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:

$$\theta_{n+1} = \theta_n + a(n)[h(\theta_n) + M_{n+1}], n \geq 0,$$

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 π_b estimate value function (i.e find the TD(0) solution) for the target policy $\pi \neq \pi_b$.

- 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

$$\theta_{n+1} = \theta_n + a(n)\rho_n\left[\delta_n(\theta_n)\phi_n - \gamma\phi'_n\phi_n^T w_n\right],$$
$$w_{n+1} = w_n + b(n)\left[(\rho_n\delta_n(\theta_n) - \phi_n^T w_n)\phi_n\right]$$
$$\phi_n = \phi(S_n), \phi'_n = \phi(S_{n+1}), \delta_n(\theta) = R_n + \gamma\theta^T\phi'_n - \theta^T\phi_n, \rho_n = \frac{\pi(A_n|S_n)}{\pi_b(A_n|S_n)}$$

- 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 equilibrium.
- Earlier convergence analysis assumed that i.i.d samples of stationary distribution available !.
- We prove that θ_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:

$$\theta_{n+1} = \theta_n + a(n)\left[h(\theta_n, w_n, Z_n^{(1)}) + M_{n+1}^{(1)}\right],$$
$$w_{n+1} = w_n + b(n)\left[g(\theta_n, w_n, Z_n^{(2)}) + M_{n+1}^{(2)}\right]$$
$$Z_{n+1}^{(i)} \sim p^{(i)}(.|Z_n^{(i)}, A_n^{(i)}, \theta_n, w_n), i = 1, 2$$

- $\frac{a(n)}{b(n)} \to 0$ makes it two timescale

Let $D^{(i)}(\theta, w), i = 1, 2$ be the set of all ergodic occupation measures for the prescribed θ and w. Define $\tilde{g}(\theta, w, \nu) = \int g(\theta, w, z)\nu(dz, U^{(2)})$ for $\nu \in P(S^{(2)} \times U^{(2)})$.

Specific Assumptions for two timescale analysis

Faster D.I. $\forall\theta \in \mathbb{R}^d$, the differential inclusion

$$\dot{w}(t) \in \hat{g}_\theta(w(t))$$

has a singleton global attractor (g.a.) $\lambda(\theta)$ where $\lambda : \mathbb{R}^d \to \mathbb{R}^k$ is a Lipschitz map with constant K. Here $\hat{g}_\theta(w) = \{\tilde{g}_\theta(w, \nu) : \nu \in D^{(2)}(\theta, w)\}$. Most important assumption as it links the fast and slow iterates.

Slower D.I. The inclusion

$$\dot{\theta}(t) \in \hat{h}(\theta(t)))$$

has a g.a. set A_0. Here $\hat{h}(\theta) = \{\tilde{h}(\theta, \lambda(\theta), \nu) : \nu \in D^{(1)}(\theta, \lambda(\theta))\}$.

Stability $\sup_n(\|\theta_n\| + \|w_n\|) < \infty$ a.s.

Main Results

Introduce Dirac Measure Process: $\mu(t) = \delta_{Z_n^{(i)}}$ when $t \in [t(n), t(n+1))$.

Lemma 1 (Tracking Lemma). *Consider the non-autonomous O.D.E.*

$$\dot{\theta}(t) = \tilde{h}(\theta(t), \lambda(\theta(t)), \mu(t)) \quad (1)$$

Let $\bar{\theta}(.)$ be the piecewise linear interpolated trajectory of the slower iterate and $\theta^s(t), t \geq s$ denote the solution to (1) with $\theta^s(s) = \bar{\theta}(s)$, for $s \geq 0$. Then $\bar{\theta}(.)$ tracks the above O.D.E.

Lemma 2 (Limit of the Dirac measure Process). *Almost surely every limit point of $(\mu(s+.), \bar{\theta}(s+.))$ as $s \to \infty$ is of the form $(\tilde{\mu}(\cdot), \tilde{\theta}(\cdot))$, where $\tilde{\mu}(\cdot)$ satisfies $\tilde{\mu}(t) \in D^{(1)}(\tilde{\theta}(t), \lambda(\tilde{\theta}(t)))$.*

Lemma 3 (Lemma linking $\tilde{\mu}(\cdot)$ and $\tilde{\theta}(\cdot)$). *$\tilde{\theta}(\cdot)$ satisfies the above mentioned O.D.E with $\mu(\cdot)$ replaced by $\tilde{\mu}(\cdot)$*

Theorem 1. *Under mentioned assumptions,*

$$(\theta_n, w_n) \to \cup_{\theta^* \in A_0}(\theta^*, \lambda(\theta^*)) \textit{a.s. as } n \to \infty.$$

Proof Outline. $\theta_n \xrightarrow{a.s.}$ an internally chain transitive invariant set of the differential inclusion

$$\dot{\theta}(t) \in \hat{h}(\theta(t)),$$

using previous three lemmas □

Lemma 3: $\dot{\tilde{\theta}}(t) = \tilde{h}(\tilde{\theta}(t), \lambda(\tilde{\theta}(t)), \tilde{\mu}(t))$

Lemma 2: $\tilde{\mu}(t) \in D^{(1)}(\tilde{\theta}(t), \lambda(\tilde{\theta}(t)))$

Final D.I. $\dot{\tilde{\theta}}(t) \in \hat{h}(\tilde{\theta}(t)), \hat{h}(\tilde{\theta}) = \{\tilde{h}(\tilde{\theta}, \lambda(\tilde{\theta}), \nu) : \nu \in D^{(1)}(\tilde{\theta}, \lambda(\tilde{\theta}))\}$

Novelty w.r.t the single timescale and 2 timescale analysis of Borkar

- The analysis is done under verifiable assumptions whereas some of the assumptions in Borkar's analysis is hard to verify.
- $\lambda(.)$ is a local attractor.
- Made the Lipschitz constant in the vector field depend on the state space.

Empirical Analysis of several off-policy learning algorithm

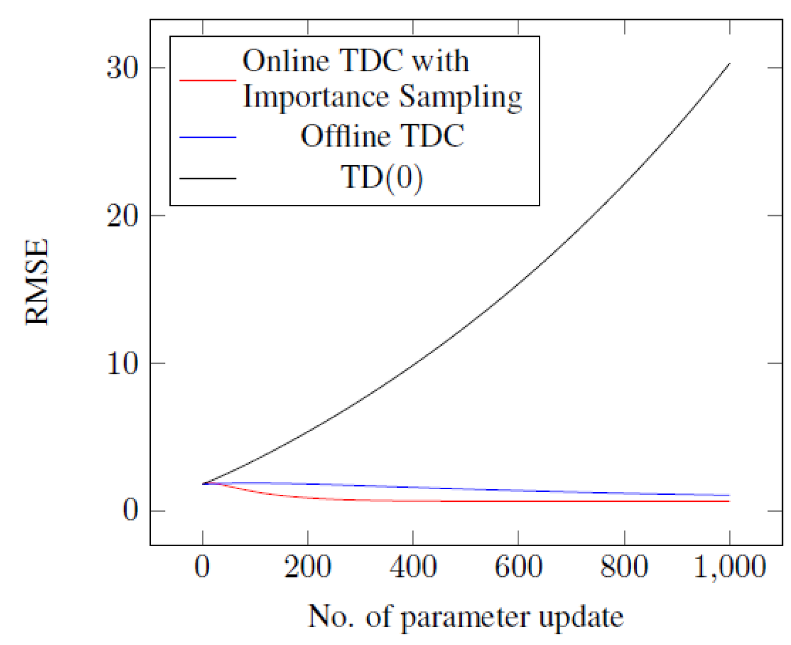


Figure 1: Comparison between TD(0), OFFTDC and ONTDC for Baird's counterexample

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_0 calculate lower bound of

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

for a compact $\bar{B} \subset G$ 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.
- 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 reinorcement learning (RL) [1]

- Risk-sensitive cost:

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

- 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

[1] P.Karmakar and S.Bhatnagar. A note on the function approximation error bound for risk-sensitive reinforcement learning. *https://arxiv.org/abs/1612.07562*.

[2] P.Karmakar and S.Bhatnagar. Two Time-scale Stochastic Approximation with Controlled Markov noise and Off-policy Temporal Difference Learning. *Mathematics of Operations Research (accepted)*, 2017.

[3] P.Karmakar, S.Bhatnagar, and A.Ramaswamy. Dynamics of stochastic approximation with Markov iterate-dependent noise with the stability of the iterates not ensured. *https://arxiv.org/abs/1601.02217*.

Analysis of Stochastic Approximation with Markov Noise and applications

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

7th April, 2017

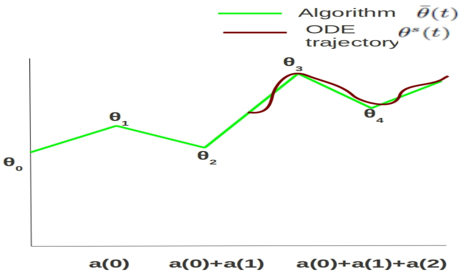
Presentation Outline

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 η unavailable.
  - But, simulated i.i.d samples $\eta_n, n \geq 1$ of η 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



For any $T > 0$, $\sup_{t\in[s,s+T]} \|\bar{\theta}(t) - \theta^s(t)\| \to 0$, *a.s. as* $s \to \infty$.

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

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.

[1] V.S.Borkar. *Stochastic Approximation : A Dynamic Systems Viewpoint. Cambridge University Press. 2008*

2 timescale stochastic approximation

- $$\text{(slow) } \theta_{n+1} = \theta_n + a(n)h(\theta_n, w_n, \eta_n^{(1)}),$$
  $$\text{(fast) } w_{n+1} = w_n + b(n)g(\theta_n, w_n, \eta_n^{(2)})$$
- $\frac{a(n)}{b(n)} \to 0$ 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. $\sup_n(\|\theta_n\| + \|w_n\|) < \infty$ 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

[3] P.Karmakar, S.Bhatnagar and A. Ramaswamy https://arxiv.org/abs/1601.02217

[4] P.Karmakar and S.Bhatnagar https://arxiv.org/abs/1612.07562

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 π_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]

$$\theta_{n+1} = \theta_n + a(n)\rho_n \left[\delta_n(\theta_n)\phi_n - \gamma\phi'_n\phi_n^T w_n\right],$$
$$w_{n+1} = w_n + b(n)\left[(\rho_n\delta_n(\theta_n) - \phi_n^T w_n)\phi_n\right]$$

$$\phi_n = \phi(S_n), \phi'_n = \phi(S_{n+1}), \delta_n(\theta) = R_n + \gamma\theta^T\phi'_n - \theta^T\phi_n, \rho_n = \frac{\pi(A_n|S_n)}{\pi_b(A_n|S_n)}$$

- We analyze in 2-timescale to make sure that the o.d.e's have globally asymptotically stable equilibrium.

[5] H. R. Maei. 2011. *Gradient temporal-difference learning algorithms.* University of Alberta.

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

Function approximation error bound for risk-sensitive RL

- Risk-sensitive cost:

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

- 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.

[6] A. Basu, T. Bhattacharya, V.S.Borkar. 2008. *A Learning Algorithm for Risk-Sensitive Cost* . Mathematics of Operations Research 33(4) 880-898.

Thank You.
Questions ?