Stochastic approximation algorithms with set-valued mean fields and their applications

Arun Selvan. R

Department of Computer Science and Automation,
Indian Institute of Science,
Bangalore.

Thesis Advisor: Prof. Shalabh Bhatnagar.

April 29, 2016

Introduction

1. Adaptive algorithms that are iterative in structure.

$$x_{n+1} = x_n + a(n)\left[h(x_n) + M_{n+1}\right], \tag{1}$$

   where $a(n)$ is the step-size, M_{n+1} is the martingale noise, h is the drift or mean field.

2. Example: stochastic gradient descent; $h = -\nabla F$.

Introduction

1. Adaptive algorithms that are iterative in structure.

$$x_{n+1} = x_n + a(n)\left[h(x_n) + M_{n+1}\right], \tag{1}$$

   where $a(n)$ is the step-size, M_{n+1} is the martingale noise, h is the drift or mean field.

2. Example: stochastic gradient descent; $h = -\nabla F$.

3. Focus of our talk: h can be set-valued.

4. Example: stochastic sub-gradient descent; $h(x) = \{-g \mid F(y) \geq F(x) + g^T(y - x)\ \forall y\}$.

Gradient based learning algorithms with errors.

1. Stochastic gradient descent to find the minimum of $F : \mathbb{R}^d \to \mathbb{R}$:

$$x_{n+1} = x_n - a(n) \left(\begin{pmatrix} \frac{F(x_n + p(n)\xi_1) - F(x_n - p(n)\xi_1)}{2p(n)} \\ \vdots \\ \frac{F(x_n + p(n)\xi_d) - F(x_n - p(n)\xi_d)}{2p(n)} \end{pmatrix} + M_{n+1} \right). \tag{2}$$

Two-sided Kiefer-Wolfowitz gradient estimator is used.

Error at stage n:

$\epsilon_n = (-\nabla F(x_n))$ − gradient estimate at stage n.

ξ_i is the vector with 1 at the i^{th} place and 0 in all others.

Errors vanish over time: $\epsilon_n \to 0$

1. Bertsekas and Tsitsiklis studied the case when $\epsilon_n \to 0$:

$$x_{n+1} = x_n + a(n)\left[g(x_n) + M_{n+1}\right],\ g(x_n) \in -\nabla F_x|_{x=x_n} + \overline{B}_{\epsilon_n}(0).$$

Bertsekas, Dimitri P and Tsitsiklis, John N. (2000) ‘Gradient convergence in gradient methods with errors.’, *SIAM Journal on Optimization, 10(3):627642, 2000.*

Errors vanish over time: $\epsilon_n \to 0$

1. Bertsekas and Tsitsiklis studied the case when $\epsilon_n \to 0$:

$$x_{n+1} = x_n + a(n)\left[g(x_n) + M_{n+1}\right],\ g(x_n) \in -\nabla F_x|_{x=x_n} + \overline{B}_{\epsilon_n}(0).$$

2. $(A1)\ \|\epsilon_n\| \leq a(n)(c + d\|\nabla F_x|_{x=x_n}\|)$, $(A2)\ \sum_n \frac{a(n)^2}{p(n)^2} < \infty$.

Bertsekas, Dimitri P and Tsitsiklis, John N. (2000) 'Gradient convergence in gradient methods with errors.', *SIAM Journal on Optimization, 10(3):627642, 2000.*

Errors vanish over time: $\epsilon_n \to 0$

1. Bertsekas and Tsitsiklis studied the case when $\epsilon_n \to 0$:

$$x_{n+1} = x_n + a(n)\left[g(x_n) + M_{n+1}\right],\ g(x_n) \in -\nabla F_x|_{x=x_n} + \overline{B}_{\epsilon_n}(0).$$

2. $(A1)\ \|\epsilon_n\| \leq a(n)(c + d\|\nabla F_x|_{x=x_n}\|)$, $(A2)\ \sum_n \frac{a(n)^2}{p(n)^2} < \infty$.

3. Main result: The iterates diverge *a.s.* or converge to the minimum *a.s.*

4. *Pros:* Stability not assumed, no "mixed" results.
Cons: couples step-sizes and estimation errors, requires estimation errors to go to zero, does not analyze Newton's method.

Bertsekas, Dimitri P and Tsitsiklis, John N. (2000) 'Gradient convergence in gradient methods with errors.', *SIAM Journal on Optimization, 10(3):627642, 2000.*

Our contributions: $\epsilon_n \nrightarrow 0$

1. In practice $p(n) := p$ at every stage *i.e.*, expect $\epsilon_n \leq \epsilon$.

Our contributions: $\epsilon_n \nrightarrow 0$

1. In practice $p(n) := p$ at every stage *i.e.*, expect $\epsilon_n \leq \epsilon$.
2. ~~(*A*1), (*A*2)~~ (Step-size and the estimation error decoupled).

Our contributions: $\epsilon_n \nrightarrow 0$

1. In practice $p(n) := p$ at every stage *i.e.*, expect $\epsilon_n \leq \epsilon$.
2. ~~(A1), (A2)~~ (Step-size and the estimation error decoupled).
3. Unified framework to analyze gradient descent, Newton's method and any gradient method with constant-errors.

$$x_{n+1} = x_n + a(n)\left[g(x_n) + M_{n+1}\right],\ g(x_n) \in G(x_n), \qquad (3)$$

$$G(x_n) = -\nabla F_x|_{x=x_n} + \overline{B}_\epsilon(0) \text{ or } -H^{-1}(x_n)\nabla F_x|_{x=x_n} + \overline{B}_\epsilon(0).$$

Main result: Gradient descent or Newton's method.

1. Sufficient conditions for stability and convergence that does not couple step-size and error.

2. Main result: Given $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that if the estimation error at each stage is at most $\epsilon(\delta)$ then the iterates are stable and converge to the δ-neighborhood of the minimum set of F.

A.R. and Shalabh Bhatnagar (2016) 'Gradient-based learning algorithms with constant-error gradient estimators: stability and convergence', *arxiv preprint: arXiv:1604.00151*.

General Borkar-Meyn Theorem for SRI

1. Easily verifiable sufficient conditions for stability and convergence of

$$x_{n+1} = x_n + a(n)\left[y_n + M_{n+1}\right], \text{ where } y_n \in h(x_n). \quad (4)$$

$\|h(x)\| \leq K(1 + \|x\|)$; $h(x)$ is convex and compact; h is upper semi-continuous.

A.R. and Shalabh Bhatnagar (2015) 'A Generalization of the Borkar-Meyn Theorem for Stochastic Recursive Inclusions.', *[arXiv:1502.01953v2]*.

General Borkar-Meyn Theorem for SRI

1. Easily verifiable sufficient conditions for stability and convergence of

$$x_{n+1} = x_n + a(n)\left[y_n + M_{n+1}\right], \text{ where } y_n \in h(x_n). \qquad (4)$$

   $\|h(x)\| \leq K(1 + \|x\|)$; $h(x)$ is convex and compact; h is upper semi-continuous.

2. Understanding unstable iterates becomes important.
3. $\dot{x}(t) \in h_\infty(x(t))$ arises naturally in such a study.

A.R. and Shalabh Bhatnagar (2015) 'A Generalization of the Borkar-Meyn Theorem for Stochastic Recursive Inclusions.', *[arXiv:1502.01953v2]*.

General Borkar-Meyn Theorem for SRI

1. Under a projective scheme with projections on the unit ball centered at origin

$$\frac{x_{n+k}}{r(n)} = \frac{x_n}{r(n)} + \sum_{i=0}^{k-1} a(n+i) \left(\frac{y(n+i)}{r(n)} + \frac{M_{n+i+1}}{r(n)} \right), \quad (5)$$

where $r(n) = \|x_n\| \vee 1$. Unstable means $r(n) \uparrow \infty$.

General Borkar-Meyn Theorem for SRI

1. Under a projective scheme with projections on the unit ball centered at origin

$$\frac{x_{n+k}}{r(n)} = \frac{x_n}{r(n)} + \sum_{i=0}^{k-1} a(n+i)\left(\frac{y(n+i)}{r(n)} + \frac{M_{n+i+1}}{r(n)}\right), \qquad (5)$$

   where $r(n) = \|x_n\| \vee 1$. Unstable means $r(n) \uparrow \infty$.

2. For $c \geq 1$ and $x \in \mathbb{R}^d$, define $h_c(x) = h(cx)/c$.
   Note $y(n+i)/r(n) \in h_{r(n)}(x_{n+i}/r(n))$.
   $h_\infty(x) := \overline{Limsup_{c\to\infty}\ h_c(x)}$.
   $Limsup_{n\to\infty} K_n := \{y \mid \underline{\lim}_{n\to\infty} d(y, K_n) = 0\}$.

General Borkar-Meyn Theorem for SRI

1. Under a projective scheme with projections on the unit ball centered at origin

$$\frac{x_{n+k}}{r(n)} = \frac{x_n}{r(n)} + \sum_{i=0}^{k-1} a(n+i)\left(\frac{y(n+i)}{r(n)} + \frac{M_{n+i+1}}{r(n)}\right), \qquad (5)$$

where $r(n) = \|x_n\| \vee 1$. Unstable means $r(n) \uparrow \infty$.

2. For $c \geq 1$ and $x \in \mathbb{R}^d$, define $h_c(x) = h(cx)/c$.
Note $y(n+i)/r(n) \in h_{r(n)}(x_{n+i}/r(n))$.
$h_\infty(x) := \overline{Limsup_{c\to\infty}\ h_c(x)}$.
$Limsup_{n\to\infty} K_n := \{y \mid \underline{lim}_{n\to\infty} d(y, K_n) = 0\}$.

3. Impose mild restrictions on $\dot{x}(t) \in h_\infty(x(t))$ for stability.

Two timescale schemes for *SRI*: Motivation

1. Constrained minimization: $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}^k$. Minimize $f(x)$ subject to the condition that $g(x) \leq 0$. Suppose strong duality holds then we may solve the following dual problem:

$$\sup_{\substack{\mu \in \mathbb{R}^k \\ \mu \geq 0}} \inf_{x \in \mathbb{R}^d} \Big(f(x) + \mu^T g(x) \Big).$$

Two timescale schemes for *SRI*: Motivation

1. Constrained minimization: $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}^k$. Minimize $f(x)$ subject to the condition that $g(x) \leq 0$. Suppose strong duality holds then we may solve the following dual problem:

$$\sup_{\substack{\mu \in \mathbb{R}^k \\ \mu \geq 0}} \inf_{x \in \mathbb{R}^d} \left(f(x) + \mu^T g(x) \right).$$

2.

$$x_{n+1} = x_n - a(n) \left[\nabla_x \left(f(x_n) + \mu_n^T g(x_n) \right) + M_{n+1}^2 \right],$$
$$\mu_{n+1} = \mu_n + b(n) \left[\nabla_\mu \left(f(x_n) + \mu_n^T g(x_n) \right) + M_{n+1}^1 \right].$$

In the above $\frac{b(n)}{a(n)} \to 0$.

Our contributions

1. To study the x iterates: $\lambda_m : \mathbb{R}^k \rightarrow \mathbb{R}^d$, where $\lambda_m(\mu_0)$ is the global attractor of $\dot{x}(t) = -\nabla_x(f(x) + \mu_0^T g(x))$.
2. Previous literature: λ_m is single valued and continuous map.

G. B. Dantzig and J. Folkman and N. Shapiro (1967) 'On the continuity of the minimum set of a continuous function', *Journal of Mathematical Analysis and Applications*.

Our contributions

1. To study the x iterates: $\lambda_m : \mathbb{R}^k \to \mathbb{R}^d$, where $\lambda_m(\mu_0)$ is the global attractor of $\dot{x}(t) = -\nabla_x(f(x) + \mu_0^T g(x))$.
2. Previous literature: λ_m is single valued and continuous map.
3. We allow λ_m to be set-valued. To show u.s.c. of λ_m we use Dantzig, Folkman and Shapiro.
4. Main result: $(x_n, \mu_n) \to (x^*, \mu^*)$ that solves the dual.

G. B. Dantzig and J. Folkman and N. Shapiro (1967) 'On the continuity of the minimum set of a continuous function', *Journal of Mathematical Analysis and Applications*.

1. More generally, we consider the following two timescale scheme:

$$x_{n+1} = x_n + a(n)\left[u_n + M^1_{n+1}\right],$$
$$y_{n+1} = y_n + b(n)\left[v_n + M^2_{n+1}\right],$$

$u_n \in h(x_n, y_n)$, $v_n \in g(x_n, y_n)$, $h : \mathbb{R}^{d+k} \to \{\textit{subsets of } \mathbb{R}^d\}$ and $g : \mathbb{R}^{d+k} \to \{\textit{subsets of } \mathbb{R}^k\}$.

A.R. and Shalabh Bhatnagar (2015) 'Stochastic recursive inclusion in two timescales with an application to the Lagrangian dual problem.', *[arXiv:1502.01956v2]*.

General theory: Two timescale for *SRI*

1. More generally, we consider the following two timescale scheme:

$$x_{n+1} = x_n + a(n)\left[u_n + M^1_{n+1}\right],$$
$$y_{n+1} = y_n + b(n)\left[v_n + M^2_{n+1}\right],$$

$u_n \in h(x_n, y_n)$, $v_n \in g(x_n, y_n)$, $h : \mathbb{R}^{d+k} \to \left\{\textit{subsets of } \mathbb{R}^d\right\}$ and $g : \mathbb{R}^{d+k} \to \left\{\textit{subsets of } \mathbb{R}^k\right\}$.

2.
   - Assume stability of the iterates.
   - $\dot{x}(t) \in h(x(t), y)$ has a globally attracting set, A_y, that is also Lyapunov stable.
   - The set-valued map $\lambda : \mathbb{R}^k \to A_y$ is upper semi-continuous.

A.R. and Shalabh Bhatnagar (2015) 'Stochastic recursive inclusion in two timescales with an application to the Lagrangian dual problem.', *[arXiv:1502.01956v2]*.

Stability of SAA with controlled Markov noise

1. Sufficient conditions for stability and convergence of *SAA* with 'controlled Markov noise'.

$$x_{n+1} = x_n + a(n) \left[h(x_n, y_n) + M_{n+1} \right], \tag{6}$$

   where $\{y_n\}_{n \geq 0}$ is an S-valued Markov process such that S is compact.
2. In reinforcement learning, the state space, S, is often finite (hence compact).
3. Our contribution: Sufficient conditions for stability and convergence including the case of non-unique stationary distributions.

   A.R. and Shalabh Bhatnagar (2015) 'Stability Theorem for Stochastic Approximation with Controlled Markov Noise with an Application to Temporal-Difference Learning .', *[arXiv:1504.06043v1]*.

Thank you. Questions?