Hybrid RL

Aarti Singh

Machine Learning 10-734
Dec 2, 2025

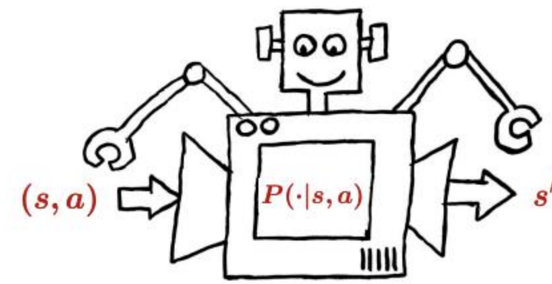
Slides courtesy: Yuda Song

Data source in RL

offline RL

"Recalculating ... recalculating ..."

online RL



generative model

The capability of exploration increases from left to right.

Offline RL

Data collected from offline policy/distribution $\mathcal{D}^{\vartheta} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{m}$
where $(s, a) \sim \vartheta$, offline distribution and $r \sim R(s, a), s' \sim P(.\,|s, a)$

Where do we get the offline data?

Collection of different (not necessarily optimal) policies

Need strong coverage – all possible optimal policies

What if we have expert demonstrations?

Can we mimic experts aka **Imitation learning**

Imitation learning

- Expert demonstrations – state, expert actions (no rewards)

1. **Behavioral cloning** – offline data from expert

   supervised learning of policy π: s-> a using (state, expert actions) data

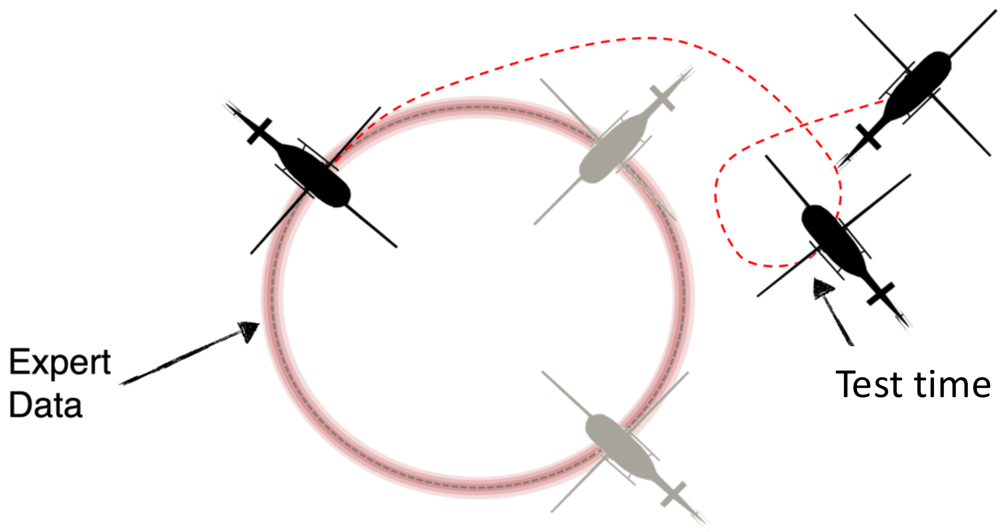
2. **Dagger** (Dataset Aggregation) – online interaction with expert

   roll out policy, collect expert actions for states visited by policy, add to dataset, then repeat

3. **Inverse RL** – first learn reward from (state, expert actions) then train policy using learnt reward

Distribution shift issue – Imitation learning

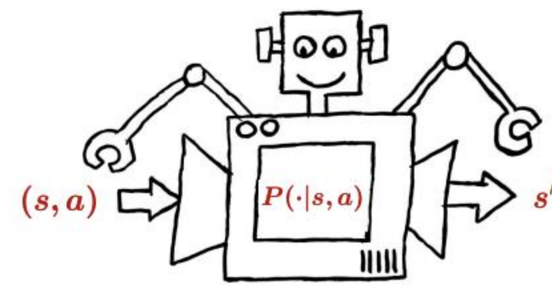
- Offline data may not have seen test time scenarios



Data source in RL

offline RL

online RL



generative model

The capability of exploration increases from left to right.

General function approximation

- Offline RL - FQI

  requires strong coverage assumption but no bonus

- RL with Generative model – Bilinear UCB

  requires generative access

- Online RL – Bilinear UCB

  No strong assumptions

  but computationally inefficient!

Fit regression for each round T under bonus
Ellipsoid constraint

Bilinear-UCB – online RL

At iteration t :

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

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

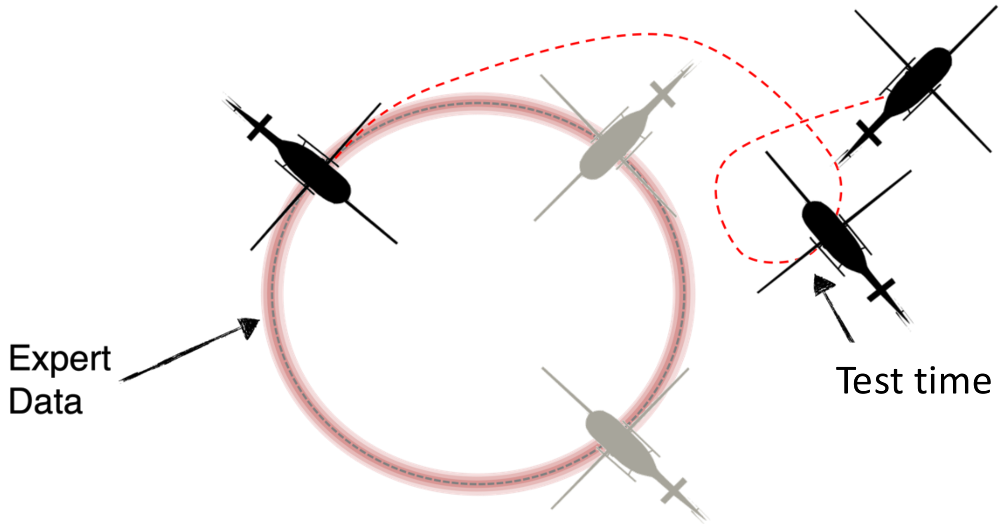
For all h, create $\mathscr{D}_{h,t} = \{s_h, a_h, s_{h+1}\}$ w/ m triples, where:

- 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 RL

- Offline data may not have seen test time scenarios

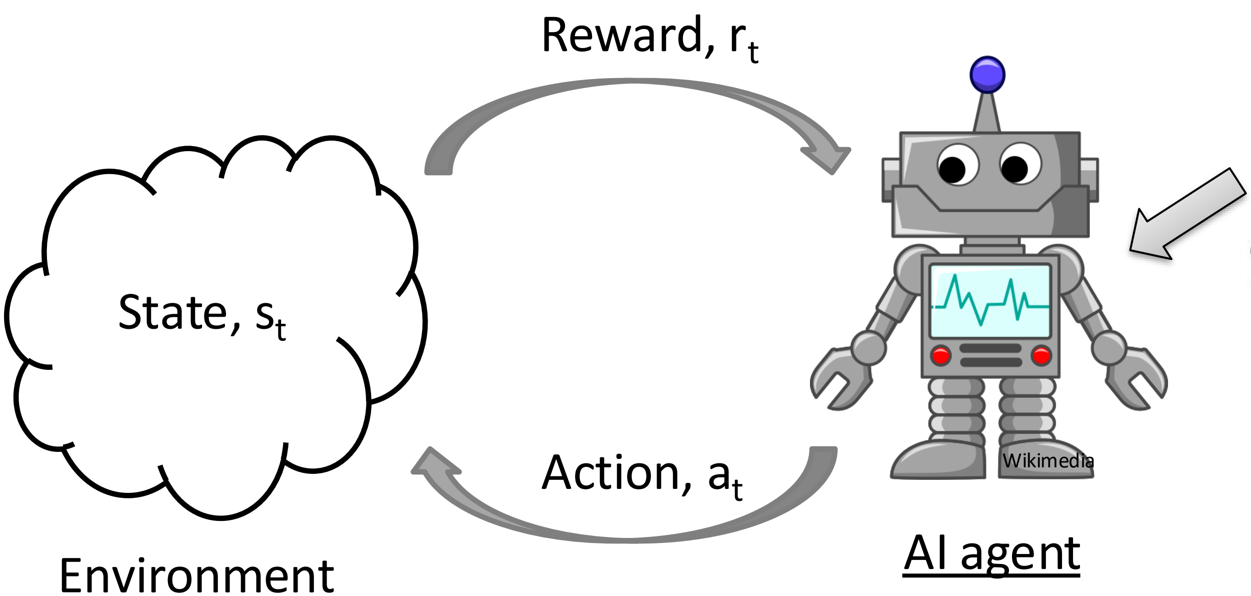


Best of both worlds?

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

Yes! Hybrid RL

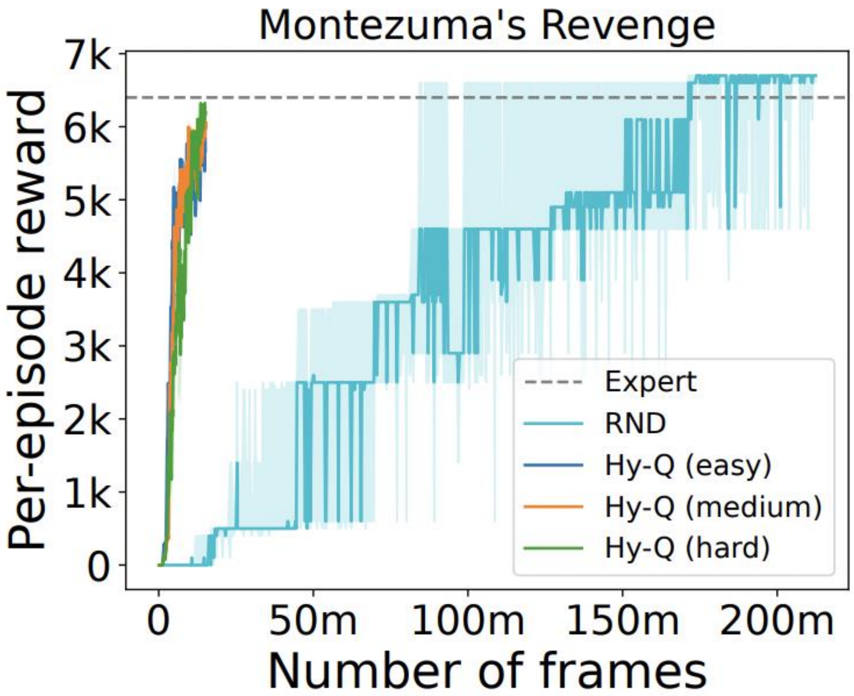
Hybrid RL



Online
interaction

Offline data

Hybrid RL



RND – DeepRL baseline

Hybrid RL – 10x faster than RND with just 0.1m samples

Easy – offline 100% expert data

Medium – 20% random + 80% expert

Hard – 50% random + 50% expert

Setup

- Finite-horizon MDP (S, A, H, P, R, d_0)

- Function approximation

$$\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \times \mathcal{F}_{H-1}$$

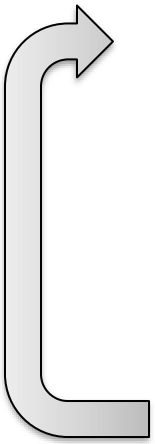
- For each h, we have iid offline dataset

$$\mathcal{D}_h^{\vartheta} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{m}$$

where $(s, a) \sim \vartheta_h$ offline distribution

and $r \sim R(s, a), s' \sim P(. | s, a)$

Hybrid Q Iteration



- Use both offline and online data to fit Q function
- Act greedily according to Q
- Collect online data

No bonus/optimism! – oracle regression efficient

Doesn’t require strong coverage!

Hybrid Q Iteration

(Song et al'23)

Hy-Q: Iterations T, Offline dataset $\mathcal{D}_h^{\nu}$ of size m = T for h = 1, ..., H-1

1: Initialize $f_h^1(s, a) = 0$.

2: **for** $t = 1, \ldots, T$ **do**

3: Let π^t be the greedy policy w.r.t. f^t i.e., $\pi_h^t(s) = \arg\max_a f_h^t(s, a)$.

4: For each h, collect $m_{\text{on}} = 1$ online tuples $\mathcal{D}_h^t \sim d_h^{\pi^t}$.

i.e. observe $s_h \sim d_h^{\pi^t}, a_h \sim \pi_h^t(\cdot|s_h), s_{h+1} \sim P(\cdot|s_h, a_h)$
and add (s_h, a_h, r_h, s_h') to $\mathcal{D}_h^t$

5-7: Run FQI using offline and online data collected so far

Hybrid Q Iteration

(Song et al'23)

Hy-Q: Iterations T, Offline dataset $\mathcal{D}_h^{\vartheta}$ of size m = T for h = 1, …, H-1

Initialize $f_h^1(s,a) = 0$.
for $t = 1, \ldots, T$ **do**
  Let π^t be the greedy policy w.r.t. f^t i.e., $\pi_h^t(s) = \arg\max_a f_h^t(s,a)$.
  For each h, collect $m_{\text{on}} = 1$ online tuples $\mathcal{D}_h^t \sim d_h^{\pi^t}$.
  Set $f_H^{t+1}(s,a) = 0$.
  for $h = H-1, \ldots, 0$ **do**
    Estimate f_h^{t+1} using least squares regression on the aggregated data $\mathcal{D}_h^t = \mathcal{D}_h^{\nu} + \sum_{\tau=1}^{t} \mathcal{D}_h^{\tau}$:

$$f_h^{t+1} \leftarrow \arg\min_{f \in \mathcal{F}_h} \left\{ \widehat{\mathbb{E}}_{\mathcal{D}_h^t} \left(f(s,a) - r - \max_{a'} f_{h+1}^{t+1}(s',a') \right)^2 \right\}$$

Key intuition

- Hy-Q ensures that f_h^t small Bellman error under both offline distribution ϑ_h and online distribution $d_h^{\pi^t}$
  - Robust to distribution shift i.e. if offline data has poor coverage
  - Still leverage offline data to reduce amount of online data
  - Computationally efficient as requires no bonus optimization (computational difficulty when implementing optimism)

Catastrophic forgetting

Why not warm-start with offline data, then switch to online?

- May result in catastrophic forgetting due to a vanishing proportion of offline samples being used for model training as we collect more online samples.
- size of the offline dataset m_{off} should be comparable to the total amount of online data, so that both have similar weight and we ensure low Bellman error on ν throughout the learning process.
- use a fixed (significant) number of offline samples for updating model even as we collect more online data, so that we do not “forget” the distribution ν.
- key practical insight - offline data should be used throughout training to avoid catastrophic forgetting.

HyQ Regret

- FQI guarantees that π_t is at least as good as any policy covered by ν.

Given any comparator policy π^e, for any $f \in \mathcal{F}$ and corresponding greedy policy π^f, we have

$$\mathbb{E}_{s_0 \sim d_0}\left[V_0^{\pi^e}(s_0) - V_0^{\pi^f}(s_0)\right] \leq \sum_{h=0}^{H-1} \underbrace{\mathbb{E}_{s_h, a_h \sim d_h^{\pi^e}}\left[\mathcal{T} f_{h+1}(s_h, a_h) - f_h(s_h, a_h)\right]}_{\textit{offline error}}$$

$$+ \underbrace{\mathbb{E}_{s_h, a_h \sim d_h^{\pi^f}}\left[f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)\right]}_{\textit{online error}}.$$

Proof:

$$\mathbb{E}_{s_0 \sim d_0}\left[V_0^{\pi^e}(s_0) - V_0^{\pi^f}(s_0)\right] = \mathbb{E}_{s_0 \sim d_0}\left[V_0^{\pi^e}(s_0) - \max_a f_0(s_0, a) + \max_a f_0(s_0, a) - V_0^{\pi^f}(s_0)\right].$$

Induction argument on each piece.

HyQ Regret

- FQI guarantees that π_t is at least as good as any policy covered by ν.

Given any comparator policy π^e, *for any* $f \in \mathcal{F}$ *and corresponding greedy policy* π^f, *we have*

$$\mathbb{E}_{s_0 \sim d_0}\left[V_0^{\pi^e}(s_0) - V_0^{\pi^f}(s_0)\right] \leq \sum_{h=0}^{H-1} \underbrace{\mathbb{E}_{s_h, a_h \sim d_h^{\pi^e}}\left[\mathcal{T} f_{h+1}(s_h, a_h) - f_h(s_h, a_h)\right]}_{\textit{offline error}}$$

$$+ \underbrace{\mathbb{E}_{s_h, a_h \sim d_h^{\pi^f}}\left[f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)\right]}_{\textit{online error}}.$$

Proof:

$$\mathbb{E}_{s \sim d_0}[\max_a f_0(s,a) - V_0^{\pi^f}(s)] = \mathbb{E}_{s \sim d_0}[\mathbb{E}_{a \sim \pi_0^f(s)} f_0(s,a) - V_0^{\pi^f}(s)]$$

$$= \mathbb{E}_{s \sim d_0}[\mathbb{E}_{a \sim \pi_0^f(s)} f_0(s,a) - \mathcal{T} f_1(s,a)] + \mathbb{E}_{s \sim d_0}[\mathbb{E}_{a \sim \pi_0^f(s)} \mathcal{T} f_1(s,a) - V_0^{\pi^f}(s)]$$

$$= \mathbb{E}_{s,a \sim d_0^{\pi^f}}[f_0(s,a) - \mathcal{T} f_1(s,a)] +$$

$$\mathbb{E}_{s \sim d_0}[\mathbb{E}_{a \sim \pi_0^f(s)}[R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(s,a)} \max_{a'} f_1(s',a') - R(s,a) + \mathbb{E}_{s' \sim \mathcal{P}(s,a)} V_1^{\pi^f}(s')]]$$

$$= \mathbb{E}_{s,a \sim d_0^{\pi^f}}[f_0(s,a) - \mathcal{T} f_1(s,a)] + \mathbb{E}_{s \sim d_1^{\pi^f}}[\max_a f_1(s,a) - V_1^{\pi^f}(s)]$$

Apply induction

HyQ Regret

(Realizability and Bellman completeness). *For any h, we have $Q_h^\star \in \mathcal{F}_h$. Additionally, for any $f_{h+1} \in \mathcal{F}_{h+1}$, we have $\mathcal{T} f_{h+1} \in \mathcal{F}_h$.*

Hy-Q ensures that f_h^t small Bellman error under both offline distribution ϑ_h and online distribution $d_h^{\pi^t}$

(Bellman error bound for FQI). *Let $\delta \in (0,1)$, with probability at least $1-\delta$, for any $h \in [H-1]$ and $t \in [T]$,*

$$\left\| f_h^{t+1} - \mathcal{T} f_{h+1}^{t+1} \right\|_{2,\nu_h}^2 \le O\left(\frac{V_{\max}^2 \log(2HT|\mathcal{F}|/\delta)}{t} \right),$$

and

$$\sum_{\tau=1}^{t} \left\| f_h^{t+1} - \mathcal{T} f_{h+1}^{t+1} \right\|_{2,d_h^{\pi^\tau}}^2 \le O\left(V_{\max}^2 \log(2HT|\mathcal{F}|/\delta) \right).$$

Standard concentration arguments

Controlling offline error

(Bellman error bound for FQI). *Let $\delta \in (0,1)$, with probability at least $1-\delta$, for any $h \in [H-1]$ and $t \in [T]$,*

$$\left\|f_h^{t+1} - \mathcal{T} f_{h+1}^{t+1}\right\|_{2,\nu_h}^2 \leq O\left(\frac{V_{\max}^2 \log(2HT|\mathcal{F}|/\delta)}{t}\right),$$

(Bellman error transfer coefficient). *For any policy π, define the transfer coefficient as*

$$C_\pi := \max\left\{0,\ \max_{f\in\mathcal{F}} \frac{\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^\pi}\left[\mathcal{T} f_{h+1}(s,a) - f_h(s,a)\right]}{\sqrt{\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim\nu_h}\left(\mathcal{T} f_{h+1}(s,a) - f_h(s,a)\right)^2}}\right\}.$$

- ratio of the worst-case expected Bellman error under policy π to the expected Bellman error under the offline data
- smaller than previous coverage used for offline FQI

Controlling offline error

(Bellman error bound for FQI). *Let* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *for any* $h \in [H-1]$ *and* $t \in [T]$,

$$\left\| f_h^{t+1} - \mathcal{T} f_{h+1}^{t+1} \right\|_{2,\nu_h}^2 \leq O\left(\frac{V_{\max}^2 \log(2HT|\mathcal{F}|/\delta)}{t} \right),$$

(Bellman error transfer coefficient). *For any policy* π, *define the transfer coefficient as*

$$C_\pi := \max\left\{ 0, \max_{f \in \mathcal{F}} \frac{\sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d_h^\pi} \left[\mathcal{T} f_{h+1}(s,a) - f_h(s,a) \right]}{\sqrt{\sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim \nu_h} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s,a) \right)^2}} \right\}.$$

For each h, with probability at least 1-δ

$$\sum_{t=1}^{T} \mathbb{E}_{s,a \sim d_h^{\pi^e}} \left[\mathcal{T} f_{h+1}^t(s,a) - f_h^t(s,a) \right] \leq \sum_{t=1}^{T} C_{\pi^e} \sqrt{\mathbb{E}_{s,a \sim \nu_h} \left(\mathcal{T} f_{h+1}^t(s,a) - f_h^t(s,a) \right)^2}$$

$$\leq \tilde{O}(\sqrt{T V_{\max}^2 \log(|\mathcal{F}|/\delta)}).$$

Controlling online error

(Bellman error bound for FQI). *Let* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *for any* $h \in [H-1]$ *and* $t \in [T]$,

$$\sum_{\tau=1}^{t} \left\| f_h^{t+1} - \mathcal{T} f_{h+1}^{t+1} \right\|_{2,d_h^{\pi^\tau}}^2 \leq O\big(V_{\max}^2 \log(2HT|\mathcal{F}|/\delta)\big).$$

Under low Bellman rank,

$$\sum_{t=1}^{T} \mathbb{E}_{s,a\sim d_h^{\pi^f}} \left[f_h^t(s,a) - \mathcal{T} f_{h+1}^t(s,a)\right] \leq \sum_{t=1}^{T} \left| \mathbb{E}_{s,a\sim d_h^{\pi^f}} \left[f_h^t(s,a) - \mathcal{T} f_{h+1}^t(s,a)\right] \right| = \sum_{t=1}^{T} \left| \langle X_h(f^t), W_h(f^t) \rangle \right|.$$

Let $\Sigma_h^t := \sum_{\tau=1}^{t} X_h(f^\tau) X_h(f^\tau)^\top + \lambda \mathbb{I}$, we get

$$\sum_{t=1}^{T} \left| \langle X_h(f^t), W_h(f^t) \rangle \right| \leq \sum_{t=1}^{T} \| X_h(f^t) \|_{\Sigma_{t-1;h}^{-1}} \sqrt{\sum_{\tau=1}^{t-1} \mathbb{E}_{s,a\sim d_h^\tau} \left[\left(f_h^t(s,a) - \mathcal{T} f_{h+1}^t(s,a) \right)^2 \right] + \lambda B_W^2}.$$

$O(\sqrt{dT})$

Using elliptical potential lemma

$\tilde{O}(\sqrt{TdV_{\max}^2 \log(|\mathcal{F}|/\delta)})$

Historical Bellman error

HyQ Regret

Given any comparator policy π^e, *for any* $f \in \mathcal{F}$ *and corresponding greedy policy* π^f, *we have with probability at least* $1-\delta$,

$$\sum_{t=1}^{T} V^{\pi^e} - V^{\pi^t} = \widetilde{O}\left(\left(\max\{C_{\pi^e}, 1\} + \sqrt{d}\right) \cdot \sqrt{V_{\max}^2 H^2 T \cdot \log(|\mathcal{F}|/\delta)}\right).$$

Comparison to online RL: Under bilinear model, regret

$$\tilde{O}(\sqrt{d V_{\max}^2 H^2 T \cdot \log(|\mathcal{F}|/\delta)})$$

So hybrid RL worse only by transfer coefficient term.

Computationally regression oracle-efficient!

Sample complexity – no advantage over online RL.