Artificial General Intelligence for *Humanity* through *Efficient Reinforcement Learning*

Yingru Li

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

October 21, 2023

Graduate Research Forum

Abstract

Abstract: We embark on a compelling journey towards Artificial General Intelligence (AGI) and emphasize its profound impact on humanity. We begin by defining AGI and its transformative potential, underlining the central role of Reinforcement Learning (RL) in achieving this aspiration. We explore the real-world applications of RL, from plasma control to ChatGPT, shedding light on the pressing need for efficient RL algorithms. Enter HyperFQI, an innovative solution to RL efficiency challenges we developed, boasting generality and scalability. Witness its remarkable efficiency in benchmark results, particularly in Atari video games. Discover the practical integration of HyperFQI, adapting seamlessly into existing RL frameworks. Delve into the theoretical guarantees of HyperFQI in tabular settings, featuring rigorous mathematical probability tools we developed. This presentation bridges theory and practice, elucidating HyperFQI's pivotal role in the expedition toward AGI, with a direct impact on realizing AGI's potential for the betterment of humanity. The talk concludes by underscoring the transformative potential of efficient RL agents and their promise for the future of AGI and, indeed, humanity.

Artificial General Intelligence (AGI)

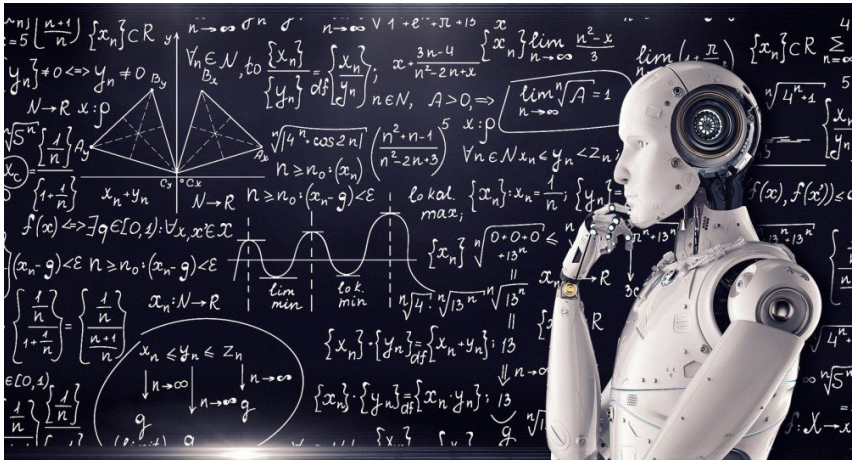
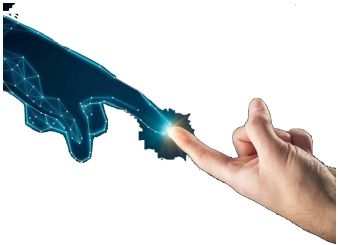


Figure: Artificial General Intelligence (AGI) is the intelligence of a machine **(agent)** that could successfully perform any intellectual task that a human being can or even smarter than human-beings.

The promise of AGI: Benefit all of humanity

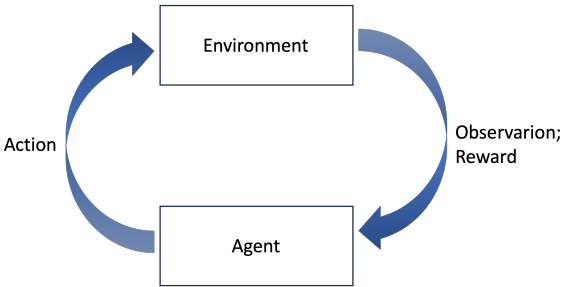
Turbocharge Economy

Aiding Scientific Discovery



Elevate Humanity

Reinforcement Learning: Fundamental methodology for AGI agents



Sequential decision-making under **uncertainty**.

Goal: **Sequentially taking actions** to maximize the **long-term** reward.

Reinforcement Learning (RL): Huge Increase in Interest

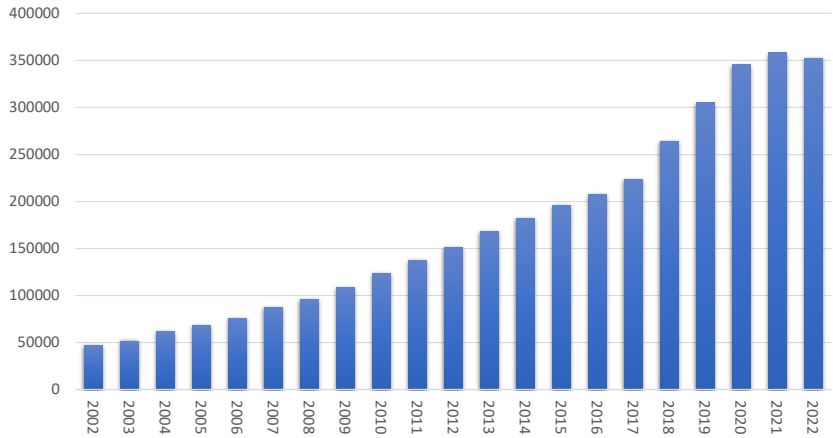


Figure: Growth of published "reinforcement learning" related papers and articles. (Data from Semantic Scholar.)

Outline

RL for Science: Learning Plasma Control for Fusion Science

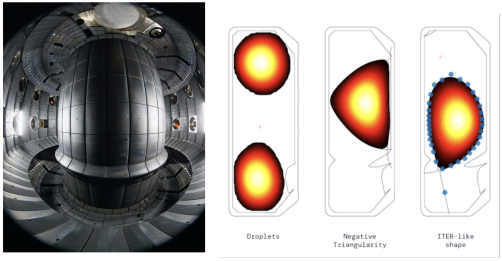


Figure: Image credits: left Alain Herzog / EPFL, right DeepMind & SPC/EPFL. Degrave et al. **Nature 22'**. "Magnetic control of tokamak plasmas through deep reinforcement learning."

- Action: Control Coil Voltages
- Observation: Physical measurements
- Uncertainty: Unknown physical dynamics in real system
- Goal
  1. Keep the Plasma Alive;
  2. Stabilize the plasma location;
  3. Shape Control.
- Ultimate Goal: **Sustainable Fusion Energy**

RL for Technology: Chip design

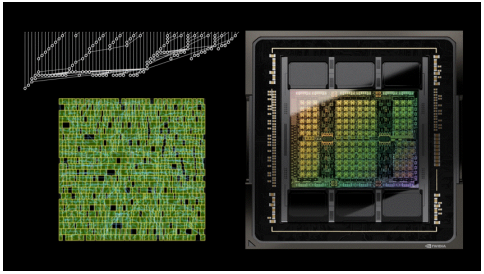


Figure: Designing Arithmetic Circuits with Deep Reinforcement Learning, Nvidia. https://developer.nvidia.com/blog/designing-arithmetic-circuits-with-deep-reinforcement-learning/

- Action: Sequentially modify the circuit design
- Observation: A graph representing circuit
- Uncertainty: Unknown circuit delay/area (Use circuit synthesis as simulator to gain information)
- Goal
  1. **Small**: A lower area so that more circuits can fit on a chip.
  2. **Fast**: A lower delay to improve the performance of the chip.
  3. **Energy-saving**: A lower power consumption of the chip.

RL for Economy: Ride-hailing Order Dispatching (Matching)

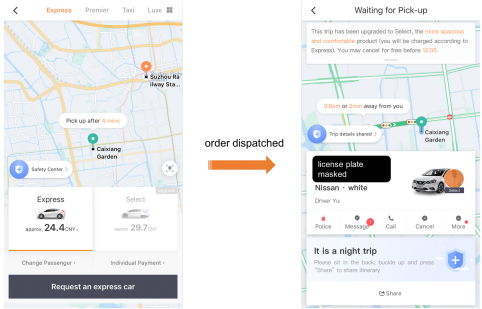


Figure: Ride-hailing Order Dispatching on DiDi via Reinforcement Learning. 2019 Informs Wagner Prize Winner; 2022 KDD paper for real-world deployment.

- Action: match user and driver
- Observation: orders and completed traffic routes
- Uncertainty: unknown traffic dynamics and user behaviors
- Goal
  1. Maximize **driver income**
  2. Maximize **order completion rate**
- Ultimate goal: **efficient urban mobility and shared economy**

RL for Health: Efficient and targeted COVID-19 border testing via RL

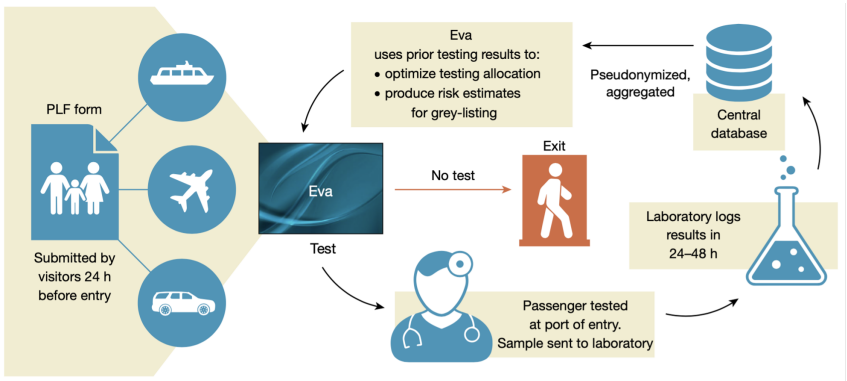


Figure: Bastani et al. **Nature 21'** https://www.nature.com/articles/s41586-021-04014-z

More applications of RL

Figure: RLHF for chatbot: **ChatGPT**.

Figure: RL for Games: AlphaGo.

Outline

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

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**.

‘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**.

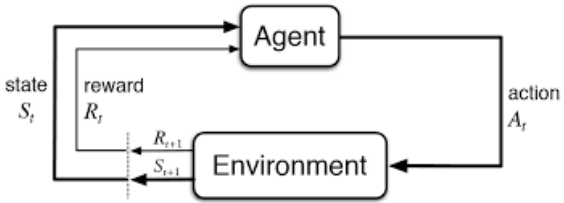
Importance of RL efficiency

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

Data-efficiency
Computational-efficiency

Outline

Mathematical formulation of episodic RL problem

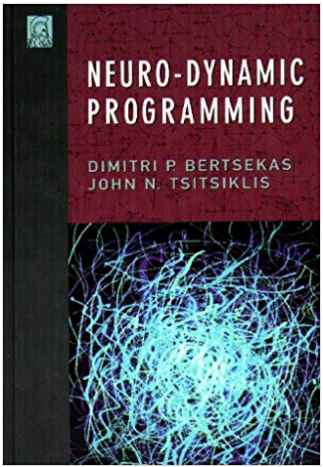


- **MDP**: $(\mathcal{S}, \mathcal{A}, P, R, s_{\text{terminal}}, \rho)$
  - $\mathcal{S}$: state space; $\mathcal{A}$: action space P: transition probability; R: reward function;
  - s_{terminal}: terminal state; ρ: initial state distribution
- Let τ be the hitting time when reaching terminal state 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

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

Outline

Value-based RL: Action-value function and greedy policy



- **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:

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

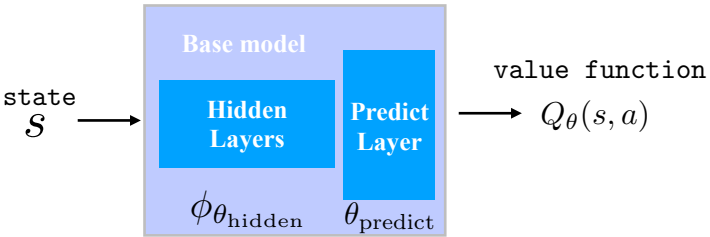
where θ is the parameter of the function.

Our solution: HyperFQI for randomized value function

HyperFQI includes Two models:

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

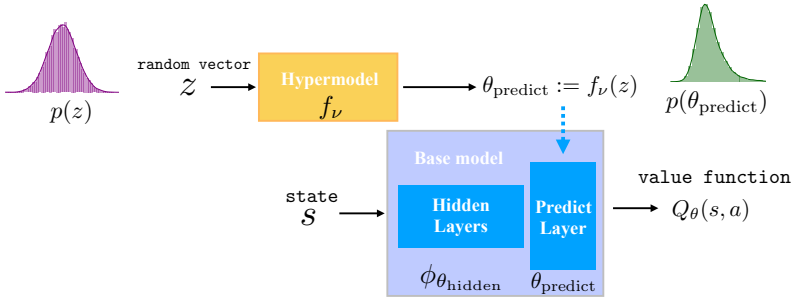
$$Q_\theta(s,a) = \langle \phi_{\theta_{\text{hidden}}}(s), \theta_{\text{predict}}(a) \rangle.$$



Our solution: HyperFQI for randomized value function

HyperFQI includes Two models:

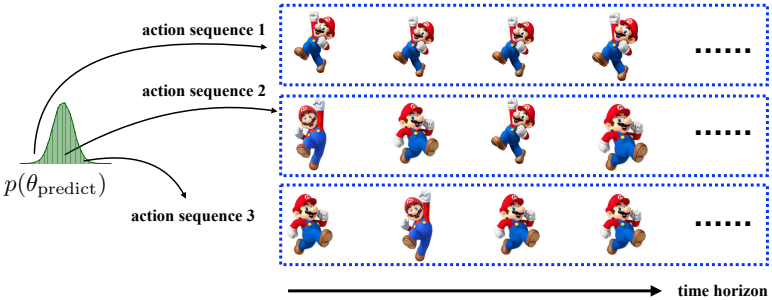
- Base model: DQN-type structure $Q_\theta(s,a) = \langle \phi_{\theta_{\text{hidden}}}(s), \theta_{\text{predict}}(a) \rangle$.
- Hypermodel: $\theta_{\text{predict}} = f_\nu(z)$ where $z \sim p(z)$. $p(z)$ is a fixed reference distribution.



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.

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.

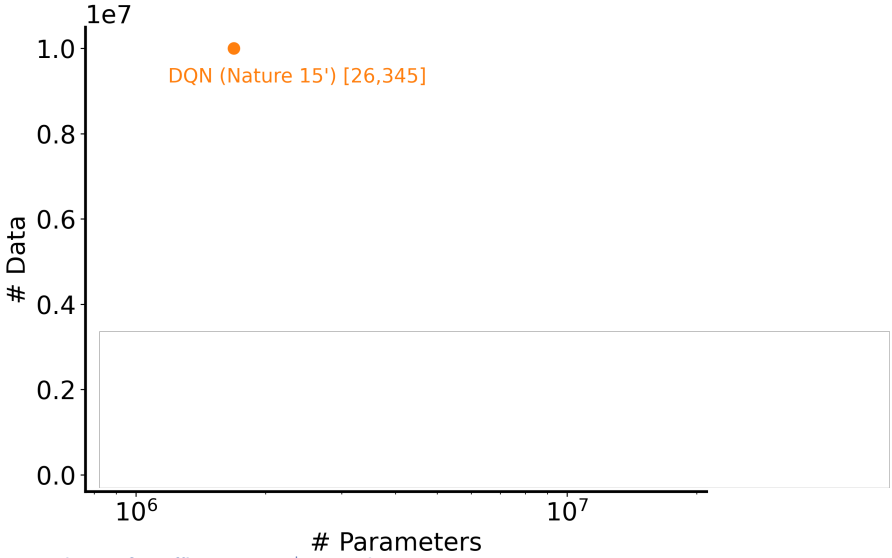
Outline

Benchmark problem in Reinforcement Learning

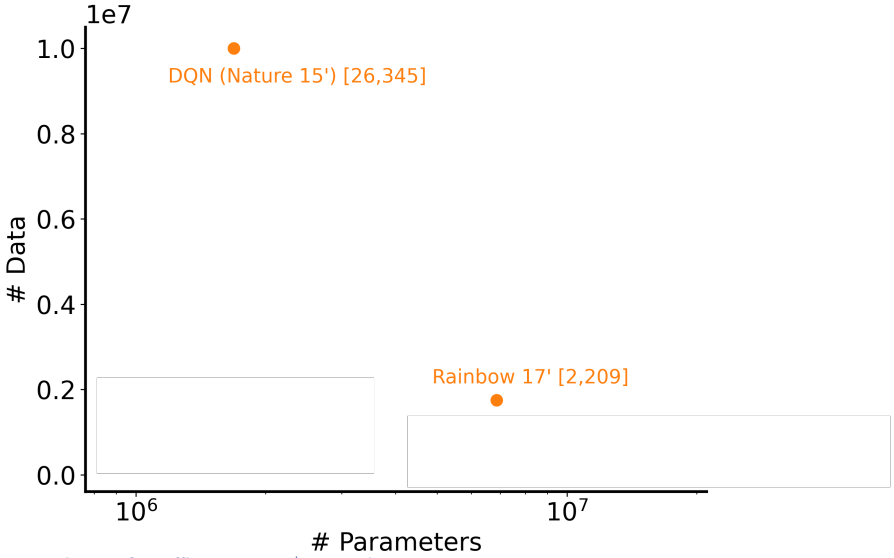
Figure: Human-level control via deep reinforcement learning (Nature 15'). Citations: 26,345.

- 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.

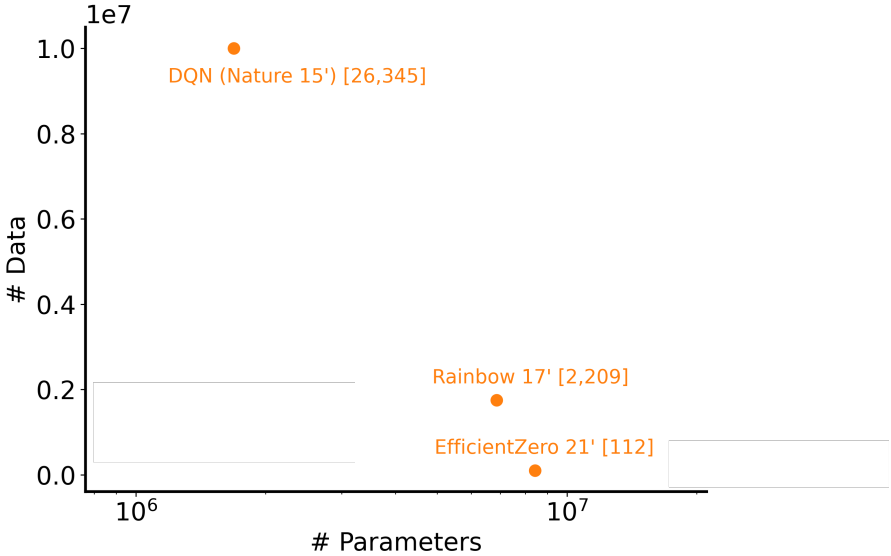
Data and computation efficiency in Deep RL benchmarks



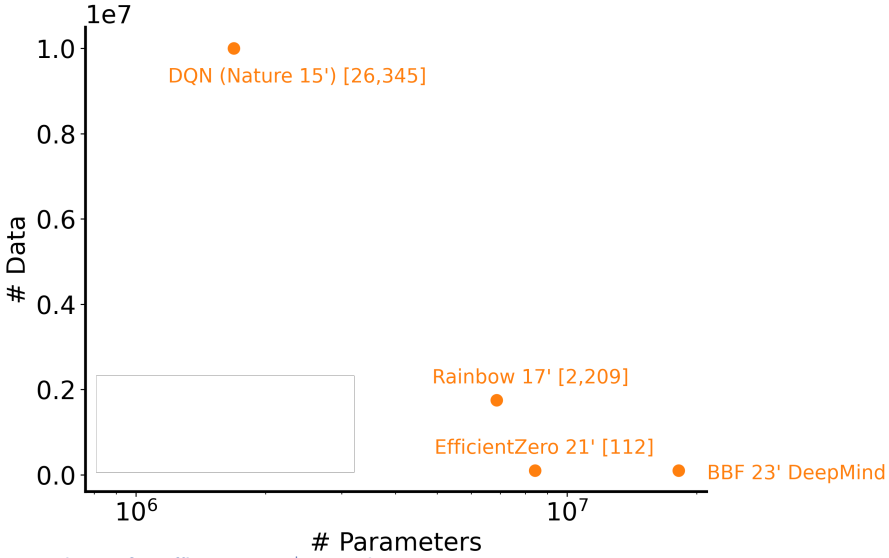
Data and computation efficiency in Deep RL benchmarks



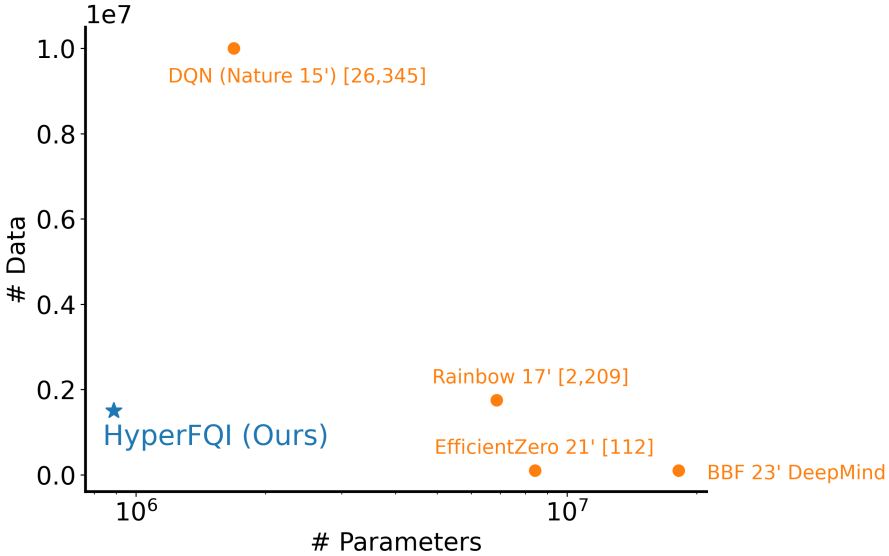
Data and computation efficiency in Deep RL benchmarks



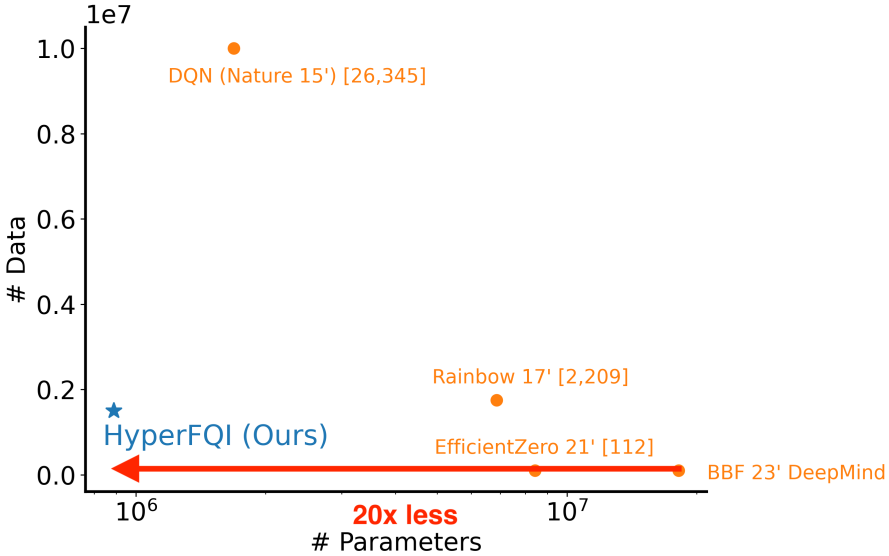
Data and computation efficiency in Deep RL benchmarks



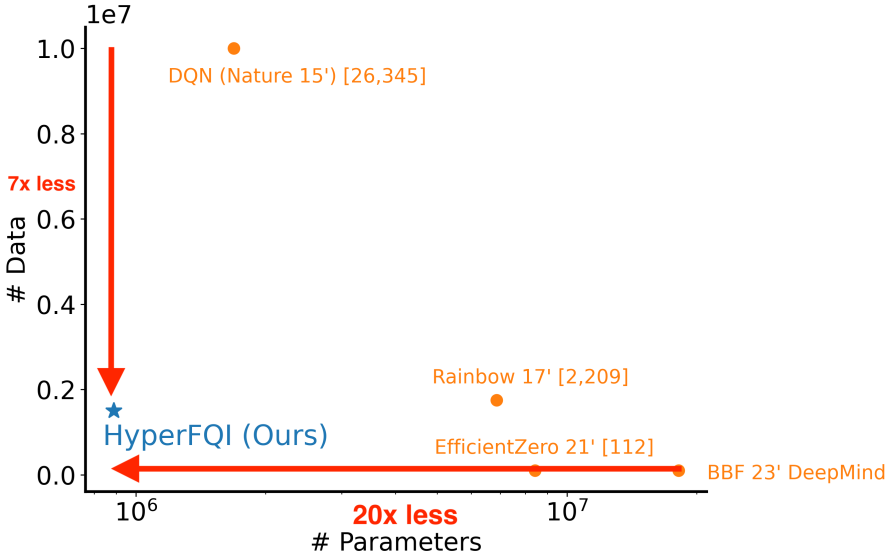
Data and computation efficiency in Deep RL benchmarks



Data and computation efficiency in Deep RL benchmarks



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 $M = \tilde{O}(\log(|\mathcal{S}||\mathcal{A}|K))$

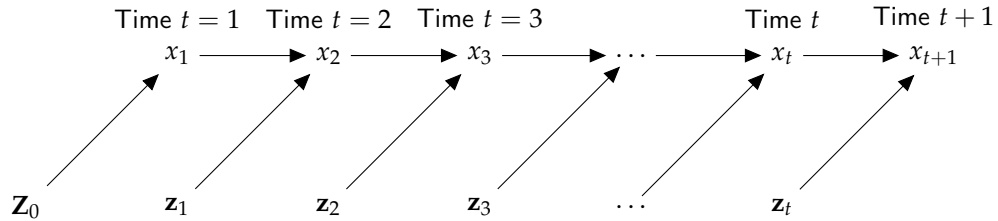
Theoretical Guarantees for HyperFQI in Tabular Setting

- 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 $M = \tilde{O}(\log(|\mathcal{S}||\mathcal{A}|K))$

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 $M = \tilde{O}(\log(|\mathcal{S}||\mathcal{A}|K))$

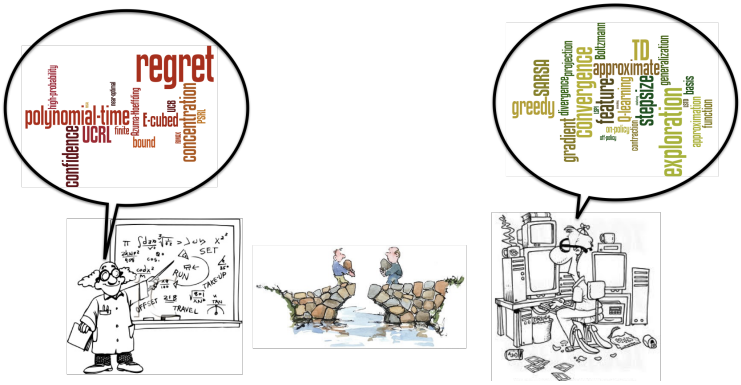
The novelty and difficulty in the mathemtical analysis



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.**

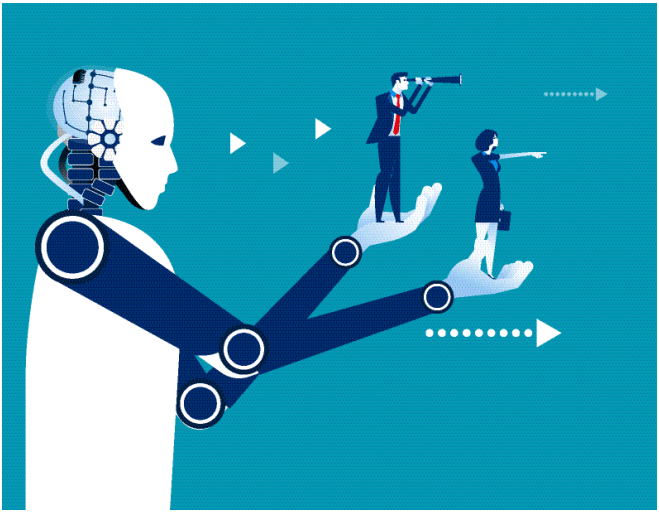
Bridging Theory and Practice



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)