Count-Based Temperature Scheduling for Maximum Entropy Reinforcement Learning

Dailin Hu,
Pieter Abbeel,
Roy Fox

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Empirical evidence from Soft Q Learning(SQL)[1] suggests that a state-independent linear scheduling can achieve good performance[2][3].

More insight can be gained from comparing two families of successful RL algorithms:

- G-Learning[2], SQL, Path Consistency Learning(PCL)[4], Soft Actor Critic[5]

$$\pi_i(a|s) \propto \pi_0(a|s) \exp \beta_i(s) Q_i(s,a) \qquad \forall s \in \mathcal{S}, a \in \mathcal{A},$$

- Relative Entropy Policy Search[6], Trust Region Policy Optimization[7],

$$\pi_i(a|s) \propto \pi_{i-1}(a|s) \exp \kappa_i(s) Q_i(s,a) \propto \pi_0(a|s) \exp \left(\sum_{j \le i} \kappa_j(s) Q_j(s,a) \right).$$

Combining the above two equations, we have:

$$\beta_i(s) \approx \sum_{j \le i} \kappa_j(s) \approx \kappa i,$$

Count-Based Soft Q Learning

We propose Count-Based Soft Q Learning based on SQL that uses a state-dependent temperature schedule in which ß grows linearly with the number of times that the algorithm updates the Q function, for any action.

Let n(s,a) be the count of sampled data points, then the inverse temperature in CBSQL is
$\beta(s) = \kappa \sum_a n(s,a)$
with κ > 0 a constant hyperparameter.

Experiments

Tabular Experiments

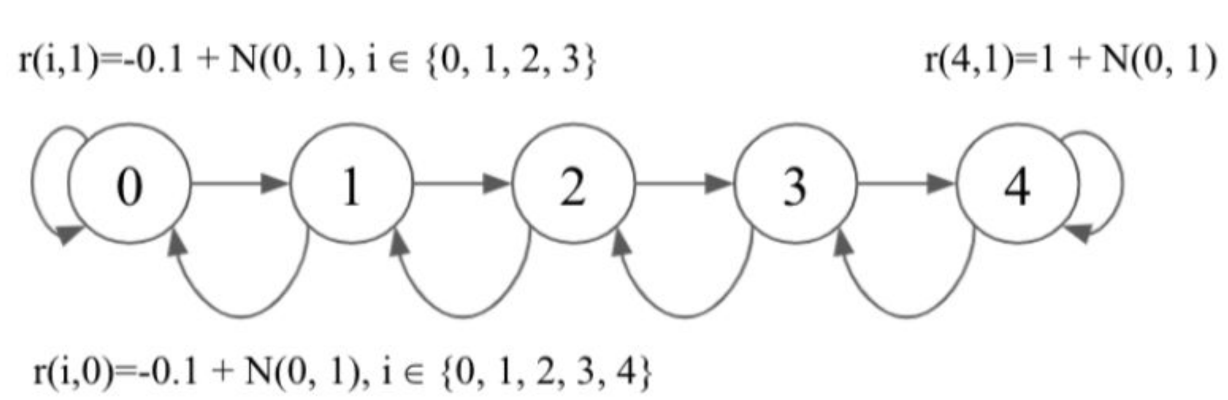

Fig1 . Noisy Chain-walk Problem

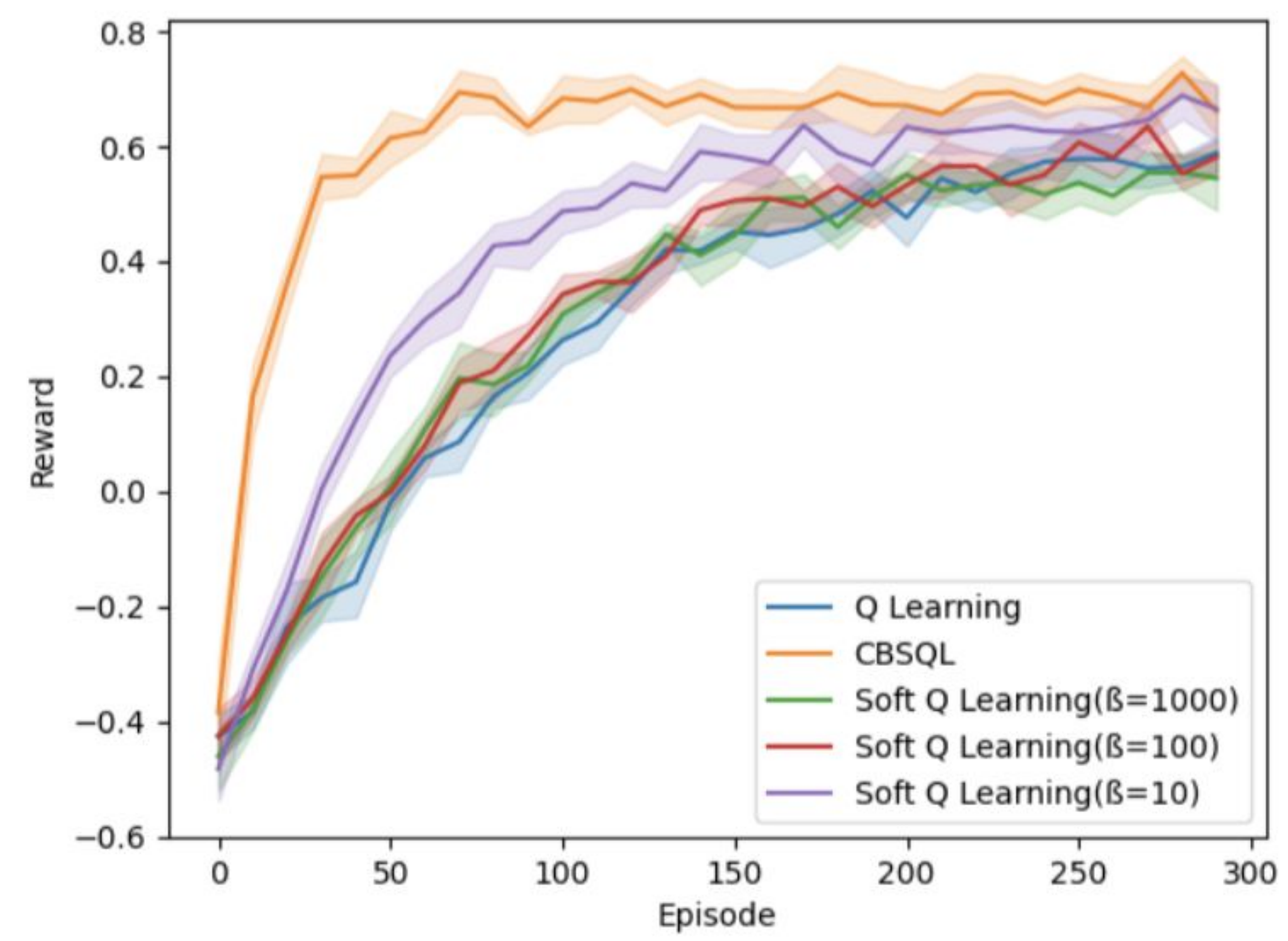

Fig2. Rewards averaged over 10000 runs on the Noisy Chain-Walk Problem

Atari Experiments

Game	DQN	SQL($\beta = 100$)	SQL($\beta = 1000$)	CBSQL
Breakout	5.9 (±5.9)	5.9 (±4.5)	5.1 (±4.7)	**8.2** (±6.1)
Freeway	21.0 (±1.5)	14.6 (±8.5)	22.56 (±4.7)	**25.82** (±4.9)
Pong	1.93 (±2.6)	**17.83** (±2.2)	16.31 (±2.7)	17.56 (±2.0)
Qbert	568.4 (±1101.9)	828 (±1411.7)	564.5 (±1097.5)	**875.3** (±1254.6)
Seaquest	13.5 (±24.1)	4 (±60.2)	17.2 (±24.0)	**84.6** (±60.2)
SpaceInvaders	132.7 (±113.2)	**158.9** (±128.5)	132.25 (±118.4)	138.9 (±112.8)

Fig3. DQN, fixed-temperature SQL and CBSQL average rewards (with standard deviation). Raw score are averaged over the last 100 testing episodes across 3 runs.

Rainbow Integrations to CBSQL

We integrate CBSQL with Rainbow DQN[8], a state-of-the-art reinforcement learning algorithm for memoryless agents including multi-step learning, double-Q learning, prioritized experience replay, dueling networks, distribution RL and noisy networks. All these methods can be straightforwardly applied to soft Q learning except multi-step learning and distributional RL.

Multi-step learning

Multi-step learning with a tuned-number of steps can lead to faster learning in on-policy RL algorithms. In SQL the n step truncated return is

$$\tilde{r}_t^{(n)} = r_t^{(n)} + \frac{1}{\beta} \sum_{k=1}^{n-1} \gamma^k \mathbb{H}[\pi(\cdot|s_{t+k})].$$

Unfortunately empirical policy entropy estimates are often very noisy and calls for further study. In this work we simply use 1-step returns for SQL and CBSQL.

Distributional RL

We adapt distributional RL to SQL and CBSQL by defining a policy distribution of

$$\pi(a'|s') = \frac{\exp \beta(s') \boldsymbol{z}^\intercal \boldsymbol{p}_{\bar{\theta}}(s',a')}{\sum_{\bar{a}'} \exp \beta(s') \boldsymbol{z}^\intercal \boldsymbol{p}_{\bar{\theta}}(s',\bar{a}')}$$

over the values

$$r + \gamma \left(\boldsymbol{z} - \tfrac{1}{\beta} \mathbb{D}[\pi(\cdot|s') \| \pi_0] \right)$$

Preliminary Results

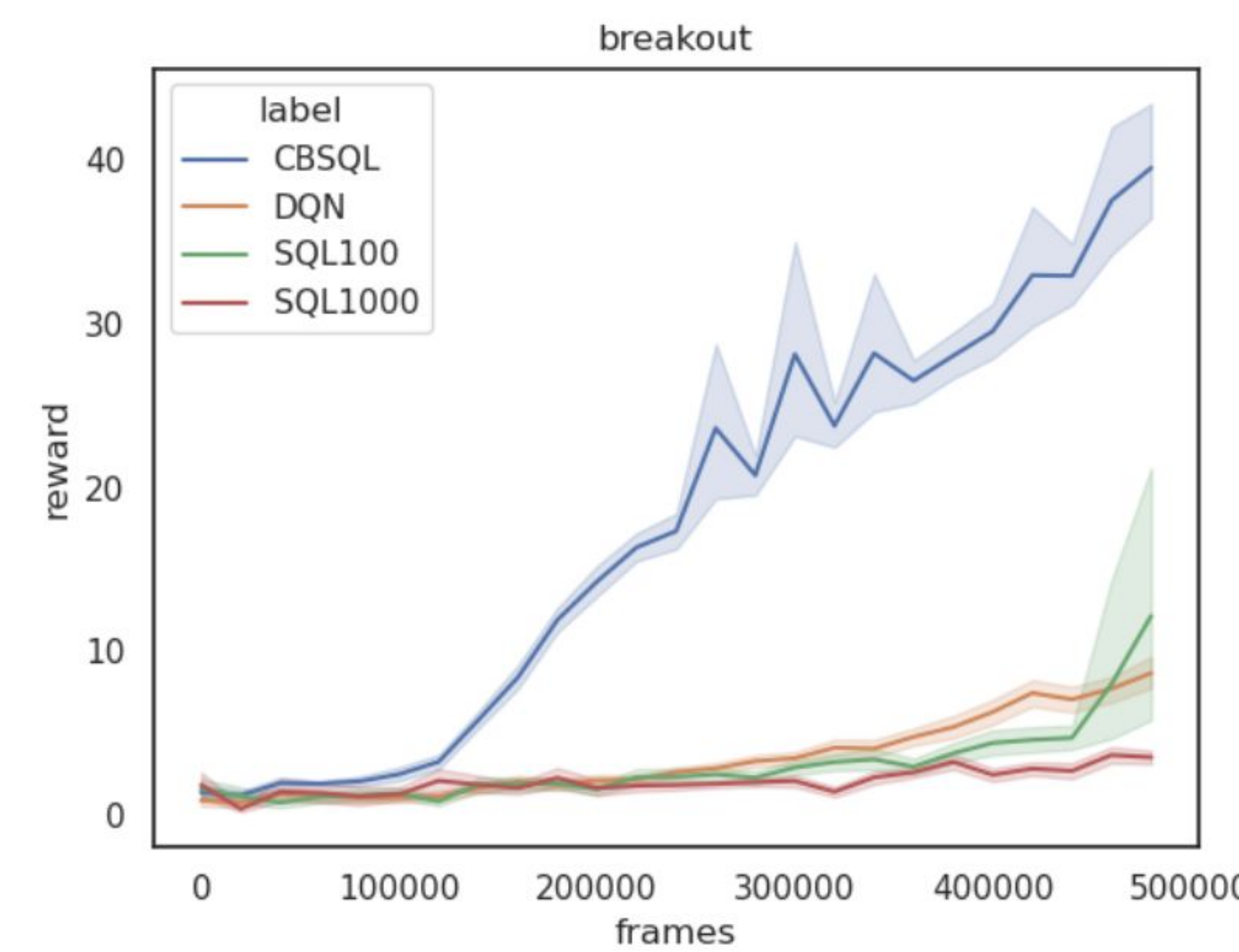

Fig4. CBSQL results compared with DQN and fixed-temperature SQL, with Rainbow. Rewards are averaged over 5 runs.

References

[1] Haarnoja T, Tang H, Abbeel P, Levine S. Reinforcement learning with deep energy-based policies. In International Conference on Machine Learning 2017 Jul 17 (pp. 1352-1361). PMLR.

[2] Fox R, Pakman A, Tishby N. Taming the noise in reinforcement learning via soft updates. arXiv preprint arXiv:1512.08562. 2016

[3] Grau-Moya J, Leibfried F, Vrancx P. Soft q-learning with mutual-information regularization. In International conference on learning representations 2018 Sep 27.

[4] Chow Y, Nachum O, Ghavamzadeh M. Path consistency learning in tsallis entropy regularized mdps. In International Conference on Machine Learning 2018 Jul 3 (pp. 979-988). PMLR.

[5] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International conference on machine learning 2018 Jul 3 (pp. 1861-1870). PMLR.

[6] Peters J, Mulling K, Altun Y. Relative entropy policy search. In Twenty-Fourth AAAI Conference on Artificial Intelligence 2010 Jul 5.

[7] Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. Trust region policy optimization. In International conference on machine learning 2015 Jun 1 (pp. 1889-1897). PMLR.

[8] Hessel M, Modayil J, Van Hasselt H, Schaul T, Ostrovski G, Dabney W, Horgan D, Piot B, Azar M, Silver D. Rainbow: Combining improvements in deep reinforcement learning. InThirty-second AAAI conference on artificial intelligence 2018 Apr 29.