Homework 3: The State-Action Polytope and the Dual Linear Program

CSE 599: Reinforcement Learning and Bandits

1 Constraints on the state-action visitation μ^π *[10 points]*

Verify the following: for all states $s \in \mathcal{S}$, μ^π satisfies:

$$\sum_a \mu^\pi(s,a) = (1-\gamma)d_0(s) + \gamma \sum_{s',a'} P(s|s',a')\mu^\pi(s',a') \,.$$

2 Completeness of the $\mu^\star$ *[30 points]*

Consider the state-action polytope $\mathcal{K}$, as defined in the notes. Now let us prove the Lemma in section 7.1.2. Specifically, after doing the previous problem, the questions asks you to show that: if $\mu \in \mathcal{K}$ then there exists a (stationary) policy π such that $\mu^\pi = \mu$.

3 Non-stationary policies and $\mu^\star$ *[40 points]*

Now consider a possibly non-stationary (i.e. time dependent) policy π, which may depend on the history of observed states and actions. The visitation distribution μ^π is also well defined in this case (exactly as before). Now prove the following stronger lemma:

Lemma 1. *For all non-stationary policies π, we have that $\mu^\pi \in \mathcal{K}$. Furthermore, if $\mu \in \mathcal{K}$, then there exists a (stationary) policy π such that $\mu^\pi = \mu$.*

(note that you already proved the second claim in the previous problem).

4 Recovering the optimal policy *[20 points]*

Suppose $\mu^\star$ is the solution to the dual LP (as defined in the notes). Define:

$$\pi(a|s) := \frac{\mu^\star(s,a)}{\sum_{a'} \mu^\star(s,a')}.$$

Show that π is an optimal policy in the MDP. Furthermore, show that $\mu^\pi = \mu^\star$. You may use the fact that $\mu^\star$ is the state-action visitation distribution of an optimal policy.