Reinforcement Learning as Probabilistic Inference

report is made by
Pavel Temirchev

based on the research of Sergey Levine's team

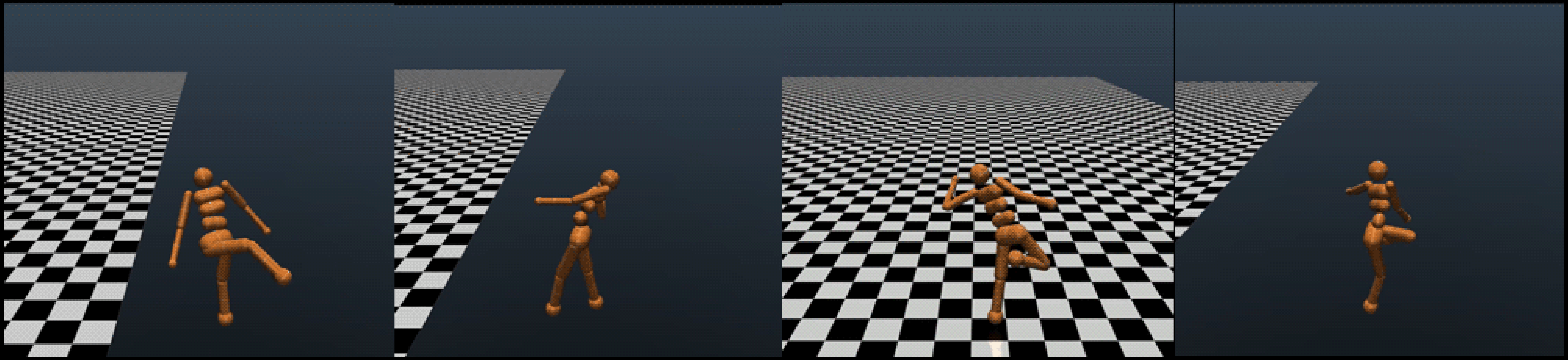
Deep RL
reading group

Motivation

The problems of standard RL:

1) Sample Complexity!

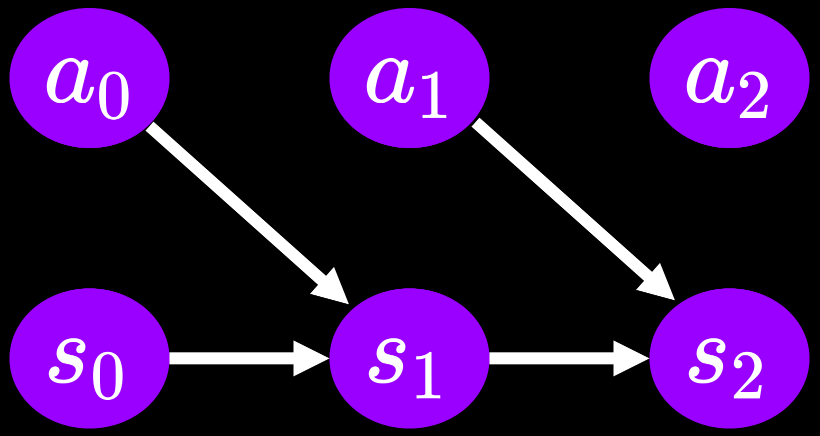
2) Convergence to local optimas



Idea: encourage an agent to investigate all the promising strategies!

REMINDER: standard RL

$$\tau = (s_0, \ldots, a_t, s_t, \ldots, a_T, s_T)$$



Markov process:

$$p(\tau) = p(s_0) \prod_{t=0}^{T} p(a_t|s_t) p(s_{t+1}|s_t, a_t)$$

Maximization problem:

$$\pi^\star = \arg\max_\pi \sum_{t=0}^{T} \mathbb{E}_{s_t, a_t \sim \pi}[r(s_t, a_t)]$$

Q-function:

$$Q^\pi(s_t, a_t) := r(s_t, a_t) + \sum_{t'=t+1}^{T} \mathbb{E}_{s_{t'}, a_{t'} \sim \pi}[r(s_{t'}, a_{t'})]$$

Bellman equality (optimal Q-function):

$$Q^\star(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{s_{t+1}} V^\star(s_{t+1})$$

$$V^\star(s_t) = \max_a Q^\star(s_t, a)$$

Maximum Entropy RL

Policy "proportional" to Q:

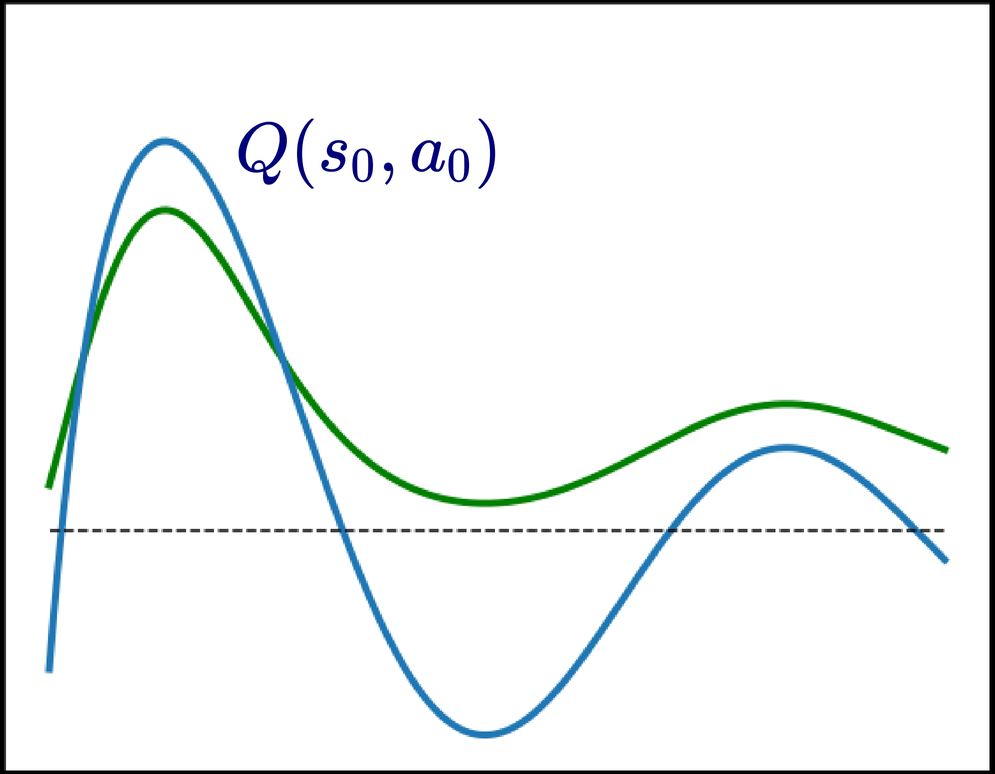
$$a_t \sim \exp Q(s_t, a_t)$$

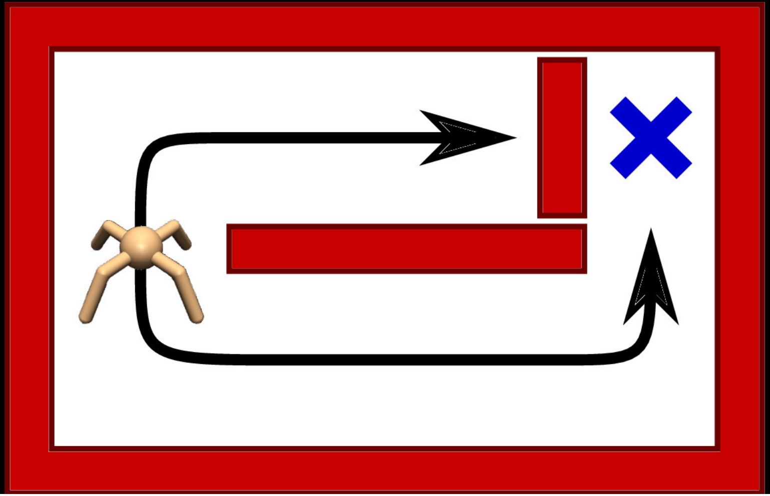
How to find such a policy?

$$\min_\pi \mathrm{KL}\Big(\pi(\cdot|s_0) || \exp Q(s_0, \cdot)\Big) =$$

$$\max_\pi \mathbb{E}_\pi\Big[Q(s_0, a_0) - \log \pi(a_0|s_0)\Big] =$$

$$\max_\pi \mathbb{E}_\pi\Big[\sum_t^T r(s_t, a_t) + \mathcal{H}\big(\pi(\cdot|s_t)\big)\Big]$$



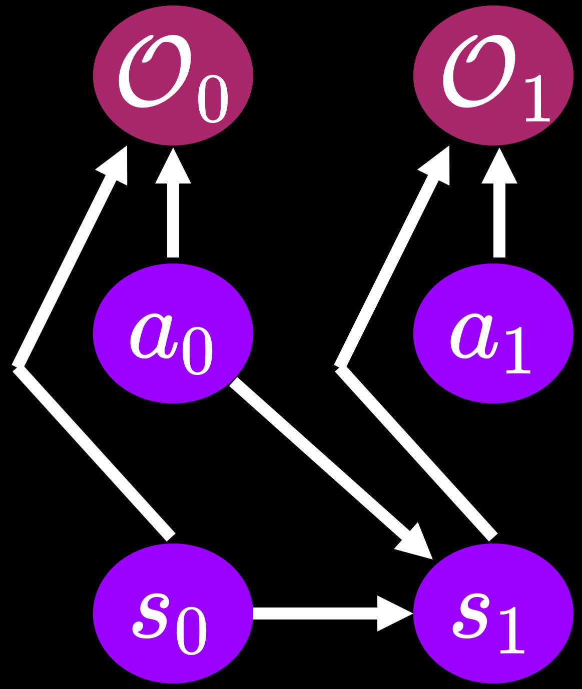


RL as Probabilistic Inference

RL:
Which actions will lead as to the optimal future?

Probabilistic Inference:
Which actions were made given that the future is optimal?

$$p(a_t|s_t\mathcal{O}_{t:T})$$



Optimality:

$$p(\mathcal{O}_t = 1|s_t, a_t) := p(\mathcal{O}_t|s_t, a_t) = \exp(r(s_t, a_t))$$

Exact Probabilistic Inference

Let's find an optimal policy:

$$p(a_t|s_t,\mathcal{O}_{t:T}) = \frac{p(s_t,a_t|\mathcal{O}_{t:T})}{p(s_t|\mathcal{O}_{t:T})} =$$ apply Bayes rule!

$$= \frac{p(\mathcal{O}_{t:T}|s_t,a_t)p(a_t|s_t)p(s_t)}{p(\mathcal{O}_{t:T})} \frac{p(\mathcal{O}_{t:T})}{p(\mathcal{O}_{t:T}|s_t)p(s_t)}$$

where $p(a_t|s_t)$ - prior policy

if $p(a_t|s_t) = \frac{1}{|\mathcal{A}|}$, then

$$p(a_t|s_t,\mathcal{O}_{t:T}) \propto \frac{p(\mathcal{O}_{t:T}|s_t,a_t)}{p(\mathcal{O}_{t:T}|s_t)}$$

Exact Probabilistic Inference

Let's introduce new notation:

$$\alpha_t(s_t, a_t) := p(\mathcal{O}_{t:T} | s_t, a_t)$$

$$\beta_t(s_t) := p(\mathcal{O}_{t:T} | s_t) = \int \alpha_t(s_t, a_t) p(a_t | s_t) da_t$$

We can find all the α_t and β_t via Message Passing algorithm:

For the timestep T :

$$\alpha_T(s_T, a_T) = \exp(r(s_T, a_T))$$

$$\beta_T(s_T) = \int \alpha_T(s_T, a_T) p(a_T | s_T) da_T$$

Recursively:

$$\alpha_t(s_t, a_t) = \int \beta_{t+1}(s_{t+1}) \exp(r(s_t, a_t)) p(s_{t+1} | s_t, a_t) ds_{t+1}$$

$$\beta_t(s_t) = \int \alpha_t(s_t, a_t) p(a_t | s_t) da_t$$

Exact Probabilistic Inference

Let's introduce new notation:

$$\alpha_t(s_t, a_t) := p(\mathcal{O}_{t:T} | s_t, a_t)$$

$$\beta_t(s_t) := p(\mathcal{O}_{t:T} | s_t) = \int \alpha_t(s_t, a_t) p(a_t | s_t) da_t$$

We can find all the α_t and β_t via Message Passing algorithm:

For the timestep T :

$$\alpha_T(s_T, a_T) = \exp(r(s_T, a_T))$$

$$\beta_T(s_T) = \int \alpha_T(s_T, a_T) p(a_T | s_T) da_T$$

Recursively:

$$\alpha_t(s_t, a_t) = \int \beta_{t+1}(s_{t+1}) \exp(r(s_t, a_t)) p(s_{t+1} | s_t, a_t) ds_{t+1}$$

$$\beta_t(s_t) = \int \alpha_t(s_t, a_t) p(a_t | s_t) da_t$$

Soft Q and V functions

We can find analogues in the log-scale:

$$Q^{soft}(s_t, a_t) := \log \alpha_t(s_t, a_t)$$

$$V^{soft}(s_t) := \log \beta_t(s_t)$$

Recursively:

$$V^{soft}(s_t) = \log \mathbb{E}_{p(a_t|s_t)}[\exp Q^{soft}(s_t, a_t)]$$ - soft maximum

$$Q^{soft}(s_t, a_t) = r(s_t, a_t) + \log \mathbb{E}_{p(s_{t+1}|s_t, a_t)}[\exp V^{soft}(s_{t+1})]$$

kinda Bellman equation

Soft and Hard Q and V functions

"Hard" Q and V functions:

$$V^\star(s_t) = \max_{a_t} Q^\star(s_t, a_t)$$

$$Q^\star(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{p(s_{t+1}|s_t,a_t)} V^\star(s_{t+1})$$

$$Q^\star(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{p(s_{t+1}|s_t,a_t)} \max_{a_{t+1}} Q^\star(s_{t+1}, a_{t+1})$$

"Soft" analogues:

$$V^{soft}(s_t) = \log \mathbb{E}_{p(a_t|s_t)}[\exp Q^{soft}(s_t, a_t)]$$

$$Q^{soft}(s_t, a_t) = r(s_t, a_t) + \log \mathbb{E}_{p(s_{t+1}|s_t,a_t)}[\exp V^{soft}(s_{t+1})]$$

$$Q^{soft}(s_t, a_t) \approx r(s_t, a_t) + \max_{s_{t+1}} \max_{a_{t+1}} Q^{soft}(s_{t+1}, a_{t+1})$$

What is being optimized?

Let's analyze an "exact variational inference" procedure:

true conditional $p(\tau|\mathcal{O}_{0:T}) = \frac{p(\tau, \mathcal{O}_{0:T})}{p(\mathcal{O}_{0:T})}$ joint / evidence

$$\arg\min_q \mathrm{KL}\big(q(\tau)||p(\tau|\mathcal{O}_{0:T})\big) =$$
$$\arg\min_{p(a_t|s_t,\mathcal{O}_{t:T})} \mathrm{KL}\big(p(\tau|\mathcal{O}_{0:T})||p(\tau|\mathcal{O}_{0:T})\big)$$
$$= \arg\min_{p(a_t|s_t,\mathcal{O}_{t:T})} \mathrm{KL}\big(p(\tau|\mathcal{O}_{0:T})||p(\tau, \mathcal{O}_{0:T})\big)$$

What is being optimized?

$$\mathrm{KL}\left(p(\tau|\mathcal{O}_{0:T})||p(\tau,\mathcal{O}_{0:T})\right) \to \min_{p(a_t|s_t,\mathcal{O}_{t:T})}$$

where the joint ("exact") distribution is:

$$p(\tau,\mathcal{O}_{0:T}) = p(s_0)\prod_{t=0}^{T} p(a_t|s_t)p(s_{t+1}|s_t,a_t)\exp\left(r(s_t,a_t)\right)$$

and the variational one is:

$$p(\tau|\mathcal{O}_{0:T}) = p(s_0|\mathcal{O}_{0:T})\prod_{t=0}^{T} p(a_t|s_t,\mathcal{O}_{0:T})p(s_{t+1}|s_t,a_t,\mathcal{O}_{0:T})$$

we tried to find a policy which is optimal only in an optimal environment!

We can fix this!

Variational Inference

Minimization problem for VI

$$\mathrm{KL}\Big(q(\tau)||p(\tau, \mathcal{O}_{0:T})\Big) \to \min_q$$

$q(\tau)$ is a distribution over ACHIEVABLE trajectories

The form of the q - is our choice

$$q(\tau) = p(s_0)\prod_{t=0}^{T}\pi(a_t|s_t)p(s_{t+1}|s_t, a_t)$$

fix the dynamics!

Variational Inference

Then:

$$\min_q \mathrm{KL}\big(q(\tau)||p(\tau, \mathcal{O}_{0:T})\big) = -\min_q \mathbb{E}_q \log \frac{p(\tau,\mathcal{O}_{0:T})}{q(\tau)} =$$

$$= \max_q \mathbb{E}_q \Big[\cancel{\log p(s_0)} + \sum_t \big(\cancel{\log p(s_{t+1}|s_t, a_t)} + r(s_t, a_t)\big) -$$
$$- \cancel{\log p(s_0)} - \sum_t \big(\cancel{\log p(s_{t+1}|s_t, a_t)} - \log \pi(a_t|s_t)\big)\Big] =$$

$$= \max_\pi \mathbb{E}_\pi \sum_t \Big[r(s_t, a_t) + \mathcal{H}\big(\pi(\cdot|s_t)\big)\Big]$$

Maximum Entropy RL Objective

Returning to the Q and V functions

This objective can be rewritten as follows:

$$\sum_{t=0}^{T} \mathbb{E}_{s_t}\left[-\mathrm{KL}\left(\pi(a_t|s_t) \middle|\middle| \frac{\exp(Q^{soft}(s_t,a_t))}{\exp(V^{soft}(s_t))}\right) + V^{soft}(s_t)\right] \to \max_\pi$$

check it yourself!

where

$$V^{soft}(s_t) = \log \int \exp Q^{soft}(s_t, a_t) da_t$$ - soft maximum

$$Q^{soft}(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{p(s_{t+1}|s_t,a_t)} V^{soft}(s_{t+1})$$ - normal Bellman equation

Then the optimal policy is: $$\pi(a_t|s_t) = \frac{\exp(Q^{soft}(s_t,a_t))}{\exp(V^{soft}(s_t))}$$

VI with function approximators

(neural nets)

- Maximum Entropy Policy Gradients
- Soft Q-learning

  https://arxiv.org/abs/1702.08165
- Soft Actor-Critic

  https://arxiv.org/abs/1801.01290

Maximum Entropy Policy Gradients

Directly maximize entropy-augmented objective

over policy parameters θ :

$$\mathbb{E}_{\tau\sim\pi_\theta}\sum_{t=0}^{T}\left[r(s_t,a_t)+\mathcal{H}\Big(\pi_\theta(\cdot|s_t)\Big)\right]\to\max_\theta$$

For gradients, use log-derivative trick:

$$\sum_{t=0}^{T}\mathbb{E}_{(s_t,a_t)\sim q_\theta}\left[\nabla_\theta\log\pi_\theta(a_t|s_t)\sum_{t'=t}^{T}\Big(r(s_{t'},a_{t'})-\log\pi_\theta(a_{t'}|s_{t'})-b(s_{t'})\Big)\right]$$

- on-policy
- unimodal policies

Soft Q-learning

Train Q-network with parameters ϕ :

$$\mathbb{E}_{(s_t, a_t, s_{t+1}) \sim \mathcal{D}} \left[Q_\phi^{soft}(s_t, a_t) - \left(r(s_t, a_t) + V_\phi^{soft}(s_{t+1}) \right) \right]^2 \to \min_\phi$$

use replay buffer

where

$$V_\phi^{soft}(s_t) = \log \int \exp Q_\phi^{soft}(s_t, a_t) da_t$$

for continuous actions use importance sampling

Policy is implicit

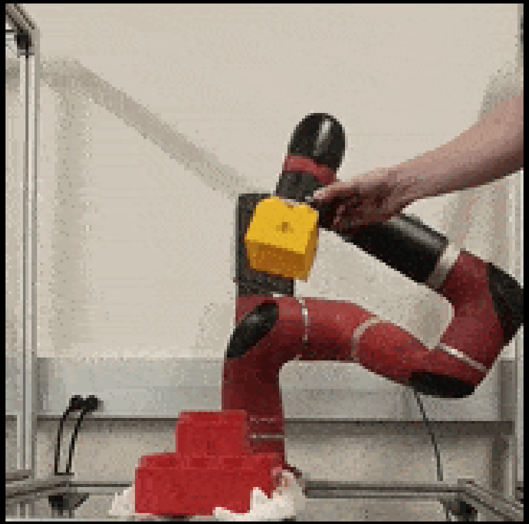
$$\pi(a_t | s_t) = \exp \left(Q_\phi^{soft}(s_t, a_t) - V_\phi^{soft}(s_t) \right)$$

for samples use SVGD or MCMC :D

Soft Q-learning

Exploration

Robustness



Multimodal Policy

Soft Actor-Critic

Train Q- and V-networks jointly with policy

Q-network loss:

$$\mathbb{E}_{(s_t,a_t,s_{t+1})\sim\mathcal{D}}\left[Q_\phi^{soft}(s_t,a_t) - \left(r(s_t,a_t) + V_\psi^{soft}(s_{t+1})\right)\right]^2 \to \min_\phi$$

V-network loss:

$$\mathbb{E}_{s_t\sim\mathcal{D}}\left[\hat{V}^{soft}(s_t) - V_\psi^{soft}(s_t)\right]^2 \to \min_\psi$$

$$\hat{V}^{soft}(s_t) = \mathbb{E}_{a_t\sim\pi_\theta}\left[Q_\phi^{soft}(s_t,a_t) - \log\pi_\theta(a_t|s_t)\right]$$

Objective for the policy:

$$\mathbb{E}_{s_t\sim\mathcal{D},\, a_t\sim\pi_\theta}\left[Q_\phi^{soft}(s_t,a_t) - \log\pi_\theta(a_{t'}|s_t)\right] \to \max_\theta$$

Soft Actor-Critic

Soft Actor-Critic

https://www.youtube.com/embed/KOObeIjzXTY?enablejsapi=1

Thank you for your attention!

and visit our seminars in RL Reading Group

telegram: https://t.me/theoreticalrl

REFERENCES:

Soft Q-learning:

https://arxiv.org/pdf/1702.08165.pdf

Soft Actor Critic:

https://arxiv.org/pdf/1801.01290.pdf

Big Review on Probabilistic Inference for RL:

https://arxiv.org/pdf/1805.00909.pdf

Implementation on TensorFlow:

https://github.com/rail-berkeley/softlearning

Implementation on **Catalyst.RL**:

https://github.com/catalyst-team/catalyst/tree/master/examples/rl_gym

Hierarchical policies (further reading):

https://arxiv.org/abs/1804.02808