Target Entropy Annealing For Discrete Soft Actor-Critic

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

University of California, Irvine

NEURAL INFORMATION PROCESSING SYSTEMS

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:

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

where

$$V_{\bar{\theta}}(s') = \mathbb{E}_{(a'|s')\sim\pi_\phi}[Q_{\bar{\theta}}(s',a') - \alpha \log \pi_\phi(a'|s')]$$

Policy update:

$$\mathrm{D}_{KL}\left[\pi_\phi(\cdot|s) \middle\| \frac{\exp\left(\frac{1}{\alpha}Q_\theta(s,\cdot)\right)}{Z_\theta(s)}\right]$$

Temperature Tuning

Temperature Lagrange Term

SAC tunes the temperature α as the dual of a policy entropy constraint.

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

Constant Target Entropy ($\bar{\mathcal{H}}$)

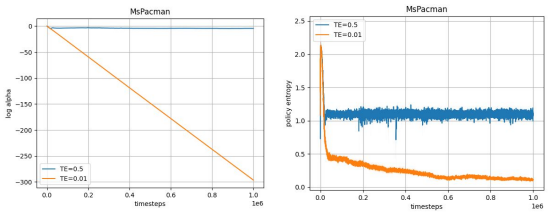
- Set a too big TE (0.98 max-ent) [Christodoulou, 2019]
  - ⇒ policy close to random
- Set a too small TE
  - Temperature α exponentially drops
  - $r(s_t,a_t) + \alpha\mathbb{H}[\pi(\cdot|s_t)] \;\rightarrow\; r(s_t,a_t)$

Environment Minimum Entropy

Why temperature α 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 α 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.



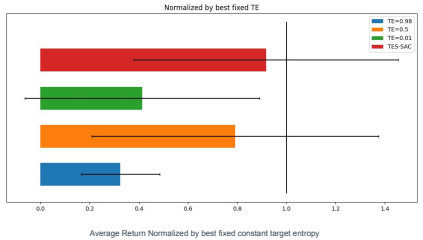
Log alpha and policy entropy in MsPacman using TE=0.5 and TE=0.01

Target Entropy Scheduled SAC

Exponential Window Entropy Estimation

$$\hat{\mu}_i = \lambda \cdot \hat{\mu}_{i-1} + (1-\lambda)\cdot e_i,$$
$$\hat{\sigma}_i = \sqrt{\lambda \cdot (\hat{\sigma}^2_{i-1} + (1-\lambda)\cdot(e_i - \hat{\mu}_{i-1})^2)}.$$

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.

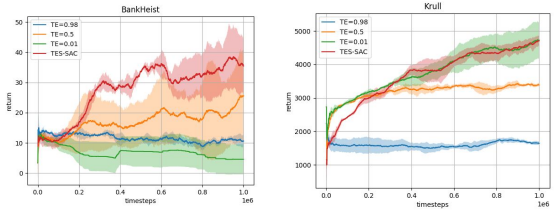


Average Return Normalized by best fixed constant target entropy

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



Average Return in BankHeist and Krull, comparing 4 different discrete SAC settings

References

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al (2018). "Soft actor-critic algorithms and applications." In: *arXiv preprint arXiv: 1812.05905*

Petros Christodoulou (2019). "Soft actor-critic for discrete action settings." In: *arXiv preprint arXiv:1910.07207*