Table 1: Summary of Temporal Difference Learning. In below $n = |S|$ is the number of states.

Remarks	Problem			
	2-step	TD(0)	LLSR	TD+LFA (On Policy)
Data	$(s_t, s'_t)_{t=0}^{\infty}$ $s_t = s$, is fixed	$(s_t, r_t, s'_t)_{t=0}^{\infty}$ $s'_t = s_{t+1}$ $s'_t \overset{iid}{\sim} p_u(s_t, \cdot)$	$(X_i, y_i)_{i=1}^{n}$	$(X(s_t), r_t, X(s'_t))_{t=0}^{\infty}$ $s'_t = s_{t+1}$ $s'_t \overset{iid}{\sim} p_u(s_t, \cdot)$
Dist.	$s'_t \overset{iid}{\sim} p(s, \cdot)$	$d_u^\top = d_u^\top \mathbb{P}_u$	$\frac{1}{n}$	$d_u^\top = d_u^\top \mathbb{P}_u$
Aim	$V_* = \mathbf{E}[\sum_{t=0}^{1} R(s_t) \vert s_0 = s,$ $s_1 \overset{iid}{\sim} p(s, \cdot)]$	$V_u(s) = \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t r_t \vert s_0 = s,$ $s_{t+1} \sim p_u(s_t, \cdot)]$	$\min_\theta \vert\vert X^\top \theta - Y \vert\vert_2^2$	$V_u \approx X^\top \theta$
Sol.	$V_* = R(s) + \sum_{s' \in S} p(s, s') R(s')$	$V_u = (I - \gamma \mathbb{P}_u)^{-1} R$	$\theta_* = (X X^\top)^{-1} X Y$	
Algo.	$V_{t+1} = V_t + \alpha_t (R(s_t)$ $+\gamma R(s'_t) - V_t)$	$V_{t+1}(s_t) = V_t(s_t) + \alpha_t \big(R(s_t)$ $+\gamma V_t(s'_t) - V_t(s_t)\big)$	$\theta_{t+1} = \theta_t + \alpha_t X_t (Y_t$ $- X_t^\top \theta_t)$	$\theta_{t+1} = \theta_t + \alpha_t X_t (R(s_t)$ $+\gamma X(s'_t)^\top \theta_t - X(s_t)^\top \theta_t)$
$\theta \in$	$\mathbb{R}^1$	$\mathbb{R}^n$	$\mathbb{R}^d$	$\mathbb{R}^d$
A	1	$D_u(I - \gamma \mathbb{P}_u)$	$\frac{1}{n} X X^\top$	$X D_u (I - \gamma \mathbb{P}_u) X^\top$
b	V_*	$D_u R$	$\frac{1}{n} X Y$	XDuR
Stab.	Simple	Greshgorin (for any D)	Symmetric Positive Definite	To be proved

Table 2: ON Policy vs OFF Policy Comparison. First Two Columns are same as Table 1.

Remarks	Problem			
	TD(0)	TD+LFA (On Policy)		
Data	$(s_t, r_t, s'_t)_{t=0}^{\infty}$ $s'_t = s_{t+1}$ $s'_t \overset{iid}{\sim} p_u(s_t, \cdot)$	$(X(s_t), r_t, X(s'_t))_{t=0}^{\infty}$ $s'_t = s_{t+1}$ $s'_t \overset{iid}{\sim} p_u(s_t, \cdot)$	$(s_t, r_t, a_t, s'_t)_{t=0}^{\infty}$ $s'_t = s_{t+1}$ $s'_t \overset{iid}{\sim} p_{a_t}(s_t, \cdot), a_t \sim \mu(s_t, \cdot)$	$(X(s_t), r_t, a_t, X(s'_t))_{t=0}^{\infty}$ $s'_t = s_{t+1}$ $s'_t \overset{iid}{\sim} p_{a_t}(s_t, \cdot), a_t \sim \mu(s_t, \cdot)$
Dist.	$d_u^\top = d_u^\top \mathbb{P}_u$	$d_u^\top = d_u^\top \mathbb{P}_u$	$d_\mu^\top = d_\mu^\top \mathbb{P}_\mu$	$d_\mu^\top = d_\mu^\top \mathbb{P}_\mu$
Aim	$V_u(s) = \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t r_t \vert s_0 = s,$ $s_{t+1} \sim p_u(s_t, \cdot)]$	$V_u \approx X^\top \theta$	$V_\pi(s), \forall s \in S$	$V_\pi \approx X^\top \theta$
Sol.	$V_u = (I - \gamma \mathbb{P}_u)^{-1} R$	$A^{-1} b$	$V_\pi = (I - \gamma \mathbb{P}_\pi)^{-1} R$	
Algo.	$V_{t+1}(s_t) = V_t(s_t) + \alpha_t \big(R(s_t)$ $+\gamma V_t(s'_t) - V_t(s_t)\big)$	$\theta_{t+1} = \theta_t + \alpha_t X_t (R(s_t)$ $+\gamma X(s'_t)^\top \theta_t - X(s_t)^\top \theta_t)$	$V_{t+1}(s_t) = V_t(s_t) + \alpha_t \rho_t \big(R(s_t)$ $+\gamma V_t(s'_t) - V_t(s_t)\big)$	$\theta_{t+1} = \theta_t + \alpha_t \rho_t X_t (R(s_t)$ $+\gamma X(s'_t)^\top \theta_t - X(s_t)^\top \theta_t)$
$\theta \in$	$\mathbb{R}^n$	$\mathbb{R}^d$	$\mathbb{R}^d$	$\mathbb{R}^d$
A	$D_u(I - \gamma \mathbb{P}_u)$	$X D_u (I - \gamma \mathbb{P}_u) X^\top$	$D_\mu (I - \gamma \mathbb{P}_\pi)$	$X D_\mu (I - \gamma \mathbb{P}_\pi) X^\top$
b	$D_u R$	$X D_u R$	$D_\mu R$	$X D_\mu R$
Stab.	Greshgorin	To be proved	Greshgorin	Does not converge

Remarks:

- First level of understanding is to notice the dimensions and symbols. Sometimes (for TD(0)) we are using V and sometimes (LLSR, TD+LFA) we are using θ. The difference between use of θ and V is that, when V is used it means we are trying to calculate values for each and every state $V(s), \forall s \in S$, this is known as ***Tabular*** column representation or ***full*** state. When we are using θ we are trying to approximate the value by $V(s) \approx X(s)^\top \theta$ (***function approximation***). When we collect all the $V(s)$ for all the states we get a vector $V \in \mathbb{R}^n$. When we collect all the $X(s)^\top$ into a matrix $X^\top = \begin{bmatrix} {}^-X^\top(s^1)^- \\ {}^-X^\top(s^2)^- \\ \vdots \\ {}^-X^\top(s^n)^- \end{bmatrix}$, we use $V \approx X^\top \theta$.
- Note that X is a $d \times n$ matrix, and $X^\top$ is an $n \times d$ matrix. Further, we can recover the tabular column case, from the function approximate case, by choosing $d = n$ and $X = I_{n \times n}$, where $I_{n \times n}$ is the $n \times n$ identity matrix.

1 Error Analysis

Our aim is to analyse the algorithms in Tables 1 and 2.

Steps in the Analysis:

- First we have to write all the algorithms in Tables 1 and 2 in a common format. Luckily, we can actually do it, see Equation (1).
- Then we analyse the error e_t, which is the difference between the final solution and the current values.
- Put down the conditions under which the algorithms will actually converge to the solution.
- Study the behaviour of the error term e_t. It will tell us the mode of convergence.

1.1 Common format

All algorithms in Tables 1 and 2 can be written in the below format:

$$\theta_{t+1} = \theta_t + \alpha_t(b - A\theta_t + N_t) \tag{1}$$

How to obtain the common format in Equation (1): If we look carefully every algorithm is $\theta_{t+1} = \theta_t + \alpha_t(\cdot)$. Say that the terms inside the bracket is $(\cdot)$. Then we do $\cdot = \mathbb{E}[\cdot] + \cdot - \mathbb{E}[\cdot]$. In all the cases, $\mathbb{E}[\cdot] = b - A\theta_t$ (the exact b and A changes from algorithm to algorithms as show in Tables 1 and 2), and $N_t = \cdot - \mathbb{E}[\cdot]$ is a zero-mean noise term. ***If noise term was not there, then algorithm will be deterministic***, i.e., it will be the same from run to run.

1.2 Dynamics of Error

We have to make some assumptions to make our analysis.

Assumption 1.1. *A is invertible.*

To make our analysis simpler, we let $\alpha_t = \alpha > 0$, i.e., a fixed constant step-size, we will look at the error terms. The constant step-size is an important assumption, i.e., we can analyse for constant step-size and later infer what will happen if we change the step-size and in fact how to even change the step-sizes (should we increase or decrease, if so, by how much etc).

Let $\theta_* = A^{-1}b$. Now define $e_t = \theta_t - \theta_*$, we need to study how this behaves.

$$\theta_{t+1} - \theta_* = \theta_t - \theta_* + \alpha\left(b - A(\theta_t - \theta_* + \theta_*) + N_t\right) \tag{2}$$

$$e_{t+1} \overset{(I)}{=} e_t + \alpha(b - Ae_t - A\theta_*) + \alpha N_t \tag{3}$$

$$e_{t+1} \overset{(II)}{=} (I - \alpha A)e_t + \alpha N_t \tag{4}$$

$$e_{t+1} \overset{(III)}{=} (I - \alpha A)^{t+1}e_0 + \alpha\sum_{s=0}^{t}(I - \alpha A)^s N_{t-s} \tag{5}$$

In the above:

(I) follows from applying definition of $e_t = \theta_t - \theta_*$.

(II) follows from $b = A\theta_*$.

(III) follows from unrolling the recursion

From Equation (5) we can infer that $(I - \alpha A)^{t+1}$ should not go to ∞ as $t \to \infty$. It is obvious that for all matrices this will not happen. So, we need further assumption to ensure convergence.

Assumption 1.2. *A has eigenvalues with positive real parts.*

Note: We will talk about eigenvalues and why they are useful a little bit later.

Assumption 1.2 ensures that the term $(I - \alpha A)^{t+1}$ does not blow up. However, it is not very clear to see how e_t actually behaves.

1.3 Learning the mean of a Radom variable

Let us say the mean of a random variable is θ_*, and each sample of the random variable is nothing by $\theta_* + N_t$. Now, we have $b = \theta_*$ and $A = 1$, so:

$$\theta_{t+1} = \theta_t + \alpha\left(\theta_t - \theta_* + N_t\right) \tag{6}$$

$$e_{t+1} \overset{(IV)}{=} (1-\alpha)^{t+1} e_0 + \alpha \sum_{s=0}^{t} (1-\alpha)^s N_{t-s} \tag{7}$$

where (IV) follows from the unrolling the recursion.

The mean behaviour of e_t:

$$\mathbb{E}[e_{t+1}] = \mathbb{E}[(1-\alpha)^{t+1} e_0 + \alpha \sum_{s=0}^{t} (1-\alpha)^s N_{t-s}] \tag{8}$$

$$\mathbb{E}[e_{t+1}] \overset{(V)}{=} (1-\alpha)^{t+1} e_0 \tag{9}$$

where (V) follows from the fact that $\mathbb{E}[N_t] = 0$.

The mean squared behaviour of e_t:

$$\mathbb{E}[e_{t+1}^2] \overset{(VI)}{=} \underbrace{\mathbb{E}[(1-\alpha)e_t^2]}_{a^2} + \underbrace{\mathbb{E}[\alpha^2 N_t^2]}_{b^2} + \underbrace{\mathbb{E}[\alpha \sum_{s=0}^{t} (1-\alpha)^s N_{t-s} (1-\alpha)^{t+1} e_t]}_{2ab} \tag{10}$$

$$\mathbb{E}[e_{t+1}] \overset{(VI)}{=} (1-\alpha)^{2(t+1)} e_0^2 + \alpha^2 \sigma^2 \underbrace{\sum_{s=0}^{t} (1-\alpha)^{2s}}_{\text{geometric series}} \tag{11}$$

$$\overset{(VII)}{<} \underbrace{\rho^t e_0^2}_{\text{forgetting initial condition}} + \underbrace{\alpha^2 \sigma^2 \frac{1}{1-\rho}}_{\text{effect of noise}} \tag{12}$$

where (VI) is like $(a+b)^2 = a^2 + b^2 + 2ab$, where $2ab$ is the cross term. The cross term is 0, because it other than N_t rest of the quantities are constants, and we know $\mathbb{E}[N_t] = 0$.

(VI) is unrolling the recursion and using $\sigma^2 = \mathbb{E}[N_t^2]$ to denote the variance of the noise term. Note that for $\alpha < 1$ we will have $\rho = (1-\alpha)^2$, and ρ^t will go down to 0 as t increases.

(VII) the inequality is because $\sum_{s=0}^{t} \rho^s < \sum_{s=0}^{\infty} \rho^s = \frac{1}{1-\rho}$

1.4 Connection between eigenvalues, recursion, learning the mean

The only difference between the expressions in Equation (7) and Equation (2) is that in the case of learning the mean the quantities are numbers and in the general case the quantities are vectors and matrices. In the learning the mean case, we have $b = 1$ and $A = 1$ which is a very special case.

Convergence: Notice that the $(1-\alpha)$ term is very similar to $(I - \alpha A)$. We know for $0 < \alpha < 2$, $(1-\alpha)^t \to 0$. We expect the general case to behave like the specific case, i.e., learning the mean in which $b = 1$, $A = 1$, and $(1 - \alpha A) = (1 - \alpha)$.

Growth of a general matrix M^t and eigenvalues: From linear algebra (Jordan decomposition) to be specific, we know that any $d \times d$ matrix can be expressed as

$$M = U \Lambda U^{-1} \tag{13}$$

$$M^t = \underbrace{U \Lambda U^{-1} U \Lambda U^{-1} \dots U \Lambda U^{-1}}_{t \text{ times}} \tag{14}$$

$$M^t = U \Lambda^t U^{-1}, \tag{15}$$

where Λ is a Jordan matrix, which contains the eigenvalues in the main diagonal and the entries immediately above the diagonal are either 0 or 1, and U contains the eigenvectors of M. Further, the M and Λ have same eigenvalues. One can show (exercise) that rate of growth of Λ^t depends on its largest eigenvalue. So, if the modulus of the largest eigenvalue is less than 1, both Λ^t and M^t will eventually go to the $\mathbf{0}$ matrix.

Growth of a symmetric matrix M^t and eigenvalues: A symmetric matrix has special structure, it can be decomposed as

$$M = U\Lambda U^{-1}, \tag{16}$$

$$\tag{17}$$

where $U^{-1} = U^\top$, i.e., $U^\top U = UU^\top = I$. Also, Λ is known to be a diagonal matrix.

Note: In our algorithms we have $\mathbf{M^t = (I - \alpha A)^t}$. This is why we need the eigenvalues of A to have positive real parts, if not, then $I - \alpha A$ will have an eigenvalue whose real part will be greater than 1 and the matrix $M^t = (I - \alpha A)^t$ will blow up.

1.5 Matrix decomposition and Error Recursion:

Let $A = U\Lambda U^{-1}$

$$U^{-1}e_{t+1} = U^{-1}(I - \alpha AUU^{-1})^{t+1}e_0 + U^{-1}\alpha \sum_{s=0}^{t}(I - \alpha A)^s UU^{-1}N_{t-s} \tag{18}$$

$$\eta_{t+1} = \Lambda^{t+1}\eta_0 + \alpha \sum_{s=0}^{t} \Lambda^s \zeta_{t-s}, \tag{19}$$

where $\eta_{t+1} = U^{-1}e_t$ and $\zeta_t = U^{-1}N_t$, are error in the new basis (given by the matrix U). Here, we use the fact that $\Lambda = U^{-1}AU$.

The speciality of Equation (19) is that it is like a diagonal system, i.e., the behaviour of the error in the new basis depends on Λ^t which is in turn dependent on the entires in its diagonal.