Some Theoretical Aspects of Reinforcement Learning
CS 285

Instructor: Aviral Kumar
UC Berkeley

Berkeley
UNIVERSITY OF CALIFORNIA

What Will We Discuss Today?

A brief introduction to some theoretical aspects of RL: In particular error/suboptimality-analysis of RL algorithms, understanding of regret, and function approximation

- Notions of Convergence in RL, Assumptions and Preliminaries
- Optimization Error in RL and Analyses of Fitted Q-Iteration Algorithms
- Regret Analyses of RL Algorithms: An Introduction
- RL with Function Approximation: When can we still obtain convergent algorithms?

This is not at all an exhaustive coverage of topics in RL theory, checkout various resources on the last slide of this lecture.

Metrics used to evaluate RL methods

Sample complexity

Used typically for measuring how easy is to infer the optimal policy assuming no exploration bottlenecks (e.g., in offline RL)

How many transitions/episodes do I need to obtain a good policy?

$$N = \mathcal{O}\left(\text{poly}\left(|S|, |A|, \frac{1}{1-\gamma}\right)\right) \quad \text{then} \quad \max_{s,a} |Q^\pi(s,a) - \hat{Q}^\pi(s,a)| \leq \varepsilon$$

Regret

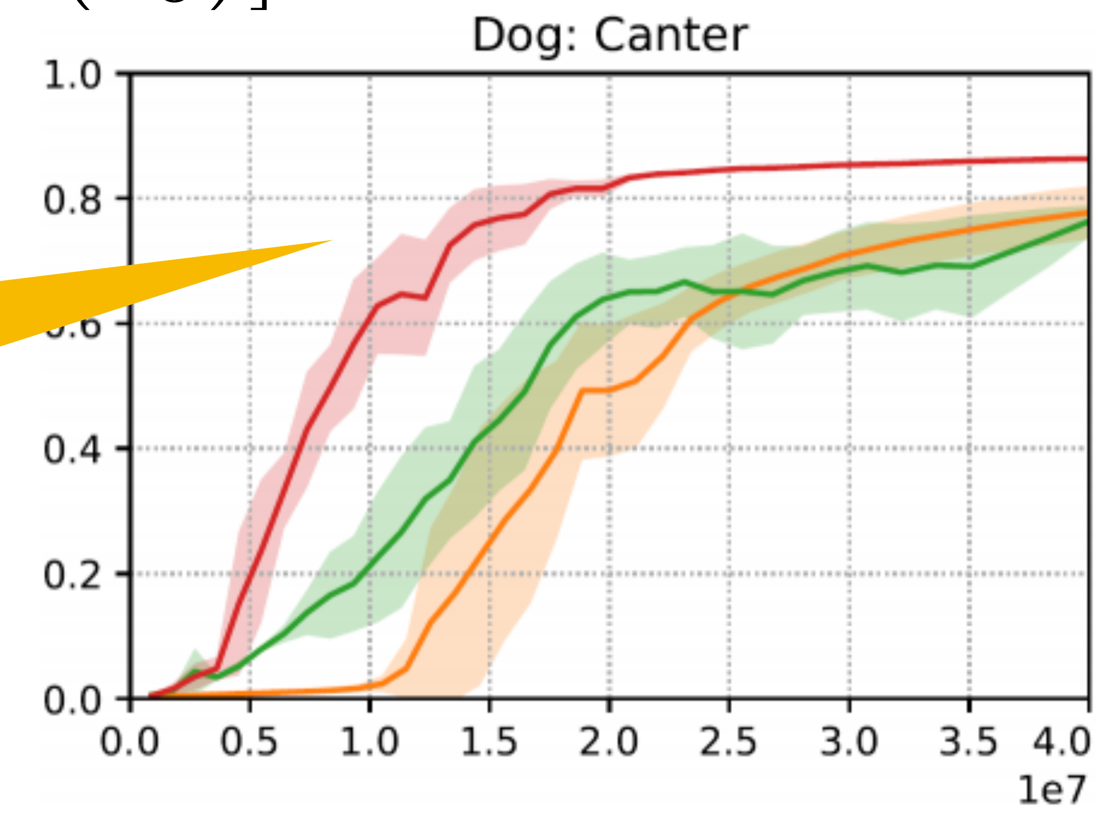
Used typically for measuring how good an exploration scheme is

$$\pi_0, \pi_1, \pi_2, \cdots, \pi_N$$

$$\text{Reg(N)} = \sum_{i=1}^{N} E_{s_0 \sim \rho}[V^*(s_0)] - E_{s_0 \sim \rho}[V^{\pi_i}(s_0)]$$

$$\text{Reg(N)} = \mathcal{O}(\sqrt{N})$$

This area



Assumptions used in RL Analyses

We can breakdown the RL into two parts:
- the exploration part
- given data from the exploration policy, we should be able to learn from it

Can we analyze these separately?

To remove the exploration aspect, perform analysis under the "generative model" assumption

$$\text{access to sampling a model} \quad s' \sim P(\cdot|s, a)$$

Suppose we can query the **true** dynamics model of the MDP for each (s, a) pair N times and construct an **empirical** dynamics model

$$\hat{P}(s'|s, a) = \frac{\#(s', a, s)}{N}$$

Goal: Approximate the Q-function or the value function

How does the approximation error of this model translate to errors in the value function?

Preliminaries

Concentration

Lemma A.1. *(Hoeffding's inequality) Suppose $X_1, X_2, \ldots X_n$ are a sequence of independent, identically distributed (i.i.d.) random variables with mean μ. Let $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. Suppose that $X_i \in [b_-, b_+]$ with probability 1, then*

$$P(\bar{X}_n \geq \mu + \epsilon) \leq e^{-2n\epsilon^2/(b_+ - b_-)^2}.$$

Similarly,

$$P(\bar{X}_n \leq \mu - \epsilon) \leq e^{-2n\epsilon^2/(b_+ - b_-)^2}.$$

$$\bar{X}_n - EX \leq (b_+ - b_-)\sqrt{\ln(1/\delta)/(2n)}\,.$$

Says that average over samples gets closer to the mean

More complex variants:

Proposition A.4. (Concentration for Discrete Distributions) Let z be a discrete random variable that takes values in $\{1, \ldots, d\}$, distributed according to q. We write q as a vector where $\vec{q} = [\Pr(z = j)]_{j=1}^{d}$. Assume we have N iid samples, and that our empirical estimate of $\vec{q}$ is $[\hat{q}]_j = \sum_{i=1}^{N} \mathbf{1}[z_i = j]/N$.

$$\Pr\left(\|\hat{q} - \vec{q}\|_1 \geq \sqrt{d}(1/\sqrt{N} + \epsilon)\right) \leq e^{-N\epsilon^2}.$$

We will use this version to obtain a worst case bound on the generative model.

Lemmas from RL Theory Textbook (Draft). Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io/

Part 1: Sampling/Optimization Error in RL

Goal: How does error in training translate to error in the value-function?

We will analyze this optimization error in two settings:
(1) generative model **(2)** Fitted Q-iteration

We want results of the form:

$$\text{if } ||\hat{P}(s'|s,a) - P(s'|s,a)||_1 \leq \varepsilon \ \text{ then } ||Q(s,a) - \hat{Q}(s,a)||_\infty \leq \delta$$

$$\text{if } ||Q(s,a) - \hat{T}Q(s,a)||_\infty \leq \varepsilon \ \text{ then } ||Q(s,a) - \hat{Q}(s,a)||_\infty \leq \delta$$

$$TQ(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(s'|s,a)} \left[\max_{a'} Q(s',a')\right]$$

$$\hat{T}Q(s,a) = \hat{r}(s,a) + \gamma \mathbb{E}_{s' \sim \hat{P}(s'|s,a)} \left[\max_{a'} Q(s',a')\right]$$

"Empirical" Bellman operator: constructed using transition samples observed by sampling the MDP

Sampling Error with Generative Model

1. Estimate $\hat{P}(s'|s,a)$
2. For a given policy, plan under this dynamics model to obtain the Q-function $\hat{Q}^{\pi}$

$$\hat{P}(s'|s,a) = \frac{\#(s',a,s)}{N}$$

First Step: Bound the difference between the learned and true dynamics model

Use concentration inequalities

with high probability greater than $1-\delta$

$$\|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1 \leq c\sqrt{\frac{|S|\log(1/\delta)}{m}}$$

m $=$ number of samples used to estimate $p(s'|s,a)$

The empirical dynamics model and the actual dynamics model are close

Sampling Error with Generative Model

Second step: Compute how the dynamics model affects the Q-function

$$V^{\pi}(s) = Q^{\pi}(s, \pi(s)).$$
$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[V^{\pi}(s')\right].$$

$$P^{\pi}_{(s,a),(s',a')} := P(s'|s,a)\pi(a'|s').$$

$$Q^{\pi} = r + \gamma P V^{\pi}$$
$$Q^{\pi} = r + \gamma P^{\pi} Q^{\pi}$$

$$\boxed{Q^{\pi} = (I - \gamma P^{\pi})^{-1} r}$$

Q-function depends on the dynamics model P(s'|s, a) via a non-linear transformation

$$\boxed{Q^{\pi} - \widehat{Q}^{\pi} = \gamma (I - \gamma \widehat{P}^{\pi})^{-1} (P - \widehat{P}) V^{\pi}}$$

$$\begin{aligned} Q^{\pi} - \widehat{Q}^{\pi} &= (I - \gamma P^{\pi})^{-1} r - (I - \gamma \widehat{P}^{\pi})^{-1} r \\ &= (I - \gamma \widehat{P}^{\pi})^{-1}((I - \gamma \widehat{P}^{\pi}) - (I - \gamma P^{\pi})) Q^{\pi} \\ &= \gamma (I - \gamma \widehat{P}^{\pi})^{-1} (P^{\pi} - \widehat{P}^{\pi}) Q^{\pi} \\ &= \gamma (I - \gamma \widehat{P}^{\pi})^{-1} (P - \widehat{P}) V^{\pi} \end{aligned}$$

1. Express Q in the vector form
2. Express the difference between the two vectors in a more closed form version and obtain ($\hat{P}$ - P) in the expression

Sampling Error with Generative Model

Third step: Understand how error in the Q-function depends on error in the model

$$\boxed{Q^\pi - \widehat{Q}^\pi \;=\; \gamma(I-\gamma\widehat{P}^\pi)^{-1}(P-\widehat{P})V^\pi}$$

For any policy π, MDP M and vector $v \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we have $\left\|(I-\gamma P^\pi)^{-1}v\right\|_\infty \leq \|v\|_\infty/(1-\gamma)$.

Define $\quad w = (I-\gamma P^\pi)^{-1}v.$

$$\|v\| = \|(I-\gamma P^\pi)w\| \geq \|w\|_\infty - \gamma\,\|P^\pi w\|_\infty \geq \|w\|_\infty - \gamma\,\|w\|_\infty$$

Triangle inequality

$||P^\pi||_\infty \leq 1$

Thus, $\quad ||w||_\infty \leq ||v||_\infty/(1-\gamma)$

$$\boxed{\|Q^\pi - \widehat{Q}^\pi\|_\infty = \|\gamma(I-\gamma\widehat{P}^\pi)^{-1}(P-\widehat{P})V^\pi\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P-\widehat{P})V^\pi\|_\infty}$$

Sampling Error with Generative Model

Final step: Completing the Proof

$$\boxed{\|Q^\pi - \widehat{Q}^\pi\|_\infty = \|\gamma(I - \gamma\widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\pi\|_\infty}$$

$$\leq \frac{\gamma}{1-\gamma}\left(\max_{s,a}\|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1\right)\|V^\pi\|_\infty$$

Bound the max element of the product by product of max elements

$$\leq \frac{\gamma}{(1-\gamma)^2}\max_{s,a}\|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1$$

Assume $R_{\max} = 1$

Now use the previous relation

$$\|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1 \leq c\sqrt{\frac{|S|\log(1/\delta)}{m}}$$

$$||Q^\pi - \hat{Q}^\pi|| \leq \frac{\gamma}{(1-\gamma)^2} c\sqrt{\frac{|S|\log(1/\delta)}{m}}$$

We want atmost eps error in Q^π, compute the minimum number of samples m needed for this..

Proof Takeaways and Summary

$$||Q^\pi - \hat{Q}^\pi|| \leq \frac{\gamma}{(1-\gamma)^2} c \sqrt{\frac{|S| \log(1/\delta)}{m}}$$

- A small error in estimating the dynamics model implies small error in the Q-function
- However, error “compounds”: Note the (1 - gamma)^2 factor in the denominator of the bound.
- The more samples we collect, the better our estimate will be, but sadly samples aren’t free!

How does optimization error manifest in model-free variants (e.g., fitted Q-iteration)?

Part 2: Optimization Error in FQI

Fitted Q-iteration runs a sequence of backups by minimizing mean-squared error

initial Q-value Q_0 $\qquad$ $Q_{k+1} \leftarrow \min_{Q} ||Q - \hat{T}Q_k||_2^2$

if we use T instead of $\hat{T}$ and $||Q_{k+1} - TQ_k|| = 0$

then FQI converges to the optimal Q-function Q^*

Which sources of error are we considering here?

- T is inexact, “sampling error” due to limited samples
- Bellman errors in that $|Q_{k+1} - TQ_k|$ may not be 0

$$\hat{T}Q(s,a) = \hat{r}(s,a) + \gamma \mathbb{E}_{s' \sim \hat{P}(s'|s,a)} \left[\max_{a'} Q(s',a') \right]$$

$$TQ(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(s'|s,a)} \left[\max_{a'} Q(s',a') \right]$$

Optimization Error in Fitted Q-Iteration

First Step: Bound the difference between the empirical and actual Bellman backup

$$|\hat{T}Q(s,a) - TQ(s,a)| \leq \Big|\hat{r}(s,a) - r(s,a) +$$

Concentration of reward

$$\gamma \left(E_{s \sim \hat{P}(s'|s,a)}[\max_{a'} Q(s',a')] - E_{s \sim P(s'|s,a)}[\max_{a'} Q(s',a')] \right) \Big|$$

Triangle inequality, bound each term separately

Concentration of dynamics

$$\leq |\hat{r}(s,a) - r(s,a)| + \gamma \left| E_{s \sim \hat{P}(s'|s,a)}[\max_{a'} Q(s',a')] - E_{s \sim P(s'|s,a)}[\max_{a'} Q(s',a')] \right|$$

Directly apply Hoeffding's

$$:= \left| \sum_{s'} (\hat{P}(s'|s,a) - P(s'|s,a)) \max_{a'} Q(s',a') \right|$$

Vector-form

$$\leq 2R_{\max} \sqrt{\frac{\log(1/\delta)}{2m}}$$

$$\leq ||\hat{P}(\cdot|s,a) - P(\cdot|s,a)||_1 \; ||Q||_\infty$$

Sum of product ≤ sum of product of absolute values, Q-values bounded by the ∞-norm

$$\bar{X}_n - EX \leq (b_+ - b_-)\sqrt{\ln(1/\delta)/(2n)}\,.$$

Optimization Error in Fitted Q-Iteration

Combining the bounds on the previous slide, and taking a max over (s, a) we get:

$$||\hat{T}Q - TQ||_\infty \leq 2R_{\max} c_1 \sqrt{\frac{\log(|S||A|/\delta)}{m}} + c_2 ||Q||_\infty \sqrt{\frac{|S| \log(1/\delta)}{m}}$$

Second step: How does error in each fitting iteration affect optimality

Let's say, we incur ε_k error in each fitting step of FQI, i.e., $||Q_{k+1} - TQ_k||_\infty \leq \varepsilon_k$

Then, what can we say about: $||Q_k - Q^*||_\infty \leq ?$

$$\begin{aligned}
||Q_k - Q^*||_\infty &\leq ||TQ_{k-1} + (Q_k - TQ_{k-1}) - TQ^*|| \\
&= ||\,(TQ_{k-1} - TQ^*) + (Q_k - TQ_{k-1})\,|| \\
&\leq ||TQ_{k-1} - TQ^*|| + ||Q_k - TQ_{k-1}|| \\
&\leq \gamma ||Q_{k-1} - Q^*||_\infty + \varepsilon_k
\end{aligned}$$

Optimization Error in Fitted Q-Iteration

$$
\begin{aligned}
||Q_k - Q^*||_\infty &\leq \gamma ||Q_{k-1} - Q^*||_\infty + \varepsilon_k \\
&\leq \gamma^2 ||Q_{k-2} - Q^*||_\infty + \gamma \varepsilon_{k-1} + \varepsilon_k \\
&\leq \gamma^k ||Q_0 - Q^*||_\infty + \sum_j \gamma^j \varepsilon_{k-j}
\end{aligned}
$$

Error from previous iteration "compounds", "propagates", etc...

Let's consider a large number of fitting iterations in FQI (so k tends ∞)

$$
\lim_{k \to \infty} ||Q_k - Q^*||_\infty \leq 0 + \lim_{k \to \infty} \sum_j \gamma^j \varepsilon_{k-j}
$$

We pay a price for each error term, and the total error in the worst-case is scaled by the (1 - gamma) factor in the denominator.

$$
\leq \left(\sum_{j=0}^{\infty} \gamma^j \right) ||\varepsilon||_\infty = \frac{||\varepsilon||_\infty}{1 - \gamma}
$$

Optimization Error in Fitted Q-Iteration

Completing the Proof

So far, we have seen how errors in the Bellman error can accumulate to form error against Q*

What is the total error in the Bellman error?
- optimization error ε_k
- "sampling error" due to limited data

$$||Q_k - TQ_{k-1}||_\infty = ||Q_k - \hat{T}Q_{k-1} + \hat{T}Q_{k-1} - TQ_{k-1}||_\infty$$
$$\leq ||Q_k - \hat{T}Q_{k-1}||_\infty + ||\hat{T}Q_{k-1} - TQ_{k-1}||_\infty$$

Optimization error: how easily can we minimize Bellman error

Sampling error: depends on number of times we see each (s, a)

$$\lim_{k\to\infty} ||Q_k - Q^*||_\infty \leq \frac{1}{1-\gamma} \max_k ||Q_k - TQ_{k-1}||_\infty \leq \cdots$$

Proof Takeaways and Summary

- Error compounds with FQI or DQN-style methods: especially a problem in offline RL settings, where the "sampling error" component is also quite high
- A stringent requirements with these bounds is that they directly ∞-norm of the error in the Q-function: but can we ever practically bound the error at the **worst** state-action pair? — Mostly not since we can't even enumerate the state or action-space!

Can we remove the dependency on the ∞-norm?

Yes! Can derive similar results for other data-distributions (μ) and L_p norms

$$||Q_k - Q^*||_p^\mu = \left(\mathbb{E}_{s,a\sim\mu(s,a)}[|Q_k(s,a) - Q^*(s,a)|^p]\right)^{1/p}$$

- So far we've looked at the generative model setting, where we have **<u>oracle</u>** MDP access to compute an approximate dynamics model. What happens in the substantially harder setting without this access, where we need exploration strategies? Coming up next…

Part 3: Analysis of Exploration Strategies

So far, we have analyzed RL algorithms in terms of optimization error and sampling error, however when the algorithm is provided with data, but we haven't seen where this data comes from. So, in the next part, we evaluate these algorithms on the cost of collecting data.

Multi-Armed Bandits

"1-step" RL

1. N possible arms/actions $a_1, a_2, \cdots, a_N$

2. Pull i-th arm in round t and observe corresponding (sampled) reward

$$r_t(a_i) \sim D(a_i), \text{ where } \mathbb{E}[r_t(a_i)] = \bar{r}(a_i)$$

3. Agent observes the resulting sampled reward and records it

$$\mathrm{Reg}(T) = T\bar{r}(a^*) - \sum_{t=1}^{T} \bar{r}(a_t)$$

Cumulative regret: How much are we losing by not picking the best arm in hindsight on the actual expected reward (not sampled reward)

If the regret grows sublinearly, then we are converging to the optimal action at infinity and thus learning "efficiently"

Exploration in Multi-Armed Bandits

UCB Algorithm / Optimistic exploration

$n^t(a_i)$ # times an arm was pulled

$\tilde{r}^t(a_i)$ Average of observed sample rewards

in round t pick arm a_t such that

$$a_t := \arg\max_{i=1,\cdots,N} \left(\tilde{r}^t(a_i) + \sqrt{\frac{\log(2NT/\delta)}{2n^t(a_i)}} \right)$$

Mean reward

Reward bonus

Where does this reward bonus come from?

$$\text{w.h.p. } \geq 1-\delta, \forall\ i \in [1,\cdots,N], t \in [1,\cdots,T] \quad |\tilde{r}^t(a_i) - \bar{r}(a_i)| \leq b(a_i)$$

$$\bar{X}_n - EX \leq (b_+ - b_-)\sqrt{\ln(1/\delta)/(2n)}\,.$$

Hoeffding inequality

Exploration in Multi-Armed Bandits

$$\tilde{r}^t(a_i) - b(a_i) \leq \quad \bar{r}(a_i) \quad \leq \tilde{r}^t(a_i) + b(a_i)$$

With high probability, the true reward for any arm lies in this interval defined by the bonus

How can we use this fact to obtain a bound on the regret?

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} (\bar{r}(a^*) - \bar{r}(a_t))$$

$$\leq \sum_{t=1}^{T} \left(\left[\tilde{r}(a^*) + b^t(a^*)\right] - \left[\tilde{r}(a_t) - b^t(a_t)\right] \right) + \delta T$$

Chosen arm maximizes this!

$$\leq \sum_{t=1}^{T} \left(\left[\tilde{r}(a_t) + b^t(a_t)\right] - \left[\tilde{r}(a_t) - b^t(a_t)\right] \right) \quad + \delta T$$

$$= 2 \sum_{i=1}^{T} b^t(a_t) = \mathcal{O}(\sqrt{T \cdot N \cdot \log\left(\frac{NT}{\delta}\right)}$$

Hint: Write down the expression for the bonus, and try to re-organize terms to bound the sum

Proof Takeaways and Summary

$$\mathrm{Reg}(T) = \mathcal{O}\left(\sqrt{T \cdot N \cdot \log\left(\frac{NT}{\delta}\right)}\right) + \delta \cdot T$$

Sublinear (sqrt)

Appears linear.. though we can set δ

- By ensuring we are optimistic (i.e. add bonuses such that suboptimal arms look more optimal) and that the optimism decays over time at the right rate, we can get good performance!
- Similar analysis also works for RL, though it is more complicated — but the skeleton is quite similar. Analysis techniques are definitely more complex.

$$\tilde{r} \to \tilde{V}$$

$$T \to \#\ \text{episodes}$$

Part 4: RL with Function Approximation

We have seen that when function approximation is used to represent the Q-function or the policy, there's not any guarantees we can give on convergence and divergence can happen

Under which special cases would RL work with function approximation?

$$Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(s'|s,a), a' \sim \pi}[Q^\pi(s', a')]$$

$$Q(s,a) \approx w^T \phi(s,a) \qquad \exists\ w^*,\ Q^\pi(s,a) = {w^*}^T \phi(s,a)$$

- Policy evaluation using TD-learning: Under **nice** data-distributions, if the linear function class can represent the desired Q-function (**realizability**), then this converges

- If the Q-function for the policy is not expressible in the linear function class, then divergence occurs generally

Remember: this is not saying anything about neural networks

RL with Function Approximation

What about actual online RL?

- Deterministic MDP + linear optimal Q-function (Wen & Van Roy, 2013)

$$\exists\ w^*,\ Q^*(s,a) = {w^*}^T \phi(s,a)$$

- Approximate linear Q^π **<u>for all</u>** π + data-distribution “covers” all policies
  (see concentrability assumption in Munos 20005, Antos et al. 2008)

  polynomial samples with “wide” initial state-distributions or generative model

- Apppproximately linear Q^*: **No! See Du et al. 2020 for counterexamples**

But when the feature representation is “informative” and “compressed enough”, this works! (see Van Roy and Dong, 2019)

…. many more: under “structural assumptions” on the MDP, we can get convergent and efficient algorithms!

Collective Table at: Du, Kakade, Wang, Yang. Is a Good Representation Sufficient for Sample Efficient RL? ICLR 2020

Suggested Readings

- Material taken from the RL Theory Book (Agarwal, Jiang, Kakade, Sun) 2020. https://rltheorybook.github.io/ — one place to find a lot of RL theory material

- Nan Jiang's statistical RL class at UIUC https://nanjiang.cs.illinois.edu/cs598/
  Wen Sun's Foundations of RL class at Cornell https://wensun.github.io/CS6789.html

- **Fitted Q-Iteration**:
  - Munos, 2003. Error Bounds for Approximate Policy Iteration.
  - Munos, 2005. Error Bounds for Approximate Value Iteration
  - Chen and Jiang, 2019. Information Theoretic Considerations in Batch RL.

- **Generative Model:**
  - Azar, Munos, Kappen, 2012. On the Sample Complexity of RL with a Generative Model.

- **Exploration:**
  - Jaksch, Ortner, Auer, 2010. Near-Optimal Regret Bounds for Reinforcement Learning
  - Osband and Van Roy, 2015. Why is Posterior Sampling Better than Optimism for RL? **(aims to answer why posterior sampling (lecture 13) is more desirable)**
  - Azar, Osband, Munos, 2017. Minimax Regret Bounds for RL **(UCB-value iteration)**