CS533
Decision Making and Intelligent Agents
Final Takehome Exam, Winter 2009
Due: Friday, March 18, 2011 by 5:00pm

Name:

1. [20pt] A standard MDP is described by a set of states S, a set of actions A, a transition function T, and a reward function R. Where $T(s, a, s')$ gives the probability of transitioning to s' after taking action a in state s, and $R(s)$ gives the immediate reward of being in state s.

   A k-order MDP is described in the same way with one exception. The transition function T depends on the current state s and also the previous $k-1$ states. That is, $T(s_{k-1}, \ldots, s_1, s, a, s') = \Pr(s'|a, s, s_1, \ldots, s_{k-1})$ gives the probability of transitioning to state s' given that action a was taken in state s and the previous $k-1$ states were $(s_{k-1}, \ldots, s_1)$.

   Given a k-order MDP $M = (S, A, T, R)$ describe how to construct a standard (first-order) MDP $M' = (S', A', T', R')$ that is equivalent to M. Here equivalent means that a solution to M' can be easily converted into a solution to M. Be sure to describe S', A', T', and R'. Give a brief justification for your construction.

2. [20pt] Consider a stochastic policy defined by the Boltzman distribution

$$\pi_\theta(s,a) = \frac{\exp\left(Q_\theta(s,a)\right)}{\sum_{a'} \exp\left(Q_\theta(s,a')\right)}$$

where $\theta = [\theta_1, \ldots, \theta_n]$ is the vector of policy parameters and

$$Q_\theta(s,a) = \sum_{i=1}^{n} \theta_i \cdot f_i(s,a)$$

Recall that one of our policy gradient methods described in class requires us to compute an analytical form for the gradient of $\log\left(\pi_\theta(s,a)\right)$ with respect to θ, which requires analytical forms for the partial derivatives of $\log\left(\pi_\theta(s,a)\right)$ with respect to each θ_i. For the Boltzman policy described above derive the follow expression for the partial derivative.

$$\frac{\partial \log\left(\pi_\theta(s,a)\right)}{\partial \theta_i} = f_i(s,a) - \sum_{a'} \pi_\theta(s,a') f_i(s,a')$$

3. [20pt] In the Monte-Carlo planning lecture we stated the following error bound between the infinite horizon Q-function of a policy $Q_\pi(s, a)$ and the finite horizon approximation $Q_\pi(s, a, h)$:

$$|Q_\pi(s, a) - Q_\pi(s, a, h)| \leq \frac{\beta^h}{1 - \beta} R_{\max}$$

where $R_{\max}$ is the maximum absolute value of any reward in the MDP.

Give a proof of this bound.

4. (a) [10pt] Consider a version of GraphPlan that is identical to the one described in the book, except that it does not utilize maintenance (or persistence) actions during forward graph construction. Is this new version of GraphPlan sound and complete? If so, then what is the purpose of maintenance actions? If not, then give an example planning problem where the new GraphPlan will fail.

   (b) [10pt] Consider a version of GraphPlan that does not compute mutex relations during forward graph construction. (Otherwise it is identical to the algorithm in the book.) Is this new version of GraphPlan sound and complete? If so, then what is the purpose of computing mutex relations? If not, then give an example planning problem where the new GraphPlan will fail.

5. [20pt] In class we introduced policy rollout. Let $\hat{\pi}$ be the rollout policy for the base policy π. That is, $\hat{\pi}(s) = \arg\max_{a \in A} \hat{Q}(s,a)$, where $\hat{Q}(s,a)$ is the estimate of $Q_\pi(s,a)$ returned by the procedure QEstimate(s,a,π,h,w) which simply averages the results of w runs of **SimQ**(s,a,π,h) for a given sampling width w and horizon h. Recall that $\hat{\pi}$ is a stochastic policy.

   We say that $\hat{\pi}(s)$ is correct if it returns an action a that maximizes the true Q-function $Q_\pi(s,a)$. Note that if $\hat{\pi}$ always returns correct actions, then it corresponds to the policy iteration improvement of π and is guaranteed to be an improvement on π when π is not optimal.

   For any state s, consider the set of possible Q-values $\{Q_\pi(s,a) | a \in A\}$ and let $Q_1(s)$ and $Q_2(s)$ be the best and second best Q-values in the set (if there is only one value then $\Delta(s) = 0$). We define the Q-advantage of state s to be $\Delta(s) = Q_1(s) - Q_2(s)$. That is the Q-advantage gives the degree to which the best Q-value dominates the others.

   Your Problem. Consider a state s with positive $\Delta(s)$. Give conditions on the horizon h and the sampling width w (e.g. lower bounds) such that there is at least a $1-\delta$ probability that $\hat{\pi}(s)$ is correct. Your conditions will likely depend on $\Delta(s)$, the maximum reward of the MDP, the discount factor, and δ. The Chernoff bound will likely be useful here. Use the following steps:

   (a) First derive conditions on w and h that will ensure that with probability at least $1-\delta'$ we have,

   $$|Q_\pi(s,a) - \text{QEstimate}(s,a,\pi,h,w)| \leq \epsilon$$

   naturally h and w will depend on δ' and ϵ. This can be done by deriving the bound that was shown in class and then selecting appropriate bounds for h and w.

   (b) To finish the proof you need to select appropriate values for δ' and ϵ and then use the above result to select the appropriate values for w and h. The value of δ' will depend on the desired δ (remember that policy rollout requires calling QEstimate $|A|$ times, thus there are $|A|$ independent chances for the bound from part a to be violated). The value for ϵ will depend on $\Delta(s)$.