HyperFQI: Efficient and Scalable RL via Hypermodel

Yingru Li

yingruli@link.cuhk.edu.cn
The Chinese University of Hong Kong, Shenzhen

December 16, 2023

Efficiency Challenges in RL

- **Data-efficiency**: Collecting data can be expensive and time-consuming.
- **Computational-efficiency**: Training Deep RL costs weeks or even months.

AlphaGo Zero as an example:

- 29 million ($> 10^7$) games of self-play training over 40 days.
- **Huge costs**: Replication would cost $\approx$ \$35,354,222
- Energy inefficient, High carbon emission, Unsustainable

‘AGI for humanity’ calls for Efficient RL

Figure: Economic Impact

Figure: Sustainability

Figure: Access and Equity

Figure: Democracy

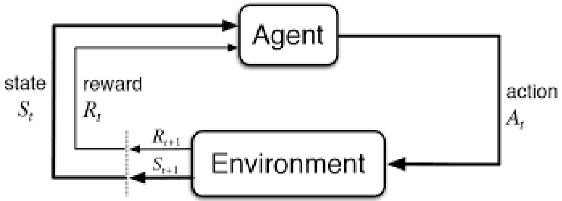
Efficiency improvements in RL pave the way for AGI that is economically viable, sustainable, accessible to all, and developed in a more democratic and inclusive manner, **ultimately benefiting humanity as a whole**.

Importance of RL efficiency

Solving efficiency challenges in RL
is the key to achieve AGI for humanity.

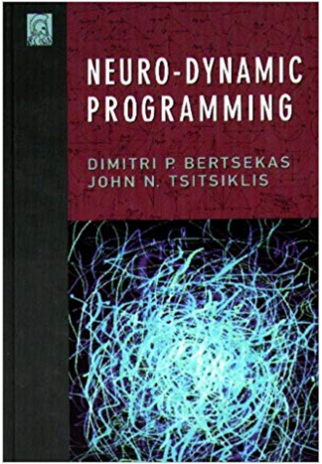
Data-efficiency
Computational-efficiency

Mathematical formulation of episodic RL problem



- **MDP**: $(\mathcal{S}, \mathcal{A}, P, R, s_{\text{terminal}}, \rho)$
  - $\mathcal{S}$: state space; $\mathcal{A}$: action space P: transition probability; R: reward function;
  - s_{terminal}: terminal state; ρ: initial state distribution
- Let τ be the hitting time when reaching terminal state s_{terminal}.
- **Episodic RL**: The agent interacts with the environment for a finite number of episodes.
- **Goal**: Find a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected total return

$$\max_{\pi} \mathbb{E}_{\pi}[\sum_{t=1}^{\tau} R(S_t, A_t)]. \tag{1}$$



- **Action-value function (Q-function)**:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}[\sum_{t=1}^{\tau} R(S_t, A_t) \mid S_1 = s, A_1 = a]$$

- **Greedy policy**: $\pi^{Q^{\pi}}(s) = \arg\max_{a \in \mathcal{A}} Q^{\pi}(s,a)$
- Agent's behavior is determined by the greedy policy π^{Q} w.r.t. a given Q-function.
- **Function approximation**: when state space is large, we use some function (say neural networks) to approximate the Q-function:

$$Q_{\theta}(s,a) \approx Q^{\pi}(s,a)$$

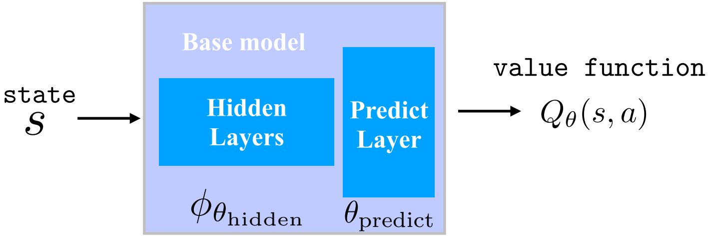
where θ is the parameter of the function.

Our solution: HyperFQI for randomized value function

HyperFQI includes Two models:

- Base model: DQN-type structure (Nature 15')

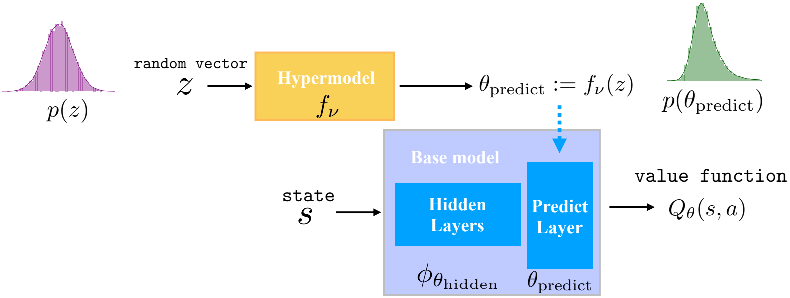
$$Q_\theta(s,a) = \langle \phi_{\theta_{\text{hidden}}}(s), \theta_{\text{predict}}(a) \rangle.$$



Our solution: HyperFQI for randomized value function

HyperFQI includes Two models:

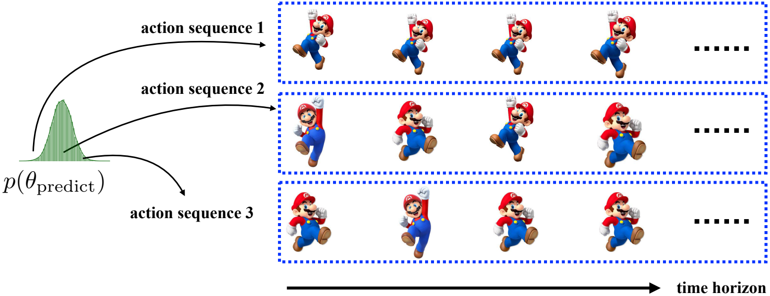
- Base model: DQN-type structure $Q_\theta(s,a) = \langle \phi_{\theta_{\text{hidden}}}(s), \theta_{\text{predict}}(a) \rangle$.
- Hypermodel: $\theta_{\text{predict}} = f_\nu(z)$ where $z \sim p(z)$. $p(z)$ is a fixed reference distribution.



Resulting model: $Q_{\theta_{\text{hidden}}, f_\nu(z)}(s,a)$ is a randomized value function depends on (s,a) and additional random variable z.

Diverse Action Sequences empowers Smart Data Collection Strategy

- In each episode, sample $z \sim p(z), \theta_{\text{predict}} = f_\nu(z)$ and use greedy policy w.r.t. randomized value function $\arg\max_a Q_{\theta_{\text{hidden}}, f_\nu(z)}(s, a)$.



- After the episode k, the agent would collect the behavior trajectory $\mathcal{O}_k = (S_{k,0}, A_{k,0}, R_{k,1}, \ldots, S_{k,\tau_k-1}, A_{k,\tau_k-1}, R_{k,\tau_k})$ into data buffer $\mathcal{D}$.

HyperFQI Adaptation with Data: Training Objective

Training objective in HyperFQI is a novel extension of fitted Q-iteration (FQI):

$$\min_{\nu,\theta_{\text{hidden}}} \int_z p(z) \Bigg[\sum_{(s,a,r,\xi,s')\in\mathcal{D}} \left(Q_{\text{target}}(s',z') + \sigma_\omega z^\top \xi - Q_{\text{prediction}}(s,a,z)\right)^2 + \frac{\sigma_\omega^2}{\sigma_p^2} \| f_\nu(z) - f_{\nu_{\text{prior}}}(z) \|^2 \Bigg] (\mathrm{d}z), \tag{2}$$

where

$$\begin{aligned} Q_{\text{prediction}}(s,a,z) &= Q_{\theta_{\text{hidden}}, f_\nu(z)}(s,a), \\ Q_{\text{target}}(s',z') &= r + \gamma \max_{a'} \| Q_{\bar{\theta}_{\text{hidden}}, f_{\bar{\nu}}(z')}(s',a') \|. \end{aligned} \tag{3}$$

- The augmented data $\xi \in \mathbb{R}^M$ is a artificially generated random vector, together with term $\sigma_\omega z^\top \xi$, for posterior approximation.
- **Joint Feature Learning and Uncertainty quantification** through Equation equation 2.

Key innovations and understandings of HyperFQI

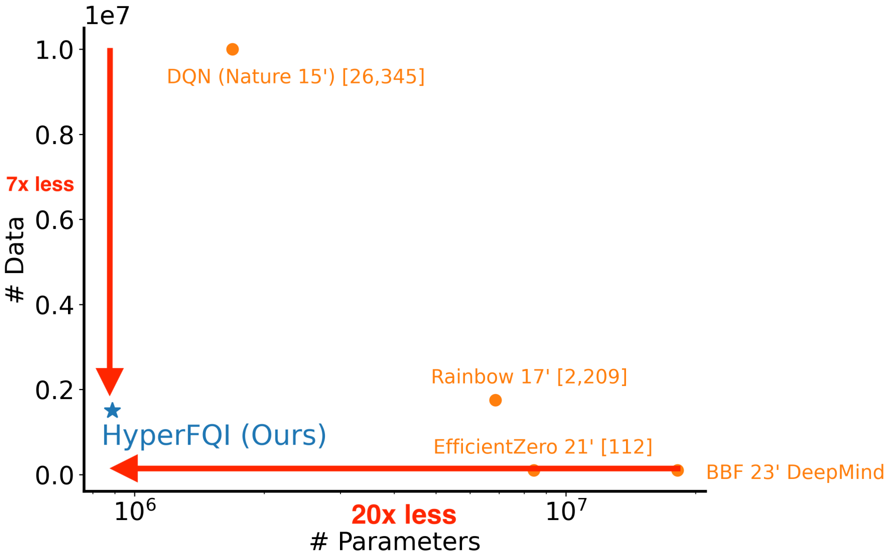
- **Hypermodel**: A novel model architecture that enables computational-efficient way of tracking the (approximate) posterior distribution of value function.
- **HyperFQI**: A novel algorithm that enables efficient RL.
  - **Smart data collection**: Diverse action sequences.
  - **Smart data usage**: Joint feature learning and uncertainty quantification.
- **Understanding**:
  - From the (approximate) posterior distribution of randomized value function, all plausible action sequences can be sampled for exploration using randomized value.
  - As more data accumulated, with the training objective, the posterior distribution of randomized value function would concentrate on the true optimal value function.

Benchmark problem in Reinforcement Learning

Figure: Human-level control via deep reinforcement learning (Nature 15'). Citations: 26,345.

- Arcade Learning Environment (ALE) (Bellemare et al. 2013): 57 Atari 2600 games.
- **State space**: raw pixel images.
- **Action space**: 18 actions.
- **Reward**: game score.
- **Goal**: Achieve human-level performance in Atari benchmark.

Data and computation efficiency in Deep RL benchmarks



Theoretical Guarantees for HyperFQI in Tabular Setting

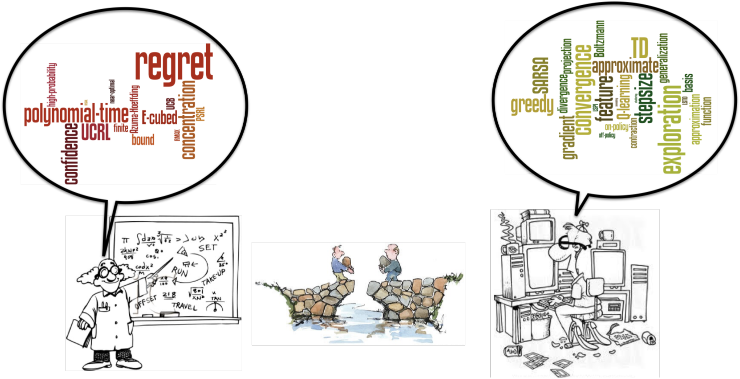
- Performance metric: **Regret** (the cumulative difference between the expected return of the optimal policy and the expected return of the learned policy).
- Finite horizon time-inhomogeneous class of MDPs.
  - # of states: $|\mathcal{S}|$
  - # of actions: $|\mathcal{A}|$
  - Problem horizons: H
  - # of episodes: K
- **Data-efficiency**: Regret upper bound $\tilde{O}(H^2\sqrt{|\mathcal{S}||\mathcal{A}|K})$ nearly match the lower bound (fundamental statistical limits) of the problem class.
- **Computational-efficiency**: The additional computation burden of HyperFQI than single point estimate is only **logarithmic** in $|\mathcal{S}|$ and $|\mathcal{A}|$ and K, i.e. the additional model dimension is $M = \tilde{O}(\log(|\mathcal{S}||\mathcal{A}|K))$

The novelty and difficulty in the mathemtical analysis

$$
\begin{tikzcd}
 & \text{Time } t=1 & \text{Time } t=2 & \text{Time } t=3 & & \text{Time } t & \text{Time } t+1 \\
 & x_1 \arrow[r] & x_2 \arrow[r] & x_3 \arrow[r] & \cdots \arrow[r] & x_t \arrow[r] & x_{t+1} \\
\mathbf{Z}_0 \arrow[ur] & \mathbf{z}_1 \arrow[ur] & \mathbf{z}_2 \arrow[ur] & \mathbf{z}_3 \arrow[ur] & \cdots \arrow[ur] & \mathbf{z}_t \arrow[ur] &
\end{tikzcd}
$$

The analysis is build upon a novel probability tool: non-asymptotic analysis of **sequential random projection**.

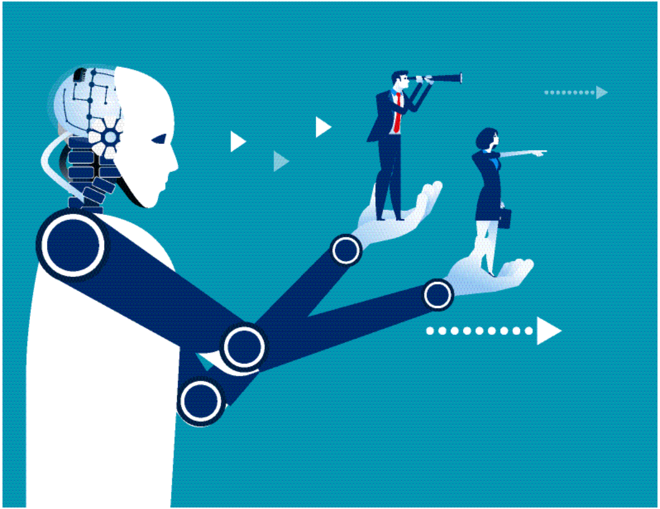
- **Difficulty**: sequential dependence of high-dimensional random variables due to the sequential nature of RL.
- **Novel solution**: A smart construction of stopped martingale and the application of 'method of mixtures' in self-normalized martingale.
- **No prior art.**



HyperFQI is the first principled RL agent that is

- **Provably efficient** in terms of both data and computation via rigorous analysis.
- **Practically scalable** to complex environments, and compatible with existing RL frameworks.

Solving efficiency challenges in RL and paving a way of AGI for Humanity



Thanks to the collaborators during this line of works

(a) Ziniu Li
CUHK(SZ)

(b) Jiawei Xu
CUHK(SZ)

(c) Tong Zhang
HKUST $\Rightarrow$ UIUC

(d) Zhi-Quan Luo
CUHK(SZ)