SIAM J. OPTIM.
Vol. 10, No. 3, pp. 627–642

GRADIENT CONVERGENCE IN GRADIENT METHODS WITH ERRORS*

DIMITRI P. BERTSEKAS† AND JOHN N. TSITSIKLIS†

Abstract. We consider the gradient method $x_{t+1} = x_t + \gamma_t(s_t + w_t)$, where s_t is a descent direction of a function $f : \Re^n \to \Re$ and w_t is a deterministic or stochastic error. We assume that ∇f is Lipschitz continuous, that the stepsize γ_t diminishes to 0, and that s_t and w_t satisfy standard conditions. We show that either $f(x_t) \to -\infty$ or $f(x_t)$ converges to a finite value and $\nabla f(x_t) \to 0$ (with probability 1 in the stochastic case), and in doing so, we remove various boundedness conditions that are assumed in existing results, such as boundedness from below of f, boundedness of $\nabla f(x_t)$, or boundedness of x_t.

Key words. gradient methods, incremental gradient methods, stochastic approximation, gradient convergence

AMS subject classifications. 62L20, 903C0

PII. S1052623497331063

1. Introduction.

We consider the problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in \Re^n, \end{aligned} \tag{1.1}$$

where $\Re^n$ denotes the n-dimensional Euclidean space and $f : \Re^n \mapsto \Re$ is a continuously differentiable function, such that for some constant L we have

$$\|\nabla f(x) - \nabla f(\overline{x})\| \le L\|x - \overline{x}\| \qquad \forall\, x, \overline{x} \in \Re^n. \tag{1.2}$$

The purpose of this paper is to sharpen the existing convergence theory for the classical descent method

$$x_{t+1} = x_t + \gamma_t(s_t + w_t), \tag{1.3}$$

where

(a) γ_t is a positive stepsize sequence satisfying

$$\sum_{t=0}^{\infty} \gamma_t = \infty, \qquad \sum_{t=0}^{\infty} \gamma_t^2 < \infty; \tag{1.4}$$

(b) s_t is a descent direction satisfying for some positive scalars c_1 and c_2, and all t,

$$c_1\|\nabla f(x_t)\|^2 \le -\nabla f(x_t)' s_t, \qquad \|s_t\| \le c_2\|\nabla f(x_t)\|; \tag{1.5}$$

(c) w_t either is a deterministic error satisfying for some positive scalars p and q, and all t,

$$\|w_t\| \le \gamma_t\big(q + p\|\nabla f(x_t)\|\big) \tag{1.6}$$

---

*Received by the editors December 4, 1997; accepted for publication (in revised form) May 24, 1999; published electronically February 29, 2000. This work was supported by NSF under grant DMI-9625489.

http://www.siam.org/journals/siopt/10-3/33106.html

†Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139 (dimitrib@mit.edu, jnt@mit.edu).

or is a stochastic error satisfying conditions that are standard in stochastic gradient and stochastic approximation methods.

Our main result is that either $f(x_t) \to -\infty$ or $f(x_t)$ converges to a finite value and $\lim_{t\to\infty} \nabla f(x_t) = 0$ (with probability 1 on the stochastic case).

The method where the errors w_t are deterministic includes as a special case the standard incremental gradient/backpropagation method for neural network training, the convergence of which has been the object of much recent analysis [Luo91], [Gai94], [Gri94], [LuT94], [MaS94], [Man93], [Ber95a] (see [BeT96] for our discussion of incremental gradient methods and their application to neural network training). The method where the errors w_t are stochastic includes as a special case the classical Robbins–Monro/stochastic gradient method, as well as methods involving scaling of the gradient and satisfying the pseudogradient condition of Poljak and Tsypkin [PoT73]; see section 4 for a precise statement of our assumptions. Basically, the entire spectrum of unconstrained gradient methods is considered, with the only restriction being the diminishing stepsize condition (1.4) (which is essential for convergence in gradient methods with errors) and the attendant Lipschitz condition (1.2) (which is necessary for showing any kind of convergence result under the stepsize condition (1.4)).

To place our analysis in perspective, we review the related results of the literature for gradient-like methods with errors and in the absence of convexity. Our results relate to two types of analysis:

(1) Results that are based on some type of deterministic or stochastic descent argument, such as the use of a Lyapunov function or a supermartingale convergence theorem. All of the results of this type known to us assume that f is bounded below and in some cases require a boundedness assumption on the sequence $\{x_t\}$ or show only that $\liminf_{t\to\infty} \|\nabla f(x_t)\| = 0$. By contrast, we show that $\lim_{t\to\infty} \|\nabla f(x_t)\| = 0$ and we also deal with the case where f is unbounded below and $\{x_t\}$ is unbounded. In fact, a principal aim of our work has been to avoid any type of boundedness assumption. For example, the classical analysis of Poljak and Tsypkin [PoT73], under essentially the same conditions as ours, shows that if f is bounded below, then $f(x_t)$ converges and $\liminf_{t\to\infty} \|\nabla f(x_t)\| = 0$ (see Poljak [Pol87, p. 51]). The analysis of Gaivoronski [Gai94], for stochastic gradient and incremental gradient methods, under similar conditions to ours shows that $\lim_{t\to\infty} \|\nabla f(x_t)\| = 0$, but it also assumes that $f(x)$ is bounded below and that $\|\nabla f(x)\|$ is bounded over $\Re^n$. The analysis of Luo and Tseng [LuT94] for the incremental gradient method shows that $\lim_{t\to\infty} \|\nabla f(x_t)\| = 0$, but it also assumes that $f(x)$ is bounded below, and it makes some additional assumptions on the stepsize γ_t. The analyses by Grippo [Gri94] and by Mangasarian and Solodov [MaS94] for the incremental gradient method (with and without a momentum term) make assumptions that are different from ours and include boundedness of the generated sequence x_t. The analysis of Walk [Wal92, p. 2] (see also Pflug [Pfl96, p. 282]) shows that $\lim_{t\to\infty} \|\nabla f(x_t)\| = 0$, assuming that $s_t = -\nabla f(x_t)$, that w_t is deterministic and satisfies somewhat different conditions than ours, and that f is bounded below. Our method of proof for the case of deterministic errors is similar to the method of Walk. (The assumption that f is bounded below is not critical for Walk's analysis.) However, in the case of stochastic errors, standard stochastic descent proofs rely critically on the boundedness of f from below, and we have used a new line of proof for our result (see the discussion in section 4).

(2) Results based on the so-called ODE analysis [Lju77], [KuC78], [BMP90], [KuY97] that relate the evolution of the algorithm to the trajectories of a differ-

ential equation $dx/dt = h(x)$. For example, if we are dealing with the stochastic steepest descent method $x_{t+1} = x_t - \gamma_t(\nabla f(x_t) - w_t)$, the corresponding ODE is $dx/dt = -\nabla f(x)$. This framework typically involves an explicit or implicit assumption that the average direction of update $h(x)$ is a well-defined function of the current iterate x. It cannot be applied, for example, to a gradient method with diagonal scaling, where the scaling may depend in a complicated way on the past history of the algorithm, unless one works with differential inclusions—rather than differential equations—for which not many results are available. For another example, an asynchronous gradient iteration that updates a single component at a time (selected by some arbitrary or hard-to-model mechanism) does not lead to a well-defined average direction of update $h(x)$, unless one makes some very special assumptions, e.g., the stepsize assumptions of Borkar [Bor95]. In addition to the above described difficulty, the ODE approach relies on the assumption that the sequence of iterates x_t is bounded or recurrent, something that must be independently verified. Let us also mention the following more recent results by Delyon [Del96], which have some similarities with ours: they are proved using a potential function argument and can establish the convergence of $\nabla f(x_t)$ to zero. Similar to the ODE approach, these results assume a well-defined average update direction $h(x)$ and are based on boundedness or recurrence assumptions.

The paper is organized as follows. In the next section, we focus on the method where there is a nonrandom error w_t satisfying the condition (1.6). The convergence result obtained is then applied in section 3 to the case of incremental gradient methods for minimizing the sum of a large number of functions. In section 4, we focus on stochastic gradient methods. Finally, in section 5, a stochastic version of the incremental gradient method is discussed.

2. Deterministic gradient methods with errors. Throughout the paper, we focus on the unconstrained minimization of a continuously differentiable function $f : \Re^n \mapsto \Re$, satisfying for some constant L

$$\|\nabla f(x) - \nabla f(\overline{x})\| \le L\|x - \overline{x}\| \qquad \forall\ x, \overline{x} \in \Re^n. \tag{2.1}$$

As mentioned in the preceding section, the line of proof of the following proposition is known, although some of our assumptions differ slightly from those in the literature. We will need the following known lemma, which we prove for completeness.

LEMMA 1. *Let Y_t, W_t, and Z_t be three sequences such that W_t is nonnegative for all t. Assume that*

$$Y_{t+1} \le Y_t - W_t + Z_t, \qquad t = 0, 1, \ldots,$$

and that the series $\sum_{t=0}^{T} Z_t$ converges as $T \to \infty$. Then either $Y_t \to -\infty$ or else Y_t converges to a finite value and $\sum_{t=0}^{\infty} W_t < \infty$.

Proof. Let $\bar{t}$ be any nonnegative integer. By adding the relation $Y_{t+1} \le Y_t + Z_t$ over all $t \ge \bar{t}$ and by taking the limit superior as $t \to \infty$, we obtain

$$\limsup_{t\to\infty} Y_t \le Y_{\bar{t}} + \sum_{t=\bar{t}}^{\infty} Z_t < \infty.$$

By taking the limit inferior of the right-hand side as $\bar{t} \to \infty$ and using the fact $\lim_{\bar{t}\to\infty} \sum_{t=\bar{t}}^{\infty} Z_t = 0$, we obtain

$$\limsup_{t\to\infty} Y_t \le \liminf_{\bar{t}\to\infty} Y_{\bar{t}} < \infty.$$

This implies that either $Y_t \to -\infty$ or else Y_t converges to a finite value. In the latter case, by adding the relation $Y_{i+1} \le Y_i - W_i + Z_i$ from $i = 0$ to $i = t$, we obtain

$$\sum_{i=0}^{t} W_i \le Y_0 + \sum_{i=0}^{t} Z_i - Y_{t+1}, \qquad t = 0, 1, \ldots,$$

which implies that $\sum_{i=0}^{\infty} W_i \le Y_0 + \sum_{i=0}^{\infty} Z_i - \lim_{t\to\infty} Y_t < \infty$. ∎

We have the following result.

PROPOSITION 1. *Let x_t be a sequence generated by the method*

$$x_{t+1} = x_t + \gamma_t(s_t + w_t),$$

where s_t is a descent direction satisfying for some positive scalars c_1 and c_2, and all t,

$$c_1\|\nabla f(x_t)\|^2 \le -\nabla f(x_t)' s_t, \qquad \|s_t\| \le c_2\bigl(1 + \|\nabla f(x_t)\|\bigr), \tag{2.2}$$

and w_t is an error vector satisfying for some positive scalars p and q, and all t,

$$\|w_t\| \le \gamma_t\bigl(q + p\|\nabla f(x_t)\|\bigr). \tag{2.3}$$

Assume that the stepsize γ_t is positive and satisfies

$$\sum_{t=0}^{\infty} \gamma_t = \infty, \qquad \sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

Then either $f(x_t) \to -\infty$ or else $f(x_t)$ converges to a finite value and $\lim_{t\to\infty} \nabla f(x_t) = 0$. Furthermore, every limit point of x_t is a stationary point of f.

Proof. Fix two vectors x and z, let ξ be a scalar parameter, and let $g(\xi) = f(x + \xi z)$. The chain rule yields $(dg/d\xi)(\xi) = z'\nabla f(x + \xi z)$. We have

$$\begin{aligned}
f(x+z) - f(x) &= g(1) - g(0) \\
&= \int_0^1 \frac{dg}{d\xi}(\xi)\, d\xi \\
&= \int_0^1 z'\nabla f(x + \xi z)\, d\xi \\
&\le \int_0^1 z'\nabla f(x)\, d\xi + \left| \int_0^1 z'\bigl(\nabla f(x+\xi z) - \nabla f(x)\bigr)\, d\xi \right| \\
&\le z'\nabla f(x) + \int_0^1 \|z\| \cdot \|\nabla f(x + \xi z) - \nabla f(x)\| d\xi \\
&\le z'\nabla f(x) + \|z\| \int_0^1 L\xi \|z\|\, d\xi \\
&= z'\nabla f(x) + \frac{L}{2}\|z\|^2.
\end{aligned} \tag{2.4}$$

We apply (2.4) with $x = x_t$ and $z = \gamma_t(s_t + w_t)$. We obtain

$$f(x_{t+1}) \le f(x_t) + \gamma_t \nabla f(x_t)'(s_t + w_t) + \frac{\gamma_t^2 L}{2}\|s_t + w_t\|^2.$$

Using our assumptions, we have

$$\begin{aligned}\nabla f(x_t)'(s_t + w_t) &\le -c_1\|\nabla f(x_t)\|^2 + \|\nabla f(x_t)\|\,\|w_t\| \\ &\le -c_1\|\nabla f(x_t)\|^2 + \gamma_t q\|\nabla f(x_t)\| + \gamma_t p\|\nabla f(x_t)\|^2.\end{aligned}$$

Furthermore, using the relations $\|s_t\|^2 \le 2c_2^2\big(1 + \|\nabla f(x_t)\|^2\big)$ and $\|w_t\|^2 \le 2\gamma_t^2\big(q^2 + p^2\|\nabla f(x_t)\|^2\big)$, which follow from (2.2) and (2.3), respectively, we have

$$\begin{aligned}\|s_t + w_t\|^2 &\le 2\|s_t\|^2 + 2\|w_t\|^2 \\ &\le 4c_2^2\big(1 + \|\nabla f(x_t)\|^2\big) + 4\gamma_t^2\big(q^2 + p^2\|\nabla f(x_t)\|^2\big).\end{aligned}$$

Combining the above relations, we obtain

$$\begin{aligned}f(x_{t+1}) \le f(x_t) &- \gamma_t(c_1 - \gamma_t p - 2\gamma_t c_2^2 L - 2\gamma_t^3 p^2 L)\|\nabla f(x_t)\|^2 \\ &+ \gamma_t^2 q\|\nabla f(x_t)\| + 2\gamma_t^2 c_2^2 L + 2\gamma_t^4 q^2 L.\end{aligned}$$

Since $\gamma_t \to 0$, we have for some positive constant c and all t sufficiently large

$$f(x_{t+1}) \le f(x_t) - \gamma_t c\|\nabla f(x_t)\|^2 + \gamma_t^2 q\|\nabla f(x_t)\| + 2\gamma_t^2 c_2^2 L + 2\gamma_t^4 q^2 L.$$

Using the inequality $\|\nabla f(x_t)\| \le 1 + \|\nabla f(x_t)\|^2$, the above relation yields for all t

$$f(x_{t+1}) \le f(x_t) - \gamma_t(c - \gamma_t q)\|\nabla f(x_t)\|^2 + \gamma_t^2(q + 2c_2^2 L) + 2\gamma_t^4 q^2 L,$$

which for sufficiently large t can be written as

$$f(x_{t+1}) \le f(x_t) - \gamma_t\beta_1\|\nabla f(x_t)\|^2 + \gamma_t^2\beta_2, \tag{2.5}$$

where β_1 and β_2 are some positive scalars.

By using (2.5), Lemma 1, and the assumption $\sum_{t=0}^{\infty}\gamma_t^2 < \infty$, we see that either $f(x_t) \to -\infty$ or else $f(x_t)$ converges and

$$\sum_{t=0}^{\infty}\gamma_t\|\nabla f(x_t)\|^2 < \infty. \tag{2.6}$$

If there existed an $\epsilon > 0$ and an integer $\bar{t}$ such that $\|\nabla f(x_t)\| \ge \epsilon$ for all $t \ge \bar{t}$, we would have

$$\sum_{t=\bar{t}}^{\infty}\gamma_t\|\nabla f(x_t)\|^2 \ge \epsilon^2\sum_{t=\bar{t}}^{\infty}\gamma_t = \infty,$$

which contradicts (2.6). Therefore, $\liminf_{t\to\infty}\|\nabla f(x_t)\| = 0$.

To show that $\lim_{t\to\infty}\nabla f(x_t) = 0$, assume the contrary; that is, $\limsup_{t\to\infty}\|\nabla f(x_t)\| > 0$. Then there exists an $\epsilon > 0$ such that $\|\nabla f(x_t)\| < \epsilon/2$ for infinitely many t and also $\|\nabla f(x_t)\| > \epsilon$ for infinitely many t. Therefore, there is an infinite subset of integers $\mathcal{T}$ such that for each $t \in \mathcal{T}$, there exists an integer $i(t) > t$ such that

$$\|\nabla f(x_t)\| < \epsilon/2, \qquad \|\nabla f(x_{i(t)})\| > \epsilon,$$

$$\epsilon/2 \le \|\nabla f(x_i)\| \le \epsilon \qquad \text{if } t < i < i(t).$$

Since

$$\begin{aligned} \|\nabla f(x_{t+1})\| - \|\nabla f(x_t)\| &\le \|\nabla f(x_{t+1}) - \nabla f(x_t)\| \\ &\le L\|x_{t+1} - x_t\| \\ &= \gamma_t L\|s_t\| \\ &\le \gamma_t L c_2\big(1 + \|\nabla f(x_t)\|\big), \end{aligned}$$

it follows that for all $t \in \mathcal{T}$ that are sufficiently large so that $\gamma_t L c_2 < \epsilon/4$, we have

$$\epsilon/4 \le \|\nabla f(x_t)\|;$$

otherwise, the condition $\epsilon/2 \le \|\nabla f(x_{t+1})\|$ would be violated. Without loss of generality, we assume that the above relations as well as (2.5) hold for all $t \in \mathcal{T}$.

We have for all $t \in \mathcal{T}$, using the condition $\|s_t\| \le c_2\big(1 + \|\nabla f(x_t)\|\big)$ and the Lipschitz condition (2.1),

$$\begin{aligned} \frac{\epsilon}{2} &\le \|\nabla f(x_{i(t)})\| - \|\nabla f(x_t)\| \\ &\le \|\nabla f(x_{i(t)}) - \nabla f(x_t)\| \\ &\le L\|x_{i(t)} - x_t\| \\ &\le L \sum_{i=t}^{i(t)-1} \gamma_i(\|s_i\| + \|w_i\|) \\ &\le L c_2 \sum_{i=t}^{i(t)-1} \gamma_i\big(1 + \|\nabla f(x_i)\|\big) + L \sum_{i=t}^{i(t)-1} \gamma_i^2\big(q + p\|\nabla f(x_i)\|\big) \\ &\le L c_2(1+\epsilon) \sum_{i=t}^{i(t)-1} \gamma_i + L(q + p\epsilon) \sum_{i=t}^{i(t)-1} \gamma_i^2. \end{aligned} \tag{2.7}$$

From this it follows that

$$\frac{1}{2Lc_2(1+\epsilon)} \le \liminf_{t\to\infty} \sum_{i=t}^{i(t)-1} \gamma_i. \tag{2.8}$$

Using (2.5), we see that

$$f\big(x_{i(t)}\big) \le f(x_t) - \beta_1 \left(\frac{\epsilon}{4}\right)^2 \sum_{i=t}^{i(t)-1} \gamma_i + \beta_2 \sum_{i=t}^{i(t)-1} \gamma_i^2 \qquad \forall\ t \in \mathcal{T}.$$

Using the convergence of $f(x_t)$ already shown and the assumption $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$, this relation implies that

$$\lim_{t\to\infty,\, t\in\mathcal{T}} \sum_{i=t}^{i(t)-1} \gamma_i = 0$$

and contradicts (2.8).

Finally, if $\overline{x}$ is a limit point of x_t, then $f(x_t)$ converges to the finite value $f(\overline{x})$. Thus we have $\nabla f(x_t) \to 0$, implying that $\nabla f(\overline{x}) = 0$. □

3. Incremental gradient methods. In this section, we apply the results of the preceding section to the case where f has the form

$$f(x) = \sum_{i=1}^{m} f_i(x),$$

where $f_i : \Re^n \mapsto \Re$ is for every i a continuously differentiable function satisfying the Lipschitz condition

$$\|\nabla f_i(x) - \nabla f_i(\overline{x})\| \leq L\|x - \overline{x}\| \qquad \forall\ x, \overline{x} \in \Re^n \tag{3.1}$$

for some constant L.

In situations where there are many component functions f_i, it may be attractive to use an incremental method that does not wait to process the entire set of components before updating x; instead, the method cycles through the components in sequence and updates the estimate of x after each component is processed. In particular, given x_t, we may obtain x_{t+1} as

$$x_{t+1} = \psi_m,$$

where ψ_m is obtained at the last step of the algorithm

$$\psi_i = \psi_{i-1} - \gamma_t \nabla f_i(\psi_{i-1}), \qquad i = 1, \ldots, m, \tag{3.2}$$

and

$$\psi_0 = x_t. \tag{3.3}$$

This method can be written as

$$x_{t+1} = x_t - \gamma_t \sum_{i=1}^{m} \nabla f_i(\psi_{i-1}). \tag{3.4}$$

It is referred to as the *incremental gradient method*, and it is used extensively in the training of neural networks. It should be compared with the ordinary gradient method, which is

$$x_{t+1} = x_t - \gamma_t \nabla f(x_t) = x_t - \gamma_t \sum_{i=1}^{m} \nabla f_i(x_t). \tag{3.5}$$

Thus, a cycle of the incremental gradient method through the components f_i differs from an ordinary gradient iteration only in that the evaluation of ∇f_i is done at the corresponding current estimates ψ_{i-1} rather than at the estimate x_t available at the start of the cycle. The advantages of incrementalism in enhancing the speed of convergence (at least in the early stages of the method) are well known; see, for example, the discussions in [Ber95a], [Ber95b], [BeT96].

The main idea of the following convergence proof is that the incremental gradient method can be viewed as the regular gradient iteration where the gradient is perturbed by an error term that is proportional to the stepsize. In particular, if we compare the incremental method (3.4) with the ordinary gradient method (3.5), we see that the error term in the gradient direction is bounded by

$$\sum_{i=1}^{m} \big\|\nabla f_i(\psi_{i-1}) - \nabla f_i(x_t)\big\|.$$

In view of our Lipschitz assumption (3.1), this term is bounded by

$$L \sum_{i=1}^{m} \|\psi_{i-1} - x_t\|,$$

which from (3.2) is seen to be proportional to γ_t. (A more precise argument is given below.)

PROPOSITION 2. *Let x_t be a sequence generated by the incremental gradient method* (3.2)–(3.4). *Assume that for some positive constants C and D, and all $i = 1, \ldots, m$, we have*

$$\|\nabla f_i(x)\| \leq C + D\|\nabla f(x)\| \qquad \forall\ x \in \Re^n. \tag{3.6}$$

Assume also that

$$\sum_{t=0}^{\infty} \gamma_t = \infty, \qquad \sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

Then either $f(x_t) \to -\infty$ or else $f(x_t)$ converges to a finite value and $\lim_{t\to\infty} \nabla f(x_t) = 0$. *Furthermore, every limit point of x_t is a stationary point of f.*

Proof. We formulate the incremental gradient method as a gradient method with errors that are proportional to the stepsize and then apply Proposition 1. For simplicity we will assume that there are only two functions f_i, that is, $m = 2$. The proof is similar when $m > 2$. We have

$$\psi_1 = x_t - \gamma_t \nabla f_1(x_t),$$

$$x_{t+1} = \psi_1 - \gamma_t \nabla f_2(\psi_1).$$

By adding these two relations, we obtain

$$x_{t+1} = x_t + \gamma_t\big(-\nabla f(x_t) + w_t\big),$$

where

$$w_t = \nabla f_2(x_t) - \nabla f_2(\psi_1).$$

We have

$$\|w_t\| \leq L\|x_t - \psi_1\| = \gamma_t L \|\nabla f_1(x_t)\| \leq \gamma_t\big(LC + LD\|\nabla f(x_t)\|\big).$$

Thus Proposition 1 applies. ∎

Condition (3.6) is guaranteed to hold if each f_k is of the form

$$f_k(x) = x' Q_k x + g_k' x + h_k,$$

where each Q_k is a positive semidefinite matrix, each g_k is a vector, and each h_k is a scalar. (This is the generic situation encountered in linear least squares problems.) If $\sum_{k=1}^{K} Q_k$ is positive definite, there exists a unique minimum to which the algorithm must converge. In the absence of positive definiteness, we obtain $\nabla f(x_t) \to 0$ if the optimal cost is finite. If, on the other hand, the optimal cost is $-\infty$, it can be shown that $\|\nabla f(x)\| \geq \alpha$ for some $\alpha > 0$ and for all x. This implies that $f(x) \to -\infty$ and that $\|x\| \to \infty$.

4. Stochastic gradient methods. In this section, we study stochastic gradient methods. Our main result is similar to Proposition 1 except that we let the noise term w_t be of a stochastic nature. Once more, we will prove that $f(x_t)$ converges and, if the limit is finite, $\nabla f(x_t)$ converges to 0. We comment on the technical issues that arise in establishing such a result. The sequence $f(x_t)$ can be shown to be approximately a supermartingale. The variance of the underlying noise is allowed to grow with $\|\nabla f(x_t)\|$ and therefore can be unbounded. While such unboundedness has been successfully handled in past works on related methods, new complications arise because no lower bound on $f(x_t)$ is assumed. For that reason, the supermartingale convergence theorem cannot be used in a simple manner. Our approach is to show that whenever $\|\nabla f(x_t)\|$ is large, it remains so for a sufficiently long time interval, guaranteeing a decrease in the value of $f(x_t)$ which is significant and dominates the noise effects.

PROPOSITION 3. *Let x_t be a sequence generated by the method*

$$x_{t+1} = x_t + \gamma_t(s_t + w_t),$$

where γ_t is a deterministic positive stepsize, s_t is a descent direction, and w_t is a random noise term. Let $\mathcal{F}_t$ be an increasing sequence of σ-fields. We assume the following:

(a) *x_t and s_t are $\mathcal{F}_t$-measurable.*

(b) *There exist positive scalars c_1 and c_2 such that*

$$c_1\|\nabla f(x_t)\|^2 \leq -\nabla f(x_t)' s_t, \qquad \|s_t\| \leq c_2(1 + \|\nabla f(x_t)\|) \qquad \forall\ t. \tag{4.1}$$

(c) *We have, for all t and with probability 1,*

$$E[w_t \mid \mathcal{F}_t] = 0, \tag{4.2}$$

$$E[\|w_t\|^2 \mid \mathcal{F}_t] \leq A(1 + \|\nabla f(x_t)\|^2), \tag{4.3}$$

where A is a positive deterministic constant.

(d) *We have*

$$\sum_{t=0}^{\infty} \gamma_t = \infty, \qquad \sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

Then, either $f(x_t) \to -\infty$ or else $f(x_t)$ converges to a finite value and $\lim_{t\to\infty}$ $\nabla f(x_t) = 0$. *Furthermore, every limit point of x_t is a stationary point of f.*

Remarks. (a) The σ-field $\mathcal{F}_t$ should be interpreted as the history of the algorithm up to time t, just before w_t is generated. In particular, conditioning on $\mathcal{F}_t$ can be thought of as conditioning on $x_0, s_0, w_0, \ldots, x_{t-1}, s_{t-1}, w_{t-1}, x_t, s_t$.

(b) Strictly speaking, the conclusions of the proposition only hold "with probability 1." For simplicity, an explicit statement of this qualification often will be omitted.

(c) Our assumptions on w_t are of the same type as those considered in [PoT73].

Proof of Proposition 3. We apply (2.4) with $x = x_t$ and $z = \gamma_t(s_t + w_t)$. We obtain

$$
\begin{aligned}
f(x_{t+1}) &\le f(x_t) + \gamma_t \nabla f(x_t)'(s_t + w_t) + \frac{\gamma_t^2 L}{2}\|s_t + w_t\|^2 \\
&\le f(x_t) - \gamma_t c_1 \|\nabla f(x_t)\|^2 + \gamma_t \nabla f(x_t)' w_t + \gamma_t^2 L(\|s_t\|^2 + \|w_t\|^2) \\
&\le f(x_t) - \gamma_t c_1 \|\nabla f(x_t)\|^2 + \gamma_t \nabla f(x_t)' w_t + \gamma_t^2 2Lc_2^2 \\
&\qquad + \gamma_t^2 2Lc_2^2 \|\nabla f(x_t)\|^2 + \gamma_t^2 L \|w_t\|^2 \\
&\le f(x_t) - \gamma_t \frac{c_1}{2}\|\nabla f(x_t)\|^2 + \gamma_t \nabla f(x_t)' w_t + \gamma_t^2 2Lc_2^2 + \gamma_t^2 L\|w_t\|^2,
\end{aligned}
\tag{4.4}
$$

where the last inequality is valid only when t is large enough so that $\gamma_t 2Lc_2^2 \le c_1/2$. Without loss of generality, we will assume that this is the case for all $t \ge 0$.

Let $\delta > 0$ be an arbitrary positive number that will be kept constant until the very end of this proof. Let η be a positive constant defined, in terms of δ, by

$$
\eta c_2 \left(\frac{1}{\delta} + 2\right) + \eta = \frac{1}{2L}. \tag{4.5}
$$

We will partition the set of all times t (the nonnegative integers) into a set S of times at which $\|\nabla f(x_t)\|$ is "small" and intervals $I_k = \{\tau_k, \tau_k + 1, \ldots, \tau'_k\}$ during which $\|\nabla f(x_t)\|$ stays "large." The definition of the times τ_k and τ'_k is recursive and is initialized by letting $\tau'_0 = -1$. We then let, for $k = 1, 2, \ldots$,

$$
\tau_k = \min\left\{ t > \tau'_{k-1} \;\middle|\; \|\nabla f(x_t)\| \ge \delta \right\}.
$$

(We leave τ_k undefined if $\|\nabla f(x_t)\| < \delta$ for all $t > \tau'_{k-1}$.) We also let

$$
\begin{aligned}
\tau'_k = \max \Bigg\{ t \ge \tau_k \;\Bigg|\; & \sum_{i=\tau_k}^{t} \gamma_i \le \eta, \text{ and} \\
& \frac{\|\nabla f(x_{\tau_k})\|}{2} \le \|\nabla f(x_r)\| \le 2\|\nabla f(x_{\tau_k})\| \; \forall\, r \text{ with } \tau_k \le r \le t \Bigg\}.
\end{aligned}
$$

We say that the interval I_k is *full* if $\sum_{t=\tau_k}^{\tau'_k+1} \gamma_t > \eta$. Let S be the set of all times that do not belong to any of the intervals I_k.

We define a sequence G_t, used to scale the noise terms w_t, by

$$
G_t = \begin{cases} \delta & \text{if } t \in S, \\ \|\nabla f(x_{\tau_k})\| = H_k & \text{if } t \in I_k, \end{cases}
$$

where the last equality should be taken as the definition of H_k. In particular, G_t is constant during an interval I_t. Note that $G_t \ge \delta$ for all t.

We now collect a few observations that are direct consequences of our definitions.

(P1) For all $t \in S$, we have $\|\nabla f(x_t)\| < \delta = G_t$.

(P2) For all $t \in I_k$, we have

$$
\frac{G_t}{2} = \frac{H_k}{2} \le \|\nabla f(x_t)\| \le 2H_k = 2G_t.
$$

Combining this with (P1), we also see that the ratio $\|\nabla f(x_t)\|/G_t$ is bounded above by 2.

(P3) If τ_k is defined and I_k is a full interval, then

$$\frac{\eta}{2} \leq \eta - \gamma_{\tau'_k+1} < \sum_{t=\tau_k}^{\tau'_k} \gamma_t \leq \eta, \tag{4.6}$$

where the leftmost inequality holds when k is large enough so that $\gamma_{\tau'_k+1} \leq \eta/2$. Without loss of generality, we will assume that this condition actually holds for all k.

(P4) The value of G_t is completely determined by $x_0, x_1, \ldots, x_t$ and is therefore $\mathcal{F}_t$-measurable. Similarly, the indicator function

$$\chi_t = \begin{cases} 1 & \text{if } t \in S, \\ 0 & \text{otherwise} \end{cases}$$

is also $\mathcal{F}_t$-measurable.

LEMMA 2. *Let r_t be a sequence of random variables with each r_t being $\mathcal{F}_{t+1}$-measurable, and suppose that $E[r_t \mid \mathcal{F}_t] = 0$ and $E[\|r_t\|^2 \mid \mathcal{F}_t] \leq B$, where B is some deterministic constant. Then, the sequences*

$$\sum_{t=0}^{T} \gamma_t r_t \qquad \text{and} \qquad \sum_{t=0}^{T} \gamma_t^2 \|r_t\|^2, \qquad T = 0, 1, \ldots,$$

converge to finite limits (with probability 1).

Proof. It is seen that $\sum_{t=0}^{T} \gamma_t r_t$ is a martingale whose variance is bounded by $B \sum_{t=0}^{\infty} \gamma_t^2$. It must therefore converge by the martingale convergence theorem. Furthermore,

$$E\left[\sum_{t=0}^{\infty} \gamma_t^2 \|r_t\|^2\right] \leq B \sum_{t=0}^{\infty} \gamma_t^2 < \infty,$$

which shows that $\sum_{t=0}^{\infty} \gamma_t^2 \|r_t\|^2$ is finite with probability 1. This establishes convergence of the second sequence. ∎

Using Lemma 2, we obtain the following.

LEMMA 3. *The following sequences converge (with probability 1):*

(a) $\displaystyle\sum_{t=0}^{T} \chi_t \gamma_t \nabla f(x_t)' w_t$;

(b) $\displaystyle\sum_{t=0}^{T} \gamma_t \frac{w_t}{G_t}$;

(c) $\displaystyle\sum_{t=0}^{T} \gamma_t \frac{\nabla f(x_t)' w_t}{G_t^2}$;

(d) $\displaystyle\sum_{t=0}^{T} \gamma_t^2 \frac{\|w_t\|^2}{G_t^2}$;

(e) $\displaystyle\sum_{t=0}^{T} \gamma_t^2 \chi_t \|w_t\|^2$.

Proof. (a) Let $r_t = \chi_t \nabla f(x_t)' w_t$. Since χ_t and $\nabla f(x_t)$ are $\mathcal{F}_t$-measurable and $E[w_t \mid \mathcal{F}_t] = 0$, we obtain $E[r_t \mid \mathcal{F}_t] = 0$. Whenever $\chi_t = 1$, we have $\|\nabla f(x_t)\| \leq \delta$

and $E[\|w_t\|^2 \mid \mathcal{F}_t] \le A(1+\delta^2)$. It follows easily that $E[|r_t|^2 \mid \mathcal{F}_t]$ is bounded. The result follows from Lemma 2.

(b) Let $r_t = w_t/G_t$. Since G_t is $\mathcal{F}_t$-measurable and $E[w_t \mid \mathcal{F}_t] = 0$, we obtain $E[r_t \mid \mathcal{F}_t] = 0$. Furthermore,

$$E[\|r_t\|^2 \mid \mathcal{F}_t] \le \frac{A(1+\|\nabla f(x_t)\|^2)}{G_t^2}.$$

Since the ratio $\|\nabla f(x_t)\|/G_t$ is bounded above [cf. observation (P2)], Lemma 2 applies and establishes the desired convergence result.

(c) Let $r_t = \nabla f(x_t)' w_t / G_t^2$. Note that

$$\frac{\nabla f(x_t)' w_t}{G_t^2} \le \frac{\|\nabla f(x_t)\| \cdot \|w_t\|}{G_t^2} \le 2\frac{\|w_t\|}{G_t}.$$

The ratio in the left-hand side has bounded conditional second moment, by the same argument as in the proof of part (b). The desired result follows from Lemma 2.

(d) This follows again from Lemma 2. The needed assumptions have already been verified while proving part (b).

(e) This follows from Lemma 2 because $\chi_t w_t$ has bounded conditional second moment, by an argument similar to the one used in the proof of part (a). ◻

We now assume that we have removed the zero probability set of sample paths for which the series in Lemma 3 does not converge. For the remainder of the proof, we will concentrate on a single sample path outside this zero probability set. Let ϵ be a positive constant that satisfies

$$\epsilon \le \eta, \qquad 2\epsilon + 2L\epsilon \le \frac{c_1\eta}{48}, \qquad 4Lc_2^2\epsilon \le \frac{c_1\delta^2\eta}{48}. \tag{4.7}$$

Let us choose some t_0 after which all of the series in Lemma 3, as well as the series $\sum_{t=0}^{T}\gamma_t^2$, stay within ϵ from their limits.

LEMMA 4. *Let t_0 be as above. If τ_k is defined and is larger than t_0, then the interval I_k is full.*

Proof. Recall that for $t \in I_k = \{\tau_k, \ldots, \tau_k'\}$ we have $G_t = H_k = \|\nabla f(x_{\tau_k})\| \ge \delta$ and $\|s_t\| \le c_2(1+\|\nabla f(x_t)\|) \le c_2(1+2H_k)$. Therefore,

$$\begin{aligned}
\|x_{\tau_k'+1} - x_{\tau_k}\| &\le \sum_{t=\tau_k}^{\tau_k'} \gamma_t \|s_t\| + \left\| \sum_{t=\tau_k}^{\tau_k'} \gamma_t w_t \right\| \\
&= \sum_{t=\tau_k}^{\tau_k'} \gamma_t \|s_t\| + H_k \left\| \sum_{t=\tau_k}^{\tau_k'} \gamma_t \frac{w_t}{G_t} \right\| \\
&\le \eta c_2(1+2H_k) + H_k\epsilon \\
&\le \eta c_2 H_k \left(\frac{1}{\delta} + 2\right) + \eta H_k \\
&= \frac{H_k}{2L},
\end{aligned}$$

where the last equality follows from our choice of η (cf. (4.5)). Thus,

$$\|\nabla f(x_{\tau_k'+1}) - \nabla f(x_{\tau_k})\| \le L\|x_{\tau_k'+1} - x_{\tau_k}\| \le \frac{H_k}{2} = \frac{\|\nabla f(x_{\tau_k})\|}{2},$$

which implies that

$$\frac{1}{2}\|\nabla f(x_{\tau_k})\| \leq \|\nabla f(x_{\tau_k'+1})\| \leq 2\|\nabla f(x_{\tau_k})\|.$$

If we also had $\sum_{t=\tau_k}^{\tau_k'+1} \gamma_t \leq \eta$, then $\tau_k' + 1$ should be an element of I_k, which it isn't. This shows that $\sum_{t=\tau_k}^{\tau_k'+1} \gamma_t > \eta$ and that I_k is a full interval. ∎

Our next lemma shows that after a certain time, $f(x_t)$ is guaranteed to decrease by at least a constant amount during full intervals.

LEMMA 5. *Let t_0 be the same as earlier. If τ_k is defined and larger than t_0, then*

$$f(x_{\tau_k'+1}) \leq f(x_{\tau_k}) - h,$$

where h is a positive constant that depends only on δ.

Proof. Note that I_k is a full interval by Lemma 4. Using (4.4), we have

$$f(x_{t+1}) - f(x_t) \leq -\gamma_t \frac{c_1}{2}\|\nabla f(x_t)\|^2 + \gamma_t \nabla f(x_t)' w_t + \gamma_t^2 2Lc_2^2 + \gamma_t^2 L\|w_t\|^2.$$

We will sum (from τ_k to τ_k') the terms in the right-hand side of the above inequality and provide suitable upper bounds. Recall that for $t \in I_k$, we have $\|\nabla f(x_t)\| \geq H_k/2$. Thus, also using (4.6),

$$-\sum_{t=\tau_k}^{\tau_k'} \gamma_t \frac{c_1}{2}\|\nabla f(x_t)\|^2 \leq -\frac{c_1 H_k^2}{8}\sum_{t=\tau_k}^{\tau_k'} \gamma_t \leq -\frac{c_1 H_k^2 \eta}{16}. \tag{4.8}$$

Furthermore,

$$\sum_{t=\tau_k}^{\tau_k'} \gamma_t \nabla f(x_t)' w_t \leq 2H_k^2 \epsilon, \tag{4.9}$$

which follows from the convergence of the series in Lemma 3(c) and the assumption that after time t_0 the series is within ϵ of its limit. By a similar argument based on Lemma 3(d), we also have

$$L\sum_{t=\tau_k}^{\tau_k'} \gamma_t^2 \|w_t\|^2 \leq 2LH_k^2 \epsilon. \tag{4.10}$$

Finally,

$$2Lc_2^2 \sum_{t=\tau_k}^{\tau_k'} \gamma_t^2 \leq 4Lc_2^2 \epsilon. \tag{4.11}$$

We add (4.8)–(4.11) and obtain

$$\begin{aligned} f(x_{\tau_k'+1}) &\leq f(x_{\tau_k}) - \frac{c_1 \eta H_k^2}{16} + (2\epsilon + 2L\epsilon)H_k^2 + 4Lc_2^2\epsilon \\ &\leq f(x_{\tau_k}) - \frac{2c_1 \eta H_k^2}{48} + \frac{c_1 \eta \delta^2}{48} \\ &\leq f(x_{\tau_k}) - \frac{c_1 \eta \delta^2}{48}. \end{aligned}$$

The second inequality made use of (4.7); the third made use of $H_k \geq \delta$. ∎

LEMMA 6. *For almost every sample path, $f(x_t)$ converges to a finite value or to $-\infty$. If $\lim_{t\to\infty} f(x_t) \neq -\infty$, then $\limsup_{t\to\infty} \|\nabla f(x_t)\| \leq \delta$.*

Proof. Suppose that there are only finitely many intervals I_k and, in particular,

$$\limsup_{t\to\infty} \|\nabla f(x_t)\| \leq \delta.$$

Let t^* be some time such that $t \in S$ for all $t \geq t^*$. We then have $\chi_t = 1$ for all $t \geq t^*$. We use (4.4) to obtain

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \gamma_t \chi_t \nabla f(x_t)' w_t + \gamma_t^2 2Lc_2^2 + \chi_t \gamma_t^2 L \|w_t\|^2 \\ &= f(x_t) + Z_t \qquad \text{for } t \geq t^*, \end{aligned}$$

where the last equality can be taken as the definition of Z_t. Using parts (a) and (e) of Lemma 3, the series $\sum_t Z_t$ converges. Lemma 1 then implies that $f(x_t)$ converges to a finite value or to $-\infty$. This proves Lemma 6 for the case where there are finitely many intervals.

We consider next the case where there are infinitely many intervals. We will prove that $f(x_t)$ converges to $-\infty$. We first establish such convergence along a particular subsequence. Let $\mathcal{T} = S \cup \{\tau_1, \tau_2, \ldots\}$. We will show that the sequence $\{f(x_t)\}_{t\in\mathcal{T}}$ converges to $-\infty$. To see why this must be the case, notice that whenever $t \in S$, we have $f(x_{t+1}) \leq f(x_t) + Z_t$, where Z_t is as in the preceding paragraph and is summable. Also, whenever $t \in \mathcal{T}$ but $t \notin S$, then $t = \tau_k$ for some k, and the next element of $\mathcal{T}$ is the time $\tau_k' + 1$. Using Lemma 5, $f(x_t)$ decreases by at least h during this interval (for k large enough). We are now in the situation captured by Lemma 1, with $W_t = h$ whenever $t = \tau_k$. The convergence of the subsequence $\{f(x_t)\}_{t\in\mathcal{T}}$ follows. Furthermore, since $W_t = h$ infinitely often, the limit can be only $-\infty$.

Having shown that $f(x_{\tau_k})$ converges to $-\infty$, it now remains to show that the fluctuations of $f(x_t)$ during intervals I_k cannot be too large. Because the technical steps involved here are very similar to those given earlier, we provide only an outline. In order to carry out this argument, we consider the events that immediately precede an interval I_k.

Let us first consider the case where I_k is preceded by an element of S, i.e., $\tau_k - 1 \in S$. By replicating the first half of the proof of Lemma 4, we can show that $x_t - x_{\tau_k - 1}$ for $t \in I_k$ is bounded by a constant multiple of δ (for k large enough). Since $\|\nabla f(x_{\tau_k-1})\| \leq \delta$, this leads to a $c\delta^2$ bound on the difference $f(x_t) - f(x_{\tau_k-1})$, where c is some absolute constant. Since $f(x_{\tau_k-1}) \to -\infty$, the same must be true for $f(x_t)$, $t \in I_k$.

Let us now consider the case where I_k is immediately preceded by an interval I_{k-1}. By replicating the proof of Lemma 5 (with a somewhat smaller choice of ϵ), we can show that (for k large enough) we will have $f(x_t) \leq f(x_{\tau_{k-1}})$ for all $t \in I_k$. Once more, since $f(x_{\tau_{k-1}})$ converges to $-\infty$, the same must be true for $f(x_t)$, $t \in I_k$. ∎

According to Lemma 6, $f(x_t)$ converges and if

$$\lim_{t\to\infty} f(x_t) \neq -\infty,$$

then $\limsup_{t\to\infty} \|\nabla f(x_t)\| \leq \delta$. Since this has been proved for an arbitrary $\delta > 0$, we conclude that if $\lim_{t\to\infty} f(x_t) \neq -\infty$, then $\limsup_{t\to\infty} \|\nabla f(x_t)\| = 0$, that is, $\nabla f(x_t) \to 0$.

Finally, if x^* is a limit point of x_t, this implies that $f(x_t)$ has a subsequence that converges to $f(x^*)$. Therefore, the limit of the entire sequence $f(x_t)$, which we have shown to exist, must be finite and equal to $f(x^*)$. We have shown that in this case $\nabla f(x_t)$ converges to zero. By taking the limit of $\nabla f(x_t)$ along a sequence of times such that x_t converges to x^*, we conclude that $\nabla f(x^*) = 0$. □

5. The incremental gradient method revisited. We now provide an alternative view of the incremental gradient method that was discussed in section 4.

Consider again a cost function f of the form

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x),$$

where each f_i is a function from $\Re^n$ into $\Re$ that satisfies the Lipschitz condition (4.1). In contrast to the setting of section 4, we now assume that each update is based on a single component function f_i, chosen at random. More specifically, let $k(t)$, $t = 1, 2, \ldots$, be a sequence of independent random variables, each distributed uniformly over the set $\{1, \ldots, m\}$. The algorithm under consideration is

$$x_{t+1} = x_t - \gamma_t \nabla f_{k(t)}(x_t), \tag{5.1}$$

where γ_t is a nonnegative scalar stepsize. We claim that this is a special case of the stochastic gradient algorithm. Indeed, the algorithm (5.1) can be rewritten as

$$x_{t+1} = x_t - \frac{\gamma_t}{m} \sum_{i=1}^{m} \nabla f_i(x_t) - \gamma_t \left(\nabla f_{k(t)}(x_t) - \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x_t) \right),$$

which is of the form

$$x_{t+1} = x_t - \gamma_t \nabla f(x_t) - \gamma_t w_t,$$

where

$$w_t = \nabla f_{k(t)}(x_t) - \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x_t).$$

We now verify that w_t satisfies the assumptions of Proposition 3. Due to the way that $k(t)$ is chosen, we have

$$E\big[\nabla f_{k(t)}(x_t) \mid \mathcal{F}_t\big] = \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x_t),$$

from which it follows that $E[w_t \mid \mathcal{F}_t] = 0$. We also have

$$\begin{aligned} E\big[\|w_t\|^2 \mid \mathcal{F}_t\big] &= E\big[\big\|\nabla f_{k(t)}(r_t)\big\|^2 \mid \mathcal{F}_t\big] - \big\|E\big[\nabla f_{k(t)}(r_t) \mid \mathcal{F}_t\big]\big\|^2 \\ &\le E\big[\big\|\nabla f_{k(t)}(r_t)\big\|^2 \mid \mathcal{F}_t\big], \end{aligned}$$

which yields

$$E\big[\|w_t\|^2 \mid \mathcal{F}_t\big] \le \max_k \big\|\nabla f_k(x_t)\big\|^2.$$

Let us assume that there exist constants C and D such that

$$\big\|\nabla f_i(x)\big\| \le C + D\big\|\nabla f(x)\big\| \qquad \forall\ i,\ x \tag{5.2}$$

(cf. the assumption of Proposition 2). It follows that

$$E\left[\|w_t\|^2 \mid \mathcal{F}_t\right] \leq 2C^2 + 2D^2\left\|\nabla f(x_t)\right\|^2$$

so that condition (4.3) is satisfied and the assertion of Proposition 3 holds.

REFERENCES

[BMP90] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, New York, 1990.

[BeT89] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[BeT96] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[Ber95a] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.

[Ber95b] D. P. Bertsekas, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim., 7 (1997), pp. 913–926.

[Bor95] V. S. Borkar, *Asynchronous stochastic approximations*, SIAM J. Control Optim., 36 (1998), pp. 840–851.

[Del96] B. Delyon, *General results on the convergence of stochastic algorithms*, IEEE Trans. Automat. Control, 41 (1996), pp. 1245–1255.

[Gai94] A. A. Gaivoronski, *Convergence analysis of parallel backpropagation algorithm for neural networks*, Optim. Methods Software, 4 (1994), pp. 117–134.

[Gri94] L. Grippo, *A class of unconstrained minimization methods for neural network training*, Optim. Methods Software, 4 (1994), pp. 135–150.

[KuC78] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1978.

[KuY97] H. J. Kushner and G. Yin, *Stochastic Approximation Methods*, Springer-Verlag, New York, 1996.

[Lju77] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, 22 (1977), pp. 551–575.

[LuT94] Z. Q. Luo and P. Tseng, *Analysis of an approximate gradient projection method with applications to the backpropagation algorithm*, Optim. Methods Software, 4 (1994), pp. 85–101.

[Luo91] Z. Q. Luo, *On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks*, Neural Comput., 3 (1991), pp. 226–245.

[MaS94] O. L. Mangasarian and M. V. Solodov, *Serial and parallel backpropagation convergence via nonmonotone perturbed minimization*, Optim. Methods Software, 4 (1994), pp. 103–116.

[Pfl96] G. Pflug, *Optimization of Stochastic Models. The Interface Between Simulation and Optimization*, Kluwer, Boston, 1996.

[PoT73] B. T. Poljak and Y. Z. Tsypkin, *Pseudogradient adaptation and training algorithms*, Automat. Remote Control, 12 (1973), pp. 83–94.

[Pol87] B. T. Poljak, *Introduction to Optimization*, Optimization Software Inc., New York, 1987.

[TBA86] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, 31 (1986), pp. 803–812.

[Wal92] H. Walk, *Foundations of stochastic approximation*, in Stochastic Approximation and Optimization of Random Systems, L. Ljung, G. Pflug, and H. Walk, eds., Birkhauser, Boston, 1992, pp. 53–93.