Introduction to Dynamic Programming

Dan Zhang
Leeds School of Business
University of Colorado at Boulder

Outline

- Examples of Sequential Decision Models
  - But Who's Counting
- Problem Definition and Notations
- Single-Product Stochastic Inventory Control

But Who's Counting

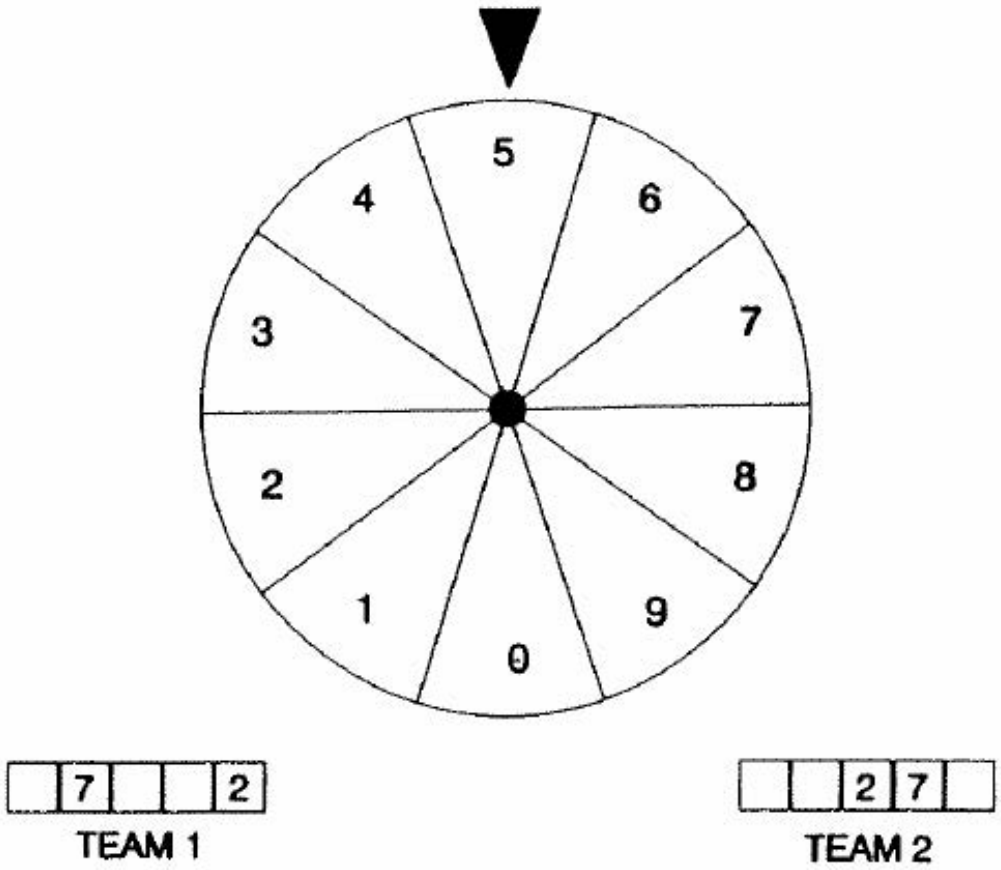


Figure 1.7.1 Spinner for "So Who's Counting."

But Who's Counting

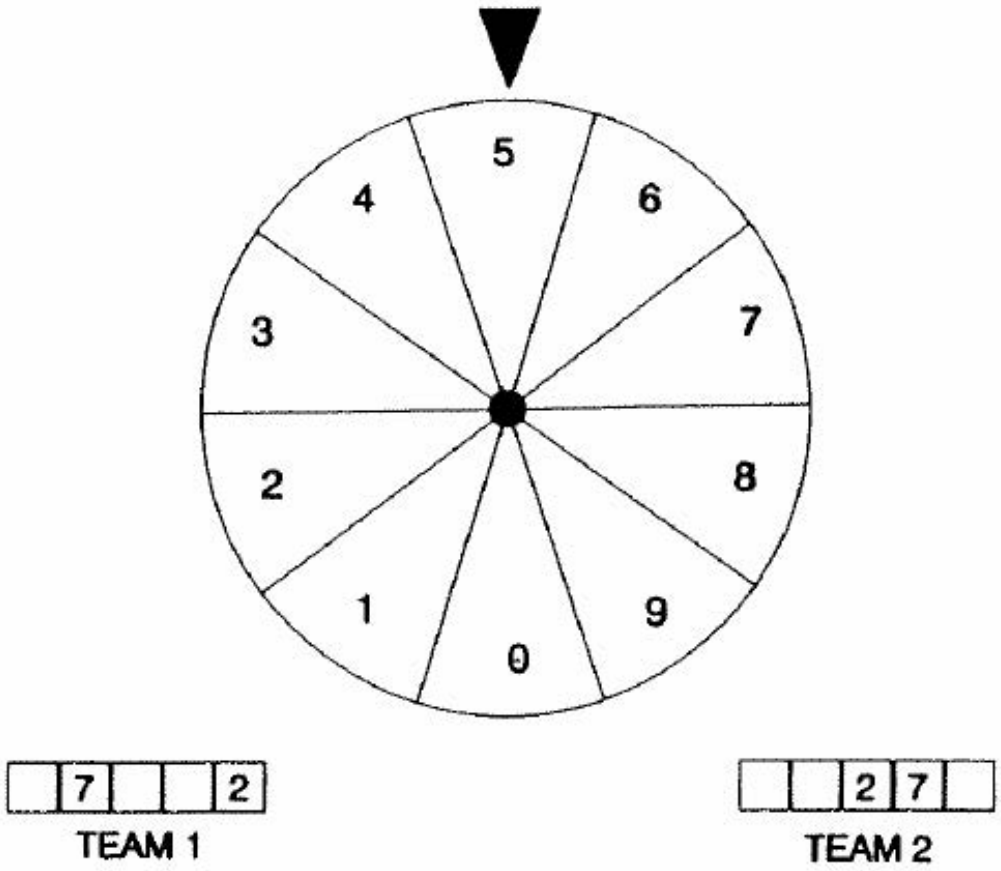


Figure 1.7.1 Spinner for "So Who's Counting."

Square One TV: But Who's Counting?

Optimal Policy for But Who's Counting

- What would you do?

Optimal Policy for But Who's Counting

- What would you do?

Table 1.7.1 Optimal Policy for "But Who's Counting."

Observed Number	Optimal Digit Locations				
	Spin 1	Spin 2	Spin 3	Spin 4	Spin 5
0	5	4	3	2	1
1	5	4	3	2	1
2	5	4	3	2	1
3	4	3	3	2	1
4	3	3	2	2	1
5	3	2	2	1	1
6	2	2	1	1	1
7	1	1	1	1	1
8	1	1	1	1	1
9	1	1	1	1	1

Discussion

- Characteristics of the problem
- How to formulate a mathematical model to find the "best" strategy?
- Should your strategy depend on your opponent's strategy if your goal is to win the game (produce a larger number than your opponent)?

Problem Definition and Notation

We focus on discrete-time models with finite state spaces and action sets.

- Decision epochs and periods
  - The set of decision epochs $T = \{1, \ldots, N\}$.
- State and action sets
  - State space S
  - For each $s \in S$, the set of allowable actions is given by A_s.
- Rewards and transition probabilities
  - A reward $r_t(s, a)$ is received when choosing action $a \in A_s$ in state s at decision epoch t.
  - The system state at the next decision epoch is determined by the probability distribution $p(\cdot|s, a)$.
- Decision rules
- Policies

Decision Rules

- A decision rule prescribes a procedure for action selection in each state at a specified decision epoch.
  - Deterministic Markovian decision rules: $d_t : S \to A$
  - Randomized Markovian decision rules: $d_t : S \to \mathcal{P}(A)$
  - Deterministic history-dependent decision rules: $d_t : H_t \to A$, where

$$H_1 = S,$$
$$H_t = H_{t-1} \times A \times S.$$

  - Randomized history-dependent decision rules: $d_t : H_t \to \mathcal{P}(A)$

Decision Rules

- A decision rule prescribes a procedure for action selection in each state at a specified decision epoch.
  - Deterministic Markovian decision rules: $d_t : S \to A$
  - Randomized Markovian decision rules: $d_t : S \to \mathcal{P}(A)$
  - Deterministic history-dependent decision rules: $d_t : H_t \to A$, where

$$H_1 = S,$$
$$H_t = H_{t-1} \times A \times S.$$

  - Randomized history-dependent decision rules: $d_t : H_t \to \mathcal{P}(A)$

For a given optimality criterion, under what conditions is it optimal to use a deterministic Markovian decision rule at each stage?

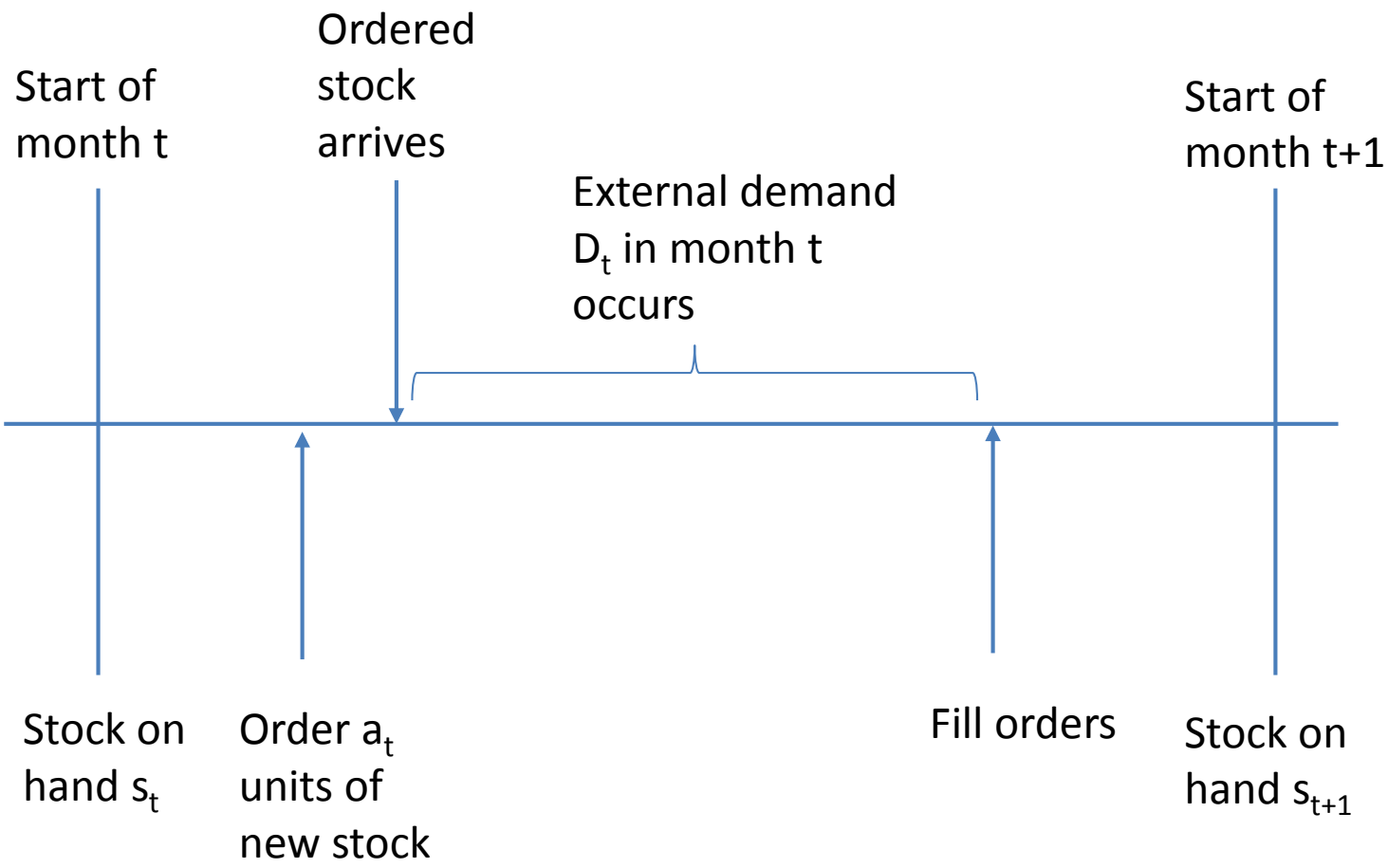
Policies

- A policy (contingency plan, plan, or strategy) specifies the decision rule to be used at all decision epochs.
  - A policy π is a sequence of decision rules; $\pi = (d_1, d_2, \ldots, d_{N-1})$.
  - A policy is stationary if $d_t = d$ for all $t \in T$; in this case, we write $\pi = d^\infty$.
  - A policy π is an element of the set of all policies of a given class; for example, we use Π^{MD} to denote the class of all deterministic Markovian policies.
  - Certain relationships hold among the various classes of policies; randomized, history-dependent policies are most general, while stationary deterministic policies are most specific.

Single-Product Stochastic Inventory Control

Each month, the manager of a warehouse determines current inventory (stock on hand) of a single product. Based on this information, he decides whether or not to order additional stock from a supplier. In doing so, he is faced with a trade-off between the costs associated with keeping inventory and the lost sales (or penalties) associated with being unable to satisfy customer demand for the product. The manager's objective is to maximize some measure of profit (sales revenue less inventory holding and ordering costs) over the decision making horizon. Demand for the product is random with a known probability distribution.

Timing of Events



Notations

- s_t: Inventory on hand at the beginning of month t
- a_t: The number of units ordered by the inventory manager
- D_t: The random demand in month t
  - $P(D_t = j) = p_j$, $j = 0, 1, \ldots$
- s_{t+1}: Inventory on hand at the beginning of month $t+1$
  - $s_{t+1} = \max\{s_t + a_t - D_t, 0\} \equiv [s_t + a_t - D_t]^+$.
- $O(u)$: Cost of ordering u units
- $h(u)$: Cost of maintaining an inventory of u units for a month
- $g(u)$: The value of remaining inventory at the end of the horizon
- $f(j)$: Revenue from satisfying j units of demand
  - $F(u) = \sum_{j=0}^{u-1} f(j)p_j + f(u)\sum_{j=u}^{\infty} p_j$.

A Markov Decision Process Formulation

- Decision epochs: $T = \{1, 2, \ldots, N\}$
- States: $S = \{0, 1, \ldots, M\}$. (Assume M is the maximum possible inventory.)
- Actions: $A_s = \{0, 1, \ldots, M - s\}$.
- Expected rewards: $r_t(s, a) = F(s + a) - O(a) - h(s + a)$, $\forall t = 1, \ldots, N - 1$.
- Terminal rewards: $r_N(s) = g(s)$.
- Transition probabilities:

$$p_t(j|s, a) = \begin{cases} 0, & \text{if } M \geq j > s + a, \\ p_{s+a-j}, & \text{if } M \geq s + a \geq j > 0, \\ \sum_{k=s+a}^{\infty} p_k, & \text{if } M \geq s + a \text{ and } j = 0. \end{cases}$$

An MDP Formulation for But Who's Counting

- Decision epochs: $T = \{1, 2, \ldots, 6\}$
- States: $S = \{(s, i) : s$ is any substring of '12345', and $i \in \{0, \ldots, 9\}\}$.
- Actions: $A_{s,i} = \{0, 1, \ldots, \text{length}(s)\}$
- Expected rewards: $r_t(s, i, a) = i \times s(a) \times 10^{s(a)-1}$, $\forall t = 1, \ldots, 5$.
- Terminal rewards: $r_6 = 0$.
- Transition probabilities:

$$p_t(s', j|s, i, a) = \begin{cases} \frac{1}{10}, & \text{if } s' = s \backslash s(a), j \in \{0, \ldots, 9\}, \\ 0, & \text{otherwise.} \end{cases}$$