A Stochastic Approximation Interpretation for Model-based Optimization Algorithms

A Dissertation Presented

by

Ping Hu

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Operations Research)

Stony Brook University

May 2012

Stony Brook University

The Graduate School

Ping Hu

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Jiaqiao Hu - Dissertation Advisor

Assistant Professor, Department of Applied Mathematics and Statistics

Eugene Feinberg - Chairperson of Defense

Professor, Department of Applied Mathematics and Statistics

Estie Arkin

Professor, Department of Applied Mathematics and Statistics

Petar Djurić

Professor, Department of Electrical and Computer Engineering

This dissertation is accepted by the Graduate School.

Charles Taber

Interim Dean of the Graduate School

Abstract of the Dissertation

A Stochastic Approximation Interpretation for Model-based Optimization Algorithms

by

Ping Hu

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Operations Research)

Stony Brook University

2012

This thesis studies a class of model-based randomized algorithms for solving general optimization problems. These are iterative algorithms that sample from and update an underlying distribution over the feasible solution space. We find that the model-based algorithms can be interpreted as the well-known stochastic approximation (SA) method. Following the connection between model-based algorithms and SA, we build a framwork to analyze the convergence and the convergence rate of these algorithms. Moreover, we present an instantiation of this framework which is the modified version of the Cross Entropy (CE) method, and analyze its convergence properties and numerical performance. In addition, we also propose a novel random search algorithm called Model-based Annealing Random Search (MARS). By exploiting its connection to SA we provide its global convergence result and analyze the asymptotic convergence rate as well. Finally, the empirical results of MARS show promising performance in comparison with some other existing methods.

To my family

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Jiaqiao Hu for his guidance, support and encouragement that help me finish my Ph.D. Degree.

I would also like to thank Dr. Eugene Feinberg, Dr. Estie Arkin, and Dr. Petar Djurić for their willingness to serve on my defense committee and for their helpful comments on my thesis.

Finally, I would like to thank my family and my girl friend Xiaomeng Lu for their endless love that supports me to conquer all difficulties I have ever met. I am also thankful to Mr. Yin Tian, Mr. Zheng Liu and all my other friends for our invaluable friendship that encourages me all the time.

Chapter 1

Introduction

In the past decades, optimization techniques have been very important in industry for improving system performance, including control systems, biostatistics, communication, scheduling, and so on. However, finding the globally optimal set of decision variables or parameter settings of the objective function is very difficult in general, especially for problems that contain many local optima. Furthermore, for many complex systems, it is often the case that the explicit relation between the objective function value and the underlying variables is unknown. In other words, the objective function is considered to be a "black-box", where we could only take observations of the outputs corresponding to some input values without knowing how these outputs are generated. For some problems, even if we know this relation explicitly, we may still be unaware of the entire structure of the problem. An example of such case is the famous traveling salesman problem (TSP), where we know the value of the cost function for every given path, but still cannot find the optimal path easily. These difficulties has inspired the development of many randomized search algorithms that only depend on the observation of the objective function value. Such methods include simulated annealing [40], [45], genetic algorithms [25], tabu search [24], nested partitions [62], pure adaptive search [77], stochastic ruler algorithm [72], stochastic comparison [26], stochastic adaptive search [76], and so on. These algorithms have been very successful for their outstanding performance on many difficult optimization problems.

Throughout this thesis, we focus on a class of randomized search algorithms called the model-based methods (see [20]). Different from other algorithms, model-based algorithms use an intermediate probability model as a guideline for the search. Typically, the algorithm works in an iterative way. In each iteration, the algorithm randomly generates a population of candidate solutions from certain probability distribution, and then use these samples to update the distribution that would be used to guide the search in the next iteration. The key idea of these algorithms is to iteratively modify the intermediate probability model based on the quality of the samples so that it will lead the search towards the promising region that contains high quality solutions. Examples of these model-based algorithms are ant colony optimization [14], [70], estimation of distribution algorithms [47], annealing adaptive search [57], the cross entropy method (CE) [60], [42], and the model reference adaptive search (MRAS) [32], [33]. Compared with other algorithms that perform the local enhancement search, the model-based algorithms explore the entire solution space at each iteration. This important feature has made these algorithms very successful for many optimization problems ([14],[2], [47], [79], [76]) as well as other applications ([59], [60], [42], [65], [48], [12], [33], [51]). However, in contrast to those existing algorithms that are well-studied, many model-based algorithms are heuristic approaches without theoretical convergence proofs, while in the meantime, the empirical study on the performance of model-based algorithms shows that there is still space for further improvement. For instance, before the collaborative work in [37], [34], there are only few convergence results of the CE method (e.g., [11]).

This thesis is based on the collaborative work by the author and Dr. Jiaqiao Hu and Dr. Hyeong Soo Chang (see [34], [37], [35], [36]). The first contribution of this thesis is that we establish a connection between the model-based algorithms and the well-known stochastic approximation (SA) method (literatures for SA can be found in [43], [44], [56], [66] and etc.). Generally speaking, a model-based algorithm can be interpreted as a stochastic approximation procedure. By exploiting this connection and the existing theories in gradient search and SA, we develop a framework to analyze the convergence properties of a class of model-based algorithms. In particular, we take the CE method as an exemplary instance. We slightly modify the standard version of the CE method (see [60], [42]) based on its SA interpretation, and prove the convergence of the modified version of CE.

Moreover, we provide some numerical examples to show that the modified CE inspired by SA has promising practical performance, and may outperform the standard version of CE. The second contribution of this thesis is that, inspired by annealing adaptive search (AAS, see [57]), CE, model reference adaptive search (MRAS, see [32], [33]), as well as the SA framework for model-based methods, we develop a novel adaptive randomized search algorithm called Model-based Annealing Random Search (MARS). By studying its connection to SA, it can be shown that MARS converges to the global optimum. Moreover, from the numerical examples it can be seen that MARS outperforms many existing searching algorithms, especially on those high dimensional optimization problems. Although most of the discussion is centered at the modified CE method and MARS, we hope that the idea can be generalized to other algorithms that also fall into this model-based category.

The rest of the thesis is structured as follows. In Chapter 2, we perform the literature review on some of the optimization algorithms, including methods for deterministic and stochastic optimizations in both continuous and discrete cases. Particularly, we briefly review the stochastic approximation method and the model-based methods. In Chapter 3, we present the stochastic approximation framework for certain model-based randomized algorithms, and then use it for analyzing the convergence properties of the cross entropy method in Chapter 4. Further in Chapter 5, based on the same framework, we present the MARS algorithm, prove its global convergence, and provide some numerical results as well. Finally, we conclude the thesis in Chapter 6. We also move the proofs for many lemmas and propositions to the Appendix at the end of this thesis.

Chapter 2

Literature Review

2.1 Deterministic Optimization

Optimization problems can be categorized into deterministic optimization and stochastic optimization. For deterministic optimization, we are concerned with finding the optimal solution to the problem of the form:

$$x^* \in \arg\max_{x \in \mathbb{X}} H(x), \tag{2.1}$$

where x is a vector of n decision variables, $\mathbb{X}$ is a non-empty set in $\Re^n$, and $H(\cdot) : \mathbb{X} \to \Re$ is a deterministic function. We assume the existence of an optimal solution x^*.

There are many algorithms designed for solving deterministic optimization problems, from the classic Newton method to the randomized algorithms. Some examples are the well-known steepest-descent methods, pure random search [9], pure adaptive search [77], simulated annealing [40], [45], tabu search [24], nested-partition [62], and genetic algorithms [25], etc. We now give a brief introduction to these optimization algorithms.

Pure Random Search (PRS) ([9]) is the algorithm that generates a sequence of uniformly distributed random points on the solution space. When the stopping criterion is met, the best candidate generated so far is used to approximate the global optimum. Though the idea behind the algorithm is easy, there are two disadvantages for its practical performance. First, the number of iterations needed for the kth best solution to come out will increase exponentially in k. Second, the complexity of the algorithm will also increase exponentially in the dimension of the solution space, see [9] for detailed discussions .

Pure Adaptive Search (PAS) ([77]) is an extension of the pure random search. It is an iterative algorithm for solving global optimization problem as well. The idea of the algorithm is simply as the following. Starting from a point $x_k \in S$, where S is the solution space, it first evaluates $Y_k = H(x_k)$, where H is the objective function, then it uniformly samples a point in the region whose function value is better than Y_k. It is very easy to show that PAS will almost surely converge to the global optimum. Moreover, under certain conditions, the complexity of this algorithm only increases at most linearly in the problem dimension (see [77]). Specifically, the complexity of the algorithm is measured by the expected number of iterations needed for a given accuracy. However, although PAS has nice theoretical properties, it is very hard to implement the algorithm because of the difficulty of sampling uniform random variates in an arbitrary region. Nevertheless, PAS provides an insight to develop an algorithm whose complexity only increases at most linearly [77] to the problem dimension. Meanwhile, other sampling techniques such as Hit-and-Run (e.g., [4], [76]) have been proposed to make the random sampling procedure computationally tractable.

The Simulated Annealing (SAN) ([40]) Algorithm is inspired by the annealing process in a physical system, where each feasible solution is analogous to a state of the system. The function value to be minimized corresponds to the internal energy of the system on that state. Doing minimization on the solution space is equivalent to bring the physical system to a state with the minimum internal energy. SAN is an iterative algorithm. At each step, the algorithm searches in the neighborhood of the current state, and probabilistically decides whether it should move to a new state. If the function evaluation at the new state is better than the current state, then it moves to the new state. Otherwise, it still moves to the new state with some probability p, or stays at the current state with probability $1-p$. In other words, the algorithm will sometimes transit to the candidate solution whose function value is worse than the current candidate. This is an important feature that leads the algorithm to escape from local optima. As the iteration number increases, the probability p will decrease to zero, and this feature guarantees the global convergence of the algorithm. In addition, under certain conditions, SAN can be extended to solve stochastic optimization problems (e.g., [21]).

Tabu Search (TS) ([24]) is a meta-heuristic approach to solve optimization problems. It is widely used in many areas such as scheduling, resource planning, telecommunications, network routing, manufacturing system, bioengineering, logistics and many others. Compared with other "memoryless" approaches such as genetic algorithms and annealing algorithms, Tabu Search introduces the "adaptive memory" feature that economically and effectively searches the feasible region (e.g., [22], [24]). This is a general framework and can be implemented differently depending on the structures of the problems. An important feature of Tabu search is that it could jump out of the neighborhood of a local optimum. For example, once the algorithm is doing local search around some local optimal solution, after visiting this area and making that area as "Tabu" (meaning to force this area to be a forbidden area), it will search towards the area outside this local optimum area to achieve the global optimum. Tabu search also has other variations such as restarting strategy, see [22] and [23] for detailed discussions.

The Nested Partition (NP) ([62]) Algorithm is a random search algorithm to solve global optimization problems. It divides the solution space into sub-regions under some predetermined scheme. At each iteration of the algorithm, we have a special sub-region which is considered to be the most promising region. Then we continue to divide it into M sub-regions while union the other regions as the surrounding region. After that, each of these regions is sampled using some random sampling scheme. Based on the function evaluations of those samples, the new promising region is determined. If the function performance (deterministic function) of one of those M subregions is found to be the best, it is the most promising region for the next iteration. However, if the function performance in the surrounding region is found to be the best, the algorithm then backtracks to the previous partition level and makes the old most promising region as the new most promising region. Then the new most promising region is partitioned in the similar fashion. It can be shown that once we consider each partition as a state, the algorithm produces a Markov chain. It also can be proved that, the algorithm will almost surely converge to the globally optimal solution, where the global optimum is considered as a singleton region (i.e., a region that contains one single solution).

Genetic Algorithm (GA) ([25]) is a meta-heuristic algorithm inspired by genetic science in biology. It inherits the idea of how the chromosomes evolve during the generation process between parents and their offsprings. In GA, each candidate solution can be encoded into a chromosome. At the beginning, the algorithm generates random population of chromosomes. After evaluating the function values, or the "fitness" of these chromosomes, the algorithm selects two chromosomes according to their fitness. Typically, the better fitness it is, the higher probability will it be chosen. Once we have chosen two chromosomes as the parents, we "crossover" these two chromosomes to generate two new chromosomes. Moreover, after the crossover step, we also do the “mutation” that mutates the new chromosomes in each position to get the offspring chromosomes. We do the same procedure until we get a new population, which is considered to be the second generation of the previous population. Iteratively, this new population continuously “evolves” under similar manners until a stopping criterion is met. Note that the practical implementation of the crossover step and mutation step will depend on the structure of the specific problem. Although there is no guarantee of the convergence of Genetic Algorithms so far, it is very popular for its good practical performance. GA falls into the class of Evolutionary Algorithm, which also includes Differential Evolution ([69]), Evolutionary Strategies ([7]) and Evolutionary Programming ([18]).

2.2 Stochastic Optimization

In practice, it is often the case that the objective function value cannot be evaluated explicitly. In other words, we could only take observations which are under the effect of noise, and the goal is to find the best decision variable that returns the maximum or minimum of the expected value of the objective function. For stochastic optimization, we are concerned to find the optimal solution to the problem of the form:

$$x^* \in \arg\max_{x \in \mathbb{X}} E_{\Psi}[F(x, \Psi)], \tag{2.2}$$

where $\mathbb{X}$ is a non-empty compact set in $\Re^n$. The quantity Ψ represents the stochastic input to the simulation, and its distribution might depend on x. We assume that $F(x, \Psi)$ is measurable and integrable with respect to the distribution of Ψ for all $x \in \mathbb{X}$ so that the expectation is well-defined. Furthermore, we let $f(x) = E_\Psi[F(x, \Psi)]$ and assume that $f(x)$ cannot be evaluated easily but the random variable $F(x, \Psi)$ can be observed via a simulation experiment at x.

There are many techniques designed to solve this stochastic optimization problem, such as the extended version of nested partition ([63]) and simulated annealing ([21]), sample average approximation [28], [41], stochastic ruler [72], stochastic comparison [26], COMPASS [31], adaptive search with resampling [3], deterministic shrinking ball and stochastic shrinking ball [3], and the well-known stochastic approximation (SA) (see [56], [43], [66], [44], and etc.). We now give a brief introduction to some of these algorithms.

The idea of the Sample Average Approximation (SAA) algorithm ([28], [41]) is simple. In the algorithm, the expected objective function value is approximated by the corresponding sample average of the simulated function values. It can be shown that, with the number of evaluations increasing to infinity on each candidate solution, the optimal solution for the sample average optimization will converge to the optimal solution of the original optimization problem. Also, the rate of convergence is provided.

The Stochastic Ruler (SR) Algorithm ([72]) is designed to solve stochastic optimization problems. It assumes a neighborhood structure on the solution space so that any two candidate solutions can be reachable from each other (i.e., we say x is reachable from y if there exists a sequence $y_0, \ldots, y_n$ such that $y_0 = y$, $y_n = x$ and $y_{k+1} \in \mathbf{N}(y_k)$ for $k = 0, \ldots, n-1$, where $\mathbf{N}(z)$ denotes the neighborhood of z. See [72]), and this property is necessary for the global convergence. The algorithm works in an iterative way and constructs a transition probability on the solution space. Specifically, in each iteration, the algorithm will randomly select a new solution in the neighborhood of the current candidate solution under certain probability distribution. Then it will compare the simulated function evaluation on the new candidate solution to a "stochastic ruler", which is uniformly distributed on the range of the objective function values. Depending on the comparison results, the algorithm decides whether it should move to the new solution from the current one. Along this procedure, the algorithm in fact generates a Markov chain on the solution space. By analyzing the limiting distribution, it can be shown that under general conditions, the algorithm will converge to the global optimum with probability one, while the rate of convergence is also provided ([72]).

The Stochastic Comparison (SC) Algorithm ([26]) is very similar to Stochastic Ruler Algorithm, except that, instead of comparing the sample average of the objective function to a "random ruler", it directly compares the sample average of the objective functions on two candidate states, and then decides whether it will transit from the old state to the new one. Similar to Stochastic Ruler Algorithm, when carefully controlling the number of simulations on each state, it will converge to the global optimum ([26]) .

Adaptive Search with Resampling (ASR) ([3]) is a framework of stochastic optimization problems on continuous domain. It consists of three important components which are a sampling strategy, a resampling strategy, and an acceptance criterion. For certain iterations, it samples a candidate solution on the solution space using the sampling strategy and takes several observations on that solution. Then, based on the acceptance criterion, the algorithm decides whether to include the new solution in the set of accepted solutions. The idea of this step is to ensure those promising solutions will eventually be evaluated enough. For other iterations, it resamples the set of accepted solutions to improve the estimator of the global optimum. This step makes the search more economically and effectively, since it spends more simulation budget on the solutions in the set that are considered more promising, and spends less on other inferior solutions. The algorithm can be proved for global convergence, under some mild assumptions.

Deterministic Shrinking Ball (DSB) and Stochastic Shrinking Ball (SSB) ([3]) are based on Pure Random Search ([9]). However in DSB and SSB, it is not necessary take several observations on each candidate solution to get the approximation of the expected function value using the sample average. Instead, the estimation of the objective function on each candidate solution x is the average of the objective function observations on all samples that are close to the candidate solution. Specifically, in DSB it constructs a “ball” that centers at x with some radius r, and the average function observations on those samples in the ball is used to estimate the expected function value on x. While in SSB, it constructs a "ball" that centers at x and contains n nearest samples inside the ball, and those n samples will be used to estimate the expected function value on x. Under certain conditions, convergence results can be provided for both DSB and SSB.

2.3 Stochastic Approximation

Stochastic Approximation (SA) is a general technique for finding approximation for roots or the minima or maxima of a given function. SA is especially useful when there is scant information of the explicit value and structure of the objective functions and one can only perform experiment or simulation to get observations which involve noise. Robbins and Monro ([56]) first discussed a stochastic approximation technique for estimating the root of a regression function, and later Kiefer and Wolfowitz ([39]) presented the procedure of finding minima and maxima of regression functions. We note that Kiefer-Wolfowitz procedure can be interpreted as an extension to the classic Steepest-Descent method for finding minima and maxima, where the true gradient is estimated by the finite difference method. We first review the literatures of the gradient estimation in Stochastic Approximation, and then discuss the general Stochastic Approximation.

2.3.1 Gradient Estimation in Stochastic Approximation

The well-known Steepest-Descent method is a classic approach which is carried out originally for searching local minima of an objective function. The method could also be easily adapted for maximization problem problem (2.1). Note that there are different ways to present the Steepest-Descent method. To be consistent with the term "descent", we present it as a method for solving minimization problems. Specifically, one starts with an initial guess x_0, and iteratively generate a sequence $x_1, x_2, ...$ such as

$$x_{n+1} = x_n - a_n \nabla f(x_n),$$

where $\{a_n\}$ is a sequence of constants and f is the objective function. For many practical problems, however, the true gradient cannot be evaluated explicitly. Therefore, some gradient estimation techniques need to be introduced. There is a vast amount of literatures that discuss the gradient estimation methods. See Fu [19] for a review. The main gradient estimation methods can be divided into two categories: Indirect gradient estimation and direct gradient estimation. For indirect gradient estimation, the estimator of the true gradient usually involves bias. Two examples of indirect gradient estimation are the finite different method [39] and the simultaneous perturbation [66] method. In the finite difference method, the gradient is estimated by a one-sided finite difference

$$\frac{\hat{f}(x + c_i e_i) - \hat{f}(x)}{c_i}$$

or a two-sided finite difference

$$\frac{\hat{f}(x + c_i e_i) - \hat{f}(x - c_i e_i)}{2c_i}$$

where $c = (c_1, ..., c_d)$ is the vector of differences (amount of perturbation), θ is the candidate solution on whose gradient is being estimated, e_i denotes the unit vector in the ith direction, and $\hat{f}$ is an observation of the objective function on x that involves noise. We note that we need to carefully choose the parameter sequence $\{a_n\}$ and $\{c_n\}$. For $\{a_n\}$ it needs to decrease to zero to make the algorithm convergent, while it cannot decrease too quickly which may cause the prematurity of the search (i.e., the search sticks at some region before it reaches a global/local optimum). A typical condition on $\{a_n\}$ is that

$$\lim_{n \to \infty} a_n = 0, \quad \sum a_n = \infty, \quad \text{and} \quad \sum a_n^2 < \infty.$$

For c_n, it also has to be chosen carefully to balance the variance and bias. A typical condition on c_n is that

$$\sum a_n c_n < \infty, \quad \sum \frac{a_n^2}{c_n^2} < \infty.$$

Note that in one-sided finite difference method, it requires d+1 evaluations on each solution; in two-sided finite difference method, it requires $2d$ evaluations on each solution, where d is the dimension.

Simultaneous Perturbation Stochastic Approximation (SPSA) method is given by Spall [66]. It is similar to the finite difference method, except that it only needs 2 evaluations on each solution. In SPSA, the gradient estimator is

$$\frac{\hat{f}(x + c\Delta) - \hat{f}(x - c\Delta)}{2c_i\Delta_i},$$

where $\Delta = (\Delta_1, ...\Delta_d)$ is a $d-$dimension vector of perturbations. Specifically, Δ is a vector of d mutually independent mean-zero random variables $(\Delta_1, ...\Delta_d)$ satisfying some conditions, see [66] for details. It can be shown that under some conditions, SPSA will converge locally, and it only requires 2 evaluations on each solution, regardless of the problem dimension. In summary, the advantage of finite difference method and SPSA is that they are model-free and easy to implement, since they only depend on the value of function evaluations. Meanwhile, the practical performance of finite difference method and SPSA may depend on the setting of parameters, see [67] for further discussions.

On the other hand, if some additional knowledge about the objective function is available, direct gradient estimation often provides the unbiased estimator which could lead to faster convergence rate when implemented in a stochastic optimization problem. Moreover, it does not need to determine the difference sequence c_n in the finite different method, and it makes the computation more efficient. A brief summary of the direct gradient estimation can be found in [19].

2.3.2 General Stochastic Approximation

A general Stochastic Approximation (SA) procedure is an iterative technique for finding roots and extreme values of an objective function. It has the following general form:

$$X_{n+1} = X_n - V_n, \quad n = 0, 1, 2, ...,$$

where V_n is a sequence of random variable and X_0 is an initial random variable (see [16]). The stochastic term V_n may have different meanings. For example (see [66]), assume $f(x)$ is the objective function to be minimized on a solution space $\mathbb{X}$, and $f(\cdot)$ is differentiable with respect to $x \in \mathbb{X}$. A steepest-descent style SA has the following form:

$$\hat{x}_{n+1} = \hat{x}_n - \alpha_k \tilde{f}_k(\hat{x}_n),$$

where $\tilde{f}_n(\hat{x}_n)$ is an estimator of the true gradient at $\hat{x}_n$, and $\{\alpha_k\}$ is a sequence of constants (gain sequence). In this case, V_n stands for the product of the gradient estimator and a constant.

A main research interest focuses on the conditions on the variables V_n

that ensure the almost sure convergence of X_n, see Robbins and Monro 1951 [56], Fabian 1968 [17], Ljung 1977 [50], Kushner and Clark 1978 [43], Evans and Weber 1986 [16], Benveniste et al. 1990 [6], Benaim 1996 [5], Kushner and Yin 1997 [44], Borkar 2008 [8] etc.

2.4 Model-Based Method

As we mentioned before, the model-based methods differs from other random search methods in that a set of population is generated at each iteration by sampling from an intermediate probability model, while this population is then used to update the model for the random sampling in the next iteration. Some examples of model-based methods are ant colony optimization [14], [70], estimation of distribution algorithms [47], annealing adaptive search [57], the cross entropy method [60] and model reference adaptive search [32].

The Ant Colony Optimization (ACO) is a classical nature-inspired algorithm ([14], [70], [13]). It is inspired by the natural phenomenon that how ants would travel between food source and home, and then it is applied to solve many difficult optimization problems such as Traveling Salesman Problem (TSP) (see [14]). It is well known that ants use pheromone as a communication media to cooperate when they are traveling. Based on that idea, in Ant Colony Optimization, it uses "Artificial Ants" as the cooperating agents to find good solutions for the optimization problem ([14], [13]).

The Annealing Adaptive Search (AAS) algorithm was introduced in Romeijn and Smith ([57]). The algorithm generates candidate solutions by sampling from a sequence of Boltzmann distributions parameterized by the temperatures that are time-dependent. As the iteration number increases, the temperature will decrease to zero so that the Boltzmann distribution will asymptotically concentrate on the region that contains the global optimum. The AAS algorithm has the nice theoretical property that the expected number of iterations only increases linearly with the number of dimension ([57], [76]). However, AAS is difficult to directly implement since it is very hard to generate random samples under a Boltzmann distribution. Fortunately, some Markov chain-based sample techniques such as Hit and Run can be used to soften the difficulty ([58], [76], [78]).

The cross entropy (CE) method was introduced to solve rare event probability estimation problem ([59]), and it then became a general approach for solving optimization problems ([60]). The CE method involves an iterative procedure where each iteration consists of two steps. First, it generates a population of random samples according to a specified mechanism; second, it updates the mechanism based on the performance of the sample generated in the first step. The CE method involves the idea of Importance Sampling (IS). Specifically, after the random mechanism generates the sample, the CE method tends to pay more attention to those "good samples". In other words, based on the performances of those random samples, CE method tends to search more on those feasible regions that seems to be more promising, and this feature may explain its excellent practical behavior. However, there are only few results for the convergence of CE (e.g., [11]).

The model reference adaptive search (MRAS) is a randomized method given by Hu et al. ([32], [33]) to solve both continuous and combinatorial optimization problems. MRAS resembles CE in that they both work with a family of parameterized distributions on the solution space ([32]). The key idea of MARS is to use a sequence of pre-determined intermediate distributions as the reference distribution models to facilitate the search. Specifically, the reference distribution is pre-specified and has convergence property. Therefore it could guide the updating of the parameterized distributions so that the search will asymptotically concentrate on the region containing the global optimum. The significance of MRAS is that it provides a framework for global optimization that allows the flexibility of choosing the reference models. Also, a proof of the global convergence for one instantiation of the framework can be found in [32].

Chapter 3

A Stochastic Approximation Method for Studying a Class of Randomized Optimization Algorithms

3.1 A Framework for Model-based Algorithms

In this chapter we focus on finding the optimal solution to the problem of the form:

$$x^* \in \arg\max_{x \in \mathbb{X}} H(x), \tag{3.1}$$

where $H(\cdot) : \mathbb{X} \rightarrow \Re$ is a deterministic objective function, $\mathbb{X} \subset \Re^n$ is a solution space that could be continuous or discrete, x is a decision variable (a vector with n entries). We assume that $H(\cdot)$ is bounded, $\mathbb{X}$ is compact, and there exists a global optimal solution x^*. Note that here we haven't addressed any other assumptions on $H(\cdot)$, i.e., $H(\cdot)$ may not necessarily be differentiable or continuous, and it may possess multiple local optima.

The framework of iterative model-based algorithms for solving (3.1) generally consists of the following steps at each iteration:

1) randomly generate candidate solutions by sampling from a distribution g_k;

2) observe the objective function values of these generated candidate solutions, update g_k to obtain a new distribution g_{k+1};

where g_k is the probability model (specifically, a probability density function or a probability mass function) at the kth iteration of the algorithm. Our key idea here is to find a desirable sequence of distributions $\{g_k\}$ that will converge to an "optimal distribution" g^* which concentrates its mass around the area containing the global optimum, as k goes to infinity. Intuitively, if we generate random samples from g^*, the samples should be close to the global optimizer with high probability. Throughout this thesis, we call the sequence $\{g_k\}$ the reference distributions, as it is used as a guideline for the random sampling procedures.

As in the existing model-based methods, different $\{g_k\}$ is chosen as the reference probability model. Some examples are:

a) proportional selection scheme which is introduced in MRAS:

$$g_{k+1}(x) = \frac{S(H(x))g_k(x)}{E_{g_k}[S(H(X))]},$$

where $S(\cdot)$ is a positive increasing function and X is a generic random variable taking values in $\mathbb{X}$,

b) important sampling scheme which is used in the CE method:

$$g_{k+1}(x) = \frac{S(H(x))f_{\theta_k}(x)}{E_{\theta_k}[S(H(X))]},$$

where f_{θ_k} is some parameterized sampling distribution, and $S(\cdot)$ is a positive increasing function as in a),

c) Boltzmann distribution with decreasing temperature schedule which is used in AAS:

$$g_{k+1}(x) = \frac{e^{H(x)/T_{k+1}}}{\int_{\mathbb{X}} e^{H(x)/T_{k+1}}dx} = \frac{e^{H(x)(\frac{1}{T_{k+1}}-\frac{1}{T_k})}g_k(x)}{E_{g_k}[e^{H(X)(\frac{1}{T_{k+1}}-\frac{1}{T_k})}]},$$

where $\{T_k\}$ is a sequence of parameters determined by an annealing schedule;

Throughout this thesis, for any reference distribution g (i.e., g_k in a) and c) above), $P_g(\cdot)$ and $E_g[\cdot]$ denote the probability and expectation taken with respect to the density/mass function g. On the other hand, for any sampling distribution f_θ that is parameterized by θ, $P_\theta(\cdot)$ and $E_\theta[\cdot]$ denote the probability and expectation taken with respect to the density/mass function f_θ as in case b).

Intuitively, in each of the updating procedure mentioned above, the distribution model is updated according to the objective function value on each solution in the feasible region. It "tilts" the current probability model to increase the probability density/mass on solutions whose objective function values are relatively large, and reduce the probability density/mass on those solutions whose performances are relatively poor. Note that in a) and b), $S(\cdot)$ is a positive increasing (possibly iteration-varying) function to keep the density/mass positive. It can be shown that, under some mild conditions, the sequence of the probability models in a) will converge to a distribution g^* that assigns all its mass on the global optimum (see [32]). This desired property also holds for case c).

An obvious difficulty in all these three cases is that, the sequence $\{g_k\}$ depends on $H(\cdot)$, therefore is unknown a priori. Moreover, even if we know g_k explicitly, it is difficult to use g_k to effectively generate random samples. To overcome this difficulty, at each iteration we try to use a surrogate sampling distribution to approximate g_k. Naturally, we would expect two features on the surrogate distribution. First, it should be easy to generate random samples on the solution space from the surrogate distribution. Second, it should be as close to g_k as possible so that the surrogate distribution could share some properties with the reference distributions, e.g., if the reference distributions converge to a limiting distribution, the surrogate distributions are expected to converge to the same limiting distribution as well. To achieve the first feature, we specify a family of parameterized distributions $\{f_\theta(\cdot), \theta \in \mathbf{\Theta}\}$, where $\mathbf{\Theta}$ is the parameter space. Once we choose a surrogate sampling distribution from this family, it should be easy for us to generate random samples from this distribution. To achieve the second feature, we would select the parameterized sampling distribution f_{θ_k} from the distribution family so that the Kullback-Leibler (KL) divergence between the reference distribution g_k and the sampling distribution f_{θ_k} is minimized, i.e.,

$$\theta_k = \arg\min_{\theta \in \mathbf{\Theta}} \mathscr{D}(g_k, f_\theta), \tag{3.2}$$

where $\mathscr{D}(g_k, f_\theta) := \arg\min_{\theta \in \mathbf{\Theta}} E_{g_k}\left[\ln \frac{g_k(X)}{f_\theta(X)}\right]$, and X here is a random variable whose probability density/mass function is g_k.

Although there are other ways to construct the surrogate distributions (see [58],[76] and [78] for AAS), this approach has the following advantages. First, by choosing a special distribution family called the Natural Exponential Families (NEFs) as the parameterized families, the random samples could be generated easily. Moreover, under NEFs, the optimization problem (3.2) can be solved analytically in closed form for any arbitrary g_{k+1}, and this attractive feature has made our approach very easy to implement. In addition, the task of updating the entire sampling distribution is simplified to the task of updating its associated parameter. We now provide the definition of NEFs as follows:

Definition 1. *A parameterized family* $\{f_\theta(\cdot), \theta \in \mathbf{\Theta} \subseteq \Re^d\}$ *on* $\mathbb{X}$ *is called a natural exponential family if*

$$f_\theta(x) = \frac{\exp\left(\theta^T \Gamma(x)\right)}{\int_{\mathbb{X}} \exp(\theta^T \Gamma(x)) \nu(dx)},$$

where $\Gamma : \Re^n \to \Re^d$ *is a continuous mapping,* ν *is the Lebesgue/discrete measure and the natural parameter space* $\mathbf{\Theta}$ *consists of all the* θ *that satisfies* $\int_{\mathbb{X}} \exp(\theta^T \Gamma(x)) \nu(dx) < \infty$*. Also,* $K(\theta) := \ln \int_{\mathbb{X}} \exp(\theta^T \Gamma(x)) \nu(dx)$ *is called the log partition function.*

Note that if X is a random variable with the probability density/mass function f_θ, then $\Gamma(X)$ turns out to be the sufficient statistics of X. Moreover, we define $m(\theta) := E_\theta[\Gamma(X)]$, i.e., $m(\theta)$ is the expected value of the sufficient statistics $\Gamma(X)$ under distribution f_θ.

Many commonly used distributions belong to NEFs, e.g., exponential distribution, univariate/multivariate normal distributions, Poisson distributions, etc. NEFs have an important property (see [53]) that the function $K(\theta)$ is strictly convex on the interior of $\mathbf{\Theta}$ with $\nabla K(\theta) = E_\theta[\Gamma(X)]$. In addition, the Hessian matrix of $K(\theta)$ is $\mathrm{Cov}_\theta[\Gamma(X)]$, where $\mathrm{Cov}_\theta[\cdot]$ is the covariance with respect to f_θ. Note that $\mathrm{Cov}_\theta[\Gamma(X)]$ is the Jacobian of $m(\theta)$ and $\mathrm{Cov}_\theta[\Gamma(X)]$ is strictly positive definite and invertible, therefore, $m(\theta)$ is also invertible by the inverse function theorem. As a consequence, there is a one-to-one mapping between θ and $m(\theta)$. In other words, for NEFs, the expected value of the sufficient statistics would be sufficient to describe the associated distribution. For instance, consider the exponential distribution with parameter λ. It is obvious that $\Gamma(x) = x$, $\theta = -\lambda$, and $m(\theta) = 1/\lambda$. Another simple example is the univariate normal distribution with mean μ and variance σ^2. It can be shown that $\Gamma(x) = (x, x^2)^T$, $\theta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})^T$ and $m(\theta) = (\mu, \sigma^2 + \mu^2)^T$. In summary, $m(\theta)$ could determine the parameterized distribution f_θ as well.

Let $g_{k+1}(x)$ be a given reference distribution (e.g., case a), b) and c) in the previous example), and f_{θ_k} be the surrogate sampling distribution obtained at the kth iteration of the algorithm. We consider a general reference distribution in the following form:

$$\widetilde{g}_{k+1}(x) = \alpha_k g_{k+1}(x) + (1 - \alpha_k) f_{\theta_k}(x), \tag{3.3}$$

where $\alpha_k \in (0, 1]$ $\forall k$ is a "smoothing" parameter ensuring that the new distribution $f_{\theta_{k+1}}$ obtained by minimizing $\mathscr{D}(\widetilde{g}_{k+1}, f_\theta)$ does not deviate too much from the current distribution f_{θ_k}.

3.2 Connection to Stochastic Approximation

When $\{\widetilde{g}_{k+1}\}$ in (3.3) is used as the reference distribution sequence, the following key lemma states a key relation between the two successive mean vectors in model-based algorithms.

Lemma 3.2.1. *If f_θ belongs to NEFs and the new parameter θ_{k+1} obtained via minimizing $\mathscr{D}(\widetilde{g}_{k+1}, f_\theta)$ is an interior point of $\mathbf{\Theta}$, i.e., $\theta_{k+1} \in int(\mathbf{\Theta})$ for all k, where $int(\mathbf{\Theta})$ denotes the set of interior points of $\mathbf{\Theta}$, then*

$$m(\theta_{k+1}) - m(\theta_k) = -\alpha_k \nabla_\theta \mathscr{D}(g_{k+1}, f_\theta)\big|_{\theta=\theta_k} \quad \forall\, k.$$

Proof. Since $\theta_{k+1} \in int(\mathbf{\Theta})$, it satisfies the first order necessary condition for optimality. It follows from (3.2) that

$$\nabla_\theta E_{\tilde{g}_{k+1}}[\ln f_\theta(X)]\big|_{\theta=\theta_{k+1}} = 0.$$

By the dominated convergence theorem, we could exchange the order of the expectation and the differential,

$$E_{\tilde{g}_{k+1}}[\nabla_\theta \ln f_\theta(X)\big|_{\theta=\theta_{k+1}}] = 0,$$

which in further gives us

$$m(\theta_{k+1}) = E_{\theta_{k+1}}[\Gamma(X)] = E_{\tilde{g}_{k+1}}[\Gamma(X)].$$

It follows from (3.3) that

$$m(\theta_{k+1}) = \alpha_k \frac{E_{g_k}[S(H(X))\Gamma(X)]}{E_{g_k}[S(H(X))]} + (1-\alpha_k)m(\theta_k).$$

Therefore, the recursion of the mean parameter vector can be written as

$$\begin{aligned} m(\theta_{k+1}) &= m(\theta_k) + \alpha_k \frac{E_{g_k}[S(H(X))(\Gamma(X) - m(\theta_k))]}{E_{g_k}[S(H(X))]} \\ &= m(\theta_k) - \alpha_k \nabla_\theta \mathscr{D}(g_{k+1}, f_\theta)\big|_{\theta=\theta_k}, \end{aligned} \tag{3.4}$$

where the last equality follows from the properties of NEFs. □

We note that Lemma 3.2.1 shows that the updating direction of the mean vector at each step is of the negative gradient of a *time-varying* objective function for the minimization problem $\min_{\theta\in\mathbf{\Theta}} \mathscr{D}(g_{k+1}, f_\theta)$ $\forall k$. As a result, the updating procedure in the model-based algorithms could be interpreted as a gradient search method. Particularly, if the sequence of the reference model $\{g_k\}$ converges to some limiting distribution function g^*, then the algorithm is implicitly solving another optimization problem, which is finding the optimal parameter θ^* so that f_{θ^*} is the best approximation for g^*. This gradient interpretation suggests that, those model-based algorithms that can be accommodated by this framework are essentially gradient recursions that implicitly transform the original optimization problem (3.1) into a new optimization problem on the parameter space with smooth structures (i.e., continuous and differentiable), and this could explain why the model-based algorithms work well for those optimization problems with little structure information. Moreover, the ample theories from gradient methods and SA enable us to perform theoretical analysis on model-based algorithms. In the next chapter, we use the (modified) CE method as a concrete example.

Chapter 4

Convergence Properties of the CE method

4.1 Connection between CE and SA

As we mentioned in the literature review, the CE method was originally motivated by the rare event probability estimation [59]. It was then found that the method could be adapted to solve combinatorial and continuous optimization problems [60]. Furthermore, the CE method works very well for multi-extremal nonlinear optimization [60] even in high dimensions. CE also belongs to model-based algorithms. As we mentioned in the previous chapter, CE uses the following model updating procedure:

$$g_{k+1}(x) = \frac{\varphi(H(x))f_{\theta_k}(x)}{E_{\theta_k}[\varphi(H(X))]},$$

where $\varphi(\cdot)$ is a positive increasing function and f_{θ_k} is the parameterized sampling distribution with the parameters calculated by (3.2). Note that throughout this chapter we use the notation φ instead of S in Chapter 3 for the analysis of the CE method.

If we apply this updating procedure to Lemma 3.2.1, then (3.4) becomes

$$\begin{aligned} m(\theta_{k+1}) &= m(\theta_k) + \alpha_k \frac{E_{\theta_k}[\varphi(H(X))(\Gamma(X) - m(\theta_k))]}{E_{\theta_k}[\varphi(H(X))]} \\ &= m(\theta_k) + \alpha_k \nabla_\theta \ln E_\theta[\varphi(H(X))]\big|_{\theta=\theta_k}, \end{aligned} \tag{4.1}$$

where the interchange of derivative and integral above is guaranteed by the dominated convergence theorem. We can see that, in CE, the updating procedure can be interpreted as the gradient method for the maximization problem $\max_{\theta\in\mathbf{\Theta}} \ln E_\theta[\varphi(H(X))]$. Because of the monotonicity of $\varphi(\cdot)$ and $\ln(\cdot)$, it can be seen that the optimal solution θ^* of this maximization problem would correspond to the sampling distribution f_{θ^*} that mostly concentrates on the area with maximum values of $H(\cdot)$.

We present the idealized version of CE for solving (3.1) as below:

Algorithm 4.1. Idealized CE Method

Step 0: Choose an initial pdf/pmf $f_{\theta_0}(\cdot)$ on $\mathbb{X}$, with $\theta_0 \in int(\mathbf{\Theta})$. Specify a non-decreasing function $\varphi(\cdot) : \Re \to \Re^+$, a gain sequence $\{\alpha_k\}$, two constants, $\rho \in (0, 1)$ and $\varepsilon > 0$. Set $k = 0$.

Step 1: Calculate the $(1-\rho)$-quantile γ_k of $H(X)$, i.e.,

$$\gamma_k := \sup_l\{l : P_{\theta_k}(H(X) \geq l) \geq \rho\},$$

where X is a random vector with respect to distribution f_{θ_k} that takes values in $\mathbb{X}$.

Step 2: Update parameter $\theta_{k+1} = \arg\min_{\theta\in\boldsymbol{\Theta}} \mathscr{D}(\widetilde{g}_{k+1}, f_\theta)$, where $\widetilde{g}_{k+1}$ is given by (4.2).

Step 3: If a stopping rule is satisfied, then terminate and return θ_{k+1}; otherwise set $k = k+1$ and go to Step 1.

Following the idea in the framework, instead of using $g_{k+1}(x) = \frac{\varphi(H(x))f_{\theta_k}(x)}{E_{\theta_k}[\varphi(H(X))]}$ as the model updating procedure, we use $\tilde{g}_{k+1}(x)$ as the reference distribution where $\tilde{g}_{k+1}(x)$ is updated by

$$\widetilde{g}_{k+1}(x) = \alpha_k \frac{\varphi(H(x))I(H(x),\gamma_k)f_{\theta_k}(x)}{E_{\theta_k}[\varphi(H(X))I(H(X),\gamma_k)]} + (1-\alpha_k)f_{\theta_k}(x) \tag{4.2}$$

with $\alpha_k \in (0,1]\ \forall\, k$, where $\varphi(\cdot)$ is a non-decreasing positive-valued function as $\varphi(\cdot)$ in the framework of model-based methods, γ_k is the (1-ρ)-quantile of $H(X)$, and $I(\cdot,\cdot)$ is a threshold function:

$$I(y,\gamma) := \begin{cases} 1 & \text{if } y \geq \gamma, \\ (y-\gamma+\varepsilon)/\varepsilon & \text{if } \gamma-\varepsilon < y < \gamma, \\ 0 & \text{if } y \leq \gamma-\varepsilon. \end{cases}$$

Intuitively, using such a threshold function could lead the computational effort to approximately focus on the top ρ-percent of the selected "elite" solutions. This approach inherits the idea of importance sampling as well as the spirit in the selection scheme employed in many population-based approaches such as genetic algorithms. Note that for pure technical reasons, we make another slight modification of the standard CE method by replacing the original indicator function with our threshold function. On the other hand, the positive-valued function $\varphi(\cdot)$ is needed to ensure the reference probability density to be positive in the case where $H(x)$ is negative. In the standard implementation of CE, $\varphi(\cdot)$ is often taken to be a constant function, i.e., $\varphi(x) \equiv 1$ for all x.

Note that the approach of using the smoothing parameter α_k here is slightly different from the standard CE method presented by Rubinstein and Kroese [60]. In [60], the smoothing parameter α_k is used for smoothing the parameter (i.e., θ_k) rather than the entire parameterized distribution as we perform here.

The following proposition is an instantiation of Lemma 3.2.1.

Proposition 4.1.1. *In Algorithm 1, if f_θ belongs to NEFs and $\theta_{k+1} \in int(\mathbf{\Theta})$ for all k, then the mean parameter vector function $m(\theta_{k+1})$ satisfies*

$$\begin{aligned} m(\theta_{k+1}) = \alpha_k \frac{E_{\theta_k}[\varphi(H(X))I(H(X),\gamma_k)\Gamma(X)]}{E_{\theta_k}[\varphi(H(X))I(H(X),\gamma_k)]} \\ + (1-\alpha_k)E_{\theta_k}[\Gamma(X)] \quad \forall\, k. \end{aligned} \tag{4.3}$$

Proof. Follows from the proof of Lemma 3.2.1. □

Once we get Proposition 3.1, we are able to write the gradient recursion of the mean vector as in Lemma 2.1. However, Algorithm 2.1 involves the expectation and the quantile which cannot be evaluated directly. Thus, based on the sample generated by the parameterized distribution, we have to use sample average to estimate the true expectation and use sample quantile to estimate the true quantile value. This brings out the following Monte Carlo version of the CE method.

Algorithm 4.2. The Monte-Carlo Version of CE

Step 0: Choose an initial pdf/pmf $f_{\widehat{\theta}_0}(\cdot)$ on $\mathbb{X}$, $\widehat{\theta}_0 \in int(\mathbf{\Theta})$. Specify a bounded non-decreasing function $\varphi(\cdot) : \Re \to \Re^+$ satisfying $\inf_y \varphi(y) > 0$, parameter sequences $\{\alpha_k\}$ and $\{\lambda_k\}$, constants $\rho \in (0,1)$ and $\varepsilon > 0$. Set $k = 0$.

Step 1: Randomly sample N_k i.i.d. solutions $\Lambda_k = \{X_1, \ldots, X_{N_k}\}$ from the distribution $f_{\widehat{\theta}_k}$.

Step 2: Calculate the sample $(1-\rho)$-quantile $\widehat{\gamma}_k = H_{(\lceil (1-\rho)N_k \rceil)}$, where $\lceil a \rceil$ is the ceiling function that returns the smallest integer greater than a, and $H_{(i)}$ is the ith order statistic of the sequence $\{H(X_i)\}_{i=1}^{N_k}$.

Step 3: Compute a new parameter $\widehat{\theta}_{k+1} = m^{-1}(\eta_{k+1})$, where $\eta_0 := m(\widehat{\theta}_0) =$

$E_{\widehat{\theta}_0}[\Gamma(X)]$, and

$$\begin{aligned}\eta_{k+1} = {} & \alpha_k \frac{\sum_{x\in\Lambda_k} \varphi(H(x)) I(H(x), \widehat{\gamma}_k)\Gamma(x)}{\sum_{x\in\Lambda_k} \varphi(H(x)) I(H(x), \widehat{\gamma}_k)} \\ & + (1-\alpha_k)\Big(\frac{\lambda_k}{N_k}\sum_{x\in\Lambda_x}\Gamma(x) + (1-\lambda_k)\eta_k\Big) \qquad (4.4)\end{aligned}$$

is an empirical estimate of recursion (4.3) based on the sampled solutions in Λ_k.

Step 4: If a stopping rule is satisfied, then return $\widehat{\theta}_{k+1}$ or the best candidate solution thus far and terminate; otherwise set $k = k+1$ and go to Step 1.

Since the main purpose of this section is to analyze the convergence properties of the CE method, we do not specify a stopping criterion. In practice, the user could terminate the algorithm whenever it reaches the computational budget or the parameter $\widehat{\theta}_k$ does not vary too much for several successive iterations. A detailed discussion of the stopping criterion can be found in [60]. On the other hand, from a practical point of view, when the algorithm terminates, we could choose to return the current best solution, instead of the solution generated in the last iteration.

Since Algorithm 4.2 is randomized, we need to clarify some probabilistic definitions as follows:

- $(\Omega, P, \mathscr{F})$: A probability space on which the random samples $\Lambda_0, \Lambda_1, \ldots$ generated by the algorithm are defined.

- $E[\cdot]$: Expectation taken with respect to $P(\cdot)$.

- $\mathscr{F}_k$: The σ-field generated by the random samples up to the kth iteration, i.e., $\mathscr{F}_k = \sigma(\{\Lambda_0, \ldots, \Lambda_k\})$. Note that $\mathscr{F}_k$ is increasing with k, and it determines all the information up to the kth iteration.

- $E_{\widehat{\theta}_k}[\cdot|\mathscr{F}_{k-1}]$: Conditional Expectation taken with respect to distribution $f_{\widehat{\theta}_k}$, where parameter $\widehat{\theta}_k$ is completely determined by σ-field $\mathscr{F}_{k-1}$.

- $P_{\widehat{\theta}_k}(\cdot|\mathscr{F}_{k-1})$: Conditional Probability taken with respect to distribution $f_{\widehat{\theta}_k}$.

- $\Lambda_k = \{X_1, \ldots, X_{N_k}\}$: The population of N_k random samples generated by distribution $f_{\widehat{\theta}_k}$ in iteration k. Note that $X_1, \ldots, X_{N_k}$ are i.i.d. and conditionally independent of the past information.

After we establish these definitions, everything should be well-defined without ambiguity. In addition, we also need to state the big O notation in the algorithm complexity:

Definition 2. *Let $f(x)$ and $g(x)$ be two functions defined on some subset of the real numbers. One writes $f(x) = O(g(x))$ if there is a positive constant M and a real number x_0 such that $|f(x)| \leq M|g(x)|$ for all $x > x_0$, or equivalently speaking, $\limsup_{k\to\infty} |\frac{f(k)}{g(k)}| < \infty$; one writes $f(x) = \Omega(g(x))$ if $g(x) = O(f(x))$; in addition, one writes $f(x) = \Theta(g(x))$ if $f(x) = O(g(x))$ and $g(x) = O(f(x))$.*

Finally, we define $\mathscr{I}_{\{A\}}$ as the indicator function of the set A.

Similar to the connection between the idealized version CE and gradient search, we expect to connect the Monte-Carlo version CE to a stochastic gradient search. To do so, we make the following assumption on the $\widehat{\theta}_{k+1}$ computed in Algorithm 5.2. Note that this assumption is not fundamentally necessary. However, it avoids the complicated projection step when the parameter $\widehat{\theta}_{k+1}$ hits or goes out of the boundary. See Kushner and Clark [43] for detailed discussion.

Assumption A1. The parameter $\widehat{\theta}_{k+1}$ computed at step 3 of Algorithm 4.2 always lies on the interior of $\mathbf{\Theta}$, i.e., $\widehat{\theta}_{k+1} \in int(\mathbf{\Theta})$ for all k.

Under the above assumption, we define the following terms:

$$\begin{aligned} L(\eta_k) &:= \frac{E_{\widehat{\theta}_k}[\varphi(H(X))I(H(X),\gamma_k)\Gamma(X)|\mathscr{F}_{k-1}]}{E_{\widehat{\theta}_k}[\varphi(H(X))I(H(X),\gamma_k)|\mathscr{F}_{k-1}]} - \eta_k \\ &= \nabla_\theta \ln E_\theta[\varphi(H(X))I(H(X),\gamma_k)]\big|_{\theta=\widehat{\theta}_k}, \\ b_k(\widehat{\theta}_k) &:= \frac{\frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi(H(x))I(H(x),\widehat{\gamma}_k)\Gamma(x)}{\frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi(H(x))I(H(x),\widehat{\gamma}_k)} \\ &\quad - \frac{E_{\widehat{\theta}_k}[\varphi(H(X))I(H(X),\gamma_k)\Gamma(X)|\mathscr{F}_{k-1}]}{E_{\widehat{\theta}_k}[\varphi(H(X))I(H(X),\gamma_k)|\mathscr{F}_{k-1}]}, \\ \xi_k(\widehat{\theta}_k) &:= \frac{\lambda_k(1-\alpha_k)}{\alpha_k}\Big(\frac{1}{N_k}\sum_{x\in\Lambda_k}\Gamma(x) - \eta_k\Big), \end{aligned}$$

where γ_k represents the true $(1-\rho)$-quantile of H under $f_{\widehat{\theta}_k}$. Using these terms, we could rewrite (4.4) as a generalized Robbins-Monro recursion algorithm as follows,

$$\eta_{k+1} = \eta_k + \alpha_k \big[L(\eta_k) + b_k(\widehat{\theta}_k) + \xi_k(\widehat{\theta}_k) \big], \tag{4.5}$$

where L stands for a gradient, b_k stands for a combined error term due to the bias and noise term from the gradient estimation, and ξ_k is an extra error term which comes from the estimation of the mean vector by the sample average.

Note that in ξ_k, $\lambda_k \in [0,1]$ is the parameter that controls the injected noise in the SA recursion (4.5) . When $\lambda_k = 0$, then $\xi_k = 0$, which brings out the basic SA recursion, while a positive λ_k has the effect of injecting noise into (4.5). The idea of using an injected noise is to allow the algorithm to escape from local optima, and this is a common way for the gradient-based algorithm to achieve global optima (e.g., [52, 74]).

4.2 Convergence of the CE method

Throughout this section, we use the standard ordinary differential equation (ODE) approach to study the convergence of Algorithm 4.2 (modified version of the CE method). This ODE approach was first proposed by Kushner and Clark [43] and Ljung [50] to study the stochastic approximation algorithm, and has been developed as a powerful tool for studying other related recursive algorithms, see [5, 6, 8, 43, 44, 66, 67]. In the proof of the convergence of

Algorithm 4.2, the basic idea is to show that the sequence of the mean vector $\{\eta_k\}$ generated by (4.5) asymptotically approaches the solution of the ODE:

$$\frac{d\eta(t)}{dt} = L(\eta),\ t \geq 0, \tag{4.6}$$

where $L(\eta) := \nabla_\theta \ln E_\theta[\varphi(H(X))I\big(H(X), \gamma(m^{-1}(\eta))\big)]\big|_{\theta=m^{-1}(\eta)}$, and $\gamma(m^{-1}(\eta))$ is the true (1-ρ)-quantile of H with respect to $f_{m^{-1}(\eta)}$. Pay attention that here in $L(\eta)$ and Ł(η_k), the differentials are calculated by freezing $\gamma(m^{-1}(\eta))$ and γ_k. Furthermore, we need to assume the continuity of $L(\eta)$ on $\Re^d$. Note that since $m^{-1}(\cdot)$ is continuous, the continuity of $L(\cdot)$ could be verified easily. Recall that we have modified the discontinuous indicator function I into a continuous threshold function $I(\cdot, \cdot)$, which is used to ensure the continuity of L. In addition, we assume that the ODE (4.6) has a unique integral curve for any initial condition.

Following this idea, if the solution curve of ODE (4.6) has a unique asymptotic stable point, then we could expect once the sequence $\{\eta_k\}$ asymptotically approaches the solution curve, it will converge to that asymptotic stable point. However, since the $L(\eta)$ involves the quantile value that essentially depends on the objective function $H(\cdot)$, it becomes difficult to analyze the structure of ODE (4.6) in an explicit way. Moreover, (4.6) may have other limiting behavior instead of have a unique equilibrium point. Therefore, before we present the convergence theorem for Algorithm 4.2, the following definitions and assumptions are needed.

Definition 3. *Let $\eta(t)$ be a flow on a metric space $(\mathbb{X}, \mathrm{d})$ and $\eta_x(\cdot)$ denote the trajectory with an initial point $x \in \mathbb{X}$, i.e., $\eta_x(0) = x$. A point $x \in \mathbb{X}$ is said to be chain recurrent if for any $\varepsilon > 0$ and $T > 0$, there exists a sequence of points $y_0, \ldots, y_k \in \mathbb{X}$ and a sequence of time $t_0, \ldots, t_{k-1} > T$, such that $y_k = x$, $\mathrm{d}(x, y_0) < \varepsilon$ and $\mathrm{d}(y_{i+1}, \eta_{y_i}(t_i)) < \varepsilon$ $\forall i = 0, ..., k-1$; A set $\mathscr{S}$ is said to be an invariant set if $\forall z \in \mathscr{S}$, the trajectory with the initial point z will stay in $\mathscr{S}$, i.e., $\eta_z(t) \subset \mathscr{S}$ $\forall t \in \Re$; Moreover, a compact invariant set $\mathscr{A}$ is said to be internally chain recurrent if every point in $\mathscr{A}$ is chain recurrent.*

Assumptions:

A2. The gain $\{\alpha_k\}$ satisfies $\alpha_k > 0 \, \forall \, k$, $\lim_{k \to \infty} \alpha_k = 0$, and $\sum_{k=0}^{\infty} \alpha_k = \infty$. On the other hand, $\lambda_k = O(k^{-\lambda})$ for some constant $\lambda \geq 0$ and $N_k = \Theta(k^{\beta})$, where $\beta > \max\{0, 1 - 2\lambda\}$.

A3. For a given $\rho \in (0, 1)$ and a distribution family $\{f_\theta(\cdot), \theta \in \mathbf{\Theta}\}$, the $(1-\rho)$-quantile of $\{H(X), X \sim f_\theta(x)\}$ is unique for each $\theta \in \mathbf{\Theta}$.

In A2, the condition on the gain sequence $\{\alpha_k\}$ is standard in the analysis of stochastic approximation. Note that the sample size parameter N_k is generally increased to infinity at a polynomial rate. The intuition here is that we need the sample size to go to infinity so that the bias and noise term $\{\beta_k\}$ will asymptotically vanish. On the other hand, A3 is necessary for showing that the estimated sample quantile sequence $\{\widehat{\gamma}_k\}$ in Algorithm 4.2 will converge to the true quantile. This assumption is satisfied for many objective functions and for many distributions in the parameterized family.

We now present the convergence theorem for Algorithm 4.2 as follows, and the same result can be obtained under other similar conditions, see [8] and [44].

Theorem 4.2.1. *Assume that $L(\eta)$ is continuous with a unique integral curve and A1-A3 hold. Then the sequence $\{\eta_k\}$ generated by (4.4) converges to a compact connected internally chain recurrent set of (4.6) w.p.1. Furthermore, if the internally chain recurrent sets of (4.6) are isolated equilibrium points, then w.p.1 $\{\eta_k\}$ converges to a unique equilibrium point.*

This theorem follows from Theorem 1.2 in [5], and the idea of the proof is to verify that conditions A1-A3 in [5] hold. A potential difficulty here is the sample $(1-\rho)$-quantile $\widehat{\gamma}_k$ that appears in $b_k(\widehat{\theta}_k)$ in (4.5). Fortunately, the following proposition and lemma ensure that, under some conditions the bias term $b_k(\widehat{\theta}_k)$ will ultimately vanish as $k \longrightarrow \infty$.

Lemma 4.2.1. *Let γ_k be the true $(1-\rho)$-quantile of $H(X)$ with respect to $f_{\widehat{\theta}_k}$ and $\widehat{\gamma}_k$ be the corresponding sample $(1-\rho)$-quantile. If A2 and A3 are satisfied, then $\widehat{\gamma}_k \to \gamma_k$ as $k \to \infty$ w.p.1.*

Proof. Similar to the proof of Lemma 7 in [32]; see also the proof of Lemma 4.3.1. □

Proposition 4.2.1. *If assumptions A2 and A3 are satisfied, then*

$$b_k(\widehat{\theta}_k) \to 0 \quad \textit{as } k \to \infty \quad \textit{w.p.1.}$$

Proof. We prove Proposition 4.2.1 in Appendix. □

Proof of Theorem 4.2.1 : To proof the theorem we could simply verify that conditions A1-A3 in [5] hold. As $\mathbb{X}$ is compact and Γ is continuous, the sequence $\{\eta_k\}$ is bounded, which ensures A1 in [5] hold. A2 in [5] is a direct consequence of Proposition 4.2.1. To establish A3 in [5], we let $M_n = \sum_{k=0}^{n} \alpha_k \xi_k(\widehat{\theta}_k)$ so that $\{M_n\}$ is a martingale. Furthermore, we have

$$
\begin{aligned}
E[\|M_n\|^2] &= E\Big[\Big\|\sum_{k=0}^{n} \alpha_k \xi_k(\widehat{\theta}_k)\Big\|^2\Big] = \sum_{k=0}^{n} \alpha_k^2 E\big[\|\xi_k(\widehat{\theta}_k)\|^2\big] \\
&= \sum_{k=0}^{n} \alpha_k^2 E\Big[E_{\widehat{\theta}_k}[\xi_k(\widehat{\theta}_k)^T \xi_k(\widehat{\theta}_k)|\mathscr{F}_{k-1}]\Big] \\
&= \sum_{k=0}^{n} (1-\alpha_k)^2 \lambda_k^2 N_k^{-1} E\Big[E_{\widehat{\theta}_k}[\Gamma(X)^T\Gamma(X)|\mathscr{F}_{k-1}] - m(\widehat{\theta}_k)^T m(\widehat{\theta}_k)\Big] \\
&= O\Big(\sum_{k=1}^{n} \frac{1}{k^{\beta+2\lambda}}\Big) < \infty
\end{aligned}
$$

where the last equation holds since $\beta + 2\lambda > 1$ by A2. Therefore, the L^2-bounded martingale convergence theorem (e.g., [64]) implies that $\{M_n\}$ converges w.p.1 to a finite random vector M_∞ which shows that condition A3 in [5] holds, and this completes the proof. □

Several remarks are needed to explain the theorem. First, the mean vector function $\eta = m(\theta)$ is invertible, therefore we could also use η as the parameter for the parameterized distribution family. Moreover, the optimal parameter η^* is expected to be the one whose associated parameterized distribution concentrates its mass on the set of optimal solutions to (3.1) (cf.

(4.5)). However, compared with case a) and c) in the examples in Chapter 3, the model updating procedure for Algorithm 4.2 and the CE method may not necessarily lead to such an optimal reference distribution g^* that only concentrates on the global optimum. This implies that Algorithm 4.2 and the CE method may not converge to the global optimum. On the other hand, Theorem 4.2.1 shows that the sequence of the iterates $\{\eta_k\}$ will asymptotically approach the limiting solution of ODE (4.6). In other words, after certain regular conditions are satisfied, the asymptotic behavior of Algorithm 4.2 merely depends on its underlying ODE. To our best knowledge, the conclusion given by Theorem 4.2.1 is the strongest result we could obtain. In particular, there exist counterexamples indicating that the CE method and its variants do not converge to the global optimum, see [32]. To better explain the statement, we provide the following example to show that, the local/global convergence properties of the CE method can be determined by the solution of the underlying ODE.

Example 3.1: Consider maximizing the function

$$H(x) = \begin{cases} 0 & x \in \{(0,1),(1,0)\} \\ 1 & x = (0,0) \\ 2 & x = (1,1) \end{cases} \tag{4.7}$$

by sampling from the parameterized p.m.f.

$$f_\theta(x) = \Big(\frac{\delta + e^{\vartheta_1}}{1+e^{\vartheta_1}}\Big)^{x_1}\Big(\frac{1-\delta}{1+e^{\vartheta_1}}\Big)^{1-x_1}\Big(\frac{\delta + e^{\vartheta_2}}{1+e^{\vartheta_2}}\Big)^{x_2}\Big(\frac{1-\delta}{1+e^{\vartheta_2}}\Big)^{1-x_2},$$

where $x := (x_1, x_2)^T \in \mathbb{X} := \{(0,0),(0,1),(1,0),(1,1)\}$, $\theta := (\vartheta_1, \vartheta_2)^T \in \Re^2$, and $\delta \in (0,1)$. By definition of NEF, it can be verified that Γ is given by

$$\Gamma(x) = (x_1, x_2)^T \qquad \text{and} \qquad \eta := (\eta_1, \eta_2)^T := \Big(\frac{\delta + e^{\vartheta_1}}{1 + e^{\vartheta_1}}, \frac{\delta + e^{\vartheta_2}}{1 + e^{\vartheta_2}}\Big)^T.$$

Take $\rho \in (0, \delta^2)$, $\varphi(x) \equiv 1$ and $\varepsilon \in (0,1)$, then since $P_\theta\big(X = (1,1)\big) \geq \delta^2 > \rho$ for all θ, we have $\gamma(m^{-1}(\eta)) = 2$. Therefore

$$E_\theta[I(H(X), \gamma(m^{-1}(\eta)))] = \eta_1\eta_2$$

and

$$E_\theta[I(H(X), \gamma(m^{-1}(\eta)))\Gamma(X)] = (\eta_1\eta_2, \eta_1\eta_2)^T.$$

Consequently,

$$L(\eta) = (1 - \eta_1, 1 - \eta_2)^T.$$

To find the equilibrium point, we simply set $L(\eta) = (0,0)^T$. Then it is obvious that $\eta^* = (1,1)^T$ is an equilibrium point of the ODE $d\eta(t)/dt = L(\eta)$. We then use the Lyapunov function approach to prove the uniqueness of the equilibrium point.

We construct the Lyapunov function as $V(\eta) = \frac{1}{2}[(\eta_1 - 1)^2 + (\eta_2 - 1)^2]$. It is easy to see that the derivative of V is $\dot{V}(\eta(t)) = (\eta_1 - 1)\dot{\eta_1} + (\eta_2 - 1)\dot{\eta_2} = -(1 - \eta_1)^2 - (1 - \eta_2)^2$. Apparently, $\dot{V}(\eta(t))$ is negative definite for all $\eta \neq \eta^*$. Therefore, $\eta^* = (1,1)^T$ is the unique globally asymptotically stable point, in which case Theorem 4.2.1 implies that the sequence of sampling distributions

$\{f_{\widehat{\theta}_k}\}$ obtained in Algorithm 4.2 will converge to a degenerate distribution that assigns unit mass to the optimal solution $x = (1, 1)$.

On the other hand, if we use the p.m.f.

$$f_\theta(x) = \Big(\frac{\delta + e^{\vartheta_1}}{1 + e^{\vartheta_1}}\Big)^{1-x_1}\Big(\frac{1-\delta}{1+e^{\vartheta_1}}\Big)^{x_1}\Big(\frac{\delta + e^{\vartheta_2}}{1 + e^{\vartheta_2}}\Big)^{1-x_2}\Big(\frac{1-\delta}{1+e^{\vartheta_2}}\Big)^{x_2}$$

for some constant $\delta \in (\frac{1}{2}, 1)$, it is straightforward to see that

$$\Gamma(x) = (x_1, x_2)^T \qquad \text{and} \qquad \eta = \Big(\frac{1-\delta}{1+e^{\vartheta_1}}, \frac{1-\delta}{1+e^{\vartheta_2}}\Big)^T.$$

If we take $\rho \in \big((1-\delta)^2, \delta^2\big)$, then since $P_\theta\big(H(X) \geq 1\big) \geq P_\theta\big(X = (0,0)\big) \geq \delta^2 > \rho$ and $P_\theta\big(H(X) \geq 2\big) = P_\theta\big(X = (1,1)\big) \leq (1-\delta)^2 < \rho$, by definition of quantiles we get $\gamma(m^{-1}(\eta)) = 1$. Therefore, we have

$$L(\eta) = \big(\frac{\eta_1\eta_2}{\eta_1\eta_2 + (1-\eta_1)(1-\eta_2)} - \eta_1, \frac{\eta_1\eta_2}{\eta_1\eta_2 + (1-\eta_1)(1-\eta_2)} - \eta_2\big)^T.$$

By setting $L(\eta) = (0,0)^T$ and solving the equation, it can be seen that the isolated equilibrium points $(0,0)^T$, $(\frac{1}{2}, \frac{1}{2})^T$ and $(1,1)^T$ are the only chain recurrent points to the ODE $d\eta(t)/dt = L(\eta)$. Therefore, Theorem 4.2.1 implies that the sequence $\{\eta_k\}$ generated by Algorithm 4.2 will converge to one of them.

In a third case, if we use the p.m.f.

$$f_\theta(x) = \Big(\frac{\delta + e^{\vartheta_1}}{1 + e^{\vartheta_1}}\Big)^{x_1}\Big(\frac{1-\delta}{1+e^{\vartheta_1}}\Big)^{1-x_1}\Big(\frac{\delta + e^{\vartheta_2}}{1 + e^{\vartheta_2}}\Big)^{1-x_2}\Big(\frac{1-\delta}{1+e^{\vartheta_2}}\Big)^{x_2}$$

for some constant $\delta \in (0,1)$, we have

$$\Gamma(x) = (x_1, x_2)^T \qquad \text{and} \qquad \eta = \Big(\frac{\delta + e^{\vartheta_1}}{1+e^{\vartheta_1}}, \frac{1-\delta}{1+e^{\vartheta_2}}\Big)^T.$$

If we take $\rho \in (1-\delta^2, 1)$, then since $P_\theta(H(X)=0) = P_\theta(X \in \{(0,1),(1,0)\}) \geq \delta^2$, it follows that $P_\theta(H(X) \geq 1) = 1 - P_\theta(H(X)=0) \leq 1-\delta^2 < \rho$ and $P_\theta(H(X) \geq 0) = 1 > \rho$. Thus, we have $\gamma(m^{-1}(\eta)) = 0$ and $L(\eta) \equiv (0,0)^T$. It follows that the sequence $\{\eta_k\}$ will converge to the set of chain recurrent points of (4.6), which is the set of all $\eta = (\eta_1, \eta_2)^T$ satisfying $\eta_1 \in [\delta, 1]$ and $\eta_2 \in [0, 1-\delta]$.

There are other ways to address the convergence analysis with the ODE approach. In [43], it is assumed that the underlying ODE has a locally asymptotically stable point η^*. Moreover, it is assumed that there exists a compact set in the domain of attraction of η^*, where the sequence $\{\eta_k\}$ will enter this domain infinitely often. For the ODE with a unique globally asymptotically stable point, Algorithm 4.2 will converge to the global optimum. However, for an arbitrary and complicated ODE, it is very difficult to verify this assumption, and this becomes a major constraint for the ODE analysis approach.

4.3 ASYMPTOTIC NORMALITY

Throughout this section, we assume the random sequence $\{\eta_k\}$ generated by Algorithm 4.2 converges to a unique limit point and then analyze the asymptotic convergence rate of Algorithm 4.2. In particular we consider the special case that the underlying ODE has a unique globally asymptotically equilibrium η^*, and

$$\eta_k \longrightarrow \eta^* \quad w.p.1 \qquad \text{as} \quad k \longrightarrow \infty.$$

For the convenience of analysis, we assume that $m^{-1}(\eta^*) \in int(\mathbf{\Theta})$. Since $m(\cdot)$ is continuously differentiable on $int(\mathbf{\Theta})$ and $m^{-1}(\cdot)$ is invertible, by inverse function theorem we know that $m(\cdot)$ is also continuously differentiable in some open neighborhood of η^*. This statement implies that, as long as the mean vector sequence $\{\eta_k\}$ converges to η^*, the sequence of the parameterized sampling distributions $\{f_{\widehat{\theta}_k}\}$ in Algorithm 4.2 will converge point-wise to a limiting distribution $f_{m^{-1}(\eta^*)}$ w.p.1.

Note that $L(\eta)$ can be viewed as the gradient of some function $F(\eta)$, therefore recursion (4.5) could be interpreted as a gradient search algorithm for solving the maximization problem $\max_\eta F(\eta)$. Once we have the convergence result, which states that $\{\eta_k\}$ converges to η^* with probability one, it is natural to expect that η^* is a local or global optimum of the objective function $F(\eta)$ in its neighborhood (see [1] for a detailed discussion). We denote the Jacobian matrix of $L(\cdot)$ in (4.6) by $J_L(\eta)$, which could essentially be viewed as the Hessian matrix of the objective function $F(\eta)$. Moreover, it is reasonable to make the following assumption on J_L.

Assumption B1. The Jacobian matrix $J_L(\eta)$ is continuous, symmetric and negative definite in a small neighborhood of η^*.

In the convergence rate analysis we choose a standard gain sequence as $\alpha_k = c/k^{\alpha}$ for constants $c > 0$, $\alpha \in (0,1)$, and let $\lambda_k = \Theta(k^{-\lambda})$ for $\lambda \geq 0$. It is very easy to verify that both α_k and λ_k satisfy A2. Once we define the difference $\delta_k := \eta_k - \eta^*$, the recursion (4.5) can be rewritten into the form:

$$\delta_{k+1} = \delta_k + \alpha_k L(\eta_k) + \alpha_k b_k(\widehat{\theta}_k) + \alpha_k \xi_k(\widehat{\theta}_k).$$

Using a Taylor expansion of $L(\eta_k)$ in a small neighborhood of η^*, we could get the following equation:

$$\delta_{k+1} = \delta_k + ck^{-\alpha} J_L(\bar{\eta}_k)\delta_k + ck^{-\alpha} b_k(\widehat{\theta}_k) + ck^{-\alpha}\xi_k(\widehat{\theta}_k),$$

where $\bar{\eta}_k$ lies on the line segment between η_k and η^*. Note that since η^* is a local optimum, we have the fact that $L(\eta^*) = 0$. Thus the Taylor expansion above only contains the second order of derivative. This equation, which can be viewed as an evolution of δ_k, also involves the bias/noise term b_k and injected noise term ξ_k. Therefore, similar to the analysis for the convergence of Algorithm 4.2, we need to carefully examine the parameters to ensure that $\{\delta_k\}$ converges to zero under certain rate.

In this thesis, we follow Fabian's approach [17] to analyze the asymptotic convergence rate of the SA recursion. The convergence rate analysis in [17] provides generally sufficient conditions to establish asymptotic normality results for SA. To get the convergence rate for Algorithm 4.2, We rewrite the above equation as the same form in [17]:

$$\delta_{k+1} = \delta_k - k^{-\alpha}\Upsilon_k\delta_k + k^{-\frac{\alpha+\tau}{2}}\Phi_k V_k + k^{-\alpha-\frac{\tau}{2}}T_k,$$

where $\tau > 0$ is a constant, $\Upsilon_k = -cJ_L(\bar{\eta}_k)$, $\Phi_k = c\mathbb{I}_{d\times d}$, $V_k = k^{-\frac{\alpha}{2}+\frac{\tau}{2}}\xi_k(\widehat{\theta}_k)$, $T_k = ck^{\frac{\tau}{2}}b_k(\widehat{\theta}_k)$, and $\mathbb{I}_{d\times d}$ denotes a d-by-d identity matrix. In addition, for a technical reason, we need the following regularity condition on the distribution function of the objective function. Let $f^H_{\widehat{\theta}_k}$ be the probability density/mass function of $H(X)$ when X is distributed with respect to $f_{\widehat{\theta}_k}$.

Assumption B2.

- *Continuous optimization*: For a given $\rho \in (0,1)$, there exist constants $\bar{\zeta} > 0$ and $\bar{\delta} > 0$ such that $f^H_{\widehat{\theta}_k}(\gamma) > \bar{\zeta}$, $\forall\, \gamma \in (\gamma_k - \bar{\delta}, \gamma_k + \bar{\delta})$ almost surely for k sufficiently large.
- *Discrete finite optimization*: For a given $\rho \in (0,1)$, there exists a constant $\bar{\zeta} > 0$ such that $P_{\widehat{\theta}_k}(H(X) \geq \gamma_k|\mathscr{F}_{k-1}) \geq \rho + \bar{\zeta}$ and $P_{\widehat{\theta}_k}(H(X) > \gamma_k\mathscr{F}_{k-1}) \leq \rho - \bar{\zeta}$ almost surely for k sufficiently large.

Generally speaking, B2 is not easy to verify a priori since the distribution of the random variable $H(X)$ depends on the objective function and the sampling distribution selected from the parameterized families. Fortunately, since in this section it is already assumed that the SA recursion (4.5) converges to a unique global optimum and the sequence of the sampling distributions also converges (point-wisely) to a limiting distribution, B2 can be considered reasonable. For example, in the continuous case, if $f^H_{m^{-1}(\eta^*)}$ is the limiting distribution of $\{f^H_{\widehat{\theta}_k}\}$ and $f^H_{m^{-1}(\eta^*)}(\gamma_*) > 0$ where γ_* is the true $(1-\rho)$-quantile of H under $f^H_{m^{-1}(\eta^*)}$, Assumption B2 holds; in the discrete finite case, if the sequence of the sampling distributions converges to a unit mass function, Assumption B2 also holds.

As the convergence rate theorem follows from the asymptotic normality result in [17], the following proposition and lemmas show that conditions (2.2.1), (2.2.2), and (2.2.3) in [17] hold in the case of Algorithm 4.2.

Lemma 4.3.1. *Let* $N_k = \Theta(k^\beta)$*. For any* $\tau > 0$*, if A3 and B2 hold, and* $\beta > 2\tau$*, then* $\lim_{k\to\infty} k^{\frac{\tau}{2}} |\widehat{\gamma}_k - \gamma_k| = 0$ *w.p.1.*

Proof. See Appendix for a proof. □

Proposition 4.3.1. *For any constant* $\tau > 0$*, let* $\beta > 2\tau$*. If A3 and B2 hold, then* $T_k \to 0$ *as* $k \to \infty$ *w.p.1.*

Proof. The proof is given in Appendix. □

Lemma 4.3.1 is a strengthened version of Proposition 4.2.1. It shows that, by carefully controlling the parameter of the sample size at each iteration of Algorithm 4.2, the sample quantile $\widehat{\gamma}_k$ would converge to the true quantile γ_k

at a desirable polynomial rate. Moreover, Proposition 4.3.1 is a strengthened version of Proposition 4.2.1, which claims that the amplified bias term T_k vanishes to zero asymptotically. Moreover, the amplified noise V_k has the following properties.

Lemma 4.3.2. *Let $\alpha_k = c/k^{\alpha}$, $\lambda_k = \Theta(k^{-\lambda})$, and $N_k = \Theta(k^{\beta})$ for constants $\alpha \in (0,1)$, $\lambda \geq 0$, and $\beta > \max\{0, 1-2\lambda\}$. If A1 and A3 hold, and in addition, $\beta \geq \alpha + \tau - 2\lambda$ for $\tau > 0$, then $E_{\widehat{\theta}_k}[V_k|\mathscr{F}_{k-1}] = 0$ and there exists a matrix Σ such that $E_{\widehat{\theta}_k}[V_k V_k^T|\mathscr{F}_{k-1}] \to \Sigma$ as $k \to \infty$ w.p.1. Moreover, the sequence $\{V_k\}$ is uniformly square integrable in the sense that $\lim_{k\to\infty} E\big[\mathscr{I}_{\{\|V_k\|^2 \geq rk^{\alpha}\}}\|V_k\|^2\big] = 0 \;\forall\, r > 0$.*

Proof. See Appendix. □

By directly applying Theorem 2.2 in [17] with lemma 4.3.2 and Proposition 4.3.1 above, we could get the asymptotic normality of Algorithm 4.2.

Theorem 4.3.1. *Let $\alpha_k = c/k^{\alpha}$, $\alpha \in (\frac{1}{2}, 1)$ and $\lambda_k = \Theta(k^{-\lambda})$, $\lambda \in [0, \alpha - \frac{1}{2})$. If A1, A3, B1, and B2 hold, $\tau \in (1-\alpha, \alpha - 2\lambda)$, and $\beta \geq \alpha + \tau - 2\lambda$, then*

$$k^{\frac{\tau}{2}}(\eta_k - \eta^*) \xrightarrow{dist} \mathcal{N}(0, QMQ^T),$$

where Q is an orthogonal matrix such that $Q^T(-J_L(\eta^*))Q = \Lambda$ with Λ being a diagonal matrix, and the (i,j)th entry of M is given by $M_{(i,j)} =$

$(Q^T \Sigma Q)_{(i,j)}(\Lambda_{(i,i)} + \Lambda_{(j,j)})^{-1}$,

$$\Sigma := \begin{cases} Cov_{m^{-1}(\eta^*)}(\Gamma(X)) & \textit{if } \beta = \alpha + \tau - 2\lambda, \\ 0 & \textit{if } \beta > \alpha + \tau - 2\lambda. \end{cases}$$

Theorem 4.3.1 indicates that the asymptotic rate for Algorithm 4.2 is bounded below by $O(1/\sqrt{k^\tau})$. This is different from $O(1/\sqrt{k})$ which is the optimal convergence rate for general stochastic approximation algorithms. Note that by choosing τ close to $\alpha - 2\lambda$, α close to 1 and λ close to 0, we could approximately achieve the optimal bound. In particular, when $\beta > \alpha + \tau - 2\lambda$, Theorem 4.3.1 implies that $k^{\tau/2}(\eta_k - \eta^*) \to 0$ in probability. However, the asymptotic rate only describes the limiting behavior of the convergence, and does not directly determine the convergence speed of the algorithm in practice. Meanwhile, the asymptotic rate is carried out in terms of the iteration number rather than the sample size. As a result, we need to carefully choose those parameters when implementing Algorithm 4.2 in order to achieve a good practical performance.

4.4 Numerical Examples

In this section, we provide some numerical examples for Algorithm 4.2 and compare it with the standard CE method in [60]. The examples would primarily focus on continuous optimization problems. We choose a set of twelve benchmark problems from [32], [46], [60].

(1) Shekel function ($n = 4$, $0 \le x_i \le 10$, $i = 1, \dots, n$):

$$H_1(x) = \sum_{j=1}^{5} \Big(\sum_{i=1}^{4} (x_i - A_{i,j})^2 + B_j \Big)^{-1} - 10.1532,$$

with $B = (0.1, 0.2, 0.2, 0.4, 0.4)^T$, $A_1 = A_3 = (4, 1, 8, 6, 3)$, and $A_2 = A_4 = (4, 1, 8, 6, 7)$, where A_i represents the ith row of A. The function has a global maxima $x^* = (4, 4, 4, 4)^T$ and $H_1(x^*) = 0$.

(2) Rosenbrock function ($n = 10$, $-10 \le x_i \le 10$):

$$H_2(x) = -1 - \sum_{i=1}^{n/2} \big[100(x_{2i} - x_{2i-1}^2)^2 + (1 - x_{2i-1})^2 \big],$$

where $H_2(x^*) = -1$.

(3) Zakharov function ($n = 20$, $-10 \le x_i \le 10$):

$$H_3(x) = -1 - \sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} 0.5 i x_i \Big)^2 - \Big(\sum_{i=1}^{n} 0.5 i x_i \Big)^4,$$

where $H_3(x^*) = -1$.

(4) Rastrigin function ($n = 30$, $-5.12 \le x_i \le 5.12$):

$$H_4(x) = -\sum_{i=1}^{n} \big(x_i^2 - 10 \cos(2\pi x_i) \big) - 10n,$$

where $H_4(x^*) = 0$.

(5) Ackley function ($n = 40$, $-32 \le x_i \le 32$):

$$H_5(x) = -20 - e + 20e^{-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}} + e^{\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)},$$

where $H_5(x^*) = 0$.

(6) levy function ($n = 50$, $-50 \le x_i \le 50$):

$$H_6(x) = -1 - \sin^2(\pi y_1) - (y_n - 1)^2(1 + \sin^2(2\pi y_n))$$

$$-\sum_{i=1}^{n-1}\big[(y_i - 1)^2(1 + 10\sin^2(\pi y_i + 1))\big],$$

where $y_i = 1 + (x_i - 1)/4,\ i = 1, \ldots, n$ and $H_6(x^*) = -1$.

(7) Trigonometric function ($n = 50$, $-50 \le x_i \le 50$):

$$H_7(x) = -1 - \sum_{i=1}^{n}\big[8\sin^2\big(7(x_i - 0.9)^2\big) + 6\sin^2\big(14(x_i - 0.9)^2\big) + (x_i - 0.9)^2\big],$$

where $H_7(x^*) = -1$.

(8) Griewank function ($n = 50$, $-50 \le x_i \le 50$):

$$H_8(x) = \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right),$$

where $H_8(x^*) = -1$.

(9) Brown function ($n = 50,\ -50 \leq x_i \leq 50$):

$$H_9(x) = -\frac{1}{25} \sum_{i=1}^{n/2} \big[(x_{2i} - 3)^2 - (x_{2i-1} - x_{2i}) + e^{20(x_{2i-1} - x_{2i})}\big]$$

$$-\frac{1}{25}\big(\sum_{i=1}^{n/2}(x_{2i-1} - 3)\big)^2,$$

where $H_9(x^*) = -1$.

(10) Powell function ($n = 50,\ -50 \leq x_i \leq 50$):

$$H_{10}(x) = -\sum_{i=1}^{(n-2)/2} \big[(x_{2i-1} + 10x_{2i})^2 + 5(x_{2i+1} - x_{2i+2})^2$$

$$+(x_{2i} - 2x_{2i+1})^4 + 10(x_{2i-1} - x_{2i+2})^4\big] - 1,$$

where $H_{10}(x^*) = -1$.

(11) Cragg and Levy function ($n = 50,\ -50 \leq x_i \leq 50,\ i = 1, \ldots, n$):

$$H_{11}(x) = -\sum_{i=1}^{(n-1)/2} \big[(e^{x_{2i-1}} - x_{2i})^2 + 100(x_{2i} - x_{2i+1})^4$$

$$+\tan^2(x_{2i+1} - x_{2i+2}) + x_{2i-1}^8 + (x_{2i+2} - 1)^4\big],$$

where $H_{11}(x^*) \approx -21.51$.

(12) Pintér function ($n = 50$, $-50 \leq x_i \leq 50$):

$$H_{12}(x) = -\sum_{i=1}^{n} ix_i^2 - 1 - \sum_{i=1}^{n} 20i \sin^2\big(x_{i-1}\sin x_i - x_i + \sin x_{i+1}\big)$$

$$-\sum_{i=1}^{n} i \log_{10}\big(1 + i(x_{i-1}^2 - 2x_i + 3x_{i+1} - \cos x_i + 1)^2\big),$$

where $x_0 = x_n$, $x_{n+1} = x_1$, and $H_{12}(x^*) = -1$.

Also, we would consider multivariate normal distributions and independent univariate normal distributions as two parameterized distribution families in Algorithm 4.2 and the standard CE. Note that when we use multivariate normal distributions $N(\widehat{\mu}_k, \widehat{\Sigma}_k)$ with $\lambda_k = 0$ for all k, the the parameter updating step (i.e., Step 3) in Algorithm 4.2 induces an explicit parameter updating procedure as the following:

$$\begin{aligned}
\widehat{\mu}_{k+1} &= \alpha_k \frac{\sum_{\Lambda_k} \varphi(H(x)) I(H(x), \widehat{\gamma}_k) x}{\sum_{\Lambda_k} \varphi(H(x)) I(H(x), \widehat{\gamma}_k)} + (1 - \alpha_k)\widehat{\mu}_k \quad \text{and} \\
\widehat{\Sigma}_{k+1} &= \alpha_k \frac{\sum_{\Lambda_k} \varphi(H(x)) I(H(x), \widehat{\gamma}_k)(x - \widehat{\mu}_{k+1})(x - \widehat{\mu}_{k+1})^T}{\sum_{\Lambda_k} \varphi(H(x)) I(H(x), \widehat{\gamma}_k)} \\
&\quad + (1 - \alpha_k)\big(\widehat{\Sigma}_k + (\widehat{\mu}_k - \widehat{\mu}_{k+1})(\widehat{\mu}_k - \widehat{\mu}_{k+1})^T\big).
\end{aligned}$$

For independent univariate normal distribution, the parameter updating procedure is similar, where each of the variances is updated in a point-wise manner.

On the other hand, the parameter updating procedure in [60] is carried out by

$$\tilde{\theta}_{k+1} := \nu_k \bar{\theta}_{k+1} + (1 - \nu_k)\tilde{\theta}_k, \quad \text{with } \tilde{\theta}_0 = \widehat{\theta}_0, \tag{4.8}$$

where $\bar{\theta}_{k+1}$ is the new parameter calculated at Step 3 of Algorithm 4.2 with $\alpha_k = 1$ in (4.4), and ν_k is a smoothing parameter imposed on the parameters of the parameterized distributions. However, as far as we can see, this smoothing procedure in the standard CE method may not have a strong theoretical support. In contrast, the smoothing parameter α_k in Algorithm 4.2 is used on the entire reference distributions in order to prevent the sampling distributions from varying too fast. By establishing the connection between Algorithm 4.2 to SA, we could see that the smoothing parameters $\{\alpha_k\}$ become the gain sequence in (4.4). Therefore, Algorithm 4.2 provides a theoretical guarantee of its convergence. Consequently, we expect that Algorithm 4.2 would show better performance in the numerical tests.

Parameter Settings

When performing Algorithm 4.2 through the test problems, we use the following parameter setting.

- *Initial parameters of the sampling distributions:* The initial means are uniformly generated from the solution space, and the initial covariance matrix is set to be a $n \times n$ diagonal matrix with diagonal entries equal to 1000.

In our preliminary experiments, we found that the practical performance of Algorithm 4.2 is insensitive to the initial parameters of the sampling distributions. Therefore, we choose a large initial variance (e.g., 1000) so that the search could cover the entire solution space at the beginning of the algorithm.

- *Smoothing parameters/gain sequence:* $\alpha_k = 2/(k+100)^{0.501}$.

  The choice of the gain sequence $\{\alpha_k\}$ reflects the trade-off between exploitation and exploration, and our preliminary results show that the practical performance of Algorithm 4.2 is mainly sensitive to $\{\alpha_k\}$. A fast decay rate for $\{\alpha_k\}$ tends to lead a rapid convergence, in which case the search may only stop at a local optimum. On the other hand, a slow decay rate tends to lead the algorithm to search more region in the solution space. Although this feature may lead the search towards the global optimum, the algorithm would keep oscillating for a long period before it converges. Generally speaking, for those high dimensional multimodal problems, a relatively slow decay rate is preferred since "exploration" is more desirable under this scenario. For our test problems, we use $\alpha_k = 2/(k+100)^{0.501}$ as a conservative choice. Note that this setting of $\{\alpha_k\}$ satisfies Assumption A2, and the constant 100 is merely used to keep the initial step size small to prevent unnecessary unstable behavior in the early stage. Further discussions on choices of the gain sequence can be found in [67].

- *Proportion for elite set:* $\rho = 0.1$.

  All $\rho \in [0.01, 0.2]$ works well in our preliminary experiments.

- *Parameter of injected noise:* $\lambda = 0$.

  In our implementation, we ignore the manner of injecting extra noise.

- *Sample size:* $N_k = max\{400, k^{1.01}\}$.

  Note that this setting also satisfies Assumption A2, where the constant 400 is to ensure that the number of the top-ρ elite samples is still enough for model updating, see [60] for a detailed discussion. Note that our settings of N_k and λ also satisfy Assumption A2.

The performance of Algorithm 4.2 was also compared with that of the standard CE with the smoothed parameter updating procedure (4.8). In standard CE, the smoothing parameter ν_k was set equal to α_k in Algorithm 4.2, while all other settings were taken to be the same as in Algorithm 4.2. Also we have put same computational budget on both algorithms for the performance comparison, i.e., for H_1 and H_8, the total number of function evaluations is set to 10^5; for H_5-H_7 the number is 3×10^5; for H_2-H_4 and H_9-H_{12}, the number is 8×10^5.

For each benchmark function, we performed 100 independent replication runs of all algorithms. Table 4.1 and 4.2 keep the report of all the tests, where for each algorithm, $\bar{H}_i^*$ is the average of the best function value H_i obtain in each run (with standard error given in parentheses), and N_ε indicates the number of replication runs (out of 100) in which it achieves ε-optimal solution (i.e., a solution whose function value is within ε of $H(x^*)$ and ε is set to 0.001 in our tests). Moreover, for test functions H_7-H_{12}, we plotted the averaged estimated optimal function value, as a function of the number of function evaluations used thus far, in Figure 4.1. The results indicate that the performance of Algorithm 4.2 that performs the smoothed reference distribution updating procedure is promising, compared with the standard CE using updating procedure (4.8). In particular, we found that in the multivariate normal case, Algorithm 4.2 achieves the ε-optimal solution in more than 90% of replication runs in H_4, H_{11} and H_{12}, and 100% for other objective functions. On the other hand, the standard CE with updating procedure (4.8) failed to find the ε-optimal solution in all replication runs for most of the problems. For the univariate normal case, although both Algorithm 4.2 and the standard CE failed in many problems (which shows that using multivariate normal distributions as the sampling distributions family is more effective), Algorithm 4.2 still had much better average of optimal estimations in H_{10}, H_{11} and H_{12}.

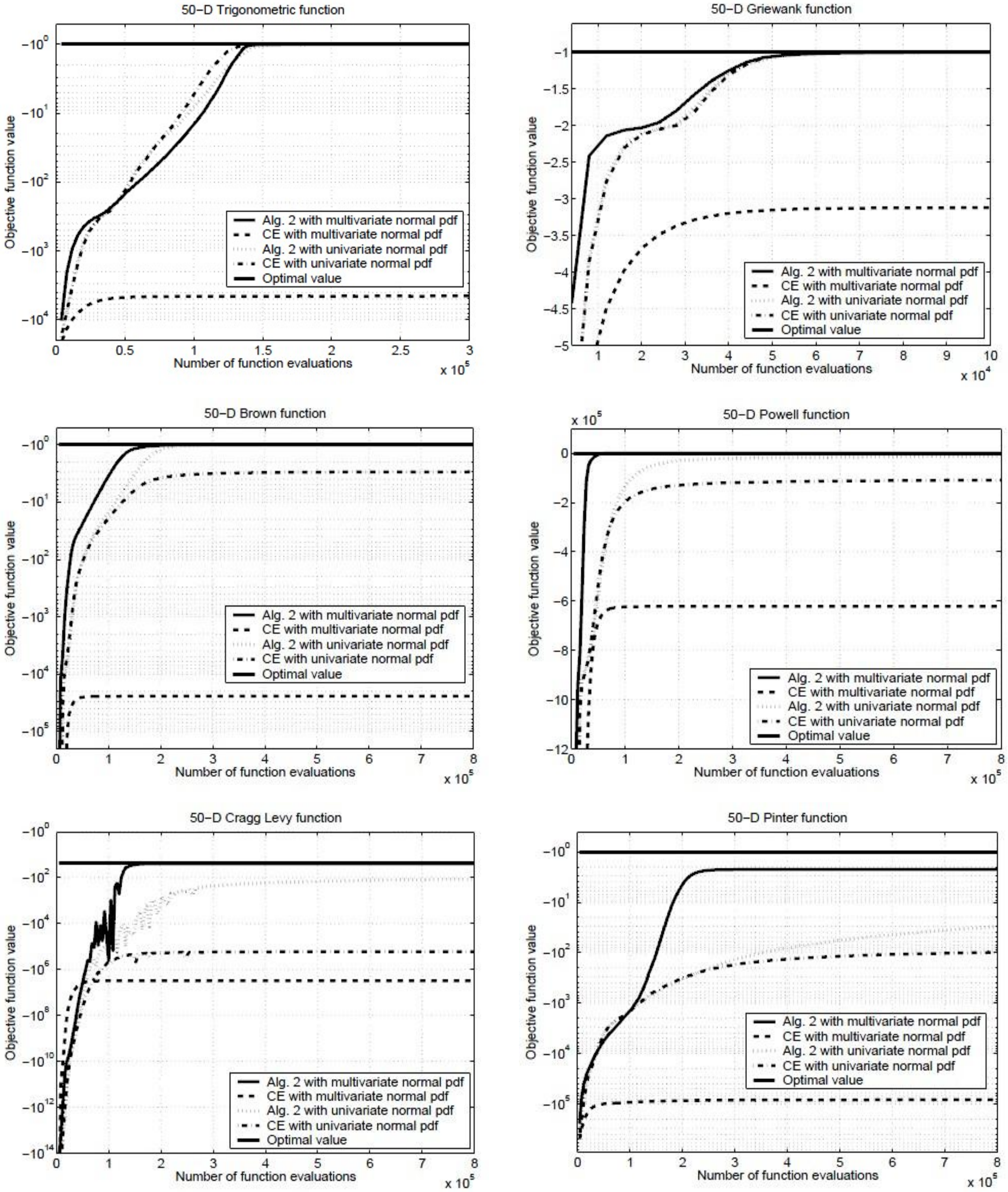
Table 4.1: Performance of Algorithm 4.2 vs. CE with smoothed parameter updating on benchmark problems $H_1 - H_{12}$, based on 100 independent replications and multivariate normal distributions (standard errors in parentheses).

Alg.	Algorithm 4.2 (multivariate normal)		CE (multivariate normal)	
Prob.	$\bar{H}_i^*$	N_ε	$\bar{H}_i^*$	N_ε
H_1	3.2e-7 (2.16e-16)	100	3.2e-7 (1.90e-16)	100
H_2	-1.00 (1.55e-06)	100	-7.92 (2.30e-2)	0
H_3	-1.00 (2.45e-12)	100	-1.00 (0e+00)	100
H_4	-1.09 (2.89e-2)	91	-2.10 (1.04e-1)	10
H_5	-6.35e-6 (1.05e-7)	100	-3.86 (0.27)	0
H_6	-1.00 (1.60e-6)	100	-2.51e+2 (1.02e+1)	0
H_7	-1.00 (1.07e-06)	100	4.55e+3 (1.69e+2)	0
H_8	-1.00 (4.47e-06)	100	-3.12 (4.82e-2)	0
H_9	-1.00 (2.40e-12)	100	-2.43e+4 (4.78e+2)	0
H_{10}	-1.00 (9.19e-17)	100	-6.22e+5 (1.18e+5)	0
H_{11}	-24.60 (0.42)	94	-3.09e+6 (7.72e+8)	0
H_{12}	-2.24 (0.48)	96	-8.23e+4 (3.13e+3)	0

Table 4.2: Performance of Algorithm 4.2 vs. CE with smoothed parameter updating on benchmark problems $H_1 - H_{12}$, based on 100 independent replications and univariate normal distributions (standard errors in parentheses).

Alg.	Algorithm 4.2		CE	
	(univariate normal)		(univariate normal)	
Prob.	$\bar{H}_i^*$	N_ε	$\bar{H}_i^*$	N_ε
H_1	-3.13 (0.37)	59	-3.43 (0.37)	54
H_2	-7.45 (2.18e-3)	0	-8.62 (5.89e-3)	0
H_3	-1.86e+2 (1.05e+2)	0	-2.88e+2 (12.30)	0
H_4	-1.00 (1.69e-6)	100	-1.00 (1.29e-6)	100
H_5	-1.28e-5 (4.07e-7)	100	-1.09e-5 (4.24e-8)	100
H_6	-1.00 (8.74e-5)	100	-1.00 (1.02e-5)	100
H_7	-1.00 (3.53e-6)	100	-1.00 (7.10e-7)	100
H_8	-1.00 (2.08e-6)	100	-1.00 (2.25e-6)	100
H_9	-1.00 (1.45e-8)	100	-2.97 (0.85)	0
H_{10}	-9.16e+3 (7.78e+2)	0	-1.10e+5 (5.60e+3)	0
H_{11}	-101.04 (3.51)	0	-1.65e+5 (3.06e+4)	0
H_{12}	-29.92 (0.22)	0	-96.20 (0.42)	0

Figure 4.1: Average Performance of Algorithm 4.2 vs. CE with smoothed parameter updating procedure on test functions H_7-H_{12}.



Chapter 5

Model-based Annealing Random Search for Global Optimization

5.1 Model-based Annealing Random Search and its connection to SA

The Model-based Annealing Random Search (MARS) for solving problem (3.1) is inspired by the Annealing Adaptive Search (AAS) [57] and the Cross Entropy method [60]. As we mentioned in the introduction, AAS uses Boltzmann distribution as the probability model, i.e., it generates candidate solutions by sampling from a sequence of Boltzmann distributions:

$$g_k(x) = \frac{e^{H(x)/T_k}}{\int_{\mathbb{X}} e^{H(x)/T_k} \nu(dx)}, \tag{5.1}$$

Note that the Boltzmann distribution has a time-dependent parameter T_k that could be interpreted as the temperature in a real physical annealing process. Intuitively speaking, as the temperature T_k decreases to zero, the Boltzmann distributions will asymptotically concentrate on the region that contains the global optima (i.e., the set on which $H(\cdot)$ achieves global maximum). Although AAS has certain nice theoretical properties [57],[76], the idealized AAS cannot be directly used to solve optimization problems in practice since it is well known that sampling directly from a Boltzmann distribution is a difficult task. Although some Markov chain Monte Carlo (MCMC) techniques can be used to approximate the sampling procedure and soften the difficulty [58, 78, 76], it is still not convenient for practical implementation.

To inherit the advantages from AAS as well as CE and other model-based methods such as model reference adaptive search (MRAS) [32], we follow the same framework as Algorithm 4.2, and choose the Boltzmann distributions as the reference probability model, see case *c*) in the example in Chapter 3. Specifically, in the idea of MARS, the target reference distribution is the Boltzmann distribution (5.1). Note that when it comes to discrete cases, the integral in (5.1) is simply replaced by summation. Intuitively, as the temperature parameter T_k decreases to a small constant $T^* \geq 0$, the sequence of the Boltzmann distributions $\{g_k\}$ will converge to a limiting distribution g^* that puts most of the weight on the region that contains promising solutions, and solutions generated with small T^* will be close to the global optima with high probability. However, an obvious difficulty is that the Boltzmann distributions involve the objective function $H(\cdot)$, which is unknown explicitly a priori. Also as we mentioned in AAS, sampling directly from this Boltzmann distribution with explicit structure is still intractable. To address this problem, as what has been done in the CE method [60], we specify a family of parameterized distributions $\{f_\theta, \theta \in \mathbf{\Theta}\}$ as the surrogate sampling distributions which are approximations of the Boltzmann distributions. The approximation follows (3.2), i.e., it minimizes the KL divergence between the reference Boltzmann distribution g_{k+1} and the parameterized family. Compared with CE and MRAS, MARS eliminates difficulty of the quantile estimation of an unknown objective function under a given distribution. At the same time, this approach avoids the task of directly sampling from the Boltzmann distribution. In contrast, generating random samples from the surrogate distributions becomes much easier, while these surrogate distributions are expected to asymptotically track the target Boltzmann distributions.

In MARS, similar to (3.3), instead of directly using the Boltzmann sequence $\{g_k\}$ in (5.1) to minimize the KL-divergence as in (3.2), we consider a more general sequence of distributions in the recursive form

$$\widetilde{g}_{k+1}(x) = \alpha_k g_{k+1}(x) + (1-\alpha_k) f_{\theta_k}(x) \quad \text{with } \alpha_k \in (0,1] \; \forall\, k = 0,1,\ldots, \tag{5.2}$$

where α_k is a smoothing parameter so that each $\widetilde{g}_{k+1}$ is a mixture of the Boltzmann density function $g_{k+1}(x) = \frac{e^{H(x)/T_{k+1}}}{\int_{\mathbb{X}} e^{H(x)/T_{k+1}} \nu(dx)}$ and the sampling distribution f_{θ_k} obtained at the kth iteration. Intuitively, such a mixture $\widetilde{g}_{k+1}$ leads the sampling distribution to be updated in a smooth way. It retains the properties of the new $k+1$ th Boltzmann distribution, while making the new sampling distribution $f_{\theta_{k+1}}$ not to deviate too much from the current sampling distribution f_{θ_k}. From our preliminary numerical implementations, this smoothing step is especially useful.

We now present the simple idealized MARS algorithm as follows:

Algorithm 5.1. Model-based Annealing Random Search (Idealized Version)

Step 0: Specify an initial parameterized density/mass function $f_{\theta_0}(x)$ on $\mathbb{X}$, $\theta_0 \in \Theta$, an annealing schedule $\{T_k\}$, and a sequence $\{\alpha_k\}$. Set iteration counter $k = 0$.

Step 1: Compute the new parameter $\theta_{k+1} = \arg\min_{\theta \in \Theta} \mathscr{D}(\widetilde{g}_{k+1}, f_\theta)$, where $\widetilde{g}_{k+1}$ is given by (5.2).

Step 2: If a stopping rule is satisfied, then terminate; otherwise set $k = k+1$ and go to Step 1.

As in CE and Algorithm 4.2, we choose the Natural Exponential Family (NEFs) to be the parameterized family. Again, a significant advantage for choosing NEFs is that the parameters generated in Step 1 could be computed analytically in a closed from for arbitrary $\widetilde{g}_k$ in step 1. When NEFs are used in Algorithm 5.1 to approximate the generalized target Boltzmann distributions (5.2), the following lemma establishes a key connection between the idealized MARS and the gradient search.

Lemma 5.1.1. *If f_θ belongs to the NEF and the new parameter θ_{k+1} obtained via minimizing $\mathscr{D}(\widetilde{g}_{k+1}, f_\theta)$ satisfies $\theta_{k+1} \in int(\Theta)$ for all k, then*

$$m(\theta_{k+1}) - m(\theta_k) = -\alpha_k \nabla_\theta \mathscr{D}(g_{k+1}, f_\theta)|_{\theta=\theta_k} \quad \forall\, k = 0, 1, 2, \ldots, \tag{5.3}$$

Proof. Note that since θ_{k+1} is an interior point of Θ, it satisfies the first order necessary condition for optimality of the problem $\min_{\theta\in\Theta} \mathscr{D}(\widetilde{g}_{k+1}, f_\theta)$. Thus, by directly applying Lemma 2 in [32], we have $m(\theta_{k+1}) = E_{f_{\theta_{k+1}}}[\Gamma(X)] = E_{\widetilde{g}_{k+1}}[\Gamma(X)]$. It follows from (5.2) that

$$m(\theta_{k+1}) = E_{\widetilde{g}_{k+1}}[\Gamma(X)] = \alpha_k E_{g_{k+1}}[\Gamma(X)] + (1 - \alpha_k) m(\theta_k).$$

Thus, the difference between the two successive mean parameter vectors can be written as

$$m(\theta_{k+1}) - m(\theta_k) = \alpha_k \Big(E_{g_{k+1}}[\Gamma(X)] - m(\theta_k) \Big)$$

$$
\begin{aligned}
&= \alpha_k E_{g_{k+1}}\left[\Gamma(X) - \frac{\int_{\mathbb{X}} e^{\theta_k^T\Gamma(x)}\Gamma(x)\nu(dx)}{\int_{\mathbb{X}} e^{\theta_k^T\Gamma(x)}\nu(dx)}\right] \\
&= -\alpha_k \nabla_\theta \mathscr{D}(g_{k+1}, f_\theta)|_{\theta=\theta_k},
\end{aligned}
$$

where the second equality above follows from the definitions of $m(\theta)$ and NEFs, and the interchange of derivative and integral in the last step is guaranteed by the dominated convergence theorem. □

Similar to the discussion of Lemma 3.2.1, Lemma 5.1.1 states that MARS implicitly interprets a deterministic gradient search method for a *time-varying* objective function on the parameter space, where the sampling distribution parameterized by the optimal parameter θ^* is the best approximation to the limiting Boltzmann distribution g^*. Moreover, the smoothing parameter α_k turns out to be the step size in the gradient recursion. As in Algorithm 4.2, MARS implicitly solves the counterpart problem of (3.1) with nice structures, and this may explain its outstanding performance on hard optimization problems in practice. As we know that a careful control of the gain sequence will lead the gradient-based method to convergence, it is natural to expect the global convergence for MARS as well.

Since the idealized MARS algorithm involves calculating the true expectation with respect to g_k, it is difficult to compute the new parameter $\widehat{\theta}_{k+1}$ in Step 2. To address this issue, we present the Monte Carlo version of MARS algorithm that uses the random samples to estimate those expected values. Before we bring out the Monte Carlo version of MARS, we need to define two additional parameter sequences. The first one is the sample size sequence $\{N_k,\ k = 0, 1, \ldots\}$, which specifies the number of candidate solutions to be generated at each iteration. The second one is a constant sequence $\{\lambda_k,\ k = 0, 1, \ldots\}$ called sample allocation rule. Specifically, in Monte Carlo MARS, we use a mixing parameter $\lambda_k \in (0, 1]$ to mix the current sampling distribution and the initial sampling distribution, and then use the mixture instead of the current sampling distribution to generate random candidate solutions.

Algorithm 5.2. Model-based Annealing Random Search

Step 0: Choose an initial density/mass function $f_{\widehat{\theta}_0}(x)$ on $\mathbb{X}$, $\widehat{\theta}_0 \in int(\Theta)$. Specify an annealing schedule $\{T_k\}$, a gain sequence $\{\alpha_k\}$, a sample size sequence $\{N_k\}$ and a sample allocation rule $\{\lambda_k\}$. Set iteration counter $k = 0$.

Step 1: Generate a population of N_k i.i.d. solutions $\Lambda_k = \{X_1, \ldots, X_{N_k}\}$ from $\widehat{f}_{\widehat{\theta}_k}(x) := (1 - \lambda_k) f_{\widehat{\theta}_k}(x) + \lambda_k f_{\widehat{\theta}_0}(x)$.

Step 2: Compute the new parameter $\widehat{\theta}_{k+1} = \arg\min_{\theta \in \Theta} \mathscr{D}(\widehat{g}_{k+1}, f_\theta)$, where $\widehat{g}_{k+1}$ is given in (5.4).

Step 3: If a stopping rule is satisfied, then terminate; otherwise set $k = k+1$ and go to Step 1.

Several remarks are listed as follows.

- *Initial sampling distribution:* In practice, if there is no information at the very beginning on where the global optima are located, the initial sampling distribution $f_{\widehat{\theta}_0}$ should be chosen in a way that it could cover the entire solution space $\mathbb{X}$. In other words, any region in $\mathbb{X}$ would have a positive probability of being sampled. For this purpose, one could simply choose an approximate uniform distribution, e.g., a normal distribution with sufficiently large variance.

- *Initial reference distribution:* The initial Boltzmann distribution g_0 is chosen in a similar way as the choice of the initial sampling distribution $f_{\widehat{\theta}_0}$, i.e., it is expected to cover the entire $\mathbb{X}$. For this purpose, we simply set the initial temperature T_0 to a sufficiently large value.

- *Sample allocation rule:* An intuitive explanation of the sample allocation rule $\{\lambda_k\}$ is that at each iteration, we tend to spend a proportion of the sampling effort to the *blind search* (i.e., it searches the entire solution space with equivalent priority) carried out by $f_{\widehat{\theta}_0}$. This effort may prevent the algorithm from premature convergence. As the iteration number increases, the distribution $f_{\widehat{\theta}_k}$ becomes more likely to concentrate on the promising region. As a result, this "back door" trick becomes less attractive and therefore we could let $\{\lambda_k\}$ vanish to zero.

- *Reference distributions in Monte Carlo MARS:* At Step 2, the KL divergence is with respect to $\widehat{g}_{k+1}$, an estimate of $\widetilde{g}_{k+1}$ (cf. (5.2)) based on the sampled solutions in Λ_k, i.e.,

$$\widehat{g}_{k+1}(x) = \alpha_k \sum_{y \in \Lambda_k} \bar{g}_{k+1}(x)\delta(x-y) + (1-\alpha_k) f_{\widehat{\theta}_k}(x), \quad x \in \mathbb{X}, \qquad (5.4)$$

where δ is the Dirac delta function and we have replaced the Boltzmann distribution g_{k+1} in (5.2) by a discrete empirical distribution

$$\bar{g}_{k+1}(x) := \frac{e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\widehat{\theta}_k}(x)}{\sum_{x \in \Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\widehat{\theta}_k}(x)} \quad \forall\, x \in \Lambda_k. \qquad (5.5)$$

Note that in Monte Carlo MARS, we use the Monte Carlo technique to estimate the true expectation as well as some other incalculable integrals. The division by $\widehat{f}_{\widehat{\theta}_k}$ in $\bar{g}_{k+1}$ is used to guarantee the sample average $\frac{1}{N_k} \sum_{x \in \Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\widehat{\theta}_k}(x)$ to be an unbiased estimator of the integral $\int_{\mathbb{X}} e^{\frac{H(x)}{T_{k+1}}} \nu(dx)$.

Similar to Assumption A1 in Chapter 4, we assume that the new parameter obtained in Algorithm 5.2 satisfies the following condition:

Assumption C1. *The parameter $\widehat{\theta}_{k+1}$ computed at Step 2 of Algorithm 5.2 satisfies $\widehat{\theta}_{k+1} \in int(\Theta)$ for all k.*

Similar to Lemma 5.1.1, the following result shows the connection between the successive mean parameter vectors obtained in Algorithm 5.2.

Lemma 5.1.2. *If C3 holds, then the mean parameter function $m(\widehat{\theta}_{k+1})$ of*

$f_{\widehat{\theta}_{k+1}}$ *satisfies*

$$m(\widehat{\theta}_{k+1}) - m(\widehat{\theta}_k) = -\alpha_k\Big(m(\widehat{\theta}_k) - E_{\bar{g}_{k+1}}[\Gamma(X)]\Big) \ \ \forall\, k = 0, 1, 2, \ldots. \qquad (5.6)$$

Proof. Follows from Lemma 2 in Hu et al. [32] and the definition of $\widehat{g}_{k+1}$. □

To state the connection between Algorithm 5.2 (the Monte Carlo version of MARS) and SA, we rewrite (5.6) as follows:

$$\begin{aligned}
m(\widehat{\theta}_{k+1}) &= m(\widehat{\theta}_k) - \alpha_k\Big(m(\widehat{\theta}_k) - E_{g_{k+1}}[\Gamma(X)] + E_{g_{k+1}}[\Gamma(X)] - E_{\bar{g}_{k+1}}[\Gamma(X)]\Big) \\
&= m(\widehat{\theta}_k) - \alpha_k \nabla_\theta \mathscr{D}(g_{k+1}, f_\theta)|_{\theta=\widehat{\theta}_k} \\
&\quad - \alpha_k\left(\frac{\int_{\mathbb{X}} e^{\frac{H(x)}{T_{k+1}}}\Gamma(x)\nu(dx)}{\int_{\mathbb{X}} e^{\frac{H(x)}{T_{k+1}}}\nu(dx)} - \frac{\frac{1}{N_k}\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}}\Gamma(x)/\widehat{f}_{\widehat{\theta}_k}(x)}{\frac{1}{N_k}\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}}/\widehat{f}_{\widehat{\theta}_k}(x)}\right). \qquad (5.7)
\end{aligned}$$

This is a typical SA recursion in Robbins-Monro's form with the true gradient of $\mathscr{D}(g_{k+1}, f_\theta)$ with respect to θ and an error term due to the bias and noise caused by Monte Carlo random sampling in MARS.

5.2 Global Convergence of MARS

Note that the stochastic approximation recursion (5.7) involves a *time-varying* function $\mathscr{D}(g_{k+1}, f_\theta)$. However, as we expect that the sequence $\{g_k\}$ will converge to some g^*, the *time-varying* function $\mathscr{D}(g_{k+1}, f_\theta)$ will finally converge to $\mathscr{D}(g^*, f_\theta)$. This desired property would imply the convergence of the optimal solution $\{\theta_k\}$ to some θ^*. Moreover, if g^* concentrates on the globally optimal solutions, f_{θ^*} is expected to concentrate on the globally optimal solutions as well. Throughout this and the next section, we follow the approach in [16] to study the convergence properties of MARS.

Since the MARS algorithm is randomized, we need to establish some probabilistic definitions. Throughout this chapter, all the definitions such as $P(\cdot)$, $E[\cdot]$, $\mathscr{F}_k = \sigma\{\Lambda_0, \Lambda_1, \ldots, \Lambda_{k-1}\}$ where $k = 1, 2, \ldots$, $P_{\widehat{f}_{\widehat{\theta}_k}}(\cdot|\mathscr{F}_k)$, $E_{\widehat{f}_{\widehat{\theta}_k}}[\cdot|\mathscr{F}_k]$ are carried out in the same manner as they are defined in the analysis of Algorithm 4.2 in Chapter 4. We also use the following shorthand notations to simplify the presentation.

$$\begin{aligned} \mathbb{U}_k = \frac{1}{N_k} \sum_{x \in \Lambda_k} e^{\frac{H(x)}{T_{k+1}}} \Gamma(x) / \widehat{f}_{\widehat{\theta}_k}(x), \;\; \bar{\mathbb{U}}_k = E_{\widehat{f}_{\widehat{\theta}_k}}[U_k|\mathscr{F}_k] = \int_{\mathbb{X}} e^{\frac{H(x)}{T_{k+1}}} \Gamma(x) \nu(dx) \\ V_k = \frac{1}{N_k} \sum_{x \in \Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\widehat{\theta}_k}(x), \;\; \bar{V}_k = E_{\widehat{f}_{\widehat{\theta}_k}}[V_k|\mathscr{F}_k] = \int_{\mathbb{X}} e^{\frac{H(x)}{T_{k+1}}} \nu(dx). \end{aligned} \quad (5.8)$$

For the convergence analysis, we make the following assumptions that will be used throughout this chapter.

Assumptions:

C2. *The problem (3.1) has a unique globally optimal solution, i.e., $\exists x^* \in \mathbb{X}$ such that $H(x) < H(x^*)$ $\forall x \neq x^*$, $x \in \mathbb{X}$. Moreover, $H(x) > 0$ $\forall x \in \mathbb{X}$.*

C3. *For any $\varepsilon < H(x^*)$, the set $\{x \in \mathbb{X} : H(x) \geq \varepsilon\}$ has a positive Lebesgue/discrete measure.*

C4. *For any $\delta > 0$, $\sup_{x \in A_\delta} H(x) < H(x^*)$, where $A_\delta := \{x \in \mathbb{X} : \|x - x^*\| \geq \delta\}$*

with $\|\cdot\|$ being the Euclidean norm in $\Re^n$, and we define the supremum over the empty set to be $-\infty$.

C5. *The mapping $\Gamma(x)$ given in Definition 1 is bounded on $\mathbb{X}$. Moreover, for any $\xi > 0$, there exists $\delta > 0$ such that $\|\Gamma(x) - \Gamma(x^*)\| \le \xi$ whenever $\|x - x^*\| \le \delta$.*

C6. *The gain sequence $\{\alpha_k\}$ satisfies $\alpha_k > 0\ \forall k$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.*

C7. *(a) The annealing schedule $\{T_k\}$ satisfies $T_k > 0\ \forall k$ and $T_k \to T^* \ge 0$ as $k \to \infty$;*

(b) The sample allocation rule $\{\lambda_k\}$ satisfies $\lambda_k > 0\ \forall k$ and $\lambda_k \to \lambda^* \in [0, 1)$ as $k \to \infty$;

(c) $\frac{e^{2H^*/T_k}}{N_k \lambda_k} \to 0$ as $k \to \infty$, where $H^* = H(x^*)$.

Assumptions C2-C4 are reasonable conditions on the objective function H. Intuitively, under C2, the limiting reference distribution g^* is expected to become degenerated, which assigns unit mass on the globally optimal solution x^*. Without loss of generality, we just assume $H(x) > 0$. C3 ensures that the area (may not be a continuous open set) containing ε-optimal solutions has a positive probability to be visited by the sampling distribution $\widehat{f}_{\widehat{\theta}_k}$. C4 ensures that the global optimum could be distinguished from other solutions outside its neighborhood by a positive difference, so that the algorithm will finally concentrate on this neighborhood and will not be disturbed by other solutions outside this area. On the other hand, Assumptions C5−C7 are regularity conditions on the input parameters. Since $\Gamma(\cdot)$ is continuous and $\mathbb{X}$ is compact, C5 becomes trivial and holds for natural exponential distributions, e.g., normal, exponential, and Gamma distribution, as well as many discrete mass functions encountered in practice, e.g., Bernoulli, Binomial and Poisson. As a result, since $\Gamma(x)$ is bounded, the initial parameterized density/mass function $f_{\widehat{\theta}_0}(x)$ is bounded away from zero on $\mathbb{X}$ for any given $\widehat{\theta}_0 \in int(\Theta)$, i.e., $f_* := \inf_{x\in\mathbb{X}} f_{\widehat{\theta}_0}(x) > 0$. C6 is a typical SA condition; it ensures that the gain sequence $\{\alpha_k\}$ will not decay too fast or too slow which may cause premature convergence or an extremely slow convergence speed (see [67] for a detailed discussion). C7(a) and (b) assume that both the temperature annealing schedule and the sample allocation rule should converge to certain limits. (Note that T_k is not necessarily monotone. In practice, some non-monotonic annealing schedules may lead to superior performance, see [45, 61, 76]). C7(c) states that all the parameters $\{T_k\}$, $\{N_k\}$ and $\{\lambda_k\}$ should be chosen in balance. Roughly speaking, the annealing schedule $\{T_k\}$ determines the convergence speed of the sequence of (idealized) Boltzmann distributions $\{g_k\}$ to the limiting distribution g^*, whereas $\{N_k\}$ determines how the surrogate distributions $\{f_{\widehat{\theta}_k}\}$ approximate the target Boltzmann distributions. Thus, if the temperature T_k decays to zero at a fast rate, it means that g_k will converge to g^* very fast. Then the sample size N_k should also increase sufficiently fast to ensure that the surrogate distributions $\{f_{\widehat{\theta}_k}\}$ can keep "tracking" the sequence of convergent Boltzmann distributions. Intuitively, as the iteration number k increases, g_k becomes more likely to concentrate on the region containing the global optimum. Therefore, we would need a better approximation between $f_{\widehat{\theta}_k}$ and g_k, in order to keep $f_{\widehat{\theta}_k}$ staying close with g_k. Moreover, we consider the following special cases of C7(c):

- *Positive limiting temperature.* $T^* > 0$.

  It is obvious that when $T^* > 0$ and $\lambda^* > 0$, C6 simply holds if $N_k \to \infty$ as $k \to \infty$. As we will see, $T^* > 0$ would bring out an easy theoretical analysis, since in this case the Boltzmann distributions just converge to a regular distribution rather than a degenerated distribution (i.e., δ-function). Moreover, from a practical point of view, by selecting a sufficiently small positive T^*, one could still get any desired level of precisions (see the remark after the proof of Lemma 5.2.1).

- *Logarithmic annealing schedule.* $T_k = \frac{T_0}{\ln(1+k)}$.

  It can be shown that when $T_k = \frac{T_0}{\ln(1+k)}$, $N_k = \Theta(k^\beta)$, and $\lambda_k = \Omega(k^{-\gamma})$ for some constants $T_0 > 0$, $\beta > 0$, and $\gamma > 0$, C6 is satisfied for $\beta > \gamma$, with T_0 sufficiently large. Here the logarithmic annealing schedule $\{T_k\}$ is frequently used in simulated annealing algorithms.

- *Polynomial annealing schedule.* $T_k = \frac{T_0}{1+ck}$.

  It can be shown that when $T_k = \frac{T_0}{1+ck}$, $N_k = \Theta(\beta^k)$, and $\lambda_k = \Omega(k^{-\gamma})$ for constants $T_0 > 0$, $c > 0$, $\beta > 1$, and $\gamma > 0$, it is easy to verify that C6 is satisfied by taking $\beta > e^{2H^* c/T_0}$. Moreover, it is easy to see that in the polynomial annealing schedule, the sequence of temperatures has a faster decay rate in contrast to the sequence in the logarithmic annealing schedule.

We present the following convergence theorem for MARS.

Theorem 5.2.1. *If Assumptions C1 to C7 hold, then*

$$m(\widehat{\theta}_k) \to \Gamma^* \quad as \quad k \to \infty \quad w.p.1,$$

where the limit is taken component-wise, $\Gamma^* := \Gamma(x^*)$ *if* $T^* = 0$, $\Gamma^* := E_{g^*}[\Gamma(X)]$ *whenever* $T^* > 0$, *and* g^* *is the limiting Boltzmann distribution parameterized by* $T^* > 0$.

The interpretation of Theorem 5.2.1 depends on the parameterized distributions chosen by MARS. For example, if we choose the independent univariate normal distribution or multivariate normal distribution for continuous optimization problems as in Section 4.4, and set T^* to zero, then Theorem 5.2.1 implies that

$$\lim_{k \to \infty} E_{f_{\widehat{\theta}_k}}[X] = x^* \quad \text{and} \quad \lim_{k \to \infty} \mathrm{Cov}_{f_{\widehat{\theta}_k}}[X] = 0 \;\; \text{w.p.1},$$

where $f_{\widehat{\theta}_k}$ is the corresponding parameterized distribution. In this case, the parameterized distribution will converge to a delta distribution that assigns unit mass on the unique globally optimal solution x^*. We take a discrete optimization problem as the second example. Assume that the feasible region $\mathbb{X}$ consists of m distinct points, and Q is a $m \times 1$ vector whose ith entry q_i

denoted the probability mass corresponding to the ith point in $\mathbb{X}$. Therefore the p.m.f of the sampling distribution can be written as

$$f_\theta(x) = \prod_{i=1}^{m} q_i^{I\{x=\mathbf{x}_i\}} := e^{\theta^T \Gamma(x)},$$

where $\theta = [\ln q_1, \ldots, \ln q_m]^T$ and $\Gamma(x) = [I\{x = \mathbf{x}_1\}, \ldots, I\{x = \mathbf{x}_m\}]^T$. With $T^* = 0$, Theorem 5.2.1 yields that

$$\lim_{k\to\infty} \sum_{x\in\mathbb{X}} \prod_{i=1}^{m} \left(q_i^k\right)^{I\{x=\mathbf{x}_i\}} I\{x = \mathbf{x}_j\} = I\{x^* = \mathbf{x}_j\} \;\; \forall\, j \;\; \text{w.p.1},$$

where q_i^k is the ith entry of the vector Q_k obtained at the kth iteration of MARS. In other words, the p.m.f Q_k will converge to a unit mass function that degenerates at the global optimum x^*. Note that this example is very easy to extend to general discrete optimization problems in $\Re^n$, where we just need to slightly adapt the corresponding p.m.f.(e.g., see Section 5.4.)

Before we prove Theorem 5.2.1, we first show a property of the Boltzmann distribution with Γ in the parameterized family.

Lemma 5.2.1. *If Assumptions C3, C4, C5, and C7(a) are satisfied, then*

$$E_{g_k}[\Gamma(X)] \to \Gamma^* \quad \textit{as } k \to \infty.$$

Proof. The $T^* > 0$ case is trivial and is thus omitted. Note that when $T^* = 0$, we have $\Gamma^* = \Gamma(x^*)$. By Assumption C5, for any $\xi > 0$, we can find a $\delta > 0$ such that $\|x - x^*\| < \delta$ implies $\|\Gamma(x) - \Gamma^*\| < \xi$. Define

$A_\delta = \{x \in \mathbb{X} : \|x - x^*\| \geq \delta\}$. We have by C4 that $\bar{H} = \sup_{x \in A_\delta} H(x) < H^*$. Take $\varepsilon = \frac{\bar{H}+H^*}{2}$. By C3, the set $B_\epsilon := \{x \in \mathbb{X} : H(x) > \varepsilon\}$ has a positive Lebesgue/discrete measure. Thus,

$$\begin{aligned}
\|E_{g_k}[\Gamma(X)] - \Gamma^*\| &\leq E_{g_k}\left[\|\Gamma(X) - \Gamma^*\|\right] \\
&= \int_{A_\delta^c} \|\Gamma(x) - \Gamma^*\| g_k(x)\nu(dx) + \int_{A_\delta} \|\Gamma(x) - \Gamma^*\| g_k(x)\nu(dx) \\
&\leq \xi + \sup_{x \in \mathbb{X}} \|\Gamma(x) - \Gamma^*\| \frac{\int_{A_\delta} e^{\frac{H(x)}{T_k}} \nu(dx)}{\int_{\mathbb{X}} e^{\frac{H(x)}{T_k}} \nu(dx)} \\
&\leq \xi + \sup_{x \in \mathbb{X}} \|\Gamma(x) - \Gamma^*\| \frac{\int_{A_\delta} e^{\frac{H(x)}{T_k}} \nu(dx)}{\int_{B_\varepsilon} e^{\frac{H(x)}{T_k}} \nu(dx)} \\
&\leq \xi + \sup_{x \in \mathbb{X}} \|\Gamma(x) - \Gamma^*\| e^{\frac{-(H^* - \bar{H})}{2T_k}} \frac{\nu(A_\delta)}{\nu(B_\varepsilon)}.
\end{aligned}$$

Since $\Gamma(x)$ is bounded, ξ is arbitrary and $H^* > \bar{H}$, we have $E_{g_k}[\Gamma(X)] \to \Gamma^*$ as $T_k \to 0$. □

From the proof of Lemma 5.2.1 we can see that for any given $\varepsilon > 0$, there exists a sufficiently small $T^* > 0$ such that $\|E_{g^*}[\Gamma(X)] - \Gamma(x^*)]\| \leq \varepsilon$. Therefore, if in practice we only search for ε-optimal solution with a desired precision ε , we could in fact set T^* to a small positive value.

The next intermediate result shows that the conditional bias of the error term in (5.7) converges to zero w.p.1.

Lemma 5.2.2. *If Assumptions C4 and C6 hold, then*

$$E_{f_{\widehat{\theta}_k}}\left[\frac{\mathbb{U}_k}{V_k}\bigg|\mathscr{F}_k\right] \to \frac{\bar{\mathbb{U}}_k}{\bar{V}_k} \text{ as } k \to \infty \text{ w.p.1,}$$

where the limit is component-wise.

Proof. See Appendix. □

Proof of Theorem 5.2.1. We rewrite (5.6) in the following recursive form:

$$\eta_{k+1} = \eta_k - \xi_k,$$

where $\eta_k := m(\widehat{\theta}_k) - \Gamma^*$, and $\xi_k = \alpha_k\Big(m(\widehat{\theta}_k) - \frac{\mathbb{U}_k}{V_k}\Big)$. Let $M_k = E_{f_{\widehat{\theta}_k}}[\xi_k|\mathscr{F}_k]$ and $Z_k = \xi_k - M_k$. To show the desired convergence result, we establish that the multivariate versions of conditions (i)-(iv) in [16] hold.

[i] First we show that for every $\epsilon > 0$, the probability that $\{\|\eta_k\| > \epsilon,\ \eta_k^T M_k < 0\}$ occurs infinitely often (i.o.) is zero. To this end, we write M_k as

$$M_k = \alpha_k\Bigg(m(\widehat{\theta}_k) - \Gamma^* + \Gamma^* - E_{g_{k+1}}[\Gamma(X)] + E_{g_{k+1}}[\Gamma(X)] - E_{f_{\widehat{\theta}_k}}\left[\frac{\mathbb{U}_k}{V_k}\bigg|\mathscr{F}_k\right]\Bigg). \tag{5.9}$$

It follows that

$$\eta_k^T M_k = \alpha_k\Bigg(\|\eta_k\|^2 + \eta_k^T\left(\Gamma^* - E_{g_{k+1}}[\Gamma(X)]\right) + \eta_k^T\Bigg(E_{g_{k+1}}[\Gamma(X)] - E_{f_{\widehat{\theta}_k}}\left[\frac{\mathbb{U}_k}{V_k}\bigg|\mathscr{F}_k\right]\Bigg)\Bigg).$$

Since η_k is bounded, by Lemma 5.2.1, the second term in the parenthesis above vanishes to zero as $k \to \infty$, whereas Lemma 5.2.2 implies that the third term also vanishes to zero w.p.1. as $k \to \infty$. Therefore, for almost every sample path generated by MARS, we must have $\eta_k^T M_k > 0$ whenever $\|\eta_k\| > \epsilon$ for k sufficiently large, i.e., $P(\|\eta_k\| > \epsilon,\ \eta_k^T M_k < 0\ i.o.) = 0$.

[ii] Note that $m(\widehat{\theta}_k) = E_{f_{\widehat{\theta}_k}}[\Gamma(X)]$ and $\frac{\mathbb{U}_k}{V_k} = E_{\bar{g}_{k+1}}[\Gamma(X)]$, where $\bar{g}_{k+1}$ is defined in (5.5). Since the mapping Γ is bounded on $\mathbb{X}$ by C5, both $m(\widehat{\theta}_k)$ and $\frac{\mathbb{U}_k}{V_k}$ are bounded. Moreover, we have from Assumption C6 that $\alpha_k \to 0$ as $k \to \infty$. Therefore, $\|M_k\|(1 + \|\eta_k\|)^{-1} \to 0$ as $k \to \infty$ w.p.1, which shows condition (ii) in [16].

[iii] By definition, we have $Z_k = \alpha_k\Big(E_{f_{\widehat{\theta}_k}}\big[\frac{\mathbb{U}_k}{V_k}|\mathscr{F}_k\big] - \frac{\mathbb{U}_k}{V_k}\Big)$. Therefore,

$$\sum_{k=1}^{\infty} E[\|Z_k\|^2] = \sum_{k=1}^{\infty} \alpha_k^2 E\Bigg[\Big(E_{f_{\widehat{\theta}_k}}\Big[\frac{\mathbb{U}_k}{V_k}|\mathscr{F}_k\Big] - \frac{\mathbb{U}_k}{V_k}\Big)^T \Big(E_{f_{\widehat{\theta}_k}}\Big[\frac{\mathbb{U}_k}{V_k}|\mathscr{F}_k\Big] - \frac{\mathbb{U}_k}{V_k}\Big)\Bigg] < \infty,$$

since $\frac{\mathbb{U}_k}{V_k}$ is bounded and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$ by C6.

[iv] Finally, we establish condition (iv) in [16] by showing that

$$P\big(\liminf_{k\to\infty} \|\eta_k\| > 0,\ \sum_{k=1}^{\infty} \|M_k\| < \infty\big) = 0.$$

From (5.9), we have

$$\|M_k\| \geq \alpha_k\Bigg(\|\eta_k\| - \|\Gamma^* - E_{g_{k+1}}[\Gamma(X)]\| - \Bigg\|E_{g_{k+1}}[\Gamma(X)] - E_{f_{\widehat{\theta}_k}}\Big[\frac{\mathbb{U}_k}{V_k}\Big|\mathscr{F}_k\Big]\Bigg\|\Bigg).$$

Let $\Omega_1 = \{\liminf_{k\to\infty} \|\eta_k\| > 0\}$ and $\Omega_2 = \{\sum_{k=1}^{\infty} \|M_k\| < \infty\}$. For every sample point $\omega \in \Omega_1$, we can find a $\delta > 0$ such that $\liminf_{k\to\infty} \|\eta_k\| >$

$\delta > 0$. This implies that there exists a $K_\delta(\omega)$ such that $\|\eta_k\| \geq \delta \; \forall \, k \geq K_\delta(\omega)$. In addition, let $\Omega_3 = \{\big\|E_{g_{k+1}}[\Gamma(X)] - E_{\widehat{f}_{\widehat{\theta}_k}}\big[\frac{\mathbb{U}_k}{V_k}\big|\mathscr{F}_k\big]\big\| \to 0\}$. Note that Lemma 5.2.2 implies $P(\Omega_3) = 1$. Since $E_{g_{k+1}}[\Gamma(X)] \to \Gamma^*$ as $k \to \infty$, there exists a $\bar{K}_{\delta/2}(\omega)$ for every $\omega \in \Omega_3$ such that

$$\big\|\Gamma^* - E_{g_{k+1}}[\Gamma(X)]\big\| + \left\|E_{g_{k+1}}[\Gamma(X)] - E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{\mathbb{U}_k}{V_k}\Big|\mathscr{F}_k\right]\right\| < \frac{\delta}{2}$$

for all $k \geq \bar{K}_{\delta/2}(\omega)$. Consequently, we have for every $\omega \in \Omega_1 \cap \Omega_3$, $\|M_k\| > \frac{\delta}{2}\alpha_k$ for all $k \geq K^*(\omega) := \max\{K_\delta(\omega), \bar{K}_{\delta/2}(\omega)\}$. Thus by C6,

$$\sum_{k=1}^{\infty} \|M_k\| \geq \sum_{k=K^*(\omega)}^{\infty} \|M_k\| \geq \frac{\delta}{2} \sum_{k=K^*(\omega)}^{\infty} \alpha_k = \infty \quad \forall \, \omega \in \Omega_1 \cap \Omega_3.$$

This implies $P(\Omega_1 \cap \Omega_2 \cap \Omega_3) = 0$. Thus, it follows that $P(\Omega_1 \cap \Omega_2) = P(\Omega_1 \cap \Omega_2 \cap \Omega_3) + P(\Omega_1 \cap \Omega_2 \cap \Omega_3^c) \leq P(\Omega_3^c) = 0$.

Finally, by directly applying the result of Evans and Weber [16], we have $\eta_k \to 0$ as $k \to \infty$ w.p.1, which completes the proof of the theorem.

5.3 Asymptotic Normality of MARS

Similar to Algorithm 4.2 and its convergence rate analysis in Section 4.3, we perform the convergence rate analysis for MARS following the asymptotic normality analysis of SA recursion in Fabian [17], i.e., we will show that the conditions (2.2.1), (2.2.2) and (2.2.3) in [17] will be satisfied under the scheme of MARS. In order to do so, we need the following two assumptions.

Assumptions:

D1. *For a given sample size sequence $N_k = \Theta(k^{\beta})$ and a gain sequence $\alpha_k = \Theta(k^{-\alpha})$, the sequence $\{T_k\}$ satisfies $T_k > T^* > 0 \,\forall\, k$ and $\lim_{k\to\infty} k^{\frac{\alpha+\beta}{2}}\big(\frac{1}{T^*} - \frac{1}{T_k}\big) = 0$, and the sequence $\{\lambda_k\}$ satisfies $\lambda_k > 0 \;\forall\, k$, $\lambda_k \to \lambda^* \in [0,1)$ as $k \to \infty$, and $\lambda_k = \Omega(k^{-\gamma})$ for some positive constant $\gamma < \frac{\beta}{2}$.*

D2. *The limit of the sequence of parameters $\{\widehat{\theta}_k\}$ generated by MARS as $k \to \infty$ is an interior point of Θ, i.e., $m^{-1}(\Gamma^*) \in int(\Theta)$.*

D1 is reasonable since it could be satisfied by carefully controlling the annealing schedule. Moreover, D1 is a strengthened version of Assumption C7. Note that Theorem 5.2.1 still holds true with Assumption C7 replaced by D1. In particular, throughout this section we set $N_k = \Theta(k^{\beta})$ for some constant $\beta > 0$. Moreover, we set $\alpha_k = c/k^{\alpha}$ for some constants $c > 0$, and $\alpha \in (\frac{1}{2}, 1)$. Note that the gain sequence $\{\alpha_k\}$ satisfies Assumption C6. On the other hand, D2 is similar to Assumption B2 in Section 4.3. By the invertibility of $m(\cdot)$, the sequence of parameters $\{\widehat{\theta}_k\}$ generated by MARS converges to a limiting parameter $m^{-1}(\Gamma^*)$ w.p.1, which is assumed to be lying on the interior of Θ.

As in the discussions in section 4.3 (following Assumption B2), by inverse function theorem and the boundness of Γ,the sequence of sampling distributions $\{f_{\widehat{\theta}_k}\}$ converges point-wise to a limiting distribution $f_{m^{-1}(\Gamma^*)}$ w.p.1.

Given the specific forms of N_k and α_k, we can rewrite (5.6) in the form of a recursion in Fabian (1968):

$$\eta_{k+1} = (1 - ck^{-\alpha})\eta_k + k^{-(2\alpha+\beta)/2} R_k + k^{-(3\alpha+\beta)/2} W_k,$$

where $\eta_k = m(\widehat{\theta}_k) - \Gamma^*$,

$$R_k = ck^{\beta/2}\left(\frac{\mathbb{U}_k}{V_k} - E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{\mathbb{U}_k}{V_k}\middle|\mathscr{F}_k\right]\right) \text{ and } W_k = ck^{(\alpha+\beta)/2}\left[E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{\mathbb{U}_k}{V_k}\middle|\mathscr{F}_k\right] - \Gamma^*\right]$$

are the amplified noise and bias caused by Monte-Carlo sampling procedure in Step 1 of MARS (Algorithm 5.2).

The term R_k has the following properties:

Lemma 5.3.1. *If Assumptions $C1-C5$, $D1$, and $D2$ hold, then there exists a symmetric positive semi-definite matrix Σ such that the conditional covariance of the amplified noise $E_{\widehat{f}_{\widehat{\theta}_k}}[R_k R_k^T|\mathscr{F}_k] \to \Sigma$ as $k \to \infty$ w.p.1. In addition, the sequence $\{R_k\}$ is uniformly square integrable in the sense that*

$$\lim_{k\to\infty} E\big[I\{\|R_k\|^2 \geq rk^\alpha\}\|R_k\|^2\big] = 0 \quad \forall\, r > 0.$$

Proof. See Appendix for the proof. □

We next show that the term W_k vanishes to zero w.p.1. We break W_k into two parts and write $W_k = ck^{(\alpha+\beta)/2}W_{1,k} + ck^{(\alpha+\beta)/2}W_{2,k}$ for better representation, where

$$W_{1,k} = \frac{E_{\widehat{f}_{\widehat{\theta}_k}}[\mathbb{U}_k|\mathscr{F}_k]}{E_{\widehat{f}_{\widehat{\theta}_k}}[V_k|\mathscr{F}_k]} - \Gamma^* \quad \text{and} \quad W_{2,k} = E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{\mathbb{U}_k}{V_k}\middle|\mathscr{F}_k\right] - \frac{E_{\widehat{f}_{\widehat{\theta}_k}}[\mathbb{U}_k|\mathscr{F}_k]}{E_{\widehat{f}_{\widehat{\theta}_k}}[V_k|\mathscr{F}_k]}.$$

The convergence of W_k is a direct consequence of the following propositions, which are strengthened versions of Lemma 5.2.1 and Lemma 5.2.2.

Proposition 5.3.1. *If Assumptions C3, C4, and D1 hold, then*

$$k^{\frac{\alpha+\beta}{2}} W_{1,k} \to 0 \quad \text{as } k \to \infty.$$

Proof. We prove Proposition 5.3.1 in Appendix. □

Proposition 5.3.2. *Assume C1−C5, D1, and D2 hold, and $\beta > \alpha$, then*

$$k^{\frac{\alpha+\beta}{2}} W_{2,k} \to 0 \quad \text{as } k \to \infty \quad w.p.1,$$

where the limit is component-wise.

Proof. It is not difficult to show that the result of Theorem 5.2.1 still holds under the conditions of Proposition 5.3.2. We can bound $\|W_{2,k}\|$ by terms [i]−[v] as in the proof of Lemma 5.2.2. Next, invoking the strong convergence of the sequence $\{\widehat{\theta}_k\}$, an argument similar to the proof of Lemma 5.3.1 implies that all terms [i]−[v] are on the order of $O(N_k^{-1})$, independent of T_k. Therefore, $k^{\frac{\alpha+\beta}{2}} W_{2,k}$ approaches zero as $k \to \infty$ by taking $N_k = \Theta(k^\beta)$ with $\beta > \alpha$. □

We have the following asymptotic convergence rate result for MARS.

Theorem 5.3.1. *Let $\alpha_k = c/k^{\alpha}$ and $N_k = \Theta(k^{\beta})$ for constants $c > 0$, $\alpha \in (\frac{1}{2}, 1)$, and $\beta > \alpha$. Assume Assumptions C1−C4, D1, and D2 hold, then*

$$k^{\frac{\alpha+\beta}{2}}\Big(m(\widehat{\theta}_k) - \Gamma^*\Big) \xrightarrow{dist} \mathcal{N}\Big(0, \Sigma\Big) \quad \text{as } k \to \infty,$$

where $\Sigma = \Upsilon\, Cov_{\widehat{f}_{m^{-1}(\Gamma^*)}}\Big[\big(\Gamma(X) - E_{g^*}[\Gamma(X)]\big)g^*(X)/\widehat{f}_{m^{-1}(\Gamma^*)}(X)\Big]$ for some constant $\Upsilon > 0$.

Proof. Follows from Proposition 5.3.1, Proposition 5.3.2, and Lemma 5.3.1 above, and then by applying Theorem 2.2 in Fabian [17]. □

Theorem 5.3.1 implies that one could achieve a sufficiently large rate by choose a large β (α is bounded above by 1). However, the asymptotic convergence rate is defined in terms of the iteration number k rather than the sample size N_k, and increasing the sample size too fast (i.e., choosing a sufficiently large β) may have a negative impact on the practical performance of the algorithm, since it needs a large number of samples in each iteration. Moreover, the asymptotic rate only describes the limiting behavior of the algorithm, and does not specify its actual speed in practice. Therefore, from a practical point of view, we need to choose an appropriate β to balance the trade-off between a fast asymptotic convergence rate and a small number of samples consumed in each iteration. On the other hand, note that the convergence of the SA recursion (5.7) can be viewd as a combinition of two "parallel" convergences, i.e., the convergence affected by the true gradient and the convergence affected by the noise and bias. Theorem 5.3.1 indicates that given $\{\alpha_k\}$ fixed, the asymptotic rate is only determined by β, which further implies that the convergence of MARS will be finally dominated by the noise and bias.

5.4 Numerical Examples

We test the MARS algorithm (Algorithm 5.2) on both continuous and discrete optimization problems, and compare MARS to simulated annealing (SAN) algorithm, Hide-and-Seek (HAS) algorithm [58, 76], and model reference adaptive search (MRAS) [32].

Continuous Optimization

In the continuous case, we consider 12 benchmark problems frequently used in global optimization literature [55], [60], [61] ,[73], and they range from highly multimodal problems to badly-scaled problems. We also test MARS in different dimensions that varies from four to one hundred.

(1) Shekel's function ($n = 4$, $0 \leq x_i \leq 10,\ i = 1, \ldots, n$)

$$H_1(x) = \sum_{j=1}^{5} \Big(\sum_{i=1}^{4} (x_i - A_{i,j})^2 + B_j \Big)^{-1} - 10.1532,$$

with $B = (0.1, 0.2, 0.2, 0.4, 0.4)^T$, $A_1 = A_3 = (4, 1, 8, 6, 3)$, and $A_2 =$

$A_4 = (4, 1, 8, 6, 7)$, where A_i represents the ith row of A. The function has a global maximizer $x^* = (4, 4, 4, 4)^T$ and $H_1(x^*) = 0$.

(2) Hartmann function ($n = 6$, $0 \leq x_i \leq 1$, $i = 1, \ldots, n$)

$$H_2(x) = \sum_{i=1}^{4} c_i \exp\Big(-\sum_{j=1}^{6} B_{i,j}(x_j - A_{i,j})^2\Big) - 3.32237,$$

with $c = (1, 1.2, 3, 3.2)^T$,

$$A = \begin{pmatrix} 0.1312 & 0.1696 & 0.5569 & 0.0124 & 0.8283 & 0.5886 \\ 0.2329 & 0.4135 & 0.8307 & 0.3736 & 0.1004 & 0.9991 \\ 0.2348 & 0.1451 & 0.3522 & 0.2883 & 0.3047 & 0.6650 \\ 0.4047 & 0.8828 & 0.8732 & 0.5743 & 0.1091 & 0.0381 \end{pmatrix},$$

$$B = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix}.$$

The global maximizer $x^* = (0.20169, 0.150011, 0.476874, 0.275332, 0.311652, 0.6573)^T$ and $H_2(x^*) = 0$.

(3) Sinusoidal function ($n = 30$, $0 \leq x_i \leq 180$, $i = 1, \ldots, n$)

$$H_3(x) = 2.5 \prod_{i=1}^{n} \sin\left(\frac{\pi x_i}{180}\right) + \prod_{i=1}^{n} \sin\left(\frac{\pi x_i}{36}\right) - 3.5,$$

where $x^* = (90, \ldots, 90)^T$ and $H_3(x^*) = 0$.

(4) Rastrigin function ($n = 50,\ -5.12 \leq x_i \leq 5.12,\ \ i = 1, \ldots, n$)

$$H_4(x) = -\sum_{i=1}^{n} \left(x_i^2 - 10\cos(2\pi x_i)\right) - 10n,$$

where $H_4(x^*) = 0$.

(5) Pinter's function ($n = 50,\ -10 \leq x_i \leq 10,\ \ i = 1, \ldots, n$)

$$\begin{aligned} H_5(x) = &-\sum_{i=1}^{n} i x_i^2 - \sum_{i=1}^{n} 20i \sin^2\left(x_{i-1}\sin x_i - x_i + \sin x_{i+1}\right) \\ &-\sum_{i=1}^{n} i \log_{10}\left(1 + i(x_{i-1}^2 - 2x_i + 3x_{i+1} - \cos x_i + 1)^2\right) - 1, \end{aligned}$$

where $x_0 = x_n$, $x_{n+1} = x_1$, $x^* = (0, \ldots, 0)^T$, $H_5(x^*) = -1$.

(6) Weighted Sphere function ($n = 100,\ -10 \leq x_i \leq 10,\ \ i = 1, \ldots, n$)

$$H_6(x) = -1 - \sum_{i=1}^{n} i x_i^2,$$

where $x^* = (0, \ldots, 0)^T$ and $H_6(x^*) = -1$.

(7) Griewank function ($n = 100,\ -10 \leq x_i \leq 10,\ i = 1, \ldots, n$)

$$H_7(x) = -\frac{1}{4000}\sum_{i=1}^{n} x_i^2 + \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right) - 1,$$

where $x^* = (0, \ldots, 0)^T$, $H_7(x^*) = 0$.

(8) Trigonometric function ($n = 100$, $-10 \leq x_i \leq 10$, $i = 1 \ldots, n$)

$$H_8(x) = -1 - \sum_{i=1}^{n} \Big[8 \sin^2 \big(7(x_i - 0.9)^2\big) + 6 \sin^2 \big(14(x_i - 0.9)^2\big) + (x_i - 0.9)^2 \Big],$$

where $x^* = (0.9, \ldots, 0.9)^T$, $H_8(x^*) = -1$.

(9) Powell function ($n = 100, \ -10 \leq x_i \leq 10, \ i = 1 \ldots, n$)

$$H_9(x) = -1 - \sum_{i=1}^{(n-2)/2} \Big[(x_{2i-1} + 10x_{2i})^2 + 5(x_{2i+1} - x_{2i+2})^2$$

$$+ (x_{2i} - 2x_{2i+1})^4 + 10(x_{2i-1} - x_{2i+2})^4 \Big],$$

where $x^* = (0, \ldots, 0)^T$ and $H_9(x^*) = -1$.

(10) Levy function ($n = 100, \ -10 \leq x_i \leq 10, \ i = 1 \ldots, n$)

$$H_{10}(x) = -10 \sin^2(\pi x_1) - \sum_{i=1}^{n-1} 100 x_i^2 (1 + 10 \sin^2(\pi x_{i+1})) - 100(x_n - 1)^2 - 1,$$

where $x^* = (0, \ldots, 0, 1)^T$, $H_{10}(x^*) = -1$.

Note that for each problem the domain is constrained in a hyperrectangle, i.e, for all $x = (x_1, \ldots, x_n)^T \in \mathbb{X}$, $l_i \leq x_i \leq u_i$ for all i, where $l_i \leq u_i$ are the respective lower and upper bounds of the ith component x_i.

For MARS and MRAS in continuous problems, we use independent multivariate normal distributions as the parameterized sampling distributions whose density function in iteration k is:

$$f_{\widehat{\theta}_k}(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(\sigma_k^i)^2}} \exp\Big(- \frac{(x_i - \mu_k^i)^2}{2(\sigma_k^i)^2} \Big).$$

It is easy to see that in Step 2 of MARS, the new parameters could be solved analytically:

$$\mu_{k+1}^i = \alpha_k \frac{\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\widehat{\theta}_k}(x) x}{\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\widehat{\theta}_k}(x)} + (1-\alpha_k)\mu_k^i$$

$$(\sigma_{k+1}^i)^2 = \alpha_k \frac{\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\widehat{\theta}_k}(x)(x-\mu_{k+1}^i)^2}{\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\widehat{\theta}_k}(x)} + (1-\alpha_k)\big((\sigma_k^i)^2 + (\mu_{k+1}^i - \mu_k^i)^2\big),$$

for all $i = 1, \ldots, n$.

To determine the Boltzmann reference distributions, we choose two standard annealing schedules for their simplicity:

- polynomial schedule (PS): $T_k = T^* + |H(x_k^*)|/(1 + k^{0.6})$;
- logarithmic schedule (LS): $T_k = T^* + 0.1|H(x_k^*)|/\log(1+k)$,

where x_k^* denotes the current best solution found so far at the kth iteration. Note that $|H(x_k^*)|$ is brought to the schedule to roughly counterbalance the effect of the magnitude of $H(x)$ in the term $e^{H(x)/T_k}$. This is useful for those badly-scaled objective functions where the Boltzmann distribution is very sensitive to the topology of the functions, in which case the sequence of the Boltzmann distributions will vary too fast. Note that we use $|H(x_k^*)|$ to systematically achieve this goal since the objective function value is unknown a priori. In both PS and LS, we set the limiting temperature $T^* = 10^{-5}$ which is considered sufficiently small to achieve the desired accuracy and could prevent the instability in parameter updating, which may happen when the distribution is very closed to be degenerated. Our preliminary result shows that, T^* also works well between 10^{-1} and 10^{-5}. In addition, it is easy to verify that both PS and LS satisfy condition C7(a).

After the annealing schedule is determined, we empirically found that the performance of MARS is also sensitive to the choice of the gain sequence $\{\alpha_k\}$, while it does not depend too much on the choices of $\{N_k\}$ and $\{\lambda_k\}$. Similar to the parameter settings and their discussions for Algorithm 4.2 in Section 4.4, for all our test problems the parameters are set as follows:

- *Gain sequence:* $\alpha_k = 1/(k+100)^{0.501}$,
- *Sample size:* $N_k = \max\{N_0, \lfloor k^{0.502} \rfloor\}$,
- *Sample allocation rule:* $\lambda_k = 1/(k+1)^{0.5}$,

  where $\lfloor a \rfloor$ is the floor function which returns the largest integer that is no greater than a, and the initial sample size N_0 is set to 10 to keep a low computational effort at initial iterations.

On the other hand, SAN and HAS both use the Boltzmann distribution with the proposed annealing schedules as a selection mechanism to accept/reject candidate solutions sequentially. In SAN, as we described in the

Table 5.1: Performance of MARS, HAS, and SAN on test problems $H_1 - H_{10}$ with polynomial schedule, based on 50 independent replications (standard errors in parentheses).

Test.	MARS	HAS	SAN
Prob.	PS	PS	PS
H_1	-5.19e-2 (0.04)	-3.77 (0.82)	-2.89 (0.66)
H_2	-2.68e-6 (1.4e-8)	-3.66e-2 (0.02)	-3.86e-2 (0.02)
H_3	-4.13e-2 (0.02)	-3.17e-4 (5.1e-5)	-1.87 (0.31)
H_4	-6.02e-2 (0.03)	-3.19e+2 (10.82)	-5.15e+2 (12.49)
H_5	-1.00 (8.2e-5)	-1.41e+4 (430.6)	-1.56e+4 (9.40e+2)
H_6	-1.00 (2.1e-6)	-1.26e+2 (8.48)	-4.11e+3 (1.07e+2)
H_7	-6.4e-9 (7.5e-11)	-7.57e-2 (1.3e-3)	-8.67e-1 (1.3e-2)
H_8	-1.00 (6.9e-7)	-9.40e+2 (21.64)	-8.04e+2 (13.05)
H_9	-1.00 (4.6e-4)	-1.01e+2 (5.78)	-5.30e+3 (3.04e+2)
H_{10}	-1.00 (4.6e-7)	-3.34e+5 (9.41e+3)	-1.17e+6 (5.8e+4)

literature review, it performs a local search on the neighborhood of the current state. For the test problems, our preliminary results suggest a neighborhood structure as $\mathcal{N}(x) = \{y \in \mathbb{X} : \max_{1\leq i\leq n} |x_i - y_i| \leq \frac{1}{20}|u_i - l_i|\}$ that empirically generates good practical performance. In contrast, HAS samples globally from the entire solution space according to an underlying Markov chain that asymptotically approximates the Boltzmann distributions. When implementing HAS in our test problems, we use a hyperspherical direction to implement the Markov chain sampler (see [76] for a detailed discussion).

For each comparison algorithm, the average value of the best estimated for $H(x^*)$ of the 50 replication runs and its standard error is shown in Table 5.1 and 5.2. Moreover, Figure 5.1 and Figure 5.2 compare the performances

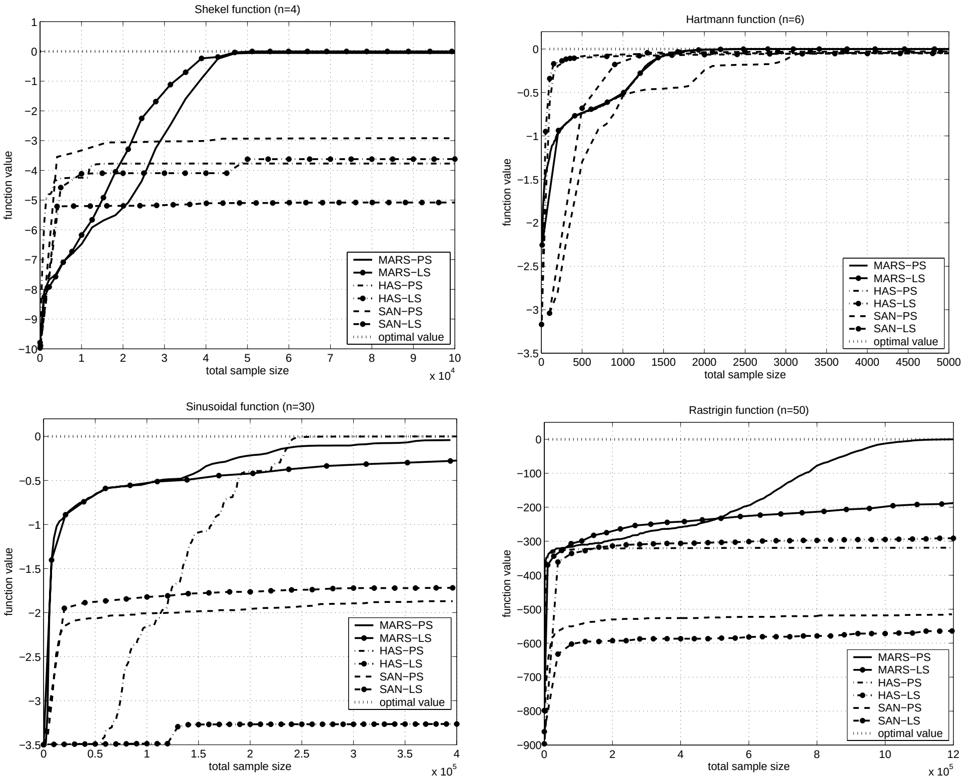
Table 5.2: Performance of MARS, HAS, and SAN on test problems $H_1 - H_{10}$ with logarithmic schedule, based on 50 independent replications (standard errors in parentheses).

Test.	MARS	HAS	SAN
Prob.	LS	LS	LS
H_1	-3.2e-7 (1.8e-12)	-3.62 (0.78)	-5.05 (0.81)
H_2	-2.0e-6 (1e-11)	-4.52e-2 (0.02)	-2.24e-2 (0.02)
H_3	-2.86e-1 (0.04)	-3.26 (0.22)	-1.72 (0.45)
H_4	-1.86e+2 (3.23)	-2.91e+2 (3.87)	-5.87e+2 (17.73)
H_5	-1.00 (3.3e-5)	-8.58e+3 (1.76e+2)	-1.48e+4 (8.06e+2)
H_6	-1.00 (2.3e-6)	-3.87e+2 (21.7)	-3.68e+3 (39.8)
H_7	-6.4e-9 (6.7e-11)	-1.27 (4.7e-3)	-7.94e-1 (7.75e-3)
H_8	-1.00 (2.1e-6)	-8.02e+2 (14.67)	-7.79e+2 (8.63)
H_9	-1.00 (3.8e-4)	-2.77e+1 (1.15)	-4.68e+3 (2.85e+2)
H_{10}	-1.00 (4.3e-2)	-4.27e+5 (9.91e+3)	-9.94e+5 (3.39e+4)

of all the algorithms by plotting the average current best values of H as a function of the number of samples that has been generated so far.

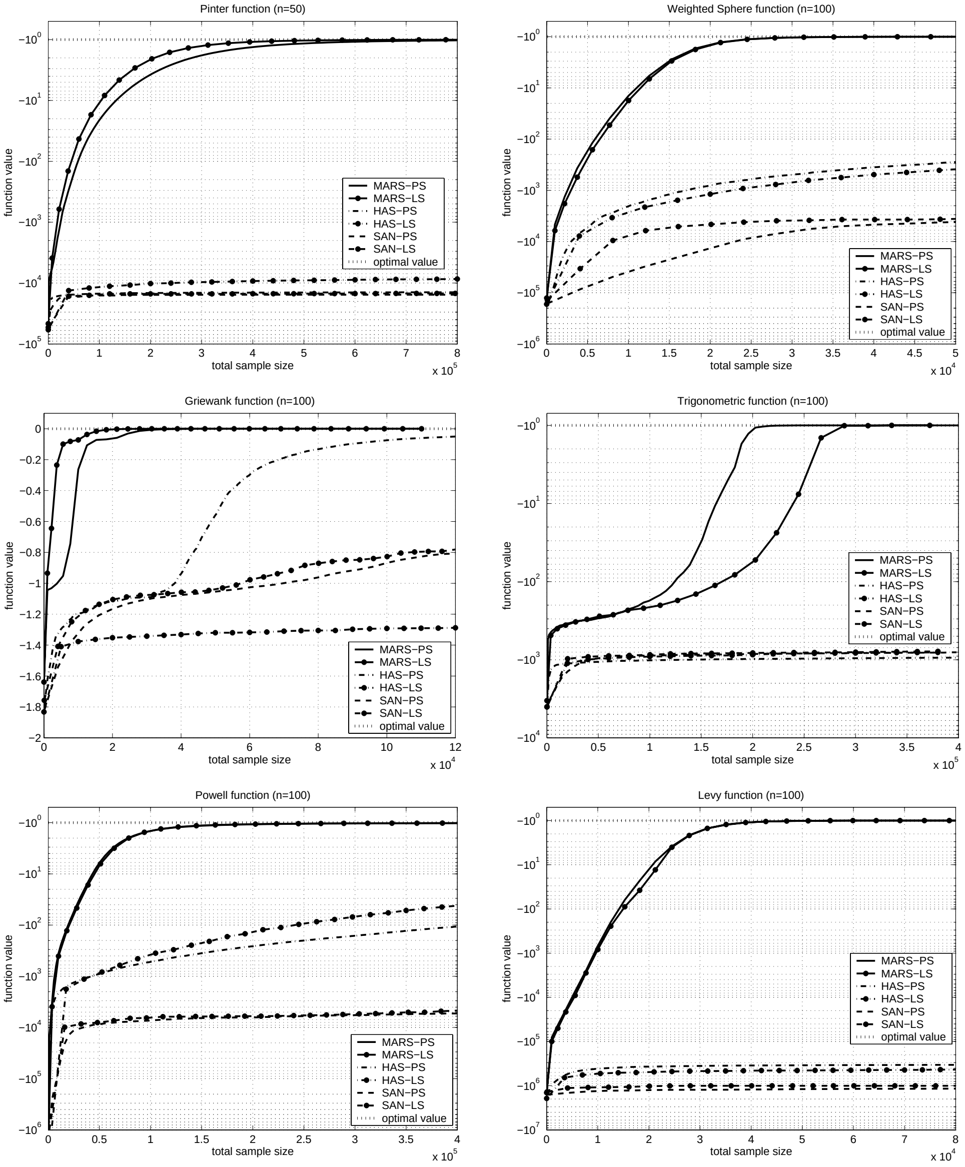
Our results indicate that MARS yields reasonably good performance with both annealing schedules in all benchmark problems. It can be seen that MARS significantly outperforms other algorithms for most high-dimensional problems. In particular, MARS-PS finds more than 90% of the ε-optimal solutions in cases H_1, H_3, and H_4, and finds ε-optimal solutions in all 50 runs in the rest seven test cases, where the specific precision $\varepsilon = 10^{-3}$. On the other hand, MARS-LS has similar performances as MARS-PS except in H_3 and H_4. A possible explanation of this behavior is that given the parameter setting, the performance of the algorithms more or less depends on the

Figure 5.1: Average Performance of MARS, HAS, and SAN on test functions H_1 to H_4.



topology of test functions. H_3 and H_4 is highly multimodal that contains a large number of local optima whose value is closed to the global optimum. Therefore a relatively slow annealing schedule may cause a slow convergence of the underlying Boltzmann distribution, which further slows down the algorithm. Preliminarily, we have extended the experiment of MARS-LS on H_4. We found that when the number of function evaluations increases to 10^7, MARS-LS could find the ε-optimal in 80% of the replication runs. Intuitively speaking, for problems in which the function values of local optima are very

Figure 5.2: Average Performance of MARS, HAS, and SAN on test functions H_5 to H_{10}.


Pinter function (n=50)

Weighted Sphere function (n=100)

Griewank function (n=100)

Trigonometric function (n=100)

Powell function (n=100)

Levy function (n=100)

closed to the global optimum, we may need a relatively faster cooling schedule to accelerate the algorithm; on the other hand, we may need a relatively slower cooling schedule to prevent the underlying Boltzmann distribution to vary too fast, and this is consistent with the idea of choosing a $|H(x_k^*)|$ in the annealing schedule. In contrast, SAN shows adequate performance on H_1 and H_2, but quickly becomes far less competitive on high dimensional problems. This is mainly because the size of the neighborhood $\mathcal{N}(x)$ to be searched at each iteration grows exponentially in the problem dimensions. The performance of HAS may be improved by careful selection of annealing schedules, which are adaptively determined so that it could have a high probability to visit a improving solution, see [57, 58, 61, 76]. However, the performance of adaptive annealing schedule often varies in practice. The reason is that the appropriate annealing schedule often depends on specific problem structures, e.g., convex quadratic function, and often assumes that the samples could be generated exactly from the Boltzmann distribution.

Discrete Optimization

We campare MARS, HAS, SAN and MRAS for discrete problems. To do so, we consider discretized versions of test functions H_1, H_3, H_4, H_6, H_8, and H_{10}. In each problem, we use the same domain constraint as their continuous version but evenly discretize the region by the same mesh size $h = 0.5$. As a result, the global optimum for H_8 becomes -9.2985, while the global optimum for other problems remain unchanged. Empirically speaking, the global optimization for discrete problems is harder than that for continuous problems, therefore we have reduced the dimension for the test functions.

To be specific, we take $n = 10$ for H_3 and $n = 50$ for problem H_4, H_6, H_8 and H_{10}. After the discretization, problem H_1 has $21^4 = 194,481$ feasible solutions, H_3 has $361^10 \approx 3.76 \times 10^{25}$ feasible solutions, while other problems have $41^{50} \approx 4.36 \times 10^{80}$ feasible solutions each. Although these problems have a large amount of feasible solutions and do not have differentiable structure, MARS still shows outstanding performance, since it implicitly does the gradient search in a parameter space with smooth structures.

In both MARS and MRAS, we generate candidate solutions by an $n \times m$ stochastic matrix Q, whereas the (i,j)th entry $q^{(i,j)}$ denotes the probability that x_i takes the jth value in the discretized set $\mathbf{X}_i := \{l_i + \frac{u_i - l_i}{m-1}(j-1),\ j = 1, \ldots, m\}$. In iteration k, step 2 in MARS updates the parameters as

$$q_{k+1}^{(i,j)} = \alpha_k \frac{\sum_{x \in \Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\widehat{\theta}_k}(x) I\{x \in \mathbf{x}_{i,j}\}}{\sum_{x \in \Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\widehat{\theta}_k}(x)} + (1 - \alpha_k) q_k^{(i,j)}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where $\mathbf{x}_{i,j}$ is the set of solutions in $\mathbb{X}$ whose ith component takes the jth value in $\mathbf{X}_i$, and $\widehat{f}_{\widehat{\theta}_k}(x) = (1 - \lambda_k) \prod_{i=1}^n \prod_{j=1}^m (q_k^{(i,j)})^{I\{x \in \mathbf{x}_{i,j}\}} + \lambda_k \prod_{i=1}^n \prod_{j=1}^m (q_0^{(i,j)})^{I\{x \in \mathbf{x}_{i,j}\}}$. The parameter updating for MRAS is in a similar manner. For the discrete problems, the initial sampling matrix Q_0 is set to the uniform stochastic matrix. In MRAS, the initial mean for each dimension is uniformly drawn between the upper and lower bounds, while the initial variances are set to 100, which is considered large enough for covering the entire region, and all other parameter settings can be found in [32]. The algorithm settings for both HAS and SAN

Table 5.3: Performance of MARS, HAS, and SAN with polynomial schedule and performance of MRAS on discrete test problems H_1, H_3, H_4, H_6, H_8, and H_{10}, based on 50 independent replications (standard errors in parentheses).

Test.	MARS	HAS	SAN	MRAS
Prob.	PS	PS	PS	N/A
H_1	-1.02 (0.34)	-3.58 (0.42)	-1.72 (0.34)	-1.21(0.31)
H_3	-3.3e-4 (1.3e-4)	-2.91e-4 (7.9e-5)	-0.34 (0.12)	-0.73(0.03)
H_4	-0.00 (0.00)	-5.93e+2 (19.2)	-5.41e+2 (11.6)	-3.11e+2(5.69)
H_6	-1.00 (0.00)	-7.36 (1.32)	-3.82e+2 (9.58)	-13.2(1.25)
H_8	-9.298 (1.2e-15)	-2.58e+2 (6.90)	-3.72e+2 (5.19)	-25.4(1.28)
H_{10}	-1.00 (0.00)	-1.44e+5 (5.52e+3)	-3.46e+5 (1.42e+4)	-2.57e+2(25.2)

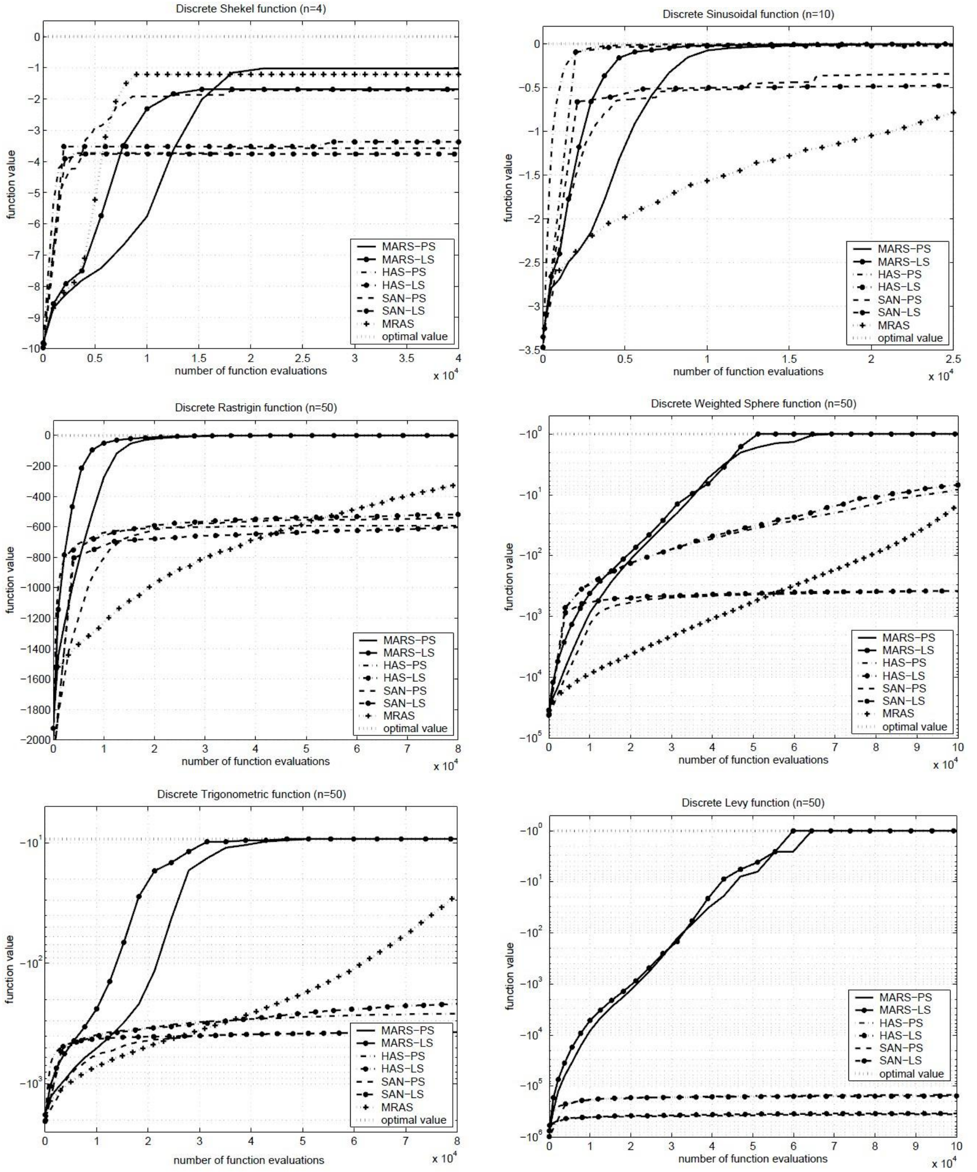
Table 5.4: Performance of MARS, HAS, and SAN with logarithmic schedule on discrete test problems H_1, H_3, H_4, H_6, H_8, and H_{10}, based on 50 independent replications (standard errors in parentheses).

Test.	MARS	HAS	SAN
Prob.	LS	LS	LS
H_1	-1.69 (0.43)	-3.37 (0.41)	-3.76 (0.46)
H_3	-1.31e-4 (8.9e-5)	-2.28e-2 (0.02)	-0.48 (0.15)
H_4	-0.00 (0.00)	-5.12e+2 (14.7)	-5.91e+2 (16.7)
H_6	-1.00 (0.00)	-6.22 (1.55)	-3.78e+2 (9.65)
H_8	-9.298 (0.00)	-2.15e+2 (7.25)	-3.67e+2 (7.41)
H_{10}	-1.00 (0.00)	-1.53e+5 (5.38e+3)	-3.40e+5 (1.18e+4)

are similar to the continuous case.

As in the continuous case, for each problem, we performed 50 independent replication runs for MARS, HAS, SAN and MRAS. The average value of the 50 sampled-optimal solutions and its stand error are reported in Table 5.3 and Table 5.4. Also, we plot in Figure 5.3 the function values (averaged over 50 runs) of the current best solution as a function of the number of samples generated so far.

Figure 5.3: Average Performance of MARS, HAS, and SAN on discretized test functions H_1, H_3, H_4, H_6, H_8 and H_{10}.



The average performance of the four algorithms is summarized in Table 5.3 and Table 5.4 In high dimensional cases, Both MARS-PS and MARS-LS consistently find the global optimum in all runs while showing a superior performance over SAN, HAS and MRAS, which is similar to the continuous case. However, we see that for low dimensional problems H_1 and H_3, MARS shows a slower improvement in early stage, but outperforms SAN and HAS on H_1, and has equal performance with HAS (while still outperforms SAN) in the later stage of $H3$. We conjecture this is due to the topology of the objective functions. After the discretization, the optimal function value $H(x^*)$ is "isolated" within the neighborhood of x^* (like a needle on the haystack), i.e., the neighbors of the global optimum x^* does not provide too much information on where x^* is located. Since MARS is a randomized algorithm, the chance for grasping the "needle" is relatively small. On the other hand, since SAN does the local enhancement search, for low dimensional problems it could quickly find the global optimum once it has been in its neighborhood.

Chapter 6

Conclusions

We have proposed a novel framework to study a class of model-based optimization algorithms by exploiting their connections to the stochastic approximation procedure. Through this connection, we proved the convergence for the CE method (modified version) and analyzed its convergence rate. At the same time, our numerical examples indicate that, our modified version of CE whose parameter updating procedure is based on the SA interpretation may have improved practical performance over the standard CE method. Moreover, inspired by CE and AAS, we proposed a novel model-based search algorithm called MARS for solving global optimization problems. As in CE, by studying the properties of Boltzmann distribution and the connection between MARS and SA, we proved the global convergence for MARS and provided the asymptotic convergence rate. In addition, MARS also shows a promising practical performance for both continuous and discrete optimization problems.

Our analysis provides new insights into the model-based algorithms, while generalizing SA procedure to problems that are not differentiable or continuous. By the SA interpretation, CE and MARS implicitly transform a target optimization problem into a counterpart optimization problem on a parameter space with smooth structures (i.e., continuous and differentiable). This may explain the fact that, for many high-dimensional multi-extremal optimization problems, model-based algorithms achieve superior performance over some of the existing algorithms.

Bibliography

[1] P. A. Absil and K. Kurdyka, "On the stable equilibrium points of gradient systems," *System & Control Letters,* vol. 55, pp. 573-577, 2006.

[2] G. Allon, D. P. Kroese, T. Raviv, and R. Y. Rubinstein, "Application of the cross-entropy method to the buffer allocation problem in a simulation-based environment," *Annals of Operations Research,* vol. 134, pp. 137-151, 2005.

[3] S.Andradóttir and A. A. Prudius, "Adaptive random search for continuous simulation optimization," *Naval Research Logistics,* vol. 57, pp. 583-604, 2010.

[4] C. J. P. Bélisle, H. E. Romeijn, and R. L. Smith, "Hit-and-Run Algorithms for Generating Multivariate Distributions", *Mathematics of Operations Research,* vol. 18, pp. 255-266, 1993.

[5] M. Benaim, "A dynamical system approach to stochastic approximations," *SIAM Journal on Control and Optimization,* vol. 34, pp. 437-472, 1996.

[6] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive algorithms and stochastic approximation,* Springer Verlag, Berlin - New York, 1990.

[7] H. -G. Beyer and H. -P. Schwefel, "Evolution Strategies: A Comprehensive Introduction," *Journal Natural Computing,* vol. 1, pp. 3-52. 2002.

[8] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint,* Cambridge University Press, New Delhi: Hindustan Book Agency, 2008.

[9] S. H. Brooks, “A discussion of random methods for seeking maxima,” *Operations Research,* vol.6, pp.244-251, 1958.

[10] Y. Cai, Y. Sun, X., and P. Jia, “Probabilistic Modeling for Continuous EDA with Boltzmann Selection and Kullback-Leibeler Divergence.” *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*, Seattle, WA, 2006, pp. 389-396.

[11] A. Costa, O. D. Jones, and D. Kroese, “Convergence properties of the Cross-Entropy method for discrete optimization,” *Operations Research Letters,* vol. 35, pp. 573-580, 2007.

[12] F. Dambreville, “Cross-Entropic learning of a machine for the decision in a partially observable universe,” *Journal of Global Optimization,* vol. 37, pp. 541-555, 2007.

[13] M. Dorigo and C. Blum, “Ant colony optimization theory: a survey,” *Theoretical Computer Science,* vol. 344, pp. 243-278, 2005.

[14] M. Dorigo and L. M. Gambardella, “Ant colony system: a cooperative learning approach to the traveling salesman problem,” *IEEE Transactions on Evolutionary Computation,* vol.1, pp. 53-66, 1997.

[15] A. Dukkipati, M. N. Murty, and S. Bhatnagar, “Cauchy Annealing Schedule: An Annealing Schedule for Boltzmann Selection Scheme in Evolutionary Algorithms,” *Proceedings of the 2004 Congress on Evolutionary Computation,* CEC 2004, June 19-23, 2004, Portland OR, USA.

[16] S.N. Evans and N.C. Weber, “On the Almost Sure Convergence of A General Stochastic Approximation Procedure,” *Bulletin of the Australian Mathematical Society,* vol. 34, pp. 335-342, 1986.

[17] V. Fabian, “On asymptotic normality in stochastic approximation,” *The Annals of Mathematical Statistics,* vol. 39, pp. 1327-1332, 1968.

[18] L. J. Fogel, A. J. Owens and M. J. Walsh, *Artificial Intelligence through Simulated Evolution,* John Wiley, 1966.

[19] M. C. Fu, "Gradient Estimation," *Handbook in OR & MS,* chapter 19, 2006.

[20] M. C. Fu, J. Hu, and S. I. Marcus, "Model-based Randomized Methods for Global Optimization," *Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems*, pp. 355-363, 2006.

[21] S. B. Gelfand and S. K. Mitter, "Simulated annealing with noisy or imprecise energy measurements," *Journal of Optimization Theory and Applications*, vol. 62, pp. 49-62, 1989.

[22] F. W. Glover, "Tabu Search - Part I", *ORSA Journal on Computing*, vol. 1, pp. 190-206, 1989.

[23] F. W. Glover, "Tabu Search - Part II", *ORSA Journal on Computing*, vol. 2, pp. 4-32, 1990.

[24] F. W. Glover, "Tabu Search: a Tutorial," *Interfaces*, vol.20, pp. 74-94, 1990.

[25] D. E. Goldberg, *Genetic algorithms in search, optimization, and machine learning,* Kluwer Academic Publishers, Boston, MA, 1989.

[26] W .Gong, Y. Ho and W. Zhai, "Stochastic Comparison Algorithm for Discrete Optimization with Estimation," *SIAM J. Optim.,* vol. 10, pp. 384-404, 1999.

[27] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association,* vol. 58, pp. 13-30, 1963.

[28] T. Homem-de-Mello, "Monto Carlo methods for discrete stochastic optimization," *Stochastic optimization: algorithms and applications,* S.Uryasev, P.M.Pardalos, eds. Kluwer Academic Publishers, Boston, MA.

[29] T. Homem-De-Mello, "A study on the Cross-Entropy method for rare event probability estimation," *INFORMS Journal on Computing,* vol. 19, pp. 381-394, 2007.

[30] T. Homem-De-Mello, “On rates of convergence for stochastic optimization problems under non-independent and identically distributed sampling,” *SIAM Journal on Optimization,* vol. 19, pp. 524-551, 2008.

[31] L. J. Hong, B. L. Nelson, “Discrete Optimization via Simulation Using COMPASS,” *Operations Research,* vol. 54, pp115-129, 2006.

[32] J. Hu, M. C. Fu, and S. I. Marcus, “A model reference adaptive search algorithm for global optimization,” *Operations Research,* vol. 55, pp. 549-568, 2007.

[33] J. Hu, M. C. Fu, and S. I. Marcus, “A model reference adaptive search algorithm for stochastic optimization with applications to Markov decision processes,” *Proceedings of the 46th IEEE Conference on Decision and Control,* pp. 975-980, New Orleans, LA, USA, 2007.

[34] J. Hu and P. Hu, “On the performance of the Cross-Entropy method,” *Winter Simulation Conference*, pp. 459-468, 2009.

[35] J. Hu and P. Hu, “An approximation annealing search algorithm to global optimization and its connection to stochastic approximation,” *Winter Simulation Conference*, pp. 1223-1234, 2010.

[36] J. Hu and P. Hu, “Annealing adaptive search, cross-entropy, and stochastic approximation in global optimization,” *Naval Research Logistics*, vol. 58, pp. 457-477, 2011.

[37] J. Hu, P. Hu and H. S. Chang, “A stochastic approximation framework for a class of randomized optimization algorithms,” *IEEE Transaction on Automatic Control,* vol. 57, pp. 165-178, 2012.

[38] A. W. Johnson and S. H. Jacobson, “A class of convergent generalized hill climbing algorithms,” *Applied Mathematics and Computation,* vol. 125, pp. 359-373, 2002.

[39] J. Kiefer and J. Wolfowitz, “Stochastic Estimation of the Maximum of a Regression Function,” *The Annals of Mathematical Statistics,* vol. 23, pp462-466, 1953.

[40] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671-680, 1983.

[41] A. Kleywegt, A. Shapiro and T. Homem-de-Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM J. Optim.* vol. 12, pp.479-502, 2011.

[42] D. P. Kroese, R. Y. Rubinstein, and T. Taimre, "Application of the Cross-Entropy method to clustering and vector quantization," *Journal of Global Optimization,* vol. 37, pp. 137-157, 2007.

[43] H. J. Kushner and D. S. Clark, *Stochastic approximation methods for constrained and unconstrained systems,* Springer-Verlag, New York, NY, 1978.

[44] H. J. Kushner and G. G. Yin, *Stochastic approximation algorithms and applications,* Springer-Verlag, New York, NY, 1997.

[45] P. J. M. Laarhoven and E. H. L. Aarts, *Simulated Annealing: Theory and Applications,* Kluwer Academic Publisher, Norwell, MA, 1987.

[46] M. Laguna and R. Marti, "Experimental testing of advanced scatter search designs for global optimization of multimodal functions," *Journal of Global Optimization,* vol. 33, pp. 235-255, 2005.

[47] P. Larrañaga and J. A. Lozano (Eds.) *Estimation of distribution algorithms: a new tool for evolutionary computation,* Kluwer Academic Publisher, Boston, MA, 2002.

[48] F. Liang, "Annealing stochastic approximation Monte Carlo for neural network training," *Machine Learning,* vol. 68, pp. 201-233, 2007.

[49] F. Liang, C. Liu, and R. J. Carroll, "Stochastic approximation in Monte Carlo computation," *Journal of the American Statistical Association,* vol. 102, pp. 305-320, 2007.

[50] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Transactions on Automatic Control*, vol. 22, pp. 551-575, 1977.

[51] S. Mannor, R. Y. Rubinstein, and Y. Gat, "The Cross-Entropy method for fast policy search," *Proceedings of the 20th International Conference on Machine Learning,* pp. 512-519, Washington, D.C., USA, 2003.

[52] J. L. Maryak and D. C. Chin, "Global random optimization by simultaneous perturbation stochastic approximation," *Proceedings of the American Control Conference,* Arlington, VA, 2001, pp. 756-762, 2001.

[53] C. N. Morris, "Natural exponential families with quadratic variance functions," *Annals of Statistics,* vol.10, pp. 65-80, 1982.

[54] H. Mühlenbein and G. Paaß, "From recombination of genes to the estimation of distributions: I. binary parameters," In Hans-Michael Voigt, Werner Ebeling, Ingo Rechenberg, and Hans-Paul Schwefel, editors, *Parallel Problem Solving from Nature - PPSN IV*, Springer, pp. 178-187, 1996.

[55] J. D. Pintér, *Global Optimization in Action.* Kluwer Academic Publisher, The Netherlands, 1996.

[56] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol.22, pp. 400-407, 1951.

[57] H. E. Romeijn and R. L. Smith, "Simulated Annealing and Adaptive Search in Global Optimization. *Probability in the Engineering and Informational Sciences* (8), pp. 571-590, 1994.

[58] H. E. Romeijn and R. L. Smith, Simulated Annealing for Constrained Global Optimization. *Journal of Global Optimization* (5) 1994, 101-126.

[59] R. Y. Rubinstein, "Optimization of computer simulation models with rare events," *European Journal of Operational Research,* vol. 99, pp. 89-112, 1997.

[60] R. Y. Rubinstein and D.P. Kroese, *The Cross-Entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning,* Springer, 2004.

[61] Y. Shen, S. Kiatsupaibul, Z. B. Zabinsky, and R. L. Smith, "An Analytically Derived Cooling Schedule for Simulated Annealing," *Journal of Global Optimization* (38) 2007, 333-365.

[62] L. Shi and S. Ólafsson, "Nested partitions method for global optimization," *Operations Research,* vol. 48, pp. 390-40, 2000.

[63] L. Shi and S. Ólafsson, "Nested Partitions Method for Stochastic Optimization," *Methodology and Computing in Applied Probability,* vol 2, pp. 271-291, 2000.

[64] A. Shiryaev, *Probability theory.* Springer-Verlag, New York, 1996.

[65] D. Sigalov and N. Shimkin, "Cross-Entropy based data association for multi-target tracking," *Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools,* Ariticle No. 30, 2008.

[66] J. C. Spall, "Multivariate stochastic approximation using simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control,* vol. 37, pp. 332-341, 1992.

[67] J. C. Spall, *Introduction to stochastic search and optimization,* John Wiley & Sons, 2003

[68] J. C. Spall and J. A. Cristion, "Model-Free Control of Nonlinear Stochastic Systems with Discrete-Time Measurements," *IEEE Transaction on Automatic Control*, vol. 43, pp. 1198-1210, 1998.

[69] R. Storn and K. Price, "Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, pp. 341-359, 1997.

[70] T. Stützle and M. Dorigo, "A short convergence proof for a class of ant colony optimization algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 6, pp. 358-365, 2002.

[71] F. Topsoe, "Bounds for Entropy and Divergence for Distributions over a Two-Element Set," *Journal of Inequalities in Pure and Applied Mathematics* (2) 2001, Article 25.

[72] D. Yan and H.Mukai, "Stochastic discrete optimization," *SIAM J. Control and Optimization*, vol. 30, No.3, pp. 594-612, May 1992.

[73] X. Yao and Y. Liu, "Fast Evolutionary Programming," *Proceedings of the 5th Annual Conference on Evolutionary Programming*, MIT Press, 1996, pp. 451-460.

[74] G. G. Yin, "Rates of convergence for a class of global stochatic optimization algorithms," *SIAM Journal on Optimization,* vol. 10, pp. 99-120, 1999.

[75] D. H. Wolpert, "Finding bounded rational equilibria part I: iterative focusing," *Proceedings of the Eleventh International Symposium on Dynamic Games and Applications.* T. Vincent (Ed.) Tucson Arizona, December 18-21, 2004.

[76] Z. B. Zabinsky, *Stochastic adaptive search for global optimization,* Kluwer Academic Publishers, 2003.

[77] Z. B. Zabinsky and R. L. Smith, "Pure Adaptive Search in Global Optimization," *Mathematical Programming* (53) 1992, 323-338.

[78] Z. B. Zabinsky, R.L. Smith, J.F. McDonald, H.E. Romeijn, and D.E. Kaufman, "Improving Hit-and-Run for Global Optimization," *Journal of Global Optimization* (3) 1993, 171-192.

[79] H. Zhang and M. C. Fu, "Applying model reference adaptive search to american-style option pricing," *Proceedings of the 38th Winter Simulation Conference*, pp. 711-718, Monterey, CA, USA, 2006.

[80] M. Zlochin, M. Birattari, N. Meuleau, and M. Dorigo. "Model-based search for combinatorial optimization: a critical survey," *Annals of Operations Research*, vol. 131, pp. 373-395, 2004.

Appendix

Proof of Lemma 4.2.1: Since $\inf_y \varphi(y) > 0$, there exists a constant $\zeta > 0$ such that $\varphi(y) \geq \zeta$ for all $y \in \Re$. We have

$$E_{\widehat{\theta}_k}[\varphi(H(X))I(H(X),\gamma_k)|\mathscr{F}_{k-1}] \geq \zeta E_{\widehat{\theta}_k}[I(H(X),\gamma_k)|\mathscr{F}_{k-1}]$$

$$\geq \zeta E_{\widehat{\theta}_k}[\mathscr{I}_{\{H(X)\geq\gamma_k\}}|\mathscr{F}_{k-1}] \geq \rho\zeta > 0, \quad (1)$$

where the second inequality follows from the definition of the threshold function $I(\cdot,\cdot)$, and the third equality follows from the definition of quantiles. Thus, it is sufficient to show that $\frac{1}{N_k}\sum_{x\in\Lambda_k}\varphi(H(x))I(H(x),\widehat{\gamma}_k)\Gamma(x) \to E_{\widehat{\theta}_k}[\varphi(H(X))I(H(X),\gamma_k)\Gamma(X)|\mathscr{F}_{k-1}]$ as $k \to \infty$ w.p.1. For notational convenience, define $\widehat{Y}_k(x) = \varphi(H(x))I(H(x),\widehat{\gamma}_k)$ and $Y_k(x) = \varphi(H(x))I(H(x),\gamma_k)$. Note that

$$\frac{1}{N_k}\sum_{x\in\Lambda_k}\widehat{Y}_k(x)\Gamma(x) - E_{\widehat{\theta}_k}[Y_k(X)\Gamma(X)|\mathscr{F}_{k-1}]$$
$$= \Big(\frac{1}{N_k}\sum_{x\in\Lambda_k}\widehat{Y}_k(x)\Gamma(x) - \frac{1}{N_k}\sum_{x\in\Lambda_k}Y_k(x)\Gamma(x)\Big)$$
$$+ \Big(\frac{1}{N_k}\sum_{x\in\Lambda_k}Y_k(x)\Gamma(x) - E_{\widehat{\theta}_k}[Y_k(X)\Gamma(X)|\mathscr{F}_{k-1}]\Big).$$

Since the mapping Γ is continuous and $\mathbb{X}$ is compact, it is clear that $\Gamma(x)$ is bounded for all $x \in \Lambda_k$. Thus, by Proposition 4.2.1 and the continuity of $I(\cdot,\cdot)$, the first term above vanishes to zero w.p.1.

To show that the second term also converges to zero, note that conditional on $\mathscr{F}_{k-1}$, $\{Y_k(x)\}_{x\in\Lambda_k}$ are i.i.d. and there exist constants a and b such that $a \leq Y_k(x)\Gamma_i(x) \leq b \;\forall\, x \in \mathbb{X}$, where $\Gamma_i(X)$ is the ith component of the vector $\Gamma(X)$. For any $\varepsilon > 0$, Let A_k be the event that $A_k = \big\{\big|\frac{1}{N_k}\sum_{x\in\Lambda_k}Y_k(x)\Gamma_i(x) -$

$E_{\widehat{\theta}_k}[Y_k(X)\Gamma_i(X)]\big| \geq \varepsilon\big\}$. We have from the Hoeffding inequality [27] that

$$P_{\widehat{\theta}_k}\big(A_k\big|\mathscr{F}_{k-1}\big) \leq 2\exp\Big(\frac{-2N_k\varepsilon^2}{(b-a)^2}\Big) = 2\exp\Big(\frac{-2Ck^\beta\varepsilon^2}{(b-a)^2}\Big),$$

for some constant $C > 0$. Next by conditioning, we get

$$P\big(A_k\big) = E\Big[P_{\widehat{\theta}_k}\big(A_k\big|\mathscr{F}_{k-1}\big)\Big] \leq 2\exp\Big(\frac{-2Ck^\beta\varepsilon^2}{(b-a)^2}\Big).$$

Moreover, we have from A2, $\sum_{k=0}^{\infty} P\big(A_k\big) \leq \sum_{k=0}^{\infty} 2\exp\Big(\frac{-2Ck^\beta\varepsilon^2}{(b-a)^2}\Big) < \infty$. Finally, the Borel-Cantelli lemma implies that $P(A_k \ i.o.) = 0$. Since this holds for arbitrary $\varepsilon > 0$, we have $\frac{1}{N_k}\sum_{x\in\Lambda_k} Y_k(x)\Gamma_i(x) \to E_{\widehat{\theta}_k}[Y_k(X)\Gamma_i(X)]$ w.p.1.

A similar argument can be used to show that $\frac{1}{N_k}\sum_{x\in\Lambda_k} Y_k(x) \to E_{\widehat{\theta}_k}[Y_k(X)]$ w.p.1. Therefore, by (1) we have $b_k(\widehat{\theta}_k) \to 0$ as $k \to \infty$ w.p.1. $\square$

Proof of Lemma 4.3.1: Our proof is based on the proof of Lemma 2.4 in [30]. Let H_l and H_u be the lower and upper bound for H. For given $f_{\widehat{\theta}_k}$ and $\rho \in (0,1)$, it can be shown that the $(1-\rho)$-quantile γ_k can be obtained as the optimal solution of the following problem (e.g., [29])

$$\min_{v\in\mathbb{V}} \ell_k(v), \tag{2}$$

where $\mathbb{V} = [H_l, H_u]$, $\ell_k(v) := E_{\widehat{\theta}_k}[\phi(H(X), v)|\mathscr{F}_{k-1}]$, and

$$\phi(H(x), v) := \begin{cases} (1-\rho)(H(x) - v) & \text{if } v \le H(x), \\ \rho(v - H(x)) & \text{if } v \ge H(x). \end{cases}$$

Similarly, the sample $(1-\rho)$-quantile $\widehat{\gamma}_k$ can be expressed as the solution to the sample average approximation of (2),

$$\min_{v \in \mathbb{V}} \widehat{\ell}_k(v), \tag{3}$$

where $\widehat{\ell}_k(v) := \frac{1}{N_k} \sum_{j=1}^{N_k} \phi(H(X_j^k), v)$ and $X_1^k, \ldots, X_{N_k}^k$ are i.i.d. with distribution function $f_{\widehat{\theta}_k}$.

For a given $\rho \in (0, 1)$ and a constant $\delta > 0$, define $r = \frac{\delta}{3 \max\{\rho, 1-\rho\}}$. Let $\{B_{v,r}, v \in \mathbb{V}\}$ be a collection of open balls centered at $v \in \mathbb{V}$ with radius r. Since $\mathbb{V}$ is compact, we can find a collection of finite points $C_v = \{v_1, \ldots, v_s\}$ such that $\mathbb{V} \subseteq \bigcup_{v \in C_v} B_{v,r}$. Moreover, for an arbitrary $v \in \mathbb{V}$, there exists $v_k \in C_v$ such that $|v - v_k| < r$. Thus, by the constructions of ℓ_k and $\widehat{\ell}_k$, we have

$$\begin{aligned} |\ell_k(v) - \ell_k(v_k)| &\le E_{\widehat{\theta}_k}\left[|\phi(H(X), v) - \phi(H(X), v_k)|\middle|\mathscr{F}_{k-1}\right] \\ &\le \max\{\rho, 1-\rho\}|v - v_k| \le \delta/3, \\ |\widehat{\ell}_k(v) - \widehat{\ell}_k(v_k)| &\le \frac{1}{N_k} \sum_{j=1}^{N_k} \left|\phi(H(X_j^k), v) - \phi(H(X_j^k), v_k)\right| \\ &\le \max\{\rho, 1-\rho\}|v - v_k| \le \delta/3. \end{aligned}$$

It follows that if $|\ell_k(v_k) - \widehat{\ell}_k(v_k)| < \delta/3$, then for all $v \in B_{v_k,r}$

$$\begin{aligned}|\ell_k(v) - \widehat{\ell}_k(v)| \leq |\ell_k(v) - \ell_k(v_k)| + |\widehat{\ell}_k(v) - \widehat{\ell}_k(v_k)| \\ + |\ell_k(v_k) - \widehat{\ell}_k(v_k)| < \delta.\end{aligned}$$

This implies that

$$P_{\widehat{\theta}_k}\big(|\ell_k(v)-\widehat{\ell}_k(v)| < \delta, \, \forall\, v \in B_{v_k,r}\big|\mathscr{F}_{k-1}\big) \geq P_{\widehat{\theta}_k}\big(\big|\ell_k(v_k)-\widehat{\ell}_k(v_k)\big| < \delta/3\big|\mathscr{F}_{k-1}\big).$$

Next, by using Bonferroni's inequality, we have

$$\begin{aligned}&P_{\widehat{\theta}_k}\Big(|\ell_k(v) - \widehat{\ell}_k(v)| < \delta, \ \forall\, v \in \mathbb{V}\Big|\mathscr{F}_{k-1}\Big) \\ &\geq P_{\widehat{\theta}_k}\Big(\big|\ell_k(v_k) - \widehat{\ell}_k(v_k)\big| < \delta/3, \ \forall\, v_k \in C_v\Big|\mathscr{F}_{k-1}\Big) \\ &\geq 1 - \sum_{k=1}^{s}\Big[1 - P_{\widehat{\theta}_k}\Big(\big|\ell_k(v_k) - \widehat{\ell}_k(v_k)\big| < \delta/3\Big|\mathscr{F}_{k-1}\Big)\Big] \\ &\geq 1 - \kappa(\delta) \max_{1\leq k\leq s} P_{\widehat{\theta}_k}\Big(\big|\ell_k(v_k) - \widehat{\ell}_k(v_k)\big| \geq \delta/3\Big|\mathscr{F}_{k-1}\Big) \qquad (4)\end{aligned}$$

where $\kappa(\delta) = \frac{3\max\{\rho,1-\rho\}(H_u-H_l)}{\delta}$. Thus, by noting that $0 \leq \phi(H(x), v) \leq \max\{\rho, 1-\rho\}(H_u-H_l) < H_u-H_l \,\forall\, v \in \mathbb{V}$ and applying Hoeffding's inequality [27] to the right-hand-size of (4), we get $P_{\widehat{\theta}_k}\big(|\ell_k(v) - \widehat{\ell}_k(v)| < \delta, \ \forall\, v \in \mathbb{V}\big|\mathscr{F}_{k-1}\big) \geq 1 - 2\kappa(\delta)\exp\big(\frac{-2N_k\delta^2}{9(H_u-H_l)^2}\big)$. Next, by unconditioning on $\mathscr{F}_{k-1}$, we have

$$P\big(|\ell_k(v) - \widehat{\ell}_k(v)| < \delta, \forall\, v \in \mathbb{V}\big) \geq 1 - 2\kappa(\delta)e^{\frac{-2N_k\delta^2}{9(H_u-H_l)^2}}. \qquad (5)$$

To complete the proof of Lemma 4.3.1, we need the following intermediate result, which states that if the two functions $\ell_k(v)$ and $\widehat{\ell}_k(v)$ are sufficiently close, then their optimal solutions will also be close.

Proposition .0.1. *Assume that A3 and B2 hold. There exists a constant*

$K > 0$ *such that*

$$K|\widehat{\gamma}_k - \gamma_k|^2 \leq \max_{v \in \mathbb{V}} |\ell_k(v) - \widehat{\ell}_k(v)| \tag{6}$$

almost surely for k sufficiently large.

Let D_k be the event that (6) holds at the kth iteration of Algorithm 4.2. We have for a sufficiently small $\epsilon > 0$,

$$\begin{aligned}
&P\big(\{|\widehat{\gamma}_k - \gamma_k| \geq \epsilon\} \cap D_k\big) \\
&\leq P\big(\{\max_{v \in \mathbb{V}} |\ell_k(v) - \widehat{\ell}_k(v)| \geq K\epsilon^2\} \cap D_k\big) \\
&\leq P\big(\max_{v \in \mathbb{V}} |\ell_k(v) - \widehat{\ell}_k(v)| \geq K\epsilon^2\big) \\
&= 1 - P(|\ell_k(v) - \widehat{\ell}_k(v)| < K\epsilon^2,\ \forall\, v \in \mathbb{V}) \\
&\leq \frac{A}{\epsilon^2} \exp(-BN_k\epsilon^4) \quad \text{ by (5),}
\end{aligned}$$

where $A = \frac{6\max\{\rho, 1-\rho\}(H_u - H_l)}{K}$ and $B = \frac{2K^2}{9(H_u - H_l)^2}$.

It follows that for a given $\tau > 0$,

$$\begin{aligned}
&P\big(\{k^{\frac{\tau}{2}}|\widehat{\gamma}_k - \gamma_k| \geq \epsilon\} \cap D_k\big) \\
&= P\big(\{|\widehat{\gamma}_k - \gamma_k| \geq k^{-\frac{\tau}{2}}\epsilon\} \cap D_k\big) \\
&\leq \frac{Ak^{\tau}}{\epsilon^2} \exp\big(-B\epsilon^4 k^{\beta - 2\tau}\big).
\end{aligned}$$

Since $\beta > 2\tau$, it is easy to verify that $\sum_{k=1}^{\infty} P\big(\{k^{\frac{\tau}{2}}|\widehat{\gamma}_k - \gamma_k| \geq \epsilon\} \cap D_k\big) \leq \sum_{k=1}^{\infty} \frac{Ak^{\tau}}{\epsilon^2} \exp\big(-B\epsilon^4 k^{\beta-2\tau}\big) < \infty$. Since $P(D_k\ i.o) = 1$, the Borel Cantelli lemma implies $P(k^{\frac{\tau}{2}}|\widehat{\gamma}_k - \gamma_k| \geq \epsilon\ i.o.) = P(k^{\frac{\tau}{2}}|\widehat{\gamma}_k - \gamma_k| \geq \epsilon \cap D_k\ i.o.) = 0$. Hence we have $k^{\frac{\tau}{2}}|\widehat{\gamma}_k - \gamma_k| \to 0$ as $k \to \infty$ w.p.1. □

Proof of Proposition .0.1: Define the difference $\Delta_k = \widehat{\gamma}_k - \gamma_k$, and let $Y = H(X)$ for notational convenience. Note that the function $\ell_k(v)$ is convex, and for a given $\rho \in (0,1)$, its subdifferential is given by $\partial_v \ell_k(v) = [\rho - P_{\widehat{\theta}_k}(Y \geq \gamma_k | \mathscr{F}_{k-1}), \rho - 1 + P_{\widehat{\theta}_k}(Y \leq \gamma_k | \mathscr{F}_{k-1})]$ (e.g., [29]). Before we proceed any further, we need to distinguish between the continuous and the discrete finite optimization cases.

Case 1: (Continuous optimization) It is easy to see that $\ell_k(v)$ is twice differentiable. Let $\bar{\zeta} > 0$ and $\bar{\delta} > 0$ be constants as defined in B2. Since $\widehat{\gamma}_k \to \gamma_k$ w.p.1 as $k \to \infty$ by Proposition 4.2.1, a Taylor expansion of $\ell_k(\widehat{\gamma}_k)$ in a small neighborhood $(\gamma_k - \bar{\delta}, \gamma_k + \bar{\delta})$ of γ_k implies that

$$\ell_k(\widehat{\gamma}_k) - \ell_k(\gamma_k) = \frac{1}{2}\ell_k^{''}(\bar{\gamma}_k)\Delta_k^2,$$

where $\bar{\gamma}_k$ lies on the line segment between γ_k and $\widehat{\gamma}_k$, and we have used the fact that $\ell_k^{'}(\gamma_k) = 0$ since γ_k is the optimal solution to the convex optimization problem (2). It is straightforward to verify that $\ell_k^{''}(\bar{\gamma}_k) = f_{\widehat{\theta}_k}^H(\bar{\gamma}_k)$. Thus, for almost every sample path generated by Algorithms 2, we have from B2 that for k sufficiently large,

$$\begin{aligned}
\frac{\bar{\zeta}}{2}\Delta_k^2 &\leq |\ell_k(\widehat{\gamma}_k) - \ell_k(\gamma_k)| \\
&\leq |\ell_k(\widehat{\gamma}_k) - \widehat{\ell}_k(\widehat{\gamma}_k)| + |\widehat{\ell}_k(\widehat{\gamma}_k) - \ell_k(\gamma_k)| \\
&\leq |\ell_k(\widehat{\gamma}_k) - \widehat{\ell}_k(\widehat{\gamma}_k)| + \max_{v \in \mathbb{V}} |\ell_k(v) - \widehat{\ell}_k(v)| \\
&\leq 2 \max_{v \in \mathbb{V}} |\ell_k(v) - \widehat{\ell}_k(v)|,
\end{aligned}$$

where the third inequality follows from the inequality $|\min_x u(x) - \min_x w(x)| \leq \max_x |u(x) - w(x)|$ for any two real-valued functions u and w. Consequently, it is clear that there exists a constant $K > 0$, such that $K\Delta_k^2 \leq \max_{v \in \mathbb{V}} |\ell_k(v) - \widehat{\ell}_k(v)|$ almost surely for all k sufficiently large.

Case 2: (Discrete finite optimization) Since the solution space $\mathbb{X}$ is finite, the function $\ell_k(v)$ is convex and piece-wise linear, and its subdifferential at γ_k can be written as $\partial_{\gamma_k}\ell_k(\gamma_k) = [\rho - P_{\widehat{\theta}_k}(Y \geq \gamma_k | \mathscr{F}_{k-1}), \rho - P_{\widehat{\theta}_k}(Y > \gamma_k | \mathscr{F}_{k-1})]$.

We have from part (ii) of B2 that for almost every sample path generated by Algorithm 4.2, $\rho - P_{\widehat{\theta}_k}(Y \geq \gamma_k|\mathscr{F}_{k-1}) \leq -\bar{\zeta}$ and $\rho - P_{\widehat{\theta}_k}(Y > \gamma_k|\mathscr{F}_{k-1}) \geq \bar{\zeta}$. For a sufficiently small Δ_k, by the definition of subderivatives, we have $\ell_k(\widehat{\gamma}_k) - \ell_k(\gamma_k) \geq C\Delta_k$ for any $C \in [-\bar{\zeta}, \bar{\zeta}] \subseteq \partial_{\gamma_k}\ell_k(\gamma_k)$. It follows that

$$\begin{aligned}\bar{\zeta}|\Delta_k| &\leq \ell_k(\widehat{\gamma}_k) - \ell_k(\gamma_k) \\ &\leq |\ell_k(\widehat{\gamma}_k) - \widehat{\ell}_k(\widehat{\gamma}_k)| + |\widehat{\ell}_k(\widehat{\gamma}_k) - \ell_k(\gamma_k)| \\ &\leq 2\max_{v\in\mathbb{V}}|\ell_k(v) - \widehat{\ell}_k(v)|.\end{aligned}$$

Hence, the desired result holds when $|\Delta_k|$ is sufficiently small. □

Proof of Proposition 4.3.1: Again, we define $\widehat{Y}_k(x) = \varphi(H(x))I(H(x), \widehat{\gamma}_k)$ and $Y_k(x) = \varphi(H(x))I(H(x), \gamma_k)$. By (1), it is sufficient to show that $k^{\frac{\tau}{2}}\big(\frac{1}{N_k}\sum_{x\in\Lambda_k}\widehat{Y}_k(x)\Gamma(x) - E_{\widehat{\theta}_k}[Y_k(X)\Gamma(X)|\mathscr{F}_{k-1}]\big) \to 0$ as $k \to \infty$ w.p.1. Note that

$$\begin{aligned}&k^{\frac{\tau}{2}}\Big(\frac{1}{N_k}\sum_{x\in\Lambda_k}\widehat{Y}_k(x)\Gamma(x) - E_{\widehat{\theta}_k}[Y_k(X)\Gamma(X)|\mathscr{F}_{k-1}]\Big) \\ =&k^{\frac{\tau}{2}}\Big(\frac{1}{N_k}\sum_{x\in\Lambda_k}\widehat{Y}_k(x)\Gamma(x) - \frac{1}{N_k}\sum_{x\in\Lambda_k}Y_k(x)\Gamma(x)\Big) \\ &+ k^{\frac{\tau}{2}}\Big(\frac{1}{N_k}\sum_{x\in\Lambda_k}Y_k(x)\Gamma(x) - E_{\widehat{\theta}_k}[Y_k(X)\Gamma(X)|\mathscr{F}_{k-1}]\Big).\end{aligned}$$

Thus, by Lemma 4.3.1 and the continuity of $I(\cdot,\cdot)$, the first term above converges to zero as $k \to \infty$ w.p.1. By using the same argument as in the proof of Lemma 4.2.1, it is easy to show that the second term also vanishes to zero as $k \to \infty$ w.p.1. □

Proof of Lemma 4.3.2: Recall that $V_k = k^{-\frac{\alpha}{2}+\frac{\tau}{2}}\lambda_k\big(\frac{1-\alpha_k}{\alpha_k}\big)\big(N_k^{-1}\sum_{x\in\Lambda_k}\Gamma(x) - m(\widehat{\theta}_k)\big)$. Note that conditional on $\mathscr{F}_{k-1}$, the solutions in Λ_k are i.i.d.. Thus it follows trivially that $E_{\widehat{\theta}_k}[V_k|\mathscr{F}_{k-1}] = 0$. To show the second claim, let $\Sigma_k = E_{\widehat{\theta}_k}[V_k V_k^T|\mathscr{F}_{k-1}]$. We have

$$\begin{aligned}
\Sigma_k &= E_{\widehat{\theta}_k}[V_k V_k^T|\mathscr{F}_{k-1}] \\
&= k^{-\alpha+\tau} E_{\widehat{\theta}_k}[\xi_k(\widehat{\theta}_k)\xi_k(\widehat{\theta}_k)^T|\mathscr{F}_{k-1}] \\
&= k^{-\alpha+\tau}\Big(\frac{1-\alpha_k}{\alpha_k}\Big)^2 \frac{\lambda_k^2}{N_k}\mathrm{Cov}_{\widehat{\theta}_k}[\Gamma(X)|\mathscr{F}_{k-1}] \\
&= k^{-\alpha-\beta+\tau-2\lambda}\Big(\frac{1-\alpha_k}{\alpha_k}\Big)^2 \mathrm{Cov}_{\widehat{\theta}_k}[\Gamma(X)|\mathscr{F}_{k-1}].
\end{aligned}$$

Since under A1, A3, and the choices of α_k and β_k, the sequence of sampling distributions $\{f_{\widehat{\theta}_k}\}$ converges point-wise to $f_{m^{-1}(\eta^*)}$ w.p.1, the dominated convergence theorem implies that the sequence $\{\Sigma_k\}$ converges w.p.1. to a limiting matrix Σ given by

$$\Sigma := \begin{cases} \mathrm{Cov}_{m^{-1}(\eta^*)}[\Gamma(X)] & \text{if } \beta = \alpha + \tau - 2\lambda, \\ 0 & \text{if } \beta > \alpha + \tau - 2\lambda. \end{cases}$$

By Hölder's inequality, for any $1 < p, q < \infty$ with $1/p + 1/q = 1$, we have

$$\begin{aligned}
&\lim_{k\to\infty} E[\mathscr{I}_{\{\|V_k\|^2 \geq rk^\alpha\}}\|V_k\|^2] \\
&\leq \limsup_{k\to\infty}\big[P(\|V_k\|^2 \geq rk^\alpha)\big]^{1/p}\big[E[\|V_k\|^{2q}]\big]^{1/q} \qquad (7)
\end{aligned}$$

Also,

$P(\|V_k\|^2 \geq rk^\alpha)$

$$
\begin{aligned}
&\leq P\Big(\Big(\frac{1-\alpha_k}{\alpha_k}\Big)^2\lambda_k^2\Big\|\frac{1}{N_k}\sum_{x\in\Lambda_k}\Gamma(x)-m(\widehat{\theta}_k)\Big\|^2 \geq rk^{2\alpha-\tau}\Big)\\
&\leq P\Big(\big\|N_k^{-1}\sum_{x\in\Lambda_k}\Gamma(x)-m(\widehat{\theta}_k)\big\| \geq C\sqrt{r}k^{-\tau/2+\lambda}\Big) \qquad \text{(for some constant } C>0\text{)}\\
&\leq \frac{E[\|N_k^{-1}\sum_{x\in\Lambda_k}\Gamma(x)-m(\widehat{\theta}_k)\|^2]}{C^2rk^{-\tau+2\lambda}} \qquad \text{(by Chebyshev's inequality)}\\
&\leq \frac{E\big[E_{\widehat{\theta}_k}[\|N_k^{-1}\sum_{x\in\Lambda_k}\Gamma(x)-m(\widehat{\theta}_k)\|^2\,|\mathscr{F}_{k-1}]\big]}{C^2rk^{-\tau+2\lambda}}\\
&= \frac{E\big[E_{\widehat{\theta}_k}[\Gamma(X)^T\Gamma(X)-m(\widehat{\theta}_k)^Tm(\widehat{\theta}_k)|\mathscr{F}_{k-1}]\big]}{C^2rk^{-\tau+2\lambda}N_k}\\
&= O(k^{\tau-\beta-2\lambda}). \qquad (8)
\end{aligned}
$$

By taking $q=2$, we have

$$
\begin{aligned}
&E[\|V_k\|^4] = E\big[(V_k^TV_k)^2\big] \leq O(k^{2(\alpha+\tau-2\lambda)})\times\\
&E\Big[\Big(\big(\frac{1}{N_k}\sum_{x\in\Lambda_k}\Gamma(x)-m(\widehat{\theta}_k)\big)^T\big(\frac{1}{N_k}\sum_{x\in\Lambda_k}\Gamma(x)-m(\widehat{\theta}_k)\big)\Big)^2\Big] \qquad (9)
\end{aligned}
$$

A straightforward calculation shows that the right-hand-size of (9) is on the order of $O\big(k^{2(\alpha+\tau-\beta-2\lambda)}\big)$. Thus, combining (8) and (9), the right-hand-side of (7) is on the order of $O(k^{\alpha+3(\tau-\beta-2\lambda)/2})$, which vanishes to zero as $k\to\infty$ by taking $\beta\geq\alpha+\tau-2\lambda$. This completes the proof of the lemma. □

Proof of Lemma 5.2.2: To simplify exposition, we focus on the ith components of $\mathbb{U}_k$ and $\bar{\mathbb{U}}_k$ ($i=1,\ldots,d$), and define

$$
U_k^i = \frac{1}{N_k}\sum_{x\in\Lambda_k}e^{\frac{H(x)}{T_{k+1}}}\Gamma_i(x)/\widehat{f}_{\widehat{\theta}_k}(x),\ \bar{U}_k^i = E_{\widehat{f}_{\widehat{\theta}_k}}[U_k^i|\mathscr{F}_k] = \int_{\mathbb{X}}e^{\frac{H(x)}{T_{k+1}}}\Gamma_i(x)\nu(dx),
$$

where $\Gamma_i(x)$ is the ith component of Γ. Denote by $\mathscr{V}$ the volume of $\mathbb{X}$. Note that since $H(x) > 0 \; \forall\, x$, we have $\bar{V}_k > \mathscr{V}$ (cf. Equation (5.8)). Moreover, by C5, since the mapping Γ is bounded on $\mathbb{X}$, there exist constants C_1 and C_2 such that $C_1 \leq \Gamma_i(x) \leq C_2 \; \forall\, x$ and $C_1 \bar{V}_k \leq \bar{U}_k^i \leq C_2 \bar{V}_k$. Define $\epsilon = \frac{\mathscr{V}}{2}$, and let $\Omega_k = \{|U_k^i - \bar{U}_k^i| < \epsilon \cap |V_k - \bar{V}_k| < \epsilon\}$ and Ω_k^c be the complement of Ω_k.

By using a second order two-variable Taylor expansion of $\frac{U_k^i}{V_k}$ around the neighborhood Ω_k of $(\bar{U}_k^i, \bar{V}_k)$, we can write, for every sample path generated by MARS,

$$
\begin{aligned}
\frac{U_k^i}{V_k} = \Bigg[& \frac{\bar{U}_k^i}{\bar{V}_k} - \frac{\bar{U}_k^i}{(\bar{V}_k)^2}(V_k - \bar{V}_k) + \frac{1}{\bar{V}_k}(U_k^i - \bar{U}_k^i) + \frac{\tilde{U}_k}{\tilde{V}_k^3}(V_k - \bar{V}_k)^2 \\
& - \frac{1}{\tilde{V}_k^2}(V_k - \bar{V}_k)(U_k^i - \bar{U}_k^i) \Bigg] I\{\Omega_k\} + \frac{U_k^i}{V_k} I\{\Omega_k^c\}, \qquad (10)
\end{aligned}
$$

where $\tilde{U}_k$ and $\tilde{V}_k$ are on the respective line segments from $\bar{U}_k^i$ to U_k^i and from $\bar{V}_k$ to V_k. By rearranging terms in (10), we have

$$
\begin{aligned}
\frac{U_k^i}{V_k} - \frac{\bar{U}_k^i}{\bar{V}_k} = & \frac{1}{\bar{V}_k}(U_k^i - \bar{U}_k^i) - \frac{\bar{U}_k^i}{(\bar{V}_k)^2}(V_k - \bar{V}_k) \\
& + \left[\frac{\tilde{U}_k}{\tilde{V}_k^3}(V_k - \bar{V}_k)^2 - \frac{1}{\tilde{V}_k^2}(V_k - \bar{V}_k)(U_k^i - \bar{U}_k^i) \right] I\{\Omega_k\} \qquad (11) \\
& + \left[\frac{U_k^i}{V_k} - \frac{\bar{U}_k^i}{\bar{V}_k} - \frac{1}{\bar{V}_k}(U_k^i - \bar{U}_k^i) + \frac{\bar{U}_k^i}{(\bar{V}_k)^2}(V_k - \bar{V}_k) \right] I\{\Omega_k^c\}.
\end{aligned}
$$

Next, by taking conditional expectations at both sides of (11), the following inequality, consisting of five terms labeled [i]$-$[v], holds w.p.1.

$$
\begin{aligned}
\left| E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{U_k^i}{V_k} \middle| \mathscr{F}_k \right] - \frac{\bar{U}_k^i}{\bar{V}_k} \right| \leq & \; E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{|\tilde{U}_k|}{|\tilde{V}_k|^3} \big(V_k - \bar{V}_k\big)^2 I\{\Omega_k\} \middle| \mathscr{F}_k \right] \\
& + E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{1}{\tilde{V}_k^2} \big|(V_k - \bar{V}_k)(U_k^i - \bar{U}_k^i)\big| I\{\Omega_k\} \middle| \mathscr{F}_k \right]
\end{aligned}
$$

$$
\begin{aligned}
&+ E_{\widehat{f}_{\widehat{\theta}_k}}\left[\left|\frac{U_k^i}{V_k} - \frac{\bar{U}_k^i}{\bar{V}_k}\right| I\{\Omega_k^c\}\Big|\mathscr{F}_k\right] \\
&+ E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{1}{|\bar{V}_k|}\big|U_k^i - \bar{U}_k^i\big| I\{\Omega_k^c\}\Big|\mathscr{F}_k\right] \\
&+ E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{|\bar{U}_k^i|}{(\bar{V}_k)^2}\big|V_k - \bar{V}_k\big| I\{\Omega_k^c\}\Big|\mathscr{F}_k\right].
\end{aligned}
$$

For every (U_k^i, V_k) pair in Ω_k, it is easy to see that $\tilde{V}_k > \bar{V}_k - \epsilon > \mathscr{V}/2$ and $|\tilde{U}_k| \leq C\bar{V}_k + \epsilon$, where $C = \max\{|C_1|, |C_2|\}$. Therefore, a bound on the first term [i] is

$$
\begin{aligned}
\text{[i]} &\leq \frac{C\bar{V}_k + \epsilon}{(\bar{V}_k - \epsilon)^3} E_{\widehat{f}_{\widehat{\theta}_k}}\left[(V_k - \bar{V}_k)^2\big|\mathscr{F}_k\right] \\
&\leq \frac{C + \epsilon/\bar{V}_k}{(1 - \epsilon/\bar{V}_k)(\bar{V}_k - \epsilon)^2}\frac{1}{N_k}\Big[\int_{\mathbb{X}} e^{\frac{2H(x)}{T_{k+1}}}/\widehat{f}_{\widehat{\theta}_k}(x)\nu(dx)\Big] \\
&\leq \frac{e^{\frac{2H^*}{T_{k+1}}}}{N_k\lambda_k}\left(\frac{C + 1/2}{\mathscr{V}^2/8}\right)\frac{\mathscr{V}}{f_*} \quad \text{since } \widehat{f}_{\widehat{\theta}_k}(x) \geq \lambda_k f_* \ \forall\, x,\ f_* := \inf\nolimits_{x\in\mathbb{X}} f_{\widehat{\theta}_0}(x) > 0 \\
&\leq \frac{e^{\frac{2H^*}{T_{k+1}}}}{N_k\lambda_k}\frac{8C + 4}{\mathscr{V} f_*}.
\end{aligned}
\tag{12}
$$

Regarding term [ii], we have

$$
\begin{aligned}
\text{[ii]} &\leq \frac{1}{(\bar{V}_k - \epsilon)^2} E_{\widehat{f}_{\widehat{\theta}_k}}\left[|V_k - \bar{V}_k| \cdot |U_k^i - \bar{U}_k^i|\big|\mathscr{F}_k\right] \\
&\leq \frac{4}{\mathscr{V}^2} E_{\widehat{f}_{\widehat{\theta}_k}}\left[(V_k - \bar{V}_k)^2\big|\mathscr{F}_k\right]^{1/2} E_{\widehat{f}_{\widehat{\theta}_k}}\left[(U_k^i - \bar{U}_k^i)^2\big|\mathscr{F}_k\right]^{1/2} \quad \text{(by Hölder's inequality)} \\
&\leq \frac{4}{\mathscr{V}^2}\frac{1}{\sqrt{N_k}}\Big[\int_{\mathbb{X}} e^{\frac{2H(x)}{T_{k+1}}}/\widehat{f}_{\widehat{\theta}_k}(x)\nu(dx)\Big]^{1/2}\frac{1}{\sqrt{N_k}}\Big[\int_{\mathbb{X}} e^{\frac{2H(x)}{T_{k+1}}}\Gamma_i^2(x)/\widehat{f}_{\widehat{\theta}_k}(x)\nu(dx)\Big]^{1/2} \\
&\leq \frac{e^{\frac{2H^*}{T_{k+1}}}}{N_k\lambda_k}\frac{4C}{\mathscr{V} f_*}.
\end{aligned}
\tag{13}
$$

For term [iii], we have

$$\begin{aligned}
\text{[iii]} &= E_{f_{\widehat{\theta}_k}}\left[\Big|E_{\bar{g}_{k+1}}[\Gamma_i(X)] - E_{g_{k+1}}[\Gamma_i(X)]\Big| I\{\Omega_k^c\}\Big|\mathscr{F}_k\right] \quad \text{where } \bar{g}_{k+1} \text{ is given by (5.5)} \\
&\leq |C_1 - C_2| P_{f_{\widehat{\theta}_k}}\left(\Omega_k^c\big|\mathscr{F}_k\right) \\
&\leq |C_1 - C_2|\Big[P_{f_{\widehat{\theta}_k}}\left(|U_k^i - \bar{U}_k^i| \geq \epsilon\big|\mathscr{F}_k\right) + P_{f_{\widehat{\theta}_k}}\left(|V_k - \bar{V}_k| \geq \epsilon\big|\mathscr{F}_k\right)\Big] \\
&\leq |C_1 - C_2|\Bigg[\frac{E_{f_{\widehat{\theta}_k}}\left[\left|U_k^i - \bar{U}_k^i\right|^2|\mathscr{F}_k\right]}{\epsilon^2} + \frac{E_{f_{\widehat{\theta}_k}}\left[\left|V_k - \bar{V}_k\right|^2|\mathscr{F}_k\right]}{\epsilon^2}\Bigg] \\
&\leq \frac{e^{\frac{2H^*}{T_{k+1}}}}{N_k\lambda_k}\frac{4|C_1 - C_2|(1 + C^2)}{\mathscr{V} f_*}. \qquad (14)
\end{aligned}$$

where the third inequality is by Chebyshev's inequality.
Also for term [iv],

$$\begin{aligned}
\text{[iv]} &\leq \frac{1}{\mathscr{V}} E_{f_{\widehat{\theta}_k}}\left[|U_k^i - \bar{U}_k^i| I\{\Omega_k^c\}\big|\mathscr{F}_k\right] \\
&\leq \frac{1}{\mathscr{V}} E_{f_{\widehat{\theta}_k}}\left[(U_k^i - \bar{U}_k^i)^2\big|\mathscr{F}_k\right]^{1/2} P_{f_{\widehat{\theta}_k}}\left(\Omega_k^c\big|\mathscr{F}_k\right)^{1/2} \quad \text{(by Hölder's inequality)} \\
&\leq \frac{C}{\sqrt{\mathscr{V} f_*}}\frac{e^{\frac{H^*}{T_{k+1}}}}{\sqrt{N_k\lambda_k}}\Big[P_{f_{\widehat{\theta}_k}}\left(|U_k^i - \bar{U}_k^i| \geq \epsilon\big|\mathscr{F}_k\right) + P_{f_{\widehat{\theta}_k}}\left(|V_k - \bar{V}_k| \geq \epsilon\big|\mathscr{F}_k\right)\Big]^{1/2} \\
&\leq \frac{e^{\frac{2H^*}{T_{k+1}}}}{N_k\lambda_k}\frac{2C\sqrt{1 + C^2}}{\mathscr{V} f_*}. \qquad (15)
\end{aligned}$$

By using a similar argument, it is straightforward to verify that term [v] is also upper bounded by

$$\text{[v]} \leq \frac{e^{\frac{2H^*}{T_{k+1}}}}{N_k\lambda_k}\frac{2C\sqrt{1 + C^2}}{\mathscr{V} f_*}. \qquad (16)$$

Finally, the proof is completed by applying C7 to (12), (13), (14), (15), and

(16). □

Proof of Lemma 5.2.2: Let $U_k^i = N_k^{-1}\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}}\Gamma_i(x)/\widehat{f}_{\widehat{\theta}_k}(x)$ and $\bar{U}_k^i = E_{\widehat{f}_{\widehat{\theta}_k}}[U_k|\mathscr{F}_k] = \int_{\mathbb{X}} e^{\frac{H(x)}{T_{k+1}}}\Gamma_i(x)\nu(dx)$ be the ith components of $\mathbb{U}_k$ and its conditional expectation. Denote by $\Sigma^k_{i,j}$ the (i,j)th entry of the matrix $\Sigma^k := E_{\widehat{f}_{\widehat{\theta}_k}}[R_kR_k^T|\mathscr{F}_k]$.

By using the same argument as in the proof of Lemma 5.2.2, we have from (11) that

$$\begin{aligned}\frac{U_k^i}{V_k} - \frac{\bar{U}_k^i}{\bar{V}_k} =& \frac{1}{\bar{V}_k}(U_k^i - \bar{U}_k^i) - \frac{\bar{U}_k^i}{(\bar{V}_k)^2}(V_k - \bar{V}_k)\\ &+ \left[\frac{\tilde{U}_k}{\tilde{V}_k^3}(V_k - \bar{V}_k)^2 - \frac{1}{\tilde{V}_k^2}(V_k - \bar{V}_k)(U_k^i - \bar{U}_k^i)\right] I\{\Omega_k\}\\ &+ \left[\frac{U_k^i}{V_k} - \frac{\bar{U}_k^i}{\bar{V}_k} - \frac{1}{\bar{V}_k}(U_k^i - \bar{U}_k^i) + \frac{\bar{U}_k^i}{(\bar{V}_k)^2}(V_k - \bar{V}_k)\right] I\{\Omega_k^c\},\end{aligned}$$

where $\tilde{U}_k$, $\tilde{V}_k$, and Ω_k are defined as in the proof of Lemma 5.2.2. Therefore, we can split $\Sigma^k_{i,j}$ into five terms labeled [i]−[iv] plus a higher-order term as follows:

$$\begin{aligned}\Sigma^k_{i,j} =& c^2k^\beta E_{\widehat{f}_{\widehat{\theta}_k}}\left[\Big(\frac{U_k^i}{V_k} - E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{U_k^i}{V_k}\Big|\mathscr{F}_k\right]\Big)\Big(\frac{U_k^j}{V_k} - E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{U_k^j}{V_k}\Big|\mathscr{F}_k\right]\Big)\Big|\mathscr{F}_k\right]\\ =& c^2k^\beta \frac{1}{\bar{V}_k^2} E_{\widehat{f}_{\widehat{\theta}_k}}\left[(U_k^i - \bar{U}_k^i)(U_k^j - \bar{U}_k^j)\big|\mathscr{F}_k\right]\\ &- c^2k^\beta \frac{\bar{U}_k^j}{\bar{V}_k^3} E_{\widehat{f}_{\widehat{\theta}_k}}\left[(U_k^i - \bar{U}_k^i)(V_k - \bar{V}_k)\big|\mathscr{F}_k\right]\\ &- c^2k^\beta \frac{\bar{U}_k^i}{\bar{V}_k^3} E_{\widehat{f}_{\widehat{\theta}_k}}\left[(U_k^j - \bar{U}_k^j)(V_k - \bar{V}_k)\big|\mathscr{F}_k\right]\end{aligned}$$

$$
\begin{aligned}
&+ c^2 k^\beta \frac{\bar{U}_k^i \bar{U}_k^j}{\bar{V}_k^4} E_{\widehat{f}_{\widehat{\theta}_k}}\left[(V_k - \bar{V}_k)^2 \middle| \mathscr{F}_k\right] \\
&+ c^2 k^\beta \mathcal{R}_k \\
= {}& [\text{i}] - [\text{ii}] - [\text{iii}] + [\text{iv}] + c^2 k^\beta \mathcal{R}_k,
\end{aligned}
$$

where $\mathcal{R}_k$ represents a remainder term.

$$
\begin{aligned}
[\text{i}] &= c^2 k^\beta \frac{1}{\bar{V}_k^2}\Big(E_{\widehat{f}_{\widehat{\theta}_k}}[U_k^i U_k^j | \mathscr{F}_k] - \bar{U}_k^i \bar{U}_k^j\Big) \\
&= c^2 k^\beta \frac{1}{\bar{V}_k^2}\Bigg[\frac{1}{N_k^2} E_{\widehat{f}_{\widehat{\theta}_k}}\Bigg[\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}}\Gamma_i(x)/\widehat{f}_{\widehat{\theta}_k}(x) \cdot \sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}}\Gamma_j(x)/\widehat{f}_{\widehat{\theta}_k}(x) \Bigg| \mathscr{F}_k\Bigg] \\
&\quad - \bar{U}_k^i \bar{U}_k^j\Bigg] \\
&= c^2 k^\beta \frac{1}{\bar{V}_k^2}\frac{1}{N_k}\Bigg(E_{\widehat{f}_{\widehat{\theta}_k}}\Big[e^{\frac{2H(X)}{T_{k+1}}}\Gamma_i(X)\Gamma_j(X)/\widehat{f}^2_{\widehat{\theta}_k}(X)\Big|\mathscr{F}_k\Big] - \bar{U}_k^i \bar{U}_k^j\Bigg) \\
&= \frac{c^2 k^\beta}{N_k}\Bigg(E_{\widehat{f}_{\widehat{\theta}_k}}\Big[\frac{1}{\bar{V}_k^2} e^{\frac{2H(X)}{T_{k+1}}}\Gamma_i(X)\Gamma_j(X)/\widehat{f}^2_{\widehat{\theta}_k}(X)\Big|\mathscr{F}_k\Big] - \frac{\bar{U}_k^i \bar{U}_k^j}{\bar{V}_k^2}\Bigg) \\
&= \frac{c^2 k^\beta}{N_k}\Bigg(E_{g_k}\Big[\Gamma_i(X)\Gamma_j(X)\frac{g_k(X)}{\widehat{f}_{\widehat{\theta}_k}(X)}\Big|\mathscr{F}_k\Big] - E_{g_k}[\Gamma_i(X)]E_{g_k}[\Gamma_j(X)]\Bigg)
\end{aligned}
$$

Similarly, we also have

$$
\begin{aligned}
[\text{ii}] &= \frac{c^2 k^\beta}{N_k}\Bigg(E_{g_k}\Big[\Gamma_j(X)\Big]E_{g_k}\Big[\Gamma_i(X)\frac{g_k(X)}{\widehat{f}_{\widehat{\theta}_k}(X)}\Big|\mathscr{F}_k\Big] - E_{g_k}[\Gamma_i(X)]E_{g_k}[\Gamma_j(X)]\Bigg), \\
[\text{iii}] &= \frac{c^2 k^\beta}{N_k}\Bigg(E_{g_k}\Big[\Gamma_i(X)\Big]E_{g_k}\Big[\Gamma_j(X)\frac{g_k(X)}{\widehat{f}_{\widehat{\theta}_k}(X)}\Big|\mathscr{F}_k\Big] - E_{g_k}[\Gamma_i(X)]E_{g_k}[\Gamma_j(X)]\Bigg), \\
[\text{iv}] &= \frac{c^2 k^\beta}{N_k}\Bigg(E_{g_k}\Big[\Gamma_i(X)\Big]E_{g_k}\Big[\Gamma_j(X)\Big]E_{g_k}\Big[\frac{g_k(X)}{\widehat{f}_{\widehat{\theta}_k}(X)}\Big|\mathscr{F}_k\Big] - E_{g_k}[\Gamma_i(X)]E_{g_k}[\Gamma_j(X)]\Bigg).
\end{aligned}
$$

Note that $|e^{\frac{H(x)}{T_{k+1}}}\Gamma_i(x)/\widehat{f}_{\widehat{\theta}_k}(x)| \leq e^{\frac{H^*}{T^*}}|\Gamma_i(x)|/\lambda_k f_*$. Thus by Assumption C5, the Hoeffding's inequality [27] shows that

$$\begin{aligned} E_{\widehat{f}_{\widehat{\theta}_k}}[I\{\Omega_k^c\}|\mathscr{F}_k] &\leq P_{\widehat{f}_{\widehat{\theta}_k}}\left(|U_k^i - \bar{U}_k^i| \geq \epsilon \big| \mathscr{F}_k\right) + P_{\widehat{f}_{\widehat{\theta}_k}}\left(|V_k - \bar{V}_k| \geq \epsilon \big| \mathscr{F}_k\right) \\ &= O(e^{-CN_k\lambda_k^2}) \text{ for some } \epsilon\text{-dependent constant } C > 0. \end{aligned}$$

This result, when combined with the conditions $N_k = \Theta(k^\beta)$, $\lambda_k = \Omega(k^{-\gamma})$, and $\gamma < \frac{\beta}{2}$ (Assumption B1), indicates that all terms containing $I\{\Omega_k^c\}$ in the remainder $\mathcal{R}_k$ are on the order of $o(k^{-\beta})$. Moreover, a straightforward calculation also shows that all terms involving $I\{\Omega_k\}$ in $\mathcal{R}_k$ are higher order terms of N_k^{-1}. Consequently, we have $ck^\beta\mathcal{R}_k = o(1)$ by taking $N_k = \Theta(k^\beta)$.

Since $T_k \to T^*$ and $T^* > 0$, it is easy to see that $\lim_{k\to\infty} g_k(x) = g^*(x)$ for all $x \in \mathbb{X}$, where $g^*(x) = \frac{e^{H(x)/T^*}}{\int_{\mathbb{X}} e^{H(x)/T^*}\nu(dx)}$. Thus, by the point-wise convergence of $\{f_{\widehat{\theta}_k}\}$ (see the discussion after Assumption B2), the dominated convergence theorem implies that the (i,j)th entry of Σ^k as $k \to \infty$ is

$$\begin{aligned} \Sigma_{i,j} = \Psi\Bigg(&E_{g^*}\left[\Gamma_i(X)\Gamma_j(X)\frac{g^*(X)}{\widehat{f}_{m^{-1}(\Gamma^*)}(X)}\right] - E_{g^*}\Big[\Gamma_j(X)\Big]E_{g^*}\left[\Gamma_i(X)\frac{g^*(X)}{\widehat{f}_{m^{-1}(\Gamma^*)}(X)}\right] \\ &+ E_{g^*}\Big[\Gamma_i(X)\Big]E_{g^*}\Big[\Gamma_j(X)\Big]E_{g^*}\left[\frac{g^*(X)}{\widehat{f}_{m^{-1}(\Gamma^*)}(X)}\right] \\ &- E_{g^*}\Big[\Gamma_i(X)\Big]E_{g^*}\left[\Gamma_j(X)\frac{g^*(X)}{\widehat{f}_{m^{-1}(\Gamma^*)}(X)}\right]\Bigg) \\ = \Psi E_{g^*}&\left[\Big(\Gamma_i(X) - E_{g^*}\left[\Gamma_i(X)\right]\Big)\Big(\Gamma_j(X) - E_{g^*}\left[\Gamma_j(X)\right]\Big)\frac{g^*(X)}{\widehat{f}_{m^{-1}(\Gamma^*)}(X)}\right] \\ = \Psi \widehat{E}_{m^{-1}(\Gamma^*)}&\left[\Big(\Gamma_i(X) - E_{g^*}\left[\Gamma_i(X)\right]\Big)\Big(\Gamma_j(X) - E_{g^*}\left[\Gamma_j(X)\right]\Big)\left(\frac{g^*(X)}{\widehat{f}_{m^{-1}(\Gamma^*)}(X)}\right)^2\right] \end{aligned}$$

for some constant $\Psi > 0$. Therefore, the limiting matrix Σ is given by

$$\Sigma = \Psi \mathrm{Cov}_{\widehat{f}_{m^{-1}(\Gamma^*)}}\left[\Big(\Gamma(X) - E_{g^*}\big[\Gamma(X)\big]\Big)\frac{g^*(X)}{\widehat{f}_{m^{-1}(\Gamma^*)}(X)}\right],$$

where $\mathrm{Cov}_{\widehat{f}_{m^{-1}(\Gamma^*)}}(\cdot)$ is the covariance under

$$\widehat{f}_{m^{-1}(\Gamma^*)}(x) = (1-\lambda^*)f_{m^{-1}(\Gamma^*)}(x) + \lambda^* f_{\widehat{\theta}_0}(x)$$

.

We now show the second claim. By Hölder's inequality, we have

$$\lim_{k\to\infty} E\big[I\{\|R_k\|^2 \geq rk^\alpha\}\|R_k\|^2\big] \leq \limsup_{k\to\infty}\Big[P\big(\|R_k\|^2 \geq rk^\alpha\big)\Big]^{1/2}\Big[E\big[\|R_k\|^4\big]\Big]^{1/2}. \tag{17}$$

By the definition of R_k, it follows that

$$\begin{aligned}
P\big(\|R_k\|^2 \geq rk^\alpha\big) &= P\left(\left\|\frac{\mathbb{U}_k}{V_k} - E_{f_{\widehat{\theta}_k}}\left[\frac{\mathbb{U}_k}{V_k}\middle|\mathscr{F}_k\right]\right\| \geq \frac{\sqrt{r}}{c}k^{\frac{\alpha-\beta}{2}}\right)\\
&\leq \frac{E\Big[\big\|\frac{\mathbb{U}_k}{V_k} - E_{f_{\widehat{\theta}_k}}\big[\frac{\mathbb{U}_k}{V_k}\big|\mathscr{F}_k\big]\big\|^2\Big]}{\frac{r}{c^2}k^{\alpha-\beta}} \quad \text{(by Chebyshev's inequality)}\\
&= \frac{E\Big[E_{f_{\widehat{\theta}_k}}\big[\big\|\frac{\mathbb{U}_k}{V_k} - E_{f_{\widehat{\theta}_k}}\big[\frac{\mathbb{U}_k}{V_k}\big|\mathscr{F}_k\big]\big\|^2\big|\mathscr{F}_k\big]\Big]}{\frac{r}{c^2}k^{\alpha-\beta}}\\
&= \frac{E\Big[E_{f_{\widehat{\theta}_k}}\big[c^2k^\beta\big\|\frac{\mathbb{U}_k}{V_k} - E_{f_{\widehat{\theta}_k}}\big[\frac{\mathbb{U}_k}{V_k}\big|\mathscr{F}_k\big]\big\|^2\big|\mathscr{F}_k\big]\Big]}{rk^\alpha}\\
&= \frac{E[\mathrm{tr}\big(\Sigma^k\big)]}{rk^\alpha}\\
&= O(k^{-\alpha})
\end{aligned}$$

by taking $N_k = \Theta(k^\beta)$ and using an argument similar to the proof of the previous part of the theorem, where $\mathrm{tr}(\Sigma^k)$ is the trace of Σ^k. On the other hand, it is tedious but straightforward to show that

$$\begin{aligned} E\big[\|R_k\|^4\big] &= c^4 k^{2\beta} E\left[\left\|\frac{\mathbb{U}_k}{V_k} - E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{\mathbb{U}_k}{V_k}\middle|\mathscr{F}_k\right]\right\|^4\right] \\ &= c^4 k^{2\beta} E\left[E_{\widehat{f}_{\widehat{\theta}_k}}\left[\left\|\frac{\mathbb{U}_k}{V_k} - E_{\widehat{f}_{\widehat{\theta}_k}}\left[\frac{\mathbb{U}_k}{V_k}\middle|\mathscr{F}_k\right]\right\|^4\middle|\mathscr{F}_k\right]\right] \\ &= c^4 k^{2\beta} O(N_k^{-2}) \\ &= O(1) \end{aligned}$$

Consequently, the right-hand-size of (17) is bounded above by $O\big(k^{-\alpha/2}\big)$, which approaches to zero as $k \to \infty$. $\square$

Proof of Proposition 5.3.1: Note that since $T_{k'} > T^* \; \forall\, k'$, for any $k > 0$, we can find a monotonically non-increasing subsequence $\{T_{k_i},\, i = 0, 1, \ldots\}$ such that $T_{k_0} = T_{k+1}$ and $\lim_{i\to\infty} T_{k_i} = T^*$. We have for any integer $N > 0$,

$$\begin{aligned} \int_{\mathbb{X}} \big|g_{k_N}(x) - g_{k+1}(x)\big|\nu(dx) &\le \int_{\mathbb{X}} \sum_{i=0}^{N-1} \big|g_{k_{i+1}}(x) - g_{k_i}(x)\big|\nu(dx) \\ &= \sum_{i=0}^{N-1} \int_{\mathbb{X}} \big|g_{k_{i+1}}(x) - g_{k_i}(x)\big|\nu(dx) \\ &\le \sum_{i=0}^{N-1} \sqrt{2\mathscr{D}(g_{k_{i+1}}, g_{k_i})} \\ &\qquad \text{(by Pinsker's inequality (e.g., [71])),} \end{aligned}$$

On the other hand, we have from the definition of KL-divergence,

$$\begin{aligned}
\mathscr{D}(g_{k_{i+1}}, g_{k_i}) &= E_{g_{k_{i+1}}}\Big[\frac{g_{k_{i+1}}(X)}{g_{k_i}(X)}\Big] \\
&= \Big(\frac{1}{T_{k_{i+1}}} - \frac{1}{T_{k_i}}\Big) E_{g_{k_{i+1}}}[H(X)] - \ln E_{g_{k_i}}\Big[e^{(\frac{1}{T_{k_{i+1}}} - \frac{1}{T_{k_i}})H(X)}\Big] \\
&\le \Big(\frac{1}{T_{k_{i+1}}} - \frac{1}{T_{k_i}}\Big)\Big[E_{g_{k_{i+1}}}[H(X)] - E_{g_{k_i}}[H(X)]\Big] \\
&\qquad\qquad\qquad\qquad \text{(by Jensen's inequality)} \\
&= \Big(\frac{1}{T_{k_{i+1}}} - \frac{1}{T_{k_i}}\Big)\Big|\int_{\mathbb{X}} H(x)\big(g_{k_{i+1}}(x) - g_{k_i}(x)\big)\nu(dx)\Big| \\
&\le \Big(\frac{1}{T_{k_{i+1}}} - \frac{1}{T_{k_i}}\Big) H^* \int_{\mathbb{X}} \big|g_{k_{i+1}}(x) - g_{k_i}(x)\big|\nu(dx) \\
&\le \Big(\frac{1}{T_{k_{i+1}}} - \frac{1}{T_{k_i}}\Big) H^* \sqrt{2\mathscr{D}(g_{k_{i+1}}, g_{k_i})} \\
&\qquad\qquad\qquad\qquad \text{(by again applying Pinsker's inequality).}
\end{aligned}$$

This implies $\sqrt{2\mathscr{D}(g_{k_{i+1}}, g_{k_i})} \le 2H^*\big(\frac{1}{T_{k_{i+1}}} - \frac{1}{T_{k_i}}\big)$. Therefore,

$$\int_{\mathbb{X}} \big|g_{k_N}(x) - g_{k+1}(x)\big|\nu(dx) \le \sum_{i=0}^{N-1} 2H^*\big(\frac{1}{T_{k_{i+1}}} - \frac{1}{T_{k_i}}\big) = 2H^*\big(\frac{1}{T_{k_N}} - \frac{1}{T_{k+1}}\big). \tag{18}$$

We now use (18) to bound $\|k^{\frac{\alpha+\beta}{2}} W_{1,k}\|$.

$$\begin{aligned}
\|k^{\frac{\alpha+\beta}{2}} W_{1,k}\| &= \big\|k^{\frac{\alpha+\beta}{2}}\big(E_{g_{k+1}}[\Gamma(X)] - E_{g^*}[\Gamma(X)]\big)\big\| \\
&= k^{\frac{\alpha+\beta}{2}}\big\|E_{g_{k+1}}[\Gamma(X)] - \lim_{N\to\infty} E_{g_{k_N}}[\Gamma(X)]\big\| \quad \text{(by Lemma 5.2.1)} \\
&\le k^{\frac{\alpha+\beta}{2}} \lim_{N\to\infty} \int_{\mathbb{X}} \|\Gamma(x)\|\big|g_{k_N}(x) - g_{k+1}(x)\big|\nu(dx) \\
&\le C k^{\frac{\alpha+\beta}{2}} \lim_{N\to\infty} \int_{\mathbb{X}} \big|g_{k_N}(x) - g_{k+1}(x)\big|\nu(dx)
\end{aligned}$$

$$
\begin{aligned}
&\text{(where } C \text{ is an upper bound for } \|\Gamma(X)\|\text{)}\\
&\leq 2H^*Ck^{\frac{\alpha+\beta}{2}}\Big(\frac{1}{T^*} - \frac{1}{T_{k+1}}\Big),
\end{aligned}
$$

which approaches zero as $k \to \infty$ by Assumption D1. □