MTA **SZTAKI**
Hungarian Academy of Sciences
Institute for Computer Science and Control

Asymptotic Analysis of the LMS Algorithm with Momentum

László Gerencsér[1] Balázs Csanád Csáji[1] Sotirios Sabanis[2]

[1]Institute for Computer Science and Control (SZTAKI), Hungarian Academy of Sciences (MTA), Hungary

[2]School of Mathematics, University of Edinburgh, UK, and Alan Turing Institute, London, UK

57th IEEE CDC, Miami Beach, Florida, December 18, 2018

Introduction

- Stochastic gradient descent (SGD) methods are popular stochastic approximation (SA) algorithms applied in a wide variety of fields.
- Here, we focus on the special case of least mean square (LMS).
- Polyak's momentum is an acceleration technique for gradient methods which has several advantages for deterministic problems.
- K. Yuan, B. Ying and A. H. Sayed (2016) argued that in the stochastic case it is "equivalent" to standard SGD, assuming fixed gains, strongly convex functions and martingale difference noises.
- For LMS, they assumed independent noises to ensure this.
- Here, we provide a significantly simpler asymptotic analysis of LMS with momentum for stationary, ergodic and mixing signals.
- We present weak convergence results and explore the trade-off between the rate of convergence and the asymptotic covariance.

Stochastic Approximation with Fixed Gain

Stochastic Approximation (SA) with Fixed Gain

$$\underbrace{\theta_{n+1}}_{\substack{\text{next}\\ \text{estimate}}} = \underbrace{\theta_n}_{\substack{\text{current}\\ \text{estimate}}} + \underbrace{\mu}_{\substack{\text{fixed}\\ \text{gain}}} \underbrace{H(\theta_n, X_{n+1})}_{\substack{\text{update}\\ \text{operator}}}$$

- $\theta_n \in \mathbb{R}^d$ is the estimate at time n.
- $X_n \in \mathbb{R}^k$ is the new data available at time n.
- $\mu \in [0, \infty)$ is the fixed gain or step-size.
- $H : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$ is the update operator.

(SA algorithms are typically applied to find roots, fixed points or extrema of functions we only observe at given points with noise.)

Stochastic Gradient Descent

– We want to minimize an unknown function, $f : \mathbb{R}^d \to \mathbb{R}$, based only on noisy queries about its gradient, ∇f, at selected points.

Stochastic Gradient Descent (SGD)

$$\theta_{n+1} \doteq \theta_n + \mu\,(-\nabla_\theta f(\theta_n) + \varepsilon_n)$$

– Polyak's heavy-ball or momentum method is defined as

SGD with Momentum Acceleration

$$\theta_{n+1} \doteq \theta_n + \mu\,(-\nabla_\theta f(\theta_n) + \varepsilon_n) + \gamma\,(\theta_n - \theta_{n-1})$$

– The added term acts both as a smoother and an accelerator.
(The extra momentum dampens oscillations and helps us getting through narrow valleys, small humps and local minima.)

Mean-Square Optimal Linear Filter

- [C0] Assume we observe a (strictly) stationary and ergodic stochastic process consisting input-output pairs $\{(x_t, y_t)\}$, where regressor (input) x_t is $\mathbb{R}^d$-valued, while output y_t is $\mathbb{R}$-valued.
- We want to find the mean-square optimal linear filter coefficients

$$\theta^* \doteq \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\frac{1}{2}\left(y_n - x_n^{\mathrm{T}}\theta\right)^2\right]$$

- Using $R_* \doteq \mathbb{E}[x_n x_n^{\mathrm{T}}]$ and $b \doteq \mathbb{E}[x_n y_n]$, the optimal solution is

Wiener-Hopf Equation

$$R_*\,\theta^* = b \qquad \Longrightarrow \qquad \theta^* = R_*^{-1} b$$

- [C1] Assume that R_* is non-singular, thus, θ^* is uniquely defined.

Least Mean Square

– The least mean square (LMS) algorithm is an SGD method

Least Mean Square (LMS)

$$\theta_{n+1} \doteq \theta_n + \mu\, x_{n+1}\,(\, y_{n+1} - x_{n+1}^{\mathrm{T}}\theta_n\,)$$

with $\mu > 0$ and some constant (non-random) initial condition θ_0.

– Introducing the observation and (coefficient) estimation errors as

$$v_n \doteq y_n - x_n^{\mathrm{T}}\theta^* \qquad \text{and} \qquad \Delta_n \doteq \theta_n - \theta^*$$

the estimation error process, $\{\Delta_n\}$, follows the dynamics

$$\Delta_{n+1} = \Delta_n - \mu\, x_{n+1}\, x_{n+1}^{\mathrm{T}}\, \Delta_n + \mu\, x_{n+1}\, v_{n+1}$$

with $\Delta_0 \doteq \theta_0 - \theta^*$. Note that $\mathbb{E}\,[\,x_n v_n\,] = 0$ for all $n \geq 0$.

The Associated ODE

– A standard tool for the analysis of SA methods is the associated ordinary differential equation (ODE). In the LMS case (for $t \geq 0$)

$$\frac{d}{dt}\,\bar{\theta}_t \;=\; h(\bar{\theta}(t)) \;=\; b \,-\, R^*\bar{\theta}_t \qquad \text{with} \qquad \bar{\theta}_0 \;\doteq\; \theta_0$$

where $h(\theta) \doteq \mathbb{E}\,[\,x_{n+1}(y_{n+1} - x_{n+1}^{\mathrm{T}}\theta)\,]$ is the mean update for θ.

– A piecewise constant extension of $\{\theta_n\}$ is defined as $\theta_t^c \;\doteq\; \theta_{[t]}$, (note that here $[\,t\,]$ denotes the integer part of t).

– LMS is modified by taking a truncation domain D, where D is the interior of a compact set; then we apply the stopping time

$$\tau \;\doteq\; \inf\{\, t : \theta_t^c \notin D \,\}.$$

– [C2] We assume that the truncation domain is such that the solution of the ODE defined above does not leave D.

The Error of the ODE

– Let us define the following error processes for the mean ODE

$$\tilde{\theta}_n \doteq \theta_n - \bar{\theta}_n \qquad \text{and} \qquad \tilde{\theta}^c_t \doteq \theta^c_t - \bar{\theta}_t$$

– The normalized and time-scaled version of the ODE error is

$$V_t(\mu) \doteq \mu^{-1/2}\,\tilde{\theta}_{[(t\wedge\tau)/\mu]} = \mu^{-1/2}\,\tilde{\theta}^c_{(t\wedge\tau)/\mu}$$

– We will also need the asymptotic covariance matrices of the empirical means of the centered correction terms, given by

$$S(\theta) \doteq \sum_{k=-\infty}^{+\infty} \mathbb{E}\left[(H_k(\theta) - h(\theta))(H_0(\theta) - h(\theta))^{\mathrm{T}}\right]$$

where $H_n(\theta) \doteq x_n(y_n - x_n^{\mathrm{T}}\theta)$, which series converges, for example, under various mixing conditions (this will be ensured by [C3]).

Weak Convergence for LMS

– [C3] We assume that the process defined by

$$L_t(\mu) \doteq \sum_{n=0}^{[t/\mu]-1} \big(H_n(\bar{\theta}_{\mu n}) - h(\bar{\theta}_{\mu n})\big)\sqrt{\mu}$$

converges weakly, as $\mu \to 0$, to a time-inhomogeneous zero-mean Brownian motion $\{L_t\}$ with local covariances $\{S(\bar{\theta}_t)\}$.

Theorem 1: Weak Convergence for LMS

Under conditions C0, C1, C2 and C3, process $\{V_t(\mu)\}$ converges weakly, as $\mu \to 0$, to a process $\{Z_t\}$ satisfying the following linear stochastic differential equation (SDE), for $t \geq 0$, with $Z_0 = 0$,

$$dZ_t = -R^* Z_t\, dt + S^{1/2}(\bar{\theta}_t)\, dW_t$$

where $\{W_t\}$ is a standard Brownian motion in $\mathbb{R}^d$.

Momentum LMS

LMS with Momentum Acceleration

$$\theta_{n+1} \doteq \theta_n + \mu\, x_{n+1}\,(\, y_{n+1} - x_{n+1}^{\mathrm{T}}\theta_n\,) + \gamma\,(\,\theta_n - \theta_{n-1}\,)$$

with $\mu > 0$, $1 > \gamma > 0$, and some non-random $\theta_0 = \theta_{-1}$.

– The filter coefficient errors now follow a 2nd order dynamics

$$\Delta_{n+1} = \Delta_n - \mu\, x_{n+1}\, x_{n+1}^{\mathrm{T}}\, \Delta_n + \mu\, x_{n+1}\, v_{n+1} + \gamma\,(\Delta_n - \Delta_{n-1})$$

with $\Delta_0 = \Delta_{-1}$ (recall that $\Delta_n \doteq \theta_n - \theta^*$ and $v_n \doteq y_n - x_n^{\mathrm{T}}\theta^*$).

– To handle higher-order dynamics, we can use a state-vector,

$$U_n \doteq \begin{bmatrix} \Delta_n \\ \Delta_{n-1} \end{bmatrix}$$

State-Space Form for Momentum LMS

– Using $U_n \doteq [\,\Delta_n,\, \Delta_{n-1}\,]^{\mathrm{T}}$, the state-space dynamics becomes

$$U_{n+1} \;=\; U_n \;+\; A_{n+1} U_n \;+\; \mu\, W_{n+1},$$

$$A_{n+1} \doteq \begin{bmatrix} \gamma I - \mu \cdot x_{n+1} x_{n+1}^{\mathrm{T}} & -\gamma I \\ I & -I \end{bmatrix}, \qquad W_{n+1} \doteq \begin{bmatrix} x_{n+1} v_{n+1} \\ 0 \end{bmatrix}$$

– This, however, does not have the canonical form of SA methods.

– We apply a state-space transformation by Yuan, Ying and Sayed,

$$T \;\doteq\; T(\gamma) \;=\; \frac{1}{1-\gamma} \begin{bmatrix} I & -\gamma I \\ I & -I \end{bmatrix}$$

$$T^{-1} \;\doteq\; T^{-1}(\gamma) \;=\; \begin{bmatrix} I & -\gamma I \\ I & -I \end{bmatrix}$$

Transformed State-Space Dynamics

– To get a standard SA form, we also need to synchronize γ and μ,

$$\frac{\mu}{1-\gamma} \;=\; c\,(1-\gamma) \qquad \text{leading to} \qquad \mu \;=\; c\,(1-\gamma)^2.$$

with some fixed constant (hyper-parameter) $c > 0$.

– After applying T, the transformed dynamics becomes an (almost) canonical SA recursion with the fixed gain $\lambda \doteq 1-\gamma$ as follows:

$$\bar{U}_{n+1} \;=\; \bar{U}_n + \lambda\left(\left[\bar{B}_{n+1} + \lambda\,\bar{D}_{n+1}\right]\bar{U}_n + \bar{W}_{n+1}\right)$$

$$\bar{B}_n \;\doteq\; \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix} + c\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \otimes x_n x_n^{\mathrm{T}},$$

$$\bar{D}_n \;\doteq\; c\begin{bmatrix} 0 & -1 \\ 0 & -1 \end{bmatrix} \otimes x_n x_n^{\mathrm{T}}, \qquad \bar{W}_n \;\doteq\; c\begin{bmatrix} x_n v_n \\ x_n v_n \end{bmatrix}.$$

The Associated ODE for Momentum LMS

– Let us introduce the notations

$$\bar{H}_n(\bar{U}) \doteq (\bar{B}_n + \lambda \bar{D}_n)\bar{U} + \bar{W}_n$$

$$h(\bar{U}) \doteq \mathbb{E}[\bar{H}_n(\bar{U})] = \bar{B}_\lambda \bar{U}$$

$$\bar{B}_\lambda \doteq \mathbb{E}[\bar{B}_n + \lambda \bar{D}_n] = \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix} + c \begin{bmatrix} -1 & 1-\lambda \\ -1 & 1-\lambda \end{bmatrix} \otimes R^*$$

Then, the associated ODE takes the form, with $\bar{\bar{U}}_0 = \bar{U}_0$,

$$\frac{d}{dt}\bar{\bar{U}}_t = \bar{h}(\bar{\bar{U}}_t) = \bar{B}_\lambda \bar{\bar{U}}_t$$

– The solution for the limit when $\lambda \downarrow 0$ is denoted by $\bar{\bar{U}}^*_t$.

– Lemma: If λ is sufficiently small, then $\bar{B}_\lambda$ is stable.

The ODE Error for Momentum LMS

- [C2'] We again introduce a truncation domain, $\bar{D}$, as an interior of a compact set, and assume that the ODE does not leave $\bar{D}$.
- We set a stopping time for leaving the domain

$$\bar{\tau} \doteq \inf\{ n : \bar{U}_n \notin \bar{D} \}$$

- And define the error process, for $n \geq 0$, as

$$\tilde{\bar{U}}_n \doteq \bar{U}_n - \bar{\bar{U}}_n$$

- Finally, the normalized and time-scaled error process is

$$\bar{V}_t(\lambda) \doteq \lambda^{-1/2}\, \tilde{\bar{U}}_{[(t\wedge\bar{\tau})/\lambda]}$$

- However, the weak convergence theorems for SA methods cannot be directly applied, because there is an extra λ term in the update.

Approximation by Standard SA Recursion

– We will approximate the original process by (of course, $\bar{U}_0^* = \bar{U}_0$)

$$\bar{U}_{n+1}^* \;=\; \bar{U}_n^* \,+\, \lambda \left(\bar{B}_{n+1} \bar{U}_n^* \,+\, \bar{W}_{n+1} \right)$$

– Using the same steps as before, we can define the normalized and time-scaled ODE error process for the approximation as

$$\bar{V}_t^*(\lambda) \;\doteq\; \lambda^{-1/2}\, \tilde{\tilde{U}}^*_{[(t \wedge \bar{\tau}^*)/\lambda]}$$

where the truncation domain $\bar{D}^*$, for $\bar{\tau}^*$, is such that $\bar{D} \subseteq \mathrm{int}(\bar{D}^*)$.

– [CW] Assume $\bar{V}_t(\lambda) - \bar{V}_t^*(\lambda)$ converges weakly to 0, as $\lambda \to 0$ (for Momentum LMS, this could be proved based on linearity).

– Thus, weak convergence results can be applied to the approximate process, $\{\bar{V}_t^*(\lambda)\}$, and the results will carry over to $\{\bar{V}_t(\lambda)\}$.

Local Covariances for Momentum LMS

– The asymptotic covariance matrices of the empirical means of the centered correction terms are (under reasonable conditions)

$$\bar{S}(\bar{U}) \doteq \sum_{k=-\infty}^{+\infty} \mathbb{E}\big[(\bar{H}_k^*(\bar{U}) - \bar{h}^*(\bar{U}))(\bar{H}_0^*(\bar{U}) - \bar{h}^*(\bar{U}))^{\mathrm{T}} \big]$$

where H_k^* and h^* denote the limit of H_k and h as $\lambda \downarrow 0$.

– [C3'] We assume that the process defined by

$$\bar{L}_t(\lambda) \doteq \sum_{n=0}^{[t/\lambda]-1} \left(\bar{H}_n^*(\bar{\bar{U}}_{\lambda n}^*) - \bar{h}^*(\bar{\bar{U}}_{\lambda n}^*) \right) \sqrt{\lambda}$$

converges weakly, as $\lambda \to 0$, to a time-inhomogeneous zero-mean Brownian motion $\{\bar{L}_t\}$ with local covariance matrices $\{\bar{S}(\bar{\bar{U}}_t^*)\}$.

Weak Convergence for Momentum LMS

Theorem 2: Weak Convergence for Momentum LMS

Under conditions C0, C1, C2', C3' and CW, process $\{\bar{V}_t(\lambda)\}$ converges weakly, as $\lambda \to 0$, to a process $\{\bar{Z}_t\}$ satisfying the following linear stochastic differential equation (SDE),

$$d\bar{Z}_t \;=\; \bar{B}_*\,\bar{Z}_t\,dt \;+\; \bar{S}^{1/2}\,(\bar{\bar{U}}_t^*)\,d\bar{W}_t$$

for $t \geq 0$, with initial condition $\bar{Z}_0 = 0$, where $\{\bar{W}_t\}$ is a standard Brownian motion in $\mathbb{R}^{2d}$ and matrix $\bar{B}_*$ is defined as

$$\bar{B}_* \;\doteq\; \lim_{\lambda \downarrow 0} \bar{B}_\lambda \;=\; \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix} + c \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \otimes R^*$$

Lyapunov Equation for Momentum LMS

– The asymptotic covariance matrix of $\{\bar{Z}_t\}$, denoted by $\bar{P}$, satisfies the Lyapunov equation (it is a transformed process)

$$\bar{B}_* \bar{P} + \bar{P} \bar{B}_*^{\mathrm{T}} + \bar{S} = 0$$

– Lemma: the solution of this Lyapunov equation is

$$\bar{P} = \frac{c}{2} \begin{bmatrix} c\,S + 2P_0 & c\,S \\ c\,S & c\,S \end{bmatrix}$$

where P_0 is the asymptotic covariance of the weak limit of LMS.

– Let us denote the asymptotic covariance matrix of $\{T_1^+ \bar{Z}_t\}$ by P, where T_1^+ is the limit of $T^{-1}(\gamma)$ as $\gamma \to 1$ (or $\lambda \to 0$). Then,

$$P = T_1^+ \bar{P} (T_1^+)^{\mathrm{T}} = c \begin{bmatrix} P_0 & P_0 \\ P_0 & P_0 \end{bmatrix}$$

Comparing LMS with and without Momentum

Theorem 3: Asymptotic Covariance of Momentum LMS

Assume C0, C1, C2, C2', C3, C3', CW and that the weak convergences carry over to $\mathcal{N}(0, P_0)$ and $\mathcal{N}(0, P)$, as $t \to \infty$, in the case of plain and Momentum LMS methods, respectively.

Then, the covariance (sub)matrix of the asymptotic distribution associated with LMS with momentum is $c \cdot P_0$, where P_0 is the corresponding covariance of plain LMS and $c = \mu/(1-\gamma)^2$.

- If $c = 1$, then the two asymptotic covariances are the same.
- But, the convergence rates are quite different, as the normalization is $\mu^{-1/2}$ for LMS and $\lambda^{-1/2}$ for Momentum LMS with $\lambda = \sqrt{\mu}$.
- Decreasing c decreases the asymptotic covariance matrix, but it also decreases the convergence rate, and vice versa, $\lambda = \sqrt{\mu/c}$.

Summary

- We have analyzed the effect of momentum acceleration on the LMS algorithm, as a special case of SGD with fixed gain.
- Momentum acceleration has many known advantages in the deterministic case, but in a stochastic setting it is found to be "equivalent" to standard SGD by Yuan, Ying and Sayed (2016).
- However, for fixed-gain LMS, they only showed this equivalence for the (restrictive) special case of independent observations.
- Here, we provided a simpler asymptotic analysis of LMS with momentum acceleration for stationary, ergodic and mixing signals.
- We presented weak convergence results and explored the trade-off between the rate of convergence and the asymptotic covariance.
- The approach can be generalized to a wide range of SA methods.

Thank you for your attention!

✉ balazs.csaji@sztaki.mta.hu