Stochastic approximation algorithms with set-valued mean fields: Theory and applications
Arunselvan. R
Thesis Advisor: Prof. Shalabh Bhatnagar
Department of Computer Science and Automation, Indian Institute of Science.
arunselvan@csa.iisc.ernet.in

Quick introduction to stochastic approximation algorithms
Consider the following recursion in \( \mathbb{R}^d \) (\( t \geq 1 \)):

\[ x_{n+1} = x_n + a_n (h(x_n) + M_{n+1}), \text{ for } n \geq 0, \]

where \( \mathbb{R}^d \rightarrow \mathbb{R}^d \) is a Lipschitz continuous function.

(i) \( (a_n) \) is a non-increasing sequence satisfying \( \sum_{n=0}^{\infty} a_n = \infty \) and \( \sum_{n=0}^{\infty} a_n^2 < \infty \).

(ii) \( (M_n) \) is a sequence of independent random variables in \( \mathbb{R}^d \) which constitute the noise.

In 1996, Benaim [1] showed that the asymptotic behavior of a stochastic recursive equation can be studied by analyzing the asymptotic behavior of the associated ODE.

Borkar-Meyn theorem for stochastic recursive inclusions, [3]

The objective is to develop sufficient conditions that are easily verifiable for both stability and convergence of set-valued dynamical systems given by:

\[ x_{n+1} = x_n + a_n (h(x_n) + M_{n+1}), \text{ for } n \geq 0, \]

where \( h \in \mathcal{L}(\mathbb{R}^d) \) and \( h : \mathbb{R}^d \rightarrow \{ \text{subsets of } \mathbb{R}^d \} \) is a Marchaud map.

Although there are two different set of assumptions in [3], we consider only one here.

Assumptions
- \( h \) is a Marchaud map. The step-size and Martingale noise sequence satisfy the standard assumptions.
- For \( \epsilon \geq 1 \) and \( x \in \mathbb{R}^d \), define \( h_\epsilon(x) := \{ y \mid y \in h(x) \} \).
- Further, for each \( x \in \mathbb{R}^d \), define \( h_\epsilon(x) := \{ y \mid \epsilon x \in h(y) \} \), i.e., the closure of the lower-limit of \( \{ h(x) \}_{\epsilon > 0} \).

(A4) \( h_\epsilon(x) \) is non-empty for all \( x \in \mathbb{R}^d \). Further, the differential inclusion \( h(x) \in h_\epsilon(x) \) has an attracting set, \( A \), with \( T_0 \{ 0 \} \) as a subset of its lower semi-continuous.

Let \( \epsilon \geq 1 \) be an increasing sequence of integers such that \( \epsilon_n \uparrow \infty \) as \( n \rightarrow \infty \). Further, let \( k_0 := 0 \) and \( y_0 := y \in y \) as \( n \rightarrow \infty \), such that \( y \in h_\epsilon(x_n) \). Now, let \( y \in h_\epsilon(x) \).

- Define \( \delta_1 := \delta_1 \) and pick real numbers \( \delta_2, \delta_3 \) such that \( \delta_1 < \delta_2 < \delta_3 < 1 \).

Outline of the proof

Application: The problem of approximate drift
- If in practice the drift function cannot be calculated accurately. A natural question is the following: Are the iterates stable? If so, where do they converge?

- We use our framework to show that the algorithm with approximate drift is stable provided the algorithm with “accurate drift” was stable. Further, we show that the algorithm converges to a neighborhood of the “intended” set, where the neighborhood is dependent on the drift errors.

Gradient based learning algorithms with constant-error gradient estimators, [6]

- Implementations of stochastic gradient search algorithms such as back propagation typically rely on finite difference (FD) approximation methods. These methods are used to approximate the objective function gradient in steepest descent algorithms as well as the gradient and Hessian inverse in Newton based schemes.

- In literature, the convergence analyses critically require that perturbation parameters in the estimators of the gradient/Hessian approach zero. However, in practice, the perturbation parameter is often held fixed to a ‘small’ constant resulting in constant-error estimates. Item [6], we present a framework to analyze the aforementioned.

- Easily verifiable conditions are presented for stability and convergence when using such FD estimators for the gradient/Hessian. In addition, our framework dispenses with a critical restriction on the step-sizes (learning rate) when using FD estimators, see [6] for details.

Stochastic recursive inclusion in two timescales with an application to the Lagrangian dual problem, [5]

We consider the following coupled iteration.

\[ x_{n+1} = x_n + a_n (h_\epsilon(x_n) + M_{n+1}), \]

\[ y_{n+1} = y_n + b_n (h_\epsilon(y_n) + M_{n+1}), \]

where \( a_n \in \mathcal{B}(x_n, y_n), x_n \in \mathcal{B}(x_n, y_n) \) such that \( b_n \) and \( y_n \) are Marchaud maps. The step-size satisfies the standard assumptions and \( \sum_{n=0}^{\infty} b_n^2 < \infty \). The iterates are assumed to be stable. The following is a key assumption that couples the \( x \) and \( y \) iterates, see [5] for more details.

(A5) The map \( \lambda, \gamma := \{ \text{globally attractive set of } x(t) \in h_\epsilon(x(t), y(t)) \} \) is upper semi-continuous.

Main result

Almost surely the set of accumulation points is given by

\[ \left\{ (x, y) \mid A = \bigcup_{\epsilon > 0} \{ x \mid \lambda_\epsilon(x) = \emptyset \} \subseteq \bigcup_{\epsilon > 0} \{ (x, y) : y \in h_\epsilon(x) \} \right\}. \]

Application: The Lagrangian dual problem

- To solve the constrained minimization problem one often constructs an associated two timescale stochastic approximation algorithm.

- The analysis involves considering a family of minimum sets. Hitherto in literature these minimum sets are assumed to be singletons.

- We extend this analysis to the general case of set-valued minimum sets.


In [4] a ‘stability theorem’ for stochastic approximation (SA) algorithms with ‘controlled Markov’ noise. The iterates are shown to track a solution to a differential inclusion defined in terms of the ergodic occupation measures associated with the ‘controlled Markov’ process.

We improve the general algorithm of Temporal Difference Learning using our framework.

References


