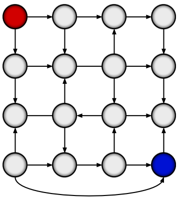
Reinforcement Learning Theory

Tor Lattimore

DeepMind, London

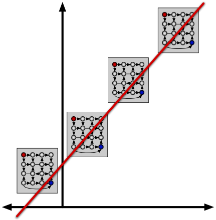
Program

Part one



- MDPs
- Policies/value functions
- Models for learning
- Learning in tabular MDPs
- Optimism

Part two



- Linear function approximation
- Experimental design
- Learning with function approximation
- Linear MDPs

Part three

- Nonlinear function approximation
- Eluder dimension
- Learning with non-linear function approximation
- Further topics

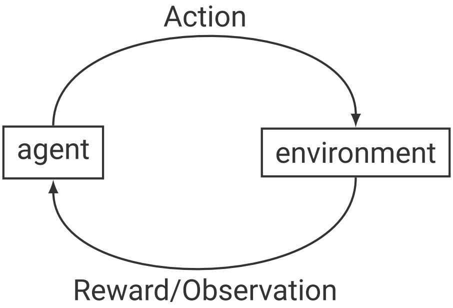
Notes before we start

- Please ask questions anytime!
- There are exercises. If you have time, you will benefit by attempting them. I will update slides with solutions at the end
- Very few prerequisites – elementary probability only
- There are some tricky concepts!
- I am around if you want to chat/ask questions out of lectures

Why RL theory?

- Use theory to guide algorithm design
- Understand what is possible
- Understand why existing algorithms work
- Understand when existing algorithms may not work

Reinforcement Learning



Learner interacts with unknown environment taking actions and receiving observations

Goal is to maximise cumulative reward in some sense

Markov Decision Processes (MDPs)

- An MDP is a tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$
- $\mathcal{S}$ is a finite or countable set of states
- $\mathcal{A}$ is a finite set of actions
- $\mathcal{P}$ is a probability kernel from $\mathcal{S} \times \mathcal{A}$ to $\mathcal{S}$
- $\mathcal{R}$ is a probability kernel from $\mathcal{S} \times \mathcal{A}$ to $[0, 1]$

Notation

$\mathcal{P}(s'|s, a)$ is the probability of transitioning to state s' when taking action a in state s

Mean reward when taking action a in state s is $r(s, a) = \int_{\mathbb{R}} r\mathcal{R}(d\, r|s, a)$

Three types value/interaction protocol

Finite horizon Learner starts in an initial state. Interacts with the MDP for H rounds and is reset to the initial state

Discounted Learner interacts with the MDP without resets. Rewards are geometrically discounted.

Average reward Learner interacts with the MDP without resets. We care about some kind of *average reward*

Discounted and finite horizon are often somehow comparable and technically similar

Average reward introduces a lot of technicalities

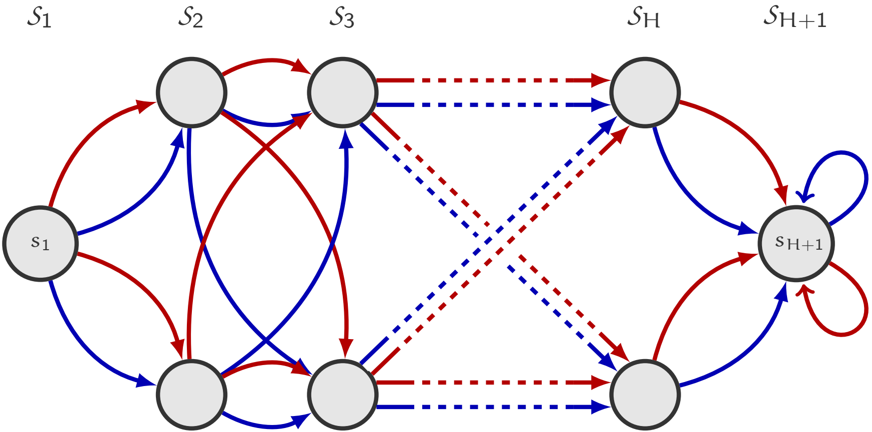
Finite horizon MDPs

- Learner starts in some initial state $s_1 \in \mathcal{S}$
- Interacts with the MDP for H rounds
- **Assume** $\mathcal{S} = \bigsqcup_{h=1}^{H+1} \mathcal{S}_h$ with $\mathcal{S}_1 = \{s_1\}$ and $\mathcal{S}_{H+1} = \{s_{H+1}\}$ and

$$\mathcal{P}(\mathcal{S}_{h+1}|s, a) = 1 \text{ for all } s \in \mathcal{S}_h,\ a \in \mathcal{A} \text{ and } h \in [H]$$

$$\mathcal{P}(s_{H+1}|s_{H+1}, a) = 1 \text{ and } r(s_{H+1}, a) = 0$$

Picture



$\mathcal{S}_1$
$\mathcal{S}_2$
$\mathcal{S}_3$
$\mathcal{S}_H$
$\mathcal{S}_{H+1}$
s_1
s_{H+1}

Finite horizon MDPs

- A stationary deterministic policy is a function $\pi : \mathcal{S} \to \mathcal{A}$
- Policy and MDP induce a probability measure on state/action/reward sequences
- The probability that π and the MDP produce interaction sequence $s_1, a_1, r_1, \ldots, s_H, a_H, r_H$ is

$$\prod_{h=1}^{H} \mathbf{1}_{\pi(s_h)=a_h} \mathcal{R}(r_h | s_h, a_h) \mathcal{P}(s_{h+1} | s_h, a_h)$$

- Expectations with respect to this measure are denoted by $\mathbb{E}_\pi$. For example,

$$v^\pi(s_1) = \mathbb{E}_\pi \left[\sum_{h=1}^{H} r_h \right]$$

is the expected cumulative reward over one episode

Value and q-value functions

Given a stationary policy π the value function $v^\pi : \mathcal{S} \to \mathbb{R}$ is the function

$$v^\pi(s) = \mathbb{E}_\pi\left[\sum_{u=h}^{H} r_u \middle| s_h = s\right] \text{ for all } s \in \mathcal{S}_h$$

Questionable formalism: What if s is not reachable under π

q-values

Value of policy π when starting in state s, taking action a and following π subsequently

$$q^\pi(s, a) = r(s, a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, a) v^\pi(s')$$

Bellman operator

- Operators on the spaces of value/q-value functions

$$(\mathcal{T}^{\pi}q)(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a)q(s',\pi(s'))$$

$$(\mathcal{T}^{\pi}v)(s) = r(s,\pi(s)) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,\pi(s))v(s')$$

Exercise 1 Prove that

1. v^{π} is the unique fixed point of $\mathcal{T}^{\pi}$ over all functions $\{q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$
2. q^{π} is the unique fixed point of $\mathcal{T}^{\pi}$ over all functions $\{v : \mathcal{S} \to \mathbb{R}\}$

Optimal policies

The optimal policy maximises v^π over all states

$$v^\star(s) = \max_\pi v^\pi(s)$$

Proposition 1 There exists a stationary policy $\pi^\star$ such that

$$v^{\pi^\star}(s) = v^\star(s) \text{ for all } s \in \mathcal{S}$$

Optimal q-value is $q^\star(s, a) = q^{\pi^\star}(s, a)$

Optimal policies

The optimal policy maximises v^π over all states

$$v^\star(s) = \max_\pi v^\pi(s)$$

Proposition 1 There exists a stationary policy $\pi^\star$ such that

$$v^{\pi^\star}(s) = v^\star(s) \text{ for all } s \in \mathcal{S}$$

Optimal q-value is $q^\star(s,a) = q^{\pi^\star}(s,a)$

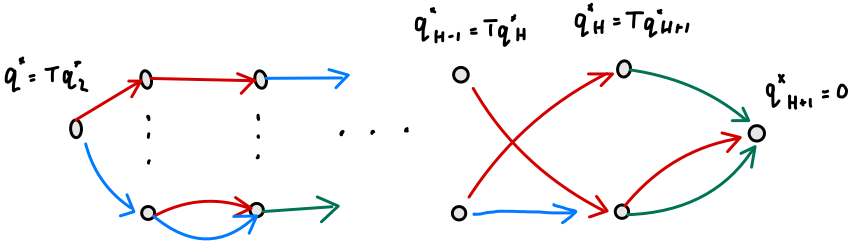
Bellman optimality operator

$$(\mathcal{T}q)(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) \max_{a' \in \mathcal{A}} q(s',a')$$

$$(\mathcal{T}v)(s) = \max_{a \in \mathcal{A}} r(s,a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) v(s')$$

Exercise 2 Show that

1. $q^\star$ is the unique fixed point of $\mathcal{T}$ over all functions $\{q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$
2. $v^\star$ is the unique fixed point of $\mathcal{T}$ over all functions $\{v : \mathcal{S} \to \mathbb{R}\}$



$$q^*_h(s,a) = r(s,a) + \sum_{s'} P(s'|s,a) \max_{a'} q^*_{h+1}(s',a')$$

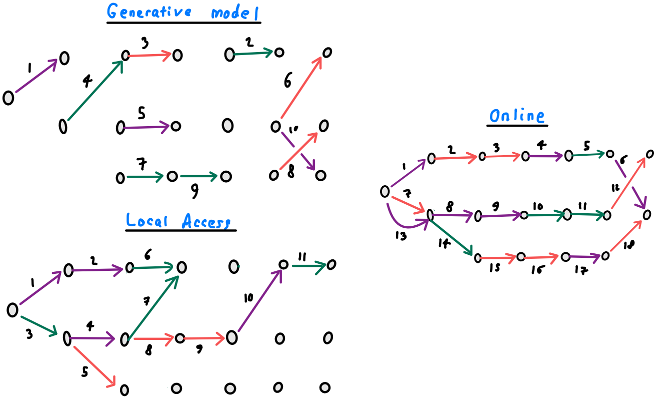
$$= (Tq_{h+1})(s,a)$$

Three Learning Models

Online RL the learner interacts with the environment as if it were in the real world

Local planning the learner can 'query' the environment at any state/action pair it has seen before

Generative model the learner can 'query' the environment at any state/action pair



Learning with a Generative Model

- Learner knows $\mathcal{S}, \mathcal{A}$ and the initial state s_1 but not $\mathcal{P}$ and r
- Learner and environment interact sequentially
- Learner chooses any state/action pair $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$
- Observes r_t, s'_t with $r_t \sim \mathcal{R}(s_t, a_t)$ and $s'_t \sim \mathcal{P}(s_t, a_t)$

How many samples are needed to find a near optimal policy?

Sometimes want a policy that is near optimal at *all* states. Sometimes only care about the initial state

Same as learning with a generative model, but learner can only query states it has observed before

- Learner knows $\mathcal{S}$, $\mathcal{A}$ and s_1 but not $\mathcal{P}$ and r
- $\mathcal{S}_1 = \{s_1\}$
- Learner and environment interact sequentially
- Learner chooses any $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$ with $s_t \in \mathcal{S}_t$
- Observes r_t, s_t' with $r_t \sim \mathcal{R}(s_t, a_t)$ and $s_t' \sim \mathcal{P}(\cdot|s_t, a_t)$
- $\mathcal{S}_{t+1} = \mathcal{S}_t \cup \{s_t'\}$

Online RL

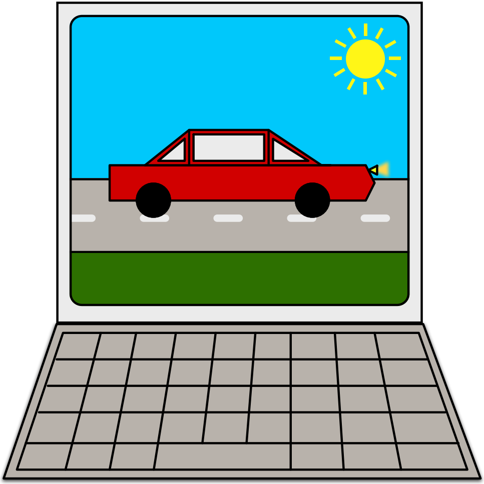
- Learner knows $\mathcal{S}$ and $\mathcal{A}$ but not $\mathcal{P}$ and r
- Learner interacts with MDP in episodes
- In episode k the learner starts in state $s_1^k = s_1$ and interacts with the MDP for H rounds producing history

$$s_1^k, a_1^k, r_1^k, s_2^k, a_2^k, \ldots, r_H^k, s_H^k$$

- a_t^k is the action played by the learner in state s_t^k
- s_t^k is sampled from $\mathcal{P}(\cdot|s_{t-1}^k, a_{t-1}^k)$
- r_t^k is sampled from $\mathcal{R}(s_t^k, a_t^k)$

How many times does the learner play a suboptimal policy? How small is the **regret**?

Example of local planning

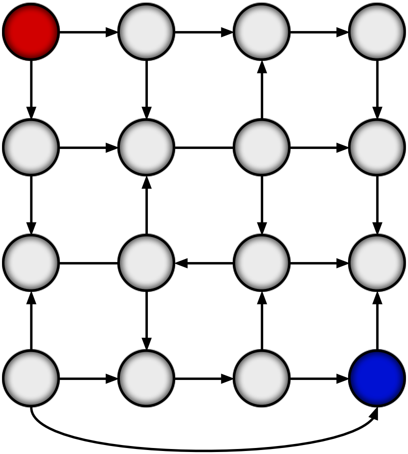


Finding optimal policies in simulators

Bandits

- A (very simple) bandit is a single-state MDP: $\mathcal{S} = \{s_1\}$
- Useful examples
- Can be analysed very deeply
- Ideas often generalise to RL – optimism, Thompson sampling and many other exploration techniques were first introduced in bandits
- Practical in their own right

Tabular MDPs



Learning tabular MDPs with a generative model

- Unknown MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathrm{r})$
- Access to a generative model

How many queries to the generative model are needed to find a near-optimal policy?

Learning tabular MDPs with a generative model

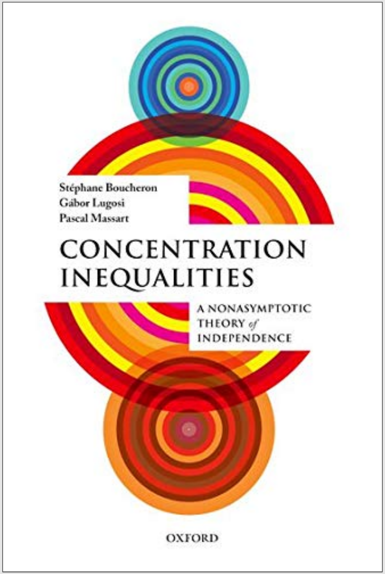
- Suppose we make m queries to the generative model from each state-action pair
- Observe data $(s_t, a_t, r_t, s_t')_{t=1}^n$ with $n = m|\mathcal{S}||\mathcal{A}|$
- Estimate p and r by

$$\hat{\mathcal{P}}(s'|s, a) = \frac{1}{m} \sum_{t=1}^{n} \mathbf{1}_{(s,a,s')=(s_t,a_t,s_t')}$$

$$\hat{r}(s, a) = \frac{1}{m} \sum_{t=1}^{n} \mathbf{1}_{(s,a)=(s_t,a_t)} r_t$$

- Output the optimal policy π for the empirical MDP $(\mathcal{S}, \mathcal{A}, \hat{\mathcal{P}}, \hat{r})$

Necessary aside



Hoeffding's bound Suppose that $X_1, \dots, X_m$ are i.d.d. random variables in $[-B, B]$ with mean μ. Then, for all $\delta \in (0,1)$,

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geqslant \text{cnst}\, B\sqrt{\frac{\log(1/\delta)}{m}}\right) \leqslant \delta$$

Categorical concentration Suppose that $\mathcal{S}_1, \dots, \mathcal{S}_m$ are i.d.d. random elements in $\mathcal{S}$ sampled from P and $\hat{P}(s) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{S_i = s}$. Then,

$$\mathbb{P}\left(\|P - \hat{P}\|_1 \geqslant \text{cnst}\sqrt{\frac{|\mathcal{S}|\log(1/\delta)}{m}}\right) \leqslant \delta\,.$$

Analysis

- π^* is true optimal policy
- π is optimal policy of empirical MDP
- $\hat{v}$ is value function of empirical MDP

$$\underbrace{v^{\pi^\star}(s_1) - v^{\pi}(s_1)}_{\text{error}} = \underbrace{v^{\pi^\star}(s_1) - \hat{v}^{\pi^\star}(s_1)}_{\text{(A)}} + \underbrace{\hat{v}^{\pi^\star}(s_1) - \hat{v}^{\pi}(s_1)}_{\text{(B)}} + \underbrace{\hat{v}^{\pi}(s_1) - v^{\pi}(s_1)}_{\text{(C)}}$$

- (B) is negative because π is optimal in empirical MDP
- (A) and (C) are the differences in value functions with a given policy

A useful lemma

- Compare the values of a single policy on different MDPs
- $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$ with value function $v : \mathcal{S} \to \mathbb{R}$
- $\hat{M} = (\mathcal{S}, \mathcal{A}, \hat{\mathcal{P}}, \hat{r})$ with value function $\hat{v} : \mathcal{S} \to \mathbb{R}$

Lemma 1 (Value decomposition lemma) For all policies π

$$v^\pi(s_1) - \hat{v}^\pi(s_1) = \mathbb{E}_\pi\left[\sum_{h=1}^{H} (r - \hat{r})(s_h, a_h) + \langle \mathcal{P}(s_h, a_h) - \hat{\mathcal{P}}(s_h, a_h), \hat{v}^\pi \rangle\right]$$

Exercise 3 Prove Lemma 1

Bounding the error

With probability at least $1-\delta$ for all $s \in \mathcal{S}_h$ and $a \in \mathcal{A}$,

$$|(r-\hat{r})(s,a)| \leqslant \text{cnst} \sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}|/\delta)}{m}}$$

$$\|\hat{\mathcal{P}}(s,a) - \mathcal{P}(s,a)\|_1 \leqslant \text{cnst} \sqrt{\frac{|\mathcal{S}_{h+1}| \log(|\mathcal{S}||\mathcal{A}|/\delta)}{m}}$$

By Lemma 1

$$v^\pi(s_1) - \hat{v}^\pi(s_1) = \mathbb{E}_\pi\left[\sum_{h=1}^H (r-\hat{r})(s_h,a_h) + \langle \mathcal{P}(s_h,a_h) - \hat{\mathcal{P}}(s_h,a_h), \hat{v}^\pi\rangle\right]$$

$$\leqslant \mathbb{E}_\pi\left[\sum_{h=1}^H (r-\hat{r})(s_h,a_h) + \|\mathcal{P}(s_h,a_h) - \hat{\mathcal{P}}(s_h,a_h)\|_1 \|\hat{v}^\pi\|_\infty\right]$$

$$\leqslant \text{cnst}\, \mathbb{E}_\pi\left[\sum_{h=1}^H \sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}|/\delta)}{m}} + H\sqrt{\frac{|\mathcal{S}_{h+1}| \log(|\mathcal{S}||\mathcal{A}|/\delta)}{m}}\right]$$

$$\leqslant \text{cnst}\, |\mathcal{S}| \sqrt{\frac{|\mathcal{A}| H^3 \log(|\mathcal{S}||\mathcal{A}|/\delta)}{\#\text{queries}}} \qquad m = \frac{\#\text{queries}}{|\mathcal{S}||\mathcal{A}|}$$

Bounding the error

By the same argument

$$\hat{v}^{\pi^\star}(s_1) - v^{\pi^\star}(s_1) \leqslant \text{cnst}\, |\mathcal{S}| \sqrt{\frac{H^3 |\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|/\delta)}{\#\text{queries}}}$$

Combining everything gives:

Theorem 2 The optimal policy in the empirical MDP satisfies with probability at least $1 - \delta$

$$v^{\pi^\star}(s_1) - v^{\pi}(s_1) \leqslant \text{cnst}\, |\mathcal{S}| \sqrt{\frac{|\mathcal{A}| H^3 \log(|\mathcal{S}||\mathcal{A}|/\delta)}{\#\text{queries}}}$$

Corollary 3 If

$$\#\text{queries} \geqslant \frac{\text{cnst}\, H^3 |\mathcal{S}|^2 |\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|/\delta)}{\varepsilon^2}$$

Then with probability at least $1 - \delta$, $v^{\pi^\star}(s_1) - v^{\pi}(s_1) \leqslant \varepsilon$

Are these bounds tight?

Number of samples needed for ε-accuracy

$$n = \frac{\text{cnst}\, H^3 |\mathcal{S}|^2 |\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|/\delta)}{\varepsilon^2}$$

- $1/\varepsilon^2$ is the standard statistical dependency – likely optimal
- $|\mathcal{A}||\mathcal{S}|^2$ parameters in the transition matrix
- Rewards scale in $[0, H]$
- A good guess would be $\frac{H^2 |\mathcal{S}|^2 |\mathcal{A}| \log(1/\delta)}{\varepsilon^2}$

Dependence on $|\mathcal{S}|$

Key inequality:

$$\langle \hat{\mathcal{P}}(s,a) - \mathcal{P}(s,a), \hat{v}^\pi \rangle \leqslant \|\hat{\mathcal{P}}(s,a) - \mathcal{P}(s,a)\|_1 \|\hat{v}^\pi\|_\infty$$

Remember,

$$\hat{\mathcal{P}}(s'|s,a) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{s_i = s'}$$

Then

$$\begin{aligned}\langle \hat{\mathcal{P}}(s,a) - \mathcal{P}(s,a), \hat{v}^\pi \rangle &= \sum_{s'} (\hat{\mathcal{P}}(s'|s,a) - \mathcal{P}(s'|s,a))\hat{v}^\pi(s') \\ &= \frac{1}{m}\sum_{i=1}^{m} \underbrace{\sum_{s'} (\mathbf{1}_{s_i=s'} - \mathcal{P}(s'|s,a))\hat{v}^\pi(s')}_{\Delta_i} \\ &\overset{\text{whp}}{\leqslant} \text{cnst}\, H\sqrt{\frac{\log(1/\delta)}{m}}\end{aligned}$$

$(\Delta_i)_{i=1}^{m}$ are independent and $|\Delta_i| \leqslant H$ and $\mathbb{E}[\Delta_i] = 0$

Dependence on $|\mathcal{S}|$

Repeating the previous analysis gives the following

Theorem 4 If

$$n = \frac{\text{cnst } H^4 |\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|/\delta)}{\varepsilon^2}$$

then with probability at least $1-\delta$, $v^{\pi^\star}(s_1) - v^{\pi}(s_1) \leqslant \varepsilon$

Exercise 4 Prove Theorem 4

Dependence on H

Dependence on H is also loose

$$\sigma_\pi^2(s,a) = \mathbb{V}_{s' \sim \mathcal{P}(s,\pi(s))}[v^\pi(s')]$$

Exercise 5 (Sobel 1982) Show that

$$\mathbb{V}_\pi\left[\sum_{h=1}^{H} r_h\right] = \mathbb{E}_\pi\left[\sum_{h=1}^{H} \sigma_\pi^2(s_h, a_h)\right]$$

Naive bounds

$$\mathbb{V}_\pi\left[\sum_{h=1}^{H} r_h\right] \leqslant H^2 \qquad \mathbb{E}_\pi\left[\sum_{h=1}^{H} \sigma_\pi^2(s_h, a_h)\right] \leqslant H^3$$

Dependence on H

Repeating the previous analysis and assuming known rewards (again...)

$$
\begin{aligned}
v^\pi(s_1) - \hat{v}^\pi(s_1) &= \mathbb{E}_\pi\left[\sum_{h=1}^{H} \langle \mathcal{P}(s_h, a_h) - \hat{\mathcal{P}}(s_h, a_h), \hat{v}^\pi \rangle\right] \\
&\lesssim \mathbb{E}_\pi\left[\sum_{h=1}^{H} \sqrt{\frac{\sigma_\pi^2(s_h, a_h)\log(|\mathcal{S}||\mathcal{A}|/\delta)}{m}}\right] \\
&\leqslant \sqrt{\frac{H}{m}\mathbb{E}_\pi\left[\sum_{h=1}^{H} \sigma_\pi^2(s_h, a_h)\right]\log(|\mathcal{S}||\mathcal{A}|/\delta)} \\
&= \sqrt{\frac{H}{m}\mathbb{V}_\pi\left[\sum_{h=1}^{H} r_h\right]\log(|\mathcal{S}||\mathcal{A}|/\delta)} \\
&\leqslant \sqrt{\frac{H^3|\mathcal{S}||\mathcal{A}|}{\#\text{queries}}\log(|\mathcal{S}||\mathcal{A}|/\delta)}
\end{aligned}
$$

Final result

Theorem 5 (Azar et al. 2012) If

$$n = \frac{\mathsf{cnst}\,|\mathcal{S}||\mathcal{A}|H^3 \log(|\mathcal{S}||\mathcal{A}|/\delta)}{\varepsilon^2} + \text{lower order}$$

then $v^{\pi^\star}(s_1) - v^{\pi}(s_1) \leqslant \varepsilon$

Matches lower bound up to constant factors and lower order terms

Learning online

- Online model
- Learner starts at initial state and interacts with the MDP for an episode
- Cannot explore arbitrary states as usual
- Our learner will choose stationary polices $\pi_1, \ldots, \pi_n$ over n episodes
- Assume the reward function is known in advance for simplicity

Regret

Regret is the difference between the expected rewards collected by the optimal policy and the rewards collected by the learner

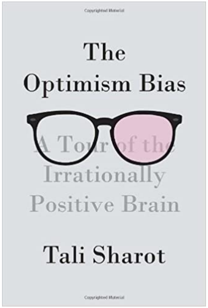
$$\mathrm{Reg}_n = \sum_{t=1}^{n} v^\star(s_1) - v^{\pi_t}(s_1)$$

Exploration/exploitation dilemma

Learner wants to play π_t with v^{π_t} close to $v^\star$ but needs to gain information as well

Optimism

Standard tool for acting in the face of uncertainty since Lai [1987] and Auer et al. [2002]



Intuition

Act as if the world is as **rewarding** as **plausibly** possible

Mathematically

$$\pi_t = \arg\max_{\pi} \max_{M \in \mathcal{C}_{t-1}} v_M^\pi(s_1)$$

where $\mathcal{C}_{t-1}$ is a confidence set containing the true MDP with high probability constructed using data from the first $t-1$ episodes

Confidence set

$$\mathcal{C}_t = \left\{Q : \|Q(s,a) - \hat{\mathcal{P}}_t(s,a)\|_1 \leqslant \text{cnst}\sqrt{\frac{|\mathcal{S}_s|\log(n|\mathcal{S}||\mathcal{A}|)}{1+N_t(s,a)}}\ \forall s,a\right\}$$

where

- $N_t(s,a)$ is the number of times the algorithm played action a in state s in the first t episodes
- $\mathcal{S}_s$ is the number of states in the layer after state s

Proposition 2 $\mathcal{P} \in \mathcal{C}_t$ for all episodes $t \in [n]$ with probability at least $1 - 1/n$

Exercise 6 Prove Proposition 2

Optimism

Regret in episode t is

$$v^\star(s_1) - v^{\pi_t}(s_1)$$

Optimistic environment/policy

$$Q_t = \operatorname*{arg\,max}_{Q \in \mathcal{C}_{t-1}} \max_\pi v_Q^\pi(s_1) \qquad \pi_t = \operatorname*{arg\,max}_\pi v_{Q_t}^\pi(s_1)$$

Key point: if $\mathcal{P} \in \mathcal{C}_{t-1}$, then

$$v_Q^{\pi_t}(s_1) \geqslant v^*(s_1)$$

Regret in episode t is

$$v^\star(s_1) - v^{\pi_t}(s_1)$$

Optimistic environment/policy

$$Q_t = \arg\max_{Q \in \mathcal{C}_{t-1}} \max_\pi v_Q^\pi(s_1) \qquad \pi_t = \arg\max_\pi v_{Q_t}^\pi(s_1)$$

Key point: if $\mathcal{P} \in \mathcal{C}_{t-1}$, then

$$v_Q^{\pi_t}(s_1) \geqslant v^*(s_1)$$

$$\boxed{v^\star(s_1) - v^{\pi_t}(s_1) \leqslant v_{Q_t}^{\pi_t}(s_1) - v^{\pi_t}(s_1)}$$

Analysis

$$\mathbb{E}[\mathrm{Reg}_n] = \mathbb{E}\left[\sum_{t=1}^{n} v^\star(s_1) - v^{\pi_t}(s_1)\right]$$

$$\lesssim \mathbb{E}\left[\sum_{t=1}^{n} v_{Q_t}^{\pi_t}(s_1) - v^{\pi_t}(s_1)\right] \qquad \text{(Optimism)}$$

$$= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H} \langle Q_t(s_h^t, a_h^t) - \mathcal{P}(s_h^t, a_h^t), v_{Q_t}^{\pi_t}\rangle\right] \qquad \text{(Lemma 1)}$$

$$= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H} \|Q_t(s_h^t, a_h^t) - \mathcal{P}(s_h^t, a_h^t)\|_1 \|v_{Q_t}^{\pi_t}\|_\infty\right] \qquad \text{(Hölder's inequality)}$$

$$\leqslant \mathrm{cnst}\, H\mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H} \sqrt{\frac{|\mathcal{S}_{h+1}|\log(1/\delta)}{1 + N_{t-1}(s_h^t, a_h^t)}}\right] \qquad \text{(Def. of conf.)}$$

Analysis (cont)

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_n] &\leqslant \mathsf{cnst}\, H \mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H}\sqrt{\frac{|\mathcal{S}_{h+1}|\log(1/\delta)}{1+N_{t-1}(s_h^t,a_h^t)}}\right] \\
&= \mathsf{cnst}\, H \mathbb{E}\left[\sum_{h=1}^{H}\sum_{s\in\mathcal{S}_h}\sum_{a\in\mathcal{A}}\sum_{t=1}^{n}\mathbf{1}_{s_h^t=s,a_h^t=a}\sqrt{\frac{|\mathcal{S}_{h+1}|\log(n|\mathcal{A}||\mathcal{S}|/\delta)}{1+N_{t-1}(s,a)}}\right] \\
&= \mathsf{cnst}\, H \mathbb{E}\left[\sum_{h=1}^{H}\sum_{s\in\mathcal{S}_h}\sum_{a\in\mathcal{A}}\sum_{u=0}^{N_{n-1}(s,a)}\sqrt{\frac{|\mathcal{S}_{h+1}|\log(n|\mathcal{A}||\mathcal{S}|/\delta)}{1+u}}\right] \\
&\leqslant \mathsf{cnst}\, H \mathbb{E}\left[\sum_{h=1}^{H}\sum_{s\in\mathcal{S}_h}\sum_{a\in\mathcal{A}}\sqrt{N_{n-1}(s,a)|\mathcal{S}_{h+1}|\log(n|\mathcal{S}||\mathcal{A}|/\delta)}\right] \\
&\leqslant \mathsf{cnst}\, H|\mathcal{S}|\sqrt{|\mathcal{A}|n\log(n|\mathcal{A}||\mathcal{S}|/\delta)}
\end{aligned}$$

Summary

- Regret of optimistic algorithm is

$$\mathbb{E}[\mathrm{Reg}_n] \leqslant \text{cnst } H|\mathcal{S}|\sqrt{|\mathcal{A}|n\log(n|\mathcal{A}||\mathcal{S}|)}$$

- With better confidence intervals and analysis

$$\mathbb{E}[\mathrm{Reg}_n] \leqslant \text{cnst } H\sqrt{|\mathcal{S}||\mathcal{A}|n\log(n|\mathcal{A}||\mathcal{S}|)}$$

Exercise 7 Show how to compute the optimistic algorithm in polynomial time

Exercise 8 Modify the algorithm to handle unknown rewards

Algorithm sometimes called UCRL (Upper Confidence for RL)

Original designed for average reward MDP setting [Auer et al., 2009]

Comparing to the bounds with a generative model

- Best regret bound: $\mathbb{E}[\mathrm{Reg}_n] \leqslant \mathsf{cnst}\, H\sqrt{|\mathcal{S}||\mathcal{A}|n \log(n|\mathcal{A}||\mathcal{S}|)}$
- Average regret

$$\frac{1}{n}\mathbb{E}[\mathrm{Reg}_n] \leqslant \mathsf{cnst}\, H\sqrt{\frac{|\mathcal{S}||\mathcal{A}| \log(n|\mathcal{A}||\mathcal{S}|)}{n}} \leqslant \varepsilon$$
$$\iff n \geqslant \frac{H^2|\mathcal{S}||\mathcal{A}| \log(n|\mathcal{A}||\mathcal{S}|)}{\varepsilon^2}$$

- H queries per episode
- Sample complexity with a generative model:

$$\frac{H^3|\mathcal{S}||\mathcal{A}| \log(|\mathcal{A}||\mathcal{S}|/\delta)}{\varepsilon^2}$$

Relation to sample complexity

An alternative notion to regret

A learner is (ε, δ)-PAC if

$$\mathbb{P}\left(\sum_{t=1}^{n} \mathbf{1}\left(v^*(s_1) - v^{\pi_t}(s_1) \geqslant \varepsilon\right) \geqslant S(\varepsilon, \delta)\right) \leqslant \delta$$

Similar optimistic algorithm has

$$S(\varepsilon, \delta) \leqslant \frac{\text{cnst}\, H^2 |\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|/\delta)}{\varepsilon^2}$$

Sample complexity bounds like this imply regret bounds

Dann et al. [2017]

Other algorithmic approaches

- UCB-VI [Azar et al., 2017]: Backwards induction in each episode

$$\tilde{q}(s,a) = \hat{r}(s,a) + \langle \hat{\mathcal{P}}(s,a), \tilde{v} \rangle + \text{bonus} \qquad \tilde{v}(s) = \max_a \tilde{q}(s,a)$$

- Thompson sampling [Ouyang et al., 2017]

$$Q_t \sim \text{Posterior}_{t-1} \qquad \pi_t = \arg\max_\pi v^\pi_{Q_t}$$

- Information-directed sampling [Lu et al., 2021]

$$\pi_t = \arg\min_\pi \frac{\text{Regret}(\pi)^2}{\text{Inf. gain}(\pi)}$$

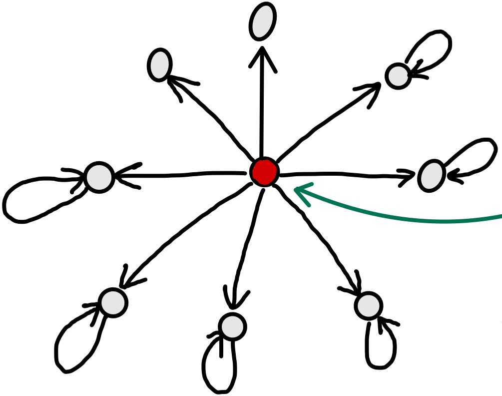
- Optimistic Q-Learning [Jin et al., 2018]

$$q(s,a) \leftarrow (1-\alpha_t) q(s,a) + \alpha_t (r + \max_{a' \in \mathcal{A}} q(s',a') + b_t)$$

- E2D [Foster et al., 2021]

Value-seeking vs information-seeking

Bandit with Graph Feedback



Informative Action

A note on conservative algorithms

Algorithms that use confidence intervals for exploration are at the mercy of their designers cleverness

Loose confidence intervals $\iff$ slow learning

Confidence intervals based on asymptotics may not be valid – can lead to linear (!) regret

Maybe the focus on minimax bounds is misguided

- Instance-dependent regret is well understood in bandits

$$\mathrm{Reg}_n = O\left(\sum_{a:\Delta_a>0} \frac{\log(n)}{\Delta_a}\right)$$

- We have asymptotic problem-dependent bounds for MDPs [Tirinzoni et al., 2021]
- Hard to tell how relevant asymptotic-style problem-dependent bounds are for MDPs

Real problem $|\mathcal{S}|$ is usually enormous

- Our hypothesis class does not encode enough structure
- Number of states in most interesting problems is enormous
- May **never** see the same state multiple times
- We need ways to impose structure on huge MDPs
- Conflicting goals:

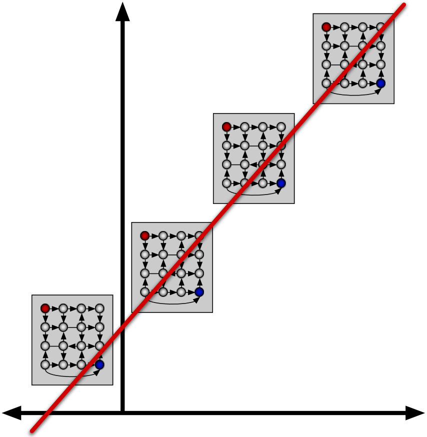
Structure needs to be

- restrictive enough that learning is possible
- flexible enough that the true environment is (approximately) in the class

Linear function approximation

Slides available at

https://tor-lattimore.com/downloads/RLTheory.pdf



Function approximation

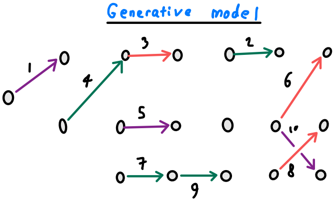
Represent (part of) huge MDP by low(er) dimension objects

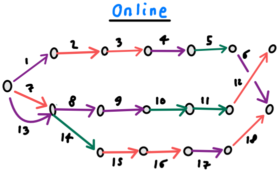
There are lots of choices

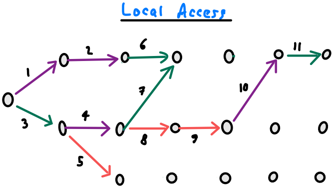
- Represent MDP dynamics and rewards (model-based)
- Represent value functions or q-value functions (model-free)

Remember

- MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$
- Value functions: $v^\pi : \mathcal{S} \to \mathbb{R}$ and $q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Optimal value functions: $v^\star$ and $q^\star$
- Rewards in $[0, 1]$. Episodes of length H
- Layered MDP assumption







Linear function approximation

- Let $\phi(s, a) \in \mathbb{R}^d$ be a feature vector associated with each state/action pair
- Assume that for all policies π there exists a θ such that

$$q^{\pi}(s, a) = \langle \phi(s, a), \theta \rangle$$

- Dynamics may still be incredibly complicated
- But generalisation across q-values is now possible

A necessary aside (linear regression)

Least squares

- Given covariates $a_1, \ldots, a_n \in \mathbb{R}^d$ and responses $y_1, \ldots, y_n$ with

$$y_t = \langle a_t, \theta_\star \rangle + \eta_t$$

$\theta_\star \in \mathbb{R}^d$ is unknown $(\eta_t)_{t=1}^n$ is noise and y_t bounded in $[-H, H]$

Least squares

- Given covariates $a_1, \ldots, a_n \in \mathbb{R}^d$ and responses $y_1, \ldots, y_n$ with

$$y_t = \langle a_t, \theta_\star \rangle + \eta_t$$

  $\theta_\star \in \mathbb{R}^d$ is unknown $(\eta_t)_{t=1}^n$ is noise and y_t bounded in $[-H, H]$

- Estimate $\theta_\star$ with least squares

$$\hat{\theta} = \arg\min_\theta \sum_{t=1}^n (\langle a_t, \theta \rangle - y_t)^2 = G^{-1} \sum_{t=1}^n a_t y_t$$

  with $G = \sum_{t=1}^n a_t a_t^\top$ the design matrix

Least squares

- Given covariates $a_1, \ldots, a_n \in \mathbb{R}^d$ and responses $y_1, \ldots, y_n$ with

$$y_t = \langle a_t, \theta_\star \rangle + \eta_t$$

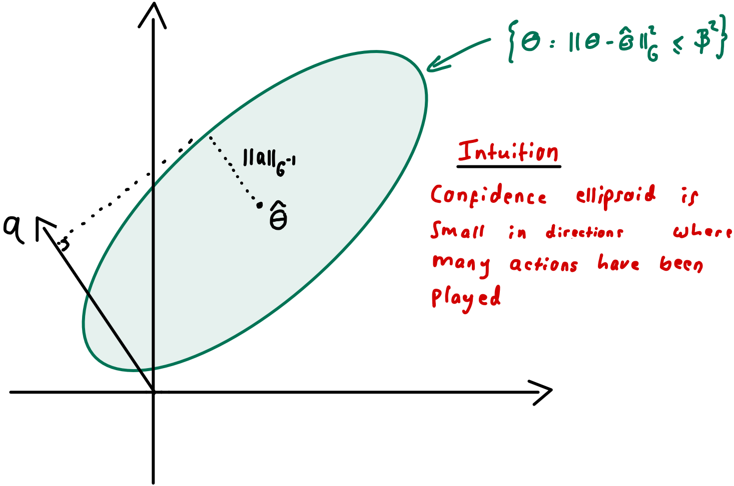
  $\theta_\star \in \mathbb{R}^d$ is unknown $(\eta_t)_{t=1}^n$ is noise and y_t bounded in $[-H, H]$
- Estimate $\theta_\star$ with least squares

$$\hat{\theta} = \arg\min_\theta \sum_{t=1}^n (\langle a_t, \theta \rangle - y_t)^2 = G^{-1} \sum_{t=1}^n a_t y_t$$

  with $G = \sum_{t=1}^n a_t a_t^\top$ the design matrix
- Conc: $\|\hat{\theta} - \theta_\star\|_G \triangleq \|G^{1/2}(\hat{\theta} - \theta_\star)\| \lesssim H\sqrt{\log(1/\delta) + d} \triangleq \beta$

$$|\langle a, \hat{\theta} - \theta_\star \rangle| \leqslant \|a\|_{G^{-1}} \|\hat{\theta} - \theta_\star\|_G \leqslant \beta \|a\|_{G^{-1}}$$

Geometric interpretation



Intuition

Confidence ellipsoid is small in directions where many actions have been played

Experimental design

- Suppose we get to choose $a_1, \ldots, a_n$ from $\mathcal{A} \subset \mathbb{R}^d$
- Estimate $\theta_\star$ by $\hat{\theta}$
- Error in direction a is proportion to $\|a\|_{G^{-1}}$
- How to choose the design to minimise $\max_{a \in \mathcal{A}} \|a\|_{G^{-1}}$

Experimental design

- Suppose we get to choose $a_1, \ldots, a_n$ from $\mathcal{A} \subset \mathbb{R}^d$
- Estimate $\theta_\star$ by $\hat{\theta}$
- Error in direction a is proportion to $\|a\|_{G^{-1}}$
- How to choose the design to minimise $\max_{a \in \mathcal{A}} \|a\|_{G^{-1}}$

Theorem 6 (Kiefer and Wolfowitz 1960) For all compact $\mathcal{A} \subset \mathbb{R}^d$ there exists a distribution ρ on $\mathcal{A}$ such that

$$\max_{a \in \mathcal{A}} \|a\|_{G(\rho)^{-1}} \leqslant \sqrt{d} \qquad\qquad G(\rho) = \sum_{a \in \mathcal{A}} \rho(a) a a^\top$$

Experimental design

- Suppose we get to choose $a_1, \ldots, a_n$ from $\mathcal{A} \subset \mathbb{R}^d$
- Estimate $\theta_\star$ by $\hat{\theta}$
- Error in direction a is proportion to $\|a\|_{G^{-1}}$
- How to choose the design to minimise $\max_{a \in \mathcal{A}} \|a\|_{G^{-1}}$

Theorem 6 (Kiefer and Wolfowitz 1960) For all compact $\mathcal{A} \subset \mathbb{R}^d$ there exists a distribution ρ on $\mathcal{A}$ such that

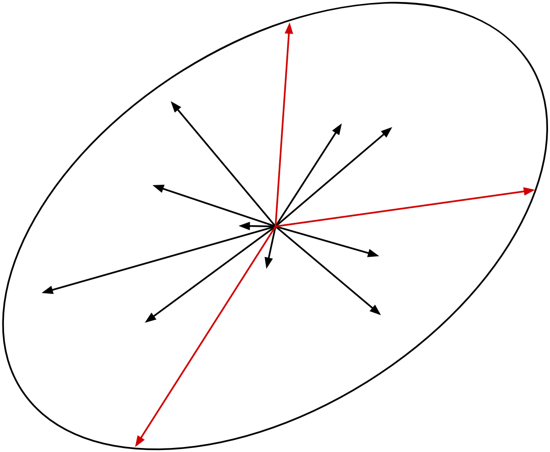
$$\max_{a \in \mathcal{A}} \|a\|_{G(\rho)^{-1}} \leqslant \sqrt{d} \qquad\qquad G(\rho) = \sum_{a \in \mathcal{A}} \rho(a) a a^\top$$

If we choose n experiments $a_1, \ldots, a_n$ in proportion to ρ, then

$$\max_{a \in \mathcal{A}} |\langle a, \hat{\theta} - \theta_\star \rangle| \overset{whp}{\leqslant} \beta \|a\|_{G^{-1}} = \beta \sqrt{\frac{\|a\|^2_{G(\rho)^{-1}}}{n}} \leqslant \beta \sqrt{\frac{d}{n}}$$

Geometric interpretation

Kiefer-Wolfowitz distribution is supported on (a subset of) the minimum volume centered ellipsoid containing $\mathcal{A}$



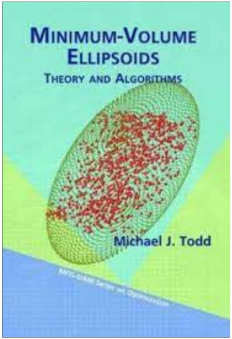
Computation

The support of the Kiefer-Wolfowitz distribution may have size $d(d+1)/2$

Theorem 7 There exists a distribution π supported on at most $\mathrm{cnst}\, d \log\log d$ points such that

$$\|a\|^2_{G(\pi)^{-1}} \leqslant 2d$$

Can be found using Frank-Wolfe and careful initialisation [Todd, 2016]



Least-squares 'policy iteration'

Generative model setting

Given feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$

Assumption 1 For all π there exists a θ such that $q^\pi(s, a) = \langle \theta, \phi(s, a) \rangle$

Start with arbitrary policy π_{H+1}
for h = H to 1

- Estimate $q^{\pi_{h+1}}$ by some $\hat{q}^{h+1}$
- Update policy $\pi_h(s) = \begin{cases} \pi_{h+1}(s) & \text{if } s \notin \mathcal{S}_h \\ \operatorname{arg\,max}_{a \in \mathcal{A}} \hat{q}^{h+1}(s, a) & \text{otherwise} \end{cases}$

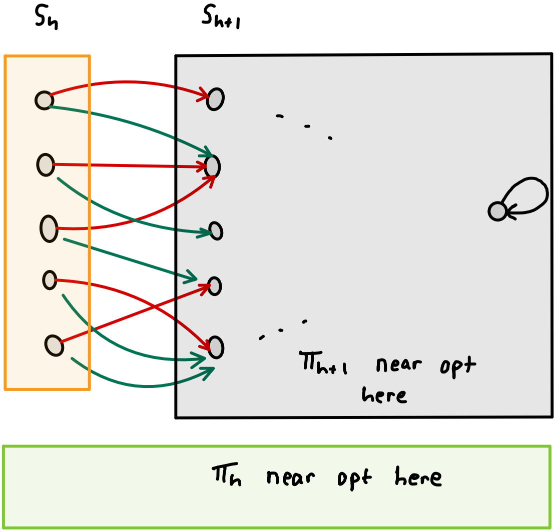
Least-squares 'policy iteration'

1. Estimate $q^{\pi_{h+1}}$ by $\hat{q}^{\pi_{h+1}}$

2. $\pi_h = \pi_{h+1}$ on □

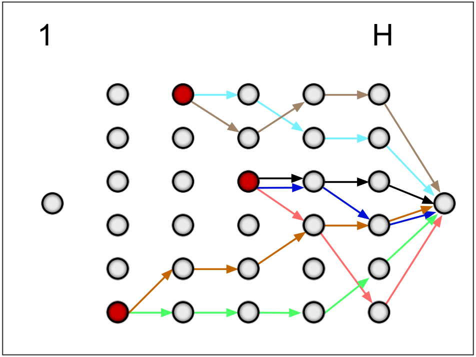
$\pi_h(s) = \arg\max_a \hat{q}^{\pi_{h+1}}(s,a)$

on □



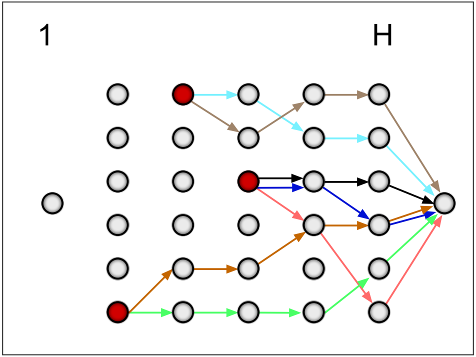
Policy evaluation

- Given a policy π, how to estimate $q^\pi(s, a)$ with a generative model and linear function approximation?
- Find an optimal design ρ on $\{\phi(s, a) : s, a \in \mathcal{S} \times \mathcal{A}\}$
- Sample **rollouts** starting from coreset of ρ in proportion to ρ following policy π to estimate $q^\pi(s, a)$ on the coreset
- Use least squares to generalise to all state/action pairs



Policy evaluation (rollouts)

- Given policy π and state-action pair (s, a) sample a rollout starting in state $s \in \mathcal{S}_h$ and taking action a and subsequently taking actions using π
- Collect cumulative rewards $r_h, \dots, r_H$ and $q = \sum_{u=h}^{H} r_u$
- Then $\mathbb{E}[q] = \mathbb{E}[\sum_{u=h}^{H} r_u] = q^\pi(s, a)$
- $|q| = |\sum_{u=h}^{H} r_u| \leqslant H$



Policy evaluation (extrapolation)

- Perform m rollouts
- Start from state (s, a) in proportion to optimal design ρ
- Collect the data:

$$(s_1, a_1, q_1), \dots, (s_m, a_m, q_m)$$

- Compute least-squares estimate

$$\hat{\theta} = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \frac{1}{2} \sum_{t=1}^{m} (\langle \theta, \phi(s_t, a_t) \rangle - q_t)^2 \qquad \hat{q}^\pi(s, a) = \langle \hat{\theta}, \phi(s, a) \rangle$$

Policy evaluation (extrapolation)

- Perform m rollouts
- Start from state (s, a) in proportion to optimal design ρ
- Collect the data:

$$(s_1, a_1, q_1), \ldots, (s_m, a_m, q_m)$$

- Compute least-squares estimate

$$\hat{\theta} = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \frac{1}{2} \sum_{t=1}^{m} (\langle \theta, \phi(s_t, a_t) \rangle - q_t)^2 \qquad \hat{q}^\pi(s, a) = \langle \hat{\theta}, \phi(s, a) \rangle$$

- With probability at least $1 - \delta$, for all $s, a \in \mathcal{S} \times \mathcal{A}$

$$\begin{aligned} |q^\pi(s, a) - \hat{q}^\pi(s, a)| &= |\langle \phi(s, a), \theta_\star - \hat{\theta} \rangle| \\ &\leqslant \beta \sqrt{\frac{d}{m}} \end{aligned}$$

Policy evaluation (summary)

- Given a policy π and mH queries to the generative model we can find an estimator $\hat{q}^\pi$ of q^π such that

$$\max_{s,a \in \mathcal{S}\times\mathcal{A}} |q^\pi(s,a) - \hat{q}^\pi(s,a)| \leqslant \text{cnst}\, dH \sqrt{\frac{\log(1/\delta)}{m}}$$

- Equivalently, with

$$n \geqslant \frac{\text{cnst}\, d^2 H^3 \log(1/\delta)}{\varepsilon^2}$$

queries to the generative model we have an estimator $\hat{q}^\pi$ of q^π such that

$$\|q^\pi - \hat{q}^\pi\|_\infty \triangleq \max_{s,a \in \mathcal{S}\times\mathcal{A}} |q^\pi(s,a) - \hat{q}^\pi(s,a)| \leqslant \varepsilon$$

Least squares policy iteration

Start with arbitrary policy π_{H+1}
for $h = H$ to 1

- Use policy evaluation and

$$n = \frac{\text{cnst } d^2 H^5 \log(H/\delta)}{\varepsilon^2}$$

queries to find $\|\hat{q}^{\pi_{h+1}} - q^{\pi_{h+1}}\|_\infty \leqslant \varepsilon/H$

- Update policy $\pi_h(s) = \begin{cases} \pi_{h+1}(s) & \text{if } s \notin \mathcal{S}_h \\ \arg\max_{a \in \mathcal{A}} \hat{q}^{\pi_{h+1}}(s, a) & \text{otherwise} \end{cases}$

Theorem 8 With probability at least $1 - \delta$, $v^{\pi_H}(s_1) - v^\star(s_1) \leqslant 2\varepsilon$

Corollary 9 With a generative model and q^π-realisable linear function approximation, sample complexity is at most

$$\frac{\text{cnst } d^2 H^6 \log(H/\delta)}{\varepsilon^2}$$

Analysis

- Same idea as backwards induction
- All policies are optimal on the last layer:

$$v^{\pi_{H+1}}(s) = v^\star(s) \text{ for } s \in \mathcal{S}_{H+1}$$

- We will prove by induction that

$$v^{\pi_h}(s) \geqslant v^*(s) - \frac{2\varepsilon(H+1-h)}{H} \text{ for all } s \in \cup_{u \geqslant h} \mathcal{S}_u$$

Analysis

For $s \in \mathcal{S}_h$

$$\pi_h(s) = \arg\max_{a \in \mathcal{A}} \hat{q}^{h+1}(s, a)$$

Hence

$$q^{\pi_{h+1}}(s, \pi_h(s))$$

Analysis

For $s \in \mathcal{S}_h$

$$\pi_h(s) = \arg\max_{a \in \mathcal{A}} \hat{q}^{h+1}(s, a)$$

Hence

$$q^{\pi_{h+1}}(s, \pi_h(s)) \geqslant \hat{q}^{\pi_{h+1}}(s, \pi_h(s)) - \frac{\varepsilon}{H} \qquad \text{(concentration)}$$

Analysis

For $s \in \mathcal{S}_h$

$$\pi_h(s) = \arg\max_{a \in \mathcal{A}} \hat{q}^{h+1}(s, a)$$

Hence

$$\begin{aligned} q^{\pi_{h+1}}(s, \pi_h(s)) &\geqslant \hat{q}^{\pi_{h+1}}(s, \pi_h(s)) - \frac{\varepsilon}{H} && \text{(concentration)} \\ &= \max_a \hat{q}^{\pi_{h+1}}(s, a) - \frac{\varepsilon}{H} && \text{(def of } \pi_h\text{)} \end{aligned}$$

Analysis

For $s \in \mathcal{S}_h$

$$\pi_h(s) = \arg\max_{a \in \mathcal{A}} \hat{q}^{h+1}(s, a)$$

Hence

$$\begin{aligned}
q^{\pi_{h+1}}(s, \pi_h(s)) &\geqslant \hat{q}^{\pi_{h+1}}(s, \pi_h(s)) - \frac{\varepsilon}{H} && \text{(concentration)} \\
&= \max_a \hat{q}^{\pi_{h+1}}(s, a) - \frac{\varepsilon}{H} && \text{(def of } \pi_h\text{)} \\
&\geqslant \max_a q^{\pi_{h+1}}(s, a) - \frac{2\varepsilon}{H} && \text{(concentration)}
\end{aligned}$$

Analysis

For $s \in \mathcal{S}_h$

$$\pi_h(s) = \arg\max_{a \in \mathcal{A}} \hat{q}^{h+1}(s, a)$$

Hence

$$\begin{aligned}
q^{\pi_{h+1}}(s, \pi_h(s)) &\geqslant \hat{q}^{\pi_{h+1}}(s, \pi_h(s)) - \frac{\varepsilon}{H} && \text{(concentration)} \\
&= \max_a \hat{q}^{\pi_{h+1}}(s, a) - \frac{\varepsilon}{H} && \text{(def of } \pi_h\text{)} \\
&\geqslant \max_a q^{\pi_{h+1}}(s, a) - \frac{2\varepsilon}{H} && \text{(concentration)} \\
&= \max_a r(s, a) + \sum_{s' \in \mathcal{S}_{h+1}} \mathcal{P}(s'|s, a) v^{\pi_{h+1}}(s') - \frac{2\varepsilon}{H}
\end{aligned}$$

Analysis

For $s \in \mathcal{S}_h$

$$\pi_h(s) = \arg\max_{a \in \mathcal{A}} \hat{q}^{h+1}(s, a)$$

Hence

$$\begin{aligned}
q^{\pi_{h+1}}(s, \pi_h(s)) &\geqslant \hat{q}^{\pi_{h+1}}(s, \pi_h(s)) - \frac{\varepsilon}{H} && \text{(concentration)} \\
&= \max_a \hat{q}^{\pi_{h+1}}(s, a) - \frac{\varepsilon}{H} && \text{(def of } \pi_h\text{)} \\
&\geqslant \max_a q^{\pi_{h+1}}(s, a) - \frac{2\varepsilon}{H} && \text{(concentration)} \\
&= \max_a r(s, a) + \sum_{s' \in \mathcal{S}_{h+1}} \mathcal{P}(s'|s, a) v^{\pi_{h+1}}(s') - \frac{2\varepsilon}{H} \\
&\geqslant \max_a r(s, a) + \sum_{s' \in \mathcal{S}_{h+1}} \mathcal{P}(s'|s, a) v^*(s') - \frac{2\varepsilon}{H} - \frac{2\varepsilon(H - h)}{H} \\
&= v^*(s) - \frac{2\varepsilon(H + 1 - h)}{H}
\end{aligned}$$

Analysis

For $s \in \mathcal{S}_h$

$$\pi_h(s) = \arg\max_{a \in \mathcal{A}} \hat{q}^{h+1}(s, a)$$

Hence

$$\begin{aligned}
q^{\pi_{h+1}}(s, \pi_h(s)) &\geqslant \hat{q}^{\pi_{h+1}}(s, \pi_h(s)) - \frac{\varepsilon}{H} && \text{(concentration)} \\
&= \max_a \hat{q}^{\pi_{h+1}}(s, a) - \frac{\varepsilon}{H} && \text{(def of } \pi_h\text{)} \\
&\geqslant \max_a q^{\pi_{h+1}}(s, a) - \frac{2\varepsilon}{H} && \text{(concentration)} \\
&= \max_a r(s, a) + \sum_{s' \in \mathcal{S}_{h+1}} \mathcal{P}(s'|s, a) v^{\pi_{h+1}}(s') - \frac{2\varepsilon}{H} \\
&\geqslant \max_a r(s, a) + \sum_{s' \in \mathcal{S}_{h+1}} \mathcal{P}(s'|s, a) v^*(s') - \frac{2\varepsilon}{H} - \frac{2\varepsilon(H - h)}{H} \\
&= v^*(s) - \frac{2\varepsilon(H + 1 - h)}{H}
\end{aligned}$$

What happens if the q-values are only *nearly* linear

Assumption 2 For all π there exists a θ such that

$$|q^\pi(s,a) - \langle \phi(s,a), \theta \rangle| \leqslant \rho \text{ for all } s, a$$

Estimating $q^\pi(s,a)$ using least squares leads to

$$|\hat{q}^\pi(s,a) - q^\pi(s,a)| \leqslant \text{cnst}\, dH \sqrt{\frac{\log(1/\delta)}{\#\text{rollouts}}} + \rho\sqrt{d}$$

Repeating the analysis before

$$v^\pi(s) \geqslant v^*(s) - \text{cnst}\, dH^3 \sqrt{\frac{\log(1/\delta)}{\#\text{queries}}} - 2\rho H \sqrt{d}$$

Lower bound huge price for beating the $\rho H \sqrt{d}$ barrier

Johnson-Lindenstraus Lemma

Lemma 10 There exists a set $\mathcal{A} \subset \mathbb{R}^d$ of size k such that

1. $\|a\| = 1$ for all $a \in \mathcal{A}$
2. $|\langle a, b \rangle| \leqslant \sqrt{8 \log(k)/(d-1)} \triangleq \gamma$ for all $a, b \in \mathcal{A}$ with $a \neq b$

Johnson-Lindenstraus Lemma

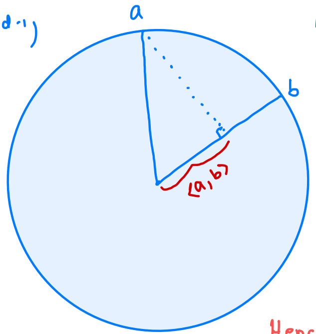
Lemma 10 There exists a set $\mathcal{A} \subset \mathbb{R}^d$ of size k such that

1. $\|a\| = 1$ for all $a \in \mathcal{A}$
2. $|\langle a, b\rangle| \leqslant \sqrt{8\log(k)/(d-1)} \triangleq \gamma$ for all $a, b \in \mathcal{A}$ with $a \neq b$

$a, b \sim U(S^{d-1})$

$\mathbb{E}[\langle a,b\rangle^2]$

$= \mathbb{E}[b_1^2]$



$1 = \mathbb{E}[\|b\|^2]$

$= \sum_i \mathbb{E}[b_i^2]$

$= d\,\mathbb{E}[b_1^2]$

Hence $\mathbb{E}[\langle a,b\rangle^2] = \frac{1}{d}$

$\mathbb{E}[|\langle a,b\rangle|] \approx \sqrt{1/d}$

Johnson-Lindenstraus Lemma

Lemma 11 There exists a set $\mathcal{A} \subset \mathbb{R}^d$ of size k such that

1. $\|a\| = 1$ for all $a \in \mathcal{A}$
2. $|\langle a, b \rangle| \leqslant \sqrt{8 \log(k)/(d-1)} \triangleq \gamma$ for all $a, b \in \mathcal{A}$ with $a \neq b$

Johnson-Lindenstraus Lemma

Lemma 11 There exists a set $\mathcal{A} \subset \mathbb{R}^d$ of size k such that

1. $\|a\| = 1$ for all $a \in \mathcal{A}$
2. $|\langle a, b \rangle| \leqslant \sqrt{8 \log(k)/(d-1)} \triangleq \gamma$ for all $a, b \in \mathcal{A}$ with $a \neq b$

Proof of lower bound

- Construct a needle-in-a-haystack
- $H = 1$ and $\mathcal{A}$ is from Lemma 11
- $\phi(s_1, a) = a$ for all a
- Let $a^\star \in \mathcal{A}$ and

$$q(s_1, a) = r(s_1, a) = \begin{cases} \varepsilon/\gamma & \text{if } a = a^\star \\ 0 & \text{otherwise} \end{cases}$$

- Sample complexity to find ε/γ-optimal action is at least k

Johnson-Lindenstraus Lemma

Lemma 11 There exists a set $\mathcal{A} \subset \mathbb{R}^d$ of size k such that

1. $\|a\| = 1$ for all $a \in \mathcal{A}$
2. $|\langle a, b\rangle| \leqslant \sqrt{8\log(k)/(d-1)} \triangleq \gamma$ for all $a, b \in \mathcal{A}$ with $a \neq b$

Proof of lower bound

- Construct a needle-in-a-haystack
- $H = 1$ and $\mathcal{A}$ is from Lemma 11
- $\phi(s_1, a) = a$ for all a
- Let $a^\star \in \mathcal{A}$ and

$$q(s_1, a) = r(s_1, a) = \begin{cases} \varepsilon/\gamma & \text{if } a = a^\star \\ 0 & \text{otherwise} \end{cases}$$

- Sample complexity to find ε/γ-optimal action is at least k
- q is ε-close to linear with $\theta_\star = \varepsilon/\gamma a^\star$
- $\langle a^\star, \varepsilon/\gamma a^\star\rangle = \varepsilon/\gamma$ and $|\langle a, \varepsilon/\gamma a^\star\rangle| \leqslant \varepsilon$

Before we assumed that q-values are nearly linear

Alternative

Assumption 3 There is a given function $\phi : \mathcal{S} \to \mathbb{R}^d$ such that for π there exists a θ such that

$$v^\pi(s) = \langle \phi(s), \theta \rangle$$

Exercise 9 What sample complexity can you achieve under Assumption 3

Local planning

- Using the generative model is simple and **statistically efficient**
- **Computationally hopeless** when $\mathcal{S}$ is big
- Finding the optimal design and extending using least-squares is impossible
- If you only care about local planning then more sophisticated algorithms can find a policy π such that

$$v^{\pi}(s_1) \geqslant v^{\star}(s_1) - \varepsilon$$

  with polynomial sample complexity [Hao et al., 2022]

High level idea

Explore using approximately optimal design on set of observed states so far

Add states to the optimal design as necessary

Online setting

Not known if polynomial sample complexity is possible with only linear q^π functions

Online setting

Not known if polynomial sample complexity is possible with only linear q^π functions

Linear MDPs

$(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$ with feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is linear if

- There exists a θ such that $r(s, a) = \langle \phi(s, a), \theta \rangle$
- There exists a signed measure $\mu : \mathcal{S} \to \mathbb{R}^d$ such that

$$\mathcal{P}(s'|s, a) = \langle \phi(s, a), \mu(s') \rangle$$

Online setting

Not known if polynomial sample complexity is possible with only linear q^π functions

Linear MDPs

$(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$ with feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is linear if

- There exists a θ such that $r(s, a) = \langle \phi(s, a), \theta \rangle$
- There exists a signed measure $\mu : \mathcal{S} \to \mathbb{R}^d$ such that

$$\mathcal{P}(s'|s, a) = \langle \phi(s, a), \mu(s') \rangle$$

Learning μ is hopeless

Why are linear MDPs learnable?

$$\mathcal{P}(s'|s,a) = \langle \phi(s,a), \mu(s') \rangle$$

Key point Never need to learn μ

All our algorithms need to learn is the Bellman operator

$$r(s,a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) v(s') = \langle \phi(s,a), \theta \rangle + \phi(s,a)^\top \sum_{s' \in \mathcal{S}} \mu(s') v(s')$$

Right-hand side only depends on (s,a) via $\phi(s,a)$

Estimating expectations

Some algorithm interacts with the MDP (online model) collecting data

$$\mathcal{D} = (s_u, a_u, r_u, s'_u)_{u=1}^m$$

Let $v : \mathcal{S} \to [0, H]$

Want to estimate from data

$$r(s,a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) f(s') = \langle \phi(s,a), \theta \rangle + \phi(s,a)^\top \sum_{s' \in \mathcal{S}} \mu(s') v(s')$$

Care about value of LHS for *all* (s, a)

Estimating expectations

$$\mathcal{D} = (s_u, a_u, r_u, s'_u)_{u=1}^m$$

$$r(s,a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) v(s') = \phi(s,a)^\top \theta + \phi(s,a)^\top \sum_{s' \in \mathcal{S}} \mu(s') v(s')$$
$$\triangleq \langle \phi(s,a), w_v \rangle$$

Estimate with least squares

$$\hat{w}_v = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \sum_{u=1}^m (\langle \phi(s_u, a_u), w \rangle - r_u - v(s'_u))^2$$

Makes sense because

$$\mathbb{E}[r_u + v(s'_u)] = r(s_u, a_u) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s_u, a_u) v(s') = \langle \phi(s_u, a_u), w_v \rangle$$

Estimating expectations

$$G = \sum_{u=1}^{m} \phi(s_u, a_u)\phi(s_u, a_u)^\top$$

$$\begin{aligned}\hat{w}_v &= \arg\min_{w \in \mathbb{R}^d} \sum_{u=1}^{m} \left(\langle \phi(s_u, a_u), w \rangle - r_u - v(s'_u)\right)^2 \\ &= G^{-1} \sum_{u=1}^{m} \phi(s_u, a_u)[r_u + v(s'_u)]\end{aligned}$$

(almost) Usual story in terms of the error

$$|\langle \phi(s, a), \hat{w}_v - w_v \rangle| \overset{whp}{\lesssim} H \|\phi(s, a)\|_{G^{-1}} \sqrt{d + \log(1/\delta)}$$

UCB-VI for Linear MDPs [Jin et al., 2020]

- Use data to construct optimistic q-values by backwards induction
- $\tilde{q}_{H+1}(s, a) = 0$ and for $h = H$ to 1

UCB-VI for Linear MDPs [Jin et al., 2020]

- Use data to construct optimistic q-values by backwards induction
- $\tilde{q}_{H+1}(s, a) = 0$ and for $h = H$ to 1

$$\tilde{q}_h(s, a) = \phi(s, a)^\top G^{-1} \sum_{u=1}^{m} \phi(s_u, a_u) \left[r_u + \tilde{v}_{h+1}(s'_u)\right] + \underbrace{\beta \|\phi(s, a)\|_{G^{-1}}}_{\text{bonus}}$$

$\tilde{v}_{h+1}(s) = \max_a \tilde{q}_{h+1}(s, a)$

UCB-VI for Linear MDPs [Jin et al., 2020]

- Use data to construct optimistic q-values by backwards induction
- $\tilde{q}_{H+1}(s, a) = 0$ and for $h = H$ to 1

$$\tilde{q}_h(s, a) = \phi(s, a)^\top G^{-1} \sum_{u=1}^{m} \phi(s_u, a_u)\left[r_u + \tilde{v}_{h+1}(s'_u)\right] + \underbrace{\beta \|\phi(s, a)\|_{G^{-1}}}_{\text{bonus}}$$

$$\tilde{v}_{h+1}(s) = \max_a \tilde{q}_{h+1}(s, a)$$

- Act greedily: for $h = 1$ to H

$$a_h = \operatorname*{arg\,max}_{a \in \mathcal{A}} \tilde{q}_h(s_h, a) \text{ and observe } s_{h+1}$$

Need to find large enough β that the algorithm is optimistic

$$\tilde{q}_h(s,a) = \phi(s,a)^\top G^{-1} \sum_{u=1}^{m} \phi(s_u,a_u)\left[r_u + \tilde{v}_{h+1}(s'_u)\right] + \underbrace{\beta \|\phi(s,a)\|_{G^{-1}}}_{\text{bonus}}$$

$$\tilde{v}_{h+1}(s) = \max_a \tilde{q}_{h+1}(s,a)$$

Caveat $\tilde{v}_{h+1}$ is not independent of the data

Where did we get?

- We have data from m interactions
- We can use it estimate $(s, a) \mapsto r(s, a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, a)v(s')$
- With probability $1 - \delta$,

$$|\langle \phi(s, a), \hat{w}_v - w_v \rangle| \lesssim H\|\phi(s, a)\|_{G^{-1}} \sqrt{d + \log(1/\delta)}$$

Where did we get?

- We have data from m interactions
- We can use it estimate $(s, a) \mapsto r(s, a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, a)v(s')$
- With probability $1 - \delta$,

$$|\langle \phi(s, a), \hat{w}_v - w_v \rangle| \lesssim H\|\phi(s, a)\|_{G^{-1}} \sqrt{d + \log(1/\delta)}$$

- We want this to hold for all possible $\tilde{v}_h$ functions

$$\tilde{v}_h \in \left\{ s \mapsto \max_a \langle \phi(s, a), w \rangle + \sqrt{\phi(s, a)^\top W \phi(s, a)} : w \in \mathbb{R}^d, W \in \mathbb{R}^{d \times d} \right\}$$

- How many functions are there of this form?

Where did we get?

- We have data from m interactions
- We can use it estimate $(s, a) \mapsto r(s, a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, a)v(s')$
- With probability $1 - \delta$,

$$|\langle \phi(s, a), \hat{w}_v - w_v \rangle| \lesssim H\|\phi(s, a)\|_{G^{-1}} \sqrt{d + \log(1/\delta)}$$

- We want this to hold for all possible $\tilde{v}_h$ functions

$$\tilde{v}_h \in \left\{ s \mapsto \max_a \langle \phi(s, a), w \rangle + \sqrt{\phi(s, a)^\top W \phi(s, a)} : w \in \mathbb{R}^d, W \in \mathbb{R}^{d \times d} \right\}$$

- How many functions are there of this form? ∞

Covering numbers and union bound

$$\mathcal{V} = \left\{ s \mapsto \max_a \langle \phi(s,a), w \rangle + \sqrt{\phi(s,a)^\top W \phi(s,a)} : w \in \mathbb{R}^d, W \in \mathbb{R}^{d \times d} \right\}$$

Covering number argument. Effective number of functions of this form is

$$N = \left(\frac{1}{\varepsilon}\right)^{d^2}$$

By a union bound, with probability at least $1 - \delta$ for all $v \in \mathcal{V}$

$$\langle \phi(s,a), \hat{w}_v - w_v \rangle \lesssim \|\phi(s,a)\|_{G^{-1}} H \sqrt{d + \log(N/\delta)} \triangleq \beta \|\phi(s,a)\|_{G^{-1}}$$

Prove by induction that $\tilde{q}_h(s,a) \geqslant q^\star(s,a)$

Start with $\tilde{q}_{H+1}(s,a) = 0$ $\qquad$ $(\tilde{v}_h(s) = \max_a \tilde{q}_h(s,a))$

$$
\begin{aligned}
\tilde{q}_h(s,a) &= \phi(s,a)^\top G^{-1} \sum_{u=1}^{m} \phi(s_u,a_u)\left[r_u + \tilde{v}_{h+1}(s'_u)\right] + \beta\|\phi(s,a)\|_{G^{-1}} \\
&= \phi(s,a)^\top \hat{w}_{\tilde{v}_{h+1}} + \beta\|\phi(s,a)\|_{G^{-1}} \\
&\geqslant \phi(s,a)^\top w_{\tilde{v}_{h+1}} \\
&= r(s,a) + \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a)\tilde{v}_{h+1}(s') \\
&\geqslant r(s,a) + \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a)v^\star(s') = q^\star(s,a)
\end{aligned}
$$

Bellman operator on q-values

- Let $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Abbreviate $v(s) = \max_{a \in \mathcal{A}} q(s, a)$
- Define $\mathcal{T} : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \to \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ by

$$(\mathcal{T} q)(s, a) = r(s, a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, a) v(s')$$

- Note: $\mathcal{T}$ depends on the dynamics/rewards of the (unknown) MDP
- We write π_q for the **greedy policy** with respect to q

$$\pi_q(s) = \arg\max_{a \in \mathcal{A}} q(s, a)$$

Policy loss decomposition

Proposition 3 Let $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $v(s) = \max_a q(s, a)$ and $\pi = \pi_q$. Then

$$v(s_1) - v^\pi(s_1) = \mathbb{E}_\pi \left[\sum_{h=1}^{H} (q - \mathcal{T} q)(s_h, a_h) \right]$$

Proof

$$\begin{aligned}
& q(s_1, \pi(s_1)) - v^\pi(s_1) \\
& = (q - \mathcal{T} q)(s_1, \pi(s_1)) + (\mathcal{T} q)(s_1, \pi(s_1)) - q^\pi(s_1, \pi(s_1)) \\
& = (q - \mathcal{T} q)(s_1, \pi(s_1)) + \mathcal{T}(q - q^\pi)(s_1, \pi(s_1)) \\
& = (q - \mathcal{T} q)(s_1, \pi(s_1)) + \sum_{s_2} P(s_2 | s_1, \pi(s_1))(q(s_2, \pi(s_2)) - q^\pi(s_2, \pi(s_2))) \\
& = \cdots \\
& = \mathbb{E}_\pi \left[\sum_{h=1}^{H} (q - \mathcal{T} q)(s_h, \pi(s_h)) \right]
\end{aligned}$$

$\square$

From optimism to regret

$$
\begin{aligned}
\mathrm{Reg}_n &= \mathbb{E}\left[\sum_{t=1}^{n} v^\star(s_1) - v^{\pi_t}(s_1)\right] \\
&\lesssim \mathbb{E}\left[\sum_{t=1}^{n} \tilde{v}^t(s_1) - v^{\pi_t}(s_1)\right] && \text{(Optimism)} \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \sum_{h=1}^{H} (\tilde{q}^t - \mathcal{T}\tilde{q}^t)(s_h^t, a_h^t)\right] && \text{(Prop 3)}
\end{aligned}
$$

Bellman error

Remember

$$\tilde{q}_t(s,a) = \phi(s,a)^\top G_{t-1}^{-1} \sum_{u=1}^{m} \phi(s_u,a_u)[r_u + \tilde{v}_t(s'_u)] + \beta \|\phi(s,a)\|_{G_{t-1}^{-1}}$$

$$\mathcal{T}\tilde{q}_t(s,a) = r(s,a) + \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a)\tilde{v}_t(s')$$

Concentration

$$\tilde{q}_t(s,a) - \mathcal{T}\tilde{q}_t(s,a) \leqslant 2\beta\|\phi(s,a)\|_{G_{t-1}^{-1}}$$

Back to the regret

$$
\begin{aligned}
\mathrm{Reg}_n &\leqslant \mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H}(\tilde{q}^{\,t} - \mathcal{T}\tilde{q}^{\,t})(s_h^t, a_h^t)\right] \\
&\leqslant 2\beta\mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|_{G_{t-1}^{-1}}\right] \\
&\leqslant 2\beta\sqrt{nH\mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|_{G_{t-1}^{-1}}^2\right]}
\end{aligned}
$$

Back to the regret

$$
\begin{aligned}
\mathrm{Reg}_n &\leqslant \mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H}(\tilde{q}^t - \mathcal{T}\tilde{q}^t)(s_h^t, a_h^t)\right] \\
&\leqslant 2\beta\mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|_{G_{t-1}^{-1}}\right] \\
&\leqslant 2\beta\sqrt{nH\mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|^2_{G_{t-1}^{-1}}\right]}
\end{aligned}
$$

Naive application of elliptical potential lemma

$$
\sum_{t=1}^{n}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|^2_{G_{t-1}^{-1}} \lesssim dH\log(n)
$$

Back to the regret

$$
\begin{aligned}
\mathrm{Reg}_n &\leqslant \mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H}(\tilde{q}^t - \mathcal{T}\tilde{q}^t)(s_h^t, a_h^t)\right] \\
&\leqslant 2\beta\mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|_{G_{t-1}^{-1}}\right] \\
&\leqslant 2\beta\sqrt{nH\mathbb{E}\left[\sum_{t=1}^{n}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|_{G_{t-1}^{-1}}^2\right]} \lesssim \sqrt{d^3H^4 n\log(n)}
\end{aligned}
$$

Naive application of elliptical potential lemma

$$
\sum_{t=1}^{n}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|_{G_{t-1}^{-1}}^2 \lesssim dH\log(n)
$$

Misspecification

- Same algorithm is robust to misspecification

$$\|\mathcal{P}(\cdot|s,a) - \langle \phi(s,a), \mu(\cdot)\rangle\|_{TV} \leqslant \varepsilon |r(s,a) - \langle \phi(s,a), \theta\rangle| \quad \leqslant \varepsilon$$

- Additive $\tilde{O}(\varepsilon n d H)$ term in the regret

Beyond linearity

Slides available at

https://tor-lattimore.com/downloads/RLTheory.pdf

Nonlinear function approximation

- Previous we assumed that for all π there exists a θ such that

$$q^\pi(s, a) = \langle \phi(s, a), \theta \rangle$$

- Equivalently, for all π, $q^\pi \in \{(s, a) \mapsto \langle \phi(s, a), \theta \rangle : \theta \in \mathbb{R}^d\}$
- Sample complexity depends on d
- **Alternative** Assume $q^\pi \in \mathcal{F}$ for some abstract function class $\mathcal{F}$
- Somehow bound sample complexity in terms of the structure of $\mathcal{F}$

Complete charactersiation of sample complexity for binary classification

$$\mathrm{SampleComplexity}(\varepsilon) = \Theta\left(\frac{\mathrm{VC}(\mathcal{H}) + \log(1/\delta)}{\varepsilon}\right)$$

Wow! Good job. What's the RL version?

Remember bandits

- Learner takes actions $a_1, \ldots, a_n$ in $\mathcal{A}$
- Observes rewards $r_1, \ldots, r_n$ with $r_t = f(a_t) + \eta_t$ for some $f : \mathcal{A} \to \mathbb{R}$
- Assume $f \in \mathcal{F}$ for some known function class $\mathcal{F}$
- Generative model, local planning and fully online are all the same for bandits

Remember bandits

- Learner takes actions $a_1, \dots, a_n$ in $\mathcal{A}$
- Observes rewards $r_1, \dots, r_n$ with $r_t = f(a_t) + \eta_t$ for some $f : \mathcal{A} \to \mathbb{R}$
- Assume $f \in \mathcal{F}$ for some known function class $\mathcal{F}$
- Generative model, local planning and fully online are all the same for bandits

Can we get a regret bound that depends on $\mathcal{F}$?

$$\mathrm{Reg}_n = \max_{a^\star \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^n f(a^\star) - f(a_t)\right]$$

Eluder dimension (intuition)

- We are going to play optimistically
- Somehow construct confidence set $\mathcal{F}_t$ based on data collected
- Optimistic value function is $f_t = \arg\max_{f \in \mathcal{F}_{t-1}} \max_{a \in \mathcal{A}} f(a)$
- Play $a_t = \arg\max_{a \in \mathcal{A}} f_t(a)$

Eluder dimension (intuition)

- We are going to play optimistically
- Somehow construct confidence set $\mathcal{F}_t$ based on data collected
- Optimistic value function is $f_t = \arg\max_{f \in \mathcal{F}_{t-1}} \max_{a \in \mathcal{A}} f(a)$
- Play $a_t = \arg\max_{a \in \mathcal{A}} f_t(a)$
- Regret

$$\begin{aligned}
\mathrm{Reg}_n &= \mathbb{E}\left[\sum_{t=1}^n f(a^\star) - f(a_t)\right] && \text{(regret def)} \\
&\lesssim \mathbb{E}\left[\sum_{t=1}^n (f_t(a_t) - f(a_t))\right] && \text{(optimism principle)}
\end{aligned}$$

Eluder dimension (intuition)

- We are going to play optimistically
- Somehow construct confidence set $\mathcal{F}_t$ based on data collected
- Optimistic value function is $f_t = \arg\max_{f \in \mathcal{F}_{t-1}} \max_{a \in \mathcal{A}} f(a)$
- Play $a_t = \arg\max_{a \in \mathcal{A}} f_t(a)$
- Regret

$$
\begin{aligned}
\mathrm{Reg}_n &= \mathbb{E}\left[\sum_{t=1}^n f(a^\star) - f(a_t)\right] && \text{(regret def)} \\
&\lesssim \mathbb{E}\left[\sum_{t=1}^n (f_t(a_t) - f(a_t))\right] && \text{(optimism principle)}
\end{aligned}
$$

Eluder dimension

measures how often $f_t(a_t) - f(a_t)$ can be large

Confidence bounds for LSE

Least squares estimator of $f_\star$ after t rounds is

$$\hat{f}_t = \operatorname*{arg\,min}_{f \in \mathcal{F}} \sum_{s=1}^{t} (r_s - f(a_s))^2$$

Lemma 12 There exists a constant $\beta^2 \lesssim \log(|\mathcal{F}|n)$ such that

$$\mathbb{P}\left(\max_{0 \leqslant t \leqslant n} \|\hat{f}_t - f_\star\|_t^2 \geqslant \beta^2\right) \leqslant \frac{1}{n} \qquad \|f - g\|_t^2 = \sum_{s=1}^{t} (f(a_s) - g(a_s))^2$$

Define confidence set $\mathcal{F}_t = \{f \in \mathcal{F}_{t-1} : \|\hat{f}_t - f\|_t^2 \leqslant \beta^2\}$

By Lemma 12, $\mathbb{P}(\exists t \in [n] : f_\star \notin \mathcal{F}_{t-1}) \leqslant 1/n$

Confidence bounds

LSE minimises the sum of squared errors by definition

$$\hat{f}_t = \operatorname*{arg\,min}_{f \in \mathcal{F}} \sum_{s=1}^{t} (r_s - f(a_s))^2$$

Rearrange some things

$$\begin{aligned}
0 &\geqslant \sum_{s=1}^{t} (r_s - f_t(a_s))^2 - \sum_{s=1}^{t} (r_s - f_\star(a_s))^2 \\
&= \sum_{s=1}^{t} (f_\star(a_s) + \eta_s - f_t(a_s))^2 - \sum_{s=1}^{t} \eta_s^2 \\
&= \underbrace{\sum_{s=1}^{t} (f_\star(a_s) - f_t(a_s))^2}_{\text{want this small}} + \underbrace{2\sum_{s=1}^{t} \eta_s (f_\star(a_s) - f_t(a_s))}_{\text{noise}}
\end{aligned}$$

CLT things

Last slide

$$\sum_{s=1}^{t}(f_\star(a_s)-f_t(a_s))^2 \leqslant \underbrace{2\sum_{s=1}^{t}\eta_s(f_\star(a_s)-f_t(a_s))}_{\text{sum of zero-mean random variables}}$$

Given fixed $f \in \mathcal{F}$. Using $\mathbb{V}[aX] = a^2\mathbb{V}[X]$

$$2\sum_{s=1}^{t}\mathbb{V}_{s-1}[\eta_s(f_\star(a_s)-f(a_s))] = 2\sum_{s=1}^{t}(f_\star(a_s)-f(a_s))^2$$

Martingale CLT

$$2\sum_{s=1}^{t}\eta_s(f_\star(a_s)-f(a_s)) \overset{whp}{\lesssim} \sqrt{\sum_{s=1}^{t}(f_\star(a_s)-f_t(a_s))^2\log(1/\delta)}$$

CLT things

Last slide

$$\sum_{s=1}^{t}(f_\star(a_s) - f_t(a_s))^2 \leqslant \underbrace{2\sum_{s=1}^{t}\eta_s(f_\star(a_s) - f_t(a_s))}_{\text{sum of zero-mean random variables}}$$

Given fixed $f \in \mathcal{F}$. Using $\mathbb{V}[aX] = a^2\mathbb{V}[X]$

$$2\sum_{s=1}^{t}\mathbb{V}_{s-1}[\eta_s(f_\star(a_s) - f(a_s))] = 2\sum_{s=1}^{t}(f_\star(a_s) - f(a_s))^2$$

Martingale CLT for all $f \in \mathcal{F}$

$$2\sum_{s=1}^{t}\eta_s(f_\star(a_s) - f(a_s)) \overset{whp}{\lesssim} \sqrt{\sum_{s=1}^{t}(f_\star(a_s) - f_t(a_s))^2 \log(|\mathcal{F}|/\delta)}$$

CLT things

Last slide

$$\sum_{s=1}^{t}(f_\star(a_s) - f_t(a_s))^2 \leqslant \underbrace{2\sum_{s=1}^{t}\eta_s(f_\star(a_s) - f_t(a_s))}_{\text{sum of zero-mean random variables}}$$

Given fixed $f \in \mathcal{F}$. Using $\mathbb{V}[aX] = a^2\mathbb{V}[X]$

$$2\sum_{s=1}^{t}\mathbb{V}_{s-1}[\eta_s(f_\star(a_s) - f(a_s))] = 2\sum_{s=1}^{t}(f_\star(a_s) - f(a_s))^2$$

Martingale CLT for LSE $f_t \in \mathcal{F}$

$$2\sum_{s=1}^{t}\eta_s(f_\star(a_s) - f_t(a_s)) \overset{whp}{\lesssim} \sqrt{\sum_{s=1}^{t}(f_\star(a_s) - f_t(a_s))^2 \log(|\mathcal{F}|/\delta)}$$

Confidence bound

$$\sum_{s=1}^{t}(f_\star(a_s) - f_t(a_s))^2 \leqslant \underbrace{2\sum_{s=1}^{t}\eta_s(f_\star(a_s) - f_t(a_s))}_{\text{sum of zero-mean random variables}}$$

$$\overset{whp}{\lesssim} \sqrt{\sum_{s=1}^{t}(f_\star(a_s) - f_t(a_s))^2 \log(|\mathcal{F}|/\delta)}$$

Confidence bound

$$\sum_{s=1}^{t}(f_\star(a_s) - f_t(a_s))^2 \leqslant \underbrace{2\sum_{s=1}^{t}\eta_s(f_\star(a_s) - f_t(a_s))}_{\text{sum of zero-mean random variables}}$$

$$\overset{whp}{\lesssim} \sqrt{\sum_{s=1}^{t}(f_\star(a_s) - f_t(a_s))^2 \log(|\mathcal{F}|/\delta)}$$

Rearranging

$$\|f_\star - f_t\|_t^2 \triangleq \sum_{s=1}^{t}(f_\star(a_s) - f_t(a_s))^2 \overset{whp}{\lesssim} \log(|\mathcal{F}|/\delta)$$

Eluder dimension

- Let $\mathcal{F}$ be a set of functions from $\mathcal{A}$ to $\mathbb{R}$
- Eluder dimension is a complexity measure of $\mathcal{F}$
- Given an $\varepsilon > 0$ and sequence $a_1, \ldots, a_n$ in $\mathcal{A}$, we say that $a \in \mathcal{A}$ is ε-dependent with respect to $(a_t)_{t=1}^n$ if

$$\forall f, g \in \mathcal{F} \text{ with } \sum_{t=1}^{n} (f(a_t) - g(a_t))^2 \leqslant \varepsilon^2, \;\; f(a) - g(a) \leqslant \varepsilon$$

- a is ε-independent with respect to $(a_t)_{t=1}^n$ if it is not ε-dependent

Definition 13 (Russo and Van Roy 2013) The Eluder dimension $\dim E(\mathcal{F}, \varepsilon)$ of $\mathcal{F}$ at level $\varepsilon > 0$ is the largest d such that there exists a sequence $(a_t)_{t=1}^d$ of ε-independent elements

Theorem 14 Let $(a_t)_{t=1}^n$ be a sequence in $\mathcal{A}$ and $(f_t)_{t=1}^n$ a sequence in $\mathcal{F}$ and

$$\mathcal{F}_t = \mathcal{F}_{t-1} \cap \{f \in \mathcal{F} : \|f - f_t\|_t^2 \leqslant \beta^2\} \qquad \underbrace{w_t(a) = \max_{f,g \in \mathcal{F}_t} f(a) - g(a)}_{\text{width of } \mathcal{F}_t \text{ wrt } a}$$

Then $\#\{t : w_{t-1}(a_t) > \varepsilon\} \leqslant \text{cnst}\, \beta^2 \,\text{dimE}(\mathcal{F}, \varepsilon)/\varepsilon^2$

Theorem 14 Let $(a_t)_{t=1}^n$ be a sequence in $\mathcal{A}$ and $(f_t)_{t=1}^n$ a sequence in $\mathcal{F}$ and

$$\mathcal{F}_t = \mathcal{F}_{t-1} \cap \{f \in \mathcal{F} : \|f - f_t\|_t^2 \leqslant \beta^2\} \qquad \underbrace{w_t(a) = \max_{f,g \in \mathcal{F}_t} f(a) - g(a)}_{\text{width of } \mathcal{F}_t \text{ wrt } a}$$

Then $\#\{t : w_{t-1}(a_t) > \varepsilon\} \leqslant \text{cnst}\, \beta^2 \dim E(\mathcal{F}, \varepsilon)/\varepsilon^2$

Proof

In round t Case 1: $w_t(a_t) \leq \varepsilon$: do nothing

Case 2: $w_t(a_t) > \varepsilon$:

$\mathcal{B}_1$ $\boxed{a_1 \; a_4}$

① $\exists\, f, g \in \mathcal{F}_{t-1}$ s.t $f(a_t) - g(a_t) > \varepsilon$

$\mathcal{B}_2$ $\boxed{a_2 \; a_5 \; a_6}$

② $\|f - g\|_t^2 \leq 2\|f - \hat{f}_t\|_t^2 + 2\|g - \hat{f}_t\|_t^2$

$\leq 4\beta^2$

$\Rightarrow \exists\, B$ s.t. $\sum_{a \in B} (f(a) - g(a))^2 \leq \varepsilon^2$

$\vdots$

$\mathcal{B}_m$ $\boxed{a_3 \; a_7}$

③ Add a_t to B

$m = \lceil 4\beta^2/\varepsilon^2 \rceil$ buckets

④ a_t is ε-ind. of elements in B. Hence at most Edim items added

Corollary 15 Let $(a_t)_{t=1}^n$ be a sequence in $\mathcal{A}$ and $(f_t)_{t=1}^n$ be a sequence in $\mathcal{F}$

$$\mathcal{F}_t = \mathcal{F}_{t-1} \cap \{f \in \mathcal{F} : \|f - f_t\|_t^2 \leqslant \beta^2\}$$

Then with $w_t(a) = \max_{f,g \in \mathcal{F}_t} f(a) - g(a)$

$$\sum_{t=1}^n w_{t-1}(a_t) \leqslant n\varepsilon + \text{cnst}\sqrt{n\beta^2 \operatorname{dimE}(\mathcal{F}, \varepsilon)}$$

Eluder dimension for bandits

- Let $\mathcal{F}$ be a set of functions from $\mathcal{A}$ to $\mathbb{R}$ and $f \in \mathcal{F}$ is unknown
- Learner plays actions $(a_t)_{t=1}^n$ observing rewards $(r_t)_{t=1}^n$

$$r_t = f(a_t) + \eta_t$$

- Regret is $\mathrm{Reg}_n = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^n f(a) - f(a_t)\right]$
- Given a confidence set $\mathcal{F}_t$ after round t, the algorithm plays

$$a_t = \operatorname*{arg\,max}_{a \in \mathcal{A}} \max_{g \in \mathcal{F}_{t-1}} g(a)$$

Bounding the regret

Theorem 16 For EluderUCB: $\mathrm{Reg}_n \lesssim \sqrt{n\beta^2 \operatorname{dimE}(\mathcal{F}, 1/n)}$

Proof Let $g_{t-1} = \arg\max_{g \in \mathcal{F}_{t-1}} \max_{a \in \mathcal{A}} g(a)$

$$\begin{aligned}
\mathrm{Reg}_n &= \mathbb{E}\left[\sum_{t=1}^{n} f(a^\star) - f(a_t)\right] \\
&\lesssim \mathbb{E}\left[\left(\sum_{t=1}^{n} g_{t-1}(a_t) - f(a_t)\right)\right] && \text{(Optimism)} \\
&\leqslant \mathbb{E}\left[\sum_{t=1}^{n} w_{t-1}(a_t)\right] \\
&\lesssim \sqrt{n \operatorname{dimE}(\mathcal{F}, 1/n) \log(|\mathcal{F}|n)} && \text{(Corollary 15)}
\end{aligned}$$

$\square$

Proposition 4 $\dim E(\mathcal{F}, \varepsilon) \leqslant |\mathcal{A}|$ for all $\varepsilon > 0$ and all $\mathcal{F}$

Proof Suppose that $a \in \{a_1, \ldots, a_n\}$. We claim that a is ε-dependent on $\{a_1, \ldots, a_n\}$

Proposition 4 $\dim E(\mathcal{F}, \varepsilon) \leqslant |\mathcal{A}|$ for all $\varepsilon > 0$ and all $\mathcal{F}$

Proof Suppose that $a \in \{a_1, \ldots, a_n\}$. We claim that a is ε-dependent on $\{a_1, \ldots, a_n\}$

Def. of independence Given an $\varepsilon > 0$ and sequence $a_1, \ldots, a_n$ in $\mathcal{A}$, we say that $a \in \mathcal{A}$ is ε-dependent with respect to $(a_t)_{t=1}^n$ if

$$\forall f, g \in \mathcal{F} \text{ with } \sum_{t=1}^{n} (f(a_t) - g(a_t))^2 \leqslant \varepsilon^2, \quad f(a) - g(a) \leqslant \varepsilon$$

Proposition 4 $\dim E(\mathcal{F}, \varepsilon) \leqslant |\mathcal{A}|$ for all $\varepsilon > 0$ and all $\mathcal{F}$

Proof Suppose that $a \in \{a_1, \ldots, a_n\}$. We claim that a is ε-dependent on $\{a_1, \ldots, a_n\}$

> **Def. of independence** Given an $\varepsilon > 0$ and sequence $a_1, \ldots, a_n$ in $\mathcal{A}$, we say that $a \in \mathcal{A}$ is ε-dependent with respect to $(a_t)_{t=1}^n$ if
>
> $$\forall f, g \in \mathcal{F} \text{ with } \sum_{t=1}^{n} (f(a_t) - g(a_t))^2 \leqslant \varepsilon^2, \;\; f(a) - g(a) \leqslant \varepsilon$$

$a \in \{a_1, \ldots, a_n\}$ so

$$\left[\sum_{t=1}^{n} (f(a_t) - g(a_t))^2 \leqslant \varepsilon^2\right] \implies [f(a) - g(a) \leqslant \varepsilon]$$

$\square$

Proposition 5 When $\mathcal{A} \subset \mathbb{R}^d$ and $\mathcal{F} = \{f : f(a) = \langle a, \theta \rangle, \theta \in \mathbb{R}^d\}$, then

$$\dim E(\mathcal{F}, \varepsilon) = O(d \log(1/\varepsilon)) \text{ for all } \varepsilon > 0$$

Proof

Complicated... Elliptical potential... Blah blah

Let $\dim = \dim E(\mathcal{F}, \varepsilon)$. By definition, there exists a sequence $(a_t)_{t=1}^{\dim}$ and $(f_t, g_t)_{t=1}^{\dim}$ such that for $t \in [\dim]$,

$$\langle f_t - g_t, a_t \rangle \geqslant \varepsilon \qquad \sum_{s=1}^{t-1} \langle f_t - g_t, a_s \rangle^2 \leqslant \varepsilon^2$$

Hence, with $G_t = \varepsilon^2 I + \sum_{s=1}^{t} a_s a_s^\top$,

$$\begin{aligned}
\varepsilon^2(1 - \|a_t\|^2_{G_t^{-1}}) &\leqslant \langle f_t - g_t, a_t \rangle^2 (1 - \|a_t\|^2_{G_t^{-1}}) \\
&\leqslant \|f_t - g_t\|^2_{G_t} \|a_t\|^2_{G_t^{-1}} - \langle f_t - g_t, a_t \rangle^2 \|a_t\|^2_{G_t^{-1}} \\
&= \|f_t - g_t\|^2_{G_{t-1}} \|a_t\|^2_{G_t^{-1}} \\
&\leqslant 2\varepsilon^2 \|a_t\|^2_{G_t^{-1}}
\end{aligned}$$

Summing: $\dim \varepsilon^2 \leqslant 3\varepsilon^2 \sum_{t=1}^{\dim} \|a_t\|^2_{G_t^{-1}} \overset{\text{Abbasi-Yadkori et al. [2011]}}{\leqslant} 3d\varepsilon^2 \log\left(1 + \frac{\dim}{d\varepsilon^2}\right)$ □

Consequences

- For k-armed bandits: $\mathcal{A} = \{1, \ldots, k\}$ and $\mathcal{F} = [0, 1]^k$,

$$\mathrm{Reg}_n \lesssim \sqrt{kn \log(|\mathcal{F}|n)}$$

- For linear bandits: $\mathcal{A} \subset \mathbb{R}^d$ and $\mathcal{F} = \{a \mapsto \langle a, \theta \rangle : \theta \in \mathbb{R}^d\}$

$$\mathrm{Reg}_n \lesssim \sqrt{dn \log(|\mathcal{F}|n)}$$

Consequences

- For k-armed bandits: $\mathcal{A} = \{1, \ldots, k\}$ and $\mathcal{F} = [0,1]^k$,

$$\mathrm{Reg}_n \lesssim \sqrt{kn \log(|\mathcal{F}|n)}$$

- For linear bandits: $\mathcal{A} \subset \mathbb{R}^d$ and $\mathcal{F} = \{a \mapsto \langle a, \theta \rangle : \theta \in \mathbb{R}^d\}$

$$\mathrm{Reg}_n \lesssim \sqrt{dn \log(|\mathcal{F}|n)}$$

$|\mathcal{F}| = \infty$ in both cases???

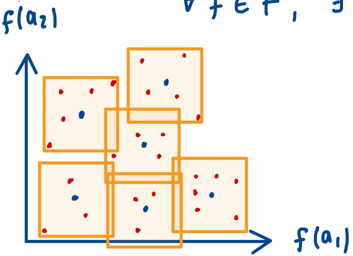
Covering

① If $\|f-g\|_\infty \lesssim \frac{1}{n}$, $\mathrm{Reg}_n(f) \approx \mathrm{Reg}_n(g)$

② If $G \subset F \Longrightarrow \dim E(G,\varepsilon) \lesssim \dim E(F,\varepsilon)$

③ Find smallest $G \subset F$ such that

$$\forall f \in F,\ \exists g \in G \ \ \|f-g\|_\infty \lesssim \frac{1}{n}$$



- Points in F
- Points in G (and F)

④ Covering number is size of G

Consequences

- For k-armed bandits: $\mathcal{A} = \{1, \ldots, k\}$ and $\mathcal{F} = [0, 1]^k$,

$$\mathrm{Reg}_n \lesssim \sqrt{kn \log(\text{Covering } n)} \lesssim k\sqrt{n \log(n))}$$

- For linear bandits: $\mathcal{A} \subset \mathbb{R}^d$ and $\mathcal{F} = \{a \mapsto \langle a, \theta \rangle : \theta \in \mathbb{R}^d\}$

$$\mathrm{Reg}_n \lesssim \sqrt{dn \log(\text{Covering } n)} \lesssim d\sqrt{n \log(n)}$$

Notes on Eluder dimension

- Definitions are made for optimistic algorithms
- Not a *real (?)* dimension – no lower bounds
- Relatively simple to work with
- further information: Li et al. [2021]
- **Alternative information-theoretic complexity measures**
- RL/Bandits: Foster et al. [2021]
- Adversarial partial monitoring: L [2022]

Global Optimism based on Local Fitting [Jin et al., 2021]

- Online RL setting
- Nonlinear function approximation
- Subsumes many existing frameworks on function approximation
- Statistically efficient
- Not computationally efficient

Assumptions

Algorithm uses a function approximation class $\mathcal{F} \subset [0, H]^{\mathcal{S}\times\mathcal{A}}$

Remember, the Bellman operator $\mathcal{T} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is

$$(\mathcal{T}f)(s, a) = r(s, a) + \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s, a)\underline{f}(s')$$

with $\underline{f}(s) = \max_{a\in\mathcal{A}} f(s, a)$

Assumption 4 (Realisability) $q^\star \in \mathcal{F}$

Assumption 5 (Closedness) $\mathcal{T}f \in \mathcal{F}$ for all $f \in \mathcal{F}$

Bellman operator on q-values

- Let $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Abbreviate $\underline{f}(s) = \max_{a \in \mathcal{A}} f(s, a)$
- Define $\mathcal{T} : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \to \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ by

$$(\mathcal{T} f)(s, a) = r(s, a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, a)\underline{f}(s')$$

- Note: $\mathcal{T}$ depends on the dynamics/rewards of the (unknown) MDP
- We write π_f for the **greedy policy** with respect to f

$$\pi_f(s) = \operatorname*{arg\,max}_{a \in \mathcal{A}} f(s, a)$$

Policy loss decomposition

Proposition 6 Let $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\underline{f}(s) = \max_a q(s, a)$ and $\pi = \pi_f$. Then

$$\underline{f}(s_1) - v^\pi(s_1) = \mathbb{E}_\pi \left[\sum_{h=1}^{H} (f - \mathcal{T} f)(s_h, a_h) \right]$$

GOLF Algorithm Intuition

- Maintain confidence set $\mathcal{F}_t$ containing $q^\star$ with high probability
- Let $f_t \in \mathcal{F}_{t-1}$ be the optimistic q-function

$$f_t = \arg\max_{f \in \mathcal{F}_{t-1}} \underline{f}(s_1)$$

- Play optimistically: $\pi_t(s) = \arg\max_{a \in \mathcal{A}} f_t(s, a)$

GOLF Algorithm Intuition

- Maintain confidence set $\mathcal{F}_t$ containing $q^\star$ with high probability
- Let $f_t \in \mathcal{F}_{t-1}$ be the optimistic q-function

$$f_t = \operatorname*{arg\,max}_{f \in \mathcal{F}_{t-1}} \underline{f}(s_1)$$

- Play optimistically: $\pi_t(s) = \arg\max_{a \in \mathcal{A}} f_t(s, a)$
- By optimism and policy loss decomposition

$$\underbrace{v^\star(s_1) - v^{\pi_t}(s_1)}_{\text{regret}} \leqslant \underline{f}_t(s_1) - v^{\pi_t}(s_1) = \mathbb{E}_{\pi_t}\left[\sum_{h=1}^{H} (f_t - \mathcal{T} f_t)(s_h, a_h)\right]$$

GOLF Algorithm Intuition

- Maintain confidence set $\mathcal{F}_t$ containing $q^\star$ with high probability
- Let $f_t \in \mathcal{F}_{t-1}$ be the optimistic q-function

$$f_t = \operatorname*{arg\,max}_{f \in \mathcal{F}_{t-1}} \underline{f}(s_1)$$

- Play optimistically: $\pi_t(s) = \arg\max_{a \in \mathcal{A}} f_t(s, a)$
- By optimism and policy loss decomposition

$$\underbrace{v^\star(s_1) - v^{\pi_t}(s_1)}_{\text{regret}} \leqslant \underline{f}_t(s_1) - v^{\pi_t}(s_1) = \mathbb{E}_{\pi_t}\left[\sum_{h=1}^{H} (f_t - \mathcal{T} f_t)(s_h, a_h)\right]$$

Acting greedily with respect to an optimistic q-value function that nearly satisfies the Bellman equation is nearly optimal

GOLF Algorithm Intuition

- Maintain confidence set $\mathcal{F}_t$ containing $q^\star$ with high probability
- Let $f_t \in \mathcal{F}_{t-1}$ be the optimistic q-function

$$f_t = \arg\max_{f \in \mathcal{F}_{t-1}} \underline{f}(s_1)$$

- Play optimistically: $\pi_t(s) = \arg\max_{a \in \mathcal{A}} f_t(s, a)$
- By optimism and policy loss decomposition

$$\underbrace{v^\star(s_1) - v^{\pi_t}(s_1)}_{\text{regret}} \leqslant \underline{f}_t(s_1) - v^{\pi_t}(s_1) = \mathbb{E}_{\pi_t}\left[\sum_{h=1}^{H} (f_t - \mathcal{T} f_t)(s_h, a_h)\right]$$

Acting greedily with respect to an optimistic q-value function that nearly satisfies the Bellman equation is nearly optimal

Need a way to eliminate functions f in the confidence set for which the Bellman error is large

Problem Bellman operator depends on unknown dynamics

GOLF Algorithm Intuition

Suppose $s' \sim \mathcal{P}(s,a)$ and $r \sim \mathcal{R}(s,a)$

$$\mathbb{E}[r + \underline{f}(s')] = (\mathcal{T}f)(s,a) \overset{\text{if } \mathcal{T}f = f}{=} f(s,a)$$

If $\mathcal{T}f - f$ is large

$$\sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \underline{f}(s')\right)^2 \gg \sum_{s,a,r,s' \in \mathcal{D}} (\underbrace{(\mathcal{T}f)(s,a)}_{\in \mathcal{F}} - r - \underline{f}(s'))^2$$

If $\mathcal{T}f = f$

$$\sum_{s,a,r,s' \in \mathcal{D}} (\underbrace{f(s,a)}_{(\mathcal{T}f)(s,a)} - r - \underline{f}(s'))^2 \overset{\text{whp}}{\lesssim} \sum_{s,a,r,s' \in \mathcal{D}} \left(g(s,a) - r - \underline{f}(s')\right)^2 + \beta^2$$

GOLF Algorithm

1. Set $\mathcal{D} = \emptyset$ and $\mathcal{C}_0 = \mathcal{F}$
2. for episode $t = 1, 2 \ldots$
3. Choose policy $\pi_t = \pi_{f^t}$ where

$$f_t = \arg\max_{f \in \mathcal{C}_{t-1}} \underline{f}(s_1)$$

4. Run π_t for one episode and add data to $\mathcal{D}$
5. Update confidence set:

$$\mathcal{C}_t = \{f \in \mathcal{C}_{k-1} : \mathcal{L}_{\mathcal{D}}(f, f) \leqslant \inf_{g \in \mathcal{F}} \mathcal{L}_{\mathcal{D}}(g, f) + \beta^2\}$$

where $\mathcal{L}_{\mathcal{D}}(g, f) = \sum_{s,a,r,s' \in \mathcal{D}} (g(s, a) - r - f(s'))^2$

Concentration analysis

Given $\mathcal{D} \subset \mathcal{S} \times \mathcal{A} \times [0,1] \times \mathcal{S}$ collected by some policy and

$$\mathcal{C}_{\mathcal{D}} = \left\{ f \in \mathcal{F} : \mathcal{L}_{\mathcal{D}}(f) \leqslant \inf_{g \in \mathcal{F}} \mathcal{L}_{\mathcal{D}}(g, f) + \beta^2 \right\} \qquad \beta^2 = \text{cnst} \log \left(\frac{|\mathcal{F}|}{\delta} \right)$$

where $\mathcal{L}_{\mathcal{D}}(g, f) = \sum_{s,a,r,s' \in \mathcal{D}} [g(s,a) - r - f(s')]^2$ and $\mathcal{L}_{\mathcal{D}}(f) = \mathcal{L}_{\mathcal{D}}(f, f)$

Proposition 7 If $f = \mathcal{T} f$, then $\mathbb{P}(f \in \mathcal{C}_{\mathcal{D}}) \geqslant 1 - \delta$

Concentration analysis

Given $\mathcal{D} \subset \mathcal{S} \times \mathcal{A} \times [0,1] \times \mathcal{S}$ collected by some policy and

$$\mathcal{C}_\mathcal{D} = \left\{ f \in \mathcal{F} : \mathcal{L}_\mathcal{D}(f) \leqslant \inf_{g \in \mathcal{F}} \mathcal{L}_\mathcal{D}(g, f) + \beta^2 \right\} \qquad \beta^2 = \text{cnst} \log\left(\frac{|\mathcal{F}|}{\delta}\right)$$

where $\mathcal{L}_\mathcal{D}(g, f) = \sum_{s,a,r,s' \in \mathcal{D}} [g(s,a) - r - f(s')]^2$ and $\mathcal{L}_\mathcal{D}(f) = \mathcal{L}_\mathcal{D}(f, f)$

Proposition 7 If $f = \mathcal{T} f$, then $\mathbb{P}(f \in \mathcal{C}_\mathcal{D}) \geqslant 1 - \delta$

Proof Let $g \in \mathcal{F}$. By concentration of measure (next slide), with probability at least $1 - \delta/|\mathcal{F}|$

$$\begin{aligned}
\mathcal{L}(f) - \mathcal{L}(g, f) &\leqslant -\sum_{s,a,r,s' \in \mathcal{D}} (f-g)^2(s,a) + \text{cnst}\left[\sqrt{\sum_{s,a,r,s' \in \mathcal{D}} (f-g)^2(s,a) \log\left(\frac{|\mathcal{F}|}{\delta}\right)} + \log\left(\frac{|\mathcal{F}|}{\delta}\right)\right] \\
&\leqslant \sup_{x \in \mathbb{R}} \left[-x^2 + \text{cnst}\, x \sqrt{\log\left(\frac{|\mathcal{F}|}{\delta}\right)} + \text{cnst} \log\left(\frac{|\mathcal{F}|}{\delta}\right)\right] \\
&\leqslant \text{cnst} \log\left(\frac{|\mathcal{F}|}{\delta}\right) = \beta^2
\end{aligned}$$

Result follows by union bound (and covering number argument) □

Concentration analysis (cont.)

$$\mathcal{L}_{\mathcal{D}}(f) - \mathcal{L}_{\mathcal{D}}(g, f) = \sum_{i=1}^{m} \underbrace{[f(s_i, a_i) - r_i - f(s_i')]^2 - [g(s_i, a_i) - r_i - f(s_i')]^2}_{X_i}$$

Given $r \sim \mathcal{R}(s, a)$ and $s' \sim \mathcal{P}(s, a)$,

$$\begin{aligned}
X &= \mathbb{E}\left[\left(f(s,a) - (r + f(s'))\right)^2 - \left(g(s,a) - (r + f(s'))\right)^2\right] \\
&= f(s,a)^2 - g(s,a)^2 + 2\mathbb{E}[r + f(s')](g(s,a) - f(s,a)) \\
&= f(s,a)^2 - g(s,a)^2 + 2(\mathcal{T} f)(s,a))(g(s,a) - f(s,a)) \\
&= f(s,a)^2 - g(s,a)^2 + 2f(s,a)(g(s,a) - f(s,a)) \\
&= -(f(s,a) - g(s,a))^2
\end{aligned}$$

Similar calculation: $\mathbb{V}[X] \leqslant \text{cnst}(f(s,a) - g(s,a))^2$

By martingale Bernstein inequality

$$\sum_{i=1}^{m} X_i \leqslant \sum_{i=1}^{m} \mathbb{E}[X_i] + \text{cnst}\sqrt{\sum_{i=1}^{m} \mathbb{V}[X_i] \log\left(\frac{1}{\delta}\right)} + \text{cnst} \log\left(\frac{1}{\delta}\right)$$

Concentration analysis (cont.)

$$\mathcal{L}_{\mathcal{D}}(f) - \mathcal{L}_{\mathcal{D}}(g, f) = \sum_{i=1}^{m} \underbrace{[f(s_i, a_i) - r_i - f(s_i')]^2 - [g(s_i, a_i) - r_i - f(s_i')]^2}_{X_i}$$

Given $r \sim \mathcal{R}(s, a)$ and $s' \sim \mathcal{P}(s, a)$ and $g = \mathcal{T} f$,

$$\begin{aligned}
X &= \mathbb{E}\left[\left(f(s,a) - (r + f(s'))\right)^2 - \left(g(s,a) - (r + f(s'))\right)^2\right] \\
&= f(s,a)^2 - g(s,a)^2 + 2\mathbb{E}[r + f(s')](g(s,a) - f(s,a)) \\
&= f(s,a)^2 - g(s,a)^2 + 2(\mathcal{T} f)(s,a))(g(s,a) - f(s,a)) \\
&= f(s,a)^2 - (\mathcal{T} f)(s,a)^2 + 2(\mathcal{T} f)(s,a))((\mathcal{T} f)(s,a) - f(s,a)) \\
&= (f(s,a) - (\mathcal{T} f)(s,a))^2
\end{aligned}$$

Similar calculation: $\mathbb{V}[X] \leqslant \text{cnst}(f(s,a) - (\mathcal{T} f)(s,a))^2$

By martingale Bernstein inequality

$$\sum_{i=1}^{m} X_i \geqslant \sum_{i=1}^{m} \mathbb{E}[X_i] - \text{cnst}\sqrt{\sum_{i=1}^{m} \mathbb{V}[X_i] \log\left(\frac{1}{\delta}\right)} - \text{cnst} \log\left(\frac{1}{\delta}\right)$$

Concentration analysis (summary)

We showed that with probability at least $1-\delta$ that any f with $\mathcal{T} f = f$

$$f \in \mathcal{C}_t$$

$\boxed{q^\star \text{ in } \mathcal{C}_t \text{ for all episodes } t \text{ with high probability}}$

$$\mathcal{L}_{\mathcal{D}}(f) - \mathcal{L}_{\mathcal{D}}(\mathcal{T} f, f)$$

$$\geqslant \sum_{s,a,r,s' \in \mathcal{D}} ((f - \mathcal{T} f)(s,a))^2 - \sqrt{\sum_{s,a,r,s' \in \mathcal{D}} ((f - \mathcal{T} f)(s,a))^2 \log \frac{1}{\delta}} - \log \frac{1}{\delta}$$

$\mathcal{L}_{\mathcal{D}_t}(f) - \mathcal{L}_{\mathcal{D}_t}(\mathcal{T} f, f) \geqslant \beta^2$ if

$$\sum_{u=1}^{t} \mathbb{E}_{\pi^u} \left[\sum_{h=1}^{H} ((f - \mathcal{T} f)(s_h^u, a_h^u))^2 \right] \geqslant \text{cnst}\, \beta^2$$

Regret analysis

1. By optimism

$$\begin{aligned}\mathrm{Reg}_n &= \sum_{t=1}^{n} v^\star(s_1) - v^{\pi_t}(s_1) \\ &\overset{\text{whp}}{\leqslant} \sum_{t=1}^{n} f^t(s_1, \pi_{f^t}(s_1)) - v^{\pi_t}(s_1) && \text{(Optimism)} \\ &= \sum_{t=1}^{n} \sum_{h=1}^{H} \mathbb{E}_{\pi_t}[(f^t - \mathcal{T} f^t)(s_h, a_h)] && \text{(Prop. 3)}\end{aligned}$$

2. Since $f^t \in \mathcal{F}_{t-1}$

$$\sum_{u=1}^{t-1} \mathbb{E}_{\pi^u}\left[\sum_{h=1}^{H} ((f^t - \mathcal{T} f^t)(s_h^u, a_h^u))^2\right] \leqslant \text{cnst}\, \beta^2$$

Bellman Eluder dimension

$$\mathrm{Reg}_n \leqslant \sum_{t=1}^{n} \sum_{h=1}^{H} \mathbb{E}_{\pi_t}[(f^t - \mathcal{T} f^t)(s_h, a_h)]$$

For all t

$$\sum_{u=1}^{t-1} \mathbb{E}_{\pi^u}\left[\sum_{h=1}^{H} ((f^t - \mathcal{T} f^t)(s_h^u, a_h^u))^2\right] \leqslant \mathsf{cnst}\, \beta^2$$

$$\mathrm{Reg}_n \leqslant \sum_{t=1}^{n} \sum_{h=1}^{H} \mathbb{E}_{\pi_t}[(f^t - \mathcal{T} f^t)(s_h, a_h)]$$

For all t

$$\sum_{u=1}^{t-1} \mathbb{E}_{\pi^u} \left[\sum_{h=1}^{H} ((f^t - \mathcal{T} f^t)(s_h^u, a_h^u))^2 \right] \leqslant \text{cnst}\, \beta^2$$

Bellman Eluder dimension

Let $\mathcal{E} = \{\mathbb{E}_{\pi_f} : f \in \mathcal{F}\}$
Given a sequence $\mathbb{E}_1, \ldots, \mathbb{E}_m$ in $\mathcal{E}$. We say $\mathbb{E} \in \mathcal{E}$ is ε-dependent if for all $f - \mathcal{T} f$,

$$\sum_{u=1}^{m} \mathbb{E}_u \left[\sum_{h=1}^{H} ((f - \mathcal{F})(s_h, a_h))^2 \right] \leqslant \varepsilon^2 \Rightarrow \mathbb{E} \left[\sum_{h=1}^{H} ((f - \mathcal{F}))(s_h, a_h) \right] \leqslant \varepsilon$$

The Bellman Eluder dimension $\mathrm{dimBE}(\mathcal{F}, \varepsilon)$ is the longest sequence of ε-independent expectation operators

Final result and applications

Theorem 17 $\mathrm{Reg}_n = \tilde{O}(H\sqrt{\mathrm{dimBE}(\mathcal{F}, \varepsilon) n \beta^2})$

What has low Bellman eluder dimension?

- Tabular (first lecture)
- Linear MDPs (last lecture)
- Generalised linear MDPs
- Kernel MDPs
- All problems with low Bellman rank
- All problems with low Eluder dimension

Comparison to other complexity measures

- Linear MDPs
- Eluder dimension [Osband and Van Roy, 2014]
- Bellman rank [Jiang et al., 2017]
- Witness rank [Sun et al., 2019]
- Bilinear rank [Du et al., 2021]

Negative results

- Last lecture we showed that you can learn with a generative model when for all π there exists a θ such that $q^{\pi}(s, a) = \langle \phi(s, a), \theta \rangle$
- What if only the **optimal policies** are linearly realisable? $q^{\star}(s, a) = \langle \phi(s, a), \theta \rangle$
- Polynomial sample complexity not possible

TensorPlan [Weisz et al., 2021]

- Finite horizon setting
- Local access planning
- Linear features $\varphi : \mathcal{S} \to \mathbb{R}^d$
- Only assume that **value function** of **optimal policy** $\pi^\star$ is realisable
- There exists a θ such that

$$v^{\pi^\star}(s) = \langle \varphi(s), \theta \rangle \text{ for all } s \in \mathcal{S}$$

- Number of samples needed for ε-accuracy is

$$\mathrm{poly}((dH/\varepsilon)^{|\mathcal{A}|})$$

Other (not covered) topics

- Information-theoretic complexity measures
- Batch RL
- RL Theory website https://rltheory.github.io/
- Draft RL Theory book by (Alekh Agarwal, Nan Jiang, Sham Kakade and Wen Sun): https://rltheorybook.github.io/

Y. Abbasi-Yadkori, D. Pál, and Cs. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320. Curran Associates, Inc., 2011.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 89–96, 2009.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1707–1714, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

D. Foster, S. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Botao Hao, Nevena Lazic, Dong Yin, Yasin Abbasi-Yadkori, and Csaba Szepesvari. Confident least square value iteration with local access to a simulator. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2420–2435. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/hao22a.html.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12(5):363–365, 1960.

Tor L. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1547–1575. PMLR, 02–05 Jul 2022.

T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.

T. Lattimore and M. Hutter. PAC bounds for discounted MDPs. In *Proceedings of the 23th International Conference on Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pages 320–334. Springer Berlin / Heidelberg, 2012.

T. Lattimore and Cs. Szepesvári. Learning with good feature representations in bandits and in RL with a generative model. arXiv:1911.07676, 2019.

Gene Li, Pritish Kamath, Dylan J Foster, and Nathan Srebro. Eluder dimension and generalized rank. *arXiv preprint arXiv:2104.06970*, 2021.

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, and Zheng Wen. Reinforcement learning, bit by bit. *arXiv preprint arXiv:2103.04047*, 2021.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. *Advances in neural information processing systems*, 30, 2017.

D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264. Curran Associates, Inc., 2013.

Matthew J Sobel. The variance of discounted markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. A fully problem-dependent regret lower bound for finite-horizon mdps. *arXiv preprint arXiv:2106.13013*, 2021.

M. J. Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Gellért Weisz, Philip Amortila, Barnabás Janzer, Yasin Abbasi-Yadkori, Nan Jiang, and Csaba Szepesvári. On query-efficient planning in mdps under linear realizability of the optimal state-value function. In *Conference on Learning Theory*, pages 4355–4385. PMLR, 2021.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.