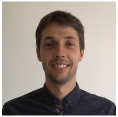
Demonstration-Regularized RL

Daniil Tiapkin[1,2], Denis Belomestny[3,2], Daniele Calandriello[4], Éric Moulines[1,5], Alexey Naumov[2], Pierre Perrault[6], Michal Valko[4], Pierre Ménard[7]

[1]CMAP - CNRS - École Polytechnique - Institut Polytechnique de Paris [2]HSE University
[3]Duisburg-Essen University [4]Google DeepMind [5] Mohamed Bin Zayed University of AI, UAE
[6]IDEMIA [7]ENS Lyon

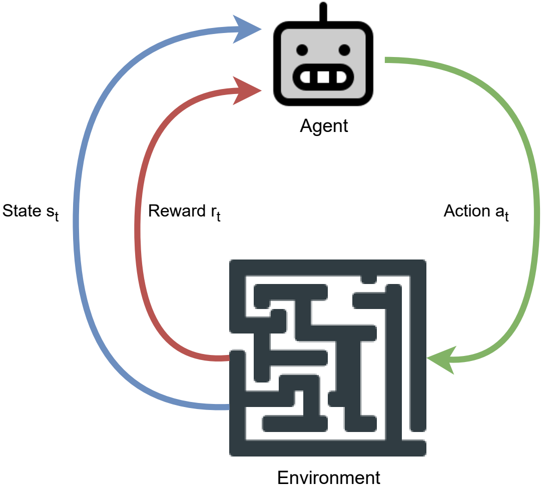
       

Statistical thinking **in the age of AI**

What is the story about?

- How to enhance reinforcement learning with an additional *expert data*?
- How to incorporate both *human preferences* and *expert data*?
  Application: ChatGPT training pipeline;
- Instruments:
  - Behaviour cloning (conditional density estimation);
  - Regularized RL algorithms.

Reinforcement Learning



Markov Decision Processes

We consider an episodic MDP

$$\mathcal{M} = \left(\mathcal{S}, s_1, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}\right)$$

where

- $\mathcal{S}$ is the set of states with initial state s_1;
- $\mathcal{A}$ is the finite set of actions of size A;
- H is the number of steps in one episode;
- $p_h(s'|s,a)$ is the probability transition from state s to state s' by performing action a in step h;
- $r_h(s,a) \in [0,1]$ is the reward obtained by taking action a in state s at step h.

Markov Decision Processes

- A policy π is a collection of functions $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ for all $h \in [H]$. We denote by Π the set of policies.
- The value functions of policy π at step h and state s,

$$V_h^\pi(s; r) = \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right].$$

- Q-functions,

$$Q_h^\pi(s, a) = r_h(s, a) + p_h V_{h+1}^\pi(s, a).$$

- The optimal value functions, denoted by $V_h^\star = \sup_{\pi\in\Pi} V_h^\pi$, are given by the optimal Bellman equations

$$Q_h^\star(s, a) = r_h(s, a) + p_h V_{h+1}^\star(s, a) \qquad V_h^\star(s) = \max_a Q_h^\star(s, a)$$

  where by definition, $V_{H+1}^\star = 0$.
- **Goal of RL**: find the best policy $\pi^\star = \arg\max_{\pi\in\Pi} V_h^\pi$, it could be described as a Dirac measure on maximal Q-value: $\pi_h^\star(s) = \arg\max_{a\in\mathcal{A}} Q_h^\star(s, a)$.

Best Policy Identification

- Before episode $t \in \mathbb{N}$, select a policy $\pi^t = \{\pi_h^t\}_{h\in[H]}$ based on all the data available before episode t;
- During the episode, start $s_1^t = s_1$ and interact with the environment as follows
  1. While in state s_h^t, choose and play action $a_h^t \sim \pi_h^t(s_h^t)$ from the policy;
  2. Receive a reward $r_h(s_h^t, a_h^t)$ and a next state $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$;
  3. Continue with s_{h+1}^t till $h \le H$.
- Decide to stop by the stopping rule is $\iota = t$;
- If agent is stopped, output an output policy $\widehat{\pi}$;

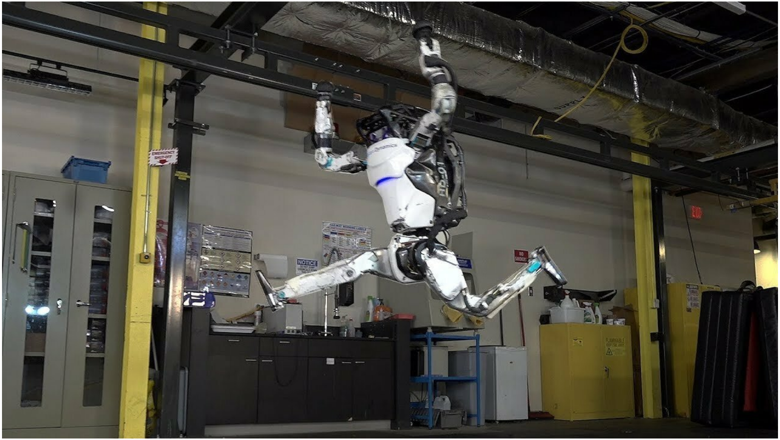
Definition

An algorithm $((\pi^t)_{t\in\mathbb{N}}, \iota, \widehat{\pi})$ is (ε, δ)-PAC for BPI with sample complexity $\mathcal{C}(\varepsilon, \lambda, \delta)$ if

$$\mathbb{P}\Big(V_1^\star(s_1) - V_1^{\widehat{\pi}}(s_1) \le \varepsilon, \quad \iota \le \mathcal{C}(\varepsilon, \delta)\Big) \ge 1 - \delta.$$

Real life: we also have data!

Observation: In the real-life applications we often have a lot of expert data.



Best Policy Identification with demonstration

- Before the interaction with MDP, we are provided an expert (demonstration) dataset

$$\mathcal{D}_{\mathrm{E}} = \{\tau_i = (s_1^i, a_1^i, \ldots, s_H^i, a_H^i),\ i \in [N^{\mathrm{E}}]\}$$

  of N^{E} independent *reward-free* trajectories sampled from a fixed unknown expert policy π^{E}.
- *Interaction phase:* for each episode $t \in \mathbb{N}$, select a policy $\pi^t = \{\pi_h^t\}_{h\in[H]}$ based on all the data available before episode t, *including* $\mathcal{D}_{\mathrm{E}}$.
- Output policy: π^{RL};

Definition

An algorithm $((\pi^t)_{t\in\mathbb{N}}, \iota, \pi^{\mathrm{RL}})$ is (ε, δ)-PAC for BPI with demonstration with sample complexity $\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta)$ if

$$\mathbb{P}\Big(V_1^\star(s_1) - V_1^{\pi^{\mathrm{RL}}}(s_1) \leq \varepsilon, \quad \iota \leq \mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta)\Big) \geq 1 - \delta.$$

Demonstration-Regularized Reinforcement Learning

Demonstration-Regularized RL

Assumption

Assume that the expert policy is close to the optimal $\pi^\star$, that is, $V_1^\star(s_1) - V_1^{\pi^{\mathrm{E}}}(s_1) \le \varepsilon_{\mathrm{E}}$ for some small $\varepsilon_{\mathrm{E}} > 0$.

Idea: reconstruct the expert policy and optimize rewards, staying close to the reconstructed expert policy.

Questions:

- How to reconstruct the expert policy and what guarantees we have?
- How to keep close to the reconstructed expert policy?

Goal: Decrease number of interactions with MDP given large enough dataset;

Behavior Cloning

Setting

In imitation learning, we are provided an expert (demonstration) dataset

$$\mathcal{D}_{\mathrm{E}} \triangleq \{\tau_i = (s_1^i, a_1^i, \ldots, s_H^i, a_H^i),\ i \in [N^{\mathrm{E}}]\}$$

of N^{E} independent *reward-free* trajectories sampled from a fixed unknown expert policy π^{E}.

Objective

Learn from these demonstrations a policy close to the optimal one.

Behavior Cloning (or conditional density estimation)

Empirical minimization

The behavior cloning policy π^{BC} is obtained by minimizing the negative-loglikelihood over a class of policies $\mathcal{F} = \{\pi \in \Pi : \pi_h \in \mathcal{F}_h\}$ with $\mathcal{F}_h$ being a class of conditional distributions $\mathcal{S} \to \mathcal{P}(\mathcal{A})$ and $\mathcal{R}_h$ some regularizer,

$$\pi^{\mathrm{BC}} \in \operatorname*{arg\,min}_{\pi \in \mathcal{F}} \sum_{h=1}^{H} \left(\sum_{i=1}^{N^{\mathrm{E}}} \log \frac{1}{\pi_h(a_h^i | s_h^i)} + \mathcal{R}_h(\pi_h) \right)$$

Trajectory Kullback-Leibler divergence

$$\mathrm{KL}_{\mathrm{traj}}(\pi \| \pi') \triangleq \mathrm{KL}(q^{\pi} \| q^{\pi'}) = \mathbb{E}_{\pi}\left[\sum_{h=1}^{H} \mathrm{KL}(\pi_h(s_h), \pi'_h(s_h)) \right],$$

where $q^{\pi}(\tau) = \pi_1(a_1|s_1) \prod_{h=1}^{H} p_h(s_{h+1}|s_h, a_h) \cdot \pi_h(a_{h+1}|s_{h+1})$.

General Guarantees

- For all $h \in [H]$, there are two positive constants $d_{\mathcal{F}}, R_{\mathcal{F}} > 0$ such that
$$\forall h \in [H], \forall \varepsilon \in (0,1) : \log \mathcal{N}(\varepsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leq d_{\mathcal{F}} \log(R_{\mathcal{F}}/\varepsilon).$$
Moreover, there is a constant $\gamma > 0$ such that for any $h \in [H]$, $\pi_h \in \mathcal{F}_h$ it holds $\pi_h(a|s) \geq \gamma$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.
- There is a constant $\kappa \in (0, 1/2)$ such that a κ-greedy version of the expert policy defined by $\pi_h^{\mathrm{E},\kappa}(a|s) = (1-\kappa)\pi_h^{\mathrm{E}}(a|s) + \kappa/A$ belongs to the hypothesis class of policies: $\pi^{\mathrm{E},\kappa} \in \mathcal{F}$.

Theorem

Let assumptions above be satisfied and let $0 \leq \mathcal{R}_h(\pi_h) \leq M$ for all $h \in [H]$ and any policy $\pi \in \mathcal{F}_h$. Then with probability at least $1-\delta$, the behavior policy π^{BC} satisfies
$$\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}}) \leq \frac{6 d_{\mathcal{F}} H \cdot (\log(A\mathrm{e}^3/(A\gamma \wedge \kappa)) \cdot \log(2HN^{\mathrm{E}} R_{\mathcal{F}}/(\gamma\delta))}{N^{\mathrm{E}}} + \frac{2HM}{N^{\mathrm{E}}} + \frac{18\kappa}{1-\kappa}.$$

Special Case: Finite MDPs

Finite MDPs

For all $N^{\mathrm{E}} \geq A$, the class of policies

$$\mathcal{F} = \{\pi \in \Pi : \pi_h(a|s) \geq 1/(N^{\mathrm{E}} + A)\}$$

and the regularizer

$$\mathcal{R}_h(\pi_h) = \sum_{s,a} \log(1/\pi_h(a|s)),$$

it holds with probability at least $1 - \delta$,

$$\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}}) \leq \frac{6\color{red}{SAH}\color{black} \cdot \log(2\mathrm{e}^4 N^{\mathrm{E}}) \cdot \log(12H(N^{\mathrm{E}})^2/\delta)}{N^{\mathrm{E}}} + \frac{18AH}{N^{\mathrm{E}}}.$$

Lower bound

$$\min_{\widehat{\pi}} \max_{\pi \in \mathcal{F}} \mathbb{E}_{\tau_1,\ldots,\tau_{N^{\mathrm{E}}} \sim \pi}[\mathrm{KL}_{\mathrm{traj}}(\pi \| \widehat{\pi})] \geq \frac{\color{red}{SAH}}{128 N^{\mathrm{E}} \log(\mathrm{e}^2(N^{\mathrm{E}} + A))}.$$

Special case: Linear MDPs

Assumptions

For $\varepsilon > 0$ and $\delta \in (0,1)$, assume that an expert policy π^{E} is $\varepsilon/8$-optimal and for all $h \in [H]$, there exists an *unknown* parameter $w_h^{\mathrm{E}} \in \mathbb{R}^d$ with $\|w_h^{\mathrm{E}}\|_2 \leq R$ for some known $R \geq 0$ such that

$$\pi_h^{\mathrm{E}}(a|s) = \frac{\exp(\psi(s,a)^\intercal w_h^{\mathrm{E}})}{\sum_{a' \in \mathcal{A}} \exp(\psi(s,a')^\intercal w_h^{\mathrm{E}})}$$

Consider

$$\mathcal{F}_h = \left\{ \pi_h(a|s) = \frac{\kappa}{A} + (1-\kappa)\frac{\exp(\psi(s,a)^\intercal w_h)}{\sum_{a' \in \mathcal{A}} \exp(\psi(s,a')^\intercal w_h)} : w_h \in \mathbb{R}^d,\ \|w_h\|_2 \leq R \right\}.$$

Corollary

Under assumption above, the function class $\mathcal{F}$ defined above and regularizer $\mathcal{R}_h = 0$ for all $h \in [H]$, it holds for all $N^{\mathrm{E}} \geq A$ with probability at least $1-\delta$,

$$\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}}) \leq \frac{8dH \cdot (\log(2\mathrm{e}^3 A N^{\mathrm{E}}) \cdot (\log(48(N^{\mathrm{E}})^2 R) + \log(H/\delta)))}{N^{\mathrm{E}}} + \frac{18AH}{N^{\mathrm{E}}}.$$

Demonstration-Regularized RL

Implementation of the initial idea:

1. Perform behavior cloning and compute π^{BC};
2. Solve RL problem with an additional regularization $\lambda \cdot \mathrm{KL}_{\mathrm{traj}}(\pi \| \pi^{\mathrm{BC}})$

Algorithm:

1: **Input:** Precision parameter $\varepsilon_{\mathrm{RL}}$, probability parameter δ_{RL}, demonstrations $\mathcal{D}_{\mathrm{E}}$, regularization parameter λ.
2: Compute behavior cloning policy $\pi^{\mathrm{BC}} = \texttt{BehaviorCloning}(\mathcal{D}_{\mathrm{E}})$.
3: Perform regularized BPI $\pi^{\mathrm{RL}} = \texttt{RegBPI}(\pi^{\mathrm{BC}}, \lambda, \varepsilon_{\mathrm{RL}}, \delta_{\mathrm{RL}})$
4: **Output:** policy π^{RL}.

Regularized best policy identification (BPI)

Setting

Given some reference policy $\widetilde{\pi}$ and some regularization parameter $\lambda > 0$, we consider the trajectory Kullback-Leibler divergence regularized value function

$$V^{\pi}_{\widetilde{\pi},\lambda,1}(s_1) = V^{\pi}_{1}(s_1) - \lambda \mathrm{KL}_{\mathrm{traj}}(\pi \| \widetilde{\pi}).$$

In this value function, the policy π is penalized for moving too far from the reference policy $\widetilde{\pi}$.

Bellman's equations

$$\begin{aligned} Q^{\pi}_{\widetilde{\pi},\lambda,h}(s,a) &= r_h(s,a) + p_h V^{\pi}_{\widetilde{\pi},\lambda,h+1}(s,a) \\ V^{\pi}_{\widetilde{\pi},\lambda,h}(s) &= \pi_h Q^{\pi}_{\widetilde{\pi},\lambda,h}(s) - \lambda\,\mathrm{KL}(\pi_h(s) \| \widetilde{\pi}_h(s))\,, \end{aligned}$$

where $V^{\pi}_{\widetilde{\pi},\lambda,H+1} = 0$.

Best Policy Identification in Regularized Finite MDPs

Optimistic planning in a regularized MDP

$$\overline{Q}_h^t(s,a) = \operatorname{clip}\Big(r_h(s,a) + \widehat{p}_h^t \overline{V}_{h+1}^t(s,a) + b_h^{p,t}(s,a), 0, H\Big)\,,$$
$$\overline{V}_h^t(s) = \max_{\pi\in\Delta_A}\Big\{\pi\overline{Q}_h^t(s) - \lambda\,\mathrm{KL}(\pi\|\widetilde{\pi}_h(s))\Big\}\,,$$
$$\bar{\pi}_h^{t+1}(s) = \operatorname*{arg\,max}_{\pi\in\Delta_A}\Big\{\pi\overline{Q}_h^t(s) - \lambda\,\mathrm{KL}(\pi\|\widetilde{\pi}_h(s))\Big\}\,,$$

with $\overline{V}_{H+1}^t = 0$ by convention, where $\widehat{p}^t$ is an estimate of the transition probabilities. Here $b^{p,t}$ is some bonus term taking into account estimation error for transition probabilities.

Best Policy Identification in Regularized Finite MDPs

Sampling rule

For $h' \in [0, H]$, the policy $\pi^{t,(h')}$ first follows the optimistic policy $\bar{\pi}^t$ until step h where it selects an action leading to the largest confidence interval for the optimal Q-value,

$$\pi_h^{t,(h')}(a|s) = \begin{cases} \pi_h^{t,(h')}(a|s) = \bar{\pi}_h^t(a|s) & \text{if } h \neq h' \\ \pi_h^{t,(h')}(a|s) = \mathbb{1}\left\{ a \in \arg\max_{a' \in \mathcal{A}} (\overline{Q}_h^t(s,a') - \underline{Q}_h^t(s,a')) \right\} & \text{if } h = h' \end{cases}$$

where $\underline{Q}^t$ is a lower bound on the optimal regularized Q-value function. The sampling rule is obtained by picking up uniformly at random one policy among the family $\pi^t = \pi^{t,(h')}$, $h' \sim \mathrm{Unif}[0, H]$.

Best Policy Identification in Regularized Finite MDPs

Stopping rule and decision rule

First recursively build an upper-bound on the difference between the value of the optimal policy and the value of the current optimistic policy $\bar{\pi}^t$,

$$W_h^t(s,a) = \left(1 + \frac{1}{H}\right)\widehat{p}_h^t G_{h+1}^t(s) + b_h^{\mathrm{gap},t}(s,a),$$

$$G_h^t(s) = \mathrm{clip}\left(\bar{\pi}_h^{t+1} W_h^t(s) + \frac{1}{2\lambda}\max_{a\in\mathcal{A}}\left(\overline{Q}_h^t(s,a) - \underline{Q}_h^t(s,a)\right)^2, 0, H\right),$$

where $b_h^{\mathrm{gap},t}$ is a bonus, $\underline{V}^t$ is a lower-bound on the optimal value function and $G_{H+1}^t = 0$ by convention.
The stopping time $\iota = \inf\{t \in \mathbb{N} : G_1^t(s_1) \le \varepsilon\}$. At this episode ι we return the policy $\widehat{\pi} = \bar{\pi}^\iota$.

Best Policy Identification in Regularized Finite MDPs

Input: Target precision ε, target probability δ, bonus functions $b^t, b^{t,\mathrm{KL}}$.
while true **do**
  Compute $\bar{\pi}^t$ by optimistic planning.
  Compute bound on the gap $G_1^t(s, a)$.
  if $G_1^t(s_1) \le \varepsilon$ **then break**
  Sample $h' \sim \mathrm{Unif}[H]$ and set $\pi^t = \pi^{t,(h')}$.
  for $h \in [H]$ **do**
    Play $a_h^t \sim \pi_h^t(s_h^t)$
    Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$
  end for
  Update transition estimates $\widehat{p}^t$.
end while
Output policy $\widehat{\pi} = \bar{\pi}^t$.

Final sample complexity for Demonstration-Regularized RL

Theorem

Assume that the expert policy is $\varepsilon_{\mathrm{E}} = \varepsilon/2$-optimal and satisfies some assumption in the linear case. Let π^{BC} be the behavior cloning policy, then demonstration-regularized RL with parameters $\varepsilon_{\mathrm{RL}} = \varepsilon/4$, $\delta_{\mathrm{RL}} = \delta/2$ and $\lambda = \widetilde{\mathcal{O}}\big(N^{\mathrm{E}}\varepsilon/(SAH)\big) \,/\, \widetilde{\mathcal{O}}\big(N^{\mathrm{E}}\varepsilon/(dH)\big)$ is (ε, δ)-PAC for BPI with demonstration in finite / linear MDPs and has sample complexity of order

$$\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 S^3 A^2}{N^{\mathrm{E}}\varepsilon^2}\right) \text{ (finite)} \qquad \mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 d^3}{N^{\mathrm{E}}\varepsilon^2}\right) \text{ (linear)}.$$

Demonstration-Regularized Reinforcement Learning with Human Feedback

Preference-Based Model

Setting

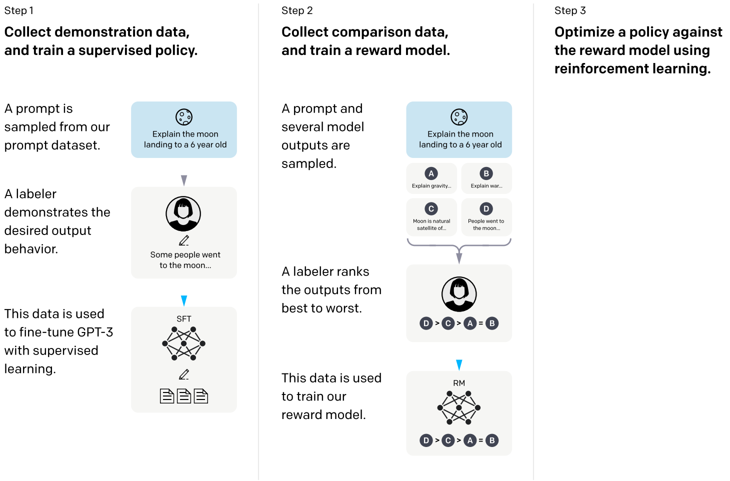
We do not know the true reward function $r^\star$ but have access to an oracle that provides a preference feedback between two trajectories.

Reward

Given a reward function $r = \{r_h\}_{h=1}^H$, we define the reward of a trajectory $\tau \in (\mathcal{S} \times \mathcal{A})^H$ as the sum of rewards collected over this trajectory $r(\tau) \triangleq \sum_{h=1}^H r_h(s_h, a_h)$.

Instruct GPT

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in Neural Information Processing Systems 35 (2022): 27730-27744.



Preference-based RL with demonstration

Assumption

Let τ_0, τ_1 be two trajectories. The preference for τ_1 over τ_0 is a Bernoulli random variable o with a parameter $q_\star(\tau_0, \tau_1) = \sigma(r^\star(\tau_1) - r^\star(\tau_0))$, where $\sigma\colon \mathbb{R} \to [0, 1]$ is a monotone increasing link function that satisfies $\inf_{x \in [-H,H]} \sigma'(x) = 1/\zeta$ for $\zeta > 0$. This function can also be viewed as a utility or preference.

Example

A sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$ leads to the Bradley-Terry-Luce (BTL) model widely used in the literature.

Preference-based RL with demonstration

The agent observes N^{E} independent trajectories $\mathcal{D}_{\mathrm{E}}$ sampled from an expert policy π^{E}. Learning is divided in two phases.

- *Preference collection.* Based on the observed expert trajectories $\mathcal{D}_{\mathrm{E}}$, the agent selects a sampling policy π^{BC} to generate a *data set of preferences* $\mathcal{D}_{\mathrm{RM}} = \{(\tau_0^k, \tau_1^k, o^k)\}_{k=1}^{N^{\mathrm{RM}}}$ consisting of pairs of trajectories and the sampled preferences.

- *Reward-free interaction.* The agent interacts with the reward-free MDP as follows: at episode t, agent selects a policy π^t based on *the collected transitions up to time t, demonstrations and preferences*. Then a new trajectory (reward-free) is sampled following the policy π^t and is observed by the agent.

  At the end of each episode, the agent can decide to stop according to a stopping rule ι, and outputs a policy π^{RLHF}.

Preference-based RL with demonstration

PAC algorithm

An algorithm $((\pi^t)_{t\in\mathbb{N}}, \pi^{\mathrm{BC}}, \iota, \pi^{\mathrm{RLHF}})$ is (ε, δ)-PAC for preference-based best policy identification (BPI) with demonstrations and sample complexity $\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta)$ if

$$\mathbb{P}\Big(V_1^\star(s_1) - V_1^{\pi^{\mathrm{RLHF}}}(s_1) \le \varepsilon,\ \iota \le \mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta)\Big) \ge 1 - \delta,$$

where the unknown true reward function $r^\star$ is used in the value-function $V^\star$.

Demonstration-regularized RLHF

Idea

The agent starts with behavior cloning applied to the expert dataset, resulting in the policy π^{BC}. During the preference collection phase, the agent generates a dataset $\mathcal{D}_{\mathrm{RM}} = \{(\tau_0^k, \tau_1^k, o^k)\}_{k=1}^{N^{\mathrm{RM}}}$ by executing the previously computed policy π^{BC}. Using this dataset, the agent can infer the reward $\hat{r}$ via MLE:

$$\max_{r\in\mathcal{G}} \sum_{k=1}^{N^{\mathrm{RM}}} o^k \log\left(\sigma\left(r(\tau_1^k) - r(\tau_0^k)\right)\right) + (1-o^k)\log\left(1-\sigma\left(r(\tau_1^k) - r(\tau_0^k)\right)\right)$$

where $\mathcal{G}$ is a function class for trajectory reward functions.

Finally, the agent computes π^{RL} by performing regularized BPI with policy π^{BC}, a properly chosen regularization parameter λ and the estimated reward $\hat{r}$.

Demonstration-regularized RLHF

Role of π^{BC}

We use the behavior cloning policy π^{BC} for two purpose.

1. First, it allows efficient offline collection of the preference dataset $\mathcal{D}_{\mathrm{RM}}$, from which a high-quality estimate of the reward can be derived.
2. Second, a regularization towards the behavior cloning policy π^{BC} enables the injection of information obtained from the demonstrations.

Demonstration-regularized RLHF

Connection RL fine-tuning for LLM

Our algorithm's policy learning phase is similar to solving an RL problem with policy-dependent rewards

$$r_h^{\mathrm{RLHF}}(s,a) = \hat{r}_h(s,a) - \lambda \log\bigl(\pi_h^{\mathrm{RLHF}}(a|s)/\pi_h^{\mathrm{BC}}(a|s)\bigr)\,.$$

This formulation, coupled with our prior stages of the behavior cloning, akin to supervised fine-tuning (SFT), and reward estimation through MLE based on trajectories generated by the SFT-policy, mirrors a simplified version of the three-phase GPT RLHF pipeline.

Demonstration-regularized RLHF

Single-policy concentrability coefficient $C_r(\mathcal{G}, \pi^{\mathrm{E}}, \pi^{\mathrm{BC}})$

$$\max\left\{0, \sup_{r \in \mathcal{G}} \frac{\mathbb{E}_{\tau_0 \sim \pi^{\mathrm{E}}, \tau_1 \sim \pi^{\mathrm{BC}}}[r^\star(\tau_0) - r^\star(\tau_1) - (r(\tau_0) - r(\tau_1))]}{\sqrt{\mathbb{E}_{\tau_0, \tau_1 \sim \pi^{\mathrm{BC}}}\left[(r^\star(\tau_0) - r^\star(\tau_1) - (r(\tau_0) - r(\tau_1)))^2\right]}}\right\}$$

Assumptions

For $\varepsilon > 0$ and $\delta \in (0,1)$, assume that an expert policy π^{E} is $\varepsilon/8$-optimal and for all $h \in [H]$, there exists an *unknown* parameter $w_h^{\mathrm{E}} \in \mathbb{R}^d$ with $\|w_h^{\mathrm{E}}\|_2 \leq R$ for some known $R \geq 0$ such that

$$\pi_h^{\mathrm{E}}(a|s) = \frac{\exp(\psi(s,a)^\intercal w_h^{\mathrm{E}})}{\sum_{a' \in \mathcal{A}} \exp(\psi(s,a')^\intercal w_h^{\mathrm{E}})}$$

Demonstration-regularized RLHF

Theorem

If the following two conditions hold

$$N^{\mathrm{E}} \cdot N^{\mathrm{RM}} \geq \widetilde{\Omega}\Big(\zeta^2 H^2 \widetilde{D}^2/\varepsilon^2\Big)$$

$$N^{\mathrm{E}} \geq \widetilde{\Omega}\Big(H^2 \widetilde{D}/\varepsilon\Big) \text{ or } N^{\mathrm{RM}} \geq \widetilde{\Omega}\Big(C_r \zeta^2 H \widetilde{D}/\varepsilon^2\Big)$$

for $\widetilde{D} = SA \,/\, d$ in finite / linear MDPs, then demonstration-regularized RLHF is (ε, δ)-PAC for BPI with demonstration in finite / linear MDPs with sample complexity

$$\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 S^3 A^2}{N^{\mathrm{E}} \varepsilon^2}\right) \text{ (finite)} \qquad \mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 d^3}{N^{\mathrm{E}} \varepsilon^2}\right) \text{ (linear)}$$

Demonstration-regularized RLHF

Remarks

- The conditions (1) and (2) control two different terms in the reward estimation error presented.
- The condition (1) shows that small size of the expert dataset should be compensated by a larger dataset used for reward estimation and vice versa.
- The condition (2) requires that at least one of these datasets is large enough to overcome sub-exponential behavior of the error in the reward estimation problem.
- The second part of the condition (2) $N^{\mathrm{RM}} \geq C_r/\varepsilon^2$ is unavoidable in the general case of offline learning even if the transitions are known due to a lower bound.
- As soon as reward estimation error is small enough, we obtain the same sample complexity guarantees as in the demonstration-regularized RL.

Takeaways & Open problems

Combine almost known 3 statistical problem $\mapsto$ real-world problem;

- Reinforcement Learning $\mapsto$ Reinforcement Learning with Demonstrations;
- Simple and implementable approach: Demonstration-Regularized RL;
- Incorporation human feedback $\mapsto$ InstructGPT pipeline;

Open questions

- Optimal sample complexity for the regularized BPI?
- Optimal sample complexity for the BPI with demonstrations?
- Optimal sample complexity for RLHF?