Distance Metrics

As part of our definition of impact, we require a measure of the distance between two states within an MDP. Multiple possible metrics are proposed here.

Random Distance

Loosely speaking, the random distance between state i and state j is the expected number of timesteps it takes to get from state i to state j, when following a purely random policy.This can be formalized as follows. Construct an episodic Markov Decision Process whose model is identical to the original MDP, where each episode begins with the agent in state s_i. Each action results in a reward of -1, and arriving at state s_j terminates the episode, andresults in a reward of 0. Then, the distance from s_i to s_j is as follows: $d_{rand}(s_i, s_j) := -V^{\pi_{rand}}(s_i)$

Direct Distance

The direct distance between state i and state j is the expected number of timesteps it takes to get from state i to state j, when following a policy that minimizes this quantity. Under the same Markov Decision Process as described for the random distance metric, the distance from s_i to s_j is quite similar: $d_{dir}(s_i, s_j) := -V^{\pi*}(s_i)$

Search Distance

The search distance between s_i and s_j is related to the direct distance, but also accounts for the difficulty of finding the optimal policy that minimizes the number of timesteps it takes to get from s_i to s_j. Consider Agent x in the same MDP as described for the random and direct distance metrics, and starting with π_{rand}. Let S_x be the set of states in Agent x's MDP. Let Π_x be the set of policies available to Agent x. Agent y is in an MDP in which the set of states is $S_x \times \Pi_x$. Agent y begins in state (s_i, π_{rand}), and all $(s_j, .)$ are terminal. The first term of Agent y's state represents the state that Agent x is in, and the second term represents the policy that Agent x is following. Agent y's actions are as follows. Agent y can have Agent x take an action according to its current policy (the policy identified by Agent y's current state). In this case, Agent y recieves the same reward as Agent x (that is, -1). Naturally, the first term of Agent y's state changes to reflect the change to Agent x's state. Alternatively, Agent y can change the policy that Agent x follows, by selecting any subset of the actions available to Agent x in its current state, such that the total probability assigned by Agent x's policy to the actions in that subset is less than or equal to $1/2$. Then, Agent x adjusts its policy to assign 0 probability to those actions, the other available actions are normalized so that their probabilities sum to 1. This amounts to Agent y providing Agent x one bit of information about which action to take from that specific state. Naturally, the second term of Agent y's state changes to reflect Agent x's new policy. Agent y recieves a reward of $-C$. Given C, the search distance as follows. $d_{search,C}(s_i, s_j) := -V_y^{\pi*}((s_i, \pi_{rand}))$

Equivalent Resistance Distance

Map the MDP to a curcuit as follows. Let each state represent a node. At a given state s_x, for each possible action a, and for each possible successor state s_y that that action can lead to, let there be a wire between those two states with resistance equal to $1/p(s_y|a)$, (along with a gate that only allows current to flow from s_x to s_y). $d_{res}(s_i, s_j)$ is equal to the equivalent resistance from the former point to the latter.

Impact

From a distance metric, we can construct an impact measure as follows. Suppose our agent is in an MDP with access to a null policy $\varnothing$. Let $a_t\varnothing$ represent the policy of taking action a at time t, and otherwise following $\varnothing$. Let $\gamma \in [0, 1)$ be a discounting factor. Let $\mu_t(s_i|\pi_{t,k})$ be the probability that $s = s_i$ at time k given whatever state is observed at time t and given following policy π from time t up until but not including time k. Let $I_t(a)$ be the impact of action a at time t. Effectively, the impact of an action a is the discounted expected distance from the state that would arise were we to take action a then follow a null policy to the state that would arise were to just follow a null policy. Formally, $I_t(a) = \sum_{k=t+1}^{\infty} \gamma^{k-(t+1)} \sum_{s_j \in S} \sum_{s_i \in S} \mu_t(s_i|(a_t\varnothing)_{t,k})\mu_t(s_j|\varnothing_{t,k})d(s_i, s_j)$. This formulation may be improved by using counterfactual probabilities as outlined by Stuart Armstrong here, so we can calculate the probability of seeing state s_i at time k after following policy $a_t\varnothing$, given that we would have seen state s_j at time k after following policy $\varnothing$. This direction is deffered for now, given the significant added difficulty for an agent to detect the underlying POMDP that gives rise to its observations. Indeed, two observationally equivalent POMDP’s can fail to be counterfactually equivalent (different counterfactual probabilities can arise).