HW7: Policy Search

CS 4300: Artificial Intelligence **Tucker Hermans**
University of Utah

---

1 Policy Gradient

In order to do policy gradient, we need to be able to compute the gradient of the value function J with respect to a parameter vector θ: $\nabla_\theta J(\theta)$. By our algebraic magic, we expressed this as:

$$\nabla_\theta J(\theta) = \sum_a \pi_\theta(s_0, a) R(a) \underbrace{\nabla_\theta \log\left(\pi_\theta(s_0, a)\right)}_{g(s_0,a)} \tag{1}$$

If we us a linear function thrown through a soft-max as our stochastic policy, we have:

$$\pi_\theta(s, a) = \frac{\exp\left(\sum_{i=1}^{n} \theta_i f_i(s, a)\right)}{\sum_{a'} \exp\left(\sum_{i=1}^{n} \theta_i f_i(s, a')\right)} \tag{2}$$

Compute a closed form solution for $g(s_0, a)$. Explain in a few sentences *why* this leads to a sensible update for gradience ascent (i.e., if we plug this in to Eq (1) and do gradient ascent, why is the derived form reasonable)?