مثال برنامه ریزی پویای قطعی

⇐ دنبال حداکثر کردن فروش را اعداد داخل جدول هستیم.

فروشنده \ ناحیه	1	2	3
1	5	2	4
2	2	1	3
3	1	4	5

هر ناحیه حداقل یک فروشنده،

هر فروشنده تنها در یک ناحیه کار میکند.

- A policy determines actions based on history and observation period.
- Define history sets:
    - $H_n = \Gamma^{n-1} \times S$ for $n > 0$.
    - $H_0 = S$.
- A policy $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$:
    - For any $n \geq 0$ and history $h_n = (i_0, a_0, \ldots, i_n) \in H_n$:
    - $\pi_n(h_n)$ is a probability distribution on $A(i_n)$.

Question 1 What is a deterministic policy?

Question 2 What is a Markov policy?

تعریف ۱: سیاست ثابت مستقل از زمان است.

تعریف ۲: $\pi_t = (\pi_t, t \geq 0) \in \Pi_m$

- For $n \geq 0$:
  - X_n: State at period n.
  - Δ_n: Action chosen at period n.
- The process $\{X_n, \Delta_n, n \geq 0\}$ is well-defined under any policy $\pi \in \Pi$.
- Under a Markov policy $\pi \in \Pi_M$ Forms a discrete-time Markov chain.
- For each $\pi \in \Pi$ and $i \in S$:
  - $P_{\pi,i}$: Probability under policy π with initial state i.
  - $E_{\pi,i}$: Expectation under policy π with initial state i.
- Reward structure:
  - Reward $r(X_n, \Delta_n)$ at period n is random.

Question How to compare different policies?

① Average reward: $V(\pi, i) = \liminf_{N \to \infty} \frac{1}{N+1} V_{1,N}(\pi, i)$

Discounted criterion/total reward criterion

$V_\beta(\pi, i) = \sum_{n=0}^{\infty} \beta^n \mathbf{E}_{\pi,i}(r(X_n, \Delta_n)), \quad i \in S, \pi \in \Pi$

In the literature, the discount rate $\beta \in [0, 1]$ is often assumed. Why?

The optimal value function for this criterion is defined by:

$$V_{\beta,N}(i) = \sup_{\pi \in \Pi} V_{\beta,N}(\pi, i), \quad i \in S$$

$\sum_{k\geq 1}^{\infty} \beta_k = \infty$ و $\sum_k \beta_k^2 < \infty$

واریانس را کم کنیم

میتوانیم β را برحسب زمان تعریف کنیم.

شروط تضمین کننده همگرایی با احتمال یک

A Gambling Problem [Ross, 2014]

At each play of the game, a gambler can bet any nonnegative amount up to his present fortune and will either win or lose that amount with probabilities p and $q = 1 - p$, respectively. The gambler is allowed to make n bets and his objective is to maximize the expectations of the logarithm of his final fortune. What strategy achieves this end?

$$x \to V_n(x) = \max_{0 \le \alpha \le 1} \left\{ p V_{n+1}(x + \alpha x) + q V_{n+1}(x - \alpha x) \right\}$$

$$V_0(x) = \log x \to V_1(x) = \max_{0 \le \alpha \le 1} \left\{ p \log(x + \alpha x) + q \log(x - \alpha x) \right\} \to V_1(x) = \max_{0 \le \alpha \le 1} \left\{ p \log(1 + x) + q \log(1 - x) \right\} + \log x$$

$$\to V_1(x) = C + \log x \Rightarrow V_2(x) = 2C + \log x \Rightarrow V_n(x) = nC + \log x$$

$$\frac{dV}{d\alpha} = \frac{p}{x + \alpha x} - \frac{q}{x - \alpha x} = 0 \to \alpha \neq 1 \to p - q = \alpha \; \& \; 2p - 1 = \alpha$$

Sequential Investment Problem

Suppose one has an amount M of money and considers investing this money over N future periods. However, the opportunity for investment is not deterministic. At each period, an investment opportunity occurs with probability p, which is independent of the past and the amount of remaining money. When an investment opportunity occurs, if he invests x, he will earn a revenue $r(x)$, including his investment. Assume that both his investment and his return at any period cannot be reinvested in the future. What is the optimal strategy for this problem?

Let $V_n(X)$ be the maximal expected profit when there are n periods remaining, X money available for future investment, and an investment opportunity occurs.

1. Write the optimality equation.
2. Assume that $r(x)$ is nondecreasing, concave, and satisfies $r(0) = 0$. Show that $V_n(X)$ is also concave in X.

$0 \le x \le M$

$\bar{V}_m(A) = pV_m(A) + q\bar{V}_{m+1}(A)$

1. $V_n(X) = \max_{0 \le x \le M} \{ r(x) + \bar{V}_{n-1}(M-x) \}$

2. $V_n(\lambda A_1 + (1-\lambda)A_2) \ge \lambda V(A_1) + (1-\lambda) V(A_2)$; $0 \le \lambda \le 1 \to \dots$