Learning to Predict by the Methods of Temporal Differences

Peng Lingwei

July 15, 2019

References

[1] Peter Dayan. The convergence of td(λ) for general λ. 1992.

[2] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, Aug 1988.

[3] John N Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. pages 1075–1081, 1997.

Contents

1 Learning to Predict by the Methods of Temporal Differences [2]

- *simgle-step*: all information about the correctness of each prediction is revealed at once.
- *multi-step*: correctness is not revealed until more than one step after the prediction is made.

In this paper, we will be concerned only with multi-step prediction problems.

We consider multi-step prediction problems in such sequences $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}, z$. For each observation-outcome sequence, we have predictions $P_1, P_2, \ldots, P_m$. P_i is a shorthand for $P_i(\mathbf{x_i}, \mathbf{w_i})$, which is an estimate of z.

In view of supervised-learning approach, the sequences is considered as $(\mathbf{x_1}, z, P_1), (\mathbf{x_2}, z, P_2), \ldots, (\mathbf{x_m}, z, P_m)$. So the empirical loss is

$$l_{sq} = \mathbb{E}_s\{(\mathbb{E}(z|s) - P(\mathbf{w})(s))^2\} = \frac{1}{2m}\sum_{t=1}^{m}(z - P_t)^2$$

If we use stochastic gradient descent method, we have

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha(z - P_t)\nabla_{\mathbf{w}_t} P_t$$

which means

$$\Delta\mathbf{w}_{t+1} = \alpha(z - P_t)\nabla_{\mathbf{w}_t} P_t \tag{1.1}$$

Let $P_{m+1} = z$, then

$$\begin{aligned}
\mathbf{w}_{m+1} =& \mathbf{w}_1 + \sum_{t=1}^{m}\alpha(z - P_t)\nabla_{\mathbf{w}_t} P_t \\
=& \mathbf{w}_1 + \sum_{t=1}^{m}\alpha\sum_{k=t}^{m}(P_{k+1} - P_k)\nabla_{\mathbf{w}_t} P_t \\
=& \mathbf{w}_1 + \sum_{k=1}^{m}\alpha(P_{k+1} - P_k)\sum_{t=1}^{k}\nabla_{\mathbf{w}_t} P_t \\
=& \mathbf{w}_1 + \sum_{t=1}^{m}\alpha(P_{t+1} - P_t)\sum_{k=1}^{t}\nabla_{\mathbf{w}_k} P_k
\end{aligned}$$

which means

$$\Delta\mathbf{w}_t = \alpha(P_{t+1} - P_t)\sum_{k=1}^{t}\nabla_{\mathbf{w}_k} P_k \tag{1.2}$$

If we update $\mathbf{w}$ immediately, it's too complex to analyse.

$$\Delta\mathbf{w}_t = \alpha\mathbf{w}_t^T(\mathbf{x}_{t+1} - \mathbf{x}_t)\sum_{k=1}^{t}\mathbf{x}_k + \alpha(\mathbf{x}_{t+1}^T\Delta\mathbf{w}_t\sum_{k=1}^{t}\mathbf{x}_k)$$

So we update $\mathbf{w}$ after each whole sequence.

$$\Delta\mathbf{w}_t = \alpha\mathbf{w}_0^T(\mathbf{x}_{t+1} - \mathbf{x}_t)\sum_{k=1}^{t}\mathbf{x}_k$$

$$\mathbf{w}_m = \mathbf{w}_0 + \sum_{t=1}^{m}\Delta\mathbf{w}_t$$

This equation can be computed incrementally.

Theorem 1.1. *On multi-step prediction problems, the linear $TD(1)$ pro cedure produces the same per-sequence weight changes as the Widrow-Hoff procedure.*

Definition 1.1. *($TD(\lambda)$)*

$$\Delta\mathbf{w}_t = \alpha(P_{t+1} - P_t)\sum_{k=1}^{t}\lambda^{t-k}\nabla_{\mathbf{w}_k} P_k \tag{1.3}$$

This also can be computed incrementally

$$\begin{aligned} e_{t+1} =& \sum_{k=1}^{t+1} \lambda^{t+1-k} \nabla_{\mathbf{w}_k} P_k \\ =& \nabla_{\mathbf{w}_{t+1}} P_{t+1} + \sum_{k=1}^{t} \lambda^{t+1-k} \nabla_{\mathbf{w}_k} P_k \\ =& \nabla_{\mathbf{w}_{t+1}} P_{t+1} + \lambda e_t \end{aligned}$$

We can get, when

$$\lambda = 0, \quad \Delta \mathbf{w}_t = \alpha (P_{t+1} - P_t) \nabla_{\mathbf{w}_t} P_t$$

$$\begin{aligned} \mathbf{w}_{m+1} =& \mathbf{w}_1 + \sum_{t=1}^{m} \alpha (P_{t+1} - P_t) \sum_{k=1}^{t} \lambda^{t-k} \nabla_{\mathbf{w}_k} P_k \\ =& \mathbf{w}_1 + \alpha \sum_{k=1}^{m} \nabla_{\mathbf{w}_k} P_k \sum_{t=k}^{m} (P_{t+1} - P_t) \lambda^{t-k} \\ & \text{(exchange the sum operations and the meanings of k and t)} \\ =& \mathbf{w}_1 + \alpha \sum_{t=1}^{m} \nabla_{\mathbf{w}_t} P_t \left\{ \lambda^{m-t} z + (1-\lambda) \sum_{k=t}^{m-1} \lambda^{k-t} P_{k+1} - P_t \right\} \\ =& \mathbf{w}_1 + \alpha \sum_{t=1}^{m} \nabla_{\mathbf{w}_t} P_t \left\{ (1-\lambda) \sum_{k=t}^{\infty} \lambda^{k-t} P_{k+1} - P_t \right\} \\ & (P_{m+1} = P_{m+2} = \cdots = z) \end{aligned}$$

This format is useful in the discussion of the convergence of $TD(\lambda)$ for General λ. [1]

We already have the proof of $TD(1)$ which is equal to the supervised learning method. Now we discuss the convergence of $TD(0)$.

Definition 1.2. *(The ideal predictions) Let μ be the probablity distribution of initial states, and Q be the probablity transfer matrix without terminal states.*

$$\mathbb{E}\{\mathbf{z}\} = \sum_{k=0}^{\infty} Q^k < \mu, \mathbf{z} >= (I - Q)^{-1} < \mu, \mathbf{z} > \tag{1.4}$$

For example, in seven states' random walk, we have $\mu = (0, 0, 0, 1, 0, 0, 0)^T$, and $Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

Proof. For simplicity, we only update $\mathbf{w}$ after each sequence.

$$\begin{aligned}\mathbf{w}_{n+1} =&\mathbf{w}_n + \sum_{t=1}^{m-1} \alpha(P_{n,t+1} - P_{n,t})\mathbf{x}_{n,t} + \alpha(z_n - P_{n,m})\mathbf{x}_{n,m}\\ =&\mathbf{w}_n + \alpha \sum_{i\in N}\sum_{j\in N} c_{ij}(P_{n,j} - P_{n,i})\mathbf{x}_{n,i} + \alpha \sum_{i\in N}\sum_{j\in T} c_{ij}(z_n - P_{n,i})\mathbf{x}_{n,i}\\ &(c_{ij} \text{ is the number of transfer from state i to state j})\\ \mathbb{E}\{\mathbf{w}_{n+1}|\mathbf{w}_n\} =&\mathbf{w}_n + \alpha \sum_{i\in N}\sum_{j\in N} d_i p_{ij}(P_{n,j} - P_{n,i})\mathbf{x}_{n,i} + \alpha \sum_{i\in N}\sum_{j\in T} d_i p_{ij}(\mathbb{E}z_n - P_{n,i})\mathbf{x}_{n,i}\end{aligned}$$

If we update w_n after each sequences, we can get simple format:

$$\mathbb{E}\{\mathbf{w}_{n+1}|\mathbf{w}_n\} = \mathbf{w}_n+\alpha \sum_{i\in N}\sum_{j\in N} d_i p_{ij}\mathbf{w}_n^T(\mathbf{x}_j-\mathbf{x}_i)\mathbf{x}_{n,i}+\alpha \sum_{i\in N}\sum_{j\in T} d_i p_{ij}(\mathbb{E}z_n-\mathbf{w}_n^T\mathbf{x}_i)\mathbf{x}_{n,i}$$

where d_i is the expected occurrence times of state i, and its easy to proof that

$$\mathbf{d}^T = \mu^T(I-Q)^{-1}.$$

Then we can reformat the equation into

$$\mathbb{E}\{\mathbf{w}_{n+1}|\mathbf{w}_n\} = \mathbf{w}_n + \alpha XD(<\mu, \mathbb{E}\mathbf{z}> + QX^T\mathbf{w}_n - X^T\mathbf{w}_n)$$

where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ and $D = diag(d_1, d_2, \ldots, d_n)$.

Taking both side's expectation, we have:

$$\mathbb{E}\{\mathbf{w}_{n+1}\} = \mathbb{E}\{\mathbf{w}_n\} + \alpha XD(<\mu, \mathbb{E}\mathbf{z}> + QX^T\mathbb{E}\{\mathbf{w}_n\} - X^T\mathbb{E}\{\mathbf{w}_n\})$$

On both side, we multiply X^T:

$$\begin{aligned}\mathbb{E}\{X^T\mathbf{w}_{n+1}\} =&\mathbb{E}\{X^T\mathbf{w}_n\} + \alpha X^TXD(<\mu, \mathbb{E}\mathbf{z}> + Q\mathbb{E}\{X^T\mathbf{w}_n\} - \mathbb{E}\{X^T\mathbf{w}_n\})\\ =&\sum_{k=0}^{n-1}(I - \alpha X^TXD(I-Q))^k \alpha X^TXD <\mu, \mathbb{E}\mathbf{z}>\\ &+ (I - \alpha X^TXD(I-Q))^n \mathbb{E}\{X^T\mathbf{w}_0\}\end{aligned}$$

If $\lim_{n\to\infty}(I - \alpha X^TXD(I-Q))^n = 0$, and D^{-1} and $(X^TX)^{-1}$ exist, then we have:

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\{X^T\mathbf{w}_{n+1}\} &= (I - (I - \alpha X^TXD(I-Q)))^{-1}\alpha X^TXD <\mu, \mathbf{z}>\\ &= (I-Q)^{-1} <\mu, \mathbf{z}>\end{aligned} \tag{1.5}$$

□

The full proof of $\lim_{n\to\infty}(I - \alpha X^TXD(I-Q))^n = 0$ is in the paper.Noticing that this proof also requires that D^{-1} and $(X^TX)^{-1}$ exist.

Why we need to show that $X^TXD(I-Q)$ has a full set of eigenvalues all of whose real parts are positive?

Proof. **Step1**: $S = D(I-Q) + (D(I-Q))^T$ is strictly diagonally dominant matrix with positive diagonal entries, so $D(I-Q)$ is positive definite. **Step2**: Let λ and $\mathbf{y}$ be any eigenvalue-eigenvector pair of matrix $X^T X D(I-Q)$, and $z = (X^T X)^{-1} y$. Then,

$$y^* D(I-Q)y = z^* X^T X D(I-Q)y = z^* \lambda y = \lambda z^* X^T X z = \lambda (Xz)^* Xz$$

So,

$$Re(y^* D(I-Q)y) = Re(\lambda (Xz)^* Xz)$$
$$a^T D(I-Q)a + b^T D(I-Q)b = (Xz)^* Xz Re(\lambda)$$
$$Re(\lambda) > 0$$

□

Here requires the indenpendence of the representation of X, not the sequences such as $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m, z)$

2 The Convergence of $TD(\lambda)$ for General λ [1]

$$\begin{aligned}\mathbf{w}^{(m+1)} =& \mathbf{w}^{(1)} + \alpha \sum_{t=1}^{m} \nabla_{\mathbf{w}^{(t)}} P_t \left\{ (1-\lambda) \sum_{k=t}^{\infty} \lambda^{k-t} P_{k+1} - P_t \right\} \\ & (P_{m+1} = P_{m+2} = \cdots = z)\end{aligned}$$

Similarly, we only update $\mathbf{w}_n$ after each sequence.So

Proof.

$$
\begin{aligned}
\mathbf{w}_{n+1} =& \mathbf{w}_n + \alpha \sum_{t=1}^{m} \mathbf{x}^{(t)} \left\{ \lambda^{m-t} z + (1-\lambda) \sum_{k=t}^{m-1} \lambda^{k-t} P_{k+1} - \mathbf{w}_n^T \mathbf{x}^{(t)} \right\} \\
=& \mathbf{w}_n + \alpha \sum_{i \in N} c_i \mathbf{x}_i \left(\begin{bmatrix} (1-\lambda) \sum_{j \in N} c_{ij|i} \mathbf{w}_n^T \mathbf{x}_j + \sum_{j \in T} c_{ij|i} z_j \\ +(1-\lambda)\lambda (\sum_{k \in N} \sum_{j \in N} c_{ikj|i} \mathbf{w}_n^T \mathbf{x}_j) + \lambda \sum_{k \in N} \sum_{j \in T} c_{ikj|i} z_j \\ + \dots \end{bmatrix} - \mathbf{w}_n^T \mathbf{x}_i \right) \\
\mathbb{E}\{\mathbf{w}_{n+1} | \mathbf{w}_n\} =& \mathbf{w}_n + \alpha \sum_{i \in N} d_i \mathbf{x}_i \left(\begin{bmatrix} (1-\lambda) \sum_{j \in N} p_{ij} \mathbf{w}_n^T \mathbf{x}_j + \sum_{j \in T} p_{ij} z_j \\ +(1-\lambda)\lambda (\sum_{k \in N} \sum_{j \in N} p_{ikj} \mathbf{w}_n^T \mathbf{x}_j) + \lambda \sum_{k \in N} \sum_{j \in T} p_{ikj} z_j \\ + \dots \end{bmatrix} - \mathbf{w}_n^T \mathbf{x}_i \right) \\
=& \mathbf{w}_n + \alpha \sum_{i \in N} d_i \mathbf{x}_i \left(\begin{bmatrix} (1-\lambda) Q \mathbf{x}^T \mathbf{w}_n + h \\ +(1-\lambda)\lambda (Q^2 \mathbf{x}^T \mathbf{w}_n) + \lambda Q h \\ + \dots \end{bmatrix} - \mathbf{w}_n^T \mathbf{x}_i \right) \\
=& \mathbf{w}_n + \alpha X D \left\{ [(I - \lambda Q)^{-1} h + (1-\lambda)(I - \lambda Q)^{-1} Q X^T \mathbf{w}^n] - X^T \mathbf{w}_n \right\}
\end{aligned}
$$

$(h =< \mathbf{p}, \mathbf{z} >)$

$$
\begin{aligned}
\mathbb{E}\{X^T \mathbf{w}_{n+1}\} =& \left\{ I - \alpha X^T X D [I - (1-\lambda)(I - \lambda Q)^{-1} Q] \right\} \mathbb{E}\{X^T \mathbf{w}_n\} \\
& + \alpha X^T X D (I - \lambda Q)^{-1} h \\
=& \left\{ I - \alpha X^T X D [I - (1-\lambda)(I - \lambda Q)^{-1} Q] \right\}^n \mathbb{E}\{X^T \mathbf{w}_1\} \\
& + \sum_{k=0}^{n} \left\{ I - \alpha X^T X D [I - (1-\lambda)(I - \lambda Q)^{-1} Q] \right\}^k \\
& \alpha X^T X D (I - \lambda Q)^{-1} h \\
\lim_{n \to \infty} \mathbb{E}\{X^T \mathbf{w}_n\} =& \left\{ \alpha X^T X D [I - (1-\lambda)(I - \lambda Q)^{-1} Q] \right\}^{-1} \alpha X^T X D (I - \lambda Q)^{-1} h \\
=& \left\{ I - (1-\lambda)(I - \lambda Q)^{-1} Q \right\}^{-1} (I - \lambda Q)^{-1} h \\
=& \left\{ (I - \lambda Q)[I - (1-\lambda)(I - \lambda Q)^{-1} Q] \right\}^{-1} h \\
=& (I - Q)^{-1} h
\end{aligned}
\tag{2.1}
$$

□

Theorem 2.1.

$$
\lim_{n \to \infty} \left\{ I - \alpha X^T X D [I - (1-\lambda)(I - \lambda Q)^{-1} Q] \right\}^n = 0 \tag{2.2}
$$

Proof.

$$
\begin{aligned}
S^r =& D(I-Q^r)+[D(I-Q^r)]^T \\
S^r_{ii} =& \begin{cases} 2d_i(1-Q^r_{ii})>0, & i=j \\ -d_iQ^r_{ij}-d_jQ^r_{ji}<0, & i\neq j \end{cases} \\
\sum_j S^r_{ij} =& d_i\left(1-\sum_j Q^r_{ij}\right)+[\mu^T(I-Q)^{-1}(I-Q^r)]_i \\
=& d_i\left(1-\sum_j Q^r_{ij}\right)+[\mu^T(I+Q+Q^2+\cdots+Q^{r-1})]_i \geq 0 \\
(1-\lambda)\sum_{r=1}^{\infty}\lambda^{r-1}S^r =& D(I-(1-\lambda)\sum_{r=1}^{n}\lambda^{t-1}Q^r)+\left[D(I-(1-\lambda)\sum_{r=1}^{n}\lambda^{t-1}Q^r)\right]^T \\
=& D(I-(1-\lambda)(I-\lambda Q)^{-1}Q)+\left[D(I-(1-\lambda)(I-\lambda Q)^{-1}Q)\right]^T \\
\succ& 0(\text{strictly diagonally dominant matrix})
\end{aligned}
\tag{2.3}
$$

Lefting to proof that the matrix has a full set of eigenvalues all of whose real parts are positive. □

3 An Analysis of Temporal-Difference Learning with Function Approximation [3]

3.1 Definition of Temporal-Difference Leraning

1. Markov chains:
   - State space: $S=\{s^1,\ldots,s^n\}$
   - Markov chain: $\{s_t|t=0,1,\ldots\}$
   - State transition matrix: $P=[p_{ij}]$
   - Reward matrix: $R=[r_{ij}]$
   - Cost-to-go function: $V^*(s)=\mathbb{E}[\sum_{t=0}^{\infty}\alpha^t r_{s_t,s_{t+1}}|s_0=s]=\sum_{t=0}^{\infty}(\alpha P)^t\langle P,R\rangle$
2. $TD(\lambda)$:
   - State matrix: $\Phi=[\phi(s_1),\phi(s_1),\ldots,\phi(s_n)]$
   - Approximation function: $V(\mathbf{w}_t)$
   - Linear approximation function $V(\mathbf{w}_t)=\Phi^T\mathbf{w}_t$
   - Temporal difference: $d_t=r_{s_t,s_{t+1}}+\alpha V(\mathbf{w}_t)(s_{t+1})-V(\mathbf{w}_t)(s_t)$
   - Update rule: $\mathbf{w}_{t+1}=\mathbf{w}_t+\gamma_t d_t\sum_{k=0}^{t}(\alpha\gamma)^{t-k}\nabla_{\mathbf{w}_t}V(\mathbf{w}_t)(s_k)$

- Eligibility vector:

$$\begin{aligned}\mathbf{z}_t =& \sum_{k=0}^{t} (\alpha\lambda)^{t-k} \nabla V(\mathbf{w}_t)(s_k) \\ =& \sum_{k=0}^{t} (\alpha\lambda)^{t-k} \phi(s_k) \\ =& \phi(s_t) + \alpha\lambda \mathbf{z}_{t-1} \quad (\mathbf{z}_{-1} = 0)\end{aligned} \tag{3.1}$$

3.2 Understanding Temporal-Difference Learning

1. Inner product space concepts and notation:
   - Steady-state: $\pi = \{\pi(1), \ldots, \pi(n)\}$, s.t. $(\pi^T P = \pi^T)$
   - Diagonal matrix: $D = diag(\pi)$
   - The set of vectors: $L_2(S, D) = \{V \in R^n | \|V\|_D < \infty\}$
   - Projection matrix: $\Pi = \Phi^T (\Phi D \Phi^T)^{-1} \Phi D$,
     where $\Pi J = \arg \min_{J' \in \{\Phi\mathbf{w} | \mathbf{w} \in R^K\}} \|J - J'\|_D$
2. The $TD(\lambda)$ Operator:
   - $TD(\lambda)$ operator:

$$\begin{aligned}(T^{(\lambda)}V)(s) =& (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \mathbb{E}\left[\sum_{t=0}^{m} \alpha^t r_{s_t, s_{t+1}} + \alpha^{m+1} V(s_{m+1}) | s_0 = s\right] \\ =& (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} (\alpha P)^t \langle P, R\rangle + (\alpha P)^{m+1} V\right)(s)\end{aligned} \tag{3.2}$$

   - For $\lambda < 1$ and $\lambda \to 1$

$$(T^{(1)}V)(s) = \lim_{\lambda \uparrow 0} (T^{(\lambda)}V)(s) = V^*(s)$$

   - Def: $\Delta(\mathbf{w}_t, X_t = (s_t, s_{t+1}, z_t)) = (r_{s_t, s_{t+1}} + \alpha V(\mathbf{w}_t)(s_{t+1}) - V(\mathbf{w}_t)(s_t))\mathbf{z}_t$
     then, $\mathbf{w}_{t+1} = \mathbf{w}_t + \gamma_t \Delta(\mathbf{w}_t, X_t)$
   - We can proof that:
     $\lim_{t\to\infty} \mathbb{E}_0[\Delta(\mathbf{w}, X_t)] = \lim_{t\to\infty} \mathbb{E}_{s\sim\pi}[\Delta(\mathbf{w}, X_t)|X_0] = \Phi D(T^{(\lambda)}(\Phi^T\mathbf{w}) - \Phi^T\mathbf{w})$
   - From preceeding result, we can get that $TD(\lambda)$ is the method for minimizing $L_{sq} = \|T^{(\lambda)}(\Phi\mathbf{w}) - V(\mathbf{w})\|_D^2$
3. Assumption1:
   - The Markov chain is irreducible and aperiodic. Furthermore, it has unique steady-state satisfying $\pi^T P = \pi^T$
   - $\mathbb{E}_0[r^2_{s_t, s_{t+1}}] = \mathbb{E}_{s\sim\pi}[r^2_{s_t, s_{t+1}} | s_0 = s] < \infty$.
     Let $R_2 = [r^2_{ij}]$, then $\mathbb{E}_0[r^2_{s_t, s_{t+1}}] = \pi^T P^t \langle P, R_2\rangle = \pi^T \langle P, R_2\rangle$
4. Assumption2:
   - Φ has full row column.
   - For every row ϕ_k in Φ, $\mathbf{E}_0[\phi_k^2(s_k)] < \infty$

3.3 Preliminaries

Lemma 3.1. $\|PV\|_D \leq \|V\|_D$

Proof.

$$\begin{aligned}\|PV\|_D^2 =& V^T P^T D P V = \sum_{i=1}^{n} \pi(i) (\sum_{j=1}^{n} p_{ij} V(j))^2 \\ \leq& \sum_{i=1}^{n} \pi(i) \sum_{j=1}^{n} p_{ij} V^2(j) = \sum_{j=1}^{n} \sum_{i=1}^{n} \pi(i) p_{ij} V^2(j) \\ =& \sum_{j=1}^{n} \pi(j) V^2(j) = \|V\|_D^2\end{aligned} \tag{3.3}$$

□

Lemma 3.2. $V^* \in L_2(S, D)$

Proof. Here
$\|\langle P, R\rangle\|_D = \sum_{i=1}^{n} \pi(i) (\sum_{j=1}^{n} p_{ij} r_{ij})^2 \leq \sum_{i=1}^{n} \pi(i) \sum_{j=1}^{n} p_{ij} r_{ij}^2 \leq \mathbb{E}_0[r_{ij}^2] < \infty$
$\|V^*\|_D \leq \sum_{t=0}^{\infty} \alpha^t \|P^t \langle P, R\rangle\|_D \leq \sum_{t=0}^{\infty} \alpha^t \|\langle P, R\rangle\|_D \leq \frac{1}{1-\alpha} \|\langle P, R\rangle\|_D < \infty$ □

Lemma 3.3. *If $V \in L_2(S, D)$, then $T^{(\lambda)} V \in L_2(S, D)$. Furthermore, $T^{(\lambda)}$ is a contraction mapping function.*

Proof.

$$\begin{aligned}T^{(\lambda)} V =& (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} (\alpha P)^t \langle P, R\rangle + (\alpha P)^{m+1} V \right) \\ \|T^{(\lambda)} V\|_D \leq& (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} \alpha^t \|\langle P, R\rangle\|_D + \alpha^{m+1} \|V\|_D \right) \leq \infty \\ \|T^{(\lambda)} V_1 - T^{(\lambda)} V_2\|_D \leq& (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \alpha^{m+1} \|V_1 - V_2\|_D \\ =& \frac{\alpha(1-\lambda)}{1-\alpha\lambda} \|V_1 - V_2\|_D\end{aligned} \tag{3.4}$$

□

Lemma 3.4. *$V^* = T^{(\lambda)} V^*$, and V^* is the uniquely solves of equations $V = T^{(\lambda)} V$.*

Proof.

$$\begin{aligned}T^{(\lambda)} V^* =& (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} (\alpha P)^t \langle P, R\rangle + (\alpha P)^{m+1} V^* \right) \\ =& (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} (\alpha P)^t \langle P, R\rangle + (\alpha P)^{m+1} \sum_{t=0}^{m} (\alpha P)^t \langle P, R\rangle \right) \\ =& (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{\infty} (\alpha P)^t \langle P, R\rangle \right) = V^*\end{aligned} \tag{3.5}$$

The uniqueness comes from preceding lemma. □

Lemma 3.5. *$\Pi T^{(\lambda)}$ is a contraction and has a unique fixed point which is of the form $\Phi \mathbf{w}^*$ for a unique choice of $\mathbf{w}^*$. Furthermore, $\mathbf{w}^*$ sarisfies the following bound:*

$$\|\Phi^T \mathbf{w}^* - V^*\|_D \leq \frac{1-\lambda\alpha}{1-\alpha}\|\Pi V^* - V^*\|_D$$

Proof. The proof of uniqueness is trivial, but the existance has no proof so far.

$$\begin{aligned}
\|\Phi^T \mathbf{w}^* - V^*\|_D \leq & \|\Phi^T \mathbf{w}^* - \Pi V^*\|_D + \|\Pi V^* - V^*\|_D \\
= & \|\Pi T^{(\lambda)} \Phi^T \mathbf{w}^* - \Pi V^*\|_D + \|\Pi V^* - V^*\|_D \\
\leq & \|T^{(\lambda)} \Phi^T \mathbf{w}^* - V^*\|_D + \|\Pi V^* - V^*\|_D \\
\leq & \frac{\alpha(1-\lambda)}{1-\alpha\lambda}\|\Phi^T \mathbf{w}^* - V^*\|_D + \|\Pi V^* - V^*\|_D \\
\|\Phi^T \mathbf{w}^* - V^*\|_D \leq & \frac{1-\lambda\alpha}{1-\alpha}\|\Pi V^* - V^*\|_D
\end{aligned} \tag{3.6}$$

□

Lemma 3.6. *Proof.*

$$\begin{aligned}
\mathbb{E}_0[\mathbf{z}_t \phi^T(s_t)] = & \pi^T \sum_{k=0}^{t} (\alpha\lambda)^{t-k} \mathbb{E}[\phi(s_k)\phi^T(s_t)|s_0 = s] \\
= & \sum_{k=0}^{t} (\alpha\lambda)^{t-k} \pi^T P^k \mathbb{E}[\phi(s_k)\phi^T(s_t)|s_k = s] \\
= & \sum_{k=0}^{t} (\alpha\lambda)^{t-k} \pi^T \mathbb{E}[\phi(s_k)\phi^T(s_t)|s_k = s] \\
= & \sum_{k=0}^{t} (\alpha\lambda)^{t-k} \pi^T \Phi D P^{t-k} \Phi^T \\
= & \sum_{m=0}^{t} (\alpha\lambda)^{m} \pi^T \Phi D P^{m} \Phi^T \quad (m = t-k)
\end{aligned} \tag{3.7}$$

(The proof of $\|\Phi D P^m \Phi^T\|_D < \infty$ is in the paper.)
Similarly, we can get

$$\begin{aligned}
\mathbb{E}_0[\Delta(\mathbf{w}, X_t)] = & \Phi D \sum_{m=0}^{t} (\alpha\lambda P)^m (\langle P, R\rangle + \alpha P \Phi^T \mathbf{w} - \Phi^T \mathbf{w}) \\
\sum_{m=0}^{\infty} (\alpha\lambda P)^m = & (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} (\alpha P)^t \\
\lim_{t\to\infty} \mathbb{E}_0[\Delta(\mathbf{w}, X_t)] = & \Phi D \sum_{m=0}^{\infty} (\alpha\lambda P)^m (\langle P, R\rangle + \alpha P \Phi^T \mathbf{w} - \Phi^T \mathbf{w}) \\
= & \Phi D \left((1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} (\alpha P)^t \langle P, R\rangle + ((1-\lambda) \sum_{m=0}^{\infty} \lambda^m (\alpha P)^{m+1} - I) \Phi^T \mathbf{w} \right) \\
= & \Phi D (T^{(\lambda)}(\Phi^T \mathbf{w}) - \Phi^T \mathbf{w})
\end{aligned} \tag{3.8}$$

□

Lemma 3.7. $t \to \infty, \forall \mathbf{w} \neq \mathbf{w}^*, (\mathbf{w} - \mathbf{w}^*)^T \mathbb{E}_0[\Delta(\mathbf{w}, X_t)] < 0$

Proof. When $t \to \infty$, because $\Phi D \Pi = \Phi D$, therefore

$$\begin{aligned}
(\mathbf{w} - \mathbf{w}^*)^T \mathbb{E}_0[\Delta(\mathbf{w}, X_t)] =& (\mathbf{w} - \mathbf{w}^*)^T \Phi D (T^{(\lambda)}(\Phi^T \mathbf{w}) - \Phi^T \mathbf{w}) \\
=& (\mathbf{w} - \mathbf{w}^*)^T \Phi D (\Pi T^{(\lambda)}(\Phi^T \mathbf{w}) - \Phi^T \mathbf{w}) \\
=& (\Phi^T \mathbf{w} - \Phi^T \mathbf{w}^*)^T D (\Pi T^{(\lambda)}(\Phi^T \mathbf{w}) - \Phi^T \mathbf{w}^* \\
& + \Phi^T \mathbf{w}^* - \Phi^T \mathbf{w}) \\
\leq& \|\Phi^T \mathbf{w} - \Phi^T \mathbf{w}^*\|_D \|\Pi T^{(\lambda)}(\Phi^T \mathbf{w}) - \Phi^T \mathbf{w}^*\|_D \\
& - \|\Phi^T \mathbf{w} - \Phi^T \mathbf{w}^*\|_D^2 \\
\leq& (\beta - 1) \|\Phi^T \mathbf{w} - \Phi^T \mathbf{w}^*\|_D^2 \\
(\exists \beta < 1, & \|\Pi T^{(\lambda)} V_1 - \Pi T^{(\lambda)} V_2\|_D \leq \beta \|V_1 - V_2\|_D)
\end{aligned} \tag{3.9}$$

□

3.4 Proof of theorem1

Theorem 3.1. *Here,*

1. *The cost-to-go function V^* is in $L_2(S, D)$.*
2. *For any $\lambda \in [0, 1]$, the $TD(\lambda)$ algorithm with linear function approximators, converges with probability one. (This needs assumption3, assumption4 and theorem2 in the paper, which is related to stochastic approximation algorithm. I'm not familiar with stochastic approximation, so this part of proof isn't in the note so far.)*
3. *The limit of convergence r^* is the unique solution of the equation*

$$\Pi T^{(\lambda)}(\Phi^T \mathbf{w}^*) = \Phi^T \mathbf{w}^*.$$

4. *Furthermore, r^* satisfies*

$$\|\Phi \mathbf{w}^* - V^*\|_D \leq \frac{1 - \lambda\alpha}{1 - \alpha} \|\Pi V^* - V^*\|_D$$

In this section, the paper proofs that the $TD(\lambda)$ algorithm makes $\mathbf{w}_t$ converging to $\mathbf{w}^*$ which solves $\lim_{t \to \infty} E_0[\Delta(\mathbf{w}^*, X_t)] = 0$:

$$\Phi D (T^{(\lambda)}(\Phi^T \mathbf{w}^*) - \Phi^T \mathbf{w}^*) = 0$$

Because ΦD has full row rank, $\Phi D = \Phi D \Pi$ and $\Pi \Phi^T = \Phi^T$, therefore

$$\Phi D (\Pi T^{(\lambda)}(\Phi^T \mathbf{w}^*) - \Phi^T \mathbf{w}^*) = 0$$

which means $\Phi^T \mathbf{w}^*$ is fixed point of $\Pi T^{(\lambda)}$.

The proof corresponds to preceeding theorem's item2.