Statistica Neerlandica (2008) Vol. 62, nr. 3, pp. 393–403
doi:10.1111/j.1467-9574.2008.00397.x

Convergence of stochastic approximation algorithms under irregular conditions

Jian Zhang*

Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF, UK

Faming Liang†

Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA

We consider a class of stochastic approximation (SA) algorithms for solving a system of estimating equations. The standard condition for the convergence of the SA algorithms is that the estimating functions are locally Lipschitz continuous. Here, we show that this condition can be relaxed to the extent that the estimating functions are bounded and continuous almost everywhere. As a consequence, the use of the SA algorithm can be extended to some problems with irregular estimating functions. Our theoretical results are illustrated by solving an estimation problem for exponential power mixture models.

Keywords and Phrases: stochastic approximation algorithm, M-estimator, exponential power mixture models.

1 Introduction

Suppose that we have p-dimensional observations $(X_1, \ldots, X_N)$ from a parametric or semiparametric model, which depends on an unknown parameter $\theta \in \Theta \subset R^d$. Inference about θ, e.g. finding its M-estimate and empirical likelihood confidence intervals, often involves a search for roots of the simultaneous equations

$$\frac{1}{N}\sum_{i=1}^{N} H_j(\theta, X_i) = 0, \quad j = 1, \ldots, d, \tag{1}$$

for given estimating functions $H_j(\cdot,\cdot): \mathbb{R}^d \times \mathbb{R}^p \mapsto \mathbb{R}$, $j = 1, \ldots, d$ (See Huber 1981; Newey, 1993; Tian *et al.*, 2004). As the partial derivative of an objective function for optimization, $H = (H_j)_{1 \le j \le d}^{\tau}$ is usually non-additive with respect to X_i. When H is smooth and $\partial H/\partial\theta$ is of full rank, the equations can be solved directly by the standard Newton–Raphson iteration. However, when H is not smooth or $\partial H/\partial\theta$ is

*J.Zhang@kent.ac.uk
†fliang@stat.tamu.edu

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

not of full rank, the stochastic approximation (SA) algorithm is found to be more appropriate for solving (1).

The subject of stochastic approximation was founded by Robbins and Monro (1951). After five decades of continual development, this method has been widely used in adaptive control, system identification and optimization problems. See Lai (2003) for a review. The standard convergence analysis of the SA algorithm is based on the condition that H is locally Lipschitz continuous, although the weaker assumptions are found as well (Tadić, 1997; Chen, 2002; Kushner and Yin, 2003). Among those conditions, Kushner (1981) required that $E[H(x, Z_n)]$ or $E[H(x, Z_n) \mid Z_{n-1}, \ldots]$ is a smooth function of x. He, Fu and Marcus (2003) required a convexity condition for the objective function. These conditions turn out to be either invalid or hard to verify. In this note, we show that for the convergence of the SA algorithm, the condition of local Lipschitz continuity can be relaxed to the extent that H is bound and continuous almost everywhere. As a consequence, the use of the SA algorithm can be extended to some problems with irregular estimating functions.

The remainder of this paper is organized as follows. In section 2, a general form for the SA algorithm with non-additive noise is introduced. In section 3, two sets of conditions are presented for proving the convergence of the SA algorithm. The application to estimating exponential power mixture models is described in section 4. A real data application is presented in section 5. The conclusions are made in section 6. Some basic lemmas are given in the last section.

2 A stochastic approximation algorithm with non-additive noise

Let $H(\cdot, \cdot)$ be an $\mathbb{R}^d \times \mathbb{R}^p \mapsto \mathbb{R}^d$ measurable function as defined before, and $\mathcal{P}$ be a probability measure for a random variable Z drawn from $\mathbb{R}^p$. The problem is to seek the roots of the following system of equations,

$$h_i(\theta) = \int H_i(\theta, z)\mathcal{P}(\mathrm{d}z) = 0, \quad i = 1, \ldots, d. \tag{2}$$

In what follows, we denote $H(\theta, Z) = (H_1(\theta, Z), \ldots, H_d(\theta, Z))^\tau$ and $h(\theta) = (h_1(\theta), \ldots, h_d(\theta))^\tau$. Assuming that solutions to (2) exist, we investigate when these roots can be found by the iterative algorithm:

$$\theta_{n+1} = \theta_n + \gamma_n H(\theta_n, Z_{n+1}), \tag{3}$$

where θ_n denotes the estimate obtained at iteration n, γ_n is the gain factor, $H(\theta_n, Z_{n+1})$ is a non-additive function of Z_{n+1} and $\{Z_n\}$ is a random process defined on $(\Omega, \mathcal{F}, \mathcal{P})$ and is referred to as noise. Here, $H(\theta_n, Z_{n+1})$ is non-additive if for any two deterministic functions H_1 and H_2, it cannot be written in the form of

$$H(\theta_n, Z_{n+1}) = H_1(\theta_n) + H_2(Z_{n+1}).$$

As $H(\theta_n, Z_{n+1})$ is a non-additive function of Z_{n+1}, algorithm (3) is referred to as a SA algorithm with non-additive noise. The SA algorithm circumvents the

differentiability requirement of the Newton–Raphson method by taking advantages of the bootstrap and the SA (Efron and Tibshinari, 1993).

3 A new convergence theorem for the SA algorithm

In some situations, the objective function (or the associated Lyapunov function) $v(\theta)$ of problem (2) may not be differentiable but have the bounded directional partial derivatives at some points. Even when $v(\theta)$ is differentiable, its derivative $\nabla v(\theta)$ may not be Hölder continuous. The question is whether the SA algorithm converges under such irregular settings. The answer is partially positive as shown in the remainder of this section.

To facilitate presentation, we define a few more notations as follows. For any $u_1 \in \mathbb{R}^d$, letting

$$A = \left\{ \lim_{u_i \to u} \frac{\partial G(u_i)}{\partial u} : \frac{\partial G(u_i)}{\partial u} \text{ exists and is finite} \right\},$$

we define the generalized gradient $\partial G(u) = \text{col } A$ as a convex closure of A (Clark, 1983). When $G(u)$ is differentiable at u, we simply set $\partial G(u) = \{\partial G(u)/\partial u\}$ as a singleton. Similarly, the generalized directional derivative is defined as

$$G^0(u; v) = \limsup_{y \to u, t \to 0^+} \frac{G(y + tv) - G(y)}{t}.$$

G is called regular if for all u: (i) G has the usual one-sided directional derivative; (ii) this derivative is equal to the corresponding generalized directional derivative.

Let Z_k be drawn from the finite set $\{X_1, \ldots, X_N\}$ randomly with replacement, $\{Z_1, \ldots, Z_n\}$ can be regarded as a random sample of P_N, where P_N is an empirical distribution defined on the set $\{X_1, \ldots, X_N\}$. Hence, $h(\theta)$ can be written as

$$h(\theta) = \frac{1}{N} \sum_{k=1}^{N} H(\theta, X_k). \tag{4}$$

In the following, we let $\mathbb{R}^+$ denote the set of non-negative reals, $\| \cdot \|$ denote the Euclidian norm in $\mathbb{R}^d$ and the Frobenius norm in $\mathbb{R}^{d \times d}$ and $d(\cdot, \cdot)$ denote the distance induced by the Euclidian norm. Assume that there is an objective function $v(\theta)$ with the set of stationary points $E_* = \{\theta \in \Theta : 0 = h(\theta)\}$. Set

$$\Theta_0 = \{\theta : h(\cdot) \text{ is not continuous at } \theta, \text{or } v(\theta) \text{ is not differentiable at } \theta\}.$$

$$\Xi = \{\{\theta_n, n \geq 0\} : \text{there is a closed subset } D \text{ of } \Theta - \Theta_0 \text{ such that } \theta_n \in D, n \geq 0\}.$$

Let $b_{n+1} = H(\theta_n, Z_{n+1}) - h(\theta_n)$ and θ_0 be any point in Θ. Then the SA algorithm has the form

$$\theta_{n+1} = \theta_n + \gamma_n (h(\theta_n) + b_{n+1}).$$

We analyze the SA algorithm under the following conditions:

B1. $0 < \lim_{n\to\infty} \gamma_n = 0, \quad \sum_{n=1}^{\infty} \gamma_n = \infty, \quad \sum_{n=1}^{\infty} \gamma_n^2 < \infty, \quad \sup_{n\geq 0} E\|b_{n+1}\|^2 < +\infty.$

B2. $v(\theta)$ is continuous with finite generalized gradient $\partial v(\theta)$. For $\theta \in \Theta$, $-h(\theta) \in \partial v(\theta)$. For any compact subset $Q \subset \Theta - E_*$,

$$\sup_{Q} -\|h(\theta)\| < 0.$$

THEOREM 1. *Assume that the conditions B_1–B_2 hold. Then, as $n \to \infty$, we have*

$$\lim_{n\to\infty} d(\theta_n, E_*) = 0 \text{ almost surely on } \Xi,$$

where $d(\theta_n, E_*) = \inf_{\theta\in E_*} \| \theta_n - \theta \|$.

Proof. Let $t_0 = 0$, $t_k = \sum_{i=0}^{k-1} \gamma_i$ and $m(t)$ be the k such that $t \in [t_k, t_{k+1})$. For $t \in \mathbb{R}^+$, define

$$\theta^k(t) = \theta_k + \sum_{i=k}^{m(t+t_k)-1} \gamma_i h(\theta_i) + \sum_{i=k}^{m(t+t_k)-1} \gamma_i b_{i+1}.$$

Then

$$\theta^k(t) = \theta_k + \int_0^t h(\theta^k(s))\mathrm{d}s + B^k(t) + \rho^k(t)$$

with

$$B^k(t) = \sum_{i=k}^{m(t+t_k)-1} \gamma_i b_i, \quad \rho^k(t) = \int_{t_{m(t+t_k)}}^{t} h(\theta^k(s))\mathrm{d}s.$$

By conditions B1 and B2, as $k \to \infty$,

$$\sup_{0\leq t<\infty} |\rho^k(t)| \leq \sup_{0\leq t<\infty} \gamma_{m(t+t_k)} \|h(\theta_{m(t+t_k)})\| \to 0.$$

Note that $\{\gamma_i b_{i+1}, \mathcal{F}_i : i \geq k\}$ is a martingale sequence and

$$\sup_{m\geq 1} E\left\|\sum_{i=1}^{m} \gamma_i b_{i+1}\right\| \leq \sup_{m\geq 1}\left[E\sum_{i=1}^{m} \gamma_i^2 b_{i+1}^2\right]^{1/2}$$
$$\leq \left[\sum_{i=1}^{m} \gamma_i^2\right]^{1/2} \sup_{n\geq 0}(E\|b_{n+1}\|^2)^{1/2} < \infty.$$

It follows from Doob's theorem (SHIRYAYEV, 1984, Theorem 1, p. 476) that

$$\lim_{k\to\infty} \sup_{m\geq k}\left\|\sum_{i=k}^{m} \gamma_i b_{i+1}\right\| = 0, \text{almost surely},$$

which yields

$$\sup_{0\le t<\infty} \|B^k(t)\| \le \sup_{m\ge k} \left\| \sum_{i=k}^{m} \gamma_i b_{i+1} \right\| \to 0$$

almost surely as $k \to \infty$.

If we define $V^k(t) = \int_0^t h(\theta^k(s))\,\mathrm{d}s$, then

$$\sup_{0\le t<\infty} \|\theta^k(t) - \theta^k(0) - V^k(t)\| = o(1), \tag{5}$$

and

$$\| V^k(t) - V^k(s)\| \le \sup_{t\le u\le s} \|h(\theta^k(u))\|\,|t-s|,$$

which implies that $\theta^k(t)$ is also equicontinuous on any finite interval in $\mathbb{R}^+$ provided $\{\theta_n\}$ is contained in a compact subset D of Θ. By the Arzelà–Ascoli theorem, there exist a subsequence, say $\{k_j\}$, and a $\mathbb{R}^d$-valued function $\theta(t)$ on $\mathbb{R}^+$ such that

$$\theta^{k_j}(\cdot) \to \theta(\cdot),$$

where $\theta(\cdot)$ is also equicontinuous on any finite interval in $\mathbb{R}^+$.

By conditions B2, $h(\cdot)$ is bounded on any closed subset of $\Theta - \Theta_0$ and $h(\theta^k(s)) \to h(\theta(s))$ when $\theta(s) \notin \Theta_0$. And the set $\{0 \le s \le t : \theta(s) \notin \Theta_0\}$ has the Lebesgue measure tending to 0. Using the dominated convergence theorem, we have

$$V^k(t) = \int_0^t h(\theta^k(s))\mathrm{d}s \to \int_0^t h(\theta(s))\,\mathrm{d}s.$$

This together with (5) yields

$$\theta(t) = \theta(0) + \int_0^t h(\theta(s))\,\mathrm{d}s.$$

Thus, the limit $\theta(s)$ of any convergent subsequence of $\{\theta_n : n \ge 0\}$ in Ξ satisfies the differential equation:

$$\dot\theta(s) = h(\theta(s)), \quad s \in \mathbb{R}^+.$$

Invoking the chain rule of derivatives and condition B2, we have

$$\dot v(\theta(s)) = -h(\theta(s))^\tau \dot\theta(s) = -\|h(\theta(s))\|^2.$$

Therefore,

$$v(\theta(t)) - v(\theta(0)) = -\int_0^t \|h(\theta(s)\|^2 \mathrm{d}s.$$

Note that $v(\theta)$ is continuous. On Ξ, for any convergent subsequence $\{\theta(t_j) : j = 1, 2, \ldots\}$ of $\{\theta(t), t \in \mathbb{R}^+\}$, with $\theta(t_j) \to \theta_*$, we have that $v(\theta(t_j)) \to v(\theta_*)$ and

$$v(\theta_*) - v(\theta(0)) = -\int_0^\infty \| h(\theta(s)) \|^2 \mathrm{d}s > -\infty,$$

which implies that for any $\delta > 0$, as $t_j \to \infty$,

$$\int_{t_j-\delta/2}^{t_j+\delta/2} \| h(\theta(s)) \|^2 \mathrm{d}s \to 0. \tag{6}$$

Note that $h(\theta)$ is continuous at θ_* and $\theta(t)$ is equicontinuous. Thus, for any $\epsilon > 0$, there existsa $\delta > 0$ such that

$$\| h(\theta(t)) - h(\theta_*) \| < \epsilon,$$

when $| t - t_j | < \delta$. This together with (6) yields that

$$(\| h(\theta_*) \| - \epsilon)\delta \leq 0.$$

Therefore $h(\theta_*) = 0$, meaning that $\theta_* \in \{\theta : 0 = h(\theta)\}$. Now using the same technique in Chen (2002, p. 15), we can easily prove the desired result. □

4 Exponential power mixture model estimation

Suppose that $(X_1, \ldots, X_N)$ is a sample from a mixture of exponential power distributions,

$$f(x \mid \theta) = \sum_{k=1}^{m} \omega_k \varphi(x \mid \mu_k, \Sigma_k, \beta_k), \tag{7}$$

where $\theta = (\mu_1, \Sigma_1, \beta_1; \ldots; \mu_m, \Sigma_m, \beta_m; \omega_1, \ldots, \omega_{m-1})$ contains all parameters of the model; m is the total number of components; ω_k is the weight of the kth component, $0 < \omega_k \leq 1$, and $\sum_{k=1}^{m} \omega_k = 1$; and $\varphi(x \mid \mu_k, \Sigma_k, \beta_k)$ is the Kotz-type distribution, which has the density

$$\varphi(x \mid \mu_k, \Sigma_k, \beta_k) = \frac{\beta_k \Gamma(p/2)}{2\pi^{p/2} \mid \Sigma \mid^{1/2} \Gamma(p/(\beta_k))} \exp\{-[(x-\mu_k)^\tau \Sigma_k^{-1} (x-\mu_k)]^{\beta_k/2}\},$$
$$\mu_k \in R^p, \quad \Sigma_k \in \mathcal{S}, \quad | \beta_k | \leq c_0,$$

where $\mathcal{S}$ is the set of all $p \times p$ positive definite matrices and c_0 is a sufficiently large constant (see Fang, Kotz and Ng, 1990, p. 76). The location parameter μ_k stands for the centre of the distribution, the matrix Σ_k for the dispersion and the shape parameter β_k for the rate of exponential decay. For $\beta_k = 2$, the distribution is $N(\mu_k, \Sigma_k/2)$, a multivariate normal distribution; for $0 < \beta_k < 1$, the distributed is heavy tailed; and for $\beta_k > 1$, the distribution is light tailed. To reduce the technical burden, we focus on the case where $\Sigma_k = \lambda_k^2 I_p$ with $\lambda_k > 0$ and I_p being a $p \times p$ unit matrix. Then

$$\varphi(x \mid \mu_k, \Sigma_k, \beta_k) = \frac{\beta_k \Gamma(p/2)}{2\pi^{p/2} \lambda_k^p \Gamma(p/\beta_k)} \exp\left\{ - \left(\frac{| \, | x - \mu_k | \, |}{\lambda_k} \right)^{\beta_k} \right\}.$$

Here λ_k is used to show the volume of the kth component and β_k is employed to describe the shape of the kth component. We re-parameterize them by $\lambda_k^* = \log(\lambda_k)$ and $\beta_k^* = \log(\beta_k)$. In addition, we re-parameterize $\omega_1, \ldots, \omega_{m-1}$ by $\omega_1^*, \ldots, \omega_{m-1}^*$ by setting $\omega_m^* = 0$ and

$$\omega_k = \frac{e^{\omega_k^*}}{\sum_{j=1}^m e^{\omega_j^*}}, \quad k = 1, \ldots, m-1.$$

For uniformity of notations, we define $\mu_k^* \equiv \mu_i$. With the above notations, model (7) can be reparametrized by $\theta = (\mu_1^*, \lambda_1^*, \beta_1^*; \ldots; \mu_m^*, \lambda_m^*, \beta_m^*; \omega_1^*, \ldots, \omega_{m-1}^*)$. In what follows, for simplicity, we still use $f(x \mid \theta)$ to denote the mixture density with $\theta \in R^d$ and $d = (p+3)m - 1$. Note that when $\beta_k^* < 0$, $\log f(z \mid \theta)$ is not differential at $\mu_k = X_k$ and $\| \partial \log f(X_k \mid \theta)/\partial \mu_k \|$ tends to infinity as $\mu_k \to X_k$. Removing these singular points from $\mathbb{R}^d$, we define

$$R_z = \{\theta \in \mathbb{R}^d : \text{ there exist } i, k \text{ such that } \mu_i \in \{X_1, \ldots, X_N\} \text{ and } \beta_k^* < 0\},$$
$$\Theta = \mathbb{R}^d \setminus R_z.$$

If there exists k such that $\beta_k^* = 0$, then the corresponding $\log(f(X_k \mid \theta)$ is not differentiable at $\mu_k = X_k$ but has a generalized gradient. For other $\theta \in \Theta$, the corresponding $\log f(x \mid \theta)$ is continuously differentiable.

For a given value of m, model (7) can be estimated by maximizing the function

$$l_N(\theta) = \sum_{k=1}^N \log f(X_k \mid \theta),$$

which is asymptotically equivalent to minimize the Kullback–Leibler distance between $f(\cdot \mid \theta)$ and the underlying density of X. Differentiating $l_N(\theta)$ with respect to θ, we have

$$\frac{\partial l_N(\theta)}{\partial \mu_k^*} = \sum_{i=1}^N P(k \mid X_i) \frac{\beta_k}{\lambda_k^{\beta_k}} \mid \mid X_i - \mu_k \mid \mid^{\beta_k - 2} (X_i - \mu_k), \tag{8}$$

$$\frac{\partial l_N(\theta)}{\partial \lambda_k^*} = \sum_{i=1}^N P(k \mid X_i) \left[p - \frac{\beta_k}{\lambda_k^{\beta_k}} \| X_i - \mu_k \|^{\beta_k} \right], \tag{9}$$

$$\frac{\partial l_N(\theta)}{\partial \beta_k^*} = \sum_{i=1}^N P(k \mid X_i) \times \left\{ 1 + \frac{p}{\beta_k} \frac{\Gamma'(p/\beta_k)}{\Gamma(p/\beta_k)} - \beta_k \left(\frac{\mid \mid X_i - \mu_k \mid \mid}{\lambda_k} \right)^{\beta_k} \log \left(\| X_i - \mu_k \| \right) \right\}, \tag{10}$$

$$\frac{\partial l_N(\theta)}{\partial \omega_k^*} = \sum_{i=1}^N \left[P(k \mid X_i) - \omega_k \right]. \tag{11}$$

where the index k ranges from 1 to m in (8)–(10) and ranges from 1 to $m-1$ in (11);

$$P(k \mid x) = \omega_k \varphi(x \mid \mu_k, \lambda_k^2 I_p, \beta_k)/f(x \mid \theta)$$

is the posterior probability that x is from the k-component; and

$$\Gamma'(x)/\Gamma(x) = -1/x - \upsilon + \sum_{k=1}^{\infty}[1/k - 1/(k+x)].$$

Here, $\upsilon \approx 0.577216$ is called Euler's constant. Note that if $\beta_k^* = 0$ (i.e. $\beta_k = 1$) and $\mu_k^* = X_j, \mu_k^* \neq X_i, i \neq j$, then $\partial l_N(\theta)/\partial \mu_k^*$ in (8) does not exist. However, under this setting, $l_N(\theta)$ is regular and has the generalized gradient

$$\partial l_N(\theta) = \sum_{i \neq j}^{N} P(k \mid X_i) \frac{1}{\lambda_k} \frac{(X_i - \mu_k)}{\| X_i - \mu_k \|} + P(k \mid X_j)) \frac{1}{\lambda_k} S_p,$$

where $S_p = \{a \in \mathbb{R}^p : \|a\| \leq 1\}$. Define

$$H_{\mu_k^*}(\theta, x) = -P(k \mid x) \frac{\beta_k}{\lambda_k^{\beta_k}} \|x - \mu_k\|^{\beta_k - 2}(x - \mu_k), \quad k = 1, \ldots, m, \tag{12}$$

$$H_{\lambda_k^*}(\theta, x) = -P(k \mid x) \left[p - \frac{\beta_k}{\lambda_k^{\beta_k}} \|x - \mu_k\|^{\beta_k} \right], \quad k = 1, \ldots, m, \tag{13}$$

$$H_{\beta_k^*}(\theta, x) = -P(k \mid x) \times \left\{ 1 + \frac{p\Gamma'(p/\beta_k)}{\beta_k \Gamma(p/\beta_k)} - \beta_k \left(\frac{\| x - \mu_k \|}{\lambda_k} \right)^{\beta_k} \log(\|x - \mu_k\|) \right\}, k = 1, \ldots, m \tag{14}$$

$$H_{\omega_k^*}(\theta, x) = P(i \mid x) - \omega_k, \quad k = 1, \ldots, m-1, \tag{15}$$

where we set

$$\frac{(x - \mu_k)}{\| x - \mu_k \|} = a \quad \text{with } \| a \| = 1 \quad \text{when } x = \mu_k.$$

Applying the stochastic approximation method defined in section 2 to solve the system of equations,

$$\begin{cases} \int H_{\mu_k^*}(\theta, z)\, \mathrm{d}F_N(z) = 0, & k = 1, \ldots, m, \\ \int H_{\lambda_k^*}(\theta, z)\, \mathrm{d}F_N(z) = 0, & k = 1, \ldots, m, \\ \int H_{\beta_k^*}(\theta, z)\, \mathrm{d}F_N(z) = 0, & k = 1, \ldots, m, \\ \int H_{\omega_k^*}(\theta, z)\, \mathrm{d}F_N(z) = 0, & k = 1, \ldots, m-1, \end{cases} \tag{16}$$

where F_N is the empirical distribution introduced by $(X_1, \ldots, X_N)$. Let

$$\theta_t = (\mu_1^{*(t)}, \lambda_1^{*(t)}, \beta_1^{*(t)}; \ldots; \mu_m^{*(t)}, \lambda_m^{*(t)}, \beta_m^{*(t)}; \omega_1^{*(t)}, \ldots, \omega_{m-1}^{*(t)})$$

denote the estimate of θ obtained at iteration t and Z_t denotes a sample randomly drawn from the set $\{X_1, \ldots, X_N\}$ at iteration t. In practice, we set the gain factor sequence

$$\gamma_t = \frac{\gamma_0 t_0}{\max(t_0, t)}, \quad t = 1, 2, \ldots, \tag{17}$$

for some values of $t_0 > 1$ and $\gamma_0 > 0$. Our default setting for them is $t_0 = 10,000$ and $\gamma_0 = 0.02$. If other values were used, the values would be specified in the context. We also set a default value for the total number of iterations. It is $500 \times N$, where N is the sample size. See Liang and Zhang (2008) for details.

To show the convergence of the algorithm, we first verify conditions B1–B2 under the restriction $\beta*_k > 0, 1 \le k \le m$. This can be performed as follows.

> By construction of the gain factor sequence, condition B1 is satisfied.
> Define $v(\theta) = -l_N(\theta)/N$. Clearly, $v(\theta)$ is bounded below, because the exponential power density function and thus its mixture are bounded above. According to the order of the components of θ, we define $H_j(\theta, x)$ to be the respective functions $H_{\mu_k^*}(\theta, x)$, $H_{\alpha_k^*}(\theta, x)$, $H_{\beta_k^*}(\theta, x)$ or $H_{\omega_k^*}(\theta, x)$ defined in (16). By definition,
>
> $$\nabla v(\theta) = -\int H(\theta, z) \mathrm{d}\mathcal{P}(z) = -h(\theta) \quad \text{and} \quad \dot{v}(\theta) = -\| h(\theta) \|^2 \le 0.$$
>
> When $\beta_k^* \ge 0,\ 1 \le k \le m,$ condition B_2 is satisfied by defining $E_* = \{\theta : h(\theta) = \mathbf{0}\}$.

5 Applications

In microarray data analysis, one often needs to test a large number of hypotheses simultaneously. To control the probability of erroneously rejecting true null hypotheses, we need to control the false discovery rate (FDR), the expected proportion of false-positive findings among all the rejected hypotheses. The main difficulty in calculating the FDR comes from modelling the observed test scores (Efron, 2004).

To set up notations, let $H_1, \ldots, H_N$ denote the collection of N null hypotheses, $P_1, \ldots, P_N$ denote the corresponding P-values of the N tests, and $Z_i = \Phi^{-1}(P_i)$ or $Z_i = \Phi^{-1}(1 - P_i)$ denote the corresponding test scores, where Φ is the cumulative distribution function (CDF) of the standard normal distribution. Liang and Zhang (2008) proposed the following model for these test scores:

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z), \tag{18}$$

where π_0 is the prior probability that a null hypothesis is true, f_0 is the empirical null distribution and f_1 is the alternative one. f_0 and f_1 are modelled by two exponential power mixture models. The unknown parameters in f_0 and f_1 are then estimated by the SA algorithm introduced in the previous section.

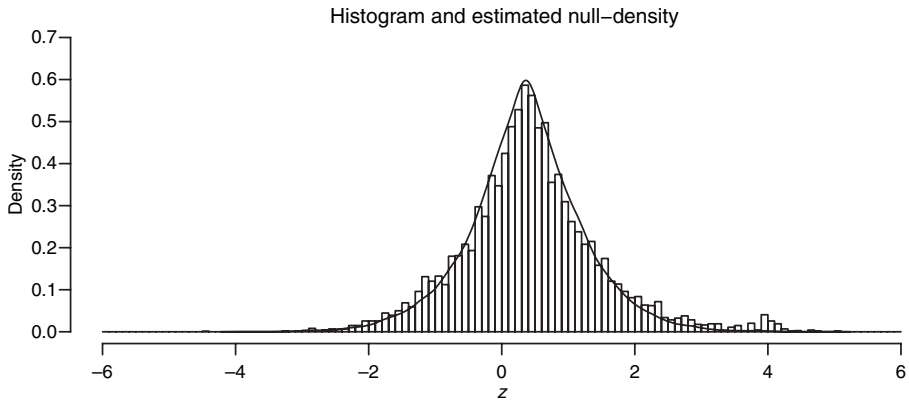

Fig. 1. z-scores of avian pineal gland gene expression data under the DD condition. The solid line is the estimated f_0 by the SA.

5.1 A real data analysis

The avian pineal gland contains both circadian oscillators and photoreceptors to produce rhythms in biosynthesis of the hormone melatonin *in vivo* and *in vitro*. It is of great interest to understand the genetic mechanism driving the rhythms. For this purpose, Dr V. Cassone's laboratory at Texas A&M University measured the expression levels of pineal gland genes under light–dark (LD) and constant darkness (DD) conditions. Under LD, the birds were killed at 2, 6, 10, 14, 18, 22 hours Zeitgeber time (ZT) to obtain mRNA to produce adequate cDNA libraries. Four microarray chips per time point were produced, and there were two replicates for each gene in each chip. The experiment was then repeated under DD. Each chip produced gene expressions for 7400+ genes. Throughout the experiment, samples from LD ZT18 were used as controls. Relative gene expression levels to the controls were recorded and processed. Our goal was to identify genes that are differentially expressed at different time points. Mixed effect analysis with the fixed effect being the different time points and the random effects corresponding to chips and biological batches were applied to the relative gene expression levels in log scale. Normalization procedures have been adopted but will not be listed here as they are not the focus of the paper. Under both LD and DD conditions, P-values, P_is, for testing the existence of different time effects were produced and transformed to test scores using $\Phi^{-1}(1 - P_i)$. Figure 1 shows the histogram of the test scores under the DD condition and the corresponding estimated density curve of f_0. It suggests that f_0 can be well estimated by the SA method.

6 Discussion

In this paper, we have introduced a stochastic approximation method for finding the roots of multiple estimating equations. One important problem is whether the

proposed algorithm converges. The classical convergence analysis is often based on the regular condition that the estimating functions are locally Lipschitz continuous. Here, we have shown that these conditions can be relaxed to the extent that the estimating functions are bounded and continuous almost everywhere.

The SA algorithm has been found a number of applications such as model-based clustering and semiparametric estimating. Here, we applied the SA algorithm to estimate the null density of the test scores in multiple testing problem. A real data application has been presented.

Acknowledgements

This paper is devoted to EURANDOM for celebrating its 10th anniversary. The first author would like to thank Professor Alessandro Di Bucchianico for his kind invitation to the jubilee conference.

References

Chen, H. F. (2002), *Stochastic approximation and its applications*, Kluwer Academic Publishers, London.

Clark, F. (1983), *Optimization and nonsmooth analysis*, Wiley, New York.

Efron, B. (2004), Large-scale simultaneous hypothesis testing: the choice of a null hypothesis, *Journal of the American Statistical Association* **99**, 96–104.

Efron, B. and R. J. Tibshirani (1993), *An introduction to the bootstrap*, Chapman & Hall, New York.

Fang, K., S. Kotz and K. Ng (1990), *Symmetric multivariate and related distributions*, Chapman & Hall, New York.

He, Y., M. C. Fu and S. I. Marcus (2003), Convergence of simultaneous perturbation stochastic approximation for nondifferentiable optimization, *IEEE Transactions on Automatic Control* **48**, 1459–1463.

Huber, P. (1981), *Robust statistics*, Wiley, New York.

Kushner, H. J. (1981), Stochastic approximation with discontinuous dynamics and state dependent noise: w.p.1 and weak convergence, *Journal of Mathematical Analysis and Applications* **82**, 527–542.

Kushner, H. J. and G. Yin (2003), *Stochastic approximation and recursive algorithms and applications*, Springer, New York.

Lai, T. L. (2003), Stochastic approximation, *The Annals of Statistics* **31**, 391–406.

Liang, F. and J. Zhang (2008), Estimating FDR using the stochastic approximation algorithm, *Biometrika*, in press.

Newey, W. K. (1993), Efficient estimation of models with conditional moment restrictions, in: G. S. Maddala, C. R. Rao and H. D. Vinod (eds), *Econometrics, Handbook of Statistics 11*, North-Holland, Amsterdam, pp. 419–453.

Robbins, H. and S. Monro (1951), A stochastic approximation method, *Annals of Mathematical Statistics* **22**, 400–407.

Shiryayev, A. N. (1984), *Probability*, Springer, New York.

Tadić, V. (1997), On the convergence of stochastic iterative algorithms and their applications to machine learning. A short version of this paper was published in *Proceedings of the 36th IEEE Conference on Decision & Control,* San Diego, CA, USA, pp. 2281–2286.

Tian, L., J. S. Liu, Y. Zhao and L. J. Wei (2004), Statistical inferences based on non-smooth estimating functions, *Biometrika* **91**, 943–954.

Received: March 2008, Revised: April 2008.