<u>Planning and Learning in Dynamical Systems (046194)</u>
<u>Homework 4, due 17/6/2020</u>

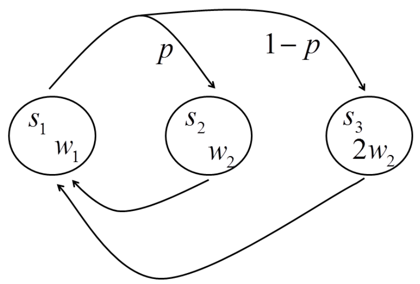
<u>Question 1</u> – Projected Bellman Operator

Consider the projection operator Π that projects a vector $V \in \mathbb{R}^n$ to a linear subspace S that is the span of k features $\phi_1(s),\ldots,\phi_k(s)$ w.r.t. the ϵ-weighted Euclidean norm. Also recall the Bellman operator for a fixed policy $T^\pi(v) \doteq r + \gamma P^\pi v$.

a. Show that for a vector $v \in S$, the vector $v' = T^\pi v$ is not necessarily in S. Choose an appropriate MDP and features to show this.

In class we have seen that ΠT^π is a contraction w.r.t. the ϵ-weighted Euclidean norm, when ϵ is the **stationary distribution** of P^π. We will now show that when ϵ is chosen differently, ΠT^π is not necessarily a contraction.

Consider the following 3-state MDP with zero rewards:



We consider a value function approximation $\tilde{V}(s) = \phi_1(s)w_1 + \phi_2(s)w_2$, given explicitly as $\tilde{V} = (w_1, w_2, 2w_2)^\top$, and we let $w = (w_1, w_2)^\top$ denote the weight vector.

b. What are the features $\phi(s)$ in this representation?

c. Write down the Bellman operator T^π explicitly. Write down $T^\pi \tilde{V}$.

c. What is the stationary distribution?

d. Write down the projection operator Π explicitly, for $\epsilon = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$.

e. Write an explicit expression for $\tilde{V}' \doteq \Pi T^\pi \tilde{V}$ in terms of w: the weight vector of $\tilde{V}$. Let $w' = (w_1', w_2')^\top$ denote the weights of $\tilde{V}'$. Write w' as a function of w .

f. Show that iteratively applying ΠT^π to $\tilde{V}$ may diverge for certain values of p.

<u>Question 2</u> – Approximate Value Iteration

(Read the lecture notes on approximate value iteration, Chapter 11.3.5)

In this question you will prove a bound on an approximate Value Iteration scheme. Recall that Value Iteration iteratively applies the optimal Bellman operator, T, on the current estimate of the value function. The value at the kth iteration is thus,

- $v_{k+1} = Tv_k$.

Consider an approximate Value Iteration scheme in which:

- $|v_{k+1} - Tv_k|_\infty = \epsilon$.

a) Let π_k be the greedy policy w.r.t v_k. Show that at the k'th iteration, we have that $|v^{\pi_k} - v^*|_\infty \leq \frac{2\gamma^{k+1}}{1-\gamma}|v_0 - v^*|_\infty + \frac{2\gamma\epsilon}{(1-\gamma)^2}$. (Hint: First analyze the behavior $|\hat{v}_k - v^*|_\infty$ in each iteration and then use the result from tutorial 12).

b) What are the guarantees at the limit of $k \to \infty$?

<u>Question 3</u>- Multiple-step return and algorithms

In class we derived and LSTD algorithm using the 1-step return. In this question we will derive such an algorithm using a multiple-step return. We consider the h-step Bellman operator $(T^\pi)^h$.

a) Based on the h-step Bellman operator, write a corresponding online TD(h) algorithm with function approximation (hint: see the relation between the 1-step return, i.e., the fixed-policy Bellman operator and the TD(0) algorithm).
b) Write down a batch algorithm for the h-step Bellman operator which generalizes LSTD algorithm. Explain how the data need to be collected.
c) Write the error bound for policy evaluation for this case. Compare with the 1-step case. What happens when $h \to \infty$? Explain.