Diss. ETH No. 19438

Black-box Landscapes: Characterization, Optimization, Sampling, and Application to Geometric Configuration Problems

A dissertation submitted to
ETH Zürich

for the degree of
Doctor of Sciences

presented by
Christian L. Müller
M.Sc., Uppsala Universitet
Dipl. Inf./Bioinf., University of Tübingen

born 14 april 1979
citizen of Schesslitz - Germany

accepted on the recommendation of
Prof. Dr. Ivo F. Sbalzarini, examiner
Dr. Nikolaus Hansen, co-examiner
Dr. Philippe H. Hünenberger, co-examiner
Prof. Dr. Emo Welzl, co-examiner
Dr. Bojan Žagrović, co-examiner

2010

Abstract

In many areas of science and engineering researchers consider systems that can be solely examined by their input and output characteristics without any knowledge of their internal workings. Such black-box systems are the topic of the present thesis. In many practical cases, a black box comprises a complex mathematical model, a computer simulation, a real-world experiment, or a combination of any of these. In this thesis we take an interdisciplinary approach to the characterization, optimization, and sampling of black-box systems. We focus on systems with high-dimensional real-valued input variables and output patterns that can be transformed by some function into a scalar real-valued quantity. Throughout this thesis we conceptualize the black-box system as a *landscape*. Inspired by our shared visual experience of natural terrains and sceneries, we consider the real-valued input variables as a high-dimensional landscape domain. Neighborhood or nearness in this landscape domain must be provided by a suitable distance metric. We interpret the scalar output quantity as a height or elevation over the landscape domain. The landscape metaphor encourages a characterization of black-box systems in terms of topographical features, such as valleys, ridges, mountain peaks, and plateaus. In order to underline that we view black-box systems as high-dimensional, complex landscapes we introduce the notion of the *black-box landscape*. After a general review of the landscape paradigm, spanning the disciplines of biology, physics, chemistry, and optimization, we present a number of statistical landscape descriptors that probe different properties of black-box landscapes. The core of the thesis is concerned with black-box optimization. We improve the performance of the arguably best state-of-the-art optimizer, the Covariance Matrix Adaptation Evolution Strategy (CMA-ES), in various aspects. The general performance is increased by considering quasi-random instead of pseudo-random sampling. For multi-funnel landscape topologies we introduce parallel CMA-ES schemes that can outperform standard CMA-ES. We also revisit Gaussian Adaptation, an optimization and sampling scheme that has been largely ignored in the black-box optimization community. Our improved Gaussian Adaptation scheme shows remarkable performance on the considered benchmarks and ranks among the best known black-box optimizers. An important conceptual result is that we can provide an explicit link between black-box optimization and black-box (or indirect) sampling through Gaussian Adaptation. We show that the same idea of adaptation has emerged in these disparate fields, and we argue that a unifying framework for sampling and optimization might constitute an important contribution. We further consider geometric configurations in two different contexts: Geometry optimization problems of atomic clusters are proposed as novel benchmarks for black-box optimization. We design a balanced set of problems that should be included in future black-box optimization benchmarks. We also revisit the configuration space of chain molecules with respect to a certain distance measure, the Root Mean Square Deviation (RMSD) after optimal superposition. Because RMSD is the most important distance metric

in structural biology, we quantify the neighborhood structure that is induced by the RMSD for the Random Walk polymer model. Based on numerical results from black-box optimization runs, we are also able to formulate a conjecture about an upper bound of the RMSD between any two Random Walks of arbitrary length. In the course of the thesis, two software libraries for black-box sampling and optimization, GaALib and pCMALib, have been developed that might prove valuable for the scientific community.

Zusammenfassung

Viele Systeme und Modelle in Wissenschaft und Technik können aufgrund ihres hohen Komplexitätsgrades nur noch bezüglich ihrer Ein- und Ausgangseigenschaften beschrieben und analysiert werden. In vielen Fällen ist detailliertes Wissen über interne Systemabläufe und -zusammenhänge nicht mehr zugänglich. Solche, so genannte Black-Box-Systeme sind das Thema der vorliegenden Arbeit. Komplexe, mathematische Modelle, Computersimulationen, aufwendige Laborexperimente sowie beliebige Kombinationen von Labor- und Computerexperimenten lassen sich als Black-Box-Systeme modellieren. Die vorliegende Arbeit präsentiert einen interdisziplinären Ansatz zur Charakterisierung, Optimierung und zum randomisierten Abtasten solcher Systeme, wobei das Hauptaugenmerk auf Modellen mit hochdimensionalen, reellwertigen Eingangsgrössen und skalaren, reellwertige Ausgangsgrössen liegt. Eine Besonderheit dieser Arbeit liegt in der Betrachtungsweise eines Black-Box-Systems als hochdimensionale, abstrakte Landschaft: die Black-Box-Landschaft. Diese Metapher ermöglicht einen anschaulichen, topographisch inspirierten Zugang zur Systemanalyse. Die reelle Eingangsgrössen definieren darin einen hochdimensionalen Raum, die skalare Ausgangsgrösse eine Höhenangabe für jeden Punkt im Raum. Nachbarschaft oder Nähe in einer solchen Landschaft wird durch ein geeignetes Abstandsmass, z.B. die Euklidische Distanz, bestimmt. Eine Charakterisierung von Black-Box-Systemen kann nun mit Hilfe topographischer Begriffe, wie zum Beispiel Täler, Grate, Gipfel oder Plateaus, erfolgen. Das Landschaftsparadigma ist ein zentraler Bestandteil der Molekularphysik, der Evolutionsbiologie sowie der kombinatorischen Optimierung. Nach einer Analyse der wichtigsten Arbeiten aus diesen Wissenschaftsgebieten stellen wir eine Reihe von statistischen Verfahren vor, mit denen sich verschiedene Merkmale von Black-Box-Landschaften beschreiben lassen. Ein wichtiger Bestandteil dieser Arbeit ist die effiziente Optimierung von Black-Box-Systemen. Wir verbessern verschiedene Komponenten einer der besten Black-Box-Optimierungsmethoden, der Evolutionsstrategie mit Kovarianzmatrixanpassung (Covariance Matrix Adaptation Evolution Strategy, CMA-ES). Das Abtastverfahren der Strategie wird durch die Verwendung von Quasi-Zufallszahlen anstelle von Pseudozufallszahlen für die Generierung von Stichproben gesteigert. Für die effiziente Exploration von Black-box-Landschaften, die mehrere tiefe, trichterförmige Täler aufweisen, d.h. für Systeme, die weit auseinander liegende Bereiche im Eingangsraum besitzen, die ähnlich optimale Ausgangsgrössen liefern, führen wir parallele CMA-ES-Suchmethoden ein. Diese Strategien können die Effizenz im Vergleichen zu sequentiellen Varianten der CMA-ES für bestimmte Modellprobleme steigern. Darüber hinaus greifen wir die Methode der Gauss'schen Anpassung (Gaussian Adaptation, GaA) wieder auf, einem Optimierungs- und Abtastverfahren, dem bislang in der Wissenschaftsgemeinde wenig Beachtung geschenkt wurde. Wir verbessern das ursprüngliche Verfahren und demonstrieren seine Effektivität auf einer grossen Klasse von Testproblemen. Darüber hinaus weisen wir nach, dass die Methode

der Gauss'schen Anpassung die Möglichkeit eröffnet, die Optimierung und Stichprobennahme für Black-Box-Systeme zu vereinheitlichen. Geometrische Konfigurationsprobleme werden in dieser Arbeit in zweierlei Hinsicht berücksichtigt. Zum einen entwerfen wir ein neuartiges Set von geometrischen Optimierungsproblemen, das auf der Energieminimierung atomarer Cluster beruht. Wir analysieren die Topographie der resultierenden Energielandschaften und zeigen, dass die behandelten Probleminstanzen als anspruchsvolle Benchmarks für Black-Box-Optimierungsmethoden dienen können. Zum zweiten beschäftigen wir uns mit dem Konfigurationsraum von Kettenmolekülen in Bezug auf eine bestimmte Distanz, die mittlere quadratische Abweichung (Root Mean Square deviation, RMSD) nach optimaler Superposition. Da RMSD die wichtigste Distanzmetrik der Strukturbiologie darstellt, quantifizieren wir die von ihr induzierte Nachbarschaftstruktur für das einfachste Polymermodell, das Random-Walk-Modell. Darüber hinaus ermöglicht eine Kombination von numerischen Black-Box-Optimierungsexperimenten und geometrischen Überlegungen das Aufstellen einer Vermutung über eine obere Schranke für den RMSD zwischen zwei beliebigen Random-Walks beliebiger Länge. Im Laufe der Arbeit wurden des weiteren zwei öffentlich zugänglich Softwarebibliotheken für Black-Box-Optimierung und Black-box-Stichprobennahme entwickelt, GaALib und pCMALib, die der Wissenschaftsgemeinde möglicherweise von Nutzen sein können.