Reinforcement Learning and Optimal Control

IFT6760C, Fall 2021

Pierre-Luc Bacon

September 14, 2021

Overview

- Infinite-horizon discounted case as random finite horizon
- Markov policies are enough
- Vector notation
- Policy evaluation
  - Policy evaluation operator
  - Application of Neumann's lemma
- Bellman equations
  - Bellman optimality equations
- Markov deterministic policies are enough
- v-improving and conserving decision rules

Infinite Horizon Discounted as Randomized Finite Horizon

Theorem Let $v_{\pi,N}$ denote the value of a policy π in an MDP with a random finite horizon N drawn from a geometric distribution with parameter γ, then:

$$v_{\pi,N}(s) = v_{\pi,\gamma}(s) \quad \forall s \in \mathcal{S} \quad ,$$

where $v_{\pi,\gamma}$ is the value of the policy π under the expected total discounted reward criterion with discount factor γ.

Proof

$$v_{\pi,N}(s) = \mathbb{E}_\pi \left[\mathbb{E}_\gamma \left[\sum_{t=1}^{N} r(S_t, A_t) \,\middle|\, S_1, A_1, \ldots \right] \,\middle|\, S_1 = s \right] =$$

$$= \mathbb{E}_\pi \left[\sum_{n=1}^{\infty} \sum_{t=1}^{n} r(S_t, A_t)(1-\gamma)\gamma^{n-1} \,\middle|\, S_1 = s \right]$$

Assuming that $\sup_{s\in\mathcal{S}} \sup_{a\in\mathcal{A}(s)} |r(s,a)| \leq M < \infty \, \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$:

$$= \mathbb{E}_\pi \left[\sum_{t=1}^{\infty} \sum_{n=t}^{\infty} r(S_t, A_t)(1-\gamma)\gamma^{n-1} \,\middle|\, S_1 = s \right]$$

$$= \mathbb{E}_\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r(S_t, A_t) \,\middle|\, S_1 = s \right] = v_{\pi,\gamma}(s) \quad .$$

Markov Policies are Enough

So far, we haven't made any assumption on whether we consider only Markovian or history-dependent policies

Theorem Let $\pi = (d_1, d_2, \ldots) \in \Pi^{\text{HR}}$, for each $s \in \mathcal{S}$ there exists a $\pi' = (d'_1, d'_2, \ldots) \in \Pi^{\text{MR}}$ such that:

$$P_{\pi'}(S_t = j, A_t = a \mid S_1 = s) = P_{\pi}(S_t = j, A_t = a \mid S_1 = s),$$
$$\forall j \in \mathcal{S}, a \in \mathcal{A}(j), t = 1, 2, \ldots$$

(both policies have the same **occupation measure**)

Proof

The full proof can be found in theorem 5.5.1 of Puterman (1994). The gist of it is that we can construct an equivalent Markov randomized policy by defining each decision rule d'_t as:

$$d'_t(a|j) \triangleq P_\pi \left(A_t = a \,|\, S_t = j, S_1 = s\right), \quad t = 1, 2, \ldots \quad .$$

for each $j \in \mathcal{S}, a \in \mathcal{A}(j)$,

Policy Evaluation

Because of the above theorem, we don't need to consider history-dependent policies when dealing with MDPs.

Therefore:

$$v_\gamma^\star(s) = \sup_{\pi \in \Pi^{HR}} v_{\pi,\gamma}(s) = \sup_{\pi \in \Pi^{MR}} v_{\pi,\gamma}(s), \quad \forall s \in \mathcal{S} \quad .$$

(there is no loss of optimality in searching over the space of Markov policies)

Recap on vector notation

Let $d \in \mathcal{D}^{MR}$, we define $P_d \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ as:

$$[P_d]_{ij} \triangleq \sum_{a\in\mathcal{A}(i)} p(j|i,a)d(a|i) \quad \forall i,j \in \mathcal{S}, a \in \mathcal{A}(i),$$

and $r_d \in \mathbb{R}^{|\mathcal{S}|}$ as:

$$[r_d]_i \triangleq \sum_{a\in\mathcal{A}(i)} r(i,a)d(a|i) \quad .$$

Furthermore, $P_\pi^t \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$:

$$P_\pi^t = P_{d_t} P_{d_{t-1}} \dots P_{d_1} \quad .$$

Policy Evaluation

For generality, assume a **nonstationary** Markov policy π:

$$v_{\pi,\gamma}(i) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r(S_t, A_t) \,\middle|\, S_1 = i\right] = \left[\sum_{t=1}^{\infty} (\gamma P_\pi)^{t-1} r_{d_t}\right]_i .$$

Or recursively:

$$\begin{aligned} v_{\pi,\gamma} &= r_{d_1} + \gamma P_{d_1} \sum_{t=1}^{\infty} (\gamma P_{\pi'})^{t-1} r_{d_{t+1}} \\ &= r_{d_1} + \gamma P_{d_1} v_{\pi',\gamma} \quad . \end{aligned}$$

where $\pi' = (d_2, d_3, \ldots)$

Stationary case

Let $d^\infty = (d, d, \ldots)$ where $d \in \mathcal{D}^{MR}$, then:

$$v_{d^\infty,\gamma} = r_d + \gamma P_d v_{\gamma,d^\infty} \quad .$$

Therefore, $v_{d^\infty,\gamma}$ satisfies the linear system of equations:

$$(I - \gamma P_d)v = r_d \quad .$$

Let L_d be a linear transformation defined by:

$$L_d v \triangleq r_d + \gamma P_d v \quad .$$

Policy evaluation operator

(Assuming the discrete state and action case)

Lemma Let $|r(s,a)| \leq M < \infty, \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$ and $0 \leq \gamma < 1$, then $r_d + \gamma P_d v \in \mathcal{V}, \quad \forall v \in \mathcal{V}, d \in \mathcal{D}^{MR}$.

Proof Both terms are bounded:

1. $\|r_d\| \leq M$ so $r_d \in \mathcal{V}$
2. $\|P_d v\| \leq \|P_d\| \|v\| = \|v\|$ so $P_d v \in \mathcal{V}$.

We can therefore write $v_{d^\infty,\gamma}$ as a fixed point of L_d:

$$v_{d^\infty,\gamma} = L_d v_{d^\infty,\gamma} \quad .$$

Application of Neumann Lemma

Theorem Let $0 \leq \gamma < 1$, then for any stationary policy d^∞, $d \in \mathcal{D}^{MR}$, $v_{d^\infty,\gamma}$ is the solution to:

$$(I - \gamma P_d)v = r_d \quad ,$$

and can be written as:

$$v_{d^\infty,\gamma} = (I - \gamma P_d)^{-1} r_d = \sum_{t=1}^{\infty} (\gamma P_d)^{t-1} r_d \quad .$$

Proof It suffices to apply Neumann's lemma by noting that $\sigma(\gamma P_d) \leq |\gamma| \|P_d\| < 1$ because $0 \leq \gamma < 1$ and P_d is a stochastic matrix (rows sum to 1).

Bellman equations

We will show that optimal policies in the expected total discounted reward setting can be characterized via the nonlinear system of equations:

$$v(s) = \sup_{a \in \mathcal{A}_s} r(s,a) + \gamma \sum_{j \in \mathcal{S}} p(j|s,a) v(j) \quad .$$

These are what we call the **optimality equations** or **Bellman equations**.

Bellman operator

Using the vector notation, we define the corresonding Bellman optimality operator as:

$$\mathcal{L}v \triangleq \sup_{d \in \mathcal{D}^{MD}} r_d + \gamma P_d v \ .$$

When the supremum is attained for all $v \in \mathcal{V}$ (for finite $\mathcal{A}(s)$ for example), then we write instead:

$$Lv \triangleq \max_{d \in \mathcal{D}^{MD}} r_d + \gamma P_d v \ .$$

Deterministic Decision Rules are enough

Theorem $\forall v \in \mathcal{V}, 0 \leq \gamma < 1,$

$$\sup_{d \in \mathcal{D}^{MD}} r_d + \gamma P_d v = \sup_{d \in \mathcal{D}^{MR}} r_d + \gamma P_d v$$

Proof Since $\mathcal{D}^{MD} \subset \mathcal{D}^{MR}$:

$$\sup_{d \in \mathcal{D}^{MD}} r_d + \gamma P_d v \leq \sup_{d \in \mathcal{D}^{MR}} r_d + \gamma P_d v$$

To show that:

$$\sup_{d \in \mathcal{D}^{MD}} r_d + \gamma P_d v \geq \sup_{d \in \mathcal{D}^{MR}} r_d + \gamma P_d v \quad ,$$

Proof (continued)

we apply lemma 4.3.1 in Puterman which says that given a real-valued function $f : \mathcal{X} \to \mathbb{R}$ on a discrete set $\mathcal{X}$ and a probability distribution p over $\mathcal{X}$, then:

$$\sup_{x \in \mathcal{X}} f(x) \geq \sum_{x \in \mathcal{X}} p(x) f(x) \quad .$$

To see this, let $x^\star = sup_{x \in \mathcal{X}} f(x)$ so that:

$$x^\star = \sum_{x \in \mathcal{X}} p(x) x^\star \geq \sum_{x \in \mathcal{X}} p(x) f(x) \quad .$$

Proof (continued)

In our context, we want to establish that:

$$\sup_{a\in\mathcal{A}(s)}\left(r(s,a)+\gamma\sum_{j\in\mathcal{S}}p(j|s,a)v(j)\right)$$
$$\geq\sum_{a\in\mathcal{A}(s)}d(a|s)\left(r(s,a)+\gamma\sum_{j\in\mathcal{S}}p(j|s,a)v(j)\right)\quad.$$

To apply the lemma, we let $f(\cdot)\triangleq r(s,\cdot)+\gamma\sum_{j\in\mathcal{S}}p(j|s,\cdot)v(j)$ and $p(\cdot)\triangleq d(\cdot|s)$ for every $s\in\mathcal{S}$.

We showed that both $\leq$, and $\geq$ hold, therefore we conclude that:

$\sup_{d\in\mathcal{D}^{MD}} r_d+\gamma P_d v=\sup_{d\in\mathcal{D}^{MR}} r_d+\gamma P_d v$.

v-improving and conserving decision rules

Definition A decision rule $d_v \in \mathcal{D}^{MD}$ is said to be **v-improving** if:

$$d_v \in \arg \max_{d \in \mathcal{D}^{MD}} \left(r_d + \gamma P_d v \right), v \in \mathcal{V} \quad .$$

Therefore, if d_v is v-improving then:

$$r_{d_v} + \gamma P_{d_v} v = \max_{d \in \mathcal{D}^{MD}} \left(r_d + \gamma P_d v \right) \qquad \text{or} \qquad L_{d_v} v = Lv \quad .$$

Definition A decision rule $d^\star$ which is $v_\gamma^\star$-improving is also said to be **conserving**.

Therefore $L_{d^\star} v_\gamma^\star = r_{d^\star} + \gamma P_{d^\star} v_\gamma^\star = v_\gamma^\star$.

Component-wise maximization

Because we have a component-wise partial order, whenever you see a sup over the space of policies, this doesn't mean that algorithmically you need to enumerate all policies to implement the given mapping.

That is:

$$\left[\sup_{d \in \mathcal{D}^{MD}} r_d + \gamma P_d v\right]_i = \sup_{a \in \mathcal{A}(i)} \left(r(i,a) + \gamma \sum_{j \in \mathcal{S}} p(j|i,a) v(j) \right) \quad .$$