Acceleration of Stochastic Approximation by Averaging

Speaker: Weihuan Huang

School of Data Science, Fudan University

The work of B.T. Polyak and A.B. Juditsky, *SICON* (1992)

Oct. 11, 2021

Overview

1. Introduction

2. Linear Problem

3. Nonlinear Problem

4. Stochastic Optimization Problem

Review of Stochastic Approximation

Let $R(x) := \mathrm{E}[Y(x,\xi)]$ be a performance measure, in which

- $Y(x,\xi)$: sample performance (observed)
- ξ: stochastic disturbance, x in a parameter space

Root-finding problem:

- Given a constant b, find the root $x^* = \{x|R(x) = b\}$

Optimization problem:

- Find $x^* = \arg\min_x R(x)$ or $x^* = \arg\max_x R(x)$

Stochastic approximation (SA) method:

- Iterates algorithm $x_{n+1} = x_n + a_n \hat{\nabla} R(x_n)$
  - RM (root-finding): R nondecreasing and $R'(x^*) > 0$, $\hat{\nabla} R(x_n) = b - y_n$
  - KW (optimization): find maxima, R concave, $\hat{\nabla} R(x_n) = \frac{y_{2n} - y_{2n-1}}{c_n}$

  in which y_n: observation from $Y(x_n,\xi)$, y_{2n}: obsv from $Y(x_n + c_n, \xi)$, and y_{2n-1}: obsv from $Y(x_n - c_n, \xi)$

Motivations of Averaging Algorithm

Wiki:" While the RM algorithm is theoretically able to achieve $O(1/n)$ under the assumption of twice continuous differentiability and strong convexity, it can perform quite poorly upon implementation. This is primarily due to the fact that the algorithm is very sensitive to the choice of the step size sequence a_n, and the supposed asymptotically optimal step size policy can be quite harmful in the beginning."

It require a large amount of a prior information on R, ∇R, and/or $\nabla^2 R$, which is hard to obtain in most situations. E.g.

- Chung (1954) and Fabian (1968): achieve optimal convergence rate $O(1/\sqrt{n})$ with $a_n = \nabla^2 R(x^*)^{-1}/n$ or $= \frac{1}{n\nabla R(x^*)}$
- $\cdots$

Averaging algorithm does not require such information

Motivations of Averaging Algorithm

Influence of step-size a_n on the finite-time performance

- If a_n too large, then iterates jump back and forth without approaching x^*
- If a_n too small, then iterates barely move

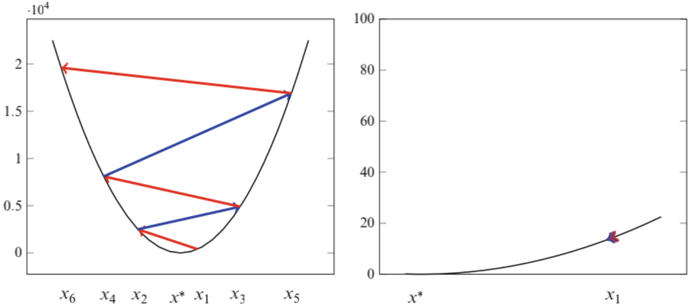


Fig. 6.1 Sensitivity of SA to step size a_n when a_n is "too large" relative to the gradient (left graph) and when a_n is "too small" relative to the gradient (right graph)

Motivations of Averaging Algorithm

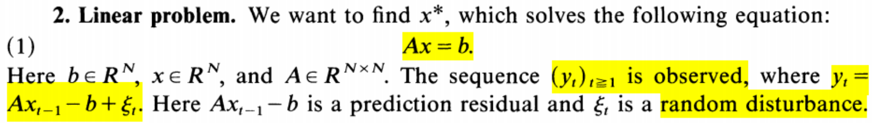
Averaging algorithm reduce sensitivity of a_n (put less emphasis on the last iterate)

- Take longer steps
- averaging of the iterates (Incorporate a subset of the iterates into the output to decrease the reliance on the last iterate)

Averaging algorithm: a more robust step size policy

Linear Problem

Linear root-searching problem:



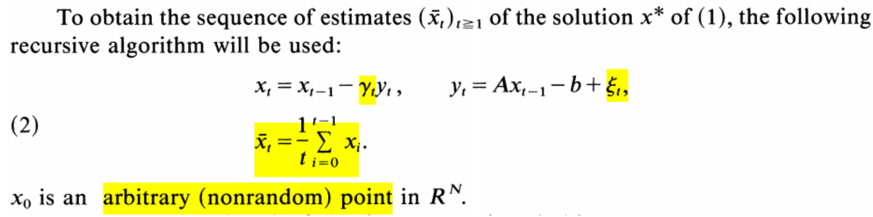
2. Linear problem. We want to find x^*, which solves the following equation:

$$Ax = b. \tag{1}$$

Here $b \in R^N$, $x \in R^N$, and $A \in R^{N \times N}$. The sequence $(y_t)_{t \geqq 1}$ is observed, where $y_t = Ax_{t-1} - b + \xi_t$. Here $Ax_{t-1} - b$ is a prediction residual and ξ_t is a random disturbance.

Linear performance measure $R(x) = Ax - b$ and the root-finding problem $R(x) = 0$, and sample performance y_t.

Averaging algorithm:

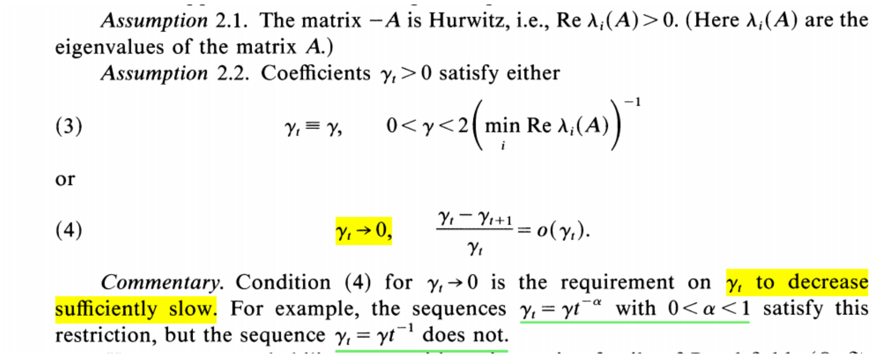


To obtain the sequence of estimates $(\bar{x}_t)_{t \geqq 1}$ of the solution x^* of (1), the following recursive algorithm will be used:

$$\begin{gathered} x_t = x_{t-1} - \gamma_t y_t, \qquad y_t = Ax_{t-1} - b + \xi_t, \\ \bar{x}_t = \frac{1}{t} \sum_{i=0}^{t-1} x_i. \end{gathered} \tag{2}$$

x_0 is an arbitrary (nonrandom) point in R^N.

Assumptions



Assumption 2.1. The matrix $-A$ is Hurwitz, i.e., $\operatorname{Re} \lambda_i(A) > 0$. (Here $\lambda_i(A)$ are the eigenvalues of the matrix A.)

Assumption 2.2. Coefficients $\gamma_t > 0$ satisfy either

$$\gamma_t \equiv \gamma, \qquad 0 < \gamma < 2\left(\min_i \operatorname{Re} \lambda_i(A)\right)^{-1} \tag{3}$$

or

$$\gamma_t \to 0, \qquad \frac{\gamma_t - \gamma_{t+1}}{\gamma_t} = o(\gamma_t). \tag{4}$$

Commentary. Condition (4) for $\gamma_t \to 0$ is the requirement on γ_t to decrease sufficiently slow. For example, the sequences $\gamma_t = \gamma t^{-\alpha}$ with $0 < \alpha < 1$ satisfy this restriction, but the sequence $\gamma_t = \gamma t^{-1}$ does not.

Why using these two assumptions?

Assumptions

Try to understand A2.1- 2.2 as follows.
Iterate algorithm:

$$\begin{aligned} x_t &= x_{t-1} - \gamma_t(Ax_{t-1} - b + \xi_t) = (I - \gamma_t A)x_{t-1} + \gamma_t(b - \xi_t) \\ &= \cdots = \prod_{i=1}^{t}(I - \gamma_i A)x_0 + \sum_{i=1}^{t}\prod_{j=i+1}^{t}(I - \gamma_j A)\gamma_i(b - \xi_i) \end{aligned} \tag{1}$$

and

$$\begin{aligned} x_t - x^* &= x_{t-1} - x^* - \gamma_t[A(x_{t-1} - x^*) + Ax^* - b + \xi_t] \\ &= (I - \gamma_t A)(x_{t-1} - x^*) - \gamma_t\xi_t \\ &= \cdots = \prod_{i=1}^{t}(I - \gamma_i A)(x_0 - x^*) - \sum_{i=1}^{t}\prod_{j=i+1}^{t}(I - \gamma_j A)\gamma_i\xi_i \end{aligned} \tag{2}$$

Assumptions

Let $\Delta_t = x_t - x^*$, and $\bar{\Delta}_t = \bar{x}_t - x^*$, then

$$\begin{aligned}
\Delta_t &= \prod_{i=1}^{t}(I - \gamma_i A)\Delta_0 - \sum_{i=1}^{t}\prod_{j=i+1}^{t}(I - \gamma_j A)\gamma_i \xi_i \\
\bar{\Delta}_t &= \frac{1}{t}\sum_{k=0}^{t-1}\Delta_k \\
&= \frac{1}{t}\sum_{k=0}^{t-1}\prod_{i=1}^{k}(I - \gamma_i A)\Delta_0 - \frac{1}{t}\sum_{k=0}^{t-1}\sum_{i=1}^{k}\prod_{j=i+1}^{k}(I - \gamma_j A)\gamma_i \xi_i \\
&= \frac{1}{t}\sum_{k=0}^{t-1}\prod_{i=1}^{k}(I - \gamma_i A)\Delta_0 - \frac{1}{t}\sum_{i=0}^{t-1}\sum_{k=i}^{t-1}\prod_{j=i+1}^{k}(I - \gamma_j A)\gamma_i \xi_i
\end{aligned}$$

Set $\alpha_j^t = \gamma_j \sum_{k=j}^{t-1}\prod_{i=j+1}^{k}(I - \gamma_i A)$, we obtain $\bar{\Delta}_t = \frac{\alpha_0^t \Delta_0}{t\gamma_0} + \frac{1}{t}\sum_{i=0}^{t-1}\alpha_i^t \xi_i$.

Assumptions

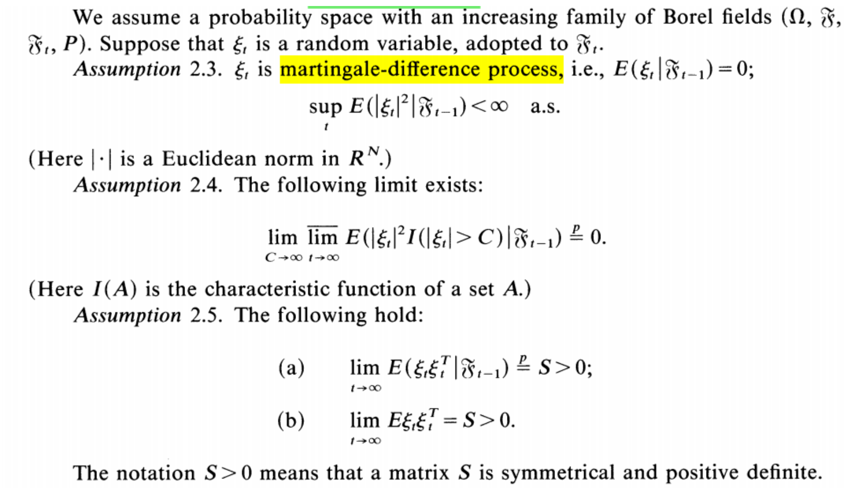
Δ_0 should has little impact on Δ_t and $\bar{\Delta}_t$. So, we need $\prod_{i=1}^{t}(I - \gamma_i A) \to 0$, as $t \to \infty$.

- When $\gamma_t = \gamma$, $\prod_{i=1}^{t}(I - \gamma_i A) = (I - \gamma A)^t \to 0$, then we need $|I - \gamma A| < 1$
  Assumption 2.2: $0 < \gamma A < 2$, then $|I - \gamma A| < 1$
- When $\gamma_t \to 0$, $\prod_{i=1}^{t}(I - \gamma_i A) \le K e^{-\alpha \sum_{i=1}^{t} \gamma_i} \to 0$ if $\sum_{i=1}^{t} \gamma_i$ divergences
  Assumption 2.2: $\gamma_t \sim t^{-\alpha}$, $0 < \alpha < 1$.
  Why α cannot be 1? ($\alpha < 1$ is a technical assumption)

$$\sum_{k=i}^{t-1} \prod_{j=i+1}^{k} (I - \gamma_j A)\gamma_i \xi_i = \gamma_j \sum_{k=i}^{t-1} \prod_{j=i+1}^{k} (I - \gamma_j A)\xi_i + \sum_{k=i}^{t-1} (\gamma_i - \gamma_j) \prod_{j=i+1}^{k} (I - \gamma_j A)\xi_i$$

$(\gamma_i - \gamma_j) = \sum_{k=j}^{i-1}(\gamma_{k+1} - \gamma_k) = \sum_{k=j}^{i-1} \gamma_k \cdot o(\gamma_k)$.
Then A2.3 guarantees the second term convergence (the proof is routine, omit here, see the proof of lemma 1 in the paper).

Assumptions



We assume a probability space with an increasing family of Borel fields $(\Omega, \mathfrak{F}, \mathfrak{F}_t, P)$. Suppose that ξ_t is a random variable, adopted to $\mathfrak{F}_t$.

Assumption 2.3. ξ_t is martingale-difference process, i.e., $E(\xi_t | \mathfrak{F}_{t-1}) = 0$;

$$\sup_t E(|\xi_t|^2 | \mathfrak{F}_{t-1}) < \infty \quad \text{a.s.}$$

(Here $|\cdot|$ is a Euclidean norm in R^N.)

Assumption 2.4. The following limit exists:

$$\lim_{C\to\infty} \overline{\lim_{t\to\infty}} E(|\xi_t|^2 I(|\xi_t| > C) | \mathfrak{F}_{t-1}) \overset{p}{=} 0.$$

(Here $I(A)$ is the characteristic function of a set A.)

Assumption 2.5. The following hold:

$$\text{(a)} \qquad \lim_{t\to\infty} E(\xi_t \xi_t^T | \mathfrak{F}_{t-1}) \overset{p}{=} S > 0;$$

$$\text{(b)} \qquad \lim_{t\to\infty} E\xi_t \xi_t^T = S > 0.$$

The notation $S > 0$ means that a matrix S is symmetrical and positive definite.

Assumptions

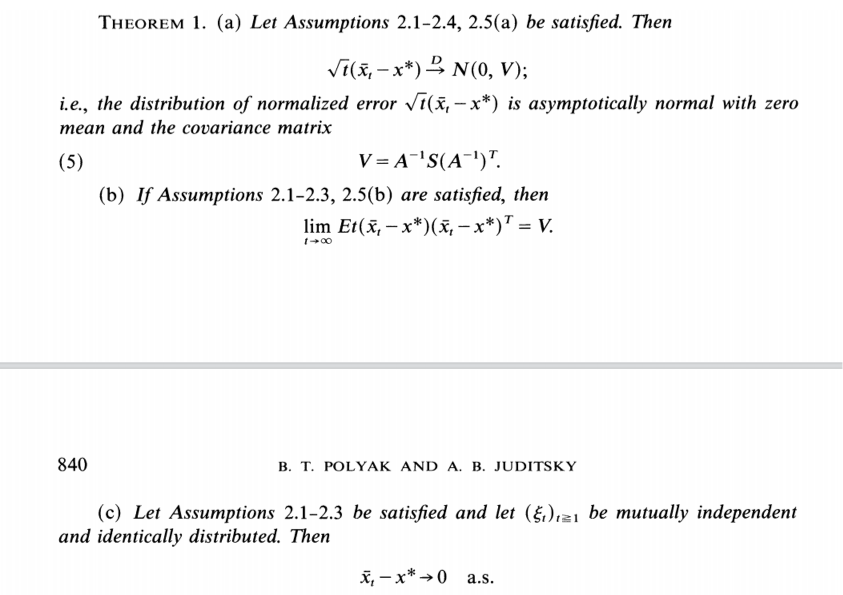
A2.3: Stochastic disturbance: random variable $\xi \rightarrow$ stochastic process ξ_t

- Time dimension, $x_n \rightarrow x_t$, and then averaging the trajectory
- Filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathrm{Pr})$ with a filtration $\mathcal{F}_t$
- $\xi_t \in m(\mathcal{F}_t)$ is a martingale-difference process

A2.4: Lindeberg condition → CLT

A2.5: There exists a covariance matrix.

Theorem 1



THEOREM 1. (a) *Let Assumptions* 2.1–2.4, 2.5(a) *be satisfied. Then*

$$\sqrt{t}(\bar{x}_t - x^*) \xrightarrow{D} N(0, V);$$

i.e., the distribution of normalized error $\sqrt{t}(\bar{x}_t - x^*)$ *is asymptotically normal with zero mean and the covariance matrix*

$$V = A^{-1}S(A^{-1})^T. \tag{5}$$

(b) *If Assumptions* 2.1–2.3, 2.5(b) *are satisfied, then*

$$\lim_{t\to\infty} Et(\bar{x}_t - x^*)(\bar{x}_t - x^*)^T = V.$$

840 B. T. POLYAK AND A. B. JUDITSKY

(c) *Let Assumptions* 2.1–2.3 *be satisfied and let* $(\xi_t)_{t\geqq 1}$ *be mutually independent and identically distributed. Then*

$$\bar{x}_t - x^* \to 0 \quad \text{a.s.}$$

Proof of (a)

Simplify without loss of generality: $N = 1$
To prove (a), we consider decomposing $\sqrt{t}\bar{\Delta}_t$ and obtain

$$\sqrt{t}\bar{\Delta}_t = \frac{\alpha_0^t \Delta_0}{\sqrt{t}\gamma_0} + \frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} \alpha_i^t \xi_i := I^{(1)} + I^{(2)} + I^{(3)}.$$

$I^{(1)} := \frac{\alpha_0^t \Delta_0}{\sqrt{t}\gamma_0} \to 0$. In fact, $\|\alpha_i^t\| \leq K$.
Idea: (we only show the case $\gamma_j = \gamma$)

$$\alpha_i^t = \gamma \sum_{k=i}^{t-1} \prod_{j=i+1}^{k} (I - \gamma A) = \gamma \sum_{k=i}^{t-1} (I - \gamma A)^{k-i-1}$$
$$= A^{-1} - (I - \gamma A)^{t-i-2} A^{-1}$$

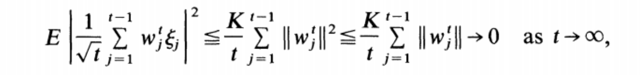
So, $\frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} \alpha_i^t \xi_i = \frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} A^{-1}\xi_i - \frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} (I - \gamma A)^{t-i-2} A^{-1} \xi_i := I^{(2)} + I^{(3)}$

Proof of (a)

We prove $I^{(3)} \to 0$.
In fact, because of $\lim_{t\to\infty}(I-\gamma A)^t = 0$, we obtain

$$\frac{1}{t}\sum_{i=0}^{t-1}(I-\gamma A)^{t-i-2}A^{-1} \to 0,$$

Denote $w_j^t := (I-\gamma A)^{t-j-2}A^{-1}$. Then $\frac{1}{t}\sum_{j=0}^{t-1}\|w_j^t\| \to 0$, and we also obtain $\|w_j^t\| \le K$.
Hence, we get that $|I^{(3)}| =$

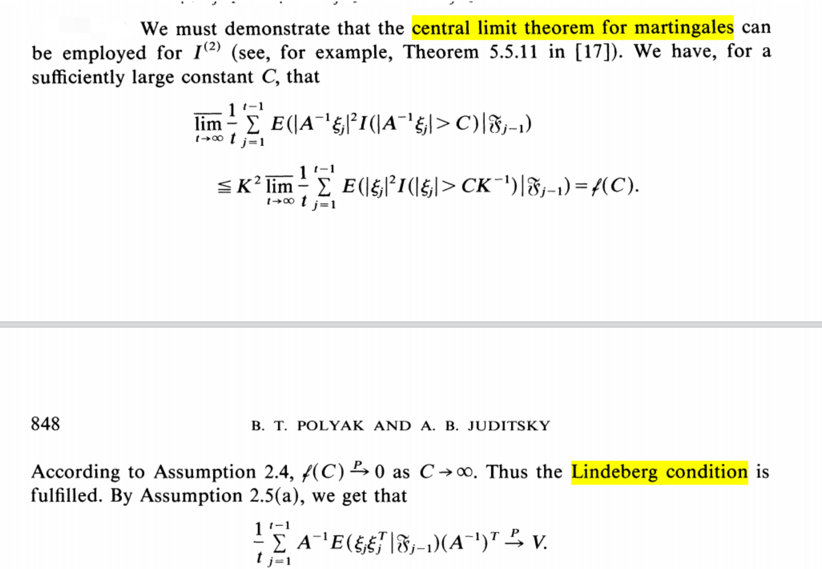


$$E\left|\frac{1}{\sqrt{t}}\sum_{j=1}^{t-1} w_j^t\xi_j\right|^2 \leqq \frac{K}{t}\sum_{j=1}^{t-1}\|w_j^t\|^2 \leqq \frac{K}{t}\sum_{j=1}^{t-1}\|w_j^t\| \to 0 \quad \text{as } t\to\infty,$$

so $|I^{(3)}| \to 0$.

Proof of (a)

We prove $I^{(2)}$ is asymptotic normality.



We must demonstrate that the central limit theorem for martingales can be employed for $I^{(2)}$ (see, for example, Theorem 5.5.11 in [17]). We have, for a sufficiently large constant C, that

$$\overline{\lim_{t\to\infty}} \frac{1}{t} \sum_{j=1}^{t-1} E(|A^{-1}\xi_j|^2 I(|A^{-1}\xi_j| > C) | \mathfrak{F}_{j-1})$$

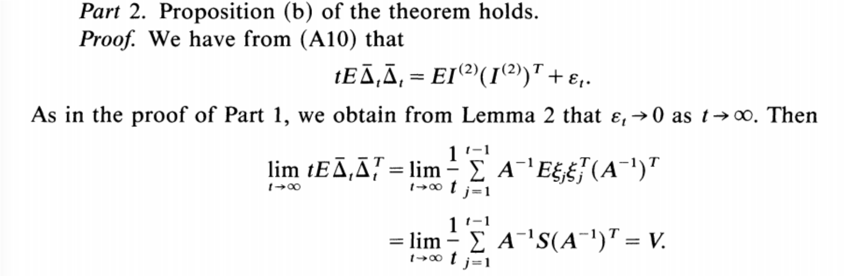
$$\leqq K^2 \overline{\lim_{t\to\infty}} \frac{1}{t} \sum_{j=1}^{t-1} E(|\xi_j|^2 I(|\xi_j| > CK^{-1}) | \mathfrak{F}_{j-1}) = f(C).$$

According to Assumption 2.4, $f(C) \xrightarrow{P} 0$ as $C \to \infty$. Thus the Lindeberg condition is fulfilled. By Assumption 2.5(a), we get that

$$\frac{1}{t} \sum_{j=1}^{t-1} A^{-1} E(\xi_j \xi_j^T | \mathfrak{F}_{j-1})(A^{-1})^T \xrightarrow{P} V.$$

Proof of (b)

$$\sqrt{t}\bar{\Delta}_t = I^{(1)} + I^{(2)} + I^{(3)}.$$
$$\mathrm{E}[t\bar{\Delta}_t\bar{\Delta}_t^\top] = \mathrm{E}[(I^{(1)} + I^{(2)} + I^{(3)})(I^{(1)} + I^{(2)} + I^{(3)})^\top]$$



Part 2. Proposition (b) of the theorem holds.

Proof. We have from (A10) that

$$tE\bar{\Delta}_t\bar{\Delta}_t = EI^{(2)}(I^{(2)})^T + \varepsilon_t.$$

As in the proof of Part 1, we obtain from Lemma 2 that $\varepsilon_t \to 0$ as $t \to \infty$. Then

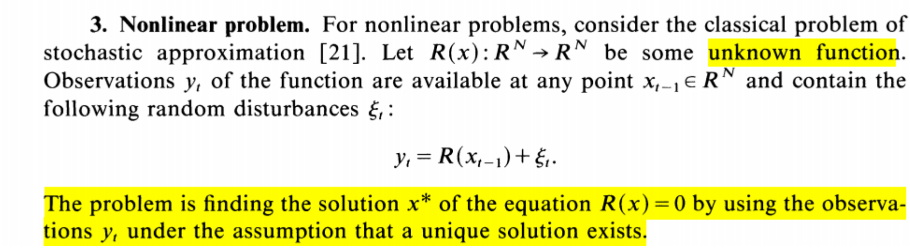
$$\lim_{t\to\infty} tE\bar{\Delta}_t\bar{\Delta}_t^T = \lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{t-1} A^{-1}E\xi_j\xi_j^T(A^{-1})^T$$
$$= \lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{t-1} A^{-1}S(A^{-1})^T = V.$$

Proof of (c)

Back to the decomposition $\bar{\Delta}_t := (I^{(1)} + I^{(2)} + I^{(3)})/\sqrt{t}$. (Consider i.i.d. ξ_i)

- $I^{(1)}/\sqrt{t} = \frac{\alpha_0^t \Delta_0}{t\gamma_0} \to 0$.
- By the law of large number, $I^{(2)}/\sqrt{t} = \frac{1}{t}\sum_{i=0}^{t-1} A^{-1}\xi_i \to 0$.
- Idea of the law of large number, $I^{(3)}/\sqrt{t} = \frac{1}{t}\sum_{i=0}^{t-1} w_i^t \xi_i \to 0$.

Nonlinear Problem

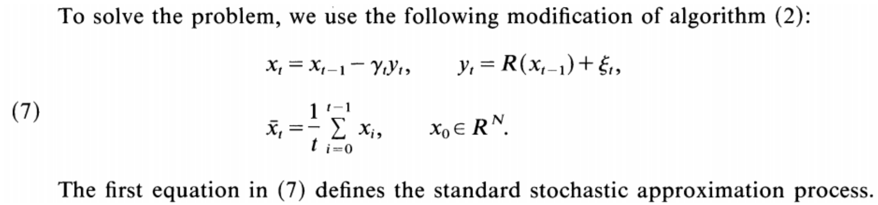


3. Nonlinear problem. For nonlinear problems, consider the classical problem of stochastic approximation [21]. Let $R(x): R^N \to R^N$ be some unknown function. Observations y_t of the function are available at any point $x_{t-1} \in R^N$ and contain the following random disturbances ξ_t:

$$y_t = R(x_{t-1}) + \xi_t.$$

The problem is finding the solution x^* of the equation $R(x) = 0$ by using the observations y_t under the assumption that a unique solution exists.

Averaging algorithm:



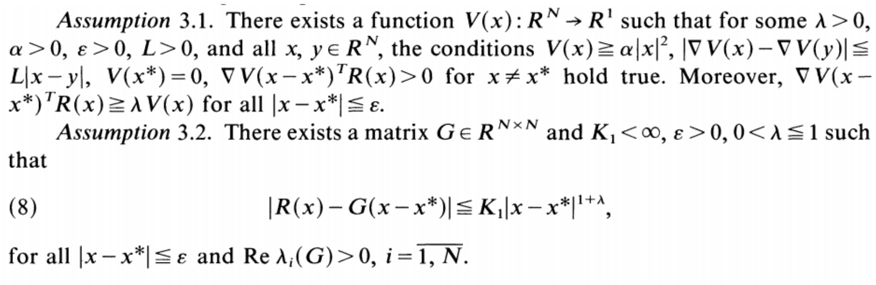
To solve the problem, we use the following modification of algorithm (2):

$$x_t = x_{t-1} - \gamma_t y_t, \qquad y_t = R(x_{t-1}) + \xi_t,$$

$$\bar{x}_t = \frac{1}{t} \sum_{i=0}^{t-1} x_i, \qquad x_0 \in R^N. \tag{7}$$

The first equation in (7) defines the standard stochastic approximation process.

Assumptions



Assumption 3.1. There exists a function $V(x): R^N \to R^1$ such that for some $\lambda > 0$, $\alpha > 0$, $\varepsilon > 0$, $L > 0$, and all $x, y \in R^N$, the conditions $V(x) \geqq \alpha |x|^2$, $|\nabla V(x) - \nabla V(y)| \leqq L|x-y|$, $V(x^*) = 0$, $\nabla V(x - x^*)^T R(x) > 0$ for $x \neq x^*$ hold true. Moreover, $\nabla V(x - x^*)^T R(x) \geqq \lambda V(x)$ for all $|x - x^*| \leqq \varepsilon$.

Assumption 3.2. There exists a matrix $G \in R^{N \times N}$ and $K_1 < \infty$, $\varepsilon > 0$, $0 < \lambda \leqq 1$ such that

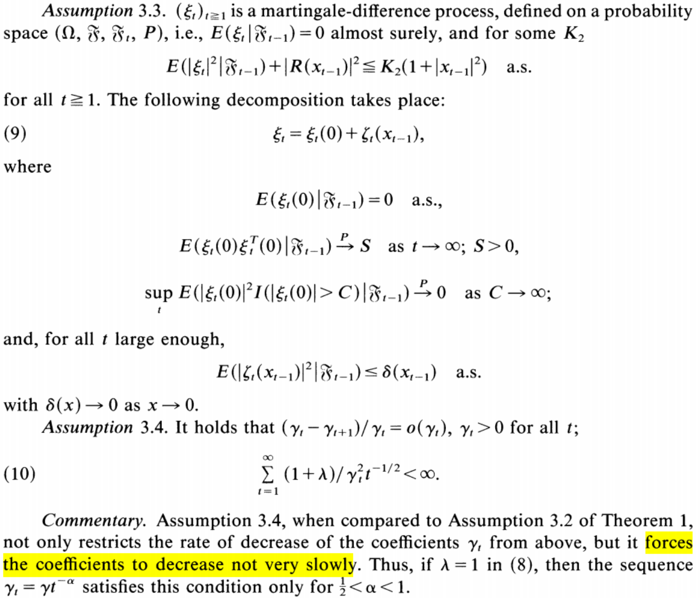
$$|R(x) - G(x - x^*)| \leqq K_1 |x - x^*|^{1+\lambda}, \tag{8}$$

for all $|x - x^*| \leqq \varepsilon$ and Re $\lambda_i(G) > 0$, $i = \overline{1, N}$.

A3.1: $V(x^*) = R(x^*) = 0$. For example, $V(x) = x^2$, $x^* = 0$, in this case $x \cdot R(x) > 0$ meaning that when $x > x^*$ then $R(x) > 0$, and when $x < x^*$ then $R(x) < 0$. Moreover, $x \cdot R(x) \geq \frac{\lambda}{2} x^2$ meaning that when $x \geq 0$ then $R(x) \geq \frac{\lambda}{2} x$, when $x < 0$ then $R(x) \leq \frac{\lambda}{2} x$.

A3.2: For example $Gx - K_1 |x|^{\lambda+1} \leq R(x) \leq Gx + K_1 |x|^{\lambda+1}$ locally around $x^* = 0$. $R(x)$ is close to Gx when λ or K_1 is small.

Assumptions



Assumption 3.3. $(\xi_t)_{t \geqq 1}$ is a martingale-difference process, defined on a probability space $(\Omega, \mathfrak{F}, \mathfrak{F}_t, P)$, i.e., $E(\xi_t | \mathfrak{F}_{t-1}) = 0$ almost surely, and for some K_2

$$E(|\xi_t|^2 | \mathfrak{F}_{t-1}) + |R(x_{t-1})|^2 \leqq K_2(1 + |x_{t-1}|^2) \quad \text{a.s.}$$

for all $t \geqq 1$. The following decomposition takes place:

$$\xi_t = \xi_t(0) + \zeta_t(x_{t-1}), \tag{9}$$

where

$$E(\xi_t(0) | \mathfrak{F}_{t-1}) = 0 \quad \text{a.s.},$$

$$E(\xi_t(0)\xi_t^T(0) | \mathfrak{F}_{t-1}) \xrightarrow{P} S \quad \text{as } t \to \infty; \; S > 0,$$

$$\sup_t E(|\xi_t(0)|^2 I(|\xi_t(0)| > C) | \mathfrak{F}_{t-1}) \xrightarrow{P} 0 \quad \text{as } C \to \infty;$$

and, for all t large enough,

$$E(|\zeta_t(x_{t-1})|^2 | \mathfrak{F}_{t-1}) \leq \delta(x_{t-1}) \quad \text{a.s.}$$

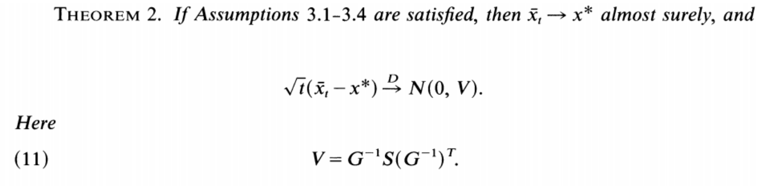
with $\delta(x) \to 0$ as $x \to 0$.

Assumption 3.4. It holds that $(\gamma_t - \gamma_{t+1})/\gamma_t = o(\gamma_t)$, $\gamma_t > 0$ for all t;

$$\sum_{t=1}^{\infty} (1+\lambda)/\gamma_t^2 t^{-1/2} < \infty. \tag{10}$$

Commentary. Assumption 3.4, when compared to Assumption 3.2 of Theorem 1, not only restricts the rate of decrease of the coefficients γ_t from above, but it forces the coefficients to decrease not very slowly. Thus, if $\lambda = 1$ in (8), then the sequence $\gamma_t = \gamma t^{-\alpha}$ satisfies this condition only for $\frac{1}{2} < \alpha < 1$.

Theorem 2



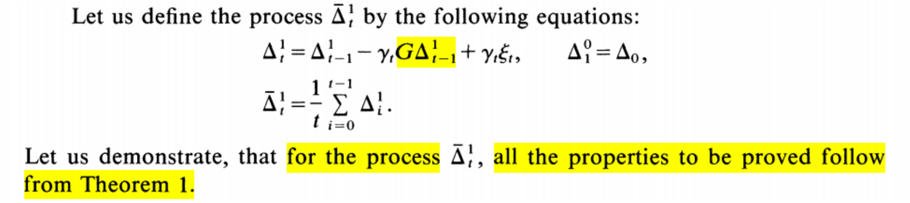
THEOREM 2. *If Assumptions* 3.1–3.4 *are satisfied, then* $\bar{x}_t \to x^*$ *almost surely, and*

$$\sqrt{t}(\bar{x}_t - x^*) \xrightarrow{D} N(0, V).$$

Here

(11) $$V = G^{-1}S(G^{-1})^T.$$

Sketch of the Proof



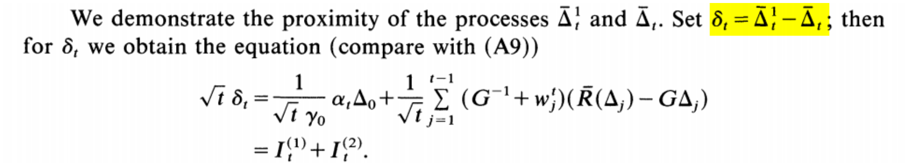
Let us define the process $\bar{\Delta}_t^1$ by the following equations:

$$\Delta_t^1 = \Delta_{t-1}^1 - \gamma_t G\Delta_{t-1}^1 + \gamma_t \xi_t, \qquad \Delta_1^0 = \Delta_0,$$

$$\bar{\Delta}_t^1 = \frac{1}{t}\sum_{i=0}^{t-1} \Delta_i^1 .$$

Let us demonstrate, that for the process $\bar{\Delta}_t^1$, all the properties to be proved follow from Theorem 1.

Denote $\bar{R}(x) := R(x - x^*)$.



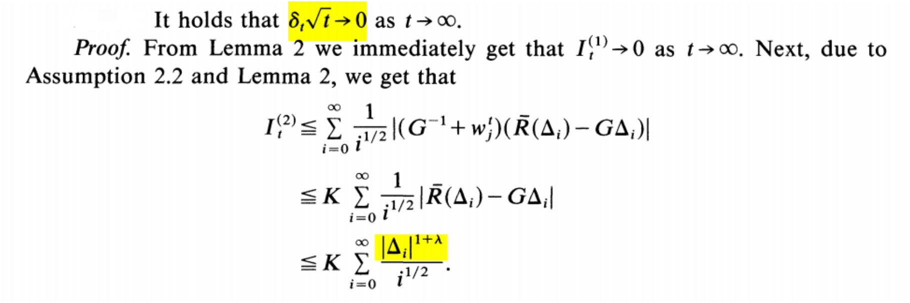
We demonstrate the proximity of the processes $\bar{\Delta}_t^1$ and $\bar{\Delta}_t$. Set $\delta_t = \bar{\Delta}_t^1 - \bar{\Delta}_t$; then for δ_t we obtain the equation (compare with (A9))

$$\sqrt{t}\,\delta_t = \frac{1}{\sqrt{t}\,\gamma_0}\alpha_t \Delta_0 + \frac{1}{\sqrt{t}}\sum_{j=1}^{t-1}(G^{-1} + w_j^t)(\bar{R}(\Delta_j) - G\Delta_j)$$
$$= I_t^{(1)} + I_t^{(2)}.$$

$\bar{\Delta}_t = \bar{\Delta}_t^1 - \delta_t$. To prove $\sqrt{t}\delta_t \to 0$.

Sketch of the Proof

It can be proved that $\sqrt{t}\delta_t \to 0$ as $t \to \infty$. So, the processes $\bar{\Delta}_t^1$ and $\bar{\Delta}_t$ are asymptotically equivalent. In fact,



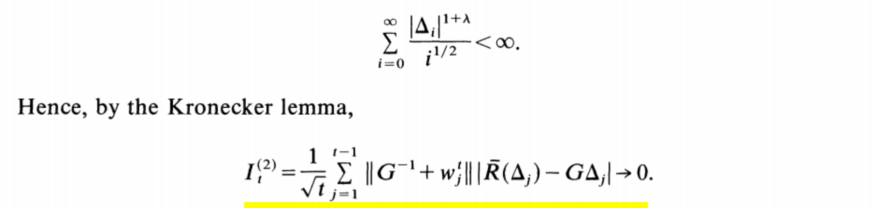
It holds that $\delta_t\sqrt{t}\to 0$ as $t\to\infty$.

Proof. From Lemma 2 we immediately get that $I_t^{(1)}\to 0$ as $t\to\infty$. Next, due to Assumption 2.2 and Lemma 2, we get that

$$\begin{aligned} I_t^{(2)} &\leqq \sum_{i=0}^{\infty} \frac{1}{i^{1/2}} |(G^{-1}+w_j^t)(\bar{R}(\Delta_i)-G\Delta_i)| \\ &\leqq K \sum_{i=0}^{\infty} \frac{1}{i^{1/2}} |\bar{R}(\Delta_i)-G\Delta_i| \\ &\leqq K \sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}}. \end{aligned}$$

Idea (not rigorous): $\Delta_i \sim i^{-1/2}$ then $\sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}} = \sum_{i=0}^{\infty} \frac{1}{i^{1+\lambda/2}} < \infty$.

Sketch of the Proof

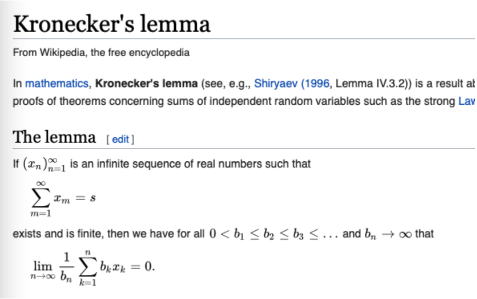


$$\sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}} < \infty.$$

Hence, by the Kronecker lemma,

$$I_t^{(2)} = \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} \|G^{-1} + w_j^t\| |\bar{R}(\Delta_j) - G\Delta_j| \to 0.$$

Remark:



Kronecker's lemma

From Wikipedia, the free encyclopedia

In mathematics, **Kronecker's lemma** (see, e.g., Shiryaev (1996, Lemma IV.3.2)) is a result ab proofs of theorems concerning sums of independent random variables such as the strong Lav

The lemma [edit]

If $(x_n)_{n=1}^{\infty}$ is an infinite sequence of real numbers such that

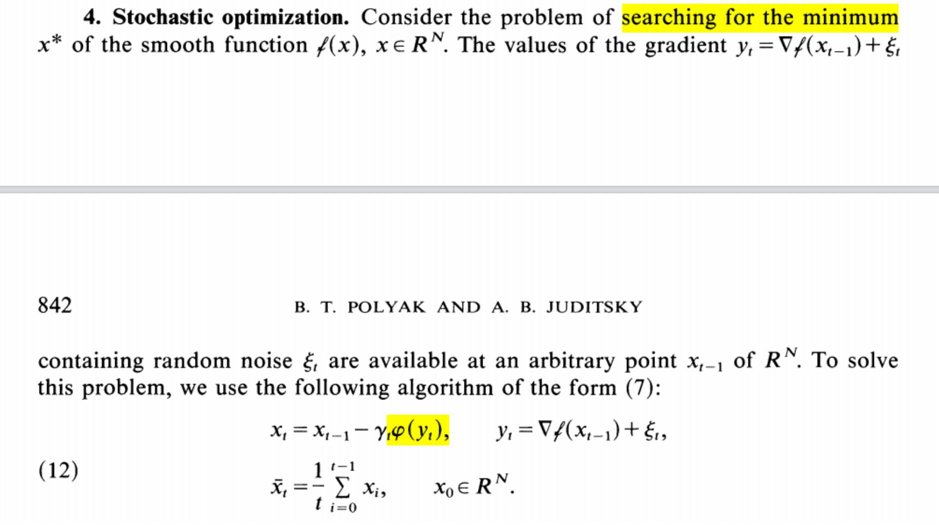
$$\sum_{m=1}^{\infty} x_m = s$$

exists and is finite, then we have for all $0 < b_1 \leq b_2 \leq b_3 \leq \ldots$ and $b_n \to \infty$ that

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^{n} b_k x_k = 0.$$

Stochastic Optimization Problem

An application of the nonlinear result:



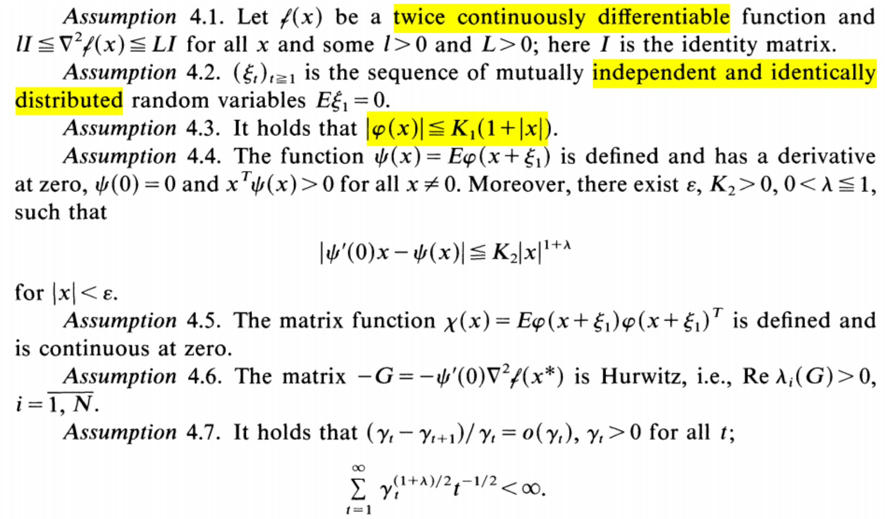
4. Stochastic optimization. Consider the problem of searching for the minimum x^* of the smooth function $f(x)$, $x \in R^N$. The values of the gradient $y_t = \nabla f(x_{t-1}) + \xi_t$

containing random noise ξ_t are available at an arbitrary point x_{t-1} of R^N. To solve this problem, we use the following algorithm of the form (7):

$$x_t = x_{t-1} - \gamma_t \varphi(y_t), \qquad y_t = \nabla f(x_{t-1}) + \xi_t,$$

$$\bar{x}_t = \frac{1}{t} \sum_{i=0}^{t-1} x_i, \qquad x_0 \in R^N. \tag{12}$$

Assumptions



Assumption 4.1. Let $f(x)$ be a twice continuously differentiable function and $lI \leqq \nabla^2 f(x) \leqq LI$ for all x and some $l > 0$ and $L > 0$; here I is the identity matrix.

Assumption 4.2. $(\xi_t)_{t \geqq 1}$ is the sequence of mutually independent and identically distributed random variables $E\xi_1 = 0$.

Assumption 4.3. It holds that $|\varphi(x)| \leqq K_1(1 + |x|)$.

Assumption 4.4. The function $\psi(x) = E\varphi(x + \xi_1)$ is defined and has a derivative at zero, $\psi(0) = 0$ and $x^T\psi(x) > 0$ for all $x \neq 0$. Moreover, there exist ε, $K_2 > 0$, $0 < \lambda \leqq 1$, such that

$$|\psi'(0)x - \psi(x)| \leqq K_2|x|^{1+\lambda}$$

for $|x| < \varepsilon$.

Assumption 4.5. The matrix function $\chi(x) = E\varphi(x + \xi_1)\varphi(x + \xi_1)^T$ is defined and is continuous at zero.

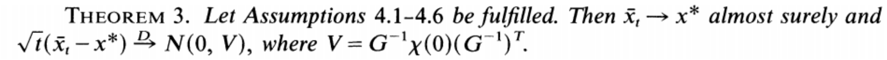
Assumption 4.6. The matrix $-G = -\psi'(0)\nabla^2 f(x^*)$ is Hurwitz, i.e., $\operatorname{Re} \lambda_i(G) > 0$, $i = \overline{1, N}$.

Assumption 4.7. It holds that $(\gamma_t - \gamma_{t+1})/\gamma_t = o(\gamma_t)$, $\gamma_t > 0$ for all t;

$$\sum_{t=1}^{\infty} \gamma_t^{(1+\lambda)/2} t^{-1/2} < \infty.$$

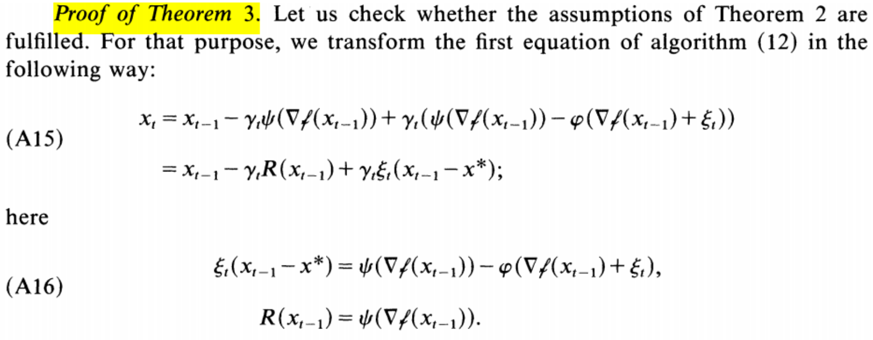
A4.1: f convex. A4.3: ϕ linear growth. A4.3+ A4.4: e.g. $\phi(x) = x$. A4.5: $\chi(0) =$ covariance matrix.

Theorem 3



THEOREM 3. *Let Assumptions* 4.1–4.6 *be fulfilled. Then* $\bar{x}_t \to x^*$ *almost surely and* $\sqrt{t}(\bar{x}_t - x^*) \xrightarrow{D} N(0, V)$, *where* $V = G^{-1}\chi(0)(G^{-1})^T$.

Sketch of the proof:



Proof of Theorem 3. Let us check whether the assumptions of Theorem 2 are fulfilled. For that purpose, we transform the first equation of algorithm (12) in the following way:

$$\begin{aligned} x_t &= x_{t-1} - \gamma_t \psi(\nabla f(x_{t-1})) + \gamma_t(\psi(\nabla f(x_{t-1})) - \varphi(\nabla f(x_{t-1}) + \xi_t)) \\ &= x_{t-1} - \gamma_t R(x_{t-1}) + \gamma_t \xi_t(x_{t-1} - x^*); \end{aligned} \tag{A15}$$

here

$$\begin{aligned} \xi_t(x_{t-1} - x^*) &= \psi(\nabla f(x_{t-1})) - \varphi(\nabla f(x_{t-1}) + \xi_t), \\ R(x_{t-1}) &= \psi(\nabla f(x_{t-1})). \end{aligned} \tag{A16}$$

Reference

B. T. Polyak and A. B. Juditsky (1992)
Acceleration of Stochastic Approximation by Averaging
SIAM Journal on Control and Optimization 30(4), 838- 855.

Marie Chau and Michael C. Fu (2005)
An Overview of Stochastic Approximation
Handbook of Simulation Optimization Chapter 6, 149- 178.

K. L. Chung (1954)
On a Stochastic Approximation Method
The Annals of Mathematical Statistics 25 (3): 463- 483.

V. Fabian (1968)
On Asymptotic Normality in Stochastic Approximation
The Annals of Mathematical Statistics 39 (4): 1327- 1332.