

Value Iteration

Reinforcement Learning

Roberto Capobianco



SAPIENZA
UNIVERSITÀ DI ROMA

Recap

Sequential Decision Making

— — —

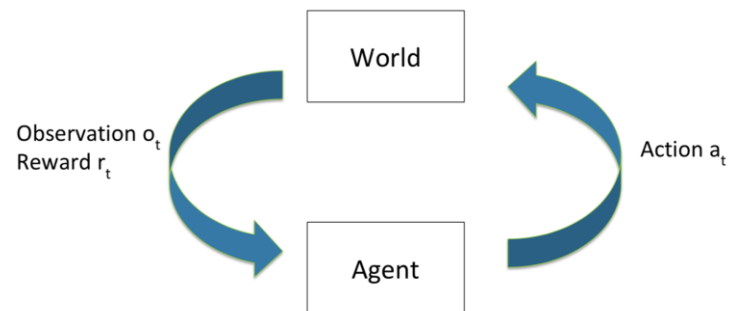
The agent interacts with the environment:

- at discrete timesteps;
- by receiving observations o_t and reward r_t from the environment;
- by taking actions a_t ;

The state is a function of the history:

$$s_t = f(h_t)$$

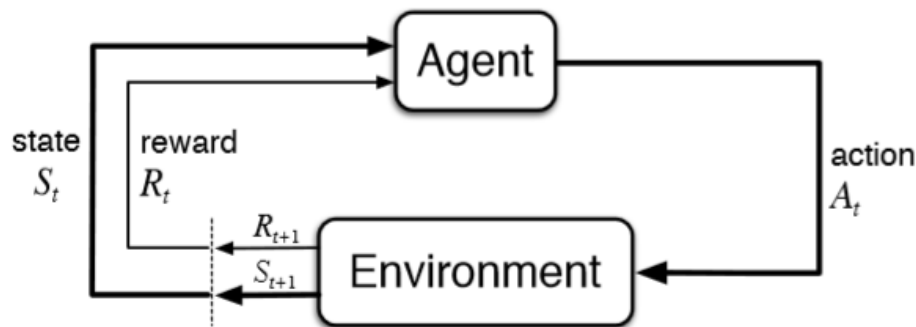
and it is typically hidden or unknown



Markov Decision Process (MDP)

— — —

- Set of states S
- Set of actions A



Alternative notation

Sequential Decision Making under Markov Assumption $s_{t+1} \sim p(\cdot | s_t, a_t)$ or

- Markovian transition dynamics
- Full Observability
- The transition dynamics T is (generally) stochastic $p(s_{t+1} | s_t, a_t)$



Policy

A policy π :

- is a mapping from (all) states to actions;
- determines how agents select actions;
- can be deterministic ($a = \pi(s)$) or stochastic ($\pi(a|s)$ or $p(a|s)$ or $a \sim \pi(.|s)$)



Value Function/Q-Function

— — —

- We estimate the goodness of states and actions based on their value
- It's also a measure to compare policies

$$V^{\pi}(s_t) = \mathbb{E}_{\pi}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots | s_t] = \mathbb{E}[\sum_{h=0}^{\infty} \gamma^h r_h | s_0 = s_t, a_h = \pi(s_h), s_{h+1} \sim p(\cdot | s_h, a_h)]$$

$$Q^{\pi}(s_t, a_t) = \mathbb{E}[\sum_{h=0}^{\infty} \gamma^h r_h | (s_0, a_0) = (s_t, a_t), a_{h+1} = \pi(s_h), s_{h+1} \sim p(\cdot | s_h, a_h)]$$

For infinite horizon MDPs there always exists a deterministic policy π^* such that

$$V^{\pi^*}(s) \geq V^{\pi}(s) \quad \forall s, \pi$$

meaning that π^* (optimal policy) dominates all other policies π in each state



Discount Factor

Setting $\gamma = 1$ for infinite tasks is a bad idea!

Note that $\sum_{h=0}^{\infty} \gamma^h$ is a geometric series and for γ in $[0,1]$ this is equivalent to $1/(1-\gamma)$

So, the value of γ approximately determines how many steps ahead we are considering

E.g., $\gamma=0.99 \rightarrow 99$ timesteps ahead



Bellman Equation

The value of a certain state is expanded in terms of the current reward and the value of the next states according to the policy

r here is function of s and $\pi(s)$

$$V^\pi(s_t) = \mathbb{E}_\pi[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots | s_t] = r_t + \gamma \mathbb{E}_{s' \sim p(\cdot | s, \pi(s))} [V^\pi(s')]$$

$$Q^\pi(s_t, a) = r_t + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a)} [V^\pi(s')]$$

r here is function of s and a

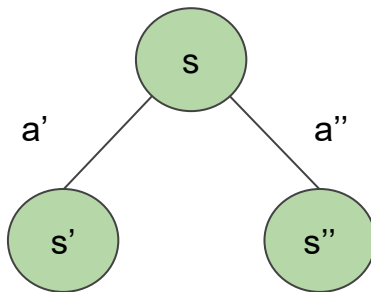
As a result $V(s) = Q(s, \pi(s))$



Bellman Optimality Example

$$V^*(s) = \max_a [r(s, a) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a)} V^*(s')]]$$

- Try a' , get $r(s, a')$,
compute
 $Q^*(s, a') = r(s, a') + \gamma V^*(s')$
- Try a'' , get $r(s, a'')$,
compute
 $Q^*(s, a'') = r(s, a'') + \gamma V^*(s'')$



Assume we know V^* at
 s' and s'' ,

$$V^*(s) = \max_{a', a''} \{ Q^*(s, a'), Q^*(s, a'') \}$$



Bellman Optimality (Theorem 1)

$$V^*(s) = \max_a [r(s, a) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a)} V^*(s')]$$

given $\hat{\pi} = \operatorname{argmax}_a Q^*(s, a)$, we can show $V^{\hat{\pi}} = V^*$

This implies $\pi^* = \operatorname{argmax}_a Q^*(s, a)$ is an optimal policy



Bellman Optimality (Theorem 2)

For any V , if $V(s) = \max_a [r(s, a) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a)} V(s')] for all s ,
then $V(s) = V^*(s)$$

This means we can focus on one step at each time (leaving the remaining “problem” to $V(s')$), and any V that satisfies this formula is in fact V^*



End - Recap



SAPIENZA
UNIVERSITÀ DI ROMA

Policy Evaluation

— — —

Question: given

- an MDP (S, A, T, R, γ)
- a policy π

how can we compute the goodness of π , i.e. V^π ?



Policy Evaluation

— — —

Question: given

- an MDP (S, A, T, R, γ)
- a policy π

how can we compute the goodness of π , i.e. V^π ?

WHY?



Policy Evaluation

— — —

Question: given

- an MDP (S, A, T, R, γ)
- a policy π

how can we compute the goodness of π , i.e. V^π ?

WHY?

There are A^S possible policies, and we want to find the optimal one! To find it, we need to be able to evaluate it



Exact Policy Evaluation

— — —

Given (S, A, T, R, γ) and π , what is V^π ?



Exact Policy Evaluation

— — —

Given (S, A, T, R, γ) and π , what is V^π ?

We know that **for ALL states**, Bellman equation holds

$$V^\pi(s) = r + \gamma \mathbb{E}_{s' \sim p(\cdot | s, \pi(s))} [V^\pi(s')]$$



Exact Policy Evaluation

Given (S, A, T, R, γ) and π , what is V^π ?

We know that **for ALL states**, Bellman equation holds

$$V^\pi(s) = r + \gamma \mathbb{E}_{s' \sim p(\cdot | s, \pi(s))} [V^\pi(s')]$$

How many linear constraints (equations) do we have?



Exact Policy Evaluation

Given (S, A, T, R, γ) and π , what is V^π ?

We know that **for ALL states**, Bellman equation holds

$$V^\pi(s) = r + \gamma \mathbb{E}_{s' \sim p(\cdot | s, \pi(s))} [V^\pi(s')]$$

How many linear constraints (equations) do we have?

S!



Exact Policy Evaluation

We know that **for ALL states**, Bellman equation holds

$$V^\pi(s) = r + \gamma \mathbb{E}_{s' \sim p(\cdot | s, \pi(s))} [V^\pi(s')]$$

We can combine all the constraints together:

$$V = R + \gamma P V$$

Credits: Wen Sun



Exact Policy Evaluation

Since we have this set of constraints

$$V = R + \gamma P V$$

we can solve for V as

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

$$\begin{matrix} \boxed{} \\ \boxed{V(s)} \\ \boxed{} \end{matrix} = \begin{matrix} \boxed{} \\ \boxed{r(s, \pi(s))} \\ \boxed{} \end{matrix} + \gamma \begin{matrix} \boxed{} & \boxed{} \\ \boxed{P(\cdot | s, \pi(s))} & \boxed{} \\ \boxed{} & \boxed{} \end{matrix} \begin{matrix} \boxed{} \\ \boxed{} \\ \boxed{} \end{matrix}$$

$V \qquad \qquad R \qquad \qquad P \qquad V$



Exact Policy Evaluation

Since we have this set of constraints

$$V = R + \gamma PV$$

we can solve for V as

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

The diagram illustrates the Bellman optimality equation $V = R + \gamma P V$ using matrix notation. On the left, a vertical vector V is shown with its middle element highlighted in red and labeled $V(s)$. This is followed by an equals sign, then a vertical vector R with its middle element highlighted in red and labeled $r(s, \pi(s))$. This is followed by a plus sign, then a scalar γ , then a matrix P (represented by a square with three rows, the middle row highlighted in red and labeled $P(\cdot | s, \pi(s))$), and finally a vertical vector V .

:(Nice but computationally expensive: inverting the matrix
is $O(S^3)$



Fixed-Point Iteration

What is a fixed-point? A point where holds

$$x = f(x)$$

How can we find such points?

- Initialize x_0
- Repeat $x_{i+1} = f(x_i)$
- Stop at convergence where x is found and does not change anymore



Contractions

Convergence to a fixed-point is possible thanks to the existence of **contraction mappings**

$f: M \rightarrow M$ (M is a metric space) is a contraction mapping if:

$$|f(x) - f(x')| \leq k|x - x'| \text{ for } k \text{ in } [0, 1)$$



Contraction Operator

In the simplest case the contraction mapping can be an operator as simple as a matrix, e.g. 0:

$$|0V - 0V'| \leq \gamma |V - V'|$$

(we can replace k with γ as they have the same range)



Iterative Policy Evaluation

We assume rewards are in
[0, 1]

- Initialize V_0 in $[0, 1/(1-\gamma)]$ (typically 0)
- Until convergence:

$$V_{i+1} = R + \gamma P V_i$$

(note: this is using matricial form because it's doing it for all states)

Iterative Policy Evaluation

- Initialize V_0 in $[0, 1/(1-\gamma)]$ (typically 0)
- Until convergence:

$$V_{i+1} = R + \gamma P V_i$$

(note: this is using matricial form because it's doing it for all states)

For each iteration it's $O(S^2)$



Iterative Policy Evaluation Theorem

— — —

At the end we have, for all s in S

$$\|V^t(s) - V^\pi(s)\| \leq \gamma^t \|V^0 - V^\pi\|$$

$$V^{t+1}(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi(s))} V^t(s')$$

$$V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi(s))} V^\pi(s')$$



$$V^{t+1}(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi(s))} V^t(s')$$

Iterative Policy Evaluation Theorem

$$V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi(s))} V^\pi(s')$$

At the end we have, for all s in S

$$\|V^t - V^\pi\| \leq \gamma^t \|V^0 - V^\pi\|$$

$$\begin{aligned} & \forall s, \left| V^{t+1}(s) - V^\pi(s) \right| \\ &= \left| r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, \pi(s))} V^t(s') - \left(r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, \pi(s))} V^\pi(s') \right) \right| \\ &= \gamma \left| \mathbb{E}_{s' \sim P(\cdot | s, \pi(s))} V^t(s') - \mathbb{E}_{s' \sim P(\cdot | s, \pi(s))} V^\pi(s') \right| \quad \text{Apply Jensen's inequality (and every norm and expectation is convex)} \\ &\leq \gamma \mathbb{E}_{s' \sim P(\cdot | s, \pi(s))} \left| V^t(s') - V^\pi(s') \right| \quad \text{Average is always smaller than } \max(\mathbb{E}|f(x)| \leq \max_x |f(x)|) \\ &\leq \gamma \|V^t - V^\pi\|_\infty \quad \text{Infinity norm: } \|V\| = \max_s |V(s)| \end{aligned}$$



Iterative Policy Evaluation Theorem

At the end we have, for all s in S

$$\|V^t(s) - V^\pi(s)\| \leq \gamma^t \|V^0 - V^\pi\|$$

$$\|V^{t+1} - V^\pi\|_\infty \leq \gamma \|V^t - V^\pi\|_\infty \leq \gamma^{t+1} \|V^0 - V^\pi\|_\infty$$



Iterative Policy Evaluation: Iterations

For iterative PE to find an ϵ accurate value function, we need a number of iterations n , with computational cost $O(S^2 \ln(1/\epsilon))$:

$$\gamma^n \|V^0 - V^\pi\| \leq \epsilon$$

$$\ln \left(\frac{\|V^0 - V^\star\|_\infty}{\epsilon} \right) / \ln(1/\gamma)$$



How to Find the Optimal Policy?

Now, what we're really interested in is finding the optimal policy π^*



How to Find the Optimal Policy?

Now, what we're really interested in is finding the optimal policy π^*

Naive approach: we know how to do policy evaluation, then

- For each possible policy, for all states
 - Do policy evaluation, and compute $V^\pi(s)$
 - Choose π' such that $V^{\pi'}(s) \geq V^\pi(s)$

How to Find the Optimal Policy?

Now, what we're really interested in is finding the optimal policy π^*

Naive approach: we know how to do policy evaluation, then

- For each possible policy, for all states
 - Do policy evaluation, and compute $V^\pi(s)$
 - Choose π' such that $V^{\pi'}(s) \geq V^\pi(s)$

If we do exact policy evaluation it's $O(\mathbf{A}^s \mathbf{S}^3)$



How to Find the Optimal Policy?

Now, what we're really interested in is finding the optimal policy π^*

Naive approach: we know how to do policy evaluation, then

- For each possible policy, for all states
 - Do policy evaluation, and compute $V^\pi(s)$
 - Choose π' such that $V^{\pi'}(s) \geq V^\pi(s)$



If we do exact policy evaluation it's $O(\mathbf{A}^s \mathbf{S}^3)$



How to Find the Optimal Policy?

Now, what we're really interested in is finding the optimal policy π^*

Let's use Bellman optimality! ...and the Bellman Operator



Bellman Backup is a Contraction

- Infinity norm: $\|V\| = \max_s |V(s)|$
- Set $\gamma < 1$
- Define the (non-linear) BV operator as a Bellman optimality equation applied to V :

$$BV = \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a)} [V(s')])$$

Alternative notation TV



Bellman Backup is a Contraction

— — —

$$\begin{aligned}\|BV_k - BV_j\| &= \left\| \max_a \left(R(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) V_k(s') \right) - \max_{a'} \left(R(s, a') + \gamma \sum_{s' \in S} P(s'|s, a') V_j(s') \right) \right\| \\ &\leq \max_a \left\| \left(R(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) V_k(s') - R(s, a) - \gamma \sum_{s' \in S} P(s'|s, a) V_j(s') \right) \right\| \\ &= \max_a \left\| \gamma \sum_{s' \in S} P(s'|s, a) (V_k(s') - V_j(s')) \right\| \\ &\leq \max_a \left\| \gamma \sum_{s' \in S} P(s'|s, a) \|V_k - V_j\| \right\| \\ &= \max_a \left\| \gamma \|V_k - V_j\| \sum_{s' \in S} P(s'|s, a) \right\| \\ &= \gamma \|V_k - V_j\|\end{aligned}$$



Bellman Backup is a Contraction

$$\begin{aligned}\|BV_k - BV_j\| &= \left\| \max_a \left(R(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) V_k(s') \right) - \max_{a'} \left(R(s, a') + \gamma \sum_{s' \in S} P(s'|s, a') V_j(s') \right) \right\| \\ &\leq \max_a \left\| \left(R(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) V_k(s') - R(s, a) - \gamma \sum_{s' \in S} P(s'|s, a) V_j(s') \right) \right\| \\ &= \max_a \left\| \gamma \sum_{s' \in S} P(s'|s, a) (V_k(s') - V_j(s')) \right\| \\ &\leq \max_a \left\| \gamma \sum_{s' \in S} P(s'|s, a) \|V_k - V_j\| \right\| \\ &= \max_a \left\| \gamma \|V_k - V_j\| \sum_{s' \in S} P(s'|s, a) \right\| \\ &= \gamma \|V_k - V_j\|\end{aligned}$$

If you apply B to two different value functions, distance between value functions shrinks after applying Bellman optimality equation to each



Bellman Operator for Q

$$TQ(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a)} \max_{a'} [Q(s', a')]$$

Since $Q: S \times A \rightarrow \mathbb{R}$, then also $TQ: S \times A \rightarrow \mathbb{R}$



Value Iteration

All of this also holds for V^*

— — —

We can obtain $Q^* = TQ^*$, since Q^* is a fixed-point solution to
 $Q = TQ$



Value Iteration

All of this also holds for V^*

We can obtain $Q^* = TQ^*$, since Q^* is a fixed-point solution to $Q = TQ$

- Initialize $||Q_0||$ in $[0, 1/(1-\gamma)]$ (typically 0)
- Until convergence, for all states and actions:

$$Q_{i+1} = TQ_i$$

We know the Bellman operator is a contraction!



Value Iteration

All of this also holds for V^*

We can obtain $Q^* = TQ^*$, since Q^* is a fixed-point solution to $Q = TQ$

- Initialize $||Q_0||$ in $[0, 1/(1-\gamma)]$ (typically 0)
- Until convergence, for all states and actions:

$$Q_{i+1} = TQ_i$$

$$||Q_{i+1} - Q^*|| = ||TQ_i - TQ^*|| \leq \gamma ||Q_i - Q^*|| \leq \gamma^{i+1} ||Q_0 - Q^*||$$



Policy from Value Iteration

We know that $\pi^*(s) = \operatorname{argmax}_a Q^*(s,a)$, and since $Q_i(s,a) \cong Q^*(s,a)$ we could choose

$$\pi_i(s) = \operatorname{argmax}_a Q_i(s,a)$$



Policy from Value Iteration

— — —

$$\pi_i(s) = \operatorname{argmax}_a Q_i(s, a)$$

What is the quality of such policy? For all states

$$V^{\pi_i}(s) \geq V^*(s) - 2\gamma^t / (1-\gamma) \|Q_0 - Q^*\|$$



Policy from Value Iteration

— — —

$$\pi_i(s) = \operatorname{argmax}_a Q_i(s, a)$$

What is the quality of such policy? For all states

$$V^{\pi_i}(s) \geq V^*(s) - 2\gamma^t / (1-\gamma) \|Q_0 - Q^*\|$$

$$\begin{aligned} V^{\pi^t}(s) - V^*(s) &= Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^*(s)) \\ &= Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^t(s)) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s)) \\ &= \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s)) \\ &\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + Q^*(s, \pi^t(s)) - Q^t(s, \pi^t(s)) + Q^t(s, \pi^*(s)) - Q^*(s, \pi^*(s)) \\ &\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) - 2\gamma^t \|Q^0 - Q^*\|_\infty \end{aligned}$$



Policy from Value Iteration

$$\pi_i(s) = \operatorname{argmax}_a Q_i(s, a)$$

What is the quality of such policy? For all states

$$V^{\pi_i}(s) \geq V^*(s) - 2\gamma^t / (1-\gamma) \|Q_0 - Q^*\|$$

$$V^{\pi^t}(s) - V^*(s) = Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^t(s)) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

Add and subtract

$$= \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + Q^*(s, \pi^t(s)) - Q^t(s, \pi^t(s)) + Q^t(s, \pi^*(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) - 2\gamma^t \|Q^0 - Q^*\|_\infty$$



Policy from Value Iteration

$$\pi_i(s) = \operatorname{argmax}_a Q_i(s, a)$$

What is the quality of such policy? For all states

$$V^{\pi_i}(s) \geq V^*(s) - 2\gamma^t / (1-\gamma) \|Q_0 - Q^*\|$$

$$V^{\pi^t}(s) - V^*(s) = Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= \cancel{Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^t(s))} + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

Expand and get rid of r

$$= \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + Q^*(s, \pi^t(s)) - Q^t(s, \pi^t(s)) + Q^t(s, \pi^*(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) - 2\gamma^t \|Q^0 - Q^*\|_\infty$$



Policy from Value Iteration

$$\pi_i(s) = \operatorname{argmax}_a Q_i(s, a)$$

What is the quality of such policy? For all states

$$V^{\pi_i}(s) \geq V^*(s) - 2\gamma^t / (1-\gamma) \|Q_0 - Q^*\|$$

$$V^{\pi^t}(s) - V^*(s) = Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^t(s)) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

...and you're left with
this

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + Q^*(s, \pi^t(s)) - Q^t(s, \pi^t(s)) + Q^t(s, \pi^*(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) - 2\gamma^t \|Q^0 - Q^*\|_\infty$$



Policy from Value Iteration

$$\pi_i(s) = \operatorname{argmax}_a Q_i(s, a)$$

What is the quality of such policy? For all states

$$V^{\pi_i}(s) \geq V^*(s) - 2\gamma^t / (1-\gamma) \|Q_0 - Q^*\|$$

$$V^{\pi^t}(s) - V^*(s) = Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^t(s)) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + \underbrace{Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))}_{\text{by definition of } Q^*}$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + \underbrace{Q^*(s, \pi^t(s)) - Q^t(s, \pi^t(s))}_{\text{by definition of } Q^*} + Q^t(s, \pi^*(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) - 2\gamma^t \|Q^0 - Q^*\|_\infty$$



Policy from Value Iteration

$$\pi_i(s) = \operatorname{argmax}_a Q_i(s, a)$$

What is the quality of such policy? For all states

$$V^{\pi_i}(s) \geq V^*(s) - 2\gamma^t / (1-\gamma) \|Q_0 - Q^*\|$$

$$V^{\pi^t}(s) - V^*(s) = Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^t(s)) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} (V^{\pi^t}(s') - V^*(s')) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

same as before

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} (V^{\pi^t}(s') - V^*(s')) + Q^*(s, \pi^t(s)) - Q^t(s, \pi^t(s)) + Q^t(s, \pi^*(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} (V^{\pi^t}(s') - V^*(s')) - 2\gamma^t \|Q^0 - Q^*\|_\infty$$



Policy from Value Iteration

$$\pi_i(s) = \operatorname{argmax}_a Q_i(s, a)$$

What is the quality of such policy? For all states

$$V^{\pi_i}(s) \geq V^*(s) - 2\gamma^i / (1-\gamma) \|Q_0 - Q^*\|$$

$$V^{\pi^t}(s) - V^*(s) = Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^t(s)) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} (V^{\pi^t}(s') - V^*(s')) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} (V^{\pi^t}(s') - V^*(s')) + \cancel{Q^*(s, \pi^t(s)) - Q^t(s, \pi^t(s))} + \cancel{Q^t(s, \pi^*(s)) - Q^*(s, \pi^*(s))}$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} (V^{\pi^t}(s') - V^*(s')) - 2\gamma^t \|Q^0 - Q^*\|_\infty$$

just exploit this

$$\|Q_i - Q^*\| \leq \gamma^i \|Q_0 - Q^*\|$$



Policy from Value Iteration

$$\pi_i(s) = \operatorname{argmax}_a Q_i(s, a)$$

What is the quality of such policy? For all states

$$V^{\pi_i}(s) \geq V^*(s) - 2\gamma^i / (1-\gamma) \|Q_0 - Q^*\|$$

$$V^{\pi^i}(s) - V^*(s) = Q^{\pi^i}(s, \pi^i(s)) - Q^*(s, \pi^*(s))$$

$$= Q^{\pi^i}(s, \pi^i(s)) - Q^*(s, \pi^i(s)) + Q^*(s, \pi^i(s)) - Q^*(s, \pi^*(s))$$

$$= \gamma \mathbb{E}_{s' \sim P(s, \pi^i(s))} (V^{\pi^i}(s') - V^*(s')) + Q^*(s, \pi^i(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^i(s))} (V^{\pi^i}(s') - V^*(s')) + Q^*(s, \pi^i(s)) - Q^i(s, \pi^i(s)) + \underbrace{Q^i(s, \pi^*(s)) - Q^*(s, \pi^*(s))}_{\text{and again}}$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^i(s))} (V^{\pi^i}(s') - V^*(s')) - \underbrace{2\gamma^i}_{\text{and again}} \|Q^0 - Q^*\|_\infty$$

$$\|Q_i - Q^*\| \leq \gamma^i \|Q_0 - Q^*\|$$



Policy from Value Iteration

$$\pi_i(s) = \operatorname{argmax}_a Q_i(s, a)$$

What is the quality of such policy? For all states

$$V^{\pi_i}(s) \geq V^*(s) - 2\gamma^t / (1-\gamma) \|Q_0 - Q^*\|$$

$$V^{\pi^t}(s) - V^*(s) = Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^t(s)) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$= \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) + Q^*(s, \pi^t(s)) - Q^t(s, \pi^t(s)) + Q^t(s, \pi^*(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^*(s') \right) - 2\gamma^t \|Q^0 - Q^*\|_\infty$$

repeat and get $1/(1-\gamma)$



Policy from Value Iteration

If we want an ϵ error on the quality of the policy, to determine the number of iterations i we can just solve for it in this equation

$$2\gamma^i / (1-\gamma) \|Q_0 - Q^*\| \leq \epsilon$$

