Stochastic Approximation with Markov Noise
Analysis and applications

Prasenjit Karmakar & Shalabh Bhatnagar
Department of Computer Science and Automation, IISC Bangalore
pkarmakar6@gmail.com

Stochastic Approximation and Ordinary Differential Equation (O.D.E) method
- Sequential methods for finding a zero or minimum of a function where only the noisy observations of the function values are available.
- Iteration:
  \[ \theta_{n+1} = \theta_n + a(n)/h(\theta_n) + M_{n+1} \]
  \( h \) Lipschitz, \( M_n \) martingale difference sequence.
- Converges to the globally asymptotically stable equilibrium of the O.D.E \( \dot{\theta}(t) = h(\theta(t)) \) under reasonable assumptions such as boundedness of the iterates.
- Questions:
  - What if the above o.d.e does not have a globally asymptotically stable equilibrium?
  - What if there is a non-additive Markov noise present in the vector field \( h \)?
  - What if the iterates are not known to be bounded beforehand?
  - Such scenario arises in off-policy learning.

Off policy TD with linear function approximation
- Given (state, action, reward) trajectory such as:
  \[ S_1, A_1, R_1, S_2, A_2, R_2, \ldots \]
  for a behaviour policy \( v \) estimate value function i.e find the TD(0) solution for the target policy \( \pi \)
- Standard temporal difference learning with linear function approximation may diverge. Also, the usual single time-scale stochastic approximation kind of argument may not be useful as the associated ordinary differential equation (o.d.e) may not have the TD(0) solution as its globally asymptotically stable equilibrium.
- Solution: TDC with importance weighting
  \[ \theta_{n+1} = \theta_n + a(n)v_\pi \frac{h(\theta_n)\pi(a_n|s_n) - \gamma \hat{g}_\pi^T \phi_n}{\sum_{a} \pi(a_n|s_n)} \phi_n \]
  \( \hat{g}_\pi = \phi(S_n) \phi_n + \phi(S_{n+1}) \hat{\lambda}(\theta) = R_n + \gamma \hat{g}_\pi^T \phi_n - \theta^T \phi_n \pi_n \thicksim A_n(s_n) \phi_n \rightarrow \pi(\phi) \pi(\phi) \]
  \( \phi_n \) as \( n \to \infty \).
- Analyzing in single time-scale requires knowledge of stationary distribution.
- Use two time-scale framework to make sure that the O.D.Es have globally asymptotically stable equilibria.
- Earlier convergence analysis assumed that i.i.d samples of stationary distribution available !.
- We prove that \( \theta_n \) converges to the TD(0) fixed point using the theory described in the next section.

Problem 1: 2 timescale stochastic approximation with controlled Markov noise [2]
- Asymptotic analysis of the following coupled iterations:
  \[ \theta_{n+1} = \theta_n + a(n)v_\pi \frac{h(\theta_n)\pi(a_n|s_n) - \gamma \hat{g}_\pi^T \phi_n}{\sum_{a} \pi(a_n|s_n)} \phi_n \]
  \[ w_{n+1} = w_n + h(n) (\phi_n(s_n) - \gamma \phi_n(S_n)) \phi_n \]
  \( \phi_n \) as \( n \to \infty \).

Specific Assumptions for two timescale analysis
Faster D.L. \( \forall \theta \in \mathbb{R}^d \), the differential inclusion
\[ u(t) \in \partial \nu(H(\theta(t))) \]
has a singleton global attractor (g.a) \( b(\theta) \) where \( \mathbb{R}^d \to \mathbb{R}^d \) is a Lipschitz map with constant \( K \). Here \( \partial \nu(\theta) = \{ \phi(\theta, v) \mid v \in D^{(1)}(\theta, \nu(\theta)) \} \).

Slower D.L. The inclusion
\[ \theta(t) \in \partial \nu(\theta(t)) \]
has a g.a. set \( A_\theta \). Here \( \nu(\theta) = \{ \phi(\theta, w) \mid w \in D^{(1)}(\theta, \nu(\theta)) \} \).

Stability \[ u_{\infty}(\theta(t)) + |w_{\infty}(\theta(t))| < \infty \] as \( t \to \infty \).

Main Results
Introduce Discrete Measure Process: \( \mu(t) = \delta_{\theta(t)} \) when \( t \in \{n, n+1, n+2, \ldots \} \).

Lemma 1 (Tracking Lemma), Consider the non-autonomous O.D.E.
\[ \dot{\theta}(t) = h(\theta(t), \nu(\theta(t), \mu(t))) \] (1)
- Let \( \dot{\theta} \) be the piece-wise linear interpolated trajectory of the slower iterates and \( \dot{\theta}(t) \), \( t \geq s \) denote the solution to (1) with \( \dot{\theta}(s) = \theta(s) \geq 0 \). Then \( \dot{\theta}(t) \) tracks the above O.D.E.

Lemma 2 (Limit of the Discrete Measure Process), Almost surely every limit point of \( (\mu(s+.), \omega(t+.) \) as \( s \to \infty \) is of the form \( (\mu(\cdot), \dot{\theta}(\cdot)) \), where \( \mu(\cdot) \) satisfies \( \mu(\cdot) \in D^{(1)}(\dot{\theta}(\cdot), \nu(\dot{\theta}(\cdot))) \). (Remark: \( \dot{\theta}(\cdot) \) satisfies the above mentioned O.D.E with \( \mu(\cdot) \) replaced by \( \mu(\cdot) \).

Lemma 3 (Limiting \( \mu(\cdot) \) and \( \dot{\theta}(\cdot) \), \( \dot{\theta}(\cdot) \) satisfies the above mentioned O.D.E with \( \mu(\cdot) \) replaced by \( \mu(\cdot) \)).

Problem 2: Relaxing the boundedness of the iterates assumption [3]
- Extension of lock-in probability to Markov noise (first single timescale and then a special case of 2 timescale).
- For sufficiently large \( n \) calculate lower bound of
  \[ f(\theta_n) \to (\theta_n) \in B \]
  for a compact \( B \subset C \) with \( H(\theta \nu) \) being an asymptotically stable attractor of the corresponding o.d.e and \( C \) is the domain of attraction.
- The boundedness of the iterates is replaced by asymptotic tightness of the iterates.
- We also give Lyapunov type dependences for asymptotic tightness.
- This, in turn, is shown to be useful in analyzing the tracking ability of general adaptive algorithms.
- We estimate sample complexity of such recursions which is used for step-size selection.

Problem 3: Function approximation error bound for risk-sensitive reinforcement learning (RL) [1]

- Risk-sensitive cost:
  \[ \lim_{n \to \infty} \sup \frac{1}{n} \ln \left( E^\pi \left[ \sum_{t=0}^{n-1} \nu(X_t) \lambda(X_t) \right] \right) \]

- The Poisson equation here is multiplicative i.e. it is a non-linear eigenvalue problem.
- The eigenvalue is the Perron-Frobenius (PF) one.
- The corresponding RL algorithm with function approximation also converges to a PF eigenvalue of a non-negative matrix.
- We give several bounds between the original cost and approximated cost.

References

![Figure 1: Comparison between TD(0), OFFTDIC, and ONTDIC for Baird’s counterexample](image)
Analysis of Stochastic Approximation with Markov Noise and applications

Prasenjit Karmakar
Advisor: Prof. Shalabh Bhatnagar
Department of Computer Science and Automation

7th April, 2017
Presentation Outline

1. Introduction

2. Application 1: Off policy TD with linear function approximation

3. Problem 2

4. Problem 3
Stochastic Approximation

- Sequential methods for finding a zero or minimum of a function where only the noisy observations of the function values are available.

- Example: find zero of the function $F(\theta) = E[g(\theta, \eta)]$
  - Distribution of $\eta$ unavailable.
  - But, simulated i.i.d samples $\eta_n$, $n \geq 1$ of $\eta$ are available.
  - Algorithm: $\theta_{n+1} = \theta_n + a(n)g(\theta_n, \eta_{n+1})$.
  - $g(\theta_n, \eta_{n+1}) = F(\theta_n) + M_{n+1}$.
  - Martingale Difference: $M_{n+1} = g(\theta_n, \eta_{n+1}) - E[g(\theta_n, \eta_{n+1})|\mathcal{F}_n]$.
  - $\mathcal{F}_n = \sigma(\theta_m, \eta_m, m \leq n)$. 
Ordinary Differential Equation (O.D.E) Method

- O.D.E: \( \dot{\theta}(t) = F(\theta(t)) \).

For any \( T > 0 \), \( \sup_{t \in [s, s+T]} \| \tilde{\theta}(t) - \theta^*(t) \| \to 0 \), a.s. as \( s \to \infty \).
Almost sure convergence of the algorithm

- possible to tell whether zero’s of $F$ are globally asymptotically stable equilibrium of the above o.d.e without knowing $F$ explicitly. e.g. $F = -\nabla f$ then $\{\nabla f = 0\}$ is the such a set.

- Conclusion[1]: Algorithm converges to the required zero of $F(.)$.

- What if the o.d.e does not have a globally asymptotically stable equilibrium?
  - sometimes (!) analyzing in 2-timescale helps.

---

2 timescale stochastic approximation

\[ (\text{slow}) \quad \theta_{n+1} = \theta_n + a(n) h(\theta_n, w_n, \eta_n^{(1)}), \]
\[ (\text{fast}) \quad w_{n+1} = w_n + b(n) g(\theta_n, w_n, \eta_n^{(2)}) \]

\[ \frac{a(n)}{b(n)} \to 0 \] makes it two timescale.

What if \( \eta_n^{(i)} \) are Markov noise, they cannot be converted to martingale difference.

Source of Markov noise

- Parametrization of value function: \( V_\theta = \theta^T \phi \).
- \( \{X_n\} \) present in the algorithm rather than \( I_{\{X_n=i\}} \) (non-parametric case).

Previous work: assumes that i.i.d samples of stationary distribution available !
Our contributions

- Convergence analysis of two time-scale stochastic approximation with controlled Markov noise assuming stability i.e. 
  \[ \sup_n (\|\theta_n\| + \|w_n\|) < \infty \text{ a.s.}[2] \]
- Apply a special case of our results to solve the well-known off-policy convergence problem for TD with linear parametrization.
- Convergence analysis of such recursions without assuming the stability of the iterates [3].
- Function Approximation error bound for risk-sensitive reinforcement learning [4].

---

2 P.Karmakar and S.Bhatnagar. accepted in Mathematics of Operations Research
Presentation Outline

1. Introduction

2. Application 1: Off policy TD with linear function approximation

3. Problem 2

4. Problem 3
What is Off-policy TD convergence problem?

- Given (state, action, reward) pairs 
  \[ S_1, A_1, R_1, S_2, A_2, R_2, \ldots \]
  for a behaviour policy \( \pi_b \) estimate value function for the target policy \( \pi \neq \pi_b \).
- Need to design an on-line algorithm which converges to the \( TD(0) \)-fixpoint.
- Algorithm: TDC with importance weighting [5]
  \[
  \begin{align*}
  \theta_{n+1} &= \theta_n + a(n) \rho_n \left[ \delta_n(\theta_n) \phi_n - \gamma \phi'_n \phi_n^T w_n \right], \\
  w_{n+1} &= w_n + b(n) \left[ (\rho_n \delta_n(\theta_n) - \phi_n^T w_n) \phi_n \right]
  \end{align*}
  \]
  \[
  \phi_n = \phi(S_n), \quad \phi'_n = \phi(S_{n+1}), \quad \delta_n(\theta) = R_n + \gamma \theta^T \phi'_n - \theta^T \phi_n, \quad \rho_n = \frac{\pi(A_n|S_n)}{\pi_b(A_n|S_n)}
  \]
- We analyze in 2-timescale to make sure that the o.d.e’s have globally asymptotically stable equilibrium.

---

Presentation Outline

1. Introduction
2. Application 1: Off policy TD with linear function approximation
3. Problem 2
4. Problem 3
Relaxing the stability of the iterates assumption

- Extension of lock-in probability to Markov noise.
- For sufficiently large $n_0$ calculate lower bound of

$$P(\theta_n \to H | \theta_{n_0} \in B)$$

for a compact $\bar{B}$ such that $H \subset \bar{B} \subset G$ with $H$ being an asymptotically stable attractor of the corresponding o.d.e and $G$ is the domain of attraction.

- The boundedness of the iterates is replaced by asymptotic tightness of the iterates.
- We also give Lyapunov type conditions for asymptotic tightness.
- We estimate sample complexity of such recursions which is used for step-size selection.
Presentation Outline

1. Introduction

2. Application 1: Off policy TD with linear function approximation

3. Problem 2

4. Problem 3
Problem 3

Function approximation error bound for risk-sensitive RL

- Risk-sensitive cost:
  \[
  \limsup_{n \to \infty} \frac{1}{n} \ln \left( \frac{1}{n} \ln \left( E[e^{\sum_{m=0}^{n-1} c(X_m, X_{m+1})}] \right) \right).
  \]

- The Poisson equation here is multiplicative i.e. it is a non-linear eigenvalue equation.

- The eigenvalue is the Perron-Frobenius (PF) one.

- The corresponding RL algorithm [6] with function approximation also converges to a PF eigenvalue of a non-negative matrix.

- We give several bounds between the original cost and approximated cost.

---

Thank You. Questions ?