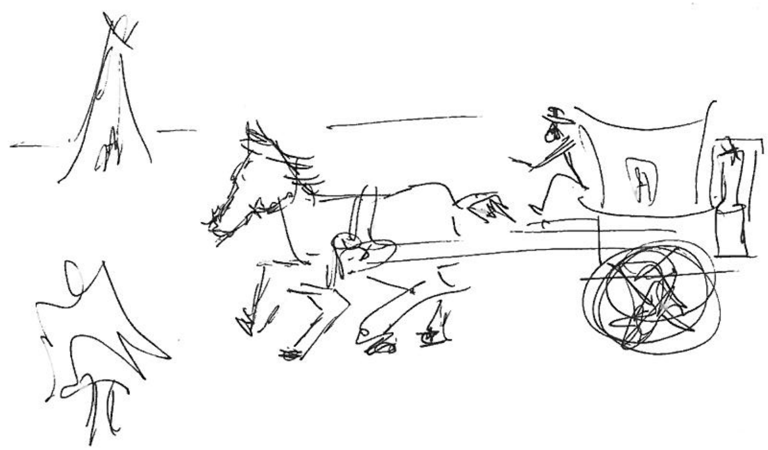
Motivation : shortest path



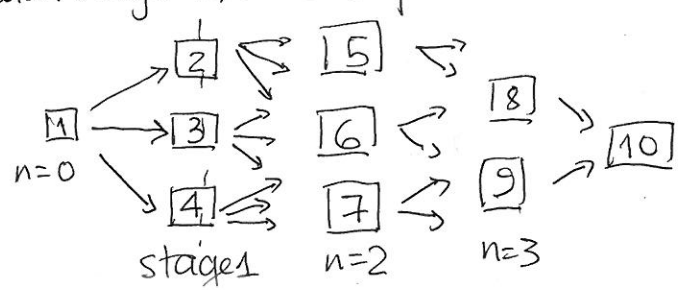
Scenario : first insurance companies "out west" had calculations about ~~the~~ the various risks in different segments of the highways, particularly concerning those that passed through indian territories.

The stage coach goes through different stages (4). [n]

At each stage the traveller may be in one of the possible states [x_n]. The coach driver must make a decision :

$u_n \in U(x_n)$

at each stage n, which depends on the state x_n.



Depending on the decision ($u_n \in U(x_n)$ represents the following state visited, so here

$x_{n+1} = u_n$)

there is a cost of travel and insurance, $r(x_n, u_n)$:

	2	3	4
1	2	4	3

	5	6	7
2	7	4	6
3	3	2	4
4	4	1	5

	8	9
5	1	4
6	6	3
7	3	3

	10
8	3
9	4

Can you figure out what is the optimal path (the one with smallest cost)?

Key idea :

If at stage n you are at state $x_n = i$, then you can evaluate the optimal decisions from i to state 10, regardless of the previous history $(x_k, u_k; k < n)$

At $n=4$ $x_n = 10$, and there is no decision.

At $n=3$:

$x_3 = 8 \Rightarrow$ cost $= 3$

$x_3 = 9 \Rightarrow$ cost $= 4$

$n=2$

x_2	$u_2=8$	$u_2=9$	u^*	$C_2^*(x_2)$
5	1+3	4+4	8	4
6	6+3	3+4	9	7
7	3+3	3+4	8	6

And, in general, let $J_n^*(x_n)$ be the optimal "cost-to-go" from x_n to the end, and u_n^* the corresponding optimal decision. We have:

$$(1) \quad J_n^*(x_n) = \min_{u_n \in U(x_n)} \left(r(x_n, u_n) + J_{n+1}^*(u_n) \right)$$

because $x_{n+1} = u_n$ in this case.

In more general models,

$$x_n = f(x_n, u_n)$$

may be a more complex function.

Example: Allocation models where u_n is how many resources (nurses, doctors, seats on an airplane) we allocate at stage n (n may represent a ward, for example, or the day) and x_n is the available resources. There are N stages, and

$$x_{n+1} = x_n - u_n$$

The cost function $f(x_n, u_n)$ will reflect the cost/gain of making that decision. An interesting application is harvesting: by stage N all products must be harvested and sold for a profit C_N/unit. But the product can be ~~harvested~~ at younger ages $1 \le n < N$ with a profit C_n/unit. Typically $C_n > C_{n'}$ if $n \le n'$, but demand is higher for older products and unsold young products go to waste. This model introduces uncertainty in the demand, so that the profit $f(x_n, u_n)$ is random. If the produce is also subject to "death", then we may have a model where "$x_{n+1} = x_n - u_n - v_n$" where v_n are random variables (may depend on x_n) representing the uncertain loss of crops due to weather and other conditions.

"So Who's Counting" Game (p. 14-15 MP)

~~General Formulation~~

The idea of equation (1) is to start at an "easy" terminal condition for which the choice is trivial (only one possibility, for example) and then work backwards. Bellman invented the methodology that he called "backwards programming" but the name was changed to "dynamic programming".

The general framework of the model is called Markov Decision Process, and one of the solution techniques (the most commonly used) is dynamic programming.

Examples (Chapter 3, MP):

Secretary Problem
Inventory
shortest path
critical path
sequential allocation
selling assets
queueing control
maintenance repair problems

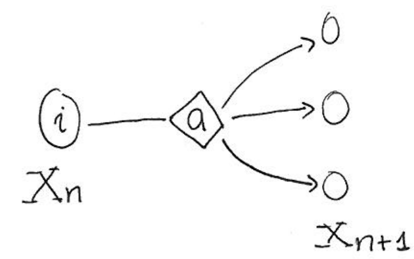
General Formulation

Let $\{X_n\}$ be a process on $(\Omega, \mathcal{F}, \mathbb{P})$. For each state $x \in S$, let $\mathcal{U}(x)$ be the set of possible actions u_n, when $X_n = x$. The simpler case considers the model where S is countable and $\bigcup_{x \in S} \mathcal{U}(x)$ is finite.

Let $\mathcal{F}_n = \sigma(X_0, u_0; X_1, u_1; \dots X_n, u_n)$ be the filtration of the process (X_n, u_n). We assume the Markov property:

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, u_n = a; \mathcal{F}_{n-1}) = P_{ij}(a) \qquad (2)$$

so that the evolution of the state is independent of the past, given the present state and action. At each stage n there is a reward function $R(X_n, u_n)$. This instantaneous reward is known and deterministic, and bounded.



Def: A decision rule is a rule that specifies the action u_n to be chosen at stage n. It is imposed that u_n be non-antipative.

Classification of decision rules:

HR: $u_n = \phi_n(X_0, u_0; \dots; X_n; \omega)$ is a randomized decision that depends on the history of the process.

MR: $u_n = \phi_n(X_n; \omega)$ is a randomized, Markovian decision.

HD: $u_n = \phi_n(X_0, u_0; \dots; X_n)$ is a deterministic function of the past trajectory (and present state).

MD: $u_n = \phi_n(X_n)$ is a deterministic function of the current state.

Given a decision rule, the corresponding policy β is a stochastic process defined by the consecutive values of the actions $(u_1, u_2, \dots)$.

Def: A policy β is called stationary if $\phi_n \equiv \phi$ is independent of the stage n.

The following theorem establishes that HR policies are "equivalent" to MR policies, so that it suffices to find optimal MR policies, rather than looking at history-dependent ones that require much more bookkeeping.

Theorem 1: Let $\{(X_n, U_n)\}_\beta$ be a MDP on a finite state space $\mathcal{S} \times \mathcal{U}$ with HR policy β. Then, there is a MR policy β' such that

$$\{(X_n, U_n)\}_{\beta'} \overset{d}{=} \{(X_n, U_n)\}_{\beta}, \text{ given } X_0 = i \in \mathcal{S}.$$

Proof: To prove the claim it suffices to show that $\forall i, j \in \mathcal{S}$ and $\forall a \in \mathcal{U}(j)$,

$$\mathbb{P}_\beta(X_n = j, U_n = a \mid X_0 = i) = \mathbb{P}_{\beta'}(X_n = j, U_n = a \mid X_0 = i).$$

Because $\{U_n\}_\beta$ follow HR policy, then

$$U_n = \phi_n(X_0, U_0; \dots ; X_n).$$

Notice that under β,

$$\mathbb{P}(U_n = a \mid X_n = j, X_0 = i) = \frac{\mathbb{E}(\mathbb{P}(U_n = a \mid X_n = j, \mathcal{F}_{n-1}, X_0 = i))}{\mathbb{P}(X_n = j \mid X_0 = i)}.$$

Define the MR policy by:

$$\mathbb{P}(U_n' = a \mid X_n = j, X_0 = i) = \mathbb{P}(U_n = a \mid X_n = j, X_0 = i) \quad (*),$$

for any $j \in \mathcal{S}$, $a \in \mathcal{U}(j)$. This defines a MR policy β'. Because of the MDP model, $\mathbb{P}(X_{n+1} = j \mid X_n = i, U_n = a) = P_{ij}(a)$ is independent of the policy, so that:

$$\mathbb{P}_\beta(X_n = j \mid X_0 = i) = \sum_{k \in S} \sum_{b \in \mathcal{U}(s)} P_{kj}(b) \underbrace{\mathbb{P}_\beta(X_{n-1} = k, U_{n-1} = b \mid X_0 = i)}$$

$$= \mathbb{P}_\beta(X_{n-1} = k \mid X_0 = i) \mathbb{P}_\beta(U_{n-1} = b \mid X_{n-1} = k, X_0 = i)$$

The proof follows by induction: $\mathbb{P}_\beta(X_1 = j \mid X_0 = i) = \mathbb{P}_{\beta'}(X_1 = j \mid X_0 = i)$. Assuming that $\mathbb{P}_\beta(X_{n-1} = k \mid X_0 = i) = \mathbb{P}_{\beta'}(X_{n-1} = k \mid X_0 = i)$, we have:

$$\mathbb{P}_\beta(X_n = j \mid X_0 = i) = \mathbb{P}_{\beta'}(X_n = j \mid X_0 = i),$$

where we have used $(*)$ and the induction hypothesis.

To finalize the claim, we notice that

$$\mathbb{P}_\beta(X_n = j, U_n = a \mid X_0 = i) = \mathbb{P}_\beta(X_n = j \mid X_0 = i)\, \mathbb{P}_\beta(U_n = a \mid X_n = j, X_0 = i)$$

$$= \mathbb{P}_{\beta'}(X_n = j \mid X_0 = i)\, \mathbb{P}_{\beta'}(U_n = a \mid X_n = j, X_0 = i)$$

$$= \mathbb{P}_{\beta'}(X_n = j, U_n = a \mid X_0 = i),$$

QED.

which proves the result.

Example pp 135-136 MP.

Excercise: Let β be a MR policy. Show that $\{(X_n, U_n)\}$ is a Markov chain. Under which condition is the chain homogeneous?

(*) INDUCED PROCESS (see p.6)

Once we have classified the type of policies, we can model decision problems where an underlying criterion is to be optimized by the strategy. MDP problems are studied according to the class of reward that we wish to maximize (or cost to minimize). [We can also add constraints.]

FINITE HORIZON PROBLEMS

Let $T(i)$ be a terminal reward, $T: S \to \mathbb{R}^+$. Consider:

$$\max_{\beta \in MR} \mathbb{E}_\beta\left(\sum_{n=0}^{N-1} R(X_n, U_n) + T(X_N)\right). \qquad (3)$$

Here, N is a deterministic integer called the horizon of the problem. Under a MR policy, we have:

$$\beta_n(i,a) = \mathbb{P}_\beta(U_n = a \mid X_n = i),$$

so the problem (3) corresponds to choosing the optimal values for $\{\beta_n\}_{n=0}^{N-1}$ such that $\sum_{a \in \mathcal{U}(i)} \beta_n(i,a) = 1 \ \forall i \in S$, $n \in \{0, \dots N-1\}$. The following is a fundamental result in MDP's and it establishes the optimality of deterministic policies.

Theorem 2: There is a DM policy that is optimal for problem (3). [notice that it may not be unique, and that there may be other optimal policies that are not deterministic]. (p. 90 MP)

Proof: Use backward programming + induction.

Recall that for a MD policy, the decision rules have the form:

$$U_n = \phi_n(X_n),$$

so that the probabilities $\beta_n(i,\cdot)$ are degenerate. For MD policies, using a first step analysis, it follows that if $J_n^*(i)$ is the optimal reward from stage $n+1$ until stage N, then:

$$\boxed{\begin{aligned} J_N^*(x) &= T_N(x) \quad \forall x \in S \\ J_n^*(i) &= \max_{a \in \mathcal{U}(i)} \left(R(i,a) + \sum_{j \in S} P_{ij}(a) J_{n+1}^*(j)\right) \end{aligned}}$$

These are called the "optimality" equations, or Bellman equations, and they lead to what we know today as dynamic programming: to solve these recursions, one starts at stage N and works backwards until stage $n=0$.

RANDOM HORIZON PROBLEMS

- Optimal stopping problems.
- Problems to maximize the probability of attaining one absorbing state ("gambling" model).
- Problems to maximize the time to reach an undesirable state (such as playing "tetris").

$$\max_{\beta \in MR} \mathbb{E}_\beta\left(\sum_{n=0}^{\tau} R(X_n, U_n)\right)$$

for a suitably defined stopping time τ and reward function R. [see MP for details]

Random horizon problems are related to absorbing Markov chains.

Stationary MR Policies for unichain models:

(1) The Linear Programming Approach

Consider $\pi \in SMR$. Under this policy, the enlarged joint process $\{(X_n, u_n)\}$ is a Markov chain (homogeneous), with:

$$\mathbb{P}(X_{n+1}=j, u_{n+1}=b \mid X_n=i, u_n=a) =$$

$$\mathbb{P}(X_{n+1}=j \mid X_n=i, u_n=a)\,\mathbb{P}(u_{n+1}=b \mid X_{n+1}=j)$$

$$= P_{ij}(a)\beta_j(b).$$

Because of the unichain assumption, for each vector of action probabilities β the above MC is irreducible. We assume wlog that it is ergodic and call $\pi_{i,a}^{(\beta)}$ the limit distribution or stationary probabilities when β is used for the decision rule:

$$\pi_{i,a}(\beta) = \lim_{n\to\infty} \mathbb{P}(X_n=i, u_n=a)$$

Then

(i) $\pi_{i,a}(\beta) \geq 0$

(ii) $\sum_{i,a} \pi_{i,a}(\beta) = 1$

(iii) $\sum_a \pi_{ja}(\beta) = \sum_i \sum_{a\in U(i)} \pi_{ia}(\beta) P_{ij}(a) \quad \forall j \in S$

(because $\pi_{ja}(\beta) = \sum_i \sum_{b\in U(i)} P_{ij}(b)\beta_j(a)\pi_{ib}(\beta)$ and $\sum_a \beta_j(a)=1$)

<u>Thm:</u> Suppose that $\{\pi_{ia}\}$ is a solution to (i),(ii),(iii). Then these are the stationary probabilities of the MDP $\{(X_n, u_n)_\beta\}$ for:

$$\beta_i(a) = \frac{\pi_{ia}}{\sum_{b\in U(i)} \pi_{ib}}.$$

The stationary average reward is therefore (by ergodicity):

$$\lim_{N\to\infty} \mathbb{E}_\beta\left(\frac{1}{N}\sum_{n=0}^{N-1} R(X_n, u_n)\right) = \sum_{i\in S}\sum_{a\in U(i)} \pi_{ia} R(i,a)$$

Therefore, the solution to the problem is the solution to the LP program:

$$\max_{\{\pi_{ia}\in\mathbb{R}^d\}} \sum_{i\in S}\sum_{a\in U(i)} \pi_{ia} R(i,a)$$

s.t. (i), (ii), (iii)

<u>Remark</u>: LP's can be solved efficiently, solution is in a vertex of the feasible set, which is defined by the constraints.

<u>Thm:</u> The solution to the LP problem corresponds to a deterministic policy $u_n = \phi(X_n)$ (SD)

(2) Policy-iteration method :

In LP theory, it is well known that the solution to the primal and dual problems is the same:

PRIMAL	DUAL
$\max \sum_{(i,a)} x_{ia} R(i,a)$	$\min \sum_j b_j y_j$
$\sum_{(i,a)} \mathbb{A}_{(i,a),j} x_{ia} = b_j$	$\sum_j \mathbb{A}_{(i,a)j} y_j \geq R(i,a)$
$x_{i,a} \geq 0$	$y_j \in \mathbb{R}$

Dual variables are associated with primal constraints, of which we have $|S|+1$. Call $y_1, \dots y_{|S|}$, and $g = y_{|S|+1}$, then we obtain the problem:

$$\min \sum_{j \in S} y_j + g$$

$$y_i + g - \sum_j P_{ij}(a) y_j \geq R(i,a)$$

$$\Rightarrow \quad y_i \geq R(i,a) - g + \sum_{j \in S} P_{ij}(a) y_j$$

and hence, because we wish to minimize, the problem can be restated as the optimality equation:

$$y_i = \max_{a \in U(i)} \left(R(i,a) - g + \sum_{j \in S} P_{ij}(a) y_j \right)$$

[Approximations, constraints, thresholds...]

[this equation can be derived directly with Potential theory, see MP pp. 337-343, and (8.4.3) p. 354].

Iteration method:

1. Choose initial ~~SMD~~ SMD $\phi_0(i)$, $i \in S$
2. Given $u_n = \phi_n(i)$, solve:

$$y_n(i) = R(i,a) - g_n + \sum_j P_{ij}[\phi_n(i)] y_n(j)$$

and set $y_n(1) = 0$. Next, with these values of y_n, g_n let

$$f_{n+1}(i) = \arg\max_{a \in U(i)} \left(R(i,a) - g_n + \sum_j P_{ij}(a) y_n(j) \right)$$

The above algorithm converges and it usually is very efficient.

(3) Value iteration method (p.p. 364-367 MP)

1. Choose J^0, $\epsilon > 0$, $n = 0$
2. $\forall s \in S$, compute

$$J^{n+1}(s) = \max_{a \in U(s)} \left(R(s,a) + \sum_{j \in S} P_{ij}(a) J^n(j) \right)$$

3. If $\| J^{n+1}(s) - J^n(s) \| \leq \epsilon \ \forall s \in S \Rightarrow$ STOP
4. $\forall i \in S$, choose

$$\phi(i) = \arg\max_{a \in U(i)} \left(R(s,a) + \sum_{j \in S} P_{ij}(a) J^n(j) \right).$$

Thm2 (p. 90 MP): If S is countable or finite and $|U(i)| < M$ for all i is finite, or (see conditions on p. 90).

Proof (idea):

For $n = N-1$ we seek

$$\max_{u_n \in \mathcal{U}(X_n)} E\left(r(X_n, u_n) \right) + E T(X_{n+1})$$

For each $i = X_n$ possible this is:

$$\max_{u_n \in \mathcal{U}(i)} \left\{ r(i, u_n) + \sum_{j \in S} T(j)\, p_{u_n}(i,j) \right\}$$

For each i, there is one optimal value of the above expression, possibly w. different actions u_n, choose any of these as the deterministic policy.

Now we know that $u^*_{N-1} = \phi(X_{N-1})$.

Use induction to show the result

(State the more general result from MP)

$$\frac{(.6)(.5)(.4) + (.1)(.8)}{(.6)(.5) + (0.1)}$$

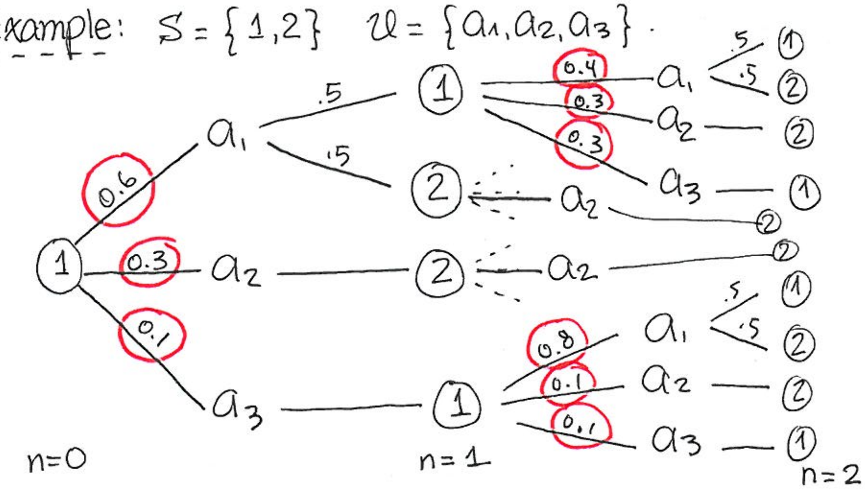
$$= \frac{(.3)(.4) + .08}{.3 + 0.1} = \frac{(.4).3 + 0.08}{.4} = \frac{0.12 + 0.08}{.4}$$

$$= \frac{0.20}{.4} = .5$$

Thm1: Let $\{(X_n, U_n)\}_\beta$ be a MDP on a finite state space $\mathcal{S} \times \mathcal{U}$ with a HR policy β. Then, there exists a MR policy β' such that $\{(X_n, U_n)\}_\beta \overset{d}{=} \{(X_n, U_n)\}_{\beta'}$. Specifically: for every n, i, j and a:

$$\mathbb{P}_\beta(X_n = j, U_n = a \mid X_0 = i) = \mathbb{P}_{\beta'}(X_n = j, U_n = a \mid X_0 = i).$$

Example: $\mathcal{S} = \{1, 2\}$ $\mathcal{U} = \{a_1, a_2, a_3\}$.



$$\mathbb{P}(X_{n+1} = 1 \mid X_n = 1, U_n = a_1) = 0.5$$

$$\mathbb{P}(X_{n+1} = 1 \mid X_n = 1, U_n = a_2) = 0$$

$$\mathbb{P}(X_{n+1} = 1 \mid X_n = 1, U_n = a_3) = 1$$

Shown in the diagram is the history-dependent policy β. Notice that $P_{ij}(a) = \mathbb{P}(X_{n+1} = j \mid U_n = a, X_n = i)$ is independent of β, determined by the probabilities circled in red. Shown in the case $X_0 = 1$, but a similar structure holds when $X_0 = 2$.

For the first stage, set:

$$\mathbb{P}(U_0 = a_1 \mid X_0 = 1) = 0.6$$

$$\mathbb{P}(U_0 = a_2 \mid X_0 = 1) = 0.3$$

$$\mathbb{P}(U_0 = a_3 \mid X_0 = 1) = 0.1$$

which is just the policy β. That is, if β is defined by:

$$U_n = \phi_n(X_0, U_0; \ldots; X_n)$$

then we use for β':

$$U_0' = \phi_0'(X_0) = \phi_0(X_0).$$

~~For the~~ For $n=1$:

$$\mathbb{P}(U_1 = a_1 \mid \overset{X_1=1}{X_0 = 1}) = \frac{(0.6)(0.5)(0.4) + (0.1)(0.8)}{(0.6)(0.5) + 0.1} = 0.5$$

so we can define

$$U_1' = \phi_1'(X_1)$$

$$\mathbb{P}(U_1 = a_1 \mid X_1 = 2, X_0 = 1) = 0$$

$$\mathbb{P}(U_1 = a_2 \mid X_1 = 2, X_0 = 1) = 1.$$

actually $\mathbb{P}(U_n = a_2 \mid X_{n-1} = 2, X_0 = i) = 1 \;\forall n > 1.$

$$\cancel{\mathbb{P}(U_1 = a_2 \mid X_1 = 1, X_0 = 1) = \frac{0.6(.5)(.3) + (.6)(.5) + (0.3)(.5) + .1(.1)}{}}$$

$$\mathbb{P}(U_1 = a_2 \mid X_1 = 1, X_0 = 1) = \frac{(.6)(.5)(.3) + (.1)(.1)}{(.6)(.5) + 0.1} = 0.25$$

$$\mathbb{P}(U_1 = a_3 \mid X_1 = 1, X_0 = 1) = \frac{(.6)(.5)(.3) + (.1)(.1)}{(.6)(.5) + 0.1} = 0.25$$