Soft Q-learning with Mutual Information Regularization

Accepted in ICLR 2019

Kyeong Ryeol, Go

M.S. Candidate of OSI Lab

Motivation

- Entropy-regularization RL

$$\max_{\pi} E\left[\sum_t \gamma^t \left(r(s_t, a_t) - \frac{1}{\beta}\log \pi(a_t|s_t)\right)\right]$$

- KL-regularized RL

$$\max_{\pi} E\left[\sum_t \gamma^t \left(r(s_t, a_t) - \frac{1}{\beta}\log \pi(a_t|s_t)/\rho(a_t)\right)\right]$$

* When some actions are simply non-useful or not frequently used
* When actions have significantly different importance depending on the task

Paper list

- When some actions are simply non-useful or not frequently used
  - Soft q-learning with mutual information regularization (ICLR 2019)
- When actions have significantly different importance
  - Information asymmetry in KL-regularized RL (ICLR 2019)
  - Exploiting hierarchy for learning and transfer in KL-regularized RL (Arxiv)

One-step decision-making…

$$\max_{\pi,\rho} E\left[r(s,a) - \frac{1}{\beta}\log \pi(a|s)\,/\rho(a)\right]$$
$$= \max_{\pi}\sum_{s,a} p(s)\pi(a|s)r(s,a) - \frac{1}{\beta}\min_{\rho}\sum_{s} p(s)KL\big(\pi(\cdot\,|s)\parallel\rho(\cdot)\big)$$
$$= \max_{\pi}\sum_{s,a} p(s)\pi(a|s)r(s,a) - \frac{1}{\beta}\mathrm{I}(\mathcal{S},\mathcal{A}) \quad \left(\because\ \rho^*(a) = \sum_{s} p(s)\pi(a|s)\right)$$

mutual-information regularization

Also hold in *Multi-step decision-making(=RL)* when $\gamma \to 1$ (See Appendix)

1. What is an optimal policy for a fixed action prior?
2. What is an optimal action prior for a fixed policy?

- Definitions
  - Transition probability
    - $P_\pi^t(s'|s) = \sum_a P(s'|a,s)\pi(a|s) \quad s.t. \quad \boldsymbol{P_\pi} \in \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^{|\mathcal{S}|}$
  - Stationary distribution over states (assumed to exist)
    - $\boldsymbol{\mu_\pi^T} \coloneqq \lim_{t\to\infty} \boldsymbol{\nu_0^T P_\pi^t} \quad s.t. \quad \mu_\pi(s') = \sum_s P_\pi(s'|s)\mu_\pi(s)\,, \quad \boldsymbol{\mu_\pi^T} = \boldsymbol{\mu_\pi^T P_\pi}$
  - Stationary distribution over actions
    - $\rho_\pi(a) \coloneqq \sum_s \mu_\pi(s)\pi(a|s)$

1. What is an optimal policy for a fixed action prior?

2. What is an optimal action prior for a fixed policy?

$$\max_{\pi} E\left[\sum_t r(s_t, a_t) - \frac{1}{\beta}\log \pi(a_t|s_t)\,/\rho(a_t)\right]$$

- $V_{\pi,\rho}(s) := E\left[\sum_t \gamma^t \left(r(s_t, a_t) - \frac{1}{\beta}\log \pi(a_t|s_t)/\rho(a_t)\right) | s_0 = s\right]$
- $Q_{\pi,\rho}(s,a) := r(s,a) + \gamma E_{s'}\left[V_{\pi,\rho}(s')\right]$

By standard variational calculus,

$$\pi^*(a|s) \propto \rho(a)\exp\left(\beta Q_{\pi^*,\rho}(s,a)\right)$$

$$\rightarrow\ \pi^*(a|s) \propto \exp\left(\beta Q_{\pi^*,\rho}(s,a)\right)\ \ if\ \ \rho(a) = 1/|\mathcal{A}|$$

$$\rightarrow\ \pi^*(a|s) = 1\quad if\ \ a = \arg\max_a Q^*(s,a)\ \ if\ \ \beta \rightarrow \infty$$

1. What is an optimal policy for a fixed action prior?

2. What is an optimal action prior for a fixed policy?

$$\arg\max_{\rho} E\left[\sum_t \gamma^t \left(r(s_t, a_t) - \frac{1}{\beta}\log \pi(a_t|s_t) / \rho(a_t)\right)\right]$$
$$= \arg\max_{\rho} \sum_t \sum_s \gamma^t v_t(s) \sum_a \pi(a|s) \left(r(s,a) - \frac{1}{\beta}\log \pi(a|s) / \rho(a)\right)$$
$$= \arg\max_{\rho} -\frac{1}{\beta}\sum_t \sum_s \gamma^t v_t(s) KL\big(\pi(\cdot\,|s) \parallel \rho(\cdot)\big)$$

- $v_t(\boldsymbol{s}) \coloneqq \sum_{s_0, a_0, \ldots, s_{t-1}, a_{t-1}} p(s_0)\left(\prod_{t'=0}^{t-2} \pi(a_{t'}|s_{t'}) P(s_{t'+1}|s_{t'}, a_{t'})\right) \pi(a_{t-1}|s_{t-1}) P(\boldsymbol{s}|s_{t-1}, a_{t-1})$

$$\rho^*(a) \propto \sum_s \sum_T \gamma^t v_t(s)\, \pi(a|s)$$

stationary distribution over actions

$\mu_\pi(s)$: stationary distribution over states

MIRL : Mutual Information RL

1. Tabular setting

- Q-functions update

$$Q(s,a) \leftarrow (1-\alpha_Q)Q(s,a) + \alpha_Q \left(T^{\rho}_{soft}Q\right)(s,a,s')$$

$$where\ \left(T^{\rho}_{soft}Q\right)(s,a,s') \coloneqq r(s,a) + \gamma \frac{1}{\beta} \log \sum_{a'} \rho(a') \exp\left(\beta Q(s',a')\right)$$

Why?

- $V_{\pi,\rho}(s) = E_a[r(s,a) - \frac{1}{\beta} \log \pi(a|s)/\rho(a) + \gamma E_{s'}[V_{\pi,\rho}(s')]]$

- $Q_{\pi,\rho}(s,a) = r(s,a) + \gamma E_{s',a'}\left[Q_{\pi,\rho}(s',a') - \frac{1}{\beta} \log \pi(a'|s')/\rho(a')\right]$

Substitute $\pi(a|s)$ to $\pi^{upd}(a|s) \propto \rho(a) \exp\left(\beta Q_{\pi,\rho}(s,a)\right)$

MIRL : Mutual Information RL

1. Tabular setting

- Prior update

$$\rho_{i+1}(a) = (1 - \alpha_\rho)\rho_i(a) + \alpha_\rho \pi_i(a|s_i)$$
$$where \;\; s_i \sim \nu_i(s) \;\; and \;\; \pi_i(a|s_i) \propto \rho_i(a) \exp(\beta Q_i(s_i, a))$$

$$\Big\downarrow \text{Converge to...}$$

$$\rho_{\pi_i}(a) = \sum_s \nu_i(s)\pi_i(a|s) \left(= \sum_s \mu_{\pi_i}(s)\pi_i(a|s) \;\; with \;\; \gamma = 1 \right)$$

Common practice
in actor-critic

- β update

$$\beta_i = c \cdot i$$

MIRL : Mutual Information RL

2. High-dimensional state space

- Q-functions update

$$L(\theta, \rho) \coloneqq E_{s,a,r,s' \sim \mathcal{M}} \left[\left(\left(T^{\rho}_{soft} Q_{\overline{\theta}} \right) (s, a, s') - Q_{\theta}(s, a) \right)^2 \right]$$

- Prior update

$$\rho_{i+1}(a) = \left(1 - \alpha_\rho\right) \rho_i(a) + \alpha_\rho \pi_i(a|s_i)$$

- β update

$$\beta_{i+1} = \left(1 - \alpha_\beta\right) \beta_i + \alpha_\beta \frac{1}{L(\theta_i, \rho_{i+1})}$$

Algorithm 1 MIRL

1: **Input:** the learning rates α_ρ, α_Q and α_β, a Q-network $Q_\theta(s,a)$, a target network $Q_{\bar{\theta}}(s,a)$, a behavioural policy π_b, an initial prior ρ_0 and parameters θ_0 at $t=0$.
2: **for** $i = 1$ to N iterations **do**
3: Get environment state s_i and apply action $a_i \sim \pi_b(\cdot|s_i)$
4: Get r_i, s_{i+1} and store (s_i, a_i, r_i, s_{i+1}) in replay memory $\mathcal{M}$
5: Update prior $\rho_{i+1}(\cdot) = \rho_i(\cdot)(1-\alpha_\rho) + \alpha_\rho \pi_i(\cdot|s_i)$
6: **if** i mod update frequency $== 0$ **then**
7: Update Q-function $\theta_{i+1} = \theta_i - \alpha_Q \nabla_\theta L(\theta, \rho_{i+1})|_{\theta_i}$ according to Equation (13)
8: Update parameter $\beta_{i+1} = (1-\alpha_\beta)\beta_i + \alpha_\beta \left(\frac{1}{L(\theta_i, \rho_{i+1})}\right)$
9: **end if**
10: **end for**

- Behavior policy $\pi_b(a|s_i)$

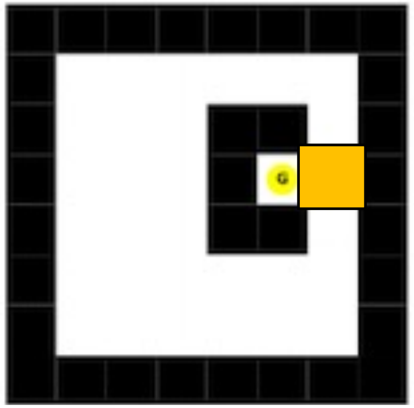
$$a_i = \begin{cases} \arg\max_a \pi_i(a|s_i) & if \quad u \geq \epsilon \quad \text{Exploitation} \\ a \sim \rho_i(a) & if \quad u \leq \epsilon \quad \text{Exploration} \end{cases}$$

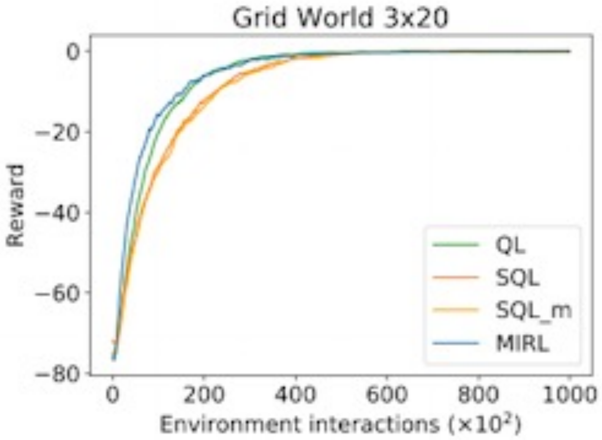
Grid-world

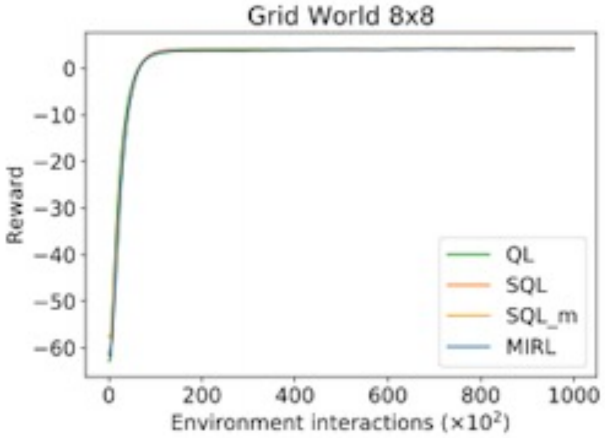
- Settings
  - Reaching a goal : reward 9
  - Else : reward -1
  - Restarted in a random locati
- Baseline
  > Q-learning (QL)
  > Soft Q-learning (SQL)
  > Soft Q-learning
  + behavior policy (SQL_m)
  > MIRL

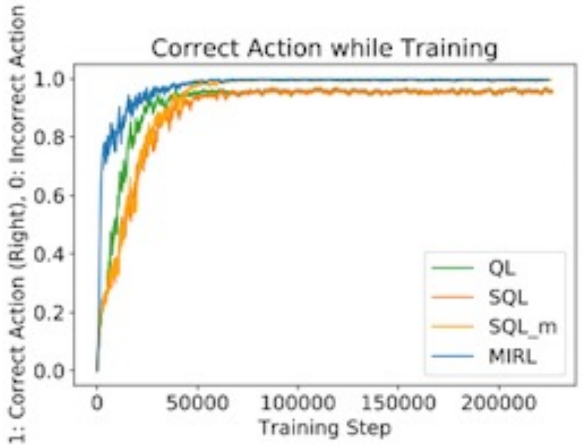
Corridor

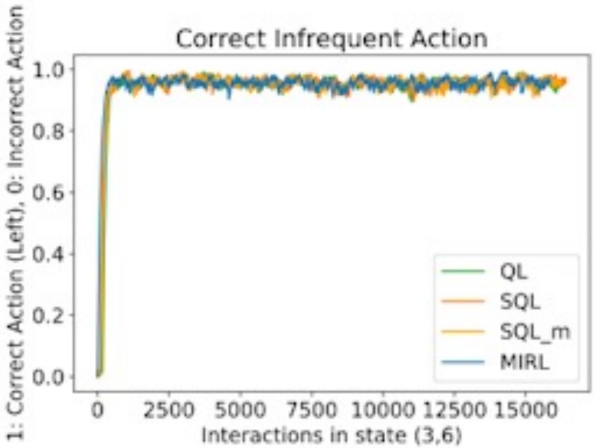
Square world







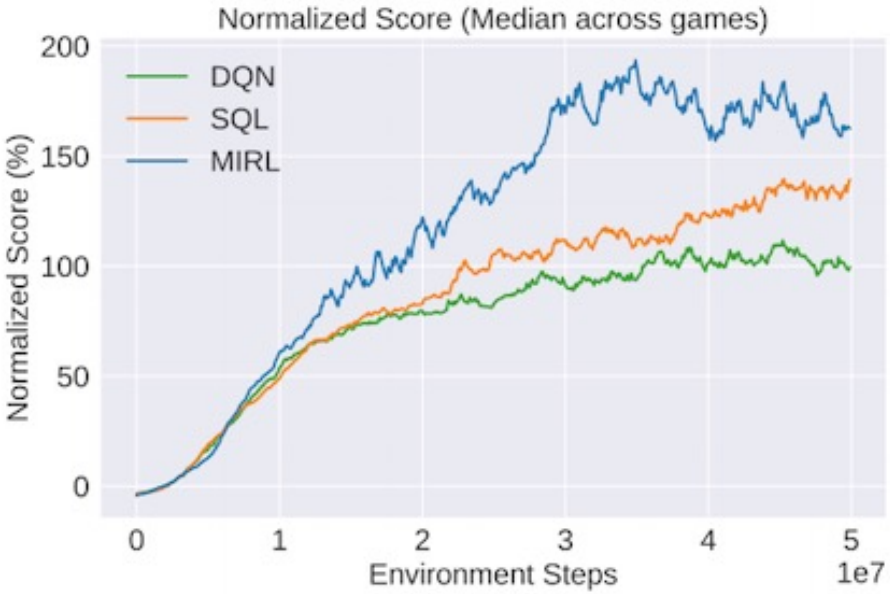




$$Z_{normalized} = \frac{z - z_{random}}{z_{human} - z_{random}} \times 100\%$$

ATARI

- Baselines
  - > Deep Q-Network (DQN)
  - > Soft Q-learning (SQL)
  - > MIRL



Game	DQN (%)	SQL (%)	MIRL (%)
Alien	**101.58**	51.02	40.23
Assault	250.61	283.62	**357.40**
Asterix	166.32	242.73	**330.19**
Asteroids	**9.74**	8.57	7.80
BankHeist	97.12	94.62	**166.26**
BeamRider	99.16	113.64	**117.21**
Boxing	2178.57	2283.33	**2338.89**
ChopperCommand	**72.71**	26.37	65.03
DemonAttack	350.95	451.78	**469.30**
Gopher	474.18	**538.87**	429.44
Kangaroo	351.48	393.16	**405.9**
Krull	843.16	886.68	**1036.04**
KungFuMaster	122.14	**142.04**	121.41
Riverraid	77.21	**109.37**	76.02
RoadRunner	548.90	613.62	**695.88**
Seaquest	21.95	36.00	**64.86**
SpaceInvaders	166.62	**200.38**	164.79
StarGunner	653.44	**681.12**	574.89
UpNDown	183.19	230.82	**394.21**
Mean	356.26	388.83	**413.46**

Table 1: Mean Normalized score in 19 Atari games for DQN, SQL and our approach MIRL.

Ablation study

- Baselines
  > Soft Q-learning (SQL)
  > Soft Q-learning
  + behavior policy (SQL_m)
  > MIRL