Reinforcement Learning

Lecture 4:
Discounted MDP

Yishay Mansour, Tel-Aviv University

Talk Announcement

- Time: Wednesday, March 20th, 1pm (After class !)
- Place: The Steinhardt Museum of Natural History.
- Speaker: Prof. Alexander Rush (Harvard University)
- Title: Controllable Text Generation with Deep Latent-Variable Models

- HW announcement: you can do it in 2 or 3 students!

Lecture 4: outline

- Discounted Return
  - Definition
  - Basic Properties
- Policy Evaluation
  - Linear equations
  - Value iteration
- Contraction Operator
  - Convergence
- Optimal Policy
  - Value Iteration
  - Policy Iteration
  - Dynamic Programing

Markov Decision Process

- Finite state space:
  - $S = \{1, \dots, n\}$
- Finite action set:
  - $A = \{1, \dots, m\}$
- Transition function:
  - $p(s'|s, a)$
- Rewards:
  - $R(s, a)$
    - can be r.v.
  - $r(s, a) = E[R(s, a)]$

Discounted Return

- Discounted return
  - Infinite horizon
  - Parameter $\gamma \in (0,1)$

$$J_\gamma^\pi(s) = E^\pi\left[\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t) | s_0 = s\right]$$

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

Optimal Policy

- Optimal return
  - $J_\gamma^*(s) = \sup_{\pi \in HR} J_\gamma^\pi(s)$
- Bounded Rewards
  - $|r(s,a)| \leq R_{max}$
  - For any policy π

$$|J_\gamma^\pi(s)| \leq \sum_{t=0}^{\infty} \gamma^t R_{max} = \frac{R_{max}}{1-\gamma}$$

Policy Evaluation

- Consider $\pi \in SD$
  - Stationary deterministic policy
  - $\pi: S \rightarrow A$
- Value function

  $\forall s \in S$:

$$V^{\pi}(s) \triangleq E^{\pi,s}\left[\sum_{t=0}^{\infty} \gamma^{t}\, r(s_t, a_t)\right] = J_{\gamma}^{\pi}(s)$$

Policy Evaluation: Value Eq

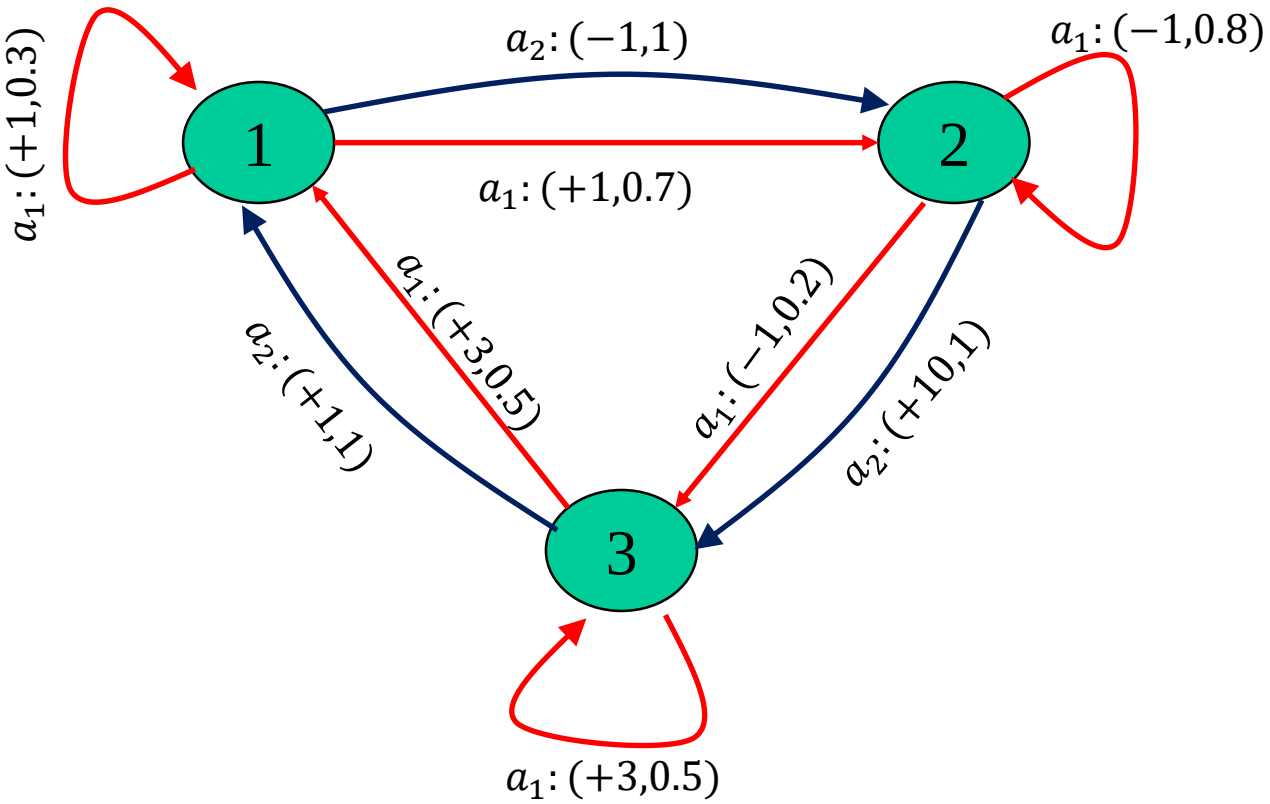
- Lemma:

  V^{π} satisfies the following $|S|$ linear equations:

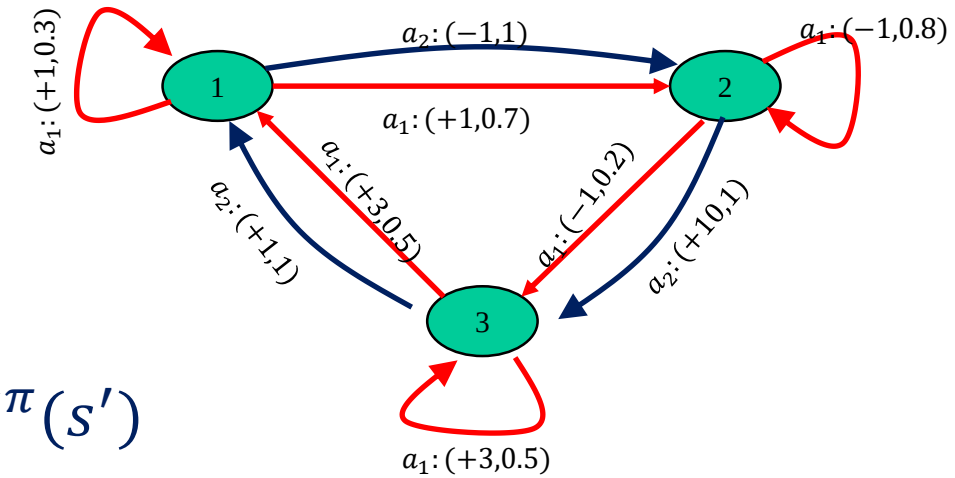
$$\forall s \in S$$

$$V^{\pi}(s) = r\big(s, \pi(s)\big) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V^{\pi}(s')$$

Example:



Example:



- Policy $\pi(s) = a_1 \;\; \forall s \in S, \gamma = 0.9$
- Equations
  - $V^{\pi}(s) = r\big(s, \pi(s)\big) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V^{\pi}(s')$
  - $V^{\pi}(1) = (+1) + 0.9(0.3\, V^{\pi}(1) + 0.7 V^{\pi}(2))$
  - $V^{\pi}(2) = (-1) + 0.9(0.8\, V^{\pi}(2) + 0.2 V^{\pi}(3))$
  - $V^{\pi}(3) = (+3) + 0.9(0.5\, V^{\pi}(3) + 0.5 V^{\pi}(1))$
- Solution
  - $V^{\pi}(1) \approx 2.4; \;\; V^{\pi}(2) \approx 1.2; V^{\pi}(3) \approx 7.42$

Policy Evaluation: Value Eq

- Proof (linear equations):
- Claim: Value function "immune" to time shift:

$$V^{\pi}(s) \triangleq E^{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t}\, r(s_t, a_t) | s_0 = s\right]$$

$$= E^{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1}\, r(s_t, a_t) | s_1 = s\right]$$

- Derivation:

$$V^{\pi}(s) = E^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \, r(s_t, a_t) | s_0 = s\right]$$

$$= r\big(s, \pi(s)\big) + E^{\pi}\left[\sum_{t=1}^{\infty} \gamma^t \, r(s_t, a_t) | s_0 = s\right]$$

- Second term:

$$E^{\pi}\left[E^{\pi}_{s' \sim p(\cdot|s_0, a_0)}\left[\sum_{t=1}^{\infty} \gamma^t \, r(s_t, a_t) | s_0 = s, s_1 = s'\right] | s_0 = s\right]$$

$$\gamma V^{\pi}(s')$$

- Completing the proof:

$$E^{\pi}[V^{\pi}(s')|s_0 = s] = \sum_{s' \in S} p(s'|s, \pi(s))V^{\pi}(s')$$

- Combining the equations:

$$V^{\pi}(s) = r\big(s, \pi(s)\big) + \gamma \sum_{s' \in S} p(s'|s, \pi(s))V^{\pi}(s')$$

- QED

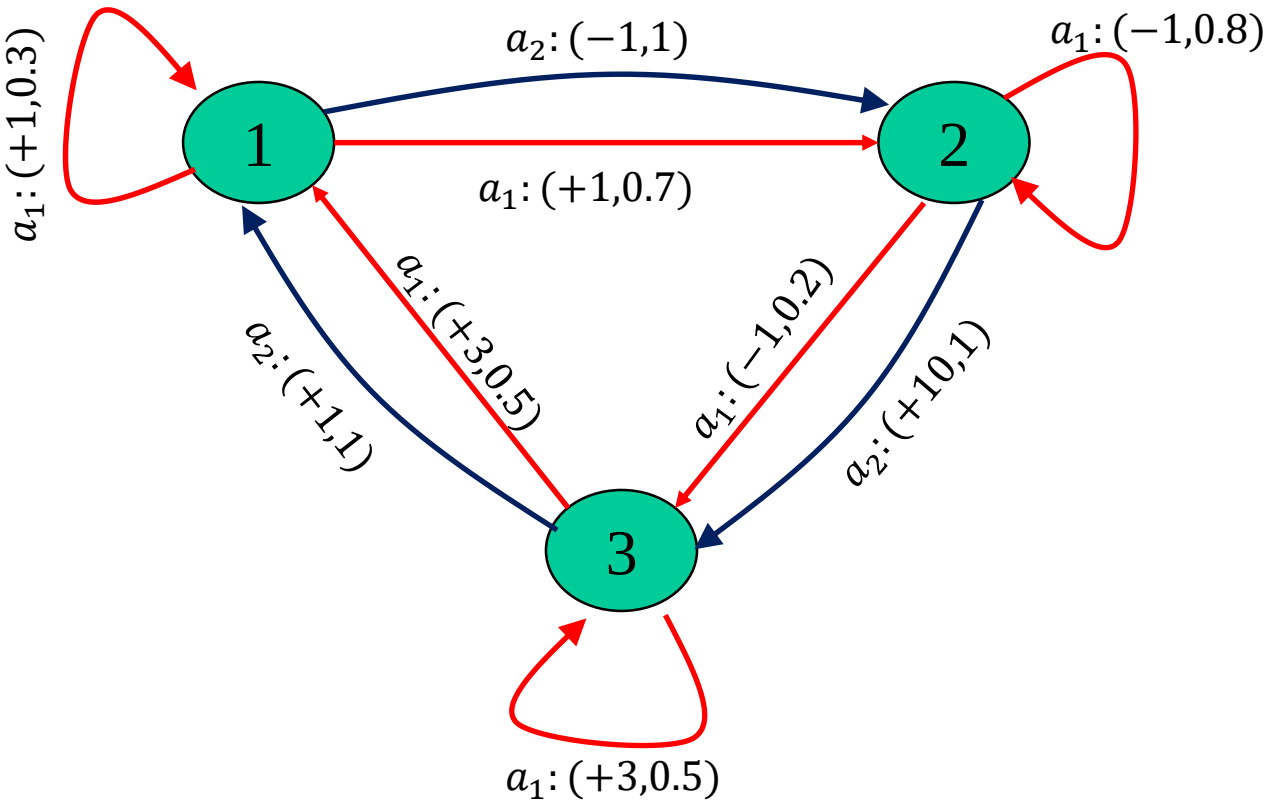
Linear equations: Matrix notation

- Using vector/matrix notation
- Let
  - $r^{\pi}(s) = r(s, \pi(s))$
  - $r^{\pi} = \left(r^{\pi}(s)\right)_{s \in S}$
  - $P^{\pi}_{s,s'} = p(s'|s, \pi(s))$
  - $V^{\pi} = \left(V^{\pi}(s)\right)_{s \in S}$

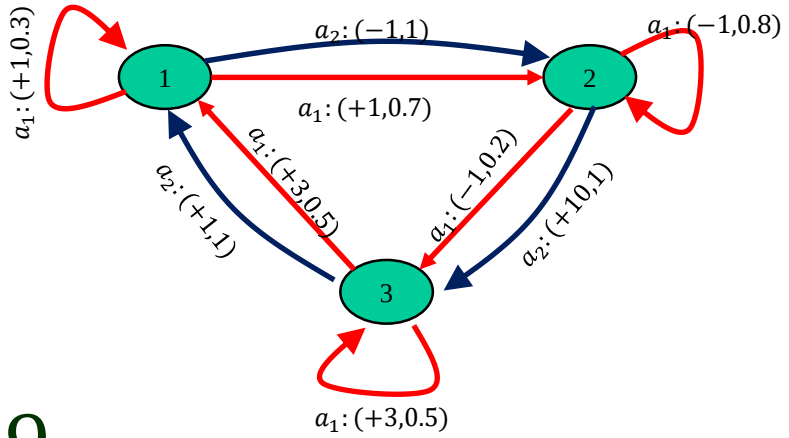
- Linear Eq:

$$V^{\pi} = r^{\pi} + \gamma P^{\pi} V^{\pi}$$

Example:



Example:



- Policy $\pi(s) = a_1 \quad \forall s \in S, \gamma = 0.9$
- Notation
  - $r^\pi = \begin{bmatrix} +1 \\ -1 \\ +3 \end{bmatrix}$ $\qquad P^\pi = \begin{bmatrix} 0.3 & 0.7 & 0.0 \\ 0.0 & 0.8 & 0.2 \\ 0.5 & 0.0 & 0.5 \end{bmatrix}$
- Equations
  - $V^\pi = r^\pi + \gamma P^\pi V^\pi = \begin{bmatrix} +1 \\ -1 \\ +3 \end{bmatrix} + 0.9 \begin{bmatrix} 0.3 & 0.7 & 0.0 \\ 0.0 & 0.8 & 0.2 \\ 0.5 & 0.0 & 0.5 \end{bmatrix} V^\pi$

Proposition:

The fixed policy value function V^{π} is given by the unique solution of:

$$V^{\pi} = (I - \gamma P^{\pi})^{-1} r^{\pi}$$

❑ Lemma:

The equations

$$V^{\pi} = r^{\pi} + \gamma P^{\pi} V^{\pi}$$

Have a unique solution:

$$V^{\pi} = (I - \gamma P^{\pi})^{-1} r^{\pi}$$

❑ Proof:

Need to show that

$$I - \gamma P^{\pi}$$

is non-singular.

Eigenvalues of stochastic matrix

- P is row stochastic
- $P\vec{1} = \vec{1}$
  - $\lambda_1 = 1$
- P can have complex λ_i

$$\begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.1 & 0.6 & 0.3 \\ 0.3 & 0.1 & 0.6 \end{bmatrix}$$

  - $\lambda_{2,3} = 0.4 \pm 0.1\sqrt{3}\boldsymbol{i}$
    - $|\lambda_{2,3}| = \sqrt{0.16 + 0.03}$

- Claim: $\forall i \ \ |\lambda_i| \leq 1$
- Assume v has $\lambda > 1$
- If P is stochastic matrix then so is P^m
- $P^m v = \lambda^m v$
- $|P^m v| = |\lambda|^m |v|$
  - $|\lambda|^m \to \infty$
  - P^m is stochastic
    - $|P^m v| < \infty$

Back to proof

- P^π is stochastic
  - Eigenvalues

    $\lambda_1 = 1, \cdots, \lambda_n$
- Matrix $I - \gamma P^\pi$
  - Eigenvalues

    $1 - \gamma\lambda_i$
  - Positivity

    $|1 - \gamma\lambda_i| \geq 1 - \gamma > 0$

- System of Eq:

  $V^\pi = r^\pi + \gamma P^\pi V^\pi$
- Unique solution:

  $V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$
- QED

Lecture 4: outline

- Discounted Return
  - Definition
  - Basic Properties
- Policy Evaluation
  - Linear equations
  - Value iteration
- Contraction Operator
  - Convergence
- Optimal Policy
  - Value Iteration
  - Policy Iteration
  - Dynamic Programing

Fixed Policy Value Iteration

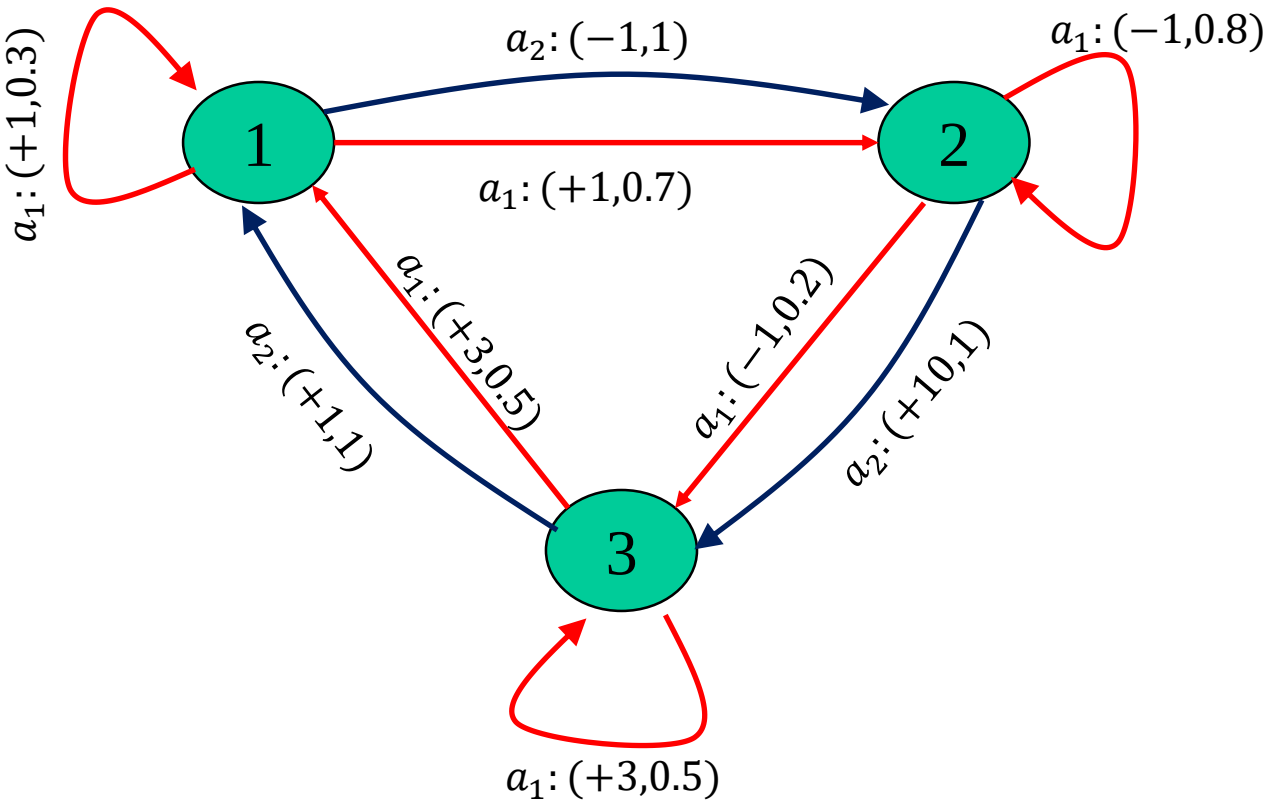
- Algorithm:
  - Let $V_0 = \left(V_0(s)\right)_{s \in S}$
    - Arbitrary
  - For $n = 0,1, \ldots$ set

$$V_{n+1}^{\pi}(s) = r\left(s, \pi(s)\right) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V_n^{\pi}(s')$$

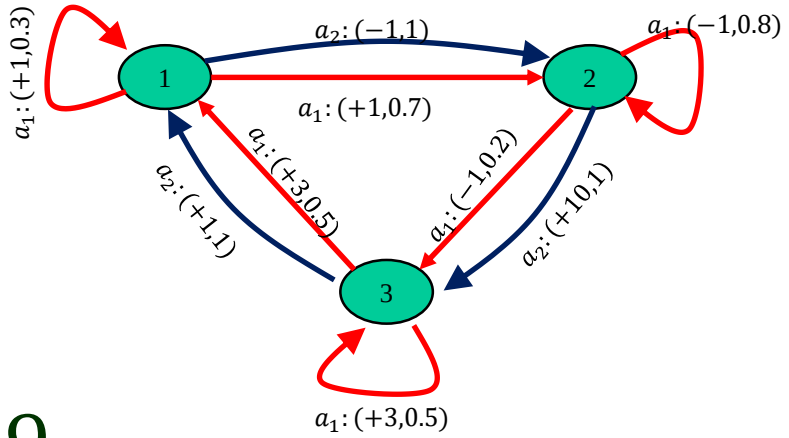
  - Or, equivalently,

$$V_{n+1}^{\pi} = r^{\pi} + \gamma P^{\pi} V_n^{\pi}$$

Example:



Example:



- Policy $\pi(s) = a_1 \quad \forall s \in S, \gamma = 0.9$
- Iteration
  - $V_{n+1}^{\pi} = r^{\pi} + \gamma P^{\pi} V_n^{\pi} = \begin{bmatrix} +1 \\ -1 \\ +3 \end{bmatrix} + 0.9 \begin{bmatrix} 0.3 & 0.7 & 0.0 \\ 0.0 & 0.8 & 0.2 \\ 0.5 & 0.0 & 0.5 \end{bmatrix} V_n^{\pi}$
  - Let $V_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}; V_1 = \begin{bmatrix} +1 \\ -1 \\ +3 \end{bmatrix}; V_2 = \begin{bmatrix} +0.64 \\ -1.18 \\ +4.8 \end{bmatrix}; V_3 = \begin{bmatrix} +0.43 \\ -0.98 \\ +5.48 \end{bmatrix}$
  - Recall
    - $V^{\pi}(1) \approx 2.4; \quad V^{\pi}(2) \approx 1.2; V^{\pi}(3) \approx 7.42$

Convergence of fixed policy value iteration

- Proposition:
  - $\lim_{n\to\infty} V_n^{\pi}(s) = V^{\pi}(s) \quad \forall s \in S$
- Proof: Note that
- $V_1(s) = r\big(s,\pi(s)\big) + \gamma \sum_{s'\in S} p(s'|s,\pi(s)) V_0(s')$
  $= E^{\pi}[r(s_0,a_0) + \gamma V_0(s_1)|s_0 = s]$
- By induction:
- $V_n^{\pi}(s) = E^{\pi}[\sum_{t=0}^{n-1} \gamma^t\, r(s_t,a_t) + \gamma^n\, V_0(s_n)|s_0 = s]$

- Comparing to V^{π}
- $V^{\pi}(s) - V_n(s) =$
  $E^{\pi}[\sum_{t=n}^{\infty} \gamma^t r(s_t, a_t) - \gamma^n V_0(s_n) | s_0 = s]$
- Let
  - $R_{max} = \max_{s,a} |r(s,a)|,\ V_{0,max} = \max_{s} |V_0(s)|$
- Now
  - $|V^{\pi}(s) - V_n^{\pi}(s)| \leq \gamma^n \left(\frac{R_{max}}{1-\gamma} + V_{0,max}\right) \rightarrow_{n\rightarrow\infty} 0$
- QED

Remarks:

- Convergence rate
  - Implicit in the proof: $O(\frac{\gamma^n}{1-\gamma})$
- Algorithms:
  - Solving $|S|$ linear equations
    - Time $O(|S|^3)$
  - Value iteration:
    - time of each iteration $O(|S|^2)$

Overview: Main DP algorithm

- Theorem (Bellman Optimality Eq)

V^* is the unique solution to the following (non-linear) equations $\forall s \in S$:

$$V^*(s) = \max_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V^*(s') \right\}$$

Overview: Main DP algorithm

- Theorem (Bellman Optimality Eq)

Any stationary policy π^* that satisfies

$$\pi^*(s) \in \arg\max_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V^*(s') \right\}$$

is an optimal policy

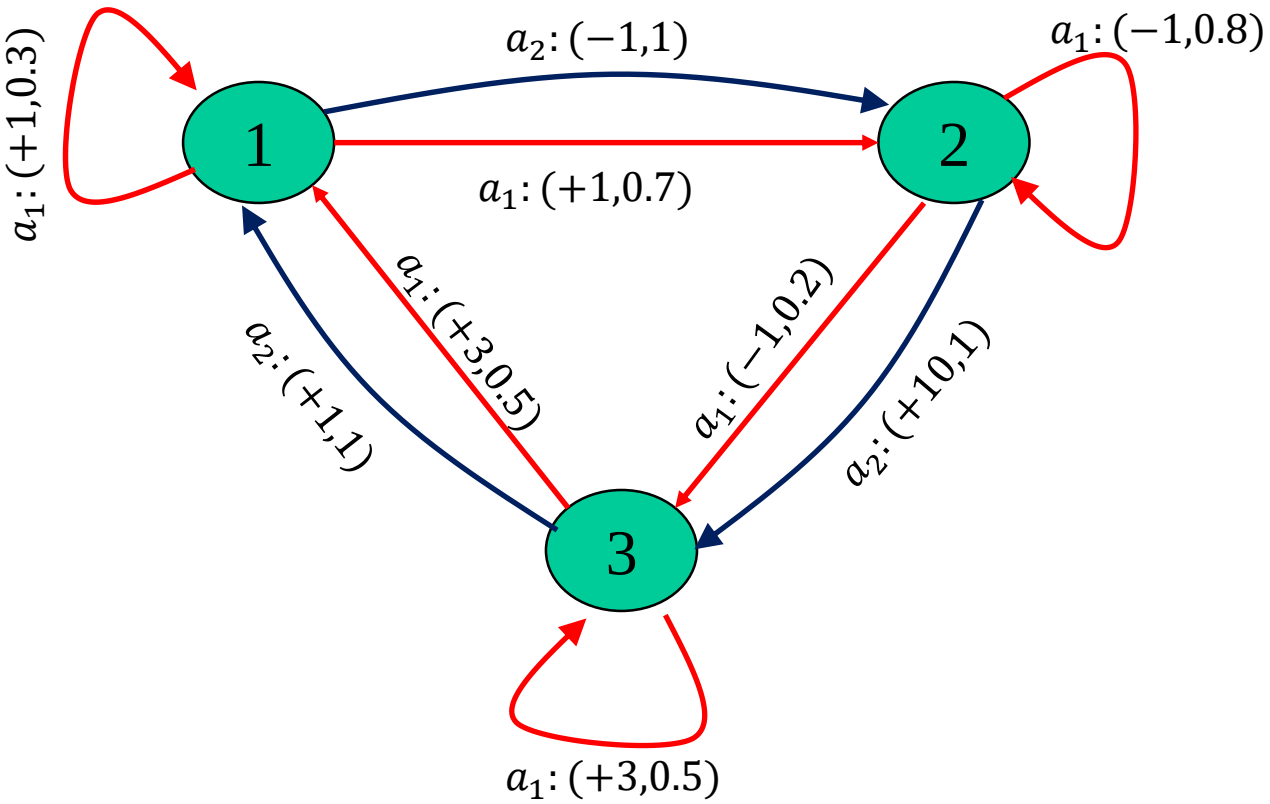
Value Iteration

- Algorithm:
  - Let $V_0 = \left(V_0(s)\right)_{s \in S}$
  - For $n = 0,1, \ldots$

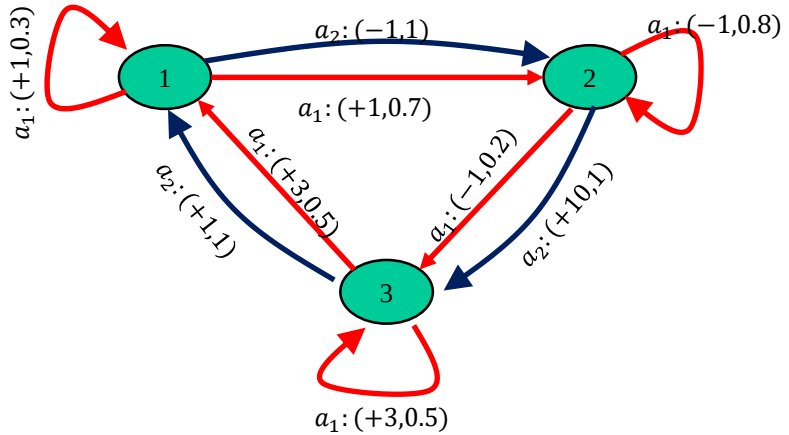
$$V_{n+1}(s) = \max_{a \in A} \left\{ r(s,a) + \gamma \sum\nolimits_{s' \in S} p(s'|s,a) V_n(s') \right\}$$

- Theorem: $\lim_{n \to \infty} V_n = V^*$
  - Convergence rate $O\left(\frac{\gamma^n}{1-\gamma}\right)$

Example:



Example:



- Let $\gamma = 0.9$
- Equations
  - $V_n(s) = \max_a \{r(s,\pi(s)) + \gamma \sum_{s' \in S} p(s'|s,\pi(s)) V_{n-1}(s')\}$
- Let $V_0(s) = 0$
  - $V_1(1) = \max\{(-1) + 0.9V_0(2),$
    $(+1) + 0.9\big(0.3V_0(1) + 0.7V_0(2)\big)\} = 1$
  - $V_1(2) = \max\{(+10) + 0.9V_0(3),$
    $(-1) + 0.9\big(0.8V_0(2) + 0.2\,V_0(3)\big)\} = 10$
  - $V_1(3) = \{(+1) + 0.9V_0(1),$
    $(+3) + 0.9\big(0.5\,V^{\pi}(3) + 0.5V^{\pi}(1)\big)\} = 3$

- Theorem: $\lim_{n\to\infty} V_n = V^*$
  - Convergence rate $O(\frac{\gamma^n}{1-\gamma})$
- Proof:
  - Consider the $T = n$ finite-horizon
    - Reward at time t: $\gamma^t R_t$
    - Final reward at state s: $\gamma^n V_0(s)$
  - V_n is the optimal value for the *n-finite-horizon*
    - $V_n(s) = \max_\pi E^{\pi,s}[\sum_{t=0}^{n-1} \gamma^t R_t + \gamma^n V_0(s_n)]$
    - $\geq \max_\pi E^{\pi,s}[\sum_{t=0}^{n-1} \gamma^t R_t] - \gamma^n V_{0,max}$

- Optimal value
  - $V^*(s) = \max_{\pi} E^{\pi,s}[\sum_{t=0}^{\infty} \gamma^t R_t]$

  $\leq \max_{\pi} E^{\pi,s}[\sum_{t=0}^{n-1} \gamma^t R_t] + \max_{\pi} E^{\pi,s}[\sum_{t=n}^{\infty} \gamma^t R_t]$

  $\leq (V_n + \gamma^n V_{0,max}) + \frac{\gamma^n}{1-\gamma} R_{max}$
- Convergence

  $|V^*(s) - V_n(s)| \leq \frac{\gamma^n}{1-\gamma} R_{max} + \gamma^n V_{0,max}$
- QED

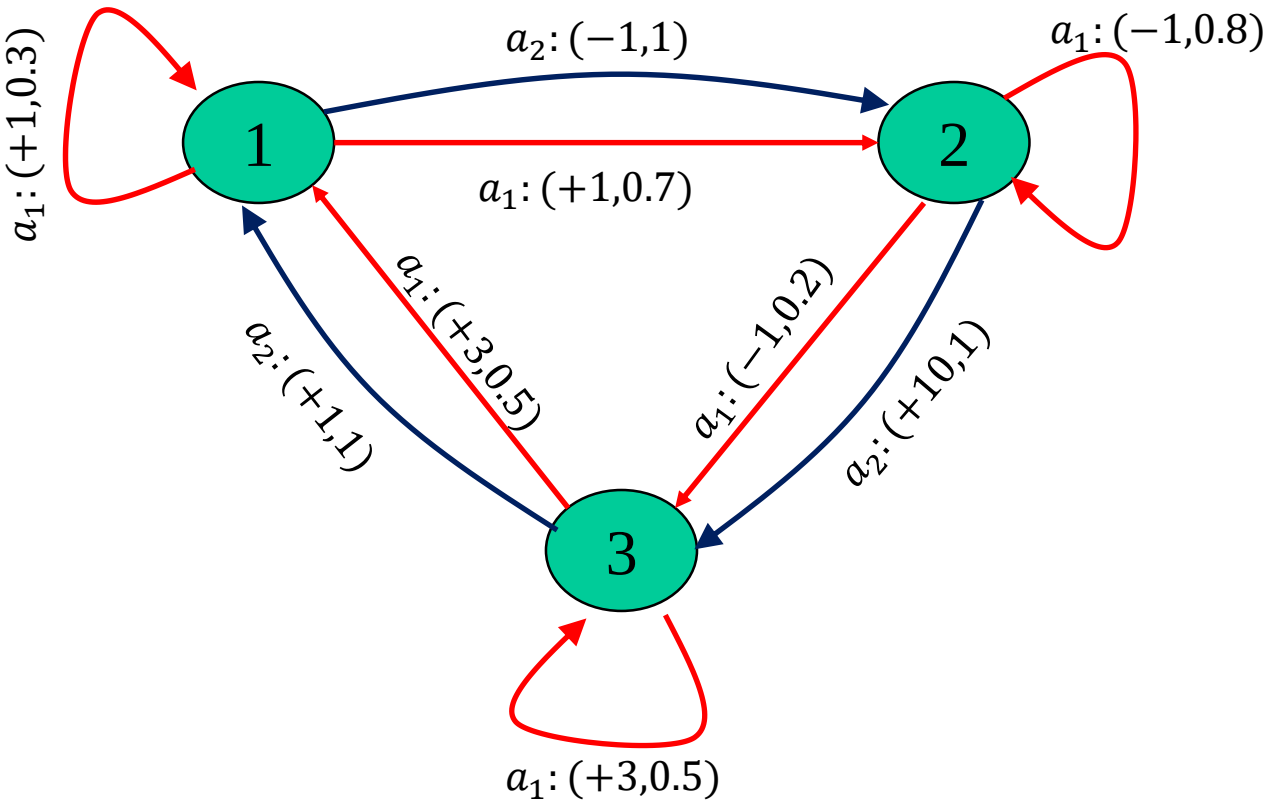
Policy Iteration

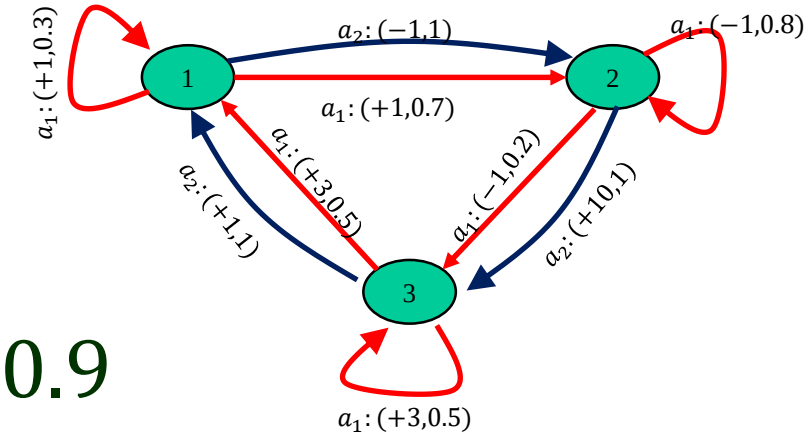
- Algorithm:
- Initialize policy π_0
- For $k = 0,1, \ldots$
  - Policy evaluation: compute V^{π_k}
  - Policy improvement: compute π_{k+1}

$$\pi_{k+1}(s) \in arg \max_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V^{\pi_k}(s') \right\}$$

  - Stop if $V^{\pi_{k+1}} = V^{\pi_k}$

Example:





Example:

- Policy $\pi_0(s) = a_1 \quad \forall s \in S, \gamma = 0.9$
- $V^{\pi_0}(1) \approx 2.4; \quad V^{\pi_0}(2) \approx 1.2; V^{\pi_0}(3) \approx 7.42$
  - $\pi_1(1) \in arg\max \{(-1) + 0.9V^{\pi_0}(2),$
    $(+1) + 0.9\big(0.3V^{\pi_0}(1) + 0.7V^{\pi_0}(2)\big)\} = \{a_1\}$
  - $\pi_1(2) \in arg\max\{(+10) + 0.9V^{\pi_0}(3),$
    $(-1) + 0.9\big(0.8V^{\pi_0}(2) + 0.2V^{\pi_0}(3)\big)\} = \{a_2\}$
  - $V_1(3) \in \arg\max\{(+1) + 0.9V^{\pi_0}(1),$
    $(+3) + 0.9\big(0.5\,V^{\pi_0}(3) + 0.5V^{\pi_0}(1)\big)\} = \{a_1\}$
- $\pi_1(1) = a_1; \pi_1(2) = a_2; \pi_1(3) = a_1$
  - $V^{\pi_1}(1) \approx 39; V^{\pi_1}(2) \approx 43; V^{\pi_1}(3) \approx 37$

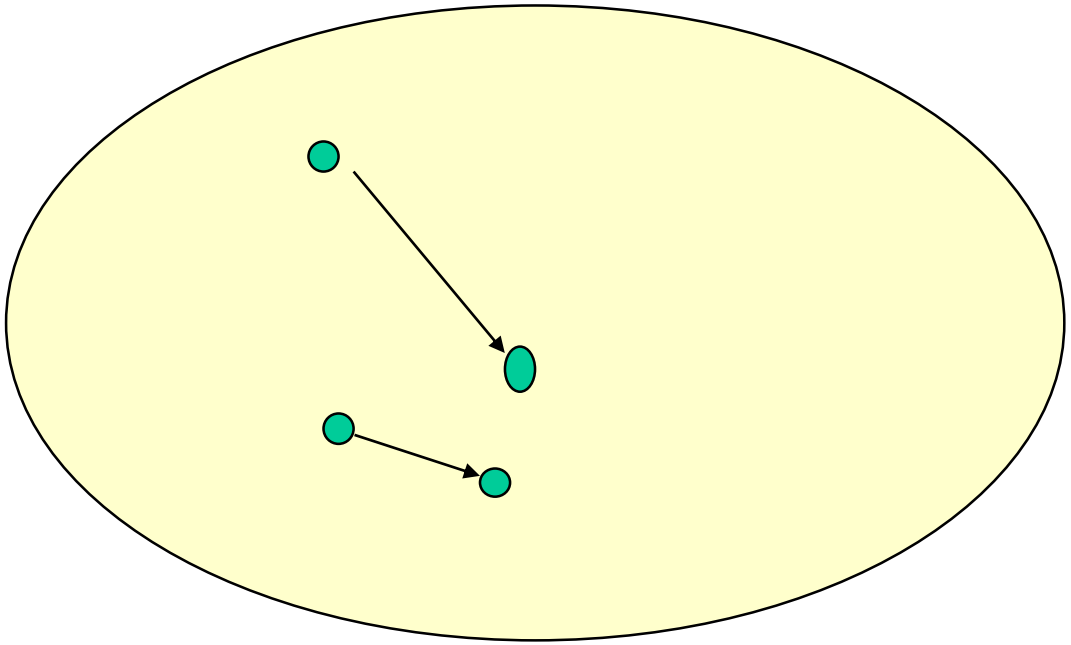
Convergence of Policy Iter.

- Theorem:
  - Each policy π_{k+1} improves over π_k
    - $\forall s \in S$: $V^{\pi_{k+1}}(s) \geq V^{\pi_k}(s)$
  - If $V^{\pi_{k+1}}(s) = V^{\pi_k}(s)$ then π_k is optimal
  - Number of iterations at most num. of policies
- Proof: latter …

Lecture 4: outline

- Discounted Return
  - Definition
  - Basic Properties
- Policy Evaluation
  - Linear equations
  - Value iteration
- Contraction Operator
  - Convergence
- Optimal Policy
  - Value Iteration
  - Policy Iteration
  - Dynamic Programing

Contraction Operator



Contraction Operator

- Norm $\|\cdot\|: \mathbb{R}^d \to \mathbb{R}$
  - For $x, y \in \mathbb{R}^d$, a scalar
  - $\|ax\| = |a| \, \|x\|$
  - $\|x + y\| \leq \|x\| + \|y\|$
  - $\|x\| = 0 \Leftrightarrow x = 0$
- Examples:
  - $\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p}$
  - $\|x\|_\infty = \max_{1 \leq x \leq d} |x_i|$
- Operator $T: \mathbb{R}^d \to \mathbb{R}^d$
  - $T^k(v) = T(T^{k-1}(v))$
- Contracting operator
  - $\|T(v_1) - T(v_2)\| \leq \beta \|v_1 - v_2\|$
    - $\beta \in (0,1)$

Banach Fixed Point Theorem

- Theorem: Let $T: \mathbb{R}^d \to \mathbb{R}^d$ be a contracting operator. Then
  - $T(v) = v$ has a unique solution $v^* \in \mathbb{R}^d$
  - For any $v_0 \in \mathbb{R}^d$: $\lim_{n\to\infty} T^n(v_0) = v^*$
    - $\|T^n(v_0) - v^*\| \leq O(\beta^n)$

- Remarks: applies to more general spaces

- Proof:
- **Existence**: consider $v_n \triangleq T^n(v_0)$
  - Show that v_n is a Cauchy sequence
    - For every $\varepsilon > 0$, exists $N > 0$, for all $m, n \geq N$:
      $$\|v_{n+m} - v_n\| \leq \varepsilon$$
  - Consider:
    - $\|v_{n+m} - v_n\| = \left\|\sum_{k=0}^{m-1} v_{n+k+1} - v_{n+k}\right\|$
    - $\leq \sum_{k=0}^{m-1} \|v_{n+k+1} - v_{n+k}\|$
    - $\leq \sum_{k=0}^{m-1} \left\|T^{n+k}(v_1) - T^{n+k}(v_0)\right\|$
    - $\leq \sum_{k=0}^{m-1} \beta^{n+k} \|v_1 - v_0\|$
    - $= \frac{\beta^n(1-\beta^m)}{1-\beta} \|v_1 - v_0\|$

❑Proof (continue)

- Every Cauchy sequence has a limit
  - ➢Let v^* be the limit
- Show that $T(v^*) = v^*$
  - ➢ $\|T(v^*) - v^*\| \leq \|T(v^*) - v_n\| + \|v_n - v^*\|$
  - ➢ $\leq \|T(v^*) - T(v_{n-1})\| + \|v_n - v^*\|$
  - ➢ $\leq \beta\|v^* - v_{n-1}\| + \|v_n - v^*\|$
  - ➢Since v^* is the limit of v_n

    $$\lim_{n\to\infty} \|v_n - v^*\| = 0$$
  - ➢Therefore $\|T(v^*) - v^*\|$=0

- Uniqueness:
  - let $T(v_1) = v_1$ and $T(v_2) = v_2$

$$\|v_1 - v_2\| = \|T(v_1) - T(v_2)\| \leq \beta\|v_1 - v_2\|$$

  - Implies $\|v_1 - v_2\| = 0$ hence $v_1 = v_2$
- Convergence rate
  - $\|v_n - v^*\| = \|T^n(v_0) - T^n(v^*)\|$
    $\leq \beta^n\|v_0 - v^*\|$
- QED

Dynamic Programing Operator

- Fixed policy π

$$\left(T^{\pi}(V)\right)(s) = r\left(s, \pi(s)\right) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V(s')$$

- Optimal policy π^*

$$\left(T^{*}(V)\right)(s) = \max_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V(s') \right\}$$

- Max-norm $\|V\|_{\infty} \triangleq \max_{s \in S} |V(s)|$
- Remark: T^{π} is linear while T^{*} is not!

DP and contraction

- Theorem (contraction property)
  - T^{π} is a γ-contraction operator w.r.t max-norm

$$\|T^{\pi}(V_1) - T^{\pi}(V_2)\| \leq \gamma \|V_1 - V_2\|_{\infty}$$

    - for any $V_1, V_2 \in \mathbb{R}^{|S|}$
  - T^{*} is a γ-contraction operator w.r.t max-norm

$$\|T^{*}(V_1) - T^{*}(V_2)\| \leq \gamma \|V_1 - V_2\|_{\infty}$$

    - for any $V_1, V_2 \in \mathbb{R}^{|S|}$

- Proof:
- For T^{π}:
  - $|T^{\pi}(V_1)(s) - T^{\pi}(V_2)(s)| =$
  - $|r\big(s,\pi(s)\big) + \gamma \sum_{s' \in S} p(s'|s,\pi(s))V_1(s')$
  - $-r\big(s,\pi(s)\big) - \gamma \sum_{s' \in S} p(s'|s,\pi(s))V_2(s')|$
  - $= |\gamma \sum_{s' \in S} p(s'|s,\pi(s))(V_1(s') - V_2(s')|$
  - $\leq \gamma \sum_{s' \in S} p(s'|s,\pi(s))|V_1(s') - V_2(s')|$
  - Hence
    $\|T^{\pi}(V_1) - T^{\pi}(V_2)\|_{\infty} \leq \gamma \|V_1 - V_2\|_{\infty}$

- For T^*:
  - $T^*(V_1)(s) - T^*(V_2)(s) =$
  - $r(s, a_s^1) + \gamma \sum_{s' \in S} p(s'|s, a_s^1) V_1(s')$
  - $-r(s, a_s^2) - \gamma \sum_{s' \in S} p(s'|s, a_s^2) V_2(s')$
  - $\leq r(s, a_s^1) + \gamma \sum_{s' \in S} p(s'|s, a_s^1) V_1(s')$
  - $- r(s, a_s^1) - \gamma \sum_{s' \in S} p(s'|s, a_s^1) V_2(s')$
  - $\leq \gamma \sum_{s' \in S} p(s'|s, a_s^1) \max_{s''} |V_1(s'') - V_2(s'')|$
  - $\leq \gamma \, \|V_1 - V_2\|_\infty$

- Similarly:
  - $T^*(V_2)(s) - T^*(V_1)(s)$
  - $\leq \gamma \, \|V_1 - V_2\|_\infty$
- Hence

$$\|T^*(V_1) - T^*(V_2)\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$$

- QED

Lecture 4: outline

- Discounted Return
  - Definition
  - Basic Properties
- Policy Evaluation
  - Linear equations
  - Value iteration
- Contraction Operator
  - Convergence
- Optimal Policy
  - Value Iteration
  - Policy Iteration
  - Dynamic Programing

Value Iteration

- Algorithm:
  - Let $V_0 = \left(V_0(s)\right)_{s \in S}$
  - For $n = 0,1, \ldots$

$$V_{n+1}(s) = \max_{a \in A} \left\{ r(s,a) + \gamma \sum\nolimits_{s' \in S} p(s'|s,a) V_n(s') \right\}$$

- Theorem: $\lim_{n \to \infty} V_n = V^*$
  - Convergence rate $O\left(\frac{\gamma^n}{1-\gamma}\right)$

Error bounds and Stopping rules

- Claim: If $\|V_{n+1} - V_n\|_\infty \leq \varepsilon \frac{1-\gamma}{2\gamma}$ then π_n , greedy w.r.t. V_n is ε-optimal
- Proof:
  - Show that $\|V^* - V^{\pi_n}\| \leq \varepsilon$
  - $\|V^* - V^{\pi_n}\| \leq \|V^* - V_{n+1}\| + \|V_{n+1} - V^{\pi_n}\|$
    - Bound each term separately

- $\|V^* - V_{n+1}\| = \|T^*(V^*) - V_{n+1}\|$

$\leq \|T^*(V^*) - T^*(V_{n+1})\| + \|T^*(V_{n+1}) - T^*(V_n)\|$

$\leq \gamma\|V^* - V_{n+1}\| + \gamma\|V_{n+1} - V_n\|$

- Hence

$$\|V^* - V_{n+1}\| \leq \frac{\gamma}{1-\gamma}\|V_{n+1} - V_n\|$$

- Similarly:
- $\|V^{\pi_n} - V_{n+1}\| = \|T^{\pi_n}(V^{\pi_n}) - V_{n+1}\| = \|T^*(V^{\pi_n}) - V_{n+1}\|$

$\leq \|T^*(V^{\pi_n}) - T^*(V_{n+1})\| + \|T^*(V_{n+1}) - T^*(V_n)\|$

$\leq \gamma\|V^{\pi_n} - V_{n+1}\| + \gamma\|V_{n+1} - V_n\|$

- Hence

$$\|V^{\pi_n} - V_{n+1}\| \leq \frac{\gamma}{1-\gamma}\|V_{n+1} - V_n\|$$

- Completing the proof:
  - $\|V^* - V^{\pi_n}\| \leq \|V^* - V_{n+1}\| + \|V_{n+1} - V^{\pi_n}\|$
  - $\|V^* - V^{\pi_n}\| \leq \frac{\gamma}{1-\gamma}\|V_{n+1} - V_n\| + \frac{\gamma}{1-\gamma}\|V_{n+1} - V_n\|$
    - Stopping criteria $\|V_{n+1} - V_n\|_\infty \leq \varepsilon\frac{1-\gamma}{2\gamma}$
  - $\|V^* - V^{\pi_n}\| \leq 2\frac{\gamma}{1-\gamma} \cdot \varepsilon\frac{1-\gamma}{2\gamma} = \varepsilon$
- QED

Policy Iteration

- Given a policy π we select a improving policy $\bar{\pi}$,
  - Greedy w.r.t. V^{π}

$$\bar{\pi}(s) \in \arg\max_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V^{\pi}(s') \right\}$$

- Theorem:
  - $V^{\bar{\pi}} \geq V^{\pi}$
  - $V^{\bar{\pi}} = V^{\pi}$ iff π is optimal

- Proof: Note that
  - $V^{\pi} = T^{\pi}(V^{\pi}) \leq T^{*}(V^{\pi}) = T^{\bar{\pi}}(V^{\pi})$
  - $T^{\bar{\pi}}$ is monotone
    - If $V_1 \leq V_2$ then $T^{\bar{\pi}}(V_1) \leq T^{\bar{\pi}}(V_2)$
  - Reuse the inequality
    - $V^{\pi} \leq T^{\bar{\pi}}(V^{\pi}) \leq \cdots \leq (T^{\bar{\pi}})^{n}(V^{\pi})$
    - Recall that $\lim_{n \to \infty} (T^{\bar{\pi}})^{n}(V^{\pi}) = V^{\bar{\pi}}$
  - We showed that $V^{\pi} \leq V^{\bar{\pi}}$
- Assume $V^{\bar{\pi}} = V^{\pi}$
  - $T^{*}(V^{\pi}) = T^{\bar{\pi}}(V^{\pi}) = V^{\bar{\pi}} = V^{\pi}$
    - V^{π} is a fix-point of T^{*}

Value vs Policy Iteration

- Computational complexity
  - Per iteration
    - VI: $O(|A|\,|S|^2)$
    - PI: $O(|A|\,|S|^2 + |S|^3)$
  - Number of iterations
    - For the same accuracy
    - PI uses less iterations

Value vs Policy Iteration

- Claim: Let V_n be the values generated by VI and U_n the values generated by PI. If $V_0 = U_0$ then $V_n \leq U_n$

- Proof:
  - by induction.
  - Base $n = 0$
    - in the statement of the claim

- Assume that $V_n \leq U_n$
  - $V_{n+1} = T^*(V_n) = T^{\pi'}(V_n)$
    - π' is the greedy policy
  - $T^{\pi'}(V_n) \leq T^{\pi'}(U_n)$
    - monotonicity
- $T^{\pi'}(U_n) \leq T^*(U_n) = T^{\pi_{n+1}}(U_n)$
  - T^* is maximizing over all policies
- $T^{\pi_{n+1}}(U_n) \leq V^{\pi_{n+1}} = U_{n+1}$
- QED

Variants of Value and Policy Iteration

- Update $V_n(s)$ sequenctially
  - Use $V_{n+1}(s')$ values for computed s'
  - Guarantee to converge "as fast" as VI
- Asynchronous Value Iteration
  - Update only a subset $S_n \subset S$ of the states
  - If each state updated infinitely often
    - $V_n \to V^*$

Variants of Value and Policy Iteration

- Asynchronous Policy Iteration
  - Update only a subset of the improving states
  - Guarantees to improve values
    - Does to cycle

Lecture 4: outline

- Discounted Return
  - Definition
  - Basic Properties
- Policy Evaluation
- Contraction Operator
  - Convergence
- Optimal Policy
  - Value Iteration
  - Policy Iteration
  - Dynamic Programing