Deep Reinforcement Learning amidst Lifelong non-stationarity

Arxiv

Kyeong Ryeol, Go
M.S. Candidate of OSI Lab

Contents

- Levine, Sergey. "Reinforcement learning and control as probabilistic inference: Tutorial and review." *arXiv preprint arXiv:1805.00909* (2018).

- Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor." *International Conference on Machine Learning*. PMLR, 2018.

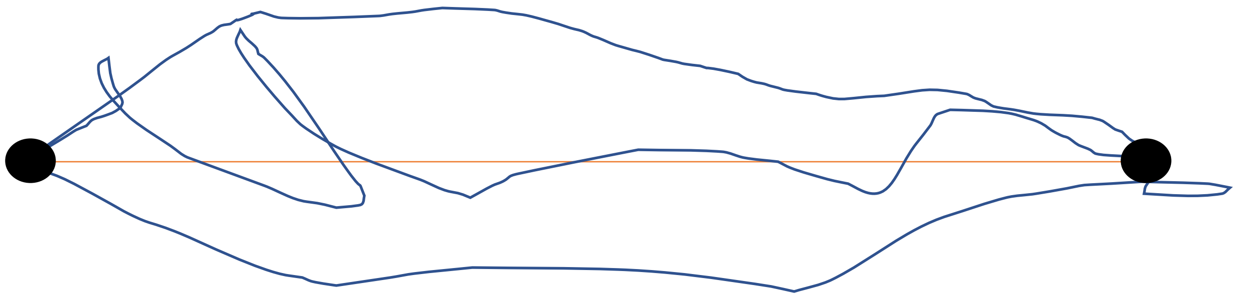
- Lee, Alex X., et al. "Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model." *arXiv preprint arXiv:1907.00953* (2019).

- Xie, Annie, James Harrison, and Chelsea Finn. "Deep reinforcement learning amidst lifelong non-stationarity." *arXiv preprint arXiv:2006.10701* (2020).

Reinforcement learning

- Typical reinforcement learning
  - $\theta = argmax_\theta \sum_t E_{s_t, a_t \sim p(\tau)}[r(s_t, a_t)]$
  - $p(\tau) = p(s_1, a_1, .., s_T, a_T|\theta) = p(s_1) \prod_t p_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t)$

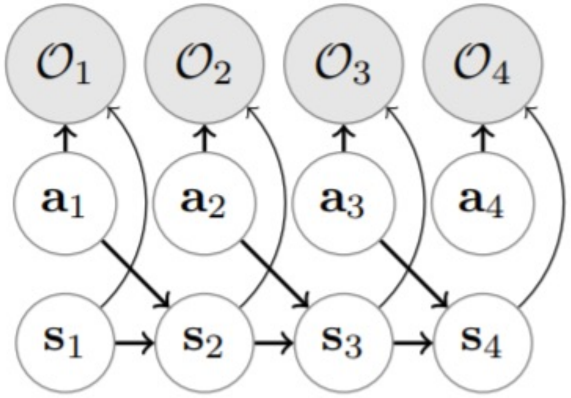
- Most probable trajectory ≈ Trajectory from the optimal policy



Optimality

- Formulate PGM s.t. Inferring posterior ≈ optimal policy
- $p(\tau, O_{1:T} = \mathbf{1})$

$$
\begin{aligned}
&= p(s_1) \prod_t p(O_t = 1 | s_t, a_t) p(s_{t+1} | s_t, a_t) \\
&= p(s_1) \prod_t \exp\big(r(s_t, a_t)\big)\, p(s_{t+1} | s_t, a_t) \\
&= p(s_1) \prod_t p(s_{t+1} | s_t, a_t) \exp\left(\sum_t r(s_t, a_t)\right)
\end{aligned}
$$



Approximate inference

- Variational distribution
  - $q(\tau) = p(s_1) \prod_t p(s_{t+1}|s_t, a_t)\pi_\phi(a_t|s_t)$

- Deriving ELBO
  - $\log p(O_{1:T} = 1) \geq E_{q(\tau)}[\log p(\tau, O_{1:T} = 1) - \log q(\tau)]$
    $$= E_{q(\tau)}\left[\sum_t r(s_t, a_t) - \log \pi_\phi(a_t|s_t)\right]$$
  - Initial state marginal and transition dynamics cancel out
  - Suffice to maximum entropy reinforcement learning

Soft Actor Critic (SAC)

- Goal : devise efficient and stable actor-critic deep RL
  - Baseline : (TRPO, PPO, A3C), (DDPG)
  - Based on maximum entropy reinforcement learning

- Soft policy iteration

  1. Soft policy evaluation
  - $V(s_t) \leftarrow E_{a_t}[Q(s_t, a_t) - \log \pi(a_t|s_t)]$
  - $Q(s_t, a_t) \leftarrow r(s_t, a_t) + \gamma E_{s_{t+1}}[V(s_{t+1})]$

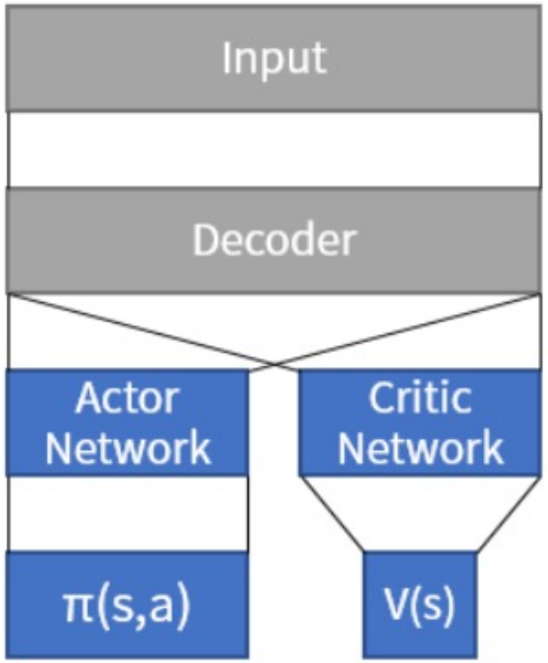
  2. Soft policy improvement
  - $\pi_{new} = argmin_{\pi}\, D_{KL}\left(\pi(\cdot\,|s_t) \parallel \frac{\exp Q^{\pi}(s_t, \cdot)}{Z(s_t)}\right)$

Lemma 1 (Soft Policy Evaluation). *Consider the soft Bellman backup operator $\mathcal{T}^{\pi}$ in Equation 2 and a mapping $Q^0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with $|\mathcal{A}| < \infty$, and define $Q^{k+1} = \mathcal{T}^{\pi} Q^k$. Then the sequence Q^k will converge to the soft Q-value of π as $k \to \infty$.*

Lemma 2 (Soft Policy Improvement). *Let $\pi_{\mathrm{old}} \in \Pi$ and let π_{new} be the optimizer of the minimization problem defined in Equation 4. Then $Q^{\pi_{\mathrm{new}}}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi_{\mathrm{old}}}(\mathbf{s}_t, \mathbf{a}_t)$ for all $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$ with $|\mathcal{A}| < \infty$.*

Soft Actor Critic (SAC)

- Recap
  - Policy based
    - $\phi \leftarrow \phi + \alpha \nabla_{\phi} E_{p(\tau)}[\sum_t r(s_t, a_t)]$
  - Value based
    1. $Q(s_t, a_t) \leftarrow r(s_t, a_t) + \gamma E_{s_{t+1}}[V(s_{t+1})]$
    2. $\pi(a_t|s_t) = 1 \; when \; a_t = argmax_{a_t} Q(s_t, a_t)$
  - Actor critic
    1. $Q(s_t, a_t) \leftarrow r(s_t, a_t) + \gamma E_{s_{t+1}}[V(s_{t+1})]$
    2. $\phi \leftarrow \phi + \alpha \nabla_{\phi} E_{p(\tau)}[Q(s_t, a_t)]$



Overall training process

- Function approximator for both the Q-function and the policy
  - $V_\psi(s_t), Q_\theta(s_t, a_t), \pi_\phi(a_t|s_t)$

- Alternate between the networks with stochastic gradient descent
  - $J_V(\psi) = E_{s_t}\left[\frac{1}{2}\left(V_\psi(s_t) - E_{a_t}\left[Q_\theta(s_t, a_t) - \log \pi_\phi(a_t|s_t)\right]\right)^2\right]$
  - $J_Q(\theta) = E_{s_t,a_t}\left[\frac{1}{2}\left(Q_\theta(s_t, a_t) - r(s_t, a_t) - \gamma E_{s_{t+1}}\left[V_{\tilde{\psi}}(s_{t+1})\right]\right)\right]$
  - $J_\pi(\phi) = E_{s_t,a_t}\left[D_{KL}\left(\pi_\phi(a_t|s_t) \parallel \frac{\exp Q_\theta(s_t,a_t)}{Z(s_t)}\right)\right]$
    $\approx E_{s_t,a_t}\left[\log \pi_\phi(a_t|s_t) - Q_\theta(s_t, a_t)\right]$

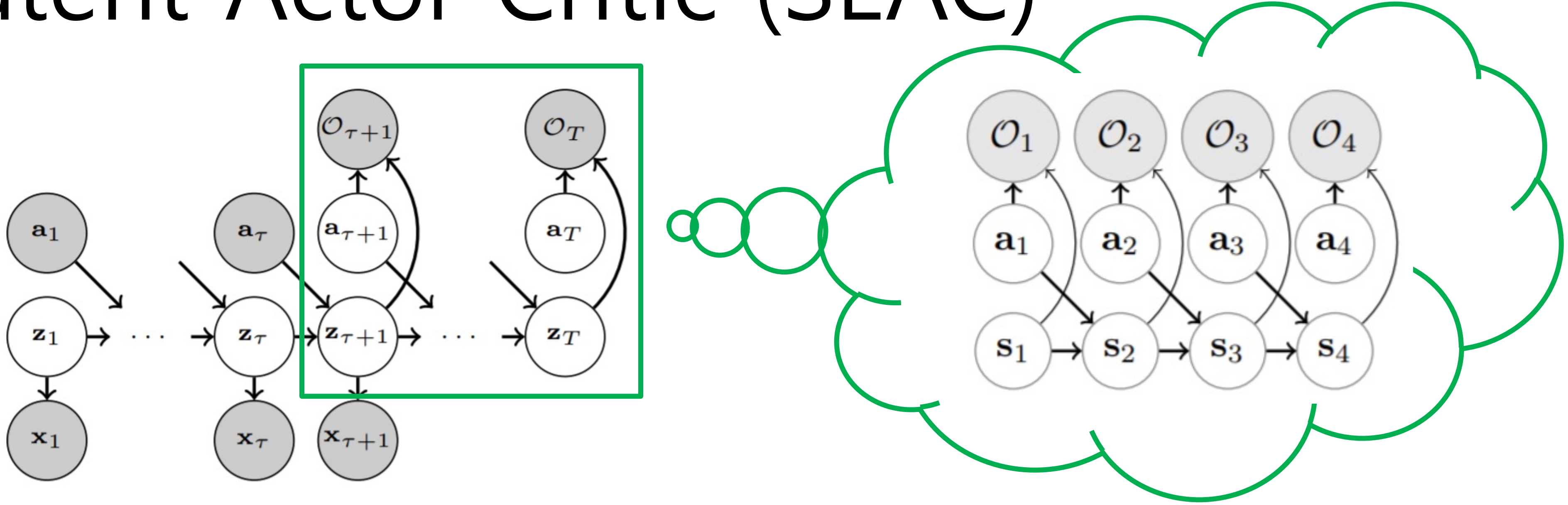
Soft Latent Actor Critic (SLAC)

- Goal : devise efficient and stable actor-critic deep RL with high dim.
  1. Acquire the explicit latent representations
  2. Train RL agent in that latent space

  $z \rightarrow x$: Work on low-dim. latent space + Handle partial observability

- Sequential latent variable model
  - Variational distribution
    - $q(z_1|x_1)$ for $p(z_1)$ and $q(z_{t+1}|x_{t+1}, z_t, a_t)$ for $p(z_{t+1}|z_t, a_t)$
  - $\log p(x_{1:\tau+1}|a_{1:\tau}) \geq$
    $E_{z_{1:\tau+1}}[\sum_{t=0}^{\tau} \log p(x_{t+1}|z_{t+1}) - D_{KL}(q(z_{t+1}|x_{t+1}, z_t, a_t) \parallel p(z_{t+1}|z_t, a_t))]$

Soft Latent Actor Critic (SLAC)



- Augment the idea with maximum entropy RL? Optimality!
- Evidence : $\log p(O_{1:T} = 1) \;\; \rightarrow \;\; \log p(O_{\tau+1:T} = 1, x_{1:\tau+1} | a_{1:\tau})$
  - Likelihood of the observed data from the past $\tau + 1$ steps
  - Optimality of the agent's actions for future steps
  - Enable joint learning of "representation learning" and "optimal control"

Soft Latent Actor Critic (SLAC)

- Variational distribution
  - $q(z_{1:T}, a_{\tau+1:T} | x_{1:\tau+1}, a_{1:\tau}) = \prod_{t=0}^{\tau} q(z_{t+1} | x_{t+1}, z_t, a_t) \prod_{t=\tau+1}^{T-1} p(z_{t+1} | z_t, a_t) \prod_{t=\tau+1}^{T} \pi(a_t | x_{1:t}, a_{1:t-1})$

- Deriving ELBO
  - $\log p(O_{\tau+1:T} = 1, x_{1:\tau+1} | a_{1:\tau})$

$$\geq E_{z_{1:\tau+1}}\left[\sum_{t=0}^{\tau} \log p(x_{t+1}|z_{t+1}) - D_{KL}(q(z_{t+1}|x_{t+1}, z_t, a_t) \parallel p(z_{t+1}|z_t, a_t))\right]$$

$$+ E_{z_{\tau+1:T}, a_{\tau+1:T}}\left[\sum_{t=\tau+1}^{T} r(z_t, a_t) - \log \pi(a_t | x_{1:t}, a_{1:t-1})\right]$$

Overall training process

- Function approximator for both the Q-function and the policy
  - $\left(p_\psi(x_{t+1}|z_{t+1}), p_\psi(z_{t+1}|z_t, a_t), q_\psi(z_{t+1}|x_{t+1}, z_t, a_t)\right), Q_\theta(s_t, a_t), \pi_\phi(a_t|s_t)$

- Alternate between the networks with stochastic gradient descent
  - $J_M(\psi) = E_{z_{1:\tau+1}}\left[\sum_{t=0}^{\tau} - \log p_\psi(x_{t+1}|z_{t+1}) + D_{KL}\left(q_\psi(z_{t+1}|x_{t+1}, z_t, a_t\right) \parallel p_\psi(z_{t+1}|z_t, a_t))\right]$
  - $J_Q(\theta) = E_{z_t,a_t}\left[\frac{1}{2}\left(Q_\theta(z_t, a_t) - r(z_t, a_t) - \gamma E_{z_{t+1}}\left[V_{\tilde{\theta}}(z_{t+1})\right]\right)\right]$
    - $V_\theta(z_{t+1}) = E_{a_{t+1}}\left[Q_\theta(z_{t+1}, a_{t+1}) - \log \pi_\phi(a_{\tau+1}|x_{1:\tau+1}, a_{1:\tau})\right]$
  - $J_\pi(\phi) = E_{z_{1:\tau+1},a_{\tau+1}}\left[D_{KL}\left(\pi_\phi(a_{\tau+1}|x_{1:\tau+1}, a_{1:\tau}) \parallel \frac{\exp Q_\theta(z_{\tau+1}, a_{\tau+1})}{Z(z_{\tau+1})}\right)\right]$
    $\approx E_{z_{1:\tau+1},a_{\tau+1}}\left[\log \pi_\phi(a_{\tau+1}|x_{1:\tau+1}, a_{1:\tau}) - Q_\theta(z_{\tau+1}, a_{\tau+1})\right]$