Markov Decision Processes II

Daisuke Oyama

Topics in Economic Theory

December 17, 2014

Review

- Finite state space S, finite action space A.
- The value of a policy $\sigma \in A^S$:

$$v_\sigma = \sum_{t=0}^{\infty} \beta^t Q_\sigma^t r_\sigma,$$

  which satisfies $v_\sigma = r_\sigma + \beta Q_\sigma v_\sigma$.
- The value function $v^* \in \mathbb{R}^S$:

$$v^*(s) = \sup_{\pi \in \Pi^{\mathrm{M}}} v_\pi(s),$$

  where Π^{M} is the set of Markov plans.
- In the end, for a v^*-greedy policy σ^* we have $v^* = v_{\sigma^*}$.

Review: Operators

- $T_\sigma \colon \mathbb{R}^S \to \mathbb{R}^S$, $\sigma \in A^S$:

$$T_\sigma v = r_\sigma + \beta Q_\sigma v.$$

  v_σ is the unique fixed point of T_σ.

- $T \colon \mathbb{R}^S \to \mathbb{R}^S$:

$$Tv = \max_{\sigma \in A^S} r_\sigma + \beta Q_\sigma v.$$

- By definition, $T_\sigma v \leq Tv$ for any σ and v.
- σ is v-*greedy* if $T_\sigma v = Tv$.

- T_σ and T are monotone.
- $T_\sigma(v + c\mathbf{1}) = T_\sigma v + \beta c\mathbf{1}$ and $T(v + c\mathbf{1}) = Tv + \beta c\mathbf{1}$.
- T_σ and T are β-contractions.
- The unique fixed point of T_σ is v_σ.
  The unique fixed point of T is v^*.
- A v^*-greedy policy (which exists) is an optimal policy.

  $(T_{\sigma^*} v^* = Tv^* = v^* \therefore v^* = v_{\sigma^*}.)$
- For any v, $T_\sigma^n v \to v_\sigma$ and $T^n v \to v^*$ as $n \to \infty$.

Policy Iteration

1. Set $n = 0$.

   Choose any σ_0; or
   choose any v_0 and let σ_0 be a v_0-greedy policy.

2. [Policy evaluation]

   Solve $(I - \beta Q_{\sigma_n})x = r_{\sigma_n}$ for x and let $v^{n+1} = x$.

3. [Policy improvement]

   Compute a v^{n+1}-greedy policy σ_{n+1}, i.e., a σ_{n+1} such that $T_{\sigma_{n+1}} v^{n+1} = T v^{n+1}$.

4. If $\sigma_{n+1} = \sigma_n$, then return $\hat{\sigma} = \sigma_n$ and $\hat{v} = v^{n+1}$.

   Otherwise, let $n = n + 1$ and go to Step 2.

Proposition 1

The policy iteration algorithm terminates in finitely many steps, and $\hat{\sigma}$ is an optimal policy and $\hat{v}$ is the optimal value.

ε-Optimality

Let v^* be the value function.

- v is a δ-*approximation* of v^* if $\|v - v^*\|_\infty < \delta$.
- σ is an ε-*optimal policy* if v_σ is an ε-approximation of v^*.

Error Bounds 1

Lemma 2

For any $v \in \mathbb{R}^S$,

$$\|v^* - Tv\| \leq \frac{\beta}{1-\beta}\|Tv - v\|.$$

Proof

- $\|v^* - Tv\| \leq \|v^* - T^m v\| + \|T^m v - Tv\|$, where

$$\begin{aligned}\text{Second term} &\leq \sum_{k=1}^{m-1} \|T^{k+1}v - T^k v\| \\ &\leq \sum_{k=1}^{m-1} \beta^k \|Tv - v\| = \frac{\beta - \beta^m}{1-\beta}\|Tv - v\|.\end{aligned}$$

Let $m \to \infty$.

Lemma 3

For any $v \in \mathbb{R}^S$ and any Tv-greedy policy σ,

$$\|v_\sigma - Tv\|_\infty \leq \frac{\beta}{1-\beta}\|Tv - v\|_\infty.$$

Proof

- Denote $u = Tv$.

  Recall that $v_\sigma = T_\sigma v_\sigma$ and $T_\sigma u = Tu$.

- Then,

$$\begin{aligned}
\|v_\sigma - u\|_\infty &= \|T_\sigma v_\sigma - u\|_\infty \\
&\leq \|T_\sigma v_\sigma - Tu\|_\infty + \|Tu - u\|_\infty \\
&= \|T_\sigma v_\sigma - T_\sigma u\|_\infty + \|Tu - Tv\|_\infty \\
&\leq \beta\|v_\sigma - u\|_\infty + \beta\|u - v\|_\infty.
\end{aligned}$$

  Rearranging terms yields the desired inequality.

Proposition 4

For any $v \in \mathbb{R}^S$ and any Tv-greedy policy σ,

$$\|v_\sigma - v^*\|_\infty \leq \frac{2\beta}{1-\beta}\|Tv - v\|_\infty.$$

Proof

- By the previous two lemmas,

$$\begin{aligned}\|v_\sigma - v^*\|_\infty &\leq \|v_\sigma - Tv\|_\infty + \|Tv - v^*\|_\infty \\ &\leq \frac{\beta}{1-\beta}\|Tv - v\|_\infty + \frac{\beta}{1-\beta}\|Tv - v\|_\infty.\end{aligned}$$

Error Bounds 2

For $x \in \mathbb{R}^S$, write $m(x) = \min_i x_i$ and $M(x) = \max_i x_i$.

Lemma 5

For any $v \in \mathbb{R}^S$ and any v-greedy policy σ,

$$\begin{aligned} & v + \frac{1}{1-\beta} m(Tv - v)\mathbf{1} \leq Tv + \frac{\beta}{1-\beta} m(Tv - v)\mathbf{1} \\ & \leq v_\sigma \leq v^* \\ & \leq Tv + \frac{\beta}{1-\beta} M(Tv - v)\mathbf{1} \leq v + \frac{1}{1-\beta} M(Tv - v)\mathbf{1}. \end{aligned}$$

For $x \in \mathbb{R}^S$, write

$$\operatorname{span}(x) = M(x) - m(x) \ (= \max_i x_i - \min_i x_i).$$

Proposition 6

For any $v \in \mathbb{R}^S$ and any v-greedy policy σ,

$$\|v^* - v_\sigma\|_\infty \leq \frac{\beta}{1-\beta} \operatorname{span}(Tv - v),$$

and

$$\left\| v^* - \left(Tv + \frac{\beta}{1-\beta} \frac{m(Tv-v) + M(Tv-v)}{2} \mathbf{1} \right) \right\|_\infty$$
$$\leq \frac{1}{2} \frac{\beta}{1-\beta} \operatorname{span}(Tv - v).$$

Proof of Lemma 5

- Take any $v \in \mathbb{R}^n$, and let σ be a v-greedy policy: $T_\sigma v = Tv$.

  (Recall $m(x) = \min_i x_i$ and $M(x) = \max_i x_i$.)

- Clearly, $T_\sigma v = Tv \geq v + m(Tv - v)\mathbf{1}$.

  By the properties of T_σ,

$$\begin{aligned} T_\sigma^2 v &\geq T_\sigma(v + m(Tv - v)\mathbf{1}) \\ &= T_\sigma v + \beta m(Tv - v)\mathbf{1} \geq v + (1 + \beta)m(Tv - v)\mathbf{1}, \end{aligned}$$

$$\begin{aligned} T_\sigma^3 v &\geq T_\sigma(v + (1 + \beta)m(Tv - v)\mathbf{1}) \\ &= T_\sigma v + \beta(1 + \beta)m(Tv - v)\mathbf{1} \\ &\geq v + (1 + \beta + \beta^2)m(Tv - v)\mathbf{1}, \end{aligned}$$

$\cdots$.

- We thus have

$$\begin{aligned} T_\sigma^n v &\geq T_\sigma v + (\beta + \cdots + \beta^{n-1}) m(Tv - v)\mathbf{1} \\ &\geq v + (1 + \beta + \cdots + \beta^{n-1}) m(Tv - v)\mathbf{1}. \end{aligned}$$

- Letting $n \to \infty$, we have

$$v_\sigma \geq T_\sigma v + \frac{\beta}{1-\beta} m(Tv - v)\mathbf{1} \geq v + \frac{1}{1-\beta} m(Tv - v)\mathbf{1}.$$

  Note that $T_\sigma v = Tv$.

- By a similar procedure, we have

$$v^* \leq Tv + \frac{\beta}{1-\beta} M(Tv - v)\mathbf{1} \leq v + \frac{1}{1-\beta} M(Tv - v)\mathbf{1}.$$

- Note finally that $v^* \geq v_\sigma$.

Remarks

- Similar estimates with $-\|Tv - v\|_\infty$ and $\|Tv - v\|_\infty$ in place of $m(Tv - v)$ and $M(Tv - v)$ hold.

  (Start with $-\|Tv - v\|_\infty \mathbf{1} \leq Tv - v \leq \|Tv - v\|_\infty \mathbf{1}$.)

- Since $-m(x) \leq \|x\|_\infty$ and $M(x) \leq \|x\|_\infty$, we have $\mathrm{span}(Tv - v) \leq 2\|Tv - v\|_\infty$.

Error Bounds and Termination Conditions

	Bound 1	Bound 2
Value iteration	$*$	
Modified policy iteration		$*$

Value Iteration with Norm Bounds

Specify $\varepsilon > 0$.

1. Set $n = 0$.

   Choose any v_0.

2. Let $v^{n+1} = Tv^n$.

3. If $\|v^{n+1} - v^n\|_\infty < \frac{1-\beta}{2\beta}\varepsilon$, then return $\hat{v} = v^{n+1}$ and a $\hat{v}$-greedy policy $\hat{\sigma}$.

   Otherwise, let $n = n + 1$ and go to Step 2.

Proposition 7

Given an $\varepsilon > 0$, the value iteration algorithm as described terminates in finitely many steps, and

- $\hat{\sigma}$ *is an ε-optimal policy and*
- $\hat{v}$ *is an $\frac{\varepsilon}{2}$-approximation of v^*.*

Modified Policy Iteration with Span Seminorm Bounds

Specify $\varepsilon > 0$ and $k \geq 1$.

1. Set $n = 0$. Choose any v_0.

2. [Policy improvement]

   Compute a v^n-greedy policy σ_{n+1}, i.e., a σ_{n+1} such that $T_{\sigma_{n+1}} v^n = T v^n$.

   Compute also $u^n = T v^n$ $(= T_{\sigma_{n+1}} v^n)$.

3. If $\operatorname{span}(u^n - v^n) < \frac{1-\beta}{\beta}\varepsilon$, then return $\hat{\sigma} = \sigma_{n+1}$ and $\hat{v} = u^n + \frac{\beta}{1-\beta}\frac{m(u^n - v^n) + M(u^n - v^n)}{2}\mathbf{1}$.

   Otherwise, go to the next step.

4. [Partial policy evaluation]

   Let $v^{n+1} = (T_{\sigma_{n+1}})^k v^n = (T_{\sigma_{n+1}})^{k-1} u^n$.

   Let $n = n + 1$ and go to Step 2.

Fact 1

For modified policy iteration, as $n \to \infty$, $v^n \to v^*$ *and hence* $\operatorname{span}(Tv^n - v^n) \to 0$.

Proposition 8

Given an $\varepsilon > 0$, *the modified policy iteration algorithm as described terminates in finitely many steps, and*

- $\hat{\sigma}$ *is an* ε*-optimal policy and*
- $\hat{v}$ *is an* $\frac{\varepsilon}{2}$*-approximation of* v^*.

References

- D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice Hall, 1987.
- M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley-Interscience, 2005.