JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 91, No. 2, pp. 439–454, NOVEMBER 1996

Convergence Analysis of Gradient Descent Stochastic Algorithms

A. SHAPIRO[1] AND Y. WARDI[2]

Communicated by W. B. Gong

Abstract. This paper proves convergence of a sample-path based stochastic gradient-descent algorithm for optimizing expected-value performance measures in discrete event systems. The algorithm uses increasing precision at successive iterations, and it moves against the direction of a generalized gradient of the computed sample performance function. Two convergence results are established: one, for the case where the expected-value function is continuously differentiable; and the other, when that function is nondifferentiable but the sample performance functions are convex. The proofs are based on a version of the uniform law of large numbers which is provable for many discrete event systems where infinitesimal perturbation analysis is known to be strongly consistent.

Key Words. Gradient descent, subdifferentials, uniform laws of large numbers, infinitesimal perturbation analysis, discrete event dynamic systems.

1. Introduction

With the advent of sample-path gradient estimation techniques in discrete event dynamic systems, like infinitesimal perturbation analysis (IPA, Ref. 1) and likelihood ratio/score functions (Ref. 2), the question of simulation-based continuous-parameter optimization of steady-state performance functions has come to the fore. One of the main theoretical aspects of this question has been how to prove convergence of an iterate

[1] Professor, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia.

[2] Associate Professor, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia.

sequence computed by an algorithm to an optimal (or suboptimal, stationary, etc.) point with probability one (w.p.1). Besides variants of the stochastic approximation (SA) method, a number of gradient descent algorithms employing increasing precision have been considered. This paper concerns a class of such algorithms where the product of the stepsize at the kth iteration and the gradient of the performance criterion at the kth iterate, converges to 0 as $k \to \infty$.

To set the stage, let $f_k(\theta) = f_k(\theta, \omega), k = 1, \dots,$ be a sequence of real-valued random functions defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{R})$, with the parameter vector θ being confined to a set $\Theta \subset \mathbb{R}^d$. Suppose that, for any fixed $\theta \in \Theta$,

$$\lim_{k \to \infty} f_k(\theta) = f(\theta), \qquad \text{w.p.1}, \tag{1}$$

where $f(\theta)$ is a deterministic function of θ. We refer to the function $f(\theta)$ as the limiting function, and to $f_k(\theta)$, $k = 1, \dots,$ as a sequence of approximating functions. In a simulation-based optimization, the approximating functions f_k often are obtained by averaging a generated (simulated) sequence of sample performance functions. In that case, Eq. (1) means that the strong law of large numbers hold pointwise with the limiting function f typically being the expected value of the corresponding steady-state distribution.

We first assume that the approximating functions f_k are locally Lipschitz continuous and that the limiting function f is continuously differentiable. Such situations happen quite often in Monte Carlo simulations of discrete-event systems where the expectation operator smooths piecewise-differentiable sample performance functions (Ref. 1). Later, we will also consider nondifferentiable limiting functions, and focus our attention on the case where the approximating functions are convex. This situation can occur in queueing networks where, in fact, the limiting function lacks a derivative (gradient) at a dense subset of the parameter space (Ref. 3).

Recall that the generalized gradient $\partial h(\theta)$, in the sense of Clarke (Ref. 4), of a locally Lipschitz function h is the convext hull of all limits of the form $\lim_{n \to \infty} \nabla h(\theta_n)$, where $\{\theta_n\}$ can be any sequence converging to θ and such that h is differentiable at every point of that sequence and the above limit exists. Note that by Rademacher's theorem, the set of points where a locally Lipschitz function fails to be differentiable has Lebesgue measure zero. If the function h is convex, then the generalized gradient coincides with the subdifferential in the sense of convex analysis (Ref. 5).

Consider the optimization problem of minimizing $f(\theta)$ over Θ. Suppose that the limiting function $f(\theta)$ lacks a closed-form analytic expression, and consequently is estimated by the approximating functions $f_n(\theta)$. These functions and their derivatives can be considered as a simulation output that is used to optimize the limiting function f. The class of algorithm that we analyze in this paper has the following form:

$$\theta_{k+1} = \theta_k - a_k g_k, \tag{2}$$

where $\{\theta_k\}_{k=1}^{\infty}$ is the iterate sequence computed by the algorithm, $a_k > 0$ is the kth stepsize, and g_k is an element of the generalized gradient $\partial f_k(\theta_k)$.

To ease the exposition of the analysis, we implicitly assume that the constraint set Θ is compact and convex, and that the sequence $\{\theta_k\}_{k=1}^{\infty}$ stays in the interior of Θ.[3] Note that $g_k = \nabla f_k(\theta_k)$ if the function f_k is continuously differentiable at θ_k, and that the algorithm imposes no restriction on the way $g_k \in \partial f_k(\theta_k)$ is chosen if the generalized gradient $\partial f_k(\theta_k)$ is not a singleton. Note also that the size n_k of the sample used to generate the approximating function f_k can be determined a priori or can be random and correlated with the iterate sequence θ_k. It only has to satisfy the condition $\lim_{k\to\infty} n_k = \infty$, w.p.1, in order to ensure the law of large numbers (1).

The stepsizes a_k can be determined a priori or can be computed in an adaptive manner, but they have to be subjected to the following two conditions w.p.1:

$$\text{(i)} \lim_{k\to\infty} a_k \|g_k\| = 0, \qquad \text{(ii)} \sum_{k=1}^{\infty} a_k = \infty.$$

In case g_k are bounded, the above assumption (i) is ensured by the condition $\lim_{k\to\infty} a_k = 0$. If a_k are determined a priori, this last condition is almost the same as the assumption (i). However, if a_k are calculated in an adaptive manner, condition (i) means that we can have a_k bounded from below by a positive constant if g_k tend to zero.

When the limiting function $f(\cdot)$ is differentiable, an important technical condition under which convergence w.p.1 of the iterate sequence will be established is that the generalized gradients $\partial f_k(\theta)$ converge $\nabla f(\theta)$ w.p.1 uniformly on Θ. Although this condition appears to be strong, we will argue that it is satisfied in many cases of interest, including just about every case where convergence of IPA-based gradient methods was proved. In particular, if the functions f_k are convex, then such uniform convergence of the subgradients follows from the pointwise convergence (1) and the assumed differentiability of the limiting function f (cf. Ref. 6). An extensive discussion of this and related results can be found in Ref. 7, and further developments will be made below. When f is not differentiable, some

[3] The latter assumption can be relaxed by extending the forthcoming analysis to constrained algorithms.

analysis is still possible in the case where the approximating functions are convex. In this case, although it is not true that $\partial f_k(\cdot) \rightarrow \partial f(\cdot)$ uniformly w.p.1, we do have a uniform convergence of $\{f_k\}$ to f over compact sets; this, together with the special properties of convex functions and the analysis of deterministic algorithms (see, e.g., Refs. 8 and 9), will give us the desired convergence proof.

We would like to summarize some known results and place our work in the context of the recently published articles concerning sample-path optimization of discrete-event dynamic systems (DEDS). Shortly after the emergence of IPA, attention has been focused on proving convergence of gradient-descent algorithms for optimizing performance of GI/G/1 queues. Typically, the performance measures considered involved the average customer-delay as a function of a parameter of the service times' distributions. Most of the early works concerned variants of the stochastic approximation (SA) technique; see Ref. 10 and the references therein for a survey. In particular, we mention the pioneering works in Refs. 11 and 12, and extensions of the former reference to regenerative systems (Ref. 13), in which, a.s. convergence of SA algorithms to minima has been proved. We point out that, regarding this problem, the delays are convex functions of the parameter θ as long as the service times are convex, and this happens in most if not all of the specific situations that have arisen in the context of the works in Refs. 11 and 13. Moreover, extensions to serial queueing networks with or without blocking also give convex system times as long as the service times are convex (Ref. 14); hence, our algorithm is probably convergent.

Regarding algorithms close in spirit to the one discussed here, Bartusek and Makowski (Ref. 15) have proved convergence w.p.1 of a similar algorithm by using the large deviation theory. They impose the conditions that the state space of the underlying DEDS be a finite-state Markov chain, that the stepsize sequence be determined a priori, and that the sample size n_k grow to infinity at least as fast as log k. These restrictions are not made here; but the assumption of uniform convergence of the subdifferentials is not made in Ref. 15.

Large deviation theory also has been used by Dupuis and Simha (Ref. 16) to prove convergence of a steepest decent method with constant stepsizes. As in Ref. 15, they require n_k to grow to infinity faster than log k. The premises in this paper permit n_k to converge to ∞ at an arbitrarily slow rate, and our assumption that $\lim_{k\rightarrow\infty} a_k \|g_k\| = 0$ does not preclude the use of constant stepsizes as long as $\lim_{k\rightarrow\infty} g_k = 0$.

Convergence of descent algorithms that compute the stepsize by line minimization was proved in Ref. 7. That reference also contains a discussion on and a justification of some of the assumptions that are made here.

Earlier related works (Refs. 17–19) concern steepest descent algorithms with Armijo stepsizes, where a convergence concept slightly weaker tha a.s. convergence is proved.

Section 2 presents the main results, namely convergence proof of the basic algorithm (2). First, we treat the case where the limiting function is continuously differentiable [in the convex case differentiability suffices, as it implies continuity of the gradients (Ref. 5)]. We then discuss the convex case, where the limiting function is not necessarily differentiable. This case can be quite important in many situations in light of the results derived in Ref. 3, ascertaining that (convex) limiting functions could lack gradients at dense sets in the parameter space. Section 3 concerns the case where f is differentiable and it discusses the crucial assumption of uniform convergence of the generalized gradients of the approximating functions to the gradient of the limiting function. Finally, Section 4 concludes the paper.

2. Convergence Results

We discuss in this section convergence properties of the considered algorithm for two cases. First, when the approximating functions are Lipschitz continuous (not necessarily differentiable or convex) and the limiting function is continuously differentiable; second, when the approximating functions are convex and the limiting function is convex but not necessarily differentiable. The arguments which we use in this section basically are deterministic. The obtained results can be easily translated into the statistical language by adding, in the assumptions as well as in the conclusions, the words "with probability one." That is, we view the assumptions and the derived implications in this section as holding for $\mathbb{P}$-almost every realization $\omega \in \Omega$.

To begin with, we will use the following assumptions.

Assumption 2.1. The approximating functions f_k are Lipschitz continuous and the limiting function f is continuously differentiable on Θ; the set Θ is convex, compact, and has a nonempty interior; and the calculated iterate points Θ_k, $k = 1, \dots,$ stay in the interior of Θ.

Assumption 2.2. The generalized gradients $\partial f_k(\theta)$ converge to $\nabla f(\theta)$ uniformly on Θ, that is,

$$\lim_{k \to \infty} \sup_{\theta \in \Theta} \sup_{v \in \partial f_k(\theta)} \|v - \nabla f(\theta)\| = 0. \tag{3}$$

Note that, by the definition of the generalized gradients, condition (3) is equivalent to

$$\lim_{k\to\infty} \sup_{\theta\in\Theta\setminus E_k} \|\nabla f_k(\theta) - \nabla f(\theta)\| = 0, \tag{4}$$

with E_k being the set of those $\theta\in\Theta$ where $\nabla f_k(\theta)$ fails to exist; in the stochastic case where $f_k(\theta) = f_k(\theta, \omega)$, $E_k = E_k(\omega)$ is generally a function of ω. Moreover, by the theory of generalized gradients (see Ref. 4), the set E_k in Eq. (4) can be enlarged to any set of Lebesgue measure zero.

As we mentioned in the introduction, the stepsizes a_k can be defined a priori or can be calculated in an adaptive manner as a function of the generated sample. We make the following assumptions about the stepsizes.

Assumption 2.3. $\lim_{k\to\infty} a_k \|g_k\| = 0$.

Assumption 2.4. $\sum_{k=1}^{\infty} a_k = \infty$.

Let S denote the set of stationary points of f over Θ, i.e.,

$$S = \{\theta\in\Theta\colon\ \nabla f(\theta) = 0\}.$$

The next assumption concerns the structure of the set S. We believe that it can be relaxed, but at the expense of greater complexity of the arguments involved with the proof.

Assumption 2.5. There exists a finite number of closed sets $S_i \subset \Theta$, $i = 1, 2, \ldots, q$, such that $S = S_1 \cup \cdots \cup S_q$, and such that $f(\theta)$ is constant on every set S_i, $i = 1, \ldots, q$; i.e., there are numbers α_i such that $f(\theta) = \alpha_i$ for any $\theta\in S_i$, $i = 1, \ldots, q$.

Note that we can assume that all of the numbers α_i, $i = 1, \ldots, q$ are different from each other.

Theorem 2.1. Suppose that Assumptions 2.1–2.5 hold. Then, there exists an $l\in\{1, \ldots, q\}$ such that every accumulation point of the sequence $\{\theta_k\}$ belongs to S_l.

Proof. By the mean-value theorem and (2), we have that

$$f(\theta_{k+1}) - f(\theta_k) = (\theta_{k+1} - \theta_k)^T \nabla f(\hat{\theta}_k) = -a_k g_k^T \nabla f(\hat{\theta}_k), \tag{5}$$

for some point $\hat{\theta}_k$ on the segment joining θ_k and θ_{k+1}. Since $\nabla f(\theta)$ is continuous, it is uniformly continuous on the compact set Θ. By Eq. (2) and Assumption 2.3, we have then that

$$\|\nabla f(\hat{\theta}_k) - \nabla f(\theta_k)\| \to 0, \qquad \text{as } k \to \infty,$$

and by Assumption 2.2,

$$\|g_k - \nabla f(\theta_k)\| \to 0.$$

Consequently and by (5), for every $\delta > 0$ there exists K such that, for every $k \geq K$, if $\|\nabla f(\theta_k)\| \geq \delta$, then

$$f(\theta_{k+1}) - f(\theta_k) \leq -a_k \delta^2/2. \tag{6}$$

Let U be an open neighborhood of the set S. Clearly, the set $\Theta \backslash U$ does not contain any stationary points of f; hence,

$$\nabla f(\theta) \neq 0, \qquad \text{for all } \theta \in \Theta \backslash U.$$

Moreover, the set $\Theta \backslash U$ is compact; hence, there exists a constant $\delta > 0$ such that

$$\|\nabla f(\theta)\| \geq \delta, \qquad \text{for all } \theta \in \Theta \backslash U.$$

It follows then from (6) that the sequence $\{\theta_k\}$ must have an infinite number of points inside the neighborhood U; for if not, by (6) and Assumption 2.4, $f(\theta_k) \to -\infty$, contradicting the assumption that the iterate sequence stays in Θ and f is continuous, and hence bounded on Θ. Since U is an arbitrary neighborhood of S, we obtain that at least one accumulation point of $\{\theta_k\}$ belongs to the set S.

Let us arrange the sets S_i in such a way that $\alpha_1 > \cdots > \alpha_q$, where $\alpha_i = f(S_i)$, $i = 1, \ldots, q$. Denote by A the set of accumulation points of $\{\theta_k\}$, and consider

$$l = \max\{i \colon S_i \cap A \neq \varnothing, 1 \leq i \leq q\}.$$

Such an l exists since $S \cap A \neq \varnothing$. We next show that every accumulation point of the iterate sequence must be in S_l. We argue by contradiction. Suppose that there exists an accumulation point $\theta^* \in A$ such that $\theta^* \notin S_l$. We consider three different cases, namely when $f(\theta^*)$ is smaller, larger, or equal to α_l. Consider the sets

$$R = S_1 \cup \cdots \cup S_{l-1} \quad \text{and} \quad T = S_{l+1} \cup \cdots \cup S_q.$$

Note that, by the construction, $T \cap A = \varnothing$. Let U_1, U_2, U_3 be open neighborhoods of the sets R, S_l, T, respectively, such that

$$f(\theta_1) > f(\theta_2) > f(\theta_3), \qquad \text{for any } \theta_1 \in U_1,\ \theta_2 \in U_2,\ \theta_3 \in U_3,$$

and let $U = U_1 \cup U_2 \cup U_3$.

Suppose that $f(\theta^*) < \alpha_l$. We choose the neighborhoods U_1, U_2, U_3 in such a way that $\theta^* \notin U$, that $f(\theta^*) + \epsilon < f(\theta)$ for all $\theta \in U_2$ and some $\epsilon > 0$, and that, for some K and all $k \geq K$, $\theta_k \notin U_3$. By (6), we can choose K large enough such that, if $k \geq K$ and $\theta_k \in \Theta \backslash U$, then $f(\theta_{k+1}) < f(\theta_k)$. Moreover, since $\theta^* \in A$, there exists an $n \geq K$ such that $f(\theta_n) \leq f(\theta^*) + \epsilon$. It follows then by induction that $f(\theta_m) \leq f(\theta^*) + \epsilon$; hence, $\theta_m \notin U_2$ for all $m \geq n$. This, of course, contradicts the existence of an accumulation point in S_l.

Suppose that $f(\theta^*) > \alpha_l$. Let us choose $\epsilon > 0$ and the neighborhoods U_1, U_2, U_3 in such a way that $\theta^* \notin U_2$, that $\alpha_l + \epsilon < f(\theta^*)$, that $f(\theta) < \alpha_l + \epsilon/2$ for all $\theta \in U_2$, and that $\alpha_l + \epsilon < f(\theta)$ for all $\theta \in U_1$. By Assumption 2.3 and Eq. (6), we can choose K large enough that, if $k \geq K$, then $|f(\theta_{k+1}) - f(\theta_k)| < \epsilon/2$, and if $k \geq K$ and $\theta_k \in \Theta \backslash U$, then $f(\theta_{k+1}) < f(\theta_k)$. Moreover, since $S_l \cap A \neq \emptyset$, there exists an $n \geq K$ such that $\theta_n \in U_2$. It follows then by induction that $f(\theta_m) \leq \alpha_l + \epsilon$ for all $m \geq n$; hence, θ^* cannot be an accumulation point, a contradiction.

Suppose that $f(\theta^*) = \alpha_l$. We show then that there exists an accumulation point $\theta' \in A$ such that $f(\theta') < f(\theta^*)$. This will bring us to the case already considered, and the proof will be completed. Denote by $B(\theta, r)$ the open ball of radius $r > 0$ and centered at θ. Since $\nabla f(\theta^*) \neq 0$, we can choose numbers $\gamma > 0$, $\delta > 0$, and K such that $B(\theta^*, 3\gamma) \cap S = \emptyset$, and the inequality $g_k^T \nabla f(\theta) \geq \delta \|g_k\|$ holds for any $\theta \in B(\theta^*, 3\gamma)$ and $k \geq K$ such that $\theta_k \in B(\theta^*, 3\gamma)$. Also, let $n \geq K$ be such that $\theta_n \in B(\theta^*, \gamma)$, and let $s(n)$ be a positive integer such that $\theta_{n+s(n)}$ first time leaves the neighborhood $B(\theta^*, 2\gamma)$; i.e., $\theta_{n+i} \in B(\theta^*, 2\gamma)$ for all $i = 1, \ldots, s(n) - 1$, and $\theta_{n+s(n)} \notin B(\theta^*, 2\gamma)$. Note that such an $s(n)$ does exist because $A \cap S \neq \emptyset$. We obtain then that

$$\sum_{i=n}^{n+s(n)-1} a_i \|g_i\| \geq \|\theta_{n+s(n)} - \theta_n\| \geq \gamma.$$

By (5), it follows that, for n large enough, if $\theta_n \in B(\theta^*, \gamma)$, then

$$f(\theta_{n+s(n)}) - f(\theta_n) \leq -\delta \sum_{i=n}^{n+s(n)-1} a_i \|g_i\| \leq -\delta\gamma.$$

By compactness arguments, this implies the existence of an accumulation point $\theta' \in A$ such that $f(\theta') \leq f(\theta^*) - \delta\gamma$. The proof is complete. □

Remark 2.1. It follows that, if in addition to the assumptions of Theorem 2.1, the set S is finite, then the sequence $\{\theta_k\}$ converges to a point $\theta^* \in S$.

Consider now the situation where the approximating functions are convex. Recall that the ϵ-subdifferential, $\epsilon \geq 0$, of a convex function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ at a point θ_0 is defined by

$$\partial_\epsilon f(\theta_0) = \{v \in \mathbb{R}^d : f(\theta) - f(\theta_0) \geq v^T(\theta - \theta_0) - \epsilon, \forall \theta \in \mathbb{R}^d\}.$$

For $\epsilon = 0$, the corresponding ϵ-subdifferential becomes the subdifferential of f at θ_0; see Refs. 9, 20, and 21 for a discussion of ϵ-subdifferentials.

Suppose that the functions $f_k(\theta)$ are convex on $\mathbb{R}^d$, and consider the iteration procedure (2) with $g_k \in \partial_{\epsilon_k} f_k(\theta_k)$, where $\epsilon_k \downarrow 0$ and g_k can be any point in the above ϵ_k-subdifferential. We need the following technical result from convex analysis.

Lemma 2.1. Let f_k: $\mathbb{R}^d \to \mathbb{R}$ be a sequence of convex functions converging pointwise to a function f: $\mathbb{R}^d \to \mathbb{R}$; i.e., $\lim_{k\to\infty} f_k(\theta) = f(\theta)$, for any $\theta \in \mathbb{R}^d$. Suppose that the set $S = \arg\min_{\theta \in \mathbb{R}^d} f(\theta)$ is nonempty and bounded. Then, for any neighborhood U of S, there exist positive constants ϵ and K such that $f_k(\theta) - f_k(\theta^*) \geq \epsilon$, for any $k \geq K$, $\theta \in \mathbb{R}^d \backslash U$, and $\theta^* \in S$.

Proof. Note that the limiting function f is convex and hence continous. Therefore, the set S is convex and compact. Let us fix a point $\theta^* \in S$. Since S is bounded there exists a number r such that $\|\theta - \theta^*\| < r$ for all $\theta \in S$. Consider

$$B(\theta^*, r) = \{\theta: \|\theta - \theta^*\| \leq r\}.$$

Since f_k are convex and converge pointwise to f, we have that f_k converge to f uniformly on the compact set $B(\theta^*, r)$; see Ref. 5. It follows then by the standard arguments of compactness that there is an $\epsilon > 0$ such that $f_k(\theta) - f_k(\theta^*) \geq \epsilon$, for all $\theta \in B(\theta^*, r) \backslash U$ and all k large enough. Also, we can choose ε in a such way that $f_k(\theta) - f_k(\theta^*) \geq \epsilon$ for all θ on the sphere $\{\theta: \|\theta - \theta^*\| = r\}$ and all k large enough. By convexity of f_k, this implies that

$$f_k(\theta) - f_k(\theta^*) \geq \epsilon, \qquad \text{for all } \theta \in \mathbb{R}^d \backslash B(\theta^*, r);$$

hence, the proof is complete. □

The following theorem and its proof are an immediate extension of the corresponding deterministic result; see, e.g., Ref. 9, Chapter 3, Section 4.

Theorem 2.2. Suppose that:

(i) $\lim_{k\to\infty} a_k |g_k|^2 = 0$ and the sequence $\{a_k\}$ is bounded from above;

(ii) $\sum_{k=1}^{\infty} a_k = \infty$;

(iii) the functions f_k are convex on $\mathbb{R}^d$ and converge pointwise to a function f: $\mathbb{R}^d \to \mathbb{R}$;

(iv) the set $S = \arg\min_{\theta \in \mathbb{R}^d} f(\theta)$ is nonempty and bounded;

(v) $\epsilon_k \downarrow 0$.

Then,

$$\lim_{k \to \to \infty} \operatorname{dist}(\theta_k, S) = 0. \tag{7}$$

Proof. By the assumption of convexity and since $g_k \in \partial_{\epsilon_k} f_k(\theta_k)$, for all $k = 1, 2, \ldots$, we have the following inequality (e.g. Lemma 1.1 in Ref. 8): for all $\theta^* \in S$,

$$\|\theta_{k+1} - \theta^*\|^2 \leq \|\theta_k - \theta^*\|^2 + a_k(a_k \|g_k\|^2 + 2[f_k(\theta^*) - f_k(\theta_k) + \epsilon_k]). \tag{8}$$

It follows from this inequality that the sequence $\{\theta_k\}$ has at least one accumulation point in the set S. Indeed, suppose that the above statement is not true. Then, there is a neighborhood U of S such that $\theta_k \notin U$ for k large enough. Let us fix a point $\theta^* \in S$. Because of the assumptions (i) and (v) and by the result of Lemma 2.1, it follows then that there exists $\epsilon > 0$ such that

$$a_k \|g_k\|^2 + 2[f_k(\theta^*) - f_k(\theta_k) + \epsilon_k] \leq -\epsilon, \tag{9}$$

for all sufficiently large k. By summing up both sides of Inequality (8) and using the assumption (ii) and Inequality (9), we obtain that eventually, for n large enough, $\|\theta_n - \theta^*\|^2$ should become negative, which of course is a contradiction.

Let us show now that (7) holds. For a given $\delta > 0$, consider the neighborhoods

$$U_1 = \{\theta: \operatorname{dist}(\theta, S) < \delta\}, \qquad U_2 = \{\theta: \operatorname{dist}(\theta, S) < 2\delta\}$$

of S, and let $\theta^* \in S$. Suppose that $\theta_k \in U_2 \backslash U_1$. Then, for k large enough and for any $\theta^* \in S$, Inequality (9) holds. Together with (8), this implies that

$$\|\theta_{k+1} - \theta^*\|^2 \leq \|\theta_k - \theta^*\|^2. \tag{10}$$

Note also that it follows from the assumption (i) that $\lim_{k \to \infty} a_k \|g_k\| = 0$. It follows then by induction that if $\theta_k \in U_2$, then $\theta_{k+1} \in U_2$ for all k large enough. Since $\{\theta_k\}$ has an accumulation point in S, we obtain that $\theta_k \in U_2$ for all sufficiently large k. Since δ is arbitrary, this completes the proof. □

3.. Uniform Convergence of the Generalized Gradients

We now return to the case where f is differentiable and discuss Assumption 2.2, which was crucial for deriving the first convergence result. The above assumption is satisfied whenever the approximating functions are convex and the limiting function is differentiable; see Refs. 5 and 6. We further establish this assumption for regenerative processes, where the functions involved are not necessarily convex.

An immediate extension of the convex case is to consider the composite functions $f_k(\theta) = h_k(A(\theta))$, where $A: \mathbb{R}^d \to \mathbb{R}^m$ is a continuously differentiable deterministic mapping and $h_k: \mathbb{R}^m \to \mathbb{R}$ are random functions that are convex w.p.1. If $h_k(\cdot)$ converge pointwise w.p.1 to a differentiable function $h(\cdot)$, and hence $f_k(\theta) \to f(\theta) = h(A(\theta))$ w.p.1, then the uniform convergence (3) follows by the chain rule from the uniform convergence w.p.1 of the subdifferentials $\partial h_k(\cdot)$ to $\nabla h(\cdot)$

We next consider a general setup of regenerative processes. Let $X_i(\theta) = X_i(\theta, \omega)$, $i = 1, 2, \ldots$, be a sequence of random functions defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$. For every $\omega \in \Omega$, we view $X_i(\cdot, \omega)$, $i = 1, 2, \ldots$, as sample paths of the considered process. The approximating functions are defined then by averaging

$$f_k(\theta) = k^{-1} \sum_{i=1}^{k} X_i(\theta). \tag{11}$$

It is possible to define the approximating functions f_k by averaging with respect to a sequence of sample sizes n_k tending to infinity as $k \to \infty$, as this will not change the subsequent convergence analysis. Let us make the following assumptions.

Assumption 3.1. For any fixed $\theta \in \Theta$, the process $X_1(\theta), X_2(\theta), \ldots$, is regenerative, with regenerative cycles of generic length $\eta(\theta)$ and finite expectations $\mathbb{E}\{|\sum_{i=1}^{\eta(\theta)} X_i(\theta)|\}$ and $\mathbb{E}\{\eta(\theta)\}$.

Assumption 3.2. For $\mathbb{P}$-almost every ω, the functions $X_i(\cdot, \omega)$, $i = 1, \ldots$, are Lipschitz continuous on Θ.

Assumption 3.3. For any fixed $\theta \in \Theta$ and for $\mathbb{P}$-almost every ω, the functions $X_i(\cdot, \omega)$, $i = 1, \ldots$, are continuously differentiable at θ.

By Assumption 3.3, the gradients $G_i(\theta) = \nabla X_i(\theta)$ exist w.p.1; hence, we can consider the vector-valued process $G_1(\theta), G_2(\theta), \ldots$.

Assumption 3.4. For any fixed $\theta \in \Theta$, the process $G_1(\theta), G_2(\theta), \ldots$ is regenerative with regenerative cycles of length $\tau(\theta) = \tau(\theta, \omega)$.

Assumption 3.5. The expectation $\mathbb{E}\{\sup_{\theta \in \Theta} \tau(\theta)\}$ is finite, and for any fixed $\theta \in \Theta$ and for $\mathbb{P}$-almost every ω, the function $\tau(\cdot, \omega)$ is continuous at θ; i.e., there is a neighborhood of θ, depending on ω, in which $\tau(\cdot, \omega)$ is constant.

Assumption 3.6. The expectation $\mathbb{E}\{\sup_{\theta\in\Theta} \|\sum_{i=1}^{\tau(\theta)} G_i(\theta)\|\}$ is finite.

By the well-known law of large numbers for regenerative processes, it follows from Assumption 3.1 that the average functions $f_k(\theta)$ converge pointwise w.p.1 as $k \to \infty$ to a deterministic function $f(\theta)$. We show now that, under the above assumptions, the function $f(\theta)$ is continuously differentiable, and the gradients $\nabla f_k(\theta)$ converge w.p.1 to $\nabla f(\theta)$ uniformly on Θ. This will extend a similar result for the iid case considered in Proposition 2.2 of Ref. 22, and a result concerning uniform strong consistency of IPA for the waiting times in GI/G/1 queues (Ref. 23), to the present regenerative setting.

Theorem 3.1. Suppose that $\Theta \subset \mathbb{R}^d$ is a convex, compact set with nonempty interior, and that Assumptions 3.1–3.6 hold. Then, the function $f(\theta)$ is continuously differentiable on Θ, and the gradients $\nabla f_k(\theta)$ converge w.p.1, as $k \to \infty$, to $\nabla f(\theta)$ uniformly on Θ; i.e., Eq. (3) and its equivalent (4) hold.

Remark 3.1. The above assumptions, or variants thereof, have become common in the literature on perturbation analysis for proving the strong consistency of the IPA estimators, i.e., that $\nabla f_k(\theta) \to \nabla f(\theta)$ as $k \to \infty$ w.p.1; see, e.g., Refs. 1 and 24, and references therein. What is new in the present result are the proofs of uniform convergence and continuity of the gradient ∇f.

The proof will make use of the following auxiliary result.

Lemma 3.1. Let $\Theta \subset \mathbb{R}^d$ be a convex set with nonempty interior, and let $\{f_k(\theta)\}$ be a sequence of real-valued, Lipschitz-continuous (deterministic) functions converging pointwise on Θ to a function $f(\theta)$. Suppose that there exists a continuous vector-valued function $Z(\theta)$ such that

$$\lim_{k \to \infty} \sup_{\theta \in \Theta \setminus E_k} \|\nabla f_k(\theta) - Z(\theta)\| = 0, \tag{12}$$

where E_k is the set of those points of Θ where $\nabla f_k(\theta)$ fails to exist. Then, $f(\theta)$ is differentiable on Θ and $\nabla f(\theta) = Z(\theta)$ for all $\theta \in \Theta$.

Proof. Fix θ in the interior of Θ, and consider another point $\bar{\theta} \in \Theta$. By the mean-value theorem for Lipschitz continuous functions (Ref. 4), we have

$$f_k(\bar{\theta}) - f_k(\theta) = g_k^T(\bar{\theta} - \theta), \tag{13}$$

where $g_k \in \partial f_k(\hat{\theta}_k)$ for some point $\hat{\theta}_k$ on the segment joining θ and $\bar{\theta}$. By passing to the limit as $k \to \infty$ and exploiting the uniform convergence condition (12), we obtain

$$f(\bar{\theta}) - f(\theta) = Z(\theta)^T(\bar{\theta} - \theta) + r(\bar{\theta}), \tag{14}$$

where

$$|r(\bar{\theta}| \le \|\bar{\theta} - \theta\| \sup_{\hat{\theta} \in [\theta, \bar{\theta}]} \|Z(\hat{\theta}) - Z(\theta)\|.$$

Since $Z(\cdot)$ is continuous at θ, it follows that $r(\bar{\theta}) = o(\|\bar{\theta} - \theta\|)$. Together with (14), this implies that f is differentiable at θ and $\nabla f(\theta) = Z(\theta)$. □

Proof of Theorem 3.1. Consider the process

$$Z_k(\theta) = k^{-1} \sum_{i=1}^{k} G_i(\theta).$$

By the renewal theory of regenerative processes (e.g., Ref. 25), it follows from Assumptions 3.4–3.6 that $Z_k(\theta)$ converge pointwise w.p.1 as $k \to \infty$ to a deterministic vector-valued function $Z(\theta)$, which can be written in the form

$$Z(\theta) = \mathbb{E}\left\{ \sum_{i=1}^{\tau(\theta)} G_i(\theta) \right\} / \mathbb{E}\{\tau(\theta)\}. \tag{15}$$

Let us observe that both functions

$$g_1(\theta) = \mathbb{E}\left\{ \sum_{i=1}^{\tau(\theta)} G_i(\theta) \right\} \quad \text{and} \quad g_2(\theta) = \mathbb{E}\{\tau(\theta)\},$$

and hence the function $Z(\theta)$, are continuous functions of θ. Indeed, by the Lebesgue dominated convergence theorem, it follows from Assumption 3.6 that

$$\lim_{\theta' \to \theta} \mathbb{E}\left\{ \sum_{i=1}^{\tau(\theta')} G_i(\theta') \right\} = \mathbb{E}\left\{ \lim_{\theta' \to \theta} \sum_{i=1}^{\tau(\theta')} G_i(\theta') \right\}.$$

Because of the almost sure continuity of $\tau(\cdot)$ and $G_i(\cdot)$ (Assumptions 3.3 and 3.5), the limit inside the expectation on the right-hand side of the above equation is equal to $\sum_{i=1}^{\tau(\theta)} G_i(\theta)$ w.p.1; hence, the corresponding expectation is equal to $g_1(\theta)$. This shows that $g_1(\cdot)$ is continuous at θ. Similarly, continuity of $g_2(\cdot)$ follows from Assumption 3.5.

Let $\tau_1(\theta), \tau_2(\theta), \ldots$ be the lengths of the regenerative cycles of the process $G_i(\theta)$, and let

$$\sigma_m(\theta) = \tau_1(\theta) + \cdots + \tau_m(\theta).$$

By a uniform version of the law of large numbers for iid processes (e.g., Ref. 2, pp. 67–69), Assumptions 3.3–3.6 imply that $m^{-1}\sum_{i=1}^{\sigma_m(\theta)} G_i(\theta)$ converge w.p.1, as $m \to \infty$, to $g_1(\theta)$ uniformly on Θ. Similarly, it follows from Assumption 3.5 that $m^{-1}\sigma_m(\theta)$ converge w.p.1, as $m \to \infty$, to $g_2(\theta)$ uniformly on Θ. Consequently, $\sigma_m(\theta)^{-1}\sum_{i=1}^{\sigma_m(\theta)} G_i(\theta)$ converge w.p.1 to $Z(\theta)$ uniformly on Θ. By standard arguments of the renewal theory, it follows that $Z_n(\theta)$ converge w.p.1, as $n \to \infty$, to $Z(\theta)$ uniformly on Θ.

Finally, by noting that $Z_k(\theta) = \nabla f_k(\theta)$ whenever the latter derivative exists, it follows directly from Lemma 3.1 that $\nabla f(\theta)$ exists and is equal to $Z(\theta)$. □

4. Conclusions

This paper has presented a convergence analysis for a stochastic, simulation-based algorithm for optimization of expected-value performance measures in discrete event systems. The algorithm is of the gradient-descent type, and it requires that the distance between two consecutive iteration points converge to zero as the iterate count goes to infinity.

In the case where the limiting function is differentiable, convergence of the iterate sequence to stationary points has been established by fairly simple arguments, based on the assumed uniform law of large numbers concerning the functions and their generalized gradients. This assumption was shown to hold when the approximating functions are convex, and otherwise, in many systems where IPA had been known to be strongly consistent. In the case where the limiting function is nondifferentiable but the approximating functions are convex, the algorithm's convergence directly follows from the established theory of deterministic optimization.

References

1. Ho, Y. C., and Cao X. R., *Perturbation Analysis of Discrete Event Dynamic Systems*, Kluwer Academic Publishers, Boston, Massachusetts, 1991.
2. Rubinstein, R. Y., and Shapiro, A., *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*, John Wiley and Sons, New York, New York, 1993.
3. Shapiro, A., and Wardi, Y., *Nondifferentiability of the Steady-State Function in Discrete Event Dynamic Systems*, IEEE Transactions on Automatic Control, Vol. 39, pp. 1707–1711, 1994.
4. Clarke, F. H., *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, New York, 1983.

5. ROCKEFELLAR, R. T., *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.
6. ROBINSON, S. M., *Convergence of Subdifferentials under Strong Stochastic Convexity*, Management Science (to appear).
7. SHAPIRO, A., and WARDI, Y., *Convergence Analysis of Stochastic Algorithms*, Mathematics of Operations Research (to appear).
8. CORREA, R., and LEMARÉCHAL, C., *Convergence of Some Algorithms for Convex Minimization*, Mathematical Programming, Vol. 62, pp. 261–275, 1993.
9. DEMYANOV, V. F., and VASILEV, L. V., *Nondifferentiable Optimization*, Optimization Software, Publications Division, New York, New York, 1985.
10. FU, M. C., *Optimization via Simulation: A Review*, Annals of Operations Research, Vol. 53, pp. 199–247, 1994.
11. CHONG, E. K. P., and RAMADGE, P. J., *Optimization of Queues Using Infinitesimal Perturbation Analysis-Based Stochastic Algorithm with General Update Times*, SIAM Journal on Control and Optimization, Vol. 31, pp. 698–732, 1993.
12. L'ECUYER, P., and GLYNN, P., *Stochastic Optimization by Simulation: Convergence Proofs for the GI/G1 Queue in Steady State*, Management Science, Vol. 40, pp. 1562–1578, 1994.
13. CHONG, E. K. P., and RAMADGE, P. J., *Stochastic Optimization of Regenerative Systems Using Infinitesimal Perturbation Analysis*, IEEE Transactions on Automatic Control, Vol. 39, pp. 1400–1410, 1994.
14. SHANTIKUMAR, J. G., and YAO, D. D., *Second-Order Stochastic Properties of Queueing Systems*, Proceedings of the IEEE, Vol. 77, pp. 162–170, 1989.
15. BARTUSEK, J. D., and MARKOWSKI, A. M., *On Stochastic Approximations Driven by Sample Averages: Convergence Results via the ODE Method*, Manuscript, Institute for Systems Research, University of Maryland, 1993.
16. DUPUIS, P., and SIMHA, R., *On Sampling Controlled Stochastic Approximation*, IEEE Transactions on Automatic Control, Vol. 36, pp. 915–924, 1991.
17. MEHESHWARI, S., and MUKAI, H., *An Optimization Algorithm Driven by Probabilistic Simulation*, Proceedings of the Conference on Decision and Control, Athens, Greece, pp. 1703–1705, 1986.
18. YAN, D., and MUKAI, H., *An Optimization Algorithm with Probabilistic Simulation*, Journal of Optimization Theory and Applications, Vol. 79, pp. 345–371, 1993.
19. WARDI, Y., *Stochastic Algorithms with Armijo Stepsizes for Minimization of functions*, Journal of Optimization Theory and Applications, Vol. 64, pp. 399–417, 1990.
20. HIRIART-URRUTY, J. B., and LEMARÉCHAL, C., *Convex Analysis and Minimization Algorithms, Part 1*, Springer Verlag, Berlin, Germany, 1993.
21. HIRIART-URRUTY, J. B., and LEMARÉCHAL, C., *Convex Analysis and Minimization Algorithms, Part 2*, Springer Verlag, Berlin, Germany, 1993.
22. SHAPIRO, A., *Asymptotic Properties of Statistical Estimators in Stochastic Programming*, Annals of Statistics, Vol. 17, pp. 841–858, 1989.

23. WARDI, Y., *Interchangeability of Expectation and Differentiation of Waiting Times in GI/G1 Queues*, Stochastic Processes and Their Applications, Vol. 45, pp. 141-154, 1993.
24. GLASSERMAN, P., *Gradient Estimation via Perturbation Analysis*, Kluwer Academic Publishers, Boston, Massachusetts, 1991.
25. ASMUSSEN, S., *Applied Probability and Queues*, John Wiley and Sons, New York, New York, 1987.