# Cauchy Sequence and Green's Equation

MDP: $S, A, p(s'|s,a), E(r|s,a,s'), \gamma$

Bellman eqn:

$$① - v^{\pi}(s) = \sum_a \pi(a/s) \sum_{s'} p(s'/s,a)\left[E(r/s,a,s') + \gamma v^{\pi}(s')\right]$$

$$v^*(s) = \max_a \sum_{s'} p(s'/s,a)\left[E(r/s,a,s') + \gamma v^*(s')\right)$$

Assumption: optimal policy is deterministic.

$\therefore$ Instead of $\pi(a/s)$; we will say $\pi(s)$

denotes a deterministic policy

$\hookrightarrow$ the action we take in state $s$

In fact, we can also say that if an MDP has an optimal policy, there exists atleast one deterministic optimal policy.

Finite MDP - $S$ & $A$ are both finite. Expectations i.e $E(r/s,a,s')$ are bounded.

Then we can think of $v^{\pi}$ as a vector with $|S|$ components.

$$\|v\| = \sup_{s \in S} |v(s)|$$
$$= \max(|v(s)|) \quad \forall s \in S$$

max norm.

$\|x\| = 0$ iff $x = 0$
$\|\alpha x\| = \alpha \|x\|$
$\|x+y\| \leq \|x\| + \|y\|$

Complete normed vector space:

Cauchy sequence: $x_1, x_2, x_3, \ldots$
For every +ve real $\epsilon > 0$, $\exists N \in Z^+$ s.t. $\forall m, n > N$
$\|x_m - x_n\| < \epsilon$.

Basically a sequence in which the successive elements are coming closer and closer to each other.

If every cauchy sequence in a normed vector space converges to a pt. in the vector space, then we call the vector space as a complete normed vector space

If every cauchy sequence is convergent, then the vector space is complete.

$$r_{\pi}(s) = \sum_a \pi(a/s) \sum_{s'} p(s'/s,a) E(r/s,a,s')$$

reward expected - one step reward - starting from $s$ & following $\pi$.

$$P_{\pi}(j/s) = \sum_a \pi(a/s) P(j/s,a)$$

prob. that I end in state $j$ in one step, starting from $s$, using policy $\pi$.

$$\sum_{s'} p(s'/s,\pi(s)) . E(r/s,\pi(s),s')$$

$$p(j/s,\pi(s))$$

for deterministic policy.

$r_\pi \rightarrow$ again $|S|$ dimensional vector.

$P_\pi \rightarrow |S| \times |S|$ dimensional stochastic matrix.

$0 \leq \gamma < 1$

all values $\geq 0$

all rows sum up to 1

like each row is a discrete prob. distr.

$$r_\pi + \gamma \underbrace{P_\pi \cdot v^\pi}_{} = v^\pi$$

immediate reward for playing an action

payoff for ending in some state $j$.

$P_\pi$ determines where we land

$v^\pi$ determines the payoff

$v^\pi$ - like a terminal cost. It's the expected return starting from the ~~state~~ state & following policy $\pi$.

Say it's like a one-step problem. We get reward $r_\pi$ for the transition along the way but also a terminal cost $v^\pi$ for landing in some state that we did.

Look at it like: my decision making problem ended with taking one decision. We made the choice according to $\pi$. We want $v^\pi$ to be the total cost.

Essentially, we want to solve:

(2) — $r_\pi + \gamma P_\pi \cdot v^\pi = v^\pi$. which actually comes from algebraically simplifying (1).

This is a matrix eqn.

$\Rightarrow v^\pi = (I - \gamma P_\pi)^{-1} r_\pi$

eigenvalues are all 1

$P_\pi$ is stochastic

$\Rightarrow$ largest eigenvalue is 1

$\Rightarrow \gamma P_\pi$ largest eigenvalue is $< 1$

$\therefore$ The entire matrix $I - \gamma P_\pi$ cannot have a 0 eigenvalue.

$\Rightarrow$ The determinant exists & matrix is invertible and non-zero

$\Rightarrow$ There is a unique soln. for $v^\pi$

Sometimes called the Green's Equation.

Another way to look at this is, take (2). Start with some $v^\pi_0$ - substitute and find $v^\pi_1$. Sub this back and keep going, find $v^\pi_2, v^\pi_3 \ldots$ and finally, it should CONVERGE to $v^\pi$ and that will be unique. We can show this.

# Banach Fixed Point Theorem:

$L_\pi : V \to V$

space of all value fns

A complete normed vector space of all value fn.

$L_\pi v = r_\pi + \gamma P_\pi v$

some element in $V$, a value fn in $V$ not necessarily a value fn, I am when I say it's a value fn, I am implicitly saying that there is a policy for which that is the expected return which need not be the case.

$L_\pi v^\pi = v^\pi$ ] This is what the Bellman eqn says.

→ $v^\pi$ is a fixed point of $L_\pi$. We apply $L_\pi$ on it and it doesn't move.

Banach Fixed Point theorem:

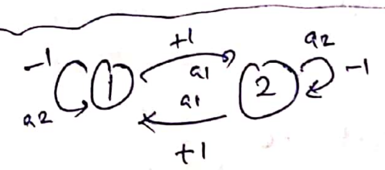
Suppose $U$ is a Banach space (complete normed vector space) and $T : U \to U$ is a contraction mapping; then.

$\underbrace{\text{contraction mapping}}_{Tu \text{ and } Tv \text{ will be close to each other than } u \,\&\, v \text{ were.}}$

~~[illegible]~~

~~for arbitrary~~ $\exists$ an unique $v^*$ in $U$, s.t. $Tv^* = v^*$; and for arbitrary $v^0$ in $U$ (~~[illegible]~~), the sequence $\{v^n\}$ defined by $v^{n+1} = Tv^n = T^{n+1} v^0$ converges to $v^*$.

---




$\pi(1) = a_1$

$\pi(2) = a_1$

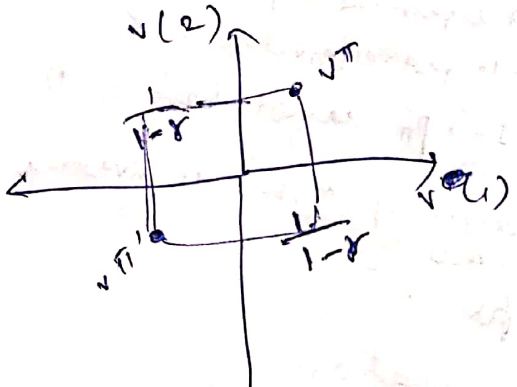
$\pi'(1) = a_2$

$\pi'(2) = a_2$

$v^\pi(1) = 1 + \gamma v^\pi(2)$

$v^\pi(2) = 1 + \gamma v^\pi(1)$

We get $v^\pi(1) = v^\pi(2) = \frac{1}{1-\gamma}$




So for any policy we choose the $\pi$, $\pi'$ and all (there are only 4 possible), we will be in one of the vertices of the square. These 4 points is our $V$.

This is for a deterministic policy. You can write the eqn and do the same thing for a stochastic policy as well.

---

<back to the theorem>

$T$ is a contraction if

$$\|Tu - Tv\| \leq \lambda \|u-v\| \; ; \; 0 \leq \lambda < 1 \;\; \forall u, v \in U$$

we need to show that $L_\pi$ is a contraction mapping. And we get then the rest of the theorem just follows. We will see. Because we know a lot of information. $v^\pi$ is a fixed point, automatically from the theorem, we get that repeatedly applying $L_\pi$ will take us to the fixed point that is also unique as per the theorem

By the way, if $\lambda = 1$, $T$ need not be identity. Just the distance has to be preserved. We can swap $u$ & $v$ also.

Proof:
$$\|v^{n+m} - v^n\| \leq \|v^{n+m} - v^{n+m/2}\| + \|v^{n+m/2} - v^n\|$$

($\triangle$ inequality)

This we can do for any three points. We don't need them to be in a sequence.

So in fact, we can keep applying this technique & the $\triangle$ inequality to say:

$$\|v^{n+m} - v^n\| \leq \sum_{k=0}^{m-1} \|v^{n+k+1} - v^{n+k}\|$$

$$= \sum_{k=0}^{m-1} \|T^{n+k} v^1 - T^{n+k} v^0\| \quad \text{(from the theorem)}$$

$$\leq \sum_{k=0}^{m-1} \lambda^{n+k} \|v^1 - v^0\|$$

$$= \underbrace{\frac{\lambda^n (1-\lambda^m)}{1-\lambda} \|v^1 - v^0\|}$$

this is going to

As $n$ & $m$ becomes large, become smaller and smaller. $\therefore$ We can say the sequence $\{v^n\}$ is cauchy.

Because it's a Banach space, $\{v^n\}$ is convergent.

## Convergence Proof

$$0 \le \|Tv^* - v^*\| \le \|Tv^* - v^n\| + \|v^n - v^*\|$$
$$= \|Tv^* - Tv^{n-1}\| + \|v^n - v^*\|$$
$$\le \lambda \underbrace{\|v^* - v^{n-1}\|}_{\text{will go to } 0} + \underbrace{\|v^n - v^*\|}_{\text{will go to } 0}$$

because $\{v^n\}$ is cauchy. As $n \to \infty$, it converges to $v^*$ and $\|v^* - v^{n-1}\| \to 0$ ||ly $\|v^n - v^*\| \to 0$ too.

We want to prove that $Tv^* = v^*$. But we are assuming that $v^*$ is the convergent pt. of $\{v^n\}$.

$\because \|v^* - v^n\| \to 0$ as $n \to \infty$

$0 \le \|Tv^* - v^*\| \le 0$

$\Rightarrow Tv^* = v^*$.

Let $u^*$ & $v^*$ be two fixed pt's

$\|Tu^* - Tv^*\| \le \lambda \|u^* - v^*\|$

$\Rightarrow \|u^* - v^*\| \le \lambda \|u^* - v^*\|$ and $0 \le \lambda < 1$

$\therefore$ For all such $\lambda$, the only possibility is

$\Rightarrow u^* = v^*$.

So now we have only shown that the Banach fixed point theorem is true.

We still need to show that $L_\pi$ is a contraction.

Let $u$ & $v$ be in $V$

$$L_\pi u(s) = r_\pi(s) + \sum_{j \in S} \gamma P_\pi(j/s)\, u(j)$$
$$L_\pi v(s) = r_\pi(s) + \sum_{j \in S} \gamma P_\pi(j/s)\, v(j)$$

let $L_\pi v(s) > L_\pi u(s)$

$$0 \le L_\pi v(s) - L_\pi u(s) \le r_\pi(s) + \gamma \sum P_\pi(j/s)\, v(j) - r_\pi(s) - \gamma \sum P_\pi(j/s) \cdot u(j)$$
$$= \gamma \sum_j p(j/s)\,(v(j) - u(j))$$
$$\le \gamma \|v - u\| \left(\sum_j p(j/s)\right) \longrightarrow 1$$
$$= \gamma \|v - u\|$$

because of max norm.

$\longrightarrow$ PTO

Illy when $L_\pi v(s) < L_\pi u(s)$ we can do

we get:

$|L_\pi v(s) - L_\pi u(s)| \le \gamma \|v-u\| \quad \forall s$

$\Rightarrow$ $L_\pi$ is a contraction.

Pointwise, it is being drawn close by $\gamma$, so even if the max difference would have come down,

Illy, we have to do this proof for the optimality eqn.

---

## $L_\pi$ Convergence

$L_\pi$ is a contraction.

$V$ – space of all fn.s that are component-wise bounded

space of bounded fn.s.

$\Longrightarrow$ vector notation

$v^* = \max_\pi \{ r_\pi + \gamma P^\pi v^* \}$

$v^*(s) = \max_a \{ E(r|s,a) + \gamma \sum_j p(j|s,a) v^*(j) \}$

call this as some operator $L$

$\therefore Lv \equiv \max_\pi \{ r_\pi + \gamma P^\pi v \}$

not a linear fn. of v. max is there.

claim is that $v^*$ is the fixed point of $L$.

claim is $L$ is a contraction.

claim is $L$ is a unique fixed pt. and if we keep

$\Rightarrow$ $v^*$ is a unique fixed pt. and if we keep applying $L$, we converge at $v^*$.

All this is true because $V$ is a Banach space.

Let $a_s^* \in \arg\max_a \{ E(r|s,a) + \gamma \sum p(j|s,a) v(j) \}$

one of the best actions to take at $s$.

$Lv(s)$

$0 \le Lv(s) - Lu(s) \le ( E(r|s,a_s^*) + \gamma \sum p(j|s,a_s^*) \cdot v(j) )$

PTO contd.

to clarify: (for $L$ & $L_\pi$)

$Lv(s)$ means:

$Lv$ is a fn.

$Lv(s)$ is the output of the fn. for argument $s$.

$L$ is not acting on $v(s)$

$L$ takes a fn. and outputs a fn. $L$ takes $v$ and outputs $Lv$.

we will proceed in the same way we did last time.

$\longrightarrow$ PTO

$$0 \le Lv(s) - Lu(s) \le E(r/s, a_s^*) + \gamma \sum p(j/s, a_s^*) \cdot v(j) - \left[ E(r/s, a_s^*) + \gamma \sum p(j/s, a_s^*) \cdot u(j) \right]$$

here in both cases, we are getting rid of the max by using $a_s^*$ as that is the best action. But in the second term, we are using the max corresponding to $v$ and not $u$. We are using $a_s^*$ in both cases. So the second term, is in a sense, not necessarily using the optimal action. The value will be $\le$ the optimal value and hence when we subtract that term, we get the $\le$ sign and not a $=$ sign.

$$= \gamma \sum p(j/s, a_s^*) [v(j) - u(j)]$$

$$\le \gamma \|v-u\| \underbrace{\cancel{\sum p(j/s, a_s^*)}}_{1} \quad (\because \text{max norm})$$

$$0 \le Lv(s) - Lu(s) \le \gamma \|v-u\|$$

|||ly we can do $Lu(s) - Lv(s)$

we finally get:

$$|Lv(s) - Lu(s)| \le \gamma \|v-u\| \quad \forall s.$$

$\therefore$ L is a contraction.

Now we have a method to solve an MDP!. Start with an arbitrary value fn, keep solving a for v repeatedly, converge at $v^*$. From an optimal value fn., how to recover an optimal policy?

Run

$$a_s^* \in \underset{a}{\arg\max} \left\{ E(r/s,a) + \gamma \sum_j p(j/s,a) \, v^*(j) \right\}$$

pick some action from the argmax set.

optimal value

→ PTO