1 Motivations

1. A policy should accomplish diverse and unknown user-specified goal at test time.
2. If the goal specified at test time is unknown, then the policy needs to perform well on all possible goals.
3. Then, the goal proposal distribution, from which we sample goals for the policy to practice, should be the uniform distribution over the set of possible goals.
4. However, we do not know this set a prior and can only observe sample from the set.
5. State and goal are assumed to be equivalent in this paper.
6. At each iteration, we can sample state by performing goal-directed exploration. How do we construct the goal proposal distribution at each iteration such that it converges to the uniform dist. over the set of possible goals over time?

2 Abstraction

1. To analyze the setting, we abstract away needless details of this process.
2. A goal $\mathbf{G} \sim p_\phi$ is sampled from the goal dist., and then the agent attempts to achieve this goal, which results in a distribution of states $\mathbf{S} \in \mathcal{S}$ seen along the trajectory.
3. The marginal dist. over $\mathbf{S}$ is written as $p\left(\mathbf{S}|p_\phi\right)$

3 Skew Fit Algorithm

1. Given a generative model p_{ϕ_t} at iteration t, they want to obtain $p_{\phi_{t+1}}$ such that p_{ϕ_t} has higher entropy than p_{ϕ_t} over the set of possible goal.
2. While they do not know the set of valid states, they can sample state from $p\left(\mathbf{S}|p_\phi\right)$, resulting in an empirical distribution over the states

$$p_{\mathrm{emp}_t}(\mathbf{s}) \triangleq \frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\left\{\mathbf{s}=\mathbf{S}_n\right\}, \quad \mathbf{S}_n \sim p\left(\mathbf{S}|p_{\phi_t}\right)$$

3. They use this empirical dist. to train the next generative model $p_{\phi_{t+1}}$.
4. However, doing this does not guarantee that the entropy of $p_{\phi_{t+1}}$ is higher than the entropy of p_{ϕ_t}
5. They thus train the generative model to approximate the uniform dist.

$$argmax_\phi \mathbb{E}_{\mathbf{S}\sim U_\mathcal{S}}\left[\log p_\phi(\mathbf{S})\right]$$

6. To get an unbiased estimate of the obj, they invoke importance sampling

$$\mathbb{E}_{\mathbf{S}\sim U_\mathcal{S}}\left[\log p_\phi(\mathbf{S})\right] = \mathbb{E}_{\mathbf{S}\sim p_{\mathrm{emp}\ t}}\left[\frac{U_\mathcal{S}(\mathbf{S})}{p_{\mathrm{emp}\ t}(\mathbf{S})}\log p_\phi(\mathbf{S})\right] \propto \mathbb{E}_{\mathbf{S}\sim p_{\mathrm{emp}\ t}}\left[\frac{1}{p_{\mathrm{emp}\ t}(\mathbf{S})}\log p_\phi(\mathbf{S})\right]$$

7. They argue that computing $p_{\mathrm{emp}\ t}$ requires marginalizing out the MDP dynamics, they thus approximate it with p_{ϕ_t}.
8. Presumably to trade off bias for lower variance, they introduce a new hyper-param α and weight each state by

$$w_{t,\alpha}(\mathbf{S}) \triangleq p_{\phi_t}(\mathbf{S})^\alpha, \quad \alpha < 0$$

9. To further reduce variance, they invoke sampling importance sampling (SIR), rather than sampling from the empirical distribution and weighting each sample, SIR explicitly defines each dist. as

$$p_{\text{skewed }t}(\mathbf{s}) \triangleq \frac{1}{Z_\alpha} p_{\text{emp }t}(\mathbf{s}) w_{t,\alpha}(\mathbf{s}), \quad Z_\alpha = \sum_{n=1}^{N} p_{\text{emp }t}\left(\mathbf{S}_n\right) w_{t,\alpha}\left(\mathbf{S}_n\right)$$

10. They then fit $p_{\phi_{t+1}}$ to $p_{\text{skewed }t}$ with MLE.

11. At iteration $t+1$, to sample goal for the goal-conditioned policy, they can either use $p_{\text{skewed }t}$ or $p_{\phi_{t+1}}$.

4 Skew Fit analysis

1. They first show that if the entropy of $p_{\phi_{t+1}}$ is always larger than the entropy of p_{ϕ_t} and is equal iff p_{ϕ_t} is the uniform dist., then p_ϕ converges to the uniform dist.

2. They then show that the condition for the entropy of $p_{\phi_{t+1}}$ to be larger than the entropy of p_{ϕ_t} is that the log density of $p_{\text{emp }t}$ and p_{ϕ_t} should be correlated.

3. They argue that this should happen for a high performing goal-conditioned policy, since $p_{\text{emp }t}$ is the distribution resulting from trying to reach goals sampled from p_{ϕ_t}.

5 Interesting bits from experimental section

1. They argue that their methods lead to diverse goals proposed to the goal-condition policy, while previous methods only propose goals closed to the initial state distribution.

2. They argue that the most common failure of prior methods is that the goal distribution collapse, resulting in the agent only learning to reach a fraction of the state.