Offline RL

Aarti Singh

Machine Learning 10-734
Nov 18, 2025

Slides courtesy: Wen Sun

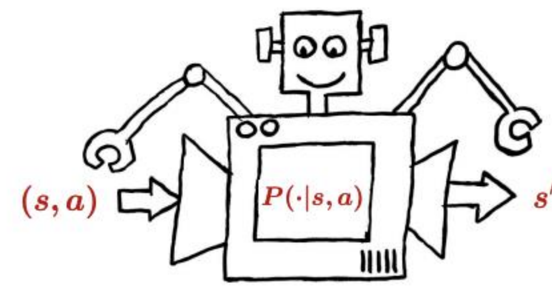
Data source in RL

Exploration

offline RL

"Recalculating ... recalculating ..."

online RL

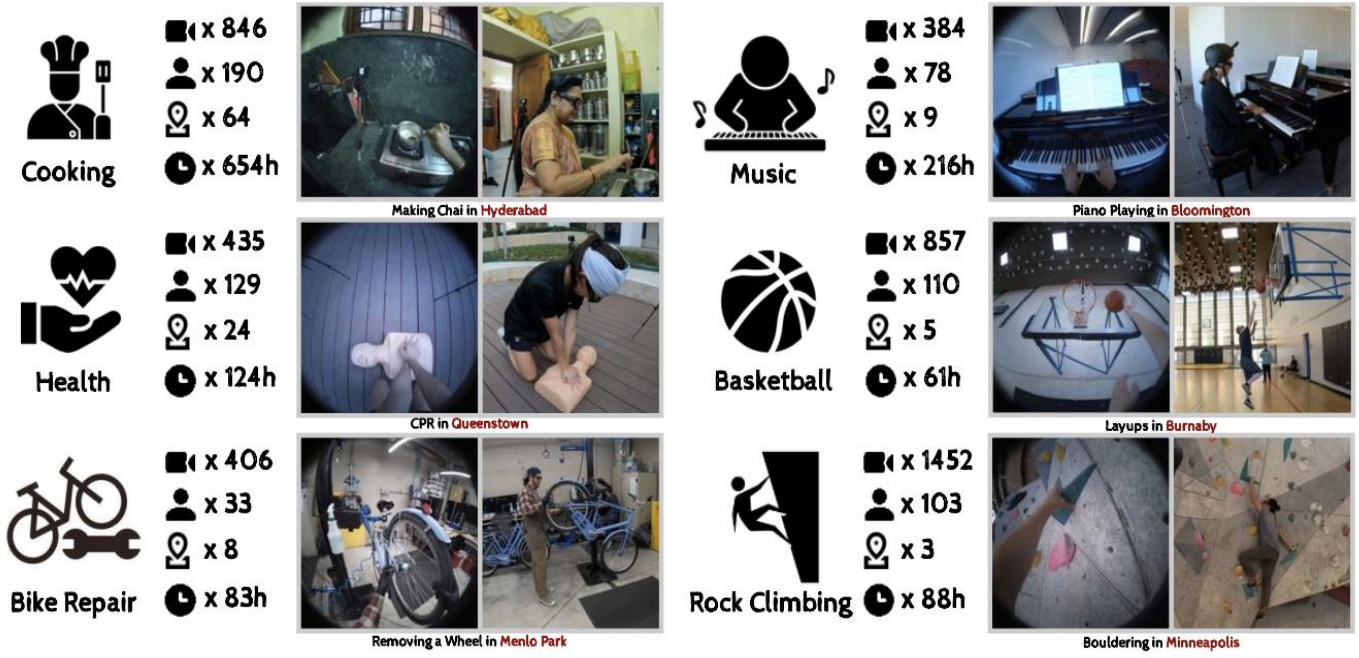


generative model

The capability of exploration increases from left to right.

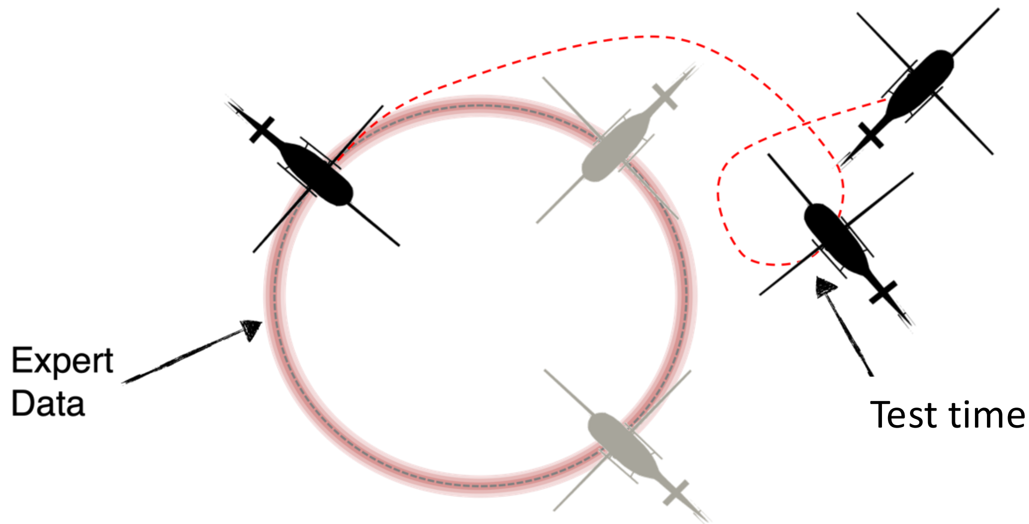
Offline RL setting

- Applications – learning from demonstrations, past experiences, observational data



Distribution shift issue

- Offline data may not have seen test time scenarios



Offline RL setting

1. Infinite horizon Discounted MDPs $\gamma \in (0,1)$

2. A given offline distribution $\nu \in \Delta(S \times A)$

3. Function class $\mathscr{F} = \{f : S \times A \mapsto [0,1/(1-\gamma)]\}$

Key assumptions

1. offline distribution ν has full coverage (i.e., diverse):

$$\max_{\pi} \max_{s,a} \frac{d^{\pi}(s,a)}{\nu(s,a)} \leq C < \infty$$

2. Small inherent Bellman error, i.e., near Bellman Completion (note it's averaged over ν):

$$\max_{g \in \mathcal{F}} \min_{f \in \mathcal{F}} \mathbb{E}_{s,a \sim \nu} \left(f(s,a) - \mathcal{T} g(s,a) \right)^2 \leq \epsilon_{approx,\nu}$$

Fitted Q-Iteration, FQI algorithm

1. offline data points obtained from ν:

$$\mathcal{D} = \{s, a, r, s'\}, \quad (s, a) \sim \nu, r = r(s, a), s' \sim P(\cdot \mid s, a)$$

2. Initialize $f_0 \in \mathcal{F}$, and iterate:

$$f_{t+1} = \arg\min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s, a) - r - \gamma \max_{a'} f_t(s', a') \right)^2$$

3. After K iterations, return $\pi(s) = \arg\max_{a} f_K(s, a), \forall s$

Note: the algorithmic idea here is similar to DQNs [Deepmind 15]

FQI – why it works

$$y := r(s,a) + \gamma \max_{a'} f_t(s',a')$$

Bayes optimal: $\underbrace{r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a} f_t(s',a')}_{(\mathcal{T} f_t)(s,a)}$

1. Near Bellman completion means regression target $\mathcal{T} f_t$ nearly belongs to $\mathcal{F}$

$$\mathbb{E}_{s,a\sim\nu} \left(f_{t+1}(s,a) - \mathcal{T} f_t(s,a)\right)^2 \approx \frac{1}{N} + \epsilon_{approx,\nu}$$

2. $f_{t+1} \approx \mathcal{T} f_t$ (under the diverse ν), i.e., it's like Value Iteration, we could hope for a convergence

FQI analysis

For simplicity, we analyze the case when $\epsilon_{approx,\nu} = 0$

The k[th] iteration of FQI guarantees that with probability $\geq 1 - \delta$

$$V^\star - V^{\pi_k} \leq \mathcal{O}\left(\frac{V_{\max}}{(1-\gamma)^2}\sqrt{\frac{C\log(|\mathcal{F}|/\delta)}{n}}\right) + \frac{2\gamma^k V_{\max}}{1-\gamma}$$

Statistical error related to regression

Optimization error related to VI convergence

$$\leq \mathcal{O}\left(\frac{1}{(1-\gamma)^3}\sqrt{\frac{C\ \log(|\mathcal{F}|/\delta)}{n}} + \frac{2\gamma^k}{(1-\gamma)^2}\right)$$

since $V_{max} \leq \frac{1}{1-\gamma}$

Statistical error

Standard Generalization Bound for regression:

Given $\{x_i, y_i\}_{i=1}^{N}$, $(x_i, y_i) \sim \nu$, $y_i = f^\star(x_i) + \epsilon_i$, where $|y_i| \leq Y$, $\|f^\star\|_\infty \leq Y$,

a function class $\mathscr{F} = \{f : \mathscr{X} \mapsto [-Y, Y]\}$, where $f^* \in \mathcal{F}$

Denote $\hat{f} := \arg\min_{f \in \mathscr{F}} \sum_{i=1}^{N} (f(x_i) - y_i)^2$ as the least square minimizer, then w/ prob $1 - \delta$:

$$\mathbb{E}_{x \sim \nu}\left(\hat{f}(x) - f^\star(x)\right)^2 \leq O\left(\frac{Y^2 \ln(|\mathscr{F}|/\delta)}{N}\right)$$

Statistical error

Recall FQI's regression problem

$$f_{t+1} = \arg\min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \gamma \max_{a'} f_t(s',a') \right)^2$$

Here define $f^\star := \mathcal{T} f_t,$

And due to small Bellman error $\min_{f \in \mathcal{F}} \mathbb{E}_{s,a\sim\nu}(f(s,a) - \mathcal{T} f_t(s,a))^2 \leq \cancelto{0}{\epsilon_{approx,\nu}}$

$$\Rightarrow \mathbb{E}_{s,a\sim\nu}(f_{t+1}(s,a) - \mathcal{T} f_t(s,a))^2 \leq \frac{1}{(1-\gamma)^2} \frac{\ln(|\mathcal{F}|/\delta)}{n}$$

with probability $\geq 1 - \delta$

Optimization error

Consider any state-action distribution $\beta(s,a)$ (induced by some policy)

$$\sqrt{\mathbb{E}_{s,a\sim\beta}(f_t(s,a) - Q^\star(s,a))^2} := \|f_t - Q^\star\|_{\beta,2}$$

$$\leq \|f_t - \mathcal{T} f_{t-1}\|_{2,\beta} + \|\mathcal{T} f_{t-1} - Q^\star\|_{2,\beta}$$

$$\leq \sqrt{C}\|f_t - \mathcal{T} f_{t-1}\|_{2,\nu} + \|\mathcal{T} f_{t-1} - Q^\star\|_{2,\beta}$$

Coverage assumption

$$\leq \sqrt{C}\epsilon_{regress} + \gamma\sqrt{\mathbb{E}_{s,a\sim\beta}\left(\mathbb{E}_{s'\sim P(\cdot|s,a)}\left(\max_{a'} f_{t-1}(s',a') - \max_{a'} Q^\star(s',a')\right)\right)^2}$$

$$\leq \sqrt{C}\epsilon_{regress} + \gamma\sqrt{\underbrace{\mathbb{E}_{s,a\sim\beta}\mathbb{E}_{s'\sim P(\cdot|s,a)}\max_{a'}}_{:=\beta'(s',a')}\left(f_{t-1}(s',a') - Q^\star(s',a')\right)^2}$$

$$= \sqrt{C}\epsilon_{regress} + \gamma\|f_{t-1} - Q^\star\|_{2,\beta'}$$

Optimization error

Consider any state-action distribution $\beta(s, a)$ (induced by some policy)

$$\sqrt{\mathbb{E}_{s,a\sim\beta}(f_t(s,a) - Q^\star(s,a))^2} := \|f_t - Q^\star\|_{\beta,2}$$

$$\le \sqrt{C}\epsilon_{regress} + \gamma\|f_{t-1} - Q^\star\|_{2,\beta'}$$

$$\le \sqrt{C}\epsilon_{regress} + \gamma\left[\sqrt{C}\epsilon_{regress} + \gamma\|f_{t-2} - Q^\star\|_{2,\beta''}\right]$$

$$\le \sqrt{C}\epsilon_{regress}\left(1 + \gamma + \ldots + \gamma^k\right) + \gamma^k\|f_0 - Q^\star\|_{2,\tilde{\beta}}$$

$$\le \frac{\sqrt{C}\epsilon_{regress}}{1-\gamma} + \gamma^k/(1-\gamma)$$

FQI policy error bound

Convert Q-error $\|f_k - Q^\star\|_{2,\beta}$ to policy error.

Denote $\pi^k(s) = \arg\max_a f_k(s,a)$

$$V^\star - V^{\pi^k} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi^k}}\left[Q^\star(s,\pi^\star(s)) - Q^\star(s,\ \pi^k(s))\right]$$

$$\leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi^k}}\left[Q^\star(s,\pi^\star(s)) - f_k(s,\pi^\star(s)) + f_k(s,\pi^k(s)) - Q^\star(s,\ \pi^k(s))\right]$$

$$\leq \frac{1}{1-\gamma}\left[\sqrt{\mathbb{E}_{s\sim d^{\pi^k}}\left(Q^\star(s,\pi^\star(s)) - f_k(s,\pi^\star(s))\right)^2} + \sqrt{\mathbb{E}_{s\sim d^{\pi^k}}\left(f_k(s,\pi^k(s)) - Q^\star(s,\pi^k(s))\right)^2}\right]$$

$$\leq \frac{2}{1-\gamma}\left(\frac{\sqrt{C}\epsilon_{regress}}{1-\gamma} + \frac{\gamma^k}{1-\gamma}\right)$$

FQI analysis – finite $\mathcal{F}$

For simplicity, we analyze the case when $\epsilon_{approx,\nu} = 0$

The k[th] iteration of FQI guarantees that with probability $\geq 1 - \delta$

$$V^{\star} - V^{\pi_k} \leq \frac{2}{1-\gamma}\left(\frac{\sqrt{C}\epsilon_{regress}}{1-\gamma} + \frac{\gamma^k}{1-\gamma}\right)$$

$$\leq \mathcal{O}\left(\frac{1}{(1-\gamma)^3}\sqrt{\frac{C\ \log(|\mathcal{F}|/\delta)}{n}} + \frac{2\gamma^k}{(1-\gamma)^2}\right)$$

Statistical error related to regression

Optimization error related to VI convergence

FQI analysis - finite $\mathcal{F}$

Now, we analyze the case when $\boldsymbol{\epsilon_{approx,\nu} \neq 0}$

The k^th iteration of FQI guarantees that with probability $\geq 1-\delta$

$$V^\star - V^{\pi_k} \leq \frac{2}{1-\gamma}\left(\frac{\sqrt{C}\epsilon_{regress}}{1-\gamma} + \frac{\gamma^k}{1-\gamma}\right)$$

$$\leq \mathcal{O}\left(\underbrace{\frac{1}{(1-\gamma)^3}\sqrt{\frac{C\ \log(|\mathcal{F}|/\delta)}{n}}}_{\text{Statistical error related to regression}} + \underbrace{\epsilon_{approx,\nu}}_{\text{Inherent Bellman error}} + \underbrace{\frac{2\gamma^k}{(1-\gamma)^2}}_{\text{Optimization error related to VI convergence}}\right)$$

FQI guarantee for low Bellman rank

Now, we analyze the case when $\boldsymbol{\epsilon_{approx,\nu} \neq 0}$

The k[th] iteration of FQI guarantees that with probability $\geq 1-\delta$

$$V^\star - V^{\pi_k} \leq \frac{2}{1-\gamma}\left(\frac{\sqrt{C}\epsilon_{regress}}{1-\gamma} + \frac{\gamma^k}{1-\gamma}\right)$$

$$\leq \mathcal{O}\left(\frac{1}{(1-\gamma)^3}\sqrt{\frac{C\, d \log(1/\delta)}{n} + \epsilon_{approx,\nu}} + \frac{2\gamma^k}{(1-\gamma)^2}\right)$$

Statistical error related to regression

Inherent Bellman error

Optimization error related to VI convergence

General function approximation

- Offline RL - FQI

  requires strong coverage assumption

  low Bellman error + finite $\mathcal{F}$ OR low Bellman rank

- RL with Generative model – Bilinear UCB

  requires generative access

  low Bellman rank

- Online RL – Bilinear UCB

  low Bellman rank

No strong assumptions, but statistically & computationally inefficient!

Worse by factor of H

Fit regression for each round T

Bilinear-UCB, online, bonus

At iteration t :

Select $f_t = \arg\max_{g \in \mathcal{F}} V_g(s_0)$

s.t., $\forall h$: $\sum_{i=0}^{t-1} \left(\mathbb{E}_{\mathcal{D}_{h,i}} [\ell(s_h, a_h, s_{h+1}, g)] \right)^2 \leq R^2$

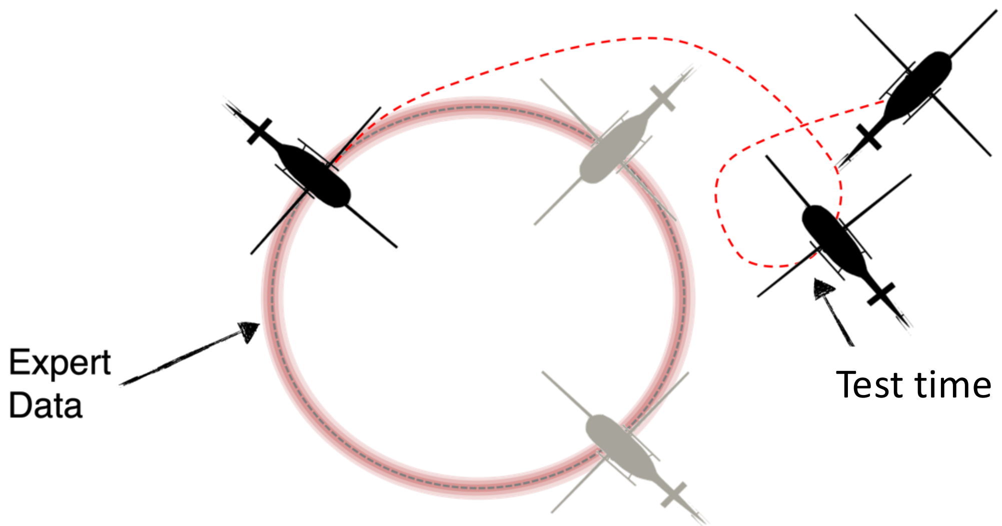
For all h, create $\mathcal{D}_{h,t} = \{s_h, a_h, s_{h+1}\}$ w/ m triples, where:

- For Q-B rank case: $s_h, a_h \sim d_h^{\pi_{f_t}}, s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$
- For V-B rank case: $s_h \sim d_h^{\pi_{f_t}}, a_h \sim U(A), s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$

Roll out π_{f_t} to collect trajectory and add to data

Distribution shift issue

- Offline data may not have seen test time scenarios



Best of both worlds?

Can we combine offline and online data in RL
to reduce sample and compute efficiency
while not requiring strong coverage assumptions?

Yes! Hybrid RL