Stochastic Optimization in Machine Learning: Inhomogeneity, Quantization and Acceleration

Balázs Csanád Csáji

SZTAKI: Institute for Computer Science and Control, Budapest, Hungary

Joint work with L. Monostori, E. Weyer, L. Gerencsér, and S. Sabanis

Data Analysis and Optimization Seminar, BME, January 21, 2021

Stochastic Approximation

Stochastic Approximation (SA)

$$\underbrace{\theta_{n+1}}_{\text{next estimate}} \doteq \underbrace{\theta_n}_{\text{current estimate}} + \underbrace{\gamma_n}_{\text{learning rate}} \underbrace{H(\theta_n, X_{n+1})}_{\text{update operator}}$$

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

(Note: $\{\theta_n\}$, $\{X_n\}$ are random vectors; $\{\gamma_n\}$ are random scalars.)

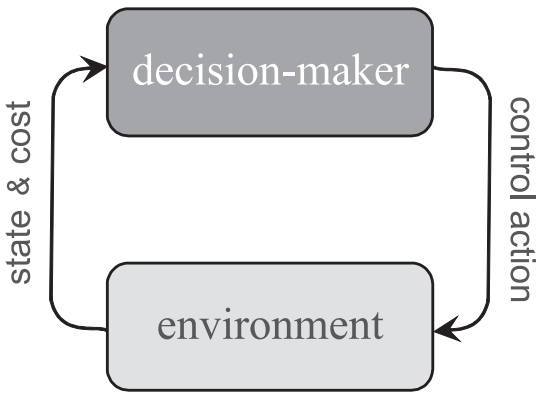
Part I: Inhomogeneity

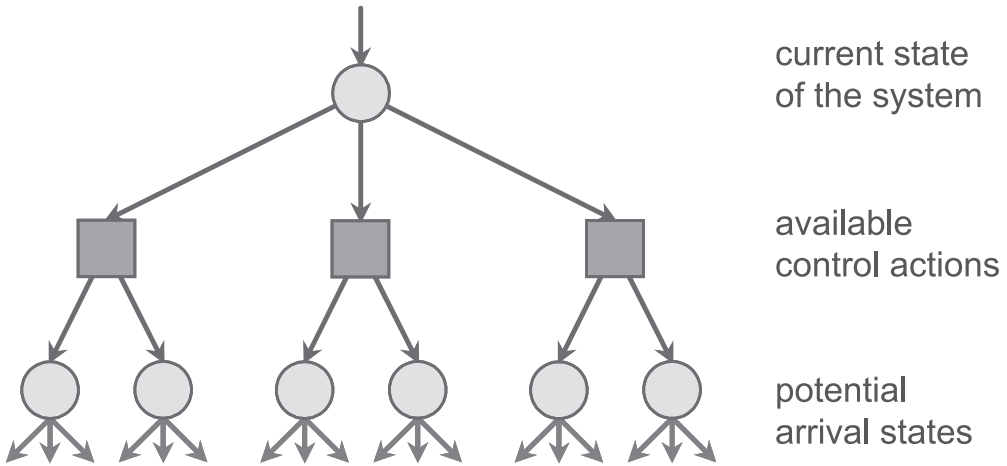
Reinforcement Learning in Time-Varying Environments

Joint work with: László Monostori (SZTAKI)

Reinforcement Learning

- Reinforcement learning (RL) is a machine learning approach to learn from interactions with an environment based on feedbacks (e.g., rewards).
- An interpretation: consider an agent acting in an uncertain environment and receiving information about the actual states and immediate costs.
- The aim is to learn an efficient behavior (control policy), such that applying this strategy minimizes the expected costs in the long run.





Applications of Reinforcement Learning

- Robot Control
- Dispatching & Scheduling
- Optimal Stopping
- Routing
- Maintenance and Repair
- Recommender Systems
- Inventory Control
- Optimal Control of Queues
- Strategic Asset Pricing
- Dynamic Options
- Insurance Risk Management
- Web System Configuration
- Bidding and Advertising
- Traffic Light Control
- Logic Games
- Communication Networks
- Dynamic Channel Allocation
- Power Grid Management
- Supply-Chain Management
- Fault Detection
- Sequential Clinical Trials
- PageRank Optimization

Markov Decision Processes

A (finite) Markov Decision Process (MDP) is characterized by

1. $\mathbb{X}$ is a (finite, non-empty) state space;
2. $\mathbb{A}$ is a (finite, non-empty) action space;
3. $\mathcal{A} : \mathbb{X} \to \mathcal{P}(\mathbb{A})$ is an action constraint function, namely, $\mathcal{A}(x)$ is the (non-empty) set of admissible actions in $x \in \mathbb{X}$;
4. $p : \mathbb{X} \times \mathbb{A} \to \Delta(\mathbb{X})$ is the transition probability function, $p_{xy}(a)$ denotes the probability of arriving at state $y \in \mathbb{X}$ after taking (admissible) action $a \in \mathcal{A}(x)$ in a state $x \in \mathbb{X}$;
5. $g : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$ is the immediate cost function, it is the cost (or reward) of taking action $a \in \mathcal{A}(x)$ in state $x \in \mathbb{X}$.

(Note that $\Delta(S)$ is the set of all probability distributions on S; and $\mathcal{P}(S)$ denotes the power set of set S: the set of all subsets of S.)

The Bellman Equation

- A (Markovian, randomized, stat.) control policy, $\pi : \mathbb{X} \to \Delta(\mathbb{A})$, is a function from states to probability distributions over actions.
- The value function, with discount factor α, of a policy π is

$$J^\pi(x) \doteq \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t g(X_t, A_t^\pi) \,\middle|\, X_0 = x\right],$$

  for all $x \in \mathbb{X}$, where $A_t^\pi \sim \pi(X_t)$, $X_{t+1} \sim p(X_t, A_t)$ and $\alpha \in (0, 1)$.
- There could be many optimal policies, but they share the same optimal value function J^*. We typically aim at estimating J^*.
- The fundamental Bellman equation is $TJ^* = J^*$, where

$$(TJ)(x) \doteq \min_{a \in \mathcal{A}(x)} \left[g(x, a) + \alpha \sum_{y \in \mathbb{X}} p(y \mid x, a) J(y) \right].$$

- Bellman operator T is a contraction with Lipschitz constant α.

Transition Probability and Cost Changes

Theorem 1: assume that two MDPs differ only in their transition-probability functions, and let these two functions be p_1 and p_2. Let the corresponding optimal value functions be J_1^* and J_2^*, then

$$\|J_1^* - J_2^*\|_\infty \leq \frac{\alpha |\mathbb{X}| \|g\|_\infty}{(1-\alpha)^2} \|p_1 - p_2\|_\infty$$

Theorem 2: assume that two MDPs differ only in the immediate-cost functions, and let these two functions be g_1 and g_2. Let the corresponding optimal value functions be J_1^* and J_2^*, then

$$\|J_1^* - J_2^*\|_\infty \leq \frac{1}{1-\alpha} \|g_1 - g_2\|_\infty$$

Improved Bound for Transition Probability Changes

Theorem 3: assume that two MDPs differ only in their transition-probability functions, and let these two functions be p_1 and p_2. Let the corresponding optimal value functions be J_1^* and J_2^*, then

$$\|J_1^* - J_2^*\|_\infty \leq \frac{\alpha \|g\|_\infty}{(1-\alpha)^2} \|p_1 - p_2\|_1,$$

where $\|\cdot\|_1$ is a norm on $f : \mathbb{X} \times \mathbb{A} \times \mathbb{X} \to \mathbb{R}$ type functions:

$$\|f\|_1 = \max_{x,a} \sum_{y \in \mathbb{X}} |f(x,a,y)|.$$

Note: since $\forall f : \|f\|_1 \leq n \|f\|_\infty$, where n is size of the state space, the bound of Theorem 3 is at least as good as that of Theorem 1.

Discount Factor Changes

Theorem 4: assume that two MDPs, $\mathcal{M}_1$ and $\mathcal{M}_2$, differ only in the discount factors, $\alpha_1, \alpha_2 \in (0,1)$. Let their corresponding optimal value functions be denoted by J_1^* and J_2^*, then

$$\|J_1^* - J_2^*\|_\infty \leq \frac{|\alpha_1 - \alpha_2|}{(1-\alpha_1)(1-\alpha_2)} \|g\|_\infty.$$

Moreover, there exists an MDP, denoted by $\mathcal{M}_3$, such that it differs only in the immediate-cost function from $\mathcal{M}_1$, thus its discount factor is α_1, and it has the same optimal value function as $\mathcal{M}_2$. The immediate-cost function of $\mathcal{M}_3$ is

$$\widehat{g}(x,a) = g(x,a) + (\alpha_2 - \alpha_1) \sum_{y \in \mathbb{X}} p(y \,|\, x,a) J_2^*(y),$$

where p is the transition function of all $\mathcal{M}_i$; g is the cost function of $\mathcal{M}_1$ and $\mathcal{M}_2$; and $J_2^*(y)$ is the optimal value function of $\mathcal{M}_2$.

Stochastic Optimization Perspective

- We denote the set of value functions by $\mathcal{V}$ which contains, in general, all bounded real-valued functions over an arbitrary set $\mathcal{X}$.
- Many (supervised and reinforcement) learning methods can be formulated as a stochastic optimization algorithm (SOA),

$$V_{t+1}(x) = (1-\gamma_t(x))V_t(x) + \gamma_t(x)\Big[(K_t V_t)(x) + W_t(x)\Big],$$

  where $V_t \in \mathcal{V}$, operator $K_t : \mathcal{V} \to \mathcal{V}$ acts on value functions, γ_t denotes the (random) stepsize and W_t is the noise at time t.
- We will consider the case when $\{K_t\}$ are pseudo-contractions, e.g., Q-learning, SARSA and TD-learning can be formulated this way.
- Note that in our formulation the update operator, K_t, is time-dependent. This will be needed to handle changing dynamics.

Main Assumptions

(A1) There exits a constant $C > 0$ such that for all x and t,

$$\mathbb{E}\big[W_t(x) \,|\, \mathcal{F}_t\big] = 0 \quad \text{and} \quad \mathbb{E}\left[W_t^2(x) \,|\, \mathcal{F}_t\right] < C < \infty,$$

where $\mathcal{F}_t = \sigma\{V_0, \ldots, V_t, W_0, \ldots, W_{t-1}, \gamma_0, \ldots, \gamma_t\}$.

(A2) For all x and t: $\gamma_t(x) \geq 0$ and we have with probability one

$$\sum_{t=0}^{\infty} \gamma_t(x) = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} \gamma_t^2(x) < \infty.$$

(A3) For all t, $K_t : \mathcal{V} \to \mathcal{V}$ is a supremum norm pseudo-contraction with Lipschitz constant $\beta_t < 1$ and with fixed point V_t^*:

$$\forall\, V \in \mathcal{V} : \|K_t V - V_t^*\|_\infty \leq \beta_t \|V - V_t^*\|_\infty.$$

Let us introduce $\beta_0 \doteq \limsup_{t\to\infty} \beta_t$, and we assume that $\beta_0 < 1$.

Approximate Convergence

Definition: a sequence of random elements $\{X_t\}$ from a normed space κ-approximates X with $\kappa > 0$ if for all $\varepsilon > 0$ there is a t_0:

$$\mathbb{P}\left(\sup_{t>t_0}(\|X_t - X\| \leq \kappa)\right) > 1 - \varepsilon$$

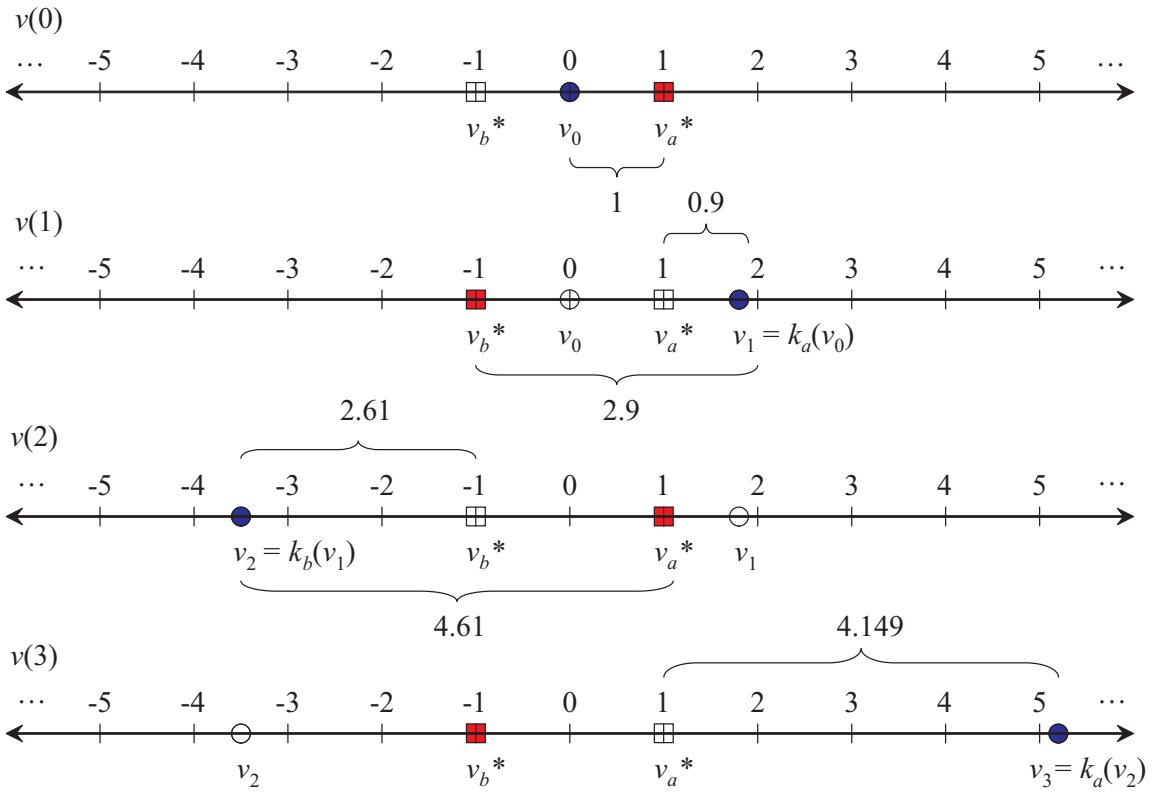
Theorem 5: suppose that Assumptions (A1), (A2) and (A3) hold and let $\{V_t\}$ be the sequence generated by a SOA. Then, for any $V_\star, V_0 \in \mathcal{V}$, the sequence V_t κ-approximates function $V_\star$ with

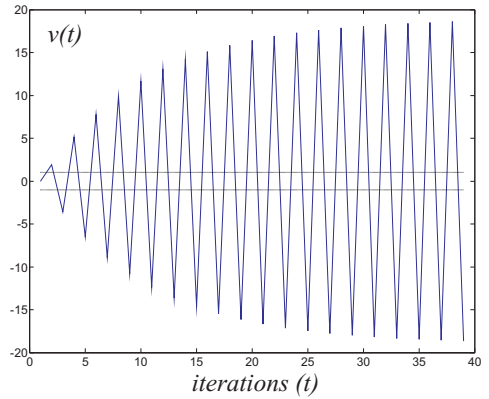
$$\kappa = \frac{4\varrho}{1-\beta_0} \qquad \text{where} \qquad \varrho = \limsup_{t\to\infty} \|V_t^* - V_\star\|_\infty$$

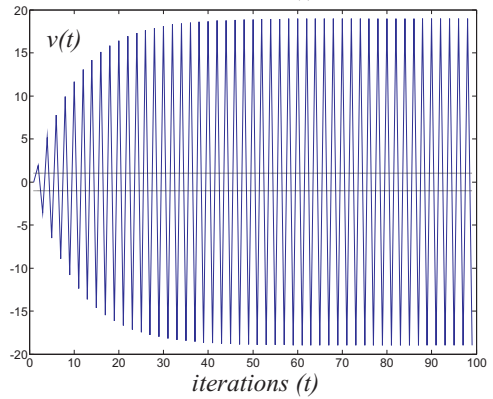
Notice that $V_\star$ can be an arbitrary function, but, naturally, the radius of the environment of $V_\star$, that the sequence $\{V_t\}$ almost surely converges to, depends on $\limsup_{t\to\infty} \|V_t^* - V_\star\|_\infty$.

A Deterministic Pathological Example

$$k_i(v) \doteq \begin{cases} v + (1 - b_i)(v_i^* - v) & \text{if } \operatorname{sign}(v_i^*) = \operatorname{sign}(v) \\ v_i^* + (v_i^* - v) + (1 - b_i)(v - v_i^*) & \text{otherwise} \end{cases}$$







Varying Environments: (ε, δ)-MDPs

– A class of non-stationary MDPs: in this model the transition-probabilities and the immediate-costs may change over time, as long as the accumulated changes remain asymptotically bounded.

Definition: a tuple $\langle \mathbb{X}, \mathbb{A}, \mathcal{A}, \{p_t\}_{t=1}^{\infty}, \{g_t\}_{t=1}^{\infty}, \alpha \rangle$, which represents a sequence of MDPs, is called an (ε, δ)-MDP where $\varepsilon, \delta > 0$, if there exists a base MDP, $\mathcal{M} = \langle \mathbb{X}, \mathbb{A}, \mathcal{A}, p, g, \alpha \rangle$, such that

$$\limsup_{t\to\infty} \|p - p_t\|_p \leq \varepsilon \qquad \text{and} \qquad \limsup_{t\to\infty} \|g - g_t\|_q \leq \delta,$$

where $1 \leq p, q \leq \infty$ (henceforth, we use $p = 1$ and $q = \infty$).

– The optimal value function of the base MDP, $\mathcal{M}$, and of the MDP at time t, $\mathcal{M}_t$, are denoted by J^* and J_t^*, respectively.

Relaxed Convergence in (ε, δ)-MDPs

Assume we have an (ε, δ)-MDP, then (using Theorems 2 and 3)

$$\limsup_{t \to \infty} \|J^* - J_t^*\|_\infty \leq d(\varepsilon, \delta)$$

$$d(\varepsilon, \delta) = \frac{\alpha\, \varepsilon\, (\|g\|_\infty + \delta)}{(1-\alpha)^2} + \frac{\delta}{1-\alpha}$$

where J_t^* and J^* are the optimal value functions of $\mathcal{M}_t$ and $\mathcal{M}$.

Corollary: consider an (ε, δ)-MDP and assume that (A1), (A2) and (A3) hold. Let $\{V_t\}$ be the sequence generated by a SOA. Assume the fixed point of each K_t is J_t^*. Then, V_t κ-approximates J^* with

$$\kappa = \frac{4\, d(\varepsilon, \delta)}{1 - \beta_0}$$

Q-learning in (ε, δ)-MDPs

- Q-learning is an arch-typical model-free and off-policy RL method.
- The one-step version of Watkins' Q-learning rule in (ε, δ)-MDPs is

$$Q_{t+1}(x,a) \doteq (1-\gamma_t(x,a))Q_t(x,a) + \gamma_t(x,a)(\widetilde{T}_t Q_t)(x,a),$$

$$(\widetilde{T}_t Q_t)(x,a) = g_t(x,a) + \alpha \min_{B \in \mathcal{A}(Y)} Q_t(Y,B),$$

  where Y is a random variable generated from (x,a) by simulation.
- The $\widetilde{T}_t$ operator can be rewritten in a form as follows

$$(\widetilde{T}_t Q)(x,a) = (\widetilde{K}_t Q)(x,a) + \widetilde{W}_t(x,a),$$

  where $\widetilde{W}_t(x,a)$ is a noise with zero mean and finite variance, and

$$(\widetilde{K}_t Q)(x,a) = g_t(x,a) + \alpha \sum_{y \in \mathbb{X}} p_t(y \mid x,a) \min_{b \in \mathcal{A}(y)} Q(y,b).$$

Q-learning in (ε, δ)-MDPs

- W_t has zero mean and finite variance $\Rightarrow$ (A1) is satisfied.
- Each $\widetilde{K}_t$ operator is an α contraction $\Rightarrow$ (A3) holds.
- Thus, (A2) $\Rightarrow \{Q_t\}$ generated by Q-learning κ-approximates Q^*, the optimal action-value function, with $\kappa = 4\,d(\varepsilon, \delta)/(1-\alpha)$.
- Similarly guarantees can be obtained for other RL methods, e.g., TD(λ) and asynchronous value iteration working in (ε, δ)-MDPs.

Lemma: assume we have two MDPs which differ only in the transition-probability functions or only in the immediate-cost functions or only in the discount factors. Let the corresponding optimal action-value functions be Q_1^* and Q_2^*, respectively. Then the bounds for $\|J_1^* - J_2^*\|_\infty$ of Theorems 2, 3 and 4 are also bounds for the optimal action-value function changes $\|Q_1^* - Q_2^*\|_\infty$.

Summary of Part I: Inhomogeneity

- The optimal (state and action) value functions of discounted MDPs Lipschitz continuously depend on the transition-probability and the immediate-cost functions. Changes in the discount factor can be traced back to changes in the immediate-costs.
- In (ε, δ)-MDPs these functions may vary over time, provided that the accumulated changes remain asymptotically bounded.
- A convergence theorem for stochastic optimization algorithms with time-dependent pseudo-contraction updates was given, which guarantees convergence to an environment of a target function.
- These results can be combined to deduce convergence theorems for reinforcement learning algorithms working in changing MDPs, which was demonstrated by studying Q-learning in (ε, δ)-MDPs.

Part II: Quantization

Recursive Estimation of ARX Systems Using Binary Sensors

Joint work with: Erik Weyer (University of Melbourne)

Binary Identification of ARX Systems

- Problem: estimating ARX systems observed via binary sensors.
- Previous (textbook) solutions typically assumed fully known noise characteristics and that the input signal can be chosen by the user.
- We try to reduce the assumptions on the noise and the input.
- Full knowledge of the noise distribution is not needed.
- The input is only assumed to be observed and not controlled.
- But, the threshold of the sensor must be controlled (which approach has similarities with dithering signal based solutions).
- Here, two recursive identification algorithms are proposed.
- Algorithm I: FIR approximation; which is strongly consistent.
- Algorithm II: simultaneous state and parameter estimation.

Problem Setting

– We observe an ARX (autoregressive exogenous) system via a binary sensor (where $\mathbb{I}$ denotes an indicator function):

$$X_t \doteq \sum_{i=1}^{p} a_i^* X_{t-i} + \sum_{i=1}^{q} b_i^* U_{t-i} + N_t,$$

$$Y_t \doteq \mathbb{I}(X_t \leq C_t),$$

where X_t — state, U_t — input, N_t — noise (at time t).

– The thresholds of the binary sensor, $\{C_t\}$, can be controlled.

– Data: the inputs $\{U_t\}$ and the binary outputs $\{Y_t\}$ are observed.

– Aim: to estimate (identify) $\theta^* = \left(a_1^*, \ldots, a_p^*, b_1^*, \ldots, b_q^*\right)^{\mathrm{T}} \in \mathbb{R}^{p+q}$

System Assumptions

- The noises $\{N_t\}$ are i.i.d., continuous, zero mean, zero median, $\mathbb{E}\left[N_t^2\right] < \infty$, and have a continuous and positive density at zero.
- The inputs $\{U_t\}$ are i.i.d., zero mean, and $0 < \mathbb{E}\left[U_t^2\right] < \infty$.
- The input $\{U_t\}$ and the noise $\{N_t\}$ sequences are independent.
- The system is stable, i.e., the roots of $A^*(z)$ lie strictly inside the unit circle; and the transfer function $B^*(z)/A^*(z)$ is irreducible,

$$A^*(z) \doteq 1 - a_1^* z^{-1} - a_2^* z^{-2} - \cdots - a_p^* z^{-p},$$

$$B^*(z) \doteq b_1^* z^{-1} + b_2^* z^{-2} + \cdots + b_q^* z^{-q},$$

  where z^{-1} is the backward shift operator (recall, $z^{-i} x_t \doteq x_{t-i}$).
- The orders (of polynomials A^* and B^*) p and q are known.

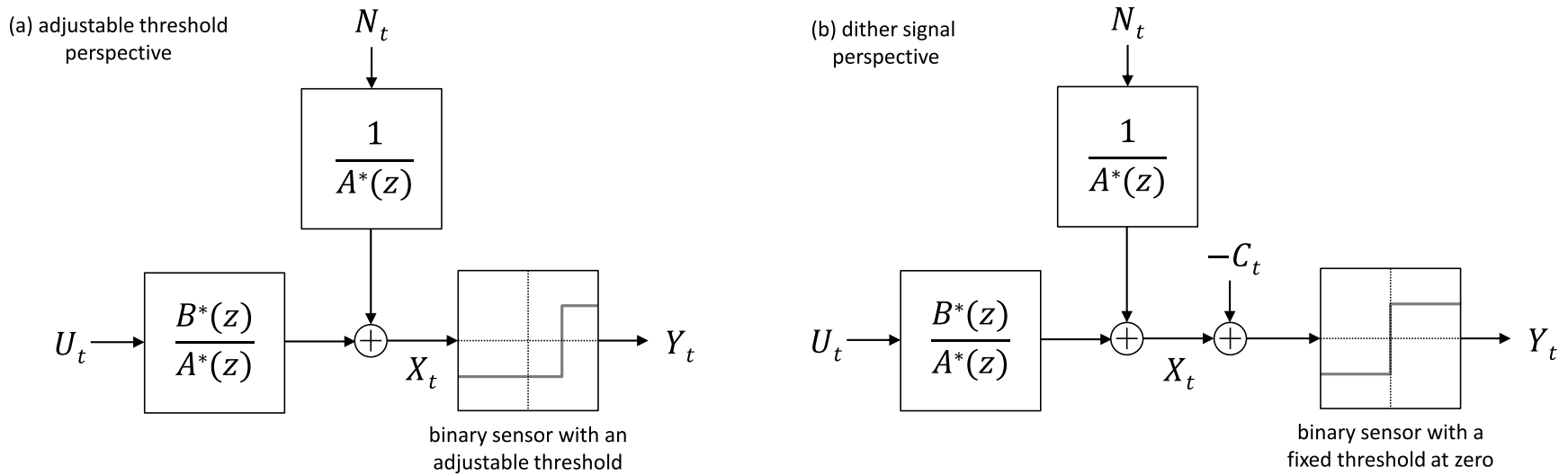
Adjustable Thresholds $\sim$ Dithering

– The binary output can be rewritten as

$$Y_t = \mathbb{I}(\varphi_t^{\mathrm{T}}\theta^* + N_t \le C_t) = \mathbb{I}(\varphi_t^{\mathrm{T}}\theta^* + N_t - C_t \le 0),$$

where $\varphi_t = (X_{t-1}, \dots, X_{t-p}, U_{t-1}, \dots, U_{t-q})$ is the regressor.

– Therefore, choosing the threshold is similar to dithering:



General Form of the Algorithms

– The general form of both proposed algorithms is

$$\hat{\theta}_{t+1} \doteq \Pi_{M_{\mu(t)}}\Big[\hat{\theta}_t + \alpha_t\,\widehat{\varphi}_t\Big(1 - 2\,\mathbb{I}(X_t \le \widehat{\varphi}_t^{\mathrm{T}}\hat{\theta}_t)\Big)\Big],$$

where $\widehat{\varphi}_t$ is a regressor defined differently in the two algorithms, $\{\alpha_t\}$ are the step-sizes and $\Pi_{M_{\mu(t)}}$ is a sequence of projections.

– As we assumed that N_t is continuous, we (almost surely) have

$$1 - 2\,\mathbb{I}(X_t \le \widehat{\varphi}_t^{\mathrm{T}}\hat{\theta}_t) = \mathrm{sign}(X_t - \widehat{\varphi}_t^{\mathrm{T}}\hat{\theta}_t).$$

– Therefore, the above algorithm will behave almost surely as

$$\hat{\theta}_{t+1} = \Pi_{M_{\mu(t)}}\Big[\hat{\theta}_t + \alpha_t\,\widehat{\varphi}_t\,\mathrm{sign}(X_t - \widehat{\varphi}_t^{\mathrm{T}}\hat{\theta}_t)\Big],$$

which is a sign-error method with expanding truncation bounds.

Step-Size Assumptions

– Typical step-size assumption of stochastic optimization algorithms

$$\sum_{t=0}^{\infty} \alpha_t = \infty,$$

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty,$$

$$\forall\, t : \alpha_t \geq 0.$$

– Henceforth, we will simply assume that for all t we use

$$\alpha_t \doteq \frac{1}{t+1}.$$

Expanding Truncation Bounds

- Let $\{M_t\}$ be a sequence of (strictly) monotone increasing positive real numbers with $M_t \to \infty$ as $t \to \infty$.
- Let $\mathbb{I}(\cdot)$ be the indicator function and define $\mu(t)$ and $\Delta\hat{\theta}_i$ as

$$\mu(t) \doteq \sum_{i=1}^{t-1} \mathbb{I}\big(|\hat{\theta}_i + \Delta\hat{\theta}_i| > M_{\mu(i)}\big),$$

$$\Delta\hat{\theta}_i \doteq \alpha_i \, \widehat{\varphi}_i \big(1 - 2\,\mathbb{I}(X_i \leq \widehat{\varphi}_i^{\mathrm{T}} \hat{\theta}_i)\big).$$

- Given a positive real M, projection Π_M is

$$\Pi_M(x) \doteq \begin{cases} x & \text{if } \|x\| \leq M, \\ 0 & \text{otherwise.} \end{cases}$$

Algorithm I: FIR Approximation

– Using impulse responses, $(c_i^*)_{i=1}^{\infty}$ and $(d_i^*)_{i=0}^{\infty}$, we have

$$X_t = \sum_{i=1}^{\infty} c_i^* U_{t-1} + \sum_{i=0}^{\infty} d_i^* N_{t-i},$$

– Let's approximate our ARX system with an FIR with order $p+q$

$$X_t = \bar{\varphi}_t^{\mathrm{T}} \bar{\theta}^* + W_t,$$

$$\bar{\varphi}_t \doteq (U_{t-1}, \dots, U_{t-p-q})^{\mathrm{T}}, \qquad \bar{\theta}^* \doteq (c_1^*, \dots, c_{p+q}^*)^{\mathrm{T}}.$$

– And W_t is simply the unmodelled part of the system

$$W_t \doteq \sum_{i=p+q+1}^{\infty} c_i^* U_{t-i} + \sum_{i=0}^{\infty} d_i^* N_{t-i}.$$

Algorithm I: FIR Approximation

- If we can estimate $\bar{\theta}^*$, we can also estimate the true θ^*.
- There is a function f, which we use for post processing, such that

$$\theta^* = f(\bar{\theta}^*),$$

- Algorithm I is defined by using $\widehat{\varphi}_t = \bar{\varphi}_t$ in the General Algorithm.

Theorem: Strong Consistency. Let $(\hat{\theta}_t)_{t=0}^{\infty}$ be the sequence generated by Algorithm I. Then, under the given assumptions, $f(\hat{\theta}_t)$ converges (a.s.) to θ^*, as $t \to \infty$, from any $\hat{\theta}_0 \in \mathbb{R}^{p+q}$.

- Moreover, one can show that $\sqrt{t}\,(\hat{\theta}_t - \bar{\theta}^*)$ is approximately normal.

Algorithm II: Simultaneous Estimation

- Main idea: to achieve a direct estimate of θ^* by simultaneously maintaining estimates for both $\widehat{X}_t$ and $\hat{\theta}_t$, at time t.
- The sequence of output estimates can be defined as

$$\widehat{X}_t \doteq \begin{cases} \sum_{i=1}^{p} \hat{a}_{t,i} \widehat{X}_{t-1} + \sum_{i=1}^{q} \hat{b}_{t,i} U_{t-i} & \text{if } t \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

  where $(\hat{a}_{t,i})_{i=1}^{p}$ and $(\hat{b}_{t,i})_{i=1}^{q}$ are estimates of the true parameters.
- Algorith II: is defined by setting the General Algorithm as

$$\begin{aligned} \widehat{\varphi}_t &\doteq (\widehat{X}_{t-1}, \ldots, \widehat{X}_{t-p}, U_{t-1}, \ldots, U_{t-q})^{\mathrm{T}}, \\ \hat{\theta}_t &\doteq (\hat{a}_{t,1}, \ldots, \hat{a}_{t,p}, \hat{b}_{t,1}, \ldots, \hat{b}_{t,q})^{\mathrm{T}}. \end{aligned}$$

Simulation Experiment: ARX(2, 2)

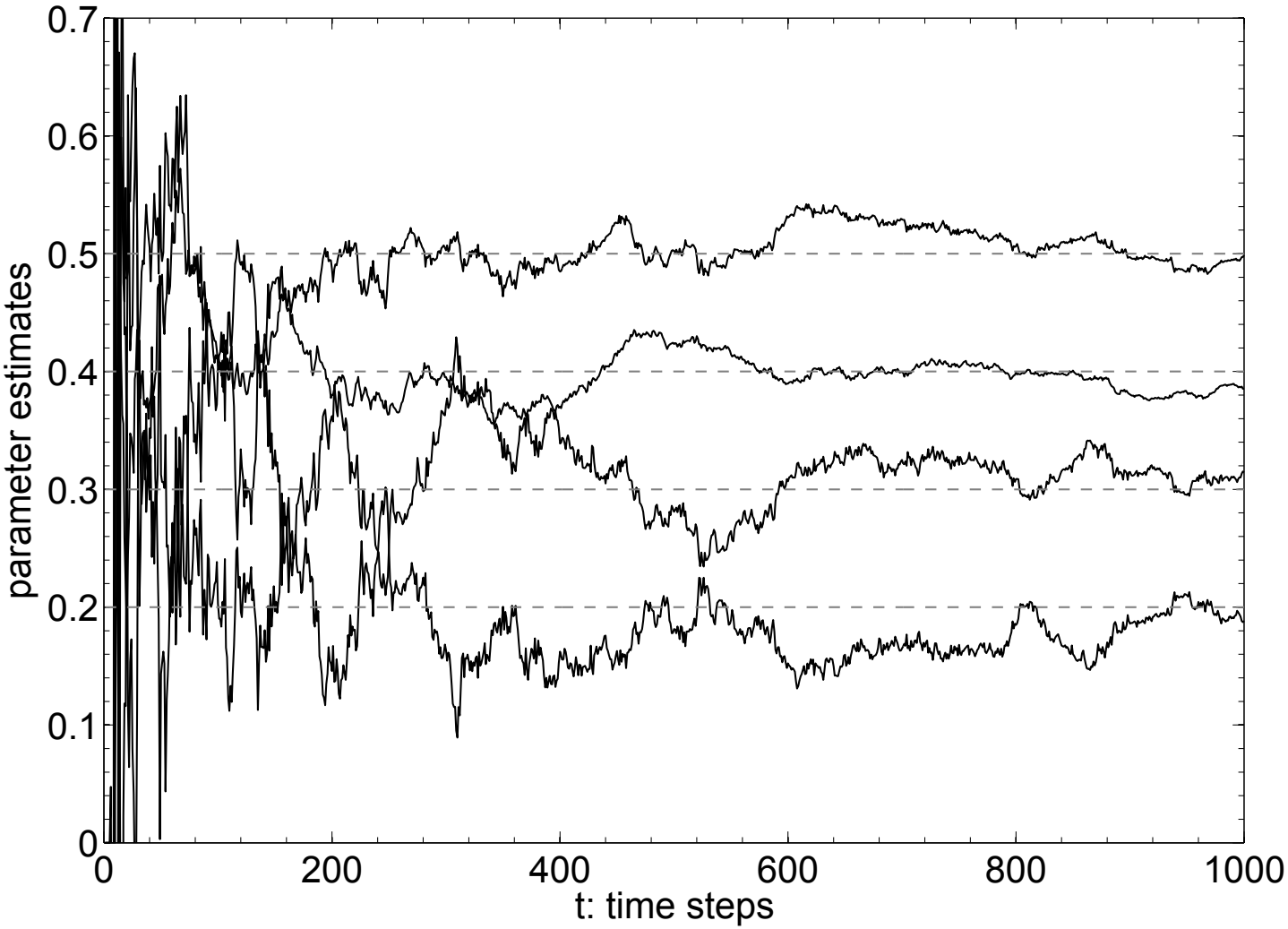


Figure: Recursive estimation with Algorithm I

Simulation Experiment: ARX(2, 2)

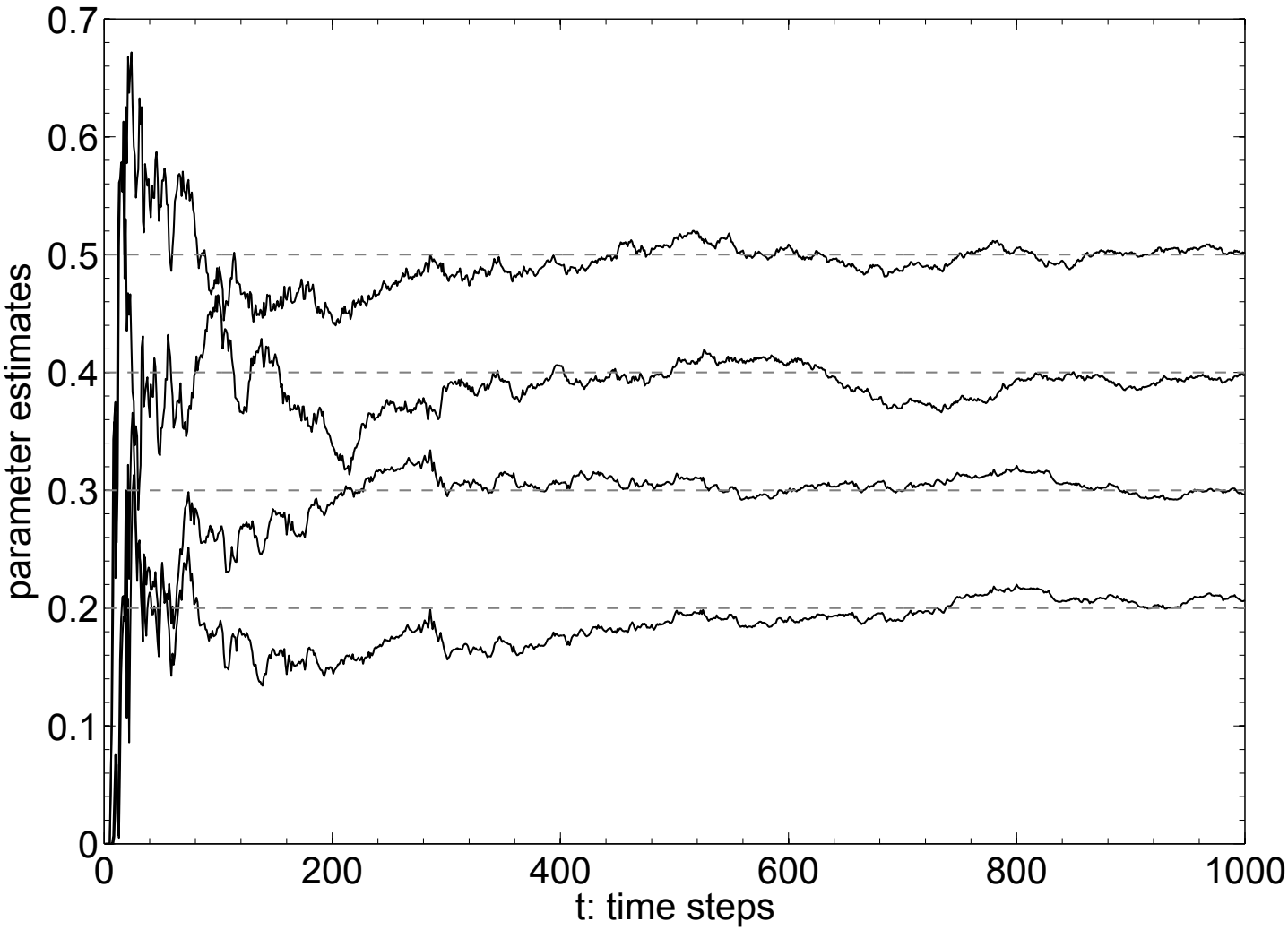


Figure: Recursive estimation with Algorithm II

Summary of Part II: Quantization

- Two recursive identification algorithms have been proposed for identifying ARX systems observed via a binary sensor.
- These algorithms neither assume the knowledge of the noise distributions, nor assume that the input signal can be chosen.
- However, we should be able to control the threshold of the sensor.
- This assumption is similar to allowing a dithering signal.
- Both algorithms are special cases of our General Algorithm that can be reformulated as a sign-error method (it is also equivalent to a stochastic gradient descent algorithm based on L1 error).
- Algorithm I: FIR approximation; which is strongly consistent.
- Algorithm II: simultaneous state and parameter estimation.
- Experimental results demonstrated that both algorithms efficiently approximated the parameters of an ARX(2,2) system.

Part III: Acceleration

Asymptotic Analysis of the LMS Algorithm with Momentum

Joint work with: László Gerencsér (SZTAKI) and Sotirios Sabanis (University of Edinburgh)

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 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)

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

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

SGD with Momentum Acceleration

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

– 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

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

- Using $R_* \doteq \mathbb{E}\left[x_n x_n^{\mathrm{T}}\right]$ and $b \doteq \mathbb{E}\left[x_n y_n\right]$, 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 $h(\theta) \doteq \mathbb{E}\left[x_{n+1}(y_{n+1} - x_{n+1}^{\mathrm{T}}\theta)\right]$ 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

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

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,

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

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

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

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

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)

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

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 of Part III: Acceleration

- 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!

www.sztaki.hu/~csaji ✉ csaji@sztaki.hu