Preface

This book aims to cover major methodological and theoretical developments in the field of stochastic global optimization. This field includes global random search and methods based on probabilistic assumptions about the objective function.

We discuss the basic ideas lying behind the main algorithmic schemes, formulate the most essential algorithms and outline the ways of their theoretical investigation. We try to be mathematically precise and sound but at the same time we do not often delve deep into the mathematical detail, referring instead to the corresponding literature. We often do not consider the most general assumptions, preferring instead simplicity of arguments. For example, we only consider continuous finite dimensional optimization despite the fact that some of the methods can easily be modified for discrete or infinite-dimensional optimization problems.

The authors' interests and the availability of good surveys on particular topics have influenced the choice of material in the book. For example, there are excellent surveys on simulated annealing (both on theoretical and implementation aspects of this method) and evolutionary algorithms (including genetic algorithms). We thus devote much less attention to these topics than they merit, concentrating instead on the issues which are not that well documented in literature. We also spend more time discussing the most recent ideas which have been proposed in the last few years.

We hope that the text of the book is accessible to a wide circle of readers and will be appreciated by those interested in theoretical aspects of global optimization as well as practitioners interested mostly in the methodology. The target audience includes graduate students and researchers in operations research, probability, statistics, engineering (especially mechanical, chemical and financial engineering). All those interested in applications of global optimization can also benefit from the book.

The structure of the book is as follows. In Chapter 1, we discuss general concepts and ideas of global optimization in general stochastic global optimization in particular. In Chapter 2, we describe basic global random search

algorithms, study them from different view-points and discuss various probabilistic and statistical aspects associated with these algorithms. In Chapter 3, we discuss and study several more sophisticated global optimization techniques including random and semi-random coverings, random multistart, stratified sampling schemes, Markovian algorithms and finally the methods of generations. In Chapter 4, techniques based on the use of statistical models about the objective function are studied. The Introduction and Chapter 1 are written by both co-authors. Chapters 2 and 3 are written by A.Zhigljavsky, Chapter 4 is written by A.Žilinskas.

A.Zhigljavsky is grateful to his colleagues at Cardiff University (V.Savani, V.Reynish, E.Hamilton) who helped with typing and editing the manuscript and patiently tolerated his monologues on different aspects of global optimization. He is also grateful to his long-term friends and collaborators Luc Pronzato and Henry Wynn for stimulating discussions and to his former colleagues from St.Petersburg University – M.Chekmasov, V.Nevzorov, S.Ermakov, and especially to M.Kondratovich, V.Nekrutkin and A.Tikhomirov. Significant parts of Sects. 2.4, 2.5 and 3.3 are based on the joint work of A.Zhiglajvsky and M.Kondratovich; Sect. 3.4 is fully based on the results of V.Nekrutkin and A.Tikhomirov who very much helped with writing a summary of their results.

A.Žilinskas thanks the Institute of Mathematics and Informatics at Vilnuis for facilitating his work on the book, and J.Mockus for introducing him to the field of global optimization many years ago. The work by A.Žilinskas has been partly supported by the Lithuanian State Science and Studies Foundation. The material on one-dimensional algorithms included into Chapter 4 is based mainly on joint publications by A.Žilinskas and J.Calvin. Before starting work on the book, the authors invited Jim Calvin to become a co-author. Although he rejected our invitation in view of his involvement in other projects, we consider him a virtual co-author of the mentioned part of the book.

Both authors thank Rebecca Haycroft and Julius Žilinskas as well as the two referees for their careful reading of the manuscript and constructive remarks. Especially, the authors are very grateful to the editor of the series Panos Pardalos for his encouragement with this project.

Cardiff, Vilnuis

Anatoly Zhigljavsky
Antanas Žilinskas

2

Global Random Search: Fundamentals and Statistical Inference

2.1 Introduction to Global Random Search

In this section, we formulate and discuss basic assumptions concerning the feasible region and the objective function, introduce a general scheme of global random search algorithms and provide a general result establishing convergence of global random search algorithms.

2.1.1 Main Assumptions

Consider a general minimization problem

$$f(x) \to \min_{x \in A}$$

with objective function $f(\cdot)$ and feasible region A. We shall always assume that the minimum value $m = \min_{x \in A} f(x)$ is attained in A. In general, $f(\cdot)$ may have more than one minimizer x_*.

Let us formulate other common assumptions about A and $f(\cdot)$. Most of these assumptions will be assumed true throughout this chapter and the next.

Assumptions concerning the feasible region A:

C1: A is a bounded closed subset of $\mathbb{R}^d$ ($d \geq 1$);

C2: $\mathrm{vol}(A) > 0$, where '$\mathrm{vol}(\cdot)$' denotes the volume (or d-dimensional Lebesgue measure) of a set;

C3: A is a finite union of the sets defined by a finite number of the inequality-type constraints $g_i(x) \leq 0$, where the functions $g_i(\cdot)$ defining the constraints are continuously differentiable (however, we do not need to know the explicit forms of these functions);

C4: there exist constants $c > 0$ and $\varepsilon_0 > 0$ such that for at least one minimizer x_* and all ε, $0 < \varepsilon < \varepsilon_0$, we have $\mathrm{vol}(B(x_*, \varepsilon)) \geq c\varepsilon^d$; that is, at least a uniformly constant proportion of a ball in $\mathbb{R}^d$ with centre at x_* and small radius must intersect A;

C5: the structure of A is simple enough for distribution sampling algorithms on A and some of its subsets, to be of acceptable complexity.

These conditions are satisfied for an extremely wide class of practically interesting sets A. We do not require, in particular, for A to be a cube; moreover, neither convexity nor connectivity for A are generally required.

Conditions C1 and C2 are very simple and natural (note that Condition C1 implies $\mathrm{vol}(A) < \infty$ so that $0 < \mathrm{vol}(A) < \infty$). Conditions C3 and C4 are needed to avoid difficulties at the boundaries of A. Thus, Condition C3 prevents fractal boundaries and Condition C4 helps avoid the configurations where random search algorithms would almost certainly fail. An example of such a configuration is shown in Figure 2.1. For such a configuration, simple random search algorithms (which are not using local descents) will not be able to approach the minimizer x_* as the path to this point is 'too narrow'.

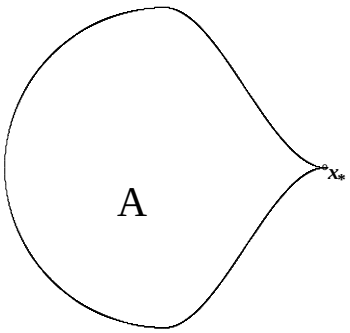

Fig. 2.1. An example of a disallowed combination of the set A and the minimizer x_*.

Of course, Condition C4 can be substituted with the following simpler but somewhat stronger condition on A:

C4′: there exist constants $c > 0$ and $\varepsilon_0 > 0$ such that $\mathrm{vol}(B(x, \varepsilon)) \geq c\varepsilon^d$ for all $x \in A$ and all ε, $0 < \varepsilon < \varepsilon_0$.

Assumptions concerning the objective function $f(\cdot)$:

C6: $f(\cdot)$ can be evaluated at any point of A without error (note, however, that we allow evaluation errors in Sects. 2.1.4, 3.5 and in some general discussions);

C7: the number of minimizers x_* is finite.

As a condition additional to C7, we shall sometimes assume that the minimizer x_* is unique.

Rather than demanding continuity of $f(x)$ for all $x \in A$, we shall demand the following two weaker conditions:

C8: function $f(\cdot)$ is bounded and piece-wise continuous on A;
C9: there exists $\delta_0 > 0$ such that for all $0 < \delta \leq \delta_0$ the sets

$$W(\delta) = \{x \in A\colon f(x) \leq m + \delta\}$$

are closed and $f(x)$ is continuous for all $x \in W(\delta_0)$.

Note that if the objective function $f(\cdot)$ is continuous for all $x \in A$ then Condition C9 holds for all δ_0 (and, in view of Condition C1, Condition C8 also holds).

The sets $W(\delta)$ and their behaviour as $\delta \to 0$ (an example)

The boundaries of the sets $W(\delta)$ are the level sets of $f(\cdot)$:

$$\partial W(\delta) = \{x \in A :\ f(x) = m + \delta\} = f^{-1}(m + \delta)\,.$$

These level sets can be easily visualized when $d \leq 2$; see e.g. Fig. 1.2 and Fig. 2.2, where the contour-plot of the function

$$g(x, y) = f_{3,3}(x) + f_{5,5}(y) + f_{3,3}(x) f_{5,5}(y), \quad (x, y) \in [0, 1] \times [0, 1]\,, \quad (2.1)$$

is provided; the function $f_{(k,l)}(\cdot)$ is defined below in (2.2).

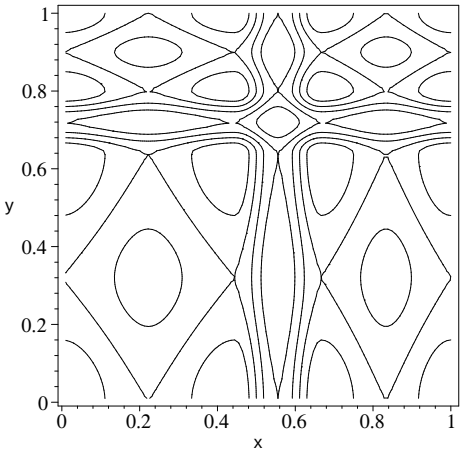

Fig. 2.2. Contour-plot of the function $g(x, y)$ defined in (2.1); the minimum value of this function is equal to 0 and is achieved at the global minimizer $(x_*, y_*) = (\frac{5}{9}, \frac{18}{25})$.

Let $A = [0, 1]$ and let $k \geq 2$ and $l \geq 2$ be some integers. The function $f_{(k,l)}(\cdot)$ is defined as

$$f_{(k,l)}(x) = \begin{cases} 1 - \frac{1}{2}\left(\sin\frac{lk\pi\,x}{(k-1)(l-1)}\right)^2 & \text{for } x \in \left[0, \frac{(k-1)(l-1)}{kl}\right] \\ 1 - \left(\sin\frac{lk\pi\,x}{l-1}\right)^2 & \text{for } x \in \left[\frac{(k-1)(l-1)}{kl}, \frac{l-1}{l}\right] \\ 1 - \frac{1}{2}\left(\sin l\pi\,x\right)^2 & \text{for } x \in \left[\frac{l-1}{l}, 1\right]. \end{cases} \tag{2.2}$$

As illustrations, the functions $f_{(12,5)}(x)$ and $f_{(3,5)}(x)$ are depicted in Figs. 1.1 and 3.2(A), respectively.

For all integers $k, l \geq 2$, the functions $f_{(k,l)}(x)$ are continuously differentiable in $A = [0,1]$ and have three local minima. These local minima are achieved at the points:

$$x_{(1)} = \frac{(k-1)(l-1)}{2k\,l}\,, \quad x_{(2)} = \frac{(2k-1)(l-1)}{2k\,l}\,, \quad \text{and } x_{(3)} = 1 - \frac{1}{2l}$$

with the point $x_* = x_{(2)}$ being the global minimizer. The values of the function $f_{(k,l)}(\cdot)$ at these points are

$$f_{(k,l)}(x_{(1)}) = f_{(k,l)}(x_{(3)}) = \frac{1}{2}\,, \quad f_{(k,l)}(x_{(2)}) = 0\,.$$

Despite the fact that the functions $f_{(k,l)}(x)$ are continuously differentiable, the problem of finding x_* is very difficult when k is large. Indeed, the complexity of the problem of global optimization is very much related to the rate of decrease of the ratio $\mathrm{vol}(W(\delta))/\mathrm{vol}(A)$ as δ decreases. Thus, for large k, the complexity of the function $f_{(k,l)}(x)$ defined in (2.2) is expressed in terms of the values of $\mathrm{vol}(W(\delta))$ which are small even for moderately large values of δ; for instance,

$$\mathrm{vol}(W(0.5)) = \frac{l-1}{2kl} < \frac{1}{2k}\,.$$

In fact, for $f(\cdot) = f_{(k,l)}(\cdot)$, we can easily compute $\mathrm{vol}(W(\delta))$ for all δ:

$$\mathrm{vol}(W(\delta)) = \begin{cases} 0 & \text{if } \delta \leq 0 \\ \frac{l-1}{kl}\left(1 - \frac{2}{\pi}\arcsin\sqrt{1-\delta}\right) & \text{if } 0 \leq \delta \leq \frac{1}{2}, \\ 1 - \frac{2}{\pi kl}\left((l-1)\arcsin\sqrt{1-\delta} + (kl-l+1)\arcsin\sqrt{2-2\delta}\right) & \text{if } \frac{1}{2} < \delta \leq 1, \\ 1 & \text{if } \delta \geq 1. \end{cases}$$

For general $f(\cdot)$, the rate of convergence of $\mathrm{vol}(W(\delta))/\mathrm{vol}(A)$ to zero as $\delta \to 0$ is studied in Sect. 2.5.3. In particular, it is shown in that section that if Conditions C7, C8 and C9 hold and if additionally for each global minimizer x_* the objective function $f(\cdot)$ is locally quadratic in the neighbourhood of x_*, then there exists a constant $c > 0$ such that

$$\mathrm{vol}(W(\delta)) = c\delta^{d/2}(1 + \mathrm{o}(1)) \;\text{ as }\; \delta \to 0\,. \tag{2.3}$$

This means that the rate of convergence of $\mathrm{vol}(W(\delta))$ to zero as $\delta \to 0$ is the same for a very broad class of objective functions. Of course, the complexity

of the function $f(\cdot)$ is also related to the value of the constant c in (2.3) and the range of values of δ, where the asymptotic relation (2.3) can be applied.

In a particular case of $f(\cdot) = f_{(k,l)}(\cdot)$, we have

$$\mathrm{vol}(W(\delta)) = \frac{2(l-1)\sqrt{\delta}}{kl\pi}\left(1 + \frac{\delta}{6} + O\left(\delta^2\right)\right), \quad \delta \to 0\,.$$

2.1.2 Formal Scheme of Global Random Search Algorithms

In a general global random search algorithm, a sequence of random points $x_1, x_2, \ldots, x_n$ is generated where for each j, $1 \le j \le n$, the point x_j has some probability distribution P_j. For each $j \ge 2$, the distribution P_j may depend on previous points $x_1, \ldots, x_{j-1}$ and the results of the objective function evaluations at these points (the function evaluations may not be noise-free). The number of points n, $1 \le n \le \infty$ (the stopping rule) can be either deterministic or random and may depend on the results of function evaluation at the points $x_1, \ldots, x_n$. For convenience, we shall refer to this general scheme as Algorithm 2.1.

Algorithm 2.1.

1. *Generate a random point x_1 according to a probability distribution P_1 on A; evaluate the objective function at x_1; set iteration number $j = 1$.*
2. *Using the points $x_1, \ldots, x_j$ and the results of the objective function evaluation at these points, check whether $j = n$; that is, check an appropriate stopping condition. If this condition holds, terminate the algorithm.*
3. *Alternatively, generate a random point x_{j+1} according to some probability distribution P_{j+1} and evaluate the objective function at x_{j+1}.*
4. *Substitute $j + 1$ for j and return to step 2.*

In the algorithm which is often called 'pure random search' all the distributions P_j are the same (that is, $P_j = P$ for all j) and the points x_j are independent. In Markovian algorithms the distribution P_{j+1} depends only on the previous point x_j and its function value $f(x_j)$. There is also a wide class of global random search algorithms where the distributions are not updated at each iteration but instead after a certain number of points have been generated. We can formally write down this scheme as follows.

Algorithm 2.2.

1. *Choose a probability distribution P_1 on the n_1–fold product set $A \times \ldots \times A$, where $n_1 \ge 1$ is a given integer. Set iteration number $j = 1$.*
2. *Obtain n_j points $x_1^{(j)}, \ldots, x_{n_j}^{(j)}$ in A by sampling from the distribution P_j. Evaluate the objective function $f(\cdot)$ at these points.*
3. *Check a stopping criterion.*

4. *Using the points* $x_{l(i)}^{(i)}$ $(l(i)=1,\ldots,n_i;\ i=1,\ldots,j)$ *and the objective function values at these points, construct a probability distribution* P_{j+1} *on the* n_{j+1}*–fold product set* $A\times\ldots\times A$*, where* n_{j+1} *is some integer that may depend on the search information.*
5. *Substitute* $j+1$ *for* j *and return to Step 2.*

Of course, if $n_j = 1$ (for all j) in Algorithm 2.2 then it becomes Algorithm 2.1. On the other hand, Algorithm 2.1 allows more freedom in defining the distributions of points where $f(\cdot)$ is evaluated and therefore can seem to be more general than Algorithm 2.2. Thus, the difference between Algorithms 2.1 and 2.2 is purely formal; sometimes one form is more convenient and in other cases the other form is more natural.

There are two important issues to deal with while constructing global random search algorithms (in either form, Algorithm 2.1 or 2.2):

(i) choosing the stopping rule n, and
(ii) choosing the way of constructing the distributions P_j.

Consider issue (i). Commonly, a fixed number of points is generated (that is, the total number of points n is fixed). A more sophisticated approach would be to estimate the closeness of the current record value of the objective function $f(\cdot)$ to its minimum value $m = \min f$. This can be done in different ways. Any of the deterministic approaches (based, for example, on the use of Lipschitz constant estimates) can be applied. An enormous advantage of many global random search algorithms is related to the fact that because of the randomness of the points where $f(\cdot)$ is evaluated, probabilistic and statistical considerations can be applied to infer about the closeness of the current record value of $f(\cdot)$ to the minimum m; many of these considerations can be used in defining the stopping rule. A large part of the present chapter is devoted to these probabilistic and statistical considerations.

Issue (ii) concerns the construction of the distributions P_j (here by 'construction' we do not mean 'giving an analytic formula' but rather 'formulating an algorithm for sampling from the distribution'). This is the issue of how we use prior information about $f(\cdot)$ and the information we obtain in the process of the search, as well as how we compromise between the globality and locality of our search. The former problem (of extracting and using information about f) is complex and versatile; significant parts of this chapter and the next deal with it. The latter problem is potentially simpler, it was briefly considered in Sect. 1.1.3.

2.1.3 Convergence of Global Random Search Algorithms

In the early stages of development of global random search theory (in the nineteen seventies and eighties), a number of papers were published establishing sufficient conditions for convergence (in probability and with probability one)

of random search algorithms; see, for example, [63, 188, 229]. The main idea in most of these, and in many other results on convergence of global random search algorithms, is the classical, in probability theory, 'zero-one law', see e.g. [226]. The following simple theorem stated and proved in [273], Sect. 3.2, illustrates this technique in a very general setup.

Let us consider a general global random search algorithm in the form of Algorithm 2.1, where the point x_j has some distribution P_j which may depend on previous points $x_1, \ldots, x_{j-1}$ and the results of the objective function evaluation at these points.

Theorem 2.1. *Let the objective function $f(\cdot)$ satisfy Condition* C7*, x_* be a global minimizer of $f(\cdot)$ and let $f(\cdot)$ be continuous in the vicinity of x_*. Assume that*

$$\sum_{j=1}^{\infty} q_j(\varepsilon) = \infty \tag{2.4}$$

for any $\varepsilon > 0$, where

$$q_j(\varepsilon) = \inf P_j(B(x_*, \varepsilon)) \,, \tag{2.5}$$

with $B(x_*, \varepsilon) = \{x \in A\colon ||x - x_*|| \le \varepsilon\}$; the infimum in (2.5) *is taken over all possible previous points and the results of the objective function evaluations at them. Then, for any $\delta > 0$, the sequence of points x_j with distributions P_j falls infinitely often into the set $W(\delta) = \{x \in A\colon f(x) - m \le \delta\}$, with probability one.*

Proof is given in Sect. 2.7; it is a simplified version of the proof given in [273].

Note that Theorem 2.1 holds in the general case where evaluations of the objective function $f(\cdot)$ can be noisy (and the noise is not necessarily random). If the function evaluations are noise-free, then the conditions of the theorem ensure that the sequence $\{x_n\}$ converges to the set $A_* = \{\arg\min f\}$ of global minimizers with probability one; similarly, the sequence of records $y_{on} = \min_{j \le n} f(x_j)$ converges to $m = \min f$ with probability one.

If for a particular sequence $\{x_j\}$ we have

$$\sum_{j=1}^{\infty} P_j(B(x_*, \varepsilon)) < \infty \,,$$

then the Borel-Cantelli lemma (see e.g. [226]) implies that the points $x_1, x_2, \ldots$ fall into $B(x_*, \varepsilon)$ only a finite number of times, with probability one. Moreover, looking at the family of functions (2.2), we conclude that (2.4) cannot be improved upon for a wide enough class $\mathcal{F}$ of objective functions. That is, if (2.4) is not satisfied then there exists $f \in \mathcal{F}$ such that for any n, none of the points $x_1, \ldots, x_n$ fall into $B(x_*, \varepsilon)$ with any fixed probability γ, $0 < \gamma < 1$.

Since the location of x_* is not known a priori, the following simple sufficient condition for (2.4) can be used:

$$\sum_{j=1}^{\infty} \inf P_j(B(x,\varepsilon)) = \infty \tag{2.6}$$

for all $x \in A$ and $\varepsilon > 0$.

In practice, a very popular rule for selecting probability measures P_j's is

$$P_{j+1} = \alpha_{j+1} P + (1 - \alpha_{j+1}) Q_j \,, \tag{2.7}$$

where $0 \le \alpha_{j+1} \le 1$, P is the uniform distribution on A (extension to other probability distributions P is straightforward) and Q_j is an arbitrary probability measure on A which may depend on the results of the evaluation of the objective function at the points $x_1, \ldots, x_j$. For example, sampling from Q_j may correspond to performing several iterations of a local descent from the current record point x_{oj}.

Sampling from the distribution (2.7) corresponds to taking a uniformly distributed random point in A with probability α_{j+1} and sampling from Q_j with probability $1 - \alpha_{j+1}$.

If the probability measures P_j in Algorithm 2.1 are chosen according to (2.7), then a simple and rather weak condition

$$\sum_{j=1}^{\infty} \alpha_j = \infty \tag{2.8}$$

is sufficient for (2.4) and (2.6) to hold.

The rate of convergence of the global random search algorithms, represented in the form of Algorithm 2.1 with distributions P_j chosen according to (2.7) and (2.8), is discussed at the end of Sect. 2.2.2.

2.1.4 Random Errors in Observations

Many global random search algorithms can easily be modified so that they can be used in the case where there are random errors in the observations of the objective function values. To give an example, several versions of the 'simulated annealing' algorithm considered in Sect. 3.3.2, have been devised for optimizing objective functions corrupted by noise, even before the simulated annealing algorithms became widely known. The corresponding algorithms are often called global (or multiextremal) stochastic approximation algorithms, see [260, 277] and [273], Sects. 3.3.3 and 3.3.4. The theoretical study of these algorithms is often related to the study of stochastic differential equations and in particular to the study of diffusion processes, see e.g. [142].

Providing the globality of search (see Sect. 1.1.3) is simple whether or not there are errors in observations. What is not that simple is recognizing the

neighbourhood of the global minimizer and making local steps (as it is difficult to estimate gradients of the objective function). However, many statistical and heuristic arguments can be employed for monitoring the arrival at the neighbourhood of the global minimizer (see e.g. Sect. 4.1 in [273]).

Rather than further developing this topic (which is not particularly challenging), we briefly consider a different problem related to the fact that there are random errors in observations of $f(\cdot)$. This is the problem of estimating the values of the objective function and its gradients in the case where the distribution of noise is known. We assume that the objective function is specified as the expectation

$$f(x) = E_x g(x, Y) = \int g(x, y)\phi_x(y)dy, \tag{2.9}$$

where $g(x, y)$ is a known function and $\phi_x(\cdot)$ is the density of the random variable Y; note that the random variable $Y = Y_x$ and the density $\phi_x(\cdot)$ may depend on x.

Assuming that the integral in (2.9) cannot be evaluated analytically, a natural way of approximating it is to use the following Monte Carlo estimator

$$f(x) \cong \frac{1}{n}\sum_{i=1}^{n} g(x, Y_x^{(i)}) \tag{2.10}$$

where $\{Y_x^{(1)}, \ldots, Y_x^{(n)}\}$ is a sample from a distribution with density $\phi_x(\cdot)$.

Assume that there exists a density $\pi(\cdot)$ such that $\phi_x(y) = 0$ whenever $\pi(y) = 0$ so that the ratio

$$w_x(y) = \phi_x(y)/\pi(y)$$

is well defined. Then, to estimate $f(x)$, we can use the method known as the *importance sampling:*

$$f(x) \cong \frac{1}{n}\sum_{i=1}^{n} g(x, Y^{(i)})w_x(Y^{(i)}) \tag{2.11}$$

where $\{Y^{(1)}, \ldots, Y^{(n)}\}$ is a sample from a distribution with density $\pi(\cdot)$. Note that this sample does not have to be independent or even random; it can be, for instance, a stratified sample, a MCMC sample or even a quasi-random sample, see Sects. 3.1, 3.2.1 and 3.3.2.

The main advantage of using (2.11) over (2.10) is the fact that we can use the same sample $\{Y^{(1)}, \ldots, Y^{(n)}\}$ for estimating values of $f(x)$ for all required values of the argument x. Moreover, using the same sample we can approximate the components of the gradient

$$\nabla f(x) = \left(\partial f(x)/\partial x_{(1)}, \ldots, \partial f(x)/\partial x_{(d)}\right), \quad x = (x_{(1)}, \ldots, x_{(d)}).$$

Indeed, the j-th derivative $\partial f(x)/\partial x_{(j)}$ can be written as

$$\frac{\partial f(x)}{\partial x_{(j)}} = \int \left(w_x(y) \frac{\partial g(x,y)}{\partial x_{(j)}} + g(x,y) \frac{\partial w_x(y)}{\partial x_{(j)}} \right) \pi(y) dy$$

and approximated by

$$\frac{\partial f(x)}{\partial x_{(j)}} \cong \frac{1}{n} \sum_{i=1}^{n} \left(w_x(Y^{(i)}) \frac{\partial g(x,Y^{(i)})}{\partial x_{(j)}} + g(x,Y^{(i)}) \frac{\partial w_x(Y^{(i)})}{\partial x_{(j)}} \right), \quad (2.12)$$

where $\{Y^{(1)}, \ldots, Y^{(n)}\}$ is the same sample as above. Similarly one can approximate higher-order derivatives of $f(\cdot)$.

The method based on (2.10) is often called the *many-samples method* as it requires a new sample $Y_x^{(1)}, \ldots, Y_x^{(n)}$ for every function evaluation. The method based on (2.11) and (2.12) is called the *single-sample method* as it only uses one sample to estimate all required function values and its derivatives. The single-sample method has numerous advantages over the more traditional many-samples method, see [92] for references and more discussion.

2.2 Pure Random and Pure Adaptive Search Algorithms

Pure random search (PRS for short) is the simplest global random search algorithm. It consists of taking a sample of n independent random points x_j $(j = 1, \ldots, n)$ in A and evaluating the objective function $f(\cdot)$ at these points. Studying this algorithm is relatively simple. However, knowing the properties of this algorithm is very important as PRS is a component of many other random search algorithms. Additionally, PRS is often a bench-mark for comparing properties of other global optimization algorithms (not necessarily random search ones).

In this section, we also consider a version of PRS which is called 'pure adaptive search' and generalize it to the 'pure adaptive search of order k'.

2.2.1 Pure Random Search and the Associated c.d.f.

The algorithm

PRS is an algorithm where n random points x_j $(j = 1, \ldots, n)$ are generated and the objective function $f(\cdot)$ at these points is evaluated. The points x_j are i.i.d.r.v. in A with common distribution P. Here n is a stopping rule which is not necessarily a fixed number (typically, however, it is a fixed number); P is some given probability measure on A, not necessarily uniform (although the case where P is uniform is the main special case).

The probability distribution P should be simple enough to sample from and must not be much different from the uniform measure on A (otherwise PRS may lose the property of being a global optimization algorithm). It is

often enough to assume that the distribution P is equivalent to the uniform distribution on A, see Condition C10 below.

Of course, PRS can be represented in the form of Algorithm 2.1, with $P_j = P$ for all $j = 1, \ldots, n$ and independent points $x_1, \ldots, x_n$.

The most common estimators (of course, they can only be used where the evaluations of $f(\cdot)$ are noise-free) of the minimum $m = \min f$ and the minimizer $x_* = \arg\min f$ are respectively the record values $y_{on} = \min_{1 \le j \le n} f(x_j)$ and the corresponding record points x_{on} which satisfy $f(x_{on}) = y_{on}$. We shall see below that the estimator y_{on} of m can often be significantly improved.

The c.d.f. of major importance

As a result of the application of PRS we obtain an independent sample $X_n = \{x_1, \ldots, x_n\}$ from a distribution P on A. Additionally, we obtain an independent sample $Y_n = \{y_1, \ldots, y_n\}$ of the objective function values at these points. The elements $y_j = f(x_j)$ of the sample Y_n are i.i.d.r.v. with the c.d.f.

$$F(t) = \Pr\{x \in A : f(x) \le t\} = \int_{f(x) \le t} P(dx)\,. \tag{2.13}$$

Fig. 2.3 displays the c.d.f. (2.13) for the case where the distribution P is unform on $A = [0, 1]$ and the objective function $f(x) = f_{(k,l)}(x)$ is as defined in (2.2) with $l = 5$ and $k = 2, 5$ and 20.

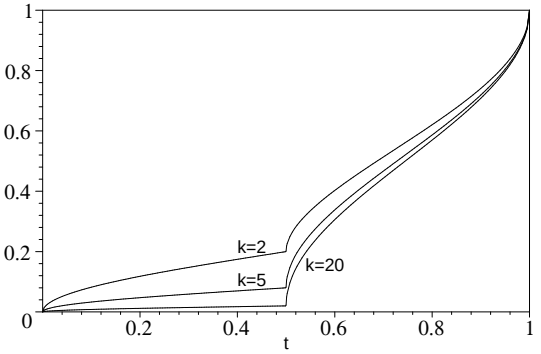

Fig. 2.3. Graphs of the c.d.f. (2.13) when P is uniform on $[0, 1]$ and $f(x) = f_{(k,l)}(x)$ is as defined in (2.2) with $l = 5$ and $k = 2, 5, 20$.

Assumption about the distribution P

Additional to the assumptions formulated in Sect. 2.1.1 we shall need an assumption about the probability distribution P. We shall assume that P is equivalent to the standard Lebesque measure on A; that is, we assume the following condition:

C10: the probability distribution P has a density $p(x)$ such that $c_{(1)} \le p(x) \le c_{(2)}$ for all $x \in A$ and some positive constants $c_{(1)}, c_{(2)}$.

Condition C10, along with condition C4 of Sect. 2.1.1, implies that for every minimizer x_* we have

$$P(B(x_*, \varepsilon)) \ge c\varepsilon^d \quad \text{for some } c > 0 \text{ and all } 0 < \varepsilon \le 1; \tag{2.14}$$

here, as usual, $B(x_*, \varepsilon) = \{z \in A : \ ||x_* - z|| \le \varepsilon\}$. As a consequence, we obtain, in particular, that the elements of the sample X belong to the vicinity of x_* with positive probability.

General properties of the c.d.f. (2.13)

The c.d.f. (2.13) is of major importance in studying PRS as well as some associated global random search algorithms. This is related to the fact that

$$F(t) = P(W(t-m)) \quad \text{for all } \ t \ge m\,, \tag{2.15}$$

where $W(\delta) = \{x \in A\colon f(x) \le m + \delta\}$, $\delta \ge 0$. Therefore, for $t \ge m$, $F(t)$ has the interpretation of the probability that a random point x_i distributed according to P falls into the set $W(t-m)$.

If the probability measure P is uniform on A, then the representation (2.15) can be written in the form

$$F(m+\delta) = \mathrm{vol}(W(\delta))/\mathrm{vol}(A)\,. \tag{2.16}$$

The importance of the ratio in the r.h.s. of (2.16) has already been discussed at the end of Sect. 2.1.1. Furthermore, as we shall see below, the behaviour of $F(m+\delta)$ for small $\delta > 0$ is a very important characteristic of the efficiency of PRS and, more generally, of the complexity of the objective function $f(\cdot)$.

Since the set $W(\delta)$ is empty for $\delta < 0$, we have $F(t) = 0$ for $t < m$. In view of Conditions C7 and C10 we also have $F(m) = 0$. On the other hand, the inequality (2.14) and Condition C9 imply that $F(t) > 0$ for all $t > m$. Moreover, Conditions C1–C4 and C7–C9 imply that the c.d.f. $F(t)$ is continuous at $t = m$.

Certain properties of the c.d.f. $F(\cdot)$, see Sect. 2.3, are different depending on whether this c.d.f. is continuous or not. In addition to Conditions C8 and C10, to guarantee the continuity of $F(\cdot)$ defined in (2.15) we have to assume that $\mathrm{vol}(f^{-1}(t)) = 0$ for all t, where

$$f^{-1}(t) = \{x \in A : f(x) = t\}\,.$$

That is, $F(\cdot)$ is continuous if the volume of every level set of $f(\cdot)$ is zero.

Let η denote a random variable with c.d.f. $F(\cdot)$. The fact that $F(m) = 0$ and $F(t) > 0$ for all $t > m$ is equivalent to the statement that the essential infimum of η is equal to m:

$$F(m) = 0 \text{ and } F(t) > 0 \text{ for all } t > m \iff \operatorname{ess\,inf} \eta = m\,.$$

Finally, let $M = \sup_{x \in A} f(x)$. Condition C8 implies that $M < \infty$. For the c.d.f. $F(\cdot)$, this means that $F(M) = 1$ and correspondingly, $F(t) = 1$ for all $t \geq M = \sup f$. The value of M is never important; however, it is sometimes important that the random variable η is concentrated on a bounded interval.

Poisson process representation

Let us follow [44] and give a representation of PRS through a Poisson process. Assume that vol(A)=1 and the distribution P is uniform on A. Let x_0 be an internal point of A (for instance, x_0 is one of the global minimizers of f). Define a sequence of point processes N_n on $\mathcal{B}$ by

$$N_n(B) = \sum_{j=1}^{n} \mathbf{1}_B(n^{1/d}(x_j - x_0))\,, \quad B \in \mathcal{B}\,,$$

where $\mathbf{1}_B(\cdot)$ is the indicator function

$$\mathbf{1}_U(z) = \begin{cases} 1 & \text{if } z \in U \\ 0 & \text{otherwise.} \end{cases}$$

That is, for a fixed measurable set B, $N_n(B)$ is defined as the number of points among $x_1, \ldots, x_n$ that belong to the set $x_0 + n^{-1/d}B$. The sequence of point processes N_n converges in distribution (as $n \to \infty$) to N, a Poisson point process with intensity 1 defined on A. For this process,

$$\Pr\{N(B) = k\} = \frac{[\mathrm{vol}(B)]^k}{k!} \exp(-\mathrm{vol}(B)), \quad k \geq 0, \quad B \in \mathcal{B};$$

additionally, for disjoint $B_1, \ldots, B_i \in \mathcal{B}$, the values $N(B_j)$ $(j = 1, \ldots, i)$ are independent random variables.

Therefore, a suitably normalized point process of observations near x_0 looks like a standard Poisson point process. This does not give us new results about the rate of convergence of PRS but permits us to look at the algorithm from a different prospective.

2.2.2 Rate of Convergence of Pure Random Search

Let us consider a PRS where x_j $(j = 1, \ldots, n)$ are i.i.d.r.v. distributed according to P and let the stopping rule n be a fixed number. In this section, our aim is to study the rate of convergence of PRS. We assume that all conditions of Sect. 2.1 concerning the feasible region A and the objective function $f(\cdot)$ are satisfied.

Rate of convergence to a neighbourhood of a global minimizer

Let $\varepsilon > 0$ be fixed, $x_* = \arg\min f$ be a global minimizer of $f(\cdot)$ and let our objective be hitting the set

$$B = B(x_*, \varepsilon, \rho) = \{x \in A\colon \rho(x, x_*) \le \varepsilon\}$$

with one or more of the points x_j $(j = 1, \dots, n)$. Let us regard the event 'a point x_j hits the set B' as success and the alternative event as a failure. Then PRS generates a sequence of independent Bernoulli trials with a success probability $P(B)$; Conditions C4 and C5 of Sect. 2.1 imply that $P(B) > 0$ for all $\varepsilon > 0$.

A sequence of independent Bernoulli trials is perhaps the most celebrated sequence in the probability theory. Below, we use some well-known results concerning this sequence to obtain results concerning the rate of convergence of PRS.

For fixed j, we have

$$\Pr\{x_j \in B\} = P(B)\,. \tag{2.17}$$

Therefore,

$$\Pr\{x_j \notin B\} = 1 - P(B), \qquad \text{for all } j\,.$$

In view of the independence of x_j,

$$\Pr\{x_1 \notin B, \dots, x_n \notin B\} = (1 - P(B))^n$$

and therefore

$$\Pr\{x_j \in B \text{ for at least one } j,\ 1 \le j \le n\} = 1 - (1 - P(B))^n\,. \tag{2.18}$$

Since $P(B) > 0$, this probability tends to one as $n \to \infty$.

Let τ_B be a random moment of first hitting the set B. Then the average number of PRS iterations required for reaching B is

$$E\tau_B = \frac{1}{P(B)}\,.$$

Typically, $P(B)$ is very small even if ε is not small (see below) and the rate of convergence of the probability (2.18) to one is very slow. Additionally, if $P(B)$ is small then $E\tau_B$ is large.

Taking $n \approx 1/P(B)$ is not enough to guarantee that B is reached with high probability. Indeed, for small $x > 0$ we have

$$(1 - x)^{\frac{1}{x}} \cong e^{-1} \cong 0.36788$$

and therefore for $n = \lceil 1/P(B) \rceil$

$$1-(1-P(B))^n \cong 0.63212 \quad \text{as } P(B) \to 0\,.$$

To achieve a probability of 0.95 for the r.h.s. of (2.18) we need to almost triple this value:

$$1-(1-P(B))^n \cong 1-\frac{1}{e^3} \cong 0.950213 \text{ for } n = \lceil 3/P(B) \rceil \text{ as } P(B) \to 0\,.$$

Furthermore, let us assume that we are required to reach the set B with probability at least $1-\gamma$ for some $0 < \gamma < 1$. This gives us the following inequality for n:

$$1-(1-P(B))^n \geq 1-\gamma\,.$$

Solving it we obtain

$$n \geq n(\gamma) = \frac{\ln \gamma}{\ln (1-P(B))}\,. \tag{2.19}$$

Since we assume that $P(B)$ is small, $\ln(1-P(B)) \cong -P(B)$, and we can replace (2.19) with

$$n \geq -\frac{\ln \gamma}{P(B)}\,; \tag{2.20}$$

that is, we need to make at least $\lceil -\ln \gamma / P(B) \rceil$ evaluations in PRS to reach the set B with probability $1-\gamma$.

Note that

$$\Pr\{x_j \in B(x_*, \varepsilon, \rho) \text{ for at least one } j, 1 \leq j \leq n\} = \Pr\left\{\min_{1\leq j\leq n} \rho(x_j, x_*) \leq \varepsilon\right\}$$

and therefore the discussion above can be considered as a discussion about the rate of convergence in probability of the sequence

$$\min_{1\leq j\leq n} \rho(x_j, x_*)$$

to zero, as $n \to \infty$.

Rate of convergence with respect to function values

If we want to study the rate of convergence with respect to the function values, that is, of

$$y_{on} - m = \min_{1\leq j\leq n} |f(x_j) - m| \quad \text{as } n \to \infty\,,$$

then in the above study we have to replace the set $B = B(x_*, \varepsilon, \rho)$, with the set

$$W(\delta) = \{x \in A\colon f(x) - m \le \delta\}$$

with some $\delta > 0$. In particular, we have $E\tau_{W(\delta)} = 1/P(W(\delta))$,

$$\Pr\{y_{on} - m \le \delta\} = 1 - (1 - P(W(\delta)))^n \to 1 \ \text{ as } \ n \to \infty,$$

and in order to reach the set $W(\delta)$ with probability $1-\gamma$, we need to perform approximately $-\ln\gamma/P(W(\delta))$ iterations of PRS.

In view of (2.16), these formulae can be expressed in terms of the c.d.f. $F(\cdot)$. We have, in particular,

$$\Pr\{y_{on} - m \le \delta\} = 1 - (1 - F(m+\delta))^n\,, \ \ E\tau_{W(\delta)} = 1/F(m+\delta)\,,$$

and to reach $W(\delta)$ with probability $1-\gamma$, we need to perform approximately $-\ln\gamma/F(m+\delta)$ iterations of PRS.

Particular case of the uniform distribution

Consider an important particular case, where the distribution P is uniform on A and ρ is the Euclidean metric (that is, $\rho = \rho_2$). Then for every $Z \in \mathcal{B}$ (this means that Z is any measurable subset of A) we have

$$P(Z) = \mathrm{vol}(Z)/\mathrm{vol}(A)$$

and for $B = B(x_*, \varepsilon)$ we have

$$P(B) = \frac{\mathrm{vol}(B)}{\mathrm{vol}(A)} \le \frac{\pi^{\frac{d}{2}}\varepsilon^d}{\Gamma\left(\frac{d}{2}+1\right)\cdot\mathrm{vol}(A)}\,, \tag{2.21}$$

where $\Gamma(\cdot)$ is the Gamma-function. If x_* is an interior point of A and ε is small enough so that the ball $\left\{x \in \mathbb{R}^d\colon \rho_2(x, x_*) \le \varepsilon\right\}$ is fully inside A, then the inequality in (2.21) becomes an equality.

The formulae (2.19) and (2.20) then say that if we want to reach the set

$$B = B(x_*, \varepsilon) = \{x \in A\colon \rho_2(x, x_*) \le \varepsilon\}$$

with probability at least $1-\gamma$, then we need to perform at least

$$n_* = \left\lceil \frac{\ln\gamma}{\ln\left(1 - \pi^{\frac{d}{2}}\varepsilon^d/(\Gamma(\frac{d}{2}+1)\cdot\mathrm{vol}(A))\right)} \right\rceil \simeq \left\lceil -\ln\gamma\cdot\frac{\Gamma\left(\frac{d}{2}+1\right)}{\pi^{\frac{d}{2}}\varepsilon^d}\cdot\mathrm{vol}(A) \right\rceil \tag{2.22}$$

iterations of PRS. Table 2.1 and Fig. 2.4 illustrate the dependence of $n_* = n_*(\gamma, \varepsilon, d)$ on γ, ε and d.

Note that in the majority of cases considered in Table 2.1, the approximation for n_* given in the r.h.s. of (2.22) over-estimates the true value of n_* by 1. Taking into account the fact that the values of n_* are typically very large, we can conclude that the r.h.s. of (2.22) gives a very good approximation for n_*.

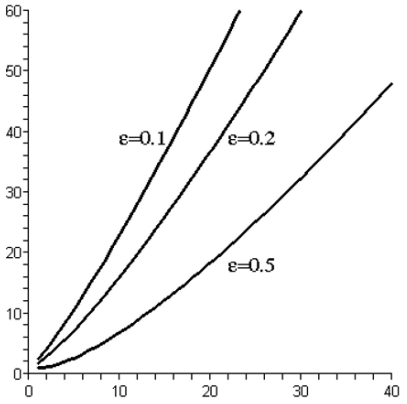

Fig. 2.4. Values of $\ln n_*$ as a function of d; here $\mathrm{vol}(A) = 1$, $\gamma = 0.1$, $\varepsilon = 0.5, 0.2, 0.1$ and n_* is as defined in (2.22).

d	$\gamma = 0.1$			$\gamma = 0.05$		
	$\varepsilon = 0.5$	$\varepsilon = 0.2$	$\varepsilon = 0.1$	$\varepsilon = 0.5$	$\varepsilon = 0.2$	$\varepsilon = 0.1$
1	0	5	11	0	6	14
2	2	18	73	2	23	94
3	4	68	549	5	88	714
4	7	291	4665	9	378	6070
5	13	1366	43743	17	1788	56911
7	62	38073	$4.9{\cdot}10^{6}$	80	49534	$6.3{\cdot}10^{6}$
10	924	$8.8{\cdot}10^{6}$	$9.0{\cdot}10^{9}$	1202	$1.1{\cdot}10^{7}$	$1.2{\cdot}10^{10}$
20	$9.4{\cdot}10^{7}$	$8.5{\cdot}10^{15}$	$8.9{\cdot}10^{21}$	$1.2{\cdot}10^{8}$	$1.1{\cdot}10^{16}$	$1.2{\cdot}10^{22}$
50	$1.5{\cdot}10^{28}$	$1.2{\cdot}10^{48}$	$1.3{\cdot}10^{63}$	$1.9{\cdot}10^{28}$	$1.5{\cdot}10^{48}$	$1.7{\cdot}10^{63}$
100	$1.2{\cdot}10^{70}$	$7.7{\cdot}10^{109}$	$9.7{\cdot}10^{139}$	$1.6{\cdot}10^{70}$	$1.0{\cdot}10^{110}$	$1.3{\cdot}10^{140}$

Table 2.1. Values of $n_* = n_*(\gamma, \varepsilon, d)$, see (2.22), for $\mathrm{vol}(A) = 1$, $\gamma = 0.1$ and 0.05, $\varepsilon = 0.5, 0.2$ and 0.1, for various d.

We can see that the dependence of n_* on γ is not crucial; on the other hand, $n_* = n_*(\gamma, \varepsilon, d)$ increases exponentially as the dimension d increases. As a matter of fact, the Stirling approximation gives for fixed $0 < \gamma < 1$ and $\varepsilon > 0$:

$$\ln n_*(\gamma, \varepsilon, d) = \frac{d+1}{2} \ln(d) - d \ln(\sqrt{2\pi e}\varepsilon) + \ln\left[\sqrt{\pi}\mathrm{vol}(A)\,(-\ln\gamma)\right] + \mathrm{O}\left(\frac{1}{d}\right)$$

as $d \to \infty$, and this approximation is extremely good even for small d.

If one is interested in the asymptotic behaviour of the value of $n_*(\gamma, \varepsilon, d)$ when d is fixed and the required precision ε tends to 0, then (2.22) implies

$$n_* = O\left(\frac{1}{\varepsilon^d}\right) \quad \text{as } \varepsilon \to 0\,. \tag{2.23}$$

Since we are not specifying the constant in (2.23), this formula holds not only for the case where $n_* = n_*(\gamma, \varepsilon, d)$ with $\rho = \rho_2$ and $P = P_0$ is the uniform distribution, but also for arbitrary $\rho = \rho_p$ $(1 \le p \le \infty)$, for any probability measure P equivalent to the uniform measure P_0 on A (see Condition C10 above). The same formula is true in the case when n_* has the meaning of the average number of iterations required to reach the ball $B(x^*, \varepsilon, \rho)$. If the objective function satisfies Conditions C7–C9 of Sect. 2.1, then we can replace the ball $B(x^*, \varepsilon, \rho)$ with the set $W(\varepsilon)$; the formula (2.23) will still hold.

Multivariate spacings

Let us mention a relevant result of S.Janson [127] on multivariate spacings. In general, the maximum (multivariate) spacing of a set of points $x_1, \ldots, x_n$ with respect to a convex set $B \subset \mathbb{R}^d$ is defined as the largest possible subset $x + rB$ of A which does not contain any of the points $x_j (j = 1, \ldots, n)$. Let $\mathrm{vol}(A) = 1$, B be either a cube or a Euclidean ball in $\mathbb{R}^d$ of unit volume: $\mathrm{vol}(B) = 1$; let also $x_1, \ldots, x_n$ be i.i.d.r.v. with the uniform distribution on A. Set

$$\triangle_n = \sup\{t\colon \text{there exists } x \in \mathbb{R}^d \text{ such that } x + tB \subset A \setminus \{x_1, \ldots, x_n\}\} \tag{2.24}$$

and define the volume of the maximum spacing as $V_n = (\triangle_n)^d$, which is the volume of the largest ball (or cube of fixed orientation) that is contained in A and avoids all n points $x_1, \ldots, x_n$. Then we have

$$\lim_{n\to\infty} \frac{nV_n - \ln n}{\ln \ln n} = d - 1 \quad \text{with probability 1} \tag{2.25}$$

(this result generalizes the result of P. Deheuvels [62]; see also [72]). Moreover, the sequence of random variables

$$nV_n - \ln n - (d-1)\ln \ln n + \beta_d$$

converges (as $n \to \infty$) in distribution to the r.v. with c.d.f. $\exp(-e^{-u})$, $u > 0$, where $\beta_d = 0$ if A is a cube and

$$\beta_d = \ln \Gamma(d+1) - (d-1)\left[\frac{1}{2}\ln \pi + \ln \Gamma\left(\frac{d}{2}+1\right) - \ln \Gamma\left(\frac{d+1}{2}\right)\right] \tag{2.26}$$

in the case when A is a ball. Since $\beta_d \ge 0$, the spherical spacings are a little bit smaller than the cubical ones. For large d, we can use the approximation

$$\beta_d = \frac{d}{2}\ln\frac{2d}{\pi} - d + \ln(\pi d) - \frac{1}{4} + \mathrm{O}\left(\frac{1}{d}\right), \quad d \to \infty\,,$$

for the quantity β_d defined in (2.26). This approximation is very accurate, especially if d is not very small (say, $d \ge 5$).

Extension to general global random search algorithms

The main results on the rate of convergence of PRS can be extended to a much wider class of global random search algorithms.

Consider the general Algorithm 2.1 of Sect. 2.1.2 and assume that the probability measures P_j are chosen according to (2.7), where the probability measure P satisfies Condition C10 of Sect. 2.2.1. Assume also that the measures Q_j are arbitrary and the condition (2.8) guaranteeing the convergence of the algorithm is met. Let us generalize the arguments that led us to the estimates of the convergence rate of PRS into this more general situation.

As a replacement of (2.17), for all $j \geq 1$ we have

$$\Pr\{x_j \in B\} \geq \alpha_j P(B)\,. \tag{2.27}$$

Arguments similar to those used in deriving (2.18) imply the inequality

$$\Pr\{x_j \in B \text{ for at least one } j,\ 1 \leq j \leq n\} \geq 1 - \prod_{j=1}^{n} (1 - \alpha_j P(B))\,. \tag{2.28}$$

In view of the condition (2.8) and the fact that $P(B) > 0$, the r.h.s. of (2.28) tends to one as $n \to \infty$.

Assume now that $P(B)$ is small and define $n(\gamma)$ as the smallest integer such that the following inequality is satisfied:

$$\sum_{j=1}^{n(\gamma)} \alpha_j \geq -\frac{\ln \gamma}{P(B)}\,.$$

Similarly to (2.20) we deduce that one has to perform at least $n(\gamma)$ iterations of Algorithm 2.1 to guarantee that at least one of the points x_j reaches the set B with probability $\geq 1 - \gamma$. Of course, $n(\gamma)$ is smallest if all $\alpha_j = 1$, that is when Algorithm 2.1 is PRS.

Extension of the main results concerning the rate of convergence with respect to function values and specialization to the case where P is the uniform distribution on A can be similarly made.

Slow rate of convergence may imply that the convergence is not practically achievable

Paying much attention to local search reduces the values of α_j's in (2.27). It may be tempting to perform many local searches leaving α_j's very small (for example, by setting $\alpha_j = 1/j$), just to guarantee the global convergence of the algorithm. Let us check what happens with the rate of convergence in the case when $\alpha_j = 1/j$. Since for large n we have $\sum_{j=1}^{n} 1/j \simeq \ln(n)$, from (2.20) we obtain

$$n(\gamma) \simeq \exp\{(-\ln \gamma)/P(B)\}\,.$$

Assuming that vol(A)=1 and that the distribution P is the uniform this, roughly speaking, implies that to compute the number of required iterations we need to exponentiate the numbers presented in Table 2.1. For instance, for very reasonable parameters $\varepsilon = 0.1$ and $d = 5$, we would need about 10^{19000} iterations of Algorithm 2.1 to guarantee that at least one of the points x_j will reach the ball $B(x_*, \varepsilon)$ with probability ≥ 0.9 (note that the total number of atoms in the universe is estimated to be smaller than 10^{81}).

The discussion above is very similar to the discussion provided by G.H.Hardy in Appendix III of his book [113]. Its consequence is that the fact of convergence of some global optimization algorithms is only a theoretical fiddle and does not mean anything in practice.

2.2.3 Pure Adaptive Search and Related Methods

In recent years there has been a great deal of activity (see e.g. papers [182, 265, 268, 269], the monograph [267] by Z.Zabinsky and references therein) related to the so-called 'pure adaptive search'. Unlike PRS, where the points x_j are independent and distributed in A with the same distribution P, at iteration $j+1$ of the pure adaptive search one chooses a random point x_{j+1} within the set

$$S_j = \{x \in A \colon f(x) < y_{oj}\}\,, \tag{2.29}$$

where $y_{oj} = \min\{f(x_1), \ldots, f(x_j)\}$ is the current record (in fact, every new point in the pure adaptive search is a new record point so that $x_j = x_{oj}$ and $y_j = y_{oj}$ for all $j \geq 1$). More precisely, x_1 has the probability distribution P and for each $j \geq 1$, x_{j+1} is a random point with the distribution P_{j+1} defined for all Borel sets $U \subset A$ by

$$P_{j+1}(U) = \frac{P(U \cap S_j)}{P(S_j)}\,, \tag{2.30}$$

where P is the original distribution and S_j is defined in (2.29). If P is the uniform distribution on A, then P_{j+1} is the uniform distribution on S_j. Of course, the points $x_1, x_2, \ldots$ generated in the pure adaptive search are dependent (unlike in PRS), see Sect. 2.3.3 for details.

If we replace the strict inequality $<$ in the definition of the sets S_j with $\leq$, then these sets are exactly the sets

$$W_{x_j} = \{x \in A : f(x) \leq f(x_j)\}\,;$$

that is, $\overline{S}_j = W_{x_j}$, where $\overline{Z}$ denotes the closure of a set Z. The corresponding method (using the sets W_{x_j} in place of S_j) is called 'weak pure adaptive search'.

The study of the sequence of function values $f(x_1), f(x_2), \ldots$ in the pure adaptive search is equivalent to the study of the record values in PRS. This

is the subject of Sect. 2.3.3; that section provides, therefore, a detailed investigation of the properties of the pure adaptive search. Note that in previous literature on the pure adaptive search this kind of investigation was lacking.

Of course, the sequence $f(x_j)$ converges to m much faster for the pure adaptive search than for PRS. The major obstacle preventing the application of the pure adaptive search to the practice is the fact that it is very hard to find points in the sets (2.29). To some extent, the problem of finding points in the sets (2.29) is one of the major objectives of all global optimization strategies. In particular, the set covering methods of Sect. 3.1 can help in removing the subregions of A that have no intersection with the sets (2.29) and thus simplify the problem of generating random points in these sets.

There are several papers fully devoted to the problem of generating random points in the sets (2.29), see e.g. [28, 194, 269] and Chapt. 5 in [267]; the corresponding methods either resemble or are fully based on the celebrated Markov Chain Monte Carlo methods. However, the problem is too difficult and cannot be resolved adequately. In general, there is no algorithmically effective way of generating (independent) random points from the sets (2.29) apart from using PRS in the first place (perhaps, with the bounding of certain subsets of A) and waiting for a new record value of the objective function (which is equivalent to obtaining a new point x_j in the pure adaptive search). If this is the way of performing the pure adaptive search then:

(a) the average waiting time of a new record is infinite for all $j > 1$, see (2.63);
(b) by discarding the k-th record values in PRS ($k > 1$) we lose an enormous amount of information contained in the evaluations made during PRS.

Taking these points into account we can state that generally, despite the fast convergence, the pure adaptive search only has theoretical interest as it is either impractical or much less efficient that PRS.

A similar conclusion can be drawn about different modifications of the pure adaptive search. These modifications include:

(i) weak pure adaptive search defined above;
(ii) 'hesitant random search' (see e.g. [28, 30]), where for all $j > 1$ the next point x_{j+1} is random and has distribution P_{j+1} with probability α_{j+1} and any other distribution on A with probability $1-\alpha_{j+1}$; here α_{j+1} ($0 \le \alpha_{j+1} \le 1$) may depend on $f(x_j)$ and the probability measures P_{j+1} are as defined in (2.30);
(iii) 'backtracking adaptive search', see [29, 265], where for all $j > 1$ the new point x_{j+1} is sampled from the sets:

$$\begin{cases} \{x \in A\colon f(x) < y_{oj}\} = S_j & \text{with probability } \alpha_{j+1} \\ \{x \in A\colon f(x) = y_{oj}\} & \text{with probability } \beta_{j+1} \\ \{x \in A\colon f(x) > y_{oj}\} & \text{with probability } 1-\alpha_{j+1}-\beta_{j+1} \end{cases}$$

for some α_{j+1} and β_{j+1} which may depend on $f(x_j)$.

2.2.4 Pure Adaptive Search of Order k

Aiming to resolve the problems (a) and (b) of the pure adaptive search, we can suggest the following extension of this algorithm (similar extensions can be suggested for its modifications (i)-(iii) above) which improves both pure random search and pure adaptive search.

Algorithm 2.3 (Pure adaptive search of order k).

1. *Choose points $x_1, \dots, x_k$ by independent random sampling from the uniform distribution on A. Compute the objective function values $y_i = f(x_i)$ $(i = 1, \dots, k)$. Set iteration number $j = k$.*
2. *For given $j \geq k$, we have points $x_1, \dots, x_j$ in A and values of the objective function at these points. Let $y_j^{(k)}$ be the k-th record value corresponding to the sample $\{y_i = f(x_i),\ i = 1, \dots, j\}$. Define the set*

$$S_j^{(k)} = \{x \in A\colon f(x) < y_j^{(k)}\}\,. \tag{2.31}$$

3. *Choose x_{j+1} as a uniform random point from the set $S_j^{(k)}$ and evaluate the objective function value at x_{j+1}.*
4. *Substitute $j + 1$ for j and return to step 2.*

For simplicity, Algorithm 2.3 is formulated under the assumption that the underlying distribution of points in A is uniform. It can be easily generalized to the case of a general distribution: in this case, $x_1, \dots, x_k$ are distributed in A according to a distribution P and x_{j+1} has the distribution $P_{j+1}^{(k)}$ defined for all Borel sets $U \subset A$ by $P_{j+1}^{(k)}(U) = P(U \cap S_j^{(k)})/P(S_j^{(k)})$; this formula is an extension of (2.30). Thus, the pure adaptive search of order 1 is just the pure adaptive search of Sect. 2.2.3.

Similarly to the case $k = 1$, we can define 'weak pure adaptive search of order k' by replacing the strict inequality $<$ in the definition of the sets $S_j^{(k)}$ with $\leq$. Analogously, we can define 'hesitant adaptive search of order k', 'backtracking adaptive search of order k' and other versions of the pure adaptive search.

Let, at iteration $j \geq k$ of Algorithm 2.3, $Y_j^{(k)} = \{y_j^{(1)}, \dots, y_j^{(k)}\}$ be the set of k record values and $X_j^{(k)} = \{x_1^{(j)}, \dots, x_j^{(k)}\}$ be the corresponding set of k record points. We have $y_j^{(1)} \leq \dots \leq y_j^{(k)}$ and these values are the k smallest values of the objective function computed so far. At iteration $j + 1$, the value $y_j^{(k)}$ is never in the set $Y_{j+1}^{(k)}$ (as $f(x_{j+1}) < y_j^{(k)}$) but the value $y_j^{(k-1)}$ always belongs to the new set of records: $y_j^{(k-1)} \in Y_{j+1}^{(k)}$ (the value $y_j^{(k-1)}$ can be either $y_{j+1}^{(k)}$ or $y_{j+1}^{(k-1)}$). Similarly, $x_j^{(k)} \notin X_{j+1}^{(k)}$ and $x_j^{(k-1)} \in X_{j+1}^{(k)}$. Thus, the pure adaptive search of order k (that is, Algorithm 2.3) is probabilistically equivalent to performing PRS and keeping k records and record points (rather

than just one record value and one record point in the original pure adaptive search).

The two main advantages of choosing $k > 1$ over $k = 1$ are:

(a) the set (2.31) is bigger than (2.29) and it is therefore easier to find random points belonging to the set (2.31). In particular, if at an iteration $j > k$ of Algorithm 2.3 we perform random sampling from A and wait for a point to arrive in the set (2.31), then the average waiting time is infinite when $k = 1$ and finite when $k > 1$, see (2.63) and (2.64), respectively;
(b) the set of records $Y_k^{(j)}$ contains much greater information about m than the set $Y_1^{(j)}$ consisting of the single record y_{oj} (see Sect. 2.4 on how to use this information).

Note also that if at each iteration of Algorithm 2.3, in order to obtain random points in the set (2.31) we sample points from A at random, then we can use the theory of k-th records (see Sect. 2.3) to devise the stopping rules (as this theory predicts the number of independent random points we need to obtain to improve the set of records $Y_k^{(j)}$). Fig.nnnn illustrates typical sequential updating of the set of records $Y_k^{(j)}$ obtained by performing random sampling of points from A. In this figure, the trajectories of three records $y_j^{(k)}$ $(k = 1, 2, 3)$ are plotted as we sequentially sample random points from A (the sample size n increases from 50 to 10000).

2.3 Order Statistics and Record Values: Probabilistic Aspects

Let $F(\cdot)$ be some c.d.f. and η be a random variable on $\mathbb{R}$ with this c.d.f. Our main particular case will be the c.d.f. (2.13) but the results of this section can be applied to many other c.d.f. as well. In this section, we shall not use the specific form of the c.d.f. (2.13) but we shall use the following two properties of this c.d.f.:

(i) the c.d.f. $F(\cdot)$ and the corresponding r.v. η have finite lower bound $m = \operatorname{ess\,inf} \eta > -\infty$ so that $F(t) = 0$ for $t < m$ and $F(t) > 0$ for $t > m$,
(ii) the c.d.f. $F(\cdot)$ is continuous at some vicinity of m.

We shall sometimes use stronger assumptions:

(i′) the c.d.f. $F(\cdot)$ has bounded support $[m, M]$ with $-\infty < m < M < \infty$ implying, additionally to (i), $F(t) = 1$ for $t \le M$,
(ii′) the c.d.f. $F(\cdot)$ is continuous.

In this section, always bearing in mind applications to the theory and methodology of global random search, we formulate and discuss numerous results of the theory of extreme order statistics and the associated theory of records.

2.3.1 Order Statistics: Non-Asymptotic Properties

Below, we collect several useful facts from the non-asymptotic theory of extreme value statistics. For more information about the theory we refer to the classical book by H.A. David [57] and to its extension [58].

Exact distributions and moments

Let $\eta_1, \eta_2, \ldots$ be i.i.d.r.v. with common c.d.f. $F(\cdot)$. If we rearrange the first n random variables $\eta_1, \ldots \eta_n$ so that $\eta_{1,n} \le \eta_{2,n} \le \cdots \le \eta_{n,n}$, then the resulting variables are called order statistics corresponding to $\eta_1, \ldots \eta_n$. Two extreme order statistics are $\eta_{1,n}$ and $\eta_{n,n}$, the minimum and maximum order statistics respectively. Their c.d.f.'s are:

$$F_{1,n}(t) = \Pr\{\eta_{1,n} \le t\} = 1 - (1 - F(t))^n \tag{2.32}$$

and

$$F_{n,n}(t) = \Pr\{\eta_{n,n} \le t\} = (F(t))^n \ .$$

The c.d.f. of $\eta_{k,n}$ with $1 \le k \le n$ can also be easily computed:

$$F_{k,n}(t) = \Pr\{\eta_{k,n} \le t\} = \sum_{m=k}^{n} \binom{n}{m} (F(t))^m (1 - F(t))^{n-m}$$

$$= \int_0^{F(t)} \frac{n!}{(k-1)!(n-k)!} u^{k-1}(1-u)^{n-k} du, \qquad -\infty < t < \infty\,. \tag{2.33}$$

The joint c.d.f of $\eta_{i,n}$ and $\eta_{j,n}$ $(1 \le i < j \le n)$ is given by $(-\infty < u < v < \infty)$:

$$\Pr\{\eta_{i,n} \le u, \eta_{j,n} \le v\} =$$

$$\sum_{s=j}^{n} \sum_{r=i}^{n} \frac{n!}{r!(s-r)!(n-s)!} (F(u))^r (F(v) - F(u))^{s-r} (1 - F(v))^{n-s}\,. \tag{2.34}$$

If η has density $p(t) = F'(t)$, then (2.34) implies the following expression for the joint density of $\eta_{i,n}$ and $\eta_{j,n}$ $(1 \le i < j \le n)$:
$p_{(i,j)}(u, v) =$

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!} (F(u))^{j-1} (F(v) - F(u))^{j-i-1} (1 - F(v))^{n-j} p(u)p(v),$$

where $u \le v$. The joint distributions of several order statistics can also be written down, if needed.

The expression for the β-th moment of $\eta_{k,n}$ easily follows from (2.33):

$$E X_{k,n}^{\beta} = \int_{-\infty}^{\infty} t^{\beta} dF_{k,n}(t)$$

$$= \frac{n!}{(k-1)!(n-k)!} \int_{-\infty}^{\infty} t^{\beta} (F(t))^{k-1} (1-F(t))^{n-k} dF(t)\,.$$

We shall also need the following expression for the joint moment $E\eta_{i,n}\eta_{j,n}$ with $1 \le i < j \le n$:

$$E\eta_{i,n}\eta_{j,n} = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \times$$
$$\int_{-\infty}^{\infty} \int_{x}^{\infty} xy (F(x))^{k-1} (F(y)-F(x))^{j-i-1} (1-F(y))^{n-j} dF(x) dF(y)\,;$$

this expression is a direct consequence of (2.34).

Two useful representations

The following representation for the order statistics has proven to be extremely useful:

$$\eta_{k,n} \stackrel{d}{=} F^{-1}\left(\exp\left\{-\left(\frac{\nu_1}{n} + \frac{\nu_2}{n-1} + \cdots + \frac{\nu_k}{n-k+1}\right)\right\}\right), \qquad (2.35)$$

where $\nu_1, \nu_2, \ldots, \nu_k$ are i.i.d.r.v. with exponential density e^{-t}, $t \ge 0$; the formula (2.35) is called the Rényi representation, and was derived in [196] (The inverse function $F^{-1}(s)$ is defined here as $F^{-1}(s) = \inf\{\, t : F(t) \ge s \,\}$ and the equality $\stackrel{d}{=}$ means that the distributions of the random variables (vectors) in the l.h.s. and r.h.s. of the equation are the same.)

When studying the joint distributions of order statistics, the following representation is often used:

$$(\eta_{1,n}, \ldots, \eta_{n,n}) \stackrel{d}{=} \left(F^{-1}(U_{1,n}), \ldots, F^{-1}(U_{n,n})\right) \qquad (2.36)$$

where $U_{1,n} \le \cdots \le U_{n,n}$ are the order statistics corresponding to the n i.i.d.r.v. with the uniform distribution on $[0,1]$.

Order statistics as a Markov chain

Of course, the order statistics $\eta_{k,n}$ are dependent random variables (we assume that n is fixed and k varies). One of their important properties is that if the original i.i.d.r.v. $\eta_1, \eta_2, \ldots$ have a continuous distribution (that is, η_j's have a common density $F'(t)$), then the order statistics $\eta_{k,n}$ form a Markov chain. The (forwards and backwards) transition probabilities of the Markov chain are:

$$\Pr\{\eta_{k,n} \le t \mid \eta_{k+1,n} = v\} = \left(\frac{F(t)}{F(v)}\right)^k, \qquad t \le v; \qquad (2.37)$$

$$\Pr\{\eta_{k+1,n} \le t \mid \eta_{k,n} = v\} = 1 - \left(\frac{1-F(t)}{1-F(v)}\right)^{n-k}, \qquad t \ge v. \qquad (2.38)$$

If we make the substitution $\eta \to -\eta$, then (2.38) will become (2.37) and vice versa.

Using the representations (2.35) and (2.36) we can express $\eta_{k+1,n}$ through $\eta_{k,n}$ as follows:

$$\eta_{k+1,n} \stackrel{d}{=} F^{-1}\left(\exp\left\{\ln F\left(\eta_{k,n}\right) - \frac{\nu_{k+1}}{n-k}\right\}\right); \tag{2.39}$$

here we assume that the c.d.f. $F(\cdot)$ is continuous, $1 \le k < n$ and ν_{k+1} is as in (2.35). Since ν_{k+1} is independent of $\eta_{k,n}$, the representation (2.39) also implies the fact that the sequence of order statistics $\{\eta_{k,n}\}$ forms a Markov chain.

If the original distribution is discrete with at least three support points, then the order statistics do not form a Markov chain.

2.3.2 Extreme Order Statistics: Asymptotic Properties

In this section, we collect classical facts from the asymptotic theory of extreme value statistics. These facts will play the key role in deriving statistical inference procedures in Sects. 2.4 and 2.5.

More information about the asymptotic theory of extreme value statistics and its numerous applications can be found in [11, 13, 68, 86, 105] and in many other books. No proofs of the classical results are given below; these proofs can easily be found in literature.

Let $\eta_1, \eta_2, \ldots$ be i.i.d.r.v. with common c.d.f. $F(\cdot)$ and $\eta_{1,n} \le \cdots \le \eta_{n,n}$ be the order statistics corresponding to the first n random variables $\eta_1, \ldots, \eta_n$. We are interested in the limiting behaviour, as $n \to \infty$, of the minimal order statistic $\eta_{1,n}$. Also, for fixed k and $n \to \infty$, we shall look at the asymptotic distributions of the k-th smallest order statistics $\eta_{k,n}$. As we are only interested in applying the theory to global random search problems, we always assume the properties (i) and (ii) stated in the beginning of Sect. 2.3 and sometimes we additionally assume one of the stronger properties (i′) or (ii′).

Note that the classical theory of extremes is usually formulated in terms of the maximum order statistics but we formulate all statements for the minimal order statistics.

Asymptotic distribution of the minimum order statistic

Consider first the asymptotic distribution of the sequence of minimum order statistics $\eta_{1,n}$, as $n \to \infty$. In the case $m = \operatorname{ess\,inf} > -\infty$ (where η has c.d.f. $F(t)$), there are two possible limiting distributions. However, in global random search applications, where $F(\cdot)$ has the form (2.13), only one asymptotic distribution arises; specifically, the Weibull distribution with the c.d.f.

$$\Psi_\alpha(z) = \begin{cases} 0 & \text{for } z < 0 \\ 1 - \exp\{-z^\alpha\} & \text{for } z \ge 0\,. \end{cases} \tag{2.40}$$

This c.d.f. only has one parameter, α, which is called the 'tail index'. The mean of the Weibull distribution with tail index α is $\Gamma(1+1/\alpha)$; the density corresponding to the c.d.f. (2.40) is

$$\psi_\alpha(t) = (\Psi_\alpha(t))' = \alpha\, t^{\alpha-1} \exp\{-t^\alpha\}\ ,\quad t > 0\,. \tag{2.41}$$

Figure 2.5 displays the density $\psi_\alpha(t)$ for $\alpha = 2, 3$ and 8.

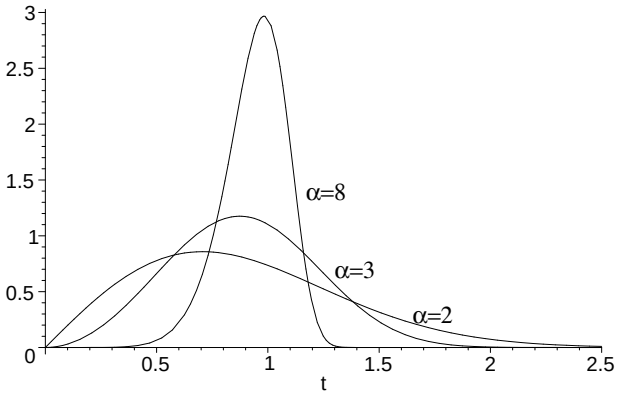

Fig. 2.5. The density $\psi_\alpha(t)$ for $\alpha = 2, 3$ and 8.

Let κ_n be the $(1/n)$-quantile of a c.d.f. $F(\cdot)$; that is, $\kappa_n = \inf\{u|F(u) \geq 1/n\}$. Note that since we assume that the c.d.f. $F(\cdot)$ is continuous in the vicinity of m, for n large enough we have $F(\kappa_n) = 1/n$. The following classical result from the theory of extreme order statistics is of primary importance to us.

Theorem 2.2. *Assume* $\operatorname{ess\,inf} \eta = m > -\infty$, *where* η *has c.d.f.* $F(t)$, *and the function*

$$V(v) = F\left(m + \frac{1}{v}\right), \qquad v > 0,$$

regularly varies at infinity with some exponent $(-\alpha)$, $0 < \alpha < \infty$; *that is,*

$$\lim_{v\to\infty} \frac{V(tv)}{V(v)} = t^{-\alpha}, \qquad \textit{for all}\ \ t > 0\,. \tag{2.42}$$

Then

$$\lim_{n\to\infty} F_{1,n}(m + (\kappa_n - m)z) = \Psi_\alpha(z)\,, \tag{2.43}$$

where $F_{1,n}$ *is the c.d.f.* (2.32), *the c.d.f.* $\Psi_\alpha(z)$ *is defined in* (2.40) *and* κ_n *is the* $(1/n)$*-quantile of* $F(\cdot)$.

The asymptotic relation (2.43) means that the distribution of the sequence of random variables $(\eta_{1,n} - m)/(\kappa_n - m)$ converges (as $n \to \infty$) to the random variable with c.d.f. $\Psi_\alpha(z)$.

The family of c.d.f.'s $\Psi_\alpha(z)$, along with its limiting case

$$\Psi_\infty(z) = \lim_{\alpha \to \infty} \Psi_\alpha(1 + z/\alpha) = 1 - \exp\{-\exp(z)\}\,,\ z > 0,$$

are the only non-degenerate limits of the c.d.f.s of the sequences $(\eta_{1,n} - a_n)/b_n$, where $\{a_n\}$ and $\{b_n\}$ are arbitrary sequences of positive numbers.

If there exist numerical sequences $\{a_n\}$ and $\{b_n\}$ such that the c.d.f.'s of $(\eta_{1,n} - a_n)/b_n$ converge to Ψ_α, then we say that $F(\cdot)$ belongs to the domain of attraction of $\Psi_\alpha(\cdot)$ and express this as $F \in D(\Psi_\alpha)$. The conditions stated in Theorem 2.2 are necessary and sufficient for $F \in D(\Psi_\alpha)$. There are two conditions: $m = \operatorname{ess\,sup} \eta < \infty$ and the condition (2.42). The first one is always valid in global random search applications. The condition (2.42) demands more attention. For example, it is never valid in discrete optimization problems since the c.d.f. $F(\cdot)$ has to be continuous in the vicinity of $m = \operatorname{ess\,inf} \eta$. In fact, for a c.d.f. with a jump at its lower end-point no non-degenerate asymptotic distribution for $\eta_{1,n}$ exists, whatever the normalization (that is, sequences $\{a_n\}$ and $\{b_n\}$).

The condition (2.42) can be written as

$$F(t) = c_0(t - m)^\alpha + \mathrm{o}((t - m)^\alpha) \qquad \text{as } t \downarrow m\,, \tag{2.44}$$

where c_0 is a function of $v = 1/(t - m)$, slowly varying at infinity as $v \to \infty$. Of course, any positive constant is a slowly varying function, but the actual range of eligible functions c_0 is much wider.

The following sufficient condition (the so-called von Mises condition) for (2.42) and (2.43) is often used: if $F(t)$ has a positive derivative $F'(t)$ for all $t \in (m, m + \varepsilon)$ for some $\varepsilon > 0$ and

$$\lim_{t \downarrow m} \frac{(t - m)F'(t)}{F(t)} = \alpha\,,$$

then (2.42) holds.

The following condition is stronger that the condition (2.44) and is often used for justifying properties of the maximum likelihood estimators:

$$F(t) = c_0(t - m)^\alpha \left(1 + \mathrm{O}((t - m)^\beta)\right) \qquad \text{as } t \downarrow m \tag{2.45}$$

for some positive constants c_0, α and β.

The quantity $\kappa_n - m$, where $m = \operatorname{ess\,inf} \eta$ and κ_n is the $(1/n)$-quantile of $F(\cdot)$ enters many formulae below and therefore its asymptotic behaviour is very important. Fortunately, the asymptotic behaviour of $\kappa_n - m$ is clear. Indeed, provided that (2.44) holds with some c_0, we have

$$\frac{1}{n} = F(\kappa_n) \sim c_0\,(\kappa_n - m)^\alpha \qquad \text{as } n \to \infty$$

implying

$$(\kappa_n - m) \sim (c_0 n)^{-1/\alpha} \quad \text{as } n \to \infty\,. \tag{2.46}$$

Extensions to k-th order statistics

There is a one-to-one correspondence between the convergence of the smallest order statistics $\eta_{1,n}$ and of the k-th smallest statistics $\eta_{k,n}$. Assume that $m = \operatorname{ess\,inf} \eta > -\infty$, $n \to \infty$ and let k be fixed. Then it is easy to prove that $F \in D(\Psi_\alpha)$ if and only if the sequence of random variables $(\eta_{k,n} - m)/(\kappa_n - m)$ converges in distribution to the random variable with c.d.f.

$$\begin{aligned}\Psi_\alpha^{(k)}(t) &= 1 - (1 - \Psi_\alpha(t)) \sum_{m=0}^{k-1} \frac{(-\ln(1 - \Psi_\alpha(t)))^m}{m!} \\ &= 1 - \exp(-t^\alpha) \sum_{m=0}^{k-1} \frac{t^{\alpha m}}{m!}, \qquad t > 0\,. \end{aligned} \tag{2.47}$$

The corresponding density is

$$\psi_\alpha^{(k)}(t) = \left(\Psi_\alpha^{(k)}(t)\right)' = \frac{\alpha}{(k-1)!}\, t^{\alpha k - 1} \exp\{-t^\alpha\}\,, \quad t > 0\,. \tag{2.48}$$

The following statement is a generalisation of this fact and reveals the joint asymptotic distribution of the k smallest order statistics: if $m > -\infty$, $F \in D(\Psi_\alpha)$, $n \to \infty$, then for any fixed k the asymptotic distribution of the random vector

$$\left(\frac{\eta_{1,n} - m}{m - \kappa_n}, \frac{\eta_{2,n} - m}{m - \kappa_n}, \cdots, \frac{\eta_{k,n} - m}{m - \kappa_n}\right) \tag{2.49}$$

converges to the distribution with density

$$\psi_\alpha(t_1, \ldots, t_k) = \alpha^k (t_1 \ldots t_k)^{(\alpha-1)} \exp(-t_k^\alpha), \quad 0 < t_1 < \cdots < t_k < \infty\,. \tag{2.50}$$

The density (2.50) is the density of the random vector

$$\left(\nu_1^{1/\alpha}, (\nu_1 + \nu_2)^{1/\alpha}, \ldots, (\nu_1 + \cdots + \nu_k)^{1/\alpha}\right)\,, \tag{2.51}$$

where $\nu_1, \ldots, \nu_k$ are i.i.d.r.v. with exponential density e^{-t}, $t > 0$.

As an important particular case, we find that the joint asymptotic density of the random vector

$$\left(\frac{\eta_{1,n} - m}{\kappa_n - m}, \frac{\eta_{k,n} - m}{\kappa_n - m}\right)$$

coincides with the joint density of the vector $(\nu_1^{1/\alpha}, (\nu_1 + \ldots + \nu_k)^{1/\alpha})$.

The following corollary of this result will be the basic tool in constructing confidence intervals for m.

Proposition 2.1. *If the conditions of Theorem 2.2 hold, then for any fixed integer $k \geq 2$ and $n \to \infty$ the sequence of random variables*

$$D_{n,k} = \frac{\eta_{1,n} - m}{\eta_{k,n} - m}$$

converges in distribution to a random variable with c.d.f.

$$F_k(u) = 1 - \left(1 - \left(\frac{u}{1+u}\right)^{\alpha}\right)^{k-1}, \qquad u \geq 0. \tag{2.52}$$

The proof of this statement is given in Sect. 2.7; it is a simplified and corrected version of the proof of Lemma 7.1.4 in [273].

In the following proposition we use the asymptotic distributions (2.47) and (2.50) to derive the asymptotic formulae for the moments of the random variables $(\eta_{k,n} - m)$ and the first joint moment $\mathrm{E}(\eta_{j,n} - m)(\eta_{k,n} - m)$.

Proposition 2.2. *Let $m = \operatorname{ess\,inf} \eta > -\infty$ and $F \in D(\Psi_\alpha)$ with $\alpha > 1$. Assume that k is either fixed or k tends to infinity as $n \to \infty$ so that $k^2/n \to 0$, $n \to \infty$. Then*

$$E(\eta_{k,n} - m)^\beta \sim (\kappa_n - m)^\beta \, \frac{\Gamma(k + \beta/\alpha)}{\Gamma(k)} \qquad as \;\; n \to \infty \tag{2.53}$$

for any $\beta > 0$ and

$$E(\eta_{j,n} - m)(\eta_{k,n} - m) \sim (\kappa_n - m)^2 \lambda_{jk} \qquad as \;\; n \to \infty, \tag{2.54}$$

where $k \geq j$ and

$$\lambda_{jk} = \frac{\Gamma(k + 2/\alpha)\,\Gamma(j + 1/\alpha)}{\Gamma(k + 1/\alpha)\,\Gamma(j)} \, . \tag{2.55}$$

We give a proof of this statement in Sect. 2.7; this proof is easier than the one given in [273], Sect. 7.1.2.

General results on the rate of convergence of the normalised minima to the extreme value distribution (see e.g. [70] and §2.10 in [86]) imply that in the case considered in Theorem 2.2 this rate is $\mathrm{O}(1/n)$ as $n \to \infty$ (note that [60] and [197], Chapt. 2 contain more sophisticated results on the rate of convergence to the extreme value distribution). This fact along with the asymptotic relation (2.46) imply that for $\alpha \leq 1$ we have

$$\mathrm{E}(\eta_{k,n} - m)^\beta = \mathrm{O}(1/n^\beta)$$

rather than (2.53). Similarly, we have to have $\alpha > 1$ for (2.54) to hold. The reasons why the condition $k^2/n \to 0$ as $n \to \infty$ must be satisfied are explained in [273], p. 245-246.

2.3.3 Record Values and Record Moments

In this section, we survey the theory of record values and record moments. The importance of this topic in global random search is related, first of all, to its link with pure adaptive search and its modifications, see Sects. 2.2.3 and 2.2.4. For all missing proofs and more information on record values and record moments we refer to [6, 171].

Definitions

Let $\eta_1, \eta_2, \ldots$ be a sequence of random variables. Define the sequences of related random variables $L(n)$ and $\eta(n)$ as follows: $L(1) = 1$, $\eta(1) = \eta_1$,

$$L(n+1) = \min\{j > L(n) \colon \eta_j < \eta_{L(n)}\},\ \eta(n) = \eta_{L(n)},\ n = 1, 2, \ldots; \quad (2.56)$$

$L(n)$ are called (lower) record moments corresponding to the sequence $\eta_1, \eta_2, \ldots$ and $\eta(n)$ are the associated (lower) record values.

If we change the inequality sign in (2.56) to be $>$, then we obtain the upper record order moments and upper record values. By changing η_j to $1/\eta_j$ or $(-\eta_j)$ for $j = 1, 2, \ldots$ we correspond the upper record moments to the lower ones. We will only consider the lower moments and values and omit the word 'lower'.

In addition to $L(n)$ and $\eta(n)$, we shall also use the following random variables: $\mathcal{N}(n)$ is the number of record values among $\eta_1, \ldots, \eta_n$ (note that $\mathcal{N}(L(n)) = n$) and $\triangle(n) = L(n) - L(n-1)$, the waiting time between $(n-1)$-th and n-th record moments.

Properties of record moments

Assume that $\eta_1, \eta_2, \ldots$ are i.i.d.r.v. with common continuous c.d.f. $F(\cdot)$. First, consider the non-asymptotic properties of the record moments $L(n)$.

P1: The distribution of $L(n)$ does not depend on $F(\cdot)$.

P2: The sequence of random variables $L(n)$ forms a Markov chain with the starting point $L(1) = 1$ and transition probabilities

$$\Pr\{L(n) = j \mid L(n-1) = i\} = \frac{i}{j(j-1)} \quad \text{for} \quad j > i \geq n-1, \quad n = 2, 3, \ldots$$

P3: The joint distribution of the record moments is

$$\Pr\{L(2) = i_2, \ldots, L(n) = i_n\} = \frac{1}{(i_2 - 1)\ldots(i_n - 1) i_n} \text{ with } 1 < i_2 < \cdots < i_n.$$

Property P3 implies

$$\Pr\{L(2) = j\} = \frac{1}{j(j-1)}, \qquad j > 1\,; \quad (2.57)$$

$$\Pr\{L(n) = j\} = \frac{|S_{j-1}^{n-1}|}{j!}, \qquad j \geq n > 1\,, \quad (2.58)$$

where S_a^b are the Stirling numbers of the first kind.

Property P3 follows from P2, and properties P1 and P2 are simple consequences of the relation between $L(n)$ and $\mathcal{N}(n)$,

$$\Pr\{\mathcal{N}(n) < n\} = \Pr\{L(n) > n\}\,, \tag{2.59}$$

and of the representation

$$\mathcal{N}(n) \stackrel{d}{=} \zeta_1 + \cdots + \zeta_n \qquad (n = 1, 2, \ldots)\,, \tag{2.60}$$

where ζ_j are independent r.v. with $\Pr\{\zeta_j = 1\} = 1/j$ and $\Pr\{\zeta_j = 0\} = 1 - 1/j$. For each j, the random variable ζ_j can be interpreted as the indicator of the event that η_j is the new record value, which is the event $\eta_j < \min\{\eta_1, \ldots \eta_{j-1}\}$.

The sequence of record times $L(n)$ has another useful representation:

$$L(1) = 1, \qquad L(n+1) \stackrel{d}{=} \left\lceil \frac{L(n)}{U_n} \right\rceil \qquad (n = 1, 2, \ldots)\,,$$

where $U_1, U_2, \ldots$ are i.i.d.r.v. uniformly distributed on [0,1].

For any integer $x > 1$, we have

$$\Pr\left\{ \frac{L(n+1)}{L(n)} > x \right\} = \frac{1}{x}\,; \tag{2.61}$$

if x is not an integer, then (2.61) holds asymptotically, as $n \to \infty$.

The representations (2.59) and (2.60) enable the application of classical techniques to obtain the law of large numbers, the central limit theorem and the law of iterated logarithm for the random variables $L(n)$ and $\mathcal{N}(n)$. In particular, as $n \to \infty$ we have

$$\Pr\left\{ \lim_{n\to\infty} \frac{1}{n} \ln L(n) = 1 \right\} = 1,$$
$$\lim_{n\to\infty} \Pr\left\{ (\ln L(n) - n) \le t\sqrt{n} \right\} = \Phi(t)$$

where $\Phi(t)$ is the c.d.f. of the standard normal distribution:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-u^2/2} du\,. \tag{2.62}$$

Consider now the moments of $L(n)$. Using (2.57) and (2.58) we obtain:

$$EL(n) = \infty \ \text{ for all } n = 2, 3, \ldots \tag{2.63}$$

(note that $L(1) = 1$). Moreover, for any $n \ge 2$, the average waiting time $\mathrm{E}[L(n) - L(n-1)]$ of a new record is infinite. This is a very unsatisfactory result for the theory of global random search: it says that on average one has to make infinitely many iterations of PRS to get any improvement over the current best value of the objective function.

The distributions of the inter-record times $\triangle(n+1) = L(n+1) - L(n)$ can be easily computed; they are:

$$\Pr\{\triangle(n+1) = j\} = F(\eta(n)) \cdot (1 - F(\eta(n)))^{j-1}, \quad j = 1, 2, \ldots, \quad n = 1, 2, \ldots$$

That is, the distribution of $\triangle(n+1)$ only depends on the n-th record value $\eta(n)$ and is in fact geometric with parameter of success $F(\eta(n))$.

The logarithmic moments of $L(n)$ can asymptotically be expressed as follows:

$$\mathrm{E} \ln L(n) = n - \gamma + \mathrm{O}\left(\frac{n^2}{2^n}\right), \qquad n \to \infty,$$

$$\mathrm{var}(\ln L(n)) = n - \frac{\pi^2}{6} + \mathrm{O}\left(\frac{n^3}{2^n}\right), \qquad n \to \infty,$$

where $\gamma = 0.5772\ldots$ is the Euler's constant.

The number of records in a given sequence

Consider again the sequence of random variables $\mathcal{N}(n)$, the number of records among the random variables $\eta_1, \ldots, \eta_n$:

$$\mathcal{N}(1) = 1, \qquad \mathcal{N}(n) = 1 + \sum_{j=2}^{n} I_{[\eta_j < \min\{\eta_1, \ldots, \eta_{j-1}\}]}.$$

In accordance with (2.60) and the fact that

$$\mathrm{E}\zeta_j = \tfrac{1}{j} \quad \text{and} \quad \mathrm{var}(\zeta_k) = \tfrac{1}{j} - \tfrac{1}{j^2},$$

for any continuous c.d.f. $F(\cdot)$ we obtain

$$\mathrm{E}\mathcal{N}(n) = \sum_{j=1}^{n} \frac{1}{j} \quad \text{and} \quad \mathrm{var}(\mathcal{N}(n)) = \sum_{j=1}^{n} \left(\frac{1}{j} - \frac{1}{j^2}\right).$$

This implies that both $\mathrm{E}\mathcal{N}(n)$ and $\mathrm{var}(\mathcal{N}(n))$ are of order $\ln n$; the approximation $\mathrm{E}\mathcal{N}(n) \cong \ln n + \gamma$ with $\gamma = 0.5772\ldots$ (the Euler's constant) is very accurate. Additionally, taking into account the asymptotic normality of $(\mathcal{N}(n) - \ln n)/\sqrt{\ln n}$ (the normalized sequence of $\mathcal{N}(n)$), we can use Table 2.2 for making good guesses (using, say, the '3σ-rule') about the number of records $\mathcal{N}(n)$ for given n. This table shows the expected number of records $\mathrm{E}\mathcal{N}(n)$ in a sequence of i.i.d.r.v. $\eta_1, \ldots, \eta_n$, along with the standard deviation of $\mathcal{N}(n)$, for some values of n. One can see that as n increases, the number of records grows very slowly.

n	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
$\mathrm{E}\mathcal{N}(n)$	2.9	5.2	7.5	9.8	12.1	14.4	16.7	19.0	21.3
$\sqrt{\mathrm{var}(\mathcal{N}(n))}$	1.2	1.9	2.4	2.8	3.2	3.6	3.9	4.2	4.4

Table 2.2. Expected number of records $\mathrm{E}\mathcal{N}(n)$ among n i.i.d.r.v. $\eta_1, \ldots, \eta_n$, along with the standard deviation of $\mathcal{N}(n)$.

Record values

Let $\eta(1), \eta(2), \ldots$ be the sequence of record values in the sequence of i.i.d.r.v. $\eta_1, \eta_2, \ldots$ Assume that the c.d.f. $F(t)$ of η_j is continuous with density $p(t)=F'(t)$. Under these assumptions, it is easy to see that the joint density of $\eta(1), \ldots, \eta(n)$ is

$$p(x_1, \ldots, x_n) = \frac{p(x_1)}{F(x_1)} \cdots \frac{p(x_{n-1})}{F(x_{n-1})} \cdot p(x_n), \qquad x_1 \le \cdots \le x_n \,.$$

This implies, in particular, that the sequence $\eta(1), \eta(2), \ldots$ is a Markov chain with transition probabilities

$$\Pr\{\eta(n+1) \le t \mid \eta(n) = u\} = \frac{F(t)}{F(u)}\,, \qquad m \le t \le u\,.$$

For each $n \ge 1$, the c.d.f. of the record value $\eta(n)$ is

$$\Pr\{\eta(n) < t\} = 1 - F(t) \sum_{j=0}^{n-1} \frac{(-\ln(F(t)))^j}{j!}\,.$$

Consider now the asymptotic distribution of the record values $\eta(n)$ corresponding to the sequence of i.i.d.r.v. $\eta_1, \eta_2, \ldots$ as $n \to \infty$, assuming that $m = \operatorname{ess\,inf} \eta_i > -\infty$ and $F \in D(\Psi_\alpha)$; that is, the c.d.f. $F(\cdot)$ belongs to the domain of attraction of the c.d.f. $\Psi_\alpha(\cdot)$ defined in (2.40). Theorem 2.2 implies that the condition $F \in D(\Psi_\alpha)$ yields that the sequence of random variables $(\eta(n) - m)/(\kappa_{L(n)} - m)$ has the asymptotic distribution with the c.d.f. $\Psi_\alpha(\cdot)$; here $L(n)$ is the sequence of record moments, and $\kappa_{L(n)}$ is $(1/L(n))$-quantile of $F(\cdot)$.

Since $L(n)$ is a random variable, $\kappa_{L(n)}$ is a random variable too, which is not very satisfactory. There is, however, another limiting law for the properly normalized record values $(\eta(n) - a(n))/b(n)$ with non-random coefficients $a(n)$ and $b(n)$. Specifically, the conditions $m > -\infty$ and $F \in D(\Psi_\alpha)$ imply that the sequence of random variables $(\eta(n) - m)/(\kappa_{\exp(n/2)} - m)$ converges in distribution to the r.v. with c.d.f.

$$\tilde{\Psi}_\alpha(z) = \begin{cases} 0, & z \le 0 \\ \Phi(-\alpha \ln(z))\,, & z > 0\,; \end{cases}$$

here $\Phi(\cdot)$ is as in (2.62).

Extension to k-th records

To perform statistical inference in global random search algorithms we need several minimal order statistics, rather than just one of them. Similarly, we can use the k-th record moments and the k-th record values. The record moments and the record values considered above are the first record moment and the first record value, respectively (in this case $k = 1$). Assume now the general case $k \geq 1$ and start with the so-called 'Gumbel's method of exceedances', see [105]. This method is aimed to answer the question: 'how many values among future observations exceed past records?'. Specifically, let the c.d.f. $F(\cdot)$ be continuous, $\eta_{1,n} \leq \cdots \leq \eta_{n,n}$ be the order statistics as usual and denote by $S_r^k(n)$ the number of exceedances of $\eta_{k,n}$ among the next r observations $\eta_{n+1}, \ldots, \eta_{n+r}$; that is,

$$S_r^k(n) = \sum_{i=1}^{r} I_{\{\eta_{n+i} < \eta_{k,n}\}} \,.$$

It is an easy consequence of (2.33) that the random variable $S_r^k(n)$ has the hypergeometric distribution with

$$\Pr\left\{S_r^k(n) = j\right\} = \frac{\binom{r+n-k-j}{n-k}\binom{j+k-1}{k-1}}{\binom{r+n}{n}}, \qquad j = 0, 1, \ldots, r \,.$$

In particular, the mean number of exceedances is equal to

$$\mathrm{E}S_r^k(n) = \frac{rk}{n+1} \,.$$

Let us now consider ways of generalizing other results discussed earlier in this section from the case $k = 1$ to the general case $k \geq 1$. We start with definitions.

For each $n \geq k$, we rearrange the random variables $\eta_1, \ldots, \eta_n$ so that

$$\eta_{1,n} \leq \eta_{2,n} \leq \cdots \leq \eta_{n,n} \,.$$

The random variable $\eta_{k,n}$ is the k-th order statistic. The sequence of k-th order statistics is

$$\eta_{k,k} \geq \eta_{k,k+1} \geq \ldots \geq \eta_{k,n} \geq \ldots$$

Let us select the indices n such that there is strict inequality in this sequence:

$$\ldots \geq \eta_{k,n-1} > \eta_{k,n} \geq \ldots$$

This gives us the sequence of k-th record moments $L^{(k)}(n)$. Formally, the sequence $L^{(k)}(n)$ can be defined as follows: $L^{(k)}(0) = 0$, $L^{(k)}(1) = k$ and

$$L^{(k)}(n+1) = \min\left\{j > L^{(k)}(n) \text{ such that } \eta_j < \eta_{k,j-1}\right\}, \quad n = 1, 2, \ldots$$

The sequence of random variables $\eta^{(k)}_{(n)} = \eta_{k,L^{(k)}(n)}$, $n = 1, 2, \ldots$, is the sequence of k-th record values; the differences $\triangle^{(k)}(n) = L^{(k)}(n) - L^{(k)}(n-1)$ are k-th record waiting times; $\mathcal{N}^{(k)}(n)$ is the number of k-th record values among $\eta_1, \ldots, \eta_n$. Of course, $L^{(1)}(n) = L(n)$, $\eta^{(1)}_{(n)} = \eta_{(n)}$, $\triangle^{(1)}(n) = \triangle(n)$ and $\mathcal{N}^{(1)}(n) = \mathcal{N}(n)$.

For each n, the k-th record waiting times have all moments $E\left(\triangle^{(k)}(n)\right)^{\beta}$ of order $0 < \beta < k$; the mean is

$$E\triangle^{(k)}(n) = (k/(k-1))^{n-1}, \; k \geq 2\,. \tag{2.64}$$

Properties of the k-th record moments $L^{(k)}(n)$ are similar to the properties of the ordinary record moments $L(n)$. In particular, for any fixed $k \geq 1$ and any continuous c.d.f. $F(\cdot)$, the sequence of $L^{(k)}(n)$ forms a Markov chain; many limit theorems for $L^{(k)}(n)$ are direct extensions of the related theorems for $L(n)$, see [171], Lectures 18–20.

A very useful tool in studying the k-th record sequences is the so-called 'Ignatov's Theorem' which says that the processes of the k-th records are independent and identically distributed copies of the same random sequence (here the process of the k-th records is defined as a sequence of time moments when the current observation of a sequence of i.i.d.r.v. has rank k), see e.g. [96, 125, 230] and [197], Sect. 4.6. The underlying c.d.f. $F(\cdot)$ of the i.i.d.r.v. is almost arbitrary; in particular, it does not have to be continuous. The Ignatov's Theorem implies that the sequence of moments when the current observation has rank k is the same for any k, in particular, for $k = 1$ where this sequence is the sequence of record moments $\{L(n)\}$ discussed above. The second extremely informative part of this theorem is the independence of the k-th record processes for all $k = 1, 2, \ldots$. Different implications of this theorem are discussed in literature; see, for example, [31].

2.4 Statistical Inference About m: Known Value of the Tail Index

In this section, we consider statistical inference about $m = \text{ess inf}\,\eta$ (in random search applications $m = \min f$) based on the asymptotic theory of extreme order statistics described in Sect. 2.3.2. These statistical procedures will be using only the k smallest order statistics

$$\eta_{1,n} \leq \eta_{2,n} \leq \ldots \leq \eta_{k,n}$$

corresponding to the independent sample $\{\eta_1, \ldots, \eta_n\}$ of the values of a random variable η with c.d.f. $F(t)$. The sample size n is assumed large (formally,

$n \to \infty$) and k is assumed small relative to n (see Sect. 2.4.3 concerning the choice of k).

We assume throughout this section that the conditions of Theorem 2.2 hold for the c.d.f. $F(\cdot)$ (in random search application this c.d.f. is defined by (2.13)) and the value of the tail index α is known. We shall also assume that $\alpha > 1$ (results of Sect. 2.5.3 show that this is the main case of interest in global optimization). As we are interested in the applications of the methodology in global random search, we always assume that the assumptions (i′) and (ii) of Sect. 2.3 are met.

Note that the statistical inference about m and the behaviour of the c.d.f. $F(\cdot)$ in the vicinity of m are much simpler when the value of α is known. Fortunately, in problems of global random search this case can be considered as typical in view of the results of Sect. 2.5.3, where the direct link between the form (2.13) of the c.d.f. $F(\cdot)$ and the value of the tail index α is considered.

A detailed consideration of the theory of asymptotic statistical inference about the bounds of random variables in the case of known α is given in [273], Chap. 7; there has not been much progress in this area since 1991, the time of publication of [273]. Hence, in this section, we only discuss the results that can be directly applied to global random search algorithms, those outlined in Sect. 2.6.1.

2.4.1 Estimation of m

The maximum likelihood estimator

The maximum likelihood estimators of m have been introduced and investigated in [109]. These estimators are constructed under the assumption that $\alpha \geq 2$ and that the distribution of the sample is the asymptotic one, which is the Weibull distribution with c.d.f. (2.40).

For fixed n and k, set

$$\beta_j(\hat{m}) = (\eta_{k,n} - \eta_{j,n})/(\eta_{j,n} - \hat{m}), \quad j < k. \tag{2.65}$$

Differentiating the logarithm of the likelihood function, see (2.89) below, with respect to m and c_0 and equating the derivatives to zero, we obtain the following likelihood equation for $\hat{m}$:

$$(\alpha - 1) \sum_{j=1}^{k-1} \beta_j(\hat{m}) = k\,. \tag{2.66}$$

Hence, when α is known, the maximum likelihood estimator $\hat{m}$ of m is the solution of the equation (2.66) under the condition $\hat{m} \leq \eta_{1,n}$; if there is no solution of this equation satisfying the inequality $\hat{m} \leq \eta_{1,n}$, then $\hat{m}$ is defined as $\eta_{1,n}$.

If the conditions (2.45), $\alpha \geq 2$, $k \to \infty$, $k/n \to 0$ (as $n \to \infty$) are satisfied, then the maximum likelihood estimators of m are asymptotically normal and

asymptotically efficient in the class of asymptotically normal estimators and their mean square error $E(\hat{m}-m)^2$ is asymptotically

$$E(\hat{m}-m)^2 \sim \begin{cases} (1-\frac{2}{\alpha})(\kappa_n-m)^2\, k^{-1+2/\alpha} & \text{for } \alpha>2, \\ (\kappa_n-m)^2 \ln k & \text{for } \alpha=2\,. \end{cases} \tag{2.67}$$

As usual, κ_n is the $(1/n)$-quantile of the c.d.f. $F(\cdot)$.

Linear estimators

Linear estimators of m are simpler than the maximum likelihood ones. However, the best linear estimators possess similar asymptotic properties.

Introduce the following notation:

$$a=(a_1,\dots,a_k)' \in \mathbb{R}^k, \quad \mathbf{1}=(1,1,\dots,1)' \in \mathbb{R}^k,$$

$$b_i = \Gamma(i+1/\alpha)\,/\,\Gamma(i), \quad b=(b_1,\dots b_k)' \in \mathbb{R}^k,$$

$$\lambda_{ji}=\lambda_{ij}=\frac{\Gamma(i+2/\alpha)\,\Gamma(j+1/\alpha)}{\Gamma(i+1/\alpha)\,\Gamma(j)} \quad \text{for } i\ge j, \quad \Lambda=\|\lambda_{ij}\|_{i,j=1}^k\,; \tag{2.68}$$

here $\Gamma(\cdot)$ is the gamma–function.

A general linear estimator of m can be written as

$$\hat{m}_{n,k}(a)=\sum_{i=1}^k a_i\eta_{i,n}\,, \tag{2.69}$$

where $a=(a_1,\dots,a_k)'$ is the vector of coefficients.

Using (2.53) with $\beta=1$, for any linear estimator $\hat{m}_{n,k}(a)$ of the form (2.69) we obtain:

$$\mathrm{E}\hat{m}_{n,k}(a)=\sum_{i=1}^k a_i\mathrm{E}\eta_{i,n}=m\sum_{i=1}^k a_i-(\kappa_n-m)a'b+\mathrm{o}(\kappa_n-m),\ n\to\infty. \tag{2.70}$$

Since $\kappa_n-m\to 0$ as $n\to\infty$, see (2.46), and the variances of all $\eta_{i,n}$ are finite (this is true, in particular, if the c.d.f. $F(\cdot)$ has bounded support, see assumption (i′) of Sect. 2.3), the estimator $\hat{m}_{n,k}(a)$ is a consistent estimator of m if and only if

$$a'\mathbf{1}=\sum_{i=1}^k a_i=1\,. \tag{2.71}$$

The additional condition

$$a'b=0 \qquad \left(\iff \sum_{i=1}^k a_ib_i=0\right) \tag{2.72}$$

guarantees that for $\alpha > 1$ the corresponding estimator $\hat{m}_{n,k}(a)$ has a bias of the order $o(\kappa_n - m) = o(n^{-1/\alpha})$, as $n \to \infty$, rather than $O(n^{-1/\alpha})$ for a general consistent linear estimator.

For a general consistent estimator $\hat{m}_{n,k}(a)$, we obtain from (2.70):

$$E\hat{m}_{n,k}(a) - m \sim (\kappa_n - m)\, a'b, \quad n \to \infty\,. \tag{2.73}$$

The mean square error of a general consistent estimator $\hat{m}_{n,k}(a)$ is obtained by applying (2.54):

$$E(\hat{m}_{n,k}(a) - m)^2 \sim (\kappa_n - m)^2\, a'\Lambda a, \quad n \to \infty\,. \tag{2.74}$$

Examples of linear estimators

For the simplest and most commonly used estimator of m, where only the minimal order statistic is used, $\hat{m}_{n,k}(a^{(0)}) = \eta_{1,n}$, where $a^{(0)}$, the vector of coefficients, is $a^{(0)} = (1, 0, \ldots, 0)'$. For this estimator we easily obtain

$$E(\hat{m}_{n,k}(a^{(0)}) - m)^2 \sim (\kappa_n - m)^2\, \Gamma(1 + 2/\alpha), \quad n \to \infty\,. \tag{2.75}$$

This is, however, a rather poor estimator, see (2.82) and Fig. 2.6.

The r.h.s. of (2.74) is a natural optimality criterion for selecting the vector a. The optimal consistent estimator $\hat{m}_{n,k}(a^*)$, we shall call it *the optimal linear estimator*, is determined by the vector of coefficients

$$a^* = \arg \min_{a: a'\mathbf{1}=1} a'\Lambda a = \frac{\Lambda^{-1}\mathbf{1}}{\mathbf{1}'\Lambda^{-1}\mathbf{1}}\,. \tag{2.76}$$

The estimator $\hat{m}_{n,k}(a^*)$ has been suggested in [52], where the form (2.76) for the vector of coefficients was obtained.

Solving the quadratic programming problem in (2.76) is straightforward. In the process of doing that, we obtain

$$\min_{a: a'\mathbf{1}=1} a'\Lambda a = (a^*)'\Lambda a^* = 1/\mathbf{1}'\Lambda^{-1}\mathbf{1}\,. \tag{2.77}$$

Lemma 7.3.4 in [273] gives us the following expression for the r.h.s. of (2.77):

$$\mathbf{1}'\Lambda^{-1}\mathbf{1} \;=\; \begin{cases} \frac{1}{\alpha-2}\left(\frac{\alpha\Gamma(k+1)}{\Gamma(k+2/\alpha)} - \frac{2}{\Gamma(1+2/\alpha)}\right) & \text{for } \alpha \neq 2, \\ \sum_{i=1}^{k} 1/i & \text{for } \alpha = 2\,; \end{cases} \tag{2.78}$$

this expression is valid for all $\alpha > 0$ and $k = 1, 2, \ldots$

The components a_i^* $(i = 1, \ldots, k)$ of the vector a^* can be evaluated explicitely: $a_i^* = u_i/\mathbf{1}'\Lambda^{-1}\mathbf{1}$ for $i = 1, \ldots, k$ with

$$\begin{aligned} u_1 &= (\alpha + 1)\,/\,\Gamma(1 + 2/\alpha), \\ u_i &= (\alpha - 1)\,\Gamma(i)/\Gamma(i + 2/\alpha) \qquad \text{for } i = 2, \ldots, k-1, \\ u_k &= -(\alpha k - \alpha + 1)\Gamma(k)\,/\,\Gamma(k + 2/\alpha). \end{aligned}$$

Deriving this expression for the coefficients of the vector a^* is far from trivial, see [273], Sect. 7.3.3.

The asymptotic properties (when both n and k are large) of the optimal linear estimators coincide with the properties of the maximum likelihood estimators and hold under the same regularity conditions (we again refer to [273], Sect. 7.3.3). In particular, the optimal linear estimators $\hat{m}_{n,k}(a^*)$ of m are asymptotically normal (as $n \to \infty$, $k \to \infty$, $k/n \to 0$) and their mean square error $E(\hat{m}_{n,k}(a^*)-m)^2$ asymptotically behaves like the r.h.s. of (2.67).

Consider two other linear estimators which have similar asymptotic properties (as $n \to \infty$, $k \to \infty$ and $k/n \to 0$).

The first one is the estimator $\hat{m}_{n,k}(a^+)$ which is optimal in the class of linear estimators satisfying the consistency condition (2.71) and the additional condition (2.72); it is determined by the vector

$$a^+ = \arg \min_{\substack{a:\, a'\mathbf{1}=1,\\ a'b=0}} a'\Lambda a = \frac{\Lambda^{-1}\mathbf{1} - (b'\Lambda^{-1}\mathbf{1})\Lambda^{-1}b/(b'\Lambda^{-1}b)}{\mathbf{1}\Lambda^{-1}\mathbf{1} - (b'\Lambda^{-1}\mathbf{1})^2/(b'\Lambda^{-1}b)} \tag{2.79}$$

(the solution to the above quadratic minimization problem is easily found using Lagrange multipliers). For the estimator $\hat{m}_{n,k}(a^+)$, the additional condition (2.72) guarantees a faster rate of decrease of the bias $E\hat{m}_{n,k}(a) - m$, as $n \to \infty$.

The Csörgő–Mason estimator $\hat{m}_{n,k}(a^{CM})$ (proposed in [55]) is determined by the vector a^{CM} with components

$$a_i = \begin{cases} v_i & \text{for } \alpha > 2, \quad i = 1, \ldots, k-1 \\ v_k + 2 - \alpha & \text{for } \alpha > 2, \quad i = k \\ 2/\ln(k) & \text{for } \alpha = 2, \quad i = 1 \\ \ln(1+1/i)/\ln(k) & \text{for } \alpha = 2, \quad i = 2, \ldots, k-1 \\ (\ln(1+1/k) - 2)/\ln(k) & \text{for } \alpha = 2, \quad i = k \end{cases}$$

with

$$v_j = (\alpha - 1)k^{2/\alpha - 1}\left(j^{1-2/\alpha} - (j-1)^{1-2/\alpha}\right).$$

The finite–sample behaviours of the optimal unbiased consistent estimator $\hat{m}_{n,k}(a^+)$ and the Csörgő–Mason estimator $\hat{m}_{n,k}(a^{CM})$ are slightly worse than that of the optimal consistent estimator $\hat{m}_{n,k}(a^*)$.

For practical use, a very simple estimator

$$\hat{m}_{n,k}(a^U) = (1 + C_k)\eta_{1,n} - C_k\eta_{k,n} \tag{2.80}$$

with $a^U = (1+C_k, 0, \ldots, 0, -C_k)'$ may be recommended, where $C_k = b_1/(b_k - b_1)$ is found from the condition $a'b = 0$. (An estimator resembling (2.80) was proposed in [257].)

For large values of α, which is an important case in global optimization practice,

$$\Gamma(k+1/\alpha)-\Gamma(k)\sim\frac{1}{\alpha}\Gamma'(k)\quad\text{ as }\ \alpha\to\infty$$

and therefore

$$C_k\sim\frac{\Gamma(1)+\frac{1}{\alpha}\Gamma'(1)}{1+\frac{1}{\alpha}\psi(k+1)-\Gamma(1)-\frac{1}{\alpha}\Gamma(1)}=\frac{\alpha-\gamma}{\psi(k+1)+\gamma},\qquad\alpha\to\infty\,,$$

where $\psi(\cdot)=\Gamma'(\cdot)/\Gamma(\cdot)$ is the psi-function and $\gamma\cong 0.5772$ is the Euler constant.

Asymptotic efficiency of the estimators

If k is fixed, then the asymptotic efficiency of any consistent linear estimator $\hat{m}_{n,k}(a)$ can naturally be defined as

$$\mathrm{eff}(\hat{m}_{n,k}(a))=\left(\min_{c\in\mathbb{R}^k:\,c'\mathbf{1}=1}c'\Lambda c\right)/a'\Lambda a\,.$$

Obviously, $0\le\mathrm{eff}(\hat{m}_{n,k}(a))\le 1$ for any $a\in\mathbb{R}^k$ satisfying the consistency condition (2.71). In view of (2.77) we obtain

$$\mathrm{eff}(\hat{m}_{n,k}(a))=\frac{1}{\mathbf{1}'\Lambda^{-1}\mathbf{1}\cdot a'\Lambda a}\,,\tag{2.81}$$

where $\mathbf{1}'\Lambda^{-1}\mathbf{1}$ can be computed using the expression (2.78).

In particular, the asymptotic efficiency of the simplest estimator $\hat{m}_{n,k}(a^{(0)})=\eta_{1,n}$ is

$$\mathrm{eff}(\hat{m}_{n,k}(a^{(0)}))=\frac{1}{\mathbf{1}'\Lambda^{-1}\mathbf{1}\cdot\Gamma(1+2/\alpha)}\,.\tag{2.82}$$

This result easily follows from (2.75) and (2.85). For large k and α, the asymptotic efficiency of the estimator $\hat{m}_{n,k}(a^{(0)})$ is low, see Fig. 2.6.

Asymptotic efficiency of the estimator $\hat{m}_{n,k}(a^U)$ is higher (especially for small k) than that of the simplest estimator $\hat{m}_{n,k}(a^{(0)})=\eta_{1,n}$. Fig. 2.7 displays this efficiency for $k=3,5$ and 20 and varying α (for $k=2$ the asymptotic efficiency of $\hat{m}_{n,k}(a^U)$ is equal to 1).

Note that the asymptotic efficiency of the optimal estimator $\hat{m}_{n,k}(a^*)$ can be low if an incorrect value of α is used to construct this estimator. This issue is considered in Sect. 2.5.2.

Finite-sample efficiency (simulation results)

Let us make a comparison of the efficiency for the maximum likelihood and linear estimators of m given finite samples of size n drawn from the Weibull distribution with tail index α. Considering the Weibull distribution means assuming that the original sample size n is large enough for the asymptotic distribution for the minimal statistics to be reached.

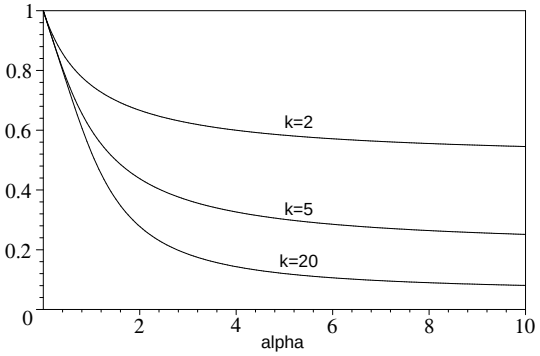

Fig. 2.6. Asymptotic efficiency of the simplest estimator $\hat{m}_{n,k}(a^{(0)}) = \eta_{1,n}$, see (2.82), for $k = 2, 5$ and 20 and varying α.

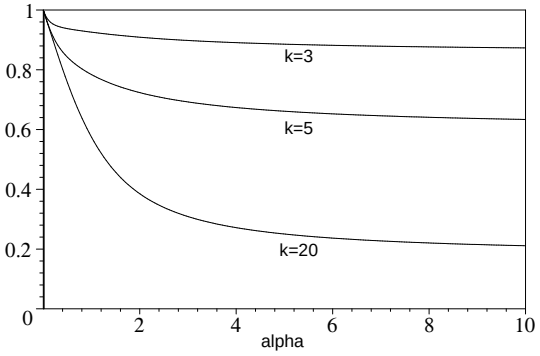

Fig. 2.7. Asymptotic efficiency of the estimator $\hat{m}_{n,k}(a^U)$ for $k = 3, 5$ and 20 and varying α.

According to the definition, for each k, the optimal linear estimator $\hat{m}_{n,k}(a^*)$, with a^* given in (2.76), provides the lowest mean square error in the class of all linear consistent estimators, as $n \to \infty$. In view of (2.74) and (2.77), we have, for the asymptotic mean square error of $\hat{m}_{n,k}(a^*)$:

$$\lim_{n\to\infty} \frac{\mathbf{1}'\Lambda^{-1}\mathbf{1}}{(\kappa_n - m)^2}\mathrm{MSE}(\hat{m}_{n,k}(a^*)) = 1\,, \tag{2.83}$$

for any k. Therefore, for fixed n and k it is natural to define the finite-sample efficiency of an estimator $\hat{m}$ as

$$\frac{(\kappa_n - m)^2}{\mathbf{1}'\Lambda^{-1}\mathbf{1}}/\mathrm{MSE}(\hat{m})\,. \tag{2.84}$$

Since we consider finite samples, it is possible for the efficiency to be slightly greater than 1.

Below, the efficiency of an estimator $\hat{m}$ will be estimated based on taking $R = 10\,000$ estimators of $\hat{m}_j$, where each $\hat{m}_j$ $(j = 1, \ldots, R)$ is estimated from

an independent sample of size n; that is,

$$\mathrm{MSE}(\hat{m}) \simeq \frac{1}{R}\sum_{j=1}^{R}(\hat{m_j} - m)^2 .$$

Thus, for fixed k, n and R, we use the following definition of efficiency of an estimator $\hat{m}$:

$$\mathrm{eff}(\hat{m}) = \left[\frac{(\kappa_n - m)^2}{\mathbf{1}'\Lambda^{-1}\mathbf{1}}\right] / \left[\frac{1}{R}\sum_{j=1}^{R}(\hat{m_j} - m)^2\right] . \tag{2.85}$$

As $R \to \infty$, the efficiency (2.85) tends to (2.84).

where in our case $m = 0$, $n = 100$, $R = 10\,000$ and k varies.

Fig. 2.8 shows the efficiencies (2.85) computed for $n = 100$, $R = 10\,000$, α=1,2,5,10, and varying k for the following estimators:

- the optimal linear estimator $\hat{m}_{n,k}(a^*)$ (depicted as circles),
- the maximum likelihood estimator (squares),
- the linear estimators $\hat{m}_{n,k}(a^+)$ defined by the vector (2.79) (triangles),
- Csörgő–Mason estimators $\hat{m}_{n,k}(a^{CM})$ (dots),
- the minimum order statistic $\eta_{1,n} = \hat{m}_{n,k}(a^{(0)})$ (bullets).

Fig. 2.8 demonstrates that the mean square error of the optimal linear estimator $\hat{m}_{n,k}(a^*)$ is very close to the asymptotically optimal value of the MSE given by (2.83) for all $\alpha \geq 1$ (sometimes it is even larger than this value). The estimator $\hat{m}_{n,k}(a^*)$ clearly provides the lowest mean square error in the class of estimators considered. The efficiency of the maximum likelihood estimator (MLE) is consistently lower than the efficiency of $\hat{m}_{n,k}(a^*)$, especially when α is small; note that MLE can only be used for $\alpha \geq 2$. Note also that the actual efficiency curves of MLE are rather uneven; they have been considerably smoothed in this figure.

The efficiency of the minimum order statistic decreases monotonically as $k \to \infty$, this is because the estimator is not using $k-1$ out of k order statistics. The efficiency of the linear estimator $\hat{m}_{n,k}(a^+)$ is poor for small k (as the unbiasedness condition (2.72) takes away one degree of freedom for the coefficients) but increases monotonically as k increases. The efficiencies of the minimum order statistic and the m^{Δ} estimators are equal for $k = 2$. This can be verified by considering the asymptotic mean square errors (as $n \to \infty$) of these two estimators at this point. The efficiency of the Csörgő–Mason estimators is poor for small α (note that this estimator is only defined for $\alpha \geq 2$) but gets better when α increases; thus, for $\alpha = 10$ the efficiency of the Csörgő–Mason estimator is basically 1.

The case of small values of α has a particular interest. Unlike the MLE and the Csörgő–Mason estimator, the linear estimators $\hat{m}_{n,k}(a^*)$ and $\hat{m}_{n,k}(a^+)$ are defined in the region $0 < \alpha < 2$ and behave rather well.

Simulation study of the bias of the estimators shows that the bias of the four main estimators (namely, MLE, $\hat{m}_{n,k}(a^*)$, $\hat{m}_{n,k}(a^+)$ and $\hat{m}_{n,k}(a^{CM})$) improves as both k and α increase; for large k and *alpha* this bias is approximately the same; for small α the bias of the Csörgő–Mason estimator is large but the bias of the other three estimators is comparable for all $\alpha \geq 2$ (note again that MLE is properly defined only for $\alpha \geq 2$). See [110] for more simulation results and related discussions.

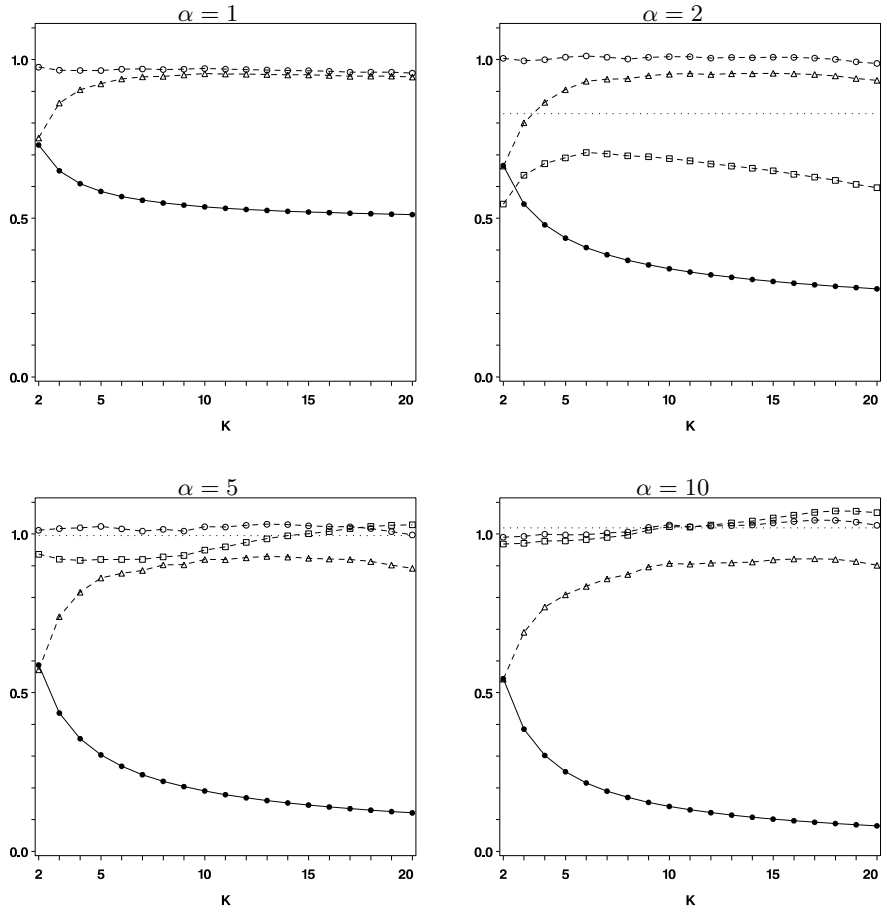

Fig. 2.8. Efficiency as defined in (2.85) of different estimators; $n = 100$, $R = 10\,000$, $\alpha = 1, 2, 5, 10$, against k.

2.4.2 Confidence Intervals and Hypothesis Testing

In global random search one of the most important statistical problems is testing the hypothesis $H_0 : m \le K$ versus the alternative $H_1 : m > K$, where K is some fixed number, $K < \eta_{1,n}$. For instance, K may be the record value of the objective function $f(\cdot)$ attained at another region, see Sect. 2.6.1.

Following the standard route, to construct a test for $H_0 : m \le K$ we construct a one-sided confidence interval for m of a fixed confidence level and reject H_0 if K does not fall into this interval.

The most convenient procedure for constructing confidence intervals for m was proposed in [51]. According to this procedure, the one-sided confidence interval for m is

$$[\eta_{1,n},\, \eta_{1,n} + r_{k,\delta}(\eta_{k,n} - \eta_{1,n})]\,. \tag{2.86}$$

Here

$$r_{k,\delta} = \left((1-\delta^{1/k})^{-1/\alpha} - 1\right)^{-1}, \tag{2.87}$$

the $(1-\delta)$-quantile of the c.d.f. $F_k(u)$ defined in (2.52). Proposition 2.1 of Sect. 2.3.2 then implies that the asymptotic (as $n \to \infty$) confidence level of the interval (2.86) equals $1 - \delta$.

The test corresponding to the confidence interval (2.86), for testing the hypothesis $H_0 : m \le K$, is defined by the rejection region

$$\{(\eta_1, \ldots, \eta_n) \,:\, (\eta_{1,n} - K)/(\eta_{k,n} - \eta_{1,n}) \ge r_{k,\delta}\}\,. \tag{2.88}$$

The first kind error probability of this test asymptotically (as $n \to \infty$) does not exceed δ; this is a consequence of Proposition 2.1 of Sect. 2.3.2.

Different asymptotic expressions for the power function of this procedure can be found in [273], Sect. 7.1.5.

Note that if n is very large (in this case, k may also be chosen large enough), then for constructing the confidence intervals and testing hypotheses one may also use the asymptotic normality (discussed above) of some estimators of m.

2.4.3 Choice of n and k

Let us briefly address the practically important problem of the choice of k and the sufficiency of the sample sample size n for applicability of the methodology considered above to the practice of global optimization.

Theoretically, n should be large enough to guarantee that there are enough, at least k, sample points in the vicinity of the global minimizer. Everything now depends on how we define 'the vicinity of the global minimizer'. This, in turn, depends on the objective function. For example, if the objective function is steep in the vicinity of the global minimizer (as an example, see Fig. 1.1), then the region of attraction of this minimizer is small and there is a high possibility that this region is completely missed. If the region of attraction of

the minimizer is not reached, then the statistical inference will be made about some other minimum (perhaps, local).

Another important point is that the vicinity of a minimizer should not be confused with the region of attraction of the minimizer, see Sects. 1.1.2 and 1.1.3. If, for example, the objective function $f(\cdot)$ is the sum of a smooth and slowly varying function $f_1(\cdot)$ and a small irregular function $f_2(\cdot)$ (see Figs. 1.3 and 1.4) then small variations in the values of $f(\cdot)$ due to the presence of $f_2(\cdot)$ can be regarded as small variations in the sample points, which are statistically insignificant provided that k is not too large. On the other hand, if the objective function changes its shape approaching the minimum within the region of attraction of the global minimizer, this would imply that Theorem 2.2 is practically useless. Theoretically, however, as $n \to \infty$ we can select $k \to \infty$ and construct a consistent estimator of the tail index.

The problem of how large the sample size n should be can be approached from the more formal point of view of the rate of convergence in (2.43). There are a number of results concerning the estimation of this convergence rate, see for instance Sect. 2.10 in [86]; note that these estimators depend on different characteristics of behaviour of the c.d.f. $F(t)$ for t close to m.

Theoretically, the value of k should be such that at least k sample points belong to the vicinity of the global minimizer. It may, of course, happen that some of these k points are far from this vicinity. This would narrow the gap between $\eta_{1,n}$ and $\eta_{k,n}$ (in the probabilistic sense, and in comparison to such a gap when all the points belong to the same vicinity of the global minimizer). This would lead, in particular, to the over-estimation of m. Since we know the location of the test points, this may sometimes be corrected as follows: using some prior information about the objective function we can define a region (say, a ball) with the centre at the record point so the points outside the region will not be able to contribute to the set of the k smallest order statistics.

From the theoretical view-point, k should be small relative to the sample size n, which tends to infinity. Typically, the theoretically optimal choice of k is $k \to \infty$ so that $k/n \to 0$ as $n \to \infty$. In practice, however, n is never large enough and therefore small or moderate values of k should be used. Theoretical results of Sect. 2.4 imply that for many procedures a reasonably small value of k, say $k = 5$, is almost as good as the theoretically optimal choice $k \to \infty$, so we do not loose much in the asymptotic efficiency by restricting k to small values.

Another argument in favour of small k is given in Sect. 2.5.2: if the value of the tail index α is not correct (for instance, α has been estimated), then an increase in k (formally, $k \to \infty$) decreases the accuracy of precision in the estimators of m.

If the tail index is unknown, the problem of the choice of k is more serious than when α is known, see Sect. 2.5.1. For example, consistency of estimators of α can only be achieved if $k \to \infty$ as $n \to \infty$. Consideration can be given

to a mixed strategy which uses a large number of extreme order statistics to estimate α and a relatively small number of these statistics for estimating m.

2.5 Unknown Value of the Tail Index

Our main objective is making statistical inferences about m based on an independent sample from the c.d.f. $F(\cdot)$ given in (2.13). In Sect. 2.4 we have shown how to make statistical inferences when the value of the tail index α is known. In Sect. 2.5.3 below, we show that the specific form of the c.d.f. (2.13) in many cases enables explicit determination of the value of α.

An alternative approach would be to find an estimator $\hat{\alpha}$ for α and use this estimator in place of the true value of α. However, we will show in Sect. 2.5.2 that this approach leads to a significant drop in precision of statistical inference procedures about m, in comparison to the case of known α.

Additionally, the requirements for the sample size seem to be unrealistic. Indeed, to construct any consistent estimator of α we must have $k = k(n)$ observations with $k(n) \to \infty$ (as $n \to \infty$) belonging to the lower tail of the c.d.f. (2.13), where the approximation (2.44) can be applied. In global optimization problems, however, obtaining more than a few observations in this region is problematic.

In global random search problems, making statistical inference about α is most useful for checking upon one of a few possible exact values of α, say $\alpha = 2/d$ or $\alpha = 1/d$; these expressions for α follow from Theorem 2.3 and related results, see Sect. 2.5.3.

2.5.1 Statistical Inference

In this section, we assume that the conditions of Theorem 2.2 are met but the value of the tail index α is unknown. Unlike the case considered in Sect. 2.4, a satisfactory precision of the statistical inference can only be guaranteed if k is large enough. Therefore, we shall suppose that the parameter k is chosen so that $k = k(n) \to \infty,\ k/n \to 0$, as $n \to \infty$. Also, we shall assume that the condition (2.45) is met.

The standard way of making statistical inference concerning m, when α is unknown, is to construct an estimator $\hat{\alpha}$ of α and to substitute $\hat{\alpha}$ for α in the formulae which determine the statistical procedures for the case of known α.

The topic of making statistical inferences about the value of the tail index is widely discussed in literature including a very recent one, see for example, [13, 32, 46, 61, 65, 133, 134, 149, 183].

Easily readable surveys of standard results concerning different estimators of α and their asymptotic properties can be found in Sect. 6.4.2 of [68] and in Sect. 2.6 of [139]. The two most known estimators are the so-called Hill estimator

$$\hat{\alpha}^{(H)} = \left(\ln \eta_{k,n} - \frac{1}{k} \sum_{j=1}^{k} \ln \eta_{j,n} \right)^{-1}$$

suggested in [118], and the Pickands estimator

$$\hat{\alpha}^{(P)} = \frac{1}{\ln 2} \ln \frac{\eta_{2k,n} - \eta_{k,n}}{\eta_{4k,n} - \eta_{2k,n}}$$

proposed in [186]. Provided that the conditions of Theorem 2.2, along with an additional regularity condition of the type (2.45), are satisfied and $k \to \infty$ as $n \to \infty$, both estimators of α are consistent and asymptotically normal. Their asymptotic properties are similar. The main practical problem is, of course, the choice of k. This problem has been addressed in a number of articles, see e.g. [65]. However, this problem can hardly be adequately resolved in global random search applications as the value of n required to achieve a reasonable precision in statistical inference about $m = \min f$ must be astronomical when the dimension d of A is not very small (recall that one of the main attractive points of the global random search methods is their applicability for solving problems with moderate or large dimension).

Results of Sect. 2.5.2 show that the linear estimators which perform well when the value of α is known become much less precise when the value of α is not known. Their asymptotic properties in the case when $\hat{\alpha}$ is noticeably different from α are poor and, consequently, it is often not worth using these estimators in the case of unknown α.

A slightly different way of making statistical inferences about m is based on making the inferences about m and α simultaneously using the maximum likelihood principle outlined below (see [109, 186, 228] for more details).

Assume that the asymptotic relation (2.42), along with the additional regularity condition (2.45) hold with some $c_0 > 0$ and $\alpha \geq 2$. Then the likelihood function depending on the unknown parameters c_0, α and m is asymptotically, as $n \to \infty$, equal to

$$\begin{aligned} &L(\eta_{1,n}, \ldots, \eta_{k,n};\, m, c_0, \alpha) \\ &= \tfrac{n!}{(n-k)!} (c_0 \alpha)^k \left(1 - c_0 \left(\eta_{k,n} - m\right)^{\alpha}\right)^{n-k} \prod_{j=1}^{k} \left(\eta_{j,n} - m\right)^{\alpha-1} . \end{aligned} \tag{2.89}$$

This asymptotic form of the likelihood function is treated as the exact one. The maximisation of (2.89), with respect to c_0 for fixed $\alpha = \hat{\alpha}$ and $m = \hat{m}$ gives the maximum likelihood estimator for c_0:

$$\hat{c}_0 = \frac{k}{n} \left(\eta_{k,n} - \hat{m}\right)^{\hat{\alpha}} . \tag{2.90}$$

The maximisation of (2.89) with respect to α for fixed $m = \hat{m}$ and the substitution (2.90) for c_0, gives the maximum likelihood estimator for α:

$$\hat{\alpha} = k/\sum_{j=1}^{k-1} \ln(1+\beta_j(\hat{m}))\,, \tag{2.91}$$

where $\beta_j(\hat{m})$ are defined in (2.65). The remaining problem is to define the maximum likelihood estimator $\hat{m}$ for m. It cannot be defined as the global maximizer of the likelihood function $L(\eta_{1,n},\dots,\eta_{n-k};\, m, \hat{c}_0,\, \hat{\alpha})$, since the global maximum is achieved at $m = \eta_{1,n}$ and equals $+\infty$ (meaning that the proper maximum likelihood estimator of m is $\eta_{1,n}$, which is a poor estimator). According to the proposal of P.Hall [109], $\hat{m}$ is defined as a solution of the likelihood equation, which is

$$1/\sum_{j=1}^{k-1} \ln(1+\beta_j(\hat{m})) \;-\; 1/\sum_{j=1}^{k-1} \beta_j(\hat{m}) \;=\; 1/k \tag{2.92}$$

provided that $\hat{m} \le \eta_{1,n}$. If there is no solution to the equation (2.92) in the half-interval $(-\infty, \eta_{1,n})$, then $\hat{m}$ is taken as $\eta_{1,n}$; if there is more than one solution of this equation in $(-\infty, \eta_{1,n})$ (that is, the likelihood function is multimodal), then the largest solution is taken. Despite the fact that the estimator does not typically maximize the likelihood function (execpt in the trivial case where $\eta_{1,n}$ is taken as the estimator), it is still called the maximum likelihood estimator. Note that the equation (2.92) is exactly the equation (2.66) with α replaced by $\hat{\alpha}$ of (2.91).

Under the regularity condition (2.45) the maximum likelihood estimator $\hat{m}$ of m is asymptotically normal with mean m and the variance

$$(\alpha-1)^2(1-2/\alpha)(\kappa_n - m)^2 k^{-1+2/\alpha}, \;\; \alpha > 2,\; n\to\infty,\; k\to\infty,\; k/n \to 0.$$

This differs from the r.h.s. of (2.67) in the multiplier $(\alpha-1)^2$ only.

Formally, we can avoid estimating α and construct confidence intervals and statistical tests for m using the result proved in [262]. This result says that if the conditions of Theorem 2.2 hold, $k\to\infty$, $k/n\to 0$, $n\to\infty$, then the sequence of random variables

$$\frac{(\ln k)\,\ln[(\eta_{2,n}-m)/(\eta_{1,n}-m)]}{\ln[(\eta_{k,n}-\eta_{3,n})/(\eta_{3,n}-\eta_{2,n})]}$$

converges in distribution to a random variable with the exponential density e^{-t}, $t \ge 0$. Some generalizations of this result can be found in [263].

2.5.2 Using an Incorrect Value of the Tail Index

Confidence intervals

Consider what happens to the level of the one-sided confidence interval (2.86) for the case where

$$r'_{k,\delta} = 1/\left((1-\delta^{1/k})^{1/\vartheta} - 1\right)$$

is being substituted for $r_{k,\delta}$ defined in (2.87); this means that ϑ is being used in place of the true α.

Proposition 2.3. *Let the conditions of Theorem 2.2 hold, $n \to \infty$, k and $\vartheta > 0$ be fixed. Then the asymptotic confidence level of the confidence interval*

$$I' = [\eta_{1,n} - r'_{k,\delta}(\eta_{k,n} - \eta_{1,n})\,,\ \eta_{1,n}] \tag{2.93}$$

is equal to

$$1 - (1 - (1-\delta^{1/k})^{\alpha/\vartheta})^k. \tag{2.94}$$

Proof is given in Sect. 2.7.

Note that if we take $\vartheta = \alpha$, then (2.94) is simplified to $1-\delta$; therefore, Proposition 2.3 generalizes the statement of Sect. 2.4.2 saying that the asymptotic confidence level of the interval (2.86) is equal to $1-\delta$.

Linear estimators of m

Let us now follow [274] and study the consequences of using incorrect values of α while constructing linear estimators of m (using incorrect values of α is inevitable when we do not know the exact value of α and use its estimator instead).

Assume that $\alpha > 1$, $\alpha \neq 2$ and start the investigation with the optimal estimator $\hat{m}_{n,k}(a^*)$. Denote by ϑ ($\vartheta \neq \alpha$) the value we use to compute $a^* = a^*(\vartheta)$ and by $\Lambda_0 = \Lambda(\vartheta)$ the matrix $\Lambda = \|\lambda_{ij}\|$ defined in (2.68) with ϑ substituted for α.

In view of (2.85) the asymptotic efficiency of the estimator $\hat{m}_{n,k}(a^*(\vartheta))$ is

$$\mathrm{eff}(\hat{m}_{n,k}(a^*(\vartheta))) = \frac{1}{\mathbf{1}'\Lambda^{-1}\mathbf{1}\cdot(a^*(\vartheta))'\Lambda a^*(\vartheta)} = \frac{(\mathbf{1}'\Lambda_0^{-1}\mathbf{1})^2}{\mathbf{1}'\Lambda^{-1}\mathbf{1}\cdot\mathbf{1}'\Lambda_0^{-1}\Lambda\Lambda_0^{-1}\mathbf{1}}\,.$$

If k is fixed and $|\vartheta-\alpha|$ is small, then the estimator $\hat{m}_{n,k}(a^*(\vartheta))$ is relatively good. For example, if $k=2$ then

$$\Lambda = \begin{pmatrix} \Gamma(1+2/\alpha) & (1+\frac{1}{\alpha})\Gamma(2+2/\alpha) \\ (1+\frac{1}{\alpha})\Gamma(2+2/\alpha) & \Gamma(2+2/\alpha) \end{pmatrix},\ \ a^*(\vartheta) = \begin{pmatrix} 1+\frac{\vartheta}{2} \\ -\frac{\vartheta}{2} \end{pmatrix},$$

$$\lambda'\Lambda^{-1}\lambda = \frac{2(\alpha+1)}{(\alpha+2)\Gamma(1+2/\alpha)} \ \text{ and } \ \mathrm{eff}(\hat{m}_{n,k}(a^*(\vartheta))) = \frac{\alpha+2}{\alpha+2+\alpha(1-\frac{\vartheta}{\alpha})^2}\,.$$

We shall say that the estimator $\hat{m}_{n,k}(a^*(\vartheta))$ is poor if

$$\mathrm{eff}(\hat{m}_{n,k}(a^*(\vartheta))) < \mathrm{eff}(\hat{m}_{n,k}(a^{(0)}))\,; \tag{2.95}$$

that is, the asymptotic efficiency of the estimator $\hat{m}_{n,k}(a^*(\vartheta))$ is worse than the asymptotic efficiency of the simplest estimator $\hat{m}_{n,k}(a^{(0)}) = \eta_{1,n}$. Note that for $k = 2$ we have

$$\mathrm{eff}(\hat{m}_{n,k}(a^{(0)})) = \alpha + 2/(2(\alpha+1))\,.$$

The inequality (2.95) cannot be true for $\vartheta < \alpha$. On the other hand, it is easy to see that the estimator $\hat{m}_{n,k}(a^*(\vartheta))$ is poor when $\vartheta > 2\alpha$.

In the case $k > 2$ the situation is not so clear. For instance, for $k = 3$ we have

$$\Lambda \;=\; \Gamma(1+2/\alpha) \begin{pmatrix} 1 & \frac{\alpha+2}{\alpha+1} & \frac{2(\alpha+2)}{2\alpha+1} \\ \frac{\alpha+2}{\alpha+1} & \frac{\alpha+2}{\alpha} & \frac{2(\alpha+1)(\alpha+2)}{\alpha(2\alpha+1)} \\ \frac{2(\alpha+2)}{2\alpha+1} & \frac{2(\alpha+1)(\alpha+2)}{\alpha(2\alpha+1)} & \frac{(\alpha+1)(\alpha+1)}{\alpha^2} \end{pmatrix},$$

$$a_1^*(\vartheta) \;=\; \frac{(\vartheta+2)(\vartheta+1)^2}{3\vartheta^2+4\vartheta+2}\,,\quad a_2^*(\vartheta) \;=\; \frac{\vartheta(\vartheta^2-1)}{3\vartheta^2+4\vartheta+2}\,,$$

$$a_3^*(\vartheta) \;=\; -\frac{\vartheta^2(2\vartheta+1)}{3\vartheta^2+4\vartheta+2}\,,\quad \mathrm{eff}(\hat{m}_{n,k}(a^{(0)})) \;=\; \frac{(\alpha+1)(\alpha+2)}{3\alpha^2+4\alpha+2}\,.$$

Thus, for given values of α and ϑ, to conclude whether the estimator $\hat{m}_{n,k}(a^*(\vartheta))$ is poor we must compute the value of the two–variate polynomial

$$(a^*(\vartheta))'\left(\frac{1}{\Gamma(1+2/\alpha)}\Lambda\right)(a^*(\vartheta))\,,$$

which depends on α and ϑ, and compare it with 1. The estimator is poor if this value is smaller than 1.

Another interesting case is where k is large. According to [274], for all $\vartheta \neq \alpha$ we have

$$(a^*(\vartheta))'\,\Lambda\,a^*(\vartheta) \sim (\vartheta-2)^2\,(\alpha-\vartheta)^2\,(\vartheta+\alpha\,\vartheta-2\alpha)^{-2}\,k^{2/\alpha} \quad \text{as} \quad k\to\infty\,.$$

In this case the estimator $\hat{m}_{n,k}(a^*(\vartheta))$ is poor (it is asymptotically less efficient than the simplest estimator $\hat{m}_{n,k}(a^{(0)}) = \eta_{1,n}$). The estimator is consistent but the order of convergence (as $k\to\infty$, $n\to\infty$, $k/n\to 0$) of the mean square error $E(m-\hat{m}_{n,k}(a^*(\vartheta)))^2$ to 0 is only $(k/n)^{2/\alpha}$ rather than $(k/n)^{2/\alpha}/k$ for the estimator $\hat{m}_{n,k}(a^{(0)})$.

Thus, if the value of the tail index α is not correct (for instance, α has been estimated), then the increase of k leads to a precision loss in the estimator $\hat{m}_{n,k}(a^*)$. A similar conclusion can be derived for the estimators $\hat{m}_{n,k}(a^+)$ and $\hat{m}_{n,k}(a^{CM})$ since these two estimators are asymptotically equivalent to $\hat{m}_{n,k}(a^*)$ (as $k\to\infty$, $n\to\infty$, $k/n\to 0$).

The situation with the estimator $\hat{m}_{n,k}(a^U(\vartheta))$ is better (that is, this estimator is less sensitive to deviations in α for large k). Indeed, we have as $k \to \infty$:

$$a_1^U(\vartheta) = \frac{b_k}{b_k - b_1} \sim 1 + k^{-1/\vartheta}\Gamma(1+1/\vartheta), \quad a_k^U(\vartheta) \sim -k^{-1/\vartheta}\,\Gamma(1+1/\vartheta),$$

$$\lambda_{11} = \Gamma(1+2/\alpha), \quad \lambda_{kk} = \frac{\Gamma(k+2/\alpha)}{\Gamma(k)} \sim k^{2/\alpha},$$

$$\lambda_{k1} = \frac{\Gamma(k+2/\alpha)\Gamma(1+1/\alpha)}{\Gamma(k+1/\alpha)} \sim k^{1/\alpha}\Gamma(1+1/\alpha),$$

$$(a^U)'\Lambda(a^U) = (a_1^U(\vartheta))^2\,\lambda_{11} + 2a_1^U(\vartheta)a_k^U(\vartheta)\lambda_{k1} + (a_k^U(\vartheta))^2\lambda_{kk}$$

$$\sim \Gamma(1+2/\alpha) - 2\Gamma(1+1/\alpha)k^{1/\alpha-1/\vartheta}\Gamma(1+1/\vartheta) + k^{2/\alpha-2/\vartheta}\Gamma^2(1+1/\vartheta)$$

$$\sim \begin{cases} \Gamma(1+2/\alpha) & \text{for } \vartheta < \alpha, \\ \Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha) & \text{for } \vartheta = \alpha, \\ k^{2/\alpha-2/\vartheta}\Gamma^2(1-1/\vartheta) & \text{for } \vartheta > \alpha. \end{cases}$$

This implies that for $\vartheta < \alpha$ the asymptotic efficiency of the estimator $\hat{m}_{n,k}(a^U(\vartheta))$ asymptotically (as $k \to \infty$) coincides with the asymptotic efficiency of the simplest estimator $\hat{m}_{n,k}(a^{(0)})$, for $\alpha = \vartheta$ the estimator $\hat{m}_{n,k}(a^U)$ is better than $\hat{m}_{n,k}(a^{(0)})$ but worse than $\hat{m}_{n,k}(a^*)$, and, finally, for $\vartheta > \alpha$ the estimator $\hat{m}_{n,k}(a^U)$ is poor but it is much more asymptotically efficient than $\hat{m}_{n,k}(a^*)$.

2.5.3 Exact Determination of the Value of the Tail Index

Recall that the c.d.f. $F(\cdot)$ arising in global random search problems has the specific form (2.13). As we show in this section, this specific form often enables the determination of the value of the tail index α explicitly. It gives us the possibility of using the simple and efficient techniques of Sect. 2.4, rather than the techniques of Sect. 2.5.1, which require a much larger sample size.

The basic result is the following theorem.

Theorem 2.3. *Assume that the global minimizer x_* of $f(\cdot)$ is unique and Conditions C1 – C4, C8 and C9 of Sect. 2.1.1 along with the condition C10 of Sect. 2.2.1 are met. Assume, in addition, that the representation*

$$f(x) - m = w(\|x - x_*\|)H(x - x_*) + \mathrm{O}(\|x - x_*\|^\beta), \quad \|x - x_*\| \to 0, \quad (2.96)$$

is valid, where $H(\cdot)$ is a positive homogeneous function on $\mathbb{R}^d \backslash \{0\}$ of order $\beta > 0$ (for $H(\cdot)$ the relation $H(\lambda z) = \lambda^\beta H(z)$ holds for all $\lambda > 0$ and $z \in \mathbb{R}^d$) and function $w : \mathbb{R} \to \mathbb{R}$ is positive and continuous. Then the conditions of Theorem 2.2 for the c.d.f. (2.13) *are fulfiled and the value of the tail index α is equal to $\alpha = d/\beta$.*

Proof of the theorem is given in Sect. 2.7.

The main condition in Theorem 2.3 is (2.96) which characterizes the behaviour of the objective function $f(\cdot)$ in the neighbourhood of its global minimizer. Let us consider two important particular cases of (2.96).

First, let us assume that $f(\cdot)$ is twice continuously differentiable in the vicinity of x_*, $\nabla f(x_*) = 0$ (here $\nabla f(x_*)$ is the gradient of $f(\cdot)$ in x_*) and the Hessian $\nabla^2 f(x_*)$ of $f(\cdot)$ at x_* is non-degenerate. In this case, we can take

$$w(\cdot) = 1, \quad H(z) = -z'[\nabla^2 f(x_*)]z\,,$$

which implies $\beta = 2$ and $\alpha = d/2$.

Assume now that all components of $\nabla f(x_*)$ are finite and non-zero which often happens if the global minimum of $f(\cdot)$ is achieved at the boundary of A. Then we may take $H(z) = z'\nabla f(x_*)$, $w(\cdot) = 1$; this gives $\beta = 1$ and $\alpha = d$.

Consider now two extensions of the basic result.

The following statement demonstrates that if we can assume that the conditions of Theorem 2.2 are met for the c.d.f. (2.13) with some α, then the value of α itself can be determined from assumptions that are weaker than those of Theorem 2.3.

Theorem 2.4. *Assume that the global minimizer x_* of $f(\cdot)$ is unique and Conditions C1–C4, C8 and C9 of Sect. 2.1.1 are met. Assume, in addition, that the conditions of Theorem 2.2 are met for some $\alpha > 0$ and there exist positive numbers ε_0, c_3 and c_4 such that for all $x \in B(x_*, \varepsilon_0)$ the inequality*

$$c_3||x_* - x||^\beta \le f(x) - m \le c_4||x_* - x||^\beta$$

is valid. Then $\alpha = d/\beta$.

Proof of the theorem is given in Sect. 2.7.

The next assertion relaxes the uniqueness requirement for the global minimizer.

Theorem 2.5. *Assume that Conditions C1–C4, C7, C8 and C9 of Sect. 2.1.1 along with Condition C10 of Sect. 2.2.1 are met. Let the global minimum $m = \min f$ of $f(\cdot)$ be attained at points $x_*^{(i)}$ $(i = 1, \dots, l)$ in whose vicinities the tail indexes α_i can be determined. Then the conditions of Theorem 2.2 for the c.d.f. (2.13) are fulfilled and the value of the tail index α is $\alpha = \min\{\alpha_1, \dots, \alpha_l\}$.*

Proof of the theorem is given in Sect. 2.7.

2.6 Some Algorithmic and Methodological Aspects

2.6.1 Using Statistical Inference in Global Random Search

In this section, we consider different ways of using statistical inference procedures in global random search algorithms, discuss the so-called branch and probability bound methods and review the statistical inference procedures in the method of random multistart.

General considerations

Many global random search algorithms consist of several iterations so that at the i-th iteration a particular probability distribution $P = P_i$ is generated to obtain the points where $f(\cdot)$ is to be evaluated – see Algorithm 2.2 of Sect. 2.1.2 and a number of methods in Sect. 3.5. At each iteration of these algorithms and for various subsets Z of A with $P(Z) > 0$, we have independent samples of points which belong to Z and are distributed according to the probability measure P_Z (for a given $Z \in \mathcal{B}$, the measure P_Z is defined as $P_Z(U) = P(U \cap Z)/P(Z)$, $U \subseteq A$), along with the values of the objective function $f(\cdot)$ at these points. For given Z, these values of $f(\cdot)$ form an independent sample from the distribution with the c.d.f.

$$F_Z(t) = P_Z\{x \in Z : f(x) \le t\}$$

and the lower bound

$$m_Z = \inf_{z \in Z} f(z)\,.$$

To guarantee that m_Z is indeed the lower bound of $F_Z(\cdot)$ it is sufficient to assume Conditions C1–C3 and C4′ of Sect. 2.1.1 for the set Z and Condition C10 of Sect. 2.2.1 for the measure P.

To decide whether it is worthwhile to place new points in Z we can draw statistical inferences concerning the parameter m_Z and the behaviour of the c.d.f. $F_Z(t)$ in the vicinity of m_Z. Since statistical procedures can be constructed for all sets Z and at various iterations of the algorithms in a similar manner, we can extend all the results of Sect. 2.4 and 2.5 formulated in the case $Z = A$ to the case of a generic $Z \subseteq A$. A wide class of global random search methods based on the statistical inference procedures developed in previous sections, is considered below.

More broadly, the statistical inference procedures of Sect. 2.4 and 2.5 aim to learn about the distance between the current record y_{on} and the unknown target $m = \min f$ and hence can be used for devising various stopping rules in any global random search algorithm presented in the form of Algorithm 2.2 of Sect. 2.1.2. For example, the estimators $\hat{m}$ of m and the confidence intervals for m can be used to define the following stopping rule: if $\hat{m}$ is close enough to the best value of $f(\cdot)$ obtained so far (alternatively, if the confidence interval is small enough), then the algorithm terminates.

The distributions for the new points in algorithms of this kind can differ from the uniform as these distributions are constantly changing. The corresponding algorithms, where the number of iterations is small but the number of points at each iteration is large, constitute a wide class of the so-called genetic random search algorithms, see Sect. 3.5; these algorithms are extremely popular in practice. As the number of points at each iteration is typically large, all the statistical procedures developed above can be used exactly as they are presented. The differences between these algorithms and the branch

and probability bound methods considered below, are:
(a) the subregions are not removed from A; instead, the distributions P_i are adapted; and
(b) the function values that were used in previous iterations cannot be used in subsequent iterations: indeed, the use of them would introduce dependence into the sample $\{f(x_i)\}$; this dependence would be difficult to handle.

Furthermore, the assumption of the independence of points $x_i^{(j)}$ at iteration j in Algorithm 2.2 of Sect. 2.1.2, which is commonly used in practice (see e.g. Sect. 3.5), can be relaxed to allow some dependence in these points and some of the statistical inference procedures developed above can be suitably modified. In Sect. 3.2 we consider in detail the problem of making statistical inference about m for the case of stratified sampling. We will show that a certain reduction in randomness typically leads to more efficient algorithms; note that improving the efficiency of algorithms by reducing the randomness of points is one of the major areas of interest in the theory of Monte-Carlo methods.

Branch and probability bound methods

Branch and bound optimisation methods are widely known. To put it briefly, they consist of several iterations, each including the following stages:

(i) branching the optimisation set into a tree of subsets (more generally, decomposing the original problem into subproblems),
(ii) making decisions about the prospectiveness of the subsets for further search, and
(iii) selecting the subsets that are recognized as prospective for further branching.

To make a decision at stage (ii) prior information about $f(\cdot)$ and values of $f(\cdot)$ at some points in A are used, deterministic lower bounds concerning the minimal values of $f(\cdot)$ on the subsets of A are constructed, and those subsets $Z \subset A$ are rejected (considered as non-prospective for further search) for which the lower bound for $m_Z = \inf_{x \in Z} f(x)$ exceeds an upper bound $\hat{m}$ for $m = \min f$; the minimum among all evaluated values of $f(\cdot)$ in A is a natural upper bound $\hat{m}$ for m. A general recommendation for improving this upper bound is to use a local descent algorithm, starting at the new record point, each time we obtain such a point.

Let us consider a version of the branch and bound technique, which we call 'branch and probability bound'; see [272] and Sect. 4.3 in [273] for a detailed description of this technique and results of numerical experiments. In the branch and probability bound methods, an independent sample from the uniform distribution in the current search region is generated at each iteration and the statistical procedures described in Sect. 2.4.2 for testing the hypothesis $H_0 : m_Z \leq \hat{m}$ are applied to make a decision concerning the prospectiveness of sets Z at stage (ii). Rejection of the hypothesis H_0

corresponds to the decision that the global minimum m can not be reached in Z. Naturally, such a rejection may be false. This may result in losing the global minimizer. An attractive feature of the branch and probability bound algorithms is that the asymptotic level for the probability of false rejection can be controlled.

The stages (i) and (iii) above can be implemented in exactly the same fashion as in the classical branch and bound methods. When the structure of A is not too complicated, the following technique has been proven to be convenient and efficient.

Let A_j be a search region at iteration j, $j \geq 1$ (so that $A_1 = A$). At iteration j, in the search region A_j we first isolate a subregion Z_{j1} with centre at the point corresponding to the record value of $f(\cdot)$. The point corresponding to the record value of $f(\cdot)$ over $A_j \backslash Z_{j1}$ is the centre of a subregion Z_{j2}. Similar subregions Z_{ji} $(i = 1, \ldots, I)$ are isolated until either A_j is covered or the hypothesis that the global minimum can occur in the residual set $A_j / \cup_{i=1}^{I} Z_{ji}$ is rejected (the hypothesis can be verified by the procedure described in Sect. 2.4.2). The search region A_{j+1} in the next $(j+1)$–th iteration is naturally either $Z^{(j+1)} = \cup_{i=1}^{I} Z_{ji}$, a hyperrectangle covering $Z^{(j+1)}$, or a union of disjoint hyperrectangles covering $Z^{(j+1)}$. In the multidimensional case the last two ways produce more computationally convenient versions of the branch and probability bound method than the first one.

As the value of the minimum of $f(\cdot)$ over these kind of subsets can often be expected to be attained at the boundary, where all the components of the gradient of the objective function are expected to be non-zero (assuming the objective function is differentiable), the results of Sect. 2.5.3 imply that $\alpha = d$ can be used as the value of the tail index α. For some subregions Z, the value d overestimates the true value of α, but this only affects the power of the test of Sect. 2.4.2 applied for testing the hypothesis $H_0 : m_Z \leq \hat{m}$. On the other hand, the fact that we do not have to estimate α significantly simplifies the problem of making statistical inferences about the minimum of $f(\cdot)$ over the subregions Z_{ji}.

Note also that at subsequent iterations all previously used points can still be used, since they follow the uniform distribution at the reduced regions.

The branch and probability bound methods are rather simple and can easily be realized as computer codes. They are both practically efficient for small or moderate values of d (say, $d < 10$) and theoretically justified in the sense that under general assumptions concerning $f(\cdot)$, they asymptotically converge with a given probability, which can be chosen close to 1. However, as d (and therefore α) increases, the efficiency of the statistical procedures of Sect. 2.4 deteriorates. Therefore, for large d the branch and probability methods are both hard to implement (this is the case for the whole family of branch and bound methods) and their efficiency is poor. As a consequence of this, the use of the branch and probability methods for large dimensions is not recommended.

2.6.2 Statistical Inference in Random Multistart

Random multistart is a global optimization method consisting of several local searches starting at random initial points. In its original form, this method is inefficient as it typically wastes much effort on repeated ascents. However, some of its modifications, such as those using cluster analysis procedures to prevent repeated ascents to the same local extrema, can be quite efficient. These modifications are widely used and have been discussed in a number of papers including [22, 23, 147, 198, 199, 210].

This section mainly follows the paper [278] by R. Zieliński and describes several statistical procedures that can be used to increase the efficiency of the simplest random multistart and some of its modifications. A number of publications have appeared developing the ideas discussed in this section, mostly using the Bayesian inference, see e.g. [16, 17, 20, 21, 114, 261]. However, all the main ideas of the approach were contained in the original paper [278] and there has not been any significant progress in the area since 1981, the time of the publication of [278].

Notation

Let $A \subset \mathbb{R}^d$ satisfy the conditions C1, C2 and C3 of Sect. 2.1.1, $f(\cdot)$ be a continuous function on A with a finite but unknown number l of local minimizers $x_*^{(1)}, \ldots, x_*^{(l)}$, P be a probability measure on A and $\mathcal{A}$ be a local descent algorithm. We shall write $\mathcal{A}(x) = x_*^{(i)}$ for $x \in A$, if when starting at the initial point x the algorithm $\mathcal{A}$ leads to the local minimizer $x_*^{(i)}$.

Set $\theta_i = P(A_i^*)$ for $i = 1, \ldots, l$, where $A_i^* = \{x \in A : \mathcal{A}(x) = x_*^{(i)}\}$ is the region of attraction of $x_*^{(i)}$ (note that A_i^* may depend on the chosen algorithm of local descent). It is clear that $\theta_i > 0$ for $i = 1, \ldots, l$ and $\sum_{i=1}^{l} \theta_i = 1$.

The method of random multistart is constructed as follows. An independent sample $X_n = \{x_1, \ldots, x_n\}$ from the distribution P is generated and a local optimization algorithm $\mathcal{A}$ is sequentially applied at each $x_j \in X_n$. Let n_i be the number of points x_j belonging to A_i^* (that is, n_i is the number of descents to $x_*^{(i)}$ from the points $x_1, \ldots, x_n$). According to the definition, $n_i \geq 0$ $(i = 1, \ldots, l)$, $\sum_{i=1}^{l} n_i = n$, and the random vector $(n_1, \ldots, n_l)$ follows the multinomial distribution

$$\Pr\{n_1 = n_1, \ldots, n_l = n_l\} = \binom{n}{n_1, \ldots, n_l} \theta_1^{n_1} \ldots \theta_l^{n_l},$$

where

$$\sum_{i=1}^{l} n_i = n, \quad \binom{n}{n_1, \ldots, n_l} = \frac{n!}{n_1! \ldots n_l!}, \quad n_i \geq 0 \quad (i = 1, \ldots, l).$$

We consider the problem of drawing statistical inferences concerning the number of local minimizers l, the parameter vector $\theta = (\theta_1, \ldots, \theta_l)$, and the

number n_* of trials that guarantees with a given probability that all local minimizers are found.

If l is known, then the problem is reduced to the standard problem of making statistical inferences about the parameters of a multinomial distribution. This problem is well documented in literature, see Chapt. 35 in [129].

The main difficulty is caused by the fact that l is usually unknown. If an upper bound for l is known, then one can apply standard statistical methods; if an upper bound for l is unknown, the Bayesian approach is a natural alternative. Let us first consider the case where the number of local minimizers is bounded.

Bounded number of local minimizers

Let L be an upper bound for l and $n \geq L$. Then $(n_1/n, \ldots, n_l/n)$ is the standard minimum variance unbiased estimate of θ, where n_i/n are the estimators of θ_i's. Of course, for all n and $l > 1$ it may happen, for some i, that $n_i = 0$ but $\theta_i > 0$. So, the above estimator non-degenerately estimates only the θ_i's for which $n_i > 0$.

Let W be the number of n_i's that are strictly positive. Then for given l and $\theta = (\theta_1, \ldots, \theta_l)$ we have

$$\Pr\{W = w \,|\, \theta\} = \sum_{\substack{n_1+\cdots+n_w=n \\ n_i>0}} \sum_{1 \leq i_1 < \cdots < i_w \leq l} \binom{n}{n_1, \ldots, n_w} \theta_{i_1}^{n_1} \ldots \theta_{i_w}^{n_w}.$$

For instance, the probability that all local descents will lead to a single local minimizer is

$$\Pr\{W = 1 \,|\, \theta\} = \sum_{i=1}^{l} \theta_i^n$$

and the probability that all local minima will be found is

$$\Pr\{W = l \,|\, \theta\} = \sum_{\substack{n_1+\cdots+n_l=n \\ n_i>0}} \sum_{1 \leq i_1 < \cdots < i_w \leq l} \binom{n}{n_1, \ldots, n_l} \theta_{i_1}^{n_1} \ldots \theta_{i_l}^{n_l}. \quad (2.97)$$

The probability (2.97) is small if at least one of the θ_i's is small. On the other hand, for any l and θ we can find n_* such that for any given $\gamma \in (0, 1)$ we will have $\Pr\{W = l \,|\theta\} \geq \gamma$ for all $n \geq n_*$. Finding $n_* = n_*(\gamma, \theta)$ is the problem of finding the (minimal) number of points in A such that the probability that all local minimizers will be found is at least γ.

Set $\delta = \min\{\theta_1, \ldots, \theta_l\} \leq 1/l$ and note that

$$\Pr\{W = l|\theta\} \geq \sum_{n_1+\ldots+n_l=n} \binom{n}{n_1, \ldots, n_l} \delta^n = (\delta l)^n \Pr\{W = l \,|\, (\tfrac{1}{l}, \ldots, \tfrac{1}{l})\}.$$

Hence the problem of finding $n_*(\gamma,\theta)$ is reduced to that of finding $n_*(\gamma,\theta_*)$, where $\theta_* = (l^{-1},\dots,l^{-1})$. The latter is easy to approximate as for large n

$$\Pr\{W = l \,|\, \theta_*\} = l^{-n} \sum_{n_1+\dots+n_l=n} \binom{n}{n_1,\dots,n_l} =$$

$$= \sum_{i=0}^{l} (-1)^i \binom{l}{i} (1-i/l)^n \sim \exp\{-l\exp\{-n/l\}\}, \qquad n\to\infty.$$

By solving the equation $\exp(-l\exp(-n/l)) = \gamma$ with respect to n we obtain the approximation

$$n_*(\gamma,\theta_*) \simeq l\ln l + l\ln(-\ln\gamma)\,. \tag{2.98}$$

This approximation is rather good even for small l and n; see Fig. 2.9, where the exact values of $n_*(\gamma,\theta_*)$ and the approximation (2.98) are given for $\gamma = 0.9$ and $l \le 20$.

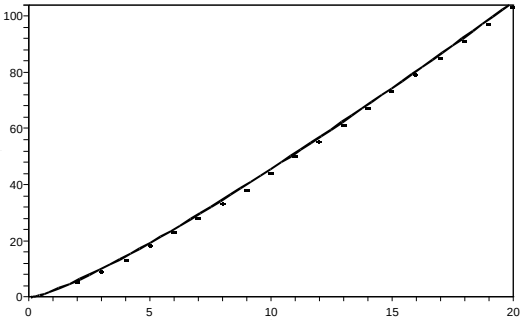

Fig. 2.9. The exact values of $n_*(\gamma,\theta_*)$ (dots) and the approximation $l\ln l + l\ln(-\ln\gamma)$ (solid line) for $\gamma = 0.9$ and $l = 2,\dots,20$.

Bayesian approach

Let α_j $(j = 1,2,\dots)$ be the prior probabilities of events that the number l of local minimizers of $f(\cdot)$ is equal to j and let $\lambda_j(d\theta^j)$ be the conditional prior measures for the parameter vector $\theta^j = (\theta_1,\dots,\theta_j)$ under the condition $l{=}j$. We shall assume that the measures $\lambda_j(d\theta^j)$ are uniform on the simplices

$$\Theta_j = \Big\{\theta^j = (\theta_1,\dots,\theta_j):\ \theta_i > 0,\ \sum_{i=1}^{j}\theta_i = 1\Big\}.$$

Thus, the parameter set Θ, on which the vector of unknown parameters $\theta = (\theta_1, \ldots, \theta_l)$ can take its values, has the form $\Theta = \cup_{j=1}^{\infty} \Theta_j$ and the prior measure $\lambda(d\theta)$ on Θ for θ equals

$$\lambda(d\theta) = \sum_{j=1}^{\infty} \alpha_j \lambda_j(d\theta^j). \tag{2.99}$$

It is natural to assume that λ is a probability measure.

Let $d = d(n_1, \ldots, n_W)$ be an estimate of l. The estimate

$$d_* = \arg\min_d \int_{\Theta} \mathrm{E}_\theta \, \mathbf{1}_{[n_1,\ldots,n_W : d \neq l]} \lambda(d\theta)$$

is the optimal Bayesian estimate of l; it can be simplified to

$$d_* = \arg\max_{j \geq W} \alpha_j Q(j, W, n), \tag{2.100}$$

where

$$Q(j, W, n) = \binom{j}{W} \Gamma(j)/\Gamma(n + j).$$

Using a quadratic loss function, the optimal Bayesian estimate for the total P-measure of the domains of attraction of the hidden $l - W$ local minimizers (i.e. of the sum of the θ_i's corresponding to the undiscovered minimizers) is given by

$$\sum_{j=W}^{\infty} \frac{j - W}{n + j} \alpha_j Q(j, W, n) \Bigg/ \sum_{j=W}^{\infty} \alpha_j Q(j, W, n)\,.$$

The optimal Bayesian procedure for testing the hypothesis $H_0 : l = W$ under the alternative $H_1 : l > W$ is constructed in a similar way. According to this procedure, H_0 is accepted if

$$c_{01} \sum_{j=W+1}^{\infty} \alpha_j Q(j, W, n) \leq c_{10} \alpha_W \Gamma(W)/\Gamma(n + W),$$

otherwise H_0 is rejected. Here c_{01} is the loss arising after accepting H_0 in the case of H_1's validity and c_{10} is the loss due to accepting the hypothesis H_1 when it is false.

2.6.3 Sampling on Surfaces

Application of any random search algorithm to an optimization problem where the feasible region A is defined by the equality-type constraints requires sampling from probability distributions on the surface defined by these

constraints. We show how to reduce the problem of distribution sampling on a surface to the problem of distribution sampling on a subset of $\mathbb{R}^k$ of positive volume (the latter problem is potentially simpler).

Let $X \subset \mathbb{R}^k$ with $0 < \mathrm{vol}(X) < \infty$ and Φ be a continuously differentiable mapping of X into $\mathbb{R}^d$ with $d \geq k$. Using the notation $x = (x_1, \ldots, x_k)$, $z = (z_1, \ldots, z_d)$ and $\Phi = (\varphi_1, \ldots, \varphi_d)$ we can write

$$\begin{cases} z_1 = & \varphi_1(x_1, \ldots, x_k) \\ \vdots & \vdots \\ z_d = & \varphi_d(x_1, \ldots, x_k) \end{cases}$$

simply as $z = \Phi(x)$. For $d > k$, the set

$$A = \Phi(X) = \{z = \Phi(x), x \in X\}$$

is a k-dimensional surface in $\mathbb{R}^d$.

For any $x \in X$ we define

$$d_{ij}(x) = \sum_{l=1}^{d} \frac{\partial \varphi_l(x)}{\partial x_i} \frac{\partial \varphi_l(x)}{\partial x_j} \qquad (i, j = 1, \ldots, k)$$

and

$$D(x) = \sqrt{\det \| d_{ij}(x) \|_{i,j=1}^{k}} \, .$$

The matrix $\| d_{ij}(x) \|_{i,j=1}^{k}$ is non-negative definite for all $x \in X$ so that its determinant is always non-negative.

If $d = k$ then $D(x) = |\partial \Phi / \partial x|$ is the Jacobian of the transformation Φ. Another important particular case is where $d = k+1$ and $\varphi_j(x) = x_j$ $(j = 1, \ldots, k)$; in this case we have

$$D(x) = \left[1 + \sum_{i=1}^{k} \left(\frac{\partial \varphi_{k+1}(x)}{\partial x_i} \right)^2 \right]^{\frac{1}{2}} .$$

Let ds denote the surface measure on the surface $A = \Phi(X)$. As follows from §10, Chapt. 4 in [211], for any Borel-measurable function p defined on A and any $B \subseteq A$ of the form $B = \Phi(U)$, where U is a measurable subset of X, we have

$$\int_B p(s) ds = \int_{\Phi^{-1}(B)} p(\Phi(x)) D(x) dx.$$

Therefore, for any measurable non-negative function $p(\cdot)$ defined on A and satisfying the condition

$$\int_X p(\Phi(x))D(x)dx = 1,$$

the probability measure with density

$$p(\Phi(x))D(x), \quad x \in X$$

induces the probability distribution $p(s)ds$ on the surface $A = \Phi(X)$. In the important particular case where

$$c = \int_X D(x)dx < \infty,$$

the probability distribution with density

$$p_0(x) = \frac{1}{c}D(x), \quad x \in X\,,$$

induces the uniform distribution ds/c on the surface A.

Thus, the problem of distribution sampling on the surface A is being reduced to the problem of distribution sampling on the set $X \subset \mathbb{R}^k$ with $\mathrm{vol}(X) > 0$. In order to obtain a realization ξ of a random vector in $\mathbb{R}^d$ with distribution $p(s)ds$ on A, it is enough to obtain a realization ζ of a random vector in $X \subset \mathbb{R}^k$ with density $p(\Phi(x))D(x)$ and compute $\xi = \Phi(\zeta)$.

This general methodology was applied in [273], Sect. 6.1, to construct distribution sampling algorithms on various surfaces including ellipsoids, hyperboloids and cones.

2.7 Proofs

Proof of Theorem 2.1.

Fix $\delta > 0$ and find some $\varepsilon > 0$ such that $B(x_*, \varepsilon) \subset W(\delta)$; this is possible as $f(\cdot)$ is continuous in the vicinity of x_*. Define the sequence of independent random variables $\{\zeta_j\}$ on the two-point set $\{0, 1\}$ so that

$$\Pr\{\zeta_j = 1\} = 1 - \Pr\{\zeta_j = 0\} = q_j(\varepsilon)$$

where $q_j(\varepsilon)$ is defined in (2.5).

For each j, the probability of the event $x_j \in B(x_*, \varepsilon)$ is larger than or equal to the probability of the event $\zeta_j = 1$. However, the first part of the Borel's 'zero-one law' (see e.g. [226]) implies that if (2.4) holds, then ζ_j infinitely often takes the value 1; this yields the assertion of the theorem. □

Proof of Proposition 2.1.

Setting $w = (1+1/u)^{\alpha}-1$ and using the fact that the joint asymptotic density of

$$\left(\frac{\eta_{1,n}-m}{\kappa_n-m}, \frac{\eta_{k,n}-m}{\kappa_n-m}\right)$$

coincides with the joint density of the random vector $(\nu_1^{1/\alpha}, (\nu_1+\ldots+\nu_k)^{1/\alpha})$, we obtain

$$\Pr\{D_{n,k}\le u\} \sim \Pr\left\{\frac{\nu_1^{1/\alpha}}{(\nu_1+\cdots+\nu_k)^{1/\alpha}-\nu_1^{1/\alpha}}\le u\right\} = \Pr\left\{\frac{\nu_1+\cdots+\nu_k}{\nu_1}\ge w\right\}$$
$$= \frac{1}{(k-2)!}\int_0^{\infty}\left[\int_{wy}^{\infty}\exp\{-x-y\}\cdot x^{k-2}dx\right]dy = 1-\left(\frac{w}{w+1}\right)^{k-1}.$$

□

Proof of Proposition 2.2.

The formula (2.53) for the asymptotic moments follows from the fact that $(\eta_{k,n}-m)/(\kappa_n-m)$ converges in distribution (as $n\to\infty$) to the random variable with density (2.48); computing the β-th moment of this distribution with this density immediately gives (2.53).

Proof of (2.54) is similar but more technical. Assume that $1\le j<k\le n$; the case $j=k$ is covered in (2.53) with $\beta=2$.

Using the fact that the sequence of random vectors (2.49) asymptotically, as $n\to\infty$, has the same density as the vector (2.51), we deduce that the random vector

$$\left(\frac{\eta_{j,n}-m}{\kappa_n-m}, \frac{\eta_{k,n}-m}{\kappa_n-m}\right)$$

asymptotically has the same density as the vector

$$\left(\zeta^{1/\alpha}, (\xi+\zeta)^{1/\alpha}\right),$$

where random variables ξ and ζ are independent and have Gamma-distributions with densities

$$p_{\zeta}(x) = \frac{1}{\Gamma(j)}x^{j-1}e^{-x} \quad \text{and} \quad p_{\xi}(x) = \frac{1}{\Gamma(k-j)}x^{k-j-1}e^{-x} \quad (x>0),$$

respectively. Therefore, as $n\to\infty$, we have

$$\frac{1}{(\kappa_n-m)^2}\mathrm{E}(\eta_{j,n}-m)(\eta_{k,n}-m) \to \mathrm{E}\zeta^{1/\alpha}(\xi+\zeta)^{1/\alpha}$$

$$= \frac{1}{\Gamma(j)\Gamma(k-j)} \int_0^\infty \int_0^\infty z^{1/\alpha}(x+z)^{1/\alpha} z^{j-1} e^{-z} x^{k-j-1} e^{-x} dx dz$$

$$= \frac{1}{\Gamma(j)\Gamma(k-j)} \int_0^\infty \int_0^\infty z^{j-1+2/\alpha}(1+x/z)^{1/\alpha} e^{-z} x^{k-j-1} e^{-x} dx dz$$

$$= \frac{1}{\Gamma(j)\Gamma(k-j)} \int_0^\infty \int_0^\infty (1+t)^{1/\alpha} t^{k-j-1} z^{k-1+2/\alpha} e^{-z(t+1)} dt dz$$

$$= \frac{1}{\Gamma(j)\Gamma(k-j)} \int_0^\infty \frac{t^{k-j-1}}{(t+1)^{k+1/\alpha}} dt \cdot \int_0^\infty u^{k-1+2/\alpha} e^{-u} du = \lambda_{kj} \,.$$

In the process of integration, we have introduced the new variables $t = x/z$ and $u = z(t+1)$; additionally, we have used the formulae

$$\int_0^\infty u^{k-1+2/\alpha} e^{-u} du = \Gamma(k+\frac{2}{\alpha}) \;\text{ and }\; \int_0^\infty \frac{t^{k-j-1}}{(t+1)^{k+1/\alpha}} dt = \frac{\Gamma(k-j)\Gamma(j+1/\alpha)}{\Gamma(k+1/\alpha)} \,.$$

□

Proof of Proposition 2.3.

According to Proposition 2.1, the sequence of random variables

$$(\eta_{1,n} - m)/(\eta_{k,n} - \eta_{1,n})$$

converges in distribution to the random variable with the c.d.f.

$$F_k(u) = 1 - \left(1 - \left(1 - \frac{1}{1+u}\right)^\alpha\right)^k$$

(note that $r_{k,\delta}$ is the (1–δ)-quantile of this c.d.f.). This implies that as $n \to \infty$, the confidence level of the interval (2.93) can be represented as

$$\Pr\{m \in I'\} = \Pr\left\{\frac{\eta_{1,n} - m}{\eta_{k,n} - \eta_{1,n}} \le r'_{k,\delta}\right\}$$

$$\sim 1 - \left(1 - \left(\frac{1}{1+r'_{k,\delta}}\right)^\alpha\right)^k = 1 - \left(1 - \left(1 - \delta^{1/k}\right)^{\alpha/\vartheta}\right)^k .$$

□