强化学习2022
第11节

涉及知识点：
离线强化学习

离线强化学习

张伟楠 – 上海交通大学

课程大纲

强化学习基础部分

1. 强化学习、探索与利用
2. MDP和动态规划
3. 值函数估计
4. 无模型控制方法
5. 规划与学习
6. 参数化的值函数和策略
7. 深度强化学习价值方法
8. 深度强化学习策略方法

强化学习前沿部分

9. 基于模型的深度强化学习
10. 模仿学习
11. 离线强化学习
12. 参数化动作空间
13. 目标导向的强化学习
14. 多智能体强化学习
15. 强化学习大模型
16. 技术交流与回顾

目录

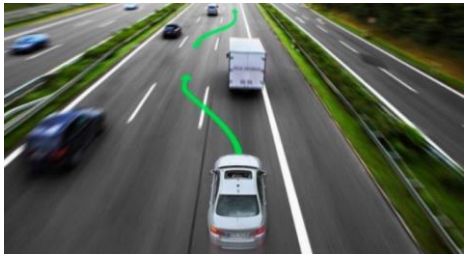
Contents

01

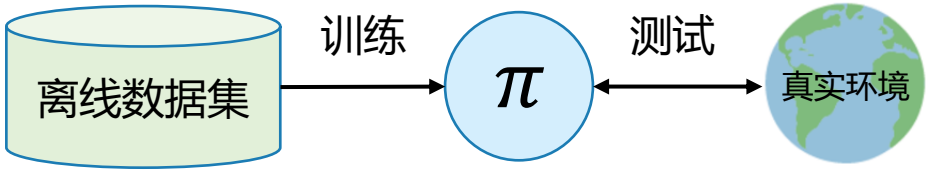
离线强化学习

离线强化学习

- 动机：在真实环境中从零开始训练一个强化学习智能体往往不可取
  - 风险较高，例如无人驾驶归控、智能医疗等
  - 十分昂贵，例如机器人控制、推荐系统等



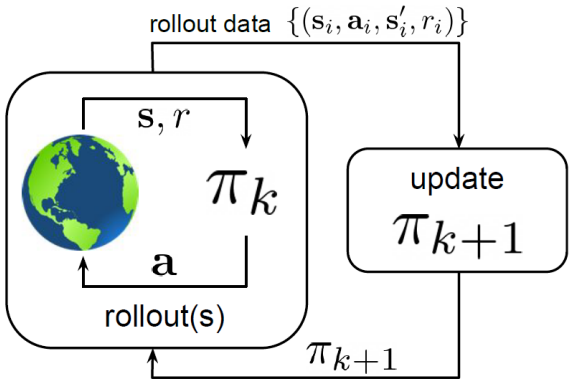
- 离线强化学习：在一个给定的离线数据集上直接训练出智能体策略，训练的过程中，智能体不得和环境做交互



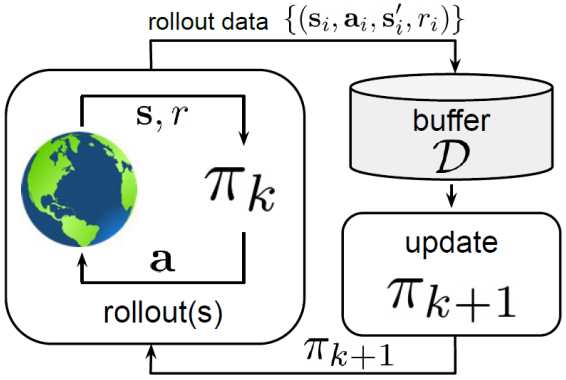
- 离线强化学习有潜力大大扩宽强化学习落地的范围

离线强化学习的不同

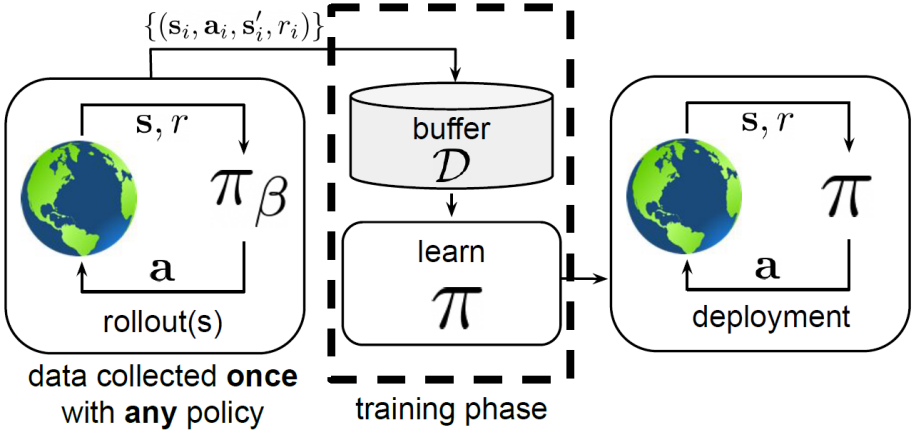
在线策略学习
(a) online reinforcement learning



离线策略学习
(b) off-policy reinforcement learning



离线强化学习
(c) offline reinforcement learning



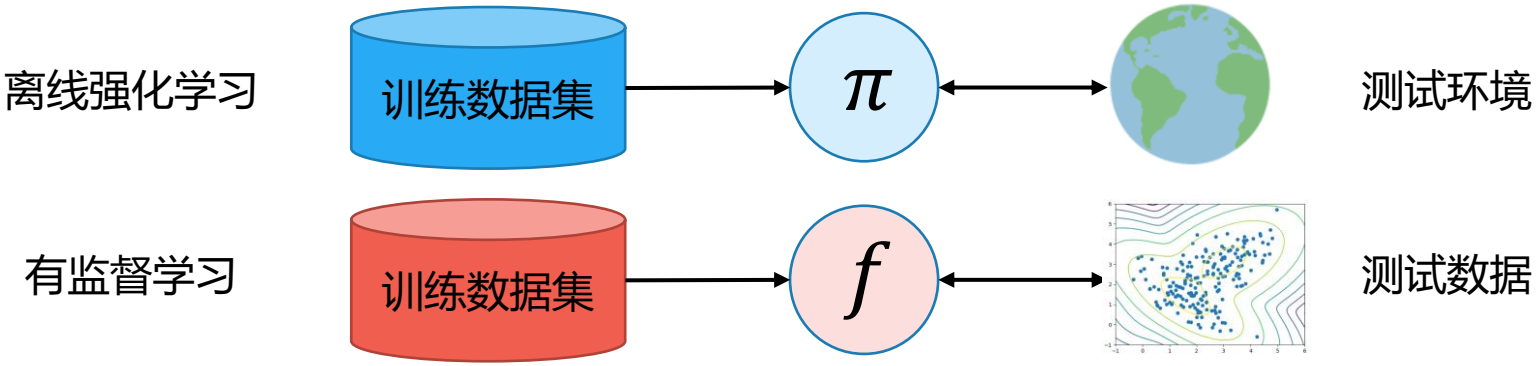
- 训练的过程中与环境交互：
  - 在线策略学习与离线策略学习的智能体可以和环境交互
  - 离线强化学习的智能体不得和环境做交互
- 训练数据是否来自别的策略交互经验：
  - Yes - 离线强化学习和离线策略学习
  - No – 在线强化学习

Levine, Sergey, et al. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv 2005.01643 (2020).

离线强化学习的优势

- 离线强化学习在以下方面带来好处
  - 基于一个已有经验数据集，预训练一个强化学习策略
  - 基于一个已有经验数据集，经验性地评测一个策略的好坏
  - 缩小学术界对强化学习的研究工作和真实世界中的落地应用的差距

- 离线强化学习让强化学习更像有监督学习



Gulcehre, Caglar, et al. RL unplugged: Benchmarks for offline reinforcement learning. arXiv preprint arXiv:2006.13888 (2020).

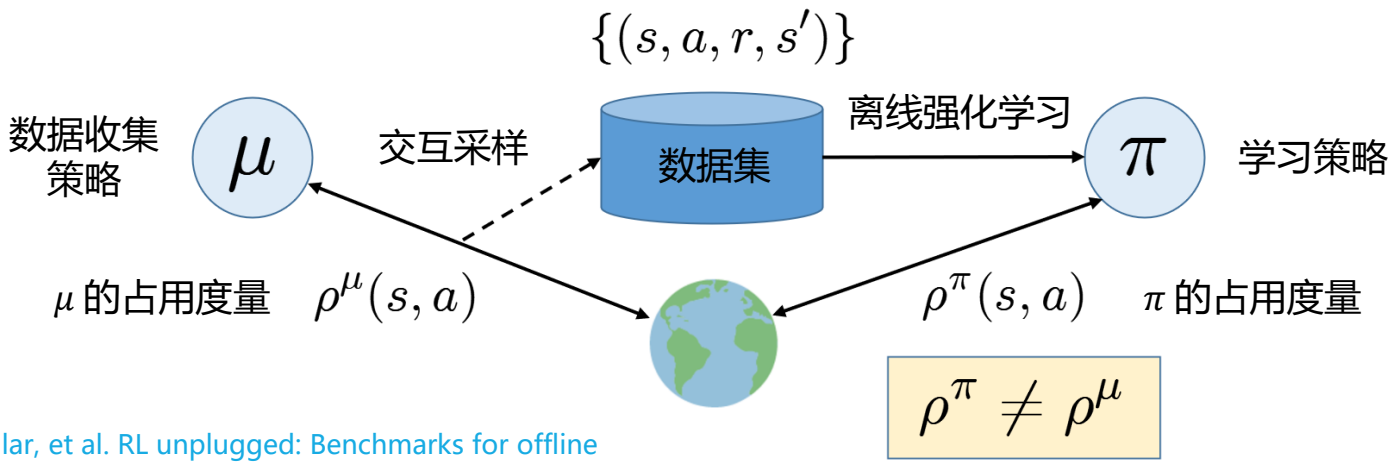
离线强化学习的主要科学问题和方法概览

- 离线强化学习面临的最重要的挑战是外延误差（Extrapolation Error）
  - 也即是处理分布外（out-of-distribution，OOD）问题
  - 智能体如果涉足到了从没有见过的、远离数据集的状态动作对，怎么办？

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a'))$$

如果a'是一个分布外的动作，怎么处理？

- 外延误差主要是由于数据集分布和当前策略的占用度量不一致导致的



Gulcehre, Caglar, et al. RL unplugged: Benchmarks for offline reinforcement learning. arXiv preprint arXiv:2006.13888 (2020).

离线强化学习的主要科学问题和方法概览

- 离线强化学习面临的最重要的挑战是外延误差（Extrapolation Error）
  - 也即是处理分布外（out-of-distribution，OOD）问题
  - 智能体如果涉足到了从没有见过的、远离数据集的状态动作对，怎么办？

- 离线强化学习的主要方法在于设计训练中的限制，从而避免分布外问题，可以大致分为无模型的方法和基于模型的方法

无模型的方法

- 显式限制
  - BCQ
  - BEAR
  - BRAC
  - CQL
- 隐式限制
  - AWR
  - REM
  - BAIL
  - …

基于模型的方法

- 使用学习的模型估计不确定性
  - MOReI
  - MOPO
  - COMBO
  - …

Gulcehre, Caglar, et al. RL unplugged: Benchmarks for offline reinforcement learning. arXiv preprint arXiv:2006.13888 (2020).

02
BCQ算法

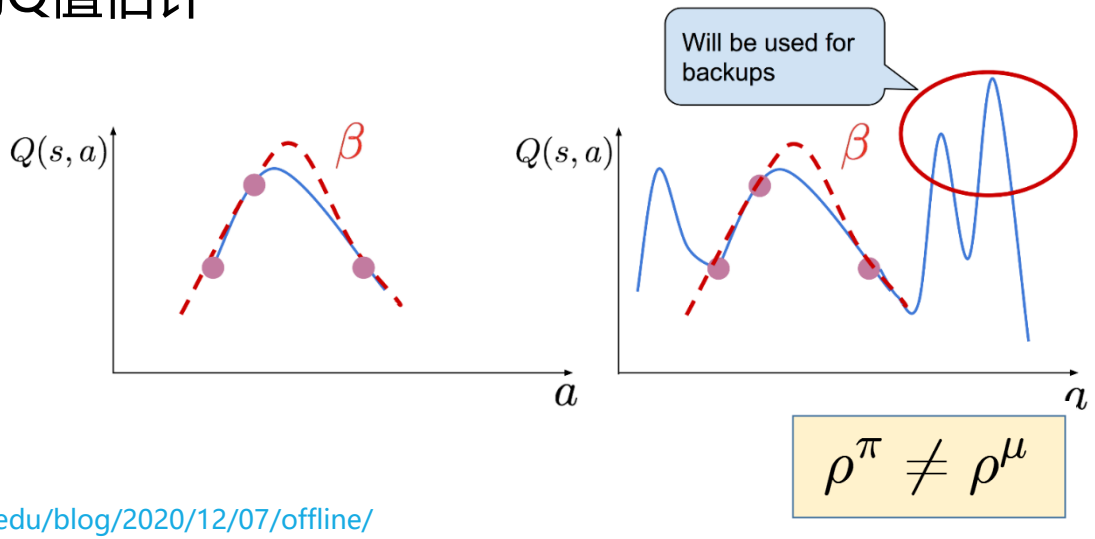
Q学习中的外延误差

- 离线强化学习面临的最重要的挑战是外延误差（Extrapolation Error）
  - 也即是处理分布外（out-of-distribution，OOD）问题
  - 智能体如果涉足到了从没有见过的、远离数据集的状态动作对，怎么办？

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a'))$$

如果a'是一个分布外的动作，怎么处理？

- 外延误差会随着时序差分公式传播到非OOD数据上的Q值估计

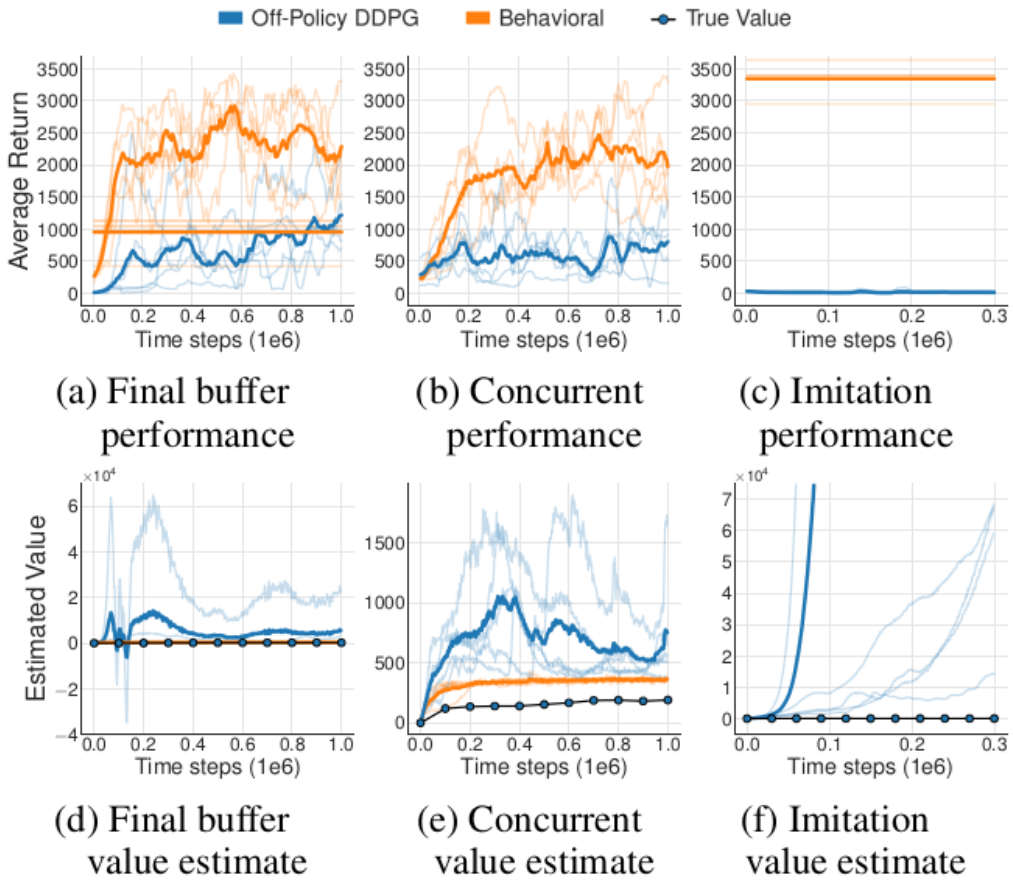


https://bair.berkeley.edu/blog/2020/12/07/offline/

BCQ：批量限制Q学习

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a'))$$

如果a'是一个分布外的动作，怎么处理?



- 外延误差

$$\rho^{\pi} \neq \rho^{\mu}$$

- 因为外延误差，甚至离线策略学习（off-policy）方法也会失效

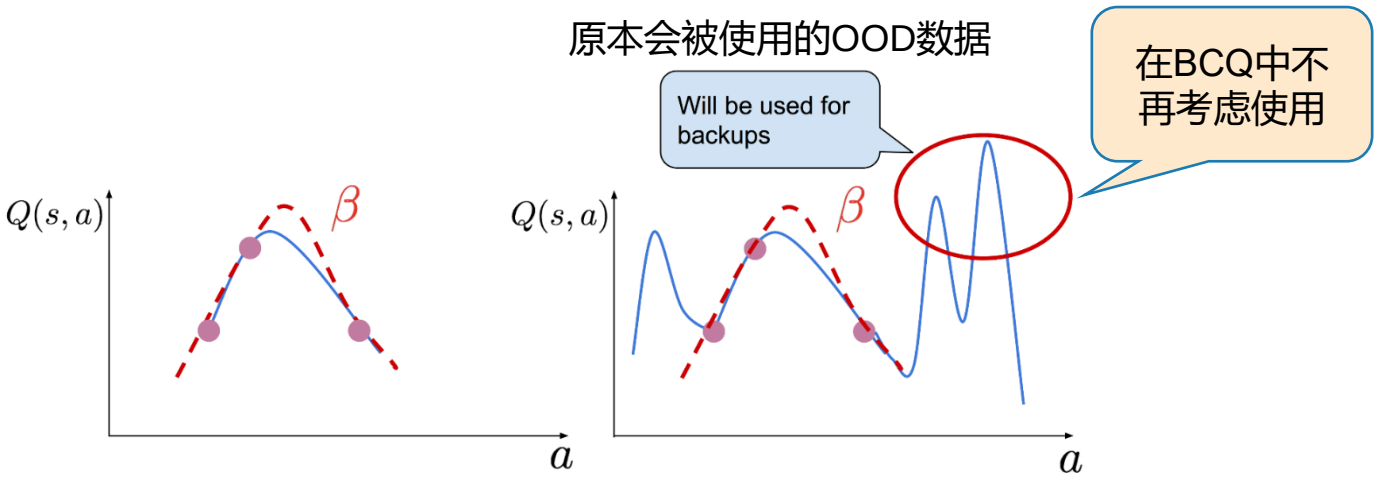
https://bair.berkeley.edu/blog/2020/12/07/offline/

BCQ：批量限制Q学习

- 对于经典表格型强化学习（Tabular RL Setting），BCQ的基本思路：仅仅使用在数据集支撑上的目标Q值做时序差分的计算

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a' \text{ s.t. } (s',a') \in \mathcal{B}} Q(s',a'))$$

仅仅考虑在数据集支撑上的(s', a')



https://bair.berkeley.edu/blog/2020/12/07/offline/

BCQ：批量限制Q学习

- 对于更广泛的连续动作强化学习设置，BCQ的基本思路“仅仅使用在数据集支撑上的目标Q值做时序差分的计算”可以如下实现：
  - 使用一个生成模型，如变分自动编码器VAE，来生成距离数据集较近的状态动作对

$$\pi(s) = \underset{a_i+\xi_\phi(s,a_i,\Phi)}{\arg\max} Q_\theta(s, a_i + \xi_\phi(s, a_i, \Phi))$$

在$[-\Phi, +\Phi]$的扰动

$$\text{where } \{a_i \sim G_\omega(s)\}_{i=1}^n$$

生成模型，如变分自动编码器VAE

  - 对于n和Φ的选择，形成了模仿学习和强化学习之间的一个权衡
    - n和Φ越小，越接近模仿学习，策略性能可能不好
    - n和Φ越大，越接近强化学习，但容易出OOD问题

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. ICML 2019.

BCQ：批量限制Q学习

Input: Batch $\mathcal{B}$, horizon T, target network update rate τ, mini-batch size N, max perturbation Φ, number of sampled actions n, minimum weighting λ.
Initialize Q-networks $Q_{\theta_1}, Q_{\theta_2}$, perturbation network ξ_ϕ, and VAE $G_\omega = \{E_{\omega_1}, D_{\omega_2}\}$, with random parameters θ_1, θ_2, ϕ, ω, and target networks $Q_{\theta'_1}, Q_{\theta'_2}$, $\xi_{\phi'}$ with $\theta'_1 \leftarrow \theta_1, \theta'_2 \leftarrow \theta_2, \phi' \leftarrow \phi$.
for $t = 1$ **to** T **do**
  Sample mini-batch of N transitions (s, a, r, s') from $\mathcal{B}$
  $\mu, \sigma = E_{\omega_1}(s, a), \quad \tilde{a} = D_{\omega_2}(s, z), \quad z \sim \mathcal{N}(\mu, \sigma)$
  $\omega \leftarrow \text{argmin}_\omega \sum (a - \tilde{a})^2 + D_{\text{KL}}(\mathcal{N}(\mu, \sigma) || \mathcal{N}(0, 1))$
  Sample n actions: $\{a_i \sim G_\omega(s')\}_{i=1}^n$
  Perturb each action: $\{a_i = a_i + \xi_\phi(s', a_i, \Phi)\}_{i=1}^n$
  Set value target y (Eqn. 13)
  $r + \gamma \max_{a_i} \left[\lambda \min_{j=1,2} Q_{\theta'_j}(s', a_i) + (1 - \lambda) \max_{j=1,2} Q_{\theta'_j}(s', a_i) \right]$
  $\theta \leftarrow \text{argmin}_\theta \sum (y - Q_\theta(s, a))^2$
  $\phi \leftarrow \text{argmax}_\phi \sum Q_{\theta_1}(s, a + \xi_\phi(s, a, \Phi)), a \sim G_\omega(s)$
  Update target networks: $\theta'_i \leftarrow \tau\theta + (1 - \tau)\theta'_i$
  $\phi' \leftarrow \tau\phi + (1 - \tau)\phi'$
end for

- BCQ方法（连续状态和动作的版本）

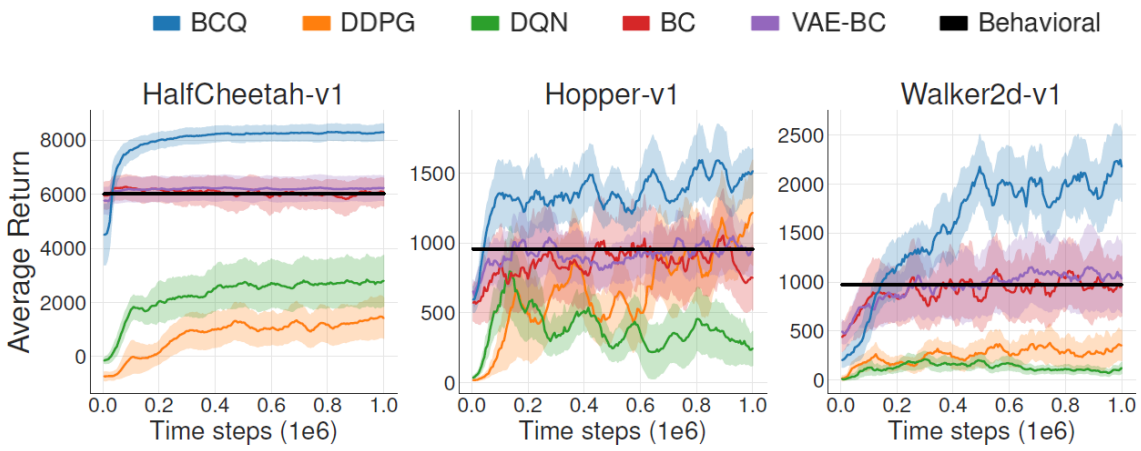
← VAE做模仿学习

← 乐观与保守估计之间的平衡

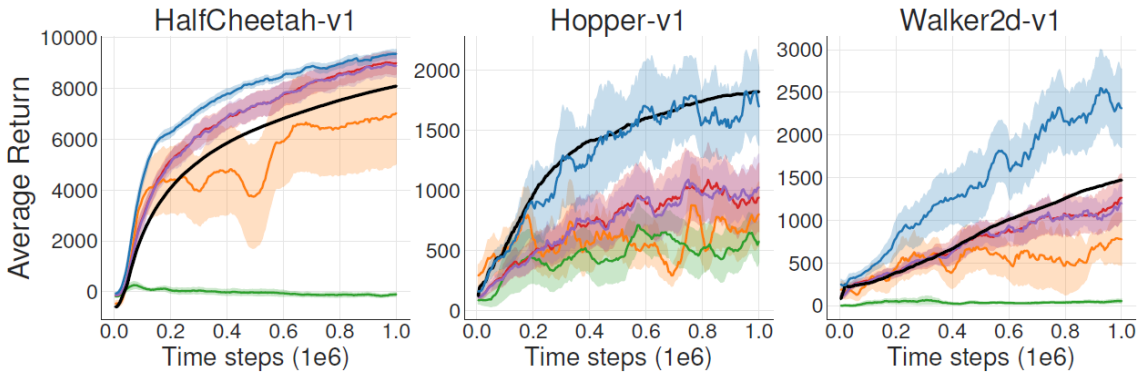
← 扰动函数ξ像是actor

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. ICML 2019.

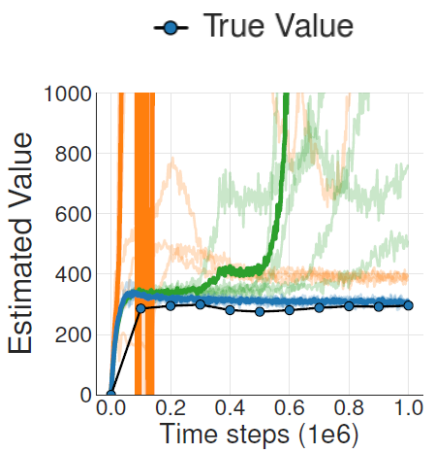
BCQ：批量限制Q学习的实验效果



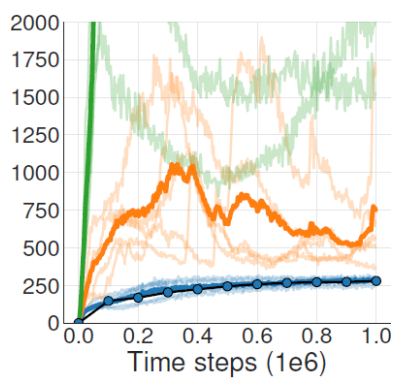
(a) Final buffer performance



(b) Concurrent performance



Full buffer上的价值估计



Concurrent上的价值估计

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. ICML 2019.

03
CQL算法

CQL：保守Q学习

- 思路：学习一个保守的、可作为价值下界的Q函数，以避免在OOD数据上的过高估计
- 于是，对于一个新的学习策略μ，需要增加一个其遇见数据上的Q函数的惩罚

排除在数据支撑范围内的Q价值惩罚

$$\hat{Q}^{k+1} \leftarrow \arg\min_{Q} \; \alpha \cdot \left(\mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\mu(\mathbf{a}|\mathbf{s})}\left[Q(\mathbf{s},\mathbf{a})\right] - \mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\hat{\pi}_\beta(\mathbf{a}|\mathbf{s})}\left[Q(\mathbf{s},\mathbf{a})\right]\right)$$
$$+\frac{1}{2}\,\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}}\left[\left(Q(\mathbf{s},\mathbf{a}) - \hat{\mathcal{B}}^{\pi}\hat{Q}^{k}(\mathbf{s},\mathbf{a})\right)^2\right]$$

新的学习策略μ的数据Q价值

仅仅在一个(s,a,s')上计算时序差分目标

- 使用max μ操作来估计当前的学习策略π，为了增加覆盖度，加上正则

$$\min_{Q}\max_{\mu}\; \alpha\left(\mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\mu(\mathbf{a}|\mathbf{s})}\left[Q(\mathbf{s},\mathbf{a})\right] - \mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\hat{\pi}_\beta(\mathbf{a}|\mathbf{s})}\left[Q(\mathbf{s},\mathbf{a})\right]\right)$$
$$+\frac{1}{2}\,\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}}\left[\left(Q(\mathbf{s},\mathbf{a}) - \hat{\mathcal{B}}^{\pi_k}\hat{Q}^{k}(\mathbf{s},\mathbf{a})\right)^2\right] + \mathcal{R}(\mu) \quad (\mathrm{CQL}(\mathcal{R}))$$

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine.
Conservative q-learning for offline reinforcement learning. NeuIPS 2020.

https://sites.google.com/view/cql-offline-rl

CQL：保守Q学习

$$\min_Q \max_\mu \alpha \left(\mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\mu(\mathbf{a}|\mathbf{s})} \left[Q(\mathbf{s},\mathbf{a}) \right] - \mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\hat{\pi}_\beta(\mathbf{a}|\mathbf{s})} \left[Q(\mathbf{s},\mathbf{a}) \right] \right)$$

$$+ \frac{1}{2} \mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}} \left[\left(Q(\mathbf{s},\mathbf{a}) - \hat{\mathcal{B}}^{\pi_k} \hat{Q}^k(\mathbf{s},\mathbf{a}) \right)^2 \right] + \mathcal{R}(\mu) \quad (\mathrm{CQL}(\mathcal{R}))$$

- 经验上使用μ和均匀分布直接的KL散度作为正则项的实现

$$\mathcal{R}(\mu) = -D_{\mathrm{KL}}(\mu, \mathrm{Unif}(\mathbf{a}))$$

- CQL可以直接做基于价值函数的训练
- 如果需要做策略训练，则在训练价值函数Q的同时，使用Soft AC算法训练出策略π

Algorithm 1 Conservative Q-Learning (both variants)

1: Initialize Q-function, Q_θ, and optionally a policy, π_ϕ.
2: **for** step t in $\{1, \ldots, \mathrm{N}\}$ **do**
3: Train the Q-function using G_Q gradient steps on objective from Equation 4
$\theta_t := \theta_{t-1} - \eta_Q \nabla_\theta \mathrm{CQL}(\mathcal{R})(\theta)$
(Use $\mathcal{B}^*$ for Q-learning, $\mathcal{B}^{\pi_{\phi_t}}$ for actor-critic)
4: (only with actor-critic) Improve policy π_ϕ via G_π gradient steps on ϕ with SAC-style entropy regularization:
$\phi_t := \phi_{t-1} + \eta_\pi \nabla_\phi \mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\pi_\phi(\cdot|\mathbf{s})} [Q_\theta(\mathbf{s},\mathbf{a}) - \log \pi_\phi(\mathbf{a}|\mathbf{s})]$
5: **end for**

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine.
Conservative q-learning for offline reinforcement learning. NeuIPS 2020.

https://sites.google.com/view/cql-offline-rl

CQL：保守Q学习的实验

Task Name	SAC	BC	BEAR	BRAC-p	BRAC-v	CQL($\mathcal{H}$)
halfcheetah-random	30.5	2.1	25.5	23.5	28.1	**35.4**
hopper-random	**11.3**	9.8	9.5	**11.1**	**12.0**	**10.8**
walker2d-random	4.1	1.6	**6.7**	0.8	0.5	**7.0**
halfcheetah-medium	-4.3	36.1	38.6	**44.0**	**45.5**	**44.4**
walker2d-medium	0.9	6.6	33.2	72.7	**81.3**	79.2
hopper-medium	0.8	29.0	47.6	31.2	32.3	**58.0**
halfcheetah-expert	-1.9	**107.0**	**108.2**	3.8	-1.1	104.8
hopper-expert	0.7	**109.0**	**110.3**	6.6	3.7	**109.9**
walker2d-expert	-0.3	125.7	106.1	-0.2	-0.0	**153.9**
halfcheetah-medium-expert	1.8	35.8	51.7	43.8	45.3	**62.4**
walker2d-medium-expert	1.9	11.3	10.8	-0.3	0.9	**98.7**
hopper-medium-expert	1.6	**111.9**	4.0	1.1	0.8	**111.0**
halfcheetah-random-expert	53.0	1.3	24.6	30.2	2.2	**92.5**
walker2d-random-expert	0.8	0.7	1.9	0.2	2.7	**91.1**
hopper-random-expert	5.6	10.1	10.1	5.8	11.1	**110.5**
halfcheetah-mixed	-2.4	38.4	36.2	**45.6**	**45.9**	**46.2**
hopper-mixed	3.5	11.8	25.3	0.7	0.8	**48.6**
walker2d-mixed	1.9	11.3	10.8	-0.3	0.9	**26.7**

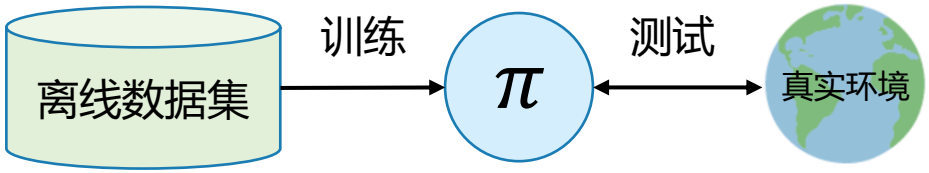
- 在多个Gym环境和不同的数据集采样设置下，CQL几乎都能取得最好的策略性能

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine.
Conservative q-learning for offline reinforcement learning. NeuIPS 2020.

https://sites.google.com/view/cql-offline-rl

总结离线强化学习

- 离线强化学习：在一个给定的离线数据集上直接训练出智能体策略，训练的过程中，智能体不得和环境做交互



- 离线强化学习面临的最重要的挑战是外延误差（Extrapolation Error）
  - 也即是处理分布外（out-of-distribution，OOD）问题
  - 智能体如果涉足到了从没有见过的、远离数据集的状态动作对，怎么办？

- 离线强化学习的主要方法在于设计训练中的限制，从而避免分布外问题，可以大致分为无模型的方法和基于模型的方法

- 离线强化学习的评测集：RL Unplugged, D4RL, NeoRL

THANK YOU