Fraunhofer IIS

Fraunhofer-Institut für Integrierte Schaltungen IIS

Reinforcement Learning

—

Exercise 12: Offline Reinforcement Learning

Nico Meyer

Overview

Exercise Content

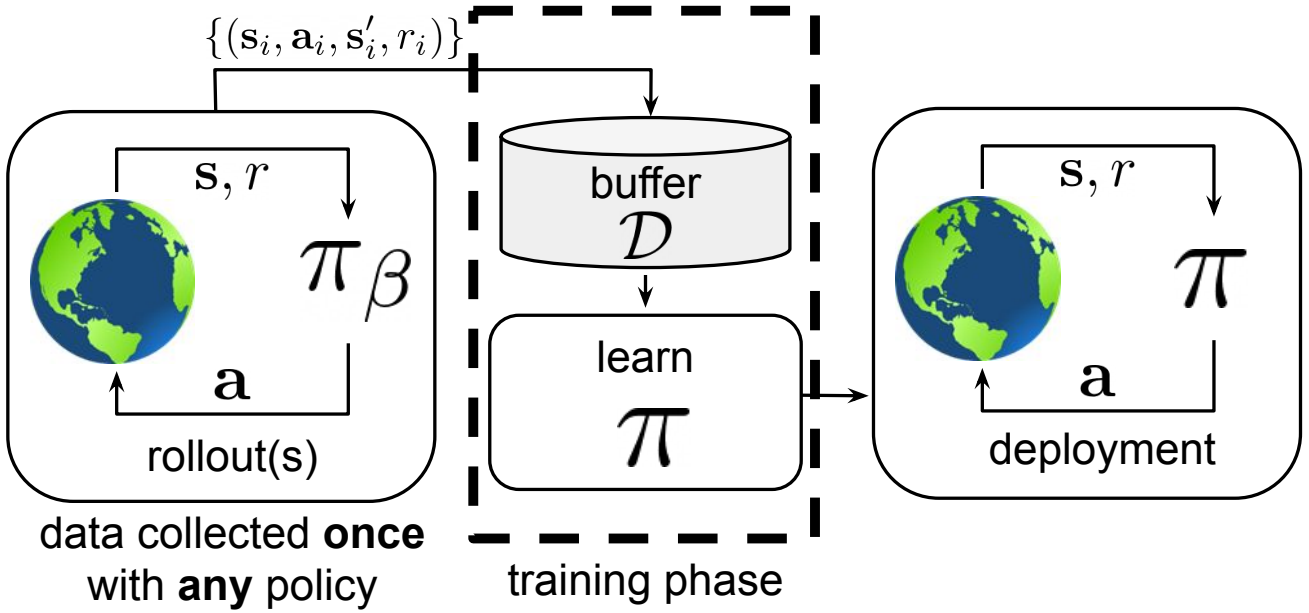
Week	*Date*	*Topic*	*Material*	*Who?*
0			*no exercises*	
1	23.04.	MDPs		Nico
2	30.04.	Dynamic Programming		Alex
3	07.05.	OpenAI Gym, PyTorch-Intro		Alex
4	14.05.	TD-Learning		Nico
5	22.05.	Practical Session (zoom@home)	**Attention: Lecture Slot!**	Nico + Alex
6	28.05.	TD-Control		Nico
7	04.06.	DQN		Nico
8	11.06.	VPG		Alex
9	18.06.	A2C		Nico
10	25.06.	Multi-armed Bandits		Alex
11	02.07.	RND/ICM		Alex
12	09.07.	MCTS		Alex
13	16.07.	BCQ		Nico

Offline Reinforcement Learning

Discrete Batch-Constraint Q-Learning (BCQ)

Offline Reinforcement Learning

Fully Offline Reinforcement Learning



- Offline RL uses a dataset $\mathcal{D}$ collected by some behavior policy π_β
  - π_β is potentially (or often assumed to be) unknown
- $\mathcal{D}$ is **collected once** and **not changed** during training
  - Transitions are sampled from $\mathcal{D}$
  - No interaction with the MDP; Policy is deployed after being fully trained.
- Policy and transitions are independent

Levine et al.: "Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems"

Offline Reinforcement Learning

Offline Policy Evaluation and Distribution Shift

What we want to do:

$$\mathbb{E}_{s\sim d_\pi,\, a\sim\pi}\left[\left(r+\gamma\mathbb{E}_{s'\sim p(s'|s,a),\, a'\sim\pi(s')}[Q_\theta(s',a')]-Q_\theta(s,a)\right)^2\right]$$

What we would naively do:

$$\mathbb{E}_{s\sim d_{\pi_\beta},\, a\sim\pi_\beta}\left[\left(r+\gamma\mathbb{E}_{s'\sim p(s'|s,a),\, a'\sim\pi(s')}[Q_\theta(s',a')]-Q_\theta(s,a)\right)^2\right]$$

- **State distribution shift:**
  - Problem arises during **test time**
  - Does not invalidate the learned strategy on the states in $\mathcal{D}$ because unobserved states are never queried during training
- **Action distribution shift:**
  - Already problematic during **training** as inaccurate action values are used as bootstrapped targets
  - Can invalidate the learned strategy even on states in $\mathcal{D}$

Offline Reinforcement Learning

Policy-Constrained Methods

- The problems arise because the maximizing action is selected without uncertainty considerations

$$\pi_{new}(s) = \arg\max_a Q_\theta(s, a)$$

- Define the admissible set of policies $\Pi_\epsilon = \{\pi \mid d(\pi, \pi_\beta) \leq \epsilon)$ where d is a distance measure

- Consider a constrained policy improvement step

$$\pi_{new} = \arg\max_{\pi \in \Pi_\epsilon} \mathbb{E}[Q_\theta(s, \pi(s))]$$

Policy-Constrained Offline RL

BCQ with Function Approximation – Discrete Case

- Q-Learning:

$$\min_{\theta}(r + \gamma \max_{a' \in A(s')} Q_\theta(s', a') - Q_\theta(s, a))^2$$

- Let us define

$$A_\epsilon^{BCQ}(s) = \left\{ a \in A(s): \frac{\hat{\pi}_\beta(a|s)}{\max_a \hat{\pi}_\beta(a|s)} \geq \epsilon \right\},$$

  where $\epsilon \in [0,1]$ is the threshold parameter and $\hat{\pi}_\beta$ an estimate for the behaviour policy

- The **constrained target** is $y = r + \gamma \cdot \max_{a' \in A_\epsilon^{BCQ}(s')} Q\,(s', a')$
  - $\epsilon = 1$ → behavioural cloning
  - $\epsilon = 0$ → Q-Learning

- The learned policy is $\pi(s) = \arg\max_{a \in A_\epsilon^{BCQ}(s')} Q(s, a)$

Policy-Constrained Offline RL

Discrete BCQ

Algorithm 1 BCQ

Input: Batch $\mathcal{B}$, number of iterations T, target_update_rate, mini-batch size N, threshold τ.
Initialize Q-network Q_θ, generative model G_ω and target network $Q_{\theta'}$ with $\theta' \leftarrow \theta$.
for $t = 1$ **to** T **do**
    Sample mini-batch M of N transitions (s, a, r, s') from $\mathcal{B}$.
    $a' = \text{argmax}_{a'|G_\omega(a'|s')/\max \hat{a}\ G_\omega(\hat{a}|s')>\tau} Q_\theta(s', a')$
    $\theta \leftarrow \text{argmin}_\theta \sum_{(s,a,r,s')\in M} l_\kappa \left(r + \gamma Q_{\theta'}(s', a') - Q_\theta(s, a)\right)$
    $\omega \leftarrow \text{argmin}_\omega - \sum_{(s,a)\in M} \log G_\omega(a|s)$
    If t mod target_update_rate $= 0$: $\theta' \leftarrow \theta$
end for

Exercise Sheet 12

Discrete BCQ

Fraunhofer
IIS

Fraunhofer-Institut für Integrierte Schaltungen IIS

Thank you for your attention!