ulm university universität **uulm**

Universität Ulm
Fakultät für Mathematik und Wirtschaftswissenschaften

Solving Markov Decision Processes by Simulation: a Model-Based Approach

Diplomarbeit

in Wirtschaftsmathematik

vorgelegt von
Kristina Steih
Juli 2008

Gutachter

Prof. Dr. Ulrich Rieder
Prof. Dr. Dieter Kalin

Contents

Symbols and Abbreviations

CE	Cross-Entropy
MDP	Markov Decision Process
MRAS	Model Reference Adaptive Search
A	action space in MDPs
$\mathcal{D}(\cdot,\cdot)$	Kullback-Leibler divergence
$\mathbb{E}_\theta$	expectation with respect to $f(\cdot\,;\theta)$
$\mathbb{E}_g$	expectation with respect to g
$f(\cdot\,;\theta)$	sample density (CE and MRAS)
$\widetilde{f}(\cdot\,;\theta)$	mixture of current and initial sample density (MRAS)
$F_{(1)},\ldots,F_{(n)}$	order statistic of $F_1,\ldots,F_n$
$\widehat{F}$	estimator for F
g	reference distribution in MRAS
g^*	reference distribution in CE (optimal importance sampling density)
$\widehat{g}$	(random) continuous idealized reference distribution in MRAS
$\widetilde{g}$	(random) discrete sample reference distribution in MRAS
M	observation size
N	sample size
R_π^N	(random) total discounted reward over N periods under policy π
S	state space in MDPs, monotone positive function in MRAS
T	sufficient statistic in an exponential family
v	smoothing factor in parameter update
$V_{N,\pi}$	expected discounted reward over N periods under policy π
$\widehat{V}_\infty^{\pi,k}$	estimated expected discounted reward under policy π obtained in iteration k
$X(\rho,N)$	sample corresponding to sample $(1-\rho)$-quantile given N samples (MRAS)
X_k°	sample corresponding to threshold in iteration k (MRAS)
α	increase rate of observation size
γ	$(1-\rho)$-quantile (CE and MRAS)
$\widehat{\gamma}$	sample $(1-\rho)$-quantile (CE and MRAS)
$\widetilde{\gamma}$	threshold level (MRAS)
ε	minimum increase in threshold level (MRAS, Adapted CE)
λ	mixture parameter of $\widetilde{f}$

Λ_k	set of samples drawn in iteration k
ν	Lebesgue (or counting) measure
θ	parameter of parameterized distribution f
$\widehat{\theta}$	estimated parameter (CE and MRAS)
Θ	parameter space of parameterized family $\{f(\cdot\,;\theta)\}$
π	policy, decision rule
ρ	quantile parameter
$\mathbb{1}_{\{A\}}$	indicator function of event A
$\widetilde{\mathbb{1}}_{\{A\}}$	modified indicator function allowing for noise (MRAS)

Introduction

Markov Decision Problems (MDPs) have proven their ability as a powerful tool for sequential decision making over the past decades and are widely used in various areas such as economics, telecommunication and robotics to solve real-world problems. Standard solution methods are usually based on dynamic programming approaches such as value iteration and policy iteration. However, these techniques assume a complete knowledge of the model, that is for example an explicit representation of the transition probabilities and a fully specified reward function.
Sometimes this is not the case and only samples generated by simulation runs of the system are available. Then one is faced with a so-called learning problem. Additionally one often resorts to simulation if exact calculations are too complex or time-comsuming but sample paths of the given problem can easily be generated. Many solution approaches are still applicable if the correct values are replaced by approximations, obtained for example through averaging over several simulation runs in Monte-Carlo simulation.
Whereas most reinforcement learning methods are concerned with efficiently testing single actions when in a certain state, we consider in this work model-based algorithms that directly search the space of complete policies. Unlike for example genetic algorithms that directly manipulate sets of solution candidates called populations, the methods discussed here are indeed model-based insofar as they work with probability distributions models on the solution space. This is similar to the idea pursued by the class of Estimation of Distribution Algorithms (EDA, see [ZM04]), only that we differentiate between the "ideal" distribution and a "practical" distribution to facilitate the sampling and updating procedures.
Our main focus is the Model Reference Adaptive Search algorithm (MRAS) introduced by Hu, Fu and Marcus in [CFHM07], [HFM07] and [HFM] in its version for stochastic optimization (called MRAS_2). Originally developed as a global optimization method, the authors stress its ability to solve Markov Decision Processes with continuous action and finite state spaces and compare it with Simulated Annealing to prove its consistently good performance. Only few algorithms directly deal with uncountable action spaces and do not use discretization - due to the curse of dimentionality more thought has been invested in the reduction of the state space. Hence an algorithm handling a continuous action space with a reasonably-sized finite state space may be quite interesting. However, it is precisely in this setting that we cannot reproduce the published results. Even more, we think that there are inherent structural problems complicating the implementation of the Model Reference Adaptive Search and disturbing its behaviour.

In addition, even though motivated by a different background, the MRAS is strikingly similar in concept and appearance to the Cross-Entropy method (CE) proposed by Rubinstein [RK04], [dBKMR05] in 1999 with adaptions suggested by Homem-de-Mello in [dM07]. The Cross-Entropy algorithm was designed (and is still used) for rare-event estimation, but has successfully been applied to optimization problems with a focus on combinatorial optimization (see e.g. [RK04], [CJK07]). Other applications include continuous deterministic optimization [KPR06], [Liu04], parameter optimization [MMS05], [Dam07] and the calculation of stochastic shortest paths [MRG03]. To our knowledge it has not as yet been applied to general Markov Decision Processes.
The similarities of CE and MRAS suggest a comparative study to explore their respective strengths. We will limit ourselves to MDPs with a focus on uncountable action spaces. This allows us to compare the Model Reference Adaptive Search to a conceptually related algorithm and at the same time to study the Cross-Entropy algorithm in the as yet not keenly explored context of continuous stochastic optimization.

The outline of this work is as follows:
Chapter 1 is a short introduction of some underlying notions and concepts needed in later sections. The basic ideas of Markov Decision Processes, Exponential Families, the Kullback-Leibler distance and the generation of random variables are presented and the corresponding notations are established.
In Chapter 2, the Cross-Entropy method is introduced. We discuss its background and motivation as well as its use as a global optimization algorithm, present some important modifications and the application of this algorithm to stochastic optimization.
The Model-Reference Adaptive Search is presented in Chapter 3. The successive adaptations from the basic idea to the stochastic optimization algorithm are explained, we highlight similarities and differences in comparison with the CE and illustrate the application to Markov Decision Processes of both algorithms.
In Chapter 4, we discuss convergence properties of the stochastic MRAS algorithm. It can be shown that if the model distribution belongs to an exponential family with sufficient statistic $T(x)$, then $\mathbb{E}_k[T(X)] \to T(x^*)$ as $k \to \infty$ (see Theorem 4.9). However, there are multiple assumptions required for this result which we discuss in detail, as well as the several corrections and clarifications of the original proof found in [CFHM07].
Our numerical experiments are described in Chapter 5. For different Markov Decision problems, we implement and test both CE and MRAS with various modifications. We show that the results concerning the behaviour of the Model Reference Adptive Search in this context as published in [CFHM07] are too positive and give reasons based on our observations. We further observe that the Cross-Entropy algorithm in all its simplicity compared to the MRAS performs in all examples at least as well, usually even better. A particularity of both algorithms, a consistent underestimation of the value function, is brought to attention and explained.
Finally, a detailed conclusion and discussion of the theoretical and practical results and observations is given in Chapter 6.

Chapter 1

General Notions

In this chapter, we will shortly introduce some basic ideas and definitions and establish the corresponding notation. Here and in subsequent chapters, ν denotes the Lebesgue measure (in discrete settings the counting measure).

1.1 Markov Decision Processes

Markov Decision Processes are an important tool for modeling and solving sequential decision making problems. There are several equivalent definitions and notations, we use the formulation from [Rie04].

Definition 1 (**Markov Decision Process**)**.** A *Markov Decision Process* (or MDP) is defined by the tuple $(S, A, D, p, r, V_0, \beta)$, where the state space S is assumed to be countable, A is the action space, $D \subseteq S \times A$ the space of admissible actions, $p : D \times S \to [0, 1]$ the transition law with $p(i, a, j)$ denoting the probability of reaching state j when choosing action a in state i, $r : D \to \mathbb{R}$ the (bounded) reward function, V_0 the terminal reward (equal to zero in infinite horizon problems) and $\beta > 0$ the discount factor.

Furthermore, a sequence $\pi = (f_0, f_1, ...f_{N-1})$ with decision rules $f : S \to A, f(i) \in D(i)\ \forall\, i$ is called a *policy*. In the case of infinite MDPs the existence of an optimal policy implies the existence of an optimal *stationary* policy $\pi^* = (f, f, ...)$.

For a fixed policy π and a given starting point i, the underlying probability space is given by $(\Omega, \mathbb{P}_{\pi i})$, where

$$\Omega := \{\omega = (i_0, i_1, \ldots, i_N)\} = S^{N+1} \text{ and}$$

$$\mathbb{P}_{\pi i}(\{\omega\}) := \delta_{i,i_0} \cdot \prod_{k=0}^{N-1} p(i_k, f_k(i_k), i_{k+1}).$$

Define $X_n : \Omega \to S$, $X_n(\omega) := i_n$. Then the *total discounted reward* is the random variable

$$R_\pi^N(\omega) := \sum_{k=0}^{N-1} \beta^k r(X_n(\omega), f(X_n(\omega)) + \beta^N V_0(X_N(\omega))$$

and

$$V_{N,\pi}(i) := \mathbb{E}_{\pi_i}(R_\pi^N)$$

is called the *expected discounted reward*. Moreover, we denote by

$$V_N(i) := \sup_\pi V_{N,\pi}(i)$$

the *maximal discounted reward* and call π^* an *optimal policy* if $V_N = V_{N,\pi^*}$.

1.2 Simulation and Random Numbers

With the emergence and further development of powerful computers in the last decades, simulation has become a widely used tool for analysis and optimization of problems in many different areas of application. It usually is employed if the underlying system is either too complex to be treated analytically or not completely known.

Random Variate Generation

To correctly and efficiently simulate a random process, we need to be able to produce realizations of random variables with a given discrete or continuous distribution F. This is not a trivial task, since computers are not only fundamentally deterministic machines but can also only store discrete numbers due to their finite memory (every number can only be represented by a finite number of bits).
The generation of random variates is usually divided into the generation of *random numbers* that are uniformly $[0,1]$-distributed and the modification of such a sequence $u_1, u_2, \cdots \sim U[0,1]$ to produce outcomes $x_1, x_2, \cdots \sim F$.
For the first part, one usually resorts to generate so-called *pseudo-random numbers*. These are elements of a deterministic sequence of numbers $\{u_n\}$ depending on some seed s_0. The idea is that even though for a fixed random generator the sequence $\{u_n\}$ is completely determined by the seed, the elements of a given sequence are uniformly distributed in that they cannot be distinguished from true realisations of a uniform distribution by statistical tests. However, since they are in fact not random, this will never be the case for all tests. The quality of such a generator can thus be measured by the choice of statistical tests the produced sequences pass. Other criteria are for example the length of period (i.e. the number of elements before the sequence starts to repeat itself), the independence of the seed (sometimes some seeds may produce "less random" sequences than others), the distribution among dimensions (e.g. no 3-dimensional patterns) and the speed with which variates are generated. In our case we will use for random number generation the *Mersenne Twister*, widely used due to its long period and good dimensional equidistribution.
There is a certain number of basic ideas to obtain random variates according to a specific distribution from uniformly distributed random numbers. For a detailed discussion refer e.g. to [Dev86]. Among the most instrumental and universal ones are for example the

- *Inversion method:*
  If the desired F is a continuous distribution and the generalized inverse $F^{-1}(y) := \inf\{x : F(x) = y\}$, $y \in (0,1)$ can be computed explicitly, the following theorem provides a general method to obtain realisations of $X \sim F$:

  Theorem 1.1. *Let F, F^{-1} as above and $U \sim U[0,1]$. Then F is the distribution function of $F^{-1}(U)$.*

  Hence $x = F^{-1}(u)$ is a realization of $X \sim F$ if u is a realization of $U \sim U[0,1]$.

- *Rejection method:*
  If there exists a distribution G such that F is absolutely continuous with respect to G with bounded density p, i.e.

  $$F(x) = \int_{-\infty}^{x} p(z)dG(z) \quad \text{and} \quad p(x) \leq c,\ 0 < c < \infty \quad \forall\, x \in \mathbb{R},$$

  and realizations of $X \sim G$ can easily be generated, the rejection (sometimes also called acceptance-rejection) method can be employed. It is based on the following theorem:

  Theorem 1.2. *Let F, G as above and $\{(X_i, U_i)\}_{i\in\mathbb{N}}$, $(X_i, U_i) \sim G \otimes U[0,1]$, a sequence of independent and identically distributed random vectors. Then the random variable $I := \min\left\{i \geq 1 : U_i \leq \frac{p(X_i)}{c}\right\}$ is geometrically distributed with parameter $\frac{1}{c}$ and the random variable X_I is distributed according to F.*

  Thus one can generate realizations of $X \sim F$ by the following algorithm:

  Algorithm 1.1 (Rejection Method)**.**

  REPEAT *Generate independently $y \sim G$ and $u \sim U[0,1]$*
  UNTIL $u \cdot c \leq p(y)$.
  RETURN y.

  Note that the $\mathbb{E}[I] = c$. Hence one should choose such a G that c is as small as possible to obtain a fast generator.

Moreover, for most common distributions there even exist highly specialized algorithms that take advantage of the concrete form of the distribution to speed up or improve in other respects the random number generation. We will shortly state here the algorithms used for our needs.

Normal Distribution

Realizations of standard normally distributed random variables are usually obtained by the so-called Box-Muller Algorithm based on the following theorem.

Theorem 1.3 (Box-Muller Method)**.**
Let $U_1, U_2 \sim U[0,1]$ *independent. Then*

$$X_1 = \sqrt{-2\ln(U_1)} \cdot \cos(2\pi U_2)$$
$$\textit{and} \quad X_2 = \sqrt{-2\ln(U_1)} \cdot \sin(2\pi U_2)$$

are independent and $\mathcal{N}(0,1)$*-distributed.*

Moreover, $\mathcal{N}(\mu,\sigma)$-distributed random variables can easily be generated from $X \sim \mathcal{N}(0,1)$, since $\sigma X + \mu \sim \mathcal{N}(\mu,\sigma)$.

Exponential Distribution

As one can easily invert the distribution function $F(x) = 1 - e^{-\lambda x}$ of the exponential distribution, this is an example for the application of the inversion method. The inverse is given by

$$F^{-1}(U) = -\frac{1}{\lambda}\log(1-U),$$

but since $1-U$ and U are identically distributed, one can simplify this and set $x = -\frac{1}{\lambda}\log(u)$ for a realization u of $U \sim U[0,1]$.

1.3 Exponential Families

Exponential Families are a large class of families of distributions widely used in statistical theory and applications due to several interesting characteristics. They admit for example natural sufficient statistics whose dimensions are independent of the sample size, form the basis for the so-called generalized linear models and are also important in Bayesian statistics. Exponential families include among others important families such as the normal, binomial, Poisson, gamma and beta distributions. We will usually consider some exponential family as model in our algorithms and especially the convergence results of the model-reference adaptive search (Chapter 4) rely heavily on the general form of these distributions. For more information refer e.g. to [BD97], [Pru89].

Definition 2 (**Exponential Family**)**.** A family of distributions $\{P_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^m$, on a space $\mathcal{X} \subseteq \mathbb{R}^n$ is called a *m-parameter exponential family* if there exist functions $h : \mathbb{R}^n \to \mathbb{R}$, $T : \mathbb{R}^n \to \mathbb{R}^m, \eta : \mathbb{R}^m \to \mathbb{R}^m$ and $B : \mathbb{R}^m \to \mathbb{R}$ such that the density $f(x,\theta)$ has the form

$$f(x,\theta) = h(x)\, \exp\{\eta(\theta)^\top T(x) - B(\theta)\} \qquad \forall\, x \in \mathcal{X}.$$

It this can be reparameterised into

$$f(x,\eta) = h(x)\, \exp\{\eta^\top T(x) - A(\eta)\} \qquad \forall\, x \in \mathcal{X},$$

with $A(\eta) = \ln \int_{\mathbb{R}^n} h(x) \exp\{\eta^\top T(x)\}\nu(\mathrm{d}x)$ finite (for discrete $\mathcal{X}$ the integral is replaced by a sum), then $\{P_\theta, \theta \in \Theta\}$ is called a *canonical m-parameter exponential family.* In that case, the set $S := \{\eta \in \mathbb{R}^m : -\infty < A(\eta) < \infty\}$ is called the *natural parameter space* and $T(X) = (T_1(X), \ldots, T_m(X))$ the *natural sufficient statistic.*

Note that the family of distributions of a random vector $X = (X_1, \ldots, X_N)$ whose components are independent and identically distributed according to some m-parameter exponential family P_θ is itself an m-parameter exponential family, since for its density holds

$$\begin{aligned} f(x,\theta) &= \prod_{i=1}^{N} \Big(h(x_i) \exp\{\eta(\theta)^\top T(x_i) - B(\theta)\} \Big) \\ &= \prod_{i=1}^{N} h(x_i) \cdot \exp\{\eta(\theta)^\top \sum_{i=1}^{N} T(x_i) - N \cdot B(\theta)\} \\ &= \widetilde{h}(x) \exp\{\eta(\theta)^\top \widetilde{T}(x) - \widetilde{B}(\theta)\}. \end{aligned}$$

Example 1.1 (Normal Distribution). Let $P_\theta = \mathcal{N}(\mu, \sigma^2)$. Then $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+$ and

$$\begin{aligned} f(x,\eta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right\} = \exp\left\{ -\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} + \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \right\} \\ &= \exp\left\{ -\frac{1}{2} \left(\frac{x^2}{\sigma^2} - \frac{2x\mu}{\sigma^2} + \frac{\mu^2}{\sigma^2} \right) - \frac{1}{2} \ln(2\pi\sigma^2) \right\} \\ &= \exp\left\{ \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{1}{2} \left(\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2) \right) \right\}. \end{aligned}$$

Hence, $\mathcal{N}(\mu, \sigma^2)$ is a two-parameter canonical exponential family with

$$\eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right), \; T(x) = \left(x, x^2 \right)^\top, \; h(x) = 1 \text{ and}$$

$$\begin{aligned} A(\eta) &= \ln \int_X h(x) \exp\{\eta^\top T(x)\} \mathrm{d}x = \ln \int_X 1 \cdot \exp\left\{ \frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} \right\} \mathrm{d}x \\ &= \ln \int_X \frac{\sqrt{2\pi\sigma^2}}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right\} \cdot \exp\left\{ -\frac{\mu^2}{2\sigma^2} \right\} \mathrm{d}x \\ &= \ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2} \frac{\mu^2}{\sigma^2} = -\frac{1}{2} \left(\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2) \right). \end{aligned}$$

However, a n-dimensional random vector X whose components X_i are independent and normally distributed with parameters (μ_i, σ_i^2) has a distribution of the form

$$\begin{aligned} p(x,\eta) &= \prod_{i=1}^{n} p(x_i, \eta_i) = \exp\left\{ \sum_{i=1}^{n} \left(\frac{\mu_i}{\sigma_i^2} x_i - \frac{1}{2\sigma_i^2} x_i^2 - \frac{1}{2} \left(\frac{\mu_i^2}{\sigma_i^2} + \ln(2\pi\sigma_i^2) \right) \right) \right\} \\ &= \exp\left\{ \sum_{i=1}^{n} \left(\frac{\mu_i}{\sigma_i^2}, -\frac{1}{2\sigma_i^2} \right)^\top \cdot \left(x_i, x_i^2 \right) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{\mu_i^2}{\sigma_i^2} + \ln(2\pi\sigma_i^2) \right) \right\} \\ &= \widetilde{h}(x) \exp\{\widetilde{\eta}^\top \widetilde{T}(x) - \widetilde{A}(\eta)\} \end{aligned}$$

with

$$\begin{aligned} \widetilde{\eta} &= \left(\frac{\mu_1}{\sigma_1^2}, \ldots, \frac{\mu_n}{\sigma_n^2}, -\frac{1}{\sigma_1^2}, \ldots, -\frac{1}{\sigma_n^2} \right) \quad \text{and} \\ \widetilde{T}(x) &= \left(x_1, \ldots, x_n, x_1^2, \ldots, x_n^2 \right). \end{aligned}$$

Since the dimension m of the sufficient statistic and of η depends on n, it is a matter of definition if this can be considered an exponential family.

Example 1.2 (Binomial Distribution). Let $P_\theta = \mathrm{Bin}(n, \theta)$, $\theta \in \Theta = (0, 1) \subset \mathbb{R}$. Then

$$\begin{aligned} p(x, \eta) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &= \binom{n}{x} \exp\left\{ \ln\left(\frac{\theta}{1-\theta}\right) x + n \ln(1-\theta) \right\}. \end{aligned}$$

Hence, $\mathrm{Bin}(n, p)$ is a canonical one-parameter exponential family with

$$\begin{aligned} \eta = \ln\left(\frac{\theta}{1-\theta}\right),\ T(x) = x,\ h(x) = \binom{n}{x} \text{ and} \\ A(\eta) = \ln \sum_{x=1}^{n} h(x) \exp\{\eta^\top T(x)\} = \ln \sum_{x=1}^{n} \binom{n}{x} \exp\left\{ \ln\left(\frac{\theta}{1-\theta}\right) x \right\} \\ = \ln \sum_{x=1}^{n} \binom{n}{x} \left(\frac{\theta}{1-\theta}\right) = \ln \sum_{x=1}^{n} \binom{n}{x} \theta^x (1-\theta)^{-x} \cdot \frac{(1-\theta)^n}{(1-\theta)^n} \\ = \ln(1 \cdot (1-\theta)^n) = -n \ln(1-\theta). \end{aligned}$$

Furthermore we consider a subset of exponential families called *Natural Exponential Families* (NEF). The definition of this class of distributions varies throughout literature [Mor82], [MN83], we will give here the rather strict one from [RK04] needed in later sections. Unfortunately we were unable to find any source confirming the definition of a NEF given in [CFHM07], which rather correponds to a canonical exponential family as defined above.

Definition 3 (**Natural Exponential Family**)**.** A one-parameter exponential family $\{P_\theta, \theta \in \Theta \subseteq R\}$ in canonical form is called *Natural Exponential Family (NEF)* if $T(x) = x$, i.e. if

$$f(x, \eta) = h(x)\ \exp\{\eta^\top x - A(\eta)\} \qquad \forall\, x \in \mathcal{X} \subseteq \mathbb{R}^n, \tag{1.1}$$

with h, A as in Definition 2.

Example 1.3. The binomial distribution is a natural exponential family (see Example 1.2). Other examples are the Poisson distribution and the Gamma distribution.
The normal distribution can be considered as a natural exponential family with $h(x) = \exp\{-\frac{x^2}{s\sigma^2}\} - \ln(2\pi\sigma^2)$ if one is only interested in the mean μ as it is often the case in linear regression. Here we consider it a two-parameter exponential family and hence not a NEF as the variance plays an important role in our context.

1.4 Kullback-Leibler divergence

The *Kullback-Leibler divergence* (also called *relative entropy*, *cross-entropy* or *information gain*) was introduced by Kullback and Leibler in [KL51] as a generalization

of Shannon's measure of information.

Shannon published in 1948 a way of measuring the uncertainty associated with a discrete random variable X, called *entropy*, as

$$H(X) := -\sum_{i=1}^{n} p_i \log p_i,$$

where $p_i = P(X = x_i)$ is the probability of outcome $x_i, i = 1, \ldots, n$ (see [Sha48]). He is since regarded as the founder of the field of information theory (information being the opposite of entropy). Consider for example the toss of a coin: the uncertainty of the outcome is maximal if both sides are equiprobable (i.e. if it is a fair coin), the uncertainty is zero if one of the sides comes up with probability one.

In contrast, the notion of Kullback and Leibler compares the information content of two different sources. In their original paper, they used it to measure the mean information in a random variable X to discriminate between two hypotheses H_1 and H_2, $H_i := \{X \text{ has probability density } f_i\}$, if H_1 is true and defined it as

$$\mathcal{D}(f_1, f_2) := \int f_1(x) \log \frac{f_1(x)}{f_2(x)} \, \nu(\mathrm{d}x).$$

Moreover, this notion can be interpreted as the information gain in using probability measure f_1 instead of f_2 and is therefore also used in Bayesian statistics to measure the improvement in changing from the prior information f_2 to the posterior information f_1.

In our context, however, we will use it primarily as a measure of discrepancy between two densities. Even though $\mathcal{D}(f_1, f_2)$ is not a true metric, since it is not symmetric and does not fulfill the triangle inequality, Kullback and Leibler showed that

$$\mathcal{D}(f_1, f_2) \geq 0$$

and

$$\mathcal{D}(f_1, f_2) = 0 \quad \Longleftrightarrow \quad f_1(x) = f_2(x) \quad \nu - \text{a.s.}$$

Note that even though the term "Kullback-Leibler divergence" today always refers to $\mathcal{D}(f_1, f_2)$, the authors themselves defined "divergence" as $\mathcal{D}(f_1, f_2) + \mathcal{D}(f_2, f_1)$, which has the added advantage of being symmetric.

1.5 Results Probability

To prove convergence results of the considered algorithms, one frequently has to assess the probability $P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \geq \epsilon\right)$. Under certain conditions, one can use the Chebychev inequality. However, often the assumptions of the following theorem are easier to verify:

Theorem 1.4 (Hoeffding's Inequality)**.** *If $X_1, \ldots, X_n$ are independent and $0 \leq X_i \leq 1 \; \forall\, i = 1, \ldots, n$, then for $0 < t < 1 - \mu$:*

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq t\right) &= \left\{\left(\frac{\mu}{\mu+t}\right)^{\mu+t}\left(\frac{1-\mu}{1-\mu-t}\right)^{1-\mu-t}\right\}^n \\ &\leq e^{-nt^2 g(\mu)} \\ &\leq e^{-2nt^2}, \end{aligned} \tag{1.2}$$

where

$$g(\mu) = \begin{cases} \frac{1}{1-2\mu}\ln\left(\frac{1-\mu}{\mu}\right) & 0 < \mu < \frac{1}{2} \\ \frac{1}{2\mu(1-\mu)} & \frac{1}{2} \leq \mu < 1. \end{cases}$$

If $a \leq X_i \leq b \; \forall\, i = 1, \ldots, n$, the inequalities hold true if one replaces t by $\frac{t}{b-a}$ and μ by $\frac{\mu-a}{b-a}$.

We use inequality (1.2) of this theorem in Chapter 4 to conclude that for independent $a \leq X_i \leq b$, $i = 1, \ldots, n$, and small $\epsilon > 0$ holds

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \geq \epsilon\right) \leq 2\exp\left\{-2n\frac{\epsilon^2}{(b-a)^2}\right\}. \tag{1.3}$$

Moreover, we need the following terms:

Definition 4. A lower semicontinuous function $I : \mathcal{X} \to \mathbb{R}$ with $I(x) \geq 0 \; \forall x \in \mathcal{X}$ is called a *rate function*. We say that a sequence $\{z_i\}_{i\in\mathbb{N}}$ satisfies a *Large Deviations Principle* with rate function I if

- For all closed sets $C \subset \mathcal{X}$ holds: $\limsup_{n\to\infty} \frac{1}{n}\log P\left(z_n \in C\right) \leq -\inf_{x\in C} I(x)$.
- For all open sets $G \subset \mathcal{X}$ holds: $\liminf_{n\to\infty} \frac{1}{n}\log P\left(z_n \in G\right) \geq -\inf_{x\in G} I(x)$.

Analogous to this definition from [SW95] we say that a stochastic function g satisfies a Large Deviations Principle with rate function I if for the observations $g_1(x), g_2(x), \ldots$ holds

$$\limsup_{n\to\infty} \frac{1}{n}\log P\left(\left|\frac{1}{n}\sum_{i=1}^{n} g_i(x) - \mathbb{E}[g]\right| \geq \epsilon\right) \leq -I(x,\epsilon) \quad \forall\, x \in \mathcal{X},$$

i.e. if there exists some N such that

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} g_i(x) - \mathbb{E}[g]\right| \geq \epsilon\right) \leq e^{-nI(x,\epsilon)} \quad \forall\, n \geq N. \tag{1.4}$$

Chapter 2

The Cross-Entropy Method

In this chapter, we will give the basic ideas and algorithms of the Cross-Entropy method as proposed by Rubinstein and Kroese in [dBKMR05] and [RK04].
The Cross-Entropy (or CE) method has originally been developed as a tool for rare-event simulation. To estimate very small probabilities (10^{-5} or smaller), simple Monte-Carlo simulation is impractical since it requires very large sample sizes to obtain accurate estimators. One popular method to avoid this problem is the so-called *Importance Sampling*: changing the original measure to increase the frequency of the rare event in the simulation. The CE method provides an algorithm to calculate the optimal parameters of this new probability by sequentially assigning more and more probability mass to the rare event.

With some small adaptations, the same algorithm can be used to solve optimization problems. We can interprete the occurence of (near-)optimal states x as a rare event. Estimating the optimal importance sampling parameters by the CE method thus results in obtaining a probability function with most of its probability mass concentrated on the solutions of the original problem.

We will now give the details of the original rare-event algorithm and its application to optimization problems.

2.1 Rare-event simulation and Importance Sampling

Let $\{f(\cdot\,;\theta), \theta \in \Theta\}$ be a parametric family of probability density functions on a set $\mathcal{X} \subseteq \mathbb{R}^n$. Here and in all subsequent chapters we assume that $\mathcal{X}$ is measurable with respect to ν. Consider the problem of estimating the probability

$$p = \mathbb{P}_{\theta_{fix}}\big(F(X) \geq \gamma\big),$$

where $F : \mathcal{X} \to \mathbb{R}$ and $X \in \mathcal{X}$ is distributed according to the given probability $f(\cdot\,;\theta_{fix})$. Since p can also be written as $\mathbb{E}_{\theta_{fix}} \mathbb{1}_{\{F(X)\geq\gamma\}}$, the usual Monte-Carlo estimator is

$$\tilde{p} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{F(X_i)\geq\gamma\}}, \qquad X_1, \ldots, X_N \sim f(\cdot\,;\theta_{fix}).$$

If p is very small, $\{F(X) \geq \gamma\}$ is called a *rare event*. In this case, the indicator function in the estimator $\tilde{p}$ is equal to zero for most of the samples $X_1, \ldots, X_N$, so that obtaining a valid estimate for p by this direct approach requires a very large sample size N. For example, to estimate a probability $p = 10^{-6}$ with a relative error (RE) of 10% the sample size has to be approximately $N \approx 10^8$, since

$$\mathrm{RE} = \frac{\sqrt{\mathrm{Var}(\tilde{p})}}{\mathbb{E}[\tilde{p}]} = \frac{\sqrt{\frac{\mathrm{Var}(\mathbb{1}_{\{F(X)\geq\gamma\}})}{N}}}{p} = \frac{\sqrt{\frac{p(1-p)}{N}}}{p} = \sqrt{\frac{1-p}{N}}$$

and thus

$$N = \frac{1-p}{p \cdot \mathrm{RE}^2} \approx \frac{1}{p \cdot \mathrm{RE}^2}.$$

To reduce the necessary sample size, one popular approach is a change of measure to some density function g, called the *importance sampling density*, under which the event is more likely to occur. Since

$$\mathbb{E}_f[H(X)] = \mathbb{E}_g\left[H(X)\ \frac{f(X)}{g(X)}\right]$$

for any function H and density g for which $f << g$, the probability p can be estimated by

$$\widehat{p} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{F(X_i)\geq\gamma\}} \frac{f(X_i;\theta_{fix})}{g(X_i)}, \quad X_1, \ldots, X_N \sim g.$$

However, finding a good importance sampling density g is often not trivial. The estimator $\widehat{p}$ is still unbiased but its variance is determined by the choice of g. The conditional density of $X \sim f(\cdot\,;\theta_{fix})$ given that the event $\{F(X) \geq \gamma\}$ occurs, i.e.

$$g^*(x) := \frac{\mathbb{1}_{\{F(x)\geq\gamma\}} f(x;\theta_{fix})}{p}, \tag{2.1}$$

is optimal in the sense that it provides a zero-variance estimator but depends unfortunately on the unknown probability p.

To facilitate the problem of explicitly calculating an optimal g^*, the Cross-Entropy approach restricts the choice of the importance sampling density to the familiy $\{f(\cdot\,;\theta)\}$. It aims to find the $g \in \{f(\cdot\,;\theta)\}$ that is closest to the optimal density g^* in the cross-entropy (or Kullback-Leibler) sense. That is, we want to

solve the problem

$$\begin{aligned}\min_\theta \mathcal{D}\left(g^*, f(\cdot\,;\theta)\right) &= \min_\theta \mathbb{E}_{g^*}\left[\ln \frac{g^*(X)}{f(X;\theta)}\right] \\ &= \min_\theta \int g^*(x)\ln g^*(x)\mathrm{d}x - \int g^*(x)\ln f(x;\theta)\mathrm{d}x \\ &= \max_\theta \int g^*(x)\ln f(x;\theta)\mathrm{d}x \\ &= \max_\theta \int \frac{\mathbb{1}_{\{F(x)\geq\gamma\}} f(x;\theta_{fix})}{p}\ln f(x;\theta)\mathrm{d}x \\ &= \max_\theta \mathbb{E}_{\theta_{fix}}\left[\mathbb{1}_{\{F(X)\geq\gamma\}}\ln f(X;\theta)\right].\end{aligned}$$

There again the problem is to obtain a good estimate for the expectation within a reasonable sample size. The CE method proposes to solve this problem iteratively. Observe that

$$\max_\theta \mathbb{E}_{\theta_{fix}}\left[\mathbb{1}_{\{F(X)\geq\gamma\}}\ln f(X;\theta)\right] = \max_\theta \mathbb{E}_w\left[\mathbb{1}_{\{F(X)\geq\gamma\}}L(X;\theta_{fix},w)\ln f(X;\theta)\right]$$

for some parameter w, where $L(x;\theta_{fix},w) := \frac{f(x;\theta_{fix})}{f(x;w)}$ is the likelihood ratio between $f(\cdot\,;\theta_{fix})$ and $f(\cdot\,;w)$.
The idea behind the algorithm is now to find the optimal important sampling parameter

$$\theta^* = \operatorname*{argmax}_\theta \ \mathbb{E}_w\left[\mathbb{1}_{\{F(X)\geq\gamma\}}L(X;\theta_{fix},w)\ln f(X;\theta)\right] \tag{2.2}$$

by first relaxing the inequality $F(X)\geq\gamma$ to $F(X)\geq\gamma_0$, i.e. using $\gamma_0<\gamma$ such that the event $\{F(X)\geq\gamma_0\}$ has a higher probability (say $\rho=0.01$). Solving (2.2) for $w=\theta_{fix}$ then gives the optimal parameter θ_1 for the estimation of the event $\{F(X)\geq\gamma_0\}$. Since under the importance sampling parameter θ_1 this event has a high probability to occur, it is most likely that we can find a new level $\gamma_0<\gamma_1<\gamma$ for which the probability that $F(X)\geq\gamma_1$ is still large enough, i.e. ρ. For this level we can again compute the importance sampling parameter θ_2 by solving (2.2) for $w=\theta_1$. Thus, the algorithm iteratively constructs a sequence of increasing levels $\{\gamma_i\}$ and corresponding parameters $\{\theta_{i+1}\}$ until the original level γ is reached. The corresponding importance sampling parameter (say θ_T) is the optimal parameter θ^* for the initial program, since (see 2.2)

$$\begin{aligned}\theta_T &= \operatorname*{argmax}_\theta \ \mathbb{E}_{\theta_{T-1}}\left[\mathbb{1}_{\{F(X)\geq\gamma\}}L(X;\theta_{fix},\theta_{T-1})\ln f(X;\theta)\right] \\ &= \operatorname*{argmax}_\theta \ \mathbb{E}_{\theta_{fix}}\left[\mathbb{1}_{\{F(X)\geq\gamma\}}\ln f(X;\theta)\right] \\ &= \operatorname*{argmin}_\theta \mathcal{D}\left(g^*, f(\cdot\,;\theta)\right).\end{aligned}$$

However, to prove that the level γ is reached in a finite number of iterations is not an easy task. Rubinstein and Kroese give in [RK04] sufficient conditions on the parameterized family, the function F and the parameter ρ to ensure convergence in maximal $M(\gamma,f)$ iterations in the one-dimensional case for some function M, but

to our knowledge, more general results have yet to be established.

The resulting algorithm is as follows:

Algorithm 2.1 (CE method for rare-event estimation)**.**

Step 1: *Initialize* $\theta_0 := \theta_{fix}$*, iteration counter* $k := 0$*.*

Step 2: *Calculate* γ_k *as the* $(1-\rho)$*-quantile of* $F(X)$ *under* $f(\cdot\,;\theta_k)$*.*
If $\gamma_k \geq \gamma$ *set* $\gamma_k := \gamma$*.*

Step 3: *Determine the new parameter* θ_{k+1} *as the solution (cf. (2.2))*

$$\theta_{k+1} = \underset{\theta}{\operatorname{argmax}}\ \mathbb{E}_{\theta_k}\left[\mathbb{1}_{\{F(X)\geq\gamma_k\}} L(X;\theta_{fix},\theta_k)\ln f(X;\theta)\right].$$

Step 4: *If* $\gamma_k = \gamma$*,* **STOP***.*
The current parameter θ_{k+1} *is the optimal parameter* θ^**.*
Else set $k := k+1$ *and proceed with* **Step 2***.*

In practice, one only estimates the quantities γ_k and θ_{k+1}. In each iteration, N samples $X_1, \cdots, X_N \sim f(\cdot\,;\widehat{\theta}_k)$ are generated and the order statistic $F_{(1)}, \ldots, F_{(N)}$ is computed. Then

$$\begin{aligned}\widehat{\gamma}_k &= F_{(\lceil(1-\rho)N\rceil)},\\ \widehat{\theta}_{k+1} &= \underset{\theta}{\operatorname{argmax}}\,\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{\{F(X_i)\geq\widehat{\gamma}_k\}} L(X_i;\theta_{fix},\widehat{\theta}_k)\ln f(X_i;\theta).\end{aligned}$$

Consequently, both the estimated quantile values $\widehat{\gamma}_k$ and the parameters $\widehat{\theta}_{k+1}$ are random variables depending on all samples drawn up to that time, i.e. on the complete history of the algorithm.
We will refer to the algorithm as given in Algorithm 2.1 as the algorithm in its *deterministic form*, whereas the *stochastic form* refers to the use of the estimates $\widehat{\gamma}_k$ and $\widehat{\theta}_{k+1}$.

Note that whereas the sequence of levels $\{\gamma_k\}$ is determined by the algorithm, the sample size N and the quantile parameter ρ have to be specified in advance. Especially the choice of ρ (usually between 0.01 and 0.1) can be critical for the performance of the algorithm in some examples (see [RK04] for more details). In our examples, however, $\rho = 0.1$ yielded consistently good results.

2.2 Application to Optimization

Consider the following optimization problem:

$$\max_{x\in\mathcal{X}} F(x), \quad F : \mathcal{X} \to \mathbb{R} \tag{2.3}$$

and denote by $x^* \in \mathcal{X}$ the optimal state and by $\gamma^* = F(x^*)$ the optimal function value.

Consider now the event $\{F(X) \geq \gamma^*\}$ or $\{F(X) \geq \gamma^* - \epsilon\}$. Interpreting this as a rare event under some arbitrary density function $f \in \{f(\cdot\,;\theta)\}$ and applying the CE algorithm for rare-event simulation will result in constructing a sequence of increasing levels $\{\gamma_k\}$ converging to γ^* and corresponding parameters $\{\theta_{k+1}\}$. Since the parameters θ_{k+1} are the optimal importance sampling parameter for estimating the event $\{F(X) \geq \gamma_k\}$, once the optimal level γ^* has been reached by the algorithm, $f(\cdot\,;\theta_{k+1})$ will be a density with most of its probability mass concentrated on the set of states $\{x \in \mathcal{X}\}$ for which $F(x) \geq \gamma^*$, i.e. on the solution set for our original optimization problem.

However, in the original rare-event sampling setting we know the value of the optimal level γ^* since it is one of the fixed parameters in the estimation problem, whereas here it is one of the unknown variables we are interested in finding. But since γ^* is the maximal level that can be reached by the algorithm - recall that it is always computed as the $(1-\rho)$-quantile of the current distribution - and the sequence $\{\gamma_k\}$ is usually at least non-decreasing, stopping the algorithm when an appropriate convergence criterion has been satisfied should yield if not the optimal γ^* then at least a good estimate of it. Even though there are few theoretical results, this behaviour has been verified by many numerical experiments.

Adapting the original Cross-Entropy method to these new requirements results in the following algorithm in its stochastic form.

Algorithm 2.2 (CE method for optimization)**.**

Step 1: *Initialize* $\widehat{\theta}_0 := \theta_{fix}$ *for an arbitrary* θ_{fix}*, iteration counter* $k := 0$*.*

Step 2: *Generate* N *samples* $X_1, \ldots, X_N \sim f(\cdot\,;\widehat{\theta}_k)$*, compute the order statistic* $F_{(1)}, \ldots, F_{(N)}$ *and estimate the* $(1-\rho)$*-quantile:*

$$\widehat{\gamma}_k \;=\; F_{(\lceil (1-\rho)N \rceil)}.$$

Step 3: *Estimate the parameter* θ_{k+1}*:*

$$\widehat{\theta}_{k+1} = \operatorname*{argmax}_{\theta} \; \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{F(X_i) \geq \widehat{\gamma}_k\}} \ln f(X_i;\theta). \tag{2.4}$$

Step 4: *If the convergence criterion is reached (say* $\widehat{\gamma}_k = \widehat{\gamma}_{k-1} = \cdots = \widehat{\gamma}_{k-d}$ *for some* $d \leq t$*),* **STOP***.*
Else set $k := k+1$ *and proceed with* **Step 2***.*

Note that the likelihood ratio term $L(\cdot\,;\theta_{fix}, w)$ is not necessary here. In the rare-event setting the parameter θ_{fix} was a fix component of the estimation problem (recall that the optimal importance sampling density g^* also depended on θ_{fix}). Here, the interpretation of the optimization program as a rare-event problem is only a tool and the initial density $f(\cdot\,;\theta_{fix})$ is quite arbitrary. Dropping $L(\cdot\,;\theta_{fix}, w)$

means that we actually change the underlying estimation problem in each step: instead of computing the optimal parameter θ^* to approximate g^* as in (2.1), we approximate in the k-th iteration

$$g_k^*(x) := \frac{\mathbb{1}_{\{F(x)\geq\gamma_k\}} f(x;\theta_k)}{\mathbb{E}_{\theta_k}\left[\mathbb{1}_{\{F(X)\geq\gamma_k\}}\right]}, \tag{2.5}$$

i.e. compute the optimal importance sampling parameter for the estimation of $\mathbb{P}_{\theta_k}(F(X) \geq \gamma_k)$. Rubinstein and Kroese mention that their numerical experiments suggest that including the likelihood ratio in the optimization algorithm will often produce less reliable estimates (see [RK04], p. 134).

The difficulty of the parameter update (2.4) depends on the choice of the parameterized family $\{f(\cdot\,;\theta), \theta \in \Theta\}$, also called *model distribution*. In some cases the argument of the maximum of $\frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\{F(X_i)\geq\widehat{\gamma}_k\}} \ln f(X_i;\theta)$ can be determined analytically, otherwise one has to solve this optimization problem numerically. These additional computations may substantially influence the algorithm's speed. Consequently the update complexity should have an impact on the choice of model distribution.

Example 2.1 (Natural exponential families). If the distributions of the components $X^{(1)}, \ldots, X^{(n)}$ of X are independent and each belongs to a natural exponential family (i.e. has a density of the form (1.1)), it can be shown (see [dMR]) that the maximizer in Step 3 of Algorithm 2.2 is given by

$$\widehat{\theta}_{k+1} = \frac{\sum_{i=1}^N \mathbb{1}_{\{F(X_i)\geq\widehat{\gamma}_k\}} X_i}{\sum_{i=1}^N \mathbb{1}_{\{F(X_i)\geq\widehat{\gamma}_k\}}}. \tag{2.6}$$

Example 2.2 (Normal Distribution). If the components $X^{(1)}, \ldots, X^{(n)}$ of X are independent and have normal distributions $\mathcal{N}(\mu^{(j)}, (\sigma^2)^{(j)})$, $j = 1, \ldots, n$, one can easily calculate that the parameter update has the following form for all components $j = 1, \ldots, n$:

$$\widehat{\mu}_{k+1}^{(j)} = \frac{\sum_{i=1}^N \mathbb{1}_{\{F(X_i)\geq\widehat{\gamma}_k\}} X_i}{\sum_{i=1}^N \mathbb{1}_{\{F(X_i)\geq\widehat{\gamma}_k\}}}, \tag{2.7}$$

$$(\widehat{\sigma}_{k+1})^{(j)} = \frac{\sum_{i=1}^N \mathbb{1}_{\{F(X_i)\geq\widehat{\gamma}_k\}} \left(X_i - \widehat{\mu}_{k+1}^{(j)}\right)^2}{\sum_{i=1}^N \mathbb{1}_{\{F(X_i)\geq\widehat{\gamma}_k\}}}. \tag{2.8}$$

2.3 Modifications of the Cross-Entropy Method

Various modifications of the basic algorithm have been proposed. They usually differ in the calculation of the new parameter vector $\widehat{\theta}_k$ in Step 3 of the algorithm or change the algorithms parameters N and ρ at runtime. We will address here the most important modifications.

Different Update Rules

Recall that when being used for optimization, the CE algorithm is supposed to appoint more and more probability mass to the set of optimal points and thus eventually end up in a degenerate distribution. However, one of the main problems of the CE method is that this process may happen too quickly and the algorithm "freezes" at a suboptimal solution.
To avoid this, Rubinstein and Kroese propose a smoothed update of the parameter vector $\widehat{\theta}_{k+1}$ in all optimization examples ([RK04], [dBKMR05]): Let $\widetilde{\theta}_{k+1}$ denote the result of the estimation (2.4). Then calculate $\widehat{\theta}_{k+1}$ as

$$\widehat{\theta}_{k+1} = v\,\widetilde{\theta}_{k+1} + (1-v)\,\widehat{\theta}_k \tag{2.9}$$

for some $v \in [0,1]$. Even though this is an additional parameter to be specified in each problem, values of $v = 0.5$ or $v = 0.7$ seem to work well consistently.
In continuous optimization, the chosen parametric family of sampling distributions is often the normal distribution with parameter vector $\theta = (\mu, \sigma^2)$. In this case, one can additionally slow down the convergence of the variance by using a dynamic smoothing parameter

$$\beta_k = \beta - \beta\left(1 - \frac{1}{k+1}\right)^q, \tag{2.10}$$

with $q \in \{5, \ldots, 10\}$ and $\beta \in [0.8, 0.99]$. Then β_k replaces v in the smoothing of the variance in iteration k.

Rubinstein and Kroese also discuss the use of different reward functions, i.e. determining $\widehat{\theta}_{k+1}$ as the solution of

$$\widehat{\theta}_{k+1} = \underset{\theta}{\operatorname{argmax}} \; \frac{1}{N}\sum_{i=1}^{N} \psi\left(F(X_i)\right) \mathbb{1}_{\{F(X_i)\geq\widehat{\gamma}_k\}} \ln f(X_i;\theta)$$

with a monotone function ψ. While they find good results for $\psi(x) = x$ for some problems, they dismiss the use of functions of the form $\psi(x) = x^\beta$ with large β or $\psi(x) = \exp(-x/\beta)$ as leading too easily to local optima.

Adaptive modification of the algorithm parameters

Since the choice of N and ρ has to be specified in advance for each problem, yet can be critical for the performance of the algorithm, other modifications of the CE method implement dynamic adaptations of these parameters.

The *Fully Adaptive CE method* (FACE) for continuous optimization basically increases the sample size N each time there is no improvement. After a certain number of successive increases the algorithm terminates, the problem is qualified as 'hard' and the current best solution as 'unreliable' (see [RK04],p.191).

A similar idea is put forward by Homem-de-Mello in the rare-event estimation setting (see [dM07], [dMR]). Since there exist convergence results for this modified

algorithm for rare event estimation under rather light assumptions and, moreover, similar ideas are implemented in the Model-Reference Adaptive Search, we will give below the algorithm in full detail but adapted to the optimization context, i.e. with an arbitrary initial parameter θ_{fix}, without the likelihood ratio $L(\cdot\,;\theta_{fix},\theta)$ and with an appropriate convergence criterion.
The underlying idea of this modification is to demand a level increase $\gamma_k - \gamma_{k-1} \geq \varepsilon$ in each iteration (and adapt ρ appropriately) or to increase the sample size to obtain a better estimate of the quantile in the next iteration.
Note that the algorithm in its stochastic form does not ensure a strict monotonicity of the sequence of levels $\{\gamma_k\}$ since the sample $(1-\rho_k)$-quantile $\widehat{\gamma}_k$ is a random variable. As N increases, however, the estimator $\widehat{\gamma}_k$ gains in accuracy and the convergence can be ensured with probability one.

Algorithm 2.3 (Adaptive CE method)**.**

Step 1: *Initialize* $\widehat{\theta}_0 := \theta_{fix}$ *for an arbitrary* θ_{fix}*, iteration counter* $k := 0$*. Define* $\rho_0 := \rho$*,* $N_0 := N$*.*

Step 2: *Generate* N_k *samples* $X_1, \ldots, X_{N_k} \sim f(\cdot\,;\widehat{\theta}_k)$*, compute the order statistic* $F_{(1)}, \ldots, F_{(N_k)}$ *and estimate the* $(1-\rho_k)$*-quantile:*

$$\widehat{\gamma}_k \;=\; F_{(\lceil (1-\rho_k)N_k \rceil)}.$$

Step 3: *Estimate the parameter* θ_{k+1}*:*

$$\widehat{\theta}_{k+1} = \operatorname*{argmax}_{\theta} \; \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{1}_{\{F(X_i) \geq \widehat{\gamma}_k\}} \ln f(X_i;\theta). \tag{2.11}$$

Step 4: *If the convergence criterion is reached,* **STOP***.*
Else proceed with **Step 5***.*

Step 5: *Check whether there exists* $\bar{\rho} \in (0, \rho_k]$ *such that the sample* $(1-\bar{\rho})$*-quantile is bigger than or equal to* $\widehat{\gamma}_{k-1} + \varepsilon$*.*

a) If so and $\bar{\rho} = \rho_k$*, set* $\rho_{k+1} = \rho_k$ *and* $N_{k+1} = N_k$*.*

b) If so and $\bar{\rho} < \rho_k$*, set* ρ_{k+1} *the largest of such* $\bar{\rho}$ *and* $N_{k+1} = N_k$*.*

c) Otherwise set $\rho_{k+1} = \rho_k$ *but increase the samplesize:* $N_{k+1} = \lceil \alpha N_k \rceil$ *for some* $\alpha > 1$*.*

Set $k := k+1$ *and proceed with* **Step 2***.*

Of course, instead of specifying N and ρ in advance, we now have to choose the parameters ε and α. This choice is not as crucial for theoretical convergence but can significantly influence the rate of convergence and the computational ressources needed.

2.4 Stochastic Optimization

Since we will be concentrating on the optimization of Markov Decision Processes, we will shortly discuss the necessary adaptations for stochastic optimization problems of the form

$$\max_{x\in\mathcal{X}} F(x) := \max_{x\in\mathcal{X}} \mathbb{E}\left[\mathcal{F}(x,Y)\right], \quad \mathcal{F}: \mathcal{X}\times\mathcal{Y}\to\mathbb{R}, \tag{2.12}$$

where $Y : \Omega_Y \to \mathcal{Y}$ is a random variable on some probability space (Ω_Y, Σ, P). We assume here that the expectation cannot be computed explicitly for every $x \in \mathcal{X}$, since otherwise this problem reduces to (2.3). This means we have to approximate $F(x)$ by an unbiased estimate $\widehat{F}(x)$, e.g.

$$\widehat{F}(x) = \frac{1}{M}\sum_{j=1}^{M} \mathcal{F}(x, Y_j).$$

We will call M the *observation size* to distinguish it from the number of samples N.

Of course, using the estimate $\widehat{F}(x)$ instead of the correct value $F(x)$ further complicates the algorithm. Whereas before the only problem was to find "good" samples (to avoid searching the complete state space $\mathcal{X}$), the evaluation of these samples is now an additional source of inaccuracy - even for the same sample two independent estimations usually differ. The parameters N and M play a crucial role here. The sample size N determines the number of different states $x \in \mathcal{X}$ visited, a large N thus increases the probability of finding the optimal point x^*; on the other hand, a large observation size M ensures that the samples are evaluated correctly and thus that good (and bad) samples are recognized as such and the algorithm is not guided in a wrong direction. Hence, both a large N and a large M are desirable. However, usually the computational budget is limited in time and/or memory space and there has to be a tradeoff between the sample and the observation size to distribute the allowed number of function evaluations. One approach is to use an adaptive scheme to determine N (as in Section 2.3 and an increasing observation size M as in the Model-Reference Adaptive Search (see Section 3.3).

Note moreover that the sequence $\{\gamma_k\}$ will usually not converge to a fixed value due to the variance of $\widehat{F}$. Therefore the convergence criterion in Step 4 has to be defined appropriately, e.g. stop when the sample variance of the moving average (i.e. the average over the last d iterations) falls under a predefined level.

2.5 Model-Based Approach

The Cross-Entropy algorithm belongs to the class of model-based optimization methods. These approaches are characterized by the use of a probability distribution modeling the current assumptions about the localisation of good solutions. In each iteration, samples from the current distribution are drawn and evaluated and then the model is updated on the basis of these samples' performance. This distinguishes

a model-based approach from for example genetic algorithms that also use multiple solution candidates called "population" in each iteration but infer the next generation of these solutions directly from the current one without the use of an intermediate distribution model.

Example 2.3. Consider the function $F(x,y) = 4\exp\{-\frac{1}{2}\left((x-4)^2+(y-4)^2\right)\} + 2\exp\{-\left((x-6.5)^2+(y-6.5)^2\right)\}$ with optimum $(x^*,y^*) = (4,4)$, plotted in Figure 2.1.

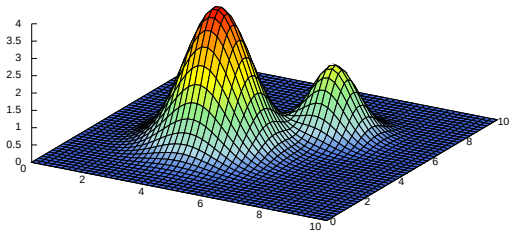
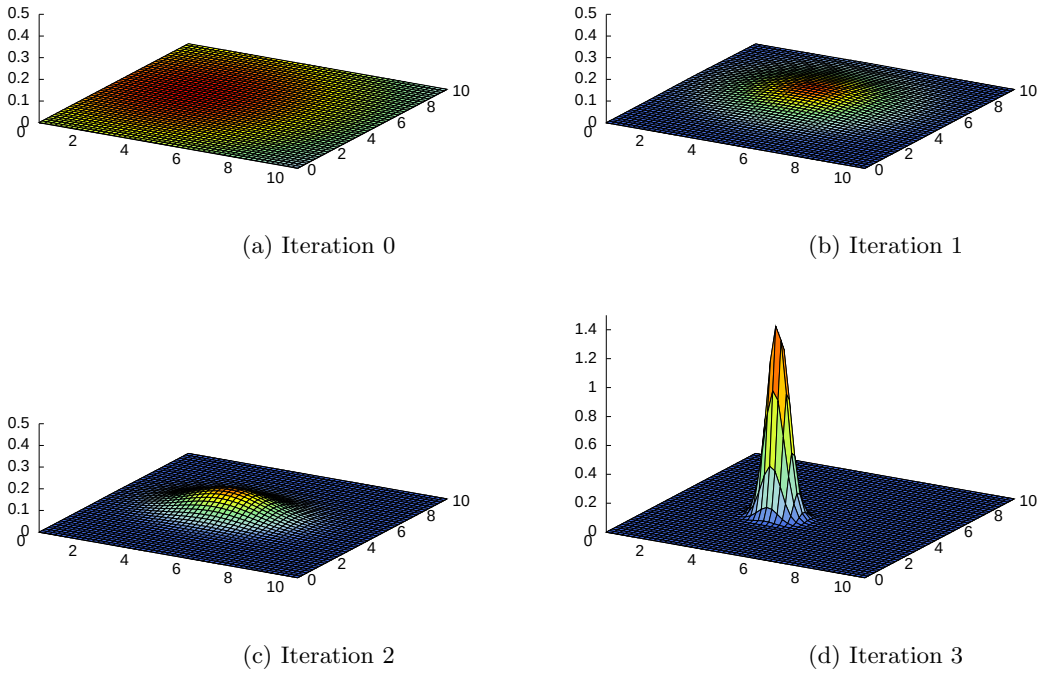


Figure 2.1: $F(x,y)$

Running the Cross-Entropy algorithm with independent normal distributions in each component yields the following sequence of parameters:

Iteration	(μ_x, μ_y)	(σ_x^2, σ_y^2)
0	$(2.79, 5.47)$	$(100, 100)$
1	$(4.61, 5.86)$	$(3.85, 3.07)$
2	$(4.43, 5.14)$	$(2.20, 1.07)$
3	$(4.13, 4.36)$	$(0.15, 0.09)$
$\vdots$		
9	$(4.00, 4.00)$	$(0.001, 0.001)$

The first iterations are shown in Figure 2.2.



(a) Iteration 0

(b) Iteration 1

(c) Iteration 2

(d) Iteration 3

Figure 2.2: Converging Density

Chapter 3

Model-Reference Adaptive Search

The Model-Reference Adaptive Search method (or MRAS) has been proposed by Hu, Fu and Marcus in [HFM07], [HFM] and [CFHM07]. Even though the authors stress the generality of their approach they only consider one explicit instantiation which is very similar to the Cross-Entropy algorithm in its modified version. Nevertheless, they are able to establish some convergence results in their framework.

3.1 Basic Idea

As in section (2.2), we will consider the problem

$$\max_{x\in\mathcal{X}} F(x), \quad F : \mathcal{X} \to \mathbb{R}.$$

Recall that in the Cross-Entropy method for optimization (Algorithm 2.2) the current parameter θ_k is calculated in each iteration as

$$\theta_k = \operatorname*{argmin}_{\theta} \mathcal{D}\left(g_k^*, f(\cdot\,;\theta)\right),$$

where

$$g_k^*(x) = \frac{\mathbb{1}_{\{F(x)\geq\gamma_k\}} f(x;\theta_k)}{\mathbb{E}_{\theta_k}\left[\mathbb{1}_{\{F(X)\geq\gamma_k\}}\right]}.$$

We can call g_k^* the *reference distribution* which we want to approximate as closely as possible by a member of the parametrized familiy of densities $\{f(\cdot\,;\theta), \theta \in \Theta\}$. In rare event estimation, this family is determined by the given problem: to estimate $\mathbb{P}_{\theta_{fix}}\left(F(X) \geq \gamma\right)$, $X \sim f(\cdot\,;\theta_{fix})$. In the optimization context, however, this family can be chosen suitably for the given problem, e.g. in such a way as to facilitate the sampling and updating processes. The reference distributions $\{g_k\}$ are then implied by the choice of $\{f(\cdot\,;\theta), \theta \in \Theta\}$.

The fundamental difference of the MRAS algorithm is now the separation of the sequence of reference distributions and the family of parameterized distributions

(which we will refer to as *sampling distributions*). The authors propose a sequence of the form

$$g_k(x) = \frac{F(x)g_{k-1}(x)}{\int_{\mathcal{X}} F(x)g_{k-1}(x)\nu(\mathrm{d}x)}, \quad x \in \mathcal{X}, k \geq 1, \tag{3.1}$$

for some initial distribution $g_0(x) > 0 \ \forall x \in \mathcal{X}$ if $F(x) > 0 \ \forall x \in \mathcal{X}$.
This is a popular approach in evolutionary algorithms, called *proportional selection*. In this form it appears in *estimation of distribution algorithms* (EDAs), a population-based class of algorithms that do not generate new populations directly from the parent generations by crossover and mutation operations but infer a probability distribution model from the populations and sample the next generation of candidate solutions from this posterior distribution. For details see e.g. [ZM04]. There it is also shown that

$$\lim_{k\to\infty} \mathbb{E}_{g_k}\left[F(X)\right] = F(x^*).$$

However, for arbitrary distributions g it can be difficult to generate samples and to perform the update (3.1). Consequently, as the CE method and unlike EDAs, the Model-Reference Adaptive Search makes use of a family of parameterized distributions $\{f(\cdot\,;\theta), \theta \in \Theta\}$:
In each iteration k, the samples are generated from the current density $f(\cdot\,;\theta_k)$ and the new parameter θ_{k+1} is obtained as

$$\theta_{k+1} = \operatorname*{argmin}_{\theta} \mathcal{D}\left(g_k, f(\cdot\,;\theta)\right)$$

for the given reference distribution g_k, which (like the optimal importance sampling density g_k^* in the CE method) is never calculated explicitly.

Since the proportional selection scheme only works with positive F, an additional strictly monotone function $S : \mathbb{R} \to \mathbb{R}_+\backslash\{0\}$ is introduced and each point $x \in \mathcal{X}$ in (3.1) weighted with the value $S\left(F(x)\right)$.

3.2 MRAS for deterministic optimization ($\mathbf{MRAS_0}$, $\mathbf{MRAS_1}$)

The specific sequence of reference distributions in the Model-Reference Adaptive Search is given by

$$g_k(x) := \frac{S\left(F(x)\right)\mathbb{1}_{\{F(x)\geq\widetilde{\gamma}_k\}}g_{k-1}(x)}{\mathbb{E}_{g_{k-1}}\left[S\left(F(X)\right)\mathbb{1}_{\{F(X)\geq\widetilde{\gamma}_k\}}\right]}, \qquad k = 1, 2, \ldots,$$

$$\text{and} \qquad g_0(x) := \frac{\mathbb{1}_{\{F(x)\geq\widetilde{\gamma}_0\}}}{\mathbb{E}_{\theta_0}\left[\frac{\mathbb{1}_{\{F(x)\geq\widetilde{\gamma}_0\}}}{f(X,\theta_0)}\right]}, \tag{3.2}$$

where the sequence of levels $\{\widetilde{\gamma}_k\}$ is ensured to be non-decreasing, which is important for the convergence results. Using this monotonicity and an iteration argument, one can show that the sequence of reference distributions can also be written as

$$g_k(x) = \frac{S\left(F(x)\right)^k\mathbb{1}_{\{F(x)\geq\widetilde{\gamma}_k\}}}{\mathbb{E}_{\theta_k}\left[\frac{S(F(X))^k}{f(X,\theta_k)}\mathbb{1}_{\{F(X)\geq\widetilde{\gamma}_k\}}\right]}, \qquad k = 0, 1, \ldots. \tag{3.3}$$

To update the parameter of the parameterized density, we therefore have to solve

$$\begin{aligned}\min_{\theta} \mathcal{D}\left(g_k, f(\cdot\,;\theta)\right) &= \min_{\theta} \mathbb{E}_{g_k}\left[\ln g_k(X)\right] - \mathbb{E}_{g_k}\left[\ln f(X;\theta)\right] \\ &= \min_{\theta} \mathbb{E}_{g_k}\left[\ln g_k(X)\right] - \frac{\mathbb{E}_{\theta_k}\left[\frac{S(F(X))^k}{f(X,\theta_k)} \mathbb{1}_{\{F(X)\geq\widetilde{\gamma}_k\}} \ln f(X,\theta)\right]}{\mathbb{E}_{\theta_k}\left[\frac{S(F(X))^k}{f(X,\theta_k)} \mathbb{1}_{\{F(X)\geq\widetilde{\gamma}_k\}}\right]} \\ &= \max_{\theta} \mathbb{E}_{\theta_k}\left[\frac{S\left(F(X)\right)^k}{f(X,\theta_k)} \mathbb{1}_{\{F(X)\geq\widetilde{\gamma}_k\}} \ln f(X,\theta)\right].\end{aligned}$$

The resulting algorithm in its deterministic form is as follows:

Algorithm 3.1 (MRAS_0 - Deterministic MRAS)**.**

Step 1: *Initialize* θ_0 *such that* $f(x,\theta_0) > 0\ \forall x \in \mathcal{X}$*, set iteration counter* $k := 0$*.*

Step 2: *Calculate* γ_k *as the* $(1-\rho)$*-quantile of* $F(X)$ *under* $f(\cdot\,;\theta_k)$*.*

Step 3: *If* $k = 0$ *or* $\gamma_k \geq \widetilde{\gamma}_{k-1} + \delta$*, set* $\widetilde{\gamma}_k = \gamma_k$*.*
Else set $\widetilde{\gamma}_k = \widetilde{\gamma}_{k-1}$*.*

Step 4: *Determine the new parameter* θ_{k+1} *as the solution*

$$\theta_{k+1} = \operatorname*{argmax}_{\theta}\ \mathbb{E}_{\theta_k}\left[\frac{S\left(F(X)\right)^k}{f(X,\theta_k)} \mathbb{1}_{\{F(X)\geq\widetilde{\gamma}_k\}} \ln f(X,\theta)\right].$$

Step 5: *If the convergence criterion is satisfied,* **STOP***. Else set* $k := k+1$ *and proceed with* **Step 2***.*

As with the CE method, one would use in practice only the stochastic form with γ_k and θ_{k+1} replaced by their respective Monte Carlo estimators. Moreover, to maintain convergence results, the authors introduce in the stochastic form algorithm (called MRAS_1) some additional features:
Like the adaptive Cross-Entropy algorithm, the MRAS_1 implements a dynamic adaptation of the sample size N and the quantile parameter ρ. Unlike in Algorithm 2.3 however, even the monotonicity of the sequence of now random levels $\{\widetilde{\gamma}_k\}$ rests assured by an exchange of the order of the parameter adaptation and the distribution update. Note that this implies that sometimes less samples are considered in the parameter vector update than it would be the case in the adaptive CE method. It might even happen that no samples at all are better than the current level $\widetilde{\gamma}_k$, so that any parameter θ is a solution of (3.4). In that case, we will set $\widehat{\theta}_{k+1} = \widehat{\theta}_k$.
We will use the notation $\widetilde{\gamma}(\rho, N)$ to denote the sample $(1-\rho)$-quantile when N samples are drawn.
Additionally, the sampling distribution $f(\cdot\,;\theta_k)$ is replaced by a mixture of the current distribution and the initial distribution $f(\cdot\,;\theta_0)$, denoted by $\widetilde{f}$. Since $f(\cdot\,;\theta_0) > 0$ $\forall x \in \mathcal{X}$, this ensures that the optimal solution always has a positive probability of

being sampled and reduces the danger of getting trapped in a local minimum. Note that in the stochastic algorithm, when only a countable number of samples is available, one actually minimizes in the parameter update the distance to a discrete distribution, defined on the set of drawn samples $\Lambda_k := \{X_1^{(k)}, \ldots, X_{N_k}^{(k)}\}$ ($X_i^{(k)}$ the i-th sample in the k-th iteration), given for $k = 0, 1, \ldots$ by

$$\widetilde{g}_k(X_i) = \begin{cases} \dfrac{\frac{S(F(X_i))^k}{\widetilde{f}(X_i,\widehat{\theta}_k)}\mathbb{1}_{\{F(X_i)\geq\widetilde{\gamma}_k\}}}{\sum_{X_i\in\Lambda_k}\frac{S(F(X_i))^k}{\widetilde{f}(X_i,\widehat{\theta}_k)}\mathbb{1}_{\{F(X_i)\geq\widetilde{\gamma}_k\}}} & \forall X_i \in \Lambda_k \quad \text{if } \{X_i \in \Lambda_k : F(X_i) \geq \widetilde{\gamma}_k\} \neq \emptyset, \\ \widetilde{g}_{k-1}(X_i) & \forall X_i \in \Lambda_{k-1} \quad \text{otherwise.} \end{cases}$$

This distribution is random, since it depends not only on the random values of $\widetilde{\gamma}_k$ and $\widehat{\theta}_k$, but also on the randomly generated set Λ_k.

Algorithm 3.2 (MRAS$_1$ - Adaptive MRAS)**.**

Step 1: *Initialize $\widehat{\theta}_0$ such that $f(x, \widehat{\theta}_0) > 0$ $\forall x \in \mathcal{X}$, set iteration counter $k := 0$. Define $\rho_0 := \rho$, $N_0 := N$.*

Step 2: *Generate N_k samples $X_1, \ldots, X_{N_k}$ according to*

$$\widetilde{f}(\cdot\,; \widehat{\theta}_k) = (1-\lambda) f(\cdot\,; \widehat{\theta}_k) + \lambda f(\cdot\,; \widehat{\theta}_0),$$

compute the order statistic $F_{(1)}, \ldots, F_{(N_k)}$ and estimate the $(1-\rho_k)$-quantile:

$$\widehat{\gamma}_k(\rho_k, N_k) \;=\; F_{(\lceil(1-\rho_k)N_k\rceil)}.$$

Step 3: *Parameter adaptation:*

If $k = 0$ or $\widehat{\gamma}_k(\rho_k, N_k) \geq \widetilde{\gamma}_{k-1} + \varepsilon$, then

a) Set $\widetilde{\gamma}_k = \widehat{\gamma}_k(\rho_k, N_k)$, $\rho_{k+1} = \rho$ and $N_{k+1} = N_k$.

Else, check whether there exists $\bar{\rho} \in (0, \rho_k]$ such that $\widehat{\gamma}_k(\bar{\rho}, N_k) \geq \widetilde{\gamma}_{k-1} + \varepsilon$.

b) If so, set $\widetilde{\gamma}_k = \widehat{\gamma}_k(\bar{\rho}, N_k)$, $\rho_{k+1} = \bar{\rho}$ and $N_{k+1} = N_k$.

c) Otherwise, set $\widetilde{\gamma}_k = \widetilde{\gamma}_{k-1}$, $\rho_{k+1} = \bar{\rho}$ and increase the sample size: $N_{k+1} = \lceil \alpha N_k \rceil$.

Step 4: *Estimate the parameter θ_{k+1}: If $\mathbb{1}_{\{F(X_i)\geq\widetilde{\gamma}_k\}} > 0$ for some X_i,*

$$\widehat{\theta}_{k+1} = \underset{\theta}{\operatorname{argmax}} \; \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{S\,(F(X_i))^k}{\widetilde{f}(X_i, \widehat{\theta}_k)} \mathbb{1}_{\{F(X_i)\geq\widetilde{\gamma}_k\}} \ln f(X_i, \theta). \tag{3.4}$$

Else set $\widehat{\theta}_{k+1} = \widehat{\theta}_k$.

Step 5: *If the convergence criterion is reached,* **STOP**.
Else set $k := k + 1$ and proceed with **Step 2**.

3.3 MRAS for stochastic optimization (MRAS_2)

To solve stochastic optimization problems of the form

$$\max_{x\in\mathcal{X}} F(x) := \max_{x\in\mathcal{X}} \mathbb{E}\left[\mathcal{F}(x,Y)\right], \quad \mathcal{F}:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}, \tag{3.5}$$

$Y:\Omega_Y\to\mathcal{Y}$ a random variable, we proceed in the same way as in the cross-entropy method (cf. Section (2.4)), i.e. replacing the expectation with its estimator

$$\widehat{F}(x) = \frac{1}{M}\sum_{j=1}^{M}\mathcal{F}(x,Y_j).$$

However, unlike the CE method, the Model Reference Adaptive Search uses an increasing sequence of observation sizes $\{M_k\}$. If $M_k\to\infty$ as $k\to\infty$, the estimation error $|F(x)-\widehat{F}(x)|$ will vanish in later iterations. Additional assumptions ensure that this will happen fast enough.
Furthermore, the authors adapt the generation of the sequence $\{\widetilde{\gamma}_k\}$ to the new setting: due to the fact that one can only use the estimate $\widehat{F}(\cdot)$ for the evaluation of each sample, there is an additional source of randomness in the stochastic level $\widehat{\gamma}_k$ in each iteration and thus also in $\widetilde{\gamma}_k$. Thus, it might be advisable to "correct" the estimate when it appears to be too good. This is done in Step 3*c* of the algorithm, i.e. when no sample in the current iteration was found to have a better performance than the last level $\widetilde{\gamma}_{k-1}$. If this is the case, the sample X°_{k-1} that achieved this level $\widetilde{\gamma}_{k-1}$ is reevaluated with the current observation size M_k. To ease notation, we will denote by $X(\rho,N)$ the sample that corresponds to the $(1-\rho)$-quantile of $\widehat{F}(x)$ if N samples are drawn, i.e.

$$X(\rho,N)\in\left\{X_i : \widehat{F}(X_i)=\widehat{F}_{(\lceil(1-\rho)N\rceil)}\right\}.$$

Note that due to this reevaluation the sequence $\{\widetilde{\gamma}_k\}$ might not be monotone increasing and we will need some assumptions on the behaviour of $\mathcal{F}$ to ensure its convergence.
Another modification in comparison with Algorithm 3.2 also accounts for the inaccuracy in the sample evaluation: as the update in Step 4 is based only on those samples that perform better than the current level $\widetilde{\gamma}_k$ but some samples with true expectation better than $\widetilde{\gamma}_k$ might have a poor score in the current iteration, we will also partly include those samples that are slightly worse than $\widetilde{\gamma}_k$. That is, instead of the indicator function $\mathbb{1}_{\{\widehat{F}(X_i)\geq\widetilde{\gamma}_k\}}$ we will use

$$\widetilde{\mathbb{1}}_{\{F,\gamma\}} := \begin{cases} 1 & \text{if } F\geq\gamma, \\ \frac{F-(\gamma-\varepsilon)}{\varepsilon} & \text{if } \gamma-\varepsilon<F<\gamma, \\ 0 & \text{if } F\leq\gamma-\varepsilon. \end{cases}$$

Then the algorithm is as follows:

Algorithm 3.3 (MRAS$_2$ - Stochastic Optimization)**.**

Step 1: *Initialize* $\widehat{\theta}_0$ *such that* $f(x,\widehat{\theta}_0) > 0\ \forall x \in \mathcal{X}$*, set iteration counter* $k := 0$*. Define* $\rho_0 := \rho$*,* $N_0 := N$*.*

Step 2: *Generate* N_k *samples* $X_1,\dots,X_{N_k}$ *according to*

$$\widetilde{f}(\cdot\,;\widehat{\theta}_k) = (1-\lambda)f(\cdot\,;\widehat{\theta}_k) + \lambda f(\cdot\,;\widehat{\theta}_0),$$

calculate

$$\widehat{F}_k(X_i) = \frac{1}{M_k}\sum_{j=1}^{M_k}\mathcal{F}(X_i,Y_j(\omega)) \quad \forall\, i = 1,\dots,N_k,$$

compute the order statistic $\widehat{F}_{k(1)},\dots,\widehat{F}_{k(N_k)}$ *and estimate the* $(1-\rho_k)$*-quantile:*

$$\widehat{\gamma}_k(\rho_k,N_k) \;=\; \widehat{F}_{k(\lceil(1-\rho_k)N_k\rceil)}.$$

Step 3: *Parameter adaptation:*

If $k=0$ *or* $\widehat{\gamma}_k(\rho_k,N_k) \geq \widetilde{\gamma}_{k-1}+\varepsilon$*, then*

a) Set $\widetilde{\gamma}_k = \widehat{\gamma}_k(\rho_k,N_k)$*,* $\rho_{k+1}=\rho$*,* $N_{k+1}=N_k$ *and* $X_k^\circ = X(\rho_k,N_k)$*.*

Else, check whether there exists $\bar{\rho}\in(0,\rho_k]$ *such that* $\widehat{\gamma}_k(\bar{\rho},N_k)\geq\widetilde{\gamma}_{k-1}+\varepsilon$*.*

b) If so, set $\widetilde{\gamma}_k = \widehat{\gamma}_k(\bar{\rho},N_k)$*,* $\rho_{k+1}=\bar{\rho}$*,* $N_{k+1}=N_k$ *and* $X_k^\circ = X(\bar{\rho},N_k)$*.*

c) Otherwise, calculate $\widetilde{\gamma}_k = \widehat{F}_k(X_{k-1}^\circ)$*, set* $\rho_{k+1}=\bar{\rho}$*,* $X_k^\circ = X_{k-1}^\circ$ *and increase the sample size:* $N_{k+1}=\lceil\alpha N_k\rceil$*.*

Step 4: *Estimate the parameter* θ_{k+1}*: If* $\widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}}>0$ *for some* X_i*,*

$$\widehat{\theta}_{k+1} = \operatorname*{argmax}_{\theta}\ \frac{1}{N_k}\sum_{i=1}^{N_k}\frac{S\left(\widehat{F}_k(X_i)\right)^k}{\widetilde{f}(X_i,\widehat{\theta}_k)}\widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}}\ln f(X_i,\theta). \tag{3.6}$$

Else set $\widehat{\theta}_{k+1}=\widehat{\theta}_k$*.*

Step 5: *If the convergence criterion is reached,* **STOP***.*
Else set $k := k+1$ *and proceed with* **Step 2***.*

As in the Cross-Entropy method for optimization, Hu, Fu and Marcus use a smoothed update of the form (2.9) in all their numerical examples.

Example 3.1. Like in the Cross-Entropy method, the optimizer in the parameter update of the model distribution (Step 4) can sometimes be calculated analytically. This is especially the case if the components of X are independent and have for example the following distributions:

- *Natural Exponential Families*:

As in Example 2.1 the maximizer in (3.6) is given by

$$\widehat{\theta}_{k+1} = \frac{\sum_{i=1}^{N} \frac{S\left(\widehat{F}_k(X_i)\right)^k}{\widetilde{f}(X_i,\widehat{\theta}_k)} \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}} X_i}{\sum_{i=1}^{N} \frac{S\left(\widehat{F}_k(X_i)\right)^k}{\widetilde{f}(X_i,\widehat{\theta}_k)} \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}}}. \tag{3.7}$$

- *Normal Distribution*:
  Here, the update formulas for mean and variance in each component $j = 1, \dots, n$ are the following (cf. Example 2.2):

$$\widehat{\mu}_{k+1}^{(j)} = \frac{\sum_{i=1}^{N} \frac{S\left(\widehat{F}_k(X_i)\right)^k}{\widetilde{f}(X_i,\widehat{\theta}_k)} \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}} X_i}{\sum_{i=1}^{N} \frac{S\left(\widehat{F}_k(X_i)\right)^k}{\widetilde{f}(X_i,\widehat{\theta}_k)} \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}}}, \tag{3.8}$$

$$\left(\widehat{\sigma}_{k+1}\right)^{(j)} = \frac{\sum_{i=1}^{N} \frac{S\left(\widehat{F}_k(X_i)\right)^k}{\widetilde{f}(X_i,\widehat{\theta}_k)} \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}} \left(X_i - \widehat{\mu}_{k+1}^{(j)}\right)^2}{\sum_{i=1}^{N} \frac{S\left(\widehat{F}_k(X_i)\right)^k}{\widetilde{f}(X_i,\widehat{\theta}_k)} \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}}}. \tag{3.9}$$

One can see that the structure of the update formulas in CE and MRAS is quite similar apart from the weight factors $\frac{S\left(\widehat{F}_k(X_i)\right)^k}{\widetilde{f}(X_i,\widehat{\theta}_k)}$ in the latter (and the less important modified indicator function). The function $S : \mathbb{R} \to \mathbb{R}_+ \backslash \{0\}$ has to be strictly increasing (and hence cannot be chosen to be 1 like in CE) and the occurence of the density $\widetilde{f}(X_i, \widehat{\theta}_k)$ in this weight factor is a direct consequence of the reference distribution's form (3.1) and has to be computed for all samples $X_1, \dots, X_{N_k}$ in each iteration. Naturally, this increases the operations needed in each iteration compared with the Cross-Entropy method, even if sample size N, quantile parameter ρ and observation size M are the same in both algorithms.

3.4 Application to MDPs

Both the Cross-Entropy method and the Model Reference Adaptive Search are principally global optimization algorithms on a multidimensional set $\mathcal{X} \subseteq \mathbb{R}^n$. However, they can quite easily be adapted to solve Markov Decision Problems through direct policy search if the state space S is finite.
Consider first the existence of a stationary policy $\pi = (f, f, \dots)$ with decision rule $f : S \to A$, $f(i) \in D(i)\ \forall\, i \in S$. In the following (particularly in Chapter 5) we will use the notation π for the policy $\pi = (f, f, \dots)$ as well as for the decision rule f itself to avoid confusion with the model density $f \in \{f(\cdot;\theta), \theta \in \Theta\}$. Since S is finite, finding an optimal policy amounts to finding a vector $\pi^* = \left(\pi^*(i_1), \dots, \pi^*(i_{|S|})\right) \in D(i_1) \times \cdots \times D(i_{|S|}) := \mathcal{X} \subseteq \mathbb{R}^{|S|}$ with $V_{\infty\pi^*}(i) = V_\infty(i)$ for some fixed initial state i. Hence the task is to maximize the function $V_{\infty\pi}(i) : \mathcal{X} \to \mathbb{R}$ over $\mathcal{X}$ - which corresponds exactly to the usual global optimization setting.
Note that it is beyond the scope of the algorithms to consider the complete value function $V_{\infty\pi} \in \mathbb{R}^{|S|}$ since they require evaluation functions $F : \mathcal{X} \to \mathbb{R}$ in order to

be able to sort candidate solutions $\pi_1, \ldots, \pi_n$ according to their performance. Consequently we have to restrict ourselves to the real-valued evaluation $F(\pi) = V_{\infty\pi}(i)$ for some i. Having fixed this initial state, we will usually even drop it in the notation and write $V_{\infty\pi} = V_{\infty\pi}(i)$.

The adaptation to non-stationary policies is now straightforward and amounts to an extension of the dimension of the search space $\mathcal{X}$. Let N denote the time horizon and $F_t := D_t(i_1) \times \cdots \times D_t(i_{|S|}) \subseteq \mathbb{R}^{|S|}$ the set of admissible policies at time $t \leq N$. Then the sought policy $\pi^* = \left(\pi_0^*, \ldots, \pi_{N-1}^*\right)$ is the maximizer of $V_{\infty\pi}(i)$ over $\mathcal{X} = F_0 \times \cdots \times F_{N-1} \subseteq \mathbb{R}^{|S| \times N}$ for some initial state i.

As described above in Sections 2.4 and 3.3, we will estimate the performance $V_{\infty\pi}$ by averaging over M_k observations of R_π^∞. Since it is impossible to calculate R_π^∞ numerically, we simulate the policy π over an adequate time horizon to obtain a good estimate $\widehat{R}_\pi^\infty$. Then $\widehat{V}_\infty^{\pi,k} := \frac{1}{M_k} \sum_{j=1}^{M_k} \widehat{R}_\pi^\infty$ denotes the value function under policy π estimated in iteration k.

Note that the algorithms are both formulated for maximization problems. Hence to minimize, we consider the equivalent problem $-V_{\infty\pi} \to \max$.

Chapter 4

Convergence Results

In this chapter, we will give convergence results for the Model Reference Adaptive Search for stochastic optimization. The statements and proofs follow [CFHM07], yet we tried to put some ideas on a more rigorous mathematical basis, corrected several formulations, were able to lessen some assumptions and to elaborate and complete some arguments.
For the Cross-Entropy method, convergence results are few and moreover only concerned with discrete and deterministic optimization. In that context, Costa, Jones and Kroese (see [CJK07]) show that Algorithm 2.2 with modification (2.9) converges with probability one to a degenerate distribution (but not necessarily to the optimum) and that the parameters can be chosen in such a way that the probability to generate the optimal solution in some iteration of the algorithm is arbitrarily close to 1. However, they are not able to show convergence to the optimum and even suppose that the conditions for the two statements cited above are mutually exclusive.

4.1 Assumptions

In the following the necessary assumptions for convergence results of Algorithm 3.3 (MRAS_2) are stated. Certainly analogical results for MRAS_0 and MRAS_1 (Algorithms 3.1 and 3.2) can be derived with far less assumptions, yet our original interest was in solving Markov Decision Processes and hence we will refrain from considering Algorithms 3.1 and 3.2 separately. In addition, the underlying ideas in the proofs of all three algorithms are similar and the deterministic optimization algorithm MRAS_1 is only a special case of the stochastic optimization handled by MRAS_2. Therefore we will only discuss the latter algorithm below. For more details on the convergence of MRAS_0 and MRAS_1 see [CFHM07].

Assumptions on the problem setting

(A1) $F(x)$ has a unique global optimal solution $x^* \in \mathcal{X}$.

(A2) For any $\xi < F(x^*)$, the set $\{x \in \mathcal{X} : F(x) \geq \xi\}$ has a strictly positive

Lebesgue (or discrete) measure.

(A3) For any $\delta > 0$, define $A_\delta := \{x \in \mathcal{X} : \|x - x^*\| \geq \delta\}$. Then it holds: $\sup_{x \in A_\delta} F(x) < F(x^*)$, where $\sup_{x \in \emptyset} := -\infty$.

(A4) There exists a compact set K_ε such that w.p.1 there exists $\mathcal{N} \in \mathbb{N}$ with: $\{x \in \mathcal{X} : F(x) \geq F(X_k^\circ) - \varepsilon\} \subseteq K_\varepsilon \ \forall k \geq \mathcal{N}$.

(A5) For each $k \in \mathbb{N}$ there exists $\Delta_k > 0$ and $L_k > 0$ such that $\frac{|S(y)^k - S(y')^k|}{|S(y)^k|} \leq L_k |y - y'|$ for all $y, y' \leq F(x^*)$ with $\|y - y'\| < \Delta_k$.

These assumptions ensure that value function and the set of admissible solutions are of such a form that the global optimum of F is detectable by the algorithm.
Assumption (A1) enables us for example to choose unimodal model distributions such as the normal distribution as underlying model. Given two or more distinct optimal solutions, such a distribution would oscillate between the different peaks, converge to only one of them or be unable to find any optimum. First approaches in the CE method to use mixtures of unimodal distributions to optimize functions with several optima are successful [KPR06], however proving their applicability is difficult. Hence this restriction to the simpler case of a unique optimum.
(A2) guarantees that one can sample (near-)optimal solution candidates $x \in \mathcal{X}$ and is automatically satisfied for discrete problems and for continuous value functions. Under (A1), (A3) is an equally mild condition which ensures that the values of points not in the neighbourhood of x^* are bounded away from the optimal value $F(x^*)$. It is not satisfied if e.g. $\mathcal{X}$ is open and $F(x) = F(x^*)$ for some $x \in \partial\mathcal{X}$.
(A4) demands the existence of some compact set to which the search can be restricted. This is necessary to obtain bounds for continuous functions on K_ε in Proposition 4.5 and Lemma 4.6. The assumption is satisfied if e.g. $\mathcal{X}$ is compact or $F(x)$ has compact level sets.
(A5) is a condition on the positive function S necessary to prove Lemma 4.7. If for example $S(y) = e^{\tau y}$ for some $\tau > 0$, it holds $\frac{|S(y)^k - S(y')^k|}{|y - y'|} = \frac{|S(ky) - S(ky')|}{|y - y'|} \leq \frac{d}{d(y+\Delta_k)} S\left(k(y + \Delta_k)\right) = \tau k e^{\tau k y} e^{\tau k \Delta_k}$ for all $y, y' : \|y - y'\| < \Delta_k$. Then $\Delta_k = \frac{1}{k}$ and $L_k = \tau k e^\tau$ fulfill Assumption (A5).

Assumptions on the distribution family

(B1) The density of the parameterized family is of the form

$$f(x, \theta) = h(x) \exp\{\theta^\top T(x) - A(\theta)\} \quad \forall \theta \in \Theta, \tag{4.1}$$

where $A(\theta) = \ln \int_{\mathcal{X}} h(x) \exp\{\theta^\top T(x)\} \nu(\mathrm{d}x)$ is finite and T continuous for continuous distributions.

(B2) It holds that $\sup_{\theta \in \Theta} \|h(x) T(x) \exp\{\theta^\top T(x)\}\|$ is integrable (summable) w.r.t. x.

(B3) The maximizer in the parameter update (Step 4 in all algorithms) is an interior point of Θ $\forall k$.

(B4) For the initial density holds: $f(x,\theta_0) > 0\ \forall x \in \mathcal{X}$ and it is bounded away from zero on the compact set K_ε defined in Assumption (A4), i.e. $f_* := \inf_{x\in K_\varepsilon} f(x,\theta_0) > 0$.

Observe that the distribution required by Assumption (B1) has the form of an exponential family (see Section 1.3). However recalling the discussion at the end of Example 1.1 we refrain from calling it such since we are more interested in the form than in the characteristics of these class of distributions.
(B3) is trivially satisfied if Θ is unbounded. Note that the parameter set is the one for the (possibly reparameterized) form given in Assumption (B1), which does not necessarily correspond to the familiar one of the considered distributions (cf. Examples 1.1, 1.2). Assumption (B4) is easily fulfilled by an appropriate choice of θ_0. In case of the normal distribution for example, the initial variance has only chosen to be large enough.

Assumptions on the available samples $\mathcal{F}(x, Y_j)$

(C1) For any $\epsilon > 0$, there exists a $\mathcal{N} > 0$ and a function $\phi : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ strictly decreasing in the first argument, non-increasing in the second argument and $\phi(n,\epsilon) \to 0$ as $n \to \infty$, such that $\forall\, n \geq \mathcal{N}$ holds:

$$\sup_{x\in\mathcal{X}} P\left(\left|\frac{1}{n}\sum_{j=1}^{n}\mathcal{F}(x,Y_j) - \mathbb{E}\big[\mathcal{F}(x,Y)\big]\right| \geq \epsilon\right) \leq \phi(n,\epsilon).$$

(C2) For any $\epsilon > 0$ and some function $\phi : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ as in (C1), there exists a $\mathcal{N} > 0$, $\mathcal{M} > 0$ such that $\forall\, n \geq \mathcal{N}, m \geq \mathcal{M}$ holds:

$$\sup_{x,y\in\mathcal{X}} P\left(\left|\frac{1}{n}\sum_{j=1}^{n}\mathcal{F}(x,Y_j) - \frac{1}{m}\sum_{j=1}^{m}\mathcal{F}(y,Y_j) - F(x) + F(y)\right| \geq \epsilon\right) \leq \phi\big(\min\{m,n\},\epsilon\big).$$

These assumptions demand not only that the samples fulfill the law of large numbers, but that they do so uniformly on $\mathcal{X}$ and that the rate of convergence can be described by some function ϕ with the characteristics stated above.

Example 4.1. The sequence $\mathcal{F}(x, Y_j)$, $j = 1, 2, \ldots$ is i.i.d with uniformly bounded variance $\sigma^2(x) < \infty$. Then by Chebyshev's inequality and $\mathrm{Var}\Big(\frac{1}{n}\sum_{j=1}^{n}\mathcal{F}(x,Y_j)\Big) = \frac{\sigma^2(x)}{n}$ the function

$$\phi(n,\epsilon) := \frac{2\,\sup_{x\in\mathcal{X}}\sigma^2(x)}{n\epsilon^2}$$

satisfies both (C1) and (C2).

Example 4.2. The sequence $\mathcal{F}(x, Y_j)$, $j = 1, 2, \ldots$ fulfills the large deviations principle (1.4) for some rate function $I(x,\epsilon)$ bounded from below on $\mathcal{X}$. Then

$$\phi(n,\epsilon) := \exp\left\{-n \cdot \inf_{x\in\mathcal{X}} I(x,\epsilon)\right\}.$$

Assumptions on the observation size

(D1) The sequence $\{M_k\}$ satisfies $M_k \geq M_{k-1}\ \forall\, k \geq 1$ and $M_k \to \infty$ as $k \to \infty$. Moreover, for some ϕ fulfilling (C1) and (C2) it holds that for any $\epsilon > 0$ there exists $\delta_\epsilon \in (0,1)$ and $\mathcal{K}_\epsilon \in \mathbb{N}$ such that

$$\phi(M_{k-1}, \epsilon) \leq \delta_\epsilon^k \quad \forall\, k \geq \mathcal{K}_\epsilon.$$

(D2) For the function ϕ from (D1) and for some Δ_k, L_k satisfying (A5), it holds: For any $\zeta > 0$ there exist $\delta_\zeta \in (0,1)$ and $\mathcal{K}_\zeta \in \mathbb{N}$ such that sequence $\{M_k\}_{k\geq 0}$ satisfies

$$\alpha^k \phi\left(M_k, \min\left\{\Delta_k, \frac{\zeta}{\alpha^{\frac{k}{2}}}, \frac{\zeta}{\alpha^{\frac{k}{2}} L_k}\right\}\right) \leq \delta_\zeta^k \quad \forall\, k \geq \mathcal{K}_\zeta.$$

Assumptions (D1) requires the observation size to grow fast enough to ensure a sufficiently good rate of convergence of the estimate $\widehat{F}_k(x) = \frac{1}{M_k}\sum_{j=1}^{M_k} \mathcal{F}(x, Y_j)$ to $F(x) = \mathbb{E}\big[\mathcal{F}(x,Y)\big]$. (D2) ensures that this convergence is even fast enough to prevent an overlarge error between $S(\widehat{F}_k(x))^k$ and $S(F(x))^k$.

Example 4.3. Let ϕ be of the form $\phi(n, \epsilon) = \frac{\sigma^2}{n\epsilon^2}$. Then

$$\begin{aligned} & \phi(M_{k-1}, \epsilon) \leq \delta_\epsilon^k \quad \forall\, k \geq \mathcal{K}_\epsilon \\ \iff \quad & M_{k-1} \geq \frac{\sigma^2}{\epsilon^2 \delta_\epsilon^k} = \frac{\sigma^2}{\epsilon^2} \cdot \left(\frac{1}{\delta_\epsilon}\right)^k \quad \forall\, k \geq \mathcal{K}_\epsilon. \end{aligned}$$

This is satisfied for some $\mathcal{K}_\epsilon$ large enough if for example $M_{k-1} = c_0 \cdot \beta^k$ with $\beta > \frac{1}{\delta_\epsilon} > 1$, $c_0 \geq 1$.

And for $\Delta_k = \frac{1}{k}$, $\text{Ł}_k = k\tau e^\tau$ holds the following, since $\min\left\{\Delta_k, \frac{\zeta}{\alpha^{\frac{k}{2}}}, \frac{\zeta}{\alpha^{\frac{k}{2}} L_k}\right\} = \frac{\zeta}{\alpha^{\frac{k}{2}} L_k}$ for large k:

$$\begin{aligned} & \alpha^k \phi\left(M_k, \min\left\{\Delta_k, \frac{\zeta}{\alpha^{\frac{k}{2}}}, \frac{\zeta}{\alpha^{\frac{k}{2}} L_k}\right\}\right) \leq \delta_\zeta^k \quad \forall\, k \geq \mathcal{K}_\zeta \\ \iff \quad M_k \ &\geq \ \frac{\sigma^2 \alpha^k}{\delta_\zeta^k (\min\{\Delta_k, \frac{\zeta}{\alpha^{\frac{k}{2}}}, \frac{\zeta}{\alpha^{\frac{k}{2}} L_k}\})^2} \\ &= \ \sigma^2 \left(\frac{\alpha}{\delta\, \zeta}\right)^k \max\left\{k^2, \frac{\alpha^k}{\zeta^2}, \frac{k^2 \tau^2 e^{2\tau} \alpha^k}{\zeta^2}\right\} \\ &= \ \sigma^2 \frac{\tau^2 e^{2\tau}}{\zeta^2} \left(\underbrace{k^{\frac{2}{k}}}_{\to 1} \frac{\alpha^2}{\delta_\zeta}\right)^k \quad \forall\, k \geq \mathcal{K}_\zeta. \end{aligned}$$

Hence for example $M_k = c_0 \cdot \beta^k$ with $c_0 > 1$, $\beta > \frac{\alpha^2}{\delta_\zeta} > \alpha^2$.

Example 4.4. The function ϕ is of the form $\phi(n,\epsilon) = \exp\{-n \cdot I(\epsilon)\}$, $0 < I(\epsilon) < \infty$. Then for $\Delta_k = \frac{1}{k}$ and $L_k = k\tau e^\tau$ it holds

$$(D1) \iff M_{k-1} \geq k \cdot \frac{\ln(\delta_\epsilon)}{-I(\epsilon)} \quad \forall\, k \geq \mathcal{K}_\epsilon,$$

$$(D2) \iff M_k \geq k \cdot \frac{\ln(\frac{\delta_\zeta}{\alpha})}{-I\left(\frac{\zeta}{\alpha^{\frac{k}{2}} k\tau e^\tau}\right)} \quad \forall\, k \geq \mathcal{K}_\zeta.$$

Consider random variables $\mathcal{F}(x, Y) \in [a, b]$ satisfying the Hoeffding inequality. Then

$$I(\epsilon) = 2\,\frac{\epsilon^2}{(b-a)^2}$$

and

$$\begin{aligned}(D2) \iff \quad M_k &\geq -k \ln\left(\frac{\delta_\zeta}{\alpha}\right)\left(\frac{(b-a)^2}{2}\right)\left(\frac{\alpha^{\frac{k}{2}} k\tau e^\tau}{\zeta}\right)^2 \\ &= \left(k^{\frac{3}{k}}\,\alpha\right)^k \cdot \left(-\ln\left(\frac{\delta_\zeta}{\alpha}\right)\frac{(b-a)^2\tau e^\tau}{2\zeta}\right) \quad \forall\, k \geq \mathcal{K}_\zeta.\end{aligned}$$

Hence a valid observation rule is for example given by $M_k = c_0 \cdot \beta^k$ with $\beta > \alpha$, $c_0 \geq 1$.

Assumptions (D1) and (D2) provide quite explicit rules for the choice of observation size M_k and its growth over time. As parameter choice may be a very difficult task, this can be quite helpful. However, the problem shifts to finding a "good" function ϕ, that is a function with very tight bounds in Assumptions (C1) and (C2) to minimize the necessary observation size M_k. In straightforward, easy-structured problems this may not present any difficulties. Nevertheless if the realizations of $\mathcal{F}(x, Y)$ are for example outcomes of a simulation of a complex or even partially unknown system, it may be almost impossible to find a proven theoretically admissible and at the same time practically realizable observation size rule.

Assumption on the sample size increase rate

(E1) Let $\varphi > 0$ be a positive constant such that $\left\{x \in \mathcal{X} : S\left(F(x)\right) \geq \frac{1}{\varphi}\right\}$ has a strictly positive Lebesque (or counting) measure and define $S^* = S\left(F(x^*)\right)$. Then assume $\alpha > (\varphi\, S^*)^2 \geq 1$.

This assumption is necessary for technical reasons in the proofs of Lemma 4.6 and 4.7 but does not impose any real constraints on the choice of the sample size increase rate α: Under Assumption (A2) such a positive constant not only alway exists but can be chosen as $\varphi = \frac{1+\epsilon}{S^*}$ for some $\epsilon > 0$ in the continuous and $\epsilon = 0$ in the discrete case. Then (E1) reduces to $\alpha > 1$ which is the only reasonable choice for the increase rate anyway.

4.2 Probability space and notations

Whereas MRAS_0 is a deterministic algorithm, both MRAS_1 and MRAS_2 are stochastic and depend on the drawn samples and in case of the MRAS_2 also on their stochastic evaluations. This means that in each iteration of MRAS_2, the current parameters $\widehat{\theta}_k$, N_k and ρ_k, the level $\widetilde{\gamma}_k$, the sample X_k° and the implicit reference distribution form a stochastic process on a probability space $(\Omega = \Omega_X \times \Omega_Y, \mathfrak{F}, P)$.
We will denote by

$$X_i^{(k)} : \Omega_X \to \mathcal{X}$$

the ith candidate solution drawn in iteration k and by

$$Y_{x,j}^{(k)} : \Omega_Y \to \mathcal{Y}$$

the stochastic component in the jth evaluation of sample x in iteration k. If we define the random vector

$$Z_k := \left(X_1^{(k)}, Y_{X_1^{(k)},1}^{(k)}, \ldots, Y_{X_1^{(k)},M_k}^{(k)}, \ldots, X_{N_k}^{(k)}, Y_{X_{N_k}^{(k)},1}^{(k)}, \ldots, Y_{X_{N_k}^{(k)},M_k}^{(k)} \right)$$

as the vector of samples drawn in iteration k, the filtration $\mathfrak{F}$ is generated by

$$\mathfrak{F} = \{\mathfrak{F}_k, k \in \mathbb{N}\}, \text{ where } \mathfrak{F}_k = \sigma\left(Z_i, i = 1, \ldots, k\right).$$

The distribution of Y may depend on x, but is independent of k and the algorithms parameters (such as θ_k). The distribution of the samples $X_1^{(k)}, \ldots, X_{N_k}^{(k)}$ depends on the random variable $\widehat{\theta}_{k-1}$, but conditioned on $\widehat{\theta}_{k-1}$ (i.e. conditioned on $\mathfrak{F}_{k-1}$) these samples are independent and identically distributed with density $\widetilde{f}(\cdot\,; \widehat{\theta}_{k-1})$. Hence, for the probability measure P it holds

$$P\left(\left(X_1^{(k)}, \ldots, X_{N_k}^{(k)}\right) \in A\right) = \int_A \prod_{i=1}^{N_k} \widetilde{f}(x_i\,; \widehat{\theta}_{k-1})\, \nu(\mathrm{d}(x_1, \ldots, x_{N_k})) \quad \forall\, A \subseteq \mathcal{X}^{N_k}$$

and if $h(\cdot\,; x)$ designates the probability density of Y under x

$$P\left(Z_k \in A\right) = \int_A \prod_{i=1}^{N_k} \widetilde{f}(x_i\,; \widehat{\theta}_{k-1}) \prod_{j=1}^{M_k} h(y_j\,; x_i) \nu(\mathrm{d}z) \quad \forall A \subseteq \mathcal{X}^{N_k} \times \mathcal{Y}^{N_k \cdot M_k}.$$

Example 4.5 (Infinite Horizon Markov Decision Problems). Consider sampling X from the space of stationary policies of some Markov Decision Problem, i.e. $X \mathrel{\widehat{=}} f$, $f : S \to A$. Then $Y = (s_0, s_1, \ldots) \in S^\infty$ can be interpreted as a random path under policy f starting in some initial state s_0. We know that

$$P_{X,x_0}(y) = \delta_{s_0 x_0} \prod_{k=0}^{\infty} p(s_k, X(s_k), s_{k+1}), \quad (s_0, s_1, \ldots) \in S^\infty$$

and thus

$$P\left(Z_k \in A\right) = \int_A \prod_{i=1}^{N_k} \widetilde{f}(x_i\,; \widehat{\theta}_{k-1}) \prod_{j=1}^{M_k} \left(\delta_{s_0^{(j)} x_0} \prod_{m=0}^{\infty} p\left(s_m^{(j)}, x_i(s_m^{(j)}), s_{m+1}^{(j)}\right) \right) \nu(\mathrm{d}z).$$

In the following, we will only consider MRAS$_2$ (Algorithm 3.3), since our original interest was in solving Markov Decision Processes. Moreover, even though convergence result for deterministic optimization (MRAS$_1$) can be obtained with far less assumptions, it is only a special case of the more general stochastic optimization.

Recall that most components in the algorithm are random, since they are computed using the random samples X and Y. This includes not only the parameters $\widehat{\gamma}_k$, N_k and ρ_k but also the levels $\widetilde{\gamma}_k$, the quantile samples X_k° and of course the set of drawn samples Λ_k. Hence, all events that are defined over one or more of those values are random events. Their occurence depends on the concrete realisation z of the sample path Z_τ, where τ is the stopping time

$$\tau := \min\{t : \text{ in iteration } t \text{ is the stopping criterion (Step 5) fulfilled}\}.$$

The events that are especially interesting for our needs are:

$$\begin{aligned}
\Omega_2 &:= \{z : \text{ Steps (3a) and (3b) are visited finitely often}\},\\
\Omega_3 &:= \left\{z : \lim_{k\to\infty}\{x \in \mathcal{X} : F(x) > F(X_k^\circ) - \varepsilon\} \subseteq K_\varepsilon\right\},\\
&\quad \text{where } K_\varepsilon \text{ is defined as in Assumption (A4)},\\
\Omega_4(\Delta_k) &:= \left\{z : \max_{x\in\Lambda_k} |\widehat{F}_k(x) - F(x)| < \Delta_k \text{ infinitely often}\right\},\\
&\quad \text{where } \Delta_k \text{ is defined as in Assumption (A5)},\\
\Omega_5(L_k) &:= \left\{z : \lim_{k\to\infty}\left(\alpha^{\frac{k}{2}} L_k \max_{x\in\Lambda_k} |\widehat{F}_k(x) - F(x)|\right) = 0\right\},\\
&\quad \text{where } L_k \text{ as in Assumption (A5)}.
\end{aligned}$$

Whenever we will use the terms "almost surely" or "with probability 1" in the following, we mean that the considered event happens for almost every sample path z.

Recall that in each iteration, the distance between the parameterized family of distributions and an implicit discrete reference distribution is minimized, where the reference distribution in iteration k is given by

$$\widetilde{g}_k(X_i) = \begin{cases} \dfrac{\frac{S\left(\widehat{F}_k(X_i)\right)^k}{\widetilde{f}(X_i,\widehat{\theta}_k)}\,\widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\,\widetilde{\gamma}_k\}}}{\sum\limits_{X_i\in\Lambda_k}\frac{S\left(\widehat{F}_k(X_i)\right)^k}{\widetilde{f}(X_i,\widehat{\theta}_k)}\,\widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\,\widetilde{\gamma}_k\}}} & \forall X_i \in \Lambda_k, \\ & \text{if } \left\{X_i \in \Lambda_k : \widehat{F}_k(X_i) \geq \widetilde{\gamma}_k\right\} \neq \emptyset, \\ \widetilde{g}_{k-1}(X_i) & \forall X_i \in \Lambda_{k-1}, \text{ otherwise.} \end{cases} \tag{4.2}$$

We will also consider a sequence of continuous reference distributions, given by

$$\widehat{g}_k(x) = \frac{S\left(F(x)\right)^k\,\widetilde{\mathbb{1}}_{\{F(x),\,F(X_k^\circ)\}}}{\int_{\mathcal{X}} S\left(F(x)\right)^k\,\widetilde{\mathbb{1}}_{\{F(x),\,F(X_k^\circ)\}}\nu(dx)} \tag{4.3}$$

and we refer to $\widetilde{g}_k$ as *sample reference distribution* and to $\widehat{g}_k$ as *idealized reference distribution*. This idealized distribution differs slightly from the one in [CFHM07] (which is defined using the modified indicator function $\widetilde{\mathbb{1}}_{\{F(x),\, F(X^\circ_{k-1})\}}$), however we find this approach more intuitive because is is closer to the original idea (cf. (3.3)). The theoretical results are easily adapted.
Note that $\widetilde{\gamma}_k = \widehat{F}_k(X^\circ_k)$ and that thus one major difference between $\widetilde{g}$ and $\widehat{g}$ is the use of correct function values F in the idealized distribution instead of the estimates $\widehat{F}$.

To abbreviate notations, we define $H_k(F(x)) := \frac{S(F(x))^k}{\widetilde{f}(x,\widehat{\theta}_k)}$. As before, we denote by $\mathbb{E}_\theta$ and $\mathbb{E}_g$ the expectations with respect to $f(\cdot\,;\theta)$ respectively g and by ν the Lebesgue or counting measure, according to context.

4.3 Results

The main result in this section is that it holds under certain assumptions for the parameter $\widehat{\theta}_k$ determined by the algorithm MRAS_2 in iteration k

$$\mathbb{E}_{\widehat{\theta}_k}\left[T(X)\right] \to T(x^*) \quad \text{as } k \to \infty,$$

where T is defined as in (4.1). If T is bijective, then we can easily determine x^* as

$$x^* = T^{-1}\left(\lim_{k\to\infty} \mathbb{E}_{\widehat{\theta}_k}\left[T(X)\right]\right).$$

For many exponential families, x is even a component of the vector $T(x)$. In those cases, the result is equivalent to

$$\mathbb{E}_{\widehat{\theta}_k}\left[X\right] \to x^* \quad \text{as } k \to \infty.$$

Example 4.6. (Normal Distribution) For $P_\theta = \mathcal{N}(\mu, \sigma^2)$ it holds that $T(x) = (x, x^2)^\top$ (see Example 1.1). Hence, we have $\mathbb{E}_{\widehat{\theta}_k}\left[X\right] \to x^*$ and $\mathbb{E}_{\widehat{\theta}_k}\left[X^2\right] \to (x^*)^2$ and thus for $k \to \infty$

$$\begin{aligned} \widehat{\mu}_k &\longrightarrow x^*, \\ \widehat{\sigma^2}_k = \mathbb{E}_{\widehat{\theta}_k}\left[X^2\right] - \left(\mathbb{E}_{\widehat{\theta}_k}\left[X\right]\right)^2 &\longrightarrow (x^*)^2 - (x^*)^2 = 0. \end{aligned}$$

Example 4.7. (Binomial Distribution) Let $P_\theta = \text{Bin}(n, \theta)$. Then we know from Example 1.2 that $T(x) = x$. Thus

$$n\widehat{p}_k \longrightarrow x^*.$$

Note however that we do not obtain any result on the variance of the model distribution.

To prove this main theorem, we show successively that almost surely

$$\mathbb{E}_{\widehat{\theta}_k}\left[T(X)\right] = \mathbb{E}_{\widetilde{g}_k}\left[T(X)\right] \to \mathbb{E}_{\widehat{g}_k}\left[T(X)\right] \to T(x^*)$$

as $k \to \infty$.
But first we state some auxiliary results that follow directly from the assumptions. The first is a basic observation that will be needed later on in almost every proof.

Lemma 4.1. *Let Assumptions (C1), (C2) and (D1) be satisfied for some ϕ. Then the sequence $\{X_k^\circ, k \in \mathbb{N}\}$ converges almost surely as $k \to \infty$.*

Proof. Recall that $X_k^\circ = X_{k-1}^\circ$ if Step (3c) is visited in the algorithm. We will show that there exists some $\mathcal{N} \in \mathbb{N}$ such that (3c) will be visited in all iterations $k \geq \mathcal{N}$. Define the event $\mathcal{A}_k := \{$ (3a) or (3b) is visited in iteration $k\}$. It holds: $\mathcal{A}_k$ implies $\{\widehat{F}_k(X_k^\circ) \geq \widetilde{\gamma}_{k-1} + \varepsilon\}$ and since $\widetilde{\gamma}_{k-1} = \widehat{F}_{k-1}(X_{k-1}^\circ)$, we have

$$\mathcal{A}_k \implies \left\{\widehat{F}_k(X_k^\circ) - \widehat{F}_{k-1}(X_{k-1}^\circ) \geq \varepsilon\right\}.$$

Define also the event $\mathcal{B}_k := \{F(X_k^\circ) - F(X_{k-1}^\circ) < \frac{\varepsilon}{2}\}$. Then $\mathcal{A}_k \cap \mathcal{B}_k$ will happen almost surely only a finite number of times, since

$$\begin{aligned} P(\mathcal{A}_k \cap \mathcal{B}_k) &\leq P\left(\left\{\widehat{F}_k(X_k^\circ) - \widehat{F}_{k-1}(X_{k-1}^\circ) \geq \varepsilon\right\} \cap \left\{F(X_k^\circ) - F(X_{k-1}^\circ) < \frac{\varepsilon}{2}\right\}\right) \\ &\leq \sup_{x,y \in \mathcal{X}} P\left(\left\{\widehat{F}_k(x) - \widehat{F}_{k-1}(y) \geq \varepsilon\right\} \cap \left\{F(x) - F(y) < \frac{\varepsilon}{2}\right\}\right) \\ &\leq \sup_{x,y \in \mathcal{X}} P\left(\widehat{F}_k(x) - \widehat{F}_{k-1}(y) + F(x) - F(y) > \frac{\varepsilon}{2}\right). \end{aligned}$$

Using Assumption (C2), we can find a $\mathcal{N}_1 \in \mathbb{N}$ such that

$$P(\mathcal{A}_k \cap \mathcal{B}_k) \leq \phi\left(M_{k-1}, \frac{\varepsilon}{2}\right) \quad \forall\, k \geq \mathcal{N}_1.$$

For $\frac{\varepsilon}{2}$ exists by Assumption (D1) $\mathcal{N}_2 \in \mathbb{N}$ and $\delta = \delta_{\frac{\varepsilon}{2}} \in (0,1)$ such that

$$\phi\left(M_{k-1}, \frac{\varepsilon}{2}\right) \leq \delta^k \quad \forall\, k \geq \mathcal{N}_2.$$

Thus

$$P(\mathcal{A}_k \cap \mathcal{B}_k) \leq \delta^k \quad \forall\, k \geq \mathcal{N} := \max\{\mathcal{N}_1, \mathcal{N}_2\}$$

and therefore

$$\sum_{k=0}^{\infty} P(\mathcal{A}_k \cap \mathcal{B}_k) \leq \mathcal{N} + \sum_{k=\mathcal{N}}^{\infty} \delta^k < \infty.$$

It follows from the lemma of Borel-Cantelli that

$$P(\mathcal{A}_k \cap \mathcal{B}_k \text{ infinitely often}) = 0.$$

Hence, if $\mathcal{A}_k$ happens infinitely often, then almost surely $\mathcal{B}_k^c = \{F(X_k^\circ) - F(X_{k-1}^\circ) \geq \frac{\varepsilon}{2}\}$ happens infinitely often. However, this is a contradiction since $F(x) \leq F(x^*)$ $\forall\, x \in \mathcal{X}$. Thus, we can conclude that almost surely $\mathcal{A}_k$ happens only a finite number of times. This means that there exists $\mathcal{N} \in \mathbb{N}$ such that in all iterations $k \geq \mathcal{N}$ Step (3c) is visited and thus $X_{k+1}^\circ = X_k^\circ$ $\forall\, k \geq \mathcal{N}$. □

With the convergence of $\{X_k^\circ\}$ and an increasing observation size it is clear that also $\widehat{F}_k(X_k^\circ) \to F(X_k^\circ)$ as $k \to \infty$. The next lemma verifies that under certain assumptions this convergence is fast enough.

Lemma 4.2. *Let Assumptions (C1), (C2), (D1) (D2) and (E1) be satisfied for some ϕ without considering Δ_k and L_k in (D2). Then*

$$\lim_{k\to\infty} \alpha^{\frac{k}{2}} |\widehat{F}_k(X_k^\circ) - F(X_k^\circ)| = \lim_{k\to\infty} \alpha^{\frac{k}{2}} |\widetilde{\gamma}_k - F(X_k^\circ)| = 0 \text{ almost surely.}$$

Proof. Let $\zeta > 0$. Then for any $k \in \mathbb{N}$

$$\begin{aligned} P\left(\alpha^{\frac{k}{2}} |\widehat{F}_k(X_k^\circ) - F(X_k^\circ)| > \zeta\right) &= P\left(|\widehat{F}_k(X_k^\circ) - F(X_k^\circ)| > \frac{\zeta}{\alpha^{\frac{k}{2}}}\right) \\ &\leq \sup_{x\in\mathcal{X}} P\left(|\widehat{F}_k(x) - F(x)| > \frac{\zeta}{\alpha^{\frac{k}{2}}}\right). \end{aligned}$$

By Assumptions (C1) and (D2) there exist $\mathcal{N}_1$, $\mathcal{N}_2 \in \mathbb{N}$ and $\delta \in (0,1)$ such that $\forall\, k \geq \max\{\mathcal{N}_1, \mathcal{N}_2\}$

$$\sup_{x\in\mathcal{X}} P\left(|\widehat{F}_k(x) - F(x)| > \frac{\zeta}{\alpha^{\frac{k}{2}}}\right) \leq \phi\left(M_k, \frac{\zeta}{\alpha^{\frac{k}{2}}}\right) \leq \alpha^k \phi\left(M_k, \frac{\zeta}{\alpha^{\frac{k}{2}}}\right) \leq \delta^k,$$

Hence, we can use Borel-Cantelli to conclude that

$$P\left(\left\{\alpha^{\frac{k}{2}} |\widehat{F}_k(X_k^\circ) - F(X_k^\circ)| > \zeta\right\} \text{ infinitely often}\right) = 0,$$

i.e. the probability that $\alpha^{\frac{k}{2}} |\widehat{F}_k(X_k^\circ) - F(X_k^\circ)|$ does not converge to zero is zero and thus the claim follows with probability one. □

Most of the following results only hold true on certain subsets of Ω. This lemma gives an overview over the assumptions needed for the subsets to be large enough to imply the respective claims on the whole of Ω almost surely.

Lemma 4.3. *We have*

(i) $P(\Omega_2) = 1$ *if (C1), (C2) and (D1) are satisfied for some ϕ.*

(ii) $P(\Omega_3) = 1$ *if (A4) is satisfied.*

(iii) If (A5), (C1) and (D2) are satisfied for some pair (Δ_k, L_k) and some ϕ, then $P(\Omega_4(\Delta_k)) = 1$.

(iv) If (A5), (C1) and (D2) are satisfied for some pair (Δ_k, L_k) and some ϕ, then $P(\Omega_5(L_k)) = 1$.

Proof.

(i) Follows directly from the proof of Lemma 4.1.

(ii) Assumption (A4).

(iii) Let Δ_k as in Assumption (A5) and consider

$$\begin{aligned} P\left(\max_{x\in\Lambda_k}|\widehat{F}_k(x)-F(x)|\geq\Delta_k\right) &\leq P\left(\bigcup_{x\in\Lambda_k}\left\{|\widehat{F}_k(x)-F(x)|\geq\Delta_k\right\}\right) \\ &\leq \sum_{x\in\Lambda_k} P\left(|\widehat{F}_k(x)-F(x)|\geq\Delta_k\right) \\ &\leq |\Lambda_k|\sup_{x\in\mathcal{X}} P\left(|\widehat{F}_k(x)-F(x)|\geq\Delta_k\right). \end{aligned}$$

To assess $|\Lambda_k|$, Hu, Fu and Marcus use the bound $N_k \leq \alpha^k N_0$. This is not quite correct, since we set $N_k = \lceil \alpha N_{k-1} \rceil$ each time Step (3c) is visited. Consider for examples $\alpha = 1.05$, $N_0 = 50$ and assume that in the first two iterations (3c) is visited. Then $\alpha^2 N_0 = 55.125$, but $N_2 = \lceil \alpha \cdot \lceil \alpha N_0 \rceil \rceil = \lceil \alpha \lceil 52.5 \rceil \rceil = \lceil \alpha \cdot 53 \rceil = \lceil 55.65 \rceil = 56 > \alpha^2 N_0$. However, since in each iteration the mistake due to rounding is bounded by 1, we can find some constant c such that $N_k \leq \alpha^k (N_0 + c)$ for all k.

It follows from Assumption (C1) that there exists $\mathcal{N}_1 \in \mathbb{N}$ such that

$$P\left(\max_{x\in\Lambda_k}|\widehat{F}_k(x)-F(x)|\geq\Delta_k\right) \leq \alpha^k (N_0+c)\,\phi(M_k,\Delta_k) \quad \forall\, k \geq \mathcal{N}_1.$$

Since $\phi(\cdot,\cdot)$ is supposed to be non-increasing in the second argument, there exists by Assumption (D2) $\delta \in (0,1)$ and $\mathcal{N}_2 \in \mathbb{N}$ such that

$$\alpha^k \phi(M_k,\Delta_k) \leq \alpha^k \phi\left(\min\left\{\Delta_k, \frac{\zeta}{\alpha^{\frac{k}{2}}}, \frac{\zeta}{\alpha^{\frac{k}{2}} L_k}\right\}\right) \leq \delta^k \quad \forall\, k \geq \mathcal{N}_2, \forall\, \zeta > 0.$$

Thus, we can conclude that

$$P\left(\max_{x\in\Lambda_k}|\widehat{F}_k(x)-F(x)|\geq\Delta_k\right) \leq (N_0+c)\,\delta^k \quad \forall\, k \geq \mathcal{N} := \min\{\mathcal{N}_1,\mathcal{N}_2\},$$

therefore

$$\sum_{k=0}^{\infty} P\left(\max_{x\in\Lambda_k}|\widehat{F}_k(x)-F(x)|\geq\Delta_k\right) \leq \mathcal{N} + (N_0+c)\sum_{k=\mathcal{N}}^{\infty}\delta^k < \infty$$

and hence by Borel-Cantelli

$$P\left(\max_{x\in\Lambda_k}|\widehat{F}_k(x)-F(x)|\geq\Delta_k \text{ infinitely often}\right) = 0.$$

If almost surely the event $\{\max_{x\in\Lambda_k}|\widehat{F}_k(x)-F(x)|\geq\Delta_k\}$ happens only a finite number of times, then with probability 1 the event $\{\max_{x\in\Lambda_k}|\widehat{F}_k(x)-F(x)|<\Delta_k\}$ must happen for almost every k. Thus $P(\Omega_4(\Delta_k)) = 1$.

(iv) Let L_k as in Assumption (A5). For an arbitrary $\zeta > 0$, we have

$$P\left(\alpha^{\frac{k}{2}} L_k \max_{x\in\Lambda_k} |\widehat{F}_k(x) - F(x)| \geq \zeta\right) \leq \sum_{x\in\Lambda_k} P\left(|\widehat{F}_k(x) - F(x)| \geq \frac{\zeta}{\alpha^{\frac{k}{2}} L_k}\right)$$

and as before, by using Borel-Cantelli and the Assumptions (C1) and (D2), we can conclude that the event $\left\{\alpha^{\frac{k}{2}} L_k \max_{x\in\Lambda_k} |\widehat{F}_k(x) - F(x)| \geq \zeta\right\}$ can almost surely happen only finitely often. This means that with probability 1 there exists some $\mathcal{N} \in \mathbb{N}$ with

$$\alpha^{\frac{k}{2}} L_k \max_{x\in\Lambda_k} |\widehat{F}_k(x) - F(x)| < \zeta \quad \forall\, k \geq \mathcal{N}$$

and, since ζ was arbitrary, $P(\Omega_5(L_k)) = 1$.

□

The following two propositions are the first two steps in the proof of the main theorem, as mentioned in the introductory paragraph of this section.

Proposition 4.4. *Assume (B1)-(B3). Then*

$$\mathbb{E}_{\widehat{\theta}_{k+1}}[T(X)] = \mathbb{E}_{\widetilde{g}_k}[T(X)] \quad \forall\, k \in \mathbb{N}.$$

Proof. We can assume that

$$\widetilde{g}_k(X_i) = \frac{H_k(\widehat{F}_k(X_i))\, \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\, \widetilde{\gamma}_k\}}}{\sum\limits_{X_i\in\Lambda_k} H_k(\widehat{F}_k(X_i))\, \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\, \widetilde{\gamma}_k\}}},$$

i.e. that $\left\{X_i \in \Lambda_k : \widehat{F}_k(X_i) \geq \widetilde{\gamma}_k\right\}$ is not empty, since otherwise we can find some $l \in \mathbb{N}$ such that $\widetilde{g}_k = \widetilde{g}_{k-1} = \cdots = \widetilde{g}_{k-l}$ has the above form. Consider for a fixed N_k the function

$$G_k(\theta) := \frac{1}{N_k} \sum_{i=1}^{N_k} H_k(\widehat{F}_k(X_i)) \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\, \widetilde{\gamma}_k\}} \ln f(X_i, \theta)$$

and recall that $\widehat{\theta}_{k+1} = \operatorname{argmax}_\theta G_k(\theta)$. Using the fact that $f(\cdot\,\theta)$ is a density of the form 4.1, we can write

$$\begin{aligned} G_k(\theta) = &\frac{1}{N_k} \sum_{i=1}^{N_k} H_k(\widehat{F}_k(X_i)) \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\, \widetilde{\gamma}_k\}} \ln h(X_i) \\ &+ \frac{1}{N_k} \sum_{i=1}^{N_k} H_k(\widehat{F}_k(X_i)) \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\, \widetilde{\gamma}_k\}} \theta^\top T(X_i) \\ &- \frac{1}{N_k} \sum_{i=1}^{N_k} H_k(\widehat{F}_k(X_i)) \widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\, \widetilde{\gamma}_k\}} \ln\left(\int_{\mathcal{X}} h(y) \exp\{\theta^\top T(y)\} \nu(\mathrm{d}y)\right). \end{aligned}$$

Since

$$\begin{aligned}\|\nabla_\theta\, h(y)\exp\{\theta^\top T(y)\}\| &= \|h(y)T(y)\exp\{\theta^\top T(y)\}\| \\ &\leq \sup_{\theta\in\Theta}\|h(y)T(y)\exp\{\theta^\top T(y)\}\|,\end{aligned}$$

which is integrable by Assumption (B2), we can interchange the derivative and the integral and have

$$\begin{aligned}\nabla_\theta\, G_k(\theta) = \frac{1}{N_k}\sum_{i=1}^{N_k} H_k(\widehat{F}_k(X_i))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}}T(X_i) \\ - \frac{\int_{\mathcal{X}} T(y)h(y)\exp\{\theta^\top T(y)\}\nu(\mathrm{d}y)}{\int_{\mathcal{X}} h(y)\exp\{\theta^\top T(y)\}\nu(\mathrm{d}y)}\frac{1}{N_k}\sum_{i=1}^{N_k} H_k(\widehat{F}_k(X_i))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}}.\end{aligned}$$

This gradient equals zero if and only if

$$\begin{aligned}\frac{\frac{1}{N_k}\sum_{i=1}^{N_k} H_k(\widehat{F}_k(X_i))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}}T(X_i)}{\frac{1}{N_k}\sum_{i=1}^{N_k} H_k(\widehat{F}_k(X_i))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(X_i),\widetilde{\gamma}_k\}}} &= \frac{\int_{\mathcal{X}} T(y)h(y)\exp\{\theta^\top T(y)\}\nu(\mathrm{d}y)}{\int_{\mathcal{X}} h(y)\exp\{\theta^\top T(y)\}\nu(\mathrm{d}y)} \\ &= \int_{\mathcal{X}} T(y)h(y)\exp\{\theta^\top T(y) - A(\theta)\}\nu(\mathrm{d}y)\end{aligned}$$

$$\iff \quad \mathbb{E}_{\widetilde{g}_k}[T(X)] = \mathbb{E}_\theta[T(X)].$$

And since $\widehat{\theta}_{k+1}$ is by Assumption (B3) an interior point of Θ, we have $\nabla_\theta\, G_k(\widehat{\theta}_{k+1}) = 0$ which implies that

$$\mathbb{E}_{\widetilde{g}_k}[T(X)] = \mathbb{E}_{\widehat{\theta}_{k+1}}[T(X)].$$

□

Proposition 4.5. *Let Assumptions (A1)-(A4), (B1) and (C1)-(D1) be satisfied for some ϕ. Then*

$$\lim_{k\to\infty}\mathbb{E}_{\widehat{g}_k}[T(X)] = T(x^*) \text{ almost surely.}$$

Proof. We will show that the claim is true for all $z \in \Omega_2 \cap \Omega_3$. Since $P(\Omega_2) = P(\Omega_3) = 1$, this implies that it is true for almost all z.
First, we will show some auxiliary results:
For all $z \in \Omega_2$, there exists $\mathcal{N}_1 \in \mathbb{N}$ such that $X^\circ_{k+1} = X^\circ_k = X^\circ_{\mathcal{N}_1}$ for all $k \geq \mathcal{N}_1$. For all $z \in \Omega_3$, there exists $\mathcal{N}_2 \in \mathbb{N}$ such that $\{x \in \mathcal{X} : F(x) \geq F(X^\circ_k) - \varepsilon\} \subseteq K_\varepsilon$ for all $k \geq \mathcal{N}_2$. Hence, for all $z \in \Omega_2 \cap \Omega_3$ we have $\{x \in \mathcal{X} : F(x) \geq F(X^\circ_{\mathcal{N}}) - \varepsilon\} \subseteq K_\varepsilon$ for all $k \geq \mathcal{N} := \max\{\mathcal{N}_1, \mathcal{N}_2\}$.

Hence, for all $z \in \Omega_2 \cap \Omega_3$ and for all $k \geq \mathcal{N}$, we can represent $\widehat{g}_k$ recursively as

$$\begin{aligned}
\widehat{g}_{k+1}(x) &= \frac{S\left(F(x)\right)^{k+1} \widetilde{\mathbb{1}}_{\{F(x),\, F(X^{\circ}_{\mathcal{N}})\}}}{\int_{\mathcal{X}} S\left(F(x)\right)^{k+1} \widetilde{\mathbb{1}}_{\{F(x),\, F(X^{\circ}_{\mathcal{N}})\}} \nu(dx)} \\
&= \frac{S\left(F(x)\right) \cdot S\left(F(x)\right)^{k} \widetilde{\mathbb{1}}_{\{F(x),\, F(X^{\circ}_{\mathcal{N}})\}}}{\int_{\mathcal{X}} S\left(F(x)\right)^{k+1} \widetilde{\mathbb{1}}_{\{F(x),\, F(X^{\circ}_{\mathcal{N}})\}} \nu(dx) \cdot \frac{\int_{\mathcal{X}} S(F(x))^{k} \widetilde{\mathbb{1}}_{\{F(x),\, F(X^{\circ}_{\mathcal{N}})\}} \nu(dx)}{\int_{\mathcal{X}} S(F(x))^{k} \widetilde{\mathbb{1}}_{\{F(x),\, F(X^{\circ}_{\mathcal{N}})\}} \nu(dx)}} \\
&= \frac{S\left(F(x)\right) \cdot \frac{S(F(x))^{k} \widetilde{\mathbb{1}}_{\{F(x),\, F(X^{\circ}_{\mathcal{N}})\}}}{\int_{\mathcal{X}} S(F(x))^{k} \widetilde{\mathbb{1}}_{\{F(x),\, F(X^{\circ}_{\mathcal{N}})\}} \nu(dx)}}{\frac{\int_{\mathcal{X}} S(F(x)) \cdot S(F(x))^{k} \widetilde{\mathbb{1}}_{\{F(x),\, F(X^{\circ}_{\mathcal{N}})\}} \nu(dx)}{\int_{\mathcal{X}} S(F(x))^{k} \widetilde{\mathbb{1}}_{\{F(x),\, F(X^{\circ}_{\mathcal{N}})\}} \nu(dx)}} \\
&= \frac{S\left(F(x)\right)\ \widehat{g}_k}{\mathbb{E}_{\widehat{g}_k}[S\left(F(x)\right)]} \qquad (4.4) \\
&= \left(\prod_{i=\mathcal{N}}^{k} \frac{S\left(F(x)\right)}{\mathbb{E}_{\widehat{g}_i}[S\left(F(x)\right)]} \right) \cdot \widehat{g}_{\mathcal{N}}(x). \qquad (4.5)
\end{aligned}$$

Using representation (4.4) and Jensen's inequality, it is easy to see that

$$\mathbb{E}_{\widehat{g}_{k+1}}[S\left(F(x)\right)] = \frac{\mathbb{E}_{\widehat{g}_k}[S\left(F(x)\right)^2]}{\mathbb{E}_{\widehat{g}_k}[S\left(F(x)\right)]} \geq \mathbb{E}_{\widehat{g}_k}[S\left(F(x)\right)]$$

for all $z \in \Omega_2 \cap \Omega_3$ and for all $k \geq \mathcal{N}$. Since

$$\mathbb{E}_{\widehat{g}_{k+1}}[S\left(F(x)\right)] \leq \mathbb{E}_{\widehat{g}_{k+1}}[S\left(F(x^*)\right)] = S(F(x^*)),$$

the sequence $\{\mathbb{E}_{\widehat{g}_k}\}$ converges for $k \to \infty$.
Assume now that $\lim_{k\to\infty} \mathbb{E}_{\widehat{g}_{k+1}}[S\left(F(x)\right)] =: S_* < S(F(x^*)) =: S^*$ and define the set

$$\begin{aligned}
\mathcal{A} &:= \left\{ x \in \mathcal{X} : F(x) \geq F(X^{\circ}_{\mathcal{N}}) - \varepsilon \right\} \cap \left\{ x \in \mathcal{X} : S(F(x)) \geq \frac{S_* - S^*}{2} \right\} \\
&= \left\{ x \in \mathcal{X} : F(x) \geq \max \left\{ F(X^{\circ}_{\mathcal{N}}) - \varepsilon,\ S^{-1}\left(\frac{S_* - S^*}{2} \right) \right\} \right\}.
\end{aligned}$$

By Assumption (A2), $\mathcal{A}$ has a strictly positive measure and also $\widehat{g}_{\mathcal{N}}(x) > 0$ for all $x \in \mathcal{A}$. For these points holds moreover that $S(F(x)) > S_*$ and thus $\lim_{k\to\infty} \frac{S(F(x))}{\mathbb{E}_{\widehat{g}_k}[S(F(x))]} = \frac{S(F(x))}{S_*} > 1$. Considering (4.5), this means that $\liminf_{k\to\infty} \widehat{g}_k(x) = \infty$ for all $x \in \mathcal{A}$, which leads to a contradiction since by Fatou's lemma

$$1 = \liminf_{k\to\infty} \int_{\mathcal{X}} \widehat{g}_k(x)\nu(\mathrm{d}x) \geq \int_{\mathcal{X}} \liminf_{k\to\infty} \widehat{g}_k(x)\nu(\mathrm{d}x) = \infty.$$

Hence, we can conclude that $\lim_{k\to\infty} \mathbb{E}_{\widehat{g}_{k+1}}[S\left(F(x)\right)] = S(F(x^*))$ for all $z \in \Omega_2 \cap \Omega_3$.

Now we can consider $\|\mathbb{E}_{\widehat{g}_k}[T(X)] - T(x^*)\|$:
Define $\mathcal{G} := \{x \in \mathcal{X} : F(x) \geq F(X^{\circ}_{\mathcal{N}}) - \varepsilon\}$ and note that this is the support of $\widehat{g}_k$

for all $z \in \Omega_2 \cap \Omega_3$ and for all $k \geq \mathcal{N}$. Since $T(x)$ is by definition continuous, there exists for any $\zeta > 0$ some $\delta > 0$ such that $\|T(x) - T(x^*)\| \leq \zeta \; \forall \, x : \|x - x^*\| \leq \delta$. Define $\mathcal{A}_\delta := \{x \in \mathcal{X} : \|x - x^*\| \geq \delta\}$. Then we have for all $z \in \Omega_2 \cap \Omega_3$ and $k \geq \mathcal{N}$:

$$\begin{aligned}
\|\mathbb{E}_{\widehat{g}_k}[T(X)] - T(x^*)\| &\leq \int_{\mathcal{X}} \|T(x) - T(x^*)\| \, \widehat{g}_k(x)\nu(\mathrm{d}x) \\
&= \int_{\mathcal{G}} \|T(x) - T(x^*)\| \, \widehat{g}_k(x)\nu(\mathrm{d}x) \\
&= \int_{\mathcal{G}\cap\mathcal{A}_\delta^c} \|T(x) - T(x^*)\| \, \widehat{g}_k(x)\nu(\mathrm{d}x) \\
&\quad + \int_{\mathcal{G}\cap\mathcal{A}_\delta} \|T(x) - T(x^*)\| \, \widehat{g}_k(x)\nu(\mathrm{d}x) \\
&\leq \zeta + \int_{\mathcal{G}\cap\mathcal{A}_\delta} \|T(x) - T(x^*)\| \, \widehat{g}_k(x)\nu(\mathrm{d}x) \\
&\leq \zeta + \sup_{\mathcal{G}\cap\mathcal{A}_\delta} \|T(x) - T(x^*)\| \int_{\mathcal{G}\cap\mathcal{A}_\delta} \widehat{g}_k(x)\nu(\mathrm{d}x).
\end{aligned}$$

Since $\mathcal{G}$ is a subset of the compact set K_ε by Assumption (A4) and T is continuous, $\sup_{\mathcal{G}\cap\mathcal{A}_\delta} \|T(x) - T(x^*)\| < \infty$.
Note that by (A3) it holds that $\sup_{x\in\mathcal{A}_\delta\cap\mathcal{G}} F(x) \leq \sup_{x\in\mathcal{A}_\delta} F(x) < F(x^*)$ and since S is monotone

$$S(F(x)) \leq S\left(\sup_{x\in\mathcal{A}_\delta\cap\mathcal{G}} F(x)\right) < S(F(x^*)) \quad \forall \, x \in \mathcal{A}_\delta \cap \mathcal{G}.$$

Moreover, since we have established that $\mathbb{E}_{\widehat{g}_{k+1}}[S\,(F(x))] \to S(F(x^*))$ as $k \to \infty$ for all $z \in \Omega_2 \cap \Omega_3$, we can find some $\mathcal{N}_3 \geq \mathcal{N} \in \mathbb{N}$ such that $\mathbb{E}_{\widehat{g}_{k+1}}[S\,(F(x))] > S(\sup_{x\in\mathcal{A}_\delta\cap\mathcal{G}} F(x)) \; \forall \, k \geq \mathcal{N}_3$.
Hence for all $z \in \Omega_2 \cap \Omega_3$ there exists $c > 0$ with

$$\frac{S(F(x))}{\mathbb{E}_{\widehat{g}_{k+1}}[S\,(F(x))]} \leq c < 1 \quad \forall x \in \mathcal{A}_\delta \cap \mathcal{G}, \forall \, k \geq \mathcal{N}_3$$

and thus

$$\widehat{g}_k(x) = \left(\prod_{i=\mathcal{N}_3}^{k} \frac{S\,(F(x))}{\mathbb{E}_{\widehat{g}_i}[S\,(F(x))]}\right) \cdot \widehat{g}_{\mathcal{N}_3}(x) \leq c^{k-\mathcal{N}_3+1} \cdot \widehat{g}_{\mathcal{N}_3}(x) \quad \forall x \in \mathcal{A}_\delta \cap \mathcal{G}.$$

This means that $\widehat{g}_k(x) \leq \zeta \, \widehat{g}_{\mathcal{N}_3}(x)$ for all $x \in \mathcal{A}_\delta \cap \mathcal{G}$ and for all k large enough (say $k \geq \mathcal{N}_4 \geq \mathcal{N}_3$). Since

$$\int_{\mathcal{G}\cap\mathcal{A}_\delta} \widehat{g}_k(x)\nu(\mathrm{d}x) \leq \int_{\mathcal{G}} \widehat{g}_k(x)\nu(\mathrm{d}x) \leq 1,$$

we can conclude the proof by observing that for all $z \in \Omega_2 \cap \Omega_3$ and for an arbitrary $\zeta > 0$ we have found $\mathcal{N}_4 \in \mathbb{N}$ such that for all $k \geq \mathcal{N}_4$

$$\begin{aligned}\|\mathbb{E}_{\widehat{g}_k}[T(X)] - T(x^*)\| &\leq \zeta + \sup_{\mathcal{G}\cap\mathcal{A}_\delta} \|T(x) - T(x^*)\| \int_{\mathcal{G}\cap\mathcal{A}_\delta} \widehat{g}_k(x)\nu(\mathrm{d}x) \\ &\leq \zeta + \sup_{\mathcal{G}\cap\mathcal{A}_\delta} \|T(x) - T(x^*)\| \int_{\mathcal{G}\cap\mathcal{A}_\delta} \zeta\, \widehat{g}_{\mathcal{N}_3}(x)\nu(\mathrm{d}x) \\ &\leq (1 + \sup_{\mathcal{G}\cap\mathcal{A}_\delta} \|T(x) - T(x^*)\|)\, \zeta.\end{aligned}$$

Hence, $\mathbb{E}_{\widehat{g}_k}[T(X)] \to T(x^*)$ on $\Omega_2 \cap \Omega_3$.

□

To prove the third and final step towards the main theorem, we need the following two results:

Lemma 4.6. *Let (A4), (B1), (B4), (C1)-(D1) and (E1) be satisfied for some ϕ. Then with probability 1 as $k \to \infty$*

(i) $\frac{1}{N_k} \sum_{x\in\Lambda_k} \varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{F(x),\, F(X_k^\circ)\}} \longrightarrow \mathbb{E}_{\widetilde{f}(\cdot,\,\widehat{\theta}_k)} \left[\varphi^k H_k(F(X))\widetilde{\mathbb{1}}_{\{F(X),\, F(X_k^\circ)\}}\right]$,

(ii) $\frac{1}{N_k} \sum_{x\in\Lambda_k} \varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{F(x),\, F(X_k^\circ)\}} T(x) \longrightarrow \mathbb{E}_{\widetilde{f}(\cdot,\,\widehat{\theta}_k)} \left[\varphi^k H_k(F(X))\widetilde{\mathbb{1}}_{\{F(X),\, F(X_k^\circ)\}} T(X)\right]$.

Proof.

(i) We consider only $z \in \Omega_2$, since under (C1), (C2) and (D1) $P(\Omega_2) = 1$.
Let $\zeta > 0$ small (cf. (1.3)) and define the event

$$\mathcal{Q}_k := \left\{ \left| \frac{1}{N_k} \sum_{x\in\Lambda_k} \varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{F(x),\, F(X_k^\circ)\}} \;-\; \mathbb{E}_{\widetilde{f}(\cdot,\,\widehat{\theta}_k)} \left[\varphi^k H_k(F(X))\widetilde{\mathbb{1}}_{\{F(X),\, F(X_k^\circ)\}}\right] \right| > \zeta \right\}.$$

We will show that $\mathcal{Q}_k$ almost surely happens only finitely often. To do so, we define for shorthand notation also

$$\mathcal{U}_k := \Big\{ \text{ Step (3a) and (3b) have been visited at most } \sqrt{k} \text{ times at iteration } k. \Big\},$$
$$\mathcal{W}_k := \Big\{ \{x \in \mathcal{X} : F(x) \geq F(X_k^\circ) - \varepsilon\} \subseteq K_\varepsilon \Big\},$$

where K_ε is defined as in Assumption (A4).
It holds that $P(\mathcal{U}_k^c \text{ infinitely often}) = 0$ (see the proof of Lemma 4.1) and $P(\mathcal{W}_k^c \text{ infinitely often}) = 0$ by Assumption (A4). Hence,

$$\begin{aligned}P(\mathcal{Q}_k \text{ infinitely often}) &= P(\,\{\mathcal{Q}_k \cap \mathcal{U}_k\} \cup \{\mathcal{Q}_k \cap \mathcal{U}_k^c\} \text{ infinitely often}) \\ &= P(\,\{\mathcal{Q}_k \cap \mathcal{U}_k\} \text{ infinitely often}) \\ &= P(\,\{\mathcal{Q}_k \cap \mathcal{U}_k \cap \mathcal{W}_k\} \cup \{\mathcal{Q}_k \cap \mathcal{U}_k \cap \mathcal{W}_k^c\} \text{ infinitely often}) \\ &= P(\,\{\mathcal{Q}_k \cap \mathcal{U}_k \cap \mathcal{W}_k\} \text{ infinitely often})\end{aligned}$$

Since we have assumed in (B4) that $f_* := \inf_{x \in K_\varepsilon} f(x, \theta_0) > 0$ and all samples $X_i \in \Lambda_k$ are drawn with probability $\widetilde{f}(\cdot\,; \widehat{\theta}_k) = (1-\lambda) f(\cdot\,; \widehat{\theta}_k) + \lambda f(\cdot\,; \theta_0)$, we know that for all $X_i \in \Lambda_k$ and for all $z \in \mathcal{W}_k$

$$\begin{aligned} 0 \;&\leq\; \varphi^k H_k(F(X_i)) \widetilde{\mathbb{1}}_{\{F(X_i),\, F(X_k^\circ)\}} \\ &=\; \varphi^k \, \frac{S(F(X_i))^k}{\widetilde{f}(X_i; \widehat{\theta}_k)} \, \widetilde{\mathbb{1}}_{\{F(X_i),\, F(X_k^\circ)\}} \;\leq\; \frac{(\varphi S^*)^k}{\lambda f_*}. \end{aligned} \tag{4.6}$$

Ω_2 implies that there exists $\mathcal{N}_1 \in \mathbb{N}$ such that $X_k^\circ = X_{k-1}^\circ$ for all $k \geq \mathcal{N}_1$. Hence for these k, $X_1^{(k)}, \ldots, X_{N_k}^{(k)}$ and thus also $\varphi^k H_k(F(X_i^{(k)})) \widetilde{\mathbb{1}}_{\{F(X_i^{(k)}),\, F(X_k^\circ)\}}$ $= \varphi^k H_k(F(X_i^{(k)})) \widetilde{\mathbb{1}}_{\{F(X_i^{(k)}),\, F(X_{k-1}^\circ)\}}$, $i = 1, \ldots, N_k$, are conditional on $\widehat{\theta}_{k-1}$ and $F(X_{k-1}^\circ)$ independent and identically distributed. Then the Hoeffding inequality yields

$$P\left(\mathcal{Q}_k \,\middle|\, \mathcal{W}_k, \widehat{\theta}_{k-1} = \theta, F(X_{k-1}^\circ) = \gamma, N_k = n\right) \leq 2 \exp\left\{\frac{-2\,n\,\zeta^2 \lambda^2 f_*^2}{(\varphi S^*)^{2k}}\right\} \quad \forall\, k \geq \mathcal{N}.$$

Then if $P_{\widehat{\theta}_{k-1}, F(X_k^\circ), N_k}$ designs the joint distribution of $\widehat{\theta}_{k-1}, F(X_k^\circ)$ and N_k (appropriately defined to allow for both the continuous and the discrete random variables), we have for all $k \geq \mathcal{N}_1$

$$\begin{aligned} P(\mathcal{Q}_k \cap \mathcal{W}_k) &= \int P(\mathcal{Q}_k \cap \mathcal{W}_k | \, \widehat{\theta}_{k-1} = \theta, F(X_k^\circ) = \gamma, N_k = n) \, \mathrm{d}P_{\widehat{\theta}_{k-1}, F(X_k^\circ), N_k}(\theta, \gamma, n) \\ &\leq \int P(\mathcal{Q}_k \,|\, \mathcal{W}_k, \widehat{\theta}_{k-1} = \theta, F(X_k^\circ) = \gamma, N_k = n) \, \mathrm{d}P_{\widehat{\theta}_{k-1}, F(X_k^\circ), N_k}(\theta, \gamma, n) \\ &\leq \int 2 \exp\left\{\frac{-2\,n\,\zeta^2 \lambda^2 f_*^2}{(\varphi S^*)^{2k}}\right\} \mathrm{d}P_{\widehat{\theta}_{k-1}, F(X_k^\circ), N_k}(\theta, \gamma, n). \end{aligned}$$

Observe that the event $\mathcal{U}_k$ implies that we visit Step (3c) at least $k - \sqrt{k}$ times and thus that $N_k \geq \alpha^{k-\sqrt{k}} N_0 \;\forall\, k$. Hence we know that

$$P_{\widehat{\theta}_{k-1}, F(X_k^\circ), N_k}\left(\widehat{\theta}_{k-1} = \theta, F(X_k^\circ) = \gamma, N_k < \alpha^{k-\sqrt{k}} N_0 \,\middle|\, \mathcal{U}_k\right) = 0 \quad \forall \theta, \forall \gamma.$$

And since

$$\exp\left\{\frac{-2\,n\,\zeta^2 \lambda^2 f_*^2}{(\varphi S^*)^{2k}}\right\} \leq \exp\left\{\frac{-2\alpha^{k-\sqrt{k}} N_0 \zeta^2 \lambda^2 f_*^2}{(\varphi S^*)^{2k}}\right\} \quad \forall\, n \geq \alpha^{k-\sqrt{k}} N_0,$$

it holds for all $k \geq \mathcal{N}_1$

$$\begin{aligned} P(\mathcal{Q}_k \cap \mathcal{W}_k \cap \mathcal{U}_k) \leq\; & P(\mathcal{Q}_k \cap \mathcal{W}_k \,|\, \mathcal{U}_k) \\ \leq\; & 2 \exp\left\{\frac{-2\alpha^{k-\sqrt{k}} N_0 \zeta^2 \lambda^2 f_*^2}{(\varphi S^*)^{2k}}\right\} \int \mathrm{d}P_{\widehat{\theta}_{k-1}, F(X_k^\circ), N_k}(\theta, \gamma, n) \\ =\; & 2 \exp\left\{\frac{-2 N_0 \zeta^2 \lambda^2 f_*^2}{\alpha^{\sqrt{k}}} \cdot \left(\frac{\alpha}{(\varphi S^*)^2}\right)^k\right\}. \end{aligned}$$

And as $e^{-x} < \frac{1}{x}\ \forall\, x$, we have for all $k \geq \mathcal{N}_1$

$$\begin{aligned} P(\mathcal{Q}_k \cap \mathcal{W}_k \cap \mathcal{U}_k) \ &< \ 2 \cdot \frac{\alpha^{\sqrt{k}}}{2N_0\zeta^2\lambda^2 f_*^2} \cdot \left(\frac{(\varphi S^*)^2}{\alpha}\right)^k \\ &= \ \frac{1}{N_0\zeta^2\lambda^2 f_*^2} \cdot \left(\frac{\alpha^{\frac{\sqrt{k}}{k}}(\varphi S^*)^2}{\alpha}\right)^k. \end{aligned}$$

By (E1) we know $\frac{(\varphi S^*)^2}{\alpha} < 1$. That means that there exists some $\delta \in (0,1)$ and some $\mathcal{N}_2 \in \mathbb{N}$ such that

$$\frac{\alpha^{\frac{\sqrt{k}}{k}}(\varphi S^*)^2}{\alpha} = \alpha^{\frac{1}{\sqrt{k}}} \cdot \frac{(\varphi S^*)^2}{\alpha} < \delta \quad \forall\, k \geq \mathcal{N}_2.$$

Hence

$$P(\mathcal{Q}_k \cap \mathcal{W}_k \cap \mathcal{U}_k) \leq \ const \cdot \delta^k \quad \forall k \geq \max\{\mathcal{N}_1, \mathcal{N}_2\}$$

and we can again use Borel-Cantelli to conclude that

$$P(\mathcal{Q}_k \cap \mathcal{W}_k \cap \mathcal{U}_k \text{ infinitely often}) = P(\mathcal{Q}_k \text{ infinitely often}) = 0,$$

and thus the claim follows with probability 1.

(ii) The proof of (ii) follows from almost the same arguments. However, we cannot as easily conclude (cf. (4.6))

$$0 \ \leq \ \varphi^k H_k(F(X_i))\widetilde{\mathbb{1}}_{\{F(X_i),\, F(X_k^\circ)\}} T(X_i) \ \leq \ \frac{(\varphi S^*)^k}{\lambda f_*}.$$

But by Assumption (B1), $T(x)$ is continuous and thus attains a minimum T_{min} and a maximum T_{max} on the compact set K_ε. Hence we know that for all $X_i \in \Lambda_k$ and for all $z \in \mathcal{W}_k$

$$\begin{aligned} \frac{(\varphi S^*)^k}{\lambda f_*} \cdot \min\{0, T_{min}\} \ &\leq \ \varphi^k H_k(F(X_i))\widetilde{\mathbb{1}}_{\{F(X_i),\, F(X_k^\circ)\}} T(X_i) \\ &\leq \ \frac{(\varphi S^*)^k}{\lambda f_*} \cdot \max\{0, T_{max}\}. \end{aligned}$$

With these bounds for the random variable $\varphi^k H_k(F(X_i))\widetilde{\mathbb{1}}_{\{F(X_i),\, F(X_k^\circ)\}} T(X_i)$, the Hoeffding inequality is applicable. Thus, the rest of the proof is analogical to (i) with an added constant $\frac{1}{\max\{0,T_{max}\} - \min\{0,T_{min}\}}$ in the argument of the exponential function.

□

Lemma 4.7. *Let (A4)-(B1), (B4), (C1)-(D2)and (E1) be satisfied for some (Δ_k, L_k) and some ϕ. Then with probability 1 as $k \to \infty$*

(i) $\frac{1}{N_k} \sum\limits_{x \in \Lambda_k} \varphi^k H_k(\widehat{F}_k(x))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\, \widehat{F}_k(X_k^\circ)\}} \ \longrightarrow \ \frac{1}{N_k} \sum\limits_{x \in \Lambda_k} \varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{F(x),\, F(X_k^\circ)\}},$

(ii) $\frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi^k H_k(\widehat{F}_k(x))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\,\widehat{F}_k(X_k^\circ)\}}T(x) \longrightarrow \frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{F(x),\,F(X_k^\circ)\}}T(x).$

Proof. We only prove (i), since the claim in (ii) can be obtained in the same way as in the proof of Lemma 4.6 (ii). Furthermore, we show that it holds for all $z \in \Omega_3 \cap \Omega_4(\Delta_k) \cap \Omega_5(L_k)$. This implies the convergence in probability 1 since $P(\Omega_3 \cap \Omega_4(\Delta_k) \cap \Omega_5(L_k)) = 1$.
It holds

$$\begin{aligned}
&\left|\frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi^k H_k(\widehat{F}_k(x))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\,\widehat{F}_k(X_k^\circ)\}} - \frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{F(x),\,F(X_k^\circ)\}}\right| \\
\leq\ &\left|\frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi^k H_k(\widehat{F}_k(x))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\,\widehat{F}_k(X_k^\circ)\}} - \frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\,\widehat{F}_k(X_k^\circ)\}}\right| \\
&+\left|\frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\,\widehat{F}_k(X_k^\circ)\}} - \frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{F(x),\,F(X_k^\circ)\}}\right|.
\end{aligned}$$

Consider the first term on the right hand side of the inequality. By (A4) and (B4) there exists $\mathcal{N}_1 \in \mathbb{N}$ such that for all $z \in \Omega_3$ and for all $k \geq \mathcal{N}_1$

$$\begin{aligned}
&\left|\frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi^k H_k(\widehat{F}_k(x))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\,\widehat{F}_k(X_k^\circ)\}} - \frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\,\widehat{F}_k(X_k^\circ)\}}\right| \\
&= \frac{\varphi^k}{N_k}\sum_{x\in\Lambda_k}\left|H_k(\widehat{F}_k(x)) - H_k(F(x))\right|\widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\,\widehat{F}_k(X_k^\circ)\}}\cdot\frac{S\left(F(x)\right)^k}{S\left(F(x)\right)^k} \\
&= \frac{\varphi^k}{N_k}\sum_{x\in\Lambda_k}\frac{\left|S\left(\widehat{F}_k(x)\right)^k - S\left(F(x)\right)^k\right|}{S\left(F(x)\right)^k}\frac{S\left(F(x)\right)^k}{\widetilde{f}(\cdot\,;\widehat{\theta}_k)}\widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\,\widehat{F}_k(X_k^\circ)\}} \\
&\leq \frac{(\varphi S^*)^k}{N_k\lambda f_*}\sum_{x\in\Lambda_k}\frac{\left|S\left(\widehat{F}_k(x)\right)^k - S\left(F(x)\right)^k\right|}{S\left(F(x)\right)^k}\widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\,\widehat{F}_k(X_k^\circ)\}}
\end{aligned}$$

and using (A5) and (E1), we further know that for all $z \in \Omega_3 \cap \Omega_4(\Delta_k)$ this is

$$\begin{aligned}
&\leq \frac{(\varphi S^*)^k}{N_k\lambda f_*}\sum_{x\in\Lambda_k}L_k|\widehat{F}_k(x) - F(x)| \\
&\leq \frac{(\varphi S^*)^k}{\lambda f_*}\max_{x\in\Lambda_k}L_k|\widehat{F}_k(x) - F(x)| \\
&\leq \frac{\alpha^{\frac{k}{2}}}{\lambda f_*}\max_{x\in\Lambda_k}L_k|\widehat{F}_k(x) - F(x)|.
\end{aligned}$$

By the definition of $\Omega_5(L_k)$, this converges to zero for all $z \in \Omega_3 \cap \Omega_4(\Delta_k) \cap \Omega_5(L_k)$.

Similarly, there exists $\mathcal{N}_2 \in \mathbb{N}$ such that for all $z \in \Omega_3$ and all $k \geq \mathcal{N}_2$

$$\begin{aligned}
&\left| \frac{1}{N_k} \sum_{x \in \Lambda_k} \varphi^k H_k(F(x)) \widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\, \widehat{F}_k(X_k^\circ)\}} \;-\; \frac{1}{N_k} \sum_{x \in \Lambda_k} \varphi^k H_k(F(x)) \widetilde{\mathbb{1}}_{\{F(x),\, F(X_k^\circ)\}} \right| \\
&= \frac{\varphi^k}{N_k} \sum_{x \in \Lambda_k} H_k(F(x)) \left| \widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\, \widehat{F}_k(X_k^\circ)\}} - \widetilde{\mathbb{1}}_{\{F(x),\, F(X_k^\circ)\}} \right| \\
&\leq \frac{(\varphi S^*)^k}{N_k \lambda f_*} \sum_{x \in \Lambda_k} \left| \widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\, \widehat{F}_k(X_k^\circ)\}} - \widetilde{\mathbb{1}}_{\{F(x),\, \widehat{F}_k(X_k^\circ)\}} \right| + \left| \widetilde{\mathbb{1}}_{\{F(x),\, \widehat{F}_k(X_k^\circ)\}} - \widetilde{\mathbb{1}}_{\{F(x),\, F(X_k^\circ)\}} \right| \\
&\leq \frac{(\varphi S^*)^k}{N_k \lambda f_*} \sum_{x \in \Lambda_k} \frac{\left| \widehat{F}_k(x) - F(x) \right|}{\varepsilon} + \frac{\left| \widehat{F}_k(X_k^\circ) - F(X_k^\circ) \right|}{\varepsilon} \\
&\leq \frac{\alpha^{\frac{k}{2}}}{\lambda f_* \varepsilon} \max_{x \in \Lambda_k} \left| \widehat{F}_k(x) - F(x) \right| + \frac{\alpha^{\frac{k}{2}}}{\lambda f_* \varepsilon} \left| \widehat{F}_k(X_k^\circ) - F(X_k^\circ) \right|.
\end{aligned}$$

Using Lemma 4.2, (D2) and a similar argument as in Lemma 4.3 (iv), we can show that this converges almost surely to zero.
This implies the claim. □

Proposition 4.8. *Let (A2), (A4), (A5), (B4) and (C1)-(E1) be satisfied for some* (Δ_k, L_k) *and some* ϕ. *Then*

$$\lim_{k \to \infty} \left| \; \mathbb{E}_{\widetilde{g}_k} [T(X)] - \mathbb{E}_{\widehat{g}_k} [T(X)] \; \right| = 0 \text{ almost surely.}$$

Proof. By a change of measure from $\widehat{g}_k$ to $\widetilde{f}(\cdot\,; \widehat{\theta}_k)$, we have

$$\begin{aligned}
\mathbb{E}_{\widehat{g}_k} [T(X)] &= \frac{\int_{\mathcal{X}} T(x) S\left(F(x)\right)^k \widetilde{\mathbb{1}}_{\{F(x), F(X_k^\circ\}} \nu(\mathrm{d}x)}{\int_{\mathcal{X}} S\left(F(x)\right)^k \widetilde{\mathbb{1}}_{\{F(x), F(X_k^\circ\}} \nu(\mathrm{d}x)} \\
&= \frac{\mathbb{E}_{\widetilde{f}(\cdot,\, \widehat{\theta}_k)} \left[T(X) H_k(F(X)) \widetilde{\mathbb{1}}_{\{F(X),\, F(X_k^\circ)\}} \right]}{\mathbb{E}_{\widetilde{f}(\cdot,\, \widehat{\theta}_k)} \left[H_k(F(X)) \widetilde{\mathbb{1}}_{\{F(X),\, F(X_k^\circ)\}} \right]}.
\end{aligned}$$

Thus it suffices to show that

$$\begin{aligned}
&\left| \frac{\sum_{x \in \Lambda_k} T(x) H_k(\widehat{F}_k(x)) \widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\, \widehat{F}_k(X_k^\circ)\}}}{\sum_{x \in \Lambda_k} H_k(\widehat{F}_k(x)) \widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\, \widehat{F}_k(X_k^\circ)\}}} - \frac{\mathbb{E}_{\widetilde{f}(\cdot,\, \widehat{\theta}_k)} \left[T(X) H_k(F(X)) \widetilde{\mathbb{1}}_{\{F(X),\, F(X_k^\circ)\}} \right]}{\mathbb{E}_{\widetilde{f}(\cdot,\, \widehat{\theta}_k)} \left[H_k(F(X)) \widetilde{\mathbb{1}}_{\{F(X),\, F(X_k^\circ)\}} \right]} \right| \\
&\leq \left| \frac{\sum_{x \in \Lambda_k} T(x) \varphi^k H_k(\widehat{F}_k(x)) \widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\, \widehat{F}_k(X_k^\circ)\}}}{\sum_{x \in \Lambda_k} \varphi^k H_k(\widehat{F}_k(x)) \widetilde{\mathbb{1}}_{\{\widehat{F}_k(x),\, \widehat{F}_k(X_k^\circ)\}}} - \frac{\sum_{x \in \Lambda_k} T(x) \varphi^k H_k(F(x)) \widetilde{\mathbb{1}}_{\{F(x),\, F(X_k^\circ)\}}}{\sum_{x \in \Lambda_k} \varphi^k H_k(F(x)) \widetilde{\mathbb{1}}_{\{F(x),\, F(X_k^\circ)\}}} \right| \\
&\quad + \left| \frac{\sum_{x \in \Lambda_k} T(x) \varphi^k H_k(F(x)) \widetilde{\mathbb{1}}_{\{F(x),\, F(X_k^\circ)\}}}{\sum_{x \in \Lambda_k} \varphi^k H_k(F(x)) \widetilde{\mathbb{1}}_{\{F(x),\, F(X_k^\circ)\}}} - \frac{\mathbb{E}_{\widetilde{f}(\cdot,\, \widehat{\theta}_k)} \left[T(X) \varphi^k H_k(F(X)) \widetilde{\mathbb{1}}_{\{F(X),\, F(X_k^\circ)\}} \right]}{\mathbb{E}_{\widetilde{f}(\cdot,\, \widehat{\theta}_k)} \left[\varphi^k H_k(F(X)) \widetilde{\mathbb{1}}_{\{F(X),\, F(X_k^\circ)\}} \right]} \right| \\
&\longrightarrow 0 \text{ almost surely.}
\end{aligned}$$

This follows from Lemma 4.6 and Lemma 4.7, if $\sum_{x\in\Lambda_k} \varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{F(x),\,F(X_k^\circ)\}} > 0$ and $\mathbb{E}_{\widetilde{f}(\cdot,\widehat{\theta}_k)}\left[\varphi^k H_k(F(X))\widetilde{\mathbb{1}}_{\{F(X),\,F(X_k^\circ)\}}\right] > 0$ for all $k \in \mathbb{N}$.

- $\mathbb{E}_{\widetilde{f}(\cdot,\widehat{\theta}_k)}\left[\varphi^k H_k(F(X))\widetilde{\mathbb{1}}_{\{F(X),\,F(X_k^\circ)\}}\right] > 0$:
  Since $F(X_k^\circ)-\varepsilon \leq F(x^*)-\varepsilon < F(x^*)\;\forall\, k$, the set $\{x \in \mathcal{X} : F(x) \geq F(X_k^\circ) - \varepsilon\}$ has by Assumption (A2) a strictly positive measure. On the other hand, φ is by (E1) such that the set $\{x \in \mathcal{X} : S\left(F(x)\right) \geq \frac{1}{\varphi}\}$ has a strictly positive measure. Thus, the measure of $\{x \in \mathcal{X} : F(x) \geq \max\{S^{-1}(\frac{1}{\varphi}), F(X_k^\circ)\}\}$ is also strictly positive for all k. Moreover, for all x in this set holds $[\varphi S(F(x))]^k \geq 1$. This and the lemma of Fatou yields

$$\begin{aligned}
&\liminf_{k\to\infty}\ \mathbb{E}_{\widetilde{f}(\cdot,\widehat{\theta}_k)}\left[\varphi^k H_k(F(X))\,\widetilde{\mathbb{1}}_{\{F(X),\,F(X_k^\circ)\}}\right] \\
&\quad= \liminf_{k\to\infty} \int_{\mathcal{X}} \varphi^k S(F(X))^k\,\widetilde{\mathbb{1}}_{\{F(X),\,F(X_k^\circ)\}}\nu(\mathrm{d}x) \\
&\quad\geq \int_{\mathcal{X}} \liminf_{k\to\infty} \varphi^k S(F(X))^k\,\widetilde{\mathbb{1}}_{\{F(X),\,F(X_k^\circ)\}}\nu(\mathrm{d}x) \\
&\quad> 0.
\end{aligned}$$

- $\sum_{x\in\Lambda_k} \varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{F(x),\,F(X_k^\circ)\}} > 0$:
  We know from Lemma 4.6 and from the calculation above that

$$\begin{aligned}
&\liminf_{k\to\infty}\ \frac{1}{N_k}\sum_{x\in\Lambda_k} \varphi^k H_k(F(x))\widetilde{\mathbb{1}}_{\{F(x),\,F(X_k^\circ)\}} \\
&\quad= \liminf_{k\to\infty}\ \mathbb{E}_{\widetilde{f}(\cdot,\widehat{\theta}_k)}\left[\varphi^k H_k(F(X))\widetilde{\mathbb{1}}_{\{F(X),\,F(X_k^\circ)\}}\right] \\
&\quad> 0.
\end{aligned}$$

Hence, the proof is completed. □

Finally, we can state the main theorem.

Theorem 4.9. *If Assumptions (A1)-(E1) are satisfied for some* (Δ_k, L_k) *and some* ϕ*, then*

$$\lim_{k\to\infty} \mathbb{E}_{\widehat{\theta}_k}\left[T(X)\right] = T(x^*) \text{ almost surely.}$$

Proof. This is now a direct consequence from Proposition 4.4, 4.5 and 4.8. □

Note that apart from the specific choice of reference distributions, the proofs rely heavily on the fact that $N_k, M_k \to \infty$ while $\widetilde{f}(x;\widehat{\theta}_k) > \lambda f(x,\theta_0) > 0\;\forall\, x \in X$. This ensures not only that the probability of drawing an ϵ-optimal sample is always positive for any $\epsilon > 0$ but that this will happen eventually for k large enough and that the near-optimal sample will be recognized as such. Hence a statement difficult to prove for the CE (cf. the beginning of this chapter) follows trivially from the assumptions here.

Chapter 5

Implementational Issues

To test and compare the Cross-Entropy and the Model Reference Adaptive Search algorithm, we implemented a flexible C++-Framework (see attached CD) and studied the algorithms' behaviour in three basic Markov Decision problems, namely a queueing, an inventory and a replacement problem. These examples were chosen for their representative character as they are all standa Figure rd problems in dynamic optimization literature. Moreover, the first two test cases allow for testing the algorithms on uncountable action spaces, a situation which is not as problematic as uncountable state spaces yet rarely considered in literature, even though standard optimization procedures such as Howard's Policy Iteration require not only discrete but even finite action spaces. Often, this problem is overcome in applications by a discretization of the action space. However, the generality of both the CE and the MRAS method leads to the very simple and direct approach of taking a continuous probability distribution over the set of all possible policies. This is often not only easier than working with discrete distributions but also avoids possible mistakes during the discretization process.
Furthermore, it has to be added that both the queueing and the inventory problem have been studied already by Hu, Fu and Marcus (see [CFHM07], [HFM]). There the authors compare the MRAS method with a simple Simulated Annealing algorithm and come to the conclusion that the Model Reference Adaptive Search shows a consistently good performance with quite small variance in the optimal value. We will reconsider exactly the same problems here, since we were not able to replicate these optimistic results. Private communication even revealed that in one of the examples a slight error in the source code seemed to be the reason for the good convergence and that the implementation of the correct algorithm yielded far worse results. A detailed discussion as well as our explanation approaches will be given below.
We restrict ourselves to infinite-horizon problems so that there will be an optimal stationary policy and our solution space has only dimension $|S|$ and not $|S| \times N$ (N denoting the horizon).

5.1 General Discussion

Before we can compare two algorithms, we should establish some criteria for "good" algorithms.
Surely, one of the most important quality features of an optimization algorithm is its ability to find optimal or near-optimal solutions. Moreover, this should happen consistently or at least sufficiently often, depending on the problem to be solved. In some cases (for example if the problem structure and complexity allow for repeated algorithm runs) it may be satisfactory if only most and not all of the found solutions are acceptable. Nevertheless, to compare the performance of two algorithms we will consider both the quality and the reliability of the solutions.
Another important aspect is the time required for an algorithm to terminate. Time refers here to CPU time, i.e. the computational effort needed. We usually measure this by the number of value function evaluations, since in each iteration the effort needed to simulate the candidate policies dominates by far the number of other operations. One should however bear in mind that for the same number of function evaluations, the MRAS algorithms has to perform more basic operations due to its more complicated parameter adaptation (Step 3 in Algorithm 3.3) and the weight factors in the model distribution's parameter update (cf. Example 3.1).
Whereas these two criteria are important in the analysis of all algorithms, we will in our case also be interested in some specific features arising from the behaviour of the two algorithms as well as from their similarities and differences.
Due to the fact that we study the Cross Entropy method and the MRAS for stochastic optimization and replace in both algorithms analytic evaluations by estimates derived from simulation, we have to differentiate between the calculated optimal policy and the respective estimated value function. Our main focus of attention is the performance (i.e. the correct value) of the found policies, but comparing the quality of the value function estimate is also of interest. To avoid confusion, the term "solution" will usually refer to the policy only.
Both algorithms are rather general in nature and are not especially designed for one single problem class. Therefore the ease of their adaptation to a given problem, the complexity of their implementation and the effort needed to find good parameter combinations are also interesting properties in our opinion.

In our experiments we performed several runs (usually 100) for each problem instance and logged in each iteration the number of value function evaluations, the sample quantile (respectively the threshold level in the MRAS) and the estimated value of the best sample. As stopping rule we usually chose a fixed number of function evaluations, i.e. we stopped after iteration T if the total sample size $\sum_{i=0}^{T} N_i \cdot M_i \geq c$ for some $c \in \mathbb{N}$. The motivation for this choice was to obtain comparable solutions of both algorithms. In "normal" applications one would surely use process-dependent stopping rules as discussed at the end of Section 2.4.
To evaluate the average behaviour of the algorithms, we considered both the threshold level and the best sample value as step functions in the total sample size and averaged at a given number of points (e.g. every 5000 samples) over these step functions. Note that for the Cross-Entropy algorithm this method corresponds to

calculating the averages in each iteration if the evaluation points are appropriately chosen, whereas in the MRAS the N_k are not fix so that for different runs the total number of function evaluations in a given iteration will differ. There are other evaluation methods: Hu, Fu and Marcus use for example in [CFHM07] the approach of plotting for each iteration the average values against the average total sample sizes. As solution π^* found in an algorithm run we chose the best-rated sample in the last iteration.

5.2 Queueing

The queueing model, taken from [CFHM07], consists of a capacitated single-server queue where during one period the customer arrival rate follows a Bernoulli distribution with given parameter $p \in [0,1]$ and one customer can be served with a service completion probability $a \in [0,1]$. Any arrival exceeding the queue capacity is lost. In each period a cost $r(i,a)$ occurs, depending on the number of customers in the system i and the service rate a. The problem is to choose the optimal service completion probabilities $a = a(i)$ such that the expected discounted costs for a given initial state i_0 are minimal.
This can be modeled as an infinite-horizon Markov Decision Problem with

- $S = \{0, 1, \ldots, L\}$. $i_t \in S$ the number of customers in the system in period t, $L \in \mathbb{N}$ the queue capacity.
- $A = [0,1]$. $a \in A$ the service completion rate, $D(i) = A$.
- $p(i,a,j) = \begin{cases} p(1-a) & j = i+1 \\ (1-p)a & j = i \\ (1-p)(1-a) + pa & j = i-1 \end{cases}$ for $i \in \{1, \ldots, L-1\}$
  and $p(0,a,j) = \begin{cases} 1-p & j = i \\ p & j = i+1 \end{cases}$, $p(L,a,j) = \begin{cases} 1-a & j = i \\ a & j = i-1 \end{cases}$ the transition probabilities.
- a cost function $r(i,a)$.
- $\beta < 1$ the discount factor.

Note that since the policy space $F = [0,1]^{|S|}$ is uncountable, it is reasonable to use a continuous parameterized family in both the CE and the MRAS algorithm.

Hu, Fu and Marcus [CFHM07] chose to employ independent univariate normal densities in each component, truncated between 0 and 1. That is, the probability of choosing action $a \in [0,1]$ when in state $i \in S$ in iteration k is determined by the density

$$f(a; \mu_k(i), \sigma_k^2(i)) = \begin{cases} \frac{\frac{1}{\sqrt{2\pi\sigma_k^2(i)}} \exp\left\{-\frac{1}{2}\frac{(a-\mu_k(i))^2}{\sigma_k^2(i)}\right\}}{\Phi\left(\frac{1-\mu_k(i)}{\sigma_k(i)}\right) - \Phi\left(\frac{-\mu_k(i)}{\sigma_k(i)}\right)} & a \in [0,1] \\ 0 & otherwise \end{cases},$$

where Φ is the standard normal distribution. Then the density of the distribution of a policy $\pi : S \to [0,1]$ is the product $\prod_{i\in S} f(\pi(i); \mu_k(i), \sigma_k^2(i))$. Moreover, as in all their examples, they employ as positive function $S(y) = e^{-\tau y}$ for $y > 0$ with some constant $\tau > 0$. As we consider the equivalent maximization problem, we have $V_{\infty\pi} < 0$ and hence we use $S(y) = e^{\tau y}$ for $\tau > 0$. Note that $S(y) \in [0,1]\ \forall\, y$. Then the update in Step 5 of Algorithm 3.3 becomes

$$\widehat{\mu}_{k+1}(i) = \frac{\sum_{\pi\in\Lambda_k} \frac{\exp\{k\tau \widehat{V}_\infty^{\pi,k}\}}{\prod_{i\in S} \widetilde{f}(\pi(i);\widehat{\mu}_k(i),\widehat{\sigma}_k^2(i))} \widetilde{\mathbb{1}}_{\{\widehat{V}_\infty^{\pi,k},\widetilde{\gamma}_k\}} \pi(i)}{\sum_{\pi\in\Lambda_k} \frac{\exp\{k\tau \widehat{V}_\infty^{\pi,k}\}}{\prod_{i\in S} \widetilde{f}(\pi(i);\widehat{\mu}_k(i),\widehat{\sigma}_k^2(i))} \widetilde{\mathbb{1}}_{\{\widehat{V}_\infty^{\pi,k},\widetilde{\gamma}_k\}}}, \tag{5.1}$$

$$\widehat{\sigma}_{k+1}(i) = \frac{\sum_{\pi\in\Lambda_k} \frac{\exp\{k\tau \widehat{V}_\infty^{\pi,k}\}}{\prod_{i\in S} \widetilde{f}(\pi(i);\widehat{\mu}_k(i),\widehat{\sigma}_k^2(i))} \widetilde{\mathbb{1}}_{\{\widehat{V}_\infty^{\pi,k},\widetilde{\gamma}_k\}} (\pi(i) - \widehat{\mu}_{k+1}(i))^2}{\sum_{\pi\in\Lambda_k} \frac{\exp\{k\tau \widehat{V}_\infty^{\pi,k}\}}{\prod_{i\in S} \widetilde{f}(\pi(i);\widehat{\mu}_k(i),\widehat{\sigma}_k^2(i))} \widetilde{\mathbb{1}}_{\{\widehat{V}_\infty^{\pi,k},\widetilde{\gamma}_k\}}}, \tag{5.2}$$

where $\widetilde{f}(\pi(i); \widehat{\mu}_k(i), \widehat{\sigma}_k^2(i)) = (1-\lambda) f(\pi(i); \widehat{\mu}_k(i), \widehat{\sigma}_k^2(i)) + \lambda f(\pi(i), \mu_0(i), \sigma_0(i))$.

We use the same model structure for the Cross-Entropy algorithm, where the parameters are updated as

$$\widehat{\mu}_{k+1}(i) = \frac{\sum_{\pi\in\Lambda_k} \mathbb{1}_{\{\widehat{V}_\infty^{\pi,k} \geq \widehat{\gamma}_k\}} \pi(i)}{\sum_{\pi\in\Lambda_k} \mathbb{1}_{\{\widehat{V}_\infty^{\pi,k} \geq \widehat{\gamma}_k\}}},$$

$$\widehat{\sigma}_{k+1}(i) = \frac{\sum_{\pi\in\Lambda_k} \mathbb{1}_{\{\widehat{V}_\infty^{\pi,k} \geq \widehat{\gamma}_k\}} (\pi(i) - \mu_{k+1}(i))^2}{\sum_{\pi\in\Lambda_k} \mathbb{1}_{\{\widehat{V}_\infty^{\pi,k} \geq \widehat{\gamma}_k\}}}.$$

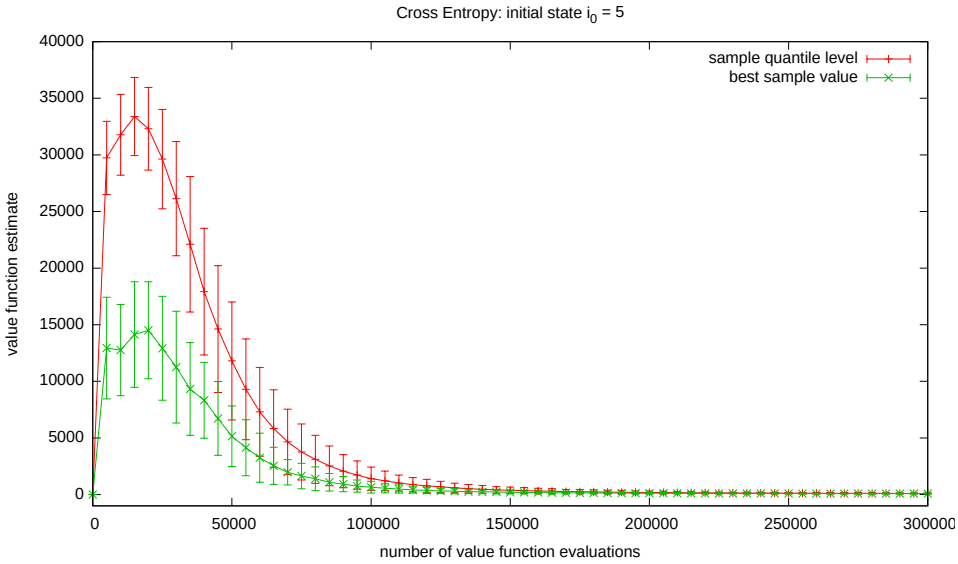
We consider in this example a queue with capacity $L = 49$, arrival probability $p = 0.2$, discount factor $\beta = 0.98$ and cost function

$$r(i,a) = i + 5\left(\frac{|S|}{2}\sin(2\pi a) - i\right)^2.$$

Hu, Fu and Marcus use a policy improvement method based on a discretization of the action space to obtain a valid estimate for the optimum and give graphically $V_\infty^* \approx 70$ for the initial state $i_0 = 5$ with optimal actions $\pi(0) \approx 0.5$ monotone decreasing in i to $\pi(25) \approx 0.3$ and $\pi(i) \approx 0.3$ for $i = 26, \ldots, 49$.
To approximate V_∞^π, each policy is simulated for 100 periods.

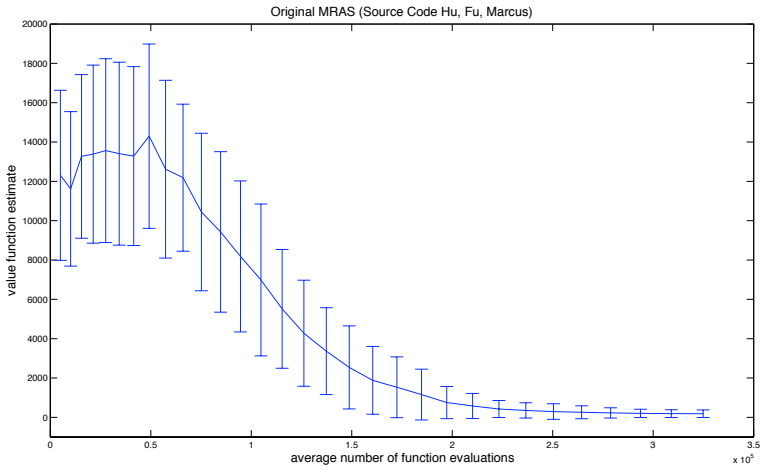
A parameters for the CE method, we choose $N = 100$, $\rho = 0.1$, observation size $M = 50$ and $v = 0.7$ for the smoothed update of μ, as suggested in [RK04]. For the variance update, we implement the additional scheme (2.10) with $q = 7$ and $\beta = 0.9$. The initial mean $\mu_0(i)$ is uniformly distributed in $[0,1]$ and the initial variance is set to be $\sigma_0^2(i) = 1$ for all i. The average results of 100 runs and the standard errors at each evaluation point are displayed in Figure 5.1.

Figure 5.1: Cross-Entropy for queueing



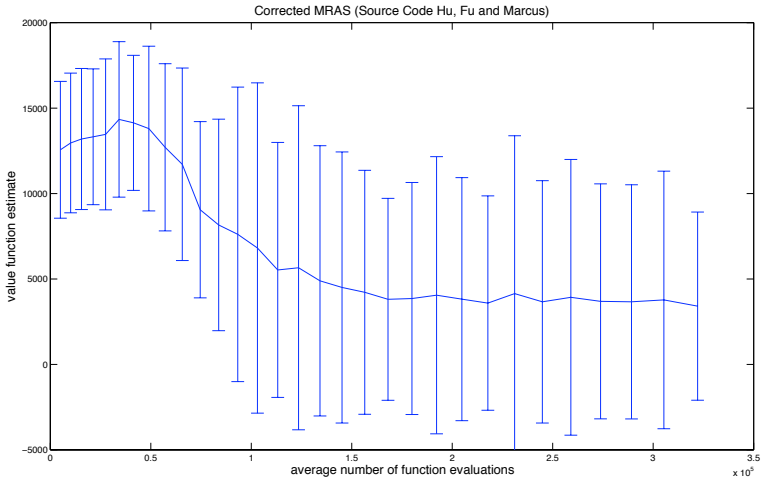
The parameters used by Hu, Fu and Marcus in [CFHM07] for the MRAS algorithm are $N_0 = 100$, $\rho_0 = 0.1$, $\lambda = 0.01$, $\alpha = 1.04$, $\tau = 0.01$, $\varepsilon = 0.1$, $v = 0.5$ and $M_k = \lceil 1.05 M_{k-1} \rceil$ with $M_0 = 50$. The initial mean and variance are as in the CE algorithm.
Using their original MATLAB-routine, we obtained as average over 100 runs the following result displayed in Figure 5.2, which is similar to those published in [CFHM07]. Note that only the behaviour of the best sample is shown.

Figure 5.2: MRAS implementation by Hu, Fu and Marcus, $i_0 = 5$



However, in their implementation they forgot to calculate the product $\prod_{i\in S} f(\pi(i);\mu_k(i),\sigma_k^2(i))$. As a result, they divide the exponential term $\exp\{k\tau\widehat{V}_\infty^{\pi,k}\}$ not by the density of the sample π, but by the density $f(\pi(i);\mu_k(i),\sigma_k^2(i))$ of some component of one of the samples in Λ_k, leading to a magnitude of mistake of several dimensions. Figure 5.3 depicts the average behaviour of the corrected implementation as obtained from 100 simulation runs.

Figure 5.3: Corrected MRAS implementation by Hu, Fu and Marcus, $i_0 = 5$



Trying to reproduce their numerical experiment in our framework, we faced some major numerical problems: Solution candidates π drawn in the first few iterations perform usually (as expected) quite poorly, i.e. the value $\widehat{V}_\infty^{\pi,k}$ is very large. In fact, $\exp\{k\tau\widehat{V}_\infty^{\pi,k}\}$ is quite frequently numerically zero for all samples considered in the update (i.e. for the approximately 10% best samples), leading to a division-by-zero error. This problem is further aggravated with each iteration by the increase of k.
A similar problem occurs in the calculation of the densities: If during the runtime of the algorithm the model distributions indeed approach degenerate distributions, the density of each drawn component $\pi(i)$ can be that large that computing the product $\prod_{i\in S} f(\pi(i);\mu_k(i),\sigma_k^2(i))$ produces a numerical overflow.
Accepting some added computational cost, these problems can technically be overcome by expanding the fractions (5.1), (5.2) in the update process by

$$\max_{\pi\in\Lambda_k} \exp\{-k\tau\widehat{V}_\infty^{\pi,k}\} \tag{5.3}$$

$$\text{and/or} \qquad \left(\max_{\pi(i)\in\Lambda_k} \widetilde{f}(\pi(i);\widehat{\mu}_k(i),\widehat{\sigma}_k^2(i))\right)^{|S|}. \tag{5.4}$$

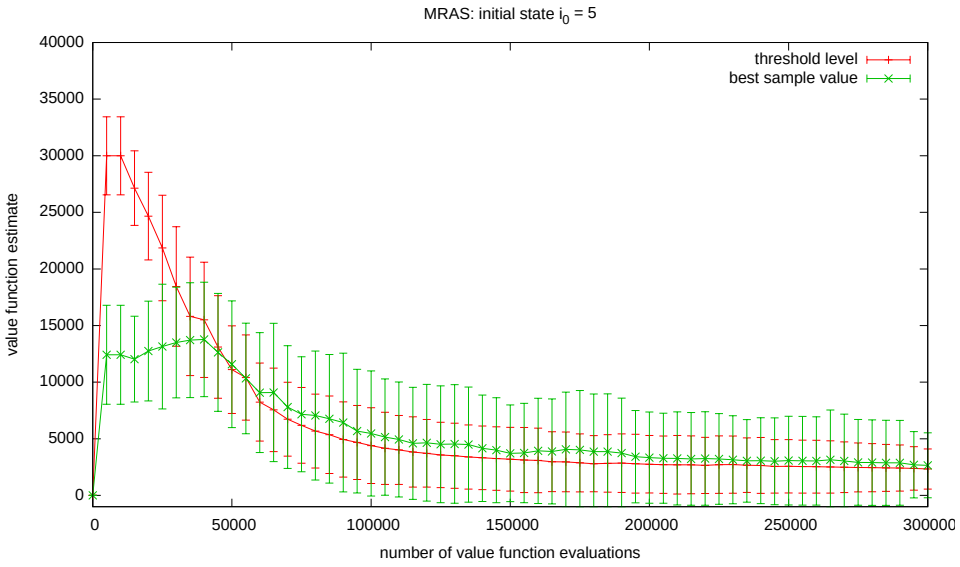
The latter circumvents the overflow by rescaling the densities to values in $[0,1]$. Unfortunately, though, especially in early iterations this may have quite the opposite effect, namely that (when $\max_{\pi(i)\in\Lambda_k} \widetilde{f}(\pi(i);\widehat{\mu}_k(i),\widehat{\sigma}_k^2(i))$ is large compared to the densities of other components) the rescaled densities are numerically zero. Finding

a useable scaling factor is then not an easy task.
Let $\pi_{max} = \text{argmax}_{\pi \in \Lambda_k} \exp\{-k\tau \widehat{V}_\infty^{\pi,k}\}$. Then the former expansion yields the larger term $\exp\{k\tau(\widehat{V}_\infty^{\pi,k} - \widehat{V}_\infty^{\pi_{max},k})\}$ in both enumerator and denominator and this ensures that at least for the one sample $\pi = \pi_{max}$ the exponential function will be positive. Note however that this modification usually produces the need for a readjustment of parameter τ. Unlike in the unmodified algorithm, where a rough knowledge of the value function dimensions could serve as a good guideline for the choice of parameters, the optimal adjustment of τ now depends on the spread in the set of best samples in each iteration, hence directly on the behaviour of the algorithm - and this can only be determined by observation!
In their source code, Hu, Fu and Marcus used even another method: they added a small positive constant to the exponential function, thus calculating

$$\left(\exp\{\tau \widehat{V}_\infty^{\pi,k}\} + 10^{-5}\right)^k \tag{5.5}$$

in each iteration. This usually works but has the disadvantage that in initial iterations the constant term will dominate any exponential weighting factor whereas in later iterations the problem is not completely excluded, since the constant factor 10^{-5k} vanishes for large k.

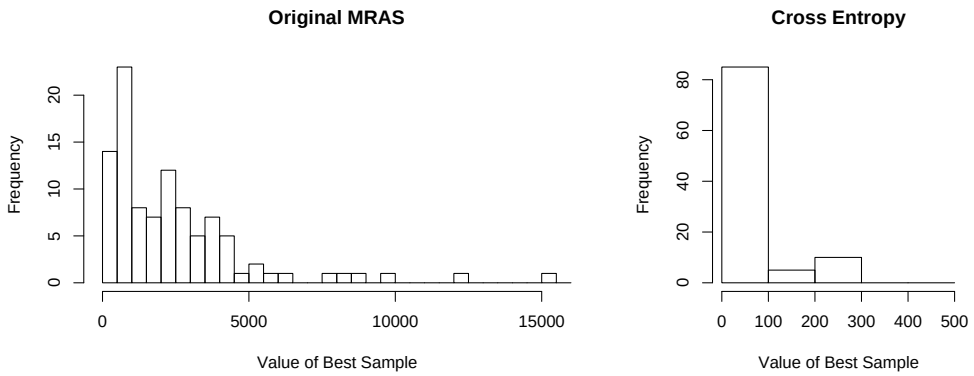
Figure 5.4: Queueing Example - Original MRAS



The behaviour of the original MRAS algorithm using the exponential function expansion and the parameters stated above is shown in Figure 5.4. The results were obtained as the average over 100 runs. We see that the mean value of the found solutions is far worse than that obtained by the Cross-Entropy algorithm whereas the variance is significantly larger. For a detailed comparison of the solutions found by both algorithms, see Figure 5.5. One can see that all values found by the CE method were less than 300, most of them even smaller than 100, whereas only 14

results of the MRAS reached a level below 500 (note that we used different grids for the two histograms).

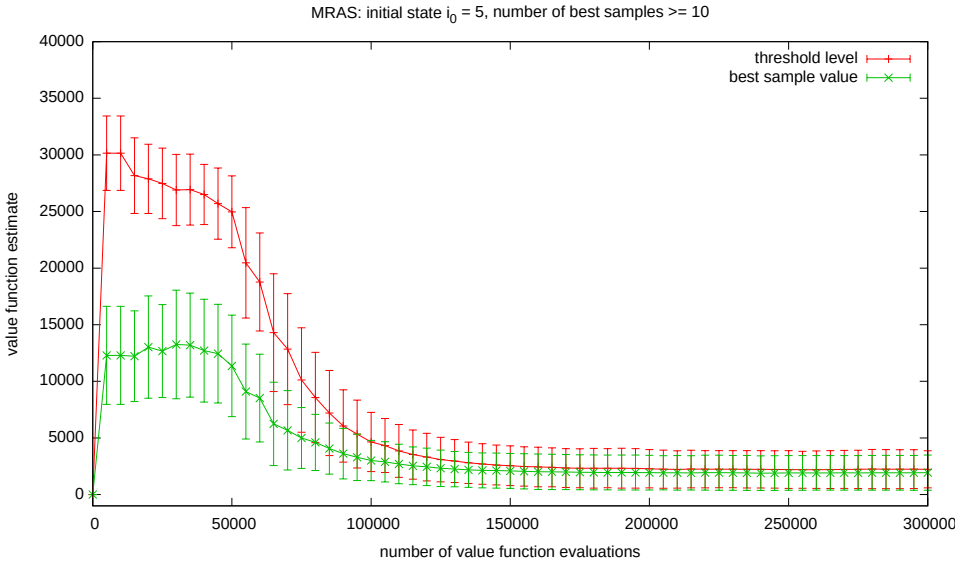
Figure 5.5: Queueing Example - Histogram of Solution Values



One problem in many of the runs seems to be a sudden and preliminary decrease in the variance of the sampling distribution, so that the algorithm freezes in some suboptimal solution. This usually happens if the adaptation of parameter ρ in step 3b is too pronounced (e.g. when only one sample exceeds the former threshold) and hence in later iterations the sample quantile is too ambitious (taking into account say only the best 1-2 % of the order statistic). On this issue, Hu, Fu and Marcus mention in [HFM07] that it may be advisable to perform the parameter update in Step 4 only if at least N_{min} samples would be considered. However, what they do is to perform the quantile update in Step 3b only if at least N_{min} samples exceed the new threshold (hence introducing a bound $\rho_{min} = \frac{N_{min}}{N} \to 0$ as $N \to \infty$) but to always update the sampling distribution parameter (unless no sample at all reaches the threshold). The results of this modification with $N_{min} = 10$ are presented in Figure 5.6, again averaging over 100 runs performed by our implementation.

A test run with $N_{min} = 5$ even yielded three solutions with values between 40 000 and 75 000. A close look at the algorithm's behaviour in these cases revealed yet another problem concerning the weighting factors $\frac{\exp\{k\tau \widehat{V}_\infty^{\pi,k}\}}{\prod_{i\in S} \widetilde{f}(\pi(i);\widehat{\mu}_k(i),\widehat{\sigma}_k^2(i))}$. Usually the range of densities varies from iteration to iteration but is quite constant between the top samples in each iteration. However, it may happen that the sampling distribution has tightened around some suboptimal solution and then a good (or better) solution π is drawn according to the initial distribution $f(\cdot\,,\theta_0)$. Then the denominator $\prod_{i\in S} \widetilde{f}(\pi(i);\widehat{\mu}_k(i),\widehat{\sigma}_k^2(i))$ is very small compared to those of the other samples (in our example e.g. of dimension 10^{-33} compared to dimension 10^{-5} or 10^5). Hence, this specific sample π will most probably have such a heavy weight compared to the others that it completely dominates the parameter update. Thus the new parameter $\widehat{\mu}_{k+1}$ will correspond almost exactly to π whereas the new variance $\widehat{\sigma}_{k+1}$ will be practically zero. Due to the fact that we use a smoothed update scheme (cf. (2.9)), the algorithm will end up with a mean somewhere between the

Figure 5.6: Queueing example - MRAS with ρ_{min}



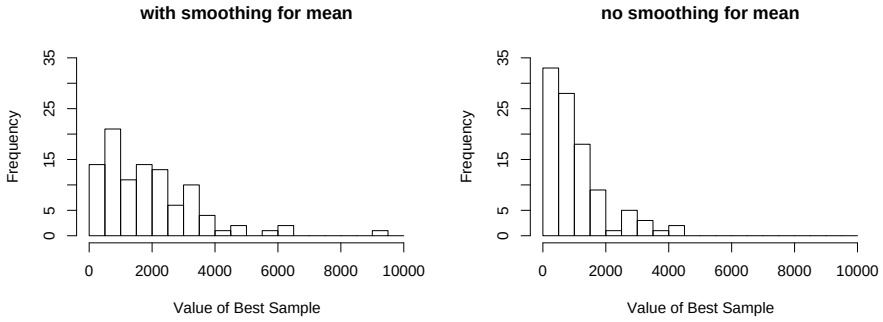
old mean and π with a very small variance. It may then be that this new mean is not only suboptimal but far worse than both the old mean $\widehat{\mu}_k$ and π and that in subsequent iterations no further sample will reach the old threshold level $\widehat{\gamma}_k$.
Note that the behaviour described above holds true if we use modification (5.3) and the sample π is actually the best sample (or nearly the best) in the iteration. Otherwise the difference in the density may be balanced out by an equally large difference in the exponential term. However, if we use (5.5), the exponential evaluation terms do not differ very much among the top samples in one iteration. Then the density is the one dominating factor in the calculation of the weights and the algorithm may just meander from one unlikely sample to another with no consideration for their performance.

The observations noted above suggest that the smoothed update may cause some of the problems in the MRAS algorithm. Yet trials runs reveal that abandoning all smoothing will cause the variance to converge to zero in about 4 - 5 iterations, leading to unmanageable numerical overflows of the densities. Hence we retain the smoothing with parameter $v = 0.5$ for the variance but set $v = 1$ (no smoothing) for the update of the mean vector (using $N_{min} = 10$, (5.3) and the other parameters as above). This produces a significant decrease of the average solution value and variance compared to the test run with only N_{min} (figure 5.6). A comparison is shown in Figure 5.7.

Additional experiments with different parameters τ (also using an adapted scheme), v and α yielded no significant further improvement.

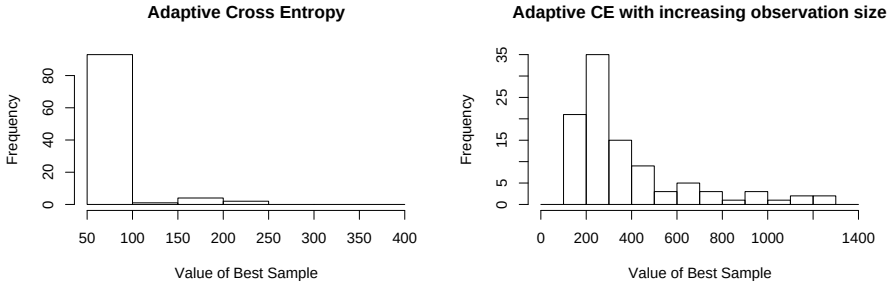
To further investigate the reasons for the different behaviour of Cross-Entropy and MRAS, we implemented the Adaptive Cross-Entropy method presented in Al-

Figure 5.7: Queueing example - Influence of smoothed update in MRAS



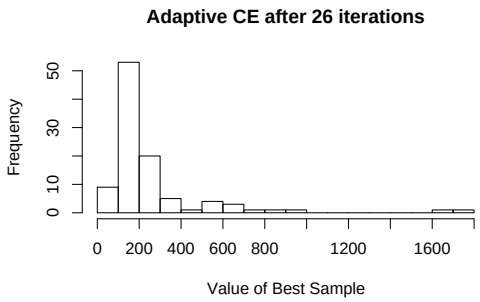
gorithm 2.3. The results of 100 simulation runs with parameters $N_0 = 100$, $\rho_0 = 0.1$, $\varepsilon = 0.1$, $\alpha = 1.04$ and $v = 0.5$ for both mean and variance as in the MRAS, stopped after 300 000 function evaluations as above, are displayed in Figure 5.8. Moreover, we ran the Adaptive Cross-Entropy separately with fixed observation size $M = 50$ as in the original algorithm and with the increasing scheme used in the MRAS (parameters $M_0 = 50$ and $M_k = \lceil 1.05 M_{k-1} \rceil$, as above). Then the only differences between MRAS and Adaptive CE lie in the use of the weighting factors and the order of parameter adaptation and distribution update. Performing the same experiments with Step 5 before Step 3 in Algorithm 2.3 (i.e. in the same order as in MRAS) yields no substantial discrepancies in the results. Hence we have excluded any differences apart from the factors $\frac{\exp\{k\tau \widehat{V}_\infty^{\pi,k}\}}{\prod_{i \in S} \widetilde{f}(\pi(i); \widehat{\mu}_k(i), \widehat{\sigma}_k^2(i))}$.

Figure 5.8: Queueing Example - Solutions of Adaptive Cross-Entropy



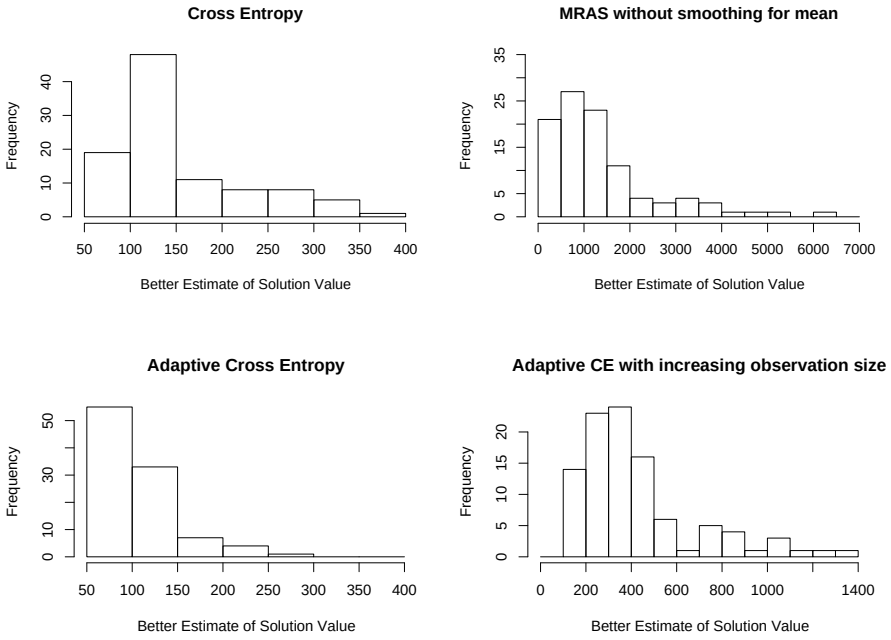
One can see in Figure 5.8 that the Adaptive CE with fixed observation size performs much better in the given scope of function evaluations. This seems to be simply because the algorithms stops after approximately 50 iterations in contrast to approximately 26 for the Adaptive CE with increasing observation size: Figure 5.9 depicts the solutions of the basic Adaptive CE (taken from the same simulation run as above) after 26 iterations which are similar to those obtained by the algorithm with increasing M_k. Yet even those solutions, though not as good as the ones found by the Cross-Entropy methods with fixed M, are far better than those of the Model Reference Adaptive Search. As we have excluded every other factor, this difference

Figure 5.9: Queueing Example - Adaptive Cross-Entropy after 26 Iterations



can only be explained by the weighting terms.
Note that the observations above are quite interesting: apparently the number of iterations has more influence on the convergence of the algorithms than the number of function evaluations. This might be explained by the fact that methods with increasing sample size use a large part of the allowed function evaluations for exact estimates and not for finding more (better) samples. One would however expect the methods with few observations allocated to each sample to converge to non-optimal solutions as they should be less able to distinguish good samples from bad ones. Nevertheless, re-evaluating all solutions found in our simulation runs by the different

Figure 5.10: Estimate based on 10 000 Observations



algorithms with 10 000 observations each to obtain more exact value estimates does not entail a different result. Even though the solutions are generally rated worse, the

overall picture remains the same (cf. Figure 5.10). Hence facing a computational budget, it is apparently advisable to invest more in sampling different solutions than in their exact evaluation. The general underestimation of the value function will be discussed in more detail in the next example.

5.3 Inventory

We consider the inventory control problem from [CFHM07] with exponentially distributed demand, no lead times, full backorders and linear holding and penalty costs. At the beginning of each period, the inventory i is reviewed and an order a is placed. Then the costs for this period are calculated, consisting of (fix and variable) ordering costs and linear holding respectively penalty costs. During the period, a stochastic demand d occurs, so that the inventory level at the beginning of the next period is $i + a - d$. This is a Markov Decision Process with uncountable state and action space, described here as a stochastic control model:

- $S = \mathbb{R}$, $i_t \in S$ the inventory position at the beginning of period t.
- $A = \mathbb{R}_+$, $a_t \in A$ the ordered amount in period t. $D(i) = A$.
- $Z = \mathbb{R}_+$ the space of disturbances, $d_t \in Z$ the demand in period t.
- Transition law $Q : D \to Z$, given by the density $q(d|i, a) = q(d) = \lambda e^{-\lambda d}$, i.e. the demand is exponentially distributed with parameter λ.
- Transition function $T : D \times Z \to S$, $T(i, a, z) = i + a - z$.
- $r(i, a, j) = \begin{cases} K + ca + hj^+ + pj^- & a > 0 \\ hj^+ + pj^- & a = 0 \end{cases}$, where K is the cost of placing an order, c the additional ordering cost per unit and h and p the holding respectively penalty costs per unit.
- $\beta = 1$, $V_0 = 0$.

The aim is to minimize the long run stationary costs $G_\pi = \int_S V_{\infty\pi}(i)\mu(i)di$ of the system, where μ is the stationary distribution of the inventory level.

It is well-known that the optimal policy for such an inventory problem is of (s, S)-form: If the inventory level i is below some threshold s the inventory is restocked to level S, i.e.

$$a_t = \begin{cases} 0 & i_t \geq s, \\ S - i_t & i_t < s. \end{cases}$$

Moreover, for exponentially distributed demand D with mean $\mathbb{E}[D] = \frac{1}{\lambda}$ and linear penalty and holding costs, the optimal levels s^* and S^* can be determined analyti-

cally (see [AKS58]) as

$$s^* = -\frac{1}{\lambda} \ln \left(\frac{h + \sqrt{2Kh\lambda}}{h+p} \right),$$
$$S^* = s^* + \sqrt{\frac{2K}{\lambda h}}.$$

Furthermore, for a given policy (s, S) we have

$$G_{(s,S)} = \frac{c}{\lambda} + \frac{K + h\left(s - \frac{1}{\lambda} + \frac{\lambda}{2}(S^2 - s^2)\right) + (h+p)\frac{1}{\lambda}e^{-\lambda s}}{1 + \lambda(S-s)}. \tag{5.6}$$

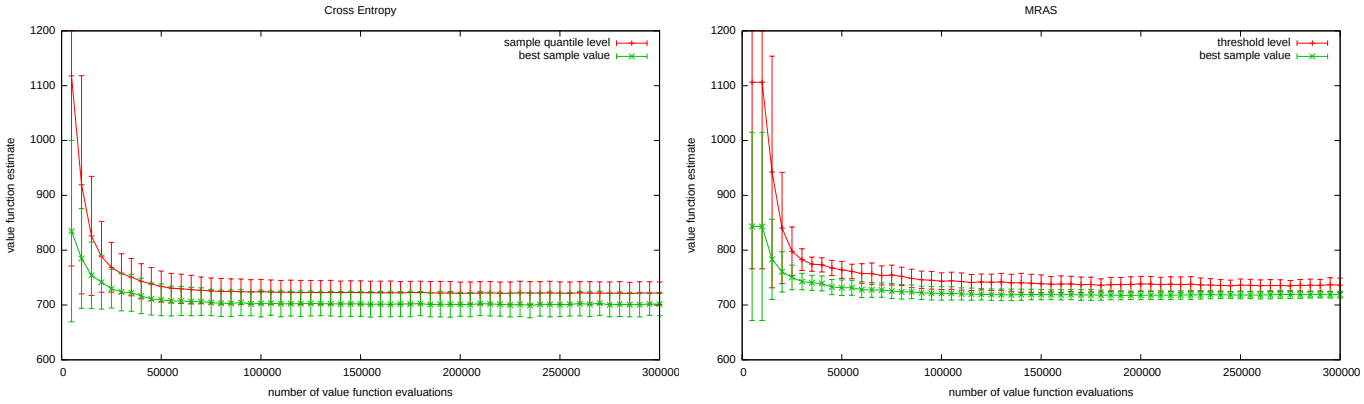
This policy structure can be used to reduce the dimension of our model distribution from $|S|$ to 2. Again, we use normal distributions in each component for both Cross-Entropy and Model Reference Adaptive Search. Since we do not truncate our distribution here, the updates are as in Examples 2.2 and 3.1. As initial parameters we use $\mu_0(s) \sim U[0, 2000]$, $\mu_0(S) \sim U[0, 4000]$ and $\sigma_0^2(i) = 10^6$ for $i = s, S$ in both algorithms. Note that in the simulation we do not restrict our policies to $s \leq S$ for ease of random number generation, but if in the final solution $s > S$, we set $s = S$, i.e. we evaluate the policy $\pi = (S, S)$.

Example 1

The first specific problem instance is given by $h = c = 1$, $p = 10$, $K = 100$ and $\mathbb{E}[D] = 200$. The optimal strategy in this case is $(s^*, S^*) = (340.95, 540.95)$ and the average costs for this policy are $G^* = 740.95$. To estimate the performance of a given policy, it is first simulated over 50 periods to reach an arbitrary system state ("warm-up phase"). Then the average costs over the next 50 periods are calculated. Averaging over 150 periods to estimate costs yielded no discernible differences.
As in the previous section, the parameters for the CE method are $N = 100$, $\rho = 0.1$, $M = 50$ and $v = 0.7$. Test runs revealed that the modified smoothing scheme (2.10) destabilizes the algorithm in this example, hence we used the simple update (2.9) with $v = 0.7$ for both mean and variance.
The parameters used in MRAS are $N_0 = 100$, $\rho_0 = 0.1$, $\lambda = 0.01$, $\alpha = 1.04$, $\tau = 0.01$, $\varepsilon = 0.01$, $M_k = \lceil 1.05 M_{k-1} \rceil$, $M_0 = 50$ and $v = 0.5$ for both mean and variance. Furthermore we used the bound for ρ_k with $N_{min} = 10$ here and in all subsequent examples.
The average results of 100 simulations runs (stopped after 300000 function evaluations) of both CE and MRAS are displayed in Figure 5.11.

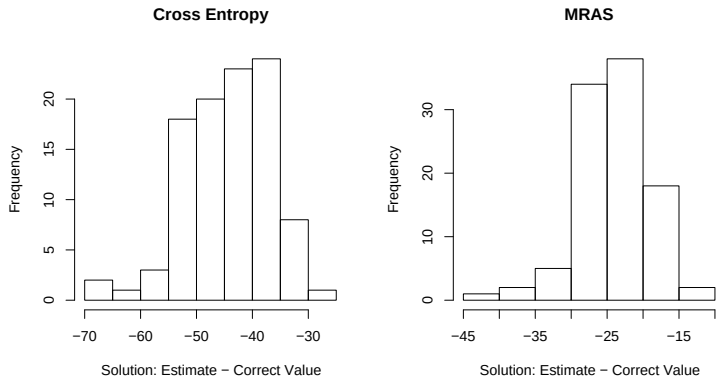
As one can see, the MRAS performs better than in the previous example. This is to be expected considering the smaller range of function values (avoiding numerical problems regarding the exponential function) and the equally smaller range of values of $\prod_{i=1}^{2} \widetilde{f}(\pi(i); \widehat{\mu}_k(i), \widehat{\sigma}_k^2(i))$ (for different samples in one iteration as well as in the course of the algorithm) due to less dimensions.

Figure 5.11: (s, S)-Inventory Problem (Example 1)



Observe that the mean estimated solution values of the Cross Entropy algorithm are far below the theoretical optimum. This deviance is also confirmed in the comparison between these estimations and the correct analytic evaluation (5.6) of the found solutions, shown in Figure 5.12. Since the differences in the estimates of the

Figure 5.12: Quality of Solution Value Estimate in (s,S)- Inventory Problem (Example 1)



MRAS are much smaller, we additionally tested the CE algorithm with the same parameters as above but an increasing sample size (as for the MRAS $M_0 = 50$, $M_k = \lceil 1.05 M_{k-1} \rceil$). As expected, this decreases the estimation error considerably (see Figure 5.13, note the different grids). However, even with increasing observation size this underevaluation is systematic in both Cross-Entropy and Model Reference Adaptive Search.

The explanation for this is straightforward: the candidate solutions are sorted according to their estimated performance. Hence those evaluations that underestimate the value (i.e. that give a score better than the real performance) are favoured in this process and, as there is always some variance in the estimator and so there are usually evaluations that suggest a better performance than it is the case, the given solution values are almost always too good. The increasing observation size

decreases the magnitude of error but cannot impede the general tendency.

Due to the observations made in the queueing example, we also tested the MRAS without the smoothed update for the vector of means. The results of all four simulations are presented in Figure 5.14 (standard errors are in parentheses). Note that unlike in Section 5.2 the reduced smoothing yielded worse results than the original update.

The standard error in the solutions of both Cross Entropy methods is far more significant than that of the MRAS algorithms. This is due to 2-3 comparatively bad solutions (values in $[800, 930]$) in both runs. However, comparing the "good" solutions - in this case all solutions with values smaller than 750 - one can see that the values of the policies found by both MRAS algorithms are more scattered than those found by the Cross-Entropy methods (Figures 5.15, 5.16). Hence the CE apparently yields better policies, but on the other hand produces more outliers.

Example 2

Similar results are obtained in the second problem instance of the (s,S)- Inventory problem with $h = 15$, $c = 20$, $p = 50$, $K = 1000$ and $\mathbb{E}[D] = 400$. The optimal strategy in this case is $(s^*, S^*) = (404.24, 635.18)$ with $G^* = 17527.65$. Again, we compare the four algorithm instances of the first example and change only the parameter τ to $\tau = 0.001$ to account for the greater function values. The results are displayed in Figures 5.17 and 5.18. Note that again the Cross-Entropy methods produce more solutions very close to the optimum than the MRAS and that in this

Figure 5.13: Quality of Solution Value Estimate in (s,S)- Inventory Problem (Example 1)

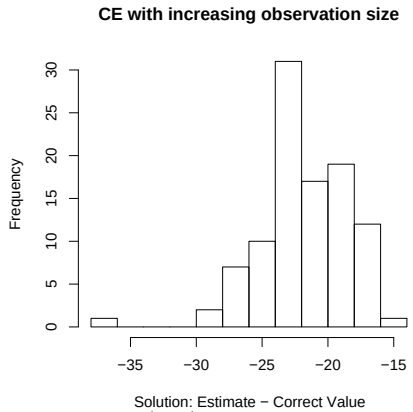


Figure 5.14: Results (s,S)-Inventory Problem Example 1

	Mean (error)	Minimum	No. < 750
MRAS	743.38 (4.38)	740.961	97
CE	746.03 (19.72)	740.953	93
MRAS (mod. smoothing)	745.2 (8.44)	740.96	94
CE (mod. observation size)	747.1 (22.73)	740.964	92

Figure 5.15: Inventory Example 1 - Comparison of good solutions

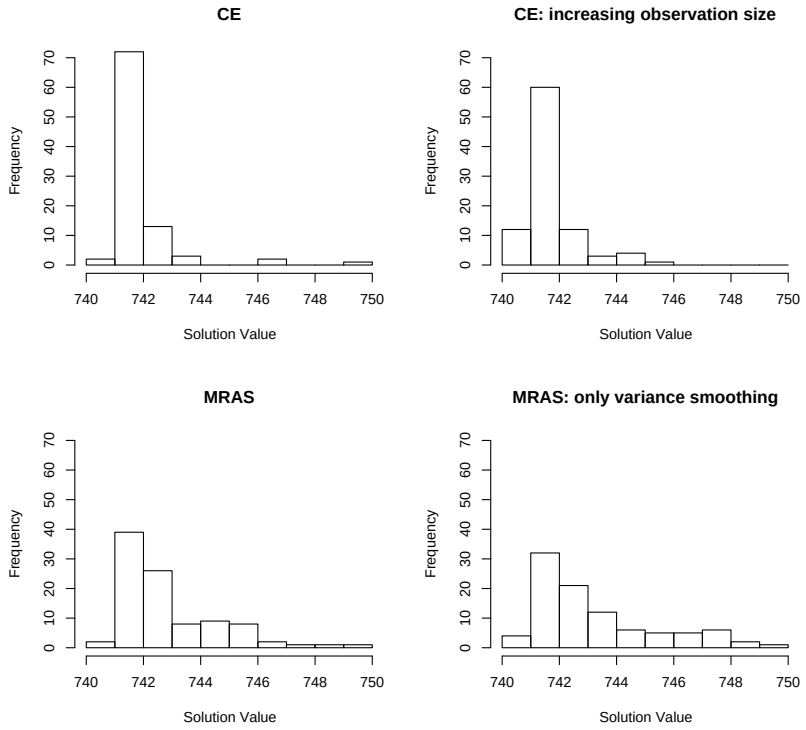
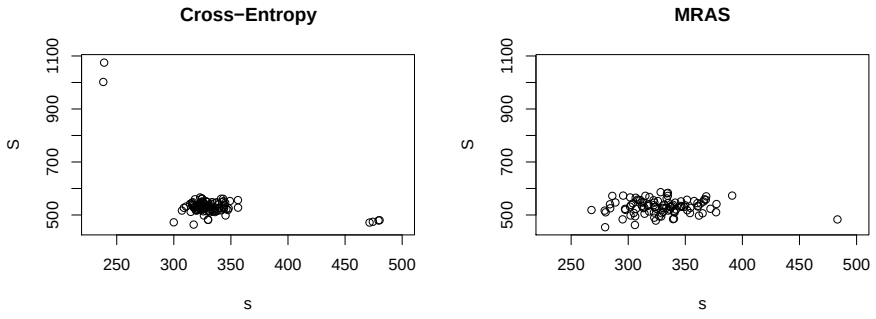


Figure 5.16: Inventory Example 1 - Solutions (Optimum $(s^*, S^*) = (340.95, 540.95)$)



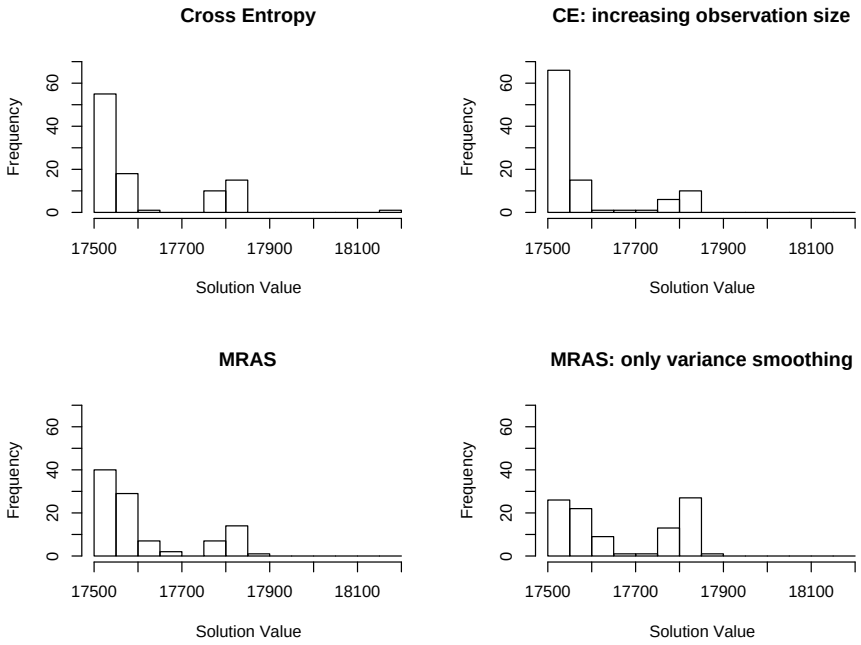
case we do not even observe a greater variance in the CE solution values.

Figure 5.17: Results (s,S)-Inventory Problem Example 2

	Mean (error)	Minimum
MRAS	17615.81 (107.15)	17528.00
CE	17615.62 (126.21)	17527.82
MRAS (mod. smoothing)	17666.30 (123.83)	17530.27
CE (mod. observation size)	17589.00 (99.33)	17527.97

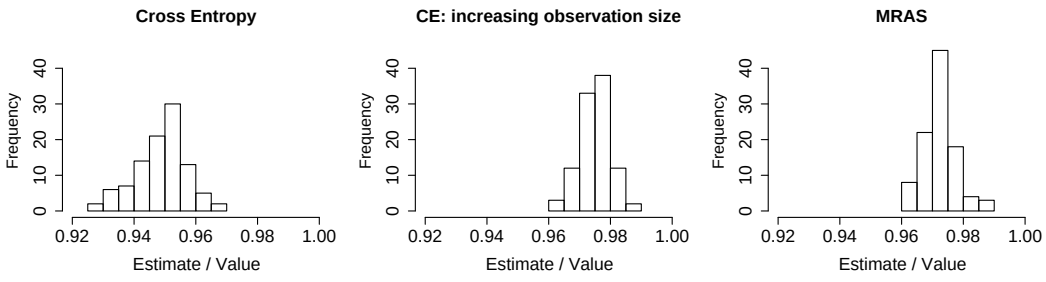
Again we are interested in the quality of the value function estimates. In Figure 5.19 we plot the ratio of each solution's value function estimate and its correct value for the different algorithms and obtain similar results as in the first example: The

Figure 5.18: Inventory Example 2 - Comparison of solutions



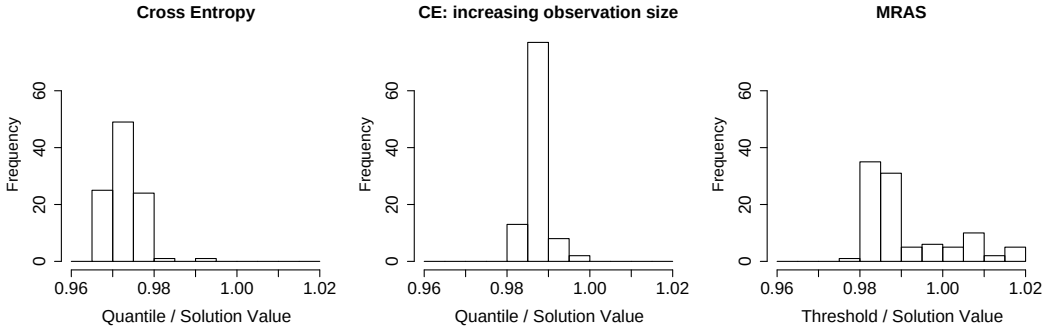
Cross-Entropy method without increasing sample size significantly underestimates the performance of the found policies (the estimate being between 92% and 97% of the correct value), whereas both MRAS and the CE with increasing observation size yield better estimates (96% - 99%) but still systematically underrate the correct values.

Figure 5.19: Inventory Example 2 - Value Function Estimates



A different picture is obtained when considering the threshold respectively sample quantile of the last iteration as estimate for the solution value. This approach is motivated by the underestimation observed before and the fact that the quantile value is of course consistently greater than that of the best sample. We see from Figure 5.20 that this yields indeed better estimates. Unlike in the CE algorithms where the quantile is still always smaller than the correct solution value (98% - 100% for increasing observation size), the ratio of threshold and correct value in the MRAS algorithm is sometimes even greater than 1. This hints at a different behaviour of

Figure 5.20: Inventory Example 2 - Threshold as Value Estimate



quantiles (in CE) and thresholds (in MRAS).
To further study this behaviour, we ran again 100 simulations for both the Cross Entropy method with increasing observation size $M_0 = 50$, $M_k = \lceil 1.1 M_{k-1} \rceil$ and the MRAS with allowed total number of function evaluations $c = 5\,000\,000$. The average results are displayed in Figures 5.21 and 5.22. Note that we could only include 99 runs in the evaluation of the MRAS since one simulation run ended with `NaN` (Not a Number) as function value and solution, evidence of a numerical over- or underflow in some denominator.

Figure 5.21: Inventory Example 2 - Long run behaviour Cross Entropy

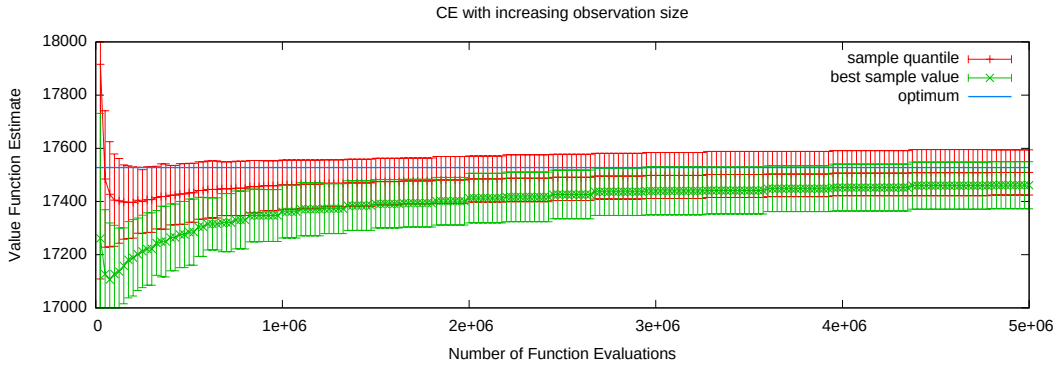


Figure 5.22: Inventory Example 2 - Long run behaviour MRAS

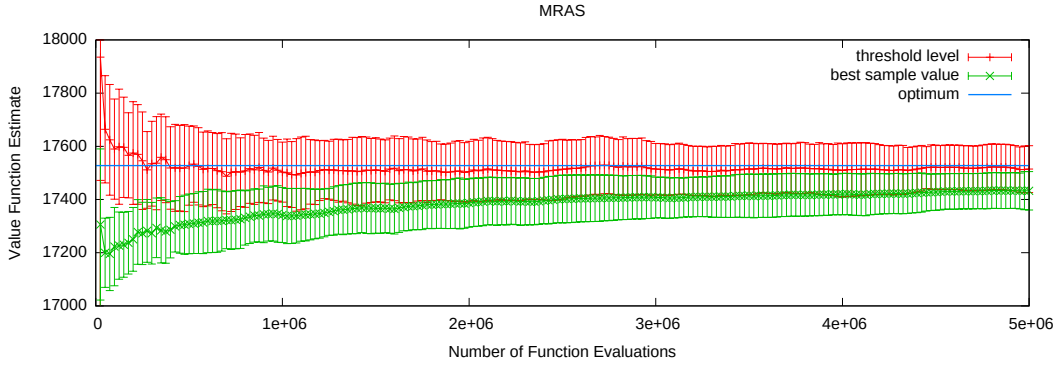
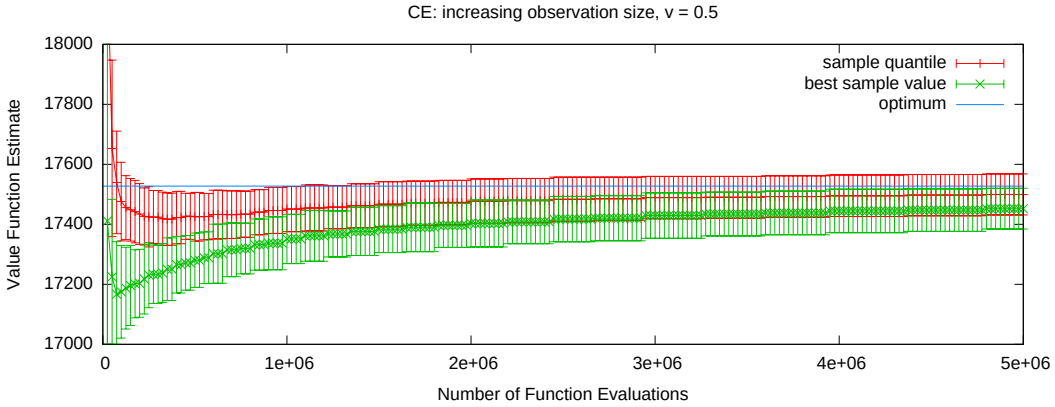


Figure 5.23: Inventory Example 2 - Long run behaviour CE with more smoothing



One can see that indeed the sample quantile level of the CE decreases faster than the sequence of thresholds in the MRAS. With increasing exactness of the estimates, the quantile levels (and best sample values) are corrected and approach the optimum from below. That these observations are not due to the different smoothing factor v shows a simulation run of the CE with $v = 0.5$ as in MRAS (Figure 5.23). Nevertheless, as expected, the smoothing factor does reduce the misjudgment in the first couple of iterations (less than $0.5 \cdot 10^6$ function evaluations), but does not substantially influence the long time behaviour.

5.4 Replacement

Consider a randomly degrading system, e.g. a machine with several components that may fail. In each period, one can either continue to run the system at a cost depending on the grade of its wearout or repair it to reduce abrasion to zero. The aim is to find a policy that minimizes the expected discounted costs for an infinite horizon. This can be modeled as a MDP with

- S countable state space, $0 \in S$, $i \in S$ the grade of wearout, $i = 0$ corresponding to a new system.
- $A = \{0, 1\}$, $a = \begin{cases} 1 & \hat{=} \text{ replace} \\ 0 & \hat{=} \text{ continue} \end{cases}$, $D(i) = A$.
- $p(i, a, j) = \begin{cases} p_{0j} & , a = 1 \\ p_{ij} & , a = 0 \end{cases}$, where (p_{ij}) is a stochastic transition matrix.
- $r(i, a) = \begin{cases} r_1(i) & , a = 1 \\ r_0(i) & , a = 0 \end{cases}$. We will consider reward functions of the form $r_1(i) = -K - c(0)$, $r_0(i) = -c(i)$ for some fix cost K and some state-dependent cost c.
- $\beta > 0$ the discount factor, $V_0 \equiv 0$.

One can show that under certain assumptions the optimal stationary policy is of the form $(0, \ldots, 0, 1, \ldots, 1)$. This is for example the case when $p_{ij} \sim b(|S| - 1, p_i)$

with $0 < p_0 \leq \cdots \leq p_{|S|-1} < 1$ and the costs $c(i)$ are isotone in i. (see exercises [Rie04]).

An obvious choice for the sampling distribution in this case are independent Bernoulli distributions $\mathrm{Bin}(1,\theta)$ in each component. As these distributions belong to a NEF, the updates are as in Examples 2.1 and 3.1. We start with $\theta_0 = \frac{1}{2}$ in all components.

We consider 20 components (i.e. $|S| = 21$), $K = 13$, $c(i) = i$, $\beta = 0.9$ and $p_{ij} \sim b(20, p_i)$ with $p_0 = 0.15$, $p_1 = \cdots = p_5 = 0.2$, $p_6 = \cdots = p_{11} = 0.3$, $p_{12} = \cdots = p_{15} = 0.5$ and $p_{16} = \cdots = p_{20} = 0.8$. Then the optimal stationary policy and its value are given by

$$\pi^*(i) = \begin{cases} 0 & i \leq 9 \\ 1 & i \geq 10 \end{cases},$$

$$V_{\pi^*} = -(39.345, 41.589, 42.589, 43.589, 44.589, 45.589, 49.2821, \\ 50.2821, 51.2821, 52.2821, 52.3498, \ldots, 52.3498).$$

We simulate every policy for 100 periods, starting in state $i_0 = 0$ to obtain an estimate for the total discounted reward $R_\pi^\infty(0)$.

For the Cross-Entropy algorithm we use again $N = 100$, $\rho = 0.1$ and $v = 0.7$. The parameters in MRAS are $N_0 = 100$, $\rho_0 = 0.1$, $v = 0.5$, $\varepsilon = 0.01$, $\alpha = 1.04$, $\tau = 0.1$, $N_{min} = 10$ and $\lambda = 0.01$. The observation size for both algorithms is $M_0 = 100$, $M_k = \lceil 1.1 M_{k-1} \rceil$.

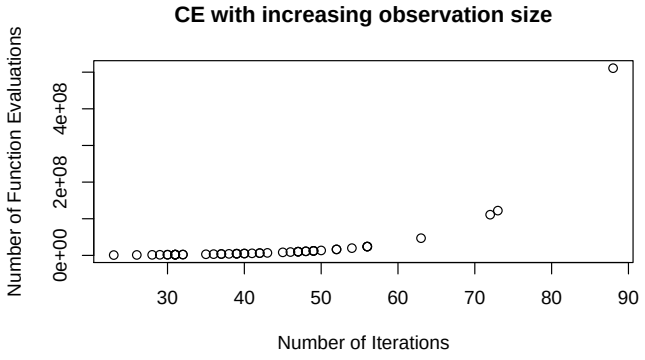
As the behaviour of both algorithms varied in test runs from simulation run to simulation run, it was quite difficult to decide on an upper bound of allowed function evaluations to assess correctly the quality of the algorithms. Hence, we decided to introduce a stopping criterion based on the convergence to a degenerate distribution in the following sense: We declared the distribution $\mathrm{Bin}(1,\theta)$ in each component as "converged" if $\theta < 0.1$ or $\theta > 0.9$ (test runs indicated that once having reached these values, the parameter θ only converged closer to 0 respectively 1). The algorithm was stopped if in iteration k

$$C := \sum_{i \in S} \mathbb{1}_{\{\theta_k(i) \in (0.1, 0.9)\}} \min\{\theta_k(i) - 0.1, 0.9 - \theta_k(i)\} \leq 0.05,$$

i.e. if basically all parameters $\theta(i)$ had reached one of the thresholds.
This worked quite well for the Cross-Entropy algorithm where most of the 50 runs converged in about 30 to 50 iterations (see Figure 5.24). Two out of those 50 runs had to be aborted as even after 84 iterations they were far from convergence ($C \approx 0.385$).

Unfortunately, even though by Example 4.7 $\mathbb{E}[X] = \theta \to x^* \in \{0,1\}$ componentwise, we could not observe such a convergence during a reasonable time intervall ($\approx$ 12 hours of real runtime). To obtain any utilizable results, we decided to stop the algorithm after 65 iterations. Over 49 simulation runs, this corresponded to a

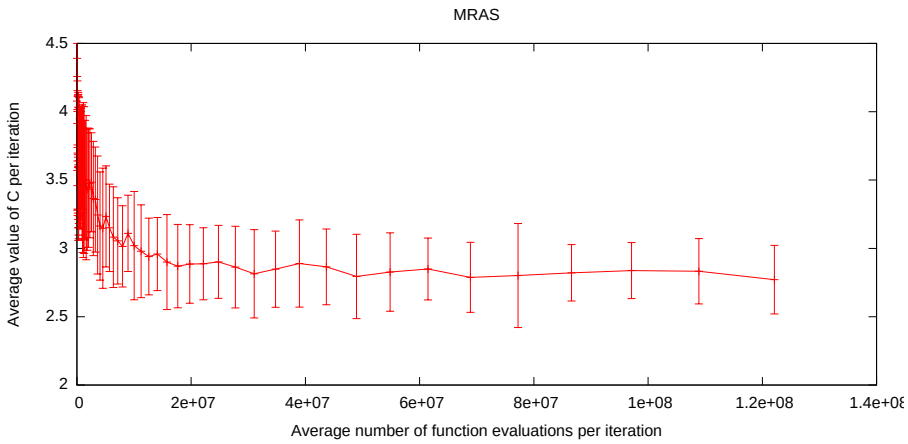
Figure 5.24: Replacement Example: Convergence behaviour Cross Entropy



	Iterations	Sample size
mean	43.33	23 525 400
error	12.97	75 567 223
median	41.50	5 957 600
IQR	17.00	9 956 300

mean number of 136 955 333 function evaluations (standard error 6 521 046), where each run took 2-3 hours. In comparison: most of the simulations runs of the Cross-Entropy algorithm that converged in 30-40 iterations took 20 minutes. In Figure

Figure 5.25: Replacement Example: Convergence behaviour MRAS



5.25 we plotted for each iteration the average value of C against the average number of function evaluations and one can observe that indeed the convergence of $C \to 0$ is very slow.

In addition, the typical evolution of a simulation run of both Cross-Entropy and MRAS (represented by the best sample π_k in each iteration k, i.e. the current solution) is displayed in figures 5.27 and 5.26. Both runs obtain very good solutions: $V_{\pi_{35}^{\mathrm{CE}}}(0) = -39.3498$ and $V_{\pi_{65}^{\mathrm{MRAS}}}(0) = -39.3499$. One can see that the CE converges in about 10 iterations in the first 13 components and has reached after 32 iterations a fix solution. Quite the contrary is observable for the MRAS: after more than 60 iterations, not even the boundary state (optimal is $i = 10$) is fix and the last 7

components do not show any convergence at all. This makes the solution extremely volatile. Had we stopped this run of MRAS after 64 iterations for example, the performance would have been $V_{\pi_{64}^{\mathrm{MRAS}}}(0) = -39.3645$.

To compare the performance of CE and MRAS in this example, we calculated the correct value $V_\pi^*(0)$ of each solution π^* through a linear program (cf. [Rie04]) where in the case of CE π^* was taken as the solution after convergence (and hence not after a fixed number of iterations or function evaluations). For the MRAS, we evaluated separately the solutions obtained after 44 iterations (corresponding to the mean number of iterations of the Cross-Entropy algorithm, see Figure 5.24) and after 65 iterations. Note that the mean number of function evaluations of the MRAS after

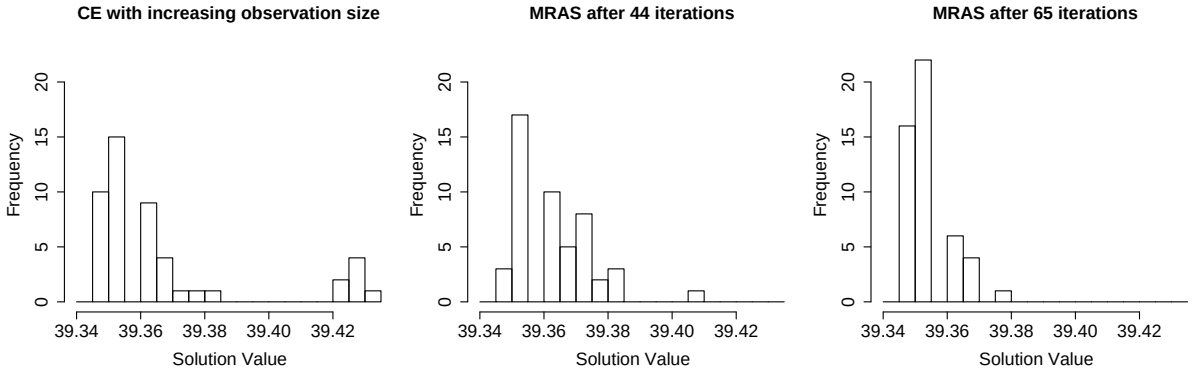
```
 0   0 0 0 0 0 0 1 1 0 0 1 1 0 1 1 1 0 1 0 0 0
 1   0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1 1 0 1 0
 2   0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 1 1 0 1 0 0
 3   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 1 0 0 1
 4   0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 0 0 1 1 0 0
 5   0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 1 1 1 0 1
 6   0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 0 0 1
 7   0 0 0 0 0 0 0 0 0 1 1 0 0 1 1 1 0 0 0 1 1
 8   0 0 0 0 0 0 0 0 0 1 1 0 0 1 1 1 0 1 1 1 0
 9   0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 1 1 1 1 0 0
10   0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 1 1 1 0 0 1
11   0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 1 1 0 1 0 1
12   0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 1 0 0 0 1 1
13   0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 0 1 1 0 1
14   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 1 0 0
15   0 0 0 0 0 0 0 0 0 1 1 0 1 1 1 1 1 1 0 1 0 1
16   0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 1 1 0 0 0 1
17   0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1 0 0 1 1 0 1
18   0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1 1 1 0 0 0
19   0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 0 1 1
20   0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 1 1 0 1 1
21   0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 1 1 1 1
22   0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 0
23   0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 1 1 1 1 0
24   0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 1 0 0
25   0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 1
...
50   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 1 1 0 0 0
51   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0
52   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 0 1 0 0 1
53   0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1 1 0 0 0 1 0
54   0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 1 1
55   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 1 0
56   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 1 1 1 0
57   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 1 1 1 0
58   0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 1 1 0 1
59   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 1 1 1 1
60   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 0
61   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 1 1 0 0
62   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 1 0 1 1 1
63   0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 1 1
64   0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 1 0 0 1 0
65   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 1 0 1 0
```

Figure 5.26: Typical solution evolution MRAS

```
 0   0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0 1 1 1 1 1
 1   0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 1 1
 2   0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 1 1 1 1 1 1 1
 3   0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 1 0 0 0 0 1 0
 4   0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 1 1 1 0 0 1
 5   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 0
 6   0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 1 0 1
 7   0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 1 0 1
 8   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 0 0 1 1
 9   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 0 0 1 1 1
10   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 0 0 0 1
11   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 0 1 0 1
12   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 0 0 1 1 1
13   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 0 0 1 1
14   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 0 1 0 1
15   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 0 0 0 1
16   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 1
17   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 0 0 0 0 0 1
18   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1 0 1 0 1
19   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 0 0 0 1
20   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 1
21   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1 0 0 0 1
22   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 0 1 0 0 1 1
23   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 1
24   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 1 0 1
25   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 0 0 0 1
26   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 0 1 0 1
27   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 1 0 1
28   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1
29   0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 0 0 0 1
30   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 1 0 1
31   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1 0 0 0 1
32   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1
33   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1
34   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1
35   0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1
```

Figure 5.27: Typical solution evolution CE

Figure 5.28: Comparison of Solutions



44 iterations was 12 575 622 (standard error 503 609). The results are displayed in Figure 5.28. One can see that the solutions of the MRAS indeed ameliorate between the 44th and the 65th iteration, hence there is some (slow) convergence. In addition, comparing the CE and the MRAS after 44 iterations one observes similar results as in the Inventory Example 1 (cf. Figure 5.15), namely that the Cross-Entropy algorithm produces more outliers on the one hand but also more very good solutions on the other hand.

Chapter 6

Conclusions

As we have seen, the Cross-Entropy method and the Model Reference Adaptive Search are conceptually very similar algorithms. Yet they differ considerably in the amount of proven theoretical statements and practical properties.

Due to several innate characteristics of the MRAS - the specific sequence of reference distributions, the parameter adaption, the inclusion of the initial density, etc. - one obtains that under certain conditions $\mathbb{E}_{\widehat{\theta}_k}[X] \to x^*$ as $k \to \infty$ under the sample density $f(\cdot\,;\widehat{\theta}_k)$. In special cases, e.g. if f is the density of a Normal distribution, additionally even $\mathrm{Var}_{\widehat{\theta}_k}[X] \to 0$. In order to verify the assumptions (which also limit the valid space for some parameters), a complete knowledge of the underlying system is required.
Similar results for the Cross-Entropy method do not exist. Even in a deterministic and discrete setting one cannot show that the sampling distribution converges to an optimal solution. And due to its relative simplicity compared to the MRAS, we cannot employ the methods used there to obtain related statements.

In contrast to what this may suggest, the Model Reference Adaptive Search yields in many cases far from optimal results in practical applications. This does not conflict with the theoretical results, as they rely basically on the fact that the complete space $\mathcal{X}$ can be searched if necessary (cf. remarks at the end of Chapter 4). Due to the limited resources of computational time and memory, this cannot be reproduced in practice.
Especially in high dimensional continuous settings (as in Section 5.2), the complicated weights in the update formula that contribute considerably to the theoretical findings cause numerous numerical problems and disturb rather than support the convergence to an optimal solution. For low dimensions or discrete problems, we observe in some cases (Section 5.3 Example 1, Section 5.4) less non-optimal solutions in the MRAS, but at the same time the Cross-Entropy method is apparently better in finding solutions very close to the optimum.
At the same time, the calculation of the weight factors (particularly the computation of the densities) and the parameter adaptation in Step 3 necessitate a significant additional amount of operations per iteration and per sample in comparison to the

CE.
Concerning the algorithm parameters, the Cross-Entropy method is apparently quite robust (compare e.g. Figures 5.21, 5.23) and consistently successfull with the same parameter set. Only the modified smoothing of the variance (2.10) is not always advisable.
The Model Reference Adaptive Search is much more susceptible to parameter choice and demands moreover good knowlegde not only of the theoretical properties (to determine e.g. the observation sizes in each iteration) but also of the range of function values in $\mathcal{X}$ and the behaviour of the method in a given example to find functioning parameters τ and v, etc.
Additionally, the implementation of the MRAS is much more difficult, as numerous numerical problems have to be dealt with, often leading to further problems in the parameter choice. Moreover, to obtain a stable behaviour one has to introduce for example bounds for the parameter ρ, based on observation and often (ideally) adapted to one specific simulation run. There were no such problems with the Cross-Entropy method.
Hence, in practice the MRAS is not better than the CE (often even worse) but more complicated in the implementation. Furthermore, it demands more computational resources and more care in the adjustment to concrete problems.

We have also observed that the Cross-Entropy method seems to be well-qualified to solve continous stochastic optimization problems, even with large solution space dimensions and independent of the prior knowledge about the system to be optimized. However, this applies rather to its ability to find good solutions than to the evaluation of their correct performance.
This is not restricted to the CE: In both algorithms the calculated function values are too optimistic, that is too small in minimization and too large in maximization problems. As this issue is less pronounced in the Model Reference Adaptive Search, we suggest a similar use of increasing observation sizes for the Cross-Entropy, even though the exponential growth necessary in the MRAS to fulfill the assumptions may be exaggerated.
Even so, using CE or MRAS for optimization one should be aware of this problem and, if need be, reevaluate the solution to obtain a more precise estimate of their value. Using the thresholds respectively sample quantiles or other estimators may also be worth considering.

Bibliography

[AKS58] Kenneth J. Arrow, Samuel Karlin, and Herbert Scarf. *Studies in the mathematical theory of inventory and production*. Stanford University Press - Stanford, California, 1958.

[Bau05] Harald Bauer. Operations Research Praktikum, Universität Ulm SS 2005.

[BD97] Peter J. Bickel and Kjell A. Doksum. *Mathematical statistics: basic ideas and selected topics*. Holden-Day series in probability and statistics. Calif. Holden Day, San Francisco, 1997.

[Ber95] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, Massachusetts, 1995.

[CFHM07] H.S. Chang, M.C. Fu, J. Hu, and S.I. Marcus. *Simulation-based Algorithms for Markov Decision Processes*. Communication and Control Engineering Series. Springer, London, 2007.

[CJK07] A. Costa, O.D. Jones, and D.P. Kroese. Convergence properties of the cross-entropy method for discrete optimization. *Operations Research Letters*, 35(5):573–580, 2007.

[Dam07] F. Dambreville. Cross-entropic learning of a machine for the decision in a partially observable universe. *Journal of Global Optimization*, 37(4):541–555, April 2007.

[dBKMR05] P.T. de Boer, D.P. Kroese, S. Mannor, and R.Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

[Dev86] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag New York, 1986.

[dM07] T. Homem de Mello. A study on the cross-entropy method for rare-event probability estimation. *INFORMS Journal on Computing*, 19(3):381–394, Summer 2007.

[dMR] T. Homem de Mello and R. Y. Rubinstein. Rare event estimation for static models via cross-entropy and importance sampling. submitted for publication.

[FH97] Michael C. Fu and Kevin J. Healy. Techniques for optimization via simulation: an experimental study on an (s,S) inventory system. *IEE Transactions*, 29:191–199, 1997.

[HFM] J. Hu, M.C. Fu, and S.I. Marcus. A model reference adaptive search method for stochastic global optimization. submitted for publication.

[HFM06] J. Hu, M.C. Fu, and S.I. Marcus. Model-based randomized methods for global optimization. In *Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems*, Kyoto, Japan, July 2006.

[HFM07] J. Hu, M.C. Fu, and S.I. Marcus. A model reference adaptive search method for global optimization. *Operations Research*, 55(3):549–568, May-June 2007.

[Hoe94] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In N. I. Fisher and P. K. Sen, editors, *Collected Works of Wassily Hoeffding*, Perspectives in Statistics. Springer-Verlag New York, 1994.

[KL51] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[KPR06] Dirk P. Kroese, Sergey Porotsky, and Reuven Y. Rubinstein. The cross-entropy method for continuous multi-extremal optimization. *Methodology and Computing in Applied Probability*, 8:383–407, 2006.

[Liu04] Jenny Liu. Global optimization techniques using cross-entropy and evolution algorithms. Master's thesis, Department of Mathematics, The University of Queensland, 2004.

[MMS05] I. Menache, S. Mannor, and N. Shimkin. Basic function adaption in temporal difference reinforcement learning. *Annals of Operations Research*, 134:215–238, 2005.

[MN83] Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1983.

[Mor82] Carl N. Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10(1):65–80, 1982.

[MRG03] S. Mannor, R. Y. Rubinstein, and Y. Gat. The cross-entropy method for fast policy search. In *The 20th International Conference on Machine Learning*, Washingtion, DC, August 2003.

[Pru89] Helmut Pruscha. *Angewandte Methoden der Mathematischen Statistik*. B. G. Teubner Stuttgart, 1989.

[Rie04] Ulrich Rieder. Operations Research II, Universität Ulm SS 2004.

[RK04] R.Y. Rubinstein and D.P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer, New York, 2004.

[Sha48] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.

[SW95] Adam Shwartz and Alan Weiss. *Large Deviations for Performance Analysis*. Stochastic Modeling Series. Chapman and Hall, 1995.

[ZM04] Quingfu Zhang and Heinz Mühlenbein. On the convergence of a class of estimation of distribution algorithms. *IEEE Transactions on Evolutionary Computation*, 8(2):127–136, April 2004.

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ich bin mir bewußt, dass eine unwahre Erklärung rechtliche Folgen haben wird.

Ulm, den 28. Juli 2008

_____

(Unterschrift)