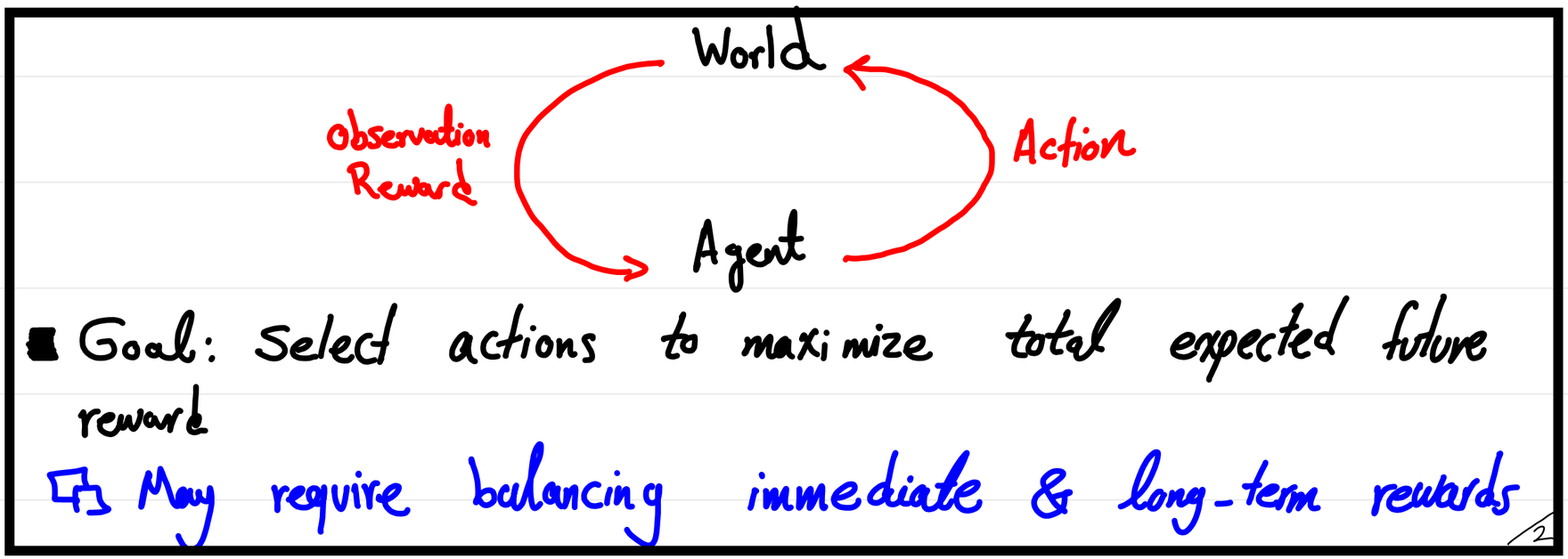
Week 11 - Part 3

- Outcome of this part
  - Return / Utility
  - Policies
  - Finding the "best" policy (i.e., solving the MDP)
    - Solving linear system of equations
    - Bellman's equations ← Next week

Returns & Episodes

- Recall:



- Goal: Select actions to maximize total expected future reward
  - ↳ May require balancing immediate & long-term rewards

- The agent's goal is to: maximize the expected return
  - ↳ Return is a function of the reward sequence

$$r_0, r_1, r_2, \ldots, r_T$$

Return

- In simplest case, the return is the sum of the rewards
  - i.e., $r_0 + r_1 + \cdots + r_T$

- Such return definition makes sense in applications with finite time steps.

- However, in applications with possibly infinite time step, i.e. $T = \infty$, the above definition of return is problematic
  - as the return can be ∞.

Return as Sum of Discounted Rewards

- That's why we use a more complex definition for return

$$U([r_0, r_1, r_2, \cdots]) = r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots = \sum_{k=0}^{\infty} \gamma^k r_k$$

This return is known a "sum of discounted rewards"

$0 \leq \gamma \leq 1$ is the discount factor

- If all $|r_i|$'s are bounded by α, Then $U([r_0, r_1, \cdots]) < \alpha \frac{1}{1-\gamma}$

- The discount parameter determines the present value of the feature rewards:
  - A reward received K time step in feature is worth only γ^K times what would be worth if it were received immediately.
- As γ approaches 1, the return objective takes into account the feature rewards more strongly, i.e., the agent becomes more farsighted
- As γ approaches 0, the return objective take into account the immediate reward more strongly, i.e., the agent is more myopic (shortsighted)

Why Discounted Return is a good Model

i) As mentioned earlier, with discount factor of $\gamma < 1$, return is always bounded (i.e., finite) even when we have infinite horizon. That's because $\sum_{k=0}^{\infty} \gamma^k r_k$ is bounded if the sequence $\{r_k\}$ is finite.

Why Discounted Return is a good Model

ii) Intuitively speaking, discount return model conforms both human and animal behaviour.

☐ We care about the reward we will get in future, but we value the immediate reward more

iii) We used return function to model agent's preference over possible outcomes, i.e., reward sequence.

A common preference assumption is stationarity, i.e.,

$$[a_1, a_2, \dots] \succ [b_1, b_2, \dots] \iff [r, a_1, a_2, \dots] \succ [r, b_1, b_2, \dots]$$

There is only one way to model a stationary preference:

$$U([r_0, r_1, \dots]) = r_0 + \gamma r_1 + \gamma^2 r_2 + \dots, \text{ where } 0 \leq \gamma \leq 1.$$

Policies

- A policy π, gives an action for each state
  - $\pi : S \rightarrow A$ ←
- Now that we defined return/utility as the sum of discounted rewards, we can compare policies by comparing the expected utilities obtained when they are followed

- For MDPs, we want to find an optimal policy $\pi^* : S \rightarrow A$
  - An optimal policy π^*, is the policy that maximizes the expected return if followed.
- Note: The above was the definition of deterministic policies.

Stochastic Polices : $\pi(a|s) = \mathbb{P}[a_t = a | s_t = s]$

Example: Erfan the undergrad student

- Recall the example "Erfan the undergrad".
  - $A = \{\underbrace{\text{"party hard"}}_{a_p}, \underbrace{\text{"work hard"}}_{a_w}, \underbrace{\text{"end"}}_{a_e}\}$
  - $S = \{\underbrace{\text{"happy"}}_{s_h}, \underbrace{\text{"tired"}}_{s_t}, \underbrace{\text{"burnt-out"}}_{s_b}\}$
- List all possible policies for the above MDP.
  - Note that the only possible action at s_b is a_e.

$\pi_1(s_h) = a_p$	$\pi_2(s_h) = a_p$	$\pi_3(s_h) = a_w$	$\pi_4(s_h) = a_w$
$\pi_1(s_t) = a_p$	$\pi_2(s_t) = a_w$	$\pi_3(s_t) = a_p$	$\pi_4(s_t) = a_w$
$\pi_1(s_b) = a_e$	$\pi_2(s_b) = a_e$	$\pi_3(s_b) = a_e$	$\pi_4(s_1) = a_e$

Expected Return

- Following a policy yields a random episode
- The utility of a policy is the sum of discounted rewards of the episode
  - This utility is a random variable

- Let's study a simple example:

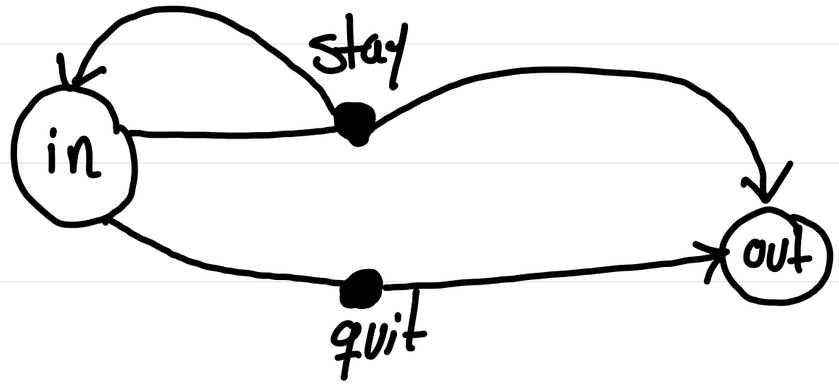
Erfan the gambler

(See the next page for the description)

Example: Erfan the Gambler

- For each round $r = 1, 2, \ldots$
  - You choose stay or quit
  - If quit, you get \$10 and we end the game
  - If stay, you get \$4 and I roll a 6-sided fair dice
    - If the dice results in 1 or 2, we end the game
    - Otherwise, continue to the next round.

- Let's draw the transition graph.



- Let the discount factor be $\gamma = 1$.
- Here are some possible episodes if Erfan's policy is to stay:

Episode	Return
[(in, stay, 4), out]	4
[(in, stay, 4), (in, stay, 4), (in, stay, 4), out]	12
[(in, stay, 4), (in, stay, 4), out]	8
⋮	⋮

- Let's formalize what we saw in the previous example for general MDP
- The value (utility) of policy π at state s, denoted by $V_\pi(s)$, is the expected return if starting at s and following policy π
- Starting at state "s" and following policy "π" can yield to different sequences of (state, action, reward), e.g.,
  - $(s, \pi(s), R(s,\pi(s))), (s', \pi(s'), R(s',\pi(s'))), (s'', \pi(s''), R(s'',\pi(s'')), \ldots$
  - $(s, \pi(s), R(s,\pi(s))), (s', \pi(s'), R(s',\pi(s'))), (s''', \pi(s'''), R(s''',\pi(s''')), \ldots$
  - $(s, \pi(s), R(s,\pi(s))), (s''', \pi(s'''), R(s''',\pi(s''))), (s', \pi(s'), R(s',\pi(s'))), \ldots$
  - $(s, \pi(s), R(s,\pi(s))), (s'', \pi(s''), R(s'',\pi(s''))), (s', \pi(s'), R(s',\pi(s'))), \ldots$
  - ⋮
- Each one of the possible episodes happens with a probability

Note: in the above MDP, we assumed a reward model of type $R(s,a)$ for simplicity

* Finding the Expected Return

■ How can we formulate $V_\pi(s)$?

$$V_\pi(s) = \mathop{\mathbb{E}}_{(s_1,s_2,\dots)\,|\,s_0=s}\left[R(s_0,\pi(s_0)) + \gamma R(s_1,\pi(s_1)) + \gamma^2 R(s_2,\pi(s_2)) + \dots \,\middle|\, (s_0=s,\pi(s))\right]$$

where $s_1, s_2, \dots$ is a random sequence of states yield by following policy π. So $s_i \in S$ for any $i \geq 1$.

■ What is the probability of observing the sequence $s_1, s_2, s_3, \dots$ Conditioned on $s_0 = s$? $P(s_1 \mid s, \pi(s)) \times P(s_2 \mid s_1, \pi(s_1)) \times \dots$

■ We can rewrite $V_\pi(s)$ as

$$V_\pi(s) = \mathop{\mathbb{E}}_{(s_1,s_2,\dots)\,|\,s_0=s}\left[R(s,\pi(s)) + \gamma\left(R(s_1,\pi(s_1)) + \gamma R(s_2,\pi(s_2)) + \dots\right) \,\middle|\, (s_0=s,\pi(s))\right]$$

$$= R(s,\pi(s)) + \gamma \mathop{\mathbb{E}}_{(s_1,s_2,\dots)\,|\,s_0=s}\left[R(s_1,\pi(s_1)) + \gamma R(s_2,\pi(s_2)) + \dots \,\middle|\, (s_0=s,\pi(s))\right]$$

* Finding the Expected Return

- We can rewrite $V_\pi(s)$ as

$$V_\pi(s) = R(s,\pi(s)) + \gamma \mathop{\mathbb{E}}_{(s_1,s_2,\dots)|s_0=s} \left[R(s_1,\pi(s_1)) + \gamma R(s_2,\pi(s_2)) + \dots \,\middle|\, (s_0=s,\pi(s)) \right]$$

$$\mathop{\mathbb{E}}_{x}[x] = \mathop{\mathbb{E}}_{y}\left[\mathop{\mathbb{E}}_{x|y}[x|y]\right]$$

- **Recall**: Law of total probability $\mathop{\mathbb{E}}_{x}(X) = \sum_i P(A_i) \mathop{\mathbb{E}}_{x|A_i}(X|A_i)$ where $\{A_i\}$ is a finite partition of the sample space.
- **Recall**: Let us rewrite the above recall with an additional condition, so that it is more suitable for us.

$$\mathbb{E}(X|B) = \sum_i P(A_i|B) \mathop{\mathbb{E}}_{x|B,A_i}(X|B,A_i)$$

- Thus, we can rewrite $V_\pi(s)$ as

$$V_\pi(s) = R(s,\pi(s)) + \gamma \sum_{s' \in S} P(s_1=s' \mid s_0=s, \pi(s_0)) \mathop{\mathbb{E}}_{(s_1,s_2,s_3,\dots)|s_1=s', s_0=s} \left[R(s_1,\pi(s_1)) + \gamma R(s_2,\pi(s_2)) + \dots \,\middle|\, (s_0=s,\pi(s_0)), (s_1=s',\pi(s')) \right]$$

* Finding the Expected Return

■ Thus, we can rewrite $V_\pi(s)$ as

$$V_\pi(s) = R(s,\pi(s)) + \gamma \sum_{s' \in S} P(s_1 = s' \mid s_0 = s, \pi(s_0)) \underset{(s_2, s_3, \ldots) \mid s_1 = s'}{\mathbb{E}} \left[R(s_1, \pi(s_1)) + \gamma R(s_2, \pi(s_2)) + \ldots \mid (s_0 = s, \pi(s_0)), (s_1 = s', \pi(s')) \right]$$

$$\Rightarrow V_\pi(s) = R(s,\pi(s)) + \gamma \sum_{s' \in S} T(s, \pi(s), s') \underset{(s_2, s_3, \ldots) \mid s_1 = s'}{\mathbb{E}} \left[R(s', \pi(s')) + \gamma R(s_2, \pi(s_2)) + \ldots \mid (s_1 = s', \pi(s')) \right]$$

$$\Rightarrow V_\pi(s) = R(s,\pi(s)) + \gamma \sum_{s' \in S} T(s, \pi(s), s') \, V_\pi(s')$$

Finding the Expected Return

- We found a recursion for $V_\pi(s)$ under the dynamics model $T(s,a,s')$ and reward model $R(s,a)$

$$V_\pi(s) = R(s,\pi(s)) + \gamma \sum_{s'} T(s,\pi(s),s')\, V_\pi(s')$$

- More generally, we can show that for dynamics model $T(s,a,s')$ and Reward model $R(s,a,s')$

$$V_\pi(s) = \sum_{s' \in S} T(s,\pi(s),s') \left[R(s,\pi(s),s') + \gamma V_\pi(s') \right]$$

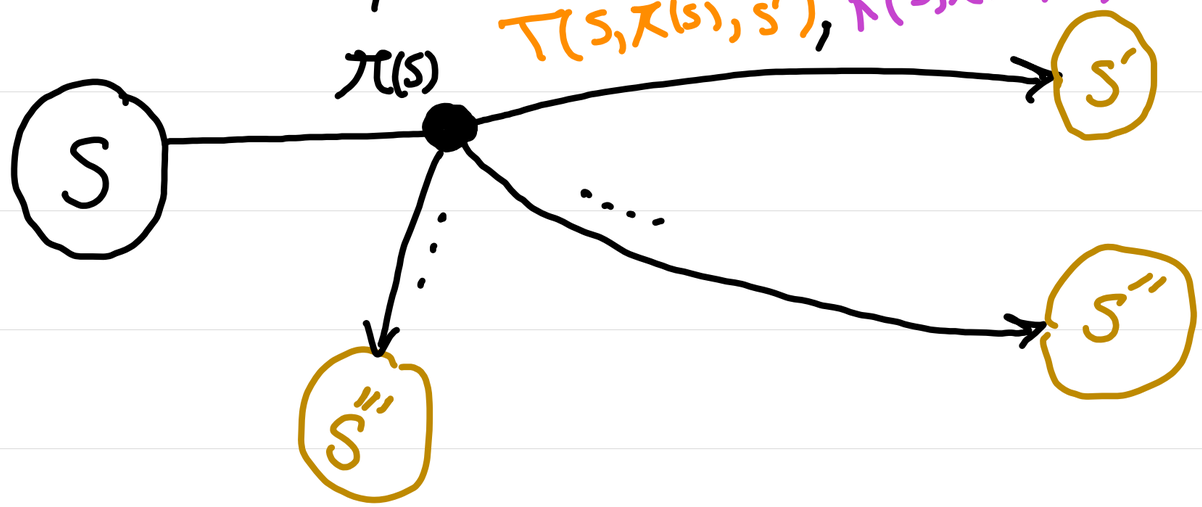
Review

- we presented Markov Decision process and how to model it
  - Transition Model (i.e., $T(s,a,s') \triangleq P(s'|s,a)$)
  - Reward Model (i.e., $R(s,a,s')$)
  - γ, distount factor
  - $V_\pi(s) \triangleq \mathbb{E}_{s_1,s_2,\ldots|s}\left[R(s,\pi(s),s_1) + \gamma R(s_1,\pi(s_1),s_2) + \gamma^2 R(s_2,\pi(s_2),s_3) + \cdots\right]$

- $V_\pi(s) = \sum_{s' \in S} T(s,\pi(s),s')\left[R(s,\pi(s),s') + \gamma V_\pi(s')\right]$

Finding the Expected Return

- An intuitive explanation



$$V_\pi(s) = T(s,\pi(s),s')\left[R(s,\pi(s),s') + \gamma V_\pi(s')\right]$$
$$+ T(s,\pi(s),s'')\left[R(s,\pi(s),s'') + \gamma V_\pi(s'')\right]$$
$$+ T(s,\pi(s),s)\left[R(s,\pi(s),s) + \gamma V_\pi(s)\right]$$

with probability $T(s,\pi(s),s')$, you will receive the immediate reward $R(s,\pi(s),s')$ and go to the next state s'. From state s', your expected utility that you would receive is $V_\pi(s')$. This is the reward you receive in the next step. Hence, you should discount it by γ.

Value and Q-value of a Policy

- We use $V_\pi(s)$ to denote the expected utility received by following policy "π" from state "s".
- We use $Q_\pi(s,a)$ to denote the expected utility of taking action "a" from state "s", and then following policy "π"
  - $V_\pi(s) = \sum_{s' \in S} T(s, \pi(s), s') \left[R(s, \pi(s), s') + \gamma V_\pi(s') \right]$
    - Value of state s in policy π
  - $Q_\pi(s,a)$ ← Q-value of state-action pair (s,a) with policy π.
  - If you had $Q_\pi(\cdot)$, how would you recover $V_\pi(s)$?

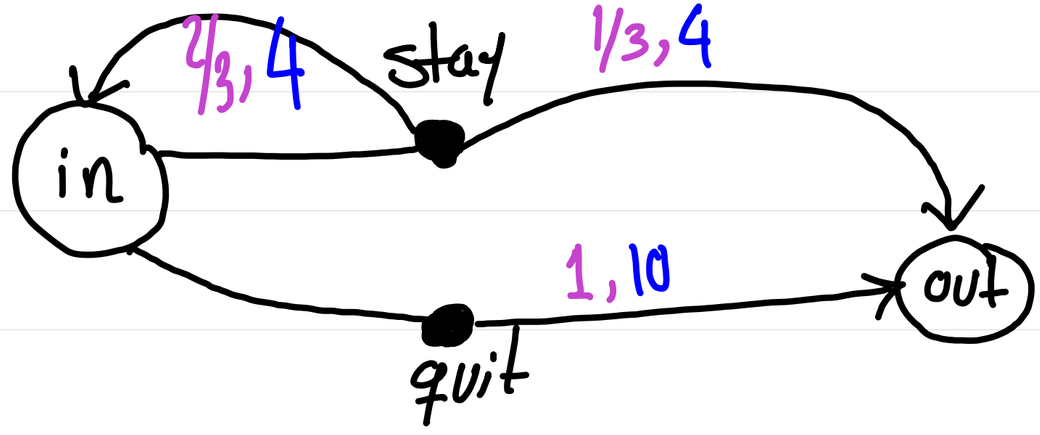
$$V_\pi(s) = Q_\pi(s, \pi(s))$$

  - If you had $V(\cdot)$, how could you recover $Q_\pi(s,a)$

$$Q_\pi(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_\pi(s') \right]$$

Example: Erfan the gambler

- Let's revisit that example and find the value of "stay" policy.



- Let π_1 denote the "stay" policy, i.e., $\pi_1(\text{"in"}) = \text{"stay"}$
  - Note that "out" is a terminal state and usually the value of terminal states is already known.
  - In this problem, when landing in "out" state, the game ends and we receive no more reward.
    - Thus, $V_\pi(\text{"out"}) = 0$, for any policy π.

- Observe that

$$V_{\pi_1}(\text{"in"}) = \sum_{s'} T(\text{"in"}, \pi_1(\text{in}), s')\left[R(\text{in}, \pi_1(\text{in}), s') + \gamma V_{\pi_1}(s')\right]$$

$$= \cancel{T}(\text{"in"}, \text{stay}, \text{"in"})\left[\cancel{R}(\text{"in"}, \text{stay}, \text{"in"}) + \cancel{\gamma} V_{\pi_1}(\text{in})\right]$$

$$+ \cancel{T}(\text{in}, \text{stay}, \text{out})\left[\cancel{R}(\text{in}, \text{stay}, \text{out}) + \cancel{\gamma} \cancel{V_{\pi_1}}(\text{out})\right]$$

$\Rightarrow V_{\pi_1}(\text{in}) = \frac{2}{3}[4 + V_{\pi_1}(\text{in})] + \frac{1}{3}[4 + 0] \Rightarrow V_{\pi_1}(\text{in}) = 12.$

$3V_{\pi_1}(\text{in}) = 2 \times 4 + 2 \times V_{\pi_1}(\text{in}) + 1 \times 4 \Rightarrow 3V_{\pi_2}(\text{in}) - 2V_{\pi_2}(\text{in}) = 12$

- Let's find the valve of "quit" policy.
- Let π_2 denote the "quit" policy, i.e., $\pi_2(\text{"in"}) = \text{"quit"}$.

$$V_{\pi_2}(\text{in}) = \sum_{s'} T(\text{in}, \pi_2(\text{in}), s') \left[R(\text{in}, \pi_2(\text{in}), s') + \gamma V_{\pi_2}(s') \right]$$

$$= T(\text{in}, \text{"quit"}, \text{out}) \left[10 + V_{\pi_2}(\text{out}) \right] = 10$$

(1 and 0 annotated above the two terms)

- Which policy do you think is better?
  - $V_{\pi_1}(\text{in}) > V_{\pi_2}(\text{in})$. Hence, π_1 is better.

Next Lecture: Policy evaluation

- We could find the value of the states in our previous example easily.
  - we just had one unknown state. Easy Peasy!

- What about larger problems? $S = \{s, s', s'', s''', \dots\}$

$$V_\pi(s) = T(s,\pi(s),s)\left[R(s,\pi(s),s)+\gamma V_\pi(s)\right] + T(s,\pi(s),s')\left[R(s,\pi(s),s')+\gamma V_\pi(s')\right]+\dots$$

$$V_\pi(s') = T(s',\pi(s'),s)\left[R(s',\pi(s'),s)+\gamma V_\pi(s)\right] + T(s',\pi(s'),s')\left[R(s',\pi(s'),s')+\gamma V_\pi(s')\right]+\dots$$

$$V_\pi(s'') = T(s'',\pi(s'),s)\left[R(s'',\pi(s''),s)+\gamma V_\pi(s)\right] + T(s'',\pi(s''),s')\left[R(s'',\pi(s''),s')+\gamma V_\pi(s')\right]+\dots$$

Next Lecture: Policy evaluation

- What about larger problems? $S = \{s, s', s'', s''', \dots\}$

$$V_\pi(s) = T(s,\pi(s),s)\left[R(s,\pi(s),s)+\gamma V_\pi(s)\right] + T(s,\pi(s),s')\left[R(s,\pi(s),s')+\gamma V_\pi(s')\right]+\dots$$

$$V_\pi(s') = T(s',\pi(s'),s)\left[R(s',\pi(s'),s)+\gamma V_\pi(s)\right] + T(s',\pi(s'),s')\left[R(s',\pi(s'),s')+\gamma V_\pi(s')\right]+\dots$$

$$V_\pi(s'') = T(s'',\pi(s'),s)\left[R(s'',\pi(s''),s)+\gamma V_\pi(s)\right] + T(s'',\pi(s''),s')\left[R(s'',\pi(s''),s')+\gamma V_\pi(s')\right]+\dots$$

$\vdots$

- Let $\underline{V} = \begin{bmatrix} V_\pi(s) \\ V_\pi(s') \\ \vdots \end{bmatrix}$, $T = \begin{bmatrix} T(s,\pi(s),s) & T(s,\pi(s),s') & \dots \\ T(s',\pi(s'),s) & T(s',\pi(s'),s') & \dots \\ \vdots & \ddots & \end{bmatrix}$, $R = \begin{bmatrix} R(s,\pi(s),s) & R(s,\pi(s),s') & \dots \\ R(s',\pi(s'),s) & R(s',\pi(s'),s') & \dots \\ \vdots & \ddots & \end{bmatrix}$

- So, $\underline{V} = (T \odot R)\underline{1} + \gamma T \underline{V}$
  - $T \odot R$ is the hadamard product (i.e., element-wise product) of T and R.
  - $(T \odot R)\underline{1} = \begin{bmatrix} T(s,\pi(s),s) R(s,\pi(s),s') + T(s,\pi(s),s') R(s,\pi(s),s') + \cdots \\ T(s',\pi(s'),s) R(s',\pi(s'),s) + T(s',\pi(s'),s') R(s',\pi(s'),s') + \cdots \\ \vdots \end{bmatrix}$
- Thus, $\underline{V} = (I - \gamma T)^{-1} \left((T \odot R)\underline{1} \right)$
  - Note that $(I - \gamma T)$ is invertible.