{x})$. Gradient or higher-order information about $f(\mathbf{x})$ are not available or do not exist. Moreover, no properties about the function $f(\mathbf{x})$ such as convexity or linearity are assumed. Optimization problems of this kind frequently arise in engineering applications where $f(\mathbf{x})$ is, e.g., the output of some complex computer simulation or a real-world experiment, and $\mathbf{x} = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$ are design variables or parameters. The aforementioned black-box optimization heuristics explore the parameter space by sampling new candidate solutions from a multivariate normal distribution $\mathcal{N}(\mathbf{m}, \mathbf{C})$ at each iteration (or generation). The information about sample positions and corresponding objective function values is then used by the respective algorithm to adapt mean $\mathbf{m} \in \mathbb{R}^n$ and covariance matrix $\mathbf{C} \in \mathbb{R}^{n\times n}$ of the distribution. This process is repeated until some user-defined convergence criteria are met.

In many practical situations, however, more refined knowledge about the search domain is available *a priori*. Variables may be restricted to the non-negative orthant $\mathbf{x} \geq \mathbf{0}$, to an n-dimensional box $\mathbf{x} \in [\mathbf{l}, \mathbf{u}] \subset \mathbb{R}^n$

(with n-dimensional vectors $\mathbf{l}, \mathbf{u} \in \mathbb{R}^n$ specifying the lower and upper bounds in each dimension), or to the n-dimensional unit simplex $\sum_{i=1}^{n} x_i \leq 1$. Feasible regions Ω of this kind can be expressed in terms of a set of m linear inequalities of the form $\Omega = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. The canonical form of a black-box optimization problem subject to linear inequality constraints thus reads:

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b}\,. \end{aligned} \tag{2}$$

Due to the fact that the multivariate normal distribution is defined over $\mathbb{R}^n$, the natural question arises how optimization heuristics that use this distribution for search should handle samples that lie outside the feasible region. In the Evolutionary Computation community, a variety of *heuristic* methods have been developed in the past decades for general constraint-handling (where the constraints may even be non-linear or not explicitly given). We refer to [7] for a general review and to [8] for a recent review related to ES. The arguably most common approaches are projection methods, penalty function methods, and rejection sampling. In the projection method, infeasible solutions are projected back onto the feasible domain. In the penalty function method, a term is added to the objective function that is proportional to the “degree of infeasibility" of the solution and vanishes if all constraints are satisfied. Both approaches are unsatisfactory because it is often not easy to either find an effective projection operator or to properly balance the influence of the penalty function on the objective function. In rejection sampling, infeasible solutions are discarded by the optimizer and resampled until only feasible solutions are present. While this approach is the only true *problem-independent* constraint-handling technique, it is, unfortunately, in many cases highly impractical. This stems from the fact that the acceptance ratio can become arbitrarily small during the optimization process, eventually leading to a complete halt of the search.

In this contribution we follow a different line of thought. Rather than sampling from an unconstrained normal distribution and repairing or resampling infeasible solutions, we present a Gibbs sampler for the truncated normal distributions $\mathcal{N}_\Omega(\mathbf{m}, \mathbf{C})$ that is guaranteed to produce only samples within the feasible domain in polynomial run time.

A Gibbs sampler for truncated normal distributions in linearly constrained domains

Gibbs sampling, originally introduced in a seminal paper by Geman and Geman [9], is a Markov Chain Monte Carlo technique that allows to efficiently sample from an n-dimensional joint distribution $\mathbf{p}(\mathbf{x})$ when knowledge about its conditional distributions is available. For instance, when the one-dimensional conditionals $\mathbf{p}(x_j|\mathbf{x}_{-j}), j = 1, \ldots, n$ with $\mathbf{x}_{-j} = [x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n]^T$ are explicitly given and easy to sample from, the Gibbs sampler generates an n-dimensional sample from the joint distribution by constructing a Markov chain that sequentially sweeps over all dimensions from some feasible starting point. The j^{th} dimension is sampled using $\mathbf{p}(x_j|\mathbf{x}_{-j})$ conditional on the current location of the chain. The computational advantage of the Gibbs sampler is the unconditional acceptance of the new sample location thus rendering the run time linear in n for *any* single sample. A set of k samples generated in this way is provably an identically independently distributed (i.i.d) but correlated sample from the distribution. In order to reduce the correlation of the chain, additional “thinning" of the chain is usually applied, i.e., only every t^{th} n-dimensional sample of the chain is added to the final sample set. We can construct such a Gibbs sampler for the following truncated multivariate normal distribution

$$\mathbf{p}(\mathbf{x}) \propto \mathcal{N}_\Omega(\mathbf{m}, \mathbf{C}) = \begin{cases} \mathcal{N}(\mathbf{m}, \mathbf{C}) & \text{if } \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

because the corresponding one-dimensional conditional distributions are truncated one-dimensional normals for which efficient sampling schemes exist. For the exact formulae of the conditional means and variances we refer to [10]. Geweke [11] was among the first to present a Gibbs sampler for a special case of Eq. (3) where $\mathbf{A}$ is allowed to contain at most n linearly independent constraints, i.e., (rotated) box constraints. Rodriguez-Yam et al. [13] developed an elegant generalization for arbitrary convex polyhedra in the context of Bayesian statistics. Let $\mathbf{x} \sim \mathcal{N}_\Omega(\mathbf{m}, \mathbf{C})$ be an n-dimensional multivariate normal vector. We first decompose $\mathbf{C} = \sigma^2 \Sigma$ with σ a scalar. Let $\mathbf{T} \in \mathbb{R}^{n \times n}$ be a matrix of full rank such that $\mathbf{T}\Sigma\mathbf{T}^T = \mathbf{I}$ where $\mathbf{I}$ denotes the n-dimensional identity matrix. $\mathbf{T}$ can be found by Cholesky or eigenvalue decomposition of Σ because the rescaled covariance matrix Σ is positive definite. Let $\mathbf{z} = \mathbf{T}\,\mathbf{x}$ and $\mathbf{c} = \mathbf{T}\,\mathbf{m}$. Sampling $\mathbf{z}$ from a truncated normal distribution thus reads:

$$\mathbf{p}(\mathbf{z}) \propto \mathcal{N}_\mathbf{S}(\mathbf{c}, \sigma^2\mathbf{I}) = \begin{cases} \mathcal{N}(\mathbf{c}, \sigma^2\mathbf{I}) & \text{if } \mathbf{D}\mathbf{z} \leq \mathbf{b} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\mathbf{D} = \mathbf{A}\,\mathbf{T}^{-1}$ and $\mathbf{S} = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{D}\mathbf{z} \leq \mathbf{b}\}$ and the original $\mathbf{x}$ can be recovered by $\mathbf{x} = \mathbf{T}^{-1}\mathbf{z}$. This reformulation drastically simplifies the sampling distribution but not the constraints. Let $\mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_j, \ldots, \mathbf{z}_n]^T$ and $\mathbf{z}_{-j} = [z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_n]^T$. A Gibbs sampler for the multivariate distribution in Eq. (4) thus generates in each sweep components z_j of $\mathbf{z}$ according to

$$\mathbf{p}(z_j | \mathbf{z}_{-j}) = \mathcal{N}_{S_j}(c_j, \sigma^2)\,, \tag{5}$$

where $S_j = \{z_j \in \mathbb{R}, \mathbf{z} \in \mathbb{R}^n : \mathbf{D}\mathbf{z} \leq \mathbf{b}\}$. Let $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_j, \ldots, , \mathbf{d}_n]$ with $\mathbf{d}_j \in \mathbb{R}^m$ and $\mathbf{D}_{-j}$ the matrix $\mathbf{D}$ without the j^{th} column $\mathbf{d}_j$. S_j can then be computed from the set of at most m linear inequalities $S_j = \{z_j \in \mathbb{R} : \mathbf{d}_j z_j \leq \mathbf{b} - \mathbf{D}_{-j}\mathbf{z}_{-j}\}$ the solution of which forms a one-dimensional convex set, i.e. either an (left/right) open interval or a closed interval. Note that the current position of the chain only appears in the sequential calculation of the intervals but not in the sample generation. Rodriguez-Yam and co-workers showed that the Markov chains produced by this Gibbs sampler exhibit excellent mixing behavior. Sequential samples are virtually uncorrelated compared to the ones produced by Geweke's sampler [13]. Moreover, our own numerical experiments revealed that, for high-conditioned covariance matrices $\mathbf{C}$, Geweke's sampler is in fact numerically unstable because the conditional one-dimensional variances cannot be calculated any more. We implemented the described Gibbs sampler both in MATLAB and Fortran 90 using BLAS/LAPACK for all relevant linear algebra calculations as well as a MATLAB-mex interface for the Fortran 90 code. The software as well as run time benchmark results will soon be made publicly available at `http://www.mosaic.ethz.ch/`.

Numerical examples for constrained black-box optimization

We now present two novel combinations of a search heuristic and Gibbs sampling and sketch their performance on two different test problems. To date, the only available study that proposes a related approach is the PolyEDA [12], a specific EDA algorithm that uses Geweke's sampler and is thus limited to problems with (rotated) box constraints. We first combine our sampler with Hansen's CMA-ES (version 2.55 of the MATLAB implementation), an ES variant that showed remarkable performance over a wide range of synthetic and real-world in the past years. We replace its standard constraint-handling and sampling mechanism with the Gibbs sampler and test the performance of this novel constrained CMA-ES algorithm on the tangent problem:

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} \quad & f_{\text{TR}}(\mathbf{x}) = \textstyle\sum_{i=1}^n x_i^2 \\ \text{s.t.} \quad & -\textstyle\sum_{i=1}^n x_i + n \leq 0\,. \end{aligned} \tag{6}$$

This problem served as a key test problem in Kramer's recent review on constraint-handling in ES [8]. The optimal solution is $\mathbf{x}^* = [1, \ldots, 1]^T$ with objective function value $f_{\text{TR}}(\mathbf{x}^*) = n$. Kramer considers the two-dimensional problem throughout his review and compares the performance of various ES as well as CMA-ES with several heuristic constraint-handling techniques. Unfortunately, initial starting points $\mathbf{x}_0$ or the initial step size σ_0 for CMA-ES are not specified for these experiments. We chose $\mathbf{x}_0 = [10, 10]^T$ and $\sigma_0 = 1$ and conducted 25 experiments. Constrained CMA-ES has been stopped when the optimum was found within $\epsilon < 1e^{-16}$. Minimum, median, and maximum number of function evaluations are $764, 992$, and 1346. For comparison, the best combination of CMA-ES and a sophisticated meta-model strategy for the constraints needs on average 3,432 function evaluations and 5,326 constraint evaluations [8].

In our second algorithm we embed the Gibbs sampler into the recently revisited GaA method [4, 5], a black-box optimizer that is rooted in Jaynes' Maximum Entropy principle. We consider the problem of solving linear programs over Klee-Minty cubes. This problem is a classic in mathematical programming because Dantzig's standard simplex method has proven *exponential* run time on this instance. We chose Kitahara and Mizuno's parameterization of the Klee-Minty cube [14]. The problem reads:

$$\begin{array}{cc} \max\limits_{\mathbf{x}\in\mathbb{R}^n} & f_{\mathrm{KM}}(\mathbf{x}) = \sum_{i=1}^{n} x_i \\ \text{s.t.} & x_1 \leq 1 \\ & 2\sum_{i=1}^{k-1} x_i + x_k \leq 2^k - 1, \quad k = 2,\ldots,n \\ & \mathbf{x} \geq 0 \end{array} \tag{7}$$

The optimal solution is located at $\mathbf{x}^* = [0, 0, \ldots, 1]^T$ with $f_{\mathrm{KM}}(\mathbf{x}^*) = 2^n - 1$. The described Klee-Minty polyhedron has an interesting shape because the axis length along the i^{th} dimension grows exponentially with i, thus leading to an increasingly anisotropic search space in higher dimensions. For $n \leq 16$, we empirically observe that constrained GaA with Gibbs sampling achieves *polynomial* convergence (in terms of function evaluations) for any fixed ϵ when started from the origin with an initial step size on the order of 2^n. A complete analysis of this test case will be provided in a future extended version of this contribution.

References

[1] Hans-Paul Paul Schwefel. *Evolution and Optimum Seeking: The Sixth Generation.* John Wiley & Sons, Inc., New York, NY, USA, 1993.

[2] Nikolaus Hansen and Andreas Ostermeier. Completely Derandomized Self-Adaption in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[3] Pedro Larranaga and Jose A. Lozano, editors. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation (Genetic Algorithms and Evolutionary Computation).* Springer, October 2001.

[4] Gregor Kjellström and Lars Taxen. Stochastic Optimization in System Design. *IEEE Trans. Circ. and Syst.*, 28(7), July 1981.

[5] Christian L. Müller and Ivo F. Sbalzarini. Gaussian adaptation revisited - an entropic view on covariance matrix adaptation. In C. Di Chio et al., editor, *EvoApplications*, volume I of *Lecture Notes in Computer Science*, pages 432–441. Springer, 2010.

[6] Reuven Y. Rubinstein and Dirk P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning (Information Science and Statistics).* Springer, 1 edition, July 2004.

[7] Zbigniew Michalewicz. A survey of constraint handling techniques in evolutionary computation methods. In *Proceedings of the 4th Annual Conference on Evolutionary Programming*, pages 135–155. MIT Press, 1995.

[8] Oliver Kramer. A Review of Constraint-Handling Techniques for Evolution Strategies. *Applied Computational Intelligence and Soft Computing*, Vol. 2010, 2010.

[9] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[10] Christian P. Robert. Simulation of truncated normal variables. *Statistics and Computing*, 5:121–125, 1995. 10.1007/BF00143942.

[11] John Geweke. Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities. Technical report, Department of Economics, University of Minnesota, Minneapolis, MN 55455, April 1991.

[12] Jörn Grahl and Franz Rothlauf. PolyEDA: Combining Estimation of Distribution Algorithms and Linear Inequality Constraints. In *GECCO '04: Proceedings of the 6th Annual Conference on Genetic and Evolutionary Computation*, volume 3102 of *Lecture Notes in Computer Science*, pages 1174–1185. Springer, 2004.

[13] Gabriel Rodriguez-Yam, Richard A. Davis, and Louis L. Scharf. Efficient Gibbs Sampling of Truncated Multivariate Normal with Application to Constrained Linear Regression. *submitted*, 2004.

[14] Tomonari Kitahara and Shinji Mizuno. Klee-minty's lp and upper bounds for dantzig's simplex method. *Operations Research Letters*, 39(2):88 – 91, 2011.