RLChina 2021

Sample Efficiency in Online RL

Zhuoran Yang

Princeton $\longrightarrow$ Yale (2022)

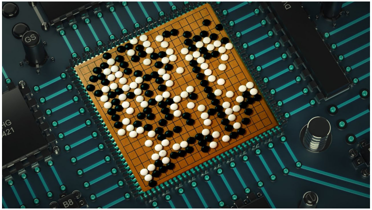
August 16, 2021

Motivation: sample complexity challenge in deep RL

Success and challenge of deep RL

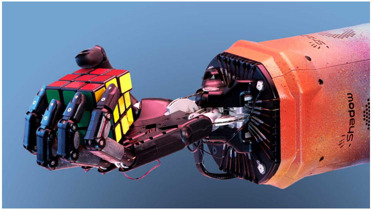
DRL = Representation (DL) + Decision Making (RL)

board games



computer games

robotic control



policy making

image source: Deepmind, OpenAI, Salesforce Research.

Success and challenge of deep RL

- AlphaGo: 3×10^7 games of self-play (data), 40 days of training (computation)
- AlphaStar: 2×10^2 years of self-play, 44 days of training
- Rubik's Cube: 10^4 years of simulation, 10 days of training

Success and challenge of deep RL

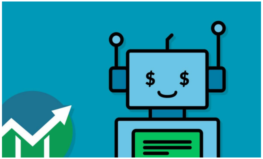
- AlphaGo: 3×10^7 games of self-play (data), 40 days of training (computation)
- AlphaStar: 2×10^2 years of self-play, 44 days of training
- Rubik's Cube: 10^4 years of simulation, 10 days of training

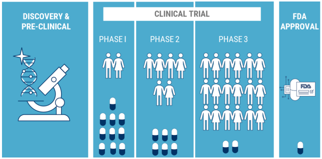
Massive Data Needed → Insufficient Data

↓

Massive Computation → Lack of Convergence

Insufficient Data, Lack of Convergence → **Unreliable AI**



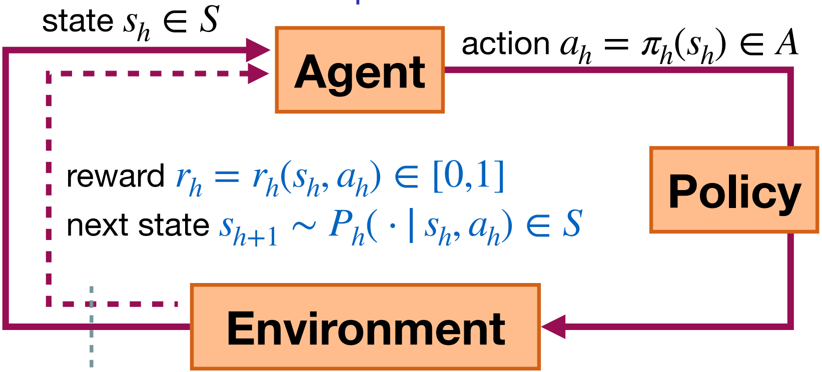


Our goal: provably efficient RL algorithms

- Sample efficiency: how many data points needed?
- Computational efficiency: how much computation needed?
- Function approximation: allow infinite number of observations?

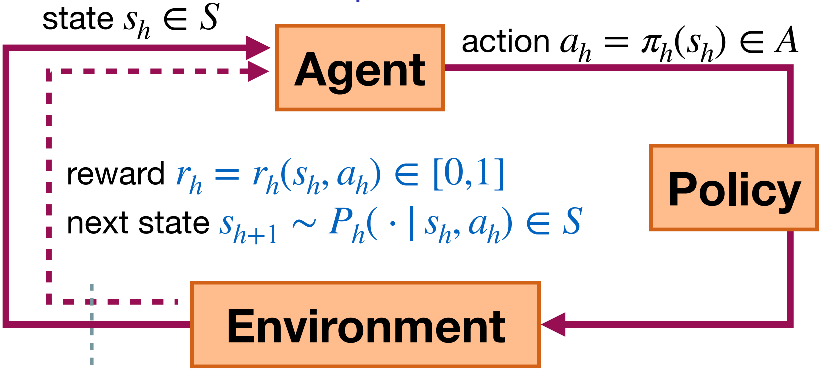
Background: Episodic Markov decision process

Episodic Markov decision process



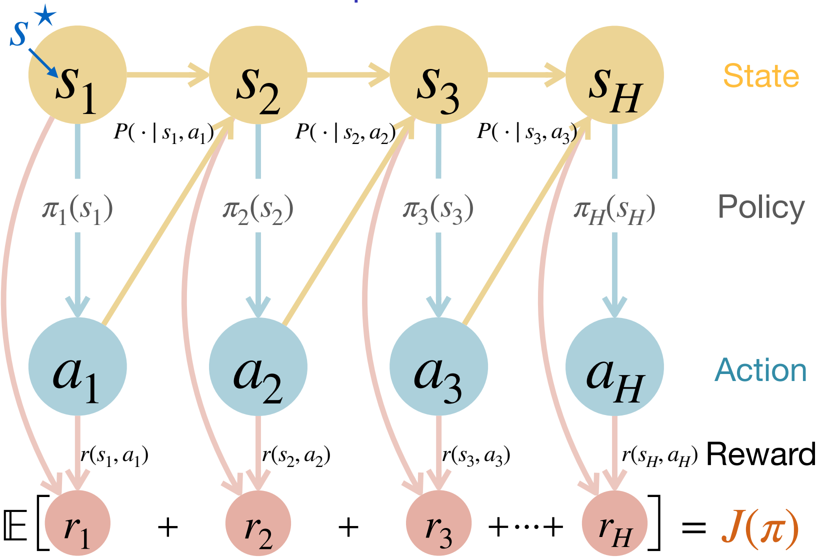
- For $h \in [H]$, in the h-th step, observe state s_h, takes action a_h
- Receive immediate reward r_h, $\mathbb{E}[r_h | s_h = s, a_h = a] = r_h(s, a)$
- Environment evolves to a new state $s_{h+1} \sim P_h(\cdot | s_h, a_h)$

Episodic Markov decision process



- For $h \in [H]$, in the h-th step, observe state s_h, takes action a_h
- Receive immediate reward r_h, $\mathbb{E}[r_h | s_h = s, a_h = a] = r_h(s, a)$
- Environment evolves to a new state $s_{h+1} \sim P_h(\cdot | s_h, a_h)$
- Policy $\pi_h \colon \mathcal{S} \to \mathcal{A}$: how agent takes action at h-th step
- Goal: find the policy $\pi^\star$ that maximizes $J(\pi) := \mathbb{E}_\pi[\sum_{h=1}^H r_h]$

Episodic Markov decision process



For simplicity, fix $s_1 = s^\star$, $r_h = r$, $P_h = P$ for all h.

Contextual bandit is a special case of MDP

- Contextual bandit (CB): observe context $s \in \mathcal{A}$, take action $a \in \mathcal{A}$, observe (random) reward r with mean $r^\star(s, a)$
- Optimal policy $\pi^\star(s) = \arg\max_{a \in \mathcal{A}} r^\star(s, a)$

Contextual bandit is a special case of MDP

- Contextual bandit (CB): observe context $s \in \mathcal{A}$, take action $a \in \mathcal{A}$, observe (random) reward r with mean $r^\star(s, a)$
- Optimal policy $\pi^\star(s) = \arg\max_{a \in \mathcal{A}} r^\star(s, a)$
- CB is a MDP with $H = 1$
- Multi-armed bandit (MAB): $H = 1$, $|A|$ finite, and $|\mathcal{S}| = 1$

MDP is significantly **more challenging** than CB due to **temporal dependency** (i.e., state transitions)

From MDP to RL

Informal definition of RL: Solve the MDP when the environment (r, P) is unknown

From MDP to RL

Informal definition of RL: Solve the MDP when the environment (r, P) is unknown

(i) generative model: able to query the reward r_h and next state s_{h+1} for any $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$

From MDP to RL

Informal definition of RL: Solve the MDP when the environment (r, P) is unknown

(i) generative model: able to query the reward r_h and next state s_{h+1} for any $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$

(ii) offline setting: given a dataset $\mathcal{D} = \{(s_h^t, a_h^t, r_h^t)\}_{h\in[H],t\in[T]}$ (T trajectories), learn $\pi^\star$ without any interactions

From MDP to RL

Informal definition of RL: Solve the MDP when the environment (r, P) is unknown

(i) generative model: able to query the reward r_h and next state s_{h+1} for any $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$

(ii) offline setting: given a dataset $\mathcal{D} = \{(s_h^t, a_h^t, r_h^t)\}_{h \in [H], t \in [T]}$ (T trajectories), learn $\pi^\star$ without any interactions

(iii) online setting: without any prior knowledge, learn $\pi^\star$ by interacting with the MDP for T episodes

From MDP to RL

Informal definition of RL: Solve the MDP when the environment (r, P) is unknown

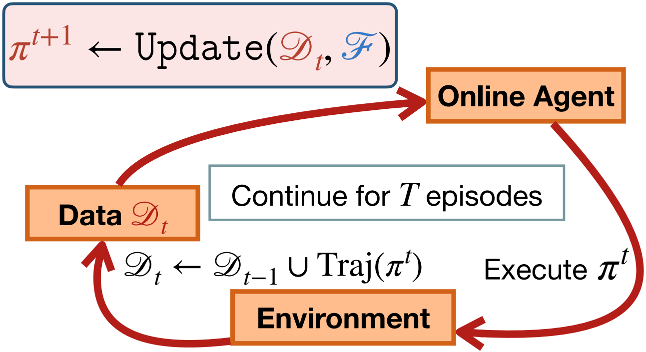
(i) generative model: able to query the reward r_h and next state s_{h+1} for any $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$

(ii) offline setting: given a dataset $\mathcal{D} = \{(s_h^t, a_h^t, r_h^t)\}_{h\in[H],t\in[T]}$ (T trajectories), learn $\pi^\star$ without any interactions

(iii) online setting: without any prior knowledge, learn $\pi^\star$ by interacting with the MDP for T episodes

(i)	learning from model	query complexity
(ii)	learning from data	sample complexity
(iii)	learning by doing	sample complexity + exploration

Online RL: Pipeline



- Initialize $\pi^1 = \{\pi_h^1\}_{h \in [H]}$ arbitrarily
- For each $t \in [T]$, execute π^t, get a trajectory $\mathrm{Traj}(\pi^t)$
- Store $\mathrm{Traj}(\pi^t)$ into dataset $\mathcal{D}_t = \{\mathrm{Traj}(\pi^i), i \le t\}$
- Update the policy via RL algorithm (need our design)

Sample efficiency in online RL: regret

- Measure sample efficiency via regret:

$$\texttt{Regret}(T) = \sum_{t=1}^{T} \underbrace{[J(\pi^\star) - J(\pi^t)]}_{\text{suboptimality in } t\text{-th episode}}$$

Sample efficiency in online RL: regret

- Measure sample efficiency via regret:

$$\texttt{Regret}(T) = \sum_{t=1}^{T} \underbrace{[J(\pi^\star) - J(\pi^t)]}_{\text{suboptimality in } t\text{-th episode}}$$

- Why regret is meaningful?
  - Define a random policy $\widetilde{\pi}^T$ uniformly sampled from $\{\pi^t, t \in [T]\}$. Suppose $\texttt{Regret}(T)$, then $J(\pi^\star) - J(\widetilde{\pi}^T) = \texttt{Regret}(T)/T$

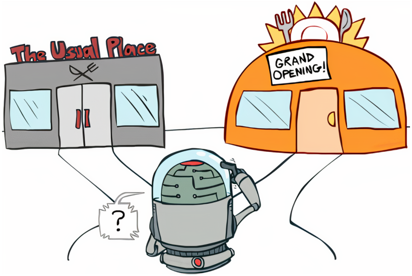
Sample efficiency in online RL: regret

- Measure sample efficiency via regret:

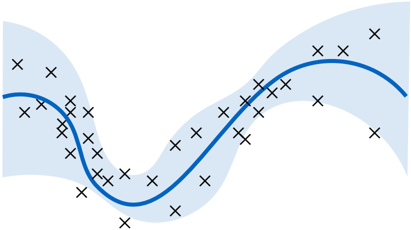
$$\texttt{Regret}(T) = \sum_{t=1}^{T} \underbrace{[J(\pi^\star) - J(\pi^t)]}_{\text{suboptimality in } t\text{-th episode}}$$

- Why regret is meaningful?
  - Define a random policy $\widetilde{\pi}^T$ uniformly sampled from $\{\pi^t, t \in [T]\}$. Suppose $\texttt{Regret}(T)$, then $J(\pi^\star) - J(\widetilde{\pi}^T) = \texttt{Regret}(T)/T$
  - If $\texttt{Regret}(T) = o(T)$, when T sufficiency large, $\widetilde{\pi}^T$ is arbitrarily close to optimality
- Goal: $\texttt{Regret}(T) = \widetilde{\mathcal{O}}\big(\sqrt{T} \cdot \mathrm{poly}(H) \cdot \mathrm{poly}(\mathrm{dim})\big)$

Challenge of Online RL: exploration & uncertainty

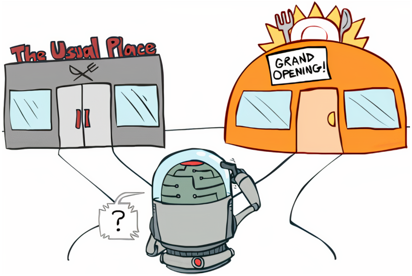


exploration vs. exploitation

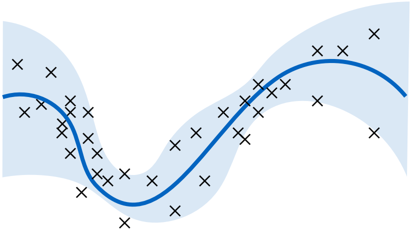


function estimation

Challenge of Online RL: exploration & uncertainty



exploration vs. exploitation



function estimation

How to assess estimation uncertainty based on adaptively acquired data?

Deep Exploration **Online RL** **Function Estimation**

How to construct exploration incentives tailored to function approximators?

Warmup example:
LinUCB for linear CB

Linear contextual bandit

- Setting of linear CB:
  - Observation structure: observe (perhaps adversarial) $s^t \in \mathcal{S}$, take action $a^t \in \mathcal{A}$, receive bounded reward $r^t \in [0, 1]$
  - $\mathbb{E}[r^t | s^t = s, a^t = a] = r^\star(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$

Linear contextual bandit

- Setting of linear CB:
  - Observation structure: observe (perhaps adversarial) $s^t \in \mathcal{S}$, take action $a^t \in \mathcal{A}$, receive bounded reward $r^t \in [0, 1]$
  - $\mathbb{E}[r^t | s^t = s, a^t = a] = r^\star(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$
  - Linear reward function: $r^\star(s, a) = \phi(s, a)^\top \theta^\star$
  - $\phi \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$: known feature mapping

Linear contextual bandit

- Setting of linear CB:
  - Observation structure: observe (perhaps adversarial) $s^t \in \mathcal{S}$, take action $a^t \in \mathcal{A}$, receive bounded reward $r^t \in [0, 1]$
  - $\mathbb{E}[r^t | s^t = s, a^t = a] = r^\star(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$
  - Linear reward function: $r^\star(s, a) = \phi(s, a)^\top \theta^\star$
  - $\phi: \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$: known feature mapping
- Normalization condition: $\|\theta^\star\|_2 \leq \sqrt{d}$, $\sup_{s,a} \|\phi(s, a)\|_2 \leq 1$
  - Why? Include MAB as a special case: $|\mathcal{S}| = 1$, $d = |A|$, $\phi(s, a) = \mathbf{e}_a$, $\theta^\star = (\mu_1, \mu_2, \cdots, \mu_A)$ with $\mu_a \in [0, 1]$ for all $a \in \mathcal{A}$

Linear contextual bandit

- Setting of linear CB:
  - Observation structure: observe (perhaps adversarial) $s^t \in \mathcal{S}$, take action $a^t \in \mathcal{A}$, receive bounded reward $r^t \in [0, 1]$
  - $\mathbb{E}[r^t | s^t = s, a^t = a] = r^\star(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$
  - Linear reward function: $r^\star(s, a) = \phi(s, a)^\top \theta^\star$
  - $\phi \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$: known feature mapping
- Normalization condition: $\|\theta^\star\|_2 \le \sqrt{d}$, $\sup_{s,a} \|\phi(s, a)\|_2 \le 1$
  - Why? Include MAB as a special case: $|\mathcal{S}| = 1$, $d = |A|$, $\phi(s, a) = \mathbf{e}_a$, $\theta^\star = (\mu_1, \mu_2, \cdots, \mu_A)$ with $\mu_a \in [0, 1]$ for all $a \in \mathcal{A}$
- $\texttt{Regret}(T) = \sum_{t=1}^{T} \big(\max_{a \in \mathcal{A}} \langle \phi(s^t, a) - \phi(s^t, a^t), \theta^\star \rangle \big)$

Exploration via optimism in the face of uncertainty (OFU)

- For each $t \in [T]$, in the t-th round, determine a^t in the following two steps

Exploration via optimism in the face of uncertainty (OFU)

- For each $t \in [T]$, in the t-th round, determine a^t in the following two steps
  - Uncertainty quantification: based on the current dataset $\mathcal{D}_{t-1}$, construct a high-probability confidence set Θ^t that contains $\theta^\star$

$$\mathbb{P}\big(\forall t \in [T], \theta^\star \in \Theta^t\big) \geq 1 - \delta \tag{1}$$

Exploration via optimism in the face of uncertainty (OFU)

- For each $t \in [T]$, in the t-th round, determine a^t in the following two steps
  - Uncertainty quantification: based on the current dataset $\mathcal{D}_{t-1}$, construct a high-probability confidence set Θ^t that contains $\theta^\star$

$$\mathbb{P}\big(\forall t \in [T], \theta^\star \in \Theta^t\big) \geq 1 - \delta \tag{1}$$

  - Optimistic planing: choose a^t to maximize the most optimistic model within Θ^t

$$a^t = \arg\max_{a \in \mathcal{A}} \Big\{ \max_{\theta \in \Theta^t} \langle \phi(s^t, a^t), \theta \rangle \Big\} \tag{2}$$

Exploration via optimism in the face of uncertainty (OFU)

- For each $t \in [T]$, in the t-th round, determine a^t in the following two steps
  - Uncertainty quantification: based on the current dataset $\mathcal{D}_{t-1}$, construct a high-probability confidence set Θ^t that contains $\theta^\star$

$$\mathbb{P}\big(\forall t \in [T], \theta^\star \in \Theta^t\big) \geq 1 - \delta \tag{1}$$

  - Optimistic planing: choose a^t to maximize the most optimistic model within Θ^t

$$a^t = \arg\max_{a \in \mathcal{A}} \Big\{ \max_{\theta \in \Theta^t} \langle \phi(s^t, a^t), \theta \rangle \Big\} \tag{2}$$

Intuition of OFU principle

- Define two functions

$$r^{+,t}(s,a) = \max_{\theta \in \Theta^t} \langle \phi(s,a), \theta \rangle$$
$$r^{-,t}(s,a) = \min_{\theta \in \Theta^t} \langle \phi(s,a), \theta \rangle$$

Intuition of OFU principle

- Define two functions

$$r^{+,t}(s,a) = \max_{\theta \in \Theta^t} \langle \phi(s,a), \theta \rangle$$
$$r^{-,t}(s,a) = \min_{\theta \in \Theta^t} \langle \phi(s,a), \theta \rangle$$

- By (1), $\theta^\star \in \Theta^t$

$$\boxed{r^{+,t}(s,a) \geq r^\star(s,a) \geq r^{-,t}(s,a),\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}}$$

Intuition of OFU principle

- Define two functions

$$r^{+,t}(s,a) = \max_{\theta \in \Theta^t} \langle \phi(s,a), \theta \rangle$$
$$r^{-,t}(s,a) = \min_{\theta \in \Theta^t} \langle \phi(s,a), \theta \rangle$$

- By (1), $\theta^\star \in \Theta^t$

$$\boxed{r^{+,t}(s,a) \geq r^\star(s,a) \geq r^{-,t}(s,a),\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}}$$

- $|r^{+,t}(s,a) - r^{-,t}(s,a)|$ reflects the uncertainty about a at context s

Intuition of OFU principle

- Define two functions

$$r^{+,t}(s,a) = \max_{\theta \in \Theta^t} \langle \phi(s,a), \theta \rangle$$
$$r^{-,t}(s,a) = \min_{\theta \in \Theta^t} \langle \phi(s,a), \theta \rangle$$

- By (1), $\theta^\star \in \Theta^t$

$$\boxed{r^{+,t}(s,a) \geq r^\star(s,a) \geq r^{-,t}(s,a), \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}}$$

- $|r^{+,t}(s,a) - r^{-,t}(s,a)|$ reflects the uncertainty about a at context s
- By (2), $a^t = \arg\max_a r^{+,t}(s^t,a)$ — choose the action with either large uncertainty (explore) or large reward (exploit)

How to construct Θ^t for linear CB?

- In the t-th round, our dataset is $\mathcal{D}_{t-1} = \{(s^i, a^i, r^i)\}_{i \leq t-1}$

How to construct Θ^t for linear CB?

- In the t-th round, our dataset is $\mathcal{D}_{t-1} = \{(s^i, a^i, r^i)\}_{i \le t-1}$
- First construct an estimator of $\theta^\star$ via ridge regression

$$\widehat{\theta}^t = \min_{\theta \in \mathbb{R}^d} \left\{ L^t(\theta) := \sum_{i=1}^{t-1} [r^i - \langle \phi(s^i, a^i), \theta \rangle]^2 + \|\theta\|_2^2 \right\}$$

How to construct Θ^t for linear CB?

- In the t-th round, our dataset is $\mathcal{D}_{t-1} = \{(s^i, a^i, r^i)\}_{i \le t-1}$
- First construct an estimator of $\theta^\star$ via ridge regression

$$\widehat{\theta}^t = \min_{\theta \in \mathbb{R}^d} \left\{ L^t(\theta) :\ = \sum_{i=1}^{t-1} [r^i - \langle \phi(s^i, a^i), \theta \rangle]^2 + \|\theta\|_2^2 \right\}$$

  - Hessian of the Ridge loss:

$$\Lambda^t = \nabla^2 L^t(\theta) = I + \sum_{i=1}^{t-1} \phi(s^i, a^i) \phi(s^i, a^i)^\top$$

How to construct Θ^t for linear CB?

- In the t-th round, our dataset is $\mathcal{D}_{t-1} = \{(s^i, a^i, r^i)\}_{i \le t-1}$
- First construct an estimator of $\theta^\star$ via ridge regression

$$\widehat{\theta}^t = \min_{\theta \in \mathbb{R}^d} \left\{ L^t(\theta) := \sum_{i=1}^{t-1} [r^i - \langle \phi(s^i, a^i), \theta \rangle]^2 + \|\theta\|_2^2 \right\}$$

  - Hessian of the Ridge loss:

$$\Lambda^t = \nabla^2 L^t(\theta) = I + \sum_{i=1}^{t-1} \phi(s^i, a^i) \phi(s^i, a^i)^\top$$

  - Closed-form solution: $\widehat{\theta}^t = (\Lambda^t)^{-1} \sum_{i=1}^{t-1} r_i \cdot \phi(s^i, a^i)$

How to construct Θ^t for linear CB?

- In the t-th round, our dataset is $\mathcal{D}_{t-1} = \{(s^i, a^i, r^i)\}_{i \le t-1}$
- First construct an estimator of $\theta^\star$ via ridge regression

$$\widehat{\theta}^t = \min_{\theta \in \mathbb{R}^d} \left\{ L^t(\theta) :\ = \sum_{i=1}^{t-1} [r^i - \langle \phi(s^i, a^i), \theta \rangle]^2 + \|\theta\|_2^2 \right\}$$

  - Hessian of the Ridge loss:

$$\Lambda^t = \nabla^2 L^t(\theta) = I + \sum_{i=1}^{t-1} \phi(s^i, a^i) \phi(s^i, a^i)^\top$$

  - Closed-form solution: $\widehat{\theta}^t = (\Lambda^t)^{-1} \sum_{i=1}^{t-1} r_i \cdot \phi(s^i, a^i)$

- Define Θ^t as the *confidence ellipsoid*:

$$\Theta^t = \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}^t\|_{\Lambda^t} \le \beta\},$$

here define $\beta = C \cdot \sqrt{d \cdot \log(1 + T/\delta)}$, $\|x\|_M = \sqrt{x^\top M x}$

LinUCB algorithm

- Confidence ellipsoid: $\Theta^t = \{\theta \in \mathbb{R}^d \colon \|\theta - \widehat{\theta}^t\|_{\Lambda^t} \leq \beta\}$

LinUCB algorithm

- Confidence ellipsoid: $\Theta^t = \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}^t\|_{\Lambda^t} \leq \beta\}$
- $r^{+,t}$ admits a closed-form:

$$r^{+,t}(s,a) = \max_{\theta \in \Theta^t} \langle \phi(s,a), \theta \rangle = \langle \phi(s,a), \widehat{\theta}^t \rangle + \underbrace{\beta \cdot \|\phi(s,a)\|_{(\Lambda^t)^{-1}}}_{\text{bonus}}$$

LinUCB algorithm

- Confidence ellipsoid: $\Theta^t = \{\theta \in \mathbb{R}^d \colon \|\theta - \widehat{\theta}^t\|_{\Lambda^t} \le \beta\}$
- $r^{+,t}$ admits a closed-form:

$$r^{+,t}(s,a) = \max_{\theta \in \Theta^t} \langle \phi(s,a), \theta \rangle = \langle \phi(s,a), \widehat{\theta}^t \rangle + \underbrace{\beta \cdot \|\phi(s,a)\|_{(\Lambda^t)^{-1}}}_{\text{bonus}}$$

- LinUCB algorithm (Dani et all. (2008), Abbasi-Yadkori et al. (2011), ...)
  - For $t = 1, \ldots, T$, in the beginning of t-th round
  - Observe context s_t

LinUCB algorithm

- Confidence ellipsoid: $\Theta^t = \{\theta \in \mathbb{R}^d \colon \|\theta - \widehat{\theta}^t\|_{\Lambda^t} \le \beta\}$
- $r^{+,t}$ admits a closed-form:

$$r^{+,t}(s,a) = \max_{\theta \in \Theta^t} \langle \phi(s,a), \theta \rangle = \langle \phi(s,a), \widehat{\theta}^t \rangle + \underbrace{\beta \cdot \|\phi(s,a)\|_{(\Lambda^t)^{-1}}}_{\text{bonus}}$$

- LinUCB algorithm (Dani et all. (2008), Abbasi-Yadkori et al. (2011), ...)
  - For $t = 1, \dots, T$, in the beginning of t-th round
  - Observe context s_t
  - Compute $\widehat{\theta}^t$ via ridge regression and compute Hessian Λ^t
  - Choose $a^t = \arg\max_{a \in \mathcal{A}} r^{+,t}(s^t, a)$ and observe r^t

Regret Bound of LinUCB

Theorem. With probability $1 - \delta$, LinUCB satisfies

$$\texttt{Regret}(T) \leq \beta \cdot \sqrt{dT} \cdot \sqrt{\log T} = \widetilde{\mathcal{O}}(d\sqrt{T})$$

Regret Bound of LinUCB

Theorem. With probability $1 - \delta$, LinUCB satisfies

$$\texttt{Regret}(T) \leq \beta \cdot \sqrt{dT} \cdot \sqrt{\log T} = \widetilde{\mathcal{O}}(d\sqrt{T})$$

- Independent of the size of context and action spaces, only depends on the dimension (linear in d)

Regret Bound of LinUCB

Theorem. With probability $1-\delta$, LinUCB satisfies

$$\texttt{Regret}(T) \le \beta \cdot \sqrt{dT} \cdot \sqrt{\log T} = \widetilde{\mathcal{O}}(d\sqrt{T})$$

- Independent of the size of context and action spaces, only depends on the dimension (linear in d)
- Depend on T through $\sqrt{T}$ (sample efficient)

Regret Bound of LinUCB

Theorem. With probability $1 - \delta$, LinUCB satisfies

$$\texttt{Regret}(T) \leq \beta \cdot \sqrt{dT} \cdot \sqrt{\log T} = \widetilde{\mathcal{O}}(d\sqrt{T})$$

- Independent of the size of context and action spaces, only depends on the dimension (linear in d)
- Depend on T through $\sqrt{T}$ (sample efficient)
- Computational cost $\mathrm{poly}(d, A, T)$ and memory requirement is $\mathrm{poly}(d, T)$ (Why?)

Regret Analysis of LinUCB (1/3)

Step 1. Conditioning on the event $\mathcal{E} = \{\forall t \in [T], \theta^\star \in \Theta^t\}$,
$r^{-,t}(s,a) \le r^\star(s,a) \le r^{+,t}(s,a), \forall (x,a) \in \mathcal{S} \times \mathcal{A}$

Regret Analysis of LinUCB (1/3)

Step 1. Conditioning on the event $\mathcal{E} = \{\forall t \in [T], \theta^\star \in \Theta^t\}$,
$r^{-,t}(s, a) \leq r^\star(s, a) \leq r^{+,t}(s, a), \forall (x, a) \in \mathcal{S} \times \mathcal{A}$

Step 2. Regret decomposition:

$$
\begin{aligned}
&\max_a r^\star(s^t, a) - r^\star(s^t, a^t) \\
&\qquad = \left(\max_a r^\star(s^t, a) - r^{+,t}(s^t, a^t)\right) + \left(r^{+,t}(s^t, a^t) - r^\star(s^t, a^t)\right) \\
&\qquad \leq r^{+,t}(s^t, a^t) - r^{-,t}(s^t, a^t) \leq 2\beta \cdot \|\phi(s^t, a^t)\|_{(\Lambda^t)^{-1}}
\end{aligned}
$$

Regret Analysis of LinUCB (1/3)

Step 1. Conditioning on the event $\mathcal{E} = \{\forall t \in [T], \theta^\star \in \Theta^t\}$, $r^{-,t}(s,a) \leq r^\star(s,a) \leq r^{+,t}(s,a), \forall (x,a) \in \mathcal{S} \times \mathcal{A}$

Step 2. Regret decomposition:

$$\begin{aligned}
&\max_a r^\star(s^t,a) - r^\star(s^t,a^t) \\
&\qquad = \big(\max_a r^\star(s^t,a) - r^{+,t}(s^t,a^t)\big) + \big(r^{+,t}(s^t,a^t) - r^\star(s^t,a^t)\big) \\
&\qquad \leq r^{+,t}(s^t,a^t) - r^{-,t}(s^t,a^t) \leq 2\beta \cdot \|\phi(s^t,a^t)\|_{(\Lambda^t)^{-1}}
\end{aligned}$$

Thus we have shown

$$\max_a r^\star(s^t,a) - r^\star(s^t,a^t) \leq \beta \cdot \min\{1, \|\phi(s^t,a^t)\|_{(\Lambda^t)^{-1}}\}$$

Regret Analysis of LinUCB (2/3)

Step 3. Telescope the bonus: Cauchy-Schwarz + elliptical potential lemma

$$
\begin{aligned}
\texttt{Regret}(T) &\le 2\beta \cdot \sum_{t=1}^{T} \cdot \min\{1, \|\phi(s^t, a^t)\|_{(\Lambda^t)^{-1}}\} \\
&\le 2\beta\sqrt{T} \cdot \left[\sum_{t=1}^{T} \min\{1, \|\phi(s^t, a^t)\|^2_{(\Lambda^t)^{-1}}\} \right]^{1/2} \\
&\le 2\beta\sqrt{T} \cdot \sqrt{2 \mathrm{logdet}(\Lambda_T)} \lesssim \beta \cdot \sqrt{T} \cdot \sqrt{d \cdot \log T}
\end{aligned}
$$

Elliptical potential lemma

Lemma

$$\sum_{i=1}^{t-1} \min\{1, \|\phi(s^i, a^i)\|^2_{(\Lambda^t)^{-1}}\} \leq 2\mathrm{logdet}(\Lambda_T) \leq d \cdot \log T$$

Elliptical potential lemma

Lemma

$$\sum_{i=1}^{t-1} \min\{1, \|\phi(s^i, a^i)\|^2_{(\Lambda^t)^{-1}}\} \leq 2\mathrm{logdet}(\Lambda_T) \leq d \cdot \log T$$

Bayesian perspective:

- Prior distribution $\theta^\star \sim N(0, I)$
- Likelihood $r_i = r^\star(s^i, a^i) + \varepsilon$, $\varepsilon \in N(0, 1)$
- Given dataset $\mathcal{D}_{t-1}$, the posterior distribution is $N(\widehat{\theta}^t, \Lambda^t)$

Elliptical potential lemma

Lemma

$$\sum_{i=1}^{t-1} \min\{1, \|\phi(s^i, a^i)\|^2_{(\Lambda^t)^{-1}}\} \le 2\mathrm{logdet}(\Lambda_T) \le d \cdot \log T$$

Bayesian perspective:

- Prior distribution $\theta^\star \sim N(0, I)$
- Likelihood $r_i = r^\star(s^i, a^i) + \varepsilon$, $\varepsilon \in N(0, 1)$
- Given dataset $\mathcal{D}_{t-1}$, the posterior distribution is $N(\widehat{\theta}^t, \Lambda^t)$
- $\|\phi(s, a)\|^2_{(\Lambda^t)^{-1}} = I(\theta^\star, (s, a)|\mathcal{D}_{t-1}) = H(\theta^\star|\mathcal{D}_{t-1}) - H(\theta^\star|\mathcal{D}_{t-1} \cup \{(s, a)\})$ — conditional mutual information gain
- Elliptical potential lemma $\leftrightarrow$ chain rule of mutual information

Regret Analysis of LinUCB (3/3)

Step 4. Show optimism: $\theta^\star \in \Theta^t = \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}^t\|_{\Lambda^t} \leq \beta\}$

$$\theta^t - \theta^\star = (\Lambda^t)^{-1}\left[\sum_{i=1}^{t-1} \underbrace{(r^\star(s^i, a^i) + \varepsilon_i)}_{r_i} \cdot \phi(s^i, a^i) - (\Lambda^t) \cdot \theta^\star\right]$$

$$= (\Lambda^t)^{-1}\theta^\star + (\Lambda^t)^{-1} \sum_{i=1}^{t-1} \underbrace{\varepsilon_i}_{\text{martingale difference}} \cdot \phi(s^i, a^i)$$

Regret Analysis of LinUCB (3/3)

Step 4. Show optimism: $\theta^\star \in \Theta^t = \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}^t\|_{\Lambda^t} \leq \beta\}$

$$\theta^t - \theta^\star = (\Lambda^t)^{-1}\left[\sum_{i=1}^{t-1} \underbrace{(r^\star(s^i, a^i) + \varepsilon_i)}_{r_i} \cdot \phi(s^i, a^i) - (\Lambda^t) \cdot \theta^\star\right]$$

$$= (\Lambda^t)^{-1}\theta^\star + (\Lambda^t)^{-1} \sum_{i=1}^{t-1} \underbrace{\varepsilon_i}_{\text{martingale difference}} \cdot \phi(s^i, a^i)$$

- $\|\theta^t - \theta^\star\|_{\Lambda^t} \approx \| \sum_{i=1}^{t-1} \varepsilon_i \cdot \phi(s^i, a^i)\|_{(\Lambda^t)^{-1}} \leq \beta$
- Concentration of self-normalized process (e.g., Abbasi-Yadkori et al, 2011)

Summary of LinUCB

- Algorithm is based on the optimism principle
- upper confidence bound $=$ ridge estimator $+$ bonus

Summary of LinUCB

- Algorithm is based on the optimism principle
- upper confidence bound = ridge estimator + bonus
- $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret via
  a. utilize optimism to bound instant regret by width of confidence region (bonus)
  b. sum up the instant regret terms by elliptical potential lemma
  c. confidence region constructed via uncertainty quantification for ridge regression

Summary of LinUCB

- Algorithm is based on the optimism principle
- upper confidence bound = ridge estimator + bonus
- $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret via
  a. utilize optimism to bound instant regret by width of confidence region (bonus)
  b. sum up the instant regret terms by elliptical potential lemma
  c. confidence region constructed via uncertainty quantification for ridge regression
- Extensions: generalized linear model, RKHS, overparameterized network, abstract function class with bounded eluder dimension, ...

From Linear CB to Linear RL

Value functions and Bellman equation

- Our goal is to find $\pi^\star = \arg\max_\pi J(\pi) = \mathbb{E}[\sum_{h=1}^{H} r_h]$

Value functions and Bellman equation

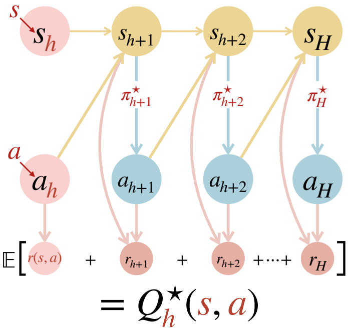
- Our goal is to find $\pi^\star = \arg\max_\pi J(\pi) = \mathbb{E}[\sum_{h=1}^{H} r_h]$
- $\pi^\star$ is characterized by $Q^\star = \{Q_h^\star\}_{h\in[H+1]} : \mathcal{S} \times \mathcal{A} \to [0, H]$

Value functions and Bellman equation

- Our goal is to find $\pi^\star = \arg\max_\pi J(\pi) = \mathbb{E}[\sum_{h=1}^{H} r_h]$
- $\pi^\star$ is characterized by $Q^\star = \{Q_h^\star\}_{h\in[H+1]} : \mathcal{S} \times \mathcal{A} \to [0, H]$
  - $\pi_h^\star(s) = \arg\max_a Q_h^\star(s, a), \forall s \in \mathcal{S}$; $V_h^\star(s) = \max_a Q_h^\star(s, a)$

Value functions and Bellman equation

- Our goal is to find $\pi^\star = \arg\max_\pi J(\pi) = \mathbb{E}[\sum_{h=1}^H r_h]$
- $\pi^\star$ is characterized by $Q^\star = \{Q_h^\star\}_{h\in[H+1]} : \mathcal{S}\times\mathcal{A}\to[0,H]$
  - $\pi_h^\star(s) = \arg\max_a Q_h^\star(s,a),\ \forall s\in\mathcal{S};\ V_h^\star(s) = \max_a Q_h^\star(s,a)$
  - $Q_h^\star(s,a)$: optimal return starting from $s_h = s$ and $a_h = a$
  - $Q^\star$ is characterized by Bellman equation



- $\pi_h^\star(s) = \arg\max_a Q_h^\star(s,\cdot)$
- $V_h^\star(s) = \max_a Q_h^\star(s,a)$
- $Q_h^\star(s,a)$ given by Bellman equation

Value functions and Bellman equation

- MDP terminates after H steps — $Q^\star_{H+1} = \mathbf{0}$, $Q^\star_H = r$

Value functions and Bellman equation

- MDP terminates after H steps — $Q^\star_{H+1} = \mathbf{0}$, $Q^\star_H = r$
- Bellman equation

$$
\begin{aligned}
Q^\star_h(s,a) &= \overbrace{r(s,a) + \mathbb{E}_{s_{h+1}\sim P}\left[V^\star_{h+1}(s_{h+1}) \middle| s_h = s, a_h = a\right]}^{\text{Bellman operator: } \mathcal{B}Q^\star_{h+1}} \\
V^\star_h(s) &= \max_a Q^\star_h(s,a)
\end{aligned}
$$

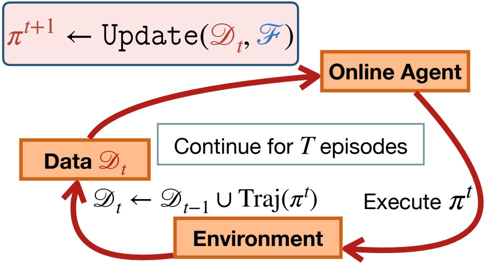
Value functions and Bellman equation

- MDP terminates after H steps — $Q^\star_{H+1} = \mathbf{0}$, $Q^\star_H = r$
- Bellman equation

$$
\begin{aligned}
Q^\star_h(s,a) &= \overbrace{r(s,a) + \mathbb{E}_{s_{h+1}\sim P}\left[V^\star_{h+1}(s_{h+1}) \middle| s_h = s, a_h = a\right]}^{\text{Bellman operator: } \mathcal{B}Q^\star_{h+1}} \\
V^\star_h(s) &= \max_a Q^\star_h(s,a)
\end{aligned}
$$

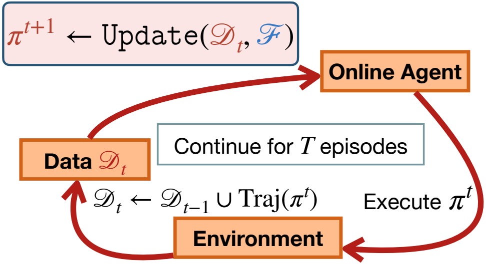
- RL with linear function approximation: use $\mathcal{F}_{\texttt{lin}} = \{\phi(s,a)^\top\theta \colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}\}$ to approximate $\{Q^\star_h\}_{h\in[H]}$
  - $\phi\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$: known feature mapping
  - Can allow known and step-varying features: $\{\phi_h\}_{h\in[H]}$

RL is more challenging than bandit



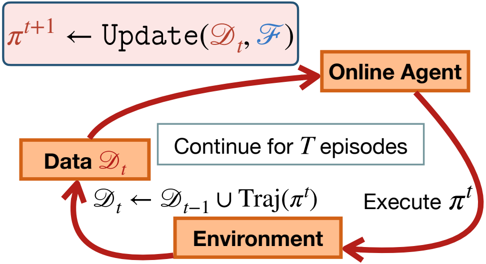
- $\texttt{Regret}(T) = \sum_{t=1}^{T}[J(\pi^\star) - J(\pi^t)]$

RL is more challenging than bandit



- $\texttt{Regret}(T) = \sum_{t=1}^{T}[J(\pi^\star) - J(\pi^t)]$
- $J(\pi^\star) = \max_a Q_1^\star(s^\star, a)$, fixed initial state for simplicity
- Can generalize to adversarial s_1

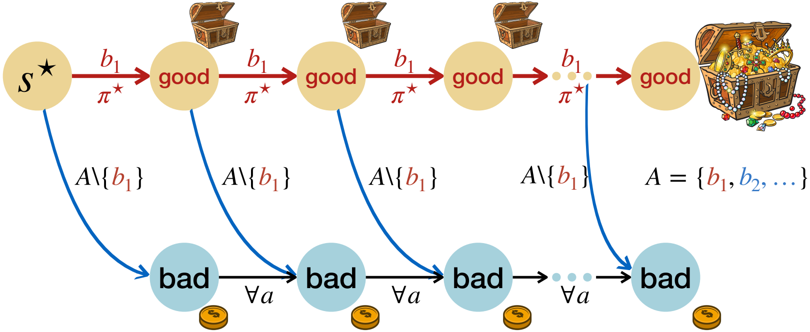
RL is more challenging than bandit



- $\texttt{Regret}(T) = \sum_{t=1}^{T}[J(\pi^\star) - J(\pi^t)]$
- $J(\pi^\star) = \max_a Q_1^\star(s^\star, a)$, fixed initial state for simplicity
- Can generalize to adversarial s_1
- Online RL $\neq$ a contextual bandit with TH rounds

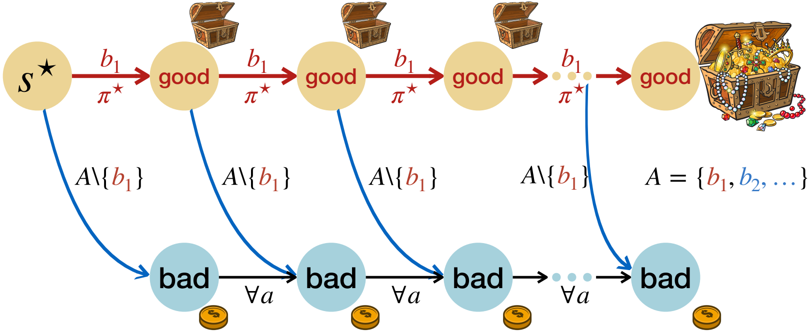
> RL requires **deep exploration** for achieving $o(T)$ regret.

Deep Exploration



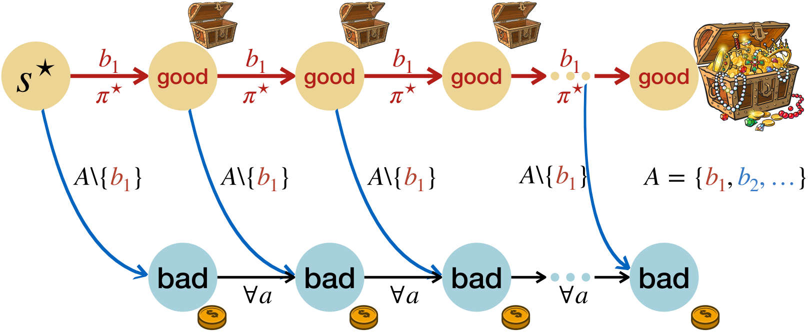
- Contextual bandit algorithm directly goes to the bad state at the initial state — $\Omega(T)$ regret

Deep Exploration



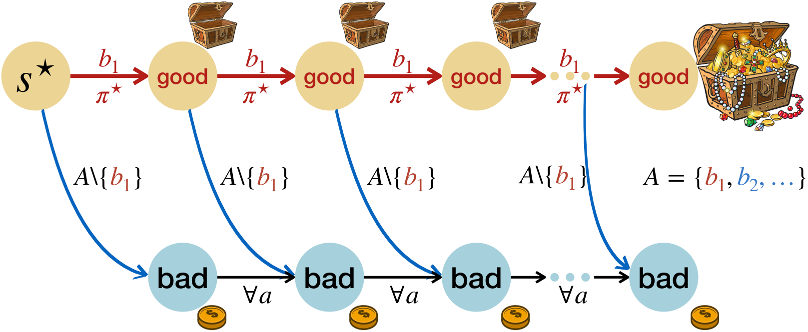
- Contextual bandit algorithm directly goes to the bad state at the initial state — $\Omega(T)$ regret
- Random sampling requires $|\mathcal{A}|^H$ samples (inefficient)

Deep Exploration



- Contextual bandit algorithm directly goes to the bad state at the initial state — $\Omega(T)$ regret
- Random sampling requires $|\mathcal{A}|^H$ samples (inefficient)
- Goal: $\texttt{Regret}(T) = \mathrm{poly}(H) \cdot \sqrt{T}$ (need deep exploration)

Deep Exploration



- Contextual bandit algorithm directly goes to the bad state at the initial state — $\Omega(T)$ regret
- Random sampling requires $|\mathcal{A}|^H$ samples (inefficient)
- Goal: $\texttt{Regret}(T) = \mathrm{poly}(H) \cdot \sqrt{T}$ (need deep exploration)
- $\mathcal{O}(SAH)$ query complexity if have a generative model

What assumption should we impose?

- In CB, $\pi^\star$ is greedy with respect to $r^\star$, we assume $r^\star$ is linear
- In RL, $\pi^\star$ is greedy with respect to $Q^\star$
- *Linear realizability*: $Q_h^\star \in \mathcal{F}_{\mathrm{lin}}, \forall h \in [H]$

What assumption should we impose?

- In CB, $\pi^\star$ is greedy with respect to $r^\star$, we assume $r^\star$ is linear
- In RL, $\pi^\star$ is greedy with respect to $Q^\star$
- *Linear realizability*: $Q_h^\star \in \mathcal{F}_{\text{lin}}, \forall h \in [H]$

> **Theorem** [Wang-Wang-Kakade-21] There exists a class of linearly realizable MDPs such that any online RL algorithm requires $\min\{\Omega(2^d), \Omega(2^H)\}$ samples to obtain a near optimal policy.

What assumption should we impose?

- In CB, $\pi^\star$ is greedy with respect to $r^\star$, we assume $r^\star$ is linear
- In RL, $\pi^\star$ is greedy with respect to $Q^\star$
- *Linear realizability*: $Q_h^\star \in \mathcal{F}_{\text{lin}}, \forall h \in [H]$

  Theorem [Wang-Wang-Kakade-21] There exists a class of linearly realizable MDPs such that any online RL algorithm requires $\min\{\Omega(2^d), \Omega(2^H)\}$ samples to obtain a near optimal policy.

- Fundamentally different from supervised learning and bandits
- To achieve efficiency in online RL, we need stronger assumptions

Assumption: $\mathrm{Image}(\mathcal{B}) \subseteq \mathcal{F}_{\mathrm{lin}}$

- We assume the image set of the Bellman operator lies in $\mathcal{F}_{\mathrm{lin}}$:

For any $Q\colon \mathcal{S} \times \mathcal{A} \to [0, H]$, there exists $\theta^Q \in \mathbb{R}^d$ s.t.

$$(\mathcal{B}Q)(s,a) = r(s,a) + \mathbb{E}[\max_{a'} Q(s',a')] = \langle \phi(s,a), \theta^Q \rangle$$

Assumption: $\mathrm{Image}(\mathcal{B}) \subseteq \mathcal{F}_{\mathrm{lin}}$

- We assume the image set of the Bellman operator lies in $\mathcal{F}_{\mathrm{lin}}$:

> For any $Q\colon \mathcal{S} \times \mathcal{A} \to [0, H]$, there exists $\theta^Q \in \mathbb{R}^d$ s.t.
>
> $$(\mathcal{B}Q)(s,a) = r(s,a) + \mathbb{E}[\max_{a'} Q(s',a')] = \langle \phi(s,a), \theta^Q \rangle$$

- Normalization condition: $\|\theta^Q\|_2 \leq 2H\sqrt{d}$, $\sup_{s,a} \|\phi(s,a)\|_2 \leq 1$

Assumption: $\mathrm{Image}(\mathcal{B}) \subseteq \mathcal{F}_{\mathrm{lin}}$

- We assume the image set of the Bellman operator lies in $\mathcal{F}_{\mathrm{lin}}$:

> For any $Q\colon \mathcal{S} \times \mathcal{A} \to [0, H]$, there exists $\theta^Q \in \mathbb{R}^d$ s.t.
>
> $$(\mathcal{B}Q)(s, a) = r(s, a) + \mathbb{E}[\max_{a'} Q(s', a')] = \langle \phi(s, a), \theta^Q \rangle$$

- Normalization condition: $\|\theta^Q\|_2 \leq 2H\sqrt{d}$, $\sup_{s,a} \|\phi(s, a)\|_2 \leq 1$
- Is there such an MDP? Yes, Linear MDP

> $r(s, a) \ = \ \langle \phi(s, a), \boldsymbol{\omega} \rangle \qquad P(s'|s, a) \ = \ \langle \phi(s, a), \boldsymbol{\mu}(s') \rangle$
> Normalization: $\|\boldsymbol{\omega}\|_2 \leq \sqrt{d}$, $\sum_{s'} \|\boldsymbol{\mu}(s')\|_2 \leq \sqrt{d}$

- Linear MDP contains tabular MDP as a special case: $\phi(s, a) = \mathbf{e}_{(s,a)}$, $d = |\mathcal{S}| \cdot |\mathcal{A}|$

Algorithm for linear RL: LSVI-UCB

Algorithm: LSVI + optimism

- In the beginning t-th episode, dataset $\mathcal{D}_{t-1} = \{(s_h^i, a_h^i, r_h^i), h \in [H]\}_{i<t}$

Algorithm: LSVI + optimism

- In the beginning t-th episode, dataset $\mathcal{D}_{t-1} = \{(s_h^i, a_h^i, r_h^i), h \in [H]\}_{i<t}$
- For $h = H, \ldots, 1$, backwardly solve ridge regression:

$$y_h^i = r_h^i + \max_a Q_{h+1}^{+,t}(s_{h+1}, a)$$

$$\widehat{\theta}_h^t = \arg\min_\theta \sum_{i=1}^{t-1} [y_h^i - \langle \phi(s_h^i, a_h^i), \theta \rangle]^2 + \|\theta\|_2^2$$

Algorithm: LSVI + optimism

- In the beginning t-th episode, dataset $\mathcal{D}_{t-1} = \{(s_h^i, a_h^i, r_h^i), h \in [H]\}_{i<t}$
- For $h = H, \ldots, 1$, backwardly solve ridge regression:

$$y_h^i = r_h^i + \max_a Q_{h+1}^{+,t}(s_{h+1}, a)$$

$$\widehat{\theta}_h^t = \arg\min_\theta \sum_{i=1}^{t-1} [y_h^i - \langle \phi(s_h^i, a_h^i), \theta \rangle]^2 + \|\theta\|_2^2$$

- Hessian of ridge loss: $\Lambda_h^t = I + \sum_{i=1}^{t-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top$

Algorithm: LSVI + optimism

- In the beginning t-th episode, dataset $\mathcal{D}_{t-1} = \{(s_h^i, a_h^i, r_h^i), h \in [H]\}_{i<t}$
- For $h = H, \ldots, 1$, backwardly solve ridge regression:

$$y_h^i = r_h^i + \max_a Q_{h+1}^{+,t}(s_{h+1}, a)$$

$$\widehat{\theta}_h^t = \arg\min_\theta \sum_{i=1}^{t-1} [y_h^i - \langle \phi(s_h^i, a_h^i), \theta \rangle]^2 + \|\theta\|_2^2$$

- Hessian of ridge loss: $\Lambda_h^t = I + \sum_{i=1}^{t-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top$
- Construct bonus $\Gamma_h^t(s, a) = \beta \cdot \|\phi(s, a)\|_{(\Lambda_h^t)^{-1}}$

Algorithm: LSVI + optimism

- In the beginning t-th episode, dataset $\mathcal{D}_{t-1} = \{(s_h^i, a_h^i, r_h^i), h \in [H]\}_{i<t}$
- For $h = H, \ldots, 1$, backwardly solve ridge regression:

$$y_h^i = r_h^i + \max_a Q_{h+1}^{+,t}(s_{h+1}, a)$$

$$\widehat{\theta}_h^t = \arg\min_\theta \sum_{i=1}^{t-1} [y_h^i - \langle \phi(s_h^i, a_h^i), \theta \rangle]^2 + \|\theta\|_2^2$$

- Hessian of ridge loss: $\Lambda_h^t = I + \sum_{i=1}^{t-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top$
- Construct bonus $\Gamma_h^t(s, a) = \beta \cdot \|\phi(s, a)\|_{(\Lambda_h^t)^{-1}}$
- UCB Q-function

$$Q_h^{+,t} = \mathrm{Trunc}_{[0,H]}\{\langle \phi(s, a), \widehat{\theta}_h^t \rangle + \Gamma_h^t(s, a)\}$$

- Execute $\pi_h^t(\cdot) = \arg\max_a Q_h^{+,t}(\cdot, a)$

A more abstract version of LSVI-UCB

- For $h = H, \ldots, 1$, backwardly solve least-squares regression:

$$y_h^i = r_h^i + \max_a Q_{h+1}^{+,t}(s_{h+1}, a)$$

$$\widehat{Q}_h^t = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} [y_h^i - f(s_h^i, a_h^i)]^2 + \mathrm{penalty}(f)$$

A more abstract version of LSVI-UCB

- For $h = H, \ldots, 1$, backwardly solve least-squares regression:

$$y_h^i = r_h^i + \max_a Q_{h+1}^{+,t}(s_{h+1}, a)$$

$$\widehat{Q}_h^t = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} [y_h^i - f(s_h^i, a_h^i)]^2 + \mathrm{penalty}(f)$$

- Uncertainty quantification: for $\widehat{Q}_h^t$: find Γ_h^t such that

$$\mathbb{P}\Big(\forall (t, h, s, a), |\widehat{Q}_h^t(s, a) - (\mathcal{B} Q_{h+1}^{+,t})(s, a)| \leq \Gamma_h^t(s, a)\Big) \geq 1 - \delta$$

A more abstract version of LSVI-UCB

- For $h = H, \ldots, 1$, backwardly solve least-squares regression:

$$y_h^i = r_h^i + \max_a Q_{h+1}^{+,t}(s_{h+1}, a)$$

$$\widehat{Q}_h^t = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} [y_h^i - f(s_h^i, a_h^i)]^2 + \mathrm{penalty}(f)$$

- Uncertainty quantification: for $\widehat{Q}_h^t$: find Γ_h^t such that

$$\mathbb{P}\Big(\forall (t, h, s, a), |\widehat{Q}_h^t(s, a) - (\mathcal{B}Q_{h+1}^{+,t})(s, a)| \le \Gamma_h^t(s, a)\Big) \ge 1 - \delta$$

- Optimism:

$$Q_h^{+,t} = \mathrm{Trunc}_{[0,H]}\{\widehat{Q}_h^t + \Gamma_h^t(s, a)\}$$

A more abstract version of LSVI-UCB

- For $h = H, \ldots, 1$, backwardly solve least-squares regression:

$$y_h^i = r_h^i + \max_a Q_{h+1}^{+,t}(s_{h+1}, a)$$

$$\widehat{Q}_h^t = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} [y_h^i - f(s_h^i, a_h^i)]^2 + \mathrm{penalty}(f)$$

- Uncertainty quantification: for $\widehat{Q}_h^t$: find Γ_h^t such that

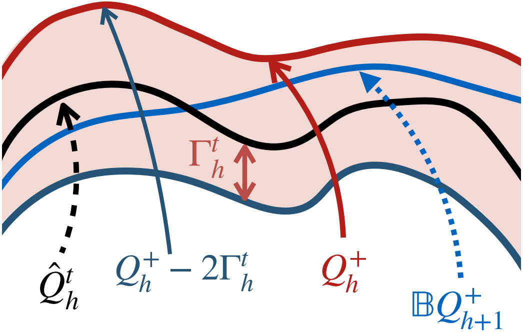
$$\mathbb{P}\Big(\forall (t, h, s, a), |\widehat{Q}_h^t(s,a) - (\mathcal{B}Q_{h+1}^{+,t})(s,a)| \leq \Gamma_h^t(s,a)\Big) \geq 1 - \delta$$

- Optimism:

$$Q_h^{+,t} = \mathrm{Trunc}_{[0,H]}\{\widehat{Q}_h^t + \Gamma_h^t(s,a)\}$$

- Computation & memory cost independent of $|\mathcal{S}|$ (even for RKHS, overparameterized NN)

A more abstract version of LSVI-UCB



$$|\hat{Q}_h^t(s,a) - (\mathbb{B}Q_{h+1}^+)(s,a)| \le \Gamma_h^t(s,a), \ \forall t, h, s, a$$

$$Q_h^+(s,a) = \underbrace{\hat{Q}_h^t(s,a)}_{\text{LSVI}} + \underbrace{\Gamma_h^t(s,a)}_{\text{bonus}}$$

- Optimism gives: $\mathcal{B}Q_{h+1}^+ \in [Q_h^+ - 2\Gamma_h^t, Q_h^+]$
- Monotonicity of Bellman operator: $Q_h^+ \ge Q_h^\star$ (UCB)

Sample efficiency of LSVI-UCB

LSVI-UCB achieves $\sqrt{T}$-regret

Theorem [Jin-Yang-Wang-Jordan-20] Choosing $\beta = \widetilde{\mathcal{O}}(dH)$, with probability at least $1 - 1/T$,

$$\mathtt{Regret}(T) = \widetilde{\mathcal{O}}(\beta \cdot H \cdot \sqrt{dT}) = \widetilde{\mathcal{O}}(H^2 \cdot \sqrt{d^3 T})$$

LSVI-UCB achieves $\sqrt{T}$-regret

Theorem [Jin-Yang-Wang-Jordan-20] Choosing $\beta = \widetilde{\mathcal{O}}(dH)$, with probability at least $1 - 1/T$,

$$\texttt{Regret}(T) = \widetilde{\mathcal{O}}(\beta \cdot H \cdot \sqrt{dT}) = \widetilde{\mathcal{O}}(H^2 \cdot \sqrt{d^3 T})$$

- Directly imply a $|\mathcal{S}|^{1.5}|\mathcal{A}|^{1.5}H^2\sqrt{T}$ regret for tabular RL

LSVI-UCB achieves $\sqrt{T}$-regret

Theorem [Jin-Yang-Wang-Jordan-20] Choosing $\beta = \widetilde{\mathcal{O}}(dH)$, with probability at least $1 - 1/T$,

$$\texttt{Regret}(T) = \widetilde{\mathcal{O}}(\beta \cdot H \cdot \sqrt{dT}) = \widetilde{\mathcal{O}}(H^2 \cdot \sqrt{d^3 T})$$

- Directly imply a $|\mathcal{S}|^{1.5}|\mathcal{A}|^{1.5}H^2\sqrt{T}$ regret for tabular RL
- First algorithm with both sample and computational efficiency in the context of RL with function approximation

LSVI-UCB achieves $\sqrt{T}$-regret

Theorem [Jin-Yang-Wang-Jordan-20] Choosing $\beta = \widetilde{\mathcal{O}}(dH)$, with probability at least $1 - 1/T$,

$$\texttt{Regret}(T) = \widetilde{\mathcal{O}}(\beta \cdot H \cdot \sqrt{dT}) = \widetilde{\mathcal{O}}(H^2 \cdot \sqrt{d^3 T})$$

- Directly imply a $|\mathcal{S}|^{1.5}|\mathcal{A}|^{1.5}H^2\sqrt{T}$ regret for tabular RL
- First algorithm with both sample and computational efficiency in the context of RL with function approximation
- Only assumption: Image set of $\mathcal{B}$ is in $\mathcal{F}_{\mathrm{lin}}$

LSVI-UCB achieves $\sqrt{T}$-regret

Theorem [Jin-Yang-Wang-Jordan-20] Choosing $\beta = \widetilde{\mathcal{O}}(dH)$, with probability at least $1 - 1/T$,

$$\texttt{Regret}(T) = \widetilde{\mathcal{O}}(\beta \cdot H \cdot \sqrt{dT}) = \widetilde{\mathcal{O}}(H^2 \cdot \sqrt{d^3 T})$$

- Directly imply a $|\mathcal{S}|^{1.5}|\mathcal{A}|^{1.5}H^2\sqrt{T}$ regret for tabular RL
- First algorithm with both sample and computational efficiency in the context of RL with function approximation
- Only assumption: Image set of $\mathcal{B}$ is in $\mathcal{F}_{\text{lin}}$
- Optimal regret $dH^2\sqrt{T}$ is achieved by [Zanette et al. 2020] with a relaxed model assumption. But the computation is intractable.

A general version of regret

Let $\mathcal{Q}_{ucb}$ be the function class containing the Q-functions constructed by LSVI-UCB

- $\mathcal{Q}_{ucb} = \left\{ \mathrm{Trunc}_{[0,H]}\{\phi(\cdot,\cdot)^\top \theta + \beta \cdot \|\phi(\cdot,\cdot)\|_{\Lambda^{-1}}\right\}$ for linear case

A general version of regret

Let $\mathcal{Q}_{ucb}$ be the function class containing the Q-functions constructed by LSVI-UCB

- $\mathcal{Q}_{ucb} = \left\{ \mathrm{Trunc}_{[0,H]}\{\phi(\cdot,\cdot)^\top \theta + \beta \cdot \|\phi(\cdot,\cdot)\|_{\Lambda^{-1}}\right\}$ for linear case

> **Theorem** [Jin-Yang-Wang-Wang-Jordan-20] Choosing $\beta = \widetilde{\mathcal{O}}(H \cdot \sqrt{\log N_\infty(\mathcal{Q}_{ucb}, T^{-2})})$, with probability at least $1 - 1/T$,
> $\mathtt{Regret}(T) = \widetilde{\mathcal{O}}(\beta H \cdot \sqrt{d_{\mathrm{eff}} \cdot T}) = \widetilde{\mathcal{O}}(H^2 \cdot \sqrt{d_{\mathrm{eff}} \cdot T \cdot \log N_\infty})$

A general version of regret

Let $\mathcal{Q}_{ucb}$ be the function class containing the Q-functions constructed by LSVI-UCB

- $\mathcal{Q}_{ucb} = \left\{ \mathrm{Trunc}_{[0,H]}\{\phi(\cdot,\cdot)^\top \theta + \beta \cdot \|\phi(\cdot,\cdot)\|_{\Lambda^{-1}}\right\}$ for linear case

Theorem [Jin-Yang-Wang-Wang-Jordan-20] Choosing $\beta = \widetilde{\mathcal{O}}(H \cdot \sqrt{\log N_\infty(\mathcal{Q}_{ucb}, T^{-2})})$, with probability at least $1 - 1/T$,

$\texttt{Regret}(T) = \widetilde{\mathcal{O}}(\beta H \cdot \sqrt{d_{\mathrm{eff}} \cdot T}) = \widetilde{\mathcal{O}}(H^2 \cdot \sqrt{d_{\mathrm{eff}} \cdot T \cdot \log N_\infty})$

- $\log N_\infty(\mathcal{Q}_{ucb}, T^{-2}) \asymp d \log T$ for linear case
- Include kernel and overparameterized neural network

A general version of regret

Let $\mathcal{Q}_{ucb}$ be the function class containing the Q-functions constructed by LSVI-UCB

- $\mathcal{Q}_{ucb} = \left\{ \mathrm{Trunc}_{[0,H]} \{ \phi(\cdot,\cdot)^\top \theta + \beta \cdot \|\phi(\cdot,\cdot)\|_{\Lambda^{-1}} \right\}$ for linear case

Theorem [Jin-Yang-Wang-Wang-Jordan-20] Choosing $\beta = \widetilde{\mathcal{O}}(H \cdot \sqrt{\log N_\infty(\mathcal{Q}_{ucb}, T^{-2})})$, with probability at least $1 - 1/T$,

$$\mathtt{Regret}(T) = \widetilde{\mathcal{O}}(\beta H \cdot \sqrt{d_{\mathrm{eff}} \cdot T}) = \widetilde{\mathcal{O}}(H^2 \cdot \sqrt{d_{\mathrm{eff}} \cdot T \cdot \log N_\infty})$$

- $\log N_\infty(\mathcal{Q}_{ucb}, T^{-2}) \asymp d \log T$ for linear case
- Include kernel and overparameterized neural network
- d_{eff} is the effective dimension of RKHS or NTK
- For an abstract function class $\mathcal{F}$, d_{eff} can be set as the Bellman-Eluder dimension [Jin-Liu-Miryoosef-21]

Regret analysis

Regret analysis: sensitivity analysis + elliptical potential

A general sensitivity analysis

$$
\begin{aligned}
J(\pi^\star) - J(\pi^t) &= \underbrace{\textstyle\sum_{h\in[H]} \mathbb{E}_{\pi^\star}\left[Q_h^{+,t}(s_h,\pi_h^\star(s_h)) - Q_h^{+,t}(s_h,\pi_h^t(s_h))\right]}_{\text{(i) policy optimization error}} \\
&\quad + \underbrace{\textstyle\sum_{h\in[H]} \mathbb{E}_{\pi^t}\left[\left(Q_h^{+,t} - \mathcal{B}Q_{h+1}^{+,t}\right)\right]}_{\text{(ii) Bellman error on } \pi^t} \\
&\quad + \underbrace{\textstyle\sum_{h\in[H]} \mathbb{E}_{\pi^\star}\left[-\left(Q_h^{+,t} - \mathcal{B}Q_{h+1}^{+,t}\right)\right]}_{\text{(iii) Bellman error on } \pi^\star}
\end{aligned}
$$

- Term (i) ≤ 0 as π^t is greedy with respect to $Q_h^{+,t}$
- Optimism: $Q_{h+1}^{+,t} - 2\Gamma_h^t \le \mathcal{B}Q_{h+1}^{+,t} \le Q_h^{+,t}$, Term (iii) ≤ 0

Regret analysis: sensitivity analysis + elliptical potential

Therefore, we have

$$
\begin{aligned}
\texttt{Regret}(T) &= \sum_{t=1}^{T} J(\pi^\star) - J(\pi^t) \le 2\sum_{h\in[H]} \mathbb{E}_{\pi^t}\left[\Gamma_h^t(s_h, a_h)\right] \\
&= 2\sum_{h\in[H]} \left[\Gamma_h^t(s_h^t, a_h^t)\right] + \mathrm{martingale} - \mathrm{diff} \\
&= 2\beta\sum_{h\in[H]} \sum_{t\in[T]} \left[\|\phi(s_h^t, a_h^t)\|_{(\Lambda_h^t)^{-1}}\right] + H \cdot \sqrt{T} \\
&= \widetilde{\mathcal{O}}(\beta H\sqrt{dT})
\end{aligned}
$$

- Second line holds because $\{(s_h^t, a_h^t), h \in [H]\} \sim \pi^t$

Regret analysis: sensitivity analysis + elliptical potential

Therefore, we have

$$
\begin{aligned}
\texttt{Regret}(T) &= \sum_{t=1}^{T} J(\pi^\star) - J(\pi^t) \le 2\sum_{h\in[H]} \mathbb{E}_{\pi^t}\left[\Gamma_h^t(s_h, a_h)\right] \\
&= 2\sum_{h\in[H]}\left[\Gamma_h^t(s_h^t, a_h^t)\right] + \mathrm{martingale-diff} \\
&= 2\beta\sum_{h\in[H]}\sum_{t\in[T]}\left[\|\phi(s_h^t, a_h^t)\|_{(\Lambda_h^t)^{-1}}\right] + H\cdot\sqrt{T} \\
&= \widetilde{\mathcal{O}}(\beta H\sqrt{dT})
\end{aligned}
$$

- Second line holds because $\{(s_h^t, a_h^t), h \in [H]\} \sim \pi^t$
- It remains to conduct uncertainty quantification:

$$
\mathbb{P}\Big(\forall(t,h,s,a), |\widehat{Q}_h^t(s,a) - (\mathcal{B}Q_{h+1}^{+,t})(s,a)| \le \Gamma_h^t(s,a)\Big) \ge 1-\delta
$$

uniform concentration over $\mathcal{Q}_{ucb}$ (new for RL)+
self-normalized concentration (same as in CB)

Summary & Extensions

- Exploration in RL is more challenging in that we need to handle deep exploration

Summary & Extensions

- Exploration in RL is more challenging in that we need to handle deep exploration
- LSVI-UCB explores by doing uncertainty quantification for LSVI estimator
    - LSVI propagates the uncertainty in larger h steps backward to $h = 1$
    - By construction, Q_1^+ contains the uncertainty about all $h \geq 1$

Summary & Extensions

- Exploration in RL is more challenging in that we need to handle deep exploration
- LSVI-UCB explores by doing uncertainty quantification for LSVI estimator
  - LSVI propagates the uncertainty in larger h steps backward to $h = 1$
  - By construction, Q_1^+ contains the uncertainty about all $h \geq 1$
- LSVI-UCB achieves sample-efficiency and computational tractability in online RL with function approximation

Summary & Extensions

- Exploration in RL is more challenging in that we need to handle deep exploration
- LSVI-UCB explores by doing uncertainty quantification for LSVI estimator
  - LSVI propagates the uncertainty in larger h steps backward to $h = 1$
  - By construction, Q_1^+ contains the uncertainty about all $h \geq 1$
- LSVI-UCB achieves sample-efficiency and computational tractability in online RL with function approximation
- Similar principle can be extended to:
  - proximal policy optimization (use soft-greedy instead of greedy)
  - zero-sum Markov game (two-player extension)
  - constrained MDP (primal dual optimization)

References (a very incomplete list)

- linear bandit
  - [Dani et al, 2008] Stochastic Linear Optimization under Bandit Feedback
  - [Abbasi-Yadkori et al, 2011] Improved algorithms for linear stochastic bandits
- optimism in tabular RL
  - [Auer & Ortner, 2007] Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning
  - [Jaksch et al, 2010] Near-optimal Regret Bounds for Reinforcement Learning
  - [Azar et al, 2017] Minimax Regret Bounds for Reinforcement Learning

References (a very incomplete list)

- RL with function approximation (incomplete list)
  - [Dann et al, 2018] On Oracle-Efficient PAC RL with Rich Observations
  - [Wang & Yang, 2019] Sample-Optimal Parametric Q-Learning Using Linearly Additive Features
  - [Jin et al, 2020] Provably Efficient Reinforcement Learning with Linear Function Approximation (this talk)
  - [Zanette et al, 2020] Learning near optimal policies with low inherent bellman error
  - [Du et al, 2020] Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning?
  - [Xie et al, 2020] Learning Zero-Sum Simultaneous-Move Markov Games Using Function Approximation and Correlated Equilibrium
  - [Yang et al, 2020] On Function Approximation in Reinforcement Learning: Optimism in the Face of Large State Spaces (this talk)
  - [Jin et al, 2021] Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms