Stochastic and deterministic algorithms for continuous black-box optimization

Wang, H.

CHAPTER 1

Introduction

Optimization problems are of fundamental importance in mathematics, statistics, Machine Learning and real-world applications (e.g., optimization of a production process). In most cases, we aim at searching for an element (called candidate solution) in some pre-determined domain (called search space) of objective functions, such that that this element "outperforms" the remaining elements according to an (partial) order structure defined in the image of objective functions. Some examples of optimization problems are: searching for the minimum of a given function, the optimal linear predictor/estimator (statistics), the optimal linear separation boundary for binary classification problems (Machine Learning) and the optimal control parameters of an industrial production line. Prior to the detailed discussions, we shall give a brief explanation on some important aspects of optimization problems.

Domain It is also referred as the *search space* in the unconstrained optimization problems. The most intuitive domain is the subset of Euclidean spaces $\mathbb{R}^d$ (d is used as the dimensionality of the domain in this thesis). Some other important ones are: Hilbert (or Banach) spaces of functions, and mixed spaces, e.g., $\mathbb{R}^{d_1} \times \{0,1\}^{d_2} \times \{\text{Mon}, \text{Tue}, \ldots\}^{d_3}$. In this thesis, the discussion is restricted to the subset of $\mathbb{R}^d$. In addition, the Euclidean metric (or the related Mahalanobis metric) is always assumed on $\mathbb{R}^d$. Under such assumptions, we shall adopt the convention "**continuous optimization**" here[1].

[1]The specification of the metric is mandatory here because $\mathbb{R}^d$ can become a discrete space if any metric that yields isolated points is equipped to $\mathbb{R}^d$.

Objective function Although there are various types of objective functions in practice, this thesis is limited to the *real-valued* functions. In addition, the well-known **black-box** assumption is set on the objective function, meaning that no additional analytical property (e.g., continuity, smoothness and differentiability) is assumed on the objective function and the only available information is the evaluation of points in its domain.

Algorithm There are many numerical algorithm for solving the optimization problem. From the perspective of randomness, those algorithms can be categorized into **deterministic** and **stochastic** optimization algorithms. The former usually refers the classical mathematical optimization techniques (e.g., the Newton's method). The latter is mainly developed for the black-box optimization problems, which relies heavily on the statistical properties of random variables. In this thesis, both categories of algorithms are studies and improved.

1.1 Stochastic Optimization

In this thesis, the discussion is restricted to the real-valued objective function of the form:

$$f : \mathrm{S} \subseteq \mathbb{R}^d \to \mathbb{R}^m, \tag{1.1}$$

where its domain S is assumed to be a subset of the d-dimensional Euclidean space and its image is $\mathbb{R}^m$. The problem of minimizing (or maximizing) f is referred as a *single objective* problem if $m = 1$. For $m > 1$, it is called a *multi-objective* optimization problem and it is typically denoted by the boldface symbol $\mathbf{f}$. Note that, in the multi-objective scenario, it is usually not possible to define a *total order* on $\mathbb{R}^m$. The result of the multi-objective optimization is typically the "best" *anti-chain* w.r.t. some partial order defined on $\mathbb{R}^m$ (see the next section). In practice, domain S could represent the so-called "feasible region" in $\mathbb{R}^d$, that is restricted by a set of constraint functions. The subject of this thesis, the stochastic optimization paradigm, targets the so-called black-box optimization problem.

Definition 1.1 (Black-Box Optimization)**.** *An objective function f as defined in Eq.* (1.1) *is called* ***black-box*** *iff no prior knowledge is available on f and the only accessible posterior information is the objective value $f(\mathbf{x})$ for every point $\mathbf{x}$ on its domain.*

Remark. With no prior knowledge on f, many mathematical/numerical optimization methods, e.g., gradient descent and Newton's method, render inapplicable because the common assumptions, e.g., analytical expressions, differentiability as well as continuity no longer hold on f. In optimization, the domain S of f is more commonly referred as **search space** or **decision space**. We shall use those two terms interchangeably in this thesis. In the context of evolutionary computation (Bäck, 1996), the so-called **fitness value** depends on $f(\mathbf{x})$.

Throughout this thesis, the objective function f is assumed to be *minimized*, without loss of generality. In the single objective case, the goal of global minimization is to solve

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in \mathrm{S}} f(\mathbf{x}),$$

where the notion of global minimum is defined as follows.

Definition 1.2 (Global minimum)**.** *In the single objective case, point* $\mathbf{x}^* \in \mathrm{S}$ *is a global minimum (or minimum for short) of* f *iff* $\forall \mathbf{x} \in \mathrm{S}, f(\mathbf{x}^*) \leq f(\mathbf{x})$.

Approaching a global minimum is generally a difficult task due to the so-called *multimodality* of the objective function. Practically, it is only possible to guarantee the convergence to the so-called local optima.

Definition 1.3 (Local minimum)**.** *In the single objective case, a point* $\mathbf{x} \in \mathrm{S}$ *is a local minimum of* f *if there exists a neighborhood* $N_{\mathbf{x}}$ *of* $\mathbf{x}$ *such that* $\forall \mathbf{x}' \in N_{\mathbf{x}}, f(\mathbf{x}) \leq f(\mathbf{x}')$.

Remark. As the search space S is a subset of the *metric space* $\mathbb{R}^d$, it is straightforward to use any metric on $\mathbb{R}^d$ to define the neighborhood. For example, when taking the Euclidean norm $\|\cdot\|$, the neighborhood can be defined as a subset of S that contains an open Euclidean ball around $\mathbf{x}$: $B_\varepsilon(\mathbf{x}) = \{\mathbf{x}' \in \mathrm{S} : \|\mathbf{x} - \mathbf{x}'\| < \varepsilon\}$ for some $\varepsilon > 0$.

For solving black-box optimization problems, a very common mechanism is to progressively refine a point $\mathbf{x}$ by evaluating other candidate points in the neighborhood of $\mathbf{x}$ and moving to the point that improves $f(\mathbf{x})$. This is called *local search*. This mechanism requires two design choices: the determination of the neighborhood and a selection method to pick points in the neighborhood. In numerical optimization, this is usually achieved by directly using gradient increments (steps) or Newton increments. However, none of those techniques is applicable under the black-box assumption. Alternatively, in stochastic optimization, random perturbations are used for the local search, where commonly a parametric distribution family centered at $\mathbf{x}$ (e.g., Gaussian) is taken to generate candidate points in the neighborhood

(typically S). More precisely, when discussing multivariate random variables in $\mathbb{R}^d$, it is common to assume the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a measurable space $(\mathbb{R}^d, \mathscr{B}^d)$, where $\mathscr{B}^d$ is the Borel algebra on $\mathbb{R}^d$. A $\mathbb{R}^d$-valued random variable (or random vector) $\mathbf{x} \in \mathbb{R}^d$ is a $\mathscr{F}$-measurable function, $\mathbf{x} : \Omega \to \mathbb{R}^d$. Then, the formal definition of stochastic optimization is given as follows.

Definition 1.4 (Stochastic Optimization)**.** *Taking the aforementioned probability settings, Stochastic Optimization is the procedure of applying one or many optimization algorithms on a black-box function f, yielding a process of $\mathbb{R}^d$-valued random variables: $\{\mathbf{x}_t : t \in \mathbb{N}_{>0}\}$, such that*

$$\forall \varepsilon > 0, \quad \lim_{t \to \infty} \Pr\left(D(\mathbf{x}_t - \tilde{\mathbf{x}}) > \varepsilon\right) = 0, \tag{1.2}$$

where D is a metric on $\mathbb{R}^d$ and $\tilde{\mathbf{x}}$ is a (local) minimum and the conditional density

$$p\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots, \mathbf{x}_1\right)$$

can be specified using the probability measure $\mathbb{P}$.

Remark. 1) The stochastic process $\{\mathbf{x}_t : t \in \mathbb{N}_{>0}\}$ could stand for the current best point or the best point found since the first iteration. 2) It is expressed explicitly that "applying one or many optimization algorithms" because in practice two or more stochastic optimizers can be combined for the task, e.g., in case of memetic algorithms (Moscato et al., 1989). 3) In the single objective case, the convergence criterion can be formulated equivalently:

$$\forall \varepsilon > 0, \quad \lim_{t \to \infty} \Pr\left(|f(\mathbf{x}_t) - f(\tilde{\mathbf{x}})| > \varepsilon\right) = 0.$$

4) If the (local) minimum $\tilde{\mathbf{x}}$ is forced to be the global minimum $\mathbf{x}^*$, then the stochastic optimization procedure is said to **converge globally**. Note that the convergence in probability (Eq. 1.2) is taken for the convergence criterion because stronger types of convergence (e.g., almost sure convergence) do not hold in some cases and thus it is generally safe to use a weaker convergence notion. 5) In the case where it is hard to verify the convergence criterion for some practically well-performing algorithms, the criterion is relaxed to the following:

$$\forall t \in \mathbb{N}_{>0}\ \exists n \in \mathbb{N}_{>0} \quad \text{s.t.} \quad \mathbb{E}\left\{f(\mathbf{x}_{t+n}) \mid f(\mathbf{x}_{t+n-1}), f(\mathbf{x}_{t+n-2}), \ldots, f(\mathbf{x}_1)\right\} \leq f(\mathbf{x}_t).$$

Or equivalently there exists a subprocess of $\{f(\mathbf{x}_t) : t \in \mathbb{N}_{>0}\}$, being a supermartingale[1]. 6) Markov property holds for some stochastic optimization algorithms, e.g., $(1+1)$-ES (Bäck, 1996), meaning that $p\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots, \mathbf{x}_1\right) = p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$.

[1]Loosely speaking, the discrete-time supermartingale indicates the situation where the conditional expectation on the whole history at each step is not bigger than the random variable at the last time step.

Many stochastic optimization algorithms has been proposed for single- (Kirkpatrick et al., 1983; Schwefel, 1993; Bäck, 1996) and multi-objective (Deb et al., 2000; Emmerich, 2005) black-box optimization problems. In the literature, some stochastic optimization algorithms are often referred to *metaheuristics* (Luke, 2009; Talbi, 2009), e.g., Particle Swarm Optimization (Kennedy and Eberhart, 1995). Those algorithms can be categorized according to different criteria:

- Local search/Global search: a well-known example of local search is the stochastic hill-climbing algorithm. Some example of global search are evolutionary algorithms (Bäck, 1996), Swarm Intelligence (Bonabeau et al., 1999) and Efficient Global Optimization (Jones et al., 1998).
- Single-point based/Population-based algorithms: if a stochastic optimizer employs only one point iteratively, it is called a single-point strategy, e.g., Simulated Annealing (Kirkpatrick et al., 1983). Otherwise, it is called a population-based algorithm, e.g., (multi-membered) evolutionary algorithms. It is worth mentioning that the so-called $(1+1)$-EAs that employ one parent and one offspring fall into the single-point category.
- Nature-inspired algorithms: examples are evolution strategies (Bäck and Schwefel, 1993; Bäck et al., 2013), genetic algorithms (Goldberg, 1989) and Swarm Intelligence (Bonabeau et al., 1999).

In this thesis, instead of focusing on some specific optimization algorithms, we illuminate and investigate several important aspects of this field, which underpin many optimization algorithms:

- *Stochastic variation* is the algorithmic component where the (local) random perturbation is generated to modify the current point. In evolutionary computation, this is typically called the *mutation operator* (Bäck and Schwefel, 1993). In $\mathbb{R}^d$, the most common method is to apply the simple random sampling method on the Gaussian distribution. Other stochastic variations include: differential vector in the Differential Evolution (Storn and Price, 1997) and polynomial mutation (Agrawal et al., 1995). In Chapter 2, we shall illustrate a drawback of the simple random sampling from a Gaussian distribution and propose an improved sampling method, whose effectiveness is validated when plugged into evolution strategies.
- *Surrogate modeling*: When the function evaluation is very *expensive* on f, e.g., due to the high time complexity, it is common to build models that are

less computational expensive on the evaluated points, in order to partially replace the actual function evaluation. The precision of the surrogate model is of vital importance when assisting the stochastic optimizer. However, this is usually a demanding requirement due to the lack of a sufficient number of data points, or irreducible modeling error when the objective function is noisy. In this scenario, it is helpful to quantify the uncertainty of the model prediction, e.g., by computing a confidence interval. We study a widely used surrogate model, Kriging/Gaussian process regression that it is naturally equipped with an uncertainty quantification.

- *Model utilization*: taking the model imprecision and uncertainty quantification into account, it is possible to determine the most trustworthy point, or alternatively which point possesses the highest potential to help the optimization procedure if the actual function evaluation were performed on it. Such decisions are made through an utility function of the surrogate model, called *infill criterion*. The infill criterion plays a vital role in many optimization paradigms, including the Surrogate-assisted Evolutionary Algorithms (Emmerich, 2005), Efficient Global Optimization (Močkus, 1975, 2012; Jones et al., 1998), Multi-armed Bandits (Auer et al., 2002) and Monte-Carlo Tree Search (Silver et al., 2016). In this thesis, we aim at summarizing the existing infill criterion and proposing a novel infill criteria that is theoretically better than the existing ones. Furthermore, the parallelization issue (Ginsbourger et al., 2010) of infill criteria is also considered in detail and several new parallelization methods are proposed and tested.

1.2 Multi-objective Optimization

In this section we introduce some definitions in the context of multi-objective problems. Due to the possibility of incomparable solutions, the notation of modality/local optimality is also modified and extended to the multi-objective scenario. The search space under consideration is $\mathrm{S} \subseteq \mathbb{R}^d$ and the objective space is $\mathbb{R}^m$. Most of our definitions can also be generalized to other spaces, however, due to space limitations, this will not be part of this section.

Now, let $\mathbf{f} : \mathrm{S} \to \mathbb{R}^m$ be a multi-objective function (which we want to "minimize") with component functions $f_i : \mathrm{S} \to \mathbb{R}$, $i = 1, \ldots, m$ and $\mathrm{S} \subseteq \mathbb{R}^d$. Given a totally

ordered set $(T, \leq)$, with total order $\leq$, the *Pareto order*[1] $\prec$ on T^k for any $k \in \mathbb{N}$ is defined as follows: Let $\mathbf{t}^{(1)} = (t_1^{(1)}, \ldots, t_k^{(1)}), \mathbf{t}^{(2)} = (t_1^{(2)}, \ldots, t_k^{(2)}) \in T^k$. We say $\mathbf{t}^{(1)} \prec \mathbf{t}^{(2)}$ if and only if (iff) $t_i^{(1)} \leq t_i^{(2)}$, $i = 1, \ldots, k$ and $\mathbf{t}^{(1)} \neq \mathbf{t}^{(2)}$. Instantiating $\leq$ to the natural total order on the real numbers, we obtain the Pareto order on $\mathbb{R}^m$. A point $\mathbf{x} \in \mathrm{S}$ is called *Pareto efficient* or *global efficient* or for short *efficient* iff there does not exist $\tilde{\mathbf{x}} \in \mathrm{S}$ such that $\mathbf{f}(\tilde{\mathbf{x}}) \prec \mathbf{f}(\mathbf{x})$. The set of all the (global) efficient points in S is denoted by $\mathcal{X}$ and is called the (Pareto) *efficient set* of $\mathbf{f}$. The image of $\mathcal{X}$ under $\mathbf{f}$ is called the *Pareto front* of $\mathbf{f}$, symbolically $P_{\mathcal{X}} = \mathbf{f}[\mathcal{X}] = \mathbf{f}(\mathbf{x}) : \mathbf{x} \in \mathcal{X}$.

Defining a locally efficient point in S (or of $\mathbf{f}$) is as straightforward as defining local minimizers (maximizers) for single-objective functions. This is in contrast to defining local efficient *sets*, which are needed for the multi-criteria setting.

Definition 1.5 (Locally Efficient Point)**.** *A point* $\mathbf{x} \in \mathrm{S}$ *is called* locally efficient point *of* $\mathbf{f}$ *if there is an open set* $U \subseteq \mathbb{R}^d$ *such that there is no point* $\mathbf{x}' \in U \cap \mathcal{X}$ *such that* $\mathbf{f}(\mathbf{x}') \prec \mathbf{f}(\mathbf{x})$*. The set of all the local efficient points in* S *is denoted by* $\mathcal{X}_L$*.*

Definition 1.6 (Globally Efficient Point)**.** *A point* $\mathbf{x} \in \mathrm{S}$ *is called* globally efficient point $\mathbf{f}$ *if there is no point* $\mathbf{x}' \in \mathbb{R}^d \cap \mathrm{S}$ *such that* $\mathbf{f}(\mathbf{x}') \prec \mathbf{f}(\mathbf{x})$*. The set of all the global efficient points in* S *is termed* efficient set *of* $\mathbf{f}$ *and denoted by* $\mathcal{X}$*.*

In order to extend the definition of the local optimality to multi-objective problems, it is necessary to first give a notation on the locality of (efficient) sets. It is defined using the so-called *connectedness*.

Definition 1.7 (Connectedness and Connected Component)**.** *Let* $A \subseteq \mathbb{R}^d$*. The subset* A *is called* connected *if and only if there do not exist two open subsets* U_1 *and* U_2 *of* $\mathbb{R}^d$ *such that* $A \subseteq U_1 \cup U_2$*,* $U_1 \cap A \neq \emptyset$*,* $U_2 \cap A \neq \emptyset$*, and* $U_1 \cap U_2 \cap A = \emptyset$*; or equivalently there do not exist two non-empty subsets* A_1 *and* A_2 *of* A *which are open in the relative topology of* A *such that* $A_1 \cup A_2 = A$ *and* $A_1 \cap A_2 = \emptyset$*. Let* B *be a non-empty subset of* $\mathbb{R}^d$*. A subset* C *of* B *is a* connected component *of* B *iff* C *is non-empty, connected, and there exists no strict superset of* C *that is connected.*

Definition 1.8 (Locally Efficient Set)**.** *A subset* $A \subseteq \mathrm{S}$ *is a* locally efficient set *of* $\mathbf{f}$ *if* A *is a connected component of* $\mathcal{X}_L$ *(= set of the locally efficient points in* S*).*

Definition 1.9 (Local Pareto Front)**.** *A subset* P *of the image of* $\mathbf{f}$ *is a* local Pareto front *of* $\mathbf{f}$*, if there exists a local efficient set* E *such that* $P = \mathbf{f}[E]$*.*

[1] It gives rise to a partial order vector space $(T^k, \prec)$.

Note that the (global) Pareto front of $\mathbf{f}$ is obtained by taking the image of the union of connected components of $\mathcal{X}$, under $\mathbf{f}$. If $\mathcal{X}$ is connected and $\mathbf{f}$ is continuous on $\mathcal{X}$, the Pareto front is also connected. In this thesis we use the notion of connectedness to define the locally efficient sets. There still remains the task of extending the notion of efficient set by looking at connectedness in the objective space. For instance it could happen that two different local efficient sets are mapped onto the same set in the objective space. This rises many questions, which need to be addressed in future work.

With a view towards algorithms that numerically approximates (locally) efficient sets and/or (local) Pareto fronts, it is necessary to generalize definition 1.8 to determine whether a *finite set* belongs to a connected component (i.e., a finite subset of $\mathcal{X}_L$ is a set of some locally efficient set). Here the issue is: a finite subset of Euclidean space is never a connected component (of some other subset) unless it consists of one point. To reconcile with definition 1.8, the notion of connectedness can be relaxed, using the ε neighborhood.

Definition 1.10 (ε-connectedness)**.** *Let $\varepsilon \in \mathbb{R}_{>0}$ and $\mathrm{S} \subseteq \mathbb{R}^d$. Set A is* ε-connected *if and only if for any distinct points $x, x' \in A$ there is a finite set of points $\{a_1, \ldots, a_k\} \subseteq A$ such that $D(x, a_1) \leq \varepsilon, D(a_1, a_2) \leq \varepsilon, \ldots, D(a_{k-1}, a_k) \leq \varepsilon, D(a_k, x') \leq \varepsilon$, where D is a metric in $\mathbb{R}^d$.*

A *finite set* $A \subseteq \mathrm{S}$ is locally efficient, if it consists of local efficient points in S and A is ε-connected: there exists $\varepsilon > 0$ on which definition 1.10 holds.

Definition 1.11 (finite ε-Local Efficient Set)**.** *Let A be a* finite *subset of $\mathcal{X}_L$. Then A is an* ε-local efficient set*, if $A \neq \emptyset$, and A is ε-connected.*

1.3 Matrix Calculus

In this thesis, the compact notation, called *Matrix Calculus* (Kollo and von Rosen, 2005) is extensively used for the derivations (e.g., Section 3.1.4, 5.2 and 5.3). For the readability of the technical part, we shall specify this notation and give some examples. Intuitively, the matrix differentiation is a collection of many partial derivatives, e.g., the gradient vector of a real-valued function $f : \mathbb{R}^d \to \mathbb{R}$:

$$\nabla f = \frac{\partial f}{\partial \mathbf{x}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d} \right)^\top .$$

The notation $\partial f/\partial \mathbf{x}$ is called a *scalar-by-vector* derivative. Throughout this thesis, the gradient is assumed to be a *column vector* and thus $\partial f/\partial \mathbf{x}$ has a column-wise layout. To avoid confusions, the *layout* of matrix derivatives like $\partial \mathbf{f}/\partial \mathbf{x}$ is determined according to that of $\mathbf{f}^\top$ or $\mathbf{x}$. This is called the *denominator layout convention*. Some common layouts are given as follows:

- *scalar-by-vector*

$$\frac{\partial f}{\partial \mathbf{x}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d}\right)^\top, \quad f : \mathbb{R}^d \to \mathbb{R},\ \mathbf{x} \in \mathbb{R}^d.$$

- *vector-by-scalar*

$$\frac{\partial \mathbf{f}}{\partial x} = \left(\frac{\partial f_1}{\partial x}, \frac{\partial f_2}{\partial x}, \ldots, \frac{\partial f_m}{\partial x}\right), \quad \mathbf{f} : \mathbb{R} \to \mathbb{R}^m,\ x \in \mathbb{R}.$$

- *vector-by-vector*

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_d} & \frac{\partial f_2}{\partial x_d} & \cdots & \frac{\partial f_m}{\partial x_d} \end{bmatrix}, \quad \mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m,\ \mathbf{x} \in \mathbb{R}^d.$$

The major benefit of using such notations is that the common rules for derivatives, e.g., chain rule, product rule and quotient rule still hold for the matrix notation. For example, consider the following functions: $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, g : \mathbb{R}^m \to \mathbb{R}$. The composition $g \circ \mathbf{f} : \mathbb{R}^d \to \mathbb{R}$ can be differentiated using the chain rule:

$$\frac{\partial (g \circ \mathbf{f})}{\partial \mathbf{x}} = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \frac{\partial g(\mathbf{f})}{\partial \mathbf{f}}$$

For example, when differentiating a quadratic form w.r.t. a vector, we have:

$$\frac{\partial \mathbf{u}^\top \mathbf{K} \mathbf{v}}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \mathbf{K} \mathbf{v} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}} \mathbf{K}^\top \mathbf{u},$$

where $\mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^n, \mathbf{K} \in \mathbb{R}^{m \times n}, \mathbf{x} \in \mathbb{R}^d$ and $\mathbf{K}$ is not a function of $\mathbf{x}$. One can easily verify that the shape of the LHS (left-hand-side) admits that of the RHS (right-hand-side), assuming the denominator layout.

1.4 Outline of the Dissertation

The outline of this thesis is as follows. The motivation, content and research questions of each chapter are briefly introduced, which is followed by a publication list on each chapter.

Chapter 2 discusses several sampling methods designed to reduce the sampling error from a multivariate Gaussian distribution. The proposed *mirrored orthogonal sampling* method is applied to Evolution Strategies. The convergence property of the resulting optimization algorithm is investigated both theoretically and empirically. In addition, the stochastic variation behind the Efficient Global Optimization algorithm is extracted and formulated as a stand-alone stochastic variation method.

- Wang, H., M. Emmerich, and T. Bäck (2014). Mirrored orthogonal sampling with pairwise selection in evolution strategies. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, New York, NY, USA, pp. 154–156. ACM.

- van Rijn, S., H. Wang, B. van Stein, and T. Bäck (2017). Algorithm configuration data mining for CMA evolution strategies. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '17, New York, NY, USA, pp. 737–744. ACM.

- Wang, H., M. Emmerich, and T. Bäck (2018). Mirrored Orthogonal Sampling for Covariance Matrix Adaptation Evolution Strategies. *Evolutionary computation* (27), to appear.

- Emmerich, M., O. M. Shir, and H. Wang (2018). *Evolution Strategies*, pp. 1–31. Cham: Springer International Publishing.

Chapter 3 aims at giving a precise and unified treatment of the commonly used surrogate modeling method, Kriging/Gaussian process regression. This estimation method is summarized and compared from many perspectives, including the theory on the best linear predictor, reproducing kernel Hilbert Space and Bayesian inference. In the second half of the chapter, a novel algorithmic framework called *Cluster Kriging* is proposed to relax the high time/space complexity of the original Kriging method, when applied to large data sets. Moreover, it is shown that Cluster Kriging can effectively support the efficient global optimization algorithm.

van Stein, B., H. Wang, W. Kowalczyk, T. Bäck, and M. Emmerich (2015). Optimally weighted cluster kriging for big data regression. In E. Fromont, T. De Bie, and M. van Leeuwen (Eds.), *Advances in Intelligent Data Analysis XIV: 14th International Symposium, IDA 2015, Saint Etienne. France, October 22 -24, 2015. Proceedings*, Cham, pp. 310–321. Springer International Publishing.

van Stein, B., H. Wang, W. Kowalczyk, M. Emmerich, and T. Bäck (2016). Fuzzy clustering for optimally weighted cluster kriging. In *Proceedings of the Conference on Evolutionary Computation*, CEC '16, pp. 154–163.

Wang, H., B. van Stein, M. Emmerich, and T. Bäck (2017b). Time complexity reduction in efficient global optimization using cluster kriging. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '17, New York, NY, USA, pp. 889–896. ACM.

van Stein, B., H. Wang, W. Kowalczyk, and T. Bäck.
A Novel Uncertainty Quantification Method for Efficient Global Optimization. In *Proceedings of 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, IPMU 2018.

Chapter 4 focuses on the issue on how to use the surrogate model properly. The utility of each location on a surrogate model is quantified by a well-defined function, called *Infill Criterion*. Various infill criteria are compared in this chapter, where the trade-offs between criterion are discovered. In addition, a novel infill criterion, *Moment-Generating Function of Improvement* (MGFI) is proposed as the extension of all improvement-based criteria. Lastly, we investigate the multi-point generalization to the existing infill criteria, allowing for the parallel evaluation of candidate solutions.

Wang, H., M. Emmerich, and T. Bäck (2016). Balancing risk and expected gain in kriging-based global optimization. In *Proceedings of the Conference on Evolutionary Computation*, CEC '16, pp. 154–163.

Emmerich, M., K. Yang, A. Deutz, H. Wang, and C. M. Fonseca (2016). *A Multicriteria Generalization of Bayesian Global Optimization*, pp. 229–242. Cham: Springer International Publishing.

Wang, H., B. van Stein, M. Emmerich, and T. Bäck (2017a, Oct). A New Acquisition Function for Bayesian Optimization based on the Moment-Generating Function. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 507–512.

Wang, H., T. Bäck, and M. T. M. Emmerich (2018). Multi-point efficient global optimization using niching evolution strategy. In A.-A. Tantar, E. Tantar, M. Emmerich, P. Legrand, L. Alboaie, and H. Luchian (Eds.), *EVOLVE - A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation VI*, Cham, pp. 146–162. Springer International Publishing.

Wang, H., M. Emmerich, and T. Bäck (2018). Cooling Strategies for the Moment-Generating Function in Bayesian Global Optimization. In *Proceedings of the Conference on Evolutionary Computation*, CEC '18, to appear.

Chapter 5 discusses numerical multi-objective optimization (MOO). The demand on this topic originates from many numerical multi-objective tasks that arise in the study of stochastic optimization, e.g., the multi-objective treatment of infill criteria in Chapter 4. The contribution in this chapter is three-fold: firstly, we mathematically analyze the so-called *Mixed-Peak* bi-objective test problem. Secondly, the gradient field and Hessian matrix of the hypervolume indicator are studied in depth. Thirdly, two novel numerical MOO algorithms, namely the hypervolume-based first- (gradient) and second-order (Hessian) methods are proposed and tested.

Kerschke, P., H. Wang, M. Preuss, C. Grimme, T. Heike, and E. Michael (2016). Towards analyzing multimodality of multiobjective landscapes. In *International Conference on Parallel Problem Solving from Nature*, pp. 206–215. Springer.

Wang, H., A. Deutz, T. Bäck, and M. Emmerich (2017). Hypervolume indicator gradient ascent multi-objective optimization. In *9th International Conference on Evolutionary Multi-Criterion Optimization - Volume 10173*, EMO 2017, New York, NY, USA, pp. 654–669. Springer-Verlag New York, Inc.

Wang, H., Y. Ren, A. Deutz, and M. Emmerich (2017). *On Steering Dominated Points in Hypervolume Indicator Gradient Ascent for Bi-Objective Optimization*, pp. 175–203. Cham: Springer International Publishing.

Kerschke, P., H. Wang, M. Preuss, C. Grimme, T. Heike, and E. Michael (2018). Search Dynamics on Multimodal Multi-Objective Problems. *Evolutionary computation* (30), to appear.

van der Blom, K., S. Boonstra, H. Wang, H. Hofmeyer, and M. Emmerich *Evaluating Memetic Building Spatial Design Optimisation Using Hypervolume Indicator Gradient Ascent*. Cham: Springer International Publishing, to appear.

In addition to this description of the chapters, some closely linked sections are shown in the dependence graph below.

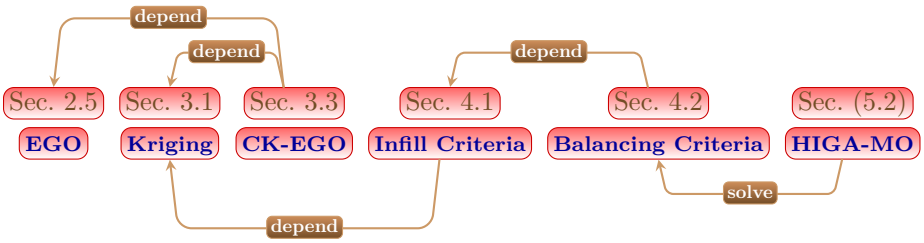

Figure 1.1: Dependences between several sections.