Stability and Sensitivity Analysis of Relative Temporal-Difference Learning: Extended Version

Masoud S. Sakha[†], Rushikesh Kamalapurkar[†] and Sean Meyn*

April 8, 2026

Abstract

Relative temporal-difference (RTD) learning was introduced to mitigate the slow convergence of TD methods when the discount factor approaches one by subtracting a baseline from the temporal-difference update. While this idea has been studied in the tabular setting, stability guarantees with function approximation remain poorly understood. This paper analyzes RTD learning with linear function approximation. We establish stability conditions for the algorithm and show that the choice of baseline distribution plays a central role. In particular, when the baseline is chosen as the empirical distribution of the state–action process, the algorithm is stable for any non-negative baseline weight and any discount factor. We also provide a sensitivity analysis of the resulting parameter estimates, characterizing both asymptotic bias and covariance. The asymptotic covariance and asymptotic bias are shown to remain uniformly bounded as the discount factor approaches one.

Extended version of manuscript submitted to the 2026 IEEE Conference on Decision and Control.

Keywords: reinforcement learning; optimal control; temporal-difference learning.

[†]University of Florida, Department of Mechanical and Aerospace Engineering (e-mails: masoud.sakha@ufl.edu, rkamalapurkar@ufl.edu).

*University of Florida, Department of Electrical and Computer Engineering (e-mail: meyn@ece.ufl.edu). Financial support from ARO award W911NF2010055, NSF award CCF 2306023, and AFRL award FA8651-24-1-0019 is gratefully acknowledged.

Contents

1 Introduction

Consider a Markov Decision Process (MDP) with state process $\boldsymbol{X} = \{X_k : k \geq 0\}$ evolving on a state space X, and input process $\boldsymbol{U} = \{U_k : k \geq 0\}$ evolving on U. In most of the technical results it is assumed that X and U are finite to avoid discussion of measurability and other technicalities.

In this paper, we consider approximation of the solution to the discounted-cost optimal control problem. For a given discount factor $\gamma \in (0, 1)$ and a cost function $c\colon \mathsf{X} \times \mathsf{U} \to \mathbb{R}$, the state-action value function is denoted by $Q^\star\colon \mathsf{X} \times \mathsf{U} \to \mathbb{R}$, and is defined by

$$Q^\star(x, u) = \min \sum_{k=0}^{\infty} \gamma^k \mathsf{E}[c(Z_k) \mid Z_0 = (x, u)] \tag{1a}$$

where $Z_k = (X_k, U_k)$, and the minimum is over all history dependent input sequences. This is the *Q-function* of Q-learning, which solves the Bellman equation,

$$Q^\star(x, u) = c(x, u) + \gamma \mathsf{E}[\underline{Q}^\star(X_1) \mid Z_0 = (x, u)] \tag{1b}$$

where $\underline{Q}^\star(x) := \min_u Q^\star(x, u)$.

One approach to approximating $Q^\star$ is through approximate policy iteration, in which case we require approximation of a fixed-policy Q-function. Let $\widetilde{\phi}$ denote a randomized stationary policy,

defined so that $\mathsf{P}\{U_k = u \mid X_0, \ldots, X_k\} = \widetilde{\phi}(u \mid x)$ when $X_k = x$. For the fixed policy $\widetilde{\phi}$, the discounted cost and Bellman equation take on a form similar to (1):

$$Q(x,u) = \sum_{k=0}^{\infty} \gamma^k \mathsf{E}[c(Z_k) \mid Z_0 = (x,u)] \tag{2a}$$

$$Q(x,u) = c(x,u) + \gamma \mathsf{E}[\underline{Q}(X_1) \mid Z_0 = (x,u)] \tag{2b}$$

where the underbar has a new interpretation: $\underline{Q}(x) = \sum_u Q(x,u)\widetilde{\phi}(u \mid x)$, and in (2a) the state-action sequence $\boldsymbol{Z}$ is obtained using the policy $\widetilde{\phi}$.

In this paper we consider linear function approximation, in which $Q^\theta(x,u) = \theta^\intercal \psi(x,u)$ for a parameter vector $\theta \in \mathbb{R}^d$, and basis vector $\psi \colon \mathsf{X} \times \mathsf{U} \to \mathbb{R}^d$. In this setting, algorithms to estimate $Q^\star$, or the fixed policy Q-function Q, can be interpreted as stochastic approximation (SA) algorithms with associated mean flows of the form

$$\tfrac{d}{dt}\vartheta_t = \bar{f}(\vartheta_t) \tag{3}$$

Stability of the algorithm follows from global asymptotic stability of the mean flow.

When convergence holds, so that $\bar{f}$ has a unique root θ^*, finer properties such as rates of convergence are governed by the Jacobian $\bar{A} = \partial \bar{f}\,(\theta^*)$. In particular, the celebrated optimal *asymptotic covariance* in the Central Limit Theorem (CLT) for SA is given by

$$\Sigma_\Theta^* = \bar{A}^{-1} \Sigma_\Delta [\bar{A}^\intercal]^{-1} \tag{4}$$

in which Σ_Δ is the $d \times d$ "noise covariance"—further discussion and history may be found below. The more recent theory of *asymptotic bias* also involves the inverse of this Jacobian:

$$\beta_\theta = \tfrac{1}{1-\rho} \bar{A}^{-1} \bar{\Upsilon} \tag{5}$$

with $\bar{\Upsilon} \in \mathbb{R}^d$; see Prop. 2.2 for details.

A fundamental challenge arises when the discount factor γ approaches one. In this regime, the matrix $\bar{A}$ is typically poorly conditioned, and an eigenvalue may approach zero as $\gamma \uparrow 1$. As a consequence, the asymptotic covariance and bias can grow large, and convergence may become arbitrarily slow. This behavior is reflected in both classical SA bounds and recent analyses of TD and Q-learning, where error bounds depend explicitly on the inverse $\bar{A}^{-1}$ associated with the linearized dynamics [5, 6, 12].

Relative temporal-difference (RTD) methods were introduced to address this issue by subtracting a baseline term from the update. These methods were originally proposed in the context of Q-learning, where they were shown empirically to improve convergence, particularly in regimes with large discount factors. There is complete theory only for tabular Q-learning—understanding of relative methods with function approximation remains limited.

The present paper provides a systematic analysis of RTD learning with linear function approximation. Our main contribution is to show that a natural choice of baseline—based on the stationary distribution of the state–action process—yields a mean flow whose linearization remains uniformly Hurwitz and well-conditioned over $\gamma \in (0,1)$. As a consequence, both asymptotic covariance and bias remain uniformly bounded. We also provide sensitivity analysis with respect to the baseline parameter.

These results connect directly with modern stochastic approximation theory, where both finite-time bounds and asymptotic covariance depend explicitly on the conditioning of the linearized mean dynamics. In this sense, RTD learning can be interpreted as a mechanism for regularizing the mean flow.

1.1 Temporal difference methods

The reason for the common notation in (1b) and (2b) is for common notation in the presentation of algorithms, which take the following form: For initialization $\theta_0 \in \mathbb{R}^d$, the sequence of estimates is defined recursively as

$$\theta_{n+1} = \theta_n + \alpha_{n+1}\mathcal{D}_{n+1}\zeta_n \tag{6a}$$

$$\mathcal{D}_{n+1} = c(Z_n) + \gamma \underline{Q}^{\theta_n}(X_{n+1}) - Q^{\theta_n}(Z_n) \tag{6b}$$

in which, $\{\alpha_n\}$ is a non-negative step-size sequence, $\{\mathcal{D}_{n+1}\}$ is known as the temporal difference sequence, and the term $\underline{Q}^{\theta_n}(x)$ is defined according to the context; see preceding discussion and (7). The d-dimensional vectors in the sequence $\{\zeta_n\}$ are known as the *eligibility vectors*. Under our assumption of linear function approximation, $Q^\theta = \theta^\intercal \psi$, a common choice of eligibility vector is defined through the d-dimensional recursion,

$$\zeta_{n+1} = \lambda\gamma\zeta_n + \psi(Z_{n+1}), \quad \zeta_0 = \psi(Z_0)\,, \tag{6c}$$

where $\lambda \in [0, 1]$ is fixed, with $\lambda = 0$ most common in applications.

The term *Q-learning* is reserved for the case in which $\underline{Q}^{\theta_n}(X_{n+1}) = \min_u Q^{\theta_n}(X_{n+1}, u)$, and the recursion (6a) reduces to Watkins' algorithm when using a tabular basis [22, 21]. See [18, 19, 12] for a range of interpretations of the algorithm. *TD-learning* is the alternative in which (6) is constructed to approximate the fixed-policy Q-function. In this case, there are three possible implementations:

$$\underline{Q}^{\theta_n}(X_{n+1}) = \sum_u Q^{\theta_n}(x, u)\widetilde{\phi}(u \mid x) \qquad \begin{array}{l}\textit{Natural}\\ x = X_{n+1}\end{array} \tag{7a}$$

$$= Q^{\theta_n}(Z_{n+1}) \qquad \text{On-policy} \tag{7b}$$

$$= Q^{\theta_n}(Z'_{n+1}) \qquad \text{Split-sampling} \tag{7c}$$

For the on-policy method (7b) it is assumed that the input $\boldsymbol{U}$ used for training in (6) is defined by the policy $\widetilde{\phi}$. In the other two approaches the only restriction on the input is that it is adapted to the state process and results in a stable algorithm. In the case of split sampling (7c) we define $Z'_{n+1} = (X_{n+1}, U'_{n+1})$ in which the random variable U'_{n+1} is drawn randomly, with distribution $U'_{n+1} \sim \widetilde{\phi}(\,\cdot \mid x)$ when $X_{n+1} = x$, independent of $\{U_{n+1}, Z_k : k \le n\}$.

1.2 Convergence rates

Any of the TD algorithms considered in this paper can be expressed in the form of a stochastic approximation (SA) recursion

$$\theta_{n+1} = \theta_n + \alpha_{n+1} f_{n+1}(\theta_n)\,, \tag{8}$$

with $f_{n+1}(\theta_n) = f(\theta_n, \Phi_{n+1})$ in which the definition of the stochastic process $\boldsymbol{\Phi}$ depends on the specifics of the algorithm.

In our treatment of TD learning it is assumed that the policy used for training is *oblivious* ($\widetilde{\phi}$ does not depend on the parameter estimate). It is further assumed that $\boldsymbol{\Phi}$ is Markovian and has a unique steady state distribution π, which for TD learning requires that $\boldsymbol{Z}$ has a unique steady state pmf, which is denoted ϖ. In on-policy TD(0) we may take $\Phi_{n+1} = (Z_n, Z_{n+1})$. For TD($\lambda$) with non-zero λ we must include the eligibility vector, so that $\Phi_{n+1} = (Z_n, Z_{n+1}, \zeta_{n+1})$ for each $n \ge 0$.

Potential limits of the SA recursion are solutions to the root finding problem $\bar{f}(\theta^*) = 0$ in which $\bar{f}(\theta) = \mathsf{E}_\pi[f(\theta, \Phi_{n+1})]$, $\theta \in \mathbb{R}^d$. Convergence theory of SA is built around stability theory for the d-dimensional mean flow $\frac{d}{dt}\vartheta_t = \bar{f}(\vartheta_t)$ introduced in (3). The strongest conclusions for SA require that the mean flow is exponentially asymptotically stable (EAS).

It is well known that the rate of convergence of the mean square error is no faster than $O(1/n)$ except in exceptional settings such as quasi-stochastic approximation [9]. When this rate is achieved we can typically also establish the following limit, which defines the *asymptotic covariance*:

$$\Sigma_\Theta := \lim_{n\to\infty} n\mathsf{E}[\tilde{\theta}_n \tilde{\theta}_n^{\intercal}]\,, \quad \textit{where } \tilde{\theta}_n = \theta_n - \theta^* \tag{9}$$

Under mild assumptions we minimize the asymptotic covariance through choice of step-size followed by averaging.

Much of the theory of SA begins with the following assumptions on the step-size sequence:

$$\sum_{n=1}^{\infty} \alpha_n = \infty\,, \qquad \sum_{n=1}^{\infty} \alpha_n^2 < \infty \tag{10}$$

In the discussion that follows and in numerical experiments we opt for the standard choice,

$$\alpha_n = \min(\alpha_0, n^{-\rho})\,, \quad \textit{with } \alpha_0 > 0 \textit{ and } 1/2 < \rho < 1. \tag{11}$$

The upper bound on ρ is imposed to justify averaging, the Polyak-Ruppert (PR) averaged estimates [15, 14], which are defined by

$$\theta_N^{\mathsf{PR}} = \frac{1}{N - N_0 + 1} \sum_{k=N_0}^{N} \theta_k\,, \qquad N \geq N_0 \tag{12}$$

in which the interval $\{0, \dots, N_0\}$ is known as the *burn-in* period.

Theory for additive white noise models is contained in [14] where it is shown that the resulting asymptotic covariance has the form (4) in which Σ_Δ is the covariance of the additive noise. It is known that the matrix Σ_Θ^* is minimal in the matricial sense. In the recent work [4] it is shown that the same conclusions hold for general SA recursions with Markovian disturbance, with an alternative definition of the disturbance covariance:

$$\Sigma_\Delta = \sum_{k=-\infty}^{\infty} R_\Delta(k) \tag{13}$$

where $R_\Delta(k)$ is the autocorrelation sequence for the zero mean sequence $\{f(\theta^*, \Phi_n) : n \in \mathbb{Z}\}$ with $\mathbf{\Phi}$ a stationary realization of the Markov chain.

We also consider the asymptotic bias, defined by

$$\beta_\theta := \lim_{n\to\infty} \frac{\mathsf{E}[\theta_n - \theta^*]}{\alpha_n}\,, \tag{14}$$

which under mild conditions admits the representation (5).

1.3 Relative TD methods

The expressions (4) and (5) tell us that we can expect slow convergence when the condition number of $\bar{A}$ is large. It is pointed out in [5, 6] that in the case of tabular Q-learning, the matrix $\bar{A}$ is

Hurwitz but has an eigenvalue that tends to zero as the discount factor tends to unity. Relative Q-learning was introduced in [6] to address this numerical instability and reduce the asymptotic covariance. Moreover, theory from [6] implies that the greedy policy obtained from relative Q-learning is optimal in the tabular setting.

A relative TD or Q-learning algorithm is defined by a recursion of the form (6a) with a slight modification of the temporal difference sequence,

$$\begin{aligned}\mathcal{D}_{n+1} = c(Z_n) + \gamma \underline{Q}^{\theta_n}(X_{n+1}) - Q^{\theta_n}(Z_n) \\ - \delta_{\mathsf{r}} \langle \mu, Q^{\theta_n} \rangle \,,\end{aligned} \tag{15}$$

in which μ is a pmf on $\mathsf{X} \times \mathsf{U}$ (the *baseline distribution*) and $\delta_{\mathsf{r}} > 0$.

It is shown under mild conditions on μ and $\delta_{\mathsf{r}} > 0$ that the resulting recursion is stable and the asymptotic covariance is uniformly bounded over $0 \le \gamma < 1$: Q-learning is treated in [6] and the simpler case of TD-learning is treated in [12]. However, to-date theory has been absent outside of the tabular setting.

Theory of stability of TD learning starting with [20] requires consideration of the autocorrelation matrices $R(k) := \mathsf{E}[\psi_{(0)}\psi_{(k)}^{\intercal}]$ for any integer k, where $\psi_{(k)} = \psi(Z_k)$, and the expectation is taken in steady state. The mean flow for the RTD(λ) algorithm is given by $\bar{f}(\theta) = \bar{A}\theta + \bar{b}$ with

$$\begin{aligned}\bar{A}(\lambda; \delta_r) &= \mathsf{E}\left[\zeta_n \left[-\psi_{(n)} - \delta_{\mathsf{r}} \bar{\psi}^{\mu} + \gamma \psi_{(n+1)}\right]^{\intercal}\right] \\ &= \bar{A}(\lambda; 0) - \frac{\delta_{\mathsf{r}}}{1 - \lambda\gamma} \bar{\psi} [\bar{\psi}^{\mu}]^{\intercal}\end{aligned} \tag{16}$$

in which $\bar{\psi} = \mathsf{E}_{\pi}[\psi_{(0)}]$, and $\bar{\psi}^{\mu}$ is the vector whose ith component is

$$\bar{\psi}_i^{\mu} = \langle \mu, \psi_i \rangle, \qquad 1 \le i \le d. \tag{17}$$

In (16) and in the following development, the dependence of $\bar{A}$ on λ and δ_{r} is made explicit when needed.

The second equality in (16) follows from (6c) and stationarity. Taking expectation in (6c) gives $\mathsf{E}[\zeta_n] = \lambda\gamma \mathsf{E}[\zeta_n] + \bar{\psi}$, and therefore

$$\mathsf{E}[\zeta_n] = \frac{1}{1 - \lambda\gamma} \bar{\psi} \,, \quad \mathsf{E}\big[\zeta_n (\bar{\psi}^{\mu})^{\intercal}\big] = \frac{1}{1 - \lambda\gamma} \bar{\psi} [\bar{\psi}^{\mu}]^{\intercal} \,,$$

which implies the second equality in (16).

In the present paper we show that the attractive results obtained for tabular Q-learning or TD learning extend to TD learning with linear function approximation, subject to assumptions on the pmf μ. The strongest conclusions are obtained when this is taken to be the steady state pmf ϖ for $\boldsymbol{Z}$.

As this is not known a priori, we can substitute the empirical distribution of the state–action process. This reasoning leads to the following algorithm:

ϖ-Relative TD(λ) learning For initialization $\theta_0 \in \mathbb{R}^d$, the sequence of estimates are defined recursively by (6a) with eligibility vectors defined in (6c) and temporal difference sequence obtained through the following:

$$\begin{aligned}\mathcal{D}_{n+1} = c(Z_n) + \gamma \underline{Q}^{\theta_n}(X_{n+1}) - Q^{\theta_n}(Z_n) \\ - \delta_{\mathsf{r}} \langle \varpi_n, Q^{\theta_n} \rangle \,,\end{aligned} \tag{18a}$$

in which $\langle \varpi_n, Q^{\theta_n} \rangle = \bar{\psi}_n^{\intercal} \theta_n$ with

$$\bar{\psi}_{n+1} = \bar{\psi}_n + \beta_{n+1}[\psi_{(n+1)} - \bar{\psi}_n], \quad \bar{\psi}_0 = \psi(Z_0), \tag{18b}$$

and $\{\beta_n\}$ is a vanishing step-size sequence also satisfying (10). For ease of exposition we assume a relatively large gain: $\lim_{n\to\infty} \beta_n/\alpha_n = \infty$.

Theory of two time-scale SA justifies consideration of a mean flow for approximation of the slow parameter sequence $\{\theta_n\}$ [3]. It remains linear, with $\bar{A}$ given by (16) using $\mu = \varpi$:

$$\bar{A}(\lambda;\delta_r) = \bar{A}(\lambda;0) - \frac{\delta_{\mathsf{r}}}{1-\lambda\gamma}\bar{\psi}\bar{\psi}^{\intercal}. \tag{19}$$

Contributions While the algorithm (15) was introduced more than five years ago, there has been no analysis outside of the tabular setting. It turns out that a good choice for δ_{r} and μ is an art in general.

Moreover, to-date there is not much theory providing sufficient conditions for stability of Q-learning outside of the tabular setting [13, 11]. Consequently, the theoretical results summarized here are restricted to RTD learning with linear function approximation. We choose the on-policy variant based on (7b) since we are assured that the mean flow for TD learning is EAS under the assumptions of [20].

The main new realization in this paper is that $\mu = \varpi$ is universally successful. For this choice of μ and other choices, subject to assumptions, we obtain in general the following dichotomy:

- For the standard TD(λ) learning algorithm the optimal asymptotic covariance (4) and asymptotic bias (14) may be unbounded as a function of discount factor $\gamma \in (0,1)$.
- Under mild assumptions the optimal asymptotic covariance and asymptotic bias for RTD learning are uniformly bounded over $\gamma \in (0,1)$, for any fixed $\lambda < 1$.

Literature Relative Q-learning for average cost appeared in [1]. One decade later it was realized that the same technique could be used to obtain uniformly bounded rates of convergence in the discounted-cost setting, uniformly over $0 < \gamma < 1$ [6]. These papers are entirely devoted to the tabular setting. Extensions of the algorithm to TD learning are contained in [12] (see Section 9.5.3 and the regenerative algorithm in Section 10.4.2). To the best of our knowledge, stability theory has been largely restricted to the tabular setting, with the exception of Theorem 10.15 of [12], which treats the TD(1) algorithm for average cost. The motivation there was for applications to actor–critic methods.

The main results of this paper rest on establishing that the SA mean flow matrix $\bar{A}$ is uniformly Hurwitz and bounded over the range of (λ,γ) considered. Bounds on the conditioning of this matrix are also required to obtain meaningful guarantees in finite-time analysis of TD learning and linear stochastic approximation. For instance, finite-time bounds for linear stochastic approximation and TD learning with constant step size are established in [17]. More recent work considers TD learning with decreasing step sizes and provides rates of convergence for averaged iterates [16]. Related finite-time high-probability bounds for Polyak–Ruppert averaged iterates of linear stochastic approximation are established in [7]. In each of these works, the quantitative behavior depends explicitly on stability of the mean flow along with conditioning properties of the mean flow matrix $\bar{A}$.

In [17], the obtained bounds are proportional to the condition number of the solution P of the Lyapunov equation $\bar{A}^{\intercal}P + P\bar{A} = -I$, which improves with improved conditioning of $\bar{A}$. In [16], bounds are obtained on the Wasserstein distance between $\sqrt{N}(\theta_N^{\mathsf{PR}} - \theta^*)$ and a zero-mean normal distribution with covariance matrix Σ_Θ^*; these bounds likewise improve when $\bar{A}$ is better conditioned. The results of [7], following [2], similarly show that finite-time concentration bounds for averaged linear SA depend on the condition number of $\bar{A}$.

The present work contributes to this line of research by showing that RTD learning, with an appropriately chosen baseline, yields a mean flow matrix $\bar{A}$ that remains uniformly Hurwitz and

well-conditioned as the discount factor approaches one. This leads to uniform bounds on both asymptotic covariance and bias, in contrast to standard TD methods where these quantities may diverge.

Organization The remainder of the paper is organized as follows. Section 2 presents the main theoretical results. Section 3 presents two simulation examples: one for a finite MDP in which all quantities (such as those defining asymptotic bias and covariance) are computable in close form, and a second example with continuous state and action space. Conclusions and directions for future research are contained in Section 4, and proofs of technical results are collected together in an appendix.

2 Main Results

The main results are restricted to on-policy TD learning (based on (7b)) in which $\boldsymbol{Z}$ is a Markov chain on the finite state space $\mathsf{Z} = \mathsf{X} \times \mathsf{U}$. The kth *autocovariance* matrix is denoted $\Sigma(k) := R(k) - \bar{\psi}\bar{\psi}^\intercal$, where the kth autocorrelation matrix was introduced earlier: $R(k) := \mathsf{E}[\psi_{(0)}\psi_{(k)}^\intercal]$ with $\psi_{(k)} = \psi(Z_k)$. Assumption (A0) justifies these definitions and (A1) was introduced in [20] in an early treatment of TD learning:

(A0) $\boldsymbol{Z}$ is a unichain and aperiodic Markov chain, with unique pmf ϖ.

(A1) $R(0) > 0$ *or* **(A1•)** $\Sigma(0) > 0$.

Proofs of the main results surveyed this section may be found in the appendix.

2.1 Bias

The following result is a consequence of (16) (which defines $\bar{A}(\lambda;\delta_r)$) combined with the Sherman–Morrison formula.

Proposition 2.1 (Bias)**.** *Suppose that (A0) holds, and for fixed $\lambda \in [0,1]$, suppose that the matrices $\bar{A}_0 := \bar{A}(\lambda;0)$ and $\bar{A}(\lambda;\delta_{\mathsf{r}})$ are invertible. Let $\theta^*(0)$, $\theta^*(\delta_{\mathsf{r}})$ denote the respective stationary points for the mean flow. Then,*

$$\theta^*(\delta_{\mathsf{r}}) = [I + G]\theta^*(0)\,, \qquad G = \frac{\delta_{\mathsf{r}}}{1-\lambda\gamma}\,\frac{\bar{A}_0^{-1}\bar{\psi}[\bar{\psi}^{\mu}]^\intercal}{1-[\bar{\psi}^{\mu}]^\intercal\bar{A}_0^{-1}\bar{\psi}}$$

■

Hence the estimate obtained from RTD-learning may be far from what is obtained using the standard algorithm. In special cases we find that the difference $Q^{\theta^*(0)}(z) - Q^{\theta^*(\delta_{\mathsf{r}})}(z)$ is independent of z, so that the corresponding $Q^{\theta^*(\delta_{\mathsf{r}})}$-greedy policies do not depend on δ_{r} [6].

We next provide justification for a finite (and in general non-zero) limit defining the asymptotic bias in (14). The following proposition follows from the general theory in [9].

Proposition 2.2 (Asymptotic bias)**.** *Consider the SA recursion* (8) *with $f_{n+1}(\theta) = A(\Phi_{n+1})\theta + b(\Phi_{n+1})$ in which $\boldsymbol{\Phi}$ is unichain and aperiodic, and the steady-state mean $\bar{A} = \mathsf{E}[A(\Phi)]$ is Hurwitz. Suppose the step-size is of the form* (11)*. Then, the limit* (14) *exists and has the form* (5) *in which $\bar{\Upsilon} = \mathsf{E}[\Upsilon^*_{n+1}]$ with expectation in steady-state, and*

$$\Upsilon^*_{n+1} = \big(A_{n+1} - \widehat{A}_{n+1}\big)\big(A_{n+1}\theta^* + b_{n+1}\big), \tag{20a}$$

where $A_{n+1} := A(\Phi_{n+1})$, $b_{n+1} := b(\Phi_{n+1})$, and $\widehat{A}_{n+1} := \widehat{A}(\Phi_{n+1})$ is a zero-mean $d \times d$ stochastic process defined through the Poisson equation $\mathsf{E}[\widehat{A}_{n+1} \mid \Phi_0^n] = \widehat{A}_n - A_n + \bar{A}$. ■

The proposition only applies to TD(0) learning since when $\lambda > 0$ we must take $\Phi_{n+1} = (Z_n, Z_{n+1}, \zeta_{n+1})$, which violates the specific geometric ergodicity assumption imposed in [9]. We conjecture that a similar result holds for general λ.

2.2 Stability and asymptotic variance

We have noted that TD-learning algorithms may exhibit very slow convergence when the discount factor is close to unity, and one explanation is that the Jacobian $\bar{A}$ may be ill conditioned. Alternatively, one finds that the estimation problem itself is fundamentally ill-conditioned. Consider for illustration the fixed policy Q-function (2a). Under mild conditions one obtains the approximation,

$$Q(x,u) = H(x,u) + \frac{\eta}{1-\gamma} + o(1-\gamma) \tag{21}$$

where $o(1-\gamma) \to 0$ as $\gamma \uparrow 1$, η is the average cost, and H a solution to Poisson's equation, $\eta + H(x,u) = c(x,u) + \mathsf{E}[H(Z_1) \mid Z_0 = (x,u)]$ [12, Chapter 6].

If the basis is chosen to approximate Q then from (21) it is not unreasonable to assume there is a vector $\xi \in \mathbb{R}^d$ satisfying

$$\xi^\intercal \psi(z) = 1\,, \quad \textit{for } z \in \mathsf{X} \times \mathsf{U} \textit{ satisfying } \varpi(z) > 0 \tag{22}$$

In the case of a tabular basis we obtain (22) using $\xi_i = 1$ for each i.

Recall that $\bar{\psi}^\mu \in \mathbb{R}^d$ is defined in (17).

Theorem 2.3 (RTD(λ) may be unstable)**.** *Suppose that* (A0)-(A1) *hold and there is a solution* $\xi \in \mathbb{R}^d$ *to* (22)*. Suppose that* μ *is the pmf on* $\mathsf{X} \times \mathsf{U}$ *used in the RTD(λ) learning algorithm. Then,*

(i) *If* $\xi^\intercal \bar{\psi}^\mu > 0$ *then for each* $\lambda \in [0,1)$ *there exists* $\delta_{\mathsf{r}}^0 > 0$ *such that the mean flow for RTD(λ) learning is EAS for any* $\gamma \in [0,1]$ *and* $\delta_{\mathsf{r}} \in (0, \delta_{\mathsf{r}}^0)$*.*

(ii) *If* $\xi^\intercal \bar{\psi}^\mu < 0$ *then for each* $\lambda \in [0,1)$ *there exists* $\delta_{\mathsf{r}}^0 > 0$ *and* $\gamma^0 > 0$ *such that* $\bar{A}(\lambda; \delta_{\mathsf{r}})$ *contains an eigenvalue in the strict right half plane whenever* $\gamma \in [\gamma^0, 1]$ *and* $\delta_{\mathsf{r}} \in (0, \delta_{\mathsf{r}}^0)$*.* ∎

To ensure stability we might design the basis so that $\xi^\intercal \psi(z) > 0$ for each z. We opt for the following alternative: To ensure stability we assume $\Sigma(0) > 0$ (i.e., (A1$^\bullet$) holds).

Observe that subject to (22) we have $\xi^\intercal \psi_{(k)} = \xi^\intercal \bar{\psi} = 1$ for all k, and hence

$$\xi^\intercal \Sigma(0) \xi = \xi^\intercal \mathsf{E}[\psi_{(k)} \psi_{(k)}^\intercal] \xi - (\xi^\intercal \bar{\psi})^2 = 0$$

Hence (A1$^\bullet$) fails. However, the existence of ξ is well motivated only when we wish to directly estimate the Q function using a standard TD learning algorithm.

Theorem 2.4 (Stability and Variance)**.** *Under* (A0)-(A1$^\bullet$)*, for any* $\varepsilon \in (0,1)$*,*

(i) *There exists* $\delta_{\mathsf{r}}^0 > 0$ *such that* $\bar{A}(\lambda; \delta_{\mathsf{r}})$ *obtained from RTD(λ) learning is Hurwitz for any* $\lambda, \gamma \in [0,1]$ *satisfying* $\gamma\lambda \le 1 - \varepsilon$*, and* $\delta_{\mathsf{r}} \in [0, \delta_{\mathsf{r}}^0)$*. Moreover, the condition number and optimal asymptotic covariance are uniformly bounded:*

$$\sup \left[\operatorname{cond}(\bar{A}(\lambda; \delta_{\mathsf{r}})) + \operatorname{trace}(\Sigma_\Theta^*) \right] < \infty \tag{23}$$

where the supremum is over this range of $\lambda, \gamma, \delta_{\mathsf{r}}$*.*

(ii) *If* $\mu = \varpi$ *then the mean flow associated with the RTD(λ) algorithm is stable for any* $\delta_{\mathsf{r}} \ge 0$*, and any* $\gamma, \lambda \in [0,1]$ *satisfying* $\gamma\lambda < 1$*. Moreover,* (23) *holds with the supremum over all such* γ, λ *satisfying* $\gamma\lambda \le 1 - \varepsilon$ *and* $\delta_{\mathsf{r}} \in [0,1]$*.* ∎

2.3 Sensitivity

The next result considers sensitivity of variance and bias of ϖ-RTD(λ) learning for small δ_{r}. We restrict to $\lambda = 0$, and recall that in this case $\delta_{\mathsf{r}} = 0$ corresponds to the standard TD(0)-learning algorithm.

Thm. 2.4 (ii) extends to this algorithm, since the joint process $\{(\theta_n; \bar{\psi}_n) : n \geq 0\}$ is the output of a two time-scale linear SA recursion in which the mean flow associated with $\{\theta_n\}$ is identical to that considered in Thm. 2.4 (ii). It is also convenient to consider an alternative to simplify analysis.

Fixed ϖ-Relative TD(0) learning The on-policy version of the algorithm is expressed as

$$\begin{aligned}\theta_{n+1} &= \theta_n + \alpha_{n+1}[A_{n+1}\theta_n + b_{n+1}] \\ A_{n+1} &= \psi_{(n)}\big[-\psi_{(n)} + \gamma\psi_{(n+1)}\big]^{\intercal} - \delta_{\mathsf{r}}\bar{\psi}\bar{\psi}^{\intercal} \\ b_{n+1} &= c_n\psi_{(n)}\end{aligned} \tag{24}$$

The vector $\bar{\psi}$ is computed a priori, or estimated through Monte-Carlo (18b) with $\beta_n = 1/n$.

We denote by $\theta^*(\delta_{\mathsf{r}})$ the stationary point for the mean flow, $\Sigma^*_{\Theta}(\delta_{\mathsf{r}})$ the asymptotic covariance matrix (4) and $\beta_{\theta}(\delta_{\mathsf{r}})$ the value of (14) for arbitrary δ_{r} in RTD-learning. In the following results, we also require explicit indication of dependency on δ_{r} for auxiliary variables such as $\bar{\Upsilon}(\delta_{\mathsf{r}})$. When $\delta_{\mathsf{r}} = 0$ we drop the dependency to simplify the resulting equations. Since the following results only apply to the $\lambda = 0$ case, we also drop this dependency from $\bar{A}(\lambda; \delta_{\mathsf{r}})$.

Theorem 2.5 (Sensitivity)**.** *Consider the ϖ-RTD(0) learning as defined in* (18) *subject to* (A0)-(A1$^{\bullet}$)*. The algorithm is convergent for any value of $\delta_{\mathsf{r}} \geq 0$ and we have the following sensitivity formulae for asymptotic covariance and bias:*

$$\begin{aligned}\frac{d}{d\delta_{\mathsf{r}}}\Sigma^*_{\Theta}(\delta_{\mathsf{r}})\big|_{\delta_{\mathsf{r}}=0} &= \bar{A}^{-1}\Sigma'_{\Delta}\bar{A}^{-\intercal} \\ &\quad + \bar{A}^{-1}\bar{\psi}\bar{\psi}^{\intercal}\Sigma^*_{\Theta} + [\bar{A}^{-1}\bar{\psi}\bar{\psi}^{\intercal}\Sigma^*_{\Theta}]^{\intercal} \qquad (25a) \\ \frac{d}{d\delta_{\mathsf{r}}}\beta_{\theta}(\delta_{\mathsf{r}})\big|_{\delta_{\mathsf{r}}=0} &= \frac{1}{1-\rho}\bar{A}^{-1}\Big[\bar{\psi}\bar{\psi}^{\intercal}\beta_{\theta} + \bar{\Upsilon}'\Big] \qquad (25b)\end{aligned}$$

The matrix $\Sigma'_{\Delta} := \frac{d}{d\delta_{\mathsf{r}}}\Sigma_{\Delta}(\delta_{\mathsf{r}})\Big|_{\delta_{\mathsf{r}}=0}$ and vector $\bar{\Upsilon}' := \frac{d}{d\delta_{\mathsf{r}}}\bar{\Upsilon}(\delta_{\mathsf{r}})\Big|_{\delta_{\mathsf{r}}=0}$ are each identically zero when using the variant (24)*.* ∎

The theorem combined with (4) implies that a bound on $\|\frac{d}{d\delta_{\mathsf{r}}}\Sigma^*_{\Theta}(\delta_{\mathsf{r}})\|$ can be expected to grow as fast as $\|\bar{A}^{-1}\|^3$ and (25b) combined with (5) tell us that a bound on $\|\frac{d}{d\delta_{\mathsf{r}}}\beta_{\theta}(\delta_{\mathsf{r}})\|$ will grow as $\|\bar{A}^{-1}\|^2$. An example of the steep rate of change in bias is illustrated in Fig. 5 in our first numerical example.

3 Simulations

We survey results from two examples to illustrate the main results of the paper. The first considers a finite-state, finite-action MDP, while the second examines a stochastic speed-scaling model with a continuous state space. In both cases, comparisons between TD and ϖ-RTD illustrate the improved stability and variance properties of the proposed method.

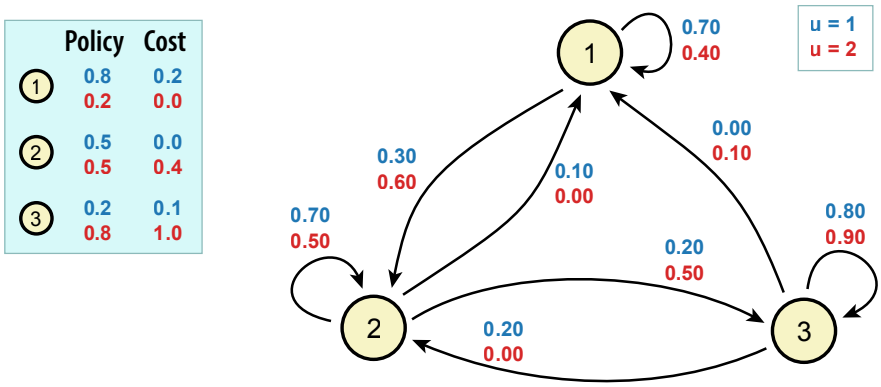

Figure 1: 3-state 2-action MDP model, with randomized policy. Variants of TD learning were performed with discount factor $\gamma = 0.99$, and step-size (11) with $\rho = 0.65$ and $\alpha_0 = 2$.

3.1 Finite-state Finite-action MDP

Fig. 1 shows an MDP example with state space $\mathsf{X} = \{1, 2, 3\}$ and action space $\mathsf{U} = \{1, 2\}$, along with the randomized policy considered in experiments. For example, $\widetilde{\phi}(1 \mid 1) = \mathsf{P}\{U_k = 1 \mid X_k = 1\} = 0.7$ and $\widetilde{\phi}(1 \mid 3) = \mathsf{P}\{U_k = 1 \mid X_k = 3\} = 0.7$.

The optimal average cost policy is unique and deterministic with $\phi^*(1) = 2$ and $\phi^*(2) = \phi^*(3) = 1$. Under this policy the transition matrix and one-stage cost vector are

$$P_{\phi^*} = \begin{bmatrix} 0.40 & 0.60 & 0.00 \\ 0.10 & 0.70 & 0.20 \\ 0.00 & 0.20 & 0.70 \end{bmatrix}, \qquad c_{\phi^*} = \begin{bmatrix} 0 \\ 0 \\ 0.1 \end{bmatrix},$$

the invariant pmf is $\pi_{\phi^*} = [1, 6, 6]/13$ and average cost $\eta^* = \pi_{\phi^*}(c_{\phi^*}) = 3/65$.

Since the optimal policy for the average-cost problem is unique, it follows that ϕ^* is also optimal for the γ-discounted problem for all γ sufficiently close to 1.

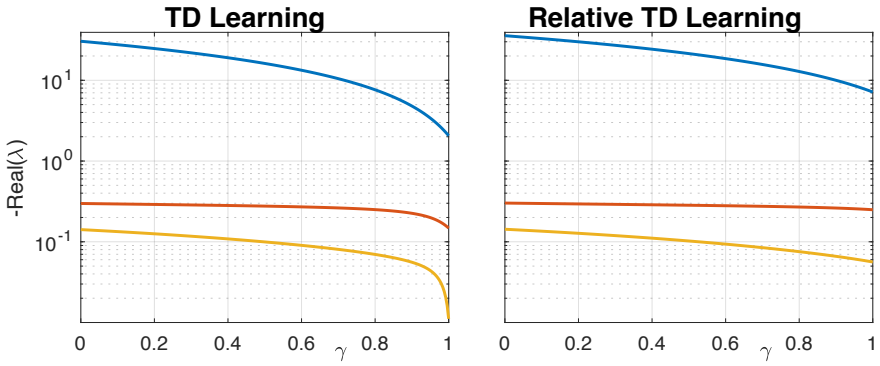

Figure 2: Eigenvalues as a function of discount factor γ. One eigenvalue approaches zero for the algorithm using $\delta_{\mathsf{r}} = 0$.

For the randomized policy shown in the figure, the induced transition matrix and one-stage cost are

$$P_{\widetilde{\phi}} = \begin{bmatrix} 0.64 & 0.36 & 0 \\ 0.05 & 0.60 & 0.35 \\ 0.08 & 0.04 & 0.78 \end{bmatrix}, \qquad c_{\widetilde{\phi}} = \begin{bmatrix} 0.16 \\ 0.20 \\ 0.72 \end{bmatrix}$$

The invariant pmf is $\pi_{\widetilde{\phi}} = [85, 108, 315]/508$, and hence the average cost is far greater than η^*:

$$\eta^{\widetilde{\phi}} = \pi_{\widetilde{\phi}}(c_{\widetilde{\phi}}) = \frac{587}{1016} \approx 0.578.$$

The 3-dimensional basis vector was chosen to be $\psi(x, u) = [x;\ u;\ xu]$, which satisfies $\Sigma(0) > 0$.

The numerical results that follow were based on $M = 200$ independent runs of length $N = 10^6$. For histograms we increased the number of samples to $n_{\mathsf{s}} \times M$ by selecting $n_{\mathsf{s}} = 5$ parameter estimates from the final 80% of each run. These were chosen approximately uniformly on the natural time scale associated with SA approximation theory, resulting in

$$N_i = \left(N_1^{1-\rho} + \frac{i-1}{n_{\mathsf{s}} - 1} \big(N_{n_{\mathsf{s}}}^{1-\rho} - N_1^{1-\rho} \big) \right)^{\frac{1}{1-\rho}}, \quad 1 \le i \le n_{\mathsf{s}}\,, \tag{26}$$

with $N_{n_{\mathsf{s}}} = N$ and $N_1 = 0.8N$. Explanation is provided in the Appendix.

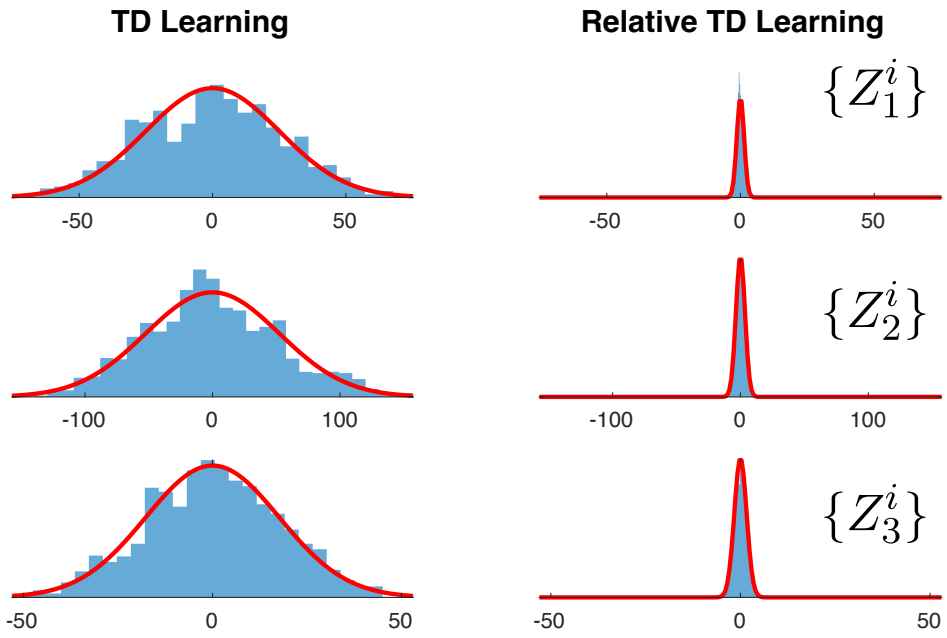

Figure 3: Histograms of the scaled and centered parameter estimates for TD(0) and ϖ-RTD(0) in the finite state-action model using Polyak–Ruppert averaging. The red solid curves show the corresponding theoretical Gaussian approximations predicted by (4) and $\{\mathcal{Z}_j^{i,m}\}$ denote the samples used to form the histogram of the j-th component.

It is known that PR averaging not only achieves the optimal asymptotic covariance, but also ensures that the CLT holds [4]. To illustrate this conclusion we construct a histogram of the centered and scaled errors: For each $1 \le m \le M$ and $1 \le i \le n_{\mathsf{s}}$ (recall (26)), denote by $\theta_{i,m}^{\mathsf{PR}}$ the value of (12) with the value of N replaced by $N_i > N_0$ in the mth independent run. The empirical mean is given by

$$\bar{\theta}^{\mathsf{PR}} = \frac{1}{M n_{\mathsf{s}}} \sum_{m=1}^{M} \sum_{i=1}^{n_{\mathsf{s}}} \theta_{i,m}^{\mathsf{PR}}\,, \tag{27}$$

Consider the centered and scaled error,

$$\mathcal{Z}^{i,m} = \sqrt{N_i - N_0} \big(\theta_{i,m}^{\mathsf{PR}} - \bar{\theta}^{\mathsf{PR}} \big)\,, \tag{28}$$

in which $1 \le m \le M$, and $1 \le i \le n_{\mathsf{s}}$. The three histograms obtained from the three components of $\{\mathcal{Z}^{i,m}\} \subset \mathbb{R}^3$ are shown in Fig. 3. Each appears to be consistent with a Gaussian random vector with covariance (4) (which was computed exactly based on model parameters). This accuracy is less tight with time horizons below $N = 10^5$.

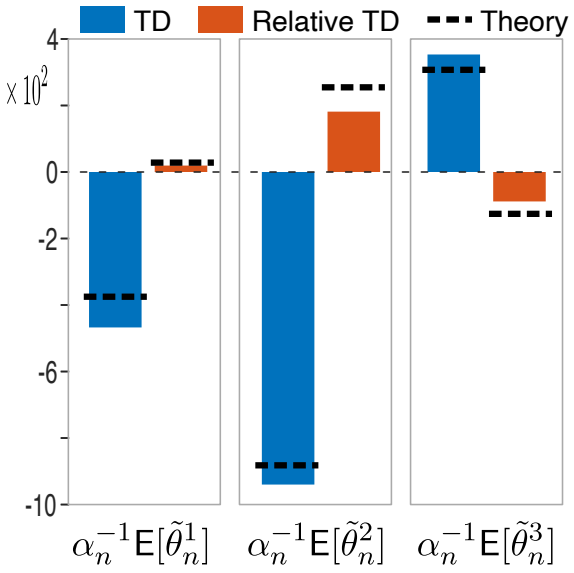

Figure 4: Bias realized vs theory.

Fig. 2 shows the eigenvalues of $\bar{A}$ as a function of the discount factor. For the standard TD(0) algorithm, the matrix $\bar{A}$ becomes close to singular as $\gamma \uparrow 1$, resulting in poor conditioning. The ϖ-RTD(0) algorithm avoids this behavior.

Fig. 4 shows the components of the empirical mean of $\{[\theta_n^i - \theta^*]/\alpha_n : 1 \le i \le M\}$ with $n = 10^6$ and $M = 200$ using ϖ-RTD learning with $\delta_{\mathsf{r}} = 1/2$. The theoretical value is obtained from the asymptotic bias formula (5).

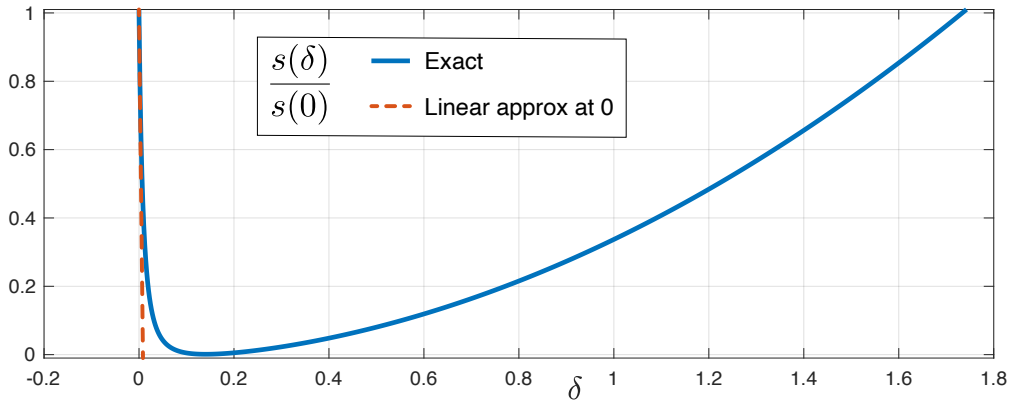

Figure 5: Squared-norm of normalized bias as a function of δ_{r}. The red dashed line is the predicted slope

The plot of $s(\delta) = \|\beta_\theta(\delta)\|^2$ shown in Fig. 5 is also obtained from (5): The red dashed line has slope $s'(0) = 2\beta_\theta^\intercal \beta'_\theta$ where β'_θ is given in (25b).

3.2 Speed scaling

The second example is the speed-scaling model, which serves as a benchmark for studying reinforcement learning with continuous state and action spaces. The state space is $\mathsf{X} = \mathbb{R}_+$ and the state-dependent action space is $\mathsf{U}(x) = [0, x]$. The dynamics are given by

$$X_{k+1} = X_k - U_k + A_{k+1} \tag{29}$$

where $\{A_k\}$ is i.i.d. with finite mean and variance. The state X_k represents the workload in a queue, and the control U_k denotes the service rate. See [12, Sec. 7.4] for history.

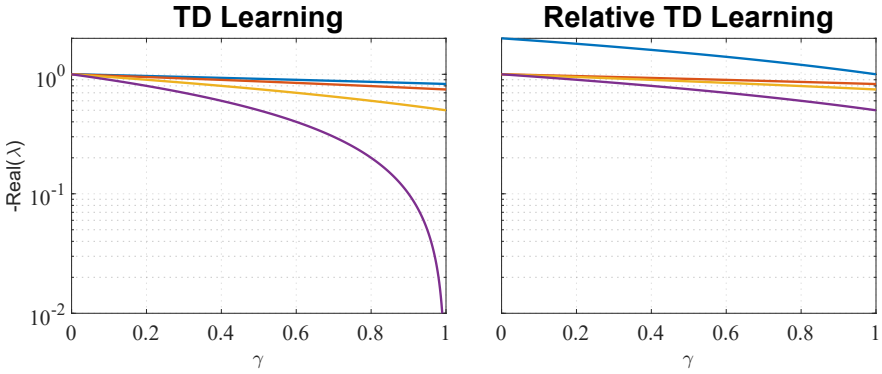

Figure 6: Eigenvalues as a function of discount factor γ for the speed-scaling model.

The one-stage cost is selected to be

$$c(x, u) = x + \frac{r}{2}u^2\,, \tag{30}$$

where $r > 0$ is a design parameter. This cost captures the trade-off between delay and power consumption: increasing the service rate reduces the queue length but incurs higher instantaneous cost.

While the main results in this paper have been established only for finite state and action models, most results easily extend to this general setting since all of the SA theory cited has been established for a general state space Markov chain $\mathbf{\Phi}$ in the recursion (8).

Our goal is to estimate the fixed policy Q-function with policy $u(x) = \kappa x$ with $0 < \kappa < 1$; The value $\kappa = 0.5$ was used in numerical experiments, $\{A_k\}$ was chosen to be i.i.d. with a Gamma distribution having mean 5 and variance 10, and $r = 10$ was used to define the cost function (30).

We adopt a linear function approximation architecture to approximate the state-action value function using the 4-dimensional basis $\psi(x, u) = [c(x, u); x^{3/2}; -(1 + \sqrt{x})u; 1 - u/\sqrt{x+1}]$. The basis is motivated by an approximation of $Q^\star$, devised through two steps. First, (21) admits the extension $Q^\star(x, u) = H^\star(x, u) + \eta^*/(1 - \gamma) + o(1 - \gamma)$ in which η^* is the optimal average cost, and $H^\star$ is the relative value function solving the average cost optimality equations. Approximation theory based on a fluid model reveals that $H^\star$ grows as $x^{3/2}$ for large $x > 0$. For the linear policy under consideration, the fixed policy Q-function grows as x^2 rather than $x^{3/2}$; however, we do not expect to obtain a tight approximation uniformly over Z, so we opt for a basis similar to what has been used in prior work.

Finally, we chose $\gamma = 0.9$, and for ϖ-RTD(0) we set $\delta_r = 1$. The step-size sequence was selected according to (11) with $\alpha_0 = 0.2$ and $\rho = 0.6$.

The main results are illustrated here for the case where $\lambda = 0$. For RTD, the baseline distribution is taken to be the empirical distribution (ϖ-RTD(0) learning). Polyak–Ruppert averaging is applied

in both cases for variance reduction, with run-length $N = 10^7$ and 20% of the trajectory treated as burn-in, so that $N_{\text{burn}} = 0.2N$. To estimate asymptotic covariance and bias we ran $M = 100$ independent simulations. For histograms we increased the number of samples to $100 \times M$ by selecting $n_{\mathsf{s}} = 100$ parameter estimates from the final 20% of each run (recall (26)).

Fig. 6 shows the eigenvalues of $\bar{A}$ as a function of the discount factor γ for both TD and RTD in the speed-scaling model, showing that the condition number of $\bar{A}$ is uniformly bounded over $0 \le \gamma \le 1$ for RTD(0).

Fig. 7 illustrates that the Polyak–Ruppert averaged parameter estimates exhibit behavior consistent with the Gaussian limit predicted by the CLT, for both TD(0) and ϖ-RTD(0).

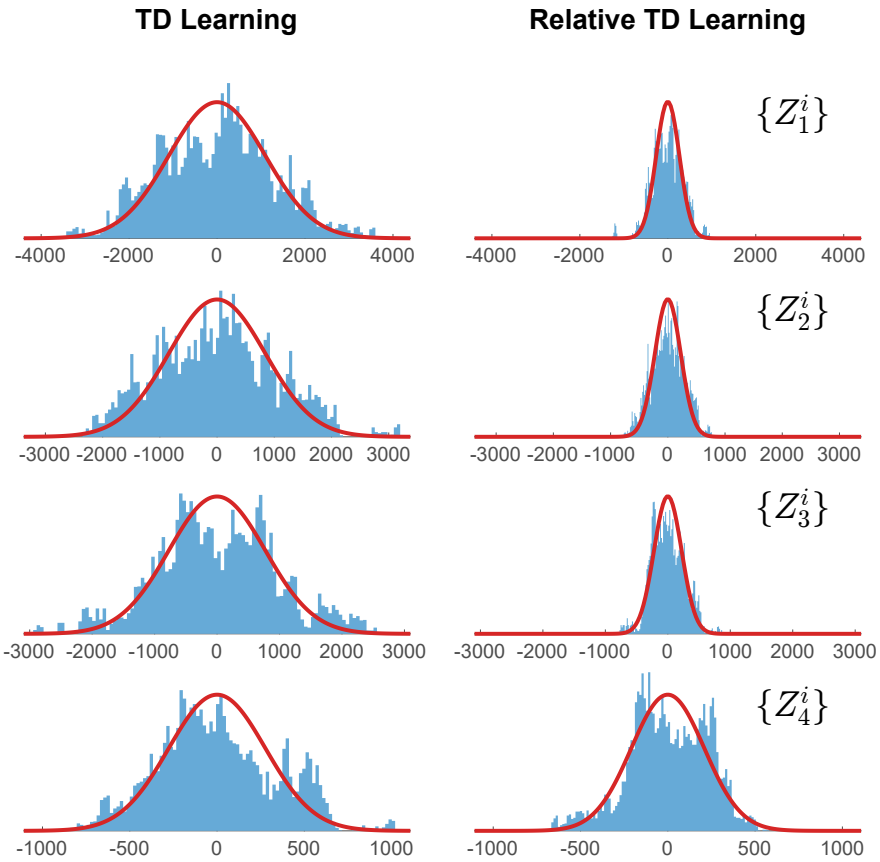

Figure 7: Histograms of the scaled and centered parameter estimates for TD(0) and ϖ-RTD(0) in the speed-scaling model using Polyak–Ruppert averaging. The red solid curves show the corresponding theoretical Gaussian approximations predicted by (4) and $\{\mathcal{Z}_j^{i,m}\}$ denote the samples used to form the histogram of the j-th component.

3.3 Discussion.

The simulation results in both examples are consistent with the theoretical analysis of stability, covariance, and bias developed in Section 2. In particular, Thm. 2.4 predicts that the ϖ-RTD(0) algorithm maintains a Hurwitz matrix $\bar{A}$ with uniformly bounded condition number. The uniform bound for ϖ-RTD(0) is illustrated in Figs. 2 and 6, for both the finite-state and speed-scaling examples. In both examples, the standard TD(0) algorithm exhibits an eigenvalue approaching zero as $\gamma \uparrow 1$, while the ϖ-RTD(0) algorithm maintains eigenvalues bounded away from the right

half plane in $\mathbb{C}$, as predicted by Thm. 2.4.

Figs. 3 and 7 show that the asymptotic covariance of ϖ-RTD(0) is reasonably well captured by the Gaussian approximation based on (4). The variance reduction for ϖ-RTD(0), compared to TD(0), as observed in Figs. 3 and 7, is attributed to the larger condition number of $\bar{A}$ under TD(0) (see Fig. 2).

The bias results in the finite-state example are also consistent with theory. As observed in Fig. 4, the empirical normalized bias closely matches the theoretical prediction in Prop. 2.2. Moreover, as observed in Fig. 5, the slope of the squared bias at $\delta_r = 0$ is consistent with the sensitivity analysis established in Thm. 2.5.

4 Conclusions

This paper introduces a relative TD learning algorithm based on an empirical baseline distribution. In addition to stability and uniform bounds on the asymptotic covariance, it is also demonstrated that this choice of baseline distribution leads to improved conditioning of the mean dynamics, uniformly over the discount factor, and results in significantly reduced variance compared to standard TD learning. In addition, we also establish sufficient conditions for instability of the mean flow. The theoretical results are supported by simulation studies in both finite-state-action MDPs and MDPs formulated on general state and action spaces.

Future work will focus on extensions to policy iteration and Q-learning.

References

[1] J. Abounadi, D. Bertsekas, and V. S. Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

[2] E. Moulines and F. R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24*, pages 451–459, 2011.

[3] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency, Delhi, India, 2nd edition, 2021.

[4] V. Borkar, S. Chen, A. Devraj, I. Kontoyiannis, and S. Meyn. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *Annals of Applied Probability (preprint at arXiv e-prints:2110.14427)*, 35(2):936–982, April 2025.

[5] A. M. Devraj and S. P. Meyn. Zap Q-learning. In *Proc. of the Intl. Conference on Neural Information Processing Systems*, pages 2232–2241, 2017.

[6] A. M. Devraj and S. P. Meyn. Q-learning with uniformly bounded variance. *IEEE Trans. on Automatic Control*, 67(11):5948–5963, 2022.

[7] A. Durmus, E. Moulines, A. Naumov, and S. Samsonov. Finite-time high-probability bounds for Polyak–Ruppert averaged iterates of linear stochastic approximation. *Mathematics of Operations Research*, 2024.

[8] J. A. Fill. Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *The Annals of Applied Probability*, 1(1):62–87, 1991.

[9] C. K. Lauand and S. Meyn. Bias in stochastic approximation cannot be eliminated with averaging. In *Allerton Conference on Communication, Control, and Computing*, pages 1–4, Sep. 2022.

[10] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2 edition, 2017.

[11] P. Mehta and S. Meyn. Optimistic training and convergence of Q-learning – extended version. *arXiv 2602.06146*, 2026.

[12] S. Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, Cambridge, 2022.

[13] S. Meyn. The projected Bellman equation in reinforcement learning. *IEEE Transactions on Automatic Control*, 69(12):8323–8337, Dec 2024.

[14] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

[15] D. Ruppert. Efficient estimators from a slowly convergent Robbins-Monro processes. Technical Report Tech. Rept. No. 781, Cornell University, School of Operations Research and Industrial Engineering, Ithaca, NY, 1988.

[16] R. Srikant. Rates of convergence in the central limit theorem for markov chains, with an application to td learning. *Math. Oper. Res.*, 2025.

[17] R. Srikant and L. Ying. Finite-time error bounds for linear stochastic approximation and td learning. In *Proc. Mach. Learn. Res.*, volume 99, pages 2803–2830, 2019.

[18] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.

[19] C. Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

[20] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Control*, 42(5):674–690, 1997.

[21] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, Cambridge, UK, 1989.

[22] C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

A Appendices

The mean flow $\bar{A} := \bar{A}(\lambda; 0)$ for the standard on-policy TD(λ) learning algorithm is linear, with

$$\begin{aligned}\bar{A} &= -\sum_{k=0}^{\infty}(\lambda\gamma)^k R_k + \gamma\sum_{k=0}^{\infty}(\lambda\gamma)^k R(k+1) \\ &= -R(0) + (1-\lambda)\gamma\sum_{k=0}^{\infty}(\lambda\gamma)^k R(k+1)\end{aligned} \tag{31}$$

It was first established in [20] that $\bar{A}$ is Hurwitz when $R(0)$ is full rank. The conclusion that is commonly applied is

$$\theta^\intercal \bar{A}\theta \le -(1-\gamma)\theta^\intercal R(0)\theta\,, \quad \theta \in \mathbb{R}^d \tag{32}$$

whose proof follows from Cauchy-Schwarz inequality, giving $\theta^\intercal R(k)\theta \le \theta^\intercal R(0)\theta$ for any k.

In some cases, such as the tabular setting, the bound is nearly tight as $\gamma \uparrow 1$, which leads to the instability result Thm. 2.3 (ii).

A.1 Dirichlet forms and eigenvalues of $\bar{A}$

See [10, 8] for background on spectral gaps and Dirichlet forms for Markov chains. This theory is set in a Hilbert space setting, in which a transition matrix is regarded as a linear operator on $L^2 = L^2(\pi)$.

Introduce two scalars $\beta = \lambda\gamma$ and $\varrho = \gamma(1-\lambda)/(1-\beta) < 1$, so that (31) becomes

$$\begin{aligned}\bar{A} &= -(1-\varrho)R(0) - \varrho M_\beta \\ \textit{with}\quad M_\beta &= R(0) - (1-\beta)\sum_{k=0}^{\infty}\beta^k R(k+1)\end{aligned} \tag{33}$$

In the special case $\lambda = 0$ we obtain $M_\beta = R(0) - \gamma R(1)$. To improve (32) we obtain a lower bound on $\theta^\intercal M_\beta \theta$ for arbitrary $\beta \in [0, 1)$.

For each $\theta \in \mathbb{R}^d$ we associate a function $g_\theta \colon \mathsf{Z} \to \mathbb{R}$ via $g_\theta(z) = \theta^\intercal\psi(z)$ for $z \in \mathsf{Z} = \mathsf{X} \times \mathsf{U}$, defined so that $\sum_z \pi(z)g_\theta(z) = \mathsf{E}_\pi[g_\theta(Z)]$.

Consider the transition matrix

$$K_\beta := (1-\beta)\sum_{k=0}^{\infty}\beta^k P^{k+1}$$

Using the operator theoretic notation $K_\beta g\big|_z = \sum_j K_\beta(z, z')g(z')$ for functions $g \colon \mathsf{Z} \to \mathbb{R}$, we obtain,

$$\theta^\intercal M_\beta\theta = \langle g_\theta, [I - K_\beta]g_\theta\rangle_{L^2} := \sum_z \pi(z)g_\theta(z)\{[I-K_\beta]g_\theta\}\big|_z \tag{34}$$

A lower bound is obtained from standard spectral-gap inequalities for finite irreducible Markov chains.

For a transition matrix T with unique invariant pmf π, its $L^2(\pi)$ adjoint is expressed $T^*(z, z') = \pi(z')T(z', z)/\pi(z)$. Letting $\mathbf{\Psi}$ denote a stationary version of the Markov chain with this transition matrix, the associated *Dirichlet form* is defined for functions $g, h \colon \mathsf{Z} \to \mathsf{Z}$ by

$$\mathcal{E}_T(g, h) := \tfrac{1}{2}\mathsf{E}_\pi[(g(\Psi_1) - h(\Psi_0))^2]$$

When $g = h$ this may be expressed,

$$\mathcal{E}_T(g, g) = \langle g, (I-T)g\rangle_{L^2} = \langle g, (I-T^*)g\rangle_{L^2} \tag{35}$$

Lemma A.1. *Let T denote an arbitrary transition matrix on Z with unique invariant pmf π, with spectral gap γ. If T is reversible then the Poincaré inequality holds: for any $g\colon \mathsf{Z} \to \mathbb{R}$,*

$$\mathcal{E}_T(g,g) \geq \gamma \|\tilde{g}\|^2_{L^2(\pi)}\,, \qquad \text{where } \tilde{g} = g - \pi(g).$$

Application of the lemma is made possible through consideration of the "reversibilization" $S_\beta = (K_\beta + K_\beta^*)/2$, where K_β^* is the $L^2(\pi)$ adjoint of K_β. The identities in (35) imply that $\mathcal{E}_{K_\beta}(g,g) = \mathcal{E}_{S_\beta}(g,g)$, allowing us to apply Lemma A.1 to obtain a lower bound on $\theta^\intercal M_\beta \theta$ that is independent of β.

Let $\gamma_\beta > 0$ denote the spectral gap of S_β.

Proposition A.2. *Suppose that $\boldsymbol{Z}$ is unichain and aperiodic. Then,*

(i) *For $\theta \in \mathbb{R}^d$,*

$$\theta^\intercal M_\beta \theta \geq \gamma_\beta\, \theta^\intercal \Sigma(0) \theta \tag{36a}$$

(ii) $\bar{\gamma}_P = \inf\{\gamma_\beta : 0 \leq \beta < 1\}$ *is strictly positive.*

(iii) *For any non-negative pair λ, γ statisfying $\lambda\gamma < 1$ the matrix $\bar{A}$ in (33) satisfies the bound*

$$\theta^\intercal \bar{A} \theta \leq -\bar{\gamma}_P \theta^\intercal \Sigma(0) \theta\,, \quad \theta \in \mathbb{R}^d \tag{36b}$$

Proof Observe that γ_β is continuous as a function of β, with $\gamma_\beta > 0$ for $0 \leq \beta < 1$ and $\gamma_\beta(\beta) \uparrow 1$ as $\beta \uparrow 1$. Conclusion (ii) thus follows, i.e., $\bar{\gamma}_P > 0$.

As for (iii), the bound (36a) combined with (33) imply (36b):

$$\theta^\intercal \bar{A} \theta = -(1-\varrho)\theta^\intercal R(0)\theta - \varrho \theta^\intercal M_\beta \theta \leq -\bar{\gamma}_P \theta^\intercal \Sigma(0)\theta$$

where the second inequality uses $R(0) \geq \Sigma(0)$. Hence the proof is complete on establishing (i).

Since S_β is reversible with respect to π, Lemma A.1 implies

$$\langle g, (I - S_\beta) g \rangle_{\widetilde{\Phi}} \geq \gamma_\beta \|\tilde{g}\|^2_{L^2(\pi)}$$

Applying this with g_θ gives the desired bound in (i):

$$\theta^\intercal M_\beta \theta \geq \gamma_\beta \|\tilde{g}_\theta\|^2_{L^2(\pi)} = \gamma_\beta\, \theta^\intercal \Sigma(0) \theta.$$

■

A.2 Proof of Thm. 2.3

Subject to the assumptions of Thm. 2.3, we will show shortly that when $\gamma = 1$ the matrix $\bar{A}$ defined in (33) has a simple eigenvalue at the origin, and all others in the strict left half plane. To prove the theorem we consider conditions under which this eigenvalue moves to the left or right left half plane for RTD(λ) learning for small $\delta_{\mathsf{r}} > 0$.

Lemma A.3. *Let $\Xi \in \mathbb{R}^{d \times d}$ have a simple eigenvalue at 0, with eigenvector $\eta \neq 0$ such that $\Xi\eta = \eta^\intercal \Xi = 0$, normalized to $\|\eta\| = 1$.*

Let $v, w \in \mathbb{R}^d$ be nonzero and denote $\Xi_t := \Xi + t\, vw^\intercal$. Then there is $\varepsilon > 0$ and a differentiable function $\{\kappa_t : 0 \leq t \leq \varepsilon\}$ such that $\kappa_0 = 0$ and κ_t is an eigenvalue of Ξ_t for $0 \leq t \leq \varepsilon$. Its derivative at the origin may be expressed,

$$\kappa_0' = (\eta^\intercal v)(w^\intercal \eta)$$

Proof An application of the implicit function theorem implies the existence of $\varepsilon > 0$ such that there is an eigenvalue-eigenvector pair (λ_t, η^t) for Ξ_t for $0 \le t \le \varepsilon$:

$$\Xi_t \eta_t = \kappa_t \, \eta_t \,, \tag{37}$$

with initialization $\lambda_0 = 0$ and $\eta^0 = \eta$, and with $\|\eta^t\| = 1$ for each t.

Differentiate (37) at $t = 0$, use $\kappa_0 = 0$ and apply the product rule to obtain

$$(vw^\intercal)\eta + \Xi\eta_0' = \kappa_0' \, \eta + \kappa_0 \, \eta_0' = \kappa_0' \, \eta,$$

Left-multiply by $\eta^\intercal$ to obtain

$$\eta^\intercal(vw^\intercal)\eta + \eta^\intercal\Xi\eta_0' = \kappa_0' \, \eta^\intercal\eta.$$

Using $\eta^\intercal\Xi = 0$ and $\eta^\intercal\eta = 1$ gives $\kappa_0' = \eta^\intercal(vw^\intercal)\eta = (\eta^\intercal v)(w^\intercal\eta)$. ∎

We have $\bar{A}(\lambda;\delta_{\mathsf{r}}) = \bar{A}(\lambda;0) - \frac{\delta_{\mathsf{r}}}{1-\lambda\gamma}\bar{\psi}[\bar{\psi}^{\mu}]^\intercal$ for RTD(λ) learning, with $\bar{A} := \bar{A}(\lambda;0)$ defined in (33). This expression invites the application of Lemma A.3 with appropriate choice of Ξ together with $t = \delta_{\mathsf{r}}/(1-\lambda\gamma)$, $v = -\bar{\psi}$, and $w = \bar{\psi}^{\mu}$.

Under the assumption that $R(0) > 0$ it follows that the null space of $\Sigma(0)$ coincides with the linear span of vectors satisfying (22). This null space has dimension no more than one: if ξ^1, ξ^2 are linearly independent solutions to (22), then their difference is in the null space of $R(0)$.

Based on this structure we may apply Lemma A.3 with Ξ equal to $\bar{A}$ with $\gamma = 1$. The bound (36b) implies that Ξ has all eigenvalues in the strict left half plane, except a simple eigenvalue at zero with eigenvector ξ.

It follows from Lemma A.3 that there exists $\delta_{\mathsf{r}}^0 > 0$ and a differentiable eigenvalue branch $\kappa(\delta_{\mathsf{r}})$ of $\bar{A}(\lambda;\delta_{\mathsf{r}})$ for $\delta_{\mathsf{r}} \in [0, \delta_{\mathsf{r}}^0)$ satisfying $\kappa(0) = 0$, with derivative

$$\kappa_0' = (\xi^\intercal v)(w^\intercal\xi) = -(\xi^\intercal\bar{\psi})(\xi^\intercal\bar{\psi}^{\mu}).$$

Since ξ satisfies (22), we have $\xi^\intercal\bar{\psi} = 1 > 0$, and hence the sign of κ_0' is determined by $\xi^\intercal\bar{\psi}^{\mu}$.

(i) If $\xi^\intercal\bar{\psi}^{\mu} > 0$, then $\kappa_0' < 0$, and hence for $\delta_{\mathsf{r}} > 0$ sufficiently small the eigenvalue $\kappa(\delta_{\mathsf{r}})$ lies in the strict left half plane. Since all other eigenvalues of $\bar{A}$ are already in the strict left half plane, it follows by continuity that $\bar{A}(\lambda;\delta_{\mathsf{r}})$ is Hurwitz for all $\delta_{\mathsf{r}} \in (0, \delta_{\mathsf{r}}^0)$ and γ sufficiently close to 1.

(ii) If $\xi^\intercal\bar{\psi}^{\mu} < 0$, then $\kappa_0' > 0$, and hence for $\delta_{\mathsf{r}} > 0$ sufficiently small the eigenvalue $\kappa(\delta_{\mathsf{r}})$ lies in the strict right half plane. Since the eigenvalue at zero persists for γ sufficiently close to 1, this implies that there exists $\gamma^0 < 1$ and $\delta_{\mathsf{r}}^0 > 0$ such that $\bar{A}(\lambda;\delta_{\mathsf{r}})$ has an eigenvalue in the strict right half plane for all $\gamma \in [\gamma^0, 1)$ and $\delta_{\mathsf{r}} \in (0, \delta_{\mathsf{r}}^0)$. ∎

A.3 Proof of Thm. 2.4

Prop. A.2 tells us that $\bar{A}(\lambda;0)$ obtained from TD(λ) learning is Hurwitz for any discount factor, including $\gamma = 1$, provided $\Sigma(0) > 0$. Part (i) immediately follows.

To establish (ii), apply Prop. A.2 to obtain $\bar{\gamma}_P > 0$ such that

$$\theta^\intercal\bar{A}(\lambda;0)\theta \le -\bar{\gamma}_P\theta^\intercal\Sigma(0)\theta, \qquad \theta \in \mathbb{R}^d,$$

uniformly over all $\gamma, \lambda \in [0,1]$ satisfying $\gamma\lambda < 1$. If $\mu = \varpi$, then

$$\bar{A}(\lambda;\delta_{\mathsf{r}}) = \bar{A}(\lambda;0) - \frac{\delta_{\mathsf{r}}}{1-\lambda\gamma}\,\bar{\psi}\bar{\psi}^\intercal.$$

Hence,

$$\begin{aligned}\theta^{\mathsf{T}}\bar{A}(\lambda;\delta_{\mathsf{r}})\theta &= \theta^{\mathsf{T}}\bar{A}(\lambda;0)\theta - \frac{\delta_{\mathsf{r}}}{1-\lambda\gamma}(\theta^{\mathsf{T}}\bar{\psi})^2 \\ &\le -\bar{\gamma}_P\theta^{\mathsf{T}}\Sigma(0)\theta - \frac{\delta_{\mathsf{r}}}{1-\lambda\gamma}(\theta^{\mathsf{T}}\bar{\psi})^2.\end{aligned}$$

Since $\Sigma(0) > 0$, there exists $\varepsilon_0 > 0$ such that

$$\theta^{\mathsf{T}}\Sigma(0)\theta \ge \varepsilon_0\|\theta\|^2, \qquad \theta \in \mathbb{R}^d.$$

Since $\delta_{\mathsf{r}} \ge 0$, $\gamma, \lambda \in [0,1]$, and $\gamma\lambda < 1$, we have

$$-\frac{\delta_{\mathsf{r}}}{1-\lambda\gamma}(\theta^{\mathsf{T}}\bar{\psi})^2 \le 0.$$

Therefore,

$$\begin{aligned}\theta^{\mathsf{T}}\bar{A}(\lambda;\delta_{\mathsf{r}})\theta &\le -\bar{\gamma}_P\varepsilon_0\|\theta\|^2 - \frac{\delta_{\mathsf{r}}}{1-\lambda\gamma}(\theta^{\mathsf{T}}\bar{\psi})^2 \\ &\le -\bar{\gamma}_P\varepsilon_0\|\theta\|^2, \qquad \theta \in \mathbb{R}^d.\end{aligned}$$

uniformly over all $\delta_{\mathsf{r}} \ge 0$ and all $\gamma, \lambda \in [0,1]$ satisfying $\gamma\lambda < 1$. It follows that $\bar{A}(\lambda;\delta_{\mathsf{r}})$ is Hurwitz for every $\delta_{\mathsf{r}} \ge 0$. Moreover, the preceding bound implies that the smallest singular value of $\bar{A}(\lambda;\delta_{\mathsf{r}})$ is bounded away from zero uniformly over this parameter range, and hence $\bar{A}(\lambda;\delta_{\mathsf{r}})^{-1}$ is uniformly bounded. Since $\bar{A}(\lambda;\delta_{\mathsf{r}})$ is also uniformly bounded, it follows that

$$\sup \operatorname{cond}\big(\bar{A}(\lambda;\delta_{\mathsf{r}})\big) < \infty.$$

Finally, since

$$\Sigma^*_\Theta = \bar{A}(\lambda;\delta_{\mathsf{r}})^{-1}\Sigma_\Delta\bar{A}(\lambda;\delta_{\mathsf{r}})^{-\mathsf{T}},$$

the uniform bounds on $\bar{A}(\lambda;\delta_{\mathsf{r}})^{-1}$ and Σ_Δ imply

$$\sup \operatorname{trace}(\Sigma^*_\Theta) < \infty.$$

Consequently,

$$\sup\ \big[\operatorname{cond}(\bar{A}(\lambda;\delta_{\mathsf{r}})) + \operatorname{trace}(\Sigma^*_\Theta)\big] < \infty,$$

where the supremum is over all $\delta_{\mathsf{r}} \ge 0$ and all $\gamma, \lambda \in [0,1]$ satisfying $\gamma\lambda < 1$. This proves (ii). ∎

A.4 Proof of Thm. 2.5

We start with a technical results on the sensitivity of $\theta^*(\delta_{\mathsf{r}})$ and $\bar{A}(\delta_{\mathsf{r}})^{-1}$ to changes in δ_{r} at $\delta_{\mathsf{r}} = 0$. These results are used in the proof of Thm. 2.5.

Lemma A.4. *Provided $\bar{A}(0)$ is full rank, we have*

$$\begin{aligned}\frac{d}{d\delta_{\mathsf{r}}}\theta^*(\delta_{\mathsf{r}})\big|_{\delta_{\mathsf{r}}=0} &= \bar{A}^{-1}\bar{A}'\theta^* \\ \bar{A}' := \frac{d}{d\delta_{\mathsf{r}}}\bar{A}(\delta_{\mathsf{r}})\big|_{\delta_{\mathsf{r}}=0} &= -\bar{\psi}\bar{\psi}^{\mathsf{T}} \\ (\bar{A}^{-1})' := \frac{d}{d\delta_{\mathsf{r}}}\bar{A}(\delta_{\mathsf{r}})^{-1}\big|_{\delta_{\mathsf{r}}=0} &= \bar{\phi}\bar{\phi}^{\mathsf{T}}, \quad \textit{with } \bar{\phi} = \bar{A}^{-1}\bar{\psi}\end{aligned}$$

Proof The derivative formula for $\bar{A}$ follows from the identity $\bar{A}(\delta_{\mathsf{r}}) = \bar{A} - \delta_{\mathsf{r}}\bar{\psi}\bar{\psi}^{\intercal}$.

Since $\theta^*(\delta_{\mathsf{r}})$ is the equilibrium of the mean flow, we have

$$\bar{A}(\delta_{\mathsf{r}})\theta^*(\delta_{\mathsf{r}}) = \bar{b},$$

with $\bar{b}$ independent of δ_{r}. Differentiating with respect to δ_{r} gives

$$\bar{A}'\theta^* + \bar{A}\frac{d}{d\delta_{\mathsf{r}}}\theta^*(\delta_{\mathsf{r}})\Big|_{\delta_{\mathsf{r}}=0} = 0.$$

Hence

$$\frac{d}{d\delta_{\mathsf{r}}}\theta^*(\delta_{\mathsf{r}})\Big|_{\delta_{\mathsf{r}}=0} = -\bar{A}^{-1}\bar{A}'\theta^*.$$

The final identity is the quotient rule for differentiation of a matrix inverse:

$$(\bar{A}^{-1})' = \bar{A}^{-1}\bar{\psi}\bar{\psi}^{\intercal}\bar{A}^{-1} = \bar{\phi}\bar{\phi}^{\intercal}, \quad \textit{with } \bar{\phi} = \bar{A}^{-1}\bar{\psi}.$$

■

From (4), for each δ_{r},

$$\Sigma^*_{\Theta}(\delta_{\mathsf{r}}) = \bar{A}(\delta_{\mathsf{r}})^{-1}\,\Sigma_{\Delta}(\delta_{\mathsf{r}})\,\bar{A}(\delta_{\mathsf{r}})^{-\intercal}.$$

Differentiating with respect to δ_{r} and applying the product rule,

$$\begin{aligned}\frac{d}{d\delta_{\mathsf{r}}}\Sigma^*_{\Theta}(\delta_{\mathsf{r}}) = (\bar{A}^{-1})'\,\Sigma_{\Delta}\,\bar{A}^{-\intercal} + \bar{A}^{-1}\,\Sigma'_{\Delta}\,\bar{A}^{-\intercal} \\ + \bar{A}^{-1}\,\Sigma_{\Delta}\,(\bar{A}^{-\intercal})'.\end{aligned}$$

Using the identities

$$(\bar{A}^{-1})' = -\bar{A}^{-1}\bar{A}'\bar{A}^{-1}, \qquad (\bar{A}^{-\intercal})' = -\bar{A}^{-\intercal}(\bar{A}')^{\intercal}\bar{A}^{-\intercal},$$

we obtain

$$\begin{aligned}\frac{d}{d\delta_{\mathsf{r}}}\Sigma^*_{\Theta}(\delta_{\mathsf{r}}) = -\bar{A}^{-1}\bar{A}'\bar{A}^{-1}\Sigma_{\Delta}\bar{A}^{-\intercal} + \bar{A}^{-1}\Sigma'_{\Delta}\bar{A}^{-\intercal} \\ - \bar{A}^{-1}\Sigma_{\Delta}\bar{A}^{-\intercal}(\bar{A}')^{\intercal}\bar{A}^{-\intercal}.\end{aligned}$$

Substituting (4) into the first and third terms yields

$$\begin{aligned}\frac{d}{d\delta_{\mathsf{r}}}\Sigma^*_{\Theta}(\delta_{\mathsf{r}})\Big|_{\delta_{\mathsf{r}}=0} = \bar{A}^{-1}\Sigma'_{\Delta}\bar{A}^{-\intercal} \\ - \bar{A}^{-1}\bar{A}'\Sigma^*_{\Theta} - \big[\bar{A}^{-1}\bar{A}'\Sigma^*_{\Theta}\big]^{\intercal}.\end{aligned}$$

To obtain the bias sensitivity formula, recall from (5) that

$$\beta_{\theta} = \frac{1}{1-\rho}\,\bar{A}^{-1}\bar{\Upsilon}.$$

Differentiating yields

$$\frac{d}{d\delta_{\mathsf{r}}}\beta_{\theta}(\delta_{\mathsf{r}}) = \frac{1}{1-\rho}\Big[(\bar{A}^{-1})'\bar{\Upsilon} + \bar{A}^{-1}\bar{\Upsilon}'\Big].$$

Substituting $(\bar{A}^{-1})' = -\bar{A}^{-1}\bar{A}'\bar{A}^{-1}$ and using $\bar{A}^{-1}\bar{\Upsilon} = (1-\rho)\beta_{\theta}$, we obtain

$$\frac{d}{d\delta_{\mathsf{r}}}\beta_{\theta}(\delta_{\mathsf{r}})\Big|_{\delta_{\mathsf{r}}=0} = \frac{1}{1-\rho}\,\bar{A}^{-1}\big[-\bar{A}'\beta_{\theta} + \bar{\Upsilon}'\big]$$

■

A.5 Sub-sampling for histograms

The motivation for (26) comes from the ODE/SDE approximation for stochastic approximation.

An SA recursion can be interpreted as a noisy Euler approximation of the mean flow (3), for which the *natural time scale* is defined by the partial sums,

$$\tau(n) = \sum_{k=1}^{n} \alpha_k.$$

The ODE approximations of SA theory are expressed,

$$\theta_n \approx \vartheta^{(k)}_{\tau(n)}\,, \quad n \geq k \geq 0\,,$$

in which $\{\vartheta^{(k)}_t : t \geq \tau(k)\}$ denotes the solution of the mean flow (3), initialized at $\vartheta^{(k)}_{\tau(k)} = \theta_k$. For vanishing stepsize, the approximation becomes increasingly tight as $k \to \infty$ [3].

For step size $\alpha_n = cn^{-\rho}$, we obtain the approximation

$$\tau(n) \approx \frac{c}{1-\rho} n^{1-\rho}. \tag{38}$$

Hence, if snapshots are chosen uniformly on the natural time scale over the interval $[N_1, N_{n_{\mathsf{s}}}]$, then for $1 \leq i \leq n_{\mathsf{s}}$ the corresponding iteration indices are approximately given by (26). Approximate uniformity is imposed in an attempt to minimize the dependency between samples for a single run.