Rishabh Indoria
24002
M.Tech AI

Reinforcement Learning

April 12, 2025

1 Introduction

- RL: An Introduction, 2^{nd} ed, `https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf`
- Neuro Dynamic Programming, Dmitri Brisckas, 1996
- Optimal Control and Dynamic Programming, Volume 1+2, 2012/13
- Reinforcement Learning and Optimal Control, 2019
- Midterm(20) around Feb 15, Quiz(5) around Jan end, Course Project(20), Final Exam(30)
- **Goal**: To select a sequence of actions depending on states of environment that maximizes "long term reward".
- We look at expectation of sum of rewards, throwing out randomness.
- Probabilistic transition between states $P(S_{t+1} = s', r_{t+1} = r | S_t = s, A_t = a)$
- **N**: Number of states of decision-making
- **Finite Horizon Problem**: $N < \infty$, and N is a deterministic number.
- **Episodic/Stochastic/Shortest path problems**: $N < \infty$ with probability 1.

$$\text{Long Term Reward: } E[\sum_{i=1}^{N} r_i | S_0 = s]$$

- **Non-Terminating Problems**: $N = \infty$

$$\text{Discounted Rewards: } \lim_{N\to\infty} E[\sum_{i=0}^{N} \gamma^i r_i | S_0 = s] \text{ where } 0 < \gamma < 1$$

$$\text{Long-run average rewards: } \lim_{N\to\infty} \frac{1}{N} E[\sum_{i=1}^{N} r_i | S_0 = s]$$

- **Key Ingredients to a RL Problem**: State(Environment), Action(Agent), and Reward
- RL problems essentially deal with **Exploration vs Exploitation**.
- Agent interacts with its environment, it learns through this interaction.
- Agent looks at the state of the environment, then takes an action.
- State of the environment is changed, and the agent gets a reward/punishment.
- **Policy**: A decision rule. In a given state, it prescribes the action to be chosen.
- Policy can be deterministic or stochastic.
- **Objective**: Find a policy(π) that maximizes the value function.
- There are two parts to an RL problem
  1. **Prediction Problem**: Given a policy(π), estimate the value function.
  2. **Control Problem**: Find the best policy
- Solving the control problem requires solving the prediction problem first.
- Policy may not be unique, however value function is unique.
- We will inherently assume Markovian Property

$$P(s_{t+1} = s' | s_t = s, a_t = a, s_{t-1} = s_1, a_{t-1} = a_1, ...) = P(s_{t+1} = s' | s_t = s, a_t = a)$$

- Read Chapter 1

2 Multi-Armed Bandit

2.1 Problem Formulation

- Consider multiple slot machines in a casino, each gives a different reward.
- We will abstract this into a single slot machine with K arms.
- This is a single state problem, where k actions can be taken.
- **Problem**: Each time we pull an arm, we get a random reward.
- **Assumption**: The rewards from various arms follow a distribution, which is different for different arms.
- Suppose $q^*(a) = E[R_{t+1}|A_t = a]$, where $a \in \{1, 2, ..., k\}$
- **Goal**: Figure out the best arm, find $a^* \in \arg\max_a q^*(a)$

2.2 Strategies to find a^*

- We define an estimate of $q^*(a)$ at time n

$$Q_n(a) = \frac{\sum_{i=1}^{n} R_i I_{A_{i-1}} = a}{\sum_{i=1}^{n} I_{A_{i-1}} = a}$$

- **Greedy Policy**: Select action a such that $a \in \arg\max_{b\in\{1,2,\dots,k\}} Q_n(b)$
- This is not a good strategy because it does not allow for exploration.
- **ϵ-Greedy Policy**: Select action a such that

$$a = \begin{cases} \arg\max_a Q_n(a), & \text{with probability } 1-\epsilon \\ \text{random action with probability } \epsilon \end{cases}$$

- A good strategy is to start at $\epsilon = 1$ and then degrade.
- Read Chapter 2.3
- Suppose we decide to select an action a always, then $Q_n(a) = \frac{\sum_{i=1}^{n} R_{i+1}}{n}$
- This can be calculated iteratively as

$$Q_{k+1}(a) = Q_k(a) + \frac{1}{k+1}(R_{k+1} - Q_k(a))$$

- By strong law of large numbers, we can say that

$$Q_n(a) \xrightarrow{a.s} q^*(a) \text{ as } n \to \infty$$

- We can also use, $Q_{k+1}(a) = \alpha R_{k+1} + (1-\alpha)Q_k(a)$ which will be a weaker convergence than before.
- Unrolling the above, we get

$$\begin{aligned} Q_{k+1}(a) =& \alpha R_{k+1} + \alpha(1-\alpha)R_{k+1} + (1-\alpha)^2 Q_{k-1}(a) \\ =& \alpha R_{k+1} + \alpha(1-\alpha)R_k + \alpha(1-\alpha)^2 R_{k-1} + ... + \alpha(1-\alpha)^n R_1 + Q_0(a)(1-\alpha)^{n+1} \end{aligned}$$

- These are called **Fading Memory** based algorithms.
- We call $\frac{1}{t+1} \triangleq \alpha_t$ as the step size or learning rate.
- Let $Q_{t+1}(a) = Q_t(a) + \alpha_t(R_{t+1} - Q_t(a))$ with arbitrary $\alpha_t > 0$ such that

$$\sum_t \alpha_t = \infty \text{ and } \sum_t \alpha_t^2 < \infty$$

- Examples of α_t, $t \geq 0$

$$\alpha_t = \frac{1}{t+1}, \quad \alpha_t = \frac{1}{(t+1)^\beta}, \beta \in (0.5, 1) \quad \alpha_t = \frac{1}{(t+1)\log(t+1)}, \quad \alpha_t = \frac{\log(t+1)}{(t+1)}$$

- It can be shown that these also satisfy $Q_n(a) \xrightarrow{a.s} q^*(a)$ as $n \to \infty$
- Such algorithms are often referred as **Stochastic Approximation Algorithm**.
- Refer to paper https://www.columbia.edu/~ww2040/8100F16/RM51.pdf.
- Consider a function $f : \mathbb{R}^d \to \mathbb{R}^d$ for which we want to find $x \in \mathbb{R}^d$ such that $f(x) = 0$.
- **Problem**: f is not known. We however have access to noisy evaluation of the function f.

$$x_{t+1} = x_t + \alpha_t(f(x_t) + \Sigma_t) \to \text{ Noisy Estimation}$$

- Start from some arbitrary $x_0 \in \mathbb{R}^d$ and iterate.
- One can show that under some conditions $x_t \to x^*$ as $t \to \infty$ such that $f(x^*) = 0$.
- Applications of this include
  1. Find a fixed point of a function $F : \mathbb{R}^d \to \mathbb{R}^d$, i.e., we want to find x^* such that $F(x^*) = x^*$, set $f(x) = F(x^*) - x^*$
  2. Find minimum of a function $F : \mathbb{R}^d \to \mathbb{R}^d$, set $f(x) = -\nabla F(x)$
- For our case,

$$\begin{aligned} Q_{t+1}(a) = Q_t(a) =& \alpha_t(R_{t+1} - Q_t(a)) \\ &= Q_t(a) + \alpha_t(E[R_{t+1}|A_t = a] + (R_{t+1} - E[R_{t+1}|A_t = a]) - Q_t(a)) \end{aligned}$$

- Here $x = Q_t(a)$, $f(x) = E[R_{t+1}|A_t = a]$ and $\Sigma_t = R_{t+1} - E[R_{t+1}|A_t = a]$ which means, this is expected to converge to $q^*(a)$ such that $E[R_{t+1}|A_t = a] = q^*(a)$.

2.3 Upper Confidence Bound

- We take actions as $A_t = \arg\max_a [Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}}]$
- $N_t(a) =$ Number of times arm a is pulled.
- $c > 0$ is the exploration parameter.
- As $N_t(a)$ increases, the effect of exploration dies down.
- Suppose $R_1, R_2, ...$ be independent and subgaussian.
- Let $Q_n(i) = \frac{1}{n}\sum_{t=1}^{n} R_t$, using Hoeffding's inequality, we get

$$\begin{aligned} P(Q_n(i) \geq \epsilon) &\leq e^{-\frac{n\epsilon^2}{2}} \triangleq \delta \\ \implies e^{\frac{n\epsilon^2}{2}} &\geq \frac{1}{\delta} \text{ or } \frac{n\epsilon^2}{2} \geq \log(\frac{1}{\delta}) \\ \implies \epsilon &\geq \sqrt{\frac{2}{n}\log(\frac{1}{\delta})} \end{aligned}$$

- If $\delta = \frac{1}{n}$, then a good candidate for estimate of mean reward is

$$Q_n(i) + \sqrt{\frac{2}{N_n(i)}\log(n)}$$

2.4 Gradient Based Algorithms

- Let's say that $H_t(a)$ is a preference for action a at time t.
- Now we can select action as **Gibb's/Boltzmann Policy**

$$P(A_t = a) \triangleq \pi_t(a) = \frac{e^{H_t(a)}}{\sum_{b=1}^{n} e^{H_t(b)}}$$

- We update $H_t(a)$ using the following gradient ascent algorithm

$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial E[R_t]}{\partial H_t(a)} \text{ where } E[R_t] = \sum_x \pi_t(x) q_*(x), \text{ and } q_*(x) = E[R_{t+1}|A_t = x]$$

$$\frac{\partial E[R_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)}[\sum_x \pi_t(x) q_*(x)] = \sum_x q_*(x)(\frac{\partial \pi_t(x)}{\partial H_t(a)}) = \sum_x (q_*(x) - B_t)(\frac{\partial \pi_t(x)}{\partial H_t(a)})$$

- B_t is independent of $H_t(a)$ and as such doesn't contribute to gradient.
- It is added because a judicial choice can decrease variance/oscillations.

3 Markov Decision Processes

- Reference: Chapter 1 of Optimal Control and Dynamic Programming Volume 1
- Key idea is **Controlled Markov Chain**
- We will work with state space S, and action space A
- Given state $s \in S$, let $A(s)$ the set of feasible actions in state s. Then,

$$A = \cup_{s \in S} A(s)$$

- Let $\{X_n\}$ be a stochastic process that depends on a control valued sequence $\{Z_n\}$. We assume that $Z_n \in A(X_n) \forall n$.
  Then $\{X_n\}$ is a controlled Markov Chain if

  $$P(X_{n+1} = j | X_n = i, Z_n = a, X_{n-1} = i_{n-1}, Z_{n-1} = z_{n-1}, ...) = P(X_{n+1} = j | X_n = i, Z_n = a) \ \forall n \triangleq P(i, a, j)$$

  Note that $P(i, a, j) \in [0, 1] \ \forall i, a, j$ and $\sum_j P(i, a, j) = 1$

- A **Markov Decision Process** is a controlled Markov chain with a cost associated with every transition, $g(i_n, a_n, i_{n+1})$ where $i_n = X_n$, $a_n = Z_n$, $i_{n+1} = X_{n+1}$

3.1 Dynamic Programming

- We will start with finite horizon problem ($N < \infty$, N is deterministic)
- A **policy** is a decision rule specified as $\pi = \{\mu_0, \mu_1, ..., \mu_{N-1}\}$ where $\mu_k : S \to A$ such that $\mu_k(s) \in A(s)$ $\forall s \in S \ \forall k$. N is called the **terminal instant**.
- Broadly, our goal is to find an "optimal" policy.
- Let, $J_\pi(x_0) = E[g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), x_{k+1}) | X_0 = x_0]$
  finite horizon cost $= E$[Terminal cost + Cost of each action/single stage cost$|X_0 = x_0$]
- The expectation in $J_\pi(x_0)$ is taken over the joint distribution of $g, X_1, X_2, ..., X_N$.
- Let Π be the set of all policies. An optimal policy $\pi^* \in \Pi$ is such that

$$\text{Optimal Cost} \to J^*(x_0) \triangleq J_{\pi^*}(x_0) = \min_{\pi \in \Pi} J_\pi(x_0) \ \forall x_0 \in S$$

- There can be multiple optimal policies.
- **Principle of Optimality**: Let $\pi^* = \{\mu_0^*, \mu_1^*, \ldots, \mu_{N-1}^*\}$ be an optimal policy. Assume that following π^*, state x_i occurs in stage i with positive probability. Consider the following subproblem starting in state x_i at time i.

$$\min_{\pi^i = \{\mu_i, \mu_{i+1}, \ldots, \mu_{N-1}\}} E\left[g_N(x_N) + \sum_{k=1}^{N-1} g_k(x_k, \mu_k(x_k), x_{k+1}) \mid X_i = x_i\right]$$

- Then the truncated policy $\pi^* = \{\mu_i^*, \mu_{i+1}^*, \ldots, \mu_{N-1}^*\}$ is optimal for this tail subproblem.
- If this is not the case, then we can replace $\{\mu_i^*, \mu_{i+1}^*, \ldots, \mu_{N-1}^*\}$ in the optimal policy to achieve a lower cost, which leads to a contradiction. This means that the policy doesn't depend on the initial state.

- **Dynamic Programming Algorithm**
  Proposition: For every initial state $x_0 \in S$, the optimal cost $J^*(x_0) = J_0(x_0)$ is obtained from the last step of the following algorithm

$$J_N(x_N) = g_N(x_N)$$
$$J_k(x_k) = \min_{a_k \in A(x_k)} E_{x_{k+1}}[g_k(x_k, a_k, x_{k+1}) + J_{k+1}(x_{k+1})] \quad \forall k = N-1, N-2, \ldots, 0$$

  Do this for all states $x_k, \ldots, x_N$.
  Here $x_{k+1} = P(\cdot|x_k, a_k)$
  Also if $u_k^* = \mu_k^*(x_k)$ minimizes the RHS of $g(z)$ $\forall x_k$, $\forall k$, then the policy $\pi^* = (\mu_0^*, ...,, \mu_{N-1}^*)$ is optimal.

- Proof: For any policy $\pi = \{\mu_0, \mu_1, \ldots, \mu_{N-1}\}$, let $\pi^k = \{\mu_k, \mu_{k+1}, \ldots, \mu_{N-1}\}$ where $k = 0, 1, \ldots, N-1$

$$\text{Let } J_k^*(x_k) = \min_{\pi_k} E_{x_{k+1}, \ldots, x_N} \left[g_N(x_N) + \sum_{i=k}^{N-1} g_i(x_i, \mu_i(x_i), x_{i+1}) \mid X_k = x_k \right]$$

- This is the optimal cost for $(N-k)$th stage subproblem.
  Let $J_N^*(x_N) = g_N(x_N) = J_N(x_N)$ $\forall x_k \in S$.

- We will show by induction that $J_k^*(x_k) = J_k(x_k)$ $\forall x_k \in S$, $\forall k \geq 0$.

- Assume that for some k and all x_{k+1}

$$J_{k+1}^*(x_{k+1}) = J_{k+1}(x_{k+1})$$

  Note that $\pi^k = \{\mu_k, \pi^{k+1}\}$. Thus, $\forall x_k$,

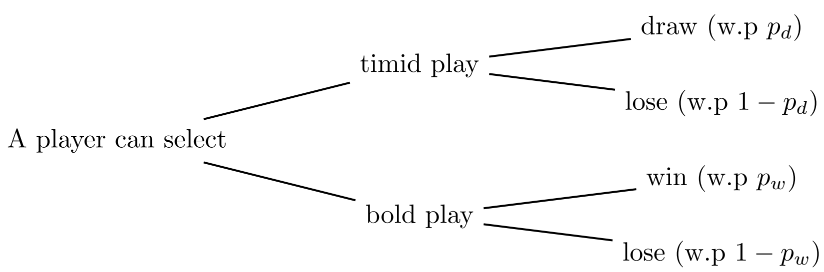
$$\begin{aligned}
J_k^*(x_k) &= \min_{\mu_k, \pi^{k+1}} E_{x_{k+1}, \ldots, x_N}[g_N(x_N) + g_k(x_k, \mu_k(x_k), x_{k+1}) + \sum_{i=k+1}^{N-1} g_i(x_i, \mu_i(x_i), x_{i+1}) | X_k = x_k] \\
&= \min_{\mu_k} E_{x_{k+1}}[g_k(x_k, \mu_k(x_k), x_{k+1}) + \min_{\pi^{k+1}} E[g_N(x_N) + \sum g_i(x_i, \mu_i(x_i), x_{i+1}) | x_{k+1}] | X_k = x_k] \\
&= \min_{\mu_k} E_{x_{k+1}}[g_k(x_k, \mu_k(x_k), x_{k+1}) + J_{k+1}^*(x_{k+1}) | X_k = x_k] \\
&= \min_{\mu_k} E_{x_{k+1}}[g_k(x_k, \mu_k(x_k), x_{k+1}) + J_{k+1}(x_{k+1}) | X_k = x_k]
\end{aligned}$$

  Here $\mu_k : S \to A$ such that $\mu_k(x) \in A(x)$ $\forall x$

$$\begin{aligned}
&= \min_{a_k \in A(x_k)} E_{x_{k+1}}[g_k(x_k, \mu_k(x_k), x_{k+1}) + J_{k+1}(x_{k+1}) | X_k = x_k] \\
&= J_k(x_k)
\end{aligned}$$

  Hence Proved

- Chess Match between player & opponent
  Aim: Formulate optimal strategy for player.



  Score assignment: win(1), draw(0.5). lose(0)
  State = Net score = points of players − points of opponent (maximization)
  Intermediate rewards $r_k(x_k, a_k, x_{k+1}) = 0$ $\forall k = 0, 1, \ldots, N-1$
  Only $r_N(x_N)$, terminal reward $J_N(x_N)$
  Optimal reward to go at kth state

$$J_k(x_k) = \max \left\{ p_d J_{k+1}(x_k) + (1 - p_d) J_{k+1}(x_k - 1), p_w J_{k+1}(x_k + 1) + (1 - p_w) J_{k+1}(x_k - 1) \right\}$$

Here,

$$J_N(x_N) = \begin{cases} 1, \text{ if } x_N > 0 \\ p_w, \text{ if } x_N = 0 \\ 0, \text{ if } x_N < 0 \end{cases}$$

Assume $p_d > p_w$, and solve for J_{N-1}

- Control of a queue (Arrivals $\rightarrow$ Departures)
  Assume buffer size $= n$
  Customers arrivals/departures happen at time $0, 1, ..., N-1$
  System can serve only one customer during a period.
  A customer can take multiple periods to service.
  Let p_m: probability of m arrivals in a period.
  Assume that the number of arrivals in a period is independent of the number in any other period.
  Types of service: Slow service (cost c_s), Fast service (cost c_f).
  With fast service, customer leaves system w.p q_f with slow service q_s.
  Same customer can be serviced with different types of service in different periods.
  Let $r(i)$: holding cost of i customers in a period.
  $R(i)$: terminal cost where i customers remain at time N.

  $$\text{Single stage cost} = \begin{cases} r(i) + c_f, \text{ fast service} \\ r(i) + c_s, \text{ slow service} \end{cases}$$

  Transition probabilities

  $$p_{0j}(u_f) = p_{0j}(u_s) = p_j, \; j = 0, 1, ..., n-1$$
  $$p_{0n}(u_f) = p_{0n}(u_s) = \sum_{j=n}^{\infty} p_j$$
  $$p_{ij}(u_f) = \begin{cases} 0, \text{ if } j < i-1 \\ q_f p_0, \text{ if } j = i-1 \\ q_f p_{j-i+1} + (1-q_f)p_{j-1}, \text{ if } i-1 < j < n-1 \\ q_f \sum_{m=n-i}^{\infty} p_m + (1-q_f)p_{n-1-i}, \text{ if } j = n-1 \\ (1-q_f)\sum_{m=n-i}^{\infty} p_m, \text{ if } j = m \end{cases}$$

  Similarly, define for slow service

  $$J_N(i) = R(i), \; i = 0, 1, ..., n$$
  $$J_k(x_k) = \min[r(i) + c_f + \sum_{j=0}^{n} p_{ij}(u_f)J_{k+1}(j), r(i) + c_s + \sum_{j=0}^{n} p_{ij}(u_s)J_{k+1}(j)]$$

3.2 Stochastic Shortest Path Problems

- Reference: chapter 2 of Optimal Control and Dynamic Programming Volume 2
- SSP problems are characterized by a goal state/terminal state.
- These are referred to as **episodic problems**.
- We assume that there is a cost free terminal state O and taking any action in state O will result in staying the same state O.

  $$P_{OO}(u) = 1, \; g(O, u, O) = 0 \; \forall u \in A(O)$$

- We will take no discounting factor, i.e., $\gamma = 1$
- **Problem**: How do we reach the terminal state with minimum expected cost

  $$J_\mu(i) \triangleq \lim_{N \to \infty} E_\mu[\sum_{n=0}^{N-1} g(s_n, \mu(s_n), s_{n+1}) | s_0 = i]$$

- We consider a stationary policy of the form $\pi = \{\mu, \mu, \ldots\} = \mu$.
- Many times it is convenient to simply call μ as our policy.

- A stationary policy is said to be **proper** if

$$P_\mu = \max_{i=1,...,n} P(s_n \neq O | s_0 = i, \mu) < 1$$

- **State space**: Non-terminal states $\{1, 2, ..., n\}$, Terminal states $\{O\}$.
- A stationary policy that is not proper is called **improper**.
- For a proper μ, in the Markov chain corresponding to μ, there is a positive probability from each state to a terminal state.
- Consider a MDP $\{X_n\}$ governed by a stationary policy μ

  $P(X_{n+1} = j | X_n = i, Z_n = a, X_{n-1} = i_{n-1}, Z_{n-1} = a_{n-1}, ..., X_0 = i_0, Z_0 = a_0) = P(X_{n+1} = j | X_n = i, Z_n = a)$

  then governed by μ, this becomes

  $P(X_{n+1} = j | X_n = i, Z_n = \mu(i), X_{n-1} = i_{n-1}, Z_{n-1} = \mu(i_{n-1}), ..., X_0 = i_0, Z_0 = \mu(i_0)) = P(X_{n+1} = j | X_n = i, Z_n = \mu$

  Let $P(X_{n+1} = j | X_n = i, Z_n = \mu(i)) \triangleq P_\mu(i, j)$
  then $P_\mu(i, j) \geq 0 \; \forall j \in S$ and $\sum_{j \in S} P_\mu(i, j) = 1 \; \forall i \in NT$.

- Since μ is independent of time, this is a homogenous Markov chain.
- If μ were changing with time, then this would be a non-homogenous Markov chain.

  $$\begin{aligned} P(s_{2n} \neq O | s_0 = u, \mu) &= P(s_{2n} \neq O | s_0 = u, s_n \neq O, \mu) P(s_n \neq O | s_0 = u, \mu) + P(s_{2n} \neq O | s_0 = u, s_n = O, \mu) P(s_n = O | s \\ &= P(s_{2n} \neq O | s_0 = u, s_n \neq O, \mu) P(s_n \neq O | s_0 = u, \mu) \\ &\leq P_\mu \qquad\qquad\qquad\qquad P_\mu \quad \text{(By Markov Property)} \\ &\leq P_\mu^2 \end{aligned}$$

- More generally, $P(s_k \neq O | s_0 = u, \mu) \leq P_\mu^{\lfloor \frac{k}{n} \rfloor}$
- From this we get,

$$\lim_{k \to \infty} P(s_k \neq O | s_0 = u, \mu) = 0$$

- Assuming μ is proper we get

$$\begin{aligned} J_\mu(i) &= \lim_{N \to \infty} E_\mu[\sum_{m=0}^{N-1} g(s_m, \mu(s_m), s_{m+1}) | S_0 = i] \\ &= E_\mu[\sum_{m=0}^{\infty} g(s_m, \mu(s_m), s_{m+1}) | S_0 = i] \end{aligned}$$

  P_μ essentially plays the role of discount factor.
  We assume that $|g(i, a, j)| \leq k \; \forall i \; \forall a \in A(i) \; \forall j$. Then,

$$\begin{aligned} |J_\mu(i)| &\leq \sum_{m=0}^{\infty} E_\mu[|g(s_m, \mu(s_m), s_{m+1})| | s_0 = i] \\ &= \sum_{m=0}^{\infty} \sum_j \sum_k P_{ij}^m(\mu) P_{jk}(\mu) |g(j, \mu(j), k)| \\ &\text{Let } \sum_k (\mu) P_{jk}(\mu) |g(j, \mu(j), k)| \triangleq \hat{g}_\mu(j) \\ |J_\mu(i)| &\leq \sum_{m=0}^{\infty} \sum_j P_{ij}^m(\mu) \hat{g}_\mu(j) \end{aligned}$$

- When $j = O$, $\hat{g}_\mu(j) = 0$, then we get

$$|J_\mu(i)| \leq \sum_{m=0}^{\infty} \sum_{j=1}^{n} P_{ij}^m(\mu) \boxed{\max_{l=1,...,n} \hat{g}_\mu(l)} \to \leq k$$

$$\sum_{j=1}^{n} P_{ij}^m(\mu) = P(s_m \neq O | s_0 = i, \mu) = P_\mu^{\lfloor \frac{m}{n} \rfloor}$$

$$\implies |J_\mu(i)| \leq \sum_{m=0}^{\infty} P_\mu^{\lfloor \frac{m}{n} \rfloor} k < \infty \text{ since } P_\mu < 1$$

- Let $\bar{g}(i,a) = \sum_{j=0}^{n} P_{ij}(\mu) g(i,a,j)$, denote the expected single stage cost in a non-terminal state i when action a is chosen.
- We define mapping T and T_μ on functions $J = (J(1), J(2), ..., J(n)),\ J : NT \to \mathbb{R}$

$$(TJ)(i) = \min_{u \in A(i)} [\bar{g}(i,u) + \sum_{j=1}^{n} P_{ij}(u) J(j)] \text{ where } i \in \{1, ..., n\}$$

$$(T_\mu J)(i) = \bar{g}_\mu(i) + \sum_{j=1}^{n} P_{ij}(\mu) J(j), \text{ where } \bar{g}_\mu(i) \triangleq \bar{g}(i, \mu(i))$$

Let $P_\mu = \begin{bmatrix} P_{11}(\mu) & P_{12}(\mu) & \cdots & P_{1n}(\mu) \\ P_{21}(\mu) & P_{22}(\mu) & \cdots & P_{2n}(\mu) \\ \vdots & & & \vdots \\ P_{n1}(\mu) & P_{n2}(\mu) & \cdots & P_{nn}(\mu) \end{bmatrix}$. Observe that $\sum_{j=1}^{n} P_{ij}(\mu) \leq 1$. Since this matrix is only on non-terminal states.

- $T_\mu J = \bar{g}_\mu + P_\mu J$ where $\bar{g}_\mu = [\bar{g}(1, \mu(1)), \bar{g}(2, \mu(2)), \ldots, \bar{g}(n, \mu(n))]^T$
- $T^k J = T(T^{k-1} J), k \geq 0$ where $T^0 = I$
- $T^k J = (T \circ T \circ ...k \text{ times}) J$
- Consider $k = 2$

$$\begin{aligned} T^2 J &= T(TJ)(i) \\ &= \min_{u \in A(i)} \left[\bar{g}(i,u) + \sum_{j=1}^{n} P_{ij}(u) TJ(j) \right] \\ &= \min_{u \in A(i)} \left[\bar{g}(i,u) + \sum_{j=1}^{n} \left(\min_{v \in A(j)} [\bar{g}(j,v) + \sum_{r=1}^{n} P_{jk}(v) J(k)] \right) \right] \end{aligned}$$

- Interpretation: $(T^2 J)(i)$ is the optimal cost for a 2-stage problem starting at i, with single stage cost $\bar{g}(\cdot, \cdot)$ and terminal cost $J(\cdot)$
- This is exactly dynamic programming
- $(T^k J)(i) \to$ Optimal cost of a k-stage problem starting at i, with single stage cost $\bar{g}$ and terminal cost J

$$(T^k J)(i) = \min_{u \in A(i)} [\bar{g}_\mu(i,u) + \sum_{j=1}^{n} P_{ij}(u) (T^{k-1} J)(j))]\ \forall i \in \{1, ..., n\}$$

- **Monotonicity Lemma**: For any $J, \bar{J} \in \mathbb{R}^{|S|}$ such that $J(i) \leq \bar{J}(i)\ \forall i = 1, ..., n$ then for any stationary policy μ

$$(T^k J)(i) \leq (T^k \bar{J})(i) \text{ and } (T_\mu^k J)(i) \leq (T_\mu^k \bar{J})(i)\ \forall k \geq 0\ \forall i = 1, ..., n$$

  Consider $k = 1$

$$\begin{aligned} TJ &= \min_{u \in A(i)} [\bar{g}(i,u) + \sum_{j=1}^{n} P_{ij}(u) TJ(j)] \\ &\leq \min_{u \in A(i)} [\bar{g}(i,u) + \sum_{j=1}^{n} P_{ij}(u) T\bar{J}(j)] \\ &\leq T\bar{J} \end{aligned}$$

  Rest follows from induction

- **Lemma 1**: $\forall k \geq 0$, vector $J = [J(1), J(2), ..., J(n)]$, stationary μ, $r > 0$

$$(T^k J + re)(i) \leq (T^{k+1} J)(i) + r, \quad \forall i = 1, ..., n$$

$$(T_\mu^k J + re)(i) \leq (T_\mu^{k+1} J)(i) + r, \quad \forall i = 1, ..., n$$

where $e = [1, 1, ..., 1]^T$
The inequality is reversed if $r < 0$
Consider $k = 1$

$$\begin{aligned}
T(J + re)(i) &= \min_{u \in A(i)} \left[\bar{g}(i, u) + \sum_{j=1}^{n} P_{ij}(u)(J + re(j)) \right] \\
&= \min_{u \in A(i)} \left[\bar{g}(i, u) + \sum_{j=1}^{n} P_{ij}(u)J(j) + r \sum_{j=1}^{n} P_{ij}(u) \right] \\
&\leq \min_{u \in A(i)} \left[\bar{g}(i, u) + \sum_{j=1}^{n} P_{ij}(u)J(j) \right] + r \\
&\leq (TJ)(i) + r
\end{aligned}$$

Again rest follows from induction.

- Assumptions
  1. There exists at least one proper policy.
  2. For every improper μ, $J_\mu(i) = \infty$ for at least one state i.
- **Proposition 1**: For a proper policy μ, the associated cost vector J_μ satisfies

$$\lim_{k \to \infty} (T_\mu^k J)(i) = J_\mu(i), \quad i = 1, ..., n, \quad \text{for any } J \in \mathbb{R}^n$$

  Moreover, $J_\mu = T_\mu J_\mu$ and J_μ is the unique solution.
- **Proposition 2**: A stationary policy μ satisfies for some vector J,

$$J(i) \geq T_\mu J(i), \quad \forall i = 1, ..., n \quad \text{then } \mu \text{ is proper}$$

  $J_\mu = T_\mu J_\mu$ means

$$J_\mu(i) = \bar{g}(i, \mu(i)) + \sum_{j=1}^{n} P_{ij}(\mu(i))J_\mu(j)$$

  This is referred to as **Bellman equation** for a policy μ
- Recall $T_\mu J = \bar{g}_\mu + P_\mu J$

$$\begin{aligned}
T_\mu J &= \bar{g}_\mu + P_\mu J \\
T_\mu^2 J &= \bar{g}_\mu + P_\mu(T_\mu J) = \bar{g}_\mu + P_\mu \bar{g}_\mu + P_\mu^2 J \\
&\vdots \\
T_\mu^k J &= P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m \bar{g}_\mu
\end{aligned}$$

  We have seen that $P(s_k \neq O | s_0 = i, \mu) \leq \rho_\mu^{\lfloor \frac{k}{n} \rfloor}, i = 1, ..., n.$

$$\begin{aligned}
(P_\mu^k J)(i) &= \sum_{j=1}^{n} P(s_k = j | s_0 = i, \mu)J(j) \\
&\leq \sum_{j=1}^{n} P(s_k = j | s_0 = i, \mu) \cdot \max_k J(k) \\
&= P(s_k \neq O | s_0 = i, \mu) \cdot \max_k J(k) \\
&\leq \rho_\mu^{\lfloor \frac{k}{n} \rfloor} \cdot \max_j J(j) \to 0, \text{ as } k \to \infty
\end{aligned}$$

  Thus, $\lim_{k \to \infty} T_\mu^k J = \lim_{k \to \infty} \sum_{m=0}^{k-1} P_\mu^m \bar{g}_\mu = J_\mu$
  By definition,

$$T_\mu^{k+1} J = \bar{g}_\mu + P_\mu T_\mu^k J$$

$$\text{Let } k \to \infty \text{ on either side}$$

$$J_\mu = \bar{g}_\mu + P_\mu J_\mu = T_\mu J_\mu$$

  Suppose $\bar{J}_\mu$ is another $\bar{J}_\mu = T_\mu \bar{J}_\mu$

$$\bar{J}_\mu = T_\mu^2 \bar{J}_\mu = \cdots = \lim_{k \to \infty} T_\mu^k \bar{J}_\mu = J_\mu$$

- By monotonicity of T_μ, $(T_\mu J)(i) \geq (T_\mu^2 J)(i)$
Upon repeating successively

$$J(i) \geq (T_\mu J)(i) > (T_\mu^2 J)(i) > ... > (T_\mu^k J)(i) = (P_\mu^k J)(i) + \sum_{m=0}^{k-1} P_\mu^m \bar{g}_\mu$$

- If μ were not proper, then by assumption, $\exists i \in S$ such that $J_\mu(i) = \infty$. This is a contradiction, since $\lim_{k\to\infty} \left(\sum_{m=0}^{k-1} P_\mu^m \bar{g}_\mu(i)\right) = J_\mu(i)$. But, $J(i) < \infty$, $\forall i \in S$
- **Proposition 3**: The optimal cost vector J^* satisfies $J^* = TJ^*$ (Bellman equation).
Moreover, J^* is the unique solution to this equation.
- **Proposition 4**: We have $\lim_{k\to\infty}(T^k J)(i) = J^*(i)$ $\forall i \in S$ and for every $J \in \mathbb{R}^n$
- **Proposition 5**: A stationary policy μ is optimal iff $T_\mu J^* = TJ^*$.
- We first show that T has at most one fixed point.
Suppose J and J' are two fixed points of T. We select μ and μ' such that

$$J = TJ = T_\mu J \text{ and } J' = TJ' = T_{\mu'} J'$$

Note $(TJ)(i) = \min_{u\in A(i)} \sum_{j\in S} P_{ij}(u)(g(i,u,j) + J(j)) \quad \forall i \in S$
Suppose for $i \to$ min action is $u_i \triangleq \mu(i)$, We can always find a policy μ.
Thus $J = T_\mu J$ and $J' = T_{\mu'} J' \implies \mu$ and μ' are proper policies.
By Proposition 1, $J = J_\mu$ and $J' = J_{\mu'}$
Now, $J = TJ = T^2 J = ... = T^k J \leq T_{\mu'}^k J$
True since T^k involves minimization over all policies while $T_{\mu'}$ involves evaluation over a given policy.
It then follows

$$J \leq \lim_{k\to\infty} T_{\mu'}^k J = J_{\mu'} = J'$$

Similarly we get $J' \leq J$, these two together $\implies J = J'$
$\implies T$ has at most one fixed point.
- We now show that T has at least one fixed point
Let μ be a proper policy and let μ' be another policy such that $T_{\mu'} J_\mu = TJ_\mu$
Then,

$$\begin{aligned} J_\mu = T_\mu J_\mu \geq TJ_\mu T_{\mu'} J_\mu \\ \implies J_\mu \geq T_{\mu'} J_\mu \implies \mu' \text{ is proper} \end{aligned}$$

$J_\mu \geq T_{\mu'} J_\mu \geq T_{\mu'}^2 J_\mu \geq \ldots \geq \lim_{k\to\infty} T_{\mu'}^k J_\mu = J_{\mu'} \implies J_\mu \geq J_{\mu'}$
Continuing in this manner, we obtain a sequence $\{\mu^k\}$ such that each μ^k is proper and
$J_{\mu^k} = T_{\mu^k} J_{\mu^k} \geq TJ_{\mu^k} = T_{\mu^{k+1} J_{\mu^k}} \geq \ldots \geq \lim_{m\to\infty} T_{\mu^{k+1}}^m J_{\mu^k} = J_{\mu^{k+1}}$
Thus, $J_{\mu^k} \geq T_{\mu^{k+1}} J_{\mu^k} \geq J_{\mu^{k+1}}$ $\forall k$ where $T_{\mu^{k+1}} J_{\mu^k} = TJ_{\mu^k}$
However, one cannot continue to improve J_{μ^k} forever and so there will exist a policy μ such that $J^\mu \geq TJ_\mu \geq J_\mu \implies J_\mu = TJ\mu$
Since the number of stationary policies is finite.
Thus, J_μ is a fixed point of T and since there are only at most one fixed point, J_μ has to be unique.
- Next we show that $J_\mu = J^*$ and $T^k J \to J^*$ as $k \to \infty$.
Let $e = (1, ..., 1)^T$ and $\delta > 0$ is a scalar. Let $\hat{J}$ be the vector that satisfies $T_\mu \hat{J} = \hat{J} - \delta e \implies \hat{J} = T_\mu \hat{J} + \delta e = (g_\mu + \delta e) + P_\mu \hat{J}$
$\implies g_\mu + \delta e$ is the new single stage cost.
$\implies \hat{J}$ is the cost vector corresponding to policy μ with g_μ replaced with $g_\mu + \delta e$.
Moreover, $J_\mu \leq \hat{J}$ since single stage costs have gone up.
$\implies J_\mu = TJ_\mu < T\hat{J} \leq T_\mu \hat{J} = \hat{J} - \delta e \leq \hat{J}$
$\implies J_\mu = T^k J_\mu \leq T^k \hat{J} \leq T^{k-1} \hat{J} \leq \hat{J}$
Thus $T^k \hat{J}$, $k \geq 1$, is a bounded monotone sequence and $T^k \hat{J} \to \tilde{J}$ as $k \to \infty$, for some $\tilde{J}$ such that

$$T\tilde{J} = T(\lim_{k\to\infty} T^k \hat{J}) = (\lim_{k\to\infty} T^{k+1} \hat{J}) = \tilde{J} \implies \tilde{J} = J_\mu$$

as J_μ is the unique fixed point of T. $J_\mu - \delta e = TJ_\mu - \delta e \leq T(J_\mu - \delta e) \leq TJ_\mu = J_\mu$
$\implies T(J_\mu - \delta e) \leq T^2(J_\mu - \delta e)$
Thus, $T^k(J_\mu - \delta e)$ is monotonically increasing and bounded above by J_μ, and $\lim_{k\to\infty} T^k(J_\mu - \delta e) = J_\mu$,

as J_μ is the unique fixed point of T.
For any $J \in \mathbb{R}^n$, we can find $\delta > 0$ such that $J_\mu - \delta e \leq J \leq \hat{J}$.
By monotonicity of T, $T^k(J_\mu - \delta e) \leq T^k J \leq T^k \hat{J}$

$$\implies J_\mu = \lim_{k\to\infty} T^k(J_\mu - \delta e) \leq \lim_{k\to\infty} T^k J \leq \lim_{k\to\infty} T^k \hat{J} = J_\mu$$

To show that $J_\mu = J^*$, take any policy $\pi = \{\mu_0, \mu_1, \mu_2, ...\}$. Then, $T_{\mu_0} T_{\mu_1} \dots T_{\mu_{k-1}} J_0 \geq T^k J_0$ where J_0 is any arbitrary vector.
Taking lim sup as $k \to \infty$ on both sides, we get $J_\pi \geq J_\mu$.
Since, π is arbitrary, μ is optimal and $J_\mu = J^*$
If μ is optimal, then $J_\mu = J^*$

$$T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^* \implies T_\mu J^* = TJ^*$$

Conversely, suppose $T_\mu J^* = TJ^*$. Then, since μ is proper, we have that $J_\mu = J^*$, since $J^* = T_\mu J^* = TJ^*$ and there exists a unique solution J_μ of the above equation.
We shall show that T and T_μ are contraction maps in a certain norm $\|\cdot\|_\zeta$ $\forall J, \bar{J} \in \mathbb{R}^{|S|}$.
Recall that $S = \{1, 2, \dots, n\}$ is the set of nonterminal states O is the terminal state, $S^+ = S \cup \{0\}$: set of all states

- **Brouwer's fixed point Theorem**: Suppose s is a complete seperable metric shape w.r.t a certain metric ρ. Suppose T is a contraction. Moreover, x^* is unique

$$\rho(Tx, T\bar{x}) \leq \beta \rho(x, \bar{x}) \quad \forall x, \bar{x}$$

Consider an interval $(0, 1]$ and consider sequence $x_n = \frac{1}{n}, \rho = 1.1$

$$x_n \to 0$$

We will show that there is a vector $\xi = (\xi(1), \dots, \xi(n))$ such that $\xi(i) > 0 \quad \forall i$, and a scalar $0 \leq \beta < 1$ such that

$$\|TJ - TJ\|_\xi \leq \beta \|J - \bar{J}\|_\xi \forall J, \bar{J} \in \mathbb{R}^n$$

where $\|J\|_\xi \triangleq \max_{i=1,\dots,n} \frac{|J(i)|}{\xi(i)}$

- **Proposition 6**: Assume all stationary policies are proper. Then, there exists a vector $\xi = (\xi(1), \dots, \xi(n))$ with $\xi(i) > 0 \quad \forall i = 1, \dots, n$ s.t. the mappings T and $T_\mu \forall$ stationary μ are contractions w.r.t $\|\cdot\|_\xi$ In particular, $\exists \beta \in (0, 1)$ such that

$$\sum_{j=1}^{n} P_{ij}(u)\xi(j) \leq \beta \xi(i) \quad \forall i, u \in A(i)$$

Consider a new SSPP where transition probabilities are same as before but transition costs are all equal to -1, except the transition state where $g(O, u, O) = 0 \quad \forall u \in A(O)$
Let $\hat{J}(i)$ = optimal cost to go from state i in the new problem

$$\begin{aligned}\hat{J}(i) &= -1 + \min_{u \in A(i))} \sum_{j \in S} P_{ij}(u)\hat{J}(j) \quad \forall i \in S \\ &\leq -1 + \sum_{j \in S} P_{ij}(\mu(i))\hat{J}(g) \forall i \in S, \text{ for any given } \mu\end{aligned}$$

Let $\xi(i) = -\hat{J}(i)$. Then, $\xi(i) \geq 1 \quad \forall i$

$$\begin{aligned}&-\hat{J}(i) \geq 1 + \sum_{j \in S} P_{ij}(\mu(i))(-\hat{J}(j)) \quad \forall i \in S \\ &\xi(i) \geq 1 + \sum_{j \in S} P_{ij}(\mu(i))\xi(j) \quad \forall i \in S \\ &\sum_{j \in S} P_{ij}(\mu(i))\xi(j) \leq \xi(i) - 1 \leq \beta \xi(i) \\ &\text{where } \beta = \max_{i=1,\dots,n} \frac{\xi(i) - 1}{\xi(i)} < 1\end{aligned}$$

Now for stationary policy μ, state i and vector $J, \bar{J} \in \mathbb{R}^n$

$$
\begin{aligned}
|(T_\mu J)(i) - (TJ)(i))| &= |\sum_{j=1}^{n} P_{ij}(\mu(i))(J(j) - \bar{J}(j))| \\
&\le (\sum_{j=1}^{n} P_{ij}(\mu(i))\xi(j))(\max_{j=1,\dots,n} \frac{|J(j) - \bar{J}(j)|}{\xi(j)}) \\
&\le \beta\xi(i)\|J - \bar{J}\|_\xi
\end{aligned}
$$

Recall, $\|J - \bar{J}\|_\xi = \max_{i=1,\dots,n}(\frac{J(i)-\bar{J}(i)}{\xi(i)})$

$$
\max_{i=1,\dots,n} \frac{|(T_\mu J)(i) - (TJ)(i)|}{\xi(i)} \le \beta\|J - \bar{J}\|_\xi
$$

$$
\implies \|T_\mu J - TJ\|_\xi \le \beta\|J - \bar{J}\|_\xi
$$

- Recall $|(T_\mu J)(i) - (TJ)(i)| \le \beta\xi(i)\|J - \bar{J}\|_\xi$

$$
\implies (T_\mu J)(i) \le (T_\mu \bar{J})(i) + \beta\xi(i)\|J - \bar{J}\|_\xi
$$

Taking minimum over μ on either side,

$$
(TJ)(i) \le (T\bar{J})(i) + \beta\xi(i)\|J - \bar{J}\|_\xi
$$

We also get $(T\bar{J})(i) \le (TJ)(i) + \beta\xi(i)\|J - \bar{J}\|_\xi$.
Combining the two inequalities, we obtain

$$
\|TJ - T\bar{J}\|_\xi \le \beta\|J - \tilde{J}\|_\xi \quad \forall J, \bar{J} \in \mathbb{R}^n
$$

3.3 Numerical Schemes for Solving MDPs

- **Value Iteration**
  Consider the optimal control problem

  1. Choose some arbitrary $J \in \mathbb{R}^n$
  2. Recursively iterate $J \leftarrow T^k J, k = 1, 2, \dots ..$

  We know that $T^k J \to J^*$ as $k \to \infty$.
  Suppose $V_0, V_1, V_2, \dots$ be the sequence of functions obtained when T is applied

$$
V_{m+1}(i) = \min_{u \in A(i)} \sum_{j \in S} P_{ij}(u)(g(i, u, j) + V_m(j)), \quad \forall i \in S,\ \forall m \ge 0
$$

  Change to max in case of reward problem.
  Start with some $V_0 \in \mathbb{R}^n$,
  We know that $V_m \to V^*$ as $m \to \infty$, where $V^* = TV^*$
  Reference: Grid world example from Sutton & Barto Ch 4
  The book uses the following update rule, Expectation, instead of minimization.

$$
V_{m+1}(i) = \sum_{u \in A(i)} \pi(u|i) \sum_{j \in S} P_{ij}(u)(g(i, u, j) + V_m(j)), \forall i \in S, \forall m \ge 0
$$

  Start from J_0 and iterate to obtain the sequence of functions $J_0, TJ_0, T^2J_0, \dots$ and $\lim_{n\to\infty} T^n J_0 = J^*$ (optimal value function)

- **Gauss Seidel Value Iteration**
  Define an operator $F : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$

$$
\begin{aligned}
(FJ)(1) &= \min_{u \in A(i)} \sum_{j=1}^{n} P_{1j}(u)(g(1, u, j) + J(j)) \\
(FJ)(i) &= \min_{u \in A(i)} \sum_{j=1}^{n} P_{ij}(u)(g(i, u, j) + J(j)) + \sum_{j=1}^{i-1} P_{ij}(u)(FJ)(j) \quad \text{for } i = 2, \dots, n
\end{aligned}
$$

  Note: $(FJ)(i) = (TJ)(i)$
  Result: $\lim_{k\to\infty} F^k J = J^* \quad \forall J \in \mathbb{R}^n$

- **Policy Iteration**
  Start with an initial policy μ_0
  Policy evaluation: Given policy μ_k, compute $J^{\mu_k}(i), \forall i \in S$ as the solution to

$$J(i) = \sum_{j=1}^{n} P_{ij}(\mu(i))(g(i, \mu(i), j) + J(j)), i = 1, \ldots, n$$

  unknowns: $J(1), J(2), \ldots, J(n)$
  Policy Improvement: Find new policy μ_{k+1} such that

$$\mu_{k+1}(i) = \arg\min_{u \in A(i)} \sum_{j=1}^{n} P_{ij}(u)(g(i, u, j) + J^{\mu_k}(j)), \forall i = 1, \ldots, n$$

  Or solve, $T_{\mu_{k+1}} J^{\mu_k} = T J^{\mu_k}$

$$\implies \sum_{j=1}^{n} P_{ij}(\mu_{k+1}(i))(g(i, \mu_{k+1}(i), j) + J^{\mu_k}(j)) = \min_{u \in A(i)} \sum_{j=1}^{n} P_{ij}(u)(g(i, u, j) + J^{\mu_k}(j)) \text{ for all } i = 1, ..., n$$

  Structure is like a nested loop

Algorithm 1 Policy Iteration

```
Policy Improvement (outer loop) {
   μ_{k+1}(i) = arg min_u Σ_{j=1}^n P_ij(u)(g(i,u,j) + J^{μ_k}(j)) ∀i = 1,...,n
   Policy Evaluation (inner loop) {
      J_{l+1}(i) = Σ_{j=1}^n P_ij(μ_{k+1}(i))(g(i,μ_{k+1}(i),j) + J_l(j)) ∀i = 1,...,n
      Starting from given J_l(·), l = 0,1,...
      Until J_l → J^{μ_{k+1}}
   }
}
```

  Repeat process if $J^{\mu_{k+1}}(i) \subset J^{\mu_k}(i)$ for at last one $i \in S$
  If $J^{\mu_{k+1}}(i) = J^{\mu_k}(i) \quad \forall i = 1, \ldots, n$ then stop and output μ_k as optimal policy
  Proposition 7: The policy iteration algorithm generates an improving sequence of proper policies, i.e., $J^{\mu_{k+1}}(i) \leq J^{\mu_k}(i) \quad \forall i$ and k, and terminates with an optimal policy in a finite number of iterations.
  Given a proper policy μ, the new $\bar{\mu}$ is obtained via policy improvement as $T_{\bar{\mu}} J^{\mu} = T J^{\mu}$
  Then, $J^{\mu} = T_{\mu} J^{\mu} \geq T J^{\mu} = T_{\bar{\mu}} J^{\mu}$

$$\implies J^{\mu} \geq T_{\bar{\mu}} J^{\mu}$$

  By monotonicity of $T_{\bar{\mu}}$,

$$J^{\mu} \geq T_{\bar{\mu}} J^{\mu} \geq T_{\bar{\mu}}^2 J^{\mu} \geq T_{\bar{\mu}}^3 J^{\mu} \geq \ldots \geq \lim_{k \to \infty} T_{\bar{\mu}}^k J^{\mu} = J^{\bar{\mu}}$$

$$\implies J^{\mu} \geq J^{\bar{\mu}}$$

  Suppose $\bar{\mu}$ is improper $\implies J^{\bar{\mu}}(i) = \infty$ for some $i \in S$

$$\implies J^{\mu}(i) = \infty \text{ for that } i \in S$$

  This is a contradiction since μ is proper
  $\implies \bar{\mu}$ is also a proper policy
  If, μ is not optimal, then $J^{\bar{\mu}}(i) < J^{\mu}(i)$ for some $i \in S$
  otherwise, $J^{\mu} = J^{\bar{\mu}} = T_{\bar{\mu}} J^{\bar{\mu}} = T_{\bar{\mu}} J^{\mu} = T J^{\mu}$

$$\implies J^{\mu} = J^*, \text{ since } J^{\mu} = T J^{\mu}$$

  $\implies \mu$ is optimal. Thus, new policy is strictly better than current policy if current policy is not optimal. Since the number of proper policies is finite, this procedure converges in a finite number of steps to an optimal policy.

- **Modified Policy Iteration**
  Select sequence of positive integers $m_0, m_1, m_2, \ldots$ and suppose $J_1, J_2, \ldots$ and stationary policies $\mu_0, \mu_1, \mu_2, \ldots$ are obtained as $T_{\mu_k} J_k = T J_k$ and $J_{k+1} = T_{\mu_k}^{m_k} J_k, k = 0, 1, \ldots, m$

- We can show that this procedure terminates in an optimal policy μ^* and optimal value function J^*. Consider,
  $m_k = 1 \forall k:$ Value iteration since $J_{k+1} = T_{\mu_k} J_k = T J_k$
  $m_k = \infty \forall k:$ Policy iteration.

- **Multi-stage lookahead iteration**
  Regular PI uses a one-step look ahead and finds optimal decision for one-stage problem with one stage cost $g(i,u,j)$ and terminal cost $J^\mu(j)$ when policy is μ.
  In m-stage lookahead problem, we find optimal policy for an m-stage DP where we start in state $i \in S$, make m subsequent decisions incurring corresponding costs of m stages and getting a terminal cost $J^\mu(j)$ where j is state after m stages.
  Claim: m-stage PI terminates with optimal policy under same conditions as PI
  Let $\{\bar{\mu}_0, \bar{\mu}_1, \ldots, \bar{\mu}_{m-1}\}$ be an optimal policy for m-stage DP with terminal cost J^μ.
  Thus, $T_{\bar{\mu}_k} T^{m-k-1} J^\mu = T^{m-k} J^\mu, k = 0, 1, \ldots, m-1$
  Suppose,

$$\begin{aligned}
&k = m-1: \ T_{\bar{\mu}_{m-1}} J^\mu = T J^\mu \\
&k = m-2: \ T_{\bar{\mu}_{m-2}} J^\mu = T J^\mu \\
&\vdots \\
&k = 0 \qquad : \ T_{\bar{\mu}_0} J^\mu = T J^\mu
\end{aligned}$$

  Now $TJ^\mu \leq T_\mu J^\mu = J^\mu$

$$\implies T^{k+1} J^\mu \leq T^k J^\mu \leq J^\mu \quad \forall k = 0, 1, \ldots$$

  Thus, $T_{\bar{\mu}_k} T^{m-k-1} J^\mu = T^{m-k} J^\mu \leq T^{m-k-1} J^\mu \quad \forall k = 0, 1, \ldots, m-1$
  Thus, $T^l_{\bar{\mu}_k} T^{m-k-1} J^\mu \leq T_{\bar{\mu}_k} T^{m-k-1} J^\mu = T^{m-k} J^\mu \quad \forall l \geq 1$
  Taking limits as $l \to \infty$, we obtain

$$J^{\bar{u}_k} = \lim_{l \to \infty} T^l_{\bar{\mu}_k} T^{m-k-1} J^\mu \leq T^{m-k} J^\mu \leq J^\mu \quad \forall k = 0, 1, \ldots, m-1$$

  Thus, for a successor policy $\bar{\mu}$ generated by m-stage PI, i.e. $\bar{\mu} = \bar{\mu}_0$, we have

$$J^{\bar{\mu}} \leq T^m J^\mu \leq J^\mu \text{ [set } k = 0 \text{ in previous equation]}$$

  $\implies \bar{\mu}$ is an improved policy relative to μ
  If $J^{\bar{\mu}} = J^\mu$, then $J^\mu = TJ^\mu$ and $J^\mu = J^*$
  $\implies$ This algorithm also terminates in an optimal policy.

3.4 Infinite Horizon Discounted Cost

- Setting involves no termination cost, $S = \{1, 2, \ldots, n\}$
- $A(i) \triangleq$ Set of feasible actions in state i
- $A = \bigcup_{i \in S} A(i) \triangleq$ Set of all actions
- $|S|, |A| < \infty$

$$J^*(i) = \min_\mu \left[\sum_{k=0}^{\infty} \alpha^k g(i_k, \mu(i_k), i_{k+1}) \mid i_0 = i \right]$$

  $0 < \alpha < 1$ is the discount factor

- $J^*(i) \equiv$ value of state i or cost-to-go from state i
- Let $J = (J(1), J(2), \ldots, J(n))$
- Define operators T and T_μ as

$$\begin{aligned}
(TJ)(i) &= \min_{u \in A(i)} \sum_{j=1}^{n} P_{ij}(u)(g(i,u,j) + \alpha J(j)), i \in S \\
(T_\mu J)(i) &= \sum_{j=1}^{n} P_{ij}(\mu(i))(g(i,\mu(i),j) + \alpha J(j)), \quad i \in S
\end{aligned}$$

- Let $P_\mu = \begin{bmatrix} P_{11}(\mu(1)) & P_{12}(\mu(1)) & \dots & P_{1n}(\mu(1)) \\ P_{21}(\mu(2)) & P_{22}(\mu(2)) & \dots & P_{2n}(\mu(2)) \\ \vdots & & & \vdots \\ P_{n1}(\mu(n)) & P_{n2}(\mu(n)) & \dots & P_{nn}(\mu(n)) \end{bmatrix}_{n\times n}$
- P_μ is a stochastic matrix because $\sum_{j\in S} P_{ij}(\mu(i)) = 1 \quad \forall i \in S$
- Let $g_\mu = \begin{bmatrix} \sum_{j=1}^{n} P_{1j}(\mu(1))g(1,\mu(1),j) \\ \sum_{j=1}^{n} P_{ij}(\mu(2))g(2,\mu(2),j) \\ \vdots \\ \sum_{j=1}^{n} P_{nj}(\mu(n))g(n,\mu(n),j) \end{bmatrix}$
- Bellman equation under a given policy μ

$$T_\mu J = g_\mu + \alpha P_\mu J = J$$

- **Monotonicity Lemma**
  For any vectors $J, \bar{J} \in \mathbb{R}^n$, such that $J(i) \le \bar{J}(i) \quad \forall i \in S$ and for any stationary policy μ
  Let $e = (1,1,\dots,1)_n$, then for any vector $J = (J(1),\dots J(n))$ and $r \in \mathbb{R}$

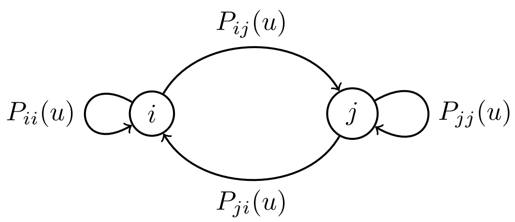
$$\begin{aligned} (T(J+re))(i) &= \min_{u\in A(i)} \sum_{j=1}^{n} P_{ij}(u)(g(i,u,j) + \alpha(J+re)(j)) \\ &= \min_{u\in A(i)} \sum_{j=1}^{n} P_{ij}(u)(g(i,u,j) + \alpha J) + \alpha r \\ &= (TJ)(i) + \alpha r \\ T(J+re) &= TJ + \alpha re \end{aligned}$$

- **Lemma 3**: For every k, J, μ & r

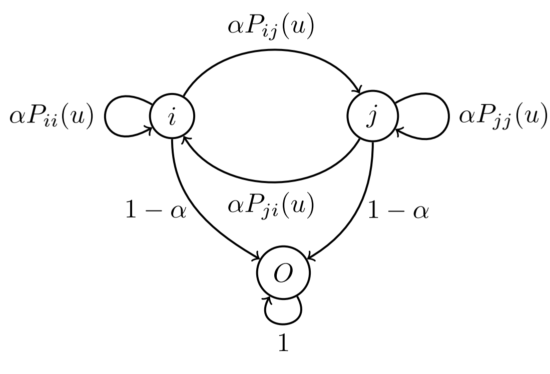
$$\begin{aligned} (T^k(J+re))(i) &= (T^kJ)(i) + \alpha^k r, \quad i = 1,\dots,n, k \ge 1 \\ (T_\mu^k(J+re))(i) &= (T_\mu^kJ)(i) + \alpha^k r, \quad i = 1,\dots,n, k \ge 1 \end{aligned}$$

  Proof from induction

- We can convert a discounted cost problem to a stochastic shortest path problem by adding a termination state



(a) DCP



(b) SSPP

Figure 1: Convert DCP to SSPP

Probability of termination in 1st stage $= 1-\alpha$
Probability of termination in 2nd stage $= \alpha(1-\alpha)$
$\vdots$
Probability of termination in k^{th} stage $= \alpha^{k-1}(1-\alpha)$
Probability of non-termination in k^{th} stage:

$$\begin{aligned} &= 1 - \left\{(1-\alpha)(1+\alpha+\alpha^2+\dots+\alpha^{k-1})\right\} \\ &= 1 - \left\{(1-\alpha)\frac{1-\alpha^k}{1-\alpha}\right\} = \alpha^k \end{aligned}$$

Expected single stage cost in k^{th} stage $= \alpha^k \sum_{j=1}^{n} P_{ij}(u) g(i,u,j)$
All policies are proper for the associated SSPP since from every state under every policy, there is a probability of $(1-\alpha)$ of termination.
Under policy μ
SSPP: $J_\mu(i) = E\left[\sum_{k=0}^{\infty} g(i_k, \mu(i_k), i_{k+1}) | i_0 = i\right]$
DCP: $J_\mu(i) = E\left[\sum_{k=0}^{\infty} \alpha^k g(i_k, \mu(i_k), i_{k+1}) \mid i_0 = i\right]$.
Consider now a DCP where $|g(i,u,j)| \leq n, \quad \forall i, j \in S, u \in A(i)$

- **DP Convergence**
For any bounded $J : S \to \mathbb{R}$, the optimal cost function satisfies

$$J^*(i) = \lim_{N \to \infty} (T^N J)(i) \quad \forall i \in S$$

Consider a policy $\pi = \{\mu_0, \mu_1, \mu_2, \ldots, \}$ with $\mu_k : S \to A$ such that $\mu_k(i) \in A(i) \quad \forall i \in S, k \geq 0$
Then,

$$\begin{aligned} J_\pi(i) &= \lim_{N\to\infty} E\left[\sum_{k=0}^{N-1} \alpha^k g(i_k + \mu_k(i_k), i_{k+1}) | i_0 = i\right] \\ &= E\left[\sum_{k=0}^{K-1} \alpha^k g(i_k, \mu_k(i_k), i_{k+1}) | i_0 = i\right] + \lim_{N\to\infty} E\left[\sum_{k=K}^{N-1} \alpha^k g(i_k, \mu_k(i_k), i_{k+1}) | i_0 = i\right] \end{aligned}$$

Since $|g(i_k, \mu_k(i_k), i_{k+1})| \leq n, \quad \forall i_{k,}, i_{k+1} \in S,\ u_k(i_k) \in A(i_k)$

$$\lim_{N\to\infty} E\left[\sum_{k=k}^{N} \alpha^k g(i_k, \mu_k(i_k), i_{k+1}) | i_0 = i\right] \leq n \sum_{k=K}^{\infty} \alpha^k = \frac{\alpha^k n}{1-\alpha}$$

Thus, $E\left[\sum_{k=0}^{K-1} \alpha^k g(i_k, \mu_k(i_k), i_{k+1}) \mid i_0 = i\right] = J_\pi(i) - \lim_{N\to\infty} E\left[\sum_{k=K}^{N-1} \alpha^k g(i_k, \mu_k(i_k), i_{k+1}) | i_0 = i\right]$

$$\begin{aligned} \implies J_\pi(i) - \frac{\alpha^K n}{1-\alpha} - \alpha^K \max_{j\in S} |J(j)| &\leq E\left[\sum_{k=0}^{K-1} g(i_k, \mu_k(i_k), i_{k+1}) + \alpha^k J(i_k) | i_0 = i\right] \\ &\leq J_\pi(i) + \frac{\alpha^K n}{1-\alpha} + \alpha^K \max_{j\in S} |J(j)| \end{aligned}$$

Taking min over π on all sides, we have for all $i \in S$ and $k > 0$,

$$J^*(i) - \frac{\alpha^k n}{1-\alpha} - \alpha^k \max_{j\in S} |J(j)| \leq (T^k J)(i) \leq J^*(i) + \frac{\alpha^k n}{1-\alpha} + \alpha^k \max_{j\in S} |J(j)|$$

Setting $k \to \infty$ over all sides,

$$\lim_{k\to\infty} (T^k J)(i) = J^*(i)$$

Corollary: DP convergence for a given policy π
For every stationary policy μ, the associated cost function satisfies

$$J_\mu(i) = \lim_{N\to\infty} (T_\mu^N J)(i) \quad \forall i \in S \text{ and any } J \in \mathbb{R}^{|S|}$$

Proof: Consider an alternative MDP where

$$A(i) = \{\mu(i)\} \quad \forall i \in S$$

Proposition 9: Bellman Equation
The optimal cost function J^* satisfies

$$J^*(i) = \min_{u\in A(i)} \sum_{j\in S} P_{ij}(u)(g(i,u,j) + \alpha J^*(j)) \quad \forall i \in S$$

or $J^* = TJ^*$
Moreover, J^* is the unique solution of this equation within the class of bounded functions.

$$\text{Recall, } J^*(i) - \frac{\alpha^k n}{1-\alpha} - \alpha^k \max_{j\in S} |J(j)| \leq (T^k J)(i) \leq J^*(i) + \frac{\alpha^k n}{1-\alpha} + \alpha^k \max_{j\in S} |J(j)|$$

Applying T on all sides

$$(TJ)^*(i) - \frac{\alpha^{k+1}n}{1-\alpha} - \alpha^{k+1}\max_{j\in S}|J(j)| \leq (T^{k+1}J)(i) \leq (TJ)^*(i) + \frac{\alpha^{k+1}n}{1-\alpha} + \alpha^{k+1}\max_{j\in S}|J(j)|$$

Let $k \to \infty$ on all sides

$$(TJ^*(i)) \leq J^*(i) \leq (TJ^*(i)) \implies TJ^* = J^*$$

For uniqueness, suppose $\bar{J} \in \mathbb{R}^n$ is another solution

$$\bar{J} = T\bar{J} = T^2\bar{J} = \cdots = \lim_{k\to\infty} T^k\bar{J} = J^* \text{ or } \bar{J} = J^*$$

- **Corollary: Bellman Equation for a given policy**
  For every stationary policy μ the associated cost function satisfies

  $$J_\mu(i) = \sum_{j\in S} P_{ij}(\mu(i))(g(i,\mu(i),j) + \alpha J_\mu(j)) \quad \forall i \in S$$

  Moreover, J is the unique solution to this equation within the class of bounded function.

- **Necessary and Sufficient conditions for optimality**
  Proposition 10: A stationary policy μ is optimal iff $\mu(i)$ attains the minimum in the Bellman equation, $\forall i \in S$,

  $$TJ^* = T_\mu J^*$$

  Suppose, $TJ^* = T_\mu J^*$
  then, $J^* = TJ^* = T_\mu J^* \implies J^* = T_\mu J^*$ or $J^* = J^\mu$
  Conversely, suppose μ is optimal. Then, $J^* = J^\mu$
  then, $J^* = T_\mu J^* = TJ^*$
  Define now max norm $\|\cdot\|_\infty$ on $\mathbb{R}^n$ by

  $$\|J\|_\infty = \max_{i\in s}|J(i)|$$

  Proposition 11: For any two bounded functions $J : S \to \mathbb{R}$ and $J' : S \to \mathbb{R}$ and for all $k = 0, 1, 2, \ldots$
  a) $\left\|T^kJ - T^kJ'\right\|_\infty \leq \alpha^k \|J - J'\|_\infty$
  b) $\left\|T^k_\mu J - T^k_u J'\right\|_\infty \leq \alpha^k \|J - J'\|_\infty$
  Let $c = \|J - J'\|_\infty = \max_{i\in S}|J(i) - J'(i)|$

  $$\begin{aligned}
  &\implies J(i) - c \leq J'(i) \leq J(i) + c, \forall i \in S \\
  &\implies (T^kJ)(i) - \alpha^k c \leq (T^kJ')(i) \leq (T^kJ)(i) + \alpha^k c \\
  &\implies \left|(T^kJ)(i) - (T^kJ')(i)\right| \leq \alpha^k \|J - J'\|_\infty \\
  &\implies \max_{i\in S} |\ (T^kJ)(i) - (T^kJ')(i)| \leq \alpha^k \|J - J'\|_\infty \\
  &\implies \left\|T^kJ - T^kJ\right\| \leq \alpha^k \|J - J'\|_\infty
  \end{aligned}$$

  b) follows similarly

4 Convergence Guarantees

4.1 Value Iteration

- **Corollary: Rate of convergence of Value Iteration**
  For any bounded function $J : S \to \mathbb{R}$, we have

  i) $\max_{i\in S}\left|(T^kJ)(i) - J^*(i)\right| \leq \alpha^k \max_{i\in S}|J(i) - J^*(i)|$
  ii) $\max_{i\in S}\left|(T^k_\mu J)(i) - J_\mu(i)\right| \leq \alpha^k \max_{i\in S}|J(i) - J_\mu(i)| \quad \forall i \in S, k = 0, 1, 2$

  Recall that

  $$\begin{aligned}
  (T_\mu J)(i) &= \sum_{j=1}^n P_{ij}(\mu(i))(g(i,\mu(i),j) + \alpha J(j)), i \in S \\
  &= \sum_{j=1}^n P_{ij}(\mu(i))g(i,\mu(i),j) + \alpha\sum_{j=1}^n P_{ij}(\mu(i))J(j), i \in S \\
  &= \bar{g}(i,\mu(i)) + \alpha\sum_{i=1}^n P_{ij}(\mu(i))J(j), i \in S
  \end{aligned}$$

Let $\bar{g}\mu = \begin{bmatrix} g_\mu(1,\mu(1)) \\ \vdots \\ \bar{g}(n,\mu(n)) \end{bmatrix}$, $P_\mu = \begin{bmatrix} P_{11}(\mu(1)) & \dots & P_{1n}(\mu(1)) \\ \vdots & \ddots & \vdots \\ P_{n1}(\mu(n)) & & P_{nn}(\mu(n)) \end{bmatrix}$
Then, $T_\mu J = \bar{g}_\mu + \alpha P_\mu J$
J^μ is the unique fixed point of this equation
Thus, $J^\mu = \bar{g}_\mu + \alpha P_\mu J^\mu$
$\implies (I - \alpha P_\mu)J^\mu = \bar{g}_\mu$
$\implies J^\mu = (I - \alpha P_\mu)^{-1}\bar{g}_\mu$
Expensive to invert matrices

- **Error Bounds**
We have shown that starting from any $J \in \mathbb{R}^n$,

$$\lim_{k\to\infty}(T^k J)(i) = J^*(i); \quad i \in S$$

Also, $\mid (T^k J)(i) - J^*(i) \mid \leq \alpha^k \mid J(i) - J^*(i) \mid \quad \forall i \in S$
Recall,

$$\begin{aligned} J^\mu(i) &= E\left[\sum_{k=0}^{\infty}\alpha^k g(i_k,\mu(i_k),i_{k+1})|i_0 = i\right] \\ &= \bar{g}(i,\mu(i)) + \sum_{k=0}^{\infty}\alpha^k E\left[|g(i_k,\mu(i_k),i_{k+1})|i_0 = i\right] \end{aligned}$$

Letting $\underline{\beta} = \min_i \bar{g}(i,\mu(i)) \quad \bar{\beta} = \max_i \bar{g}(i,\mu(i))$

$$\implies \bar{g}_\mu + \left(\frac{\alpha\underline{\beta}}{1-\alpha}\right)e \leq J_\mu \leq \bar{g}_\mu + \left(\frac{\alpha\bar{\beta}}{1-\alpha}\right)e$$

Since, $\underline{\beta} \leq \bar{g}_\mu \leq \bar{\beta}$, we have

$$\frac{\underline{\beta}}{1-\alpha}e \leq \bar{g}_\mu + (\frac{\alpha\underline{\beta}}{1-\alpha})e \leq J_\mu \leq \bar{g}_\mu + (\frac{\alpha\bar{\beta}}{1-\alpha})e \leq \frac{\bar{\beta}}{1-\alpha}e$$

Given a vector J, we know that $T_\mu J = \bar{g}_\mu + \alpha P_\mu J$
Subtracting above from $J^\mu = \bar{g}_\mu + \alpha P_\mu J^\mu$, we get

$$\begin{aligned} J^\mu - T_\mu J &= \alpha P_\mu(J^\mu - J) \\ \implies (J^\mu - J) &= (T_\mu J - J) + \alpha P_\mu(J^\mu - J) \end{aligned}$$

Thus, if cost-per-stage vector is $T_\mu J - J$, then $J^\mu - J$ is the cost-to-go vector. Then,

$$\frac{\underline{r}}{1-\alpha}e \leq T_\mu J - J + \frac{\alpha\underline{r}}{1-\alpha}e \leq J^\mu - J \leq T_\mu J - J + \frac{\alpha\bar{r}}{1-\alpha}e \leq \frac{\bar{r}}{1-\alpha}e$$
where $\underline{r} = \min_i[(T_\mu J)(i) - J(i)]$ and $\bar{r} = \max_i[(T_\mu J)(i) - J(i)]$

Adding J on all sides, we get

$$J + \frac{\alpha\underline{r}}{1-\alpha}e \leq T_\mu J + \frac{\alpha\underline{r}}{1-\alpha}e \leq J^\mu \leq T_\mu J + \frac{\alpha\bar{r}}{1-\alpha} \leq J + \frac{\bar{r}}{1-\alpha}e$$
$$J + \frac{\underline{c}}{\alpha}e \leq T_\mu J + \underline{c}e \leq J^\mu \leq T_\mu J + \bar{c}e \leq J + \frac{\bar{c}}{\alpha}e$$
where $\underline{c} = \dfrac{\alpha\underline{r}}{1-\alpha}$ and $\bar{c} = \dfrac{\alpha\bar{r}}{1-\alpha}$

- **Proposition 12**: For every function $J : S \to \mathbb{R}$, state i and $k \geq 0$,

$$\begin{aligned} &(T^k J)(i) + \underline{c}_k \leq (T^{k+1}J)(i) + \underline{c}_{k+1} \leq J^*(i) \leq (T^{k+1})(i) + \bar{c}_{k+1} \leq (T^k J)(i) + \bar{c}_k \\ &\underline{c}_k = \frac{\alpha}{1-\alpha}\min_{i=1,\dots,n}\left[(T^k J)(i) - (T^{k-1}J)(i)\right] \\ &\bar{c}_k = \frac{\alpha}{1-\alpha}\max_{i=1,\dots,n}\left[(T^k J)(i) - (T^{k-1}J)(i)\right] \end{aligned}$$

4.2 Policy Iteration

- **Proposition 12: Policy Iteration**
  Let μ and $\bar{\mu}$ be two stationary policies such that $T_{\bar{\mu}}J^{\mu} = TJ^{\mu}$ or equivalently

$$g(i,\bar{\mu}(i)) + \alpha \sum_{j=1}^{n} P_{ij}(\bar{\mu}(i))J^{\mu}(j) = \min_{u \in A(i)} \left[g(i,u) + \alpha \sum_{j=1}^{n} P_{ij}(u)J^{\mu}(j) \right] \quad i = 1,2,\ldots,n$$

  Then $J_{\bar{\mu}}(i) \leq J_{\mu}(i) \quad \forall i = 1,\ldots,n$
  Moreover, if μ is not optimal, strict inequality holds in the above for at least one state i.
  Since, $J^{\mu} = T_{\mu}J^{\mu}$ and by hypothesis $T_{\bar{\mu}}J^{\mu} = TJ^{\mu}$

$$J^{\mu}(i) = g(i,\mu(i)) + \alpha \sum_{j=1}^{n} P_{ij}(\mu(i))J^{\mu}(j) \geq g(i,\bar{\mu}(i)) + \alpha \sum_{j=1}^{n} P_{ij}(\bar{\mu}(i))J^{\mu}(j) = (T_{\bar{\mu}}J^{\mu})(i)$$

  Thus, $J^{\mu} = T_{\mu}J^{\mu} \geq TJ^{\mu} = T_{\bar{\mu}}J^{\mu} \implies J^{\mu} \geq T_{\bar{\mu}}J^{\mu}$
  Applying $T_{\bar{\mu}}$ repeatedly above and using monotonicity,

$$J^{\mu} \geq T_{\bar{\mu}}J^{\mu} \geq T_{\bar{\mu}}^2 J^{\mu} \geq \ldots \geq \lim_{k\to\infty} T_{\bar{\mu}}^k J^{\mu} = J^{\bar{\mu}} \implies J^{\mu} \geq J^{\bar{\mu}}$$

  If $J^{\mu} = J^{\bar{\mu}}$, then since $T_{\bar{\mu}}J^{\mu} = TJ^{\mu}$, we have

$$\begin{gathered} J^{\mu} = J^{\bar{\mu}} = T_{\bar{\mu}}J^{\bar{\mu}} = T_{\bar{\mu}}J^{\mu} = TJ^{\mu} = TJ^{\bar{\mu}} \\ J^{\mu} = TJ^{\mu} \text{ and } J^{\bar{\mu}} = TJ^{\bar{\mu}} \end{gathered}$$

  Since T has a unique fixed point, $J^{\mu} = J^{\bar{\mu}} = J^*$
  $\implies \mu$ and $\bar{\mu}$ are optimal policies
  Thus, if μ is not optimal, then $J^{\mu}(i) > J^{\bar{\mu}}(i)$ for at least one $i \in S$

- **Policy Iteration Algorithm**

Step 1: Initialize a stationary policy μ^0

Step 2: Policy Evaluation $\rightarrow$ Given a stationary policy u^k, compute the corresponding cost function J^{μ^k} from the linear system of equations

$$(I - \alpha P_{\mu^k})J^{\mu^k} = \bar{g}_{\mu^k} \text{ or } J^{\mu^k} = \bar{g}_{\mu^k} + \alpha P_{\mu^k}J^{\mu^k} \text{ or } J^{\mu^k} = TJ^{\mu^k}$$

Step 3: Policy Improvement $\rightarrow$ Obtain a new stationary policy μ^{k+1} satisfying $T_{\mu^{k+1}}J^{\mu^k} = TJ^{\mu^k}$

Step 4 : If $J^{\mu^k} = TJ^{\mu^k}$, stop, else go back to step 2 and repeat the process.

- So for we assumed knowledge of system model, transition probabilities and reword function.
- From now on we shall assume no knowledge of system model, in return we will have access to some data
  Data: $S_0, A_0, R_1, S_1, A_1, R_2, S_2, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
- Here, we consider tuples of data

$$(S_0, A_0, R_1, S_1), (S_1, A_1, R_2, S_2), \ldots, (S_{T-1}, A_{T-1}, R_T, S_T)$$

4.3 Monte Carlo Schemes

- Recall $J^{\mu}(i) = E\left[\sum_{k=1}^{T} r(S_k, \mu(S_k), S_{k+1}) \mid S_0 = i\right]$, cost-to-go under μ
- Monte Carlo schemes largely work with sample averages of collected data trajectories
- **First visit**: Ignore states on an episode that are not visited for the first time.
  Suppose we have n estimates for a state s (n episodes)

$$\hat{V}_1(s), \hat{V}_2(s), \ldots, \hat{V}_n(s) \implies V(s) \approx \frac{1}{n}\sum_{k=1}^{n} \hat{V}_k(s)$$

- **Every visit**: Similar to first visit, except that we do not ignore states on an episode that are not visited the first time.
  $G_m \triangleq r_{m+1} + r_{m+2} + \cdots + r_N =$ the return starting at time m when state $= S_m$

- MC can also be written as an update rule

$$V_n(s) = \frac{1}{n}\sum_{m=1}^{n} G_m, \quad n \geq 1 \text{ when } S_0 = s$$

- Then,

$$\begin{aligned}
V_{n+1}(s) &= \frac{1}{n+1}\sum_{m=1}^{n+1} G_m = \frac{1}{n+1}\left(\sum_{m=1}^{n} G_m + G_{n+1}\right) = \frac{n}{n+1}\frac{1}{n}\sum_{m=1}^{n} G_m + \frac{1}{n+1}G_{n+1} \\
&= \frac{n}{n+1}V_n(s) + \frac{1}{n+1}G_{n+1} \\
V_{n+1}(s) &= V_n(s) + \frac{1}{n+1}(G_{n+1} - V_n(s))
\end{aligned}$$

- In general, one may let $V_{n+1}(s) = V_n(s) + \alpha_n(G_{n+1} - V_n(s))$ where $\alpha_n, n \geq 0$ are step sizes such that

$$\sum_n \alpha_n = \infty, \quad \sum_n \alpha_n^2 < \infty$$

- One can show that as $n \to \infty$

$$V_n(s) \to E_\mu\left[G_{n+1} \mid S_{n+1} = s\right] = J^\mu(s)$$

- Online version of the algorithm

$$\begin{aligned}
&V_{n+1}(s_n) = V_n(s_n) + \alpha_n(G_{n+1} - V_n(s_n)) \\
&\text{with } V_{n+1}(s) = V_n(s) \quad \forall s \neq s_n \\
&V_{n+1}(s) = V_n(s) + \alpha_n \mathbf{1}_{\{s=s_n\}}(G_{n+1} - V_n(s_n))
\end{aligned}$$

- Recall that

$$\begin{aligned}
V_{n+1}(s_n) &= V_n(s_n) + \alpha_n(G_{n+!} - V_n(s_n)) \\
&= V_n(s_n) + \alpha_n(R_{n+1} + R_{n+2} + \ldots + R_N - V_n(s_n)) \\
&= V_n(s_n) + \alpha_n(R_{n+1} + V_n(s_{n+1}) - V_n(s_n) \\
&\qquad\qquad + R_{n+2} + V_n(s_{n+2}) - V_n(s_{n+1}) \\
&\qquad\qquad\quad \vdots \\
&\qquad\qquad + R_N + V_n(s_N) - V_n(s_{N-1}))
\end{aligned}$$

- Define

$$\begin{aligned}
d_n &= R_{n+1} + V_n(s_{n+1}) - V_n(s_n) \\
d_{n+1} &= R_{n+2} + V_n(s_{n+2}) - V_n(s_{n+1}) \\
&\vdots \\
d_{N-1} &= R_N + V_n(s_N) - V_n(s_{N-1})
\end{aligned}$$

  $d_n, d_{n+1}, \ldots, d_{N-1}$ are referred to as temporal difference(TD) terms.

- Then our update rule becomes, $V_{n+1}(s_n) = V_n(s_n) + \alpha_n(d_n + d_{n+1} + \cdots + d_{N-1})$

$$V_{n+l+1}(s_n) = V_{n+l}(s_n) + \alpha_n d_{n+l}, \quad l = 0, 1, \ldots, N-n$$

4.4 Temporal Difference

- Instead of looking at $V_\pi(s) = E_\pi[G_n \mid S_n = s]$, we look at Bellman equation

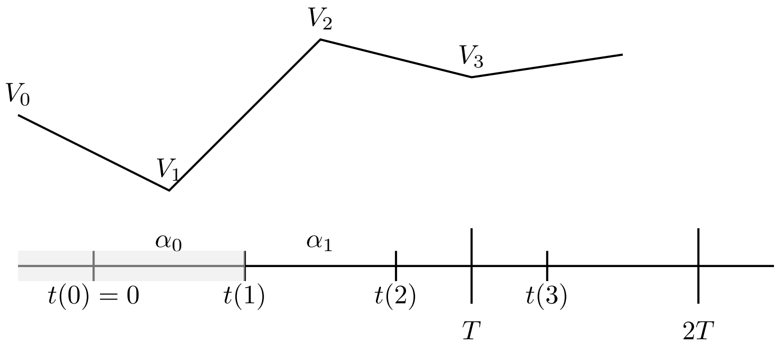
$$V_\pi(s) = E_\pi\left[R_{n+1} + V_\pi(s_{n+1}) \mid S_n = s\right] \text{ or } E_\pi\left[R_{n+1} + V_\pi(s_{n+1}) - V_\pi(s_n) \mid S_n = s\right] = 0$$

- **TD Recursion**: $V_{n+1}(s_n) = V_n s S_n) + \alpha_n(R_{n+1} + V_n(s_{n+1}) - V_n(s_n))$ for $n \geq 0$ with $V_{n+1}(s) = V_n(s) \quad \forall s \neq s_n$

- Alternatively, $V_{n+1}(s) = V_n(s) + \alpha_n \mathbf{1}_{\{s=s_n\}}(R_{n+1} + V_n(s_{n+1}) - V_n(s_n))$

- Suppose the Markov chain $\{S_n\}$ under policy π is episodic, i.e., irreducible, aperiodic, and positive recurrent.
- Then starting from any initial distribution, $\{S_n\}$ will settle into a steady state or stationary distribution that will be unique V.
- Form a sequence of finite points $\{t(n)\}$ as follows

$$t(0) = 0,\ t(1) = \alpha_0\ t(2) = \alpha_0 + \alpha_1, \ldots$$



Conditions on $\{\alpha\}$

$$\alpha_n > 0\ \forall n,\ \sum_n \alpha_n = \infty,\ \sum_n \alpha_n^2 < \infty$$

$$\implies t(n) \to \infty \text{ as } n \to \infty$$

$$\implies \alpha_n \to 0 \text{ as } n \to \infty$$

- One can show that

$$\lim_{n\to\infty} \sup_{t\in[T_n, T_{n+1}]} \|\bar{V}(t) - V^{T_n}(t)\| = 0 \text{ w.p.1}$$

Here, $\bar{V}(t)$, $t \geq 0$: Algorithm's trajectory (continuously interpolated)
$V^{T_n}(t)$, $t \in [T_n, T_{n+1}]$: ODE trajectory where $V^{T_n}(T_n) = \bar{V}(T_n)$

- Suppose the ODE has v^* as a globally asymptotically stable equilibrium. Then, the algorithm will satisfy $V_n \to v^*$ a.s. as $n \to \infty$ (under same conditions).
- ODE: $\dot{V}(t) = D\bar{V}$

$$\text{where } D = \begin{bmatrix} v(1) & & & 0 \\ & v(2) & & \\ & & \ddots & \\ 0 & & & v(3) \end{bmatrix}_{n\times n}$$

$$\text{where } \bar{V} = \begin{bmatrix} \sum_{j=1}^n P_{1j}(\pi(1))(R_\pi(1,j)) + V_\pi(j) - V_\pi(1) \\ \vdots \\ \sum_{j=1}^n P_{nj}(\pi(n))(R_\pi(n,j)) + V_\pi(j) - V_\pi(n) \end{bmatrix}$$

$V(t) = 0$ will satisfy the Bellman Equation.
Then it can be shown that

$$V_n \to V_\pi \text{ where } V_\pi = \begin{bmatrix} V_\pi(1) \\ \vdots \\ V_\pi(n) \end{bmatrix}$$

- A good reference for ODE approach: Chapter 2 of Stochastic approximation: A dynamical systems viewpoint, 2022
- Usual Stochastic Approximation algorithm

$$\text{Algorithm: } x_{n+1} = x_n + a(n)(h(x_n) + M_{n+1})$$

$$\text{ODE: } \dot{x}(t) = h(x(t))$$

Let $h(x) = \nabla f(x)$, $f : \mathbb{R}^d \to \mathbb{R}$

- In many cases, we are faced with algorithms, such as

$$x_{n+1} = x_n + a(n)(y_n + M_{n+1}) \text{ where } y_n \in h(x_n) \text{ (Now a set of points)}$$

$$\dot{x}(t) \in h(x(t))$$

Since, we have episodic tasks, each set of $h(x_n)$ can be thought of as an episode.

- **TD(λ) algorithm**: Consider the $(l+1)$ step Bellman equation

$$V_\pi(i_k) = E_\pi\left[\sum_{m=0}^{l} r(i_{k+m}, i_{k+m+1} + V_\pi(i_{k+l+1}))\right]$$

Since, value of l is arbitrary, we can form a weighted average y for all such Bellman equations.

$$\text{Let } 0 \le \lambda < 1. \text{ Since } \sum_{l=0}^{\infty}(1-\lambda)\lambda^l = 1$$

we can write the following Bellman equation

$$\begin{aligned}
V_\pi(i_k) &= (1-\lambda)E_\pi\left[\sum_{l=0}^{\infty}\lambda^l\left(\sum_{m=0}^{l} r(i_{k+m}, i_{k+m+1}) + V_\pi(i_{k+l+1})\right)\right] \\
&= (1-\lambda)E_\pi\left[\sum_{l=0}^{\infty}\lambda^l\sum_{m=0}^{l} r(i_{k+m}, i_{k+m+1})\right] + (1-\lambda)E_\pi\left[\sum_{l=0}^{\infty}\lambda^l V_\pi(i_{k+l+1})\right] \\
&= (1-\lambda)E_\pi\left[\sum_{l=0}^{\infty}\lambda^l\sum_{m=0}^{l} r(i_{k+m}, i_{k+m+1})\right] \quad \text{I} \\
&+ E_\pi\left[\sum_{l=0}^{\infty}(\lambda^l - \lambda^{l+1})V_\pi(i_{k+l+1})\right] \quad \text{II}
\end{aligned}$$

Now,

$$\begin{aligned}
\text{I} &= (1-\lambda)E_\pi\left[\sum_{m=0}^{\infty} r(i_{k+m}, i_{k+m+1})\sum_{l=m}^{\infty}\lambda^l\right] \\
&= E_\pi\left[\sum_{m=0}^{\infty}\lambda^m r(i_{k+m}, i_{k+m+1})\right]
\end{aligned}$$

$$\begin{aligned}
\text{II} &= E_\pi\left[\sum_{l=0}^{\infty}(\lambda^l - \lambda^{l+1})V_\pi(i_{k+l+1})\right] \\
&= E_\pi\left[(1-\lambda)V_\pi(i_{k+1}) + (\lambda-\lambda^2)V_\pi(i_{k+2}) + \cdots\right] \\
&= E_\pi[V_\pi(i_{k+1}) - V_\pi(i_k) \\
&\quad + \lambda(V_\pi(i_{k+2}) - V_\pi(i_{k+1})) \\
&\quad + \lambda^2(V_\pi(i_{k+3}) - V_\pi(i_{k+2})) \\
&\quad \vdots \\
&\quad + V_\pi(i_k)] \\
&= E\pi\left[\sum_{m=0}^{\infty}\lambda^m(V_\pi(i_{k+m+1}) - V_\pi(i_{k+m}))\right] + V_\pi(i_k)
\end{aligned}$$

Combining I and II, we get

$$V_\pi(i_k) = E_\pi\left[\sum_{m=0}^{\infty}\lambda^m(r(i_{k+m}, i_{k+m+1}) + V_\pi(i_{k+m+1}) - V_\pi(i_{k+m}))\right] + V_\pi(i_k)$$

Recall here that $\forall k \ge N$ (terminal instant)

$$i_k = 0,\ r(i_k, i_{k+1}) = 0,\ V_\pi(i_k) = 0$$

Letting $d_m = r(i_m, i_{m+1}) + V_\pi(i_{m+1}) - V_\pi(i_m)$ (temporal difference terms). Then,

$$\begin{aligned} V_\pi(i_k) &= E_\pi\left[\sum_{m=0}^{\infty} \lambda^m d_{m+k}\right] + V_\pi(i_k) \\ &= E_\pi\left[\sum_{m=k}^{\infty} \lambda^{m-k} d_m\right] + V_\pi(i_k) \\ &= V_\pi(i_k) \end{aligned}$$

Since, from Bellman equation: $E_\pi[d_m] = 0$
Stochastic approximation version

$$V(i_k) = V(i_k) + \alpha \sum_{m=k}^{\infty} \lambda^{m-k} \bar{d}_m$$

where $\bar{d}_m = r(i_m, i_{m+1}) + V(i_{m+1}) - V(i_m)$
Here α: Step size or learning rate
As number of iterates $\to \infty$, $V(i_k) \to V_\pi(i_k)$

Case 1: $\lambda = 0$ gives TD(0) algorithm

$$V(i_k) = V(i_k) + \alpha \bar{d}_k$$

where $\bar{d}_k = r(i_k, i_{k+1}) + V(i_{k+1}) - V(i_k)$

Case 2: $\lambda = 1$ gives TD(1) algorithm

$$V(i_k) = V(i_k) + \alpha \sum_m \bar{d}_k$$

- We have seen earlier that the sum of TD terms = Sum of rewards until termination
- This gives us Monte Carlo

4.5 Q-Learning

- Consider SSPP (Stochastic Shortest Path Problem)
- Recall Bellman Equation

$$J^*(i) = \min_{u \in A(i)} \left(\sum_{j=1}^{n} P_{ij}(u) \left(g(i,u,j) + J^*(j) \right) \right) \quad i \in S \implies Q^*(i,u)$$

- Let $Q^*(i,u) = \sum_{j=1}^{n} P_{ij}(u) \left(g(i,u,j) + J^*(j) \right) \quad \forall i \in S, u \in A(i)$
  Then, $J^*(i) = \min_{u \in A(i)} Q^*(i,u)$
- **Q-Bellman Equation**

$$Q^*(i,u) = \sum_{j=1}^{n} P_{ij}(u) \left(g(i,u,j) + \min_{v \in A(j)} Q^*(j,v) \right) \quad \forall i \in S, u \in A(i)$$

  A numerical scheme for its solution

$$Q_{m+1}(i,u) = \sum_{j=1}^{n} P_{ij}(u) \left(g(i,u,j) + \min_{v \in A(j)} Q_m(j,v) \right) \quad \forall i \in S, u \in A(i), m \geq 0$$

  It can be shown that $Q_n(i,u) \to Q^*(i,u)$ as $n \to \infty \quad \forall i \in S, \forall u \in A(i)$
- Suppose now that we do not have access to $P_{ij}(u) \quad \forall i,j \in S, u \in A(i)$
- But suppose we have access to states $j \sim P_i(u) \quad \forall i \in S, u \in A(i)$
- **Learning algorithm (Q-Learning)**

$$Q_{m+1}(i,u) = Q_m(i,u) + \gamma_m \left(g(i,u,j) + \min_{v \in A(j)} Q_m(j,v) - Q_m(i,u) \right) \quad \forall i, u \in A(i)$$

  γ_m(Learning Rate/Step size) should be selected such that $\sum_m \gamma_m = \infty, \sum_m \gamma_m^2 < \infty$

- **Proposition 13**: General proposition on convergence
  Consider the following algorithm

$$\gamma_{t+1}(i) = (1 - \gamma_t(i))\gamma_t(i) + \gamma_t(i)((H\gamma_t)(i) + w_t(i))$$

where

$$\text{(a)} \quad \sum_t \gamma_t(i) = \infty, \;\; \sum_t \gamma_t^2(i) < \infty$$

$$\text{(b)} \quad \forall i, t \quad E[w_t(i)|F_t] = 0 \text{ where } F_t = \sigma(\gamma_s, s \leq t, w_s, s < t)$$

$$\exists A, B > 0 \text{ such that } E\left[w_t^2(i) \mid F_t\right] \leq A + B\|\gamma_t\|^2 \quad \forall i, \forall t$$

(c) $H : \mathbb{R}^n \to \mathbb{R}^n$ is a weighted max norm pseudo contraction, i.e. $\exists r^* \in \mathbb{R}^n$, a positive vector $\xi = (\xi(1), \ldots, \xi(n))^\top$ and a constant $\beta \in [0, 1)$ such that $\|Hr - r^*\|_\xi \leq \beta\|r - r^*\|_\xi$ where $\|x\|_\xi = \max_{i=1,\ldots n} \frac{|x(i)|}{\xi(i)}$ for any $x = (x(1), \ldots x(n))^T$.
Then, $r_t \to r^*$ as $t \to \infty$ w.p. 1.
We will not prove this result, but use it to prove convergence of Q-Learning.

- **Q-Learning Convergence**: Recall the Q-Learning algorithm

$$Q_{t+1}(i, u) = (1 - \gamma_t(i, u))\, Q_t(i, u) + \gamma_t(i, u) \left(g(i, u, \bar{i}) + \min_{v \in A(\bar{i})} Q_t(\bar{i}, v)\right) \quad \forall i \in S, u \in A(i)$$

where $\bar{i} \sim P_i(u)$
Let $Q_t(0, u) = 0 \quad \forall u \in A(0)$, terminal states.
Let $T^{i,u} \triangleq$ set of times at which $Q(i, u)$ is updated.
Let $\gamma_t(i, u) = 0 \quad \forall t \notin T^{i,u}$ and $\sum_t \gamma_t(i, u) = \infty \quad \sum_t \gamma_t^2(i, u) < \infty$
Then, $Q_t(i, u) \to Q^*(i, u)$ w.p. 1 $\forall i, u$ in both the following cases

  (i) All policies are proper
  (ii) Assumptions a and b hold

Proof: Define the mapping H as follows

$$(HQ)(i, u) = \sum_{j=1} P_{ij}(u)(g(i, u, j) + \min_{v \in A(j)} Q(j, v)) \quad \forall i \neq 0, u \in A(i)$$

The Q-Learning algorithm can then be rewritten as

$$Q_{t+1}(i, u) = (1 - \gamma_t(i, u))Q_t(i, u) + \gamma_t(i, u)((HQ_t)(i, u) + w_t(i, u))$$

$$\text{Here, } w_t(i, u) = g(i, u, \bar{i}) + \min_{v \in A(\bar{i})} Q_t(\bar{i}, v) - \sum_{j=1}^{n} P_{ij}(u)(g(i, u, j) + \min_{v \in A(j)} Q_t(j, v))$$

Note: $E[w_t(i, u)|F_t] = 0$

$$E[w_t^2(i, u)|F_t] \leq K(1 + \max_{j,v} Q_t^2(j, v))$$

Then, assumption (b) holds
Suppose now that all policies are proper. Then, we have seen that $\exists \xi(i) \quad \forall i \neq 0$ and $\beta \in [0, 1)$ such that

$$\sum_{j=1}^{n} P_{ij}(u)\xi(j) \leq \beta\xi(i) \quad \forall i \neq 0, u \in A(i)$$

Let $Q = (Q(i, u), i \in S, u \in A(i))^\top$

$$\text{Let } \|Q\|_\xi = \max_{i \in S, u \in A(i)} \frac{|Q(i, u)|}{\xi(i)}$$

Consider 2 vectors Q and $\bar{Q}$. Then,

$$
\begin{aligned}
|(HQ)(i,u) - (H\bar{Q})(i,u)| &\leq \sum_{j=1}^{n} P_{ij}(u) \left| \min_{v \in A(j)} Q(j,v) - \min_{v \in A(j)} \bar{Q}(j,v) \right| \\
&\leq \sum_{j=1}^{n} P_{ij}(u) \max_{v \in A(j)} |Q(j,v) - \bar{Q}(j,v)| \text{ (To be proved below)} \\
&\leq \sum_{j=1}^{n} P_{ij}(u) \max_{v \in A(j)} \left(\frac{|Q(j,v) - \bar{Q}(j,v)|}{\xi(j)} \right) \xi(j) \\
&\leq \sum_{j=1}^{n} P_{ij}(u) \|Q - \bar{Q}\|_{\xi} \xi(j) \\
&\leq \beta \|Q - \bar{Q}\|_{\xi} \xi(i) \quad \left(\text{Since } \sum_{j=1}^{n} P_{ij}(u)\xi(j) \leq \beta \xi(i)\right)
\end{aligned}
$$

Divide both sides by $\xi(i)$

$$
\begin{aligned}
\frac{|(HQ)(i,u) - (H\bar{Q})(i,u)|}{\xi(i)} &\leq \beta \|Q - \bar{Q}\|_{\xi} \quad \forall i \in S, u \in A(i) \\
\implies \|HQ - H\bar{Q}\|_{\xi} &\leq \beta \|Q - \bar{Q}\|_{\xi}
\end{aligned}
$$

By the general proposition on convergence, Q-learning algorithm converges.
Note that if $A \subset B$, then

$$
\begin{aligned}
\inf_{x \in A} f(x) &\geq \inf_{x \in B} f(x) \\
\inf_{x \in A} (f(x) + g(x)) &= \inf_{x,y \in A, x=y} (f(x) + g(y)) \\
&\geq \inf_{x,y \in A} (f(x) + g(y))
\end{aligned}
$$

$$
\begin{aligned}
\text{Thus, } \inf_{x \in A} (f(x) + g(x)) &\geq \inf_{x \in A} f(x) + \inf_{x \in A} g(x) \\
\inf_{x \in A} ((f-g)(x) + g(x)) &\geq \inf_{x \in A} (f-g)(x) + \inf_{x \in A} g(x) \\
\text{or } \inf_{x \in A} f(x) &\geq \inf_{x \in A} (f(x) - g(x)) + \inf_{x \in A} g(x) \\
\inf_{x \in A} (f(x) - g(x)) &\leq \inf_{x \in A} f(x) - \inf_{x \in A} g(x) \quad (*) \\
\text{Let } h(x) &= -g(x) \quad \forall x \\
\text{Then, } \sup_{x \in A} h(x) &= \sup_{x \in A} (-g(x)) = -\inf_{x \in A} g(x) \\
\inf_{x \in A} (f(x) + h(x)) &\leq \inf_{x \in A} f(x) + \sup_{x \in A} h(x) \quad \text{from } (*) \\
\inf_{x \in A} (f(x) + h(x)) - \inf_{x \in A} f(x) &\leq \sup_{x \in A} h(x)
\end{aligned}
$$

Let $h(x) = g(x) - f(x)$

$$
\begin{aligned}
\implies \inf_{x \in A} g(x) - \inf_{x \in A} f(x) &\leq \sup_{x \in A} (g(x) - f(x)) \\
&\leq \sup_{x \in A} |g(x) - f(x)|
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\inf_{x \in A} f(x) - \inf_{x \in A} g(x) &\leq \sup_{x \in A} |g(x) - f(x)| \\
\implies \left| \inf_{x \in A} f(x) - \inf_{x \in A} g(x) \right| &\leq \sup_{x \in A} |g(x) - f(x)| \\
\text{Thus, } \left| \min_{v \in A(j)} Q(j,v) - \min_{v \in A(j)} \bar{Q}(j,v) \right| &\leq \max_{v \in A(j)} |Q(j,v) - \bar{Q}(j,v)|
\end{aligned}
$$

- Suppose the state S_t is visited at time t.
  Q-learning algorithm (in the **online setting**)

$$
Q_{t+1}(S_t, A_t) = Q_t(S_t, A_t) + \gamma_t(S_t, A_t) \left(g(S_t, A_t, S_{t+1}) + \min_{v \in A(S_{t+1})} Q_t(S_{t+1}, v) - Q_t(S_t, A_t) \right)
$$

with $Q_{t+1}(s,a) = Q_t(s,a) \quad \forall s \neq S_t$ or $a \neq A_t$
Question: How do we select A_t in the update rule?
Answer: A_t is selected randomly from the set $A(S_t)$
An alternative way of rewriting the above is

$$Q_{t+1}(S_t, A_t) = Q_t(S_t, A_t) + \gamma\,(S_t, A_t)(g(S_t, A_t, S_{t+1}) + Q_t(S_{t+1}, A_{t+1}) - Q_t(S_t, A_t)) \quad \dagger$$

one possibility

$$A_t = \begin{cases} \arg\min_{u \in A(S_t)} Q_t(S_t, u) & \text{w.p. } 1-\epsilon \\ \text{random action w.p. } \epsilon \end{cases}$$

$$A_{t+1} = \underset{v \in A(S_{t+1})}{\arg\min}\; Q_t(S_{t+1}, v)$$

This is an Off policy algorithm

- **SARSA (State-Action-Reward-State-Action)**
  Use $\dagger$ update rule but with

$$A_t = \begin{cases} \arg\min_{u \in A(S_t)} Q_t(S_t, u) \text{ w.p. } 1-\epsilon \\ \text{random action w.p. } \epsilon \end{cases}$$

  Online algorithm

$$A_{t+1} = \begin{cases} \arg\min_{v \in A(S_{t+1})} Q_t\,(S_{t+1}, v) \text{ w.p. } 1-\epsilon \\ \text{random action w.p. } \epsilon \end{cases}$$

  This is an Online algorithm

- Other methods: Double Q-learning, Expected SARSA

5 Function Approximation

- Tabular RL $\rightarrow$ MDP $\quad (S, A, P, r, \gamma)$ can be solved $\forall (s,a) \in S \times A$
- Model-based $\rightarrow P, r$ known.
- Model-free $\rightarrow P, r$ unknown $\rightarrow$ Estimating P, r is hard.
- Estimating $Q_* \in \mathbb{R}^{|S||A|}$ is easier
- when S or A is very large, need RL with approximation $\rightarrow$ Deep RL
- Advantage Function

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$$

5.1 Stochastic Approximation

- Recall the general update equation

$$x_{n+1} = x_n + \alpha_n[\overbrace{h(x_n) + M_{n+1}}^{\text{Noisy estimate of the true driving function}}]$$

$x_n \rightarrow$ Current Estimate
$\alpha_n \rightarrow$ Step size
$h(x_n) \rightarrow$ "True" driving function

$$\begin{aligned} \rightarrow x_{n+1} - x_{n_0} &= \sum_{j=n_0}^{n} (x_{j+1} - x_j) \\ &= \sum_{j=n_0}^{n} \alpha_j \left[h\left(x_j\right) + M_{j+1}\right] \\ &= \sum_{j=n_0}^{n} \alpha_j h\left(x_j\right) + \underbrace{\sum_{j=n_0}^{n} \alpha_j M_{j+1}}_{\substack{\text{If cumulative error due to noise} \\ \text{is very large, it will be troublesome}}} \end{aligned}$$

- We want to identify conditions under which the cumulative error is negligible.
- Theorems [Chapter 2 of Borkar, Stochastic Approximation Theory: A Dynamical Systems Perspective]

  (A1) $h : \mathbb{R}^d \to \mathbb{R}^f$ is Lipschitz continuous: $\exists L \geq 0$ such that

  $$\|h(x) - h(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^d$$

  (A2) $\sum_{n=0}^{\infty} \alpha_n = \infty, \sum_{n=0}^{\infty} \alpha_n^2 < \infty \implies \alpha_n \to 0$

  $$\alpha_n = \frac{1}{n+1}, \quad \alpha_n = \frac{1}{(n+1)^k} \quad \frac{1}{2} < k \leq 1$$

  (A3) $(M_n)_{n\geq 0}$ needs to be square integrable martingale difference sequence i.e., $\exists F_n$ of σ-fields(collection of events $F_n \subset F_{n+1}$) such that $\mathbb{E}\left[M_{n+1}|F_n\right] = 0 \quad \forall n$

  $$\implies \mathbb{E}\left\|M_n\right\|^2 < \infty \quad \forall n \geq 1$$

  F_n can be viewed as information available at time n
  A sequence $\{y_n\}$ of i.i.d zero-mean random variables

  $$F_n = \sigma(y_0, \ldots, y_n)$$
  $$E[y_{n+1} \mid F_n] = E[y_{n+1}] = 0$$

  Martingale difference is a nice generalization of i.i.d

  (A4) Let $h_c(x) = \frac{h(cx)}{c}$ for $c \geq 1$
  Suppose $\exists$ a continuous function $h_\infty : \mathbb{R}^d \to \mathbb{R}^d$ such that $h_c(x) \to h_\infty(x)$ as $c \to \infty$ uniformly on compact sets.
  Furthermore, the ODE $\dot{x}(t) = h_\infty(x(t))$ has origin as its globally asymptotically stable equilibrium.
  $\to$ Then (x_n) generated by $x_{n+1} = x_n + \alpha_n[h(x_n) + M_{n+1}]$ converges to a compact, connected invariant set of the ODE $\dot{x}(t) = h(x(t))$.
  $A \subset \mathbb{R}^d$ is positively invariant if for any $x_0 \in A$,

  $$\frac{dx(t, 0, x_0)}{dt} = h(x(t, 0, x_0)) \text{ solution trajectory is in } A$$

- **Read**: Invariant sets of ODEs

5.2 TD Learning

- Given a policy μ, find the state value function, $J_\mu \in \mathbb{R}^{|S|}$

  $$J_\mu(s) = \mathbb{E}\left[g_n(x_n) + \sum_{i=0}^{n-1} g_i(x_i, \mu(x_i), x_{i+1}) \mid x_0 = s\right] \to \text{ more general finite horizon}$$
  $$\mu : S \to \Delta(A)$$
  $$s_1 \sim P(\cdot \mid s_0, a_0), \quad a_0 \sim \mu(\cdot \mid s_0) \quad a_1 \sim \mu(\cdot \mid s_1)$$
  $$J_\mu(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \underbrace{\gamma^t r(s_t, a_t)}_{\text{reward at time } t \text{ valued at time } 0}\right]$$

- We also consider policies from a risk perspective. e.g. a route takes 30 min on average, 3 hr at most; another route takes 45 min on average, 1.5 hr at most

  $$\underbrace{\begin{aligned} \to J_\mu(s) &= \mathbb{E}_{a,s'}[r(s, a) + \gamma J_\mu(s') \mid s_0 = s] \\ J_\mu(s) &= \sum_{a,s'} \mu(a \mid s)P(s' \mid s, a)[r(s, a) + \gamma J_\mu(s')] \end{aligned}}_{\text{Bellman Equation}}$$

- Suppose we know $P(s' \mid s, a)$, we want to find J_μ: Planning setup/Model-based setup, Bellman equation gives many linear equations

- Suppose we don't know $P(s' \mid s, a)$, want to find J_μ: Model-free setup, Each interaction with environment gives $(s, a, r(s,a), s')$
  If we have a lot of these, we can exploit the Law of Large Numbers to approximate expectation
$$\begin{aligned} & x_1, x_2, \ldots, x_n \rightarrow \text{ Samples} \\ & \frac{x_1 + \ldots + x_n}{n} \xrightarrow{\text{a.s.}} \mathbb{E}(x) \\ & 1 \text{ run: } x_1, \frac{x_1 + x_2}{2}, \ldots \end{aligned}$$
- Pick state s, In finite horizon setting, start at s, interact with μ, we get
$$\begin{gathered} s_0, a_0, r\left(s_0, a_0\right), s_1, \ldots, s_T \\ \downarrow \\ \sum_{t=0}^{T-1} \gamma^t r\left(s_t, a_t\right) + \gamma^T r\left(s_T\right) \rightarrow \text{ one sample} \end{gathered}$$
  So, for every state s, collect k samples of $\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) + \gamma^T r(s_t)$ where $s_0 = s$. Let these k samples be $C_1(s), \ldots, C_k(s)$
$$J_\mu(s) \approx \frac{1}{k} \sum_{i=1}^{k} C_i(s)$$
  This is the naive approach
  Issues with this approach are
  1. Space requirement grows linearly with n
  2. Time complexity grows linearly with n
  3. The process is non-incremental, Each time we get a new sample, we do all the calculations from scratch
- A somewhat better approach:
$$\begin{aligned} x_n &= \frac{x_1 + \ldots + x_n}{n} \\ x_{n+1} &= \frac{n x_n + x_{n+1}}{n+1} \\ &= x_n - \frac{1}{n+1} x_n + \frac{x_{n+1}}{n+1} \end{aligned}$$
  New Estimate:
$$x_{n+1} = x_n + \alpha_n [x_{n+1} - x_n] \quad , \alpha_n = \frac{1}{n+1} \rightarrow \text{Stochastic Approximation}$$
  Example: Let $f(x) = \frac{1}{2}(x - \mathbb{E}x)^2$
$$\begin{aligned} \nabla f(x) &= (x - \mathbb{E}x) \\ \implies x_{n+1} &= x_n + \alpha_n \left(-\nabla f\left(x_n\right)\right) \\ &= x_n + \alpha_n \left(\mathbb{E}x - x_n\right) \\ x_{n+1} &= x_n + \alpha_n \left[x_{n+1} - x_n\right] \\ &\uparrow \\ \hat{J}_\mu^{n+1}(s) &= \hat{J}_\mu^n(s) + \alpha_n [C_{n+1}(s) - \hat{J}_\mu(s)] \\ x_{n+1} &= x_n + \underbrace{\alpha_n [r(s_n, a_n) + \gamma x_n(s_{n+1}) - x_n(s_n)]}_{\in \mathbb{R}} e_{s_n} \end{aligned}$$
  where $x_{n+1}, x_n, e_{s_n} \in \mathbb{R}^{|S|}$
- **Temporal Difference Learning**: Try to derive the algorithm
$$\begin{aligned} f(x) =& \frac{1}{2} \left\| J_\mu - x \right\|_D^2 = \frac{1}{2} \sum_s d(s) [J_\mu(s) - x(s)]^2 \\ \nabla f(x) =& -\sum_s d(s)(J_\mu(s) - x(s)) e_s \\ x_{n+1} =& x_n + \alpha_n [-\nabla f(x_n)] \\ =& x_n + \alpha_n \left[\sum_s d(s)(J_\mu(s) - x_n(s)) e_s\right] \end{aligned}$$

Use Bellman equation for $J_\mu(s)$:

$$x_{n+1} = x_n + \alpha_n \sum_{s,a,s'} d(s)\mu(a \mid s)p(s' \mid s,a)[r(s,a) + \gamma J_\mu(s') - x_n(s)]e_s$$

J_μ is an expectation $\to$ infinite sum. X_n is current estimate of J_μ, so lazily put $x_n(s)$:

$$\approx x_n + \alpha_n \sum_{s,a,s'} d(s)\mu(a \mid s)P(s' \mid s,a)[r(s,a) + \gamma x_n(s') - x_n(s)]e_s$$

TD(0) Algorithm from a generative model with Markov sampling

$$x_{n+1} = x_n + \alpha_n[r(s_n,a_n) + \gamma x_n(s'_n) - x_n(s_n)]e_{s_n}$$

This is not a stochastic gradient descent, Gradient descent analysis can only be done with SA Theory

$$\begin{gathered} s_n \sim d \\ a_n \sim \mu(\cdot \mid s_n) \\ s'_n \sim P(\cdot \mid s_n, a_n) \end{gathered}$$

Using this in practice is difficult.

- **Policy Evaluation with Function Approximation**: Given y, we want to find J_μ

$$J_\mu(s) = \mathbb{E}\left[\sum_t^\infty \gamma^t r(s_t,a_t) \mid s_0 = s\right]$$

$s_0, a_0, s_1, \ldots, a_t \sim \mu(\cdot|s_t)$

- Recall the Bellman Equation

$$\begin{aligned} J_{\mu(s)} &= \mathbb{E}_{a,s'}[r(s,a) + \gamma J_\mu(s')] \\ &= \sum_{s',a} \mu(a \mid s)P(s' \mid s,a)[r(s,a) + J_\mu(s')] \end{aligned}$$

Given the tuple $(s, a, r(s,a), s')$, the tabular TD(0) is

$$x_{n+1} = x_n + \alpha_n[r(s_n,a_n) + \gamma x_n(s'_n) - x_n(s_n)]e_{s_n}$$

$$\begin{gathered} s_n \sim d_\mu(.) \to \text{some stationary distribution associated with a Markov chain induced} \\ a_n \sim \mu(\cdot \mid s_n) \\ s'_n \sim P(\cdot \mid s_n, a_n) \end{gathered}$$

Consider an example: $S = \{1,2\}$, $A = \{L,R,U,D\}$

$$x_n \in \mathbb{R}^2 \quad d_\mu = (0.2, 0.8)$$

$$\mu : \begin{bmatrix} 0.1 & 0.1 & 0.2 & 0.6 \\ 0.1 & 0.3 & 0.15 & 0.15 \end{bmatrix}$$

For next state we refer to $P(\cdot|s_n,a_n)$ row in P matrix, P is a $|S||A| \times |S|$ sized matrix.
Say $n = 5$, $s_n = 2$

$$x_6 = x_5 + \alpha_5[5 + 0.9x_5(1) - x_5(2)]e_2$$

MDP $= (S, A, P, r, \gamma)$, where $P \to |S||A| \times |S|$
MC $= (S, P_\mu)$, where $P_\mu \to |S| \times |S|$
P_μ is transition matrix induced by policy μ

$$P_\mu(s' \mid s) = \sum_a \mu(a \mid s)P(s' \mid s,a)$$

A stationary distribution d_μ is one such that

$$\begin{aligned} &d_\mu^\top P_\mu = d_\mu \\ &s_0 \sim d_\mu \\ &s_1 \sim d_\mu \end{aligned}$$

- The above algorithm is for a tabular setting. For a function approximation setting:
  Given: μ, x
  Goal: Find an approximation of J_μ in x

- **Linear Function Approximation**
$$\Phi \in \mathbb{R}^{S\times d} \quad , d \ll S$$
  $x =$ columns of Φ
  New goal: Find θ^* such that $J_\mu \approx \Phi\theta^*$, θ is a vector in $\mathbb{R}^d$
$$f(\theta) = \frac{1}{2}\|\Phi\theta - J_\mu\|^2_{D_\mu}, \quad D_\mu \in \mathbb{R}^{|S|\times|S|} = \text{diag}(d_\mu)$$
$$f(\theta) = \sum_s \frac{1}{2} d_\mu(s)(\Phi^\top(s)\theta - J_\mu(s))^2$$
  where $\Phi^\top(s)$ is the s^{th} row of Φ
$$\text{Gradient Descent:} \quad \theta_{n+1} = \theta_n + \alpha_n(-\nabla f(\theta_n))$$
$$\nabla f(\theta) = \sum_s d_\mu(s)\left(\Phi^\top(s)\theta - J_\mu(s)\right)\Phi(s)$$
$$\theta_{n+1} = \theta_n + \alpha_n \sum_s d_\mu(s)\left(J_\mu(s) - \Phi^\top(s)\theta_n\right)\Phi(s)$$
  Use Bellman equation for J_μ
$$= \theta_n + \alpha_n \sum_{s,a,s'} d_\mu(s)\mu(a \mid s)P(s' \mid s,a)\,[r(s,a)+?$$
  New algorithm becomes, **TD(0) with Function Approximation**
$$\theta_{n+1} = \theta_n + \alpha_n[r(s_n,a_n) + \gamma\Phi^\top(s'_n)\theta_n - \Phi^\top(s_n)\theta_n]\Phi(s_n)$$
  Entire operation happening in d-dimensional space
$$s_n \sim d_\mu(\cdot)$$
$$a_n \sim \mu(\cdot \mid s_n)$$
$$s'_n \sim P(\cdot \mid s_n, a_n)$$
$$(s_n, a_n, s'_n) \text{ is i.i.d.}$$

- **Analysis for Algorithm**
$$\theta_{n+1} = \theta_n + \alpha_n[h(\theta_n) + M_{n+1}]$$
$$F_n = \sigma(\theta_0, s_0, a_0, r(s_0,a_0), s'_0, s_1, a_1, r(s_1,a_1), s'_1, \ldots, s_{n-1}, a_{n-1}, r(s_{n-1},a_{n-1}), s'_{n-1}) \to \sigma\text{-field}$$
  $\theta_0, \ldots, \theta_n \in F_n$ are measurable w.r.t. F_n.
  Let $\delta_n = r(s_n,a_n) + \gamma\Phi^\top(s'_n)\theta_n - \Phi^\top(s_n)\theta_n$
$$\theta_{n+1} = \theta_n + \alpha_n\delta_n\Phi(s_n)$$
$$h(\theta_n) = \mathbb{E}[\delta_n\Phi(s_n) \mid F_n] \quad (\mathbb{E}[c] = c)$$
$$= \mathbb{E}[r(s_n,a_n)\Phi(s_n) + \gamma\Phi(s_n)\Phi^\top(s'_n)\theta_n - \Phi(s_n)\Phi(s_n)^\top\theta_n \mid F_n]$$
  By linearity of conditional expectation
$$= \mathbb{E}[r(s_n,a_n)\Phi(s_n) \mid F_n] + \gamma\mathbb{E}[\Phi(s_n)\Phi^\top(s'_n) \mid F_n]\theta_n - \mathbb{E}[\Phi(s_n)\Phi^\top(s_n) \mid F_n]\theta_n$$
  $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$ if X is independent of Y
$$= \mathbb{E}[r(s_n,a_n)\Phi(s_n)] + \gamma\mathbb{E}[\Phi(s_n)\Phi^\top(s'_n)]\theta_n - \mathbb{E}[\Phi(s_n)\Phi^\top(s_n)]\theta_n \to \text{since } (s_n,a_n,s'_n) \text{ is independent of } F_n$$
$$\mathbb{E}[r(s_n,a_n)\Phi(s_n)] = \sum_{s,a} d_\mu(s)\sum_a \mu(a \mid s)r(s,a)\Phi(s) + \ldots$$
$$= \Phi^\top D_\mu r_\mu + \gamma\Phi^\top D_\mu P_\mu \Phi\theta_n - \Phi^\top D_\mu\Phi\theta_n$$
$$\to \theta_{n+1} = \theta_n + \alpha_n[h(\theta_n) + M_{n+1}]$$
  where $M_{n+1} = \delta_n\Phi(s_n) - h(\theta_n)$
  and $h(\theta_n) = \mathbb{E}[\delta_n\Phi(s_n) \mid F_n]$
$$\mathbb{E}[M_{n+1} \mid F_n] = \mathbb{E}[\delta_n\Phi(s_n) \mid F_n] - \mathbb{E}[h(\theta_n) \mid F_n] = \mathbb{E}[\delta_n\Phi(s_n) \mid F_n] - h(\theta_n) = 0$$
  $h(\theta) = b - A\theta$, where
$$b = \Phi^\top D_\mu r_\mu \quad A = \Phi^\top D_\mu(I - \gamma P_\mu)\Phi$$

- First we analyze the noiseless version of this algorithm. This algorithm and $\dot{\theta}(t) = h(\theta(t))$ have the following relation

$$\begin{aligned}\theta(t_2) - \theta(t_1) &= \int_{t_1}^{t_2} \dot{\theta}(t)dt \\ &= \int_{t_1}^{t_2} h(\theta(t))dt \\ &\approx h(\theta(t))(t_2 - t_1) \\ \theta(t_2) &= \theta(t_1) + (t_2 - t_1)h(\theta(t))\end{aligned}$$

Also, we can write

$$\begin{aligned}\dot{\theta}(t) &= b - A\theta(t) \\ b - A\theta &= 0 \\ \theta'_* &= A^{-1}b\end{aligned}$$

- We will answer two questions
  Is θ'_* asymptotically stable
  Waht is the relation between θ_* and θ'_*

- Answer to first question can be given using Lyapunov function
  Let $V : \mathbb{R}^d \to \mathbb{R}$ be given by

$$V(\theta) = \frac{1}{2}\|\theta - \theta'_*\|^2$$

Suppose,

$$\begin{aligned}\nabla V(\theta)^T h(\theta) &< 0 \quad \forall \theta \neq \theta_* \\ \frac{dV(\theta(t))}{dt} &= \nabla V^T(\theta(t))\dot{\theta}(t) \\ &= \nabla V^T(\theta(t))h(\theta(t)) < 0\end{aligned}$$

$$\begin{aligned}\nabla V(\theta) &= (\theta - \theta'_*) \\ \nabla V(\theta)^T h(\theta) &= (\theta - \theta'_*)^T(b - A\theta) \\ &= (\theta - \theta'_*)^T A(\theta'_* - \theta) \\ &= -(\theta - \theta'_*)^T A(\theta - \theta'_*)\end{aligned}$$

Claim: $\theta^T A\theta > 0 \; \forall \theta \neq 0$
Recall that $A = \Phi^T D_\mu(I - \gamma P_\mu)\Phi \in \mathbb{R}^{d\times d}$
Assume Φ has full column rank and μ is such that $d_\mu > 0$

$$\begin{aligned}\theta^T A\theta &= \theta^T \Phi^T D_\mu(I - \gamma P_\mu)\Phi\theta \\ &= y^T \underbrace{D_\mu(I - \gamma P_\mu)}_{B} y\end{aligned}$$

We will show that B is PD

$$y^T D_\mu - \gamma y^T D_\mu P_\mu y > 0 \quad \forall y \neq 0$$
$$\text{To Prove: } y^T D_\mu P_\mu y \leq y^T D_\mu y$$

$$\begin{aligned}y^T D_\mu P_\mu y &= y^T D_\mu^{\frac{1}{2}} D_\mu^{\frac{1}{2}} P_\mu y \\ &\leq \|D_\mu^{\frac{1}{2}} y\|_2 \|D_\mu^{\frac{1}{2}} P_\mu y\|_2 \quad \{\text{From Cauchy Schwarz}\end{aligned}$$

We can write

$$\begin{aligned}\|D_\mu^{\frac{1}{2}} y\|_2 = \sqrt{y^T D_\mu^{\frac{1}{2}} D_\mu^{\frac{1}{2}} y} = \sqrt{y^T D_\mu y} = \sqrt{\|y\|_{D_\mu}^2} &= \|y\|_{D_\mu} \\ &= \|y\|_{D_\mu}\|P_\mu y\|_{D_\mu}\end{aligned}$$

Now, we will prove $\|P_\mu y\|^2_{D_\mu} \leq \|y\|^2_{D_\mu}$

$$\begin{aligned}
\text{L.H.S}^2 &= \sum_s d_\mu(s)(P_\mu^T(s,\cdot)y)^2 \\
&= \sum_s d_\mu(s) \sum_{s'} P_\mu(s'|s) y^2(s') \\
&\leq \sum_{s'} y^2(s') \sum_s d_\mu(s) P_\mu(s'|s) \\
&= \sum_{s'} y^2(s') d_\mu(s') \\
&= \|y\|^2_{D_\mu}
\end{aligned}$$

Overall, this becomes

$$\begin{aligned}
y^T D_\mu P_\mu y &\leq \|y\|_{D_\mu} \|P_\mu y\|_{D_\mu} \\
&\leq \|y\|^2_{D_\mu} \\
&\leq y^T D_\mu y
\end{aligned}$$

- Claim: If $\theta'_* = A^{-1}b$ then,

$$\pi T_\mu \Phi \theta'_* = \Phi \theta'_*$$

where $\pi = \Phi(\Phi^T D_\mu \Phi)^{-1} \Phi^T D_\mu$, and πT_μ is the projected bellman operator
$\Phi\theta'_*$ is the fixed point of πT_μ
From the formulation,

$$\|J - \pi J\|^2_{D_\mu} = \min_\theta \frac{1}{2} \|J - \Phi\theta\|^2_{D_\mu}$$

- Recall the formulation

$$\begin{aligned}
\nabla f(\theta) &= 0 \\
\implies \sum_s d_\mu(s) J_\mu(s) \Phi(s) &= \sum_s d_\mu(s) \Phi^\top(s) \theta \Phi(s) \\
\text{LHS} = \Phi^\top D_\mu J_\mu \quad \text{RHS} &= \sum_s d_\mu(s) \Phi^\top(s) \Phi(s) \theta = \Phi^\top D_\mu \Phi \theta \\
\Phi^\top D_\mu J_\mu &= \Phi^\top D_\mu \Phi \theta \\
\theta_* &= (\Phi^\top D_\mu \Phi)^{-1} \Phi^\top D_\mu J_\mu \\
\Phi\theta_* &= \Phi(\Phi^\top D_\mu \Phi)^{-1} \Phi^\top D_\mu J_\mu
\end{aligned}$$

The above is the closest approximation to J in col(Φ)
Try:

$$\pi T_\mu \Phi \theta'_* = \Phi \theta'_*$$

- Claim: Distance after projection will not increase

$$\|\pi V - \pi V'\|_{D_\mu} \leq \|V - V'\|_{D_\mu}$$

- Claim: $\|J_\mu - \Phi\theta_*\|_{D_\mu} \leq \|J_\mu - \Phi\theta'_*\|_{D_\mu}$ as θ_* was obtained to minimize $\|J_\mu - \Phi\theta_*\|_{D_\mu}$
Attempt to bound:

$$\begin{aligned}
\|J_\mu - \Phi\theta'_*\|_{D_\mu} &\leq \|J_\mu - \Phi\theta_*\|_{D_\mu} + \|\pi J_\mu - \Phi\theta'_*\|_{D_\mu} \\
&\leq \|J_\mu - \Phi\theta_*\|_{D_\mu} + \|\pi J_\mu - \pi T_\mu \Phi\theta'_*\|_{D_\mu} \\
&\leq \|J_\mu - \Phi\theta_*\|_{D_\mu} + \|J_\mu - T_\mu \Phi\theta'_*\|_{D_\mu} \\
J_\mu \text{ is the fixed the point of } T_\mu &\quad (T_\mu J_\mu = J_\mu) \\
&\leq \|J_\mu - \Phi\theta_*\|_{D_\mu} + \|T_\mu J_\mu - T_\mu \Phi\theta'_*\|_{D_\mu} \\
&\leq \|J_\mu - \Phi\theta_*\|_{D_\mu} + \gamma \|J_\mu - \Phi\theta'_*\|_{D_\mu} \quad (\text{as } T_\mu \text{ is a } \gamma\text{-contraction w.r.t } \|\cdot\|_{D_\mu}) \\
\|J_\mu - \Phi\theta'_*\|_{D_\mu} &\leq \frac{1}{1-\gamma} \|J_\mu - \Phi\theta_*\|_{D_\mu}
\end{aligned}$$

- **Summary**: Given μ, J_μ, Φ

$$\theta_{n+1} = \theta_n + \alpha_n \left[r(s_n, a_n) + \gamma \Phi^\top (s'_n) \theta_n - \Phi^\top (s_n) \theta_n \right] \Phi(s_n)$$

sample: (s_n, a_n, s'_n)

$$\theta_{n+1} = \theta_n + \alpha_n \left[b - A\theta_n + M_{n+1} \right] \text{ Noisy, Stochastic } \quad 1$$

Noiseless: $\dot{\theta}(t) = b - A\theta(t) \to \theta'_* = A^{-1}b$ Given initial time t_0 and initial point θ_0

$$\theta(t, t_0, \theta_0) \to \theta_*$$

- **Claim**: $(\theta_n)_{n\geq 0}$ generated using 1 converges almost surely to θ'_* i.e. $\theta_n \xrightarrow{\text{a.s.}} \theta'_*$
Proof: We verify the four assumptions of the convergence result proved by Michel Benaim in 1996 (Ch 2 of Borkar)

(A1) h is Lipschitz continuous

$$\|h(x) - h(y)\| \leq L\|x - y\|$$

$h(x) = b - Ax$, and $h(y) = b - Ay$

$$\|h(x) - h(y)\| = \|A(x - y)\| \leq \|A\|\|x - y\|$$
$$\|A\| = \sup \frac{\|Ax\|}{\|x\|}$$

(A2) $\sum_{n\geq 0} \alpha_n = \infty, \sum_{n\geq 0} \alpha_n^2 < \infty$
We can choose α_n to satisfy these

$$\alpha_n = \frac{1}{n+1}, \quad \alpha_n = \frac{1}{n^\sigma} \quad \sigma \in \left(\frac{1}{2}, 1\right], \quad \alpha_n = \frac{1}{n \ln n}$$

In practice, we can keep α_n constant for a while, then decay it then keep it constant again for a while. (Example)
There exist better strategies for different domains

(A3) $(M_n)_{n\geq 1}$ be a square-integrable martingale difference sequence, i.e

$$\begin{aligned} &\text{(a)} \quad \mathbb{E}\|M_n\|^2 < \infty \quad \forall n \\ &\text{(b)} \quad \mathbb{E}[M_{n+1} \mid F_n] = 0 \\ &\text{(c)} \quad \mathbb{E}[\|M_{n+1}\|^2 \mid F_n] \leq K[1 + \|\theta_n\|^2] \text{ for some } K \geq 1 \end{aligned}$$

$$\begin{aligned} (b) &= \mathbb{E}[r(s_n, a_n)\Phi(s_n) \mid F_n] \\ &= \mathbb{E}[r(s_n, a_n)\Phi(s_n)] \\ &= \Phi^\top D_\mu r_\mu, \quad r_\mu(s) = \sum_a \mu(a \mid s) r(s, a) \end{aligned}$$

Extra: Prob. with martingales with David Williams

$$\begin{aligned} \text{(c)} \quad M_{n+1} &= \delta_n \Phi(s_n) - (b - A\theta_n) \\ \|r(s_n, a_n)\Phi(s_n)\| &= |r(s_n, a_n)|\|\Phi(s_n)\| \\ &\leq R_{max} \times 1 \quad \text{WLoG, assume } \|\Phi(s_n)\| \text{ upper bounded by 1} \\ b &= \mathbb{E}[r(s_n, a_n)\Phi(s_n) \mid F_n] \\ \|b\| &\leq R_{\max} \cdot 1 \end{aligned}$$

$$A = \mathbb{E}[\underbrace{\gamma \Phi(s_n)\Phi^\top(s'_n)}_{d\times d} - \underbrace{\Phi(s_n)\Phi^\top(s_n)}_{d\times d} \mid F_n]$$

$$\begin{aligned} \|\gamma \Phi(s_n)\Phi^\top(s'_n)\| &= \gamma \|\Phi(s_n)\Phi^\top(s'_n)\| \\ &\leq \gamma \|\Phi(s_n)\|\|\Phi(s'_n)\| \end{aligned}$$

$$\|uv^\top\| = \sup_{x\neq 0} \frac{\|uv^\top x\|}{\|x\|} = \|v^\top x\| \frac{\|u\|}{\|x\|} \leq \frac{\|v\|\|x\|\|u\|}{\|x\|} = \|v\|\|u\|$$

We can think of $\mathbb{E}[\times|G]$ as the best representation/guess of $\times$ given the information in G.

$$M_{n+1} = \delta_n \Phi(s_n) + (b - A\theta_n)$$
$$M_{n+1} = r(s_n, a_n)\Phi(s_n) - b + [A - (\Phi(s_n)\Phi^\top(s_n) - \gamma\Phi(s_n)\Phi(s_n'))]\theta_n$$

$$\|M_{n+1}\|^2 \leq 2\|\overbrace{r(s_n,a_n)}^{\text{scalar}}\overbrace{\Phi(s_n)}^{\text{vector}} - b\|^2 + 2\|A - \underbrace{(\underbrace{\Phi(s_n)\Phi^\top(s_n)}_{\text{finite norm}} - \gamma\Phi(s_n)\Phi^\top(s_n'))}_{\text{bounded/finite norm}}\|^2\|\theta_n\|^2 \quad (a+b)^2 \leq 2a^2 + 2b^2$$

$$\|M_{n+1}\|^2 \leq K(1 + \|\theta_n\|^2) \quad 1$$

Monotonicity Property of Expectation (also works for conditionals)

$$X \leq Y \implies \mathbb{E}[X] \leq \mathbb{E}[Y]$$

From the monotonicity property of conditional expectations, eq 1

$$\begin{aligned}\implies \mathbb{E}[\|M_{n+1}\|^2 \mid F_n] &\leq \mathbb{E}[K(1 + \|\theta_n\|^2) \mid F_n] \\ &= K(1 + \|\theta_n\|^2)\end{aligned}$$

We now verify (a)

$$\begin{aligned}\theta_{n+1} &= \theta_n + \alpha_n[b - A\theta_n + M_{n+1}] \\ M_{n+1} &= \delta_n\Phi(s_n) - [b - A\theta_n] \\ &= [r(s_n, a_n)\Phi(s_n) - b] + [A - (\Phi(s_n)\Phi^\top(s_n) - \gamma\Phi(s_n)\Phi^\top(s_n))]\theta_n \\ \|M_{n+1}\| &\leq K_1 + K_2\|\theta_n\|\end{aligned}$$

How much can θ_n grow to?

$$\begin{aligned}\theta_1 &= \theta_0 + \alpha_0 Z_0 \\ \theta_2 &= \theta_1 + \alpha_1 Z_1 \\ &\vdots\end{aligned}$$

If Z_0 is bounded, then θ_1 is bounded
If Z_1 depending on θ_1 is finite, then θ_2 is finite

$$\begin{aligned}\|M_{n+1}\| &\leq K'[1 + \|\theta_n\|] \text{ where } K' = \max\{K_1, K_2\} \\ \theta_1 &= \theta_0 + \alpha_0\delta_0\Phi(s_0) \\ &= \theta_0 + \alpha_0[r(s_0, a_0)\Phi(s_0) + (\gamma\Phi(s_0)\Phi^\top(s_0) - \Phi(s_0)\Phi^\top(s_0))\theta_0 \\ \|\theta_1\| &\leq \|\theta_0\| + 1 \cdot [R_{\max} \cdot 1 + (\gamma + 1)\|\theta_0\|] \\ \|\theta_{n+1}\| &\leq \|\theta_n\| + 1 \cdot [R_{\max} \cdot 1 + (\gamma + 1)\|\theta_n\|]\end{aligned}$$

If $\|\theta_0\| \leq C_0$

$$\begin{aligned}\|\theta_1\| \leq C_0 + [R_{\max} + (1 + \gamma)C_0] = C_1 \\ \implies \|\theta_n\| \leq C_n < \infty\end{aligned}$$

The bound can grow with n, but will be finite

(A4) $h_c(\theta) = \frac{h(c\theta)}{c}$, $\lim_{c\to\infty} h_c(\theta) = h_\infty(\theta)$
$\dot{\theta}(t) = h_\infty(\theta(t))$, origin be globally asymptotically stable equilibrium

$$\begin{aligned}h(\theta) &= b - A\theta \\ h_c(\theta) &= \frac{b - A(c\theta)}{c} = \frac{b}{c} - A\theta \\ h_\infty(\theta) &= \lim_{c\to\infty} h_c(\theta) = -A\theta \\ \dot{\theta}(t) &= -A\theta(t)\end{aligned}$$

Is the origin a globally asymptotically stable equilibrium?
Construct Lyapunov function

$$\begin{aligned}V(\theta) = \|\theta\|^2 \quad V(\theta) = 0 \text{ iff } \|\theta\| = 0 \\ \nabla V^\top(\theta)h_\infty(\theta) = -\theta^\top A\theta < 0 \quad \forall\theta \neq 0 \\ \frac{dV(\theta(t))}{dt} = \nabla V^\top(\theta(t)) \cdot h_\infty(\theta(t)) < 0\end{aligned}$$

- **Conclusion**: $\theta_n \xrightarrow{a.s.} \theta'_* = A^{-1}b$

5.3 Q-Learning

- Q_* satisfies Bellman Equations

$$\begin{aligned} Q_* &= TQ_* \quad \text{(Bellman optimality Equation)} \\ Q_* &= T_{\pi_*} Q_* \text{ (Bellman Equation)} \end{aligned}$$

π_* is greedy w.r.t. Q_*

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha_n[r(s_n, a_n) + \gamma \max_{a'} Q_n(s'_n, a') - Q_n(s_n, a_n)]e_{s_n,a_n} \quad [e \text{ is column vector of size } |S||A| \times 1] \\ \to f(\theta) &= \frac{1}{2}\|Q - Q^*\|_2^2 \\ Q_{n+1} &= Q_n + \alpha_n[Q_* - Q_n] \\ &= Q_n + \alpha_n[TQ_* - Q_n] \\ TQ_*(s,a) &= \mathbb{E}[r(s,a) + \gamma \max_{a'} Q_*(s', a')] \quad s' \sim P(\cdot \mid s, a) \\ Q_{n+1}(s,a) &= Q_n(s,a) + \alpha_n[\mathbb{E}[r(s,a) + \gamma \max_{a'}[Q_*(s',a')] - Q_n(s,a)]] \quad \forall s, a \\ Q_{n+1} &= Q_n + \alpha_n[r(s_n, a_n) + \gamma \max_{a'} Q_*(s'_n, a') - Q_n(s_n, a_n)]e_{s_n,a_n} \end{aligned}$$

- **Behavior Policy**: Fixed Behavior Policy π_b

$$\begin{aligned} a_n &\sim \pi_b(\cdot \mid s_n) \\ s_n &\sim d_{\pi_b} \end{aligned}$$

- **Experience Replay Buffer**: Store (s, a, s') in a buffer, and sample randomly, used to avoid correlation between samples.

- If the max wasn't there and $a' \sim \pi_b(\cdot \mid s_n)$, we would have written as:

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha_n[b - AQ_n + M_{n+1}] \\ b &= \mathbb{E}[r(s_n, a_n)e(s_n, a_n)] \\ A &= D_{SA \times SA}(s,a) = d_{\pi_b}(s)\pi_b(a \mid s) \\ A^{-1}b &= Q_{\pi_b} \end{aligned}$$

- But because of the max, the update rule is non-linear

- **Question**: Are the iterates stable?

$$\text{bounded} \quad \sup_n \|Q_n\| < \infty \quad \text{a.s.}$$

$$Q_{n+1}(s,a) = \begin{cases} Q_n(s,a) & \forall (s,a) \neq (s_n, a_n) \\ (1-\alpha_n)Q_n(s,a) + \alpha_n[r(s,a) + \gamma \max_{a'} Q_n(s',a')] & \forall (s,a) = (s_n, a_n) \text{ and } s' = s'_n \end{cases}$$

Goal: To show that if $\|Q_n\|_\infty < C$, then $\|Q_{n+1}\|_\infty \leq C$
$\forall (s,a) = (s_n, a_n)$ and $(s' = s'_n)$
Using triangle inequality

$$\begin{aligned} |Q_{n+1}(s,a)| &\leq (1-\alpha_n)|Q_n(s,a)| + \alpha_n[|r(s,a)| + \gamma \max_{a'} |Q_n(s',a')|] \\ &\leq (1-\alpha_n)C + \alpha_n[R_{\max} + \gamma C] \leq C \end{aligned}$$

$$\begin{aligned} \alpha_n(\gamma - 1)C + \alpha_n R_{\max} + C &\leq C \\ \alpha_n[(\gamma-1)C + R_{\max}] &\leq 0 \\ R_{\max} &\leq (1-\gamma)C \\ C &\geq \frac{R_{\max}}{1-\gamma} \\ \text{So,} \quad C &= \max\{\|Q_0\|_\infty, \frac{R_{\max}}{1-\gamma}\} \end{aligned}$$

Then induction follows and hence

$$\|Q_n\|_\infty \leq C \quad \forall n$$

This works in the tabular setting

- Extra - Sutton's phrase: "Deadly Triad"
  1. Function Approximation
  2. TD Learning
  3. Off-Policy Learning
- **Tabular Q-learning (switched ODE Theory)**: Refer `https://arxiv.org/abs/1912.02270`

$$Q_{n+1} = Q_n + \alpha_n[r(s_n, a_n) + \gamma \max_{a'} Q_n(s'_n, a')]e_{s_n,a_n}$$

$$Q_n, e_{s_n,a_n} \in \mathbb{R}^{|S||A|} \quad \sup_{n\geq 0} \|Q_n\|_\infty < \infty \quad \text{Convergence of Q-Learning}$$

- Formal Description of Q-learning

$$\text{Let } \delta_n = r(s_n, a_n) + \gamma \max_{a'} Q_n(s'_n, a^l) - Q_n(s_n, a_n)$$
$$\text{And } F_n = \sigma(Q_0, s_0, a_0, r(s_0, a_0), s'_0, \ldots, s_{n-1}, a_{n-1}, r(s_{n-1}, a_{n-1}), s'_{n-1}) \quad Q_n \in F_n$$

$$\begin{aligned}
\mathbb{E}[\delta_n e_{s_n,a_n} \mid F_n] &= \mathbb{E}[r(s_n, a_n)e_{s_n,a_n} \mid F_n] + \mathbb{E}[\gamma \max_{a'} Q_n(s'_n, a') - Q_n(s_n, a_n)e_{s_n,a_n} | F_n] \\
&= \mathbb{E}[r(s_n, a_n)e_{s_n,a_n}] + \text{ Second term } \to B \\
&= \sum_{s,a} r(s, a)e_{s,a}d_{\pi_b}(s)\pi_b(a \mid s) + B \\
&= (D_{\pi_b})_{|S||A|\times|S||A|}(r)_{|S||A|\times 1} + B \\
&\qquad (D_{\pi_b})_{(s,a)(s,a)} = d_{\pi_b}(s)\pi_b(a \mid s)
\end{aligned}$$

Need to handle second term (B) now

$$\begin{aligned}
B &= \sum_{s,a,s'} d_{\pi_b}(s)\pi_b(a|s)P(s'|s, a)[\gamma \max_{a'} Q_n(s', a') - Q_n(s, a)]e_{s,a} \\
&= \gamma(D_{\pi_b})_{SA\times SA}(P)_{SA\times S}(\Pi_{Q_n})_{S\times SA}(Q_n)_{SA\times 1} - (D_{\pi_b})_{SA\times SA}(Q_n)_{SA\times 1}
\end{aligned}$$

where

$$\Pi_{Q_n} = \begin{cases} 1 & \text{if } s' = s \text{ and } Q(s', a') = \max_a Q(s, a) \\ 0 & \text{otherwise} \end{cases}$$

Every row will have a single 1 (with a tie-breaking rule in case of ties in max values)

$$Q_{n+1} = Q_n + \alpha_n[D_{\pi_b}r + \gamma D_{\pi_b}P\pi_{Q_n}Q_n - D_{\pi_b}Q_n + M_{n+1}]$$

where

$$M_{n+1} = \delta_n e_{s_n,a_n} - [D_{\pi_b}r + \gamma D_{\pi_b}P\Pi_{Q_n}Q_n - D_{\pi_b}Q_n]$$
$$\mathbb{E}[M_{n+1} \mid F_n] = 0$$

- What can we say about the noiseless part?

$$\dot{Q}(t) = D_{\pi_b}r + \gamma D_{\pi_b}P\pi_{Q(t)}Q(t) - D_{\pi_b}Q(t)$$

Let

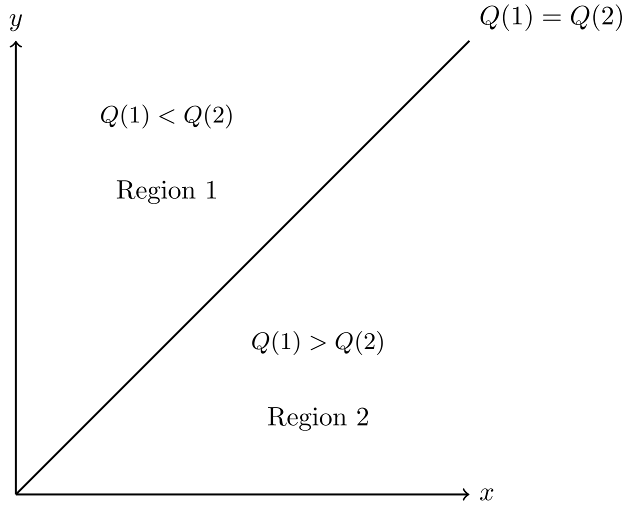
$$x(t) = Q(t) - Q_* \implies \dot{x}(t) = \dot{Q}(t)$$
$$\underbrace{TQ_*}_{(s,a)^{th}\text{ coordinate is}} = Q_*$$
$$r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) \max_{a'} Q_*(s', a')$$
$$r + \gamma P\pi_{Q_*}Q_* = Q_*$$
$$D_{\pi_b}r + \gamma D_{\pi_b}P\Pi_{Q_*}Q_* = D_{\pi_b}Q_*$$

$$\begin{aligned}
\dot{x}(t) &= D_{\pi_b}Q_* - \gamma D_{\pi_b}\Pi_{Q_*}Q_* + \gamma D_{\pi_b}P\Pi_{Q(t)}Q(t) - D_{\pi_b}Q(t) \\
&= [\gamma D_{\pi_b}P\Pi_{Q(t)} - D_{\pi_b}]x(t) + \gamma D_{\pi_b}P[\Pi_{Q(t)} - \Pi_{Q_*}]Q_* \\
\dot{x}(t) &= A_{\sigma(x(t))}x(t) + b_{\sigma(x(t))}
\end{aligned}$$

Because of max operation, this is not exactly affine, "Switched Affine".

- Example with $S = 1, A = 2$:

$$Q = \begin{pmatrix} Q(1) \\ Q(2) \end{pmatrix}$$



$$\dot{x}(t) = \begin{cases} A_1 x(t) + b_1 x(t) & \text{if } x(t) \in R_1 \\ A_2 x(t) + b_2 x(t) & \text{if } x(t) \in R_2 \end{cases}$$

Question: What happens at the boundary?
Answer: Continuity (Exercise)
Generalizing this idea: Affine dynamics with different A and b in different regions

- **Lemma 3 (from paper)**: Switched Linear System. $\dot{x}(t) = A_{\sigma(x(t))} x(t)$ 1, $\sigma : \mathbb{R}^{SA} \to M$

  The origin is the globally asymptotically stable equilibrium of 1 under arbitrary switchings, $\sigma_t \in M$, if and only if $\exists$ a full column rank matrix L of size $m \times d$ with $m \geq d$ and a family of matrices $\bar{A}_\sigma, \sigma \in \mathcal{M}$ that satisfy the strictly negative row dominating diagonal condition.
  Extra: Lyapunov function, $V(x) = \sum_{i=1}^{d} |x(i)|$, i.e.,

$$[\bar{A}_\sigma]_{ii} + \sum_{j \neq i} |[\bar{A}_\sigma]_{ij}| < 0 \quad \forall i$$

$$\text{and } (L)_{m \times d} (A_\sigma)_{d \times d} = (\bar{A}_\sigma)_{m \times m} (L)_{m \times d} \quad \text{Check correctness of dimensions}$$

- Example:

$$\dot{x}(t) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -5 \end{bmatrix} x(t)$$

$$\dot{x}(t) = \begin{bmatrix} -3 & 0 & 0 \\ 0 & -7 & 0 \\ 0 & 0 & -1 \end{bmatrix} x(t)$$

  Pull towards origin at different rates, but origin is equilibrium.

- **Analysis of Tabular Q-earning**
  Stability: $\sup_{n \geq 0} \|Q_n\| \leq C$

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha_n [(D_{\pi_b})_{SA \times SA} (r + \gamma \Pi_u Q_n - Q_n) + M_{n+1}] \\ (D_{\pi_b})_{s,a} &= d_{\pi_b}(s) \Pi_b(a|s) \\ Q_{n+1} &= Q_n + \alpha_n [D_{\pi_b}(TQ_n - Q_n) + M_{n+1}] \end{aligned}$$

$$
\begin{aligned}
\dot{Q}(t) &= D_{\pi_b}(TQ(t) - Q(t)) \\
x(t) &= Q(t) - Q_* \\
\dot{x}(t) &= D_{\pi_b}(TQ(t) - Q(t)) = D_{\pi_b}(TQ(t) - TQ_* + Q_* - Q(t)) \\
TQ &= r + \gamma P \Pi_Q Q \\
TQ(s,a) &= r(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a') \\
TQ_* &= r + \gamma P \Pi_{Q_*} Q_* = Q_* \\
TQ_*(s,a) &= r(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q_*(s',a') \\
TQ - TQ_* &= \gamma P \Pi_Q Q - \gamma P \Pi_{Q_*} Q_* \\
\dot{x}(t) &= D_{\pi_b}(\gamma P \Pi_{Q(t)} Q(t) - \gamma P \Pi_{Q_*} Q_* - (Q(t) - Q_*)) \\
&= D_{\pi_b}(\gamma P \Pi_{Q(t)} Q(t) x(t) + \gamma P(\Pi_{Q(t)} - \Pi_{Q_*}) Q_* - x(t)) \\
&= D_{\pi_b}(\gamma P \Pi_{Q(t)} - I) x(t) + \gamma D_{\pi_b} P (\Pi_{Q(t)} - \Pi_{Q_*}) Q_*
\end{aligned}
$$

Let $\mu : S \to A$ be a deterministic policy

$$R_\mu = \{Q : \mu \text{ is greedy w.r.t. } Q\}$$

- **Claim**: R_μ is a cone
  i.e. $Q \in R_\mu \implies cQ \in R_\mu \quad \forall c > 0$
  Q is a vector in $\mathbb{R}^{SA}$ space
  This leads to piecewise linear dynamics
  Depending on which region $Q(t)$ is, the matrix

  $$D_{\pi_b}(\gamma P \Pi_{Q(t)} - I) \text{ is different}$$

  So, we need Switched Dynamics

- **Lemma 2 (Vector comparison Principle)**: Suppose $\bar{f}, f : \mathbb{R}^d \to \mathbb{R}^d$ be global Lipschitz functions. Let $\bar{f}$ be quasi-monotone, i.e., if $x \le y$ with $x(i) = y(i)$ for at least one i, then $\bar{f}_i(x) \le \bar{f}_i(y)$ for all such i.
  Further, suppose $f(x) \le \bar{f}(x)$
  If $u(t)$ is a solution to $\dot{x}(t) = \bar{f}(x(t))$
  and $l(t)$ is a solution to $\dot{x}(t) = \underline{f}(x(t))$
  and $l(0) \le u(0)$,
  then $l(t) \le u(t) \quad \forall t \ge 0$

  $$\text{Let } \underline{f}(x) = D_{\pi_b}(\gamma P \Pi_{\underbrace{x + Q_*}_{Q(t)}} - I)x + \gamma D_{\pi_b} P(\Pi_{x+Q_*} - \Pi_{Q_*})Q_*$$

  $$\text{and } \bar{f}(x) = D_{\pi_b}(\gamma P \Pi_x - I)x$$

  Consider $\underline{f}$

  $$
  \begin{aligned}
  (\Pi_{x+Q_*} - \Pi_{Q_*})Q_* &\le 0 \\
  a'(s) &= \arg\max_a [x(s,a) + Q_*(s,a)] \\
  \Pi_{x+Q_*} Q_*(s) &= Q_*(s, a'(s)) - \max_a Q_*(s,a) \\
  \implies \bar{f}(x) &\le D_{\pi_b}(\gamma P \Pi_{x+Q_*} - I)x \\
  &\le D_{\pi_b}(\gamma P \Pi_x - I)x = \bar{f}(x)
  \end{aligned}
  $$

  The linear system now

  $$\dot{x}(t) = \underbrace{D_{\pi_b}(\gamma P \Pi_{x(t)} - I)x(t)}_{A\sigma(x(t))} \quad \sigma : \mathbb{R}^{SA} \to \{1, \ldots, M\}$$

  GASE: Globally Asymptotically stable Equilibrium
  The origin is GASE
  $LA_i = \bar{A}_i L$ and $\bar{A}_i$ were negative diagonal dominant

- **Claim**: A_i is negative diagonal dominant. Hence, $L = I$ works
  Proof: $[A_i]_{(s,a)\times(s,a)} + \sum_{(s',a') \ne (s,a)} |[A_i]_{(s,a),(s',a')}| < 0$
  Let $\Pi_x = \Pi_u$ for some u:

  $$\implies \Pi_x(s, (s', a')) = \begin{cases} 1 & \text{if } s' = s \text{ and } a = u(s) \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Diagonal } (\gamma P \Pi_x)_{(s,a),(s,a)} = \gamma \sum_{s'} P(s'|s,a)\Pi_u(s,a|s')$$

$$\text{Off-diagonal } (\gamma P \Pi_x)_{(s,a),(s",a")} = \gamma \sum_{s",a"} \sum_{s'} P(s'|s,a)\Pi_u(s",a"|s')$$

$$\text{Sum} = \gamma \sum_{s'} P(s'|s,a) \underbrace{[\Pi_u(s,a|s') + \sum_{s",a"} \Pi_u(s",a"|s')]}_{\text{Should come out to 1}} = \gamma$$

We can conclude that solution of $\dot{x}(t) = \underline{f}(x(t))$ will be suitably upper bounded by solution of $\dot{x}(t) = \bar{g}(x(t))$, every solution of which goes to origin

- Paper also shows a lower bound (which also goes to origin)
- So, noiseless ODE for Q-earning also goes to origin ($x(t) \to 0$)
- i.e. $Q(t) - Q_* \to 0$
- **Linear ODE**: "noiseless" variant of the algorithm
- **Switched ODE Perspective**: We took the ODE and an upper comparison ODE (and Lower comparison ODE) and used vector comparison lemma to get upper (and lower bounds)
- **Q-Learning with Function Approximation**: In FA, we have, $\Pi_\mu T_\mu$(Projection operator). (Presume we know the policy, we will see if this carries over to Q-Learning)

  $$\Pi = \Phi(\Phi^\top D_\mu \Phi)^{-1} D_\mu$$

  This is a γ-contraction (ΠT_μ)

  $$\|\Pi\| \leq 1 \implies \|\Pi J\|_{D_\mu} \leq \|J\|_{D_\mu} \quad \text{(Non-expansive Property)}$$
  $$\|T_\mu J - T_\mu J'\|_{D_\mu} \leq \gamma \|J - J'\|_{D_\mu} \quad \text{(contraction as } \gamma < 1)$$

  Combining these, we get

  $$\|\Pi T_\mu J - \Pi T_\mu J'\|_{D_\mu} \leq \|T_\mu J - T_\mu J'\|_{D_\mu} \leq \gamma \|J - J'\|_{D_\mu}$$

  We would ideally like to find fixed point of ΠT, but:

  1. Projection w.r.t. which policy?
  2. Contraction w.r.t. which norm?

  We know T is a contraction w.r.t. $\|\cdot\|_\infty$ norm
  T is contraction w.r.t. both $\|\cdot\|_\infty$ and $\|\cdot\|_{D_\mu}$ norms

- **Policy Iteration**: This needs to compute Q_{μ_k} exactly. Works because of Policy Improvement Lemma:

  $$\mu_{k+1}(s) = \arg\max_a Q_{\mu_k}(s,a)$$
  $$Q_{\mu_{k+1}} \geq Q_{\mu_k}$$

  When they are equal at all coordinates, it satisfies Bellman equations, and we have convergence
  Reference: Bruno Scherrer Paper, approximate $Q_{\mu_0} \approx \Phi\theta_0$

- **Q-Learning with Function Approximation**
- Reference: Deep Q-Learning from DeepMind
- We look at Q-learning with linear function approximation

  $$Q_* \approx \Phi_{|S||A| \times d} (\theta_n)_{d \times 1}$$
  $$\theta_{n+1} = \theta_n + \alpha_n [r(s_n,a_n) + \gamma \max_{a'} \Phi^\top(s'_n, a')\theta_n - \Phi^\top(s_n,a_n)\theta_n]\Phi(s_n,a_n)$$

- Instead of working with fixed behavior policy, we work with adaptive behavior policy, $\Phi\theta_n \approx Q_*$
  Policy that is greedy w.r.t. $\Phi\theta_n$

  $$\mu_n(s) = \arg\max_a \Phi^\top(s,a)\theta_n$$

  But this can get stuck
  Instead of greedy, we take policy that is ϵ-greedy w.r.t. $\Phi\theta_n$:

  $$\mu_n(s) = \begin{cases} \text{random action} & \text{w.p. } \epsilon \\ \arg\max_a \Phi^\top(s,a)\theta_n & \text{w.p. } 1-\epsilon \end{cases}$$

  ϵ-greedy behavior policy, How to sample this policy?

- **Experience Replay Buffer**: Experience defined as (s_n, a_n, s'_n). Store experiences in buffer and sample uniformly from the buffer to update the equation ($\theta_{n+1} = \theta_n + \dots$)
It empirically works. Why? $\rightarrow$ We don't know.

- **Idealized Replay Buffer**:

$$\begin{aligned} s_n &\sim d_{\mu_n} \\ a_n &\sim \mu_n(\cdot \mid s_n) \\ s'_n &\sim P(\cdot \mid s_n, a_n) \end{aligned}$$

s_n sampled from stationary distribution of $\Phi\theta_n$

$$\theta_{n+1} = \theta_n + \alpha_n[b_n - A_n\theta_n + M_{n+1}]$$

$\bar{a} : S \rightarrow A$

$$\mathcal{R}_{\bar{a}} = \{\theta \in \mathbb{R}^d ; \bar{a} \text{ is greedy w.r.t. } \Phi\theta_n\}$$

$\mathcal{R}_{\bar{a}}$ is a cone

$$\theta_{n+1} = \theta_n + \alpha_n[\sum_{\bar{a}}(b_{\bar{a}} - A_{\bar{a}}\theta_n)\mathbb{1}_{\{\theta_n \in R_{\bar{a}}\}} + M_{n+1}]$$

- In rare cases, DQN may converge to worst-case policies.

5.4 Policy Gradient Methods

- Here also we will use Function Approximation

- FA is used for **Parameterizing Policies**, $\theta \in \mathbb{R}^d$
Consider this example

$$\pi_\theta(a \mid s) = \frac{e^{\Phi^\top(s,a)\theta}}{\sum_{a'} e^{\Phi^\top(s,a')\theta}} \quad \text{where } \Phi^\top(s,a) \in \mathbb{R}^{1\times d} \text{ and } \theta \in \mathbb{R}^{d \times 1}$$

In general:

$$\pi_\theta(a \mid s) = \frac{e^{h(s,a,\theta)}}{\sum_{a'} e^{h(s,a',\theta)}}$$

where $h(s, a, \theta)$ gives a scalar

- **Gradient**: $J(\theta) = V_{\pi_\theta}(s_0)$
V_π is a vector, so, we consider $V_\pi(s_0)$

$$\begin{aligned} V_\pi(s) &= \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s] \\ a_t &\sim \pi(\cdot \mid s_t) \\ s_{t+1} &\sim P(\cdot \mid s_t, a_t) \end{aligned}$$

Goal: Solve $\max_\theta J(\theta)$
Solution: Gradient Ascent, but need $\nabla J(\theta)$
Then, $\theta_{n+1} = \theta_n + \alpha_n \nabla J(\theta_n)$

- **Policy Gradient Theorem**
For Episodic tasks, we have

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) \mid s_0 = s]$$

$$T = \inf\{t \geq 0 : s_t \in \{\text{terminal states}\}\}$$

$$J(\theta) = V_{\pi_\theta}(s_0) \implies \nabla J(\theta) = \nabla V_{\pi_\theta}(s_0)$$

$$\text{Use the relation } V_\pi(s) = \sum_a \pi(a \mid s) Q_\pi(s,a)$$

$$\begin{aligned} \nabla V_{\pi_\theta}(s) &= \nabla[\sum_a \pi_\theta(a \mid s) Q_{\pi_\theta}(s,a)] \\ &= \sum_a \nabla[\pi_\theta(a \mid s) Q_{\pi_\theta}(s,a)] \\ &= \sum_a [\nabla \pi_\theta(a \mid s) Q_{\pi_\theta}(s,a) + \pi_\theta(a \mid s) \underbrace{\nabla Q_{\pi_\theta}(s,a)}_{\text{hard part}}] \end{aligned}$$

How will we do this?
Suppose $\gamma = 1$

$$Q_{\pi_\theta}(s,a) = r(s,a) + \sum_{s'} P(s' \mid s,a) V_{\pi_\theta}(s')$$

Only $V_{\pi_\theta}(s')$ this depends on θ

$$\nabla Q_{\pi_\theta}(s,a) = 0 + \sum_{s'} p(s' \mid s,a) \nabla V_{\pi_\theta}(s')$$

Combine with above, we get

$$\begin{aligned}
\nabla V_{\pi_\theta}(s) &= \sum_a [\nabla \pi_\theta(a \mid s) Q_{\pi_\theta}(s,a) + \pi_\theta(a \mid s) \sum_{s'} P(s' \mid s,a) \qquad \underbrace{\nabla V_{\pi_\theta}(s')}_{\text{Substitute formula here again}} \qquad] \\
&= \sum_a \nabla \pi_\theta(a \mid s) Q_{\pi_\theta}(s,a) + \sum_{a,s'} \pi_\theta(a|s) P(s'|s,a) \Bigg[\sum_{a'} \nabla \pi_\theta(a'|s') Q_{\pi_\theta}(s',a') + \\
&\qquad \sum_{a',s"} \pi_\theta(a'|s') P(s^{"}|s',a') \nabla V_{\pi_\theta}(s^{"}) \Bigg]
\end{aligned}$$

If we keep doing this recurrence. Look at:

$$\sum_{a,s'} \pi_\theta(a \mid s) P(s' \mid s,a) = \sum_{s'} P\{s_1 = s' \mid s_0 = s, \theta\}$$

Putting this in above equations, we get

$$= \sum_{s'} P\{s_1 = s' | s_0 = s, \theta\} \left(\sum_{a'} \nabla \pi_\theta(a'|s) Q_{\pi_\theta}(s',a') + \sum_{s",a'} P\{s_2 = s^{"} | s_1 = s', \theta\} + \nabla V_{\pi_\theta}(s^{"}) \right)$$

On solving, we get

$$\nabla V_{\pi_\theta}(s) = \sum_{k=0}^{\infty} \sum_{s'} \mathbb{P}\{s_k = s', k \le T \mid s_0 = s, \theta\} \sum_{a'} \nabla \pi_\theta(a' \mid s') Q_{\pi_\theta}(s',a')$$

This only has $\nabla \pi_\theta$
Can we write this in a simple form?
Cannot blindly switch summation order
We can switch if we ensure the term is absolutely summable
Justification for switching the order of summation

$$\begin{aligned}
Q_{\pi_\theta}(s,a) &= \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t,a_t) | s_0 = s, a_0 = a \right] \\
|Q_{\pi_\theta}(s,a)| &\le \mathbb{E}\left[\sum_{t=0}^{T-1} |r(s_t,a_t)| \Bigg| s_0 = s, a = a_0 \right] \\
\text{Suppose } |r(s,a)| \le r_{max} \\
&\le r_{max} \mathbb{E}[T | s_0 = s, a_0 = a]
\end{aligned}$$

where $\mathbb{E}[T|s_0 = s, a_0 = a]$ is the expected length of trajectory starting from state s_0, taking action a_0.

Now, using this, we can show that the series is convergent

$$\left\| \sum_{k=0}^{\infty} \sum_{s'} P\{s_k = s', k \leq T | s_0 = s, \theta\} \sum_{a'} \nabla \pi_\theta(a'|s') Q_{\pi_\theta}(s', a') \right\|$$

$$\leq \sum_{k=0}^{\infty} \sum_{s'} P\{s_k = s', k \leq T | s_0 = s, \theta\} \sum_{a'} \overbrace{\pi_\theta(a'|s') \underbrace{\|\nabla \ln \pi_\theta(a'|s')\|}_{\leq C}}^{\text{Log Trick}} \underbrace{|Q_{\pi_\theta}(s', a')|}_{\leq r_{max}\mathbb{E}[T|s',a',\theta]}$$

$$\leq C \sum_{k=0}^{\infty} \sum_{s}^{\prime} P(s_k = s', T > k | s_0 = s, \theta) \left(\sum_{a'} \pi_\theta(a', s') |Q_{\pi_\theta}(s', a')| \right)$$

$$\leq C r_{max} \sum_{k=0}^{\infty} \sum_{s'} P(s_k = s', T > k | s_0 = s, \theta) \sum_{a'} \pi_\theta(a'|s') \mathbb{E}[T|s', a', \theta]$$

$$\leq C r_{max} \sum_{k=0}^{\infty} \sum_{s'} P(s_k = s', T > k | s_0 = s, \theta) \mathbb{E}[T|s', \theta]$$

Change order of summation $\rightarrow$ doable as everything here is non-negative

$$= C r_{max} \sum_{s'} \sum_{k=0}^{\infty} P(s_k = s', T > k | s_0 = s, \theta) \mathbb{E}[T|s', \theta]$$

$$= C r_{max} \sum_{s'} \mathbb{E} \left[\sum_{k=0}^{\infty} \mathbb{1}_{s_k = s', T > k} | s_0 = s, \theta \right] \mathbb{E}[T|s', \theta]$$

Changing summation in original,

$$= \sum_{s'} \underbrace{\sum_{k=0}^{\infty} \mathbb{P}\{s_k = s', k \leq T \mid s_0 = s, \theta\}}_{\eta(s')} \underbrace{\sum_{a'} \nabla \pi_\theta(a' \mid s') Q_{\pi_\theta}(s', a')}_{\text{doesn't depend on k}}$$

$$= \sum_{s'} \mathbb{E} \left[\sum_{k=0}^{\infty} \mathbb{1}_{s_k = s', k \leq T} | s_0 = s, \theta \right] \sum_{a'} \nabla \pi_\theta(a' \mid s') Q_{\pi_\theta}(s', a')$$

Expectation of indicator is its probability $\rightarrow$ Expected number of times s' is visited

$$= \sum_{s'} \eta(s') \sum_{a'} \nabla \pi_\theta(a' \mid s') Q_{\pi_\theta}(s', a')$$

Scale η to make a distribution

$$= (\sum_{j} \eta(j)) \sum_{s'} \left(\frac{\eta(s')}{\sum_j \eta(j)} \right) \sum_{a'} \nabla \pi_\theta(a' \mid s') Q_{\pi_\theta}(s', a')$$

$$= (\sum_{j} \eta(j)) \sum_{s'} \mu(s') \sum_{a'} \nabla \pi_\theta(a' \mid s') Q_{\pi_\theta}(s', a')$$

Log derivative trick. Need to ensure all $\pi_\theta(a' \mid s')$ are positive in parameterization

$$= (\sum_{j} \eta(j)) \sum_{s', a'} \underbrace{\mu(s') \pi_\theta(a' \mid s')}_{\text{distribution over states and actions}} \nabla \ln \pi_\theta(a' \mid s') Q_{\pi_\theta}(s', a')$$

$$= (\sum_{j} \eta(j)) \mathbb{E}_{s', a'} \nabla \ln \pi_\theta(a' \mid s') Q_{\pi_\theta}(s', a')$$

Now, summation over η can be written as

$$\sum_{s'} \eta(s') = \text{Expected number of visits to } s'$$

$$= \text{Expected length of trajectory}$$

$$= \mathbb{E}[T \mid s_0 = s, \theta]$$

Finally, we have

$$\nabla J(\theta) = \nabla V_{\pi_\theta}(s_0) \propto \mathbb{E}_{s', a'} \nabla \ln \pi_\theta(a' \mid s') Q_{\pi_\theta}(s', a')$$

- We can use this derivative using **REINFORCE** algorithm
  Start with some $\theta = \theta_0$

The we loop the following forever
Simulate a trajectory using π_θ to get

$$s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{T-1}, a_{T-1}, r_{T-1}$$

For $k = 0, \ldots, T-1$ do

$$G_k = \sum_{t=k}^{T-1} r_t$$
$$\theta \leftarrow \theta + \alpha_k [G_k \nabla \ln \pi_\theta(a_k|s_k)]$$

- What about $\sum_j \eta(j)$?
  Assumed to be implicitly present in choice of α_k
- $a_k \sim \pi_\theta(\cdot|s_k)$, but what about s_k, is it sampled correctly?
  Number of times s' is visited in given batch of trajectories $\propto \mu(s')$
- This need episode-level information, whereas in TD algorithm we just looked at (s, a, s')
- Using this in infinite horizon setting is tricky $\implies$ Actor-Critic
- **Actor-Critic Methods**
  Actor: Improve/Update the policy
  Actor uses gradient ascent to improve
  Critic: Evaluate the policy's value function, given π_θ, get Q_{π_θ}
  Critic tries to implement some kind of TD type algorithm to get Q_{π_θ}
- **Trust Region Policy Optimization(TRPO)**

$$\eta(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$$

Given π, can we get a better policy?

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_s d_{\tilde{\pi}}(s) \tilde{\pi}(a|s) A_\pi(s, a)$$

where $A_\pi(s, a)$ is the advantage function, i.e.

$$d_{\tilde{\pi}}(s) = (1-\gamma) \sum_{k=0}^{\infty} \gamma^k P_{\tilde{\pi}}\{s_k = s\} \leftarrow \text{Discounted state-visitation distribution}$$
$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$$

Computing $d_{\tilde{\pi}}(s)$ is difficult. So, we find an alternative form

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_{s,a} d_\pi(s) \tilde{\pi}(a|s) A_\pi(s, a)$$

One can show that

$$\eta(\tilde{\pi}) \Big|_{\tilde{\pi}=\pi} = L_\pi(\tilde{\pi}) \Big|_{\tilde{\pi}=\pi}$$
$$\nabla L_\pi(\tilde{\pi}) \Big|_{\tilde{\pi}=\pi} = \nabla \eta(\tilde{\pi}) \Big|_{\tilde{\pi}=\pi}$$

$L_\pi(\tilde{\pi})$ gives a first order approximation to $\eta(\tilde{\pi})$, only in the neighborhood of π