Review: Soft Actor Critic (Haarnoja et al., 2018a)

In this work, Haarnoja et al. (2018a) build upon previous works in maximum entropy reinforcement learning (RL), most notably soft Q-learning (Haarnoja et al., 2017), to formulate a novel theoretical framework and a related practical algorithm for robust and sample-efficient reinforcement learning.

The motivation for maximum entropy RL is intuitive. Adding an entropy term to the objective to be maximised encourages learned polices to retain some randomness aiding exploration and the learning of more robust behaviour. In this way the entropy term can be considered as a regulariser which encourages agents to learn policies which display a range of 'skills' and therefore are more robust to changes in the environment and more generalisable.

This paper is not the first to introduce an entropy term to the learning objective. The contribution of this work is a formalised framework and derivation of maximum-entropy RL from policy iteration. The authors prove that their approach converges an optimal policy in a similar way to the convergence of policy iteration. The authors then construct a practical algorithm, known as Soft Actor-Critic (SAC), based upon the theoretical framework they lay out. SAC is the first off-policy actor-critic algorithm proposed in the maximum-entropy framework. In the experiments the authors empirically show that SAC achieves state-of-the-art performance in several MuJoCo locomotion domains. The sample efficiency and the robustness to hyperparameter settings demonstrated by the authors empirically are a crucial step forwards for future work scaling RL for application to real-world problems (e.g. robotics) where sample efficiency and robustness historically limited the efficacy of practical RL.

The authors focus on the case of continuous actions. We note that SAC can be adapted to a discrete action space by updating the objective. Equation 10 of the paper lays out the objective to be maximised with a term in the Kullback-Leibler divergence between the policy and an approximation to an 'optimal play' distribution. Since they work with continuous actions they are forced to approximate the integral using the reparameterisation trick. They then appeal to policy being parameterised as a differentiable neural network to make the objective differentiable. In the case of discrete actions, the integral which required approximation becomes a sum and we may therefore work with the objective in Equation 10 directly. Otherwise the working of the paper holds for both discrete and continuous actions.

Finally, we consider the limitations of SAC and propose possible solutions. We consider two key limitations to the work as presented in the paper i) their choice of entropy measure (Shannon entropy) for regularisation never eliminates actions from a policy even where they are repeatedly experienced as leading to negative outcomes and ii) the use of entropy regularisation leads to sensitivity of results to reward scaling or equivalently the temperature parameter which weights the entropy term in the objective.

The first limitation is highlighted and handled by the recent work of Chen and Peng (2019). Appealing to previous works (e.g. (Chen et al., 2018)) they show that utilising Tsallis entropy in the place of Shannon entropy for maximum-entropy RL leads to agents that ignore completely unpromising actions and therefore explore more efficiently. Chen and Peng (2019) develop a new framework to be able to perform optimisation with different entropy measures within maximum entropy RL. They empirically show that maximising objectives with such alternative entropy measures which enable learners to remove unpromising actions from consideration allows agents to outperform those trained using SAC on a similar tasks to those used by Haarnoja et al. (2018a).

Turning to the second highlighted limitation, Haarnoja et al. (2018b) improve upon the initial implementation of SAC by integrating a gradient-based learning process for the temperature parameter. They achieve this by updating the optimisation problem to include an inequality constraint such that reward is maximised subject to the entropy exceeding some value. This requires the user to determine a desired minimum entropy level. While this is another hyperparameter it is easier to set and determine from limited experimental experience and allows the temperature parameter to be tuned automatically.

References

[Chen and Peng 2019] CHEN, Gang ; PENG, Yiming: Off-Policy Actor-Critic in an Ensemble: Achieving Maximum General Entropy and Effective Environment Exploration in Deep Reinforcement Learning. In: *arXiv preprint arXiv:1902.05551* (2019)

[Chen et al. 2018] CHEN, Gang ; PENG, Yiming ; ZHANG, Mengjie: Effective exploration for deep reinforcement learning via bootstrapped q-ensembles under tsallis entropy regularization. In: *arXiv preprint arXiv:1809.00403* (2018)

[Haarnoja et al. 2017] HAARNOJA, Tuomas ; TANG, Haoran ; ABBEEL, Pieter ; LEVINE, Sergey: Reinforcement learning with deep energy-based policies. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70* JMLR. org (Veranst.), 2017, S. 1352–1361

[Haarnoja et al. 2018a] HAARNOJA, Tuomas ; ZHOU, Aurick ; ABBEEL, Pieter ; LEVINE, Sergey: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: *arXiv preprint arXiv:1801.01290* (2018)

[Haarnoja et al. 2018b] HAARNOJA, Tuomas ; ZHOU, Aurick ; HARTIKAINEN, Kristian ; TUCKER, George ; HA, Sehoon ; TAN, Jie ; KUMAR, Vikash ; ZHU, Henry ; GUPTA, Abhishek ; ABBEEL, Pieter et al.: Soft actor-critic algorithms and applications. In: *arXiv preprint arXiv:1812.05905* (2018)