CS7011 : Topics in Reinforcement Learning
Summary of RL with Deep Energy-Based Polices : Haarnoja,Tang, Abbeel, Levine
By Sahana Ramnath : EE15B109

1. Abstract
This paper (Haarnoja et al., 2017) proposes a method to learn expressive *energy-based* policies for continuous states and actions(has been done only in tabular domains so far). This method learns maximum entropy policies using a new algorithm called *soft Q-Learning* that expresses the optimal policy via a Boltzman distribution. It uses amortized Stein variational gradient descent to learn a stochastic sampling network that approximates samples from the distribution. This new algorithm provides improved exploration and compositionality that facilitates transfer of skills between tasks; this was confirmed experimentally.

2. Notion of Optimality
Conventionally, the optimal solution is always a deterministic policy; although stochastic policies are preferred for exploration, exploration is usually attained(in deep RL so far) by injecting noise into the deterministic policy. But, there are cases where stochastic policies are explicitly preferred : (exploration and final policy)multimodal objectives, getting compositionality via pretraining, uncertain system dynamics etc.

A stochastic policy is the optimal solution when considering optimal control with probabilistic inference. Framing control as inference produces policies that capture not one single deterministic 'best' way to do the task, but tries to learn **all** the ways to do this task. This serves as a good solution, and also as a good initialization to a more specific task(eg:-first learning all the ways that a robot could move forward, and then using this as an initialization to learn separate running and bounding skills). This leads to discovery of the best mode(s) in a multimodal setup and an ability to perform the same task in different ways(and so, can help the agent recover from adversarial perturbations).

To achieve the above, the framework of maximum entropy policy search has to extend to arbitrary policy distributions. This paper formulates a stochastic policy as a conditional energy based model(EBM) with the energy function corresponding to the 'soft' Q-function obtained when optimizing the maximum entropy objective. It uses an approximate sampling procedure based on training a separate sampling network, which is optimized to produce unbiased samples from the policy EBM; this sampling network can then be used for both updating the EBM and action selection. This approximate Q-Learning method can be seen as parallel to entropy regularized actor critic algorithms with the actor serving the role of an approximate sampler from an intractable posterior.

3. Soft Value Functions and EBMs
To express multimodal behaviour, the policy can have an energy based form : $\pi(a_t|s_t) = exp(-\epsilon(s_t, a_t))$, where ϵ is an energy function that could be represented by a neural network. If a universal function approximator is chosen as ϵ, any $\pi(a_t|s_t)$ can be represented. There is a close connection between such energy based models and soft versions of Q-functions when $\epsilon(s_t, a_t) = -\frac{1}{\alpha}Q_{soft}(s_t, a_t)$.

$Q^*_{soft}(s_t, a_t) = r_t + \mathrm{E}_{\rho_\pi}[\sum_{l=1}^{\infty}\gamma^l(r_{t+l} + \alpha\mathcal{H}(\pi^*_{MaxEnt}(.|s_{t+l})))]$

$V^*_{soft}(s_t) = \alpha log \int_{\mathcal{A}} exp(\frac{1}{\alpha}Q^*_{soft}(s_t, a'))da'$

$\pi^*_{MaxEnt}(a_t|s_t) = exp(\frac{1}{\alpha}(Q^*_{soft}(s_t, a_t) - V^*_{soft}(s_t)))$

3.1. Training EBMs via Soft Q-Learning
The optimal policy can be obtained by iteratively updating estimations of V^*_{soft} and Q^*_{soft}. This leads to a fixed point iteration that similar to Q-iteration(and is hence called "soft" Q-iteration or soft Bellman backup). However, this cannot be performed in high dimensional or continuous state and action spaces, and sampling from an EBM is in general intractable. So, 'soft Q-Learning', a practical implementation with a function approximator θ representing $Q^\theta_{soft}(s_t, a_t)$ and a SGD update procedure for θ is proposed. The soft value function is expressed as an expectation via importance sampling : $V^\theta_{soft}(s_t) = \alpha log \mathrm{E}_{q_{a'}}[\frac{exp(\frac{1}{\alpha}Q^\theta_{soft}(s_t,a'))}{q_{a'}(a')}]$.

This stochastic optimization problem can be solved approximately using stochastic gradient descent using states and actions sampled from rollouts of the current policy $\pi(a_t|s_t)$. The sampling network is based on Stein Variational Gradient Descent(SVGD) and amortized SVGD.

3.2. Algorithm - Short Version
This paper proposes a soft Q-Learning algorithm to learn maximum entropy policies in continuous domains. The algorithm alternates between collecting new experience and updating the soft Q-function and sampling network parameters. The experience is stored in a replay memory and parameters are updated using random minibatches from this, and the updates use a delayed version of the target values(both similar to a DQN).

4. Experiments
4.1. Multigoal Environment
This experiment aimed to test whether the proposed model shows multimodal behaviour. The environment had 4 symmetrically placed goals, and the reward distribution is a mixture of Gaussians with means at the four goal positions. A policy trained with a deterministic algorithm such as DDPG randomly committed to a single goal, but the policy trained with EBMs learnt Q-values with complex shapes, being multimodal in regions between the goals.

4.2. Multimodal Policies for Exploration
Not all tasks have a clear multigoal environment; but multimodality is present in many tasks. Learning unimodal action distributions may lead to prematurely committing to one mode and converging to suboptimal behaviour. To demontrate this, the EBM model and DDPG were tested on the swimmer snake task and the quadrupedal robot maze task.

4.3. Speedup training using pretraining
One way to speedup deep NN training is task-specific initialization. But in RL, near optimal policies of source tasks are often deterministic which make them poor initializers for new tasks. But, energy based policies can be trained with fairly broad policies to produce a good initializer for more specific tasks. This is demonstrated on a variant of the quadrupedal robot task.The pretraining phase involves learning to locomote in an arbitrary direction, with a reward that simply equals the speed of the COM. The pretrained policy explores the space extensively and in all directions and gives a good initialization for the policy as compared to one given by DDPG. Deterministic pretraining chooses an

arbitrary but consistent direction in the training environment, providing a poor initialization for finetuning to a specific target task.

References

Haarnoja, Tuomas, Tang, Haoran, Abbeel, Pieter, and Levine, Sergey. Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*, 2017.