Offline Reinforcement Learning for Recommendation

Xin Xin, PhD, University of Glasgow

About Me

- Basic Information
  - Xin Xin (辛鑫)，PhD in School of Computing Science, University of Glasgow
  - Supervised by: Prof.Joemon Jose and Dr. Alexandros Karatzoglou
- Research Interest
  - Recommender systems, information retrieval, machine learning & reinforcement learning
- Contact Me
  - E-mail: x.xin.1@research.gla.ac.uk
  - Homepage: https://xinxin-me.github.io/
  - Wechat ID: xin_glazt

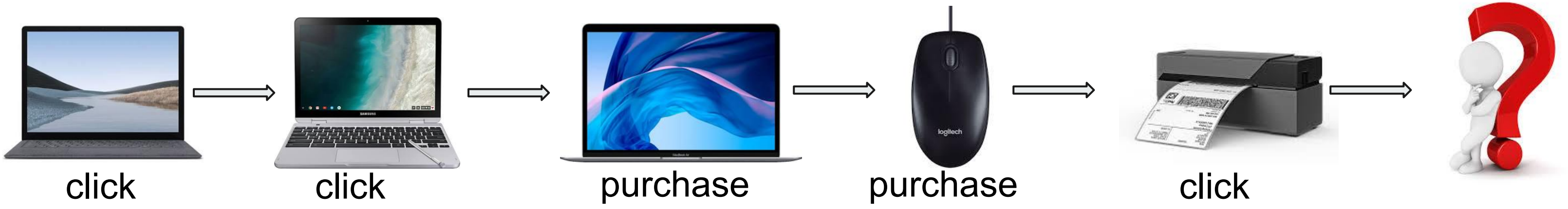
Outline

- Background and Motivation
- Research Challenges
- Offline RL for Recommendation
- Promising Directions

Outline

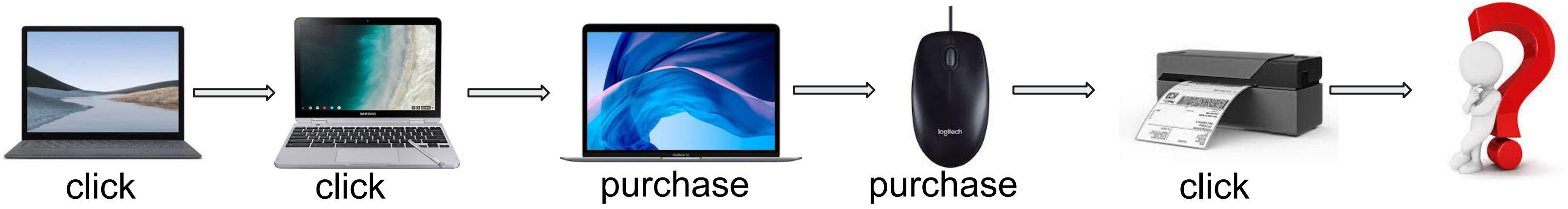
- Background and Motivation

Background and Motivation



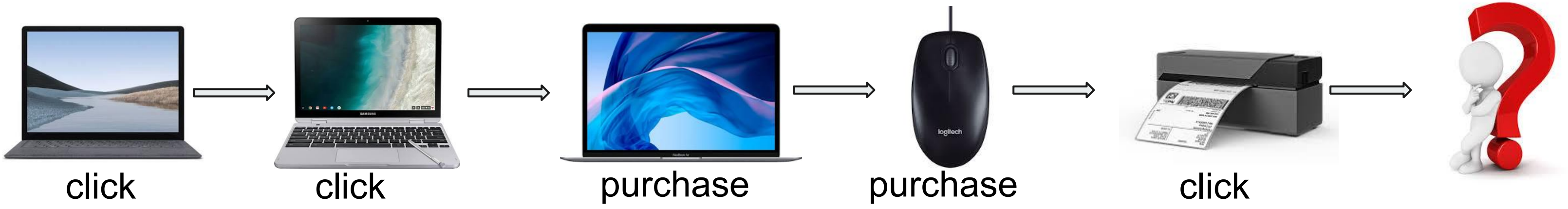
- Recommender systems (RS) aim to provide interesting items to users according to previous interactions

Background and Motivation



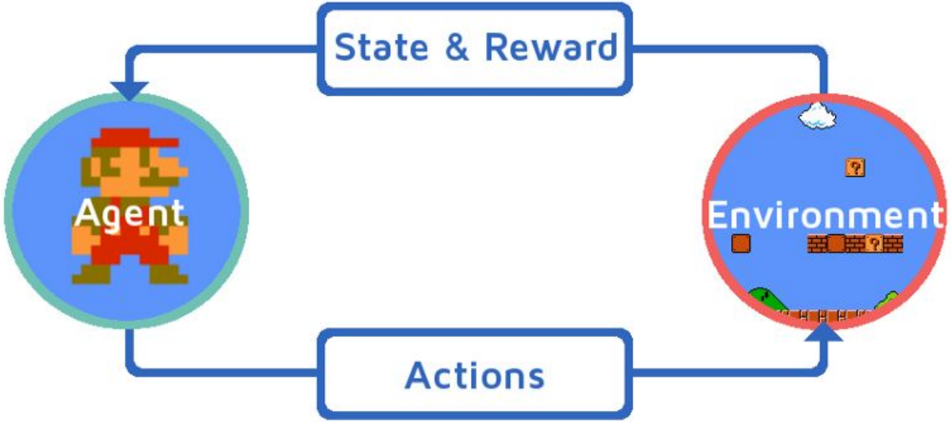
- Recommender systems (RS) aim to provide interesting items to users according to previous interactions
- A typical training method of RS is supervised learning

Background and Motivation



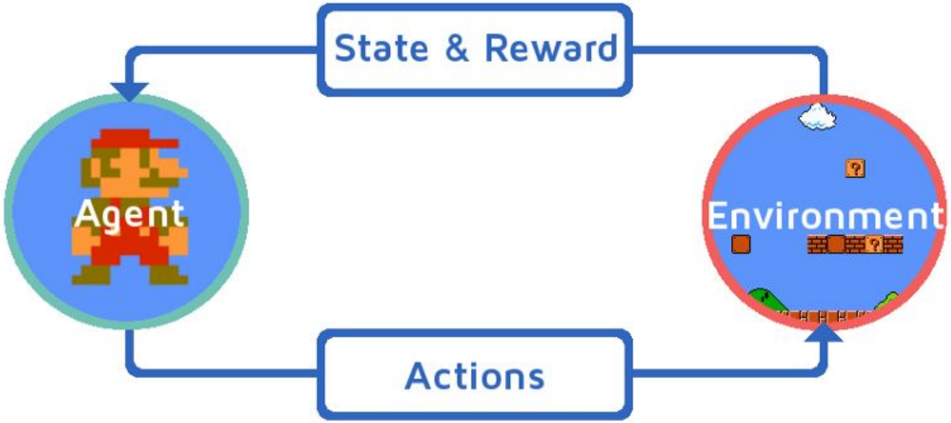
- Recommender systems (RS) aim to provide interesting items to users according to previous interactions
- A typical training method of RS is supervised learning (SL)
- There are some practical needs which SL may be ineffective to model
    - long-term user engagement
    - promoting purchases
    - longer dwell time, etc.

Background and Motivation



The **RL agent** is trained to take **actions** given the **state** of the **environment** with the objective of getting the **maximum long-term rewards.**

Background and Motivation

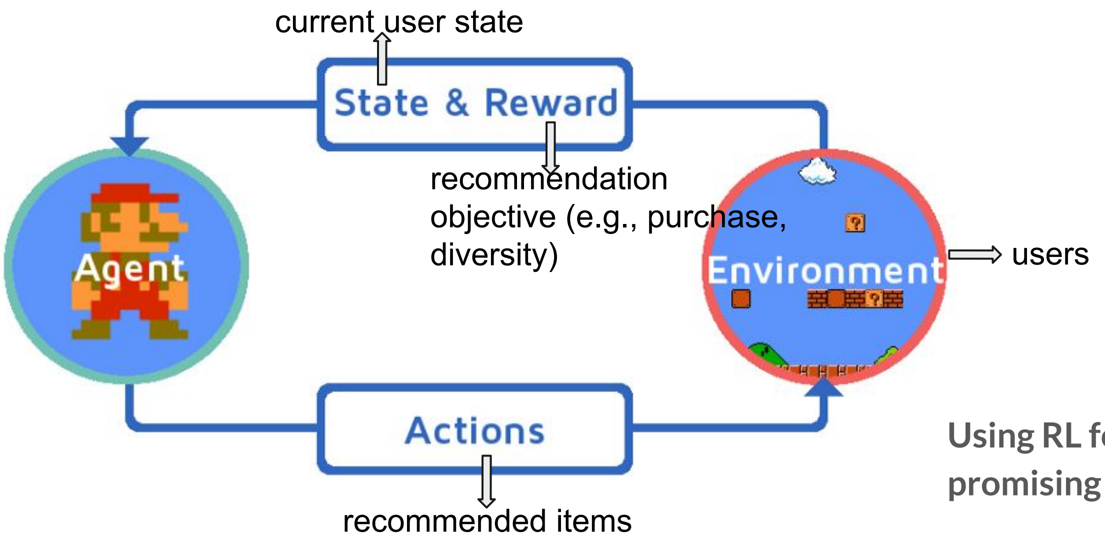


The **RL agent** is trained to take **actions** given the **state** of the **environment** with the objective of getting the **maximum long-term rewards.**

Advantage of RL:

- Flexible reward setting
- Long-term optimization

Background and Motivation



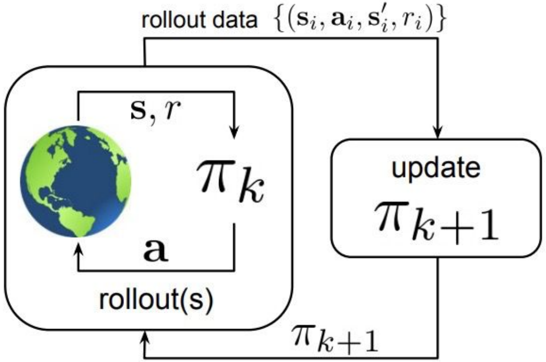
Using RL for recommendation is a promising direction

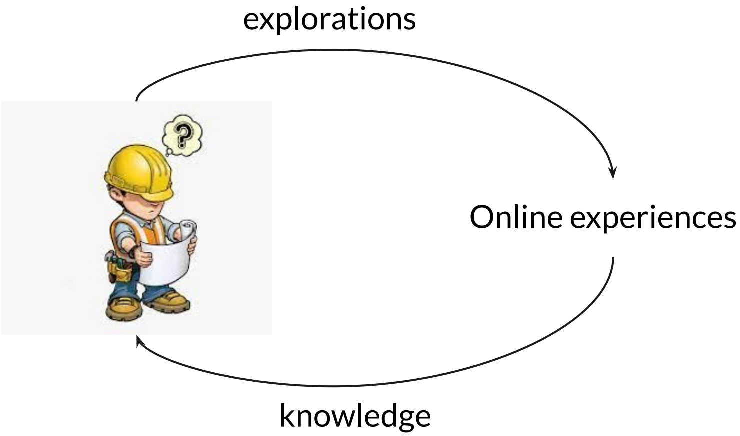
Outline

- Background and Motivation
- Research Challenges
  - RL Overview

RL Overview

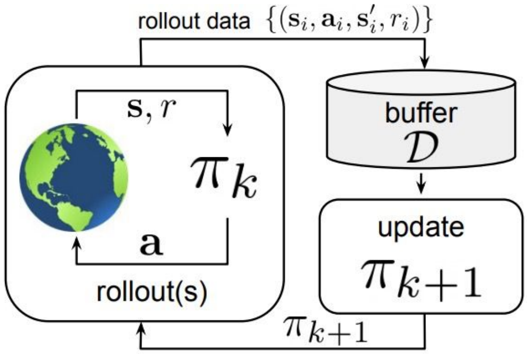
- RL is a more human-like learning approach
    - Learning through ‘making errors’
- On-policy(online) RL one of the most widely used methods
    - Policy Gradient, Monte-Carlo estimator

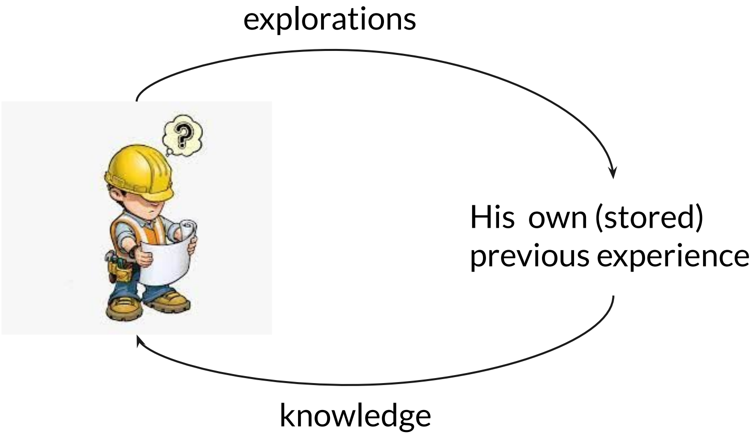




RL Overview

- On-policy RL is data inefficient
- Space for data reuse (improved data efficiency)-->off-policy RL
    - Q-learning, actor-critic, soft actor-critic, etc.



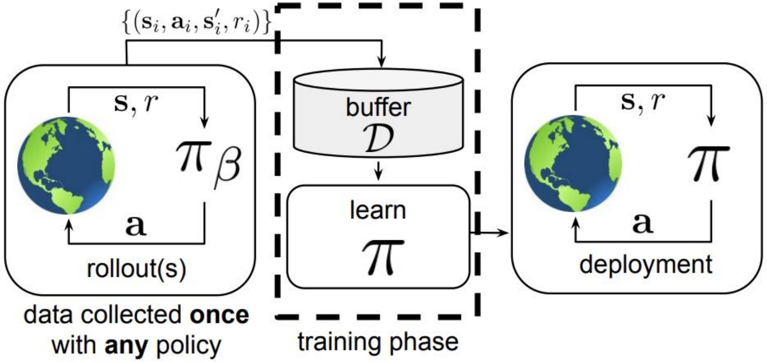


RL Overview

- Recommendation is a user-oriented task
    - Making ‘errors’ is too expensive
        - On-policy learning could be infeasible
    - We hope that the agent can learn a good policy without affecting the real user experience
        - Learning from historical logged data
        - Off-policy methods learn from ‘own’ experiences
            - Logged data is not controlled by the agent
            - Still needs plenty of new online interactions

RL Overview

- Offline RL is pure ‘off-policy’

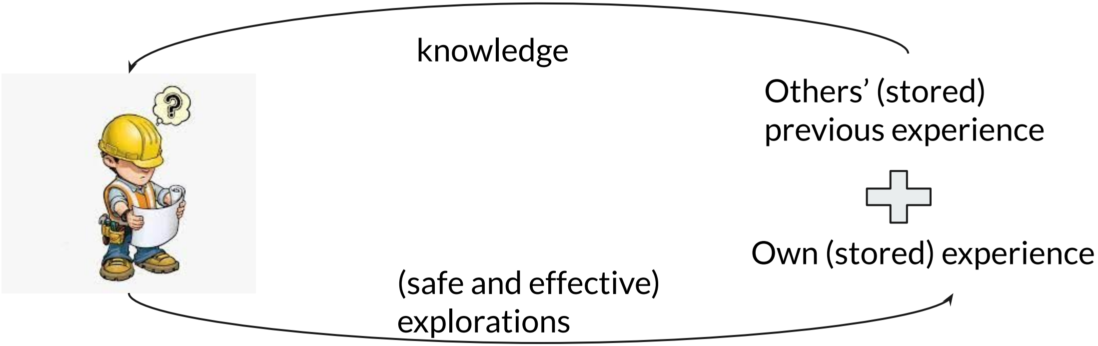


knowledge

Others’ (stored) experience

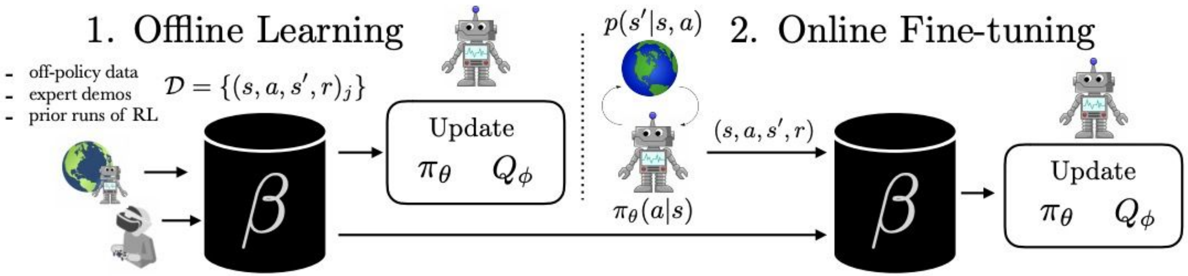
RL Overview

- An expected RL-based recommendation learning approach
  - Offline learning+online fine-tuning



RL Overview

- Offline RL provides an acceptable starting point for RL-based recommendation
  - Offline RL is expected to overperform supervised learning (behaviour cloning)



Outline

- Background and Motivation
- Research Challenges
    - RL Overview
    - Challenge Analysis

Challenge Analysis

- Offline RL vs Supervised Learning
  - SL trains a model that attains minimum supervised loss on data coming from the same distribution as the training data
    - regression to training data
  - Offline RL is about making counterfactual inferences, i.e., “what if” questions.
    - what might happen if the agent were to carry out actions different from the training data
    - if we want the learned policy to perform better than the SL, we must execute ‘something new’
    - ‘Making new’ is not so easy....

Challenge Analysis

- Distribution Shift
    - RL objective: $\max_{\pi_\theta} \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)], \text{ where } R(\tau) = \sum_{t=0}^{|\tau|} \gamma^t r(\mathbf{s}_t, a_t),$

Challenge Analysis

- Distribution Shift
  - RL objective: $\max_{\pi_\theta} \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)]$, where $R(\tau) = \sum_{t=0}^{|\tau|} \gamma^t r(\mathbf{s}_t, a_t)$,
  - Desired gradient (using "log-trick")

$$\mathbb{E}_{\tau \sim \pi_\theta}[R(\tau) \nabla_\theta \log \pi_\theta(\tau)]$$

Challenge Analysis

- Distribution Shift
  - RL objective: $\max_{\pi_\theta} \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$, where $R(\tau) = \sum_{t=0}^{|\tau|} \gamma^t r(\mathbf{s}_t, a_t)$,
  - Desired gradient (using "log-trick")

$$\mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)\nabla_\theta \log \pi_\theta(\tau)]$$

  - Estimated gradient (off-policy)

$$\mathbb{E}_{\tau\sim\beta}[R(\tau)\nabla_\theta \log \pi_\theta(\tau)]$$

Challenge Analysis

- Distribution Shift
  - RL objective: $\max_{\pi_\theta} \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$, where $R(\tau) = \sum_{t=0}^{|\tau|} \gamma^t r(\mathbf{s}_t, a_t)$,
  - Desired gradient (using "log-trick")

$$\mathbb{E}_{\tau\boxed{\sim\pi_\theta}}[R(\tau)\nabla_\theta \log \pi_\theta(\tau)]$$

  - Estimated gradient (off-policy)

$$\mathbb{E}_{\tau\boxed{\sim\beta}}[R(\tau)\nabla_\theta \log \pi_\theta(\tau)]$$

There is the distribution discrepancy.

Challenge Analysis

- Lack of Rewards
  - Rewards is sparse in RS
  - Lack of negative rewards

$$x_{1:t}\ \{\overset{click}{x_1},\ \overset{purchase}{x_2},\ \ldots,\ \overset{click}{x_{t-1}},\ \overset{click}{x_t}\}$$

$$Q(\boldsymbol{s}_0, x_1) = \text{reward of click} + \max_a Q(\boldsymbol{s}_1, a)$$

$$Q(\boldsymbol{s}_1, x_2) = \text{reward of purchase} + \max_a Q(\boldsymbol{s}_2, a)$$

$$Q(\boldsymbol{s}_0, x_1^-) = ?\ \ Q(\boldsymbol{s}_1, x_2^-) = ?$$ ** *no learning constraints* **

$$argmaxQ(\boldsymbol{s}, a) = ?$$ ** *fails to perform ranking* **

Outline

- Background and Motivation
- Research Challenges
- Offline RL for Recommendation
    - Inverse Propensity Score

Inverse Propensity Score

- Explicitly correct the distribution discrepancy
  - Trade-off between bias and variance
  - Smoothing and Cliping
  - Estimation of behavior policy

$$\nabla_\theta \mathcal{J}(\pi_\theta) = \sum_{s_t \sim d_t^\beta(\cdot), a_t \sim \beta(\cdot|s_t)} \omega(s_t, a_t) R_t \nabla_\theta \log \pi_\theta(\tau) \qquad \omega(s_t, a_t) = \frac{d_t^\pi(s_t)}{d_t^\beta(s_t)} \times \frac{\pi_\theta(a_t|s_t)}{\beta(a_t|s_t)}$$

[1]Chen, Minmin, et al. "Top-k off-policy correction for a REINFORCE recommender system." *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 2019.

Outline

- Background and Motivation
- Research Challenges
- Offline RL for Recommendation
    - Inverse Propensity Score
    - Batch Constrained Q-learning

Batch Constrained Q-learning

- Explicitly restrict the action space to the logged data
  - through a supervised generative model
  - allow explorations in a trust region

Algorithm 1 Batch Constrained Q-Learning

1. Input: Batch data $\mathcal{B}$, horizon T, target network update rate τ, mini-batch size N, threshold τ.
2. Initialize Q-networks Q_θ, generative model G_ω (trained in a standard supervised learning fashion, with a cross-entropy loss) and target networks $Q_{\theta'}$,with $\theta' \leftarrow \theta$.
3. for $t = 1\, to\, T$ do
4. Sample mini-batch M of N transitions (s, a, r, s') from $\mathcal{B}$
5. $a' = \arg\max_{a' \mid \frac{G_\omega(a'|s')}{max_{\hat{a}} G_\omega(\hat{a}|s')} > \tau} Q_\theta(s', a')$
6. $\theta \leftarrow \arg\min_\theta \sum_{(s,a,r,s') \in M} l_\kappa\left(r + \gamma Q_{\theta'}(s', a') - Q_{\theta'}(s, a)\right)$
7. $\omega \leftarrow \arg\min_\omega - \sum_{(s,a) \in M} \log G_\omega(a|s)$
8. If t mod $\tau = 0$: $\theta' \leftarrow \theta$
9. end for

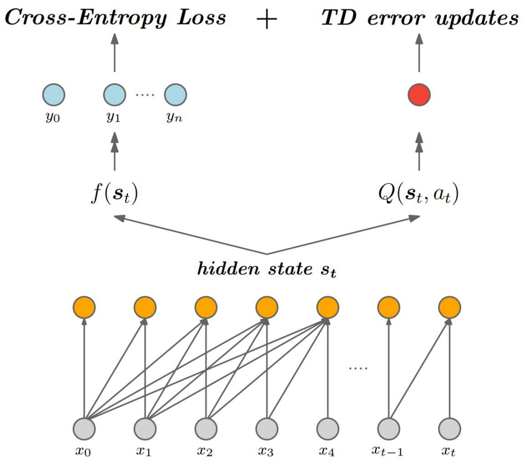
[2] Fujimoto, Scott, David Meger, and Doina Precup. "Off-policy deep reinforcement learning without exploration." *International Conference on Machine Learning*. PMLR, 2019.
[3] Dynamic Personalized Pricing Using Batch Deep Reinforcement Learning: An Application to LiveStream Shopping," DRL4IR 2020

Outline

- Background and Motivation
- Research Challenges
- Offline RL for Recommendation
    - Inverse Propensity Score
    - Batch Constrained Q-learning
    - Self-Supervised RL

Self-Supervised RL

- Self-Supervised Q-learning
  - The Q-value estimator acts as a regularizer
  - The recommendation is still generated from SL
  - A shared base model for knowledge transfer between SL and RL



Cross-Entropy loss provides ranking (negative) gradient signals

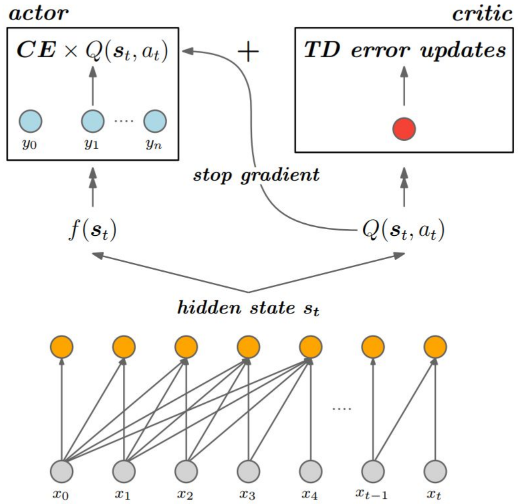
$$L_{SQN} = L_s + L_q.$$

RL loss introduces desired reward settings and long-term perspective

[4] Xin, Xin, et al. "Self-supervised reinforcement learning for recommender systems." *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020.

Self-Supervised RL

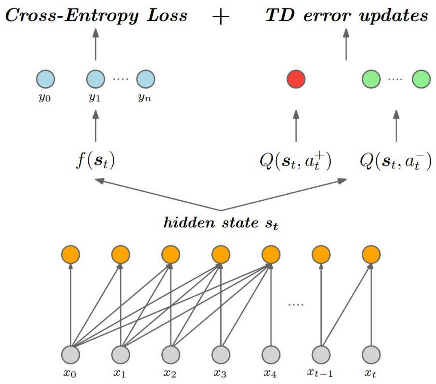
- Self-Supervised Actor-Critic
  - SL as the actor
  - Q-learning output as the critic
  - A shared base model for knowledge transfer between SL and RL
  - Stop gradient when Q-values are used as weights



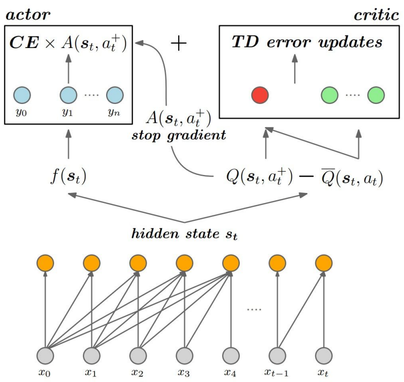
[4] Xin, Xin, et al. "Self-supervised reinforcement learning for recommender systems." *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020.

Self-Supervised RL

- Negative Sampling+RL
  - Supervised Negative Q-learning
  - Supervised Advantage Actor-Critic



(a) SNQN architecture.



(b) SA2C architecture. CE is short for cross-entropy.

Outline

- Background and Motivation
- Research Challenges
- Offline RL for Recommendation
    - Inverse Propensity Score
    - Batch Constrained Q-learning
    - Self-Supervised RL
    - Uncertainty-Based RL

Uncertainty-Based RL

- Conservative Q-learning
  - we expect the uncertainty to be substantially larger for out-of-distribution actions
  - introduce the uncertainty into Q-value estimation
  - Uncertainty can be defined as the difference between behavior policy and target policy

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

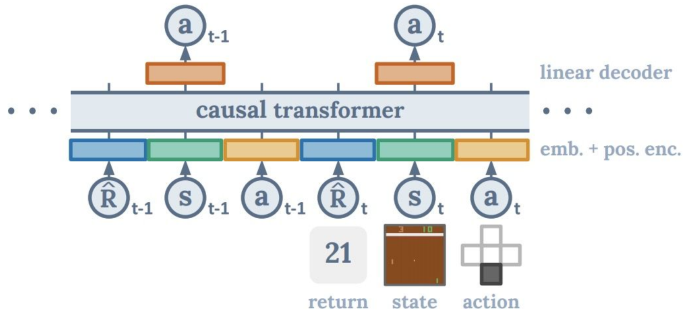
[5] Kumar, Aviral, et al. "Conservative q-learning for offline reinforcement learning." *arXiv preprint arXiv:2006.04779* (2020).

Outline

- Background and Motivation
- Research Challenges
- Offline RL for Recommendation
- Promising Directions
    - Self- Supervised Learning + RL

Self-Supervised Learning + RL

- SSL helps to improve the representation learning
  - Increase the data efficiency of RL
  - Enough offline data+powerful model→ Decision Transformer



[6] Chen, Lili, et al. "Decision transformer: Reinforcement learning via sequence modeling." *arXiv preprint arXiv:2106.01345* (2021).

Outline

- Background and Motivation
- Research Challenges
- Offline RL for Recommendation
- Promising Directions
    - Self- Supervised Learning + RL
    - Model-Based offline RL

Model-Based Offline RL

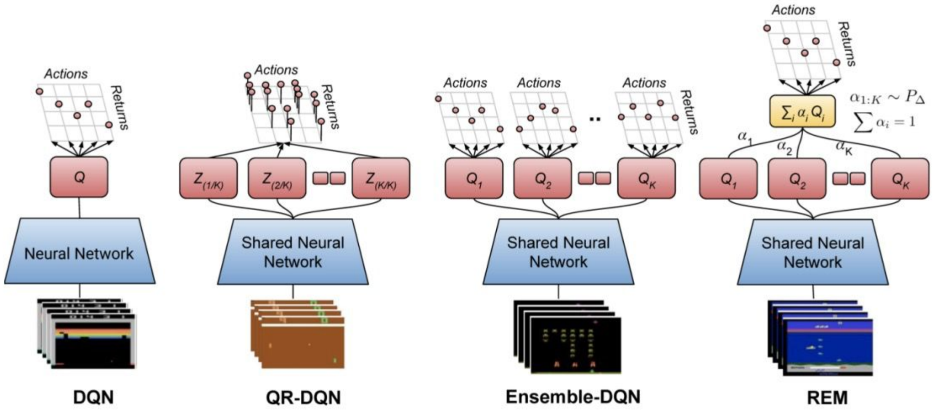
- Model-Based RL
    - Learning a simulator from historical records
    - Data augmentation from simulated interactions
    - Distribution shift (bias) in the simulator
    - Debias of the simulator
        - Causal inference, disentangle learning

Outline

- Background and Motivation
- Research Challenges
- Offline RL for Recommendation
- Promising Directions
    - Self- Supervised Learning + RL
    - Model-Based offline RL
    - Ensemble RL

Ensemble RL

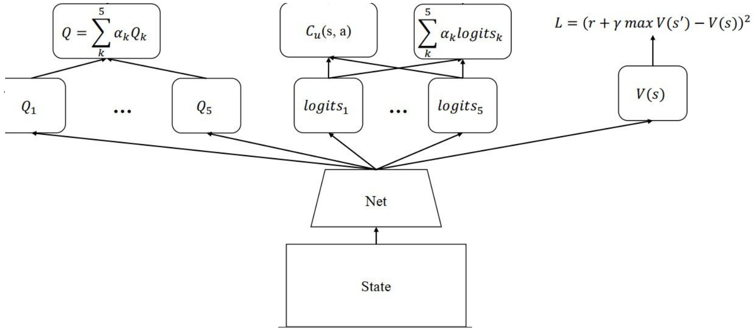
- Ensemble Q-learning for better exploration



[7] Agarwal, Rishabh, Dale Schuurmans, and Mohammad Norouzi. "An optimistic perspective on offline reinforcement learning." *International Conference on Machine Learning*. PMLR, 2020.

Ensemble RL

- Ensemble learning+Uncertainty estimation
  - Exploitation and exploration trade-off: for good actions, we encourage exploitation; for bad actions, we encourage exploration（uncertainty estimation）
  - So how to determine good or bad?
    - 1.from data itself (observed reward)
    - 2.from the comparison between different ensembled models



Outline

- Background and Motivation
- Research Challenges
- Offline RL for Recommendation
- Promising Directions
    - Self- Supervised Learning + RL
    - Model-Based offline RL
    - Ensemble RL
    - Online fine-tuning

Outline

- Background and Motivation
- Research Challenges
- Offline RL for Recommendation
- Promising Directions
    - Self- Supervised Learning + RL
    - Model-Based offline RL
    - Ensemble RL
    - Online fine-tuning
    - Offline evaluation

Q&A

Thanks for your listening!