Markov Decision Processes

Sequential decision-making with perfect observation

Aditya Mahajan
McGill University

Lecture Notes for ECSE 506: Stochastic Control and Decision Theory
February 8, 2016

Examples of Markov decision processes

These examples illustrate how to use Markov decision theory to establish qualitative properties of optimal strategies. Such properties are useful because:

- they appeal to decision makers,
- they enable efficient computation,
- they are easy to implement.

Optimal gambling

Optimal inventory management

Power-delay trade-off in wireless

Call options

Optimal choice

Energy storage

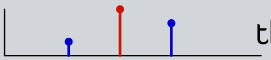
MDP Example: Optimal gambling

Image credit: http://commons.wikimedia.org/wiki/File:Gambling-ca-1800.jpg

Description of an optimization problem faced by a gambler

Optimal gambling

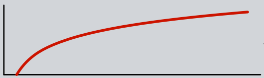
A gambler goes to a casino with an initial fortune of $\$x_1$ and places bets over time and must leave after T bets. Let X_t denote the gambler's fortune after t bets. In this example, time denotes the number of times that the gambler has bet.



Probability of winning and loosing starting at X_t

At time t, the gambler may place a bet for any amount U_t less than his current fortune X_t. If he wins the bet (denoted by the event $W_t = 1$), the casino gives him the amount that he had bet. If he loses the bet (denoted by the event $W_t = -1$), he pays the casino the amount that he had bet.

The outcomes of the bets $\{W_t\}_{t=1}^{T}$ are primitive random variables, i.e., they are independent of each other, of the gambler's initial fortune, and the gambler's betting strategy. Let $\mathbb{P}(W_t = 1) = p$.


Utility of gambler: $\log x$

The gambler's payoff is $\log X_T$. Find the optimal gambling strategy for the gambler that maximizes the expected value of his payoff.

Mathematical setup of optimal gambling problem

Notation

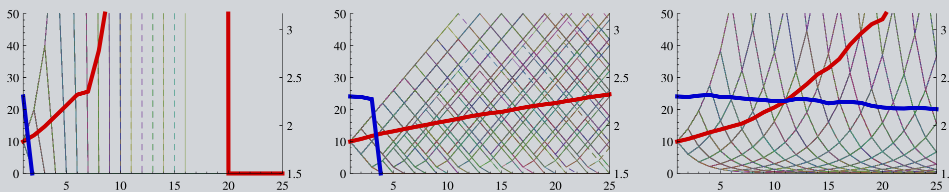
- State : $X_t \in \mathbb{R}_{\geqslant 0}$
- Action : $U_t \in \mathbb{R}_{\geqslant 0}$
- Feasible actions: $\mathcal{U}_t(x_t) = \{u_t \in \mathbb{R}_{\geqslant 0} : u_t \leqslant x_t\}$

Dynamics $X_{t+1} = X_t + W_t U_t$ where $U_t = g_t(X_{1:t}, U_{1:t-1})$

Rewards

- Per step reward: $r_t(x_t, u_t) = 0$
- Terminal reward: $r_T(x_T) = \log x_T$

Illustration Fortune of gambler over time for three possible strategies for $x_1 = 10$, $p = 0.6$, $T = 25$ (1000 sample paths).



Strategy $U_t = X_t$ | Strategy $U_t = \min\{4, X_t\}$ | Strategy $U_t = 0.4X_t$

(red line) denotes $\mathbb{E}[X_t]$; (blue line) denotes $30\,\mathbb{E}[\log X_t]$.

The optimal gambling problem is a special case of a MDP

	MDP Dynamic Model	Optimal Gambling
System Dynamics	$X_{t+1} = f_t(X_t, U_t, W_t)$	$X_{t+1} = X_t + W_t U_t$
Information Structure	$U_t = g_t(X_{1:t}, U_{1:t-1})$	$U_t = g_t(X_{1:t}, U_{1:t-1})$
Objective Function	$\mathbb{E}\left[\sum_{t=1}^{T-1} r_t(X_t, U_t) + r_T(X_T)\right]$	$\mathbb{E}[\log X_T]$

Structure of Controller

Using Markov strategies does not entail any loss of optimality

$$U_t = g_t(X_t)$$

Dynamic program

$$V_T(x_T) = r_T(x_T);$$

$$V_t(x_t) = \max_{u_t \in \mathcal{U}_t(x_t)} \Big\{ r_t(x_t, u_t) + \mathbb{E}[V_{t+1}(f_t(x_t, u_t, W_t))] \Big\},$$

$$t = T-1, \dots, 1.$$

Closed form solution of optimal gambling

Theorem When $p \leqslant 0.5$:

- the optimal strategy is to not gamble, specifically, $g_t(x) = 0$;
- the value function is $V_t(x) = \log x$.

When $p > 0.5$:

- the optimal strategy is to bet a fraction of the current fortune, specifically, $g_t(x) = (2p - 1)x$;
- the value function is $V_t(x) = \log x + (T - t)C$ where $C = \log 2 + p \log p + (1 - p) \log(1 - p)$.

Backward induction proof of the solution $(p \leqslant 0.5)$

Proof of Case 1: $p \leqslant 0.5$

Let $p = \mathbb{P}(W_t = 1)$ and $q = \mathbb{P}(W_t = -1)$. Then $p \leqslant 0.5$ implies $p \leqslant q$. Proceed by backward induction.

- **Basis**: For $t = T$, $V_T(x) = \log x$.
- **Induction hypothesis**: For $t = t+1$, $V_{t+1}(x) = \log x$, and $g_{t+1}(x) = 0$.
- **Induction step**: Define $Q_t(x, u) = pV_{t+1}(x+u) + qV_{t+1}(x-u)$.

$$\frac{\partial Q_t(x,u)}{\partial u} = \frac{p}{x+u} - \frac{q}{x-u} < 0; \implies Q_t(x,u) \text{ is decreasing in } u$$

$$\therefore \quad g_t(x) = \arg\max_{u \in [0,x]} Q_t(x,u) = 0; \implies V_t(x) = Q_t(x, g_t(x)) = \log x.$$

Backward induction proof of the solution $(p > 0.5)$

Proof of Case 2: $p > 0.5$

Let $p = \mathbb{P}(W_t = 1)$ and $q = \mathbb{P}(W_t = -1)$. Then $p > 0.5$ implies $p > q$.

Proceed by backward induction.

- Basis: For $t = T$, $V_T(x) = \log x$.
- Induction hypothesis: For $t = t + 1$,

$$V_{t+1}(x) = \log x + (T - t - 1)C, \quad \text{and} \quad g_{t+1}(x) = (p - q)x;$$

where $C = \log 2 + p \log p + q \log q$.

- Induction step: Define $Q_t(x, u) = pV_{t+1}(x + u) + qV_{t+1}(x - u)$.

$$\frac{\partial Q_t(x, u)}{\partial u} = \frac{p}{x + u} - \frac{q}{x - u}; \implies \text{Extremum } u = (p - q)x.$$

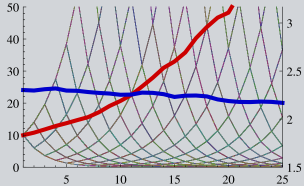
$$\text{and} \quad \frac{\partial^2 Q_t(x, u)}{\partial u^2} = -\frac{p}{(x + u)^2} - \frac{q}{(x - u)^2} < 0;$$

$$\therefore \quad g_t(x) = \arg \max_{u \in [0, x]} Q_t(x, u) = (p - q)x;$$

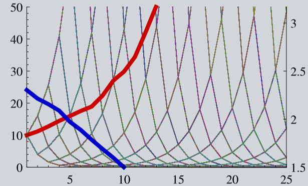
$$\implies V_t(x) = Q_t(x, g_t(x)) = \log x + (T - t)C.$$

Maximizing $\mathbb{E}[\log X_T]$ does not maximize $\mathbb{E}[X_T]$

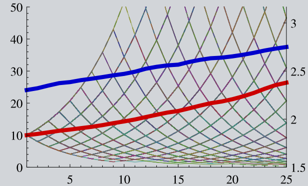
Illustration Recall previous setup: $x_1 = 10$, $p = 0.6$, $T = 25$ (1000 sample paths).



Strategy $U_t = 0.4X_t$

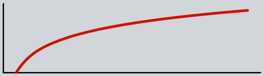


Strategy $U_t = 0.6X_t$



Strategy $U_t = 0.2X_t$

(red line) denotes $\mathbb{E}[X_t]$; (blue line) denotes $30\,\mathbb{E}[\log X_t]$.



Utility of gambler: $\log x$

The strategy $g_t(x) = (p = q)x = 0.2x$ maximizes $\mathbb{E}[\log X_T]$.

It does **not maximize** $\mathbb{E}[X_T]$ or $\mathbb{E}[\log X_T]$.

Generalized model: If terminal reward is increasing in x, then value function is increasing in x and decreasing t

Generalization The terminal reward $r_T(x)$ is monotone increasing in x

Theorem For the generalized optimal gambling problem

- For each x, the value function $V_t(x)$ is monotone decreasing in t.
- For each t, the value function $V_t(x)$ is monotone increasing in x.

Proof: $V_t(x)$ is monotone in t Let $p = \mathbb{P}(W_t = 1)$ and $Q_t(x, u) = pV_{t+1}(x+u) + (1-p)V_{t+1}(x-u)$.

$$\text{Then,}\quad V_t(x) = \max_{u \in [0,x]} Q_t(x,u) \geqslant Q_t(x,0) = V_{t+1}(x).$$

Proof: $V_t(x)$ is monotone in x Proceed by backward induction.

- Basis: By assumption, $r_T(x)$ is monotone increasing in x.
- Induction hypothesis: $V_{t+1}(x)$ is monotone increasing in x.
- Induction step: $\forall x_1, x_2, u \in \mathbb{R}_{\geqslant 0}$, such that $x_1 \leqslant x_2$ and $u \leqslant x_1$,

$$V_{t+1}(x_1) \leqslant V_{t+1}(x_2) \implies Q_t(x_1, u) \leqslant Q_t(x_2, u).$$

$$\therefore V_t(x_1) = \max_{u \in [0, x_1]} Q_t(x_1, u) \leqslant \max_{u \in [0, x_1]} Q_t(x_2, u) \leqslant \max_{u \in [0, x_2]} Q_t(x_2, u) = V_t(x_2)$$

Exercises and further reading on optimal gambling

1. For generalization of this problem, read: Sheldon M. Ross, "Dynamic Programming and Gambling Models", Advances in Applied Probability, Vol. 6, No. 3 (Sep., 1974), pp. 593-606. http://www.jstor.org/stable/1426236

2. Find the expected reward of using the all-in strategy $g_t(x) = x$.

3. Find the expected reward of using the proportional-betting strategy $g_t(x) = \alpha x$ as a function of α. Use this expression to optimize over the value of α.

4. Bonus question: Find conditions on the terminal reward function r_T such that the optimal gambling strategy is increasing in x.

MDP Example: Optimal inventory management

Image credit: http://commons.wikimedia.org/wiki/File:Modern_warehouse_with_pallet_rack_storage_system.jpg

Description of an optimization problem faced by online retailers in managing inventory

Inventory management

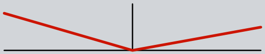
Retail stores stockpile products in warehouses to meet the random demand. Additional stocks are procured at regular intervals. Let X_t denote the amount of stock before the t-th procurement. In this example, time denotes the number of additional stock procurements.

At time t, the store may procure an addition stock U_t units at a cost of \$$p$ per unit. Thus the total procurement cost is pU_t.

The random demand W_t is i.i.d. with distribution P_W. The stock available at the next time is $X_{t+1} = X_t + U_t - W_t$, where a negative stock denotes backlogged demand.

Holding cost $h(x)$

$$\begin{cases} ax, & \text{if } x \geqslant 0 \\ -bx, & \text{if } x < 0 \end{cases}$$



Holding cost: $h(x)$

The holding cost for the stock is given by $h(x)$ where a is the per-unit storage cost and b is the per-unit backlog cost.

Per-stage cost is $c(X_{t+1}, U_t) = h(X_{t+1}) + pU_t$. Find the optimal inventory control strategy to minimize the expected total cost over a finite horizon.

Mathematical setup of the inventory management problem

Notation State : $X_t \in \mathbb{Z}$

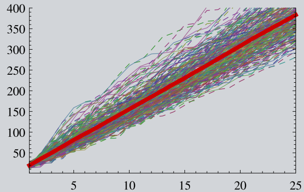
Action: $U_t \in \mathbb{Z}_{\geqslant 0}$

Dynamics $X_{t+1} = X_t + U_t - W_t$, where $U_t = g_t(X_{1:t}, U_{1:t-1})$.

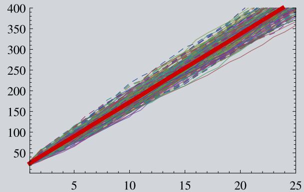
Cost Per-stage cost: $c_t(x_{t+1}, u_t) = h(x_{t+1}) + p u_t$

Terminal cost : $c_T(x_{T+1}) = h(x_{T+1})$

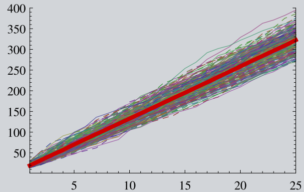
Illustration Cost incurred by the retail store for three possible strategies for $x_1 = 0$, $p = 1$, $a = 2$, $b = 3$, $P_W = \text{Unif}[0, 10]$, $T = 25$ (250 sample paths)



Strategy $U_t = 10\, \mathbb{1}_{\{x \leqslant 0\}}$



Strategy $U_t = 10\, \mathbb{1}_{\{x \leqslant 5\}}$



Strategy $U_t = (7 - x)\, \mathbb{1}_{\{x \leqslant 7\}}$

Optimal inventory management is a special case of a MDP

	MDP Dynamic Model	**Optimal inventory management**
System Dynamics	$X_{t+1} = f_t(X_t, U_t, W_t)$	$X_{t+1} = X_t + U_t - W_t$
Information Structure	$U_t = g_t(X_{1:t}, U_{1:t-1})$	$U_t = g_t(X_{1:t}, U_{1:t-1})$
Objective Function	$\mathbb{E}\Big[\sum_{t=1}^{T} c_t(X_t, U_t, X_{t+1})\Big]$	$\mathbb{E}\Big[\sum_{t=1}^{T} pU_t + h(X_{t+1}))\Big]$

Structure of Controller

Using Markov strategies does not entail any loss of optimality

$$U_t = g_t(X_t)$$

Dynamic program

$$V_{T+1}(x_{T+1}) = 0;$$

$$V_t(x_t) = \min_{u_t \in \mathcal{U}_t(x_t)} \mathbb{E}[c_t(x_t, u_t, X_{t+1}) + V_{t+1}(X_{t+1}) \mid X_t = x_t, U_t = u_t], \qquad t = T, \dots, 1.$$

Qualitative properties of the value function

Definition

$$Y_t = X_t + U_t$$

$$L(y_t) = \mathbb{E}\left[a[y_t - W_t]^+ + b[W_t - y_t]^+\right], \qquad \text{where } [x]^+ = \max(0, x).$$

$$Q_t(y_t) = py_t + L(y_t) + \mathbb{E}[V_{t+1}(y_t - W_t)]$$

$$S_t = \arg\min_{y_t \in \mathbb{R}} Q_t(y_t)$$

Lemma $L(y)$ is convex in y.

This result is true as long as the holding cost is convex.

Lemma $V_t(x) = \min_{y \geqslant x} Q_t(y) - px$ and $g_t(x) = y_t^* - x_t$ where $y_t^* = \arg\min_{y \geqslant x} Q_t(y)$.

Theorem

- $\forall x, y$, the functions $Q_t(y)$ and $V_t(x)$ are decreasing in t.
- $\forall t$, $V_t(x) + px$ is increasing in x.

Backward induction proof of qualitative properties

Proof of monotonicity in t

Proceed by backward induction.

- Basis: For completeness, define $Q_{T+1}(y) \equiv py$.
  - By definition, $Q_{T+1}(y) = py \leqslant py + L(y) = Q_T(y)$.
  - By definition, $V_{T+1}(x) = 0 \leqslant V_T(x)$.
- Induction hypothesis: $V_{t+1}(x) \leqslant V_{t+2}(x)$ for all x.
- Induction step:

$$Q_t(y) = py + L(y) + \mathbb{E}[V_{t+1}(y - W)]$$

$$\geqslant py + L(y) + \mathbb{E}[V_{t+2}(y - W)] = Q_{t+1}(y)$$

Similarly,

$$V_t(x) = \min_{y \geqslant x} Q_t(y) - px$$

$$\geqslant \min_{y \geqslant x} Q_{t+1}(y) - px = V_{t+1}(x)$$

Proof of monotonicity in x

In the next Theorem, we show that $Q_t(y)$ is convex in y for all t. Therefore, $V_t(x) + px = \min_{y \geqslant x} Q_t(y)$ is increasing in x.

A base stock strategy is optimal

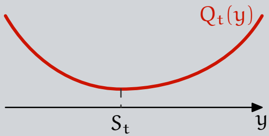
Theorem For all t, $Q_t(y)$ and $V_t(x)$ are convex in y and x respectively. Furthermore, V_t is given by

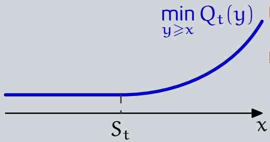
$$V_t(x) = \begin{cases} Q_t(S_t) - px, & \text{if } x \leqslant S_t \\ Q_t(x) - px, & \text{otherwise} \end{cases}$$

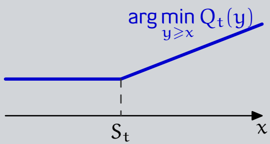
and the optimal strategy is a base stock strategy given by

$$g_t^*(x_t) = [S_t - x_t]^+.$$

Backward induction proof of the optimal strategy

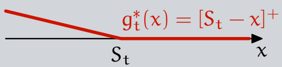






Proof Proceed by backward induction.

- **Basis**:
  - $Q_T(y) = py + L(y)$ and is, therefore, convex.
  - $V_T(x) = \min_{y \geqslant x} Q_T(y) - px$. The minimizing $y = \max(x, S_T)$.
  - $V_T(x)$ is convex and has the desired form.
- **Induction hypothesis**: $V_{t+1}(x)$ is convex and has the desired form.
- **Induction step**:
  - $Q_t(y) = py + L(y) + \mathbb{E}[V_{t+1}(y - W_t)]$ is convex.
  - $V_t(x) = \min_{y \geqslant x} Q_t(y) - px$. The minimizing $y = \max(x, S_t)$.
  - $V_t(x)$ is convex and has the desired form.



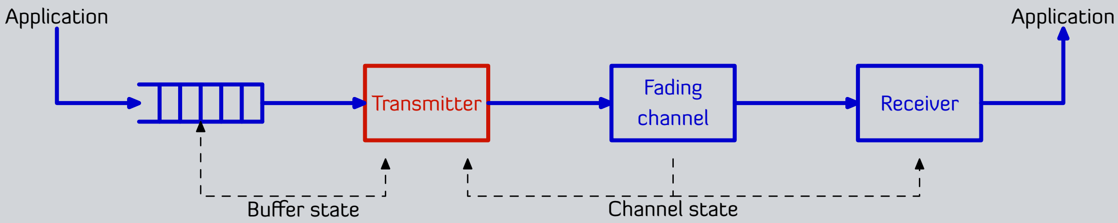
Further reading on optimal inventory management

1. The mathematical model of inventory management considered here was originally proposed in the following seminal paper: Kenneth J. Arrow, Theodore Harris, Jacob Marschak "Optimal Inventory Policy", Econometrica, pp 250-272, Jul 1951. http://www.jstor.org/stable/1906813

2. The optimality of base-stock policy was first presented in R. Bellman, I. Glicksberg and O. Gross, "On the optimal inventory equation", Management Science, pp 83-104, 1955. http://www.jstor.org/stable/2627240

3. Bonus question: Find conditions under which the optimal thresholds S_t are decreasing in t.

MDP Example: Optimal power-delay trade-off in wireless communication



Design of rate-allocation protocol in wireless communication

Rate allocation in MAC layer

In a cell phone, higher layer applications (voice, email, etc.) data packets; these packets are buffered in a queue and the transmission protocol decides how many packets to transmit at each step.

At time t, $X_t \in \mathbb{Z}_{\geqslant 0}$ packets are buffered in the queue; the transmission protocol transmits $U_t \leqslant X_t$, $U_t \in \mathbb{Z}_{\geqslant 0}$ packets, and $W_t \in \mathbb{Z}_{\geqslant 0}$ new packets arrive. Thus, $X_{t+1} = X_t - U_t + W_t$. The delay incurred by the packets are proportional to $d(X_t - U_t)$, where

- $d(\cdot)$ is strictly increasing and convex; moreover $d(0) = 0$.

Power-allocation in physical layer

The transmission protocol sets the transmit power such that the signal to noise ratio (SNR) at the receiver, which depends on channel fading, is sufficiently high.

At time t, $S_t \in \mathcal{S}$ denotes the state of channel fading. The transmit power is proportional to $p(U_t) \cdot q(S_t)$, where

- $p(\cdot)$ is strictly increasing and convex; moreover $p(0) = 0$.
- $q(\cdot)$ is strictly decreasing and convex.

Design of rate-allocation protocol in wireless communication

Primitive variables

- The initial state X_1 has distribution P_X.
- The arrival process $\{W_t\}_{t=1}^T$ is an i.i.d. process with distribution P_W.
- The channel state $\{S_t\}_{t=1}^T$ is a stochastic monotone Markov process, i.e., for any increasing function $f\colon \mathcal{S} \to \mathbb{R}$,

  $$h(s) = \mathbb{E}[f(S_{t+1}) \mid S_t = s] \text{ is increasing.}$$
- X_1, $\{W_t\}_{t=1}^T$, and $\{S_t\}_{t=1}^T$ are mutually independent.

A Markov process with transition matrix P is stochastic monotone if

$$Q_{ik} = \sum_{j \geqslant k} P_{ij}$$

is increasing in i for all k.

Objective

The objective is to choose a transmission strategy $(g_1, \dots, g_T)$ where

$$U_t = g_t(X_{1:t}, S_{1:t}, U_{1:t-1})$$

to minimize the total expected cost

$$\mathbb{E}\Big[\sum_{t=1}^{T} c(X_t, S_t, U_t)\Big]$$

where $c(X_t, S_t, U_t) = p(U_t) \cdot q(S_t) + d(X_t - U_t)$.

Power-delay trade-off as a special case of MDP

	MDP Dynamic Model	**Power-delay trade-off**
System Dynamics	$X_{t+1} = f_t(X_t, U_t, W_t)$	$X_{t+1} = X_t - U_t + W_t$ $\{S_t\}_{t=1}^{T}$ independent Markov process
Information Structure	$U_t = g_t(X_{1:t}, U_{1:t-1})$	$U_t = g_t(X_{1:t}, S_{1:t}, U_{1:t-1})$
Objective Function	$\mathbb{E}\Big[\sum_{t=1}^{T} c_t(X_t, U_t, X_{t+1})\Big]$	$\mathbb{E}\Big[\sum_{t=1}^{T} d(X_t - U_t) + p(U_t)q(S_t)\Big]$

Structure of Controller Using Markov strategies does not entail any loss of optimality

$$U_t = g_t(X_t, S_t)$$

Dynamic program

$$V_{T+1}(x_{T+1}, s_{T+1}) = 0;$$

$$V_t(x_t, s_t) = \min_{u_t \in \mathcal{U}_t} \Big\{ c_t(x_t, s_t, u_t) + \mathbb{E}[V_{t+1}(X_{t+1}, S_{t+1}) \mid X_t = x_t, S_t = s_t, U_t = u_t] \Big\}, \qquad t = T, \dots, 1.$$

Qualitative properties of the value function

Theorem (Monotonicity and convexity)

- $\forall t$: $V_t(x, s)$ is increasing in x for all s; and decreasing in s for all x.
- $\forall t$: $V_t(x, s)$ is convex in x for all s.

Theorem (Structural property)

Let $g^* = (g_1^*, \dots, g_T^*)$ be an optimal strategy. Then,

- $\forall t$: $g_t^*(x, s)$ is increasing in x for all s.

Thus, the optimal strategy is monotone.

Proof of monotonicity of value function

Proof We show that the model satisfies the sufficient conditions (C1) and (C2) for monotonicity of value function (see notes on MDP theory).

First note that $\mathcal{U}(x) = [0, x]$ satisfies:

- $\mathcal{U}(x) \subseteq \mathcal{U}(x')$ for all $x' > x$
- For any $x \in \mathcal{X}$, $u, u' \in \mathcal{U}$ such that $u' < u$, if $u \in \mathcal{U}(x)$ then $u' \in \mathcal{U}(x)$.

Therefore, we can use the Theorem.

C1. For fixed x and u , $c_t(x, s, u)$ is decreasing in s
For fixed s and u, $c_t(x, s, u)$ is increasing in x.

C2. For a fixed x, $\{S_t\}_{t \geqslant 1}$ is stochastic monotone.
For a fixed s and u, $\{X_t\}_{t \geqslant 1}$ is stochastic monotone because for any increasing v and $x' \geqslant x$

$$\mathbb{E}[v(x' - u + W)] \geqslant \mathbb{E}[v(x - u + W)]$$

Hence, the result follows from the theorem on monotonicity of value functions.

Backward induction proof of convexity of value function

Proof (Convexity)

A direct proof of convexity does not work. For fixed s and u, we can show that $Q_t(x, s, u)$ is convex. But minimum of convex functions is not convex.

Proceed by backward induction.

- Basis: Fix s and $x > 1$. Let $\underline{u} = g^*_T(x-1, s)$ and $\bar{u} = g^*_T(x+1, s)$.

$$V_T(x-1, s) + V_T(x+1, s) = Q_T(x-1, s, \underline{u}) + Q_T(x+1, s, \bar{u})$$
$$= d(x-1-\underline{u}) + d(x+1+\bar{u}) + \big[p(\underline{u}) + p(\bar{u})\big] q(s)$$

by convexity of $d(\cdot)$ and $p(\cdot)$

$$\geqslant d(x-\underline{v}) + d(x-\bar{v}) + \big[p(\underline{v}) + p(\bar{v})\big] q(s)$$
$$= Q_T(x, s, \underline{v}) + Q_T(x, s, \bar{v})$$
$$\geqslant 2 \min_{u \geqslant 0} Q_T(x, s, u) = 2V_T(x, s)$$

where $\underline{v} = \lfloor (\underline{u} + \bar{u})/2 \rfloor$ and $\bar{v} = \lceil (\underline{u} + \bar{u})/2 \rceil$.

Thus, for a fixed s, $V_T(x, s)$ is convex in x.

- Induction hypothesis: For fixed s, $V_{t+1}(x, s)$ is convex in x.
- Induction step:Follow the same argument as above with $d(x-u)$ replaced by

$$d(x-u) + \mathbb{E}[V_{t+1}(x-u+W, S_{t+1}) \mid S_t = s].$$

which is convex in x.

Since $\underline{u} \leqslant x-1$ and $\bar{u} \leqslant x+1$, we have that $\underline{v} \leqslant x$ and $\bar{v} \leqslant x$. Thus, the next state is given by $x - u + W$.

Proof that optimal strategy is monotone

Proof Define:

$$h_t(x-u, s) = d(x-u) + \mathbb{E}[V_{t+1}(x-u+W, S_{t+1}) \mid S_t = s]$$

which is increasing and convex in $x-u$. Since $h_t(x-u,s)$ is convex in $x-u$, $\partial^2 h_t(x-u,s)/\partial x \partial u \leqslant 0$. Hence, h_t is submodular in (x,u). Thus, for a fixed u,

$$Q_t(x,s,u) = p(u) \cdot q(s) + h_t(x-u,s)$$

is submodular in (x,u). Therefore, the optimal strategy is increasing in x. (See notes on MDP theory).

Additional properties for i.i.d. fading

Theorem (i.i.d. fading) Suppose $\{S_t\}_{t=1}^{T}$ is an i.i.d. process. Then

- $\forall t$: $V_t(x, s)$ is convex in s for all x.
- $\forall t$: $g_t^*(x, s)$ is increasing in s for all x.

Proof Since $\{S_t\}_{t \geqslant 1}$ is i.i.d., $h_t(x-u, s)$ (defined on prev. slide) does not depend on s and we can write it as $h_t(x-u)$. Thus,

$$Q_t(x, s, u) = p(u)q(s) + h_t(x-u)$$

- $\partial^2 Q_t/\partial s^2 = p(u)\partial^2 q/\partial s^2 \geqslant 0$. Hence, Q_t is convex in s.

  Hence, $V_t(x, s)$ is convex in s.

- $\partial^2 Q_t/\partial s \partial u = \dot{p}(u)\dot{q}(s) \leqslant 0$. Hence, Q_t is submodular in (s, u).

  Hence, $g_t(x, s)$ is increasing in s (for a fixed x).

Exercises and further reading on power-delay trade-off

1. The mathematical model of power-delay trade-off considered here is from: Randall Berry, "Power and Delay Trade-offs in Fading Channels," Phd Thesis, MIT, June, 2000, http://www.ece.northwestern.edu/~rberry/thesis.pdf

2. For a more detailed characterization of the optimal transmission strategy when the average power goes to zero, see: Randall Berry and Robert Gallager, "Communication over fading channels with delay constraints," IEEE Transactions on Information Theory, vol. 48, pp. 1135–1149, May 2002.

3. For a more detailed characterization of the optimal transmission strategy when the average delay goes to zero, see: Randall Berry, "Optimal power-delay trade-offs in fading channels—small delay asymptotics," IEEE Transactions on Information Theory, vol. 59, no. 6, pp. 3939-3952, June 2013.

Optimal Stopping Example: Call options

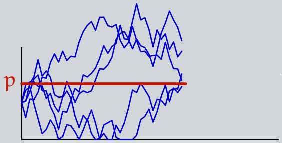
Image credit: http://commons.wikimedia.org/wiki/File:Sao_Paulo_Stock_Exchange.jpg

An optimization problem arising in trading of call options

Call options

An investor has a call option to buy one share of a stock at a fixed price of $\$p$ and has T days to exercise this option. For simplicity, assume that the investor makes a decision at the beginning of each day.

If the investor exercises the option when the stock price is $\$x$, he gets $\$(x - p)$. The investor may decide not to exercise this option at all.



Price of a stock
with independent increments

Assume that the price of the stock varies with independent increments. More precisely, the value X_t of the stock on day t is

$$X_t = X_0 + \sum_{k=1}^{t} W_t$$

where $\{W_t\}_{t=1}^{T}$ is an i.i.d. process independent of X_0. Assume that $\mathbb{E}[W_t] = \mu_W < \infty$.

Let τ denote the day stopping time when the investor exercises his option. Find the optimal investment strategy for the investor that maximizes $\mathbb{E}[(X_\tau - p)\, \mathbb{1}[\tau \leqslant T]]$.

Mathematical setup of call options

This problem is an optimal stopping problem in which a single stopping decision has to be made: when to exercise the option.

Notation

State : $X_t \in \mathbb{R}_{\geqslant 0}$

Action: $U_t \in \{0, 1\}$

- $U_t = 0$ means do not exercise the option;
- $U_t = 1$ means exercise the option.

Dynamics

$X_{t+1} = X_t + W_t$, where $U_t = g_t(X_{1:t}, U_{1:t-1})$

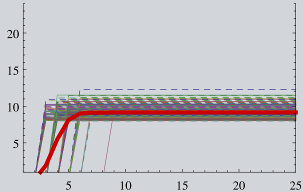
Cost

Continuation reward: $c_t(X_t) = 0$

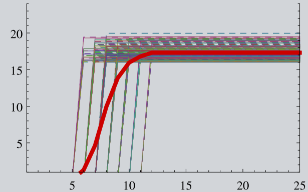
Stopping reward : $c^*(X_t) = X_t - p$

Illustration

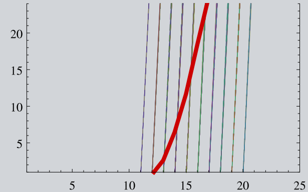
Profit earner by the investor for three possible strategies for $p = 50$, $\mu = 2$, $\sigma^2 = 1$, $x0 \sim \mathcal{N}(p, \sigma^2)$, $W \sim \mathcal{N}(\mu, \sigma^2)$ (250 sample paths)



Strategy $U_t = \mathbb{1}[X_t > p + 8\sigma^2]$



Strategy $U_t = \mathbb{1}[X_t > p + 16\sigma^2]$



Strategy $U_t = \mathbb{1}[X_t > p + 32\sigma^2]$

Call options is a special case of a MDP

	MDP Dynamic Model	**Call Options**
System Dynamics	$X_{t+1} = f_t(X_t, U_t, W_t)$	$X_{t+1} = X_t + W_t U_t$
Information Structure	$U_t = g_t(X_{1:t}, U_{1:t-1})$	$U_t = g_t(X_{1:t}, U_{1:t-1})$
Objective Function	$\mathbb{E}\left[\sum_{t=1}^{T} r_t(X_t, U_t)\right]$	$\mathbb{E}\left[X_\tau - p\right]$

Structure of Controller

Using Markov strategies does not entail any loss of optimality

$$U_t = g_t(X_t)$$

Dynamic program

$$V_{T+1}(x_{T+1}) = 0;$$

$$V_t(x_t) = \max_{u_t \in \mathcal{U}_t} \Big\{ r_t(x_t, u_t) + \mathbb{E}[V_{t+1}(X_{t+1}) \mid X_t = x_t, U_t = u_t] \Big\}, \qquad t = T, \dots, 1.$$

Qualitative properties of the value function

Dynamic Porgram

$$V_{T+1}(x) = 0$$
$$V_t(x) = \max\{x - p, \mathbb{E}[V_{t+1}(x + W)]\}$$

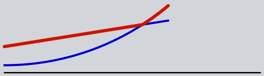
Theorem (Monotonicity properties)

- $\forall t$: $V_t(x)$ is increasing in x
- $\forall t$: $V_t(x) - x$ is decreasing in x.
- $\forall x$: $V_t(x)$ is decreasing in t.

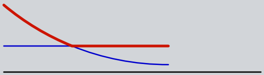
Theorem (Structural properties) There exist numbers $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_T$ such that it is optimal to exercise an option at time t iff $x_t \geqslant s_t$. Hence, the optimal strategy is of threshold type.

Backward induction proof of monotonicity properties

Proof of monotonicity properties



maximum of two increasing functions



maximum of a constant and a decreasing function

Proceed by backward induction.

- Basis:
  - $V_T(x) = \max\{x - p, 0\}$ is increasing in x.
  - $V_T(x) - x = \max\{-p, -x\}$ is decreasing in x.
  - $V_T(x) = \max\{x - p, 0\} \geqslant V_{T+1}(x)$.
- Induction hypothesis: Assume that all results are true for $t = t + 1$.
- Induction step:
  - $V_t(x) = \max\{\underbrace{x - p}_{\text{increasing in } x}, \underbrace{\mathbb{E}[V_{t+1}(x + W)]}_{\text{increasing in } x}\}$ is increasing in x.
  - $V_t(x) - x$ is decreasing in x because

$$V_t(x) - x = \max\{\underbrace{-p}_{\text{const}}, \underbrace{\mathbb{E}[V_{t+1}(x+W) - (x+W)]}_{\text{decreasing in } x} + \mu_W\}$$

  - By the induction hypothesis $V_{t+1}(x) \geqslant V_{t+2}(x)$. Thus,

$$\begin{aligned} V_t(x) &= \max\{x - p, \mathbb{E}[V_{t+1}(x + W)]\} \\ &\geqslant \max\{x - p, \mathbb{E}[V_{t+2}(x + W)]\} \\ &= V_{t+1}(x) \end{aligned}$$

Backward induction proof of the structural properties

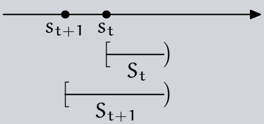
Lemma If the selling action is optimal at x°, then it is optimal at all $x \geqslant x^\circ$.

Proof Let $x \geqslant x^\circ$. Since the selling action is optimal at x°.

$$-p \geqslant \mathbb{E}[V_{t+1}(x^\circ + W)] - x^\circ \geqslant \mathbb{E}[V_{t+1}(x + W)] - x$$

where the second inequality follows from monotonicity of $V_t(x) - x$.

Proof of the structural property Define $S_t = \{x : g_t(x) = 1\}$ or equivalently, $\{x : x - p \geqslant \mathbb{E}[V_{t+1}(x + W)\}$. The previous lemma shows that S_t is of the form $[s_t, \infty)$ where $s_t = \min S_t$. This proves the structural result.



To show that $\{s_t\}_{t=1}^{T}$ is decreasing, we show that $S_t \subseteq S_{t+1}$. Let $x \in S_t$, then $x - p \geqslant \mathbb{E}[V_t(x + W)] \geqslant \mathbb{E}[V_{t+1}(x + W)]$ Hence, $x \in S_{t+1}$.

Exercises and further reading on option pricing

1. The mathematical model of option pricing considered here was originally investigated in the following paper: Howard M. Taylor, "Evaluating a Call Option and Optimal Timing Strategy in the Stock Market", Management Science, Vol. 14, No. 1, pp. 111-120, Sep 1967.
   http://www.jstor.org/stable/2628546

2. Show that if $\mu_W > 0$, then $s_t = \infty$ for all t. Thus, the result presented here is useful only when the mean drift is negative.

3. Bonus question: Find a closed form expression for $V_t(x)$ when $W \sim \mathcal{N}(\mu, \sigma^2)$.

Optimal Stopping Example: Optimal choice

Image credit: http://www.americanscientist.org/issues/feature/2009/2/knowing-when-to-stop/

Optimal choice of the best alternative

Optimal choice

A decision maker (DM) wants to select the best alternative from a set of T alternatives. The DM evaluates the alternatives sequentially. After evaluating alternative t, the DM knows whether alternative t was the best alternative so far or not. Based on this information, the DM must decide whether to choose alternative t and stop the search, or to permanently reject alternative t and evaluate remaining alternatives. The DM may reject the last alternative and not make a choice at all.

All alternatives are equally likely to be the best.

Find the optimal choice strategy that maximize the probability of picking the best alternative.

This optimization problem is known by different names including secretary problem (in which the alternatives correspond to finding the best candidate as a secretary), marriage problem (in which the alternatives correspond of find the best spouse), Googol (in which the alternatives consist of finding the biggest number), parking problem (in which the alternatives correspond to finding the nearest parking spot) and so on.

Optimal choice of the best alternative

The problem is an optimal stopping problem in which a single stopping decision has to be made: when to select the current alternative.

Notation

State : $X_t \in \{0, 1\}$.

- $X_t = 1$ means that the current alternative is the best so far.

Action: $U_t \in \{0, 1\}$.

- $U_t = 1$ means to choose alternative t
- $U_t = 0$ means to reject alternative t

Dynamics

$\{X_t\}_{t=1}^{T}$ independent with $\mathbb{P}(X_t = 1) = 1/t$.

Reward

The continuation reward is zero.

The DM receives a stopping reward only if the current alternative is the best (i.e., better than all previous alternatives (so $X_t = 1$) and better than all future alternatives (so $X_{t+1:T} = 0$).

The expected stopping reward conditioned on X_t is

$$r_t^*(X_t) = X_t \cdot \mathbb{P}(X_{t+1:T} = 0 \mid X_t = 1) = X_t \cdot \frac{t}{T}.$$

Optimal choice is a special case of a MDP

	MDP Dynamic Model	Optimal choice
System Dynamics	$X_{t+1} = f_t(X_t, U_t, W_t)$	X_{t+1} independent
Information Structure	$U_t = g_t(X_{1:t}, U_{1:t-1})$	$U_t = g_t(X_{1:t}, U_{1:t-1})$
Objective Function	$\mathbb{E}\Big[\sum_{t=1}^{T} r_t(X_t, U_t)\Big]$	$\mathbb{E}\Big[X_\tau \cdot \tau/T\Big]$

Structure of Controller

Using Markov strategies does not entail any loss of optimality

$$U_t = g_t(X_t)$$

Dynamic program

$$V_{T+1}(x_{T+1}) = 0;$$

$$V_t(x_t) = \max_{u_t \in \mathcal{U}_t} \Big\{ r_t(x_t, u_t) + \mathbb{E}[V_{t+1}(X_{t+1}) \mid X_t = x_t, U_t = u_t] \Big\}, \qquad t = T, \dots, 1.$$

Qualitative properties of the value function

Dynamic Program

$$V_{T+1}(x) = 0$$

$$V_t(x) = \max\left\{x \cdot \frac{t}{T}, \mathbb{E}[V_{t+1}(X_{t+1})]\right\}$$

Lemma Define

$$L_t = V_t(0) = \frac{t}{t+1}V_{t+1}(0) + \frac{1}{t+1}V_{t+1}(1).$$

Then:

$$V_t(1) = \max\left\{\frac{t}{T}, L_t\right\}$$

and therefore:

$$L_t - L_{t+1} = \left[\frac{1}{T} - \frac{L_{t+1}}{t+1}\right]^+ \quad \text{with} \quad L_T = 0.$$

Note that it is never optimal to select an alternative if it is not the best so far (i.e., $X_t = 0$). Thus, we can completely characterize an optimal strategy by solving for $\{L_t\}_{t=1}^{T}$ in a backward manner.

Structure of optimal strategy

Theorem (Critical time)

- There exists a critical time t_0, $t_0 < T$, such that it is optimal to reject all alternatives until $t_0 - 1$.
- The critical time is the smallest integer t such that

$$\sum_{k=t}^{T-1} \frac{1}{k} < 1$$

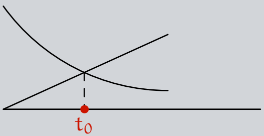
- The value functions are given by

$$L_t = \begin{cases} \frac{t}{T} \sum_{k=t}^{T-1} \frac{1}{k} & \text{for } t \geqslant t_0 \\ L_{t_0} & \text{for } t < t_0 \end{cases}$$

- The optimal strategy is reject the first $t_0 - 1$ alternatives and then select the first alternative superior to all predecessors, if one such occurs.
- For large T, $t_0 \approx T/e$ and the probability of selecting the best candidate is $\approx 1/e$.

Proof of structural properties

Proof



- $L_t - L_{t-1} \geqslant 0$, thus L_t is non-increasing with t.
- $V_t(1) = \max\{t/T, L_t\}$ where t/T is increasing with t and L_t is non-increasing with t. Thus, the critical time t_0 is the first time when $t/T \geqslant L_t$. Since $L_T = 0$ and $T/T = 1$, such a $t_0 < T$.
- For any t such that $t/T < L_t$,

$$L_{t-1} = L_t + \left[\frac{1}{T} - \frac{L_t}{t}\right]^+ = L_t.$$

- For any t such that $t/T \geqslant L_t$, we have that $(t+1)/T \geqslant L_{t+1}$. Thus,

$$L_t = L_{t+1} + \frac{1}{T} - \frac{L_{t+1}}{t+1} = \frac{t}{T}\left[\frac{1}{t} + \frac{T}{t+1}L_{t+1}\right]$$

- For large T,

$$\sum_{k=t}^{T-1} \frac{1}{k} \approx \int_{k=t}^{T} \frac{1}{k}\,dk = \log\left(\frac{T}{t}\right)$$

Thus, $t_0 = T/e$. Moreover,

$$V_1(0) = V_1(1) = L_1 = L_{t_0} \approx \frac{t_0}{T} = \frac{1}{e}.$$

Exercises and further reading on optimal choice

1. The mathematical model of optimal choice considered here is adapted from John P. Gilbert and Frederick Mosteller, "Recognizing the Maximum of a Sequence," Journal of the American Statistical Association Vol. 61, No. 313, pp. 35-73, Mar 1966. http://www.jstor.org/stable/2283044

2. For a history of the variations of this problem, see Thomas S. Ferguson, "Who Solved the Secretary Problem?," Statistical Science, vol. 4, no. 3, 282-289, 1989. http://projecteuclid.org/euclid.ss/1177012493

3. Let $\{W_t\}_{t=1}^{T}$ be continuous valued i.i.d. random variables with PDF f_W. Let X_t be the indicator function of the event that $\{W_t \geqslant \max\{W_{1:t-1}\}\}$. Then show that $\{X_t\}_{t=1}^{T}$ are independent and $\mathbb{P}(X_t = 1) = 1/t$.

MDP Example: Energy storage in renewable generation

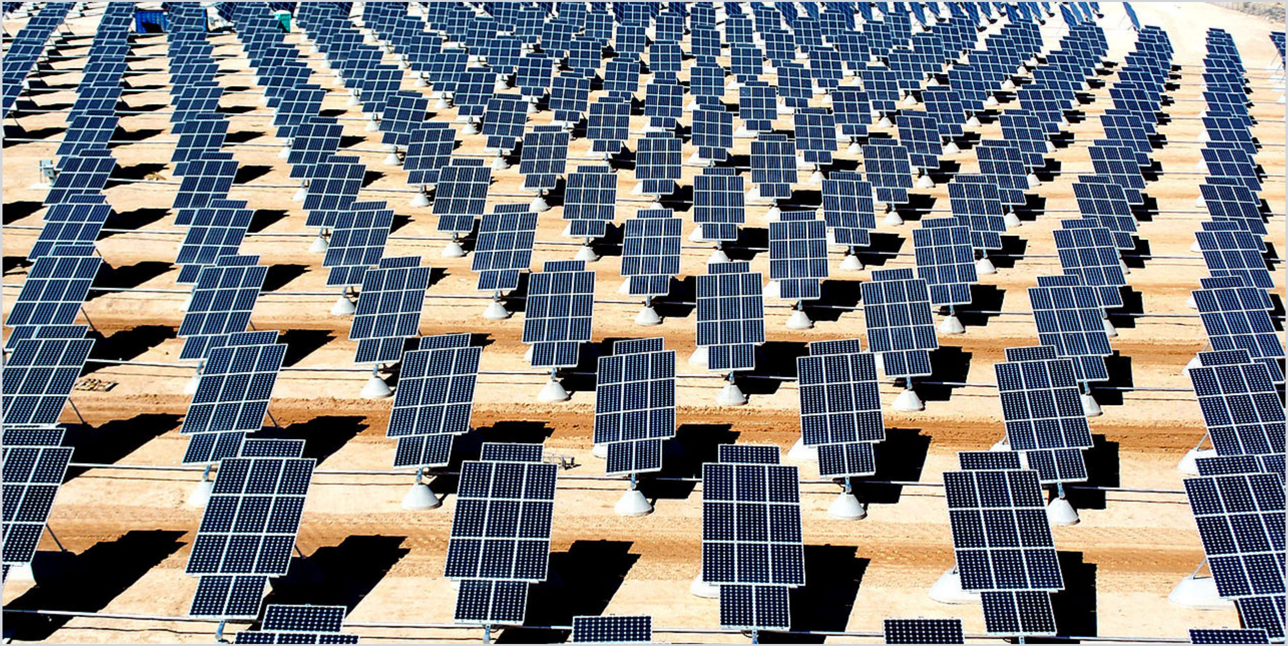


Image credit: http://en.wikipedia.org/wiki/File:Giant_photovoltaic_array.jpg

To be written

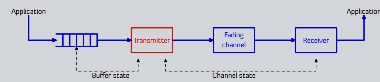
MDP Example: Optimal gambling

Image credit: http://commons.wikimedia.org/wiki/File:Gambling-ca-1800.jpg

MDP Example: Optimal inventory management

Image credit: http://commons.wikimedia.org/wiki/File:Modern_warehouse_with_pallet_rack_storage_system.jpg

MDP Example: Optimal power-delay trade-off in wireless communication



Optimal Stopping Example: Call options

Image credit: http://commons.wikimedia.org/wiki/File:Sao_Paulo_Stock_Exchange.jpg

Optimal Stopping Example: Optimal choice

Image credit: http://www.americanscientist.org/issues/feature/2009/2/knowing-when-to-stop/

MDP Example: Energy storage in renewable generation

Image credit: http://en.wikipedia.org/wiki/File:Giant_photovoltaic_array.jpg