# Neuro Dynamic Programming

Peng Lingwei

September 6, 2019

## Contents

# 1 Introduction

# 2 Dynamic Programming

**Definition 1.** ***(Proper stationary policy).*** *(Reach termination state 0 w.p.1)*

$$\rho^\pi = \max_{i=1,\ldots,n} P^\pi \{s_n \neq 0 | i_0 = i\} < 1$$

In stochastic shortest path problems, we have two assumptions:

- There exists at least one proper policy;
- For every improper policy $\pi$, the corresponding cost-to-oo $J^\pi(i)$ is infinite for at least one state i.

**Policy Iteration as an Actor-Critic System**

- Critic: policy evaluation;
- actor: policy improvement.

# 3 Neural Network Architectures and Training

Trivial. Using some function to approximate $V^\pi, V^*, Q^\pi, Q^*$. This book uses neural network.

# 4 Stochastic Iterative Algorithms

Suppose that we are interested in solving a system of equations of the form

$$Hr = r,$$

where $H$ is a function from $\mathbb{R}^n$ into itself. If $Hr = r - \nabla f(r)$, the solution of the system $Hr = r$ is of the form

$$\nabla f(r) = 0,$$

Then it's sometime minimize the cost function $f$.

One possible algorithm for solving the system $Hr = r$ is provided by the iteration

$$r_{t+1} = Hr_t, \quad or \quad r_{t+1} = (1-\gamma)r_t + \gamma Hr_t.$$

the second method reduces to the gradient method if $Hr = r - \nabla f(r)$.

Sometimes an exact evaluation of $Hr$ is difficult but that we have access to a random variable s of the form $s = Hr + w$, where w is a random noise term. Then we obtain stochastic iterative or stochastic approximation algorithm

$$r_{t+1} = (1-\gamma)r + \gamma(Hr + w).$$

A more concrete setting is obtained as follows. Let $v$ be a random variable with a known probability distribution $p(v|r)$ that depends on r. Suppose that we are interested in solving:

$$\mathbb{E}_{v\sim p(v|r)}\left[g(r,v)\right] = r,$$

where g is a known function. We can use preceeding algorithm:

$$r_{t+1} = (1-\gamma)r_t + \gamma \mathbb{E}_{v\sim p(v|r)}\left[g(r,v)\right].$$

We can estimate $\mathbb{E}_{v\sim p(v|r)}\left[g(r,v)\right] \approx \frac{1}{k}\sum_{i=1}^{k} g(r,\tilde{v}_i)$. We get Robbins-Monro stochastic approximation algorithm ($k=1$),

$$r_{t+1} = (1-\gamma)r_t + \gamma g(r,\tilde{v}),$$

which is a special case of the algorithm $r_{t+1} = (1-\gamma)r_t + \gamma\left(Hr_t + w\right)$, where $Hr = \mathbb{E}_{v\sim p(v|r)}\left[g(r,v)\right]$, and $w = g(r,\tilde{v}) - \mathbb{E}\left[g(r,v)\right]$.

## 4.1 THE BASIC MODEL

Let $T^i$ be the set of times at which $r(i)$ updates:

$$r_{t+1}(i) = \begin{cases} r_t(i), & t \notin T^i \\ (1-\gamma_t(i))r_t(i) + \gamma_t(i)\left((Hr_t)(i) + w_t(i)\right), & t \in T^i \end{cases}$$

Assumption: $\sum_{t=0}^{\infty}\gamma_t(i) = \infty$ and $\sum_{t=0}^{\infty}\gamma_t^2(i) < \infty$.

## 4.2 CONVERGENCE BASED ON A SMOOTH POTHENTIAL FUNCTION

$$r_{t+1} = r_t + \gamma_t\delta_t, \quad \delta_t = Hr_t - r_t + w_t.$$

Let $\mathcal{H}_t$ denote the history of the algorithm

$$\mathcal{H}_t = \left\{r_0,\ldots,r_t,\delta_0,\ldots,\delta_{t-1},\gamma_0,\ldots,\gamma_t\right\}.$$

**Assumption 1.** *Exist function $f:\mathbb{R}^n \mapsto \mathbb{R}$, with the following properties:*

1. $\forall \mathbb{R}^n, f(r) \geq 0$;
2. $\|\nabla f(r_1) - \nabla f(r_2)\| \leq L\|r_1 - r_2\|_2$;
3. *(Pseudogradient property)* $c\|\nabla f(r_t)\|_2^2 + \langle \nabla f(r_t), \mathbb{E}\left[\delta_t|\mathcal{H}_t\right]\rangle \leq 0$
4. $\mathbb{E}\left[\|\delta_t\|_2^2|\mathcal{H}_t\right] \leq K_1 + K_2\|\nabla f(r_t)\|_2^2$

**Proposition 1.** *Consider the algorithm $r_{t+1} = r_t + \gamma_t s_t$, if $\sum_{t=0}^{\infty}\gamma_t = \infty$ and $\sum_{t=0}^{\infty}\gamma_t^2 < \infty$. Under preceeding assumption, the following hold with probability 1:*

- *The sequence $f(r_t)$ converges;*
- $\lim_{t\to\infty}\nabla f(r_t) = 0$;
- *Every limit point of $r_t$ is a stationary point of f.*

**Example 1.** ***(Stochastic Gradient Algorithm).***

$$r_{t+1} = r_t + \gamma_t\delta_t, \quad \delta_t = -\left(\nabla f(r_t) + w_t\right)$$

*Assumption:*

1. $\sum_{t=0}^{\infty} \gamma_t = \infty, \quad \sum_{t=0}^{\infty} \gamma_t^2 < \infty$;

2. $f$ *is nonnegative and has a Lipschitz continuous gradient;*

3. $\mathbb{E}[w_t|\mathcal{H}_t] = 0, \quad \mathbb{E}\left[\|w_t\|^2|\mathcal{H}_t\right] \le A + B\|\nabla f(r_t)\|_2^2$;

*We proof Assumption 1 is satisfied.*

$$\langle \nabla f(r_t), \mathbb{E}[\delta_t|\mathcal{H}_t] \rangle = \langle \nabla f(r_t), -\nabla f(x_t) - \mathbb{E}[w_t|\mathcal{H}_t] \rangle = -\|\nabla f(r_t)\|_2^2$$

$$\begin{aligned}\mathbb{E}\left[\|\delta_t\|_2^2|\mathcal{H}_t\right] =& \|\nabla f(r_t)\|_2^2 + \mathbb{E}\left[\|w_t\|_2^2|\mathcal{H}_t\right] + \langle 2\nabla f(r_t), \mathbb{E}[w_t|\mathcal{H}_t] \rangle \\ =& \|\nabla f(r_t)\|_2^2 + A + B\|\nabla f(r_t)\|_2^2 \\ =& A + (B+1)\|\nabla f(r_t)\|_2^2\end{aligned}$$

**Example 2.** ***(Estimate of an Unknown Mean).*** *For random variables $v$ with unknow mean $\mu$ and unit variance.*

$$r_{t+1} = (1-\gamma_t)r_t + \gamma_t v_t.$$

*with assumption*

1. $\sum_{t=0}^{\infty} \gamma_t = \infty$ *and* $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$;

*Proof.*

$$r_{t+1} = r_t - \gamma_t(r_t - \mu) + \gamma_t(v_t - \mu)$$

where $f(r) = (r-\mu)^2/2$, $\nabla f(r_t) = (r_t - \mu)$. (The other assumptions in stochastic gradient algorithm are sastified naturally.) □

**Example 3.** ***(Euclidean Norm Pseudo-Contractions).***

$$r_{t+1} = (1-\gamma_t)r_t + \gamma_t(Hr_t + w_t),$$

*Assuming:*

1. $\|Hr - r^*\|_2 \le \beta\|r - r^*\|_2, \forall r \in \mathbb{R}^n$, $0 \le \beta < 1$;

2. $\mathbb{E}[w_t|\mathcal{H}_t] = 0$;

3. $\mathbb{E}\left[\|w_t\|_2^2|\mathcal{H}_t\right] \le A + B\|r_t - r^*\|_2^2$

*The potential function is* $f(r) = \frac{1}{2}\|r - r^*\|_2^2.$, $\delta_t = -r_t + Hr_t + w_t$, *then* $\mathbb{E}[\delta_t|\mathcal{H}_t] = Hr_t - r_t$.

$$\begin{aligned}\langle Hr - r^*, r - r^* \rangle \le& \|Hr - r^*\|_2\|r - r^*\|_2 \le \beta\|r - r^*\|_2^2 \\ \langle Hr - r, r - r^* \rangle \le& -(1-\beta)\|r - r^*\|_2^2 \\ \langle \mathbb{E}[\delta_t|\mathcal{H}_t], \nabla f(r_t) \rangle \le& -(1-\beta)\|\nabla f(r_t)\|_2^2\end{aligned}$$

$$\mathbb{E}\left[\delta_t^2|\mathcal{H}_t\right] = \mathbb{E}\left[(-r_t + Hr_t)^2|\mathcal{H}_t\right] + \mathbb{E}\left[\|w_t\|^2|\mathcal{H}_t\right] \le (Hr_t - r_t)^2 + A + B\|r_t - r^*\|_2^2$$

### 4.2.1 Convergence Proofs

In this section, we discarded a suitable set of measure zero, and don't keep repeating the qualification " with probability 1".

**Theorem 1.** ***(Supermartingale Convergence Theorem).*** *Here is three sequences of random variables $\{X_t\}, \{Y_t\}$ and $\{Z_t\}$. And let $\mathcal{F}_t$ be set of random variables and $\mathcal{F}_t \subset \mathcal{F}_{t+1}$. Suppose that*

1. *$X_t, Y_t, Z_t$ are nonegative, and are functions of the random variables in $\mathcal{F}_t$;*
2. *$\forall t$, we have $\mathbb{E}\left[Y_{t+1}|\mathcal{F}_t\right] \leq Y_t - X_t + Z_t$;*
3. *$\sum_{t=0}^{\infty} Z_t < \infty$.*

*Then we have $\sum_{t=0}^{\infty} X_t < \infty$, and the sequence $Y_t$ converges to a nonegative random variable $Y$, w.p.1.*

**Theorem 2.** ***(Martigale Convergence Theorem)*** *Let $\{X_t\}$ be a sequence of random variables and let $\mathcal{F}_t$ be set of random variables such that $\mathcal{F}_t \subset \mathcal{F}_{t+1}$. Suppose that:*

1. *The random variable $X_t$ is a function of the random variable in $\mathcal{F}_t$;*
2. *$\mathbb{E}\left[X_{t+1}|\mathcal{F}_t\right] = X_t$,*
3. *$\exists M < \infty$ such that $\mathbb{E}\left[|X_t|\right] \leq M$.*

*Then, the sequence $X_t$ converges to a random variable $X$, w.p.1.*

*Now we begin proof the preceeding section.*

*Proof.* By assumption, we have $\|\nabla f(r_1) - \nabla f(r_2)\|_2 \leq L\|r_1 - r_2\|$, we have

$$f(r_{t+1}) \leq f(r_t) + \gamma_t \langle \nabla f(r), \delta_t \rangle + \frac{L}{2}\gamma_t^2 \|\delta_t\|_2^2$$

$$\begin{aligned}
\mathbb{E}\left[f(r_{t+1}|\mathcal{F}_t)\right] \leq& f(r_t) + \gamma_t \langle \nabla f(r_t), \mathbb{E}\left[\delta_t|\mathcal{F}_t\right]\rangle + \frac{L}{2}\gamma_t^2\left(K_1 + K_2\|\nabla f(r_t)\|_2^2\right) \\
\leq& f(r_t) - \gamma_t\left(c - \frac{LK_2\gamma_t}{2}\right)\|\nabla f(r_t)\|_2^2 + \frac{LK_1\gamma_t^2}{2} \\
=& f(r_t) - X_t + Z_t,
\end{aligned}$$

where

$$X_t = \begin{cases} \gamma_t\left(c - \frac{LK_2\gamma_t}{2}\right)\|\nabla f(r_t)\|_2^2, & if\ LK_2\gamma_t \leq 2c, \\ 0, & otherwise. \end{cases}$$

and

$$Z_t = \begin{cases} \frac{LK_1\gamma_t^2}{2}, & if\ LK_2\gamma_t \leq 2c, \\ \frac{LK_1\gamma_t^2}{2} - \gamma_t\left(c - \frac{LK_2\gamma_t}{2}\right)\|\nabla f(r_t)\|_2^2, & otherwise \end{cases}$$

Because $\sum_{t=0}^{\infty}\gamma_t^2 < \infty$, so after some finite time $LK_2\gamma_t \leq 2c$, and $Z_t = LK_1\gamma_t^2/2$, and therefore $\sum_{t=0}^{\infty} Z_t < \infty$. Thus, the supermartingale convergence theorem applies and shows that $f(r_t)$ converges and $\sum_{t=0}^{\infty} X_t < \infty$.

Because $X_t = \gamma_t \left(c - \frac{LK_2\gamma_t}{2}\right) \|\nabla f(r_t)\|_2^2 \geq \frac{c}{2}\gamma_t \|\nabla f(r_t)\|_2^2$ after some finite time. Hence

$$\sum_{t=0}^{\infty} \gamma_t \|\nabla f(r_t)\|_2^2 < \infty$$

Because $\sum_{t=0}^{\infty} \gamma_t = \infty$, $\liminf_{t\to\infty} \|\nabla f(r_t)\|_2 = 0$

Let us denote $\bar{s}_t = \mathbb{E}\left[s_t|\mathcal{F}_t\right]$ and $w_t = s_t - \bar{s}_t$, then

$$\|\bar{s}_t\|_2^2 + \mathbb{E}\left[\|w_t\|_2^2|\mathcal{F}_t\right] = \mathbb{E}\left[\|s_t\|_2^2|\mathcal{F}_t\right] \leq K_1 + K_2\|\nabla f(r_t)\|_2^2$$

We need take a break and proof another lemma

**Lemma 1.** *$u_t = \sum_{\tau=0}^{t-1} \chi_\tau \gamma_\tau w_\tau$, converges w.p.1. where $\chi_t = 1_{\left[\|\nabla f(r_t)\|_2 \leq \epsilon\right]}$.*

*Proof.* We start the assumption $\sum_{t=0}^{\infty} \gamma_t^2 \leq A < \infty$.

$$\mathbb{E}\left[\chi_t \gamma_t w_t|\mathcal{F}_t\right] = \chi_t \gamma_t \mathbb{E}\left[w_t|\mathcal{F}_t\right] = 0 \Rightarrow \mathbb{E}\left[u_{t+1}|\mathcal{F}_t\right] = u_t$$

If $\chi_t = 0$, then $\mathbb{E}\left[\|u_{t+1}\|_2^2|\mathcal{F}_t\right] = \|u_t\|^2$. If $\chi_t = 1$, we have

$$\mathbb{E}\left[\|u_{t+1}\|_2^2|\mathcal{F}_t\right] = \|u_t\|_2^2 + \gamma_t^2 \mathbb{E}\left[\|w_t\|_2^2|\mathcal{F}_t\right] \leq \|u_t\|_2^2 + \gamma_t^2(K_1 + K_2\epsilon^2)$$

$$\mathbb{E}\left[\|u_t\|_2^2\right] \leq (K_1 + K_2\epsilon^2)\mathbb{E}\left[\sum_{\tau=0}^{t-1} \gamma_\tau^2\right] \leq (K_1 + K_2\epsilon^2)A$$

$$\sup_t \mathbb{E}\left[\|u_t\|^2\right] \leq \sup_t \mathbb{E}\left[1 + \|u_t\|_2^2\right] < \infty$$

Then we can use Martigale convergence theorem to $u_t$ and get that $u_t$ converges, w.p.1.

We can assume that $\sum_{\tau=0}^{t-1} \gamma_\tau^2 \leq A < \infty$ and get the same result. □

**I give up today.** □

# 5 Sinulation Methods for a Lookup Table Representation

## 5.1 SOME ASPECTS OF MONTE CARLO SIMULATION

Suppose that v is a random variable with an unknown mean $m$ that we wish to estimate. Monte-Carlo simulation is to generate a number of samples $\{v_1, \ldots, v_N\}$, and estimate the mean of v by forming the sample mean

$$M_N = \frac{1}{N}\sum_{k=1}^{N} v_k = M_{N-1} + \frac{1}{N}(v_N - M_N).$$

**(The Case of i.i.d. Samples)**:

- $\mathbb{E}\left[M_N\right] = \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}\left[v_k\right] = m$;

- $Var(M_N) = \frac{1}{N^2}\sum_{k=1}^{N} Var(v_k) = \frac{\sigma^2}{N}$.

**(The Case of Dependent Samples)**:

- The estimator of mean remains unbiased;
- We can use a weighted average to lower the variance.

**(The Case of a Random Sample Size)**:
In this case, the number of samples N is itself a random variable.

If N is correlated with $\{v_1, \dots, v_N\}$. Then the sample mean is unbiased, and the variance is $\sigma^2\mathbb{E}[1/N]$.

If N is correlated with $\{v_1, \dots, v_N\}$, the sample mean maybe biased.

**Theorem 3.** *If $\{v_1, \dots, v_N\}$ is i.i.d. with finite mean, N depends upon given sequence, and $\mathbb{E}[N] < \infty$.*

$$\mathbb{E}\left[\sum_{k=1}^{N} v_k\right] = \sum_{k=1}^{\infty} P(N \geq k)\mathbb{E}[v_k|N \geq k] = \mathbb{E}[v_1]\sum_{k=1}^{\infty} P(N \geq k)$$
$$=\mathbb{E}[v_1]\sum_{k=1}^{\infty}\sum_{n=k}^{\infty} P(N = n) = \mathbb{E}[v_1]\sum_{n=1}^{\infty}\sum_{k=1}^{n} P(N = n) = \mathbb{E}[v_1]\mathbb{E}[N].$$

## 5.2 POLICY EVALUATION BY MONTE CARLO SIMULATION

- State space: $\{0, 1, \dots, n\}$, where 0 is a cost-free absorbing state;
- The policy $\pi$ is proper;
- For $m$th trajectory is $(s_0^m, s_1^m, \dots, s_N^m)$;
- Cost of trajectory is $c(s_0^m, m) = \sum_{t=1}^{N-1} g(s_t^m, s_{t+1}^m)$;
- Policy value: $J^\pi(s) = \mathbb{E}^\pi[c(s, m)]$.

**Algorithm1**:

$$\tilde{J}(i) = \frac{1}{K}\sum_{m=1}^{K} c(i, m)$$

$$\tilde{J}^m(i) = \tilde{J}^{m-1}(i) + \frac{1}{m}(c(i, m) - \tilde{J}^{m-1}(i)), \quad s.t. J^0(i) = 0.$$

**Algorithm2**: Use the full trajectory.

$$J(i_k) = J(i_k) + \gamma(i_k)(g(i_k, i_k + 1) + \cdots + g(i_{N-1}, i_N) - J(i_k))$$

**Every-visit method** provides a biased estimator.

**Consistency of the Every-Visit Method**

Let $c(s, k, m)$ mean, in $m$th trajectory, the cost after visiting state $s$ $k$th times. And $n_s^m$ means the total number of state s in $m$th trajectory. Then every-visit method estimator is

$$\tilde{J}(s) = \frac{\sum_{m=1}^{K}\sum_{k=1}^{n_s^m} c(s, k, m)}{\sum_{m=1}^{K} n_s^m}$$

When K is fixed, the extimator is biased. But if $K \to \infty$, the estimator is unbiased:

$$\mathbb{E}\left[\tilde{J}(s)\right] = \frac{\mathbb{E}\left[\sum_{k=1}^{n_k^m} c(s,k,m) | n_k \geq 1\right]}{\mathbb{E}\left[n_k | n_k \geq 1\right]} = \mathbb{E}\left[c(s,1,m)|n_k \geq 1\right] = J^\pi(s)$$

**The First-Visit Method**

$$\tilde{J}(s) = \frac{\sum_{m:n_s^m \geq 1} c(s,1,m)}{\sum_{m=1}^{K} 1_{[n_s^m \geq 1]}}$$

which is unbiased.

The significance of the comparison of the every-visit and first-visit methods should not be overemphasized. For problems with large state space, the likelihood of a trajectory visiting the same state twice is usually quite small.

### 5.2.1 Q-Factors and Policy Iteration

$$Q^\pi(s,a) = \sum_{s' \in S} p_{s,s'}(a)(g(s,a,s') + J^\pi(s'))$$

## 5.3 TEMPORAL DIFFERENCE METHODS

TD: policy evaluation.

$J_{k+1}^m(s_k^m) = J_k^m(s_k^m) + \gamma(g(s_k^m, s_{k+1}^m) + \cdots + g(s_{N-1}^m, s_N^m) - J_k^m(i_k))$

$$\begin{aligned} J_{k+1}^m(s_k^m) = J_k^m(s_k^m) + \gamma[&(g(s_k^m, s_{k+1}^m) + J_{k+1}^m(s_{k+1}^m) - J_k^m(s_k^m)) \\ &+ (g(s_{k+1}^m, s_{k+2}^m) + J_{k+2}^m(s_{k+2}^m) - J_{k+1}^m(s_{k+1}^m)) \\ &+ \cdots \\ &+ (g(s_{N-1}^m, s_N^m) + J_N^m(s_N^m) - J_{N-1}^m(s_{N-1}^m))] \end{aligned}$$

where we have made use of $J(s_N) = 0$.

$$J_{k+1}^m(s_k^m) = J_k^m(s_k^m) + \gamma(d_k^m + d_{k+1}^m + \cdots + d_{N-1}^m),$$

$$d_k^m = g(s_k^m, s_{k+1}^m) + J_{k+1}^m(s_{k+1}^m) - J_k^m(s_k^m)$$

Make explanation complex, maybe approximation TD is more easy to understanding.

# 6 Approximate DP with Cost-to-Go Function Approximation

## 6.1 GENERIC ISSUES-FROM PARAMETERS TO POLICIES

Most of the methods discussed in this chapter lead to an approximate cost-to-go function $\tilde{J}_w(s)$, which is meant to be a good approximation of the optimal cost-to-go function $J^*(s)$. Such $\tilde{J}_w(s)$ leads to a corresponding greedy policy $\tilde{\pi}$ defined by

$$\tilde{\pi}(s) = \arg\min_{a \in A_s} \sum_{s'} p_{ss'}(a)(g(s,a,s') + \tilde{J}_w(s'))$$

(Here is an undiscounted problem. Critic period.)

**(Approximation of Q-Factors)** We approximate $\hat{Q}_{w_2}$:

$$\tilde{Q}_{w_1}(s,a) = \sum_{s'} p_{ss'}(a)(g(s,a,s') + \tilde{J}_{w_1}(s'))$$

$$\min_{w_2} \sum_{(s,a)} \left(\hat{Q}_{w_2}(s,a) - \tilde{Q}_{w_1}(s,a)\right)^2$$

**(Policy Approximation)**

$$\min_{w_3} \sum_{s \in \hat{S}} \|\hat{\pi}_{w_3}(s) - \tilde{\pi}(s)\|^2$$

(Actor period.)

### 6.1.1 Generic Error Bounds

Approximate DP is based on the hypothesis that: If $\tilde{J}_w$ is a good approximation of $J^*$, then the greedy policy based on $\tilde{J}_w$ is close to optimal.

**Theorem 4.** *Consider a discounted problem, with discount factor $0 \le \gamma < 1$,*

$$\|\tilde{J}_w - J^*\|_\infty \le \epsilon \Rightarrow \|J^{\tilde{\pi}} - J^*\|_\infty \le \frac{2\gamma\epsilon}{1-\gamma}$$

*which means that $\tilde{\pi}$ can be arbitrary $\frac{2\gamma\epsilon}{1-\gamma}$-optimal policy.*

*Proof.*

$$\begin{aligned}
\|J^{\tilde{\pi}} - J^*\|_\infty =& \|T^{\tilde{\pi}} J^{\tilde{\pi}} - J^*\|_\infty \\
\le& \|T^{\tilde{\pi}} J^{\tilde{\pi}} - T^{\tilde{\pi}} \tilde{J}_w\|_\infty + \|T\tilde{J}_w - TJ^*\|_\infty \\
\le& \gamma\|J^{\tilde{\pi}} - \tilde{J}_w\|_\infty + \gamma\|\tilde{J}_w - J^*\|_\infty \\
\le& \gamma\|J^{\tilde{\pi}} - J^*\|_\infty + 2\gamma\|\tilde{J}_w - J^*\|_\infty
\end{aligned}$$

□

### 6.1.2 Multistage Lookahead Variations

### 6.1.3 Rollout Policies

### 6.1.4 Trading off Control Space Complexity with State Space Complexity

## 6.2 APPROXIMATE POLICY ITERATION

- Approximate policy evaluation;
- Policy update.

### 6.2.1 Approximate Policy Iteration Based on Monte Carlo Simulation

A variant of approximate policy iteration that combines Monte Carlo simulation and approximation for the purpose of policy evaluation.

First, we sample M trajectories, and minimize $w$ as supervised learning.

$$\min_w \sum_{s\in S} \sum_{m=1}^{M(s)} \left(\tilde{J}_w(s) - c(s,m)\right)^2$$

Then we get statistical approximation $\tilde{Q}_w(s,a) = \sum_{s'} p_{ss'}(a) \left(g(s,a,s') + \tilde{J}_w(s')\right)$, and get greedy policy $\tilde{\pi}$.

### 6.2.2 Error Bounds for Approximate Policy Iteration

We assume that all policy evaluations and all policy updates are performed with a certain error tolerance of $\epsilon$ and $\delta$.

$$\|\tilde{J}_{w_k} - J^{\pi_k}\|_\infty \leq \epsilon$$
$$\|T^{\pi_{k+1}}\tilde{J}_{w_k} - T\tilde{J}_{w_k}\|_\infty \leq \delta$$

**Discounted Problems**

**Lemma 2.** *If $\|J - J^\pi\|_\infty \leq \epsilon$, and $T^{\bar{\pi}}J - TJ \preceq \delta \cdot \vec{1}$, then*

$$J^{\bar{\pi}} - J^\pi \preceq \frac{\delta + 2\gamma\epsilon}{1-\gamma} \cdot \vec{1}$$

*Proof.* Let $\xi = \max_{s\in S} J^{\bar{\pi}}(s) - J^\pi(s)$.

$$\begin{aligned}
J^{\bar{\pi}} - J^\pi =& J^{\bar{\pi}} - T^{\bar{\pi}}J^\pi + T^{\bar{\pi}}J^\pi - T^{\bar{\pi}}J + T^{\bar{\pi}}J - T^\pi J + T^\pi J - J^\pi \\
\preceq& \gamma\xi \cdot \vec{1} + \gamma\epsilon \cdot \vec{1} + T^{\bar{\pi}}J - T^\pi J + \gamma\epsilon \cdot \vec{1} \\
(1-\gamma)\xi \preceq& 2\gamma\epsilon \cdot \vec{1} + T^{\bar{\pi}}J - TJ \preceq (2\gamma\epsilon + \delta) \cdot \vec{1}
\end{aligned}$$

□

Then we have $J^{\pi_{k+1}} - J^{\pi_k} \preceq \frac{\delta+2\gamma\epsilon}{1-\gamma} \cdot \vec{1}$.

**Lemma 3.** *Let $\sigma_k = \max_{s\in S} (J^{\pi_k}(s) - J^*(s))$*

$$\sigma_{k+1} \leq \gamma\sigma_k + \gamma\xi_k + \delta + 2\gamma\epsilon$$

*Proof.*

$$\begin{aligned}
J^{\pi_{k+1}} - J^* =& (J^{\pi_{k+1}} - T^{\pi_{k+1}}J^{\pi_k}) + (T^{\pi_{k+1}}J^{\pi_k} - T^{\pi_{k+1}}\tilde{J}_{w_k}) \\
&+ (T^{\pi_{k+1}}\tilde{J}_{w_k} - T\tilde{J}_{w_k}) + (T\tilde{J}_{w_k} - TJ^{\pi_k}) + (TJ^{\pi_k} - J^*) \\
\preceq& (\gamma\xi_k + \gamma\epsilon + \delta + \gamma\epsilon + \gamma\sigma_k) \cdot \vec{1} = (\gamma\sigma_k + \gamma\xi_k + \delta + 2\gamma\epsilon) \cdot \vec{1}
\end{aligned}$$

□

**Theorem 5.**
$$\limsup_{k\to\infty} \|J^{\pi_k} - J^*\|_\infty \le \frac{\delta + 2\gamma\epsilon}{(1-\gamma)^2}$$

*Proof.*

$$(1-\gamma)\limsup_{k\to\infty} \sigma_k \le \gamma\frac{\delta + 2\gamma\epsilon}{1-\gamma} + \delta + 2\gamma\epsilon = \frac{\delta + 2\gamma\epsilon}{1-\gamma}$$

□

**Stochastic Shortest Path Problems**

In this problem (undiscounted), the policy can be proper of improper. If $\pi_k$ is improper, $J^{\pi_k}$ has infinity part and the preceeding algorithm breaks down.

$$S = \{0, 1, 2, \ldots\}$$

**Theorem 6.** *Let $\rho = \max_{i=1,2,\ldots,n;\pi:proper} P^\pi(s_n \neq 0 | s_0 = i)$. Assume that the preceeding algorithm generates proper policies. Then*

$$\limsup_{k\to\infty} \|J^{\pi_k} - J^*\|_\infty \le \frac{n(1-\rho+n)(\delta + 2\epsilon)}{(1-\rho)^2}$$

*Proof.*

$$\begin{aligned} J^{\pi_{k+1}} - J^{\pi_k} =& (T^{\pi_{k+1}} J^{\pi_{k+1}} - T^{\pi_{k+1}} J^{\pi_k}) + (T^{\pi_{k+1}} J^{\pi_k} - T^{\pi_{k+1}} \tilde{J}^{w_k}) \\ &+ (T^{\pi_{k+1}} \tilde{J}^{w_k} - T\tilde{J}_{w_k}) + (T\tilde{J}_{w_k} - TJ^{\pi_k}) + (TJ^{\pi_k} - T^{\pi_k} J^{\pi_k}) \\ \preceq& P^{\pi_{k+1}}(J^{\pi_{k+1}} - J^{\pi_k}) + (\epsilon + \delta + \epsilon) \cdot \vec{1} \end{aligned}$$

□

**Lemma 4.** *1.* $x \preceq Px + c\vec{1} \Rightarrow x \preceq \frac{nc}{1-\rho} \cdot \vec{1}$

*2.* $x_{k+1} \preceq Px_k + c\vec{1} \Rightarrow \limsup_{k\to\infty} x_k \preceq \frac{nc}{1-\rho} \cdot \vec{1}$

*Proof.* Let $y(i) = \max\{0, x(i)\}, i = 1, \ldots, n.$

$$x \preceq Px + c\vec{1} \preceq Py + c\vec{1} \Rightarrow \max\{0, x\} \preceq \max\left\{0, Py + c\vec{1}\right\} = Py + c\vec{1}$$

$$y \preceq P^n y + nc\vec{1} \preceq \rho(\max y) \cdot \vec{1} + nc\vec{1}$$
$$x \preceq \max y \cdot \vec{1} \preceq \frac{nc}{1-\rho}$$

Similarly, we obtain $\max y_{k+n} \le \rho \max y_k + nc$. Hence,

$$\limsup_{k\to\infty} (\max y_{k+n}) \le \rho \limsup_{k\to\infty} (\max y_k) + nc$$

□

*From preceeding lemma, we can get* $J^{\pi_{k+1}} - J^{\pi_k} \preceq \frac{n(\delta+2\epsilon)}{1-\rho}$.

$$
\begin{aligned}
J^{\pi_{k+1}} - J^* =& (J^{\pi_{k+1}} - T^{\pi_{k+1}} J^{\pi_k}) + (T^{\pi_{k+1}} J^{\pi_k} - T^{\pi^*} J^{\pi_k}) + (T^{\pi^*} J^{\pi_k} - T^{\pi^*} J^*) \\
\preceq& P^{\pi_{k+1}} (J^{\pi_{k+1}} - J^{\pi_k}) + (T^{\pi_{k+1}} J^{\pi_k} - T^{\pi_k} J^{\pi_k}) + P^{\pi^*} (J^{\pi_k} - J^*) \\
\preceq& P^{\pi^*} (J^{\pi_k} - J^*) + \frac{n(\delta+2\epsilon)}{1-\rho} \cdot \vec{1} + (\delta+2\epsilon) \cdot \vec{1} \\
=& P^{\pi^*} (J^{\pi_k} - J^*) + \frac{(1-\rho+n)(\delta+2\epsilon)}{1-\rho} \cdot \vec{1}
\end{aligned}
$$

$$
\limsup_{k\to\infty} \|J^{\pi_{k+1}} - J^*\|_\infty \leq \frac{n(1-\rho+n)(\delta+2\epsilon)}{(1-\rho)^2}
$$

The preceeding theorem uses the stochastic shortest path problems's property, which shows that the probability of termination at state 0 after n transformation is positive. We can get better estimate.

$$
\rho_m = \max_{i=1,\dots,n;\pi:proper} P^\pi (s_m \neq 0 | s_0 = i)
$$

Then

$$
\limsup_{k\to\infty} \|J^{\pi_{k+1}} - J^*\|_\infty \leq \frac{m(1-\rho_m+m)(\delta+2\epsilon)}{(1-\rho_m)^2}
$$

If we can guarantee termination occurs within at most N stages for all proper policies, then $\rho_N = 0$, and we obtain

$$
\limsup_{k\to\infty} \|J^{\pi_{k+1}} - J^*\|_\infty \leq N(1+N)(\delta+2\epsilon)
$$

If policies converge, we can obtain

$$
k \to \infty, \quad J^{\pi_{k+1}} - J^* \preceq P^{\pi^*} (J^{\pi_k} - J^*) + (\delta+2\epsilon) \cdot \vec{1}
$$

$$
\limsup_{k\to\infty} \|J^{\pi_{k+1}} - J^*\|_\infty \leq \frac{n(\delta+2\epsilon)}{1-\rho}
$$

### 6.2.3 Tightness of the Error Bounds and Empirical Behavior

Here is an example showing that the bound is tight.

**Example 4.** *Environment is discounted MDP:*

1. $S = \{1, 2, \dots, n\}$;
2. $A_1 = \{stay\}$, *and* $A_i = \{stay, goto\ s_{i-1}\}$;
3. $r(s=1, a=stay) = 0$, $r(s=i, a=stay) = r(s=i-1, a=stay) + \delta + 2\gamma\epsilon$, *otherwise* $r = 0$.

*Proof.* Here are two cases.

1. Case 1: $\pi_k = \{a_1 = stay, a_i = \text{goto } s_{i-1} | i \geq 2\}$. Then, $J^{\pi_k} = \{0, 0, \dots, 0\}$. And we let $\tilde{J}_{w_k} = \{\epsilon, -\epsilon, 0, \dots, 0\}$.
If $\pi_{k+1} = \{a_1 = stay, a_2 = stay, a_i = \text{goto } s_{i-1} | i > 2\}$,
$$T^{\pi_{k+1}} \tilde{J}_{w_k} = \left\{\epsilon\lambda, \delta + \epsilon\lambda, -\epsilon\lambda, \dots, -\epsilon\lambda^{n-2}\right\}$$
$$T\tilde{J}_{w_k} = \left\{\lambda\epsilon, \lambda\epsilon, -\epsilon\lambda, \dots, \epsilon\lambda^{n-2}\right\} \Rightarrow \|T^{\pi_{k+1}}\tilde{J}_{w_k} - T\tilde{J}_{w_k}\|_\infty \leq \delta$$
which satisfies the error condition.

2. Case 2: $\pi_k = \{a_1 = stay, a_j = stay, a_{otherwise} = \text{goto preceeding state}\}$
$$J^{\pi_k} = \left\{0, 0, \dots, \frac{g_j}{1-\gamma}, \frac{\gamma g_j}{1-\gamma}, \dots, \frac{\gamma^{n-j} g_j}{1-\gamma}\right\}$$
$$\tilde{J}_{w_k} = \left\{0, 0, \dots, \epsilon + \frac{g_j}{1-\gamma}, -\epsilon + \frac{\gamma g_j}{1-\gamma}, \dots, \frac{\gamma^{n-j} g_j}{1-\gamma}\right\}$$
Let $\pi_{k+1} = \{a_1 = stay, a_{j+1} = stay, a_{otherwise} = \text{goto preceeding state}\}$
$$T^{\pi_{k+1}}\tilde{J}_{w_k} = \left\{0, 0, \dots, 0, g_{j+1} - \gamma\epsilon + \frac{\gamma^2 g_j}{1-\gamma}, \left(-\epsilon + \frac{\gamma g_j}{1-\gamma}\right)\gamma, \dots, \left(-\epsilon + \frac{\gamma g_j}{1-\gamma}\right)\gamma^{n-j-1}\right\}$$
$$T\tilde{J}_{w_k} = \left\{0, 0, \dots, 0, T_{j+1}, \left(-\epsilon + \frac{\gamma g_j}{1-\gamma}\right)\gamma, \dots, \left(-\epsilon + \frac{\gamma g_j}{1-\gamma}\right)\gamma^{n-j-1}\right\}$$
where $T_{j+1} = \min\left\{g_{j+1} - \gamma\epsilon + \frac{\gamma^2 g_j}{1-\gamma}, \gamma\epsilon + \frac{\gamma^2 g_k}{1-\gamma}\right\}$
$$\|T^{\pi_{k+1}}\tilde{J}_{w_k} - T\tilde{J}_{w_k}\|_\infty \leq \left|g_{j+1} - \gamma\epsilon + \frac{\gamma^2 g_j}{1-\gamma} - \gamma\epsilon - \frac{\gamma^2 g_k}{1-\gamma}\right| = \delta$$
which satisfies the error condition.

3. If $\pi_k = \{a_1 = stay, a_n = stay, a_{otherwise} = \text{goto preceeding state}\}$,
$$J^{\pi_k} = \left\{0, 0, \dots, \frac{g_n}{1-\gamma}\right\}$$
For $n \to \infty, g_n \to \frac{\delta + 2\lambda\epsilon}{1-\lambda}$,
$$\|J^{\pi_k} - J^*\|_\infty = \frac{\delta + 2\lambda\epsilon}{(1-\lambda)^2}$$

4. Overall, the algorithm will go into oscillatory pattern.

□

**Doubt**: The same example can be also viewed as a stochastic shortest path problem, by interpreting $1-\gamma$ as a termination probability, we have $m = 1$ and $\rho_m = \gamma$. We thus conclude that the bound in stochastic shortest path problem is also tight, within a small constant factor.

## 6.3 APPROXIMATION POLICY EVALUATION USING USING TD($\lambda$)

- If the cost-to-go have large variance, $TD(\lambda)$ converges faster and leads to better performance than that obtained from $TD(1)$;
- $TD(\lambda)$ may converge to a different limit for different values of $\lambda$;
- Approximation $TD(\lambda)$ convergence's issue is much more complex.

### 6.3.1 Approximate Policy Evaluation Using $TD(1)$

We consider a stochastic shrotest path problem, with 0 being a cost-free absorbing state, and we assume that $\mu$ is a proper policy. And we use $\tilde{J}_w(s)$ to approximate $J^\pi(s)$, and fixed $J(0) = \tilde{J}_w(0) = 0$.

For $m$th trajectory $(s_0^m, s_1^m, \ldots, s_N^m)$, we update $w$ by Monte Carlo method.

$$w_{m+1} = \arg\min_w \frac{1}{2} \sum_{t=0}^{N-1} \left( \tilde{J}_{w_m}(s_t^m) - \sum_{k=t}^{N-1} g(s_k^m, \pi(s_k^m), s_{k+1}^m) \right)^2$$

$$w_{m+1} = w_m - \alpha \sum_{t=0}^{N-1} \nabla_w \tilde{J}_{w_m}(s_t^m) \left( \tilde{J}_{w_m}(s_t^m) - \sum_{k=t}^{N-1} g(s_k^m, \pi(s_k^m), s_{k+1}^m) \right)$$

Let $d_k^m = g(s_k^m, s_{k+1}^m) + \tilde{J}_{w_m}(s_{k+1}^m) - \tilde{J}_{w_m}(s_k^m)$, thus

$$\begin{aligned} w_{m+1} =& w_m + \alpha \sum_{t=0}^{N-1} \nabla_w \tilde{J}_{w_m}(s_t^m) \sum_{k=t}^{N-1} d_k^m \\ =& w_m + \alpha \sum_{k=0}^{N-1} d_k^m \sum_{t=0}^{k} \nabla_w \tilde{J}_{w_m}(s_t^m) \\ =& w_m + \alpha \sum_{t=0}^{N-1} d_t^m \sum_{k=0}^{t} \nabla_w \tilde{J}_{w_m}(s_k^m) \end{aligned}$$

- Off-line: all updates of the vector w are performed at the end of a trajectory;
- On-line: an update is performed subsequent to each transition.

$$w_{t+1}^m = w_t^m + \alpha d_t^m \sum_{k=0}^{t} \nabla_w \tilde{J}_{w_t^m}(s_k^m) \quad s.t. \quad w_0^m = w_m \wedge w_N^m = w_{m+1}$$

In linear approximation and first-visit method, off-line and on-line are same. The difference between the two variants is of second order in the stepsize $\alpha$ and is therefore inconsequenctial as the stepsize diminishes to zero.

#### 6.3.2 $TD(\lambda)$ for General $\lambda$

$$\begin{aligned}
w_{m+1} =& w_m + \alpha \sum_{t=0}^{N-1} \nabla_w \tilde{J}_{w_m}(s_t^m) \sum_{k=t}^{N-1} d_k \lambda^{k-t} \\
=& w_m + \alpha \sum_{k=0}^{N-1} d_k \sum_{t=0}^{k} \nabla_w \tilde{J}_{w_m}(s_t^m) \cdot \lambda^{k-t} \\
=& w_m + \alpha \sum_{t=0}^{N-1} d_t \sum_{k=0}^{t} \nabla_w \tilde{J}_{w_m}(s_k^m) \cdot \lambda^{t-k}
\end{aligned}$$

$$w_{t+1}^m = w_t^m + \alpha d_t \sum_{k=0}^{t} \lambda^{t-k} \nabla_w \tilde{J}_{w_t^m}(s_k^m)$$

Its convergence behavior is unclear. (?)
We now look deeper in $TD(0)$.

$$w_{t+1}^m = w_t^m + \alpha d_t \nabla_w \tilde{J}_{w_t^m}(s_t^m)$$

$TD(0)$ can be thought of as a stochastic approximation method for solving the Bellman equations

$$\forall s \in S, J(s) = \sum_{s' \in S} p_{ss'}^{\pi}(g^{\pi}(s, s') + J(s'))$$

for $S = \{0, 1, \dots, n\}$, we are trying to minimize

$$\sum_{s \in S} \left( \tilde{J}_w(s) - \sum_{s' \in S} p_{ss'}^{\pi}(g^{\pi}(s, s') + \tilde{J}_w(s')) \right)^2$$

An incremental gradient update based on the state s,

$$\begin{aligned}
w' =& w + \alpha \left( \nabla_w \tilde{J}_w(s) - \sum_{s' \in S} p_{ss'}^{\pi} \nabla_w \tilde{J}_w(s') \right) \left( \tilde{J}_w(s) - \sum_{s' \in S} p_{ss'}^{\pi}(g^{\pi}(s, s') + \tilde{J}_w(s')) \right) \\
=& w + \alpha \sum_{s' \in S} p_{ss'}^{\pi} \left( \tilde{J}_w(s) - g^{\pi}(s, s') - \tilde{J}_w(s') \right) \left( \nabla_w \tilde{J}_w(s) - \sum_{s' \in S} p_{ss'}^{\pi} \nabla_w \tilde{J}_w(s') \right) \\
=& w + \alpha \mathbb{E}_{s' \sim P_{s\cdot}^{\pi}} [d_w(s, s')] \left( \nabla_w \tilde{J}_w(s) - \sum_{s' \in S} p_{ss'}^{\pi} \nabla_w \tilde{J}_w(s') \right)
\end{aligned}$$

Thus, $TD(0)$ could be explained as an stochastically incremental gradient algorithm, but the term $d_w(s, s') \sum_{s' \in S} p_{ss'} \nabla_w \tilde{J}_w(s')$ is omitted, because it's not easy to predict.

Here is an example that $TD(\lambda)$ preforms badly.

**Example 5.**
- *$s = \{0, 1, 2, \dots, n\}$;*
- *There is only one policy $\pi = (s_0 = stay, a_i = goto\ s_{i-1} | i \geq 2)$;*

- *So the costs are fixed as $g_i$;*

- *We use a poor linear approximation:*$\tilde{J}_w(s) = ws$.

*Then we have* $d_t^m = g_{s_t^m} + \tilde{J}_{w_t^m}(s_{t+1}^m) - \tilde{J}_{w_t^m}(s_t^m) = g_{s_t^m} - w_t^m$. *If we always start sampling from state $n$, a complete trajectory* $\tau_{m-1} = \tau_m = \tau_{m+1} = (n, n-1, \ldots, 0)$

$$\begin{aligned}
w_{m+1} =& w_m + \gamma \sum_{t=0}^{n-1} d_t \sum_{k=0}^{t} \nabla_w \tilde{J}_{w_m}(s_k^m) \cdot \lambda^{t-k} \\
=& w_m + \gamma \sum_{t=0}^{n-1} (g_{n-t} - w_m) \sum_{k=0}^{t} (n-k) \lambda^{t-k} \\
=& w_m + \gamma \sum_{t=1}^{n} (g_t - w_m) \sum_{k=t}^{n} k \lambda^{k-t} \\
=& w_m \cdot \left( 1 - \gamma \sum_{t=1}^{n} \sum_{k=t}^{n} k \lambda^{k-t} \right) + \gamma \sum_{t=1}^{n} g_t \sum_{k=t}^{n} k \lambda^{k-t}
\end{aligned}$$

*If* $0 < \gamma < 2\big(\sum_{t=1}^{n} \sum_{k=t}^{n} k\lambda^{k-t}\big)^{-1}$, *the sequence $w_m$ is a contraction sequence, and converges to the scalar $\hat{w}(\lambda)$, which satisfies that*

$$\sum_{k=1}^{n} (g_k - \hat{w}(\lambda)) \sum_{k=t}^{n} k \lambda^{k-t} = 0$$

$$\hat{w}(1) = \frac{\sum_{t=1}^{n} g_t \sum_{k=t}^{n} k}{\sum_{t=1}^{n} \sum_{k=t}^{n} k} = \frac{\sum_{t=1}^{n} t \sum_{k=1}^{t} g_k}{\sum_{t=1}^{n} t^2}, \quad \hat{w}(0) = \frac{\sum_{t=1}^{n} t g_t}{\sum_{t=1}^{n} t}$$

*We go back to the sum of squared errors*

$$\sum_{s=1}^{n} \left( J(s) - \tilde{J}_w(s) \right)^2 \Rightarrow \sum_{s=1}^{n} s(J(s) - ws) = 0 \Rightarrow w = \frac{\sum_{t=1}^{n} t J(t)}{\sum_{t=1}^{n} t^2}$$

*Because* $J(t) = \sum_{k=1}^{t} g_t$, *therefore $\hat{w}(1)$ is the minimization of the sum of squared errors. In this problem, $TD(1)$ is poor approximation because of the approximation function and $TD(0)$ is worse because of $\lambda$.*

### $\gamma$ Discounted Problems

In the absence of an absorbing termination state, the trajectory never terminates and the entire algorithm involves a single infinitely long trajectory. It's necessary to gradually reduce $\gamma$ towards zero as the algorithm progresses.

$$d_t^m = g^\pi(s_t, s_{t+1}) + \gamma \tilde{J}_{w_t^m}(s_{t+1}^m) - \tilde{J}_{w_t^m}(s_t^m)$$

$$w_{t+1}^m = w_t^m + \alpha d_t^m \sum_{k=0}^{t} (\gamma\lambda)^{t-k} \nabla_w \tilde{J}_{w_t^m}(s_t^m)$$

$TD(\lambda)$ **Can Diverge for Nonlinear Architectures**

**Example 6.**

$$P^{\pi} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix} \quad Q = \begin{bmatrix} 1 & 1/2 & 3/2 \\ 3/2 & 1 & 1/2 \\ 1/2 & 3/2 & 1 \end{bmatrix}$$

*And the cost is zero.*

*Let* $\tilde{J}_w = (f_1(w), f_2(w), f_3(w))$, *and* $\frac{d\tilde{J}_w}{dw} = (Q + \epsilon I)\tilde{J}_w$, *s.t.* $f_1(0) + f_2(0) + f_3(0) = 0$. *Let* $F(w) = f_1(w)+f_2(w)+f_3(w)$, *then* $\frac{dF(w)}{dw} = (3+\epsilon)F(w), s.t. F(0) = 0$. *We can get* $F(w) = 0$.

*Because* $Q + Q^T = 2\vec{1}\vec{1}^T$, *therefore* $\tilde{J}_w^T(Q + Q^T)\tilde{J}_w = 0 \Rightarrow \tilde{J}_w^T Q \tilde{J}_w = 0$

$$\frac{d}{dw}\|\tilde{J}_w\|_2^2 = \tilde{J}_w^T(Q + Q^T)\tilde{J}_w + 2\epsilon\|\tilde{J}_w\|_2^2 = 2\epsilon\|\tilde{J}_w\|_2^2$$

*For a single infinitely long trajectory leads to the update equation*

$$w_{t+1} = w_t + \alpha_t \frac{d\tilde{J}_{w_t}(s_t)}{dw}(\gamma\tilde{J}_{w_t}(s_{t+1}) - \tilde{J}_{w_t}(s_t))$$

$$\begin{aligned}\mathbb{E}\left[\frac{d\tilde{J}_{w_t}(s_t)}{dw}(\gamma\tilde{J}_{w_t}(s_{t+1}) - \tilde{J}_{w_t}(s_t))\right] =& \frac{1}{3}\sum_{i=1}^{3}\frac{d\tilde{J}_{w_t}(i)}{dw}\left(\gamma\sum_{j=1}^{3}p_{ij}\tilde{J}_{w_t}(j) - \tilde{J}_{w_t}(i)\right) \\ =& \frac{1}{3}\Big((Q + \epsilon I)\tilde{J}_{w_t}\Big)^T(\gamma P - I)\tilde{J}_{w_t} \\ =& \frac{\gamma}{3}\tilde{J}_{w_t}^T Q^T P \tilde{J}_{w_t}^T + \frac{\epsilon}{3}\tilde{J}_{w_t}^T(\gamma P - I)\tilde{J}_{w_t} = \frac{dw}{dt}\end{aligned}$$

*If* $\epsilon = 0$,

$$\frac{dw}{dt} = \frac{\gamma}{6}\tilde{J}_{w_t}^T\left(Q^T P + P^T Q\right)\tilde{J}_{w_t}$$

*It's easy to verify that* $Q^T P + P^T Q \succ 0$, *which means*

$$\frac{dr}{dt} \geq c\|\tilde{J}_{w_t}\|_2^2$$

I have some question about this example. I will refer to Chapter4 ODE.

### 6.3.3 $TD(\lambda)$ with Linear Architectures-Discounted Problems

For $\vec{w} \in \mathbb{R}^K$, $\tilde{J}_{\vec{w}}(s) = \langle \vec{w}, \phi(s)\rangle$, where $s \in S = \{1, 2, \ldots, n\}$. And let

$$\Phi = [\phi_1, \ldots, \phi_K] = [\phi(1), \ldots, \phi(n)]^T$$

Then

$$\tilde{J}_{\vec{w}} = \Phi\vec{w} \Rightarrow \nabla_{\vec{w}}\tilde{J}_{\vec{w}}(s) = \phi(s)$$

$$\begin{aligned} w_{t+1}^m =& w_t^m + \alpha_t d_t \sum_{k=0}^{t}(\gamma\lambda)^{t-k}\nabla_w \tilde{J}_{w_t^m}(s_k) \\ =& w_t^m + \alpha_t d_t \sum_{k=0}^{t}(\gamma\lambda)^{t-k}\phi(s_k)\end{aligned}$$

Let $z_t = \sum_{k=0}^{t}(\gamma\lambda)^{t-k}\phi(s_k)$, then $w_{t+1}^m = r_t^m + \alpha_t d_t z_t$ and $z_{t+1} = \gamma\lambda z_t + \phi(s_{t+1})$

**Assumption 2.** *1.* $\sum_{t=0}^{\infty} \alpha_t = \infty$, *and* $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$;

*2.* $\forall s, s' \in S, \pi_\infty(s') = \forall \lim_{t\to\infty} P(s_t = s'|s_0 = s) > 0$;

*3.* $K \le n$ *and* $\Phi$ *has full colomn rank.*

We denote a steady-state probabilties $\pi_\infty = (\pi(1), \dots, \pi(n))$, which satisfies

$$\pi_\infty^T P = \pi_\infty^T$$

Define: $\|J\|_D^2 = J^T D J = \sum_{i=1}^n \pi_\infty(i) J^2(i)$

**Lemma 5.**

$$\|PJ\|_D \le \|J\|_D$$

$$\begin{aligned}\|PJ\|_D^2 =& J^T P^T D P J = \sum_{i=1}^n \pi_\infty(i) \left( \sum_{j=1}^n p_{ij} J(j) \right)^2 \le \sum_{i=1}^n \pi_\infty(i) \sum_{j=1}^n p_{ij} J^2(j) \\ &\le \sum_{j=1}^n \pi_\infty(j) J^2(j) = J^T D J = \|J\|_D^2\end{aligned}$$

$$d_t z_t = (g(s_t, s_{t+1}) + \gamma \phi(s_{t+1})^T w_t - \phi(s_t)^T w_t) z_t = z_t \left( \gamma \phi(s_{t+1})^T - \phi(s_t)^T \right) w_t + z_t g(s_t, s_{t+1})$$

We donte

$$d_t z_t = A(X_t) w_t + b(X_t)$$

The following is trying to calculate $\mathbb{E}_{\pi_\infty}[A(X_t)]$ and $\mathbb{E}_{\pi_\infty}[b(X_t)]$. Some trivial result:

1. $\mathbb{E}_{\pi_\infty}\left[J^T(s_0) J(s_m)\right] = J^T D P^m J$;
2. $\mathbb{E}_{\pi_\infty}\left[\phi(s_0) \phi^T(s_m)\right] = \Phi^T D P^m \Phi$;
3. 

$$\begin{aligned}\lim_{t\to\infty} \mathbb{E}_{\pi_\infty}[A(X_t)] =& \lim_{t\to\infty} \mathbb{E}_{\pi_\infty}\left[ \sum_{k=0}^t (\gamma\lambda)^{t-k} \phi(s_k)(\gamma \phi^T(s_{t+1}) - \phi^T(s_t)) \right] \\ =& \lim_{t\to\infty} \sum_{k=0}^t (\gamma\lambda)^{t-k} \Phi^T D \left[ \gamma P^{t-k+1} - P^{t-k} \right] \Phi \\ =& \lim_{t\to\infty} \sum_{m=0}^t (\gamma\lambda)^m \Phi^T D \left[ \gamma P^{m+1} - P^m \right] \Phi \\ =& \Phi^T D \left( (1-\lambda) \sum_{m=0}^{\infty} \lambda^m (\gamma P)^{m+1} - I \right) \Phi\end{aligned}$$

4.

$$\begin{aligned}
\lim_{t\to\infty}\mathbb{E}_{\pi_\infty}\left[b(X_t)\right] &= \lim_{t\to\infty}\mathbb{E}_{\pi_\infty}\left[\sum_{k=0}^{t}(\gamma\lambda)^{t-k}\phi(s_k)g(s_t,s_{t+1})\right]\\
&=\lim_{t\to\infty}\sum_{k=0}^{t}(\gamma\lambda)^{t-k}\mathbb{E}_{\pi_\infty}\left[\phi(s_k)g(s_t,s_{t+1})\right]\\
&=\lim_{t\to\infty}\sum_{k=0}^{t}(\gamma\lambda)^{t-k}\Phi^T DP^{t-k}\sum_{s'\in S}g(s,s')\\
&=\lim_{t\to\infty}\sum_{m=0}^{t}(\gamma\lambda)^{m}\Phi^T DP^{m}\sum_{s'\in S}g(s,s')\\
&=\Phi^T D\sum_{m=0}^{\infty}(\gamma\lambda P)^m\bar{g}, \quad \left(where\ \bar{g}=\sum_{s'\in S}g(s,s')\right)
\end{aligned}$$

5. Denote $M=(1-\lambda)\sum_{m=0}^{\infty}\lambda^m(\gamma P)^{m+1}$.

$$\|MJ\|_D \leq (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\|P^{m+1}J\|_D \leq \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\|J\|_D$$

6. $\mathbb{E}_{\pi_\infty}\left[A(X_t)\right]=A\prec 0$

$$\begin{aligned}
J^TDMJ =& J^TD^{1/2}D^{1/2}MJ \leq \|D^{1/2}J\|_2\cdot\|D^{1/2}MJ\|_2=\|J\|_D\|MJ\|_D\\
\leq&\frac{\gamma(1-\lambda)}{1-\gamma\lambda}\|J\|_D\cdot\|J\|_D=\frac{\gamma(1-\lambda)}{1-\gamma\lambda}J^TDJ\leq\gamma J^TDJ
\end{aligned}$$

$$J^TD(M-I)J\leq(\gamma-1)J^TDJ<0, \quad \forall J\neq 0$$

I need go back to Chapter4. Here I give up again. Martingales theorem is too ugly. To use convergence theorem in chapter4, we need to proof: $\exists C,\rho$

- $$\|\mathbb{E}\left[A(X_t)|X_0=X\right]-A\|\leq C\rho^t.$$

  *Proof.*

$$\mathbb{E}\left[A(X_t)|X_0=X\right]=\mathbb{E}\left[z_t\phi^T(s_t)|s_0,s_1,z_0\right]=\mathbb{E}\left[\sum_{k=0}^{t}\phi(s_m)(\gamma\lambda)^{t-k}\phi^T(s_t)|s_0,s_1,z_0\right]$$

$$=A+\mathbb{E}\left[\sum_{k=0}^{t}\phi(s_m)(\gamma\lambda)^{t-k}\phi^T(s_t)|s_0,s_1,z_0\right]-\lim_{t'\to\infty}\mathbb{E}_{\pi_\infty}\left[\sum_{k=0}^{t'}\phi(s_m)(\gamma\lambda)^{t'-k}\phi^T(s_{t'})\right]$$

$$\lim_{t'\to\infty}\mathbb{E}_{\pi_\infty}\left[\sum_{k=t+1}^{t'}\phi(s_m)(\gamma\lambda)^{t'-k}\phi^T(s_{t'})\right]\leq M(\gamma\lambda)^t\sum_{m=1}^{\infty}(\lambda\gamma)^m=\frac{\gamma\lambda M}{1-\gamma\lambda}(\gamma\lambda)^t$$

(S is finite?)

$$\sum_{k=0}^{t}\sum_{j=1}^{n}\left(P\left(s_m=j|s_1\right)-\pi_\infty(j)\right)\phi(j)\mathbb{E}\left[\phi^T(s_t)|s_m=s\right]$$

And $P\left(s_m=j|s_1\right)$ converges to $\pi_\infty$ exponentially fast in m. $\square$

- $$\|\mathbb{E}\left[b(X_t)|X_0=X\right]-b\|\leq C\rho^t$$

*Proof.*

$$\mathbb{E}\left[b(X_t)|X_0=X\right]=b+\mathbb{E}\left[\sum_{k=0}^{t}(\gamma\lambda)^{t-k}\phi(s_k)g(s_t,s_{t+1})|s_0,s_1,z_0\right]-b$$

$\lim_{k\to\infty}\left[\sum_{k=t}^{t'}(\gamma\lambda)^{t-k}\phi(s_k)g(s_t,s_{t+1})\right]$ convergences exponentially. And

$$\sum_{k=0}^{t}\sum_{j=1}^{n}\left(P(s_m=j|s_1)-\pi_\infty(j)\right)\phi(j)\mathbb{E}\left[g(_t,s_{t+1})|s_m=s\right]$$

also convergences exponentially. $\square$

The $TD(\lambda)$ algorithm convergence to $Ar^\infty+b=0$.

$$\Phi^TD\left((1-\lambda)\sum_{m=0}^{\infty}\lambda^m(\gamma P)^{m+1}-I\right)\Phi\cdot r^\infty+\Phi^TD\sum_{m=0}^{\infty}(\gamma\lambda P)^m\sum_{s'\in S}g(s,s')=0$$

**Definition 2.** *Let* $\sum_{s'\in S}g(s,s')=\bar{g}$

$$T_\pi^{(\lambda)}J=(1-\lambda)\sum_{m=0}^{\infty}\lambda^m(\lambda P^\pi)^{m+1}J+\sum_{m=0}^{\infty}(\gamma\lambda P^\pi)^m\bar{g}=MJ+q$$

It's easy to verify that $T_\pi^{(\lambda)}J=J^\pi$

$$Ar^\infty+b=\Phi^TDT_\pi^{(\lambda)}(\Phi r^\infty)-\Phi^TD\Phi r^\infty=0$$

$$\Rightarrow\Phi r^\infty=\Pi T_\pi^{(\lambda)}(\Phi r^\infty),\quad\Pi=\Phi(\Phi^TD\Phi)^{-1}\Phi^TD$$

**Lemma 6.**

$$\|\Pi J-J\|_D=\min_r\|\Phi r-J\|_D$$

$J^\Pi=\Phi r^\infty$ is the fixed point of $\Pi T_\pi^{(\lambda)}$, $J^\pi$ is the fixed point of $T_\pi^{(\lambda)}$. Then we want estimate

**Lemma 7.**

$$\|\Phi r^\infty-J^\pi\|_D=\|J^\Pi-J^\pi\|_D\leq\frac{1-\gamma\lambda}{1-\gamma}\|\Pi J^\pi-J^\pi\|_D$$

$$
\begin{aligned}
\|J^\Pi - J^\pi\|_D \leq & \|J^\Pi - \Pi J^\pi\|_D + \|\Pi J^\pi - J^\pi\|_D \\
\leq & \|\Pi T_\pi^{(\lambda)} J^\Pi - \Pi T_\pi^{(\lambda)} J^\pi\|_D + \|\Pi J^\pi - J^\pi\|_D \\
\leq & \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\|J^\Pi - J^\pi\|_D + \|\Pi J^\pi - J^\pi\|_D
\end{aligned}
$$

The case of an infinite state space (either discrete or continuous) has note been addressed before in this book.

## 6.4 $TD(\lambda)$ with Linear Architectures — Stochastic Shortes Path Problems

Assumption

1. $\sum_{k=0}^\infty \gamma_k = \infty, \sum_{k=0}^\infty \gamma_k^2 < \infty$;
2. all states have positive probability of being visited by the algorithm;
3. $\Phi$ has full columns ranks.

We use off-line method.

$$
\begin{aligned}
A =& \mathbb{E}^\pi \left[ \sum_{t=0}^{N-1} z_t(\phi^T(s_{t+1}) - \phi^T(s_t)) \right] \\
=& \mathbb{E}^\pi \left[ \sum_{t=0}^{N-1} \sum_{k=0}^{t} \lambda^{t-m} \phi(s_m)(\phi^T(s_{i+1}) - \phi^T(s_t)) \right] \\
=& \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \sum_{k=0}^{t} \lambda^{t-m} \phi(s_m)(\phi^T(s_{i+1}) - \phi^T(s_t)) \right] \\
=& \Phi^T \sum_{t=0}^{\infty} \sum_{k=0}^{t} Q_k (\lambda P)^{t-k} (P - I)\Phi
\end{aligned}
$$

$$
B = \sum_{t=0}^\infty Q_t \sum_{k=1}^\infty (\lambda P)^k (P-I) = \sum_{t=0}^\infty Q_t \left( (1-\lambda) \sum_{k=0}^\infty \lambda^k P^{k+1} - I \right) = Q(M-I)
$$

**Lemma 8.**

$$B \prec 0.$$

*Proof.* 1. $\lambda = 1$, $B = -Q \prec 0$.

2. $0 < \lambda < 1$.
   Let $q_t = diag(Q_t)$ and $q = diag(Q)$.
   $q_t^T P = q_{t+1}^T \Rightarrow q^T P = q - q_0 \preceq q$.
   $\|PJ\|_Q \leq \|J\|_Q, P^k J \to 0 \Rightarrow \|MJ\|_Q < \|J\|_Q$.
   More specifically, $\exists \rho > 0, \|MJ\|_Q \leq \rho \|J\|_Q$.

$$J^T QMJ \leq \|J\|_Q \|MJ\|_Q \leq \rho J^T DJ$$

$$J^T Q(M-I)J \preceq -(1-\rho) J^T DJ \prec 0$$

3. $\lambda = 0$: $M = P$. For Markov matrix property, $P$'s eigenvalues are all nonnegative. And for assumption $P^k \to 0$, the eigenvalues are all smaller than 1. Then $Q(P - I)$ is negative definite.

$\square$

**Error Bounds**

$$\|\Phi r^\infty - J^\pi\|_Q \leq \frac{\|\Pi J^\pi - J^\pi\|_Q}{1 - \beta}$$

where $\beta$ is the contraction factor of the operator $T^{(\lambda)}$, which is equal to the contraction factor of matrix

$$M = (1 - \lambda) \sum_{k=0}^{\infty} \lambda^k P^{k+1}.$$

If $\lambda = 1$, we have $M = 0$ and $\beta = 0$, and we obtain the most favorable bound.

## 6.5 OPTIMISTIC POLICY ITERATION