

Gradient Temporal Difference with Momentum: Stability and Convergence

Rohan Deb, Shalabh Bhatnagar

Department of Computer Science and Automation,
 Indian Institute of Science,
 Bangalore
 rohandeb@iisc.ac.in, shalabh@iisc.ac.in

Abstract

Gradient temporal difference (Gradient TD) algorithms are a popular class of stochastic approximation (SA) algorithms used for policy evaluation in reinforcement learning. Here, we consider Gradient TD algorithms with an additional heavy ball momentum term and provide choice of step size and momentum parameter that ensures almost sure convergence of these algorithms asymptotically. In doing so, we decompose the heavy ball Gradient TD iterates into three separate iterates with different step sizes. We first analyze these iterates under one-timescale SA setting using results from current literature. However, the one-timescale case is restrictive and a more general analysis can be provided by looking at a three-timescale decomposition of the iterates. In the process we provide the first conditions for stability and convergence of general three-timescale SA. We then prove that the heavy ball Gradient TD algorithm is convergent using our three-timescale SA analysis. Finally, we evaluate these algorithms on standard RL problems and report improvement in performance over the vanilla algorithms.

1 Introduction

In reinforcement learning (RL), the goal of the learner or the agent is to maximize its long term accumulated reward by interacting with the environment. One important task in most of RL algorithms is that of *policy evaluation*. It predicts the average accumulated reward an agent would receive from a state (called *value function*) if it follows the given policy. In *model-free learning*, the agent does not have access to the underlying dynamics of the environment and has to learn the *value function* from samples of the form (state, action, reward, next-state). Two very popular algorithms in the *model-free* setting are *Monte-Carlo* (MC) and *temporal difference* (TD) learning (see Sutton and Barto (2018), Sutton (1988)). It is a well known fact that TD learning diverges in the off-policy setting (see Baird (1995)). A class of algorithms called *gradient temporal difference* (Gradient TD) were introduced in (Sutton, Maei, and Szepesvári 2009) and (Sutton et al. 2009) which are convergent even in the off-policy setting. These algorithms fall under a larger class of algorithms called linear stochastic approximation (SA) algorithms.

A lot of literature is dedicated to studying the asymptotic behaviour of SA algorithms starting from the work of (Robbins and Monro 1951). In recent times, the ODE method to analyze asymptotic behaviour of SA (Ljung 1977; Kushner and Clark 1978; Borkar 2008; Borkar and Meyn 2000) has become quite popular in the RL community. The Gradient TD methods were shown to be convergent using the ODE approach. A generic one-timescale (One-TS) SA iterate has the following form:

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

where $x \in \mathbb{R}^{d_1}$ are the iterates. The function $h : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_1}$ is assumed to be Lipschitz continuous. M_{n+1} is a Martingale difference noise sequence and $a(n)$ is the step-size at time-step n . Under some mild assumptions, the iterate given by (1) converges (see Borkar 2008; Borkar and Meyn 2000). When h is a linear map of the form $b - Ax_n$, the matrix A is often called the driving matrix. The three Gradient TD algorithms: GTD (Sutton, Maei, and Szepesvári 2009), GTD2 and TDC (Sutton et al. 2009) consist two iterates of the following form:

$$x_{n+1} = x_n + a(n)(h(x_n, y_n) + M_{n+1}^{(1)}), \quad (2)$$

$$y_{n+1} = y_n + b(n)(g(x_n, y_n) + M_{n+1}^{(2)}), \quad (3)$$

where $x_n \in \mathbb{R}^{d_1}$, $y_n \in \mathbb{R}^{d_2}$, $\forall n \geq 0$. See Section 2 for exact form of the iterates. The two iterates still form a One-TS SA scheme if $\lim_{n \rightarrow \infty} \frac{b(n)}{a(n)} = c$, where c is a constant and a two-timescale (Two-TS) scheme if $\lim_{n \rightarrow \infty} \frac{b(n)}{a(n)} = 0$.

Separately, adding a momentum term to accelerate the convergence of iterates is a popular technique in stochastic gradient descent (SGD). The two most popular schemes are the Polyak's Heavy ball method (Polyak 1964), and Nesterov's accelerated gradient method (Nesterov 1983). A lot of literature is dedicated to studying momentum with SGD. Some recent works include (Ghadimi, Feyzmahdavian, and Johansson 2014; Loizou and Richtárik 2020; Gitman et al. 2019; Ma and Yarats 2019; Assran and Rabbat 2020). Momentum in the SA setting, which is the focus of the current work, has limited results. Very few works study the effect of momentum in the SA setting. A recent work by (Mou et al. 2020) studies SA with momentum briefly and shows an improvement of mixing rate. However, the setting considered is restricted to linear SA and the driving matrix is assumed

to be symmetric. Further, the iterates involve an additional Polyak-Ruppert averaging (Polyak 1990). Here, in contrast, we analyze the asymptotic behaviour of the algorithm and make none of the above assumptions. A somewhat distant paper is by (Devraj, Bušić, and Meyn 2019) that introduces Matrix momentum in SA and is not equivalent to heavy ball momentum.

A very recent work by (Avrachenkov, Patil, and Thoppe 2020) studied One-TS SA with heavy ball momentum in the univariate case (i.e., $d = 1$ in iterate (1)) in the context of web-page crawling. The iterates took the following form:

$$x_{n+1} = x_n + a(n)(h(x_n) + M_{n+1}) + \eta_n(x_n - x_{n-1}). \quad (4)$$

The momentum parameter η_n was chosen to decompose the iterate into two recursions of the form given by (2) and (3). We use such a decomposition for Gradient TD methods with momentum. This leads to three separate iterates with three step-sizes. We analyze these three iterates and provide stability (iterates remain bounded throughout) and almost sure (a.s.) convergence guarantees.

1.1 Our Contribution

- We first consider the One-TS decomposition of Gradient TD with momentum iterates and show that the driving matrix is Hurwitz (all eigen values are negative). Thereafter we use the theory of One-TS SA to show that the iterates are stable and convergent to the same TD solution.
- Next, we consider the Three-TS decomposition. We provide the first stability and convergence conditions for general Three-TS recursions. We then show that the iterates under consideration satisfy these conditions.
- Finally, we evaluate these algorithms for different choice of step-size and momentum parameters on standard RL problems and report an improvement in performance over their vanilla counterparts.

2 Preliminaries

In the standard RL setup, an agent interacts with the environment which is a Markov Decision Process (MDP). At each discrete time step t , the agent is in state $s_t \in \mathcal{S}$, takes an action $a_t \in \mathcal{A}$, receives a reward $r_{t+1} \equiv r(s_t, a_t, s_{t+1}) \in \mathbb{R}$ and moves to another state $s_{t+1} \in \mathcal{S}$. Here \mathcal{S} and \mathcal{A} are finite sets of possible states and actions respectively. The transitions are governed by a kernel \mathbb{P} . A policy $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ is a mapping that defines the probability of picking an action in a state. We let $P^\pi(s'|s)$ be the transition probability matrix induced by π . Also, $\{d^\pi(s)\}_{s \in \mathcal{S}}$ represents the steady-state distribution for the Markov chain induced by π and the matrix D is a diagonal matrix of dimension $n \times n$ with the entries $d^\pi(s)$ on its diagonals. The state-value function associated with a policy π for state s is

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s \right],$$

where $\gamma \in [0, 1)$ is the discount factor.

In the linear architecture setting, *policy evaluation* deals with estimating $V^\pi(s)$ through a linear model $V_\theta(s) =$

$\theta^T \phi(s)$, where $\phi(s) \equiv \phi_s$ is a feature associated with the state s and θ is the parameter vector. We define the TD-error as $\delta_t = r_{t+1} + \gamma \theta_t^T \phi_{t+1} - \theta_t^T \phi_t$ and Φ as an $n \times d$ matrix where the s^{th} row is $\phi(s)^T$. In the i.i.d setting it is assumed that the tuple (ϕ_t, ϕ'_t) (where $\phi_{t+1} \equiv \phi'_t$) is drawn independently from the stationary distribution of the Markov chain induced by π . Let $\bar{A} = \mathbb{E}[\phi_t(\gamma \phi'_t - \phi_t)^T]$ and $\bar{b} = \mathbb{E}[r_{t+1} \phi_t]$, where the expectations are w.r.t. the stationary distribution of the induced chain. The matrix \bar{A} is negative definite (see Maei (2011); Tsitsiklis and Van Roy (1997)). In the off-policy case, the importance weight is given by $\rho_t = \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$, where π and μ are the target and behaviour policies respectively. Introduced in (Sutton, Maei, and Szepesvári 2009), Gradient TD are a class of TD algorithms that are convergent even in the off-policy setting. Next, we present the iterates associated with the algorithms GTD (Sutton, Maei, and Szepesvári 2009), GTD2, TDC (Sutton et al. 2009).

1. GTD:

$$\theta_{t+1} = \theta_t + \alpha_t(\phi_t - \gamma \phi'_t) \phi_t^T u_t, \quad (5)$$

$$u_{t+1} = u_t + \beta_t(\delta_t \phi_t - u_t). \quad (6)$$

2. GTD2:

$$\theta_{t+1} = \theta_t + \alpha_t(\phi_t - \gamma \phi'_t) \phi_t^T u_t, \quad (7)$$

$$u_{t+1} = u_t + \beta_t(\delta_t - \phi_t^T u_t) \phi_t. \quad (8)$$

3. TDC:

$$\theta_{t+1} = \theta_t + \alpha_t \delta_t \phi_t - \alpha_t \gamma \phi'_t (\phi_t^T u_t), \quad (9)$$

$$u_{t+1} = u_t + \beta_t(\delta_t - \phi_t^T u_t) \phi_t. \quad (10)$$

The objective function for GTD is Norm of Expected Error defined as $NEU(\theta) = \mathbb{E}[\delta \phi]$. The GTD algorithm is derived by expressing the gradient direction as $-\frac{1}{2} \nabla NEU(\theta) = \mathbb{E}[(\phi - \gamma \phi') \phi^T] \mathbb{E}[\delta(\theta) \phi]$. Here $\phi' \equiv \phi(s')$. If both the expectations are sampled together, then the term would be biased by their correlation. An estimate of the second expectation is maintained as a long-term quasi-stationary estimate (see (5)) and the first expectation is sampled (see (6)). For GTD2 and TDC, a similar approach is used on the objective function Mean Square Projected Bellman Error defined as $MSPBE(\theta) = \|V_\theta - \Pi T^\pi V_\theta\|_D$. Here, Π is the projection operator that projects vectors to the subspace $\{\Phi \theta | \theta \in \mathbb{R}^d\}$ and T^π is the Bellman operator defined as $T^\pi V = R^\pi + \gamma P^\pi V$. As originally presented, GTD and GTD2 are one-timescale algorithms ($\frac{\alpha_t}{\beta_t}$ is constant) while TDC is a two-timescale algorithm ($\frac{\alpha_t}{\beta_t} \rightarrow 0$). It was shown in all the three cases that $\theta_n \rightarrow \theta^* = -\bar{A}^{-1} \bar{b}$.

3 Gradient TD with Momentum

Although, Gradient TD starts with a gradient descent based approach, it ends up with Two-TS SA recursions. Momentum methods are known to accelerate the convergence of SGD iterates. Motivated by this, we examine momentum in the SA setting, and ask if the SA recursions for Gradient TD with momentum even converge to the same TD solution. We probe the heavy ball extension of the three Gradient TD algorithms

where, we keep an accumulation of the previous gradient values in ζ_t . Then, at time step $t + 1$ the new gradient value multiplied by the step size is added to the current accumulation vector ζ_t multiplied by the momentum parameter η_t as below:

$$\zeta_{t+1} = \eta_t \zeta_t + \alpha_t (\phi_t - \gamma \phi'_t) \phi_t^T u_t.$$

The parameter θ is then updated in the negative of the direction ζ_{t+1} , i.e., $\theta_{t+1} = \theta_t - \zeta_{t+1}$. Since u_{t+1} is computed as a long-term estimate of $\mathbb{E}[\delta(\theta)\phi]$, its update rule remains same. The momentum parameter η_t is usually set to a constant in the stochastic gradient setting. An exception to this can however be found in (Gitman et al. 2019; Gadat, Panloup, and Saadane 2016), where $\eta_t \rightarrow 1$. Here, we consider the latter case. Substituting ζ_{t+1} into the iteration of θ_{t+1} and noting that $\zeta_t = \theta_t - \theta_{t-1}$, the iterates for GTD with Momentum (**GTD-M**) can be written as:

$$\theta_{t+1} = \theta_t + \alpha_t (\phi_t - \gamma \phi'_t) \phi_t^T u_t + \eta_t (\theta_t - \theta_{t-1}), \quad (11)$$

$$u_{t+1} = u_t + \beta_t (\delta_t \phi_t - u_t). \quad (12)$$

Similarly the iterates for **GTD2-M** are given by:

$$\theta_{t+1} = \theta_t + \alpha_t (\phi_t - \gamma \phi'_t) \phi_t^T u_t + \eta_t (\theta_t - \theta_{t-1}), \quad (13)$$

$$u_{t+1} = u_t + \beta_t (\delta_t - \phi_t^T u_t) \phi_t. \quad (14)$$

Finally, the iterates for **TDC-M** are given by:

$$\theta_{t+1} = \theta_t + \alpha_t (\delta_t \phi_t - \gamma \phi'_t) \phi_t^T u_t + \eta_t (\theta_t - \theta_{t-1}), \quad (15)$$

$$u_{t+1} = u_t + \beta_t (\delta_t - \phi_t^T u_t) \phi_t. \quad (16)$$

We choose the momentum parameter η_t as in (Avrachenkov, Patil, and Thoppe 2020) as follows: $\eta_t = \frac{\varrho_t - w\alpha_t}{\varrho_{t-1}}$, where $\{\varrho_t\}$ is a positive sequence s.t. $\varrho_t \rightarrow 0$ as $t \rightarrow \infty$ and $w \in \mathbb{R}$ is a constant. Note that $\eta_t \rightarrow 1$ as $t \rightarrow \infty$. We later provide conditions on ϱ_t and w to ensure a.s. convergence. As we would see in section 4, the condition on w in the One-TS setting is restrictive. Specifically, it depends on the norm of the driving matrix \bar{A} . This motivates us to look at the Three-TS setting and then the corresponding condition on w is less restrictive. Using the momentum parameter as above,

$$\theta_{t+1} = \theta_t + \alpha_t (\phi_t - \gamma \phi'_t) \phi_t^T u_t + \frac{\varrho_t - w\alpha_t}{\varrho_{t-1}} (\theta_t - \theta_{t-1})$$

Rearranging the terms and dividing by ϱ_t , we get:

$$\begin{aligned} \frac{\theta_{t+1} - \theta_t}{\varrho_t} &= \frac{\theta_t - \theta_{t-1}}{\varrho_{t-1}} \\ &+ \frac{\alpha_t}{\varrho_t} \left((\phi_t - \gamma \phi'_t) \phi_t^T u_t - w \left(\frac{\theta_t - \theta_{t-1}}{\varrho_{t-1}} \right) \right). \end{aligned}$$

We let $\frac{\theta_{t+1} - \theta_t}{\varrho_t} = v_{t+1}$, $\xi_t = \frac{\alpha_t}{\varrho_t}$ and $\varepsilon_t = v_{t+1} - v_t$. Then, the GTD-M iterates in (11) and (12) can be re-written with the following three iterates:

$$v_{t+1} = v_t + \xi_t ((\phi_t - \gamma \phi'_t) \phi_t^T u_t - w v_t) \quad (17)$$

$$u_{t+1} = u_t + \beta_t (\delta_t \phi_t - u_t) \quad (18)$$

$$\theta_{t+1} = \theta_t + \varrho_t (v_t + \varepsilon_t) \quad (19)$$

A similar decomposition can be done for the GTD2-M and TDC-M iterates.

The work by (Devraj, Bušić, and Meyn 2019) come closest to ours in the sense that they look at momentum in stochastic approximation. Although motivated from Polyak and Nesterov's momentum schemes they explore a different direction where instead of a scalar momentum parameter, they consider a matrix momentum parameter.

4 Convergence Analysis

In this section we analyze the asymptotic behaviour of the GTD-M iterates given by (17), (18) and (19). Throughout the section, we consider $v_t, u_t, \theta_t \in \mathbb{R}^d$. We first consider the One-TS case when $\beta_t = c_1 \xi_t$ and $\varrho_t = c_2 \xi_t \forall t$, for some real constants $c_1, c_2 > 0$. Subsequently, we consider the Three-TS setting where $\frac{\beta_t}{\xi_t} \rightarrow 0$ and $\frac{\varrho_t}{\beta_t} \rightarrow 0$ as $t \rightarrow \infty$.

4.1 One-Timescale Setting

We begin by analyzing GTD-M using a one-timescale SA setting. We let $c_1 = c_2 = 1$ for simplicity. The iterates of GTD-M can then be re-written as:

$$\psi_{t+1} = \psi_t + \xi_t (G_t \psi_t + g_t + \varepsilon_t), \quad (20)$$

where,

$$\psi_t = \begin{pmatrix} v_t \\ u_t \\ \theta_t \end{pmatrix}, g_t = \begin{pmatrix} 0 \\ r_{t+1} \phi_t \\ 0 \end{pmatrix}, \bar{\varepsilon}_t = \begin{pmatrix} 0 \\ 0 \\ \varepsilon_t \end{pmatrix},$$

$$G_t = \begin{pmatrix} -wI & (\phi_t - \gamma \phi'_t) \phi_t^T & 0 \\ 0 & -I & \phi_t (\gamma \phi'_t - \phi_t)^T \\ I & 0 & 0 \end{pmatrix}.$$

Equation (20) can be re-written in the general SA scheme as:

$$\psi_{t+1} = \psi_t + \xi_t (h(\psi_t) + M_{t+1} + \bar{\varepsilon}_t). \quad (21)$$

Here $h(\psi) = g + G\psi$, $g = \mathbb{E}[g_t]$, $G = \mathbb{E}[G_t]$, where the expectations are w.r.t. the stationary distribution of the Markov chain induced by the target policy π . $M_{t+1} = (G_{t+1} - G)\psi_t + (g_{t+1} - g)$. In particular,

$$G = \begin{pmatrix} -wI & -\bar{A}^T & 0 \\ 0 & -I & \bar{A} \\ I & 0 & 0 \end{pmatrix}, g = \begin{pmatrix} 0 \\ \bar{b} \\ 0 \end{pmatrix},$$

where recall that $\bar{A} = \mathbb{E}[\phi(\gamma \phi' - \phi)^T]$ and $\bar{b} = \mathbb{E}[r\phi]$

Lemma 1. Assume $w(w+1) > \|\bar{A}\|^2$. Then G is Hurwitz.

Proof. Let λ be an eigenvalue of G . The characteristic equation of the matrix G is given by:

$$\begin{vmatrix} -wI - \lambda I & -\bar{A}^T & 0 \\ 0 & -I - \lambda I & \bar{A} \\ I & 0 & -\lambda I \end{vmatrix} = 0$$

$$\begin{vmatrix} wI + \lambda I & \bar{A}^T & 0 \\ 0 & I + \lambda I & -\bar{A} \\ -I & 0 & \lambda I \end{vmatrix} = 0$$

Using the following formula for determinant of block matrices

$$\begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{vmatrix} = |A_{11}| \left| \begin{pmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{pmatrix} - \begin{pmatrix} A_{21} \\ A_{31} \end{pmatrix} A_{11}^{-1} \begin{pmatrix} A_{12} & A_{13} \end{pmatrix} \right|$$

we have,

$$\begin{aligned}
& \left| \begin{matrix} wI + \lambda I & \bar{A}^T & 0 \\ 0 & I + \lambda I & -\bar{A} \\ -I & 0 & \lambda I \end{matrix} \right| = \\
& |(w + \lambda)I| \left| \begin{pmatrix} I + \lambda I & -\bar{A} \\ 0 & \lambda I \end{pmatrix} - \frac{1}{w+\lambda} \begin{pmatrix} 0 \\ -I \end{pmatrix} (\bar{A}^T \quad 0) \right| \\
& = (w + \lambda)^d \left| \begin{matrix} I + \lambda I & -\bar{A} \\ \frac{\bar{A}^T}{w+\lambda} & \lambda I \end{matrix} \right| \\
& = (w + \lambda)^d |(1 + \lambda)I| \left| \lambda I + \frac{1}{(1+\lambda)(w+\lambda)} \bar{A}^T \bar{A} \right| \\
& = \frac{(w + \lambda)^d (1 + \lambda)^d}{(w + \lambda)^d (1 + \lambda)^d} |\lambda(1 + \lambda)(w + \lambda)I + \bar{A}^T \bar{A}| \\
& = |\lambda(1 + \lambda)(w + \lambda)I + \bar{A}^T \bar{A}|
\end{aligned}$$

Therefore, from the characteristic equation of G , we have that

$$|\lambda(1 + \lambda)(w + \lambda)I + \bar{A}^T \bar{A}| = 0.$$

There must exist a non-zero vector $x \in \mathbb{C}^d$, such that $x^*(\lambda(1 + \lambda)(w + \lambda)I + \bar{A}^T \bar{A})x = 0$, where x^* is the conjugate transpose of the vector x and $x^*x = \|x\|^2 > 0$. The above equation reduces to the following cubic-polynomial equation: $\lambda^3 \|x\|^2 + (w+1)\lambda^2 \|x\|^2 + w\lambda \|x\|^2 + \|\bar{A}x\|^2 = 0$, where $\|\bar{A}x\|^2 = x^* \bar{A}^T \bar{A}x$. Using Routh-Hurwitz criterion, a cubic polynomial $a_3\lambda^3 + a_2\lambda^2 + a_1\lambda + a_0$ has all roots with negative real parts iff $a_3, a_2, a_1, a_0 > 0$ and $a_1a_2 > a_0a_3$. In our case, $a_3 = \|x\|^2 > 0, a_2 = (w+1)\|x\|^2 > 0, a_1 = w\|x\|^2 > 0$ and $a_0 = \|\bar{A}x\|^2 > 0$. The last inequality follows from the fact that \bar{A} is negative definite and therefore $x^* \bar{A}^T \bar{A}x > 0$. Finally, $a_1a_2 = w(w+1)\|x\|^4, a_0a_3 = \|x\|^2 \|\bar{A}x\|^2$ and $a_1a_2 > a_0a_3$ follows from $\frac{\|\bar{A}x\|^2}{\|x\|^2} < \|\bar{A}\|^2 < w(w+1)$. Therefore $\text{Re}(\lambda) < 0$ and the claim follows. \square

Consider the following assumptions:

A1. All rewards and features satisfy $|r(s, s')| \leq 1$ and $\|\phi(s)\| \leq 1 \forall s, s' \in \mathcal{S}$. Also, the matrix Φ has full rank, where Φ is an $n \times d$ matrix where the s^{th} row is $\phi(s)^T$.

A2. The step-sizes satisfy $\xi_t = \beta_t = \varrho_t > 0$,

$$\sum_t \xi_t = \infty \sum_t \xi_t^2 < \infty, \text{ where } \xi_t = \frac{\alpha_t}{\varrho_t}$$

and the momentum parameter satisfies: $\eta_t = \frac{\varrho_t - w\alpha_t}{\varrho_{t-1}}$.

A3. The samples (ϕ_t, ϕ'_t) are drawn i.i.d from the stationary distribution of the Markov chain induced by target policy π .

Theorem 2. Assume **A1**, **A2** and **A3** hold and let $w \geq 1$. Then, the GTD-M iterates given by (11) and (12) satisfy $\theta_n \rightarrow \theta^* = -\bar{A}^{-1}\bar{b}$ a.s. as $n \rightarrow \infty$.

Proof. Assumption **A1** ensures that $\|\bar{A}\|^2 < w(w+1)$ and **A3** ensures that the function $h(\cdot)$ is well defined. Now, using Lemma 1 and (Borkar and Meyn 2000) we can show that the iterates in (20) remain stable. Then using the third extension from (Chapter-2 pp. 17, Borkar (2008)) we can show that $\psi_n \rightarrow -G^{-1}g$ as $n \rightarrow \infty$. Thereafter using the formula for inverse of block matrices it can be shown that $\theta_n \rightarrow -\bar{A}^{-1}\bar{b}$

as $n \rightarrow \infty$. See Appendix A1 in (Deb and Bhatnagar 2021) for a detailed proof. \square

Similar results can be proved for the GTD2-M and TDC-M.

Remark 1. If w is large, the initial values of the momentum parameter is small. The condition on w in lemma 1 is large compared to the condition on w in (Avrachenkov, Patil, and Thoppe 2020), where the condition is $w > 0$. Motivated by this, we look at the three-TS case of the iterates.

4.2 Three Timescale Setting

We consider the three iterates for GTD-M in (17), (18) and (19) under the following criteria for step-sizes: $\frac{\xi_t}{\beta_t} \rightarrow 0$ and $\frac{\varrho_t}{\xi_t} \rightarrow 0$ as $t \rightarrow \infty$. We provide the first conditions for stability and a.s. convergence of generic three-TS SA recursions. We emphasize that the setting we look at in Theorem 3 is more general than the setting at hand of GTD-M iterates. Although stability and convergence results exist for One-TS and Two-TS cases, this is the first time such results have been provided for the case of three-TS recursions. We next provide the general iterates for a three-TS recursion along with the assumptions used while analyzing them. Consider the following three iterates:

$$x_{n+1} = x_n + a(n) \left(h(x_n, y_n, z_n) + M_{n+1}^{(1)} + \varepsilon_n^{(1)} \right), \quad (22)$$

$$y_{n+1} = y_n + b(n) \left(g(x_n, y_n, z_n) + M_{n+1}^{(2)} + \varepsilon_n^{(2)} \right), \quad (23)$$

$$z_{n+1} = z_n + c(n) \left(f(x_n, y_n, z_n) + M_{n+1}^{(3)} + \varepsilon_n^{(3)} \right), \quad (24)$$

and the following assumptions:

(B1) $h : \mathbb{R}^{d_1+d_2+d_3} \rightarrow \mathbb{R}^{d_1}, g : \mathbb{R}^{d_1+d_2+d_3} \rightarrow \mathbb{R}^{d_2}, f : \mathbb{R}^{d_1+d_2+d_3} \rightarrow \mathbb{R}^{d_3}$ are Lipchitz continuous, with Lipchitz constants L_1, L_2 and L_3 respectively.

(B2) $\{a(n)\}, \{b(n)\}, \{c(n)\}$ are step-size sequences that satisfy $a(n) > 0, b(n) > 0, c(n) > 0, \forall n > 0$,

$$\begin{aligned}
\sum_n a(n) &= \sum_n b(n) = \sum_n c(n) = \infty, \\
\sum_n (a(n)^2 + b(n)^2 + c(n)^2) &< \infty, \\
\frac{b(n)}{a(n)} &\rightarrow 0, \frac{c(n)}{b(n)} \rightarrow 0 \text{ as } n \rightarrow \infty.
\end{aligned}$$

(B3) $\{M_n^{(1)}\}, \{M_n^{(2)}\}, \{M_n^{(3)}\}$ are Martingale difference sequences w.r.t. the filtration $\{\mathcal{F}_n\}$ where,

$$\begin{aligned}
\mathcal{F}_n &= \sigma \left(x_m, y_m, z_m, M_m^{(1)}, M_m^{(2)}, M_m^{(3)}, m \leq n \right) \\
\mathbb{E} \left[\|M_{n+1}^{(i)}\|^2 | \mathcal{F}_n \right] &\leq K_i (1 + \|x_n\|^2 + \|y_n\|^2 + \|z_n\|^2); \\
\forall n \geq 0, i = 1, 2, 3 \text{ and constants } 0 < K_i < \infty. \text{ The terms } \varepsilon_t^{(i)} &\text{ satisfy } \|\varepsilon_n^{(1)}\| + \|\varepsilon_n^{(2)}\| + \|\varepsilon_n^{(3)}\| \rightarrow 0 \text{ as } n \rightarrow \infty.
\end{aligned}$$

- (B4)**
- (i) The ode $\dot{x}(t) = h(x(t), y, z), y \in \mathbb{R}^{d_2}, z \in \mathbb{R}^{d_3}$ has a globally asymptotically stable equilibrium (g.a.s.e) $\lambda(y, z)$, and $\lambda : \mathbb{R}^{d_2 \times d_3} \rightarrow \mathbb{R}^{d_1}$ is Lipchitz continuous.
 - (ii) The ode $\dot{y}(t) = g(\lambda(y(t), z), y(t), z), z \in \mathbb{R}^{d_3}$ has a globally asymptotically stable equilibrium $\Gamma(z)$, where $\Gamma : \mathbb{R}^{d_3} \rightarrow \mathbb{R}^{d_2}$ is Lipchitz continuous.

- (iii) The ode $\dot{z}(t) = f(\lambda(\Gamma(z(t)), z(t)), \Gamma(z(t)), z(t))$, has a globally asymptotically stable equilibrium $z^* \in \mathbb{R}^{d_3}$.
- (B5)** The functions $h_c(x, y, z) = \frac{h(cx, cy, cz)}{c}$, $c \geq 1$ satisfy $h_c \rightarrow h_\infty$ as $c \rightarrow \infty$ uniformly on compacts. The ODE: $\dot{x}(t) = h_\infty(x(t), y, z)$, has a unique globally asymptotically stable equilibrium $\lambda_\infty(y, z)$, where $\lambda_\infty : \mathbb{R}^{d_2+d_3} \rightarrow \mathbb{R}^{d_1}$ is Lipschitz continuous. Further, $\lambda_\infty(0, 0) = 0$.
- (B6)** The functions $g_c(y, z) = \frac{g(c\lambda_\infty(y, z), cy, cz)}{c}$, $c \geq 1$ satisfy $g_c \rightarrow g_\infty$ as $c \rightarrow \infty$ uniformly on compacts. The ODE: $\dot{y}(t) = g_\infty(y(t), z)$, has a unique globally asymptotically stable equilibrium $\Gamma_\infty(z)$, where $\Gamma_\infty : \mathbb{R}^{d_3} \rightarrow \mathbb{R}^{d_2}$ is Lipschitz continuous. Further, $\Gamma_\infty(0) = 0$.
- (B7)** The functions $f_c(z) = \frac{g(c\lambda_\infty(\Gamma_\infty(z), z), c\Gamma_\infty(z), cz)}{c}$, $c \geq 1$ satisfy $f_c \rightarrow f_\infty$ as $c \rightarrow \infty$ uniformly on compacts. The ODE: $\dot{z}(t) = f_\infty(z(t))$, has the origin in \mathbb{R}^{d_3} as its unique globally asymptotically stable equilibrium.

Remark 2. Conditions **(B5)-(B7)** (along with **(B1)-(B3)**) give sufficient conditions that ensure that the iterates remain stable. Specifically it ensures that $\sup_n(\|x_n\| + \|y_n\| + \|z_n\|) < \infty$ a.s. Conditions **(B1)-(B4)** along with the stability of iterates ensures a.s. convergence of the iterates.

Theorem 3. Under assumptions **(B1)-(B7)**, the iterates given by (22) satisfy (23) and (24),

$$(x_n, y_n, z_n) \rightarrow (\lambda(\Gamma(z^*), z^*), \Gamma(z^*), z^*) \text{ as } n \rightarrow \infty$$

Proof. See Appendix A2 in (Deb and Bhatnagar 2021). \square

Next we use theorem 3, to show that the iterates of GTD-M a.s. converge to the TD solution $-\bar{A}^{-1}\bar{b}$. Consider the following assumption on step-size sequences instead of **A2**.

A4. The step-sizes satisfy $\xi_t > 0, \beta_t > 0, \varrho_t > 0 \forall t$,

$$\sum_t \xi_t = \sum_t \beta_t = \sum_t \varrho_t = \infty, \sum_t (\xi_t^2 + \beta_t^2 + \varrho_t^2) < \infty,$$

$$\frac{\beta_t}{\xi_t} \rightarrow 0, \frac{\varrho_t}{\beta_t} \rightarrow 0 \text{ as } t \rightarrow \infty$$

and the momentum parameter satisfies: $\eta_t = \frac{\varrho_t - w\alpha_t}{\varrho_{t-1}}$.

Theorem 4. Assume **A1**, **A3** and **A4** hold and let $w > 0$. Then, the GTD-M iterates given by (11) and (12) satisfy $\theta_n \rightarrow \theta^* = -\bar{A}^{-1}\bar{b}$ a.s. as $n \rightarrow \infty$.

Proof. We transform (17), (18) and (19) into the standard SA form given by (22), (23) and (24). Let $\mathcal{F}_t = \sigma(u_0, v_0, \theta_0, r_{j+1}, \phi_j, \phi'_j : j < t)$. Let, $A_t = \phi_t(\gamma\phi'_t - \phi_t)^T$ and $b_t = r_{t+1}\phi_t$. Then, (17) can be re-written as:

$$v_{t+1} = v_t + \xi_t \left(h(v_t, u_t, \theta_t) + M_{t+1}^{(1)} \right)$$

where, $h(v_t, u_t, \theta_t) = \mathbb{E}[(\phi_t - \gamma\phi'_t)\phi_t^T u_t - wv_t | \mathcal{F}_t] = -A_t^T u_t - wv_t - h(v_t, u_t, \theta_t) = (A_t^T - A_t^T)u_t$. Next, (18) can be re-written as:

$$u_{t+1} = u_t + \beta_t \left(g(v_t, u_t, \theta_t) + M_{t+1}^{(2)} \right)$$

where, $g(v_t, u_t, \theta_t) = \mathbb{E}[\delta_t\phi_t - u_t | \mathcal{F}_t] = \bar{A}\theta_t + \bar{b} - u_t$ and $M_{t+1}^{(2)} = A_t\theta_t + b_t - u_t - g(v_t, u_t, \theta_t) = (A_t - \bar{A})\theta_t + (b_t - \bar{b})$. Finally, (19) can be re-written as:

$$\theta_{t+1} = \theta_t + \varrho_t \left(f(v_t, u_t, \theta_t) + \varepsilon_t + M_{t+1}^{(3)} \right)$$

where, $f(v_t, u_t, \theta_t) = v_t$ and $M_{t+1}^{(3)} = 0$. The functions h, g, f are linear in v, u, θ and hence Lipchitz continuous, therefore satisfying **(B1)**. We choose the step-size sequences such that they satisfy **(B2)**. One popular choice is $\xi_t = \frac{1}{(t+1)\xi}, \beta_t = \frac{1}{(t+1)\beta}, \varrho_t = \frac{1}{(t+1)\varrho}, \frac{1}{2} < \xi < \beta < \varrho \leq 1$. Next, $M_{t+1}^{(1)}, M_{t+1}^{(2)}$ and $M_{t+1}^{(3)}$ $t \geq 0$, are martingale difference sequences w.r.t \mathcal{F}_t by construction. $\mathbb{E}[\|M_{t+1}^{(1)}\|^2 | \mathcal{F}_t] \leq \|(\bar{A}^T - A_t^T)\|^2 \|u_t\|^2, \mathbb{E}[\|M_{t+1}^{(2)}\|^2 | \mathcal{F}_t] \leq 2(\|(A_t - \bar{A})\|^2 \|\theta_t\|^2 + \|(b_t - \bar{b})\|^2)$. The first part of **(B3)** is satisfied with $K_1 = \|(\bar{A}^T - A_t^T)\|^2, K_2 = 2 \max(\|A_t - \bar{A}\|^2, \|b_t - \bar{b}\|^2)$ and any $K_3 > 0$. The fact that $K_1, K_2 < \infty$ follows from the bounded features and bounded rewards assumption in **A1**. Next, observe that $\|\varepsilon_t^{(3)}\| = \xi_t \|((\phi_t - \gamma\phi'_t)\phi_t^T u_t - wv_t)\| \rightarrow 0$ since $\xi_t \rightarrow 0$ as $t \rightarrow \infty$. For a fixed $u, \theta \in \mathbb{R}^d$, consider the ODE $\dot{v}(t) = -\bar{A}^T u - wv(t)$. For $w > 0$, $\lambda(u, \theta) = -\frac{\bar{A}^T u}{w}$ is the unique g.a.s.e, is linear and therefore Lipchitz continuous. This satisfies **(B4)(i)**. Next, for a fixed $\theta \in \mathbb{R}^d$, $\dot{u}(t) = \bar{A}\theta + \bar{b} - u(t)$, has $\Gamma(\theta) = \bar{A}\theta + \bar{b}$ as its unique g.a.s.e and is Lipschitz. This satisfies **(B4)(ii)**. Finally, to satisfy **(B4)(iii)**, consider,

$$\dot{\theta}(t) = \frac{-\bar{A}^T \bar{A}\theta(t) - \bar{A}^T \bar{b}}{w}.$$

Since, \bar{A} is negative definite, therefore, $-\bar{A}^T \bar{A}$ is negative definite. Therefore, $\theta^* = -\bar{A}^{-1}\bar{b}$ is the unique g.a.s.e. Next, we show that the sufficient conditions for stability of the three iterates are satisfied. The function, $h_c(v, u, \theta) = \frac{-c\bar{A}^T u - wcv}{c} = -\bar{A}^T u - wv \rightarrow h_\infty(v, u, \theta) = -\bar{A}^T u - wv$ uniformly on compacts as $c \rightarrow \infty$. The limiting ODE: $\dot{v}(t) = -\bar{A}^T u - wv(t)$ has $\lambda_\infty(u, \theta) = -\frac{\bar{A}^T u}{w}$ as its unique g.a.s.e. λ_∞ is Lipschitz with $\lambda_\infty(0, 0) = 0$, thus satisfying assumption **(B5)**. The function, $g_c(u, \theta) = \frac{c\bar{A}\theta + \bar{b} - cu}{c} = \bar{A}\theta - u + \frac{\bar{b}}{c} \rightarrow g_\infty(u, \theta) = -\bar{A}\theta - u$ uniformly on compacts as $c \rightarrow \infty$. The limiting ODE $\dot{u}(t) = -\bar{A}\theta - u(t)$ has $\Gamma_\infty(\theta) = \bar{A}\theta$ as its unique g.a.s.e. Γ_∞ is Lipchitz with $\Gamma_\infty(0) = 0$. Thus assumption **(B6)** is satisfied. Finally, $f_c(\theta) = \frac{-c\bar{A}^T \bar{A}\theta}{cw} \rightarrow f_\infty = \frac{-\bar{A}^T \bar{A}\theta}{w}$ uniformly on compacts as $c \rightarrow \infty$ and the ODE: $\dot{\theta}(t) = -\frac{\bar{A}^T \bar{A}\theta(t)}{w}$ has origin as its unique g.a.s.e. This ensures the final condition **(B7)**. By theorem 3,

$$\begin{pmatrix} v_t \\ u_t \\ \theta_t \end{pmatrix} \rightarrow \begin{pmatrix} \lambda(\Gamma(-\bar{A}^{-1}\bar{b}), -\bar{A}^{-1}\bar{b}) \\ \Gamma(-\bar{A}^{-1}\bar{b}) \\ -\bar{A}^{-1}\bar{b} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -\bar{A}^{-1}\bar{b} \end{pmatrix}$$

Specifically, $\theta_t \rightarrow -\bar{A}^{-1}\bar{b}$. \square

Similar analysis can be provided for GTD2-M and TDC-M. See Appendix A3 in (Deb and Bhatnagar 2021) for details.

Remark 3. Convergence with importance-weighting: The Gradient TD with momentum iterates with importance weight ρ_t multiplied with the TD error can also be shown to converge along similar lines. The Martingale noise needs to be substituted with a Markov noise and results along the lines of (Ramaswamy and Bhatnagar 2019; Karmakar and Bhatnagar 2018) can be used to show convergence of these schemes.

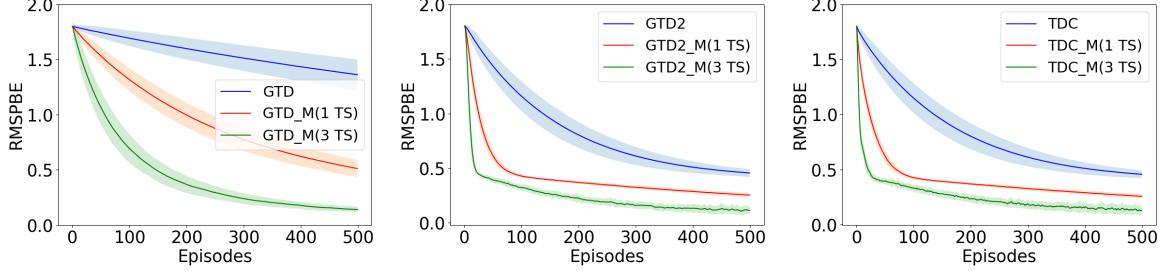


Figure 1: RMSPBE (averaged over 100 independent runs) across episodes for Boyan Chain. The features used are the standard spiked features of size 4 used in Boyan chain (see (Dann, Neumann, and Peters 2014)).

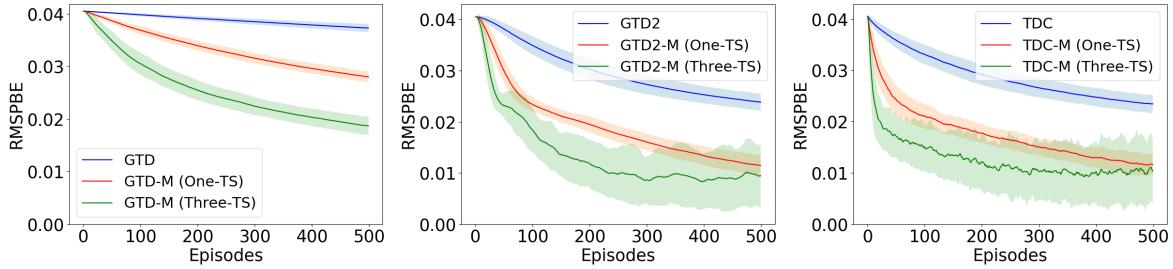


Figure 2: RMSPBE (averaged over 100 independent runs) across episodes for the 5-State Random Chain problem. The features used are the *Dependent* features used in (Sutton et al. 2009).

5 Experiments

We evaluate the momentum based GTD algorithms defined in section 3 to four standard problems of policy evaluation in reinforcement learning namely, Boyan Chain (Boyan 1999), 5-State random walk (Sutton et al. 2009), 19-State Random Walk (Sutton and Barto 2018) and Random MDP (Sutton et al. 2009). See Appendix A4 in (Deb and Bhattacharjee 2021) for a detailed description of the MDP settings and (Dann, Neumann, and Peters 2014) for details on implementation. We run the three algorithms, GTD, GTD2 and TDC along with their heavy ball momentum variants in One-TS and Three-TS settings and compare the RMSPBE (Root of MSPBE) across episodes. Figure-1 to Figure-4 plot these results. We consider decreasing step-sizes of the form: $\xi_t = \frac{1}{(t+1)\xi}$, $\beta_t = \frac{1}{(t+1)\beta}$, $\varrho_t = \frac{1}{(t+1)\varrho}$, $\alpha_t = \frac{1}{(t+1)\alpha}$ in all the examples. Table 1 summarizes the different step-size sequences used in our experiment.

In One-TS setting, we require $\xi = \beta = \varrho$. Since $\xi_t = \frac{\alpha_t}{\varrho_t}$, we must have $\alpha = 2\varrho$. In the Three-TS setting, $\xi < \beta < \varrho$ thus implying, $\alpha < \varrho + \beta$ and $\beta < \varrho$. Although our analysis requires square summability: $\xi, \beta, \varrho > 0.5$, such choice of step-size makes the algorithms converge very slowly. Recently, (Dalal et al. 2018a) showed convergence rate results for Gradient TD schemes with non-square summable step-sizes also (See Remark 2 of (Dalal et al. 2018a)). Therefore, we look at non-square summable step-sizes here, and observe the iterates do converge. The momentum parameter is chosen as in **A2**. In all the examples considered, the momentum methods outperform their vanilla counterparts.

Table 1: Choice of step-size parameters

Boyan Chain	α	β	ϱ	w
Vanilla	0.25	0.125	-	-
One-TS	0.25	0.125	0.125	1
Three-TS	0.25	0.125	0.2	0.1
5-state RW	α	β	ϱ	w
Vanilla	0.25	0.125	-	-
One-TS	0.25	0.125	0.125	1
Three-TS	0.25	0.125	0.2	0.1
19-State RW	α	β	ϱ	w
Vanilla	0.125	0.0625	-	-
One-TS	0.125	0.0625	0.0625	1
Three-TS	0.125	0.0625	0.1	0.1
Random Chain	α	β	ϱ	w
Vanilla	0.5	0.25	-	-
One-TS	0.5	0.25	0.25	1
Three-TS	0.5	0.25	0.3	0.1

Since, in the Three-TS setting, a lower value of w can be chosen, this ensures that the momentum parameter is not small in the initial phase of the algorithm as in the One-TS setting. This in turn helps to reduce the RMSPBE faster in the initial phase of the algorithm as is evident from the experiments.

Remark 4. Increased Variance: Increase in variance with momentum schemes is a known fact in the SGD literature and

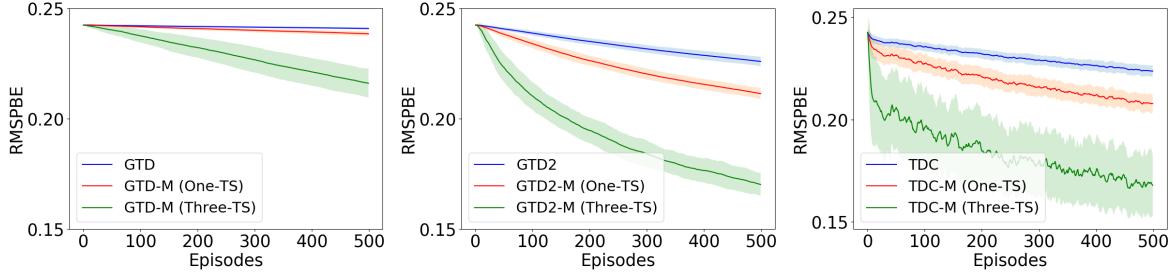


Figure 3: RMSPBE (averaged over 100 independent runs) across episodes for the 19-State Random Walk problem. The features used are an extension of the *Dependent* features used in (Sutton et al. 2009).

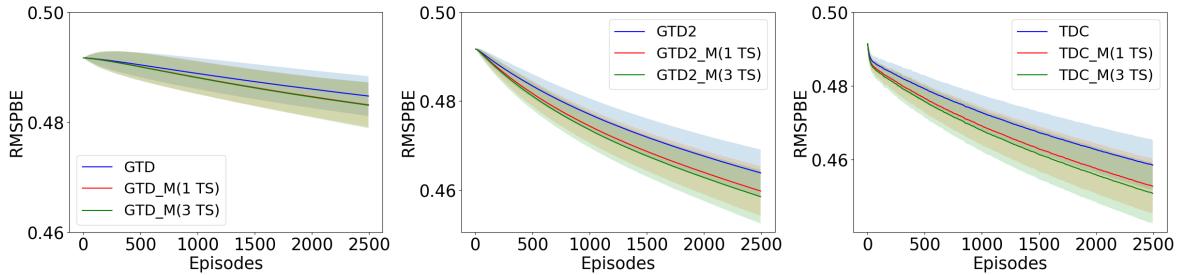


Figure 4: RMSPBE (averaged over 100 independent runs) across episodes for 20-state Random MDP with 5 random actions. The features used are *Linear random* of size 10 (see (Dann, Neumann, and Peters 2014)). For each state, the value of the feature vector at 10th position is 1 and all the values in all other 9 positions is chosen randomly from 0 to 10 and are then normalized.

is evident in our experiments too. Since there is an added push at each step the path becomes jittery. A similar reasoning holds for our algorithms. We expect the initial distribution of the parameter to have a large co-variance matrix.

6 Related Work and Conclusion

To the best of our knowledge no previous work has specifically looked at Gradient TD methods with an added heavy ball term. The use of momentum specifically in the SA setting is very limited. Section 4.1 of (Mou et al. 2020) does talk about momentum; however the problem looked at is that of SGD with momentum and the driving matrix is assumed to be symmetric (see Appendix H of their paper). We do not make any such assumption here. The work of (Devraj, Bušić, and Meyn 2019), indeed looks at momentum in SA setting. However, they introduce a matrix momentum term which is not equivalent to heavy ball momentum. Acceleration in Gradient TD methods has been looked at in (Pan, White, and White 2017). The authors provide a new algorithm called ATD and the acceleration is in form of better data efficiency. However, they do not make use of momentum methods.

In this work we have introduced heavy ball momentum in Gradient Temporal difference algorithms for the first time. We decompose the two iterates of these algorithms into three separate iterates and provide asymptotic convergence guarantees of these new schemes under the same assumptions made by their vanilla counterparts. Specifically, we show convergence in the One-TS regime as well as Three-TS regime.

In both the cases, the momentum parameter gradually goes 1. Three-TS formulation gives us more flexibility in choosing the momentum parameter. Specifically, compared to the One-TS setting, a larger momentum parameter can be chosen during the initial phase in the Three-TS case. We observe improved performance with these new schemes when compared with the original algorithms.

As a step forward from this work, the natural direction would be to look at more sophisticated momentum methods such as Nesterov’s accelerated method (Nesterov 1983). Also, here we only provide the convergence guarantees of these new momentum methods. A particularly interesting step would be to quantify the benefits of using momentum in SA settings. Specifically, it would be interesting to extend weak convergence rate analysis of (Konda and Tsitsiklis 2004; Mokkadem and Pelletier 2006) to Three-TS regime. For the One-TS versions, we expect the convergence rate to remain same. In the Three-TS case, since noise in the third iterate is missing, there could be a possibility of improvement in convergence rate and requires further exploration. Also, extending the recent convergence rate results in expectation and high probability of GTD methods (Dalal et al. 2018b; Gupta, Srikant, and Ying 2019; Kaledin et al. 2019; Dalal, Szorenyi, and Thoppe 2020) to these momentum settings would be interesting works for the future. Finally a recent work by (Thoppe et al. 2021) looked at sample complexity of momentum schemes with a single iterate. It could be interesting to extend these results to our case.

Acknowledgments

This work was supported in part by a J.C.Bose Fellowship, a project from the Department of Science and Technology, Government of India and the RBCCPS, IISc. The authors would also like to thank Gugan Thoppe for some useful discussions.

References

- Assran, M.; and Rabbat, M. 2020. On the Convergence of Nesterov's Accelerated Gradient Method in Stochastic Settings. *Proceedings of the 37th International Conference on Machine Learning, PMLR*, 119: 410–420.
- Avrachenkov, K.; Patil, K.; and Thoppe, G. 2020. Online Algorithms for Estimating Change Rates of Web Pages. *arXiv*, 2009.08142.
- Baird, L. 1995. Residual Algorithms: Reinforcement Learning with Function Approximation. In *In Proceedings of the Twelfth International Conference on Machine Learning*, 30–37. Morgan Kaufmann.
- Borkar, V. S. 2008. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press. ISBN 9780521515924.
- Borkar, V. S.; and Meyn, S. P. 2000. The O.D.E. Method for Convergence of Stochastic Approximation and Reinforcement Learning. *SIAM Journal on Control and Optimization*, 38(2): 447–469.
- Boyan, J. 1999. Least-Squares Temporal Difference Learning. In *ICML*.
- Dalal, G.; Szorenyi, B.; and Thoppe, G. 2020. A Tale of Two-Timescale Reinforcement Learning with the Tightest Finite-Time Bound. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 3701–3708.
- Dalal, G.; Szorenyi, B.; Thoppe, G.; and Mannor, S. 2018a. Finite Sample Analysis of Two-Timescale Stochastic Approximation with Applications to Reinforcement Learning. *arXiv*:1703.05376.
- Dalal, G.; Thoppe, G.; Szörényi, B.; and Mannor, S. 2018b. Finite Sample Analysis of Two-Timescale Stochastic Approximation with Applications to Reinforcement Learning. In Bubeck, S.; Perchet, V.; and Rigollet, P., eds., *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, 1199–1233. PMLR.
- Dann, C.; Neumann, G.; and Peters, J. 2014. Policy Evaluation with Temporal Differences: A Survey and Comparison. *Journal of Machine Learning Research*, 15(24): 809–883.
- Deb, R.; and Bhatnagar, S. 2021. Gradient Temporal Difference with Momentum: Stability and Convergence. *arXiv*:2111.11004.
- Devraj, A. M.; Bušić, A.; and Meyn, S. 2019. On Matrix Momentum Stochastic Approximation and Applications to Q-learning. *57th Annual Allerton Conference on Communication, Control, and Computing*, 749–756.
- Gadat, S.; Panloup, F.; and Saadane, S. 2016. Stochastic Heavy ball. *Electronic Journal of Statistics*, 12: 461–529.
- Ghadimi, E.; Feyzmahdavian, H. R.; and Johansson, M. 2014. Global convergence of the Heavy-ball method for convex optimization. *arXiv*:1412.7457.
- Gitman, I.; Lang, H.; Zhang, P.; and Xiao, L. 2019. Understanding the role of momentum in stochastic gradient methods. *Advances in Neural Information Processing Systems*, 9630–9640.
- Gupta, H.; Srikant, R.; and Ying, L. 2019. Finite-Time Performance Bounds and Adaptive Learning Rate Selection for Two Time-Scale Reinforcement Learning. *arXiv*:1907.06290.
- Kaledin, M.; Moulines, E.; Naumov, A.; Tadic, V.; and Wai, H. 2019. Finite Time Analysis of Linear Two-timescale Stochastic Approximation with Markovian Noise. *Conference on Learning Theory*, 125: 2144–2203.
- Karmakar, P.; and Bhatnagar, S. 2018. Two Time-Scale Stochastic Approximation with Controlled Markov Noise and Off-Policy Temporal-Difference Learning. *Mathematics of Operations Research*, 43(1): 130–151.
- Konda, V.; and Tsitsiklis, J. 2004. Convergence rate of linear two-time-scale stochastic approximation. *Annals of Applied Probability*, 14.
- Kushner, H.; and Clark, D. 1978. *Stochastic Approximation Methods for constrained and unconstrained systems*. Springer.
- Ljung, L. 1977. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(4): 551–575.
- Loizou, N.; and Richtárik, P. 2020. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77: 653–710.
- Ma, J.; and Yarats, D. 2019. Quasi-hyperbolic momentum and adam for deep learning. *International Conference on Learning Representations*.
- Maei, H. R. 2011. *Gradient Temporal-Difference Learning Algorithms*. Ph.D. thesis, University of Alberta, CAN. AAINR89455.
- Mokkadem, A.; and Pelletier, M. 2006. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *The Annals of Applied Probability*, 16(3): 1671 – 1702.
- Mou, W.; Li, C. J.; Wainwright, M. J.; Bartlett, P. L.; and Jordan, M. I. 2020. On Linear Stochastic Approximation: Fine-grained Polyak-Ruppert and Non-Asymptotic Concentration. *Proceedings of Thirty Third Conference on Learning Theory, PMLR*, 125: 2947–2997.
- Nesterov, Y. 1983. A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$. *Soviet Mathematics Doklady*, 269: 543–547.
- Pan, Y.; White, A.; and White, M. 2017. Accelerated Gradient Temporal Difference Learning. *arXiv*:1611.09328.
- Polyak, B. 1964. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4: 1–17.
- Polyak, B. 1990. New stochastic approximation type procedures. *Avtomática i Telemekhanika*, 7: 98–107.

Ramaswamy, A.; and Bhatnagar, S. 2019. Stability of Stochastic Approximations With “Controlled Markov” Noise and Temporal Difference Learning. *IEEE Transactions on Automatic Control*, 64(6): 2614–2620.

Robbins, H.; and Monro, S. 1951. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3): 400 – 407.

Sutton, R.; and Barto, A. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book. ISBN 0262039249.

Sutton, R.; Maei, H.; Precup, D.; Bhatnagar, S.; Silver, D.; Szepesvári, C.; and Wiewiora, E. 2009. Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML ’09, 993–1000. New York, NY, USA: Association for Computing Machinery. ISBN 9781605585161.

Sutton, R. S. 1988. Learning to Predict By the Methods of Temporal Differences. *Machine Learning*, 3(1): 9–44.

Sutton, R. S.; Maei, H.; and Szepesvári, C. 2009. A Convergent O(n) Temporal-difference Algorithm for Off-policy Learning with Linear Function Approximation. In Koller, D.; Schuurmans, D.; Bengio, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Thoppe, G.; Deb, R.; Ganesh, S.; and Budhiraja, A. 2021. Does Momentum Help? A Sample Complexity Analysis. arXiv:2110.15547.

Tsitsiklis, J.; and Van Roy, B. 1997. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5): 674–690.