FUNCTION OPTIMIZATION BY SIMULTANEOUS PERTURBATION STOCHASTIC APPROXIMATION WITH RANDOMLY VARYING TRUNCATIONS

Katsuji Uosaki[*], **Toshiharu Hatanaka**[*], **Akihiko Yonemochi**[†], **Han-Fu Chen**[‡]

[*] Department of Information and Physical Sciences, Graduate School of Information Science and Technology,
Osaka University, Suita, Osaka 565-0871 Japan, E-mail: { uosaki, hatanaka }@ist.osaka-u.ac.jp
[†] Seiko Epson Corp., Suwa, Nagano 392–8502 Japan
[‡] Institute of Systems Science, Chinese Academy of Science,
Beijing 100080, P. R. China, E-mail: hfchen@iss03.iss.ac.cn

Keywords: Simultaneous perturbation stochastic approximation, Randomly varying truncations, Optimization, Finite difference stochastic approximation.

Abstract

A new recursive algorithm is proposed for finding the minimum of an objective function whose gradient is not obtainable directly but is approximated from the noisy observations of the function. The algorithm is based on the simultaneous perturbation stochastic approximation method (SPSA) combined with randomly varying truncations, and provides the estimate, which is convergent under weaker conditions than the conventional SPSA. Numerical simulation studies illustrate the applicability of the proposed algorithm.

1 Introduction

Many engineering problems such as system design, modeling and control can be reduced to optimization (maximization or minimization) problems of a certain mathematical objective function under some constraints. Generally, the solution to the optimization problem corresponds to finding the parameters where the gradient of the function with respect to the concerned parameters is zero. Sometimes, the information of the gradient of the objective function cannot be available exactly or difficult to compute. In such cases, we have to consider approaches that rely on the gradient approximations evaluated from (noisy) observations of the objective function.

One of such approaches is Kiefer-Wolfowitz stochastic approximation [3]. It is an optimization version of Robbins-Monro stochastic approximation method [4] developed to find a root of a regression function. Since Kiefer-Wolfowitz stochastic approximation is based on standard finite difference gradient approximations, it may sometimes be called by finite difference stochastic approximation (FDSA). It requires $2p$ observations of the objective functions at each iteration, and there may be difficulties in computation burden for higher dimensional problems. In contrast to this FDSA, Spall [5] proposed a new stochastic approximation method called by simultaneous perturbation stochastic approximation (SPSA) that is based on the gradient approximation relied on only a pair of observations. Hence, in this approach, the required number of observations is independent of dimension of the problem.

Though these approaches are applicable to wide class of objective functions and observation noise processes, the idea of a randomly varying truncations introduced by Chen [2] is employed, in this paper, to extend the applicable class of objective functions and observation noise processes. Numerical simulation studies illustrate the usefulness of the proposed idea.

2 Function Optimization by Finite Difference Stochastic Approximation and Simultaneous Perturbation Stochastic Approximation

Consider function minimization problem for an objective function $f(\theta) : R^p \to R^1,\ (p \geq 1)$. Here, we assume the value of the objective function $f(\theta)$ cannot be obtained directly but with observation noise w as follows.

$$y(\theta) = f(\theta) + w \tag{1}$$

Let $g(\theta)$ be the gradient of the objective function $f(\theta)$, if available, i.e.,

$$g(\theta) = \frac{\partial f(\theta)}{\partial \theta} \tag{2}$$

then the parameters θ^* that provides the minimum of $f(\theta)$ satisfies $g(\theta^*) = 0$. Stochastic approximation method provides the estimate θ recursively by using the approximate of the gradient $\hat{g}(\theta)$ constructed by the noisy observations of $f(\theta)$

$$\begin{aligned} &\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}(\hat{\theta}_k) \qquad (k = 0, 1, \cdots) \\ &\hat{\theta}_0 : \text{initial estimate} \end{aligned} \tag{3}$$

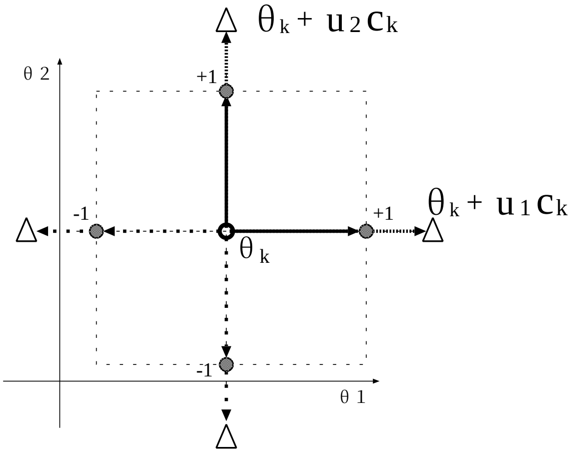


Figure 1: Observation points in FDSA

with a suitable gain sequence $\{a_k\}$.

2.1 Finite Difference Stochastic Approximation

Conventional stochastic approximation method for function optimization, which was first proposed by Kiefer and Wolfowitz [3] and then extended by Blum [1] to multi-dimensional functions, uses the approximate of gradient $g(\theta): R^p \to R^p$ by the finite difference,

$$\hat{g}_k(\hat{\theta}_k) = \frac{1}{2c_k}\begin{pmatrix} \tilde{y}_k^{(1+)} - \tilde{y}_k^{(1-)} \\ \vdots \\ \tilde{y}_k^{(p+)} - \tilde{y}_k^{(p-)} \end{pmatrix} \tag{4}$$

with the noisy observations of $f(\theta)$ at the observation points $\hat{\theta}_k \pm c_k u_l,\ (l = 1, 2, \cdots,\ p)$

$$\begin{aligned} \tilde{y}_k^{(l+)} &= f(\hat{\theta}_k + c_k u_l) + w_k^{(l+)} \\ \tilde{y}_k^{(l-)} &= f(\hat{\theta}_k - c_k u_l) - w_k^{(l-)} \qquad (l = 1, 2, \cdots,\ p) \end{aligned} \tag{5}$$

where u_l is a unit vector of the direction of the lth coordinate in R^p $(l = 1, 2, \cdots,\ p)$ and $\{c_k\}$ is a positive scalar number sequence. We call this stochastic approximation method as the finite difference stochastic approximation (FDSA). Figure 1 shows the observation points for the case of $p = 2$ in FDSA. Theorem 1 gives sufficient conditions for the estimate $\hat{\theta}_k$ by FDSA (3), (4) to converge almost surely to the point θ^* that provides the minimum of the objective function $f(\theta)$.

[Theorem 1]
Assume

A0: The third derivative of $f(\theta)$ is finite.

A1: The point $\theta^* \in R^p$ is an asymptotically stable solution of the differential equation

$$\frac{d\theta(t)}{dt} = -g(\theta)$$

where $g = \nabla f$.

A2: Let $\theta(t|\theta_0)$ be a solution of the differential equation in A1 with initial condition $\theta(0) = \theta_0$, and $D(\theta^*) = \{\theta_0 : \lim_{t\to\infty} \theta(t|\theta_0) = \theta^*\}$ (i.e., this means that $D(\theta^*)$ is the domain of attraction of the differential equation in A1). Then, there exists a subset $S \subseteq D(\theta^*)$ such that $\hat{\theta}_k \in S$ infinitely often for almost all sample processes.

A3: $\sup_k \|\hat{\theta}_k\| < \infty$ a.s.

A4: $\{w_k^{(+)} - w_k^{(-)}\}$ is a martingale difference sequence, i.e., $E[w_k^{(+)} - w_k^{(-)}|\hat{\theta}_0, \hat{\theta}_1, \cdots, \hat{\theta}_{k-1}] = 0$.

A5: $\{a_k\}, \{c_k\}$ are positive number sequences satisfying

$$\lim_{k\to\infty} a_k = 0, \quad \lim_{k\to 0} c_k = 0,$$
$$\sum_{k=1}^{\infty} a_k = \infty, \quad \sum_{k=1}^{\infty} (\frac{a_k}{c_k})^2 < \infty$$

Then the estimate $\hat{\theta}_k$ by FDSA converges the minimum point θ^* with probability one, i.e.,

$$\Pr\{\lim_{k\to\infty} \hat{\theta}_k = \theta^*\} = 1$$

2.2 Simultaneous Perturbation Stochastic Approximation

Since FDSA needs $2p$ observations to estimate gradient vector $g(\theta)$ in each iteration, the computation burden to estimate the gradients becomes larger for higher dimension problems. Spall [5] proposed a new stochastic approximation method called SPSA (simultaneous perturbation stochastic approximation), which approximates the gradient vector with only a pair of observations of the objective function in each iteration. The difference of the number of observations between FDSA and SPSA is p-fold and is important in higher dimension problems. The SPSA gradient is approximated by

$$\hat{g}_k(\hat{\theta}_k) = \frac{1}{2c_k}\begin{pmatrix} \Delta_{k1}^{-1} \\ \vdots \\ \Delta_{kp}^{-1} \end{pmatrix} (\tilde{y}_k^{(+)} - \tilde{y}_k^{(-)}) \tag{6}$$

with a pair of observations

$$\begin{aligned} \tilde{y}_k^{(+)} &= f(\hat{\theta}_k + c_k\Delta_k) + w_k^{(+)} \\ \tilde{y}_k^{(-)} &= f(\hat{\theta}_k - c_k\Delta_k) - w_k^{(-)} \end{aligned} \tag{7}$$

where the components Δ_{kl} $(l = 1, 2, \cdots,\ p)$ of the perturbation Δ_k are independent Bernoulli random variables taking values $+1$ and -1 with probability $1/2$. The observation points for the case of $p = 2$ in FDSA is shown in Fig.2. Convergence conditions for SPSA are given in Theorem 2.

[Theorem 2]
In addition to the conditions of Theorem 1, we assume the following.

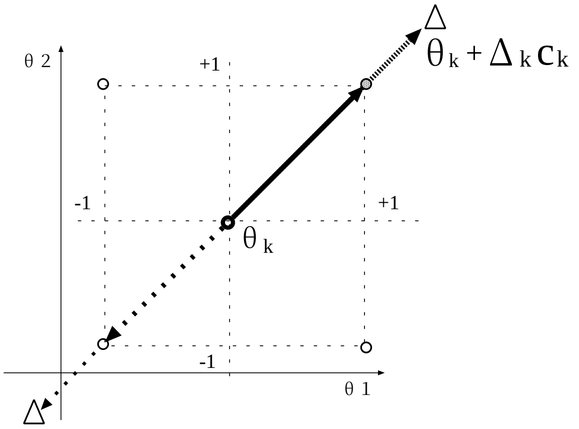


Figure 2: Observation points in SPSA

A6: There exist positive constants C_0 and C_1 such that

$$E[w_k^2] < C_0,\ \ E[f^2(\hat{\theta}_k \pm c_k\Delta_k)] < C_1$$

Then the estimate $\hat{\theta}_k$ by FDSA converges to the minimum point θ^* with probability one, i.e.,

$$\Pr\{\lim_{k\to\infty} \hat{\theta}_k = \theta^*\} = 1$$

3 SPSA with Randomly Varying Truncations

To relax the convergence conditions given in Theorem 2, we combine Chen's idea of randomly varying truncations [2] originally developed to the Robbins- Monro type stochastic approximation method for finding the solution of regression functions with SPSA for function minimization. The new stochastic approximation method, SPSA with randomly varying truncations, will be called, hereinafter, as RTSPSA.

Let $\{K_j\}$ be an increasing positive number sequence satisfying

$$K_j > 0, \quad K_j < K_{j+1}, \quad \lim_{j\to\infty} K_j = \infty \tag{8}$$

and $\hat{\theta}^*$ be a rough estimate of θ^*. In RTSPSA, a preliminary estimate $\{\tilde{\theta}_k\}$ is generated by the conventional SPSA,

$$\begin{aligned} \tilde{\theta}_{k+1} &= \hat{\theta}_k - a_k\hat{g}(\hat{\theta}_k) \qquad (k = 0, 1, \cdots) \\ \hat{g}_k(\hat{\theta}_k) &= \frac{1}{2c_k}\begin{pmatrix} \Delta_{k1}^{-1} \\ \vdots \\ \Delta_{kp}^{-1} \end{pmatrix} (\tilde{y}_k^{(+)} - \tilde{y}_k^{(-)}) \\ \tilde{y}_k^{(+)} &= f(\hat{\theta}_k + c_k\Delta_k) + w_k^{(+)} \\ \tilde{y}_k^{(-)} &= f(\hat{\theta}_k - c_k\Delta_k) - w_k^{(-)} \end{aligned} \tag{9}$$

Then, the estimate is obtained by

$$\begin{aligned} \hat{\theta}_{k+1} &= \tilde{\theta}_{k+1} I_{[\|\tilde{\theta}_k - \hat{\theta}^*\| \le K_{\sigma_k}]} + \hat{\theta}^* I_{[\|\tilde{\theta}_k - \hat{\theta}^*\| > K_{\sigma_k}]} \\ \sigma_k &= \sum_{i=1}^{k-1} I_{[\|\tilde{\theta}_k - \hat{\theta}^*\| > K_{\sigma_i}]} \\ \sigma_0 &= 0 \end{aligned} \tag{10}$$

where I_x is the indicator function such that $I_x = 1$ (if x is true), $I_x = 0$ (if x is false). This implies that the preliminary estimate $\tilde{\theta}_{k+1}$ is replaced by a priori estimate $\hat{\theta}^*$ if the deviation between them exceeds the threshold K_{σ_k} determined by the number of truncations until the iteration and the preliminary estimate is used as the estimate otherwise.

The convergence conditions are relaxed by applying the randomly varying truncations as in the following theorem.

[Theorem 3]

Under the following conditions C0 through C4, RTSPSA estimate converges to the maximum point θ^* with probability one, i.e.,

$$\Pr\{\lim_{k\to\infty} \hat{\theta}_k = \theta^*\} = 1$$

C0: $f(\theta)$ is twice continuously differentiable, and θ^* is a point that gives the global minimum of $f(\theta)$. Moreover, $\inf_{\|\theta - \theta^*\| = s} f(\theta) > f(\tilde{\theta})$ holds for $\tilde{\theta}$ such that $\|\tilde{\theta} - \theta^*\| < s$.

C1: Let $\{a_k\}$ and $\{c_k\}$ are positive number sequences satisfying.

$$\lim_{k\to\infty} a_k = 0, \quad \lim_{k\to 0} c_k = 0,$$
$$\sum_{k=1}^{\infty} a_k = \infty, \quad \sum_{k=1}^{\infty} a_k^r < \infty, \quad r \in (1, 2]$$

C2: Components of the perturbation Δ_k are mutually independent identically distributed random variables satisfying $0 < |\Delta_{kl}| < C$, $E[\Delta_{kl}] = 0$ for a suitable positive constant C.

C3: Observation noise w_k is decomposed into sum of two components as

$$w_{kl} = e_{kl} + v_{kl}$$

such that

$$\sum_{k=1}^{\infty} \frac{a_k e_{k+1,l}}{c_k \Delta_{k+1,l}} < \infty,$$
$$\lim_{k\to\infty} \frac{v_{k+1,l}}{c_k \Delta_{kl}} = 0 \quad (l = 1, 2, \cdots, p)$$

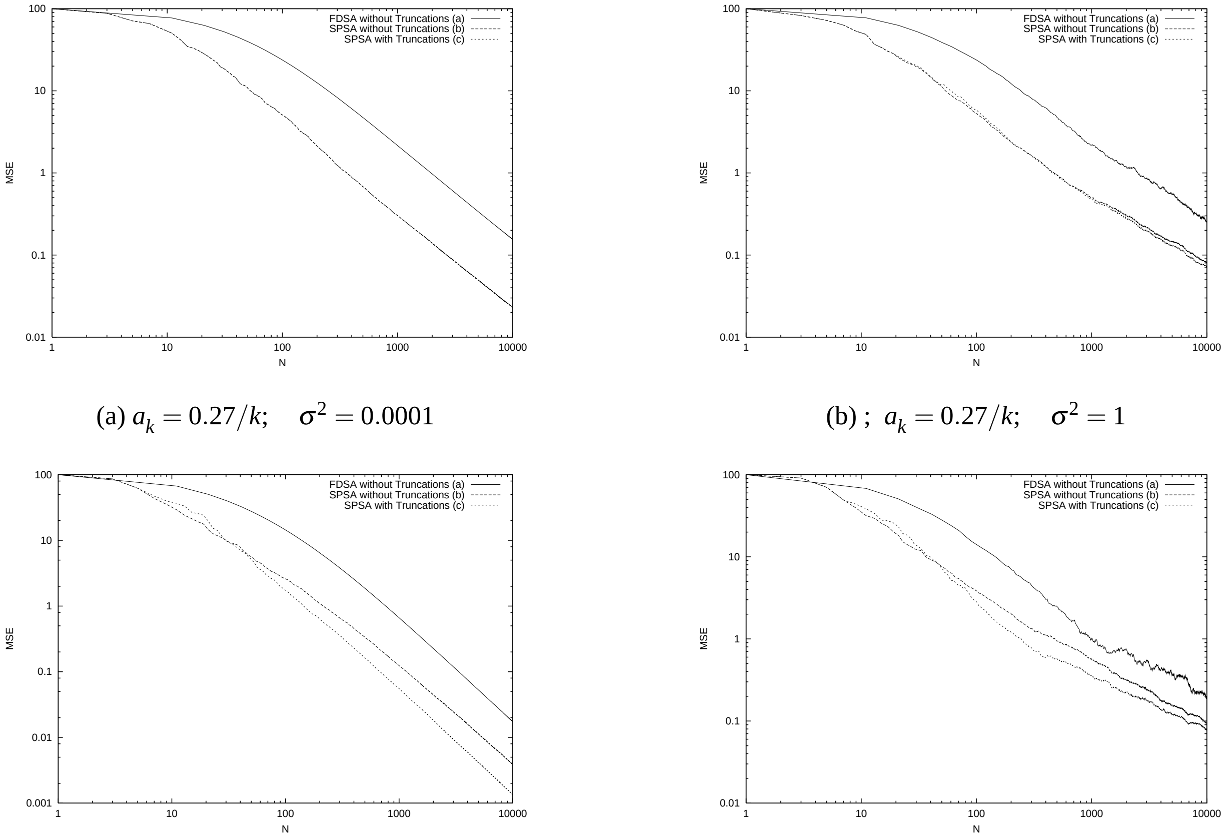
4 Numerical Simulations

The proposed RTSPSA is applied to find the minimum of the functions.

[Example 1]

Consider the following function:

$$f(\theta) = \|\theta\|^2 + 0.1\sum_{i=1}^{5} \theta_i^3 + 0.01\sum_{i=1}^{5} \theta_i^4$$



(a) $a_k = 0.27/k; \quad \sigma^2 = 0.0001$

(b) ; $a_k = 0.27/k; \quad \sigma^2 = 1$

(c) $a_k = 0.4/k; \quad \sigma^2 = 0.0001$

(d) ; $a_k = 0.4/k; \quad \sigma^2 = 1$

Figure 3: Behaviors of mean square errors (Example 1)

whose minimum is attained at $\theta^* = (0, \cdots, 0)^T$. Observation noise w_k is normally distributed with mean 0 and variance σ^2. Initial estimate and the rough estimate are both set as $\theta_0 = \hat{\theta}^* = (10, \cdots, 10)^T$, and gain constants are chosen as $a_k = 0.27/k$or $a_k = 0.4/k \quad c_k = 0.06^{-1/6}, \quad K_j = 3.6^{j+1}$,

[Example 2]

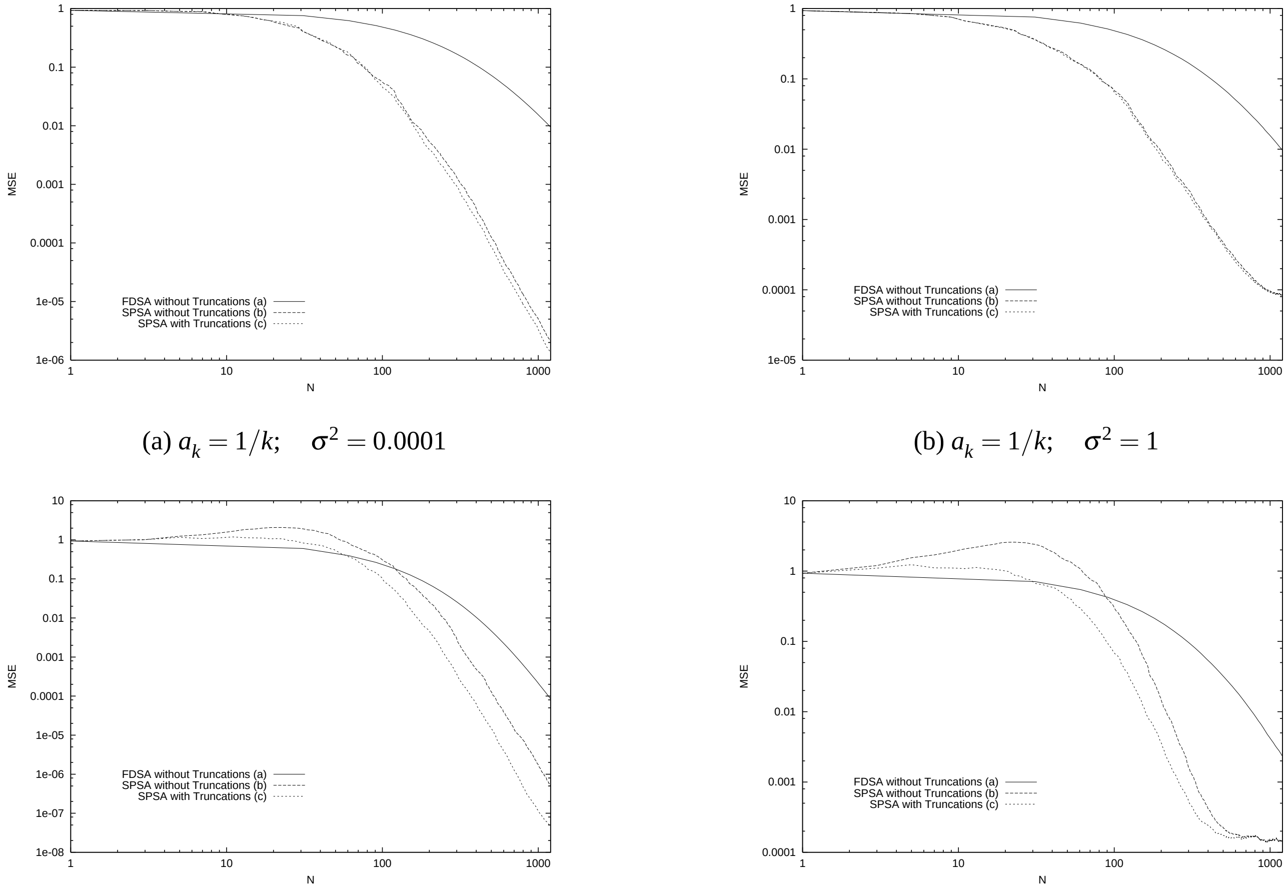
We consider a higher dimensional function,

$$f(\theta) = \|\theta\|^2 + 0.1 \sum_{i=1}^{15} \exp(\frac{\theta_i}{15})$$

Its minimum is attained at $\theta^* = (-0.033259, \cdots, -0.033259)^T$. Observation noise w_k is normally distributed with mean 0 and variance σ^2. Initial estimate and the rough estimate are set as $\theta_0 = \hat{\theta}^* = (-1, \cdots, -1)^T$ and gain constants are chosen as $a_k = 1/k$or $a_k = 2/k$, $c_k = 25.22k^{-5/12}$, $K_j = 1.3^{j+1}$. Mean squared errors for 100 simulation runs are compared with those of (a) FDSA and (b) conventional SPSA. Simulation results indicate that RTSPSA is superior to both FDSA and conventional SPSA in convergence speed and also in stability for larger observation noise cases. Furthermore, it is found that introduction of the randomly varying truncations prevents over-correction by using larger gain as well. It is founded from other simulation results not shown here that randomly varying truncation prevents the estimate from divergence for bigger gain constant case.

5 Conclusions

This paper presents a new stochastic approximation procedure, simultaneous perturbation stochastic approximation with randomly varying truncations (RTSPSA), for finding the minimum of the objective function. This approach is applicable when the gradient of the objective function is not available exactly but is approximated with noisy observations. By introducing the idea of randomly varying truncations, the convergence conditions for the applicable objective functions and observation noise processes are relaxed. Numerical simulations confirm this. Investigation of the optimal choice for the design parameters in the RTSPSA is necessary for practical use, and it is under study by considering the asymptotic distribution of the estimation estimate[6].



(a) $a_k = 1/k; \quad \sigma^2 = 0.0001$

(b) $a_k = 1/k; \quad \sigma^2 = 1$

(c) $a_k = 2/k; \quad \sigma^2 = 0.0001$

(d) $a_k = 2/k; \quad \sigma^2 = 1$

Figure 4: Behaviors of mean square errors (Example 2)

Acknowledgements

This work is supported in part by the Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (C)(2)10650432 and (C)(2)14550447.

References

[1] Blum, J. R., "Multidimensional stochastic approximation methods," *Ann. Math. Statist.*,Vol. 25,pp. 737–744,1954.

[2] Chen, H.-F. and Y. Zhu, "Stochastic approximation procedures with randomly varying truncations," *Scientia Sinica* (Series A), Vol. 29, pp. 914–926 ,1986.

[3] Kiefer, J. and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Statist.*, Vol. 23, pp. 462–466, 1952.

[4] Robbins, H. and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, Vol. 22, pp.400–407, 1951.

[5] Spall, J. C., "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Automatic Contr.*, Vo. 37,pp. 332–341,1992.

[6] Yonemochi, A., *Stochastic Approximation with Random Varying Truncations and Its Applications to Nonlinear Filtering and Function Optimization*, MS Thesis, Department of Information and Knowledge Engineering, Tottori University, 2001.