Stochastic Approximation Algorithms with Set-Valued Maps

Shalabh Bhatnagar

Computer Science and Automation

Indian Institute of Science

Bangalore

November 10, 2019

Outline

1. Optimization under noise
2. Stochastic approximation algorithms
3. Random Directions Stochastic Approximation (RDSA)
4. Gradient algorithms with set-valued maps and their analysis
5. Ongoing and future work

Example: Parameter Optimization[1]

- Consider a repeated experiment that gives i.i.d input-output pairs $(X_n, Y_n), n \geq 0$, in real time
- **Goal:** Find a best parameterized fit

$$Y_n = f_w(X_n) + \epsilon_n,$$

i.e., one with the least $g(w) = \frac{1}{2}E[\| \epsilon_n \|^2]$

- f_w could correspond to polynomials, neural networks, splines, wavelets etc.
- Note $\nabla g(w) = -E[< Y_n - f_w(X_n), \nabla f_w(X_n) >]$

[1] V.S.Borkar, Stochastic Approximation: A Dynamical Systems Viewpoint, Cambridge Univ Press, 2008

Example: Parameter Optimization

- **Problem:** Cannot find the expectation
- **Solution:** Drop the expectation!
- Gradient Scheme with Noise:

$$w_{n+1} = w_n + a(n) < Y_n - f_{w_n}(X_n), \nabla f_{w_n}(X_n) >$$

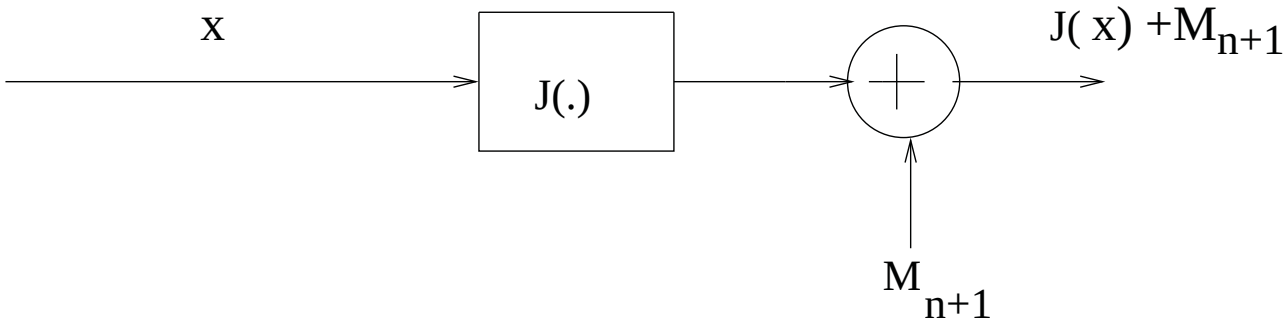
$$= w_n + a(n)(-\nabla g(w_n) + M_{n+1}),$$

  where M_{n+1} is the noise term

- Algorithms of this type are called *Stochastic Approximation Algorithms*

Stochastic Approximation[2]

- Objective: Solve the equation $J(x) = 0$ when analytical form of J is not known, however, 'noisy' measurements $J(x) + M_{n+1}$ can be obtained



- The Robbins-Monro Algorithm:

$$x_{n+1} = x_n + a(n)(J(x_n) + M_{n+1}) \tag{1}$$

[2] H.Robbins and S.Monro *Annals of Mathematical Statistics*, 22: 400–407, 1951

A Convergence Result[3]

(C1) $J : \mathcal{R}^N \rightarrow \mathcal{R}^N$ is Lipschitz continuous

(C2) $\sum_n a(n) = \infty, \ \sum_n a(n)^2 < \infty$

(C3) $M_{n+1}, n \geq 0$ is a martingale difference w.r.t. $\{\mathcal{F}_n \stackrel{\triangle}{=} \sigma(x_m, M_m, m \leq n)\}$. Further, for some $K > 0$,

$$E[\| M_{n+1} \|^2 | \mathcal{F}_n] \leq K(1 + \| x_n \|^2)$$

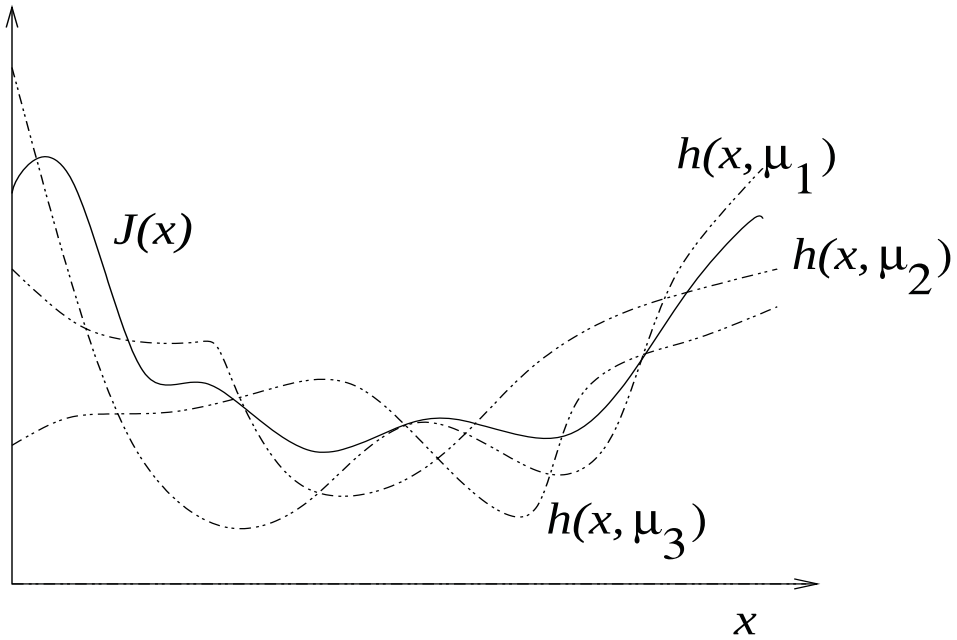
(C4) $\sup_n \| x_n \| < \infty$ almost surely

- Let x^* be the unique globally asymptotically stable attractor for the ODE $\dot{x}(t) = J(x(t))$. Then
- **Theorem** (Convergence of SAA): Under (C1)–(C4), $\{x_n\}$ converges almost surely to x^*.

[3] V.S.Borkar, *Stochastic Approximation Algorithms: A Dynamical Systems Viewpoint*, Cambridge Univ. Press, 2008

Optimization under Noise

- Let $J : \mathcal{R}^N \rightarrow \mathcal{R}$ be a given objective function having the form $J(x) = E_\mu[h(x, \mu)]$, where μ denotes 'noise' and $E_\mu[\cdot]$ is the expectation under that noise



- **Goal:** Find x^* s.t. $J(x^*) = \min_{x \in \mathcal{R}^N} J(x)$

Gradient Estimation using RDSA[4]

- Run two simulations with parameters

$$x + \delta d = \begin{pmatrix} x^1 + \delta d^1 \\ x^2 + \delta d^2 \\ \cdot \\ \cdot \\ x^N + \delta d^N \end{pmatrix}, \; x - \delta d = \begin{pmatrix} x^1 - \delta d^1 \\ x^2 - \delta d^2 \\ \cdot \\ \cdot \\ x^N - \delta d^N \end{pmatrix}$$

  where $d^1, \ldots, d^N$ are independent random variables with distribution $U[-\eta, \eta]$

- Gradient Estimator:

$$\hat{\nabla} J(x) = \frac{3}{\eta^2} d \frac{J(x + \delta d) - J(x - \delta d)}{2\delta}$$

[4] L.A.Prashanth, S.Bhatnagar, M.Fu and S.Marcus, *IEEE Transactions on Automatic Control*, 62(5):2223-2238, 2017

Hessian Estimator for RDSA

- Hessian Estimator:

$$\hat{\nabla}^2 J(x) = \frac{9}{2\eta^4} R \left(\frac{J(x+\delta d) + J(x-\delta d) - 2J(x)}{\delta^2} \right),$$

where

$$R = \begin{bmatrix} \frac{5}{2}(d^1)^2 - \eta^2/3 & \cdots & d^1 d^N \\ d^2 d^1 & \cdots & d^2 d^N \\ \cdots & \cdots & \cdots \\ d^N d^1 & \cdots & \frac{5}{2}(d^N)^2 - \eta^2/3 \end{bmatrix}.$$

Main Convergence Result

- RDSA Algorithm:

$$x_{n+1} = x_n - a(n)\Gamma(\hat{\nabla}^2 J(x_n))^{-1}\hat{\nabla} J(x_n)$$

  except that δ is replaced with $\delta_n \downarrow 0$ and

$$\sum_n a(n) = \infty; \ \sum_n \left(\frac{a(n)}{\delta_n}\right)^2 < \infty \tag{2}$$

- Let x^* be the unique globally asymptotically stable equilibrium of the ODE

$$\dot{x}(t) = -\Gamma(\nabla^2 J(x))^{-1}\nabla J(x)$$

- Let $a(n) = 1/n^\alpha$ and $\delta_n = 1/n^\gamma$ with $\alpha - \gamma > 0.5$ and $\beta \stackrel{\triangle}{=} \alpha - 2\gamma > 0$
- **Theorem:** Under (C1) on ∇J, (2), (C3) and (C4)
  1. $x_n \stackrel{a.s.}{\to} x^*$
  2. $n^{\beta/2}(x_n - x^*) \stackrel{dist}{\to} \mathcal{N}(\mu, \Omega)$

Sufficient Conditions for Stability of SA[5]

- (C5)
  - (i) $J_c(x) \stackrel{\triangle}{=} J(cx)/c$, $c \geq 1$ satisfies $J_c \to J_\infty$, for some $J_\infty : \mathcal{R}^N \to \mathcal{R}^N$ uniformly on compacts
  - (ii) The origin in $\mathcal{R}^N$ is a unique globally asymptotically stable equilibrium for the ODE $\dot{x}(t) = J_\infty(x(t))$
  - (iii) There is a unique globally asymptotically stable equilibrium $x^* \in \mathcal{R}^N$ for the ODE $\dot{x}(t) = J(x(t))$

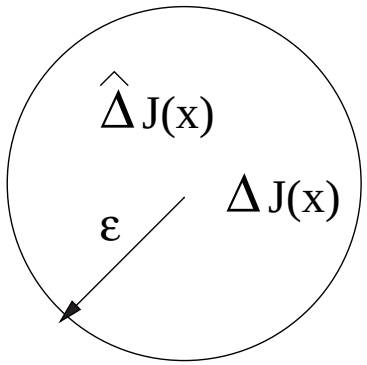
- The Stability Theorem: Under (C1)-(C3), (C5), for any initial condition $x_0 \in \mathcal{R}^N$, $\sup_n \| x_n \| < \infty$ a.s. Further, $x_n \stackrel{a.s.}{\to} x^*$.

[5] V. S. Borkar and S. P. Meyn, *SIAM Journal of Control and Optimization*, 38(2):447-469, 2000

Analysis of Algorithms with Biases[6][7]

- Let $\hat{\nabla} J(x)$ denote an estimator for $\nabla J(x)$ s.t.

$$\| \hat{\nabla} J(x) - \nabla J(x) \| \leq \epsilon(\delta) \rightarrow 0 \text{ as } \delta \rightarrow 0$$

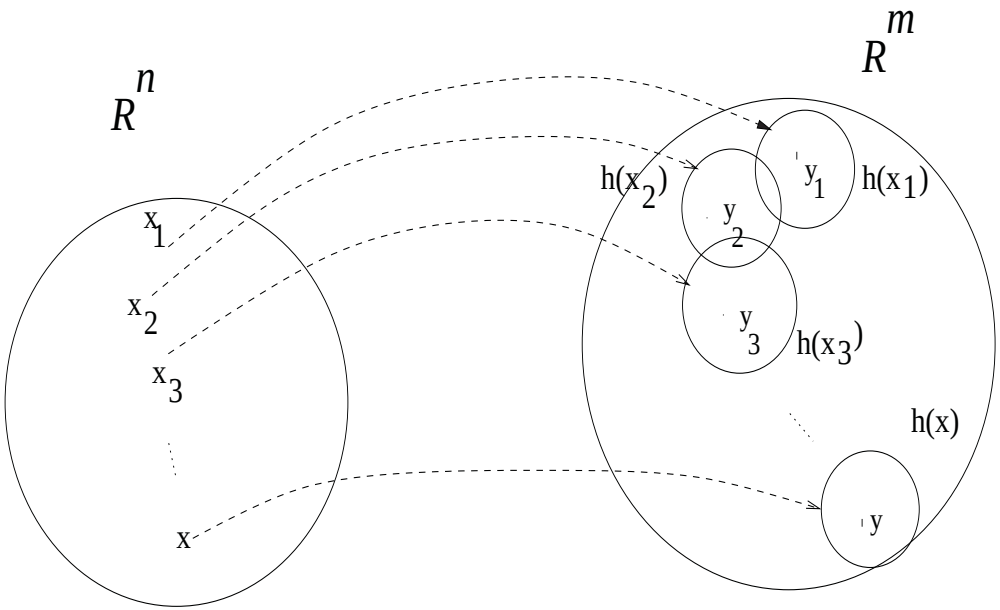


- Consider the recursion $x_{n+1} = x_n - a(n)(\nabla J(x_n) + \epsilon_n)$, where $\| \epsilon_n \| \leq \epsilon \; \forall n$

[6] A.Ramaswamy and S.Bhatnagar, *IEEE Transactions on Automatic Control*, 63(5):1465-1471, 2018

[7] A.Ramaswamy and S.Bhatnagar, *Mathematics of Operations Research*, 42(3):648-661, 2017

Marchaud Map



- A set-valued map h is called Marchaud if
  - $h(x)$ is convex and compact for each x
  - $\sup_{w \in h(x)} \| w \| \leq K(1 + \| x \|)$ for each x
  - h is upper-semicontinuous, i.e., given $\{x_n\} \subset \mathcal{R}^n$ and $\{y_n\} \subset \mathcal{R}^m$ with $x_n \to x$ and $y_n \to y$ with $y_n \in h(x_n), \forall n$, we have $y \in h(x)$

Differential Inclusion

- Consider the differential inclusion (DI) in $\mathcal{R}^d$:

$$\dot{x}(t) \in H(x(t)), \tag{3}$$

where $H : \mathcal{R}^d \to \{\text{subsets of } \mathcal{R}^d\}$ is Marchaud. Then the above DI has at least one solution **x** and each solution is absolutely continuous.[8]

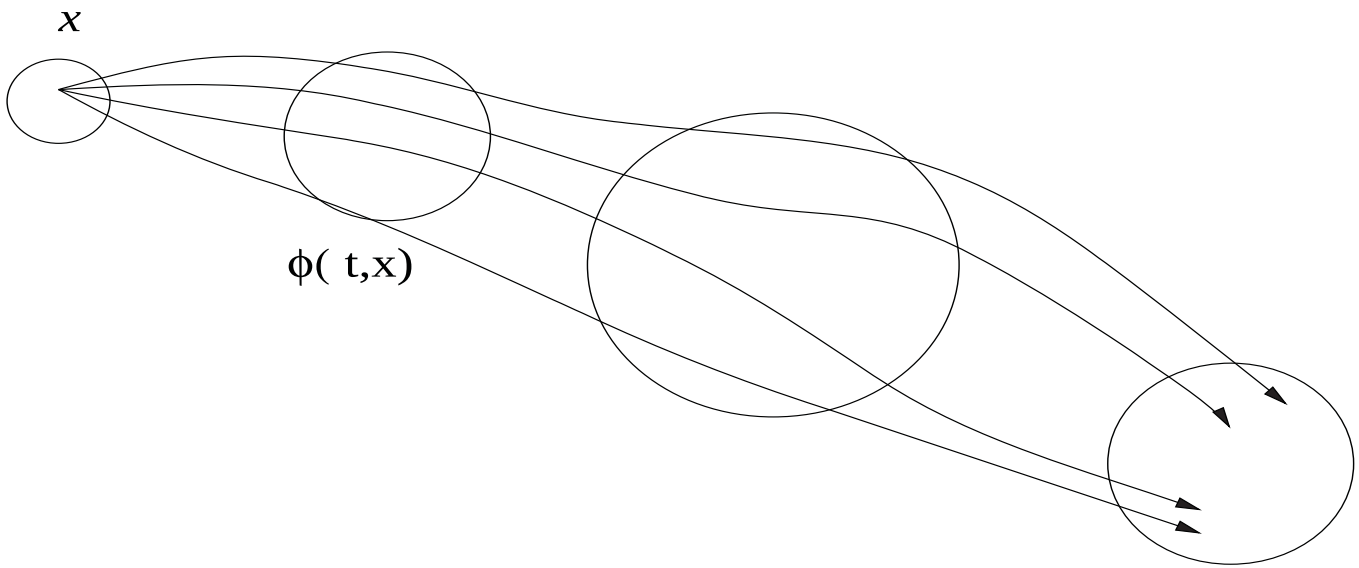
[8] J.Aubin and A.Cellina, Differential Inclusions: Set-Valued Maps and Viability Theory, Springer, 1984

Semiflow

- The Set-Valued Semiflow Φ associated with (3) is defined on $[0, \infty) \times \mathcal{R}^d$ as

$$\Phi(t, x) = \{x(t) \mid \mathbf{x} \in \Sigma, \mathbf{x}(0) = x\},$$

where Σ is the set of all absolutely continuous solutions to (3).

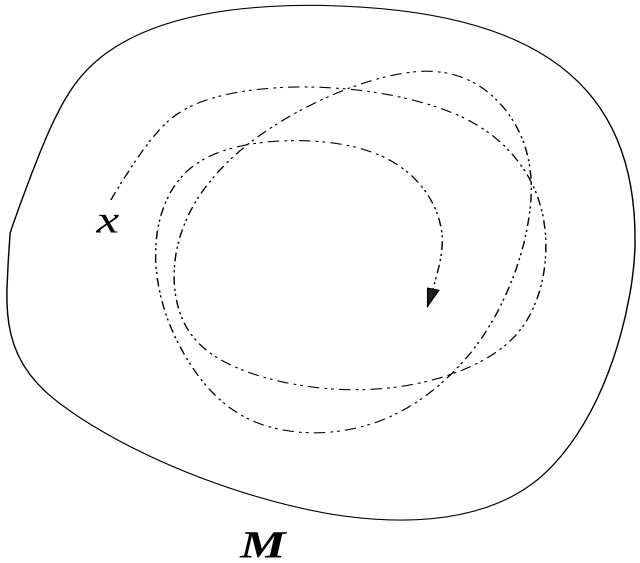


- For $B \times L \subset [0, \infty) \times \mathcal{R}^d$, let

$$\Phi(B, L) = \cup_{t \in B, x \in L} \Phi(t, x)$$

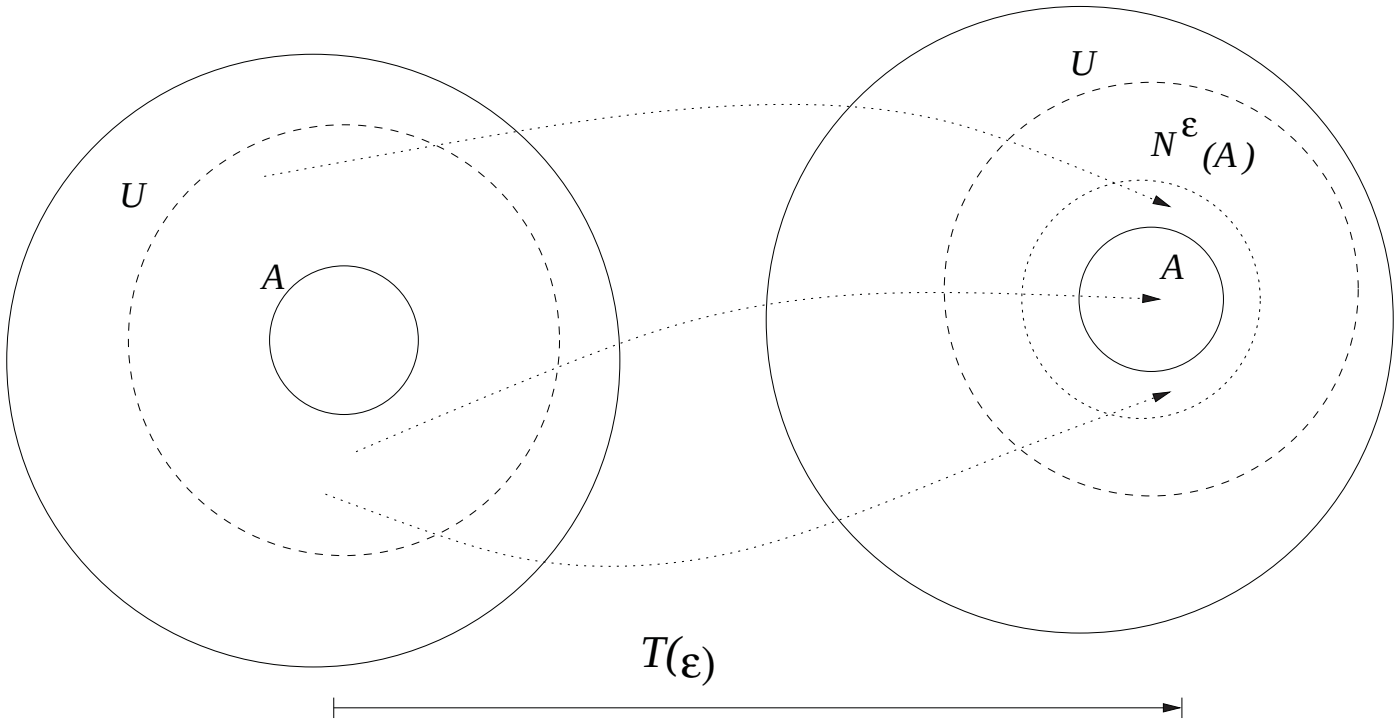
Invariant Set

- $M \subset \mathcal{R}^d$ is invariant if for every $x \in M$, there exists $\mathbf{x} \in \Sigma$ s.t. $x(t) \in M \; \forall t$ with $x(0) = x$



Attractor of a DI

- $A \subset \mathcal{R}^d$ is attracting if it is compact and there exists a neighborhood U such that for any $\epsilon > 0$, $\exists T(\epsilon) \geq 0$ with $\Phi([T(\epsilon), \infty), U) \subset N^{\epsilon}(A)$



- If the above A is invariant, it is called an attractor

An Alternative View of Algorithm

- Recall the recursion

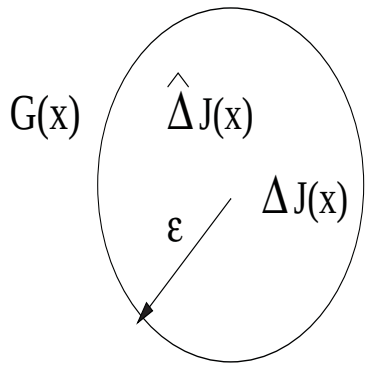
$$x_{n+1} = x_n - a(n)(\nabla J(x_n) + \epsilon_n),$$

where $\| \epsilon_n \| \leq \epsilon \; \forall n$

- Alternatively consider

$$x_{n+1} = x_n - a(n) g(x_n), \tag{4}$$

where $g(x_n) \in G(x_n) \forall n$ and $G(x) \stackrel{\triangle}{=} \nabla J(x) + \bar{B}_\epsilon(0)$, i.e., gradient estimate lies in an ϵ-ball around true gradient



Assumptions

- (A1) ∇J is a continuous function s.t. $\| \nabla J(x) \| \leq K(1+ \| x \|)$ for all $x \in \mathcal{R}^d$, $K > 0$
- (A2) $a(n) > 0 \forall n$ with

$$\sum_n a(n) = \infty, \ \sum_n a(n)^2 < \infty$$

- Can show that G is upper-semicontinuous
- Let $G_c(x) \stackrel{\triangle}{=} \{\frac{y}{c} \mid y \in G(cx)\}$
- Let $G_\infty(x) \stackrel{\triangle}{=} \bar{co}(Limsup_{c\to\infty} G_c(x))$, where $Limsup_{c\to\infty} G_c(x) \stackrel{\triangle}{=} \{y \mid \liminf_{c\to\infty} d(y, G_c(x)) = 0\}$

An Important Result

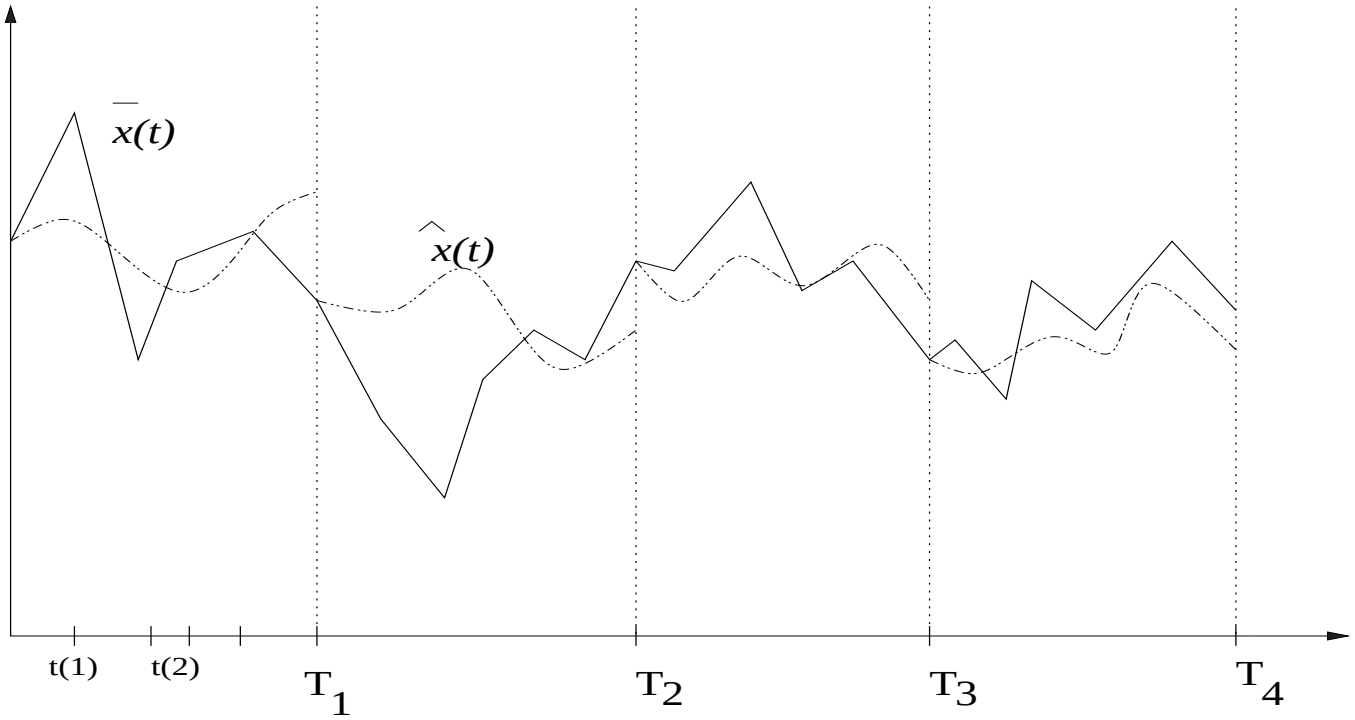
- Lemma 1 The map $x \mapsto G_\infty(x)$ is Marchaud
- Thus, $\dot{x}(t) \in -G_\infty(x(t))$ has at least one solution and which is absolutely continuous
- (A3) $\dot{x}(t) \in -G_\infty(x(t))$ has an attractor set A such that $A \subset B_a(0)$ for some $a > 0$ and $\bar{B}_a(0)$ is a fundamental neighborhood of A
- (A4) Let $c_n \geq 1$ be an increasing sequence of integers such that $c_n \uparrow \infty$ as $n \uparrow \infty$. Let $x_n \to x$ and $y_n \to y$ as $n \uparrow \infty$, such that $y_n \in G_{c_n}(x_n), \forall n$, then $y \in G_\infty(x)$

The Stability Result

- Theorem 1 Under (A1)-(A4), the iterates (4) are stable i.e., $\sup_n \| x_n \| < \infty$ a.s.
- Now recall that $G(x) = \nabla J(x) + \bar{B}_\epsilon(0)$
- Let the minimum set M of J be the global attractor of $\dot{x}(t) = -\nabla J(x(t))$
- It can be shown that any compact set $\mathcal{K}$ with $M \subset \mathcal{K} \subset \mathcal{R}^d$ is a fundamental neighborhood of M
- From Theorem 1, $\bar{x}(t) \in \mathcal{K}_0 \ \forall t \geq 0$ for some (possibly sample path dependent) compact set $\mathcal{K}_0$ which then is a fundamental neighborhood of M

The Main Result

- Theorem 2 Given $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that (4) converges to $N^{\delta}(M)$ provided $\epsilon \leq \epsilon(\delta)/2$



Related Recent Work

- A general stochastic recursion with set-valued maps and Markov noise

$$x_{n+1} = x_n + a(n)(h(x_n, Z_n) + M_{n+1})$$

- General convergence with Z_n non-ergodic, iterate-dependent, Markov process [V.Yaji and SB, *Stochastics*, 2018]

- Two-timescale stochastic recursions with Markov noise

$$\begin{aligned} x_{n+1} &= x_n + a(n)(h(x_n, y_n, Z_n^1) + M_{n+1}^1) \\ y_{n+1} &= y_n + b(n)(g(x_n, y_n, Z_n^2) + M_{n+1}^2) \end{aligned}$$

- Z_n^1, Z_n^2 independent non-ergodic iterate-dependent Markov processes, M_{n+1}^1, M_{n+1}^2 independent martingale differences, $a(n) = o(b(n))$, h, g point-to-point maps – analysis and application to reinforcement learning [P.Karmakar and SB, *Math of OR*, 2018]

- Analysis under set-valued h, g [V.Yaji and SB, *Math of OR*, 2019]

Ongoing and Future Work

- Finding minima of non-differentiable functions under noise
- Algorithms for convergence to global minima
- Asynchronous update algorithms
- Reinforcement learning algorithms for partially observed Markov decision processes
- Analysis of deep reinforcement learning algorithms
- Applications in robotics, microgrids, vehicular traffic control etc.