QUÁ TRÌNH QUYẾT ĐỊNH MARKOV

Ngô Quốc Trần Hiếu
Nguyễn Đức Minh
Lê Thị Duyên

Ngày 5 tháng 1 năm 2020

Lời mở đầu

Quá trình quyết định Markov (MDP) cung cấp một nền tảng toán học cho việc mô hình hóa việc ra quyết định trong các tình huống mà kết quả là một phần ngẫu nhiên và một phần dưới sự điều khiển của một người ra quyết định. MDP rất hữu dụng cho việc học một loạt bài toán tối ưu hóa được giải quyết thông qua quy hoạch động và học tăng cường. MDP được biết đến sớm nhất là vào những năm 1950 (cf. Bellman 1957). Một cốt lõi của nghiên cứu về quá trình ra quyết định Markov là từ kết quả của cuốn sách của Ronald A. Howard xuất bản năm 1960, Quy hoạch động và quá trình Markov. Chúng được sử dụng trong rất nhiều các lĩnh vực khác nhau, bao gồm robot, điều khiển tự động, kinh tế, và chế tạo.

Quá trình quyết định Markov là một phần mở rộng của chuỗi Markov; khác biệt là ở sự bổ sung của các hành động (cho phép lựa chọn). Ngược lại, nếu chỉ có một hành động tồn tại cho mỗi trạng thái, thì một quá trình Markov chính là xích Markov.

Báo cáo này sẽ trình bày những kiến thức cơ bản về quá trình quyết định Markov và ứng dụng. Chúng em xin cảm ơn cô Nguyễn Thị Ngọc Anh đã tận tình hướng dẫn, giúp chúng em hoàn thành báo cáo này. Do thời gian gấp gáp, báo cáo không tránh khỏi sai sót. Chúng em mong nhận được những lời đóng của cô để bài báo cáo hoàn thiện hơn.

Hà Nội, tháng 1 năm 2020

Mục lục

Chương 1

Khái niệm cơ bản

1.1 Quá trình Quyết định Markov (Markov Decision Process)

Ví dụ 1.1 Cho X là một quá trình ngẫu nhiên với bốn trạng thái $E = \{a, b, c, d\}$. Quá trình này đại diện cho một cỗ máy có thể vận hành ở các điều kiện khác nhau từ a đến d với chi phí tăng dần. Khi máy xuống cấp, không chỉ chi phí vận hành đắt hơn mà còn bị mất sản phẩm. Do đó, các hoạt động bảo trì luôn được thực hiện ở các trạng thái từ b đến d. Ngoài không gian trạng thái, còn có một *không gian hành động* đưa ra quyết định cho bước theo. Ví dụ, ta giả sử không gian hành động $A = 1, 2$. Tức là, ở mỗi bước tiếp theo, một trong hai hành động có thể được thực hiện: Dùng một người điều hành thiếu kinh nghiệm, chi phí thấp (hành động 1) và dùng một người điều hành kinh nghiệm, chi phí cao (hành động 2). Cho véc tơ chi phí và ma trận xác suất chuyển tương ứng với mỗi hành động trong không gian hàng động

$$f_1 = (100, 125, 150, 500)^T,$$
$$f_2 = (300, 325, 350, 600)^T,$$
$$P_1 = \begin{bmatrix} 0.1 & 0.3 & 0.6 & 0.0 \\ 0.0 & 0.2 & 0.5 & 0.3 \\ 0.1 & 0.1 & 0.2 & 0.7 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{bmatrix},$$
$$P_2 = \begin{bmatrix} 0.6 & 0.3 & 0.1 & 0.0 \\ 0.75 & 0.1 & 0.1 & 0.05 \\ 0.8 & 0.2 & 0.0 & 0.0 \\ 0.9 & 0.1 & 0.0 & 0.0 \end{bmatrix}.$$

Quá trình vận hành được mô tả như sau: Nếu ở thời điểm n, quá trình đang ở trạng thái $i(X_n = i)$, quyết định k được đưa ra, thì chi phí phải bỏ ra là $f_k(i)$ và xác suất để trạng thái bước tiếp theo là j là $P_k(i, j)$. Ví dụ, nếu $X_n = a$ và quyết định 1 được đưa ra, thì chi phí phải bỏ ra là \$100 và $P(X_{n+1} = a) = 0.1$; hay $X_n = d$ và quyết định 2 được đưa ra thì chi phí phải bỏ ra là \$600 và $P(X_{n+1} = a) = 0.9$.

Định nghĩa 1.1. Cho X là một quá trình mô tả hệ thống (system description process) với không gian trạng thái E và cho D là một quá trình quyết định (decision process) với không gian hành động A. Quá trình (X, D) là *Quá trình Quyết định Markov* nếu, với mọi $j \in E$ và $n = 0, 1, \ldots$ ta đều có

$$P(X_{n+1} = j | X_0, D_0, \ldots, X_n, D_n) = P(X_{n+1} = j | X_0, D_0).$$

Hơn nữa, với mỗi $k \in A$, cho f_k là một véc tơ chi phí, P là ma trận Markov. Khi đó:

$$P(X_{n+1} = j | X_n = i, D_n = k) = P_k(i, j)$$

và chi phí $f_k(i)$ phát sinh mỗi khi $X_n = i$ và $D_n = k$.

Định nghĩa 1.2. Một *chính sách (policy)* là một tập quy tắc, sử dụng thông tin hiện tại, thông tin quá khứ, và/hoặc ngẫu nhiên chỉ định hành động nào được thực hiện tại mỗi thời điểm. Tập tất cả các chính sách được biểu diễn bởi $\mathcal{M}$.

Sau đây là một số chính sách cho bài toán trên:

Chính sách 1. Luôn chọn hành động 1, không phụ thuộc vào trạng thái của X, tức $D_n \equiv 1$ với mọi n

Chính sách 2. Nếu X_n ở trạng thái a hoặc b, cho $D_n = 1$; nếu X_n ở trạng thái c hoặc d, cho $D_n = 2$.

Chính sách 3. Nếu X_n ở trạng thái a hoặc b, cho $D_n = 1$; nếu X_n ở trạng thái c, tung một đồng xu và cho $D_n = 1$ nếu đồng xu xấp, cho $D_n = 2$ nếu đồng xu ngửa; nếu X_n ở trạng thái d thì cho $D_n = 2$.

Chính sách 4. Cho $D_n \equiv 1$ nếu $n = 0$ và 1. Với $n \geq 2$, nếu $X_n > X_{n-1}$ và $X_{n-1} = a$, cho $D_n = 1$. Nếu $X_n > X_{n-1}, X_{n-2} = b$ và $D_{n-1} = 2$ cho $D_n = 1$, ngược lại, $D_n = 2$.

Dễ thấy, nếu áp dụng chính sách 1, quá trình Quyết định Markov (X, D) là xích Markov với ma trận xác suất chuyển là P_1 và véc tơ chi phí f_1. Nếu áp dụng chính sách 2, quá trình (X, D) là xích Markov với ma trận xác suất chuyển và véc tơ chi phí là

$$P = \begin{bmatrix} 0.1 & 0.3 & 0.6 & 0.0 \\ 0.0 & 0.2 & 0.5 & 0.3 \\ 0.8 & 0.2 & 0.0 & 0.0 \\ 0.9 & 0.1 & 0.0 & 0.0 \end{bmatrix},$$
$$f = (100, 125, 350, 600)^T,$$

Còn nếu áp dụng chính sách 4, quá trình (X, D) không phải là xích Markov do trạng thái trạng thái tại một thời điểm phụ thuộc vào lịch sử.

1.2 Kỳ vọng tổng chi phí có chiết khấu

Cho α là tỷ lệ chiết khấu sao cho 1 tại thời điểm $n = 1$ có giá trị hiện tại bằng α tại thời điểm $n = 0$. (Trong kinh tế, thông thường $\alpha = \dfrac{1}{r+1}$ với r là lãi suất). Kỳ vọng tổng chi phí có chiết khấu cho một quá trình quyết định Markov được cho bởi công thức

$$E(\sum_{n=0}^{\infty} \alpha^n f_{D_n}(X_n)).$$

Từ công thức trên, dễ thấy với mỗi chính sách ta sẽ có một kỳ vọng tổng chi phí có chiết khấu khác nhau. Ví dụ, áp dụng chính sách 1 cho Ví dụ 1.1, $\alpha = 0.95$, thì quá trình (X, D) là xích Markov có ma trận xác suất chuyển P_1 và véc tơ chi phí f_1, khi đó tổng chí phí có chiết khấu là $(I - \alpha P_1)^{-1} f_1 = v = (4502, 4591, 4676, 4815)^T$ (Theo [1], trang 326). Nói cách khác, nếu xuất phát từ trạng thái a, tổng chi phí là 4502, xuất phát từ b tổng chi phí là 4591. Ký hiệu $E_d[.]$ là kỳ vọng tổng chi phí có chiết khấu nếu sử dụng chính sách $d, d \in \mathcal{M}$. Đặt

$$v_d^\alpha = E_d(\sum_{n=0}^{\infty} \alpha^n f_{D_n}(X_n) | X_0 = i)$$

với mọi $i \in E$ và $0 < \alpha < 1$. Từ đó nảy sinh bài toán tìm $d^\alpha \in \mathcal{M}$ sao cho

$$v_{d^\alpha}^\alpha(i) = v^\alpha(i) = \min_{d \in \mathcal{M}} v_d^\alpha(i) \ \forall i \in E. \tag{1.1}$$

1.3 Chi phí trung bình dài hạn

Chí phí trung bình dài hạn của một quá trình quyết định Markov khi áp dụng chính sách $d \in \mathcal{M}$ được đưa ra bởi công thức

$$\varphi_d = \lim_{m\to\infty} \frac{1}{m} \sum_{n=0}^{m-1} f_{D_n}(X_n).$$

Từ đó, nảy sinh bài toán tối ưu: Tìm $d^* \in \mathcal{M}$ sao cho

$$\varphi^* = \varphi_{d^*} = \min_{d\in\mathcal{M}} \varphi_d \tag{1.2}$$

1.4 Chính sách Cố định (Stationary Policies)

Bài toán Markov quyết định 1.1 và 1.2 rất khó giải do không gian chính sách $\mathcal{M}$ có cấu trúc khó nắm bắt. Tuy nhiên, người ta chứng minh được rằng chính sách tối ưu luôn có cấu trúc khá đẹp. Như trong Ví dụ 1.1, chính sách 3 và chính sách 4 không thể là chính sách tối ưu.

Định nghĩa 1.3. Một *hàm hành động* là một véc tơ ánh xạ từ không gian trạng thái vào không gian hành động.

Nói cách khác, nếu a là một hàm hành động, thì $a(i) \in A$ với mọi $i \in E$. Xét Ví dụ 1.1, $a_1 = (1, 1, 1, 1), a_2 = (1, 1, 2, 2)$ là các hàm hành động.

Định nghĩa 1.4. Một *chính sách cố định* là một chính sách có thể được biểu diễn bằng một hàm hành động. Chính sách cố định được biểu diễn bằng hàm hành động a thực hiện hành động $a(i)$ tại thời điểm n nếu $X_n = i$, độc lập với các trạng thái trước, hành động trước, và thời điểm n.

Trong Ví dụ 1.1, chính sách 1 được biểu diễn bằng hàm hành động a_1 và chính sách 2 được biểu diễn bằng hàm hành động a_2.

Ý tưởng của chính sách cố định là nó độc lập với thời điểm, và là một chính sách không ngẫu nhiên chỉ phụ thuộc vào trạng thái hiện tại của quá trình, và do đó, không phụ thuộc vào lịch sử. Chính sách cố định thuận lợi cho tính toán ở chỗ, một quá trình quyết định Markov theo chính sách cố định luôn là một xích Markov. Ví dụ, theo chính sách cố định được định nghĩa bằng hàm hành động a thì quá trình quyết định Markov là xích Markov có ma trận xác xuất chuyển và véc tơ chi phí là

$$P^a(i,j) = P_{a(i)}(i,j) \ \forall i,j \in E, \tag{1.3}$$
$$f^a(i) = f_{a(i)}(i) \ \forall i \in E. \tag{1.4}$$

Tính chất 1.1. *Nếu không gian trạng thái E hữu hạn, tồn tại một chính sách cố định là nghiệm của Bài toán (1.1). Thêm nữa, nếu mỗi chính sách cố định sinh ra một xích Markov không giảm, thì tồn tại một chính sách cố định là nghiệm của Bài toán (1.2). (Chính sách tối ưu phụ thuộc vào chiết khấu và có thể khác nhau đối với hai Bài toán (1.1) và (1.2.)*

Chú ý. Xích Markov không giảm là xích chỉ có một lớp liên thông.

Chương 2

Các thuật toán chi phí có chiết khấu

Mục này sẽ giới thiệu ba thủ tục tìm chính sách tối ưu cho tiêu chí tổng chi phí có chiết khấu. Các thủ tục này dựa trên tính chất điểm cố định.

Trong toán học, một hàm được gọi là bất biến đối với một phép toán nếu phép toán không làm thay đổi hàm. Ví dụ, trạng thái dừng π, còn được gọi là véc tơ bất biến cho ma trận Markov P bởi vì phép toán πP không làm thay đổi véc tơ π. Cho quá trình quyết định Markov, phép toán sẽ phức tạp hơn phép nhân ma trận, nhưng ý tưởng cơ bản về hàm bất biến không thay đổi. Nếu hàm bất biến là duy nhất, thì nó được gọi là điểm cố định cho phép toán

Tính chất 2.1. ***Định lý điểm cố định cho quá trình quyết định Markov.*** *Cho v^α là hàm giá trị tối ưu của Bài toán (1.1) với $0 < \alpha < 1$. Hàm v^α thỏa mãn, với mọi $i \in E$, ta đều có*

$$v^\alpha(i) = \min_{k \in A}\{f_k(i) + \alpha \sum_{j \in E} P_k(i,j) v^\alpha(j)\}. \qquad (2.1)$$

Hơn nữa, nó là hàm duy nhất thỏa mãn tính chất này.

Tính chất 2.1 cung cấp cho ta một cách để kiểm tra xem một hàm có phải giá trị tối ưu của Bài toán (1.1) hay không. Nếu ta có được một hàm giá trị tối ưu, ta có thể tìm được chính sách tối ưu bằng tính chất sau

Tính chất 2.2. *Cho v^α là hàm giá trị tối ưu của Bài toán (1.1) với $0 < \alpha < 1$. Định nghĩa một hàm hành động, với mỗi $i \in E$, ta có*

$$a(i) = \text{argmin}_{k \in A}\{f_k(i) + \alpha \sum_{j \in E} P_k(i,j) v^\alpha(j)\} \tag{2.2}$$

Chính sách cố định được định nghĩa bằng hàm hành động a là chính sách tối ưu.

Bây giờ, chúng ta kiểm chứng xem $v^\alpha = (4287, 4382, 4441, 4613)$ với $\alpha = 0.95$ có phải là giá trị tối ưu của Bài toán (1.1) hay không.

$$\begin{aligned} v^\alpha(a) = \min\{&100 + 0.95(0.1, 0.3, 0.6, 0.0) \begin{bmatrix} 4287 \\ 4382 \\ 4441 \\ 4613 \end{bmatrix}; \\ &300 + 0.95(0.6, 0.3, 0.1, 0.0) \begin{bmatrix} 4287 \\ 4382 \\ 4441 \\ 4613 \end{bmatrix}\} \\ &= 4287 \end{aligned}$$

$$v^{\alpha}(b) = \min\{125 + 0.95(0.0, 0.2, 0.5, 0.3) \begin{bmatrix} 4287 \\ 4382 \\ 4441 \\ 4613 \end{bmatrix};$$
$$325 + 0.95(0.75, 0.1, 0.1, 0.05) \begin{bmatrix} 4287 \\ 4382 \\ 4441 \\ 4613 \end{bmatrix}\}$$
$$= 4382.$$

$$v^{\alpha}(c) = \min\{150 + 0.95(0.0, 0.1, 0.2, 0.7) \begin{bmatrix} 4287 \\ 4382 \\ 4441 \\ 4613 \end{bmatrix};$$
$$350 + 0.95(0.8, 0.2, 0.0, 0.0) \begin{bmatrix} 4287 \\ 4382 \\ 4441 \\ 4613 \end{bmatrix}\}$$
$$= 4441.$$

$$v^{\alpha}(d) = \min\{500 + 0.95(0.8, 0.1, 0.0, 0.1) \begin{bmatrix} 4287 \\ 4382 \\ 4441 \\ 4613 \end{bmatrix};$$
$$600 + 0.95(0.9, 0.1, 0.0, 0.0) \begin{bmatrix} 4287 \\ 4382 \\ 4441 \\ 4613 \end{bmatrix}\}$$
$$= 4613.$$

Từ đó kết luận v^α là giá trị tối ưu của Bài toán 1. Sử dụng Tính chất 2.2, ta xác định được chính sách tối ưu được biểu diễn bằng hành động $a = (1, 1, 2, 1)$.

2.1 Cải thiện giá trị cho chi phí có chiết khấu

Tính chất 2.3. Thuật toán cải thiện giá trị. *Thủ tục sau sẽ cho phép ta tìm được xấp xỉ giá trị tối ưu của Bài toán (1.1)*

Bước 1. Cho $\alpha < 1$, chọn một giá trị dương đủ nhỏ ϵ, đặt $n = 0$, và đặt $v_0(i) = 0$ với mỗi $i \in E$. (Ta đặt $v_0 = 0$ để tiện tính toán, ta có thể chọn v_0 bất kỳ).

Bước 2. Với mỗi $i \in E$, xác định $v_{n+1}(i)$ như sau

$$v_{n+1}(i) = \min_{k \in A}\{f_k(i) + \alpha \sum_{j \in E} P_k(i, j)v_n(j)\}.$$

Bước 3. Tính δ

$$\delta = max\{|v_{n+1}(i) - v_n(i)|\}.$$

Bước 4. Nếu $\delta < epsilon$, đặt $v^\alpha = v_{n+1}$ và dừng thuật toán; ngược lại, tăng n thêm 1 và quay lại Bước 2.

Thuật toán cải thiện giá trị có hai vấn đề lớn: (1) nó có thể chậm hội tụ và (2) không có quy tắc đơn giản nào để thiết lập tiêu chí hội tụ (đặt giá trị cho ϵ. Về lý thuyết, khi số lần lặp vô hạn, hàm giá trị trở thành tối ưu, tuy nhiên, trong thực tế ta phải dừng lại sau hữu hạn bước khi hàm giá trị không giảm đáng kể.)

Áp dụng thuật toán cải thiện giá trị cho Ví dụ 1.1, ta có kết

quả các bước lặp như sau

$$\begin{aligned}
&v_0 = (0, 0, 0, 0)\\
&v_1 = (100, 125, 150, 500)\\
&v_2 = (230.62, 362.50, 449.75, 635.38)\\
&v_3 = (481.58, 588.59, 594.15, 770.07)\\
&\vdots
\end{aligned}$$

2.2 Cải tiến chính sách cho chi phí có chiết khấu

Thuật toán của phần trước tập trung vào giá trị. Trong phần này, ta xét thuật toán tập trung vào chính sách và sau đó tính toán giá trị gắn liền với chính sách đó. Kết quả là tốc độ hội tụ nhanh hơn đáng kể, nhưng tính toán trong mỗi bước lặp phức tạp hơn

Tính chất 2.4. Thuật toán cải tiến chính sách. *Thủ tục sau sẽ cho phép ta tìm được chính sách tối ưu của Bài toán (1.1)*

Bước 1. Cho $\alpha < 1$, đặt $n = 0$, định nghĩa hàm hành động a_0

$$a_0(i) = argmin_{k \in A} f_k(i)$$

với mọi $i \in A$.

Bước 2. Xác định ma trận P và véc tơ f

$$\begin{aligned}
f(i) &= f_{a_n(i)}(i)\\
P(i, j) &= P_{a_n(i)}(i, j)
\end{aligned}$$

với mỗi $i, j \in E$.

Bước 3. Tính véc tơ v

$$v = (I - \alpha P)^{-1} f.$$

Bước 4. Xác định hàm hành động a_{n+1} như sau

$$a_{n+1}(i) = \text{argmin}_{k \in A}\{f_k(i) + \alpha \sum_{j \in E} P_k(i,j) v(j)\}$$

Bước 5. Nếu $a_{n+1} = a_n$, đặt $v^\alpha = v, a^\alpha = a_n$ và dừng thuật toán, ngược lại tăng n lên 1 và quay lại Bước 2.

Ý tưởng cơ bản của Thuật toán cải tiến chính sách là lấy một chính sách cố định, tính toán véc tơ chi phí và ma trận chuyển tiếp liên quan đến chính đó, sau đó xác định tổng chi phí dự kiến chiết khấu dự kiến tương ứng với véc tơ chi phí và ma trận xác suất chuyển. Nếu chính sách đó là chính sách tối ưu, ta sẽ tìm nó thông qua tính chất 2.2 và véc tơ chi phí. Nếu chính sách Bước 1 không tối ưu, thì chính sách tìm được ở Bước 4 sẽ có hàm giá trị tốt hơn chính sách ở Bước 1

Ta sẽ tìm chính sách tối ưu với tiêu chí là tổng chi phí có chiết khấu trong Ví dụ 1.1

Bước lập 1.
Bước 1.

$$a_0 = (1, 1, 1, 1)$$

Bước 2.

$$f = (100, 125, 150, 500)^T$$

$$P = \begin{bmatrix} 0.1 & 0.3 & 0.6 & 0.0 \\ 0.0 & 0.2 & 0.5 & 0.3 \\ 0.1 & 0.1 & 0.2 & 0.7 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{bmatrix},$$

Bước 3.

$$v = (10108.12, 1006462, 10841.11, 10097.58)^T$$

Bước 4.

$$a_1(a) = \text{argmin}\{100 + 0.95(0.1, 0.3, 0.6, 0.0) \begin{bmatrix} 10108.12 \\ 10064.62 \\ 10841.11 \\ 10097.58 \end{bmatrix},$$
$$300 + 0.95(0.6, 0.3, 0.1, 0.0) \begin{bmatrix} 10108.12 \\ 10064.62 \\ 10841.11 \\ 10097.58 \end{bmatrix}\}$$
$$= 2.$$

$$a_1(b) = \text{argmin}\{125 + 0.95(0.0, 0.2, 0.5, 0.3) \begin{bmatrix} 10108.12 \\ 10064.62 \\ 10841.11 \\ 10097.58 \end{bmatrix},$$
$$325 + 0.9255(0.75, 0.1, 0.1, 0.05) \begin{bmatrix} 10108.12 \\ 10064.62 \\ 10841.11 \\ 10097.58 \end{bmatrix}\}$$
$$= 2.$$

$$a_1(c) = \text{argmin}\{150 + 0.95(0.0, 0.1, 0.2, 0.7) \begin{bmatrix} 10108.12 \\ 10064.62 \\ 10841.11 \\ 10097.58 \end{bmatrix},$$
$$350 + 0.9255(0.8, 0.2, 0.0, 0.0) \begin{bmatrix} 10108.12 \\ 10064.62 \\ 10841.11 \\ 10097.58 \end{bmatrix}\}$$
$$= 2.$$

$$a_1(d) = \text{argmin}\{500 + 0.95(0.8, 0.1, 0.0, 0.1) \begin{bmatrix} 10108.12 \\ 10064.62 \\ 10841.11 \\ 10097.58 \end{bmatrix},$$
$$600 + 0.9255(0.9, 0.1, 0.0, 0.0) \begin{bmatrix} 10108.12 \\ 10064.62 \\ 10841.11 \\ 10097.58 \end{bmatrix}\}$$
$$= 1.$$

Từ đó ta có $a_1 = (2, 2, 2, 1)$.

Bước 5. Do $a_1 \neq a_0$, đặt $n = 1$ và quay lại Bước 2.

Bước lặp 2.

Bước 2.

$$f = (300, 325, 350, 500)^T$$
$$P = \begin{bmatrix} 0.6 & 0.3 & 0.1 & 0.0 \\ 0.75 & 0.1 & 0.1 & 0.05 \\ 0.8 & 0.2 & 0.0 & 0.0 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{bmatrix},$$

Bước 3.

$$v = (6249.51, 6278.91, 6292.62, 6459.80)^T.$$

Bước 4. Tương tự trong Bước lặp 1

$$a_2 = (1, 1, 2, 1).$$

Bước 5. Do $a_2 \neq a_1$ nên đặt $n = 2$ và quay lại Bước 2.

Bước 2.

$$f = (100, 125, 350, 500)^T$$

$$P = \begin{bmatrix} 0.1 & 0.3 & 0.6 & 0.0 \\ 0.0 & 0.2 & 0.5 & 0.3 \\ 0.1 & 0.1 & 0.2 & 0.7 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{bmatrix},$$

Bước 3.

$$v = (4287.40, 4381.63, 4440.93, 4612.90)^T.$$

Bước 4.

$$a_3 = (1, 1, 2, 1).$$

Bước 5. Do $a_3 = a_2$ nên dừng thuật toán, nghiệm tối ưu là $a^\alpha = a_3$ và giá trị tối ưu là $v^\alpha = v$.

Lưu ý rằng thay vì tính $v = (I - \alpha P)^{-1} f$ ta có thể giải hệ phương trình tuyến tính

$$(I - \alpha P)v = f.$$

2.3 Quy hoạch tuyến tính cho chi phí có chiết khấu

Tính chất 2.5. ***Bổ đề về quy hoạch tuyến tính*** *Cho v^α là giá trị tối ưu của Bài toán (1.1) với $0 < \alpha < 1$, và cho u là một hàm thực trên không gian E (hữu hạn). Nếu u thỏa mãn*

$$u(i) \leq \min_{k \in A}\{f_k(i) + \alpha \sum_{j \in E} P_k(i)u(j)\}$$

với mọi $i \in E$, thì $u \leq v^\alpha$

Xét tập được tạo bởi tất cả các hàm thỏa mãn Tính chất 2.5, sau đó từ Tính chât 2.1 ta biết rằng giá trị tối ưu v^α thuộc tập trên. Hơn nữa v^α là hàm lớn nhất trong tập trên. Nói cách khác, v^α là nghiệm tối ưu của bài toán

$$\begin{aligned} &\max\ u \\ &\text{với điều kiện} \\ &\qquad\qquad u \leq \min\{f_k + \alpha P_k u\} \end{aligned}$$

Tính chất 2.6. ***Quy hoạch tuyến tính cho chi phí có chiết khấu*** *Nghiệm tối ưu của bài toán sau đây chính là nghiệm tối ưu v^α của Bài toán (1.1) với $0 < \alpha < 1$*

$$\begin{aligned} &\max \sum_{i \in E} u(i) \\ &\textit{với điều kiện} \\ &u(i) \leq f_k(i) + \alpha \sum_{j \in E} P_k(i,j)u(j) \ \textit{với mọi } i \in E, k \in A. \end{aligned}$$

Để minh hoạ cho tính chất 2.6, ta sẽ quay trở lại với Ví dụ 1.

$$
\begin{array}{llrrr}
\max & z & = u_a + u_b & +u_c + u_d & \\
\text{với điều kiện} & & & & \\
& u_a \leq 100 & +0.095u_a + 0.285u_b & +0.57u_c & \\
& u_a \leq 300 & +0.57u_a + 0.285u_b & +0.095u_c & \\
& u_b \leq 125 & + 0.19u_b & +0.475u_c + 0.285u_d & \\
& u_b \leq 325 & +0.7125u_a + 0.095u_b & +0.095u_c + 0.0475u_d & \\
& u_c \leq 150 & + 0.095u_b & +0.19u_c + 0.665u_d & \\
& u_c \leq 350 & +0.76u_a + 0.19u_b & & \\
& u_d \leq 500 & +0.76u_a + 0.095u_d & & \\
& u_d \leq 600 & +0.855u_a + 0.095u_b & &
\end{array}
$$

Giải bài toán trên, ta được kết quả là $u_a = 4287, u_b = 4382, u_c = 4441, u_d = 4613$.

Chương 3

Các thuật toán chi phí trung bình

Tính chất 3.1. *Giả sử rằng mọi chính sách cố định đều sinh ra mỗi xích Markov với tập không giảm. Khi đó, tồn tại một đại lượng vô hướng φ^* và một véc tơ h thỏa mãn, với mọi $i \in E$,*

$$\varphi^* + h(i) = \min_{k \in A}\{f_k(i) + \sum_{j \in E} P_k(i,j)h(j)\}. \qquad (3.1)$$

Đại lượng vô hướng φ^* là giá trị tối ưu của Bài toán (1.2), và hàm hành động tối ưu là

$$a(i) = \text{argmin}_{k \in A}\{f_k(i) + \sum_{j \in E} P_k(i,j)h(j)\}$$

Véc tơ h là duy nhất sai khác một hằng số.

Từ Tính chất 3.1, ta có

$$\varphi^* + h(i) = f_{a(i)}(i) + \sum_{j \in E} P_{a(i)}(i,j)h(j) \qquad (3.2)$$

$$\Leftrightarrow \varphi^* + h(i) = f_{a(i)}(i) + (P_{a(i)}h)(i). \qquad (3.3)$$

Bây giờ, ta sẽ chứng minh chính sách xác định bởi hàm hành động $a = (1, 1, 2, 1)$ là chính sách tối ưu. Đầu tiên, ta giải hệ phương trình (3.3)

$$\begin{aligned}
\varphi^* + h_a &= 100 & +0.1h_a + 0.3h_b & & +0.6h_c \\
\varphi^* + h_b &= 125 & + 0.2h_b & & +0.5h_c + 0.3h_d \\
\varphi^* + h_c &= 350 & +0.8h_a + 0.2h_b & & \\
\varphi^* + h_d &= 500 & +0.8h_a + 0.1h_b & &
\end{aligned}$$

Cho $h_a = 0$, giải hệ ta được nghiệm $h = (0.0, 97.10, 150.18, 322.75)^T$ và $\varphi^* = 219.24$. Bây giờ ta phải chứng minh

$$\varphi^* + h(i) \leq (P_k h)(i).$$

với mọi $k \neq a(i)$.

$$\begin{aligned}
\varphi^* + h_a &\leq 300 & +0.6h_a + 0.3h_b & & +0.1h_c \\
\varphi^* + h_b &\leq 125 & +0.75h_a + 0.1h_b & & +0.1h_c + 0.05h_d \\
\varphi^* + h_c &\leq 150 & + 0.1h_b & & + 0.05h_d \\
\varphi^* + h_d &\leq 500 & +0.9h_a + 0.1h_b & &
\end{aligned}$$

Do đó, chính sách xác định bởi hàm hành động a là chính sách tối ưu.

Tính chất 3.2. *Cho v^α là giá trị tối ưu của Bài toán (1.1),φ^* là giá trị tối ưu của Bài toán (1.2) giả sử mọi trạng thái cố định đều sinh ra xích Markov với một tập không giảm. Khi đó*

$$\lim_{\alpha \to 1}(1 - \alpha)v^\alpha(i) = \varphi^*$$

với mọi $i \in E$.

Tính chất 3.3. *Cho v^α là giá trị tối ưu của Bài toán (1.1), φ^* là giá trị tối ưu của Bài toán (1.2), h là véc tơ được xác định ở Tính chất (3.1). Khi đó*

$$\lim_{\alpha \to 1}[v^\alpha(i) - v^\alpha(j)] = h(i) - h(j)$$

với mọi $i, j \in E$.

3.1 Cải tiến chính sách cho chi phí trung bình

Tính chất 3.4. Thuật toán cải tiến chính sách. *Thủ tục sau sẽ cho phép ta tìm được nghiệm tối ưu của Bài toán (1.2)*

Bước 1. Cho $n = 0$, ký hiệu trạng thái đầu tiên trong không gian trạng thái là 1, ta xác định hàm hành động a_0

$$a_0(i) = \text{argmin}_{k \in A} f_k(i)$$

với mỗi $i \in E$.

Bước 2. Xác định ma trận P và véc tơ f như sau

$$f(i) = f_{a_n(i)}(i)$$
$$P(i,j) = P_{a_n(i)}(i,j)$$

với mỗi $i, j \in E$.

Bước 3. Xác định giá trị φ và h bằng việc giải hệ

$$\varphi + h = f + Ph$$

trong đó $h(1) = 0$.

Bước 4. Xác định hàm hành động a_{n+1} như sau

$$a_{n+1}(i) = \text{argmin}_{k\in A}\{f_k(i) + \sum_{j\in E} P_k(i,j)h(j)\}$$

với mỗi $i \in E$.

Bước 5. Nếu $a_{n+1} = a_n$, đặt $\varphi^* = \varphi, a^* = a_n$, và dừng thuật toán, ngược lại tăng n thêm một và quay lại Bước 2.

Để minh họa cho thuật toán trên chúng ta sẽ áp dụng nó để cho Ví dụ 1.1.

Bước lặp 1

Bước 1.

$$a_0 = (1, 1, 1, 1).$$

Bước 2.

$$f = (100, 125, 150, 500)^T$$
$$P = \begin{bmatrix} 0.1 & 0.3 & 0.6 & 0.0 \\ 0.0 & 0.2 & 0.5 & 0.3 \\ 0.1 & 0.1 & 0.2 & 0.7 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{bmatrix},$$

Bước 3. Giải hệ phương trình ($h_a = 0$.)

$$\begin{array}{llll} \varphi & = 100 & +0.3h_b + 0.6h_c & \\ \varphi & +h_b = 125 & +0.2h_b + 0.5h_c & +0.3h_d \\ \varphi & +h_c = 150 & +0.1h_b + 0.2h_c & +0.7h_d \\ \varphi & +h_d = 500 & +0.1h_b & +0.1h_d \end{array}$$

ta được nghiệm $\varphi = 232.86$ và $h = (0, 90.40, 176.23, 306.87)^T$.

Bước 4.

$$a_1(a) = \text{argmin}\{100 + (0.1, 0.3, 0.6, 0.0) \begin{bmatrix} 0.0 \\ 90.40 \\ 176.23 \\ 306.87 \end{bmatrix}; \\ 300 + (0.6, 0.3, 0.1, 0.0) \begin{bmatrix} 0.0 \\ 90.40 \\ 176.23 \\ 306.87 \end{bmatrix}\} \\ = 1$$

$$a_1(b) = \text{argmin}\{125 + (0.0, 0.2, 0.5, 0.3) \begin{bmatrix} 0.0 \\ 90.40 \\ 176.23 \\ 306.87 \end{bmatrix}; \\ 325 + (0.75, 0.1, 0.1, 0.05) \begin{bmatrix} 0.0 \\ 90.40 \\ 176.23 \\ 306.87 \end{bmatrix}\} \\ = 1$$

$$a_1(c) = \text{argmin}\{150 + (0.0, 0.1, 0.2, 0.7) \begin{bmatrix} 0.0 \\ 90.40 \\ 176.23 \\ 306.87 \end{bmatrix}; \\ 350 + (0.8, 0.2, 0.0, 0.0) \begin{bmatrix} 0.0 \\ 90.40 \\ 176.23 \\ 306.87 \end{bmatrix}\} \\ = 2$$

$$a_1(d) = \text{argmin}\{500 + (0.8, 0.1, 0.0, 0.1) \begin{bmatrix} 0.0 \\ 90.40 \\ 176.23 \\ 306.87 \end{bmatrix};$$
$$600 + (0.9, 0.1, 0.0, 0.0) \begin{bmatrix} 0.0 \\ 90.40 \\ 176.23 \\ 306.87 \end{bmatrix}\}$$
$$= 1$$

Từ đó, $a_1 = (1, 1, 2, 1)$.

Bước 5. Do $a_1 \neq a_0$, tăng n thêm một và quay lại bước 2

Bước lặp 2.

Bước 2.

$$f = (100, 125, 350, 500)$$
$$P = \begin{bmatrix} 0.1 & 0.3 & 0.6 & 0.0 \\ 0.0 & 0.2 & 0.5 & 0.3 \\ 0.8 & 0.2 & 0.0 & 0.0 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{bmatrix},$$

Bước 3. Giải hệ phương trình trang 20 được nghiệm là $\varphi^* = 219.24$ và $h = (0.0, 97.10, 150.18, 322.75)$

Bước 4.

$$a_2(a) = \text{argmin}\{100 + (0.1, 0.3, 0.6, 0.0) \begin{bmatrix} 0.0 \\ 97.10 \\ 150.18 \\ 322.75 \end{bmatrix};$$
$$300 + (0.6, 0.3, 0.1, 0.0) \begin{bmatrix} 0.0 \\ 97.10 \\ 150.18 \\ 322.75 \end{bmatrix}\}$$
$$= 1$$

$$a_2(b) = \text{argmin}\{125 + (0.0, 0.2, 0.5, 0.3) \begin{bmatrix} 0.0 \\ 97.10 \\ 150.18 \\ 322.75 \end{bmatrix};$$
$$325 + (0.75, 0.1, 0.1, 0.05) \begin{bmatrix} 0.0 \\ 97.10 \\ 150.18 \\ 322.75 \end{bmatrix}\}$$
$$= 1$$

$$a_2(c) = \text{argmin}\{150 + (0.0, 0.1, 0.2, 0.7) \begin{bmatrix} 0.0 \\ 97.10 \\ 150.18 \\ 322.75 \end{bmatrix};$$
$$350 + (0.8, 0.2, 0.0, 0.0) \begin{bmatrix} 0.0 \\ 97.10 \\ 150.18 \\ 322.75 \end{bmatrix}\}$$
$$= 2$$

$$a_2(d) = \text{argmin}\{500 + (0.8, 0.1, 0.0, 0.1) \begin{bmatrix} 0.0 \\ 97.10 \\ 150.18 \\ 322.75 \end{bmatrix};$$
$$600 + (0.9, 0.1, 0.0, 0.0) \begin{bmatrix} 0.0 \\ 97.10 \\ 150.18 \\ 322.75 \end{bmatrix}\}$$
$$= 1.$$

Do đó, $a_2 = (1, 1, 2, 1)$.

Bước 5. $a_2 = a_1$ nên dừng thuật toán. Chính sách tối ưu là a_2 và chi phí trung bình tối ưu là φ.

3.2 Quy hoạch tuyến tính cho chi phí trung bình

Mô hình quy hoạch tuyến tính cho bài toán chi phí trung bình dài hạn có cách tiếp cận hoàn toàn khác so với bài toán tổng chi phí có chiết khấu. Để đảm bảo tập chính sách chấp nhận được là tập lồi, ta xét cả chính sách cố định và chính sách ngẫu nhiên. Một chính sách ngẫu nhiên, ví dụ chính sách 3 trong Ví dụ 1.1. Ký hiệu

$$v_i(k) = P(D_n = k | X_n = i).$$

Tất nhiên, một chính sách cố định xác định bởi hàm hành động a cũng là một chính sách ngẫu nhiên với $v_i(k) = 1$ nếu $k = a(i)$ và bằng không nếu ngược lại.

Mỗi chính sách ngẫu nhiên sẽ sinh ra một véc tơ phân phối dừng π. Ký hiệu

$$x(i, k) = v_i(k)\pi(i) = \lim_{n\to\infty} P(X_n = i | D_n = k). \qquad (3.4)$$

Với mỗi $i \in E$ cố định,

$$\sum_{k \in E} x(i,k) = \sum_{k \in E} v_i(k)\pi(i) = \pi(i). \tag{3.5}$$

Do đó, kỳ vọng chi phí trung bình dài hạn được tính theo công thức

$$\varphi = \sum_{i \in E} \sum_{k \in A} x(i,k) f_k(i)$$

và ba điều kiện sau phải thỏa mãn:

1. $\sum_{i \in E} \sum_{k \in E} = 1$.
2. $x(i,k) \geq 0, \forall i \in E, \forall k \in A$.
3. Điều kiện phân phối dừng $\pi P = \pi$.

Diều kiện 3 có một số điểm cần bàn.

$$P(i,j) = \sum_{k \in A} v_i(k) P_k(i,j)$$

với mọi $i, j \in E$. Do đó phương trình $\pi P = \pi$ trở thành

$$\pi(j) = \sum_{i \in E} \pi(i) \sum_{k \in A} v_i(k) P_k(i,j)$$

với mọi $j \in E$. Kết hợp phương trình (3.4), (3.5) với phương trình ta có

$$\sum_{k \in A} x(j,k) = \sum_{i \in E} \sum_{k \in A} x(i,k) P_k(i,j)$$

Tính chất 3.5. ***Quy hoạch tuyến tính cho chi phí trung bình*** *Nghiệm tối ưu của bài toán quy hoạch tuyến tính dưới đây là giá trị tối ưu của bài toán (1.2)*

$$\min \varphi = \sum_{i \in E} \sum_{k \in A} x(i,k) f_k(i)$$

với điều kiện

$$\sum_{k \in A} x(j,k) = \sum_{i \in E} \sum_{k \in A} x(i,k) P_k(i,j) \text{ với mọi } j \in E$$

$$\sum_{i \in E} \sum_{k \in A} x(i,k) = 1$$

$$x(i,k) \geq 0 \text{ với mọi } i \in E \text{ và } k \in A$$

Chính sách tối ưu là chọn hành động k cho trạng thái i sao cho $x(i,k) > 0$

Do ta đã biết nghiệm tối ưu của Bài toán (1.2) là chính sách cố định, nên sẽ chỉ có một giá trị dương $x(i,k)$ với mỗi $i \in E$. Do đó, hàm hành động tối ưu là a với $a(i)$ bằng giá trị k mà $x(i,k) > 0$. Đồng thời ta cũng tìm được phân phối dừng ứng với chính sách tối ưu là π với $\pi(i) = x(i, a(i))$ với mọi $i \in E$.

Để minh họa cho mô hình quy hoạch tuyến tính cho chi phí trung bình dài hạn, ta quay lại Ví dụ 1. Giải bài toán quy hoạch

tuyến tính sau

$$\begin{aligned}
\min\ \varphi &= 100x_{a1} + 125x_{b1} + 150x_{c1} + 500x_{d1} \\
&\quad + 300x_{a2} + 325x_{b2} + 350x_{c2} + 600x_{d2} \\
&\text{với điều kiện} \\
&x_{a1} + x_{a2} = 0.1x_{a1} + 0.6x_{a2} + 0.75x_{b2} + 0.8x_{c2} + 0.8x_{d1} + 0.9x_{d2} \\
&x_{b1} + x_{b2} = 0.3x_{a1} + 0.3x_{a2} + 0.2x_{b1} + 0.1x_{b2} + 0.1x_{c1} + 0.2x_{c2} \\
&\quad + 0.1x_{d1} + 0.1x_{d2} \\
&x_{d1} + x_{d2} = 0.3x_{b1} + 0.05x_{b2} + 0.7x_{c1} + 0.1x_{d1} \\
&x_{a1} + x_{a2} + x_{b1} + x_{b2} + x_{c1} + x_{c2} + x_{d1} + x_{d2} = 1 \\
&x_{ik} \geq 0 \text{ với mọi } i, k.
\end{aligned}$$

Giải bài toán trên ta được nghiệm là $x_{a1} = 0.363, x_{b1} = 0.229, x_{c2} = 0.332, x_{d1} = 0.076$ và tất cả các biến còn lại bằng không, giá trị tối ưu là $\varphi^* = 219.24$. Do đó chính sách tối ưu là $a = (1, 1, 2, 1)$ và chi phí trung bình dài hạn nhỏ nhất là $\varphi^* = 219.24$, phân phối dừng $\pi = (0.363, 0.229, 0.332, 0.076)$.

Chương 4

Kết luận

Vừa rồi chúng em đã trình bày những khái niệm cơ bản, một số bài toán, thuật toán và ứng dụng về quá trình quyết định Markov. Trong chương tiếp theo, chúng em sẽ một vài chương trình máy tính giải các bài toán về quá trình quyết định Markov.

Chương 5

Phụ lục

Mục này trình bày kết quả chương trình thử nghiệm viết bằng Python trên máy tính Dell, vi xử lý Intel(R) Core(TM) i3-5005U CPU @ 2.00GHz 2.00GHz, RAM 4GB, hệ điều hành Window 10 64 bit.

Ví dụ. Cho quá trình quyết định Markov

$$E = \{a, b, c, d, g\}$$
$$A = \{1, 2, 3\}$$
$$P_1 = \begin{bmatrix} 0.1 & 0.3 & 0.3 & 0.0 & 0.3 \\ 0.0 & 0.2 & 0.3 & 0.3 & 0.2 \\ 0.1 & 0.1 & 0.2 & 0.4 & 0.3 \\ 0.5 & 0.1 & 0.0 & 0.1 & 0.3 \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{bmatrix},$$
$$P_2 = \begin{bmatrix} 0.1 & 0.3 & 0.3 & 0.0 & 0.3 \\ 0.0 & 0.2 & 0.3 & 0.3 & 0.2 \\ 0.1 & 0.1 & 0.2 & 0.4 & 0.2 \\ 0.5 & 0.1 & 0.0 & 0.1 & 0.3 \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{bmatrix},$$

$$P_3 = \begin{bmatrix} 0.2 & 0.1 & 0.3 & 0.4 & 0.0 \\ 0.0 & 0.2 & 0.2 & 0.3 & 0.3 \\ 0.1 & 0.2 & 0.4 & 0.0 & 0.1 \\ 0.3 & 0.3 & 0.0 & 0.1 & 0.3 \\ 0.2 & 0.2 & 0.2 & 0.3 & 0.1 \end{bmatrix},$$

$f_1 = (200, 250, 300, 500)^T,$

$f_2 = (250, 350, 500, 700)^T,$

$f_3 = (240, 300, 400, 600)^T,$

$\alpha = 0.95.$

5.1 Thuật toán cải thiện giá trị chi phí có chiết khấu

$v^\alpha = [8023.77098803, 8097.41421759, 8145.25632824,$
$8301.5438769, 8534.57032443]^T$

$a^\alpha = [1, 1, 3, 1, 1]$

5.2 Thuật toán cải thiện chính sách theo tiêu chí tổng chi phí có chiết khấu

$a_0 = [1, 1, 1, 1, 1]$

$v = [8229.33829685, 8299.96373136, 8394.16036451,$
$8494.09727824, 8735.87996956]^T$

$a_1 = [1, 1, 3, 1, 1]$

$v = [8023.78906518, 8097.43229474, 8145.27440539,$
$8301.56195404, 8534.58840157]^T$

$a_2 = a_1 \ \Rightarrow a^\alpha = a_2; v^\alpha = v.$

5.3 Thuật toán cải thiện chính sách theo tiêu chí chi phí trung bình dài hạn

$a_0 = [1, 1, 1, 1, 1]$
$h = [0, 71.25274433, 166.72210764, 261.92535427, 501.99587519]^T$
$\varphi = 421.99121815$
$a_1 = [1, 1, 3, 1, 1]$
$h = [0, 75.00676956, 117.88067515, 277.37160394, 507.0403466]^T$
$\varphi = 409.9783374$
$a_2 = a_1 \Rightarrow a^* = a_2, \varphi^* = \varphi.$

Chương 6

Tài liệu tham khảo

1. Richard M. Feldman - Ciriac Valdez-Flores, *Applied Probability and Stochastic Processes, Second Edition* , Springer-Verlag Berlin Heidelberg, 2010.