

CONVERGENCE ANALYSIS OF STOCHASTIC ALGORITHMS

A. SHAPIRO AND Y. WARDI

This paper investigates asymptotic convergence to stationary points of gradient descent algorithms where the functions involved are not available in closed form but are approximated by sequences of random functions. The algorithms take large stepsizes and use progressively finer precision at successive iterations. Two kinds of optimization algorithms are considered: one, where the limiting function (to be minimized) is differentiable but the random approximating functions can be nondifferentiable, and the other, where both the limiting and approximating functions are nondifferentiable and convex. Such functions often arise in discrete event dynamic system-models in various application areas.

The analysis brings together ideas and techniques from the disciplines of nonlinear programming and nondifferentiable optimization on one hand, and stochastic processes and especially uniform laws of large numbers, on the other hand. A general algorithmic framework is first developed and analyzed, and then applied to the specific algorithms considered. The analysis extends the results derived to date for similar algorithms, and has a potential for further extensions in proving convergence of other algorithms as well.

1. Introduction. This paper analyzes convergence properties of nonlinear programming algorithms for optimization problems where the involved functions (and their derivatives) are evaluated by a Monte Carlo simulation. The paper's focus is on descent methods, but the developed concepts and techniques may prove to be useful in investigating other algorithms as well.

Our main motivation comes from optimization of expected-value performance measures (functions) in discrete event dynamic systems (DEDS), like queueing networks. Often lacking closed-form expressions, such functions have to be estimated, say by a Monte Carlo simulation. Sensitivity measures like gradients, or subgradients, of such functions can be evaluated by the perturbation analysis (PA) (Ho and Cao, 1991) or the score function (SF) (Rubinstein and Shapiro, 1993) methods. In aiming at a greater generality, the discussion in this paper is in terms of general stochastic systems not necessarily having the special structure of DEDS.

Stochastic approximation (SA) has been a widely-used technique for simulation-based optimization; see Kushner and Clark (1978), for a general discussion, and Fu (1994), and the references therein, for analysis and applications in DEDS. Due to the mandate for small stepsizes inherent in SA, alternative algorithmic techniques have been sought that use larger stepsizes and longer simulation times per iteration than SA. In particular two classes of descent algorithms were proposed in Dupuis and Simha (1991) and Simha and Kurose (1989) and in Meheshwari and Mukai (1986), Mukai and Yan (1991), Wardi (1990), Wardi and Lee (1991) and Yan and Mukai (1995), respectively. Both are based on descent against the direction of a simulated gradient, and they increase the simulation times at successive iterations to infinity. They differ in the employed stepsizes, constant stepsize in Dupuis and Simha (1991),

Received May 27, 1994; revised February 20, 1995.

AMS 1991 subject classification. Primary: 90C30.

OR/MS Index 1978 subject classification. Primary: 652 Programming/Nondifferentiable.

Key words. Gradient descent methods, stochastic optimization, nondifferentiable optimization, convex analysis, uniform laws of large numbers.

Simha and Kurose (1989), and Armijo stepsize in Meheshwari and Mukai (1986), Mukai and Yan (1991), Wardi (1990), Wardi and Lee (1991), Yan and Mukai (1995). Asymptotic convergence of these algorithms to stationary points has been analyzed in Dupuis and Simha (1991) and in Wardi (1990) and Yan and Mukai (1995), respectively, and numerical experiments on DEDS optimization problems have testified to their viability. The analysis in this paper includes convergence proofs for these algorithms, but its scope is considerably broader as it will be pointed out in the sequel. (Instead of the Armijo stepsize we consider precise line minimization, where the exposition is simpler while the main ideas are similar.)

Expected-value performance functions in DEDS can be *nondifferentiable* at a dense set of points in the parameter space (Shapiro and Wardi, 1994), and therefore, algorithms for their optimization may have to be based on nondifferentiable programming techniques. Indeed, Plambeck, Fu, Robinson and Suri (1995) have recently developed a bundle-type algorithm for convex, possibly nondifferentiable, problems based on PA, and applied it with notable success to optimization of production systems.

To set the stage for our discussion, let $f(\theta)$ be a real-valued function, and consider the problem of minimization of $f(\theta)$ over a set $\Theta \subset \mathbb{R}^m$. Let $\{f_n(\theta)\}$ be a sequence of stochastic real-valued functions viewed as approximations or estimates of $f(\theta)$. We assume that the approximating sequence $\{f_n(\theta)\}$ converges in some sense to the (deterministic) function $f(\theta)$ referred to as the *true* or *limiting* function. We also assume that the approximating functions are defined on a common probability space (Ω, \mathcal{F}, P) . That is, for a fixed θ , $f_n(\theta) = f_n(\theta, \omega)$ is considered as a random variable over the probability space (Ω, \mathcal{F}, P) . In order to keep the exposition as simple as possible we discuss in this paper the case where the feasible set Θ is fixed (deterministic). We believe that the main ideas can be extended to situations where the feasible set is defined by constraints which also should be approximated.

The above setup is quite common in simulation environments where the expected-value function $f(\theta)$ cannot be directly computed and it is estimated by a simulation output, $f_n(\theta)$. For a fixed $\omega \in \Omega$, the functions $f_n(\theta, \omega)$ are called realizations or *sample path* functions. The PA or SF methods can be used to compute the gradients (subgradients) $\nabla f_n(\theta)$ of the sample path functions which, in some cases, can serve as approximations to the gradients (subgradients) $\nabla f(\theta)$ of the limiting function.

In this paper we focus on *Markovian* type algorithms. The term "Markovian" is understood in the sense that the considered algorithm calculates the next iteration point θ_{n+1} by using only the current iteration point θ_n and the current approximation f_n , and it does not accumulate information from previous approximations. Please note that one can have $f_n = f_{n+1}$, i.e., one can use the same approximation in several steps of the algorithm.

The descent methods considered in Dupuis and Simha (1991), Meheshwari and Mukai (1986), Mukai and Yan (1991), Simha and Kurose (1989), Wardi (1990), Wardi and Lee (1991) and Yan and Mukai (1995) are Markovian, and hence our analysis applies to them. The bundle type algorithm developed in Plambeck, Fu, Robinson and Suri (1995), on the other hand, is not Markovian and our results cannot be applied to its analysis in a straightforward manner. Hopefully the general techniques developed in this paper can be extended to non-Markovian algorithms as well.

A broad class of descent-type algorithms can be written in the form

$$(1.1) \quad \theta_{n+1} = \theta_n - t_n g_n,$$

where $t_n \geq 0$ is a suitable stepsize and $(-g_n)$ is a direction of descent. For example, if the approximating functions are smooth and the considered problem is uncon-

strained, one can take $g_n = \nabla f_n(\theta_n)$. Let $M(n)$ denote the sample size used in the simulation for computing f_n . The main convergence result proved in Dupuis and Simha (1991) is that for a fixed stepsize $t_n = \alpha > 0$ and $g_n = \nabla f_n(\theta_n)$, the generated sequence of iteration points $\{\theta_n\}$ converges to the set of stationary points of f with probability one (w.p.1) if $M(n)$ grows to infinity at least at a rate of $\log(n)$. The convergence analyses in Wardi (1990) and Yan and Mukai (1995), for the case of Armijo stepsizes, require no lower bound on the rate at which $M(n) \rightarrow \infty$, but they fail to prove convergence w.p.1 to stationary points of f ; only a mildly weaker, and nonstandard convergence mode is proved there. Results of the analysis presented in this paper include convergence w.p.1 of both kinds of algorithms under no restrictions on the rate at which $M(n) \rightarrow \infty$, as well as extensions to some kinds of nondifferentiable programs.

The rest of the paper is organized as follows. Section 2 gives a convergence analysis for general descent-type algorithms. Section 3 contains mainly a survey of some results on uniform convergence w.p.1 of the sample-path functions to the limiting function. The described uniform versions of the Law of Large Numbers provide a basis for applications of the general result given in §2 to specific situations. In §4, we show that the required assumptions hold in several important cases arising in queueing networks. Namely, when the sample-path functions are nondifferentiable but the limiting function is differentiable, or when the sample-path functions as well as the limiting function are convex and nondifferentiable. Finally, §5 concludes the paper and suggests directions for future research.

2. Convergence of descent-type algorithms. In this section we give a general analysis of descent-type algorithms. Our analysis is based on the concept of descent functions. All functions are assumed to be deterministic in this section.

We assume that the considered functions belong to a functional space \mathcal{X} equipped with a norm $\|\cdot\|$. We denote a generic element of \mathcal{X} by ϕ in order to distinguish it from the limiting function f . For example, if we work with continuous functions and the convergence is uniform, we may use the space $\mathcal{X} = C(\Theta)$ of bounded continuous functions $\phi : \Theta \rightarrow \mathbb{R}$ with the sup-norm

$$\|\phi\| = \sup_{\theta \in \Theta} |\phi(\theta)|.$$

If, in addition, the functions are continuously differentiable, we can work in the space $C^1(\Theta)$ of continuously differentiable functions with the norm

$$\|\phi\| = \sup_{\theta \in \Theta} |\phi(\theta)| + \sup_{\theta \in \Theta} \|\nabla\phi(\theta)\|.$$

We also will use the space $\text{Lip}(\Theta)$ of Lipschitz continuous functions with the norm

$$\|\phi\| = \sup_{\theta \in \Theta} |\phi(\theta)| + \sigma(\phi),$$

where

$$\sigma(\phi) = \sup \left\{ \frac{|\phi(\theta_1) - \phi(\theta_2)|}{\|\theta_1 - \theta_2\|} : \theta_1, \theta_2 \in \Theta, \theta_1 \neq \theta_2 \right\}$$

is the Lipschitz constant of ϕ . (This space is denoted by $W^{1,\infty}(\Theta)$ in the literature on functional analysis.) Clearly $C^1(\Theta) \subset \text{Lip}(\Theta) \subset C(\Theta)$. Note that it follows from the

theory of the generalized gradient of Clarke (1983) that

$$\sigma(\phi) = \sup_{\theta \in \Theta \setminus E} \|\nabla\phi(\theta)\|,$$

where E can be any set of Lebesgue measure zero containing the set of points where $\nabla\phi(\theta)$ fails to exist.

We say that a sequence of points $\{\theta_n\}$ in \mathbb{R}^m is generated by a (Markovian) algorithm if the next iteration point θ_{n+1} depends only on the current iteration point θ_n and a current approximation function $f_n \in \mathcal{Z}$. That is, there exists a point-to-set mapping $\mathcal{A}: \mathcal{Z} \times \Theta \rightarrow 2^{\mathbb{R}^m}$ such that $\theta_{n+1} \in \mathcal{A}(f_n, \theta_n)$ for all n .

Let us recall now the following definition of a closed mapping.

DEFINITION 2.1. The point-to-set mapping \mathcal{A} is said to be *closed* at $(f, \theta) \in \mathcal{Z} \times \Theta$ if $(f_n, \theta_n) \rightarrow (f, \theta)$, $(f_n, \theta_n) \in \mathcal{Z} \times \Theta$ and $\xi_n \rightarrow \xi$, $\xi_n \in \mathcal{A}(f_n, \theta_n)$, imply $\xi \in \mathcal{A}(f, \theta)$. For a given $f \in \mathcal{Z}$ we say that \mathcal{A} is closed at f if it is closed at (f, θ) for every $\theta \in \Theta$, and \mathcal{A} is said to be closed on $\mathcal{Z} \times \Theta$ if it is closed at each point of $\mathcal{Z} \times \Theta$.

The following concept of a descent function is well known in nonlinear programming (e.g., Luenberger, 1989, p. 184).

DEFINITION 2.2. Given a (limiting) function $f \in \mathcal{Z}$ and a set $S \subset \Theta$, we say that a continuous real-valued function $\Delta_f(\cdot)$ is a descent function for \mathcal{A} with respect to f and S if:

- (i) If $\theta \notin S$ and $\xi \in \mathcal{A}(f, \theta)$, then $\Delta_f(\xi) < \Delta_f(\theta)$.
- (ii) If $\theta \in S$ and $\xi \in \mathcal{A}(f, \theta)$, then $\Delta_f(\xi) \leq \Delta_f(\theta)$.

We refer to the set S as a solution set or a set of stationary points of the function f , and say that \mathcal{A} is a descent algorithm if it possesses a descent function. For the examples later discussed, we will take the descent function to be $\Delta_f(\theta) = f(\theta)$. However, we develop the main theoretical results in this section in terms of a more general descent function $\Delta_f(\theta)$ because, we believe, the results may be applicable to other algorithms (e.g., Plambeck, Fu, Robinson and Suri, 1995) where $f(\theta)$ will not be a descent function.

The following is an extension of a convergence theorem due to Zangwill (1967).

THEOREM 2.1. Let \mathcal{A} be a descent algorithm with a corresponding descent function $\Delta_f(\cdot)$, with respect to a limiting function $f \in \mathcal{Z}$ and a solution set S . Let f_n be a sequence of approximating functions converging, in the norm topology of \mathcal{Z} , to f and let θ_n be a generated sequence of points satisfying $\theta_{n+1} \in \mathcal{A}(f_n, \theta_n)$. Suppose that:

- (i) The solution set S is closed and there is a constant $M > \sup_{\theta \in S} \Delta_f(\theta)$ such that the set $L = \{\theta : \Delta_f(\theta) \leq M\}$ is compact and $L \subset \Theta$.
- (ii) All points θ_n stay in a compact subset of Θ .
- (iii) The point-to-set mapping \mathcal{A} is closed at f .

Then

$$(2.1) \quad \limsup_{n \rightarrow \infty} \Delta_f(\theta_n) \leq \sup_{\theta \in S} \Delta_f(\theta).$$

PROOF. Consider the set A of accumulation points of the sequence $\{\theta_n\}$. First we show that at least one accumulation point of $\{\theta_n\}$ belongs to the solution set S , i.e., A has a nonempty intersection with S . We argue by a contradiction. Suppose that $A \cap S = \emptyset$. Since the set A is closed, and by the assumption (ii) is bounded, and since the function $\Delta_f(\cdot)$ is continuous, this function attains its minimum over A , say at a point θ^* . Let $\theta_{n(k)}$ be a subsequence of θ_n converging to θ^* . Consider the subsequence $\theta_{n(k)+1}$. By passing to a further subsequence if necessary we can assume

that $\theta_{n(k)+1}$ converges, say to a point $\bar{\theta}$, and hence $\bar{\theta} \in A$. Since $\theta_{n(k)+1} \in \mathcal{A}(f_{n(k)}, \theta_{n(k)})$ and \mathcal{A} is closed at f , we obtain that $\bar{\theta} \in \mathcal{A}(f, \theta^*)$. Because $\theta^* \notin S$, it follows then by the property (i) in Definition 2.2 of a descent function that $\Delta_f(\bar{\theta}) < \Delta_f(\theta^*)$, which contradicts the minimality property of the point θ^* .

We now prove (2.1). It follows from the assumption (i) that the set S is compact, and since the function $\Delta_f(\cdot)$ is continuous, it attains its maximum over S . Denote this maximum by μ , i.e., $\mu = \max_{\theta \in S} \Delta_f(\theta)$. We have to show that μ is greater or equal to the left-hand side of (2.1).

Let us fix a positive number ε and consider the following sets: $S_1 = \{\theta \in \Theta : \Delta_f(\theta) \leq \mu + \varepsilon/2\}$, $S_2 = \{\theta \in \Theta : \Delta_f(\theta) \leq \mu + \varepsilon\}$, and $D = \{\theta \in \Theta : \mu + \varepsilon/2 \leq \Delta_f(\theta) \leq \mu + \varepsilon\}$. Clearly $S \subset S_1 \subset S_2$. We take ε small enough such that $\mu + \varepsilon < M$ and hence $S_2 \subset L$. Moreover, $D \subset \Theta$ and D is compact. We show now that for sufficiently large n , if $\theta_n \in S_1$, then $\theta_{n+1} \in S_2$. We argue by a contradiction. Suppose that the assertion is false. Then there exists a subsequence $\theta_{n(k)}$ such that $\theta_{n(k)} \in S_1$ and $\theta_{n(k)+1} \notin S_2$. By passing to further subsequences if necessary we can assume that both $\theta_{n(k)}$ and $\theta_{n(k)+1}$ do converge, say to θ^* and $\bar{\theta}$ respectively. Since \mathcal{A} is closed at f we obtain that $\bar{\theta} \in \mathcal{A}(f, \theta^*)$. Then using continuity and descent properties of $\Delta_f(\cdot)$ we can write, for $n(k)$ large enough.

$$\Delta_f(\theta_{n(k)+1}) \leq \Delta_f(\bar{\theta}) + \varepsilon/3 \leq \Delta_f(\theta^*) + \varepsilon/3 \leq \Delta_f(\theta_{n(k)}) + \varepsilon/2.$$

Consequently

$$\Delta_f(\theta_{n(k)+1}) \leq \Delta_f(\theta_{n(k)}) + \varepsilon/2 \leq \mu + \varepsilon,$$

and hence $\theta_{n(k)+1} \in S_2$, a contradiction.

Now let us show that for sufficiently large n , if $\theta_n \in D$, then $\Delta_f(\theta_{n+1}) \leq \Delta_f(\theta_n)$. Again we argue by contradiction. Suppose that the above assertion is false. Then by compactness arguments there is a subsequence $\theta_{n(k)} \in D$, converging to a point $\theta^* \in D$, and such that $\Delta_f(\theta_{n(k)+1}) > \Delta_f(\theta_{n(k)})$ and $\theta_{n(k)+1}$ converges to a point $\bar{\theta}$. By continuity of $\Delta_f(\cdot)$ we obtain then that $\Delta_f(\bar{\theta}) \geq \Delta_f(\theta^*)$. On the other hand, since $\theta_{n(k)+1} \in \mathcal{A}(f_{n(k)}, \theta_{n(k)})$ and \mathcal{A} is closed at f , we have that $\bar{\theta} \in \mathcal{A}(f, \theta^*)$. Since $\theta^* \notin S$, it follows that $\Delta_f(\bar{\theta}) < \Delta_f(\theta^*)$, a contradiction.

Let us summarize. We showed that there exists an integer N such that, for any $n \geq N$ the following properties follow. If $\theta_n \in S_1$, then $\theta_{n+1} \in S_2$. If $\theta_n \in D$, then $\theta_{n+1} \in S_2$. Moreover, since the sequence $\{\theta_n\}$ has at least one accumulation point in S and $\Delta_f(\cdot)$ is continuous, we can choose N large enough such that $\theta_N \in S_1$, and since $S_1 \subset S_2$, $\theta_N \in S_2$. Since $S_2 = S_1 \cup D$ it follows then by induction that θ_n stays in S_2 for any $n \geq N$. By the definition of S_2 , we get that $\limsup_{n \rightarrow \infty} \Delta_f(\theta_n) \leq \mu + \varepsilon$. Since ε was arbitrarily small, we obtain that

$$\limsup_{n \rightarrow \infty} \Delta_f(\theta_n) \leq \mu,$$

which completes the proof. \square

REMARKS. (a). It follows that if, in addition to the assumptions of Theorem 2.1, the function $\Delta_f(\cdot)$ is constant on the set S , then the distance from θ_n to S tends to zero as $n \rightarrow \infty$. In particular, if $S = \{\theta_0\}$ is a singleton, we obtain that θ_n converges to θ_0 .

(b). The above deterministic result can be easily translated into the probabilistic language by using a uniform version of the strong Law of Large Numbers. This will be further discussed in the next section.

(c). It is somewhat surprising that the descent properties of the algorithm \mathcal{A} in Theorem 2.1 were used only in the limit, i.e., with respect to the limiting function f , and were not required for the approximating functions. That is, the algorithms do not have to “descent” at each iteration.

(d). In some situations the considered approximating as well as the limiting functions belong to a certain (closed) subset of the space \mathcal{X} . For example, one may work with convex functions forming a (closed) subset of the space $C(\Theta)$. In that case the mapping \mathcal{A} needs only to be defined for functions belonging to that subset.

3. Uniform laws of large numbers. In this section we discuss some uniform versions of the Law of Large Numbers. Consider a sequence of random functions $f_n(\cdot) = f_n(\cdot, \omega)$, $n = 1, 2, \dots$, defined on a common probability space (Ω, \mathcal{F}, P) .

DEFINITION 3.1. We say that the strong Law of Large Numbers (LLN) holds *uniformly* on a set Θ if there is a (deterministic) function $f(\cdot)$ such that the random variables $X_n = \sup_{\theta \in \Theta} |f_n(\theta) - f(\theta)|$ converge with probability one (w.p.1) to zero as $n \rightarrow \infty$.

It follows from $X_n \rightarrow 0$ w.p.1 that the random variables X_n tend to zero in probability, i.e., for any $\varepsilon > 0$,

$$(3.1) \quad \lim_{n \rightarrow \infty} P \left\{ \sup_{\theta \in \Theta} |f_n(\theta) - f(\theta)| \geq \varepsilon \right\} = 0.$$

For random functions satisfying condition (3.1) we say that the weak Law of Large Numbers holds *uniformly* on Θ . Since the weak LLN follows from the strong LLN we discuss only the strong version of the LLN.

Classical results for the strong LLN are available. Let $\zeta_i = \zeta_i(\omega)$, $i = 1, 2, \dots$, be a sequence of independent identically distributed (iid) random variates, where ζ_i can be real valued random variables, random vectors or, more generally, can take values in an abstract probability space (Ξ, \mathcal{S}, Q) . Let $g(\theta, \xi) : \Theta \times \Xi \rightarrow \mathbb{R}$ be a real-valued function and consider the average functions

$$(3.2) \quad f_n(\theta) = n^{-1} \sum_{i=1}^n g(\theta, \zeta_i), \quad n = 1, 2, \dots$$

Suppose that for any $\theta \in \Theta$, the function $g(\theta, \cdot)$ is measurable and the expectation $\mathbb{E}\{|g(\theta, \zeta_1)|\}$ is finite and consider the expected value function $f(\theta) = \mathbb{E}\{g(\theta, \zeta_1)\}$. Then by the Kolmogorov's LLN, the sequence $f_n(\theta)$ converges w.p.1 to $f(\theta)$ for any fixed $\theta \in \Theta$. This pointwise convergence w.p.1 can be easily extended to the uniform convergence under mild additional conditions. In the following proposition all probabilistic statements are made with respect to the distribution (probability measure) Q of ζ_1 .

PROPOSITION 3.1. *Suppose that:*

- (i) *The set Θ is compact.*
- (ii) *For Q -almost every ξ , the function $g(\cdot, \xi)$ is continuous on Θ .*
- (iii) *The family $\{|g(\theta, \cdot)|, \theta \in \Theta\}$ is dominated by an integrable function, i.e., $\mathbb{E}\{\sup_{\theta \in \Theta} |g(\theta, \zeta_1)|\} < \infty$.*

Then $f(\theta)$ is continuous on Θ and $f_n(\theta)$ converges to $f(\theta)$ uniformly on Θ w.p.1.

The above uniform version of the LLN was used by various authors. For an elementary proof of this result see, e.g., Rubinstein and Shapiro (1993). If, in addition to the assumptions (i)–(iii) of Proposition 3.1, the function $g(\cdot, \xi)$ is continuously differentiable for Q -almost every ξ and the family $\{\|\nabla g(\theta, \cdot)\|, \theta \in \Theta\}$ is dominated

by an integrable function, then the expected value function $f(\theta)$ is continuously differentiable, $\nabla f(\theta) = \mathbb{E}\{\nabla g(\theta, \zeta_1)\}$ and $\nabla f_n(\theta)$ converges to $\nabla f(\theta)$ uniformly on Θ w.p.1. Similar statements can be made for the second-order derivatives, etc.

This can be extended to nondifferentiable situations as follows. Recall that the generalized gradient $\partial f(\theta)$, in the sense of Clarke (1983), of a locally Lipschitz function $f(\cdot)$ is the convex hull of all limits of the form $\lim_{n \rightarrow \infty} \nabla f(\theta_n)$, where $\{\theta_n\}$ is any sequence converging to θ and such that f is differentiable at θ_n , $n = 1, 2, \dots$, and the above limit exists. The generalized gradient $\partial f(\theta)$ is a singleton, i.e., contains only one element, iff the function f is continuously differentiable at θ . If f is convex, then the generalized gradient coincides with the subdifferential in the sense of convex analysis (Rockafellar, 1970).

Consider next a situation when the function $g(\cdot, \xi)$ is locally Lipschitz continuous, but not necessarily convex. We have the following result (Shapiro, 1989):

PROPOSITION 3.2. *Suppose that:*

(i) *The set Θ is convex and compact.*

(ii) *There exists a Q -integrable function $k(\xi)$ such that for all $\theta_1, \theta_2 \in \Theta$ and Q -almost every ξ ,*

$$|g(\theta_1, \xi) - g(\theta_2, \xi)| \leq k(\xi) \|\theta_1 - \theta_2\|.$$

(iii) *For any given (fixed) $\theta \in \Theta$, the function $g(\cdot, \zeta_1)$ is continuously differentiable at θ w.p.1.*

Then

(a) *The expected value function $f(\cdot)$ is continuously differentiable on Θ and*

$$\nabla f(\theta) = \mathbb{E}\{\nabla g(\theta, \zeta_1)\}.$$

(b) *The functions $f_n(\cdot)$ are Lipschitz continuous on Θ and $\partial f_n(\theta)$ converges to $\nabla f(\theta)$ w.p.1 uniformly on Θ , i.e.,*

$$\lim_{n \rightarrow \infty} \sup_{\theta \in \Theta} \sup_{z \in \partial f_n(\theta)} \|z - \nabla f(\theta)\| = 0, \quad \text{w.p.1.}$$

A somewhat different approach to the uniform LLN is based on convex analysis. That is, suppose that w.p.1 the random functions $f_n(\theta)$ are convex and that for any fixed $\theta \in \mathbb{R}^m$, the sequence $f_n(\theta)$ converges w.p.1 to a finite limit $f(\theta)$, i.e., the LLN holds pointwise. It follows then that the function $f(\theta)$ is convex and the convergence is uniform w.p.1 on any compact set Θ . Indeed, we have then that for almost every ω , $f_n(\cdot, \omega)$ converges to $f(\cdot)$ on a countable dense subset of \mathbb{R}^m and hence the uniform convergence follows by convex analysis. This result was noticed by several authors (see, e.g., Robinson 1995, and references therein). Moreover, the following result holds with respect to the derivatives of the considered functions.

For two subsets S and T of \mathbb{R}^m , the excess of S with respect to T is defined by

$$e[S, T] = \sup_{s \in S} \text{dist}(s, T),$$

where $\text{dist}(s, T) = \inf_{t \in T} \|s - t\|$ is the distance from s to T . The Hausdorff distance between S and T is defined by

$$\rho[S, T] = \max\{e[S, T], e[T, S]\}.$$

Let Θ be a relatively open convex subset of \mathbb{R}^m . (A convex set is said to be relatively open if it is open relative to the linear space it generates.)

PROPOSITION 3.3. *Suppose that:*

- (i) *The set Θ is convex and relatively open.*
- (ii) *For each n , $f_n(\cdot)$ is a finite convex function on Θ w.p.1.*
- (iii) *For every $\theta \in \Theta$ the sequence $f_n(\theta)$ converges w.p.1 to a finite (deterministic) limit $f(\theta)$.*

Then

- (a) *The limiting function $f(\cdot)$ is a convex function on Θ .*
- (b) *$f_n(\cdot)$ converges to $f(\cdot)$ w.p.1 uniformly on any compact subset of Θ .*
- (c) *If $\{\theta_n\}$ is a sequence in Θ converging to $\theta^* \in \Theta$, then $e[\partial f_n(\theta_n), \partial f(\theta^*)]$ tends to zero w.p.1 as $n \rightarrow \infty$.*

Property (c) of Proposition 3.3 was derived in Robinson (1995). Note that this property does not imply the uniform convergence to zero of $e[\partial f_n(\cdot), \partial f(\cdot)]$ w.p.1. Consider, for example, $f_n(\theta) = \max\{|\theta| - n^{-1}, 0\}$, $\theta \in \mathbb{R}$. Then $f_n(\theta)$ converges to $f(\theta) = |\theta|$. However, for $\theta \in (0, n^{-1})$, $\nabla f_n(\theta) = 0$, while $\nabla f(\theta) = 1$. Therefore $e[\partial f_n(\theta), \partial f(\theta)]$ does not converge uniformly to zero on any interval containing zero.

It is known from convex analysis (Rockafellar, 1970) that a convex function $f(\cdot)$ is differentiable at a point θ if and only if its subdifferential $\partial f(\theta)$ is a singleton in which case $\partial f(\theta) = \{\nabla f(\theta)\}$ and $\nabla f(\cdot)$ is continuous at θ provided θ belongs to the relative interior of Θ . (The latter statement means that, if $\theta_n \rightarrow \theta$ and $\nabla f(\theta_n)$ exists, then $\nabla f(\theta_n) \rightarrow \nabla f(\theta)$.) Therefore, under the assumptions of Proposition 3.3, it follows by compactness arguments that, if the limiting function $f(\cdot)$ is differentiable, then, for any choice of the subgradient $\nabla f_n(\theta)$, $\nabla f_n(\cdot)$ converges w.p.1 to $\nabla f(\cdot)$ uniformly on any compact subset of Θ .

Let $f(\cdot)$ be a finite convex function on \mathbb{R}^m . For $\varepsilon \geq 0$ its ε -subdifferential at θ_0 is defined by

$$\partial_\varepsilon f(\theta_0) = \{v \in \mathbb{R}^m : f(\theta) - f(\theta_0) \geq v^T(\theta - \theta_0) - \varepsilon \text{ for all } \theta \in \mathbb{R}^m\}.$$

We show now that for any fixed $\varepsilon > 0$, the uniform LLN holds for ε -subdifferentials with respect to the Hausdorff distance.

PROPOSITION 3.4. *Suppose that:*

- (i) *For each n , $f_n(\cdot)$ is a finite convex function on \mathbb{R}^m w.p.1.*
- (ii) *For every $\theta \in \mathbb{R}^m$ the sequence $f_n(\theta)$ converges w.p.1 to a finite (deterministic) limit $f(\theta)$.*

Then, for any $\varepsilon > 0$, $\rho[\partial_\varepsilon f_n(\theta), \partial_\varepsilon f(\theta)]$ tends to zero w.p.1 uniformly on any compact subset of \mathbb{R}^m .

PROOF. First, let us observe that for $\varepsilon > 0$, the ε -subdifferential $\partial_\varepsilon f(\theta)$ is continuous in θ (in fact even locally Lipschitz continuous) with respect to the Hausdorff metric $\rho(\cdot, \cdot)$ (e.g., Dem'yanov and Vasil'ev, 1985, p. 91, Hiriart-Urruty and Lemaréchal, 1993b, p. 128). Therefore it will be sufficient to show that if $f_n(\cdot)$ is a sequence of convex functions converging (pointwise) to a function $f(\cdot)$ and θ_n is a sequence of points converging to a point θ^* , then $\rho[\partial_\varepsilon f_n(\theta_n), \partial_\varepsilon f(\theta^*)]$ tends to zero.

Consider the ε -directional derivative of f at θ ,

$$f'_\varepsilon(\theta, d) = \inf_{t>0} \frac{f(\theta + td) - f(\theta) + \varepsilon}{t}.$$

It is known (Hiriart-Urruty and Lemaréchal, 1993b, p. 102) that $f'_\varepsilon(\theta, \cdot)$ is the support function of the set $\partial_\varepsilon f(\theta)$. Therefore, since convergence of a sequence of

nonempty convex compact sets in the Hausdorff metric is equivalent to the pointwise convergence of the corresponding support functions (Hiriart-Urruty and Lemaréchal, 1993a, p. 232), it will be sufficient to show that for any given $d \in \mathbb{R}^m$,

$$\lim_{n \rightarrow \infty} f'_{n\varepsilon}(\theta_n, d) = f'_\varepsilon(\theta^*, d).$$

Let us fix $t > 0$. Then

$$\limsup_{n \rightarrow \infty} f'_{n\varepsilon}(\theta_n, d) \leq \limsup_{n \rightarrow \infty} \frac{f_n(\theta_n + td) - f_n(\theta_n) + \varepsilon}{t} = \frac{f(\theta^* + td) - f(\theta^*) + \varepsilon}{t}.$$

Since $t > 0$ was arbitrary this implies that

$$\limsup_{n \rightarrow \infty} f'_{n\varepsilon}(\theta_n, d) \leq f'_\varepsilon(\theta^*, d).$$

Now let us suppose for a moment that the minimum of $t^{-1}[f(\theta^* + td) - f(\theta^*) + \varepsilon]$ over $t \in \mathbb{R}_+$ is attained on a bounded set T_ε . It follows then by convexity that for n large enough $t^{-1}[f_n(\theta_n + td) - f_n(\theta_n) + \varepsilon]$ attains its minimum over \mathbb{R}_+ , say at a point t_n , and $\text{dist}(t_n, T_\varepsilon) \rightarrow 0$. Note that $\inf T_\varepsilon > 0$. Consequently

$$\liminf_{n \rightarrow \infty} f'_{n\varepsilon}(\theta_n, d) = \liminf_{n \rightarrow \infty} \frac{f_n(\theta_n + t_n d) - f_n(\theta_n) + \varepsilon}{t_n} \geq f'_\varepsilon(\theta^*, d).$$

In the general case the proof can be completed by adding the term $\alpha \|\theta - \theta^*\|^2$, $\alpha > 0$, to the functions $f_n(\theta)$ and $f(\theta)$ and passing to the limit $\alpha \rightarrow 0^+$ (cf. Dem'yanov and Vasil'ev, 1985, pp. 93–94). \square

4. Applications and examples. In this section we apply the general result of Theorem 2.1 to various particular algorithms. Let $\Theta \subset \mathbb{R}^m$ be a convex, bounded open set, and consider a sequence of random Lipschitz functions $f_n \in \text{Lip}(\Theta)$ converging (in the norm topology of $\text{Lip}(\Theta)$) to a function f w.p.1. Note that this implies uniform convergence of $f_n(\theta)$ to $f(\theta)$ and if, in addition, f is continuously differentiable, uniform convergence of $\partial f_n(\theta)$ to $\nabla f(\theta)$ follows as well. We refer to Proposition 3.2 and 3.3 for conditions ensuring such convergence w.p.1 in nonconvex and convex cases, respectively.

Let us start by considering the steepest descent algorithm with a fixed stepsize $\alpha > 0$. That is, $\theta_{n+1} = \theta_n - \alpha g_n$, where $g_n = \nabla f_n(\theta_n)$ if f_n is differentiable at θ_n , and g_n can be any element of the generalized gradient $\partial f_n(\theta_n)$ otherwise. The corresponding mapping \mathcal{A} is defined by

$$\mathcal{A}(\phi, \theta) = \{ \xi = \theta - \alpha g : g \in \partial \phi(\theta) \}, \quad \phi \in \text{Lip}(\Theta), \theta \in \Theta.$$

(Recall that ϕ denotes a generic element of the space $\mathcal{X} = \text{Lip}(\Theta)$.) Suppose that the limiting function f is continuously differentiable and has a unique minimizer θ_0 .

Consider the function $\Delta_f(\theta) = f(\theta)$. Let us observe that Δ_f is a descent function in a neighborhood of θ_0 for the above algorithm, provided f is twice continuously differentiable with nonsingular Hessian matrix $H = \nabla^2 f(\theta_0)$ and the stepsize α is small enough. Indeed, suppose for a moment that the function f is quadratic and hence coincides with its second-order Taylor expansion. In that case the number t^* which minimizes $\psi(t) = f(\theta_n - td_n)$, where $d_n = \nabla f(\theta_n)$, is given by $t^* = d_n^T d_n / (d_n^T H d_n)$. Consequently $t^* \geq \lambda^{-1}$, where λ is the largest eigenvalue of the

Hessian matrix H . Therefore, in that case Δ_f is a descent function if the stepsize α is less than $2\lambda^{-1}$. In general, for a twice continuously differentiable function f , under such a choice of α , Δ_f becomes a descent function in a sufficiently small neighborhood of θ_0 . In the following proposition we postulate existence of such a neighborhood (assumption (ii)) and assume that the generated sequence eventually stays in that neighborhood w.p.1 (assumption (iii)).

It is clear that the corresponding mapping \mathcal{A} is closed at the limiting function f . This follows from the fact that, in the considered norm topology of $\text{Lip}(\Theta)$, if $\theta_n \rightarrow \theta^*$ and $f_n \rightarrow f$, then $g_n \rightarrow \nabla f(\theta^*)$ and hence $(\theta_n - \alpha g_n) = \theta_{n+1} \rightarrow (\theta^* - \alpha \nabla f(\theta^*))$. We obtain the following result.

PROPOSITION 4.1. *Consider the steepest descent algorithm with a fixed stepsize $\alpha > 0$. Suppose that:*

(i) *The approximating functions $f_n(\theta)$ are Lipschitz continuous on Θ and converge w.p.1, in the norm topology of $\text{Lip}(\Theta)$, to a continuously differentiable function $f(\theta)$.*

(ii) *The limiting function $f(\theta)$ has a locally unique minimizer θ_0 and there exists a convex compact set $C \subset \Theta$ such that θ_0 is an interior point of C and for any $\theta \in C$, $\theta \neq \theta_0$, the inequality $f(\theta - \alpha \nabla f(\theta)) < f(\theta)$ holds.*

(iii) *For all n large enough, the generated sequence $\{\theta_n\}$ stays w.p.1 in the set C .*

Then θ_n converges to θ_0 w.p.1.

It is interesting to note that the above descent algorithm, with a fixed stepsize, does not guarantee descent of a current approximating function f_n at each iteration. The descent properties, postulated in the assumption (ii), were used only for the limiting function f . Note also that the associated mapping \mathcal{A} is closed at the limiting function f because f is assumed to be continuously differentiable, and is not necessarily closed at other elements ϕ of the space $\text{Lip}(\Theta)$.

Another possible choice of the stepsize is suggested by the line search. That is, $t_n = \arg \min_{t \geq 0} f_n(\theta_n - t g_n)$ and $\theta_{n+1} = \theta_n - t_n g_n$. In that case the function $\Delta_f(\theta) = f(\theta)$ is clearly a descent function and, provided the limiting function f is continuously differentiable, closedness at f of the corresponding mapping \mathcal{A} follows from the uniform convergence of $f_n(\theta)$ and $\partial f_n(\theta)$ to $f(\theta)$ and $\nabla f(\theta)$, respectively. Practically, this line search has to be approximated, e.g., by the Armijo stepsize (Armijo, 1966). To keep the exposition as simple as possible we consider here the conceptual algorithm where the line search can be precisely computed; extensions to implementable algorithms are obvious.

PROPOSITION 4.2. *Consider the steepest descent algorithm with the stepsize calculated by the line search. Suppose that:*

(i) *The approximating functions $f_n(\theta)$ are Lipschitz continuous on Θ and converge w.p.1, in the norm topology of $\text{Lip}(\Theta)$, to a continuously differentiable function $f(\theta)$.*

(ii) *The limiting function $f(\theta)$ has a nonempty connected compact set S of optimal solutions.*

(iii) *Every stationary point of f is an optimal solution, i.e., if $\nabla f(\theta) = 0$, $\theta \in \Theta$, then $\theta \in S$.*

(iv) *The generated sequence $\{\theta_n\}$ stays w.p.1 in a compact subset of Θ .*

Then

$$\lim_{n \rightarrow \infty} \text{dist}(\theta_n, S) = 0, \quad \text{w.p.1.}$$

Finally, we discuss a steepest descent method for nondifferentiable, convex problems. That is, we work now in the subset of $\mathcal{Z} = C(\Theta)$ formed by convex functions. Suppose that $f_n(\theta)$ and $f(\theta)$ are convex, but possibly nondifferentiable, and let

$f_n(\theta) \rightarrow f(\theta)$ w.p.1 for any $\theta \in \Theta$. By Propositions 3.3(b) and 3.4, it follows that $f_n(\cdot) \rightarrow f(\cdot)$ uniformly w.p.1, and for every $\varepsilon > 0$, $\partial_\varepsilon f_n(\cdot) \rightarrow \partial_\varepsilon f(\cdot)$ (in the Hausdorff metric) uniformly w.p.1 on any compact subset of Θ .

The steepest descent algorithm studied in the remainder of this section follows some well-established principles of nondifferentiable optimization (see, e.g., Hiriart-Urruty and Lemaréchal, 1993a, b, Polak 1987, and the references therein). In order to construct a descent direction the algorithm finds the point nearest to 0 in the ε -subdifferential, for a suitably chosen ε , and then goes in the opposite direction. The precision ε is updated during the iteration process and is determined in the following way.

Let $\varepsilon_1, \varepsilon_2, \dots$ be a monotonically decreasing to zero sequence of positive numbers and consider the set $\Upsilon = \{\varepsilon_1, \varepsilon_2, \dots\} \cup \{0\}$. Given a convex compact set $C \subset \mathbb{R}^m$, we denote by $\text{Nr}(C)$ the point in C closest to 0. Note that $\|\text{Nr}(C)\|$ is the distance $\text{dist}(0, C)$ from 0 to C . Now for $\theta \in \Theta$ and a convex function f , we define

$$\delta_f(\theta) = \max\{\varepsilon \in \Upsilon : \|\text{Nr}(\partial_\varepsilon f(\theta))\| \geq \varepsilon\}.$$

It is not difficult to see that $\delta_f(\theta) = 0$ iff $0 \in \partial f(\theta)$, i.e., θ is a minimizer of $f(\cdot)$.

For every $\varepsilon \geq 0$ such that $0 \notin \partial_\varepsilon f(\theta)$, the direction $-\text{Nr}(\partial_\varepsilon f(\theta))$ is a descent direction for f from θ . That is, proceeding from θ in this direction, f goes down for a small enough stepsize. The value of ε commonly used in descent algorithms is $\varepsilon = \delta_f(\theta)$. In our algorithm, as applied to the function f , the next iteration point is determined by minimizing the function f along that descent direction. Formally, given $\theta \in \Theta$, $\phi \in \mathcal{Z}$ and a vector $h \in \mathbb{R}^m$, denote

$$T_\phi(\theta, h) = \arg \min\{\phi(\theta - th) : t \geq 0\}.$$

Now, given a convex function ϕ , the next iteration point computed from θ has the form $\theta - th$, where $h = \text{Nr}(\partial_\varepsilon \phi(\theta))$, $\varepsilon = \delta_\phi(\theta)$ and $t \in T_\phi(\theta, h)$. The algorithm thus has the following form.

ALGORITHM 4.1. Given a current iteration point θ_n , compute θ_{n+1} as follows.

(1) Compute $\delta_n := \delta_{f_n}(\theta_n)$, and compute $h_n := \text{Nr}(\partial_{\delta_n} f_n(\theta_n))$.

(2) Compute an (arbitrarily chosen) point $t \in T_{f_n}(\theta_n, h_n)$, denote it by t_n , and set $\theta_{n+1} = \theta_n - t_n h_n$.

Now it is natural to define the associated point-to-set mapping as $\mathcal{A} : (\phi, \theta) \rightarrow \{\theta - th : t \in T_\phi(\theta, h)\}$, where $h = \text{Nr}(\partial_{\delta_\phi(\theta)} \phi(\theta))$. With the descent function $\Delta_f(\theta) = f(\theta)$, such mapping \mathcal{A} indeed defines a descent algorithm. However, \mathcal{A} is not closed at the limiting function f . To see this, consider a sequence $\{\theta_n\}$ such that $\theta_n \rightarrow \theta$, and suppose that for a fixed $\varepsilon_j \in \Upsilon$, $\delta_n = \varepsilon_j$ for all n . We then have that $h_n = \text{Nr}(\partial_{\varepsilon_j} f_n(\theta_n))$. By Proposition 3.4, we have $\lim_{n \rightarrow \infty} h_n = h$, where $h = \text{Nr}(\partial_{\varepsilon_j} f(\theta))$. However, $-h$ might not be the algorithm's descent direction for f at θ , because $\delta_f(\theta)$ need not be equal to ε_j . In other words, the descent direction computed by the algorithm is not continuous in the topology of $\mathcal{Z} \times \Theta$. The following, however, holds true.

LEMMA 4.1. Let $\{\theta_n\}$ be a sequence, generated by the algorithm 4.1, having an accumulation point θ .

(i) If there is an $\varepsilon_j \in \Upsilon$ such that $\delta_n = \varepsilon_j$ for all $n = 1, 2, \dots$, then $\delta_f(\theta) \in \{\varepsilon_j, \varepsilon_{j-1}\}$.

(ii) If $\delta_n \rightarrow 0$ as $n \rightarrow \infty$, then $\delta_f(\theta) = 0$.

PROOF. It suffices to assume, without loss of generality, that $\lim_{n \rightarrow \infty} \theta_n = \theta$.

(i) We have that $\|h_n\| \geq \varepsilon_j$ for all n , and hence (and by Proposition 3.4), $\|\text{Nr}(\partial_{\varepsilon_j} f(\theta))\| \geq \varepsilon_j$. Consequently, $\delta_f(\theta) \geq \varepsilon_j$. On the other hand, for all n , we have $\|\text{Nr}(\partial_{\varepsilon_{j-1}} f_n(\theta_n))\| < \varepsilon_{j-1}$, and therefore $\|\text{Nr}(\partial_{\varepsilon_{j-1}} f(\theta))\| \leq \varepsilon_{j-1}$. This implies that

$$\|\text{Nr}(\partial_{\varepsilon_{j-2}} f(\theta))\| \leq \|\text{Nr}(\partial_{\varepsilon_{j-1}} f(\theta))\| \leq \varepsilon_{j-1} < \varepsilon_{j-2},$$

and hence $\delta_f(\theta) < \varepsilon_{j-2}$. We thus have seen that $\varepsilon_j \leq \delta_f(\theta) < \varepsilon_{j-2}$, and hence either $\delta_f(\theta) = \varepsilon_j$ or $\delta_f(\theta) = \varepsilon_{j-1}$.

(ii) Suppose, for the sake of contradiction, that $\delta_f(\theta) > 0$. Let $\delta_f(\theta) = \varepsilon_j$ for some $j = 1, 2, \dots$. Then,

$$\|\text{Nr}(\partial_{\varepsilon_{j+1}} f(\theta))\| \geq \|\text{Nr}(\partial_{\varepsilon_j} f(\theta))\| \geq \varepsilon_j > \varepsilon_{j+1}.$$

Consequently, and by Proposition 3.4, for n large enough, $\|\text{Nr}(\partial_{\varepsilon_{j+1}} f_n(\theta_n))\| > \varepsilon_{j+1}$. This implies that $\delta_n \geq \varepsilon_{j+1}$, contradicting the assertion of the lemma. \square

Let us modify now the mapping \mathcal{A} in the following way. Consider a convex function ϕ and the set $S_\phi = \{\theta : 0 \in \partial\phi(\theta)\}$. Since ϕ is convex, S_ϕ is the set of minimizers of ϕ . The mapping $\mathcal{A}(\phi, \theta)$ is defined according to three separate cases, as follows.

(a) If $\theta \in S_\phi$, then define $\mathcal{A}(\phi, \theta) = S_\phi$.

(b) If $\delta_\phi(\theta) = \varepsilon_j$ for some $j = 1, 2, \dots$, and $\|\text{Nr}(\partial_{\varepsilon_j} \phi(\theta))\| > \varepsilon_j$, then define $\mathcal{A}(\phi, \theta) = \{\theta - th : t \in T_\phi(\theta, h)\}$, with $h := \text{Nr}(\partial_{\varepsilon_j} \phi(\theta))$.

(c) If $\delta_\phi(\theta) = \varepsilon_j$ and $\|\text{Nr}(\partial_{\varepsilon_j} \phi(\theta))\| = \varepsilon_j$, then define

$$\mathcal{A}(\phi, \theta) = \{\theta - th : t \in T_\phi(\theta, h)\} \cup \{\theta - \bar{t}h : t \in T_\phi(\theta, \bar{h})\},$$

with $h := \text{Nr}(\partial_{\varepsilon_j} \phi(\theta))$ and $\bar{h} := \text{Nr}(\partial_{\varepsilon_{j+1}} \phi(\theta))$.

REMARKS. It is evident that $\Delta_f(\theta) := f(\theta)$ is a descent function for \mathcal{A} with respect to f and $S = S_f$. Considering the above definition of \mathcal{A} , we observe the following. If θ is a minimizer of f , then $\mathcal{A}(f, \theta)$ is the minimizer set. Case (b) is obvious in light of Algorithm 4.1. Case (c) has been introduced because the descent directions are discontinuous, and, as we shall see, it makes the mapping \mathcal{A} a closed mapping.

LEMMA 4.2. *The mapping \mathcal{A} is closed at f .*

PROOF. Given converging sequences $\theta_n \rightarrow \theta$ and $\xi_n \rightarrow \xi$, suppose that, for all n , $\xi_n \in \mathcal{A}(f_n, \theta_n)$. We will prove that $\xi \in \mathcal{A}(f, \theta)$, which will establish the closedness of \mathcal{A} . Two separate cases will be considered. Namely (i) $\limsup_{n \rightarrow \infty} \delta_n > 0$, and (ii) $\lim_{n \rightarrow \infty} \delta_n = 0$.

Let us start with the case (i). By taking a subsequence if necessary, we can assume that, for all n and for some $j = 1, 2, \dots$, $\delta_n = \varepsilon_j$. By the first part of Lemma 4.1, $\delta_f(\theta) \in \{\varepsilon_j, \varepsilon_{j-1}\}$. We will use the notation $h_n = \text{Nr}(\partial_{\varepsilon_j} f_n(\theta_n))$, $\tilde{h}_n = \text{Nr}(\partial_{\varepsilon_{j-1}} f_n(\theta_n))$, $h = \text{Nr}(\partial_{\varepsilon_j} f(\theta))$, and $\tilde{h} = \text{Nr}(\partial_{\varepsilon_{j-1}} f(\theta))$. By further taking subsequences, we need only to consider the following two situations: first $\|h_n\| > \varepsilon_j$ for all n , and second $\|h_n\| = \varepsilon_j$ for all n . In the first situation, $\xi_n \in \{\theta_n - th_n : t \in T_{f_n}(\theta_n, h_n)\}$, and therefore (and by convexity and Proposition 3.4), $\xi \in \{\theta - th : t \in T_f(\theta, h)\}$. Now if $\delta_f(\theta) = \varepsilon_j$, then clearly $\xi \in \mathcal{A}(f, \theta)$. If, on the other hand, $\delta_f(\theta) = \varepsilon_{j-1}$, then $\|\tilde{h}\| = \varepsilon_{j-1}$, since $\|\tilde{h}_n\| < \varepsilon_{j-1}$ and $\tilde{h}_n \rightarrow \tilde{h}$, excluding the possibility that $\tilde{h} > \varepsilon_{j-1}$. In this case, $\xi \in \mathcal{A}(\theta, f)$ by part (c) of the definition of \mathcal{A} .

In the second situation, $\|h_n\| = \varepsilon_j$ for all n , and hence, $\|h\| = \varepsilon_j$, $\delta_f(\theta) = \varepsilon_j$, and $\mathcal{A}(f_n, \theta_n) \rightarrow \mathcal{A}(f, \theta)$ in the Hausdorff metric. It then follows that $\xi \in \mathcal{A}(f, \theta)$.

Consider now the case (ii). We have that $\delta_n \rightarrow 0$, and by the second part of Lemma 4.1, $\delta_f(\theta) = 0$. This means that $\theta \in S$. We need only to show that $\xi \in S$; by part (a) of the definition of \mathcal{A} , this implies that $\xi \in \mathcal{A}(f, \theta)$.

For all n , we have that $f_n(\xi_n) \leq f_n(\theta_n)$. Since $\theta_n \rightarrow \theta$ and $\xi_n \rightarrow \xi$, and by the uniform convergence of $f_n(\cdot)$ to $f(\cdot)$, we have that $f(\xi) \leq f(\theta)$. Now θ is a stationary point and hence a minimizer of f , implying that $f(\xi) = f(\theta)$ and $\xi \in S$. \square

We thus have the following result.

PROPOSITION 4.3. *Consider Algorithm 4.1, and suppose that:*

(i) *The approximating functions $f_n(\theta)$ are convex and they converge pointwise w.p.1 to the limiting function $f(\theta)$.*

(ii) *The solution set S is contained in the interior of Θ and the generated sequence $\{\theta_n\}$ stays w.p.1 in a compact subset of Θ .*

Then, w.p.1, every accumulation point of the sequence $\{\theta_n\}$ is in S .

5. Conclusions and some open research directions. This paper has proved almost sure convergence to stationary points of various optimization algorithms where the functions involved and their subgradients are estimated by Monte Carlo simulation. The algorithms considered are adaptations of well-known nonlinear programming methods to the stochastic environment of simulation. They choose steepest descent directions, their stepsizes are either constant or are computed by line minimization, and the approximations they employ become progressively finer at successive iterations.

Such algorithms for differentiable programs were proposed and analyzed in the past, but our analysis has extended the results that had been obtained. In particular, the treatment of nondifferentiable programs is novel.

We first developed a general theoretical framework based on the concepts of descent functions and algorithm mappings, within which a convergence result was proved. This was later used for establishing convergence of some specific algorithms.

The main results, including those of the general theory, pertain only to Markovian algorithms, namely where the next iteration point depends only on the current iteration point and the current approximating function. A number of interesting algorithms, however, are not Markovian, and proving their convergence remains an open question. Examples include the bundle-type method developed by Plambeck, Fu, Robinson and Suri (1995), and descent-type algorithms with adaptive precision levels. An additional class of non-Markovian algorithms of potential interest consists of second-order methods like quasi-Newton.

Another open question concerns the convergence rate of simulation-based algorithms. The concept of convergence rate is differently understood in the contexts of (deterministic) nonlinear programming on one hand, and stochastic optimization on the other hand. Our algorithms, after all, are adaptations of nonlinear programming techniques to the stochastic setup, and therefore it seems possible to define the concept of convergence rate in a way that adequately captures an algorithm's efficiency within the context of Monte Carlo simulation.

References

- Armijo, L. (1966). Minimization of functions having continuous partial derivatives. *Pacific J. Math.* **16** 1–3.
- Clarke, F. H. (1983). *Optimization and Nonsmooth Analysis*, Wiley, New York.
- Dem'yaov, V. F., L. V. Vasil'ev (1985). *Nondifferentiable Optimization*, Optimization Software, New York.
- Dupuis, P., R. Simha (1991). On sampling controlled stochastic approximation. *IEEE Trans. Automat. Control* **AC-36** 915–924.
- Fu, M. C. (1994). Optimization via simulation: A review. *Ann. Oper. Res.* **53** 199–247.
- Hiriart-Urruty, J. B., C. Lemaréchal (1993a). *Convex Analysis and Minimization Algorithms*. I. Springer-Verlag, Berlin.

- _____, _____. (1993b). *Convex Analysis and Minimization Algorithms. II*, Springer-Verlag, Berlin.
- Ho, Y. C., X. R. Cao (1991). *Perturbation Analysis of Discrete Event Dynamic Systems*, Kluwer Academic Publishers, Boston, MA.
- Kushner, H. J., D. S. Clark (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York.
- Luenberger, D. G. (1989). *Linear and Nonlinear Programming*, Addison-Wesley Publishing Company, Second Edition, London.
- Meheshwari, S., H. Mukai (1986). An optimization algorithm driven by probabilistic simulation. *Proc. Conference on Decision and Control*, Athens, Greece, 1703–1705.
- Mukai, H., D. Yan (1991). *Methods for minimizing the expectation*. Presented at the NSF Workshop on Discrete Event Dynamic Systems, Amherst, MA.
- Plambeck, E. L., B. R. Fu, S. M. Robinson, R. Suri (1995). Optimizing performance functions in stochastic systems. *Math. Programming* (to appear).
- Polak, E. (1987). On the mathematical foundations of nondifferentiable optimization in engineering design. *SIAM Rev.* **29** 21–89.
- Robinson, S. M. (1995). Convergence of subdifferentials under strong stochastic convexity. *Management Sci.* **41** 1397–1401.
- Rockafellar, R. T. (1970). *Convex Analysis*, Princeton University Press, NJ.
- Rubinstein, R. Y., A. Shapiro (1993). *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley and Sons, New York.
- Shapiro, A. (1989). Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.* **17** 841–858.
- _____, Y. Wardi (1994). Nondifferentiability of the steady state function in discrete event dynamic systems. *IEEE Trans. Automat. Control* **39** 1707–1711.
- Simha, R., J. F. Kurose (1989). Stochastic approximation schemes for a load balancing problem. *Proc. 27th Annual Allerton Conf.*, Allerton, IL.
- Wardi, Y. (1990). Stochastic algorithms with Armijo step sizes for minimization of functions. *J. Optim. Theory Appl.* **64** 399–417.
- _____, K. Lee (1991). Application of descent algorithms with Armijo stepsizes to simulation-based optimization of queueing networks. *Proc. 30th Conf. on Decision and Control*, Brighton, England, 110–115.
- _____, M. W. McKinnon, R. Schuckle (1991). On perturbation analysis of queueing networks with finitely supported service time distributions. *IEEE Trans. Automat. Control* **AC-36** 863–867.
- Yan, D., H. Mukai (1995). An optimization algorithm with probabilistic simulation. *J. Optim. Theory Appl.* (to appear).
- Zangwill, W. J. (1967). *Nonlinear Programming, A Unified Approach*, Prentice-Hall, NJ.

A. Shapiro: School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332; e-mail: ashapiro@isye.gatech.edu

Y. Wardi: School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332