MATH4406 (Control Theory), HW2 (Unit 2)
MDP Modelling, Simulation and Policy Evaluation.

Prepared by Yoni Nazarathy, Last Updated: August 22, 2014

This homework project is mostly about getting a feel for MDP models. The emphasis is on modelling and model formulation. You will also get some practice in evaluating policies using simulation and policy evaluation. In this homework you are still not "optimising" in any structured way. You are rather getting a feel for the behaviours of different policies.

- There are 5 problems here. You can do either 4 of them (of your choice), or do all 5, in which case your chances of getting a good grade are probably higher.
- A good part of the assignment relies on Chapter 3 of [Put94]. You can get the chapter on-line from the library.
- For some of you the programming aspect may be a bit of a challenge. Make sure to allow enough time for this. Break up each programming task into well-defined sub-tasks.
- Please make sure to present your results in a clear and organised manner. Numerical output results should always be well explained and documented. Labels on graphs, diagrams, tables etc...
- Hand-in all code (preferably as an appendix).

Problem 1: Inventory Control MDP
Read Section 3.2 of [Put94]. Answer Problem 3.2 on page 68.

Problem 2: Queueing Control MDP
Read Section 3.7 of [Put94]. Answer *one* of 3.21, 3.22 or 3.23 from page 72 (only one).

Problem 3: More MDP Examples
Carry out *one* of the following (pp. 70–73): 3.15, 3.16, 3.17, 3.18, 3.19 or 3.26.

Extra, not mandatory: If the associated problem really interests you, look at the associated paper (e.g. Love, 1985 for problem 3.15). Attach the paper with your assignment together with any comments you have.

Problem 4: The Promotion MDP
Consider the "promotion MDP" as presented in class and appearing in PromotionMDP-StateSpace.pdf. You start at $(0, 0, 0)$ and wish to maximise,

$$v = \sum_{t=1}^{10} L_t,$$

where L_t is the level at time t. The transition probabilities of applying are $1 - q^\tau$ for "success" and q^τ for failure. Here τ is the number of years in the level and $q = 3/4$.

2a: Formalise this problem as an MDP. I.e. what is the time set, the state space, the action sets, the transition probabilities and the rewards. Be formal and precise in your specification.

2b: Assume you are considering *stationary Markov* policies. How many policies are there? Try to neatly specify the policy (decision rule) set.

2c: Consider the policy "apply as soon as you can" (denote it π_1) and the policy "apply as soon as you can, only for $\tau \geq 3$" (denote it π_3). The latter policy implies that when-ever $\tau \geq 3$ you apply if allowed to. Run (Monte-Carlo) simulations to compare the performance of these two policies. Do this by for e.g. running 1000 runs of each policy, and compare the resulting v (you can also give confidence bounds, but this isn't mandatory). Which policy is better?

2d: Now implement and use the policy evaluation algorithm (backward recursion) to compare π_1 and π_3. The values of v which you get should agree with the previous item.

2e: Now use the policy evaluation algorithm on all[1] stationary Markov policies (many of these - this may take a while to run!). What is the optimal policy?

2f (not mandatory): Maybe the optimal policy you found above is a "threshold policy", meaning that for any level, ℓ, when $\tau < \tau_\ell$ you don't apply and when $\tau \geq \tau_\ell$ you do apply. Is this the case? If so, would you like to conjecture that for any q there is a threshold policy? Try this maybe for a few other q values? Why is a threshold policy appealing?

[1] This is a very inefficient way to find the best policy. But use it now. In the sequel, we will work with dynamic programming algorithms, but not yet.

Problem 5: Restless bandits
This problem deals with a type of MDP known as a *restless multi-armed bandit.* We don't describe the problem in generality here; we rather consider a useful common real life example: *Breastfeeding triplets.*

Consider a family that has just expanded from 2 (wife and husband) to 5 (wife, husband and triplets). Call the new babies, bandit 1, bandit 2 and bandit 3. The husband (who now became father), is rushing around all day, trying to make decisions on how and what to do. The wife (new mother), still just after birth, is sitting stationary, and most of the time breast-feeding... But as you might have figured out, she can't handle all three bandits at once. It is then up to the father to schedule the babies, moving them from their cot, to their mother's comforting nutrition, and back.

With an aim of being synchronised and efficient, every 30 minutes the father makes a transition, exchanging the bandits between the mother and their cots. So his decision at every such time-slot is: "where to place the bandits", with the obvious constraint of no more than 2 with mummy.

When a bandit is in his cot, refer to the bandit as *passive.* When the bandit is with mummy, refer to him as *active.* At this early point in their life, the bandits don't appear to be doing much. It is then sensible to model the *state* of each bandit by,

$$1 \equiv \text{deep sleep}, \quad 2 \equiv \text{awake and happy}, \quad 3 \equiv \text{screaming the house down.}$$

As the new parents recently discovered, the bandits are not deterministic objects, but nevertheless, they seem to take well to breastfeeding and thus when active, a bandit changes state according to the following transition probability matrix:

$$P_a = \begin{bmatrix} 0.6 & 0.1 & 0.3 \\ 0.5 & 0.2 & 0.3 \\ 0.3 & 0.6 & 0.1 \end{bmatrix}.$$

On the other hand, a *passive* bandit changes state as follows:

$$P_p = \begin{bmatrix} 0.6 & 0.1 & 0.3 \\ 0.1 & 0.3 & 0.6 \\ 0.2 & 0.1 & 0.7 \end{bmatrix}.$$

In view of this data, the brave scheduling father has two main goals: (1) Minimise screaming. (2) Let the mother occasionally get some breastfeeding relief. For this he formulated the following cost structure per time interval:

$$\text{cost} = \text{number of screaming bandits} \quad + \quad \text{number of feeding bandits.}$$

For you: Formulate the problem as an MDP (state-space, action sets, transitions, rewards, etc...). Based on the problem data, think of some stationary Markov policy which you think is sensible. Specify your decision rule precisely. Run a long time horizon simulation (e.g. $10,000$ time steps) to evaluate the average cost per step of your policy.