Target Entropy Annealing For Discrete Soft Actor-Critic

Yaosheng Xu, Dailin Hu, Lilian Liang, Stephen McAleer, Pieter Abbeel, Roy Fox

Introduction

Soft Actor-Critic (SAC) is considered the state-of-the-art algorithm in continuous action space settings. It uses the maximum entropy framework for efficiency and stability, and applies a heuristic temperature Lagrange term to tune the temperature, which determines how “soft” the policy should be. It is counter-intuitive that empirical evidence shows SAC does not perform well in discrete domains. In this paper, we investigate the possible explanations for this phenomenon and propose Target Entropy Scheduled SAC (TES-SAC), an annealing method for the target entropy parameter applied on SAC.

Recap on SAC

Soft Bellman backup:

\[ J_Q(\theta) = \frac{1}{2} (r + \gamma V_{\theta}(s') - Q_{\theta}(s, a))^2 \]

where

\[ V_{\theta}(s') = E_{(a'|s') \sim \pi_{\theta}, (s', a') \sim \pi_{\theta}}[Q_{\theta}(s', a') - \alpha \log \pi_{\theta}(a'|s')] \]

Policy update:

\[ D_{KL}[\pi_{\theta}(\cdot|s) \parallel \exp(\frac{1}{\alpha} Q_{\theta}(s, \cdot)) / Z_{\theta}(s)] \]

Temperature Tuning

Temperature Lagrange Term

SAC tunes the temperature \( \alpha \) as the dual of a policy entropy constraint.

\[ \nabla \alpha = \mathcal{H}[\pi_{\theta}] - \mathcal{H} \quad [\text{Haarnoja et al.}, 2018] \]

Constant Target Entropy (TE)

- Set a too big TE (0.98 max-ent) [Christodoulou, 2019]
  - \( \Rightarrow \) policy close to random
- Set a too small TE
  - Temperature \( \alpha \) exponentially drops
  - \( \Rightarrow \) Policy entropy never reaches target entropy

Why temperature \( \alpha \) exponentially drops when setting a small constant TE?

- Some environments have a minimum entropy level not close to 0
  - Some actions are really similar
- Policy entropy never reaches target entropy

As a result, temperature \( \alpha \) exponentially drops. Below is an example in MsPacman. Orange line is for target entropy = 0.01 * max-ent, blue line is for target entropy = 0.5 * max-ent. The left figure shows the log alpha, in which orange line drops quickly, because the policy entropy cannot reach target entropy. The right figure shows the policy entropy, in which orange line reaches around 0.1 without further dropping, whereas the target entropy is set to 0.01 * log|9| = 0.022 < 0.1.

Comparing 4 different discrete SAC settings

- Target entropy = 0.98 max-ent = 0.98 log|A|
- Target entropy = 0.5 max-ent
- Target entropy = 0.01 max-ent
- TES-SAC with Avg threshold=0.01, Std threshold=0.05, drop multiplier=0.9

Recap on SAC

Soft Bellman backup:

\[ J_Q(\theta) = \frac{1}{2} (r + \gamma V_{\theta}(s') - Q_{\theta}(s, a))^2 \]

where

\[ V_{\theta}(s') = E_{(a'|s') \sim \pi_{\theta}, (s', a') \sim \pi_{\theta}}[Q_{\theta}(s', a') - \alpha \log \pi_{\theta}(a'|s')] \]

Policy update:

\[ D_{KL}[\pi_{\theta}(\cdot|s) \parallel \exp(\frac{1}{\alpha} Q_{\theta}(s, \cdot)) / Z_{\theta}(s)] \]

Temperature Tuning

Temperature Lagrange Term

SAC tunes the temperature \( \alpha \) as the dual of a policy entropy constraint.

\[ \nabla \alpha = \mathcal{H}[\pi_{\theta}] - \mathcal{H} \quad [\text{Haarnoja et al.}, 2018] \]

Constant Target Entropy (TE)

- Set a too big TE (0.98 max-ent) [Christodoulou, 2019]
  - \( \Rightarrow \) policy close to random
- Set a too small TE
  - Temperature \( \alpha \) exponentially drops
  - \( \Rightarrow \) Policy entropy never reaches target entropy

We calculate moving average and moving std of empirical policy entropy.

If Avg is close to target, and Std is small, we drop the target entropy.

Environment Minimum Entropy

Comparing 4 different discrete SAC settings

- Target entropy = 0.98 max-ent = 0.98 log|A|
- Target entropy = 0.5 max-ent
- Target entropy = 0.01 max-ent
- TES-SAC with Avg threshold=0.01, Std threshold=0.05, drop multiplier=0.9

Experiments

Comparing 4 different discrete SAC settings

- Target entropy = 0.98 max-ent = 0.98 log|A|
- Target entropy = 0.5 max-ent
- Target entropy = 0.01 max-ent
- TES-SAC with Avg threshold=0.01, Std threshold=0.05, drop multiplier=0.9

Target Entropy Scheduled SAC

Exponential Window Entropy Estimation

\[ \hat{\mu}_t = \lambda \cdot \hat{\mu}_{t-1} + (1 - \lambda) \cdot e_t, \]

\[ \hat{\sigma}_t = \sqrt{\lambda \cdot (\hat{\sigma}_{t-1}^2 + (1 - \lambda) \cdot (e_t - \hat{\mu}_{t-1})^2)} \]

We calculate moving average and moving std of empirical policy entropy.

If Avg is close to target, and Std is small, we drop the target entropy.

References
