Variable-Sample Methods and Simulated Annealing for Discrete Stochastic Optimization

Tito Homem-de-Mello*

Department of Industrial, Welding and Systems Engineering
The Ohio State University
1971 Neil Ave., Columbus, Ohio 43210-1271, USA
e-mail: homem-de-mello.1@osu.edu

Abstract

In this paper we discuss the application of a certain class of Monte Carlo methods to stochastic optimization problems. Particularly, we study *variable-sample* techniques, in which the objective function is replaced, *at each iteration*, by a sample average approximation. We first provide general results on the *schedule* of sample sizes, under which variable-sample methods yield consistent estimators as well as bounds on the estimation error. Because the convergence analysis is performed sample-path wise, we are able to obtain our results in a flexible setting, which includes the possibility of using different sampling distributions along the algorithm, without making strong assumptions on the underlying distributions. In particular, we allow the distributions to depend on the decision variables x. We illustrate these ideas by studying a modification of the well-known *simulated annealing* method, adapting it to the variable-sample scheme, and show conditions for convergence of the algorithm.

Keywords: Stochastic optimization, Monte Carlo methods, simulated annealing, Markov chains, sample-path bounds.

*supported by a SEED grant from The Ohio State University

1 Introduction

In the past twenty years a great deal of attention has been devoted to theoretical and practical aspects of optimization of systems under *uncertainty*. Many practical problems involve some type of randomness, which can originate from a variety of sources such as unknown demand or failures of machines, to name a few examples.

Perhaps the most common way to obtain a model that captures the existing randomness is by defining a *random function* of the underlying parameters on a proper probability space and then optimizing the *expected value* of such function with respect to the decision variables. More formally, we have a probability space $(\Omega, \mathcal{F}, P)$, a subset $X \subset I\!R^m$, a (measurable) function $G : X \times \Omega \to I\!R$ and we want to solve

$$\min_{x \in X} \left\{ g(x) := \mathbb{E}\left[G(x)\right] = \int_{\Omega} G(x,\omega)\, P(d\omega) \right\}. \tag{1.1}$$

Typically, the expected value in problem (1.1) cannot be computed exactly, so approximation methods are required. One such approach is to resort to *Monte Carlo methods*: in its basic form, the idea is to replace the expected value function with its corresponding Monte Carlo approximation

$$\hat{g}_N(x) \;=\; \frac{1}{N} \sum_{i=1}^{N} G(x,\omega_i),$$

(where $\omega_1, \ldots, \omega_N$ form an i.i.d. sample) and then solve the resulting deterministic problem. This type of algorithm (sometimes called "sample path optimization", or "sample average approximation") has been well studied, see for instance [23, 27, 29]. One advantage of such method is its nice convergence properties; indeed, it is possible to show convergence of optimal solutions and optimal values under fairly general assumptions (see, e.g., [12, 31, 34, 35]). In some cases, the solution of the approximating problem converges *exponentially fast* on the sample size N to a solution of the original problem. This type of phenomenon was observed by Shapiro and Homem-de-Mello [37] in the context of piecewise linear convex stochastic programs, and studied by Kleywegt, Shapiro and Homem-de-Mello [23] in the context of discrete optimization.

The basic idea in the Monte Carlo method yields several possible variations. For example, suppose we have at hand an iterative method to solve the deterministic problem. Instead of fixing a sample from the beginning and then minimizing the resulting deterministic function, one may consider using *different samples* along the algorithm. That is, the idea is to use, at iteration k, the approximating function

$$\hat{g}_k(x) \;:=\; \frac{G(x,\omega_1^k) + \ldots + G(x,\omega_{N_k}^k)}{N_k},$$

where $\omega_1^k, \ldots, \omega_{N_k}^k$ is a sample from some distribution P_k close to P. Notice that we use a new sample at every iteration (hence the term "variable-sample method" used throughout the paper) as opposed to fixing a large sample at the beginning and then optimizing the resulting deterministic function.

One apparent advantage of a variable-sample scheme over the basic Monte Carlo method is that, since we generate independent estimates of the objective function at different iterations, we avoid getting "trapped" in a single sample-path. For example, as the results in [37, 23] show, some problems have the property that, for almost all ω, there exists a number $N_0 = N(\omega)$ such that the solution $x^*_{N_0}$ of the approximating problem $\min \hat{g}_{N_0}(x)$ *coincides* with the solution x^* of the original problem (1.1). Such N_0, however, is difficult to determine in practice, so for a given sample size N there may be a positive probability that x^*_N is actually far away from x^* — which in turn implies the existence of "bad" sample-paths. This effect tends to be minimized once we generate independent estimates of the objective function.

Another advantage of a variable-sample scheme is that the sample sizes can increase along the algorithm, so that sampling effort is not wasted at the initial iterations of the algorithm. Also, because the estimates at different iterations are independent, one can perform *statistical tests* to compare those estimates, which in turn can lead to *stopping criteria* for the algorithm. Indeed, this type of approach has been successfully used in some gradient-based methods for continuous stochastic optimization; see, for instance, [20, 36].

The price to pay for the flexibility provided by a variable-sample scheme, of course, is that the function being optimized changes at every iteration. Therefore, the convergence results developed for the sample average approximation described above are no longer valid. For example, it is important to ensure that $\hat{g}_k(x) \to g(x)$ with probability one (w.p.1) — i.e., it is desirable that $\hat{g}_k(x)$ be a *consistent* estimator of $g(x)$. Perhaps surprisingly, it turns out that, for such property to hold, it is not enough that the sequence of sample sizes $\{N_k\}$ be increasing; as we show in section 2, N_k must grow at a certain rate.

In fact, we need more than consistency of estimators: in order to obtain convergence of a method adapted to the variable-sample scheme, we must ensure that the error from the deterministic algorithm dominates the stochastic error $|\hat{g}_k(x) - g(x)|$, so that the convergence properties of the deterministic algorithm are preserved. While this assertion is quite intuitive, showing that such property holds for a given algorithm can be a difficult task. This can be made easier by imposing a proper *schedule* of sample sizes, so that we can bound the stochastic error $|\hat{g}_k(x) - g(x)|$.

In this paper we address these issues. We propose a framework to analyze methods that use the variable-sample scheme. In particular, we focus on algorithms that use function evaluations only. We provide general results on consistency of estimators as well as bounds on $|\hat{g}_k(x) - g(x)|$ under the variable-sample scheme. These goals are accomplished by exploiting the fact the estimates $\hat{g}_k(x)$ of the objective function $g(x)$ are obtained via *averaging*, which allows us to use some classical results from probability theory. We obtain generic bounds on the deviation $|\hat{g}_k(x) - g(x)|$, which can then be used to show convergence of a specific method. Such conditions translate directly to the choice of *sample size* N_k used to compute the average at each iteration. An important aspect of our study is that the analysis is completely performed in terms of *sample paths*. This is why it is necessary to resort to tools such as the law of the iterated logarithm instead of the more commonly used Central Limit Theorem. By doing so, we do not need to assume any distribution for the error $|\hat{g}_k(x) - g(x)|$.

Another feature of our analysis is that we allow the use of different sampling distributions P_k at each iteration to obtain the estimate $\hat{g}_k$. This feature can be exploited in several ways, for example by using sampling methods that yield *variance reduction* for the resulting estimators. Notice that the sampling distributions P_k can even yield *biased* estimators of $g(x)$, as long as the bias goes to zero at a specific rate. An important particular case occurs when the underlying distributions *depend on the decision variables* x; in that case, we have $P_k = P_{x_k}$, where x_k is the point obtained in the kth iteration.

As an application of the general framework described above, we consider a method for *discrete stochastic optimization* problems. This class of models of the form (1.1) consists of problems where the feasibility set X is *finite* but typically very large, so that explicit enumeration is not feasible. We focus on problems of the form (1.1) in which the *exact* evaluation of $g(x)$ for a given x is difficult or even impossible. Such difficulty appears for example when the integral in (1.1) cannot be computed exactly (e.g. multidimensional integration), or when G lacks a closed form and can only be evaluated through a "black box" whose inputs are x and ω. In the latter case one cannot make use of methods that exploit the structure of the problem, so it is necessary to resort to general techniques.

Several methods have proposed in the literature to handle the above type of problems. Here we can mention general *random search* procedures such as the ones discussed in Yan and Mukai [42], Gong, Ho and Zhai [16], Andradóttir [3, 4] and Alreafaei and Andradóttir [6]. Another approach is the *ordinal optimization*, proposed by Ho, Sreenivas and Vakili [19], where the order of the function values are estimated, rather than the function

values themselves. Yakowitz, L'Ecuyer and Vásquez-Abad [41] discuss a method where quasi-Monte Carlo techniques are used to select *low-dispersion* points in the feasibility set. An adaptation of the classical *branch-and-bound* method to the context of stochastic optimization is studied by Norkin, Pflug and Ruszczynski [28]. Boesel and Nelson [7] present an alternative procedure based on the combination of *genetic algorithms* with *ranking and selection* techniques. In the particular case where the function $G(\cdot, \omega)$ is the optimal value of a linear programming problem and the set X is polyhedral, *stochastic integer programming* techniques can be applied; see for instance the bibliography on stochastic programming compiled by Maarten Van der Vlerk [39].

The basic Monte Carlo approach described above has also been applied to discrete stochastic optimization problems. As seen earlier, in that case the expected value function is replaced by its corresponding sample average approximation, and the resulting deterministic problem is solved by some discrete optimization method. Morton and Wood [27] use this approach to derive upper and lower bounds to the optimal value, and show that the gap decreases with the sample size. Kleywegt, Shapiro and Homem-de-Mello [23] show some theoretical properties of the method. Besides showing convergence of optimal values, they resort to large deviations techniques to show that the solution of the approximating problem converges *exponentially fast* on the sample size N to a solution of the original problem. On the implementation side, they propose solving a sample average approximation of the problem a few times and then using ranking and selection procedures as a second step. We refer to [23] for details.

In this paper we study the use of the variable-sample framework described above to adapt the *simulated annealing* (SA) method for discrete stochastic optimization. The SA method originates in the work of Metropolis et al. [25] in the fifties to simulate the physical process of annealing, but it was not until the eighties that its use as an optimization tool for *deterministic* problems was proposed by Kirkpatrick, Gelatt and Vecchi [22]. Since then, a large volume of research has been devoted to the study of theoretical properties as well as implementation aspects of the method. There are also quite a few papers reporting successful use of SA techniques in applications. The book by Van Laarhoven and Aarts [40] describes the SA method in detail as well as many applications of this technique; for a more recent review, we refer to Aarts, Korst and Van Laarhoven [1].

The basic mechanism of SA algorithms is the following: let x_k denote the (feasible) point visited on iteration k. Then, choose a *neighbor* of x_k, say y, with probability $R(x_k, y)$, and compare the value of the objective function g at x_k and y. If $g(y) \leq g(x_k)$ (i.e. y is a

better point), then visit point y; otherwise, visit y with a probability that depends on the "temperature" control T_k, and which goes to zero as k goes to infinity. The idea is to allow "uphill" moves to escape local minima, but to decrease more and more the probability of moving to a worse point. The sequence of states visited at each iteration forms a Markov Chain, and it is possible to show that if T_k goes to zero at a specific rate, then the stationary distribution of this Markov Chain is concentrated on the set of optimal solutions. Variations of this basic mechanism have been proposed in the literature, we refer again to [1] for details.

Most of the literature on simulated annealing, however, focuses on *deterministic* optimization problems where the objective function can be evaluated exactly. Few papers have been devoted to the study of the case when such objective is the expected value of a random function and thus must be approximated. Perhaps the first work to touch this subject was the paper by Gelfand and Mitter [15]. They analyze the case where the objective function $g(x)$ can only be computed with "noise" W_k, so that the estimator $g(x) + W_k$ is used at iteration k. By assuming that the noise W_k is normally distributed with mean zero and variance σ_k^2 (which is independent of x), Gelfand and Mitter impose conditions on σ_k^2 to ensure convergence of the method. A similar approach is taken by Gutjahr and Pflug [17], who are able to weaken the normality assumption by considering distributions which are "more peaked around zero" than the normal distribution. Again, conditions are imposed on the variances σ_k^2.

Another type of analysis is done by Fox and Heine [14]. They do not make any normality assumptions; however, they assume that there exist consistent estimators $g_k(x)$ of the objective function g such that the estimators $g_k(x)$ *coincide* with g after a *finite* (almost surely) time N. Fox and Heine suggest that one way to enforce this assumption is by considering only computer-representable numbers as the range of the functions.

Finally, another approach to the problem is studied by Alrefaei and Andradóttir [5]. Their idea is to use a variant of SA where the temperature T_k is kept *constant*. Of course, in such case the Markov Chain defined by the states visited at each iteration does not converge at all to the set of optimal solutions; the idea in [5] is then to consider the sequence of points defined by "points with best estimated objective function so far", and then show that, w.p.1, all accumulation points of such sequence belong to the set of optimal solutions.

Our work differs from those existing approaches. We incorporate a variable-sample method into the standard simulated annealing algorithm, and prove its convergence by applying the general techniques developed for that framework. In particular, we derive a schedule of sample sizes that ensures that the error from the simulated annealing algorithm

dominates the error $|\hat{g}_k(x) - g(x)|$, so that the convergence properties of SA are preserved.

Our setting allows for some degrees of flexibility. As mentioned above, the results we obtain do not assume any particular distribution. Moreover, we allow the use of different sampling distributions P_k at each iteration to obtain the estimate $\hat{g}_k$. Also, we consider the possibility of changing the *selection distributions* $R(x, y)$ — i.e. the probability of selecting the point y for comparison with the current point x — along the iterations, as long as those distributions converge at some specific rate. Some possible ways to take advantage of this flexibility are by using *dynamic neighborhoods* or setting $R(x, y)$ according to previous estimates of g. Both features are mentioned by Fox [13] as essential to obtain a fast implementation of SA.

We must emphasize here that it is not the aim of this paper to provide a new algorithm for discrete stochastic optimization problems; rather, our goal is to establish some general results that can be used by someone who wishes to show convergence of a variable-sample method. In that sense, the SA algorithm is presented here for illustration purposes only — which explains the lack of definitive algorithmic statements as well as the absence of numerical results in the text. Nevertheless, to the best of our knowledge the proof of convergence of a "simulation-based SA" in a general setting, as detailed in section 3, is new and therefore constitutes an additional contribution in itself.

The remaining of this paper is organized as follows: in section 2 we formally introduce *variable-sample* methods. We provide general conditions under which those methods produce consistent estimates as well as pathwise bounds on the estimation error $|\hat{g}_k(x) - g(x)|$. The analysis in these sections is general, in that it does not depend on the particular algorithm being used for optimization. Then, in section 3 we present a variable-sample modification of the SA algorithm, and show that the modification preserves the convergence properties of SA for deterministic problems, as long as the sample sizes N_k grow at a specific rate. Finally, in section 4 we present some concluding remarks.

2 Variable-sample methods

In this section we establish a framework to analyze simulation-based methods that utilize different samples along the iterations. The framework presented is general in that we do not assume any particular structure for the problem, which could be either discrete or continuous.

We start with a few definitions. Let $N_1, N_2, \ldots$ be an increasing sequence of integer

numbers representing the size of the sample used at each iteration. We will call $\{N_k\}$ the *schedule* of sample sizes associated with the algorithm under scrutiny. *We assume that the sample used at any given iteration is i.i.d., and that this sample is independent of previous samples*. Notice that the i.i.d. assumption regards only the sample at a given iteration — that is, samples at different iterations can be drawn from different distributions. With that setting, let Ω^{N_k} denote the N_k-fold cartesian product of the sample space Ω, and let P_k be a probability measure on Ω^{N_k}. Also, let $\widetilde{\Omega} = \Omega^{N_1} \times \Omega^{N_2} \times \ldots$, and let $\widetilde{P}$ denote the corresponding probability distribution on $\widetilde{\Omega}$ generated by the P_ks. Notice that such construction is valid because of the assumption of independence between samples at different iterations. In particular, it implies that $\widetilde{P}(A_k) = P_k(A_k)$, where A_k is any event in the σ-algebra corresponding to iteration k.

It is worth observing here that, in case the original problem is discrete, we can also allow the original distribution P to *depend on the decision variables* x, i.e., we can have

$$g(x) \;=\; \int_{\Omega} G(x,\omega)\, P_x(d\omega).$$

This, is turn, yields considerable flexibility to the model, since in that case we can have $P_k = P_{x_k}$, i.e. P_k may depend on the point selected at iteration k. It is important to notice that such setting is valid with a finite set X, which means that there is actually only a finite number of distinct P_k's. In continuous problems, the dependence of P on x imposes some difficulties, especially if some derivative-based method is to be used. In that case, one needs to resort to techniques such as *likelihood ratio* (see e.g. [32]) in order to compute derivatives. In the discrete case, however, the dependence on x can be easily implemented in a variable-sample context, the only condition being that we must be able to generate samples from the different distributions P_x, $x \in X$, using any stream of *uniform* random numbers. Such condition is imposed in order to guarantee that all random variables $G(x,\omega)$ are defined on a common probability space. It is clear that this condition imposes hardly any constraint on the distributions.

Notice that a point $\omega = (\omega_1^1, \ldots, \omega_{N_1}^1, \omega_1^2, \ldots, \omega_{N_2}^2, \ldots) \in \widetilde{\Omega}$ represents a sample-path followed along the iterations of the algorithm. Define now the following random variables on $(\widetilde{\Omega}, \widetilde{P})$:

$$G_i^k(x,\omega) \;:=\; G(x,\omega_i^k), \quad k = 1,2,\ldots,\; i = 1,\ldots,N_k.$$

Now, for each $\omega \in \widetilde{\Omega}$, define the approximating functions

$$\hat{g}_k(x) \;:=\; \frac{G_1^k(x,\omega) + \ldots + G_{N_k}^k(x,\omega)}{N_k}, \quad k = 1,2,\ldots \tag{2.1}$$

(we omit the dependence of $\hat{g}_k(x)$ on ω for brevity). The function $\hat{g}_k(x)$ is the approximation to the original function $g(x)$ that is used in the kth iteration. Notice that, conditionally on x, the estimates $\{\hat{g}_k(x)\}$, $k = 1, 2, \ldots$, are all *independent* of each other.

2.1 Consistency of estimators

One approach to show convergence of a variable-sample simulation-based method is to show that the algorithm converges for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$. It is natural to think that, in order to have convergence for almost all sample paths, we must have that that the estimators used at each iteration are *consistent*, i.e., for all $x \in X$ and $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$ we must have that

$$\lim_{k\to\infty} \hat{g}_k(x) \;=\; g(x). \tag{2.2}$$

It is interesting to notice that, although condition (2.2) may seem a direct consequence of the law of large numbers, this is not the case. Two factors contribute to that: first, we do not impose that $\mathbb{E}\hat{g}_k(x) = g(x)$, i.e., the estimator $\hat{g}_k(x)$ is allowed to be *biased*. Second, even when this is not the case — for example, when all measures P_k are identical — it could happen that, in principle, $\hat{g}_k(x, \omega)$ does not get close to $g(x)$ with a sample of size N_k. To illustrate the latter point, consider the function $G(x, z) = z$, where z is 0 or 1 with probability 1/2 each, and a sequence ω formed by 2ℓ ones followed by $2\ell + 1$ zeros, $\ell = 0, 1, 2, \ldots$. That is, $\omega = (0, 1, 1, 0, 0, 0, 1, 1, 1, 1, \ldots)$. Suppose that $N_k = k$, $k = 1, 2, \ldots$. Then, after a little algebra we see that

$$\frac{\ell(n)}{2\ell(n)+1} \;\le\; \frac{\sum_{j=1}^{n} \omega_j}{n} \;\le\; \frac{\ell(n)+1}{2\ell(n)+1}, \tag{2.3}$$

where $\ell(n)$ is the smallest nonnegative integer ℓ such that $n \le (\ell + 1)(2\ell + 1)$. The above inequalities imply that

$$\lim_{n\to\infty} \frac{\sum_{j=1}^{n} \omega_j}{n} \;=\; \frac{1}{2} \;=\; \mathbb{E}\omega,$$

so this sample path satisfies the law of large numbers for the overall sequence. However, we have that $\omega_i^k = u_k$, $i = 1, \ldots, N_k$, where $u_k = 0$ if k is odd and $u_k = 1$ otherwise. Hence, we have that $(\omega_1^k + \ldots + \omega_{N_k}^k)/N_k = u_k$ and thus the limit in (2.2) does not exist.

Our task is therefore to show that pathological cases such as the one described above happen only on a set of $\widetilde{P}$-probability zero. Moreover, we must impose conditions on the bias $\mathbb{E}\hat{g}_k(x) - g(x)$. Then, we will be able conclude that (2.2) holds. We start with the following assumptions:

Assumption A1: For each $x \in X$, there exists a positive constant $M = M(x) > 0$ such that $\sup_{k,i} G_i^k(x) \leq M$ w.p.1.

Assumption A2: For each $x \in X$, we have that $\lim_{k\to\infty} \mathbb{E}\hat{g}_k(x) = g(x)$.

A few words about the above assumptions. Assumption A1 says that all random variables are *uniformly bounded* w.p.1. As we shall see below, this assumption can be relaxed, at a certain expense. Assumption A2 says that the estimators $\hat{g}_k(x)$ are *asymptotically unbiased*, and obviously holds in case all $\hat{g}_k(x)$ are unbiased estimators of $g(x)$. We must also notice that the term "constant" in assumption A1 refers to ω rather than x, i.e. constant means "non-random quantity". This terminology is used throughout the paper.

We consider now the following alternative assumptions A1' and A2'. Assumption A1' is clearly weaker than assumption A1; assumption A2', on the other hand, deals with the special case when all probability measures P_k are identical.

Assumption A1': For each $x \in X$, there exists a positive constant $M = M(x) > 0$ such that $\sup_k \mathrm{Var}[G_1^k(x)] \leq M$.

Assumption A2': All probability measures P_k are identical and the estimator $\hat{g}_k(x)$ is unbiased.

Before proceeding with the results, let us recall some basic facts from *large deviations* theory. Let $Y_1,\ Y_2, \ldots$ be a sequence of i.i.d. random variables with finite expectation μ, and for all $N \geq 1$ define $\bar{Y}_N = \sum_{i=1}^N Y_i/N$; then, the weak law of large numbers says that, for any $\delta > 0$,

$$\lim_{N\to\infty} P\left(|\bar{Y}_N - \mu| \geq \delta\right) \ = \ 0, \tag{2.4}$$

and the large deviation theory asserts that the above probability converges to zero exponentially fast. Indeed, for any $N \geq 1$, Chernoff's bound yields

$$P\left(\bar{Y}_N - \mu \geq \delta\right) \ \leq \ e^{-NI(\mu+\delta)}. \tag{2.5}$$

Here, $I(\cdot)$ is the so-called *rate function* corresponding to the distribution of Y_1, which is defined by $I(z) := \sup_{t\in\mathbb{R}}\{tz - \log M(t)\}$, where $M(t) := \mathbb{E}[e^{tY_1}]$ is the *moment generating function* of Y_1 (which is assumed to be finite in a neighborhood of zero). It is possible to show that $I(\cdot)$ is non-negative, strictly convex and attains it minimum at μ, so that the

exponent in the right-hand side of (2.5) is strictly negative. Next, by applying inequality (2.5) to the process $\{-Y_i\}$, we have that

$$P\left(\bar{Y}_N - \mu \leq -\delta\right) \;\leq\; e^{-NI(\mu-\delta)},$$

whence

$$P\left(|\bar{Y}_N - \mu| \geq \delta\right) \;\leq\; 2e^{-N\gamma(\delta)}, \tag{2.6}$$

where $\gamma(\delta) := \min(I(\mu+\delta), I(\mu-\delta))$. This implies (2.4). It is possible to show that the exponential bound in the above inequality is asymptotically sharp, in the sense that

$$P\left(|\bar{Y}_N - \mu| > \delta\right) \;=\; e^{-N\gamma(\delta)+o(N)}. \tag{2.7}$$

The literature on large deviations theory is quite ample; we refer the reader to the books by Dembo and Zeitouni [11] and Shwartz and Weiss [38], for example, for comprehensive discussions. For our purposes, the results in (2.6) and (2.7) will suffice.

The above results, while very important from a *qualitative* point of view, are difficult to use directly since typically it is very hard to compute the rate function $I(\cdot)$. Thus, estimates for $I(\cdot)$ are needed. A useful one can be derived when the corresponding random variables are *uniformly bounded*, i.e., $|Y_i| \leq M$ w.p.1. In that case, we have that

$$I(z) \;\geq\; \frac{(z-\mu)^2}{2M^2} \qquad \text{for all } z \in I\!R \tag{2.8}$$

(see for example Shapiro and Homem-de-Mello [37] for a proof). A similar result can be derived under the weaker assumption that the corresponding random variables have finite variance σ^2. Then, there exists a neighborhood $\mathcal{N}$ of μ such that

$$I(z) \;\geq\; \frac{(z-\mu)^2}{3\sigma^2} \qquad \text{for all } z \in \mathcal{N}. \tag{2.9}$$

This follows directly from the Taylor expansion of the function I; see for example Kleywegt et al. [23] for a proof.

Another estimate of the deviation probability $P(|\bar{Y} - \mu| > \delta)$ can be obtained by a variant of the Central Limit Theorem. In that case, δ goes to zero with n, so that the deviations are never very large. In Chung [8, Thm. 7.1.3], the following result is proved: suppose the i.i.d. sequence $\{Y_i\}$ has finite variance σ^2 and finite third central moment $\gamma^3 := I\!E(|Y_1 - \mu|^3)$. Let a_n be a sequence of real numbers increasing to infinity, and subject to the following growth condition:

$$\lim_{n\to\infty} \log \frac{n\gamma^3}{(n\sigma^2)^{3/2}} + \frac{a_n^2}{2}(1+\varepsilon) \;=\; -\infty \qquad \text{for some } \varepsilon > 0. \tag{2.10}$$

Then, for this ε, there exists N such that for all $n \geq N$ we have

$$e^{-a_n^2(1+\varepsilon)/2} \;\leq\; P\left(\bar{Y} - \mu \geq \frac{a_n \sigma}{\sqrt{n}}\right) \;\leq\; e^{-a_n^2(1-\varepsilon)/2}. \tag{2.11}$$

We can now state the results:

Proposition 2.1 *Suppose that assumptions A1 and A2 hold. Suppose also that the schedule* $\{N_k\}$ *satisfies the following property:*

$$\sum_{k=1}^{\infty} \alpha^{N_k} \;<\; \infty \qquad \textit{for all } \alpha \in (0,1). \tag{2.12}$$

Then, $\hat{g}_k(x) \to g(x)$ *for* $\widetilde{P}$*-almost all* $\omega \in \widetilde{\Omega}$.

Proof: Fix $x \in X$, let $k \geq 1$ and $\delta > 0$. To simplify the notation, let $g_k := \mathbb{E}\hat{g}_k(x)$. Then, by inequality (2.6) above, we have that

$$P_k\left(|\hat{g}_k(x) - g_k| \geq \delta\right) \;\leq\; 2e^{-N_k \gamma_k(\delta)}, \tag{2.13}$$

where $\gamma_k(\delta) := \min(I_k(g_k + \delta), I_k(g_k - \delta))$, and I_k is the rate function of $G_i^k(x)$. By assumption A1, we have that $G_i^k(x) \leq M$ w.p.1 for all k and i and thus, from (2.8), we have

$$\gamma_k(\delta) \;\geq\; \frac{\delta^2}{2M^2}$$

and hence

$$\widetilde{P}\left(|\hat{g}_k(x) - g_k| \geq \delta\right) \;=\; P_k\left(|\hat{g}_k(x) - g_k| \geq \delta\right) \;\leq\; 2e^{-N_k \delta^2/(2M^2)}, \tag{2.14}$$

It follows that

$$\sum_{k=1}^{\infty} \widetilde{P}\left(|\hat{g}_k(x) - g_k| \geq \delta\right) \;\leq\; \sum_{k=1}^{\infty} 2\left(e^{-\delta^2/(2M^2)}\right)^{N_k}. \tag{2.15}$$

Notice that, if condition (2.12) holds, then the expression on the right hand side of the above inequality is finite for all $\delta > 0$. By the Borel-Cantelli lemma (see, e.g., Chung [8, Thm. 4.2.1]), we then have that

$$\widetilde{P}\left(|\hat{g}_k(x) - g_k| \geq \delta \text{ infinitely often}\right) \;=\; 0 \;\; \forall \delta > 0.$$

Finally, assumption A2 implies that, given $\delta > 0$, $|g_k - g(x)| < \delta/2$ for k large enough. It follows that

$$\widetilde{P}\left(|\hat{g}_k(x) - g(x)| \geq \delta/2 \text{ infinitely often}\right) \;=\; 0 \;\; \forall \delta > 0 \tag{2.16}$$

and therefore $\hat{g}_k(x) \to g(x)$ for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$. ∎

The above result can be strengthened in case the measures P_k are identical. As the proposition below shows, in that case condition (2.12) is also *necessary* for convergence of $\hat{g}_k(x)$ to $g(x)$.

Proposition 2.2 *Suppose that assumptions A1' and A2' hold. Then, a sufficient condition to have $\hat{g}_k(x) \to g(x)$ for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$ is that the schedule $\{N_k\}$ satisfies condition (2.12). If, in addition,* $\mathrm{Var}[G_1^1(x)] > 0$*, then condition (2.12) is also* necessary.

Proof: Fix $x \in X$, let $k \geq 1$ and $\delta > 0$. Then, (2.6) and (2.7), together with assumption A2', imply that there exists a sequence $\{c_k\}$ such that $c_k \to 0$ and

$$P_k\left(|\hat{g}_k(x) - g(x)| \geq \delta\right) \;\leq\; 2e^{-N_k(\gamma(\delta)} \tag{2.17}$$

$$P_k\left(|\hat{g}_k(x) - g(x)| > \delta\right) \;\geq\; e^{-N_k(\gamma(\delta)+c_k)}, \tag{2.18}$$

where $\gamma(\delta) := \min(I(g(x)+\delta), I(g(x)-\delta))$, and I is the rate function of $G_i^k(x)$. Now, from (2.9) and assumption A1' we have that there exists a neighborhood $\mathcal{N}$ of zero such that

$$\gamma(\delta) \;\geq\; \frac{\delta^2}{3M^2} \text{ for all } \delta \in \mathcal{N}$$

and thus, for δ small enough and all k,

$$\widetilde{P}\left(|\hat{g}_k(x) - g(x)| \geq \delta\right) \;=\; P_k\left(|\hat{g}_k(x) - g(x)| \geq \delta\right) \;\leq\; 2e^{-N_k\delta^2/(3M^2)}. \tag{2.19}$$

On the other hand, since the random variables $G_i^k(x)$ are assumed to have positive variance, it follows that the rate function I is finite in a neighborhood of $g(x)$, i.e., $\gamma(\delta) < \infty$ for δ small enough. Moreover, since the sequence $\{c_k\}$ goes to zero, we have from (2.18) that, for k large enough,

$$\widetilde{P}\left(|\hat{g}_k(x) - g(x)| > \delta\right) \;=\; P_k\left(|\hat{g}_k(x) - g(x)| > \delta\right) \;\geq\; e^{-N_k C} \tag{2.20}$$

for some $C > 0$.

Together, (2.19) and (2.20) imply that $\sum_{k=1}^{\infty} \widetilde{P}\left(|\hat{g}_k(x) - g_k| > \delta\right)$ is finite for all δ sufficiently small *if and only if* condition (2.12) holds. By applying the full statement of the Borel-Cantelli lemma (see, e.g., Chung [8, Thms. 4.2.1 and 4.2.4]), we conclude that, under assumption A1', we have

$$\begin{aligned} \text{condition (2.12) holds} \;&\Longrightarrow\; \widetilde{P}\left(|\hat{g}_k(x) - g(x)| > \delta \text{ infinitely often}\right) \;=\; 0 \\ \text{condition (2.12) does not hold} \;&\Longrightarrow\; \widetilde{P}\left(|\hat{g}_k(x) - g(x)| > \delta \text{ infinitely often}\right) \;=\; 1. \end{aligned}$$

It follows from the above implications that a necessary and sufficient condition to have $\hat{g}_k(x) \to g(x)$ for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$ is that condition (2.12) holds. ∎

Some remarks about the above results are now in order. First, notice that condition (2.12) imposes a mild constraint on the schedule of sample sizes. Indeed, it is evident that such condition holds if $N_k \geq ck$, where c is any positive constant. Even a sublinear growth such as $N_k = \sqrt{k}$ satisfies (2.12). Notice however not all increasing schedules satisfy (2.12): for example, with $N_k = \log k$ we have, for any $\alpha > 0$,

$$\sum_{k=1}^{\infty} \alpha^{\log k} \;=\; \sum_{k=1}^{\infty} e^{\log k \log \alpha} \;=\; \sum_{k=1}^{\infty} k^{\log \alpha}$$

which converges if and only if $\alpha < 1/e$. Therefore, condition (2.12) does not hold in that case. A somewhat surprising consequence of Proposition 2.2 is that, when the measures P_k are identical and $N_k = \log k$, $\hat{g}_k(x)$ does *not* converge to $g(x)$ for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$. The proof of Proposition 2.2 shows why this happens — for any given δ, the deviation $|\hat{g}_k(x) - g(x)|$ is larger than δ infinitely often w.p.1. In that case, we can only conclude that there exists a *subsequence* of $\{\hat{g}_k(x)\}$ converging to $g(x)$.

Another remark concerns the necessity of condition (2.12) in Proposition 2.2. Observe the importance of the assumption of positive variance in that proposition, since otherwise $G_i^k(x) \equiv g(x)$ and thus (2.12) would not be necessary. Moreover, condition (2.12) is not necessary under the conditions of Proposition 2.1. Indeed, suppose for example that $G_i^k(x)$ has distribution with mean $g(x)$ and variance $\sigma_k^2 = 1/k$; clearly, $G_i^k(x)$ approaches the constant $g(x)$ w.p.1 as k grows and therefore any nondecreasing schedule $\{N_k\}$ (for example, $N_k = 1$ for all k) guarantees that $\hat{g}_k(x)$ converges to $g(x)$ w.p.1.

We conclude this subsection by proposing yet another alternative to Propositions 2.1 and 2.2. It requires a stronger assumption on the schedule $\{N_k\}$ but it requires weaker assumptions on the underlying random variables. A related result was derived by Cooper, Henderson and Lewis [9] in a different context.

Proposition 2.3 *Suppose that assumptions A1' and A2 hold. Suppose also that the schedule $\{N_k\}$ satisfies the following property:*

$$\sum_{k=1}^{\infty} \frac{1}{N_k} \;<\; \infty. \tag{2.21}$$

Then, $\hat{g}_k(x) \to g(x)$ for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$.

Proof: Fix $x \in X$, let $k \geq 1$ and $\delta > 0$. We apply Chebyshev's inequality to obtain

$$P_k\left(|\hat{g}_k(x) - \mathbb{E}\hat{g}_k(x)| \geq \delta\right) \;\leq\; \frac{\mathrm{Var}[\hat{g}_k(x)]}{\delta^2} \;=\; \frac{\mathrm{Var}[G_1^k(x)]/\delta^2}{N_k} \;\leq\; \frac{M/\delta^2}{N_k}. \tag{2.22}$$

Thus, $\sum_{k=1}^{\infty} \widetilde{P}\left(|\hat{g}_k(x) - \mathbb{E}\hat{g}_k(x)| \geq \delta\right)$ converges if and only if (2.21) holds. The remainder of the proof is identical to that of Proposition 2.1 and therefore is omitted. ∎

2.2 Sample-path bounds

The results in the previous subsection ensure consistency of the estimators used at each iteration. Intuitively, this guarantees that, if k (and therefore N_k) is large enough, then $\hat{g}_k(x)$ is close to $g(x)$ and so, in principle, when k is large a variable-sample method should not behave too differently from a hypothetical method that could solve the original problem (1.1). Notice however that the basic argument for using a variable-sample method is to update the sample sizes *as the algorithm progresses*; therefore, we need stronger results than just consistency. In particular, we need to derive *bounds* on the deviation $|\hat{g}_k(x) - g(x)|$.

The theorem below provides such bound. Notice that the result is not stated in terms of distributions, but rather for each sample path ω. In a sense, it corresponds to the law of the iterated logarithm in the standard i.i.d. case. Observe that conditions imposed on the schedule $\{N_k\}$ are stronger than before. We shall also impose the following assumption:

Assumption A3: For each $x \in X$, there exist a positive constant $M_1 = M_1(x)$ such that $\sup_k \mathbb{E}[|G_1^k(x) - \mathbb{E}\hat{g}_k(x)|^3]/(\mathrm{Var}[G_1^k(x)])^{3/2} \leq M_1$.

Assumption A3 holds, for example, if the random variables $G_1^k(x)$ have uniformly bounded third moment (for all k) and their variances are uniformly bounded away from zero.

Theorem 2.1 *Suppose that assumption A3 holds. Suppose also that the schedule $\{N_k\}$ satisfies the following property:*

$$N_k \;\geq\; ck^{\rho} \quad \text{for some } c > 0 \text{ and some } \rho > 0. \tag{2.23}$$

If $\rho > 2$, then for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$ there exists $K = K(\omega) > 0$ such that

$$|\hat{g}_k(x) - g(x)| \;\leq\; \sigma_k(x)\sqrt{\frac{\log N_k}{N_k}} + |\mathbb{E}\hat{g}_k(x) - g(x)| \tag{2.24}$$

for all $k > K$, where $\sigma_k^2(x) := \mathrm{Var}[G_1^k(x)]$.

Proof: Fix $x \in X$ and let $k \geq 1$. Our goal is to apply inequalities (2.11) to estimate the deviation probabilities of $\hat{g}_k(x) - \mathbb{E}\hat{g}_k(x)$. Let $\sigma_k^2 := \mathrm{Var}[G_i^k(x)]$ and $\gamma_k^3 := \mathbb{E}[|G_i^k(x) - \mathbb{E}\hat{g}_k(x)|^3]$. Let $\delta > 0$ be arbitrarily chosen, and define $a_k := \sqrt{\frac{1}{(1+\delta)} \log N_k}$. Clearly, $a_k \to \infty$. Moreover, the sequence $\{a_k\}$ satisfies the growth condition (2.10), since

$$\begin{aligned}
\limsup_{k\to\infty} \log \frac{N_k \gamma_k^3}{(N_k \sigma_k^2)^{3/2}} + \frac{a_k^2}{2}(1+\varepsilon) &= \limsup_{k\to\infty} \log \frac{\gamma_k^3}{\sigma_k^3} + \log N_k^{-1/2} + \frac{1+\varepsilon}{2(1+\delta)} \log N_k \\
&\leq \log M_1 + \limsup_{k\to\infty} \left(\frac{1+\varepsilon}{2(1+\delta)} - \frac{1}{2} \right) \log N_k \\
&= -\infty \quad \text{for } \varepsilon < \delta.
\end{aligned}$$

Therefore, the conditions for (2.11) are satisfied when $\varepsilon < \delta$. Fix now an $\varepsilon < \delta/(1+2\delta)$. Thus, from (2.11) we have

$$2e^{-\frac{1+\varepsilon}{2(1+\delta)} \log N_k} \leq P_k \left(|\hat{g}_k(x) - \mathbb{E}\hat{g}_k(x)| \geq \frac{\sigma_k}{\sqrt{1+\delta}} \sqrt{\frac{\log N_k}{N_k}} \right) \leq 2e^{-\frac{1-\varepsilon}{2(1+\delta)} \log N_k},$$

that is,

$$2N_k^{-\frac{1+\varepsilon}{2(1+\delta)}} \leq \widetilde{P} \left(|\hat{g}_k(x) - \mathbb{E}\hat{g}_k(x)| \geq \frac{\sigma_k}{\sqrt{1+\delta}} \sqrt{\frac{\log N_k}{N_k}} \right) \leq 2N_k^{-\frac{1-\varepsilon}{2(1+\delta)}}$$

for k large enough, say $k \geq K$.

Suppose now that the schedule $\{N_k\}$ satisfies (2.23). By summing over k in the above inequalities, we obtain

$$\begin{aligned}
\sum_{k=K}^{\infty} C_1 k^{-\frac{(1+\varepsilon)\rho}{2(1+\delta)}} &\leq \sum_{k=1}^{\infty} \widetilde{P} \left(|\hat{g}_k(x) - \mathbb{E}\hat{g}_k(x)| \geq \frac{\sigma_k}{\sqrt{1+\delta}} \sqrt{\frac{\log N_k}{N_k}} \right) \qquad (2.25) \\
&\leq \sum_{k=K}^{\infty} C_2 k^{-\frac{(1-\varepsilon)\rho}{2(1+\delta)}} \qquad (2.26)
\end{aligned}$$

where C_1 and C_2 are positive constants. Next, recall that ε was fixed above to be smaller than δ. It follows from (2.25) that

$$\sum_{k=K}^{\infty} C_1 k^{-\rho/2} \leq \sum_{k=K}^{\infty} \widetilde{P} \left(|\hat{g}_k(x) - \mathbb{E}\hat{g}_k(x)| \geq \frac{\sigma_k}{\sqrt{1+\delta}} \sqrt{\frac{\log N_k}{N_k}} \right)$$

and so the series on the right-hand side of the above inequality *diverges* if $\rho \leq 2$.

Assume now that $\rho > 2$. Then, let $\delta \leq (\rho - 2)/4 > 0$. We have $\rho \geq 2(1+2\delta)$ and thus in (2.26) we obtain

$$\sum_{k=K}^{\infty} \widetilde{P} \left(|\hat{g}_k(x) - \mathbb{E}\hat{g}_k(x)| \geq \frac{\sigma_k}{\sqrt{1+\delta}} \sqrt{\frac{\log N_k}{N_k}} \right) \leq \sum_{k=K}^{\infty} C_1 k^{-\frac{(1-\varepsilon)(1+2\delta)}{1+\delta}}. \qquad (2.27)$$

The expression on the right hand side of the above inequality is finite if and only if the exponent of k is less than -1, i.e. if and only if $\varepsilon < \delta/(1+2\delta)$. Since ε was fixed above to satisfy such condition, it follows that the expression on the left hand side of (2.27) is finite when $\rho > 2$.

The above conclusions, together with the Borel-Cantelli lemma, imply that

$$\widetilde{P}\left(|\hat{g}_k(x) - \mathbb{E}\hat{g}_k(x)| \geq \frac{\sigma_k}{\sqrt{1+\delta}}\sqrt{\frac{\log N_k}{N_k}} \text{ infinitely often}\right) = 0 \quad \text{if } \rho > 2$$

$$\widetilde{P}\left(|\hat{g}_k(x) - \mathbb{E}\hat{g}_k(x)| \geq \frac{\sigma_k}{\sqrt{1+\delta}}\sqrt{\frac{\log N_k}{N_k}} \text{ infinitely often}\right) = 1 \quad \text{if } \rho \leq 2.$$

The first equation holds for all $0 < \delta \leq (\rho-2)/4$, whereas the second holds for any $\delta > 0$. Therefore, the assertion of the theorem follows. ∎

Remark: The proof of the theorem shows that, in a sense, $\rho > 2$ is the weakest requirement on ρ that yields a bound of order $\sqrt{\log N_k/N_k}$. Indeed, if $\rho \leq 2$ then for any $\delta > 0$ and $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$ there does not exist $K = K(\omega) > 0$ such that

$$|\hat{g}_k(x) - \mathbb{E}\hat{g}_k(x)| \leq \frac{\sigma_k}{\sqrt{1+\delta}}\sqrt{\frac{\log N_k}{N_k}}$$

for all $k > K$.

Notice that the right-hand side in (2.24) has a component due to the bias $\mathbb{E}[G_1^k(x)] - g(x)$. If this bias dominates the term $\sqrt{\log N_k/N_k}$, then of course the error $|\hat{g}_k(x) - g(x)|$ will be the order of the bias. Under assumption A2" below, Theorem 2.1 yields a direct consequence.

Assumption A2": For each $x \in X$, there exists a positive constant $D = D(x)$ such that

$$\left|\mathbb{E}\left[G_i^k(x)\right] - g(x)\right| \leq D\sqrt{\frac{\log N_k}{N_k}} \quad \text{for all } k \geq 1. \tag{2.28}$$

Corollary 2.1 *Suppose that assumptions A1',A2" and A3 hold. Suppose also that the schedule $\{N_k\}$ satisfies the following property:*

$$N_k \geq ck^{2+\delta} \quad \textit{for some } c > 0 \textit{ and some } \delta > 0. \tag{2.29}$$

Then, there exists a constant $C = C(x) > 0$ such that, for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$, there exists $K = K(\omega) > 0$ such that

$$|\hat{g}_k(x) - g(x)| \leq C\sqrt{\frac{\log N_k}{N_k}} \tag{2.30}$$

for all $k > K$.

The above results provide the desired bound on the deviation $|\hat{g}_k(x) - g(x)|$. Note that no assumptions were made on the distribution of $G_i^k(x)$, other than some boundedness assumptions on the first three moments. This underscores the generality of the results. Another remark is that, clearly, (2.30) implies that $\hat{g}_k(x) \to g(x)$ w.p.1, which was the conclusion of Propositions 2.1-2.3. Those propositions, however, use weaker assumptions on the schedule $\{N_k\}$, so we stated them for the sake of completeness even though we shall assume in the next section that the conditions for validity of Corollary 2.1 hold.

2.3 Cumulative samples

The results in the previous sections focus on a certain sampling structure — namely, it is assumed that samples used at different iterations are independent. It is natural to think of an alternative scheme, where at each iteration a new sample is *appended* to the previous one. In other words, using the notation defined earlier, if the sample used at the first iteration is $\omega_1^1, \ldots, \omega_{N_1}^1$, then the sample used at the second iteration is $\omega_1^1, \ldots, \omega_{N_1}^1, \omega_1^2, \ldots, \omega_{N_2}^2$ and so on. Thus, the estimator $\bar{g}_k(x)$ used at iteration k is defined as

$$\bar{g}_k(x) \;:=\; \frac{G_1^1(x,\omega) + \ldots + G_{N_k}^k(x,\omega)}{N_1 + \ldots + N_k}, \quad k = 1, 2, \ldots \tag{2.31}$$

for each $\omega \in \widetilde{\Omega}$.

It is clear that, under such scheme, consistency follows immediately from the strong law of large numbers when the measures P_k are identical. If the measures P_k are not identical, then we need extra conditions on the bias $|\mathbb{E}[G_i^k(x)] - g(x)]$, as the proposition below shows:

Proposition 2.4 *Suppose that assumptions A1' and A2" hold. Then, for all $x \in X$,*

$$\lim_{k\to\infty} \bar{g}_k(x) \;=\; g(x) \tag{2.32}$$

for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$, provided that $\lim_{k\to\infty} N_1 + \ldots + N_k = \infty$.

Proof: Fix $x \in X$, and let $Y_i^k = G_i^k(x) - \mathbb{E}[G_i^k(x)]$, $k = 1, 2, \ldots$, $i = 1, \ldots, N_k$. Then, $\mathbb{E}[Y_i^k] = 0$ and, by assumption A1', $\mathbb{E}|Y_i^k|^2 \leq M$. It follows from a classical result in Probability that $\lim_{k\to\infty}(Y_1^1 + \ldots + Y_{N_k}^k)/(N_1 + \ldots + N_k) = 0$ w.p.1 (see e.g. Chung[8, p.125]) and hence, by assumption A2", we have that

$$\begin{aligned}
\lim_{k\to\infty} \frac{G_1^1(x) + \ldots + G_{N_k}^k(x)}{N_1 + \ldots + N_k} \;&=\; \lim_{k\to\infty} \frac{\mathbb{E}\left[G_1^1(x)\right] + \ldots + \mathbb{E}\left[G_{N_k}^k(x)\right]}{N_1 + \ldots + N_k} \\
&\leq\; \lim_{k\to\infty} \frac{N_1\left(g(x) + D\sqrt{\frac{\log N_1}{N_1}}\right) + \ldots + N_k\left(g(x) + D\sqrt{\frac{\log N_k}{N_k}}\right)}{N_1 + \ldots + N_k}
\end{aligned}$$

$$\begin{aligned} &\leq g(x) + \lim_{k\to\infty} \frac{(\sqrt{N_1 \log N_1} + \ldots + \sqrt{N_k \log N_k})D}{N_1 + \ldots + N_k} \\ &= g(x) \end{aligned}$$

for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$. Similarly, one can show that $\lim_{k\to\infty}(G_1^1(x) + \ldots + G_{N_k}^k(x))/(N_1 + \ldots + N_k) \geq g(x)$, so the assertion of the proposition follows. ∎

A bound similar to the one given by Theorem 2.1 can also be derived in this case. It is a direct consequence of the *law of the iterated logarithm* — which can be seen from the "log log" term on the bound.

Proposition 2.5 *Suppose that any of the conditions below is satisfied:*

i. Assumption A2' holds;

ii. Assumption A1 holds;

iii. There exist constants $A > 0$ and $0 < \varepsilon < 1$ such that

$$\frac{\Gamma_k}{\Sigma_k^3} \leq \frac{A}{(\log \Sigma_k)^{1+\varepsilon}},$$

where $\Sigma_k^2 := \sum_{i=1}^k N_i \sigma_i^2$, $\sigma_i^2 := \mathrm{Var}[G_1^i(x)]$, $\Gamma_k = \sum_{i=1}^k N_i \gamma_i^3$, and also $\gamma_i^3 :=$ $\mathbb{E}\left[\left|G_1^i(x) - \mathbb{E}[G_1^i(x)]\right|^3\right]$.

Suppose also that assumption A1' holds and that $\mathbb{E}[G_1^k(x)] = g(x)$ for all $x \in X$ and all $k = 1, 2, \ldots$. Then, for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$ there exist positive constants C and $K = K(\omega)$ such that

$$|\bar{g}_k(x) - g(x)| \leq C\sqrt{\frac{\log\log(N_1 + \ldots + N_k)}{N_1 + \ldots + N_k}} \qquad \forall\, k \geq K, \tag{2.33}$$

provided that $\lim_{k\to\infty} N_1 + \ldots + N_k = \infty$.

Proof: Fix $x \in X$. Then, conditions (i)-(iii) above, together with independence of the variables $G_i^k(x)$, imply that we can use the law of the iterated logarithm for the sequence $G_1^1(x), \ldots, G_{N_k}^k(x)$, so that, w.p.1,

$$\limsup_{k\to\infty} \frac{(G_1^1(x) - \mathbb{E}[G_1^1(x)]) + \ldots + (G_{N_k}^k(x) - \mathbb{E}[G_{N_k}^k(x)])}{\sqrt{2\Sigma_k^2 \log\log \Sigma_k}} = 1 \tag{2.34}$$

$$\liminf_{k\to\infty} \frac{(G_1^1(x) - \mathbb{E}[G_1^1(x)]) + \ldots + (G_{N_k}^k(x) - \mathbb{E}[G_{N_k}^k(x)])}{\sqrt{2\Sigma_k^2 \log\log \Sigma_k}} = -1 \tag{2.35}$$

(see, for instance, Chung [8] and Rao [30]). Since $\mathbb{E}[G_1^k(x)] = g(x)$ for all k and $\Sigma_k^2 = \sum_{i=1}^k N_i \sigma_i^2 \leq M \sum_{i=1}^k N_i$, it follows from the above inequalities that, given $\delta > 0$, there exists $K = K(\omega)$ such that

$$\left|(G_1^1(x) - g(x)) + \ldots + (G_{N_k}^k(x) - g(x))\right| \leq (1+\delta)\sqrt{2\Sigma_k^2 \log\log \Sigma_k} \quad \forall\, k \geq K$$

and thus

$$\left|\frac{G_1^1(x)+\ldots+G_{N_k}^k(x)}{N_1+\ldots+N_k}-g(x)\right| \leq (1+\delta)\sqrt{\frac{2M\log\log\sqrt{M(N_1+\ldots+N_k)}}{N_1+\ldots+N_k}}$$
$$\leq C\sqrt{\frac{\log\log(N_1+\ldots+N_k)}{N_1+\ldots+N_k}}$$

for some $C>0$ and k large enough. ∎

The task of showing convergence of a variable-sample simulation-based method is facilitated by the above results. Such proof depends of course on the specific deterministic algorithm being used, but all is needed now is a proof that the convergence properties of the deterministic algorithm are kept when, at each iteration k, one replaces the original function $g(x)$ by an approximating function $\hat{g}_k(x)$ such that $\hat{g}_k(x) \to g(x)$ w.p.1 as k goes to infinity.

The main task is to show that, in some sense, the deterministic error *dominates* the stochastic error resulting from approximating $g(x)$ by $\hat{g}_k(x)$. In that sense, Theorem 2.1 and Proposition 2.5 are crucial, as they provide upper bounds on the error $|\hat{g}_k(x)-g(x)|$. An important aspect of those bounds is that they are *distribution-free*, which allows for applicability of those results in fairly general contexts. Moreover, because those bounds are derived for sample paths, one can analyze the underlying algorithm for each individual sample path, which typically leads to stronger "w.p.1" results. Finally, we emphasize that the properties derived in the previous section hold both when new samples are drawn at each iteration as well as when samples are accumulated from one iteration to the next. Therefore, the convergence results will be valid under either sampling scheme.

In the next section we will see an example of analysis of a specific algorithm, a modified version of simulated annealing.

3 Simulated Annealing

In this section we describe a variable-sample modification of the basic SA method to adapt it to discrete stochastic optimization problems. We then proceed to prove its convergence. Throughout this section, we will call this modified SA "stochastic simulated annealing", as opposed to "deterministic simulated annealing", which is understood as SA applied to deterministic optimization.

3.1 Description of the algorithm

The basic stochastic simulated annealing algorithm takes the form below. Here, $\{N_k\}$ is the sequence of sample sizes used at each iteration, $\{T_k\}$ is the sequence of values for the "temperature" control, and $V(x)$ denotes the set of the neighbors of x. We assume that the neighborhood structure is symmetric, i.e., if $y \in V(x)$ then $x \in V(y)$.

Algorithm

$x_0 :=$ *initial state;*
$N_0 :=$ *initial sample size;*
$k := 0$*;*
Repeat
Choose a state y *from* $V(x_k)$ *according to the selection distribution* $R_{x_k,\cdot}(k)$*;*
Generate a sample $\omega_1^k, \ldots, \omega_{N_k}^k$ *from a distribution* P_k*;*
Compute $\hat{g}_k(x_k)$*,* $\hat{g}_k(y)$ *according to (2.1) or (2.31);*
If $\hat{g}_k(y) \leq \hat{g}_k(x_k)$
then $x_{k+1} := y$
else generate a uniform random number U *between 0 and 1;*
if $U < \exp([\hat{g}_k(x_k) - \hat{g}_k(y)]/T_k)$
then $x_{k+1} := y$*;*
Update T_k*,* N_k*;*
$k := k + 1$*;*
Until {*stopping criterion is satisfied*}.

The above description is of course very loose, since it does not specify how to perform some of the steps. We discuss now some of these issues.

The choice of a state y in the neighborhood structure of the current point x_k is made randomly, according to some pre-specified distribution $R_{x_k,\cdot}(k)$, which we will call the *selection distribution*. That is, point $y \in V(x_k)$ is chosen with probability $R_{x_k,y}(k)$. A common choice for the selection distribution is $R_{xy}(k) = 1/|V(x)|$, i.e. all neighbors of a point x are equally likely. In general, however, different neighbors may be assigned different probabilities. Notice that, unlike the usual description of simulated annealing, we allow the selection distribution to vary with k (as we shall see later, such variation is allowed as long as the selection distributions $R(k)$ converge at a certain rate). This feature yields another degree of flexibility, and in particular allows the implementation of *dynamic neighborhoods*, so that

the neighborhood structure changes with k, since by setting $R_{x_k,y}(k) = 0$ (and therefore $R_{y,x_k}(k) = 0$) we prevent y to be chosen. This can be potentially used to enhance the speed of convergence (cf. Fox [13]).

The next issue is the generation of a sample $\omega_1^k, \ldots, \omega_{N_k}^k$. This is an i.i.d. sample which is also independent of previous samples. As remarked in section 2, the idea is to "reset" the estimates so that the behavior of the algorithm is not influenced by a particular sample. Notice that we also allow for the use of cumulative samples, as discussed in section 2.3. In some cases, it may be useful to drop the i.i.d. assumption, especially if one is implementing some type of *variance reduction* techniques. Using non-i.i.d. samples does not affect the validity of the theoretical convergence results, as long as it is possible to guarantee that the estimators $\hat{g}_k(x)$ converge pointwise to $g(x)$ at a rate satisfying (2.30).

Once the value of the approximating function $\hat{g}_k$ is compared at the points x_k and a chosen neighbor y, the algorithm moves from x_k to y with "probability one" if $\hat{g}_k(y) < \hat{g}_k(x_k)$, and with "probability $\exp((\hat{g}_k(x_k) - \hat{g}_k(y))/T_k)$" otherwise. Notice the use of quotes, since the probability depends on $\hat{g}_k$ and thus is a random measure. In summary, the Markov chain $\{Z_k\}$ generated by the algorithm has the (random) transition probability matrix

$$P_{ij}(k) \;=\; R_{ij}(k) \exp(-[\hat{g}_k(j) - \hat{g}_k(i)]^+/T_k), \tag{3.1}$$

where $[a]^+ = \max(a, 0)$.

The issue of how to update N_k and T_k is discussed below. As we shall see, it is necessary to impose some growth conditions on N_k and T_k, as well as on the rate of convergence of the selection distributions $R(k)$, in order to guarantee convergence of the overall algorithm. Our goal is then to show that, under those conditions and when (2.30) holds (as well as additional assumptions), the simulated annealing algorithm converges to an optimal solution w.p.1. Here "w.p.1" refers to the probability measure $\widetilde{P}$ corresponding to the sample space $\widetilde{\Omega}$, which in turn represents the set of possible sample paths to be followed along the algorithm (see section 2).

We now discuss briefly the issue of choosing a stopping criterion. In the deterministic context, a criterion commonly used is $T_k < \varepsilon$, where ε is a pre-specified constant; here, we suggest comparing the values of the objective function estimates at the current iteration with values at previous iterations, and stop if no improvement has been obtained for some time. Notice that we can perform a *t-test* to compare the estimates used at two different iterations, or more generally, we can perform an *analysis of variance* to compare the estimates corresponding to several iterations. This type of idea was used by Shapiro and Homem-de-Mello [36] in the context of continuous stochastic optimization.

A final remark about the above algorithm: in the way it is stated, a new sample is generated and the parameters N_k and T_k are updated every iteration. In practice, however, we can "freeze" those values for a few iterations, so that the algorithm uses the same sample and same temperature for a few iterations before being updated again. The benefit of such approach is to take advantage of the fact that, once the sample is fixed, the resulting approximating function is deterministic and hence the algorithm behaves as its deterministic version during those iterations.

3.2 Convergence of the algorithm

We proceed now to show convergence of the algorithm. The main result is formalized in Theorem 3.1 below. Before that, however, we need to establish a few definitions and preliminary results.

We follow the approach in Mitra, Romeo and Sangiovanni-Vicentelli [26]. The idea is to show that the stochastic process consisting of the feasible points visited on each iteration forms a time-inhomogeneous Markov Chain, and then to show that this chain is *strongly ergodic* and converges to the points in the solution set.

In what follows, we shall assume that the sampling scheme employed is that of generating new samples at each iteration, i.e. estimator (2.1) is used. This is done just for simplicity, and the same results can be easily derived for the cumulative sampling scheme.

We shall impose the following assumption:

Assumption B1: The selection distributions $R(k)$ are *reversible*, that is, for each $k = 0, 1, \ldots$ there exists a probability distribution $w(k)$ such that

$$w_i(k)R_{ij}(k) \;=\; w_j(k)R_{ji}(k) \quad \text{for all } i, j \in X. \tag{3.2}$$

Notice that, in particular, the uniform selection distribution $R_{ij}(k) = 1/|V(i)|$ for $j \in V(i)$ satisfies assumption B1, since if we define $w(k)$ by $w_i(k) = |V(i)|/\sum_{\ell \in X} |V(\ell)|$, then $w(k)$ satisfies (3.2). Since we assume that the feasible set X is *finite*, without loss of generality we shall write $X = \{1, \ldots, S\}$. Define now, for $i = 1, \ldots, S$,

$$\pi_i(k) \;=\; \frac{w_i(k)\exp(-\hat{g}_k(i)/T_k)}{\Gamma(k)} \tag{3.3}$$

where $\Gamma(k) = \sum_{i=1}^{S} w_i(k)\exp(-\hat{g}_k(i)/T_k)$ is a normalizing function to ensure that $\|\pi(k)\| =$

1, where $\pi(k) = (\pi_i(k), \ldots, \pi_S(k))$ and

$$\|\pi(k)\| \ := \ \sum_{i=1}^{S} \pi_i(k). \tag{3.4}$$

It is important to observe that all quantities defined above depend on the sample path $\omega \in \widetilde{\Omega}$. Recall that our goal is to prove convergence of the algorithm for $\widetilde{P}$-almost all ω; thus, we shall fix from now on some $\omega \in \widetilde{\Omega}$ for which (2.30) holds, and show that the algorithm converges for that ω. In this setting, we can then omit the dependence on ω to ease the notation.

The following lemma illustrates a property of the probability distribution $\pi(k)$ defined in (3.3). It shows that $\pi(k)$ is a left-eigenvector of the matrix $P(k)$. Such property is called *quasi-stationarity* and will play an important role in showing convergence of the algorithm.

Lemma 3.1 *Suppose that assumption B1 holds. Then, we have*

$$\pi(k)^T P(k) \ = \ \pi(k)^T, \qquad k = 0, 1, \ldots \ .$$

Proof: The proof is similar to that of Mitra et al. [26, Proposition 3.1]. We have that, for any $i, j \in X$ and any $k \geq 0$,

$$\begin{aligned} \frac{P_{ij}(k)}{P_{ji}(k)} &= \frac{R_{ij}(k)\exp(-[\hat{g}_k(j) - \hat{g}_k(i)]^+/T_k)}{R_{ji}(k)\exp(-[\hat{g}_k(i) - \hat{g}_k(j)]^+/T_k)} \\ &= \frac{R_{ij}(k)}{R_{ji}(k)}\exp((\hat{g}_k(i) - \hat{g}_k(j))/T_k) \qquad (3.5) \\ &= \frac{w_j(k)\exp(-\hat{g}_k(j)/T_k)}{w_i(k)\exp(-\hat{g}_k(i)/T_k)} \ = \ \frac{\pi_j(k)}{\pi_i(k)}, \end{aligned}$$

where the equality in (3.5) follows from the fact that $[a-b]^+ - [b-a]^+ = a - b$ for any a, b. Thus, we have that $\pi_i(k)P_{ij}(k) = \pi_j(k)P_{ji}(k)$, whence the assertion of the lemma follows. ∎

The next result shows that the quasi-stationary probability vector $\pi(k)$ converges, as k goes to infinity, to a distribution supported on the set X^* of optimal solutions. Unlike the deterministic result in Mitra et al. [26, Proposition 3.2], however, here we need to impose conditions on the schedule $\{N_k\}$. We shall also need the following assumption:

Assumption B2: The selection distribution matrices $R(k)$ converge (as $k \to \infty$) to a matrix R such that R defines an *irreducible* Markov chain.

Notice that, because of the symmetry of the neighborhood structure, it is natural to impose the irreducibility assumption. Indeed, if R is not irreducible then it must be composed of disconnected classes — which means that the feasible set X is partitioned into "clusters" that do not communicate with each other. Assumption B2 prevents such situation from happening.

We recall some notation. For two sequences $\{a_k\}$ and $\{b_k\}$ of positive numbers, we use the "asymptotic lower bound" notation $a_k = \omega(b_k)$ and $a_k = \Omega(b_k)$, meaning respectively that $\lim_{k\to\infty} a_k/b_k = \infty$ and $\liminf_{k\to\infty} a_k/b_k \geq c$ for some $c \geq 0$. We use the "asymptotic upper bound" notation $a_k = o(b_k)$ and $a_k = O(b_k)$, meaning respectively that $\lim_{k\to\infty} a_k/b_k = 0$ and $\limsup_{k\to\infty} a_k/b_k \leq c$ for some $c \geq 0$. Finally, we use the "asymptotically tight bound" notation $a_k = \Theta(b_k)$, meaning that there exist positive constants c_1, c_2 and K such that $c_1 \leq a_k/b_k \leq c_2$ for all $k \geq K$. See e.g. Cormen, Leiserson and Rivest [10] for further discussion on this topic.

Lemma 3.2 *Suppose that: i) assumptions B1 and B2 hold, ii) the control sequence $\{T_k\}$ converges to zero, and iii) the schedule $\{N_k\}$ satisfies $N_k = \Omega((1/T_k)^{2+\delta})$ for some constant $\delta > 0$. Suppose also that (2.30) holds. Then, for all $i = 1, \ldots, S$ we have that*

$$\lim_{k\to\infty} \pi_i(k) \;=\; \pi_i^* := \begin{cases} \frac{w_i}{\sum_{j\in X^*} w_j} & \text{if } i \in X^* \\ 0 & \text{otherwise,} \end{cases} \tag{3.6}$$

where $w = \lim_{k\to\infty} w(k)$.

Proof: Notice initially that (3.2) implies that $w(k) = w(k)^T R(k)$ for all k. By taking limits on both sides we obtain that

$$\lim_{k\to\infty} w(k) \;=\; (\lim_{k\to\infty} w(k))^T R$$

and hence, since R is irreducible with finite number of states, it follows that R is positive recurrent and thus $\lim_{k\to\infty} w(k) = w$, where w is the unique probability vector that satisfies $w = w^T R$. Next, notice that

$$\begin{aligned} \frac{1}{\pi_i(k)} &= \frac{\sum_{j=1}^S w_j(k) \exp(-\hat{g}_k(j)/T_k)}{w_i(k) \exp(-\hat{g}_k(i)/T_k)} \\ &= \sum_{j=1}^S \frac{w_j(k)}{w_i(k)} \exp([\hat{g}_k(i) - \hat{g}_k(j)]/T_k). \end{aligned} \tag{3.7}$$

The latter sum can be broken into three pieces, namely, $A_< = \{j : g(i) < g(j)\}$, $A_> = \{j : g(i) > g(j)\}$ and $A_= = \{j : g(i) = g(j)\}$. By assumption, we have that $\hat{g}_k(j) \to g(j)$ for all

j and hence there exists constants $c, K > 0$ such that, for all $k > K$,

$$\begin{aligned} \hat{g}_k(i) - \hat{g}_k(j) &< -c \quad \text{when } j \in A_< \\ \hat{g}_k(i) - \hat{g}_k(j) &> c \quad \text{when } j \in A_>. \end{aligned}$$

This in turn implies that

$$\lim_{k\to\infty} \exp([\hat{g}_k(i) - \hat{g}_k(j)]/T_k) = 0 \quad \text{for } j \in A_< \tag{3.8}$$

$$\lim_{k\to\infty} \exp([\hat{g}_k(i) - \hat{g}_k(j)]/T_k) = \infty \quad \text{for } j \in A_>. \tag{3.9}$$

Now, observe that when $j \in A_=$ we have $g(i) - g(j) = 0$ and hence it follows from (2.30) that $\hat{g}_k(i) - \hat{g}_k(j) = O(\sqrt{\log N_k/N_k})$. By assumption, $N_k = \Omega((1/T_k)^{2+\delta})$, so we have that $N_k T_k^2 = \Omega\left(1/T_k^{\delta}\right)$. Therefore, $N_k T_k^2 \to \infty$ and hence

$$\begin{aligned} 0 &= \lim_{k\to\infty} \frac{\log(N_k T_k^2)}{N_k T_k^2} = \lim_{k\to\infty} \frac{\log N_k}{N_k T_k^2} + \frac{\log T_k^2}{N_k T_k^2} \\ &= \lim_{k\to\infty} \frac{\log N_k}{N_k T_k^2}, \end{aligned} \tag{3.10}$$

where the last equality follows from the fact that

$$\frac{\log T_k^2}{N_k T_k^2} = O\left(\frac{\log T_k}{1/T_k^{\delta}}\right) = o(1).$$

From (3.10) we conclude that $\log N_k/N_k = o(T_k^2)$ and thus $\hat{g}_k(i) - \hat{g}_k(j) = o(T_k)$. Hence,

$$\lim_{k\to\infty} \exp([\hat{g}_k(i) - \hat{g}_k(j)]/T_k) = 1 \quad \text{for } j \in A_=. \tag{3.11}$$

Finally, as seen earlier we have that $\lim_{k\to\infty} w(k) = w$ and hence $\lim_{k\to\infty} w_j(k)/w_i(k) = w_j/w_i$. From (3.7)-(3.11) we conclude that

$$\lim_{k\to\infty} \frac{1}{\pi_i(k)} = \begin{cases} \infty & \text{if } A_> \neq \emptyset \\ \sum_{j\in A_=} w_j/w_i & \text{otherwise} \end{cases}$$

and thus

$$\lim_{k\to\infty} \pi_i(k) = \begin{cases} 0 & \text{if } \exists\, j : g(j) < g(i) \\ \frac{w_i}{\sum_{j\in X^*} w_j} & \text{otherwise} \end{cases}$$

as asserted by the lemma. ∎

The above result, although useful, is not sufficient to prove convergence of the simulated annealing algorithm. The reason is that the vector $\pi(k)$, which was shown in Lemma 3.2 to converge to a distribution supported on the set X^* of optimal solutions, does not correspond

to the probabilities $P(Z_k = i)$, $i = 1, \ldots, S$ (recall that $\{Z_k\}$ is the Markov chain representing the states visited at each iteration). As seen in Lemma 3.1, $\pi(k)$ is a *quasi-stationary* distribution corresponding to the transition probabilities $P(k)$.

What is necessary here is the concept of *ergodicity* for time-inhomogeneous Markov chains. This topic has been largely studied in the literature on Markov chains (see e.g. Iosifescu [21], Madsen and Isaacson [24] and Seneta [33]), and its use as a tool to show convergence of simulated annealing in the deterministic context was proposed in Mitra et al. [26].

Let $p(k) = (p_1(k), \ldots, p_S(k))$ denote the state probability vector after k transitions, i.e. $p_i(k) = P(Z_k = i)$. Also, let $P(n, k)$ denote the k-step transition probability matrix starting at step n, i.e. $P(n, k) = \prod_{i=0}^{k-1} P(n + i)$. We need now the following definitions:

Definition: A time-inhomogeneous Markov chain is *weakly ergodic* if, for all i, j, r and n,

$$\lim_{k \to \infty} |P_{ir}(n, k) - P_{jr}(n, k)| \;=\; 0. \tag{3.12}$$

The chain is *strongly ergodic* if there exists a vector $\nu = (\nu_1, \ldots, \nu_S)$ such that $\nu \geq 0$, $\|\nu\| = 1$ and, for all i, r and n,

$$\lim_{k \to \infty} |P_{ir}(n, k) - \nu_r| \;=\; 0. \tag{3.13}$$

Thus, in a weakly ergodic chain the matrix $P(n, k)$ tends (as $k \to \infty$) to have all rows identical, although those rows depend on k. In a strongly ergodic chain, all rows converge to vector ν. In the latter case we have, for all n and j,

$$\lim_{k \to \infty} p_j(n + k) \;=\; \lim_{k \to \infty} \sum_{i=1}^{S} P_{ij}(n, k) p_i(n) \;=\; \sum_{i=1}^{S} \nu_j p_i(n) \;=\; \nu_j, \tag{3.14}$$

so ν is actually the limiting probability distribution of the chain.

We consider initially weak ergodicity. It is well known (see e.g. Iosifescu [21, Theorem 1]) that a necessary and sufficient condition for a time-inhomogeneous Markov chain to be weakly ergodic is the existence of an increasing sequence of numbers $\{n_i\}$, $i = 0, 1, \ldots$ such that

$$\sum_{i=0}^{\infty} \alpha(P(n_i, n_{i+1} - n_i)) \;=\; \infty, \tag{3.15}$$

where $\alpha(Q)$ is the *coefficient of ergodicity* of the matrix Q, which is defined as

$$\alpha(Q) \;=\; \min_{i,j \in \{1,\ldots,S\}} \sum_{\ell=1}^{S} \min(Q_{i\ell}, Q_{j\ell}) \tag{3.16}$$

$$= \quad 1 - \frac{1}{2} \max_{i,j \in \{1,\ldots,S\}} \sum_{\ell=1}^{S} |Q_{i\ell} - Q_{j\ell}|.$$

Notice that $1-\alpha(Q)$ measures the maximum difference (in L_1-norm) between any two rows of Q.

We proceed now as in Mitra et al [26]. The idea is to show that there exists a number r and a node ℓ such that ℓ can be reached from any other node in r steps. Thus, the matrix $P(n,r)$ has only positive entries in the ℓth column. It is possible to show that, for n large enough,

$$P_{i\ell}(n,r) \;\geq\; \kappa^r \exp(-rL/T_{n+r-1}), \tag{3.17}$$

where $\kappa := \min_i \min_{j \in V(i)} R_{ij} - \varepsilon$ and $L = \max_i \max_{j \in V(i)} |g(j) - g(i)| + \varepsilon$, for some $\varepsilon > 0$.

The proof of the validity of the above inequality follows similar steps as in [26], with two additional observations. The first one is that, by assumption B2, we have that $\min_i \min_{j \in V(i)} R_{ij} > 0$ and thus, since $R_k \to R$, it follows that there exists $\varepsilon > 0$ such that $\min_i \min_{j \in V(i)} R_{ij}(k) \geq \min_i \min_{j \in V(i)} R_{ij} - \varepsilon > 0$ for k large enough. The second observation is that, since $\hat{g}_k(i) \to g(i)$ for all i, we have that $|\hat{g}_k(j) - \hat{g}_k(i)| \leq L$ for k large enough. Therefore, $P_{ij}(k) \geq \kappa \exp(-L/T_k)$ for all i,j such that $j \in V(i)$ and all sufficiently large k.

From (3.17) it follows that, if

$$T_k \;=\; \frac{\Lambda}{\log(k+k_0)}, \quad k = 0,1,2,\ldots, \tag{3.18}$$

where $\Lambda \geq rL$ and $k_0 \geq 1$, then the Markov chain defined by the simulated annealing algorithm is weakly ergodic. We refer again to [26] for details. Notice that the logarithmic schedule defined by (3.18) appears often in the literature on deterministic simulated annealing, and has been shown to suffice for convergence of the algorithm in that context (see e.g. Hajek[18]).

It remains to show strong ergodicity. As shown below, this is again guaranteed by imposing appropriate conditions on the schedule $\{N_k\}$. Also, as seen earlier, strong ergodicity implies existence of the limiting probabilities $P(Z_k = i)$. Together with Lemma 3.2, this shows that the stochastic simulated annealing algorithm does converge to the set optimal solutions. The theorem below states precisely this result:

Theorem 3.1 *Suppose that the control sequence $\{T_k\}$ satisfies (3.18), whereas the schedule $\{N_k\}$ satisfies*

$$N_k \;=\; \Omega\left(k^{2(1+\delta)}\right), \tag{3.19}$$

for some constant $\delta > 0$. Suppose also that assumptions B1 and B2 hold, and that the convergence of $R(k)$ to R occurs at a rate of at least $1/k^{1+\delta}$. Finally, assume that assumptions A1',A2" and A3 hold. Then, the Markov chain defined by the stochastic simulated annealing algorithm is strongly ergodic and we have, for $\widetilde{P}$-almost all $\omega \in \widetilde{\Omega}$,

$$\lim_{k\to\infty} P(Z_k = i) \;=\; \begin{cases} \frac{w_i}{\sum_{j\in X^*} w_j} & \text{if } i \in X^* \\ 0 & \text{otherwise.} \end{cases} \tag{3.20}$$

Proof: We shall prove that

$$\sum_{k=1}^{\infty} \|\pi(k+1) - \pi(k)\| \;<\; \infty, \tag{3.21}$$

where $\pi(k)$ and $\|\pi(k)\|$ are defined respectively in (3.3) and (3.4). Since the chain is weakly ergodic under the assumptions of the theorem, strong ergodicity follows from (3.21) — a result due to Madsen and Isaacson [24].

In order to show (3.21), consider the Markov chain $\{Z_n^d\}$ defined by the deterministic simulated annealing algorithm which uses the original function g rather than approximating functions $\hat{g}_k$. Such algorithm is of course theoretical, since g cannot be evaluated exactly. Let $P^d(k)$ denote the transition matrix of this chain at iteration k. Next, define the vector

$$\pi_i^d(k) \;=\; \frac{w_i \exp(-g(i)/T_k)}{\Gamma^d(k)}, \tag{3.22}$$

where w is the limit of $w(k)$ and $\Gamma^d(k) = \sum_{j=1}^{S} w(j)\exp(-g(j)/T_k)$. Notice that $\pi^d(k)$ is a deterministic function, whereas $\pi(k)$ is defined in terms of the approximating functions $\hat{g}_k$ and thus depends on the sample path $\omega \in \widetilde{\Omega}$. We can then apply the (deterministic) results in Mitra et al [26] to conclude that:

$$\begin{aligned} \pi^d(k)P^d(k) &= \pi^d(k), \qquad k = 0, 1, \ldots \\ \sum_{k=1}^{\infty} \|\pi^d(k+1) - \pi^d(k)\| &< \infty. \end{aligned} \tag{3.23}$$

Now, since

$$\|\pi(k+1) - \pi(k)\| \;\le\; \|\pi(k+1) - \pi^d(k+1)\| + \|\pi^d(k+1) - \pi^d(k)\| + \|\pi^d(k) - \pi(k)\|,$$

it follows from (3.23) that, to show (3.21), it suffices to show that

$$\sum_{k=1}^{\infty} \|\pi(k) - \pi^d(k)\| < \infty. \tag{3.24}$$

By the definition of $\pi(k)$ and $\pi^d(k)$ we have that

$$\|\pi(k) - \pi^d(k)\| \;=\; \sum_{i=1}^{S} \left| \frac{w_i(k)\exp(-\hat{g}_k(i)/T_k)}{\Gamma(k)} - \frac{w_i \exp(-g(i)/T_k)}{\Gamma^d(k)} \right|. \tag{3.25}$$

Notice that, by assumption, we have that $R(k)$ converges to R at the rate of at least $1/k^{1+\delta}$ and hence the same property holds for $w(k)$. Thus, there exists a constant C such that, for k large enough and each $i \in X$,

$$|w_i(k) - w_i| \;\leq\; C/k^{1+\delta}.$$

Now, assumption B2 implies that $w_i > 0$ for all i, since w is the stationary distribution of a positive recurrent chain. Since $w(k) \to w$, it follows that there exists $\alpha > 0$ such that $w_i \geq \alpha$ and $w_i(k) \geq \alpha$ for k large enough and all i. Hence, we have that, for sufficiently large k,

$$\begin{aligned} \frac{w_i(k)}{w_i} &\leq 1 + \frac{C}{\alpha k^{1+\delta}} \\ \frac{w_i}{w_i(k)} &\leq 1 + \frac{C}{\alpha k^{1+\delta}} \end{aligned}$$

which (since $T_k \to 0$) in turn implies that

$$|\log w_i(k) - \log w_i| \;\leq\; \log\left(1 + \frac{C}{\alpha k^{1+\delta}}\right) \;\leq\; \frac{1}{T_k}\log\left(1 + \frac{C}{\alpha k^{1+\delta}}\right) \tag{3.26}$$

for k large enough, say $k > K_1$.

Next, since the assumptions of Corollary 2.1 are satisfied, (2.30) holds and so $|\hat{g}_k(i) - g(i)| = O(\sqrt{\log N_k/N_k})$. Let $\gamma = 1 + (\delta/2) > 1$. From (3.19) we obtain

$$\frac{N_k}{\log N_k} \;=\; \Omega\left(\frac{k^{2(1+\delta)}}{\log k^{2(1+\delta)}}\right) \;=\; \Omega\left(k^{2\gamma}\frac{k^{\delta}}{2(1+\delta)\log k}\right) \;=\; \omega(k^{2\gamma}),$$

since $k^{\delta}/\log k \to \infty$. Thus, $\sqrt{\log N_k/N_k} = o(1/k^{\gamma})$. Moreover, from the series expansion $\log(1+x) = x - x^2/2 + x^3/3 - x^4/4 + \ldots$ (for small $x > 0$), it follows that $1/k^{\gamma} = \Theta(\log(1 + 1/k^{\gamma}))$. Thus, we have $|\hat{g}_k(i) - g(i)| = o(\log(1 + 1/k^{\gamma}))$ and hence there is some $K_2 > 0$ such that

$$\left|\frac{\hat{g}_k(i) - g(i)}{T_k}\right| \;<\; \frac{1}{T_k}\log\left(1 + \frac{1}{k^{\gamma}}\right) \qquad \forall\, k \geq K_2. \tag{3.27}$$

Inequalities (3.26) and (3.27) together imply that

$$\left|\log w_i(k) - \log w_i - \frac{\hat{g}_k(i) - g(i)}{T_k}\right| \;<\; 2\beta_k \qquad \forall\, k \geq K := \max(K_1, K_2),$$

where $\beta_k = \log(1 + \mu/k^{\gamma})/T_k$ and $\mu = \max(1, C/\alpha)$. It follows that, when $k \geq K$,

$$\log w_i - \frac{g(i)}{T_k} - 2\beta_k \;<\; \log w_i(k) - \frac{\hat{g}_k(i)}{T_k} \;<\; \log w_i - \frac{g(i)}{T_k} + 2\beta_k.$$

This in turn implies that

$$w_i e^{-g(i)/T_k} e^{-2\beta_k} \;<\; w_i(k) e^{-\hat{g}_k(i)/T_k} \;<\; w_i e^{-g(i)/T_k} e^{2\beta_k} \tag{3.28}$$

and thus

$$\Gamma^d(k)e^{-2\beta_k} \;<\; \Gamma(k) \;<\; \Gamma^d(k)e^{2\beta_k}. \tag{3.29}$$

Inequalities (3.28) and (3.29) together imply that

$$\pi_i^d(k)e^{-4\beta_k} \;<\; \pi_i(k) \;<\; \pi_i^d(k)e^{4\beta_k}.$$

Therefore,

$$\left|\pi_i^d(k) - \pi_i(k)\right| \;<\; \pi_i^d(k)\left[e^{4\beta_k} - e^{-4\beta_k}\right]$$

and hence

$$\|\pi(k) - \pi^d(k)\| \;<\; e^{4\beta_k} - e^{-4\beta_k}.$$

for all $k > K$.

We now show that the series $\sum_{k=1}^{\infty} e^{4\beta_k} - e^{-4\beta_k}$ converges. Indeed, from the series expansion $e^x = 1 + x + x^2/2! + x^3/3! + \ldots$ we have

$$e^{4\beta_k} - e^{-4\beta_k} \;=\; 8\beta_k\left(1 + \frac{(4\beta_k)^2}{3!} + \frac{(4\beta_k)^4}{5!} + \frac{(4\beta_k)^6}{7!} + \ldots\right) \;\leq\; 8\beta_k e^{4\beta_k}.$$

From (3.18) and the fact that $\log(1 + \mu/k^\gamma) = \Theta(1/k^\gamma)$ we have that $\beta_k = \Theta(\log k/k^\gamma)$. Thus, $\beta_k \to 0$ and hence it follows that for k large enough we have

$$e^{4\beta_k} - e^{-4\beta_k} \;<\; 9\beta_k.$$

Since $\sum_{k=1}^{\infty} \log k/k^\gamma < \infty$, we have that $\sum_{k=1}^{\infty} e^{4\beta_k} - e^{-4\beta_k} < \infty$. It follows that (3.24) holds and consequently (3.21) holds.

Now, the aforementioned result by Madsen and Isaacson [24] implies that the chain is strongly ergodic and, moreover, $\lim_{k\to\infty} |P_{ir}(n,k) - \pi_i^*| = 0$, where π^* (defined in (3.6)) is the limit of $\pi(k)$, which exists by the virtue of Lemma 3.2 (notice that the assumption on the schedule $\{N_k\}$ used in Lemma 3.2 is implied by (3.19)). From (3.14) we see that $\lim_{k\to\infty} P(Z_k = i) = \pi_i^*$, which completes the proof. ∎

Some remarks about the above theorem. First, notice that, as mentioned before, Theorem 3.1 states convergence in distribution with respect to the random measure corresponding to the selected point $\widetilde{\omega}$ (for $\widetilde{P}$-almost all $\widetilde{\omega}$), i.e. the measure obtained *conditionally* on $\widetilde{\omega}$. By unconditioning on $\widetilde{\omega}$ and applying a bounded convergence theorem, we can easily obtain convergence in distribution with respect to the original measure $\widetilde{P}$. Moreover, if the original problem has a unique optimal solution x^* — i.e., if $X^* = \{x^*\}$ — then Theorem 3.1 implies convergence in probability to x^*.

Another remark is that, in terms of implementation, it is worth keeping not only the current point x_k but also *the best solution obtained so far*, say y_k. Clearly, the Markov chain defined by the y_k's will not "oscillate" as much the chain defined by the x_k's. However, because of the stochastic error that is present when comparing two function evaluations, with this procedure we still have only convergence in probability at best.

Finally, notice that Theorem 3.1 agrees with a result obtained by Gutjahr and Pflug [17], which states that SA for functions with normal noise $N(0, \sigma_k^2)$ converges if $\sigma_k = O(k^{-\gamma})$ for some $\gamma > 1$. Indeed, if $G(x, \omega)$ has normal distribution $N(g(x), \sigma^2)$, then $\hat{g}_k(x) - g(x)$ has normal distribution $N(0, \sigma^2/N_k)$ and hence $\sqrt{\sigma^2/N_k} = O(k^{-\gamma})$ if and only if $N_k = \Omega(k^{2\gamma})$, which is the condition (3.19). As pointed out earlier, however, Theorem 3.1 does not make use of any normality assumptions.

4 Conclusions

We have presented a general framework to show convergence of a certain class of methods to solve stochastic optimization problems, which we called *variable-sample* methods. Such procedures essentially consist of incorporating sampling into deterministic algorithms that use function evaluations only. Although a complete proof of convergence will depend on the method under scrutiny, we have provided general results to aid in that task. In particular, we have given conditions on the schedule of sample sizes $\{N_k\}$ that ensure consistency of the estimators and also guarantee some bounds on the deviation from true values. The results provided are general, in that no particular distribution is assumed.

To illustrate the type of analysis made possible by this framework, we presented a modification of the simulated annealing algorithm that can be used to solve general discrete stochastic optimization problems. Our results provide a schedule of sample sizes that guarantees convergence of the algorithm, without making strong assumptions on the underlying distributions. They also allow some degree of flexibility in the choice of neighborhood structures and sampling distributions, which can vary along the algorithm and can also depend on the feasible points.

Some questions of course remain open: on the theoretical side, the study of *rates of convergence* is very important to provide some intuition on the behavior of the algorithm. Unfortunately, however, the very definition of rate of convergence is not standard in the stochastic optimization literature, so some further study is required. Perhaps an analysis of *finite time* behavior of simulated annealing, which has been studied for SA applied to

deterministic problems, can be the way to go.

On the practical side, the implementation of an algorithm based on the techniques described in this paper would be of interest, as well as its behavior in a real application. In that respect, it would be important to derive appropriate stopping criteria, perhaps using analysis of variance techniques as discussed in section 3.1.

References

[1] E. Aarts, J. Korst and P. Van Laarhoven, "Simulated Annealing", in *Local Search in Combinatorial Optimization*, E. Aarts and J.K. Lenstra (eds.), 91-120, John Wiley & Sons, New York , 1997.

[2] E. Aarts and J.K. Lenstra (eds.), *Local Search in Combinatorial Optimization*, John Wiley & Sons, New York , 1997.

[3] S. Andradóttir, "A method for discrete stochastic optimization", *Management Science* 41 (1995), 1946-1961.

[4] S. Andradóttir, "A global search method for discrete stochastic optimization", *SIAM Journal of Optimization* 6 (1996), 513-530.

[5] M. Alrefaei and S. Andradóttir, "A simulated annealing algorithm with constant temperature for discrete stochastic optimization", *Management Sci.*, 45: 748–764, 1999.

[6] M. Alrefaei and S. Andradóttir, "A modification of the stochastic ruler method for discrete stochastic optimization", accepted for publication in *The European Journal of Operational Research* , 2000.

[7] J. Boesel and B. L. Nelson, "Accounting for randomness in heuristic simulation optimization", preprint, 1999.

[8] K.L. Chung, *A Course in Probability Theory*, 2nd ed., Academic Press, 1974.

[9] W.L. Cooper, S.G. Henderson and M.E. Lewis, "Perfect simulation and the policy iteration algorithm", manuscript, Department of Industrial and Operations Engineering, University of Michigan, 2001.

[10] T.H. Cormen, C.E. Leiserson and R.L. Rivest, *Introduction to Algorithms*, MIT Press, 1990.

[11] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd. ed., Springer-Verlag, New York, 1998.

[12] J. Dupačová and R. Wets, "Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems", *The Annals of Statistics*, 16 (1988), 1517-1549.

[13] B.L. Fox, "Simulated annealing: folklore, facts, and directions", in *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, Lecture Notes in Statistics 106, 17-48, 1995.

[14] B.L. Fox and G.W. Heine, "Probabilistic search with overrides", *Annals of Applied Probability*, 5:1087-1094, 1995.

[15] S.B. Gelfand and S.K. Mitter, "Simulated annealing with noisy or imprecise energy measurements", *Journal of Optimization Theory and Applications*, vol. 62, no. 1, 49-62, 1989.

[16] W.-B. Gong, Y.-C. Ho and W. Zhai, "Stochastic comparison algorithm for discrete optimization with estimation", *Proc. 31st. Conf. Decision Control*, 795-800, 1992.

[17] W.J. Gutjahr and G.Ch. Pflug, "Simulated annealing for noisy cost functions", *Journal of Global Optimization*, 8:1-13, 1996.

[18] B. Hajek, "Cooling schedules for optimal annealing", *Mathematics of Operations Research*, 13 (1988), 311-329.

[19] Y.-C. Ho, R.S. Sreenivas and P. Vakili, "Ordinal optimization of DEDS", *Discrete Event Dyn. Systems*, 2, 61-88, 1992.

[20] T. Homem-de-Mello, A. Shapiro and M.L. Spearman, "Finding optimal material release times using simulation based optimization", *Management Science* 1, 45 (1999), 86-102.

[21] M. Iosifescu, "On two recent papers on ergodicity in nonhomogeneous Markov chains", *Annals of Math. Stat.*, vol. 43, no. 5, 1732-1736, 1972.

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

[23] A. Kleywegt, A. Shapiro and T. Homem-de-Mello, "The sample average approximation method for stochastic discrete optimization", to appear in *SIAM Journal on Optimization.*

[24] R.W. Madsen and D.L. Isaacson, "Strongly ergodic behavior for non-stationary Markov processes", *Annals of Probability*, vol. 1, no. 2, 329-335, 1973.

[25] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, "Equation of state calculations by fast computing machines", *J. Chem. Phys.*, 21, 1087-1092, 1953.

[26] D. Mitra, F. Romeo and A. Sangiovanni-Vicentelli, "Convergence and finite-time behavior of simulated annealing", *Adv. Appl. Prob.* 18, 747-771, 1986.

[27] D.P. Morton and R.K. Wood, "On a stochastic knapsack problem and generalization", in *Interfaces in Computer Science and Operations Research*, D.L. Woodruff (ed.), Kluwer Academic Publishers, 1998.

[28] V.I. Norkin, G.Ch. Pflug and A. Ruszczynski, "A branch-and-bound method for stochastic global optimization", *Mathematical Programming*, vol. 83 (1998), no. 3, 425-450.

[29] E.L. Plambeck, B.R. Fu, S.M. Robinson and R. Suri, "Sample-path optimization of convex stochastic performance functions", *Mathematical Programming, Series B*, 75 (1996), 137-176.

[30] M.M. Rao, *Probability Theory with Applications*, Academic Press, 1984.

[31] S.M. Robinson, "Analysis of sample-path optimization", *Mathematics of Operations Research*, 21 (1996), 513-528.

[32] R.Y. Rubinstein and A. Shapiro, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. Wiley, New York, NY, 1993.

[33] E. Seneta, *Nonnegative Matrices and Markov Chains*, 2nd. ed., Springer-Verlag, New York, 1980.

[34] A. Shapiro, "Asymptotic analysis of stochastic programs," *Annals of Operations Research*, 30 (1991), 169-186.

[35] A. Shapiro, "Asymptotic behavior of optimal solutions in stochastic programming", *Mathematics of Operations Research*, 18 (1993), 829-845.

[36] A. Shapiro and T. Homem-de-Mello, "A simulation-based approach to stochastic programming with recourse", *Mathematical Programming*, vol. 81 (1998), no. 3, 301-325.

[37] A. Shapiro and T. Homem-de-Mello, "On the rate of convergence of Monte Carlo approximations of stochastic programs", *SIAM Journal on Optimization*, Vol. 11, No. 1, 70-86, 2000.

[38] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis*, Chapman & Hall, London, 1995.

[39] M. Van der Vlerk, *Stochastic programming bibliography*, available at `http://129.125.9.214/BIBLIO/STOPROG.HTML`.

[40] P. Van Laarhoven and E. Aarts, *Simulated Annealing: Theory and Applications*, Kluwer Academic Publishers, 1987.

[41] S. Yakowitz, P. L'Ecuyer and F. Vásquez-Abad, "Global stochastic optimization with low-dispersion point sets", to appear in *Operations Research*, 1998.

[42] D. Yan and H. Mukai, "Stochastic discrete optimization", *SIAM Journal of Control and Optimization*, 30, 594-612, 1992.