Taxi Problem

MDP Example

©Dennis Bricker, 2001
Dept of Industrial Engineering
The University of Iowa

A taxi serves three adjacent towns: A, B, and C.

Each time the taxi discharges a passenger, the driver must choose from three possible actions:

(1) "Cruise" the streets looking for a passenger.
(2) Go to the nearest taxi stand (hotel, train station, etc.)
(3) Wait for a radio call from the dispatcher with instructions (but not possible in town B because of distance and poor reception).

MDP model:
States: {A, B, C}

Action sets:
K_A = {1,2,3}, K_B = {1,2,3}, K_C = {1,2}

Transition probability matrices

Cruising streets	Waiting at taxi stand	Waiting for dispatch
$P^1 = \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$	$P^2 = \begin{bmatrix} 1/16 & 3/4 & 3/16 \\ 1/16 & 7/8 & 1/16 \\ 1/8 & 3/4 & 1/8 \end{bmatrix}$	$P^3 = \begin{bmatrix} 1/4 & 1/8 & 5/8 \\ 0 & 1 & 0 \\ 3/4 & 1/16 & 3/16 \end{bmatrix}$

Payoff matrices (expected profit per passenger):

R_{ij}^k = expected profit if action k is selected, and passenger wishes to travel from town i to town j

Cruising streets	Waiting at taxi stand	Dispatch call
$R^1 = \begin{bmatrix} 10 & 4 & 8 \\ 14 & 0 & 18 \\ 10 & 2 & 8 \end{bmatrix}$	$R^2 = \begin{bmatrix} 8 & 2 & 4 \\ 8 & 16 & 8 \\ 6 & 4 & 2 \end{bmatrix}$	$R^3 = \begin{bmatrix} 4 & 6 & 4 \\ 0 & 0 & 0 \\ 4 & 0 & 8 \end{bmatrix}$

Since our model assumes *minimization* of cost, we use

$$C_i^k = -\sum_j P_{ij}^k R_{ij}^k$$

Note: This example was introduced by Ron Howard in his textbook, *Dynamic Programming and Markov Processes*, MIT Press (1960), in which no consideration was given to the variable amount of time per stage (trip) in the optimization model.

States:

i	state
1	town A
2	town B
3	town C

Actions:

k	action
1	CRUISE
2	TAXISTAND
3	RADIO CALL

Cost Matrix

i	state	1	2	3
1	town A	¯8	¯2.75	¯4.25
2	town B	¯16	¯15	999
3	town C	¯7	¯4	¯4.5

(Rows ~ states, Columns ~ actions)

A value of 999 above signals an infeasible action in a state.

Note that the algorithm assumes minimization, and so the "cost" is the negative of the expected payoffs!

Transition Probabilities

Action: **CRUISE**

from \ to	1	2	3
1	0.5	0.25	0.25
2	0.5	0	0.5
3	0.25	0.25	0.5

Action: **TAXISTAND**

from \ to	1	2	3
1	0.0625	0.75	0.1875
2	0.0625	0.875	0.0625
3	0.125	0.75	0.125

Action: **RADIO CALL**

from \ to	1	2	3
1	0.25	0.125	0.625
2	0	1	0
3	0.75	0.0625	0.1875

Let's first use the criterion: **Maximize average reward per trip**

$$\text{Minimize } \sum_i \sum_a c_i^a x_i^a$$

$$\text{subject to } \sum_j x_j^a = \sum_i \sum_a p_{ij}^a x_i^a \quad \text{for all states } j$$

$$\sum_i \sum_a x_i^a = 1, \quad x_i^a \geq 0 \quad \text{for all states } i \text{ and actions } a \in A_i$$

LP Tableau for MDP

k:	1	2	3	1	2	1	2	3	
i:	1	1	1	2	2	3	3	3	RHS
Min	¯8	¯2.75	¯4.25	¯16	¯15	¯7	¯4	¯4.5	0
	0.5	0.9375	0.75	¯0.5	¯0.0625	¯0.25	¯0.125	¯0.75	0
	¯0.25	¯0.75	¯0.125	1	0.125	¯0.25	¯0.75	¯0.0625	0
	1	1	1	1	1	1	1	1	1

Note that one of the "steady-state" equations (for state C) was eliminated because of redundancy.

Phase One procedure was used to find an **initial basic feasible** solution

Iteration 0

Policy: (Cost= ¯8)

State		Action		P{i}	R{i}
1)	town A	3)	RADIO CALL	0.283186	4
2)	town B	1)	CRUISE	0.327434	6
3)	town C	2)	TAXISTAND	0.389381	8

k:	1	2	**3**	**1**	2	1	**2**	3	
i:	1	1	**1**	**2**	2	3	**3**	3	rhs
Min	¯3.5	¯3	0	0	**¯6**	0.5	0	6.125	8
	0.725664	1.0531	1	0	0.247788	¯0.0176991	0	¯0.473451	0.283186
	0.0265487	¯0.376106	0	1	**0.411504**	0.292035	0	0.561947	0.327434
	0.247788	0.323009	0	0	0.340708	0.725664	1	0.911504	0.389381

Initially the basic variables are $\left\{X_A^3, X_B^1, X_C^2\right\}$ (exactly one per state).

The values of these variables are the *steady-state probabilities* of the Markov chain corresponding to the policy (3, 1, 2).

The "most negative" reduced cost is **–6** (of variable X_B^2), and so that variable should enter the basis, replacing X_B^1. *(The pivot element is 0.411504, indicated above.)*

Iteration 1

Policy: (Cost= ¯12.7742)

State		Action		P{i}	R{i}
1)	town A	3)	RADIO CALL	0.0860215	¯9.16129
2)	town B	2)	TAXISTAND	0.795699	13.2258
3)	town C	2)	TAXISTAND	0.11828	12.7742

k:	1	2	3	1	2	1	2	3	
i:	1	1	1	2	2	3	3	3	rhs
Min	¯3.1129	**¯8.48387**	0	14.5806	0	4.75806	0	14.3185	12.7742
	0.709677	**1.27957**	1	¯0.602151	0	¯0.193548	0	¯0.811828	0.0860215
	0.0645161	¯0.913978	0	2.43011	1	0.709677	0	1.36559	0.795699
	0.225806	0.634409	0	¯0.827957	0	0.483871	1	0.446237	0.11828

The next pivot should enter X_A^2 into the basis, replacing X_A^3.

Iteration 2

Policy: (Cost= ¯13.3445)

State		Action		P{i}	R{i}
1)	town A	2)	TAXISTAND	0.0672269	¯1.17647
2)	town B	2)	TAXISTAND	0.857143	12.6555
3)	town C	2)	TAXISTAND	0.0756303	13.3445

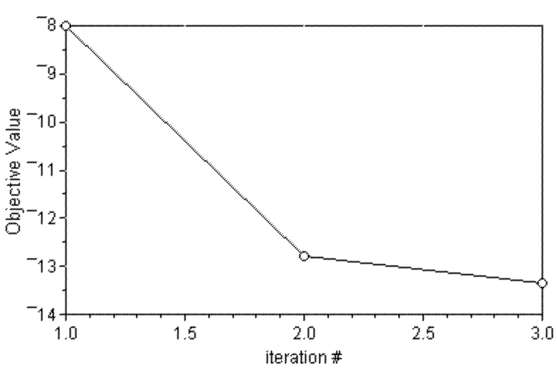
k:	1	2	3	1	2	1	2	3	
i:	1	1	1	2	2	3	3	3	rhs
Min	1.59244	0	6.63025	10.5882	0	3.47479	0	8.93592	13.3445
	0.554622	1	0.781513	¯0.470588	0	¯0.151261	0	¯0.634454	0.0672269
	0.571429	0	0.714286	2	1	0.571429	0	0.785714	0.857143
	¯0.12605	0	¯0.495798	¯0.529412	0	0.579832	1	0.848739	0.0756303

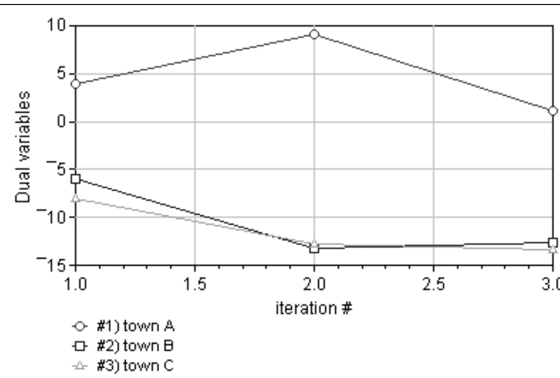
All reduced costs are nonnegative!

Optimal Policy

State	Action	P{i}	R{i}
1) town A	2) TAXISTAND	0.0672269	¯1.17647
2) town B	2) TAXISTAND	0.857143	12.6555
3) town C	2) TAXISTAND	0.0756303	13.3445

Average cost/stage = ¯13.3445





Value Iteration Method

(Note: objective: maximize average reward per passenger)

We want to compute $\lim_{n\to\infty} \frac{f_n(i)}{n}$

where $f_n(i) = \min_{a\in A_i}\left\{C_i^a + \sum_j p_{ij}^a f_{n-1}(j)\right\}$

Since $\lim_{n\to\infty} \frac{f_n(i)}{n}$ should be independent of the state i, our convergence criterion is to compute

$$\Delta f(i) = \left|\frac{f_n(i)}{n} - \frac{f_{n-1}(i)}{n-1}\right|$$

and terminate when $\max_i \left\{\Delta f_n(i)\right\} - \min_i \left\{\Delta f_n(i)\right\} \le \varepsilon$

Tolerance: 1.00E¯6

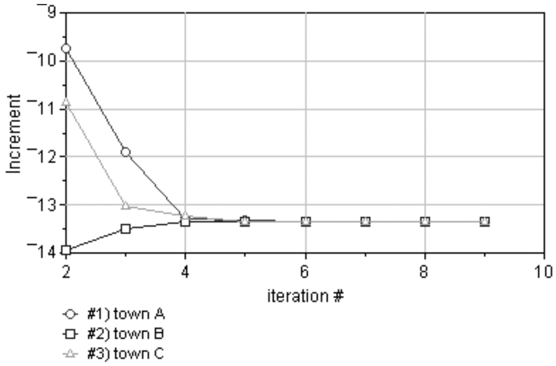
Minimizing average cost/period

iteration	Max ΔV	Min ΔV	gap (%)
1	¯9.75000E0	¯1.39375E1	3.00448E1
2	¯1.19141E1	¯1.34844E1	1.16454E1
3	¯1.32314E1	¯1.33579E1	9.46741E¯1
4	¯1.33307E1	¯1.33465E1	1.18444E¯1
5	¯1.33431E1	¯1.33448E1	1.27635E¯2
6	¯1.33444E1	¯1.33446E1	1.59546E¯3
7	¯1.33445E1	¯1.33445E1	1.91449E¯4
8	¯1.33445E1	¯1.33445E1	2.39312E¯5

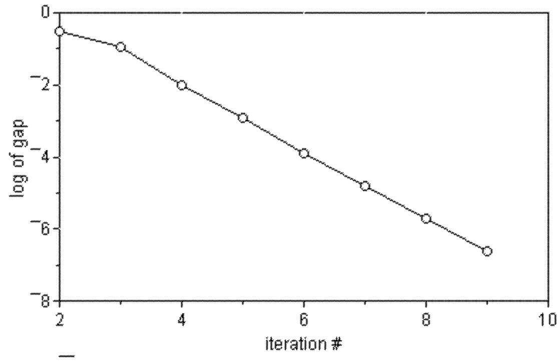
***Converged! with gap = 0.0000239312%

Solution:

state	action	Value
1	2	¯13.3445
2	2	¯13.3445
3	2	¯13.3445



Plot of $f_n(i) - f_{n-1}(i)$



Plot of $\log_{10}\left[\max_i \{\Delta f_n(i)\} - \min_i \{\Delta f_n(i)\}\right]$

Next we will solve the problem with the objective of maximizing the total discounted payoff.

Criterion: Discounted Total Cost, with $\beta = (1.2)^{-1} = 0.833333$

k:	1	2	3	1	2	1	2	3	
i:	1	1	1	2	2	3	3	3	RHS
Min	¯8	¯2.75	¯4.25	¯16	¯15	¯7	¯4	¯4.5	0
	0.583	0.948	0.792	¯0.417	¯0.052	¯0.208	¯0.104	¯0.625	1
	¯0.208	¯0.625	¯0.104	1	0.271	¯0.208	¯0.625	¯0.052	0
	¯0.208	¯0.156	¯0.521	¯0.417	¯0.052	0.583	0.895	0.843	0

Note: Specifying initial conditions to be deterministic, with town A as initial state

$$\text{Minimize} \sum_i \sum_{a \in A_i} C_i^a X_i^a$$

$$\text{subject to} \quad \sum_{a \in A_j} X_j^a = \beta \sum_i \sum_{a \in A_i} P_{ij}^a X_i^a \qquad \forall j$$

$$X_i^a \geq 0 \quad \forall a \in A_i, \forall i$$

Note that X_i^a is **not a probability** in this model, and so the equation

$$\sum_i \sum_a X_i^a = 1$$

is **not** included in the LP tableau.

There is **one equation for each state** (not including the objective row), with no redundancy as in the average cost/stage LP model, so the total number of variables, as before, is equal to the number of states, and as before, in a basic feasible solution there is **one basic variable per state**.

*Phase One procedure was used to find **initial basic feasible** solution*

Iteration 0

Policy: (Cost= ¯50.12)

State		Action		X{i}	V{i}
1)	town A	1)	CRUISE	3.783	¯50.12
2)	town B	1)	CRUISE	0.8589	¯56.01
3)	town C	3)	RADIO CALL	1.358	¯45.91

k:	1	2	3	1	2	1	2	3	
i:	1	1	1	2	2	3	3	3	rhs
Min	0	2.574	5.677	0	¯4.831	¯2.327	¯3.097	0	50.12
	1	1.361	1.151	0	0.4107	0.3612	0.5118	0	3.783
	0	¯0.3425	0.1215	1	0.3679	¯0.09487	¯0.4685	0	0.8589
	0	¯0.01831	¯0.273	0	0.2214	0.7337	0.9567	1	1.358

i~state, k~action

Iteration 1

Policy: (Cost= ¯61.4)

State		Action		X{i}	V{i}
1)	town A	1)	CRUISE	2.824	¯61.4
2)	town B	2)	TAXISTAND	2.334	¯77.89
3)	town C	3)	RADIO CALL	0.8414	¯55.62

k:	1	2	3	1	2	1	2	3	
i:	1	1	1	2	2	3	3	3	rhs
Min	0	¯1.923	7.273	13.13	0	¯3.573	¯9.249	0	61.4
	1	1.743	1.016	¯1.116	0	0.4671	1.035	0	2.824
	0	¯0.9308	0.3302	2.718	1	¯0.2579	¯1.273	0	2.334
	0	0.1877	¯0.3461	¯0.6017	0	0.7908	1.239	1	0.8414

i~state, k~action

Iteration 2

Policy: (Cost= ¯67.68)

State		Action		X{i}	V{i}
1)	town A	1)	CRUISE	2.121	¯67.68
2)	town B	2)	TAXISTAND	3.199	¯81.74
3)	town C	2)	TAXISTAND	0.6793	¯69.36

k:	1	2	3	1	2	1	2	3	
i:	1	1	1	2	2	3	3	3	rhs
Min	0	¯0.5206	4.689	8.638	0	2.332	0	7.467	67.68
	1	1.586	1.305	¯0.6136	0	¯0.1936	0	¯0.8355	2.121
	0	¯0.7378	¯0.02557	2.099	1	0.5551	0	1.028	3.199
	0	0.1516	¯0.2794	¯0.4858	0	0.6384	1	0.8073	0.6793

i~state, k~action

Iteration 3

Policy: (Cost= ¯68.37)

State		Action		X{i}	V{i}
1)	town A	2)	TAXISTAND	1.337	¯68.37
2)	town B	2)	TAXISTAND	4.186	¯81.91
3)	town C	2)	TAXISTAND	0.4766	¯69.56

k:	1	2	3	1	2	1	2	3	
i:	1	1	1	2	2	3	3	3	rhs
Min	0.3282	0	5.117	8.437	0	2.269	0	7.193	68.37
	0.6304	1	0.8227	¯0.3868	0	¯0.122	0	¯0.5267	1.337
	0.4651	0	0.5814	1.814	1	0.4651	0	0.6395	4.186
	¯0.09556	0	¯0.4041	¯0.4271	0	0.6569	1	0.8872	0.4766

i~state, k~action

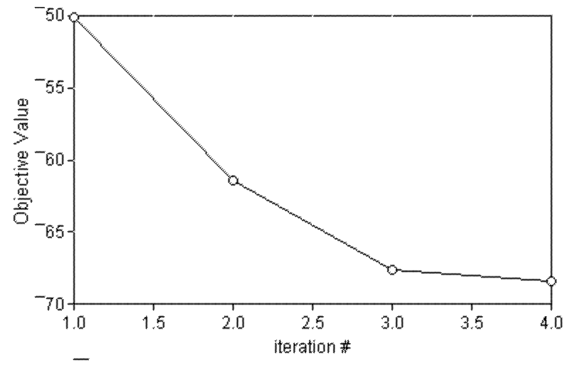
Reduced costs are now nonnegative!

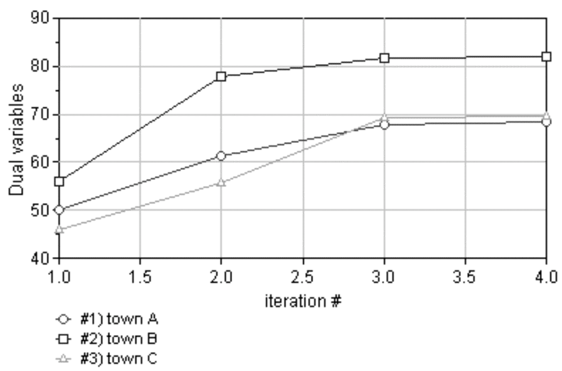
Optimal Policy

	State	Action	X{i}	V{i}	alpha{i}
1)	town A	2) TAXISTAND	1.337	¯68.37	1
2)	town B	2) TAXISTAND	4.186	¯81.91	0
3)	town C	2) TAXISTAND	0.4766	¯69.56	0

Alpha is initial distribution of the state

Discounted future costs = ¯68.37





Value Iteration Method

Tolerance: 1.00E¯6

Minimizing discounted future costs

iteration	Max ΔV	Min ΔV	gap (%)
1	¯8.12500E0	¯1.14479E1	2.90264E1
2	¯7.55425E0	¯9.21658E0	1.80363E1
3	¯7.04056E0	¯7.57706E0	7.08063E0
4	¯6.24756E0	¯6.28089E0	5.30648E¯1
5	¯5.22831E0	¯5.23178E0	6.63601E¯2
6	¯4.35921E0	¯4.35951E0	6.89203E¯3
7	¯3.63287E0	¯3.63290E0	8.61510E¯4
8	¯3.02741E0	¯3.02741E0	1.02206E¯4
9	¯2.52284E0	¯2.52284E0	1.27758E¯5

***Converged! with gap = 0.00001278%

Solution:

state	action	Value
1	2	¯55.76
2	2	¯69.30
3	2	¯56.95

Solving again....

Tolerance: 1.00E¯12 ***Reduced the tolerance!***

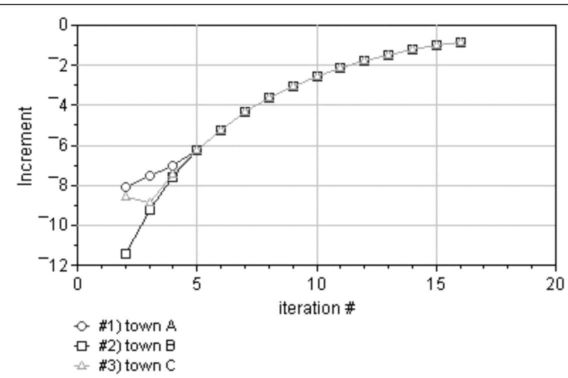
iteration	Max ΔV	Min ΔV	gap (%)
1	¯8.12500E0	¯1.14479E1	2.90264E1
2	¯7.55425E0	¯9.21658E0	1.80363E1
3	¯7.04056E0	¯7.57706E0	7.08063E0
4	¯6.24756E0	¯6.28089E0	5.30648E¯1
5	¯5.22831E0	¯5.23178E0	6.63601E¯2
6	¯4.35921E0	¯4.35951E0	6.89203E¯3
7	¯3.63287E0	¯3.63290E0	8.61510E¯4
8	¯3.02741E0	¯3.02741E0	1.02206E¯4
9	¯2.52284E0	¯2.52284E0	1.27758E¯5
10	¯2.10237E0	¯2.10237E0	1.57556E¯6
11	¯1.75198E0	¯1.75198E0	1.96944E¯7
12	¯1.45998E0	¯1.45998E0	2.45345E¯8
13	¯1.21665E0	¯1.21665E0	3.06725E¯9
14	¯1.01387E0	¯1.01387E0	3.82647E¯10
15	¯8.44896E¯1	¯8.44896E¯1	4.79360E¯11

***Converged! with gap = 4.794E¯11%

Solution:

state	action	Value
1	2	¯64.15
2	2	¯77.69
3	2	¯65.34

Note: Policy is same as the earlier run with larger tolerance, but objective value is nearer to true value.



Because the duration (in minutes) of a stage (trip) will depend upon the policy which we select,
the objective of maximizing the reward per trip is inappropriate--
we should instead maximize the **reward per unit time**.
This requires that we treat this as a

Semi-Markov Decision Process (SMDP).

Suppose we have the additional data:
Expected time to obtain passenger:

W_i^k = expected waiting time (minutes) in town i when action k is selected

Town \ Action	Cruising	Taxi stand	Dispatch call
A	15	20	20
B	10	25	∞
C	20	25	20

Expected travel time between towns:

T_{ij}= expected travel time (minutes) from town i to town j

Town \ Town	A	B	C
A	10	20	30
B	20	10	20
C	30	20	10

Expected travel time (minutes) of trip = $\sum_j P_{ij}^k T_{ij} =$

	Cruise	Taxi-stand	Radio call
town A	17.5	21.25	23.75
town B	20	11.25	10
town C	17.5	20	25.625

$v_i^a \triangleq E\left[t_i^a\right]$ = expected total duration of a trip (waiting + traveling) when in town i if action a is selected, i.e., $E\left[t_i^k\right] = W_i^k + \sum_j P_{ij}^k T_{ij} =$

	Cruise	Taxi-stand	Radio call
town A	32.5	41.25	43.75
town B	30	36.25	0
town C	37.5	45	45.625

Average reward per minute for the optimal policy (2,2,2) found by the MDP:

$$\frac{\sum_i \sum_a c_i^a x_i^a}{\sum_i \sum_a v_i^a x_i^a} = \frac{\text{average reward/trip}}{\text{average duration of trip}} = \frac{\$13.3445}{37.2479 \text{ min.}} = \$0.358262/\text{min.}$$

If we treat this as a Semi-Markov Decision Process (**SMDP**), then we can find the policy which maximizes our reward per minute by solving the LP:

$$\text{Minimize} \quad \sum_i \sum_a c_i^a u_i^a$$

$$\text{subject to} \quad \sum_j u_j^a = \sum_i \sum_a p_{ij}^a u_i^a \quad \text{for all states } j$$

$$\sum_i \sum_a v_i^a u_i^a = 1$$

$$u_i^a \geq 0 \quad \text{for all states } i \text{ and actions } a \in A_i$$


```
k:  |   1      2        3      1       2       1       2        3   |
i:  |   1      1        1      2       2       3       3        3   | RHS
Min | ¯8    ¯2.75    ¯4.25  ¯16   ¯15        ¯7     ¯4       ¯4.5   |  0
    |  0.5    0.9375  0.75  ¯0.5   ¯0.0625 ¯0.25 ¯0.125 ¯0.75   |  0
    | ¯0.25 ¯0.75    ¯0.125  1       0.125  ¯0.25 ¯0.75   ¯0.0625|  0
    | 32.5  41.25    43.75  30      36.25   37.5  45       45.625 |  1
```

*Note that this tableau differs from that of the LP for MDP only in the **last row**!*

Optimal Policy:

```
     State          |  Action              |  U{i}
       1)    town A |   1)     CRUISE      |   0.00331263
       2)    town B |   2)     TAXISTAND   |   0.0215321
       3)    town C |   2)     TAXISTAND   |   0.00248447

Average cost/unit time = ¯ 0.35942
```

The optimal policy (1,2,2) of the SMDP is different in town A, and the average reward per minute is slightly larger ($0.35942) than that ($0.358262) of the earlier policy (2,2,2).

In reality, of course, the infinite horizon is a much more problematic assumption in this particular problem!)