

PSRL for motion planning

I'm going to make two suggestions on the reward formulation, one containing the gist of what I think the reward should contain, and the other simpler one that comes from counterarguing the first one.

Let G be a graph representing an MDP, sampled from $G \sim \text{distribution}(G)$. Let an *episode* indicate N rollouts spent on each MDP, where each rollout is a path evaluation on the true (unknown) G^* . Let ξ be a trajectory, and let $|\xi|$ be the length (cost) of the trajectory.

Reward per episode:

$$R = \sum_{n=1}^N \left[\alpha e^{-|\xi_n|} - \lambda \frac{\text{False edges on } \xi_n}{\text{True edges on } \xi_n} + \gamma |E_{\text{new}}^{(n)}| \right]$$

where $|E_{\text{new}}^{(n)}|$ is the number of new edges evaluated on the n th rollout.

Per episode, this reward encourages the policy to rollout as many shortest paths as possible while minimizing the catastrophic event of putting “false” edges on a path being evaluated and encouraging new edge exploration. At the high level, the objective is to succeed in as many rollouts with short paths while identifying the true G^* .

Suppose $\gamma = 0$. This is a greedy algorithm. For $n = 1$, the optimal policy is to choose the shortest path. If all edges on ξ_1 are true, the optimal policy is to rollout the same path for the remaining $N - 1$ trials. This is what the policy in the true G^* would look like, along with other G s whose ξ_1 turn out to be feasible in G^* . If ξ_1 had k true edges and $L - k$ false edges, from $n \geq 2$, the optimal policy is to use as many known edges as possible while minimizing the total path length. The balance of α and λ indicates how much the policy is inclined to succeed in the true G^* while maintaining the initial belief that the current G is correct. If λ is too small, the policy will be inclined to try the shortest path in G regardless of its continued failure on the true G^* . With an appropriate λ the policy should not put any known false edges on a new rollout and utilize only the remaining unchecked edges + checked true edges to compute the next shortest path.

Suppose $\lambda = 0$. This is an explorative algorithm. Even for the true G^* , at every rollout, it would try to find the next shortest path with minimal overlap with previous paths, where the amount of overlap is determined by the balance of α and γ .

Counterargument In a typical PSRL, the MDP sampled at episode K remains to be the “assumed true MDP” until the end of that episode. However, the above formulation does not hold this assumption because it penalized “false” edges from $n = 2$. From then, the algorithm is already using a different G . Such an approach is equivalent to doing only 1 rollout per episode and using the evaluated edges to update the MDP distribution. This leads to the following reward formulation:

Reward per episode:

$$R = \sum_{n=1}^N \left[\alpha e^{-|\xi_n|} + \gamma |E_{\text{new}}^{(n)}| \right]$$

Here, the policy is encouraged to evaluate G 's short(est) path in G^* while validating as many new edges as possible. The balance of α and γ determines how explorative this algorithm is. If all edges in the shortest path in G have already been evaluated, with a proper γ , the optimal algorithm may choose to try a completely new, short-enough path.