Abstract

In this thesis, we develop a novel class of discrete-time derivative-free optimization algorithms for unconstrained optimization problems. Our key idea is a new procedure to extract gradient information of an objective function using compositions of non-commutative maps. Those are defined by function evaluations and applied in such a way that gradient descent steps are approximated.

The procedure to construct a gradient approximation is based on two main ingredients: 1) a periodic exploration sequence that defines where the objective function is to be evaluated; 2) so-called gradient-generating functions that are composed with the objective function in such a way that an approximation of the gradient is obtained. Both ingredients can be characterized by a set of nonlinear equations. We propose a way to solve these equations, and we show how this leads to a derivative-free optimization algorithm with semi-global convergence properties.

The theoretical findings are supplemented with numerical results. A qualitative and quantitative simulation study is presented in which we investigate suitable design parameters, convergence speed, and gradient approximation errors of the proposed algorithm class. Further, we introduce various tuning rules such as variable step size schemes and adaptive exploration sequences. The algorithms we have developed are applied to challenging benchmarking problems and we compare them with other derivative-free optimization algorithms in their class. We validate that the presented algorithms are robust against noisy function evaluations, are able to deal with discontinuous objective functions, and potentially overcome local minima.

Deutsche Kurzfassung

In der vorliegenden Arbeit wird eine neue Klasse an zeitdiskreten ableitungsfreien Optimierungsalgorithmen für unbeschränkte Optimierungsprobleme entwickelt. Unsere Hauptidee ist ein neues Verfahren zur Bestimmung von Gradienteninformationen einer Funktion mittels Kompositionen nichtkommutativer Abbildungen. Diese sind durch Funktionsauswertungen definiert und ihre Anwendung führt zur Approximation von Gradientenabstiegsschritten.

Die Konstruktion dieses Verfahrens zur Gradientenapproximation basiert auf den folgenden zwei Bestandteilen: 1) periodische Explorationssequenzen, welche den nächsten Punkt zur Funktionsauswertung bestimmen; 2) Funktionen zur Gradientenerzeugung, bestehend aus Funktionsauswertungen der Optimierungsfunktion und einer analytischen Funktion, sodass ein Gradientenabstiegsschritt approximiert wird. Beide Bestandteile werden durch ein System von nichtlinearen Gleichungen charakterisiert. Wir stellen eine Lösung für diese Gleichungen vor und leiten damit einen ableitungsfreien Optimierungsalgorithmus mit semi-globalen Konvergenzeigenschaften her.

Die theoretischen Ergebnisse werden mit numerischen Resultaten ergänzt. Wir präsentieren eine qualitative und quantitative numerische Studie des Algorithmus, in der wir geeignete Designparameter sowie Konvergenzgeschwindigkeit und Gradientenapproximationsfehler untersuchen. Zur Performanzsteigerung wird eine abnehmende Schrittweitensteuerung und eine Adaptierung der Explorationssequenz vorgestellt. Außerdem werden numerische Experimente hinsichtlich verschiedener Benchmark-Probleme und Anwendungen diskutiert. Wir validieren, dass die vorgestellten Algorithmen ein robustes Verhalten gegenüber verrauschten Funktionsauswertungen haben, nicht stetiger Funktionen handhaben und potentiell lokale Minima überwinden können.

1 Introduction

1.1 Motivation and Background

Powerful optimization algorithms are key ingredients in science and engineering applications. Over the last decades, advances in machine learning, big-data-driven decision making, and real-time control methods have been accelerated by sophisticated optimization algorithms and the increasing computing power of microprocessors. Optimization in such applications is often very challenging, e.g. they are high-dimensional, non-convex, non-smooth, or of stochastic nature. Thus, improving existing optimization algorithms and developing novel algorithms is of central importance to master those challenges and therewith enhance technologies.

The need to solve increasingly complex optimization problems has particularly enabled the development of so-called *derivative-free or zeroth order optimization algorithms*, i.e., methods where no derivative information of the objective function is required—only function evaluations. This is especially appealing when the value of the objective function is obtained by simulations or other black box oracles, or where the calculation of the objective's derivative is computationally too expensive and only function evaluations are affordable. Such scenarios arise constantly in almost every area of modern technology and research: identifying and constructing the next drug against a disease, planning and scheduling the traffic flow in big cities, calculating the flight trajectory of a space mission, or developing an human-like artificial intelligence application, to name only a few potential applications.

The vast number of technology-driven applications and the increasing need for efficient optimization algorithms is emerging along with a growing number of publications in the area of derivative-free optimization algorithms. Historically, one of the earliest implementations of derivative-free algorithms was carried out on the von Neumann architecture-based computer *MANIAC 1*—an approximated solution of a six-dimensional non-linear least-squares problem calculated by utilizing derivative-free coordinate search (cf. Fermi (1952)). In the same year the well-known derivative-free optimization algorithm based on a gradient-approximation scheme by Kiefer, Wolfowitz, et al. (1952) was presented. Continuously, several extensions and improvements in the class of gradient approximations with so-called sample-averaging were developed, e.g in the work of Spall (1992) or Kushner and Clark (2012). The algorithms presented in this work are also derivative-free optimization algorithms based on gradient approximation ideas but are closely related to so-called *ex-*

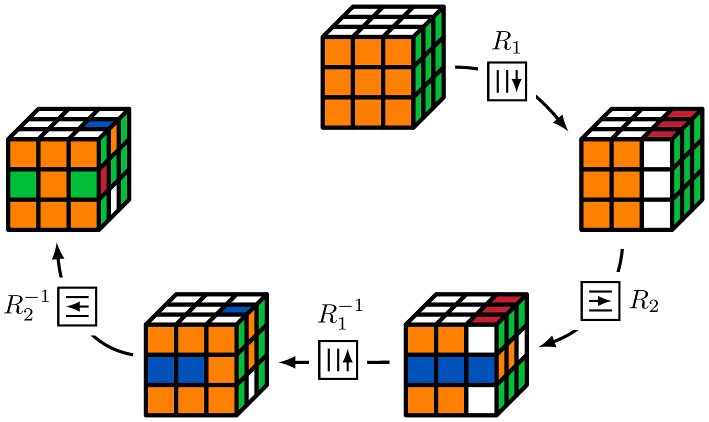

Figure 1.1. A figurative illustration of the mathematical concept of non-commutativity based on the magic cube. Rotations R_1 and R_2 and then their backward counterparts R_1^{-1} and R_2^{-1} are sequentially performed. Apparently, the initial configuration differs from the final configuration as a result of the non-commutativity of rotations R_1 and R_2.

tremum seeking control, a real-time (derivative-free) optimization method in control systems. Extremum seeking control can be traced back as far as 1922 to the work of Leblanc (1922), but only in the last two decades the field has regained new interest. Today, sophisticated tools for real-time optimization problems are available; see Krstić and Wang (2000), Teel and Popovic (2001), Guay and Zhang (2003), Tan, Moase, Manzie, Nešić, and Mareels (2010), and Benosman (2016), to mention just a few. One methodology in this field utilizes so-called *Lie brackets* from nonlinear geometric control (cf. Dürr, Stanković, Ebenbauer, and Johansson (2013) and Grushkovskaya, Zuyev, and Ebenbauer (2018)). Conceptually, this method is related to the approach we present in this thesis. However, as is quite common in control theory, extremum seeking schemes are stated as continuous time dynamical systems and not as discrete-time algorithmic optimization schemes.

In the present thesis, we develop a novel class of discrete-time, derivative-free optimization algorithms relying on gradient approximations based on *non-commutative maps*—inspired by the aforementioned Lie bracket approximation ideas in extremum seeking control systems. The introduced algorithm class has several interesting features. It shows robustness against noisy function evaluations, is able to deal with discontinuous objective functions, and potentially overcomes local minima. The main idea is to construct non-commutative maps with function evaluations to extract gradient information of the objective function. Conceptually speaking, two maps R_1, R_2 with a composition rule $\circ$ are commutative if their permuted compositions are identical (i.e., if $R_1 \circ R_2 = R_2 \circ R_1$). For example, multiplication of linear maps in the form of matrices is generally non-commutative, i.e., the result depends on the order in which they are multiplied. The way we utilize non-commutativity for optimization can be illustrated by the Magic Cube, known also as *Rubik's Cube* (Rubik (1975)), as depicted in Figure 1.1. Let R_i represent a rotation around an axis and R_i^{-1} its inverse counterpart; then it is obvious that the composition $R_1 \circ R_2 \circ R_1^{-1} \circ R_2^{-1}$

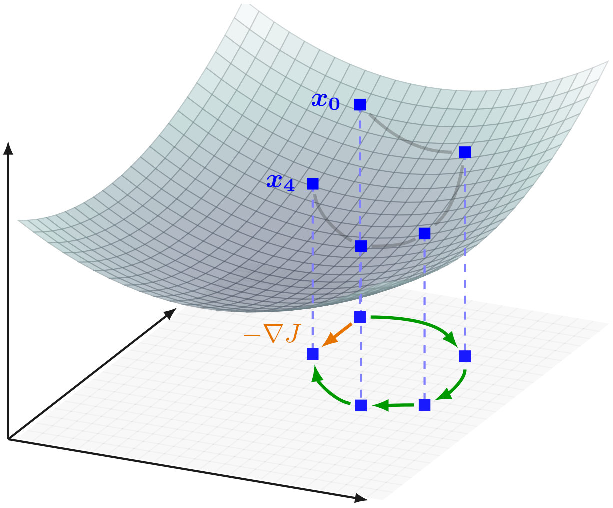

Figure 1.2. An illustration of the presented optimization algorithm class based on non-commutative maps. Effects of non-commutativity are utilized to approximate the negative gradient of the objective function and therewith the direction to a local minimum. The algorithm is initialized at x_0, and in four steps the algorithm performs an (approximated) negative gradient step to x_4.

does not commute as illustrated in Figure 1.1—the initial configuration is not equal to the final configuration. Figuratively speaking, the main idea and result of this work is the derivation of suitable mathematical definitions for maps R_i such that the difference between the initial and final configuration approximates the gradient of an objective function, as visualized in Figure 1.2. As a special case, two non-commutative maps and their inverse counterparts, similar to the Magic Cube illustration in Figure 1.1, are applied w.r.t. a point x_0 of an objective function. The gap between the first and last point, i.e., x_0 and x_4, is an approximation of the negative gradient at x_0 of the objective function (see Figure 1.1).

In a nutshell, this thesis is motivated by the idea to utilize the concept of non-commutative maps to approximate discrete-time gradient descent algorithms, i.e., designing a class of novel derivative-free optimization algorithms with convergence guarantees, various tuning parameters, and a scope of applications ranges from extremum seeking control problems to general (derivative-free) optimization problems.

1.2 Problem Statement

We consider unconstrained minimization problems of the form

$$\min_{x \in \mathbb{R}^n} J(x), \tag{1.1}$$

where only function evaluations of the objective $J : \mathbb{R}^n \to \mathbb{R}$ can be utilized to find a local minimum $x^* \in \mathbb{R}^n$ of J. The class of algorithms we propose is of the form

$$x_{k+1} = M_k^{\sqrt{h}}(x_k, J(x_k)), \quad k \geq 0 \tag{1.2}$$

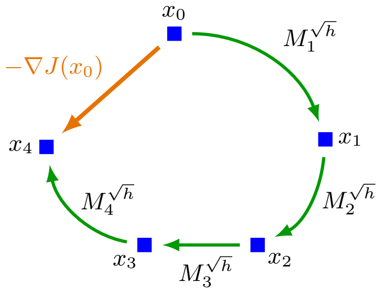

Figure 1.3. Principle of the algorithm class presented in this thesis. Composition of non-commutative maps as stated in (1.3) for $m = 4$ and $k = 0$ such that $x_4 - x_0 \approx -\nabla J(x_0)$, see (1.4).

where we call $M_k^{\sqrt{h}} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ the *transition map* and $h \in \mathbb{R}_{>0}$ the step size. The main idea is to design the transition maps in such a way that for every $k \in \mathbb{N}_0$, the m-fold composition of these maps

$$x_{k+m} = \left(M_{k+m-1}^{\sqrt{h}} \circ \cdots \circ M_k^{\sqrt{h}} \right) (x_k, J(x_k)) \tag{1.3}$$

approximates a gradient descent step

$$x_{k+m} = x_k - h\nabla J(x_k) + \mathcal{O}(h^{3/2}), \tag{1.4}$$

as illuminated in Figure 1.3 (b). Conceptually, the "non-commutativity gap" between x_k and x_{k+m} represents the negative gradient of the objective function with an approximation error of order $h^{3/2}$. Thus, an approximated gradient-descent optimization procedure, well-known for example from finite-difference approximations (cf. Kiefer et al. (1952)). The term $\mathcal{O}(h^{3/2})$ in (1.4) represents a function $R : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}^n$ such that for every compact convex set $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{J} \subseteq \mathbb{R}$ there exist an $L \in \mathbb{R}_{>0}$ and $\bar{h} \in \mathbb{R}_{>0}$ such that for all $x_k \in \mathcal{X}$, $J(x_k) \in \mathcal{J}$, and $h \in [0, \bar{h}]$, $\|R(x_k, J(x_k); h^{3/2})\|_2 \leq Lh^{3/2}$.

For the analysis of the algorithm (1.2) we impose the following assumptions.

Assumption 1. The objective function in the optimization problem (1.1) fulfills the following properties:

[A1] $J : \mathbb{R}^n \to \mathbb{R}$ is of class $C^2(\mathbb{R}^n, \mathbb{R})$, and transition maps $M_k^{\sqrt{h}} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ for all $k \geq 0$ are of class $\mathcal{C}^2(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$.

[A2] $J : \mathbb{R}^n \to \mathbb{R}$ is radially unbounded, and there exists an $x^* \in \mathbb{R}^n$ such that $\nabla J(x)^\top (x - x^*) > 0$ for all $x \in \mathbb{R}^n \backslash \{x^*\}$. ●

We note that **[A2]** will be required only for the analysis of the convergence properties. The implementation of the algorithms, however, is not limited to the class of objective functions satisfying Assumption 1.

1.3 Contributions and Outline

In this thesis, we provide a novel class of derivative-free optimization algorithms that is based on non-commutative maps. In particular, we present a new procedure to extract gradient information from an objective function by constructing compositions of non-commutative maps. Those are defined by function evaluations and applied in such a way that gradient descent steps are approximated and semi-global convergence guarantees are given. We supplement our theoretical findings with numerical results. Therein, we provide several algorithm parameter studies and tuning rules, as well as the results of applying our algorithm to challenging benchmarking problems. The outline of the presented thesis is as follows:

- In Chapter 2, we first give an overview of related work. We highlight five main classes of derivative-free optimization algorithms and present established gradient approximation methods. Further, various applications where derivative-free optimization show good performance are discussed. Section 2.2 is dedicated to the concept of extremum seeking control that inspired the algorithms developed in this thesis. Finally, we present our main idea in Section 2.3. We derive our novel derivative-free algorithm class by two ingredients, namely the so-called *exploration sequence* and *gradient-generating functions*. The results of this chapter are based on Feiling, Zeller, and Ebenbauer (2018) and Feiling et al. (2019).

- In Chapter 3, we study the theoretical problem of the presented algorithm class. First, a general algorithmic scheme is proposed and its gradient approximation behavior is analyzed (Theorem 1). The two main ingredients are the generalized periodic exploration sequence that indicates where the objective is to be evaluated and a set of gradient-generating functions, which are composed with the objective function in such a way that an approximation of a gradient descent step is obtained. Based on that, the problem in Section 1.2 is approached by solving i) a quadratic system of equations and ii) a set of functional equations. This is related to the following two problems: 1) construction of the exploration sequence (Theorem 4) and 2) various cases of generating function pairs (Theorem 5 and Theorem 6), respectively. Given the solutions of i) and ii) by 1) and 2), the semi-global practical asymptotic convergence (Theorem 2) and, by some extension, semi-global asymptotic convergence (Theorem 3) to a local minimum of the objective function is proven. Eventually, we discuss the algorithm's design parameters and functions. This chapter's results are based on Feiling, Belabbas, and Ebenbauer (2020).

- In Chapter 4, we study the numerical problem of the presented algorithm class. A qualitative numerical study of the design parameters and functions is provided, as well as a quantitative numerical study with convergence speed and gradient approximation error as performance evaluation metric. Based on that, several performance tuning rules are discussed, namely variable and adaptive step sizes and an adaption of the exploration sequence. Finally, numerical experiments are carried out. In particular, a benchmarking study on challenging derivative-free optimization problems is presented, and we discuss potential applications of our algorithm class.

- In Chapter 5, we summarize the results of this thesis and conclude with an outlook of potential research directions.

Some technical background, notation, and preliminary lemmas are summarized in Appendix A. All technical proofs are gathered in Appendix B. Additional information on numerical results is provided in Appendix C and a step-by-step construction of the exploration sequence is presented in Appendix D. A list of nomenclature in this work is presented on page 123.

2
Related Work and Main Idea

In this chapter we establish the main idea of the presented algorithm class, which paves the way for the upcoming chapter. Before we present our main idea in Section 2.3, we give a brief overview of research in derivative-free optimization, gradient approximation schemes, and various applications in these fields (Section 2.1); because of the vast literature available in these research fields, we do not aim for a complete overview but refer to several survey and overview articles in the dedicated sections below. Since the presented algorithm class and its gradient approximation scheme is inspired by concepts from nonlinear geometric control and the continuous-time control method extremum seeking, we provide an introduction of these concepts in Section 2.2.

2.1 Related Work and Literature Review

Derivative-free optimization. This class of optimization algorithms requires no derivative of the objective function to find local minima (or a global minimum)—only function evaluations. Over the last decade, derivative-free optimization (as well as its naming twins black-box optimization, gradient-free optimization, optimization without derivatives, simulation-based optimization, and zeroth-order optimization) has been an active research field, that regained new interest after the early outstanding work in the sixties to eighties, e.g. in Fletcher (1965); Karmanov (1974); Matyas (1965); Nelder and Mead (1965); Polyak (1987); Rastrigin (1963); Rosenbrock (1960); **?**. The acceleration in computational power in the last decades and the simplicity of applying derivative-free optimization methods were the main triggers for an increase in research publications in this field. Clearly, this class of algorithms is limited by accuracy, computational cost, or problem size, because of its strong correlation on the problem dimension and potential computationally expensive function evaluations. Nevertheless, the algorithms are known for their simple formulations and, thus, efficient implementation approaches. In this view, it is a class of algorithms that is very appealing for practical applications.

An overview of well-established, newly developed, and improved derivative-free optimization algorithms is presented in Conn, Scheinberg, and Vicente (2009), Audet and Hare (2017), and Rios and Sahinidis (2013) with a focus on software implementations for applications and industry problems. The algorithms can be clustered in five main categories:

- Direct search methods, first presented in Hooke and Jeeves (1961): only function evaluations are utilized, while no approximation of the gradient or the objective function is developed. Famous algorithms in this class are random and grid search (cf. Rastrigin (1963); J. Bergstra and Bengio (2012)), as well as the simplex optimization algorithm by Nelder and Mead (1965).

- Model-based methods, so-called surrogate or merit functions of the objective function serve as prediction models to calculate an update step of the algorithm by applying, for example, convex optimization principles and algorithms (cf. Boyd and Vandenberghe (2004)). A well-known method in this category is the trust region method (cf. Moré and Sorensen (1983); Conn, Gould, and Toint (2000)), of which various approaches to derive surrogate functions can be found in literature, e.g. polynomial models (cf. M. J. Powell (2003)), quadratic interpolation (cf. Winfield (1973)), or radial basis function interpolations (cf. Buhmann (2003)).

- Meta-heuristics, first mentioned in Fogel, Owens, and Walsh (1966): algorithms that mimic processes in natural selection, statistical mechanics, and population dynamics (cf. Holland et al. (1992)), for example, processes inspired by natural phenomena, such as grey wolf hunting behavior (cf. Mirjalili, Mirjalili, and Lewis (2014)) or the social behavior of humpback whales (cf. Mirjalili and Lewis (2016)). Famous and well-performing algorithms in this category are, for example, simulated annealing (cf. Kirkpatrick, Gelatt, and Vecchi (1983)), genetic algorithms (cf. Bonabeau, Dorigo, Marco, Theraulaz, and Théraulaz (1999)), or particle swarm optimization (cf. Eberhart and Kennedy (1995)).

- Bayesian optimization: this method is (often) based on Gaussian processes and sequential strategies motivated by statistical analysis (see, for example, Jones, Schonlau, and Welch (1998); Brochu, Cora, and De Freitas (2010); Shahriari, Swersky, Wang, Adams, and De Freitas (2015)). Conceptually, the objective function is treated as a random function with a prior distribution over the objective function beliefs. Those prior beliefs are updated via function evaluations to build a posterior distribution over the objective function beliefs. Based on that, the next search step is calculated via different versions of sampling criteria.

- (Stochastic) Approximation methods: derivatives, specifically first order information, i.e., gradients of the objective function, are approximated by so-called sample averaging of function evaluations. The first well-known algorithm in this class is the (scalar) method of Kiefer et al. (1952). Because this is the category of the presented optimization algorithm class of this thesis, we provide a more detailed discussion in the paragraph about gradient approximations below.

Note that many subcategories and various hybrid versions of algorithm classes in between the five categories, as stated above, exist. For a more detailed categorization we refer to Audet and Hare (2017) and Conn et al. (2009).

In our view, a further class of derivative-free optimization methods is extremum seeking control, which is derived as feedback controller for dynamical systems. Erroneously, this