Reinforcement Learning with Deep Energy-Based Policies

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, Sergey Levine

***Abstract*—The paper proposes a novel algorithm called *soft Q-learning* for learning *energy based* maximum entropy policies for continuous state and action spaces (previously solved only in the tabular domain) and expresses the optimal policy using a Boltzmann distribution.**

I. Preliminaries

Maximum Entropy RL : The entropy augmented objective is given by :

$$\pi^*_{std} = \arg\max_{\pi} \sum_t \mathbf{E}_{(s_t,a_t)\sim\rho_\pi}\left[r(s_t,a_t) + \alpha\mathcal{H}(\cdot|s_t))\right] \quad (1)$$

where α controls the stochasticity of the optimal policy, $\rho_\pi(s_t)$ and $\rho_\pi(s_t,a_t)$ denotes the state and state-action marginals of the trajectory distribution under policy $\pi(a_t|s_t)$

Soft Value Functions and Energy-Based Models : Energy-based policies are of the form :

$$\pi(a_t|s_t) \propto \exp(-\mathcal{E}(s_t,a_t)) \quad (2)$$

where $\mathcal{E}$ is the energy function, which if represented by Universal function approximator can represent any policy distribution. Connecting between energy function, maximum entropy objective and soft versions of value and Q functions give by : $Q^*_{soft}(s_t,a_t) =$ $r_t + \mathbf{E}_{(s_{t+1},\dots)\sim\rho_\pi}\left[\sum_{l=1}^{\infty}\gamma^l(r_{t+l} + \alpha\mathcal{H}(\pi^*_{MaxEnt}(\cdot|s_{t+1})))\right]$ and $V^*_{soft}(s_t) = \alpha\log\int_{\mathcal{A}}\exp\left(\frac{1}{\alpha}Q^*_{soft}(s_t,a')\right)da'$, then the optimal policy is given by : $\pi^*_{MaxEnt}(a_t|s_t)) = \exp\left(\frac{1}{\alpha}(Q^*_{soft}(s_t,a_t) - V^*_{soft}(s_t))\right)$ where $\frac{1}{\alpha}Q^*_{soft}(s_t,a_t)$ acts as the negative energy, and $\frac{1}{\alpha}V^*_{soft}(s_t)$ serves as the log-partition function and the Q-function and the value function at a future state is related via a soft Bellman equation : $Q^*_{soft}(s_t,a_t) = r_t + \gamma\mathbf{E}_{s_{t+1}\sim p_\pi}[V^*_{soft}(s_{t+1})]$ which becomes the hard maximum over actions when $\alpha \rightarrow 0$.

II. Theory

The problem of learning energy-based policies for continuous states and actions is defined as policy search in an infinite-horizon Markov decision process (MDP) where maximum entropy formulation provides improved exploration and better pretraining. The authors formulate the multimodal stochastic policy as an energy-based model, where the energy function corresponds to Q^*_{soft} obtained by optimizing the maximum entropy objective. A separate sampling network is maintained which is optimized to produce unbiased samples from the policy of the EBM and is used for action selection and to update the EBM. This entropy regularized actor-critic algorithm is more like an approximate Q-learning, with the actor as the sampling network (sampling from an intractable posterior).

III. Algorithm in Detail

Soft Q-Iteration is fixed point iteration analogous to Q-iteration where the V^*_{soft} and Q^*_{soft} are iteratively estimated using the soft Bellman backup operator $\mathcal{T}$. The soft Bellman backup is a contraction with the soft Bellman error $|\mathcal{T}Q - Q|$. The **Soft Q-Learning** algorithm is used for learning maximum entropy policies in continuous domains. The soft Q-function is approximated using a function approximator with parameters θ ($Q^\theta_{soft}(s_t,a_t)$) and the soft value function is computed as the expectation using importance sampling

$$V^\theta_{soft}(s_t) = \alpha\log\mathbf{E}_{q_{a'}}\left[\frac{\exp\left(\frac{1}{\alpha}Q^\theta_{soft}(s_t,a')\right)}{q_{a'}(a')}\right]$$

and the target Q-value is given by

$$\hat{Q}^{\bar{\theta}}_{soft}(s_t,a_t) = r_t + \gamma\mathbf{E}_{s_{t+1}\sim p_s}[V^{\bar{\theta}}_{soft}(s_{t+1})]$$

where θ is replaced by target parameters $\bar{\theta}$ to compute $V^{\bar{\theta}}_{soft}(s_{t+1})$. The algorithm alternates between collecting experience from the environment and updating the soft Q-function and sampling network parameters. The samples are drawn from rollouts using the current policy $\pi(a_t|s_t) \propto \exp\left(\frac{1}{\alpha}Q^\theta_{soft}(s_t,a_t)\right)$ using approximate sampling procedure by utilizing a sampling network based on **Stein Variational Gradient Descent (SVGD)** and **Amortized SVGD**. The stochastic sampling network can be queried to give samples generated extremely fast and it converges to an accurate estimate of the posterior distribution of an EBM. The experience is stored in a replay memory buffer and the parameters are updated using random minibatches from this memory. The resulting algorithm strongly resembles actor-critic algorithm, leading to a simple and computationally efficient implementation. Stein variational gradient descent gives the most greedy directions ($\Delta f^\phi(\zeta^{(i)};s_t)$) to perturb the independent samples $a_t^{(i)} = f^\phi(\zeta^{(i)};s_t)$ (where $\zeta^{(i)}$ are noise samples drawn from a normal Gaussian) which effectively reduces the KL divergence between the policy parameterised by ϕ and the energy-based distribution $\exp\left(\frac{1}{\alpha}(Q^\theta_{soft}(s_t,\cdot) - V^\theta_{soft})\right)$. Hence, any gradient-based optimization method can be used to learn the optimal sampling network parameters. The sampling network f^ϕ is viewed as the actor in the actor-critic algorithm. It is to be noted that the energy-based policies can be trained with fairly broad objectives to produce an initializer for accelerated learning on more specific tasks.

References

[1] Tuomas Haarnoja et al. “Reinforcement learning with deep energy-based policies”. In: *arXiv preprint arXiv:1702.08165* (2017).