Optimality and Approximations for a Variable Discounted Infinite-Horizon Control Problem

Fernando M. Vidal , Eugenio M. Della Vecchia

Enero de 2016

Universidad Nacional de Rosario

Contenido

Markov Decision Problems

- Discrete time system.
- An action chosen which makes
  - An instant reward.
  - A probability distribution on the state space.
- Rewards and distributions may depend on the decision epoch.

Elements of the **MDP**

$$\mathscr{M} := (\mathscr{S}, \mathscr{A}, \{\mathscr{A}_s : s \in \mathscr{S}\}, \{Q_t\}, \{r_t\}, \{\lambda_t\})$$

- $\mathscr{S}$, *state* space.
- $\mathscr{A}$, *action* space.
- $\mathscr{A}_s$. $\mathbb{K} = \{(s,a) : s \in \mathscr{S}, a \in \mathscr{A}_s\}$.
- Q_t, *transition law* on $\mathscr{S}$, $Q_t^{a,b}(z|s)$.
- r_t, *reward function* . $r_t^{a,b}(s)$.
- λ_t, *discount factor*.

Strategies

- History up to time n,

$$h_n = (s_0, a_0, s_1, a_1, \ldots, s_{n-1}, a_{n-1}, s_n).$$

- Markov strategy $\pi = \{f_n\}, f_n \in \mathbb{P}(\mathscr{A} | H_n)$,

$$\mathrm{P}(a|h_n) = \mathrm{P}(a|s_n)\ .$$

- Stationary strategy: $\pi = \{f, f, \ldots\} = f$.
- Π, Π_{stat}.

Performance Criteria

$$V_N^{\pi}(s) = \mathbb{E}_s^{\pi}\left[r_0^{A_0}(s) + \sum_{t=0}^{N-1} \lambda_{t-1}\ r_t^{A_t}(S_t) + \lambda_{N-1}\ r_N(S_N)\right] .$$

With the convention $\lambda_{-1} = 1$, .

$$V_N^{\pi}(s) = \mathbb{E}_s^{\pi}\left[\sum_{t=0}^{N-1} \lambda_{t-1}\ r_t^{A_t}(S_t) + \lambda_{N-1}\ r_N(S_N)\right] .$$

$$V^{\pi}(s) = \mathbb{E}_s^{\pi}\left[\sum_{t=0}^{\infty} \lambda_{t-1}\ r_t^{A_t}(S_t)\right] ,$$

Objectives on the **MDP**

- Infinite-horizon,

$$\pi^*(s) \in \arg\max_{\pi} V^{\pi}(s) \ .$$

- π^* optimal.
- $V^*(s) = \sup_{\pi \in \Pi} V^{\pi}(s)$ value function.
- $\varepsilon > 0$, π_ε, ε-optimal strategies

$$V^*(s) - V^{\pi_\varepsilon}(s) \leqq \varepsilon \ .$$

- The same in finite horizon problems.

Works with Constant Discounts

- Discrete spaces,
  - Kallenberg L., *Finite state and action MDPS. Handbook of Markov Decision Processes. Methods and applications*, 2002.
  - Puterman L., *Markov Decision Processes* , 2005.
- General spaces
  - Ross, S., *Applied Probability Models with Optimization Applications*, 1970.
  - Hernández-Lerma O., Lasserre J.B., *Discrete-Time Markov Control Processes*, 1996.

Assumptions

Hipótesis

(a) $\mathscr{S}$, *Borel.*

(b) $\mathscr{A}_s$, *compact.*

(c) $r_t^{\cdot}(s)$, *upper semicontinuous on* $\mathscr{A}_s$.

(d) $|r_t^a(s)| \leqq M_t$ *and* $|r_N(s)| \leqq M_N$.

(e) v *bounded and meassurable* $a \mapsto \int v(y) Q_t^a(dz|s)$ *continuous on* $\mathscr{A}_s$.

Idea

- Constant discount, particular case, $\lambda_\tau = (\alpha)^\tau$.
-
$$(Tv)(s) = \sup_{a \in \mathscr{A}_s} \left\{ r_t^a(s) + \alpha \int_{\mathscr{S}} v(z) Q_t^a(dz|s) \right\} ,$$
- $\alpha = \frac{\alpha^\tau}{\alpha^{\tau-1}}$.
- Propose dynamic programming with factors $\frac{\lambda_\tau}{\lambda_{\tau-1}}$.

Finite Horizon Result

Theorem 1

V_0, V_1,...,V_N functions on $\mathscr{S}$

$$V_N(s) = r_N(s) \; ,$$

$$V_n(s) = \sup_{a \in \mathscr{A}_s} \left\{ r_n^a(s) + \frac{\lambda_n}{\lambda_{n-1}} \int_{\mathscr{S}} V_{n+1}(z) Q_n^a(dz|s) \right\} \; . \tag{1}$$

- *There exist f_n^*, which in s in time n, maximizes* (1).
- $\pi^* = \{f_0^*, f_1^*, \ldots, f_{N-1}^*\}$ *optimal.*
- $V_N^* = V_0$.

Augmented Model

Hipótesis

(a) $|r_t^a(s)| \leqq M.$

(b) $\lambda_t \leqq \rho \, \lambda_{t-1}, \, \rho < 1 \, .$

Augmented Model

Hipótesis

(a) $|r_t^a(s)| \leqq M$.

(b) $\lambda_t \ \leqq \ \rho \ \lambda_{t-1}$, $\rho < 1$.

$$\tilde{\mathscr{M}} := (\tilde{\mathscr{S}}, \tilde{\mathscr{A}}, \{\tilde{\mathscr{A}}_{\tilde{s}} : \tilde{s} \in \tilde{\mathscr{S}}\}, \tilde{Q}, \tilde{r}, \{\tilde{\lambda}_{\tilde{s}}, \tilde{s} \in \tilde{\mathscr{S}}\})$$

- $\tilde{\mathscr{S}} = \mathscr{S} \times \mathbb{N}_0$
- $\tilde{\mathscr{A}} = \mathscr{A}$, $\tilde{\mathscr{A}}_{(s,\tau)} = \mathscr{A}_s$.
- $\tilde{r}^a(s,\tau) = r_\tau^a(s)$.
- $\tilde{Q}^a(z,\tau'|s,\tau) = \begin{cases} Q^a(z|s) & \text{if } \ \tau' = \tau + 1 \\ 0 & \text{si no} \end{cases}$
- $\tilde{\lambda}_{(s,\tau)} = \lambda_\tau$.

Augmented Model

- $\tilde{\Pi}$. $\tilde{\Pi}_{\text{stat}}$.
- 1-1 correspondence strategies stationary on $\tilde{\mathscr{M}}$ and Markov on $\mathscr{M}$.

Augmented Model

- $\tilde{\Pi}$. $\tilde{\Pi}_{\text{stat}}$.
- 1-1 correspondence strategies stationary on $\tilde{\mathscr{M}}$ and Markov on $\mathscr{M}$.

$$\begin{aligned}
\tilde{V}^{\tilde{\pi}}(s,\tau) &:= \frac{1}{\tilde{\lambda}_{(s,\tau-1)}}\mathbb{E}^{\tilde{\pi}}_{(s,\tau)}\left[\tilde{\lambda}_{(s,\tau-1)}\tilde{r}^{A_\tau}(s,\tau)+\sum_{t=\tau+1}^{\infty}\tilde{\lambda}_{\tilde{S}_{t-1}}\tilde{r}^{A_t}(\tilde{S}_t)\right] \\
&= \tilde{r}^{\tilde{f}_\tau}(s,\tau)+\mathbb{E}^{\tilde{\pi}}_{(s,\tau)}\left[\sum_{t=\tau+1}^{\infty}\frac{\tilde{\lambda}_{\tilde{S}_{t-1}}}{\tilde{\lambda}_{(s,\tau-1)}}\tilde{r}^{A_t}(\tilde{S}_t)\right].
\end{aligned}$$

- Optimal strategies and value function .
- $\tilde{V}^*(s,\tau)$ expected optimal value from time τ, in s.
- $\tilde{V}^*(s,0)=V^*(s)$.

Dynamic Programming Operators

- $\mathscr{M}(\tilde{\mathscr{S}})$, $\mathscr{B}(\tilde{\mathscr{S}})$.
- $||v||_\infty = \sup_{(s,\tau)\in\tilde{\mathscr{S}}} \left|v(s,\tau)\right|$, $(\mathscr{B}(\tilde{\mathscr{S}}), ||\cdot||_\infty)$ Banach space.
- $T, T^f : \mathscr{B}(\tilde{\mathscr{S}}) \to \mathscr{B}(\tilde{\mathscr{S}})$.

$$(Tv)(s,\tau) = \sup_{a\in\mathscr{A}_s} \left\{ \tilde{r}^a(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathscr{S}}} v(z,\tau')\tilde{Q}^a(dz,\tau'|s,\tau) \right\} ,$$

$$(T^{\tilde{f}}v)(s,\tau) = \tilde{r}^{\tilde{f}}(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathscr{S}}} v(z,\tau')\tilde{Q}^{\tilde{f}}(dz,\tau'|s,\tau) .$$

Results

Lema 1

T, $T^{\tilde{f}}$ *monotone and contractive on* $\mathscr{B}(\tilde{\mathscr{S}})$, *modulus* ρ.

Lema 2

$\tilde{f} \in \tilde{\Pi}_{\text{stat}}$, $\tilde{V}^{\tilde{f}}$ *unique fixed point of* $T^{\tilde{f}}$ *on* $\mathscr{B}(\tilde{\mathscr{S}})$.

Results

Lema 1

T, $T^{\tilde{f}}$ monotone and contractive on $\mathscr{B}(\tilde{\mathscr{S}})$, modulus ρ.

Lema 2

$\tilde{f} \in \tilde{\Pi}_{\text{stat}}$, $\tilde{V}^{\tilde{f}}$ unique fixed point of $T^{\tilde{f}}$ on $\mathscr{B}(\tilde{\mathscr{S}})$.

Theorem 2

- *$\tilde{V}^*$ unique fixed point of T,*

$$\tilde{V}^*(s,\tau) = \sup_{a \in \mathscr{A}_s} \left\{ \tilde{r}^a(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathscr{S}}} \tilde{V}^*(z,\tau')\tilde{Q}^a(dz,\tau'|s,\tau) \right\} .$$

- *There exists $\tilde{f}^* \in \tilde{\Pi}_{\text{stat}}$, which in $(s,\tau) \in \tilde{\mathscr{S}}$ maximizes rhs.*
- *$\tilde{f}^*$ optimal,*

$$T^{\tilde{f}^*}\tilde{V}^* = \tilde{V}^* , \text{ and } \tilde{V}^{\tilde{f}^*} = \tilde{V}^* .$$

Policy Iteration Algorithm

PI1 $n = 0, \tilde{f}_0 \in \tilde{\Pi}_{\text{stat}}$.

PI2 $\tilde{u}_n = \tilde{V}^{f_n}$, punto fijo de $T^{\tilde{f}_n}$.

PI3 $\tilde{f}_{n+1} \in \tilde{\Pi}_{\text{stat}}, T^{\tilde{f}_{n+1}} \tilde{u}_n = T \tilde{u}_n$.

PI4 $n := n + 1$, ir a **PI2**.

Policy Iteration Algorithm

PI1 $n = 0, \tilde{f}_0 \in \tilde{\Pi}_{\text{stat}}$.

PI2 $\tilde{u}_n = \tilde{V}^{f_n}$, punto fijo de $T^{\tilde{f}_n}$.

PI3 $\tilde{f}_{n+1} \in \tilde{\Pi}_{\text{stat}}, T^{\tilde{f}_{n+1}} \tilde{u}_n = T\tilde{u}_n$.

PI4 $n := n+1$, ir a **PI2**.

Theorem 3

- $\tilde{u}_n \uparrow \tilde{V}^*$.
- *If for some* $n \in \mathbb{N}$, $\tilde{u}_{n+1} = \tilde{u}_n$, $\tilde{u}_n = \tilde{V}^*$ *and* $\tilde{f}_n$ *optimal.*

Rolling Horizon Procedure

RH1 In τ, s_τ, buscar

$$V^*_{\tau,N}(s) \;=\; \max_{\pi}\; \mathbb{E}^{\pi}_{s}\left[\sum_{t=\tau}^{\tau+N-1} \lambda_{t-1} r_t^{A_t}(S_t)\right]$$

$s = s_\tau$ initial state. Get $f_{N-1}(s_\tau)$.

RH2 Apply $a_\tau = f_{N-1}(s_\tau)$.

RH3 Observe the state in $\tau+1$: $s_{\tau+1}$.

RH4 Put $\tau := \tau+1$ and go to **RH1**.

Rolling Horizon Procedure

RH1 In τ, s_τ, buscar

$$V^*_{\tau,N}(s) \;=\; \max_{\pi}\; \mathbb{E}^{\pi}_{s}\left[\sum_{t=\tau}^{\tau+N-1} \lambda_{t-1} r_t^{A_t}(S_t)\right]$$

$s = s_\tau$ initial state. Get $f_{N-1}(s_\tau)$.

RH2 Apply $a_\tau = f_{N-1}(s_\tau)$.

RH3 Observe the state in $\tau+1$: $s_{\tau+1}$.

RH4 Put $\tau := \tau+1$ and go to **RH1**.

Theorem 4

$$||\tilde{V}^* - \tilde{U}_N||_\infty \leq \frac{2M\rho^N}{1-\rho}\ .$$

Concluding Remarks

- In Finite Horizon,
  - Characterization of the value funtion, Markov optimal strategy, recurrence method.
- Infinite-Horizon,
  - Existence and characterization of the value function and existence of stationary strategies.
  - Approximation schemes to obtain value function and ε-optimal strategies. Policy Iteration Algorithm, Rolling Horizon Procedure.

Referencias

- Heal G., *Valuing the Future: Economic Theory and Sustainability*. Columbia University Press (2000)
- Hernández-Lerma O., Lasserre J.B., *Error Bounds for Rolling Horizon Policies in Discrete-Time Markov Control Processes*, IEEE Trans. Automat. Control, **35**, 10, 1990, pp. 1118–1124. (1991)
- Hernández-Lerma O., Lasserre J.B., *Discrete-Time Markov Control Processes*. Springer-Verlag (1996)
- Mas-Colell A., Whinston M., Green J., *Microeconomic Theory*. Oxford University Press (1995)
- Puterman L., *Markov Decision Processes*. Wiley and Sons (2005)
- Ross, S., *Applied Probability Models with Optimization Applications*. Holden-Day (1970)
- Sethi T., Sorger G., *A theory of rolling horizon decision making*, Ann. Ops. Res., **29**, 1, pp. 387–415 (1991).
- Tidball M., Altman E., *Approximations in dynamics zero-sum games*. SIAM J. Control and Optimization. **34**, 1, pp. 311–328 (1996)