Markov Decision Processes (Part II)

Lodewijk Kallenberg

kallenberg@math.leidenuniv.nl

Outline of Part II

- Week 1: Short Review Part I; Sensitive optimality criteria: theory
- Week 2: Sensitive optimality criteria: methods
- Week 3: Applications: gambling; optimal stopping; replacement
- Week 4: Multi-armed bandit problems; stochastic games

Course Material

- Sheets (download from www.lnmb.nl/courses/MDP.html)
- Lecture Notes (download from www.lnmb.nl/courses/MDP.html)
- Book: M.L. Puterman, *Markov decision processes*, Wiley, 1994
- Papers (for references see Lecture Notes)

Review of Part I (1)

The Model

- Discrete time stochastic process $t = 1, 2, \ldots$
- State space X (finite); $N = \#X$
- Action sets $A(x)$, $x \in X$ (finite)
- Rewards $r(x, a)$ (stationary)
- Transition probabilities $p(y \mid x, a)$ (stationary)
- $X_t =$ state at time t (stochastic variable)
- $Y_t =$ action at time t (stochastic variable)

Review of Part I (2)

Utility functions:

- Total discounted expected rewards ($\alpha \in (0,1)$ *discount factor*)
  $v_s^{\alpha}(\pi) = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{x,a} \mathbb{P}_{s,\pi}\{X_t = x, A_t = a\} \cdot r(x,a)$.
  For $\pi \in \Pi^S$ (stationary policy): $v^{\alpha}(\pi) = \{I - \alpha P(\pi)\}^{-1} r(\pi)$.

- Long-run average expected rewards
  $g_s(\pi) = liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{x,a} \mathbb{P}_{s,\pi}\{X_t = x, A_t = a\} \cdot r(x,a)$.
  For $\pi \in \Pi^S$ (stationary policy): $g(\pi) = P^*(\pi) r(\pi)$.

Property:

For both criteria deterministic optimal policies exist.

Proof: see Puterman p.153 and p.451 respectively.

Review of Part I (3)

Markov Chains

- $lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k = lim_{\alpha\uparrow 1} (1-\alpha) \sum_{k=0}^{\infty} (\alpha P)^k = P^*$, the *stationary matrix* P^*.
- The stationary matrix satisfies $P^* = P^*P^* = PP^* = P^*P$.
- The matrix $I - P + P^*$ is nonsingular and $Z = (I - P + P^*)^{-1} = lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \{\sum_{i=0}^{k-1} (P - P^*)^i\}$ is the *fundamental matrix*.
- $D = Z - P^*$ is the *deviation matrix* and satisfies $De = 0$ and $P^*D = DP^* = (I - P)D + P^* - I = D(I - P) + P^* - I = 0$.
- For $\pi \in \Pi^S$ (stationary policy): $g(\pi) = P^*(\pi) r(\pi)$.

Proofs: Veinott (1974) (reference in Lecture Notes).

Sensitive criteria (1): Laurent Expansion

- For $\pi \in \Pi^S$ (stationary policy), we have the following expansion:

  $\alpha v^\alpha(\pi) = \frac{\alpha}{1-\alpha} g(\pi) + u^0(\pi) + \frac{1-\alpha}{\alpha} u^1(\pi) + \cdots + (\frac{1-\alpha}{\alpha})^k u^k(\pi) + \cdots,$

  where $u^k(\pi) = (-1)^k D(\pi)^{k+1} r(\pi),\ k \geq 0.$

- *Proof:*

  The proof is based on the property that $v^\alpha(\pi)$ is the unique solution of the linear system $r(\pi) + \alpha P(\pi)x = x$ and on the properties of the matrices P^* and D.

  For details see Miller and Veinott (1969) (reference in Lecture Notes).

- *Corollaries*:

  1. $v^\alpha(\pi) = \frac{g(\pi)}{1-\alpha} + u^0(\pi) + \varepsilon(1-\alpha)$, where $lim_{\alpha \to 1}\, \varepsilon(1-\alpha) = 0.$
  2. $g(\pi) = lim_{\alpha \to 1}\, (1-\alpha) v^\alpha(\pi).$

Sensitive criteria (2): n-discount optimality ($n = -1, 0, 1 \ldots$)

- π^* is n-discount optimal $\Leftrightarrow$ $lim_{\alpha \uparrow 1} \frac{1}{(1-\alpha)^n} \{v^\alpha(\pi^*) - v^\alpha(\pi\} \geq 0 \; \forall \; \pi$.
- There exists a deterministic n-discount optimal policy ($\forall \, n = -1, 0, \ldots$).

  Proof: see Veinott (1969) (reference in Lecture Notes).
- $n = -1$: $\pi^* \in \Pi^D$ is (-1)-discount optimal $\Leftrightarrow$

  $lim_{\alpha \uparrow 1} (1 - \alpha) \{v^\alpha(\pi^*) - v^\alpha(\pi\} \geq 0 \; \forall \; \pi \in \Pi^D \; \Leftrightarrow$

  $\forall \; \pi \in \Pi^D$: $g(\pi^*) - g(\pi) \geq 0 \Leftrightarrow$ π^* is average optimal.
- $n = 0$: $\pi^* \in \Pi^D$ is 0-discount optimal $\Leftrightarrow$

  $lim_{\alpha \uparrow 1} \{v^\alpha(\pi^*) - v^\alpha(\pi\} \geq 0 \; \forall \; \pi \in \Pi^D \; \Leftrightarrow$

  $\forall \; \pi \in \Pi^D$: $g(\pi^*) \geq g(\pi)$ and if $g(\pi^*) = g(\pi)$, then $u^0(\pi^*) \geq u^0(\pi)$.

  0-discount optimality is also called *bias optimality*.
- Proofs for $n = -1$ and $n = 0$: directly from the Laurent expansion.

Sensitive criteria (3): n-average optimality ($n = -1, 0, 1 \ldots$)

- π^* is n-average optimal $\Leftrightarrow$

  $liminf_{T\to\infty} \frac{1}{T} \{v^{n,T}(\pi^*) - v^{n,T}(\pi)\} \geq 0 \; \forall \; \pi,$

  where $v^{n,t}(\pi) = \begin{cases} \sum_{t=1}^{T} P^{t-1}(\pi) r(\pi) & n = -1 \\ \sum_{t=1}^{T} v^{n-1,t}(\pi) & n \geq 0 \end{cases}$

- *Theorem:* $\forall n$: n-discount optimality equivalent to n-average optimality.

  Proof: see Sladky (1974) (reference in Lecture Notes).

- $n = -1$:

  $liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \{P^{t-1}(\pi^*) r(\pi^*) - P^{t-1}(\pi) r(\pi)\} \geq 0 \; \forall \pi$

  $\Leftrightarrow$ average optimality.

- $n = 0$: 0-average optimality $\Leftrightarrow$

$0 \leq liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{t} \{P^{s-1}(\pi^*)r(\pi^*) - P^{s-1}(\pi)r(\pi)\} \; \forall \; \pi.$

$\sum_{s=1}^{t} P^{s-1}(\pi)r(\pi) = tg(\pi) + u^0(\pi) - P^t(\pi)D(\pi)r(\pi) \;\Rightarrow$

$$\begin{aligned} 0 \leq\; & liminf_{T\to\infty} \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \sum_{s=1}^{t} \{P^{s-1}(\pi^*)r(\pi^*) - P^{s-1}(\pi)r(\pi)\} \\ =\; & liminf_{T\to\infty} [\tfrac{1}{2}(T+1)\{g(\pi^*) - g(\pi)\} + \{u^0(\pi^*) - u^0(\pi)\} \\ & -\tfrac{1}{T} \textstyle\sum_{t=1}^{T} \{P^t(\pi^*)D(\pi^*)r(\pi^*) - P^t(\pi)D(\pi)r(\pi)\}] \end{aligned}$$

$lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P^t(\pi)D(\pi)r(\pi) = P^*(\pi)D(\pi)r(\pi) = 0 \; \forall \; \pi \;\Rightarrow$

0-average optimality $\Leftrightarrow g(\pi^*) \geq g(\pi)$ and $\forall \pi$ with $g(\pi^*) = g(\pi)$:

$u^0(\pi^*) \geq u^0(\pi)$: 0-average optimality equivalent to bias optimality.

Sensitive criteria (3): Blackwell optimality

- π^* Blackwell optimal $\Leftrightarrow$ π^* discounted optimal $\forall \alpha \in [\alpha_0, 1)$ for some α_0.
- There exists a deterministic Blackwell optimal policy π^0.

  Proof (outline):

  $v^\alpha(\pi)$ is the unique solution of the linear system $r(\pi) + \alpha P(\pi)x = x$.
  Solve the system by Cramer's rule: $v_i^\alpha(\pi)$ is rational in α.

  Suppose that a deterministic Blackwell optimal policy does not exist.

  Then for each increasing sequence $\{\alpha_n,\, n \in \mathbb{N}\}$ with $lim_{n\to\infty}\, \alpha_n = 1$,

  there exists a sequence of deterministic policies $\{\pi_n,\, n \in \mathbb{N}\}$ such that

  $v^\alpha = v^\alpha(\pi_{n+1}) > v^\alpha(\pi_n)$ for $\alpha = \alpha_{n+1},\, n \in \mathbb{N}$.

  Because Π^D is finite, there are two policies, say π^1 and π^2,

such that for some subsequence $\{\alpha_{n_k},\, k \in \mathbb{N}\}$

$$\begin{cases} v^{\alpha_{n_k}}(\pi^1) > v^{\alpha_{n_k}}(\pi^2) \text{ for } k = 1, 3, 5, \ldots \\ v^{\alpha_{n_k}}(\pi^1) < v^{\alpha_{n_k}}(\pi^2) \text{ for } k = 2, 4, 6, \ldots \end{cases}$$

Let $h(\alpha) = v^{\alpha}(\pi^1) - v^{\alpha}(\pi^2)$: $h_i(\alpha)$ continuous rational function in α with an infinite number of zero's: $h_i(\alpha) \equiv 0$: contradiction.

- π^0 Blackwell optimal $\rightarrow$ π^0 n-discount optimal $\forall n = -1, 0, \ldots$.

  Proof: Let π^0 is Blackwell optimal $\rightarrow$

  $\{v^{\alpha}(\pi^0) - v^{\alpha}(\pi)\} \geq 0 \; \forall \alpha \in [\alpha_0, 1)$ and $\forall \pi \; \rightarrow$

  $\frac{1}{(1-\alpha)^n} \{v^{\alpha}(\pi^0) - v^{\alpha}(\pi\} \geq 0 \; \forall \alpha \in [\alpha_0, 1)$ and $\forall \pi \; \rightarrow$

  $lim_{\alpha \uparrow 1} \frac{1}{(1-\alpha)^n} \{v^{\alpha}(\pi^0) - v^{\alpha}(\pi\} \geq 0 \; \forall \pi$.

- π^0 $(N-1)$-discount optimal $\Leftrightarrow$ π^0 Blackwell optimal.

  Proof: see the papers of Veinott (1969) and Miller & Veinott (1969)

Exercise 1: Stationary, fundamental and deviation matrix

Deadline: February 27, 2006.

Let P be a Markov matrix with $P^*,\ Z,\ D$ as corresponding stationary, fundamental and deviation matrices.

Show the following properties:

a. $lim_{\alpha\uparrow 1}\,(1-\alpha)\sum_{k=0}^{\infty}\alpha^k(P^k-P^*)=0$;

b. $lim_{\alpha\uparrow 1}\sum_{k=0}^{\infty}\alpha^k(P-P^*)^k=Z$;

c. $lim_{\alpha\uparrow 1}\sum_{k=0}^{\infty}\alpha^k(P^k-P^*)=D$.

Hint:

Use that convergence in Cesaro sense implies convergence in Abel sense (for a proof see: J.Korevaar, *Tauberian Theory*, Springer 2004, p. 4)

Exercise 2: Laurent expansion

Deadline: February 27, 2006.

Consider the following model:

$X = \{1, 2\};\ A(1) = A(2) = \{1, 2\}.\ p(1|1, 1) = 1,\ p(2|1, 1) = 0;$

$p(1|1, 2) = 0,\ p(2|1, 2) = 1;\ p(1|2, 1) = 1;\ p(2|2, 1) = 0;\ p(1|2, 2) = \frac{1}{2};$

$p(2|2, 2) = \frac{1}{2}.\ r(1, 1) = 4,\ r(1, 2) = 0;\ r(2, 1) = 8,\ r(2, 2) = 6.$

Let π_{22} the deterministic policy that takes action 2 in state 1 and state 2.

a. Compute for the policy π_{22} the stationary matrix $P^*(\pi_{22})$, the average reward vector $g(\pi_{22})$ and the deviation matrix $D(\pi_{22})$.

b. Determine the vectors $u^k(\pi_{22})$ for $k = -1, 0, 1, \dots$ and the Laurent expansion for $v_1^\alpha(\pi_{22})$, i.e. determine the coefficients a_k $(k = -1, 0, 1, \dots)$ such that $v_1^\alpha(\pi_{22}) = \sum_{k=-1}^{\infty} a_k(1 - \alpha)^k$.

Exercise 3: Césaro and Abel convergence

Deadline: February 27, 2006.

A sequence $a_1, a_2, a_3, \ldots$ is Césaro convergent if $lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n} a_i$ exists; the sequence is said to be Abel convergent if $lim_{\alpha\uparrow 1}(1-\alpha)\sum_{i=0}^{\infty} \alpha^i a_i$ exists.

a. Prove that ordinary convergence implies Césaro convergence.

b. Prove that ordinary convergence implies Abel convergence.

c. Give a counterexample that Césaro convergence doesn't imply, in general, ordinary convergence.

d. Give a counterexample that Abel convergence doesn't imply, in general, ordinary convergence.

e. Give a counterexample that Abel convergence doesn't imply, in general, Césaro convergence.