Reinforcement Learning Theory Basics

CS 285

Instructor: Sergey Levine
UC Berkeley

What questions do we ask in RL theory?

Lots of different questions! But here are a few common ones:

If I use this algorithm with N samples, k iterations, how good is the result?

Let's say we're doing Q-learning...

$||\hat{Q}_k - Q^\star|| \leq \epsilon$ with probability at least $1-\delta$ if $N \geq f(\epsilon, \delta)$

$||Q^{\pi_k} - Q^\star|| \leq \epsilon$

We'll focus on these types of questions today

not the same thing! Q^{π_k} is the *true* Q-function of policy at iteration k

If I use this exploration algorithm, how high is my regret?

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

But there are many others!

What kinds of assumptions do we make?

Effective analysis is **very** hard in RL without strong assumptions

The trick is to make assumptions that admit interesting conclusions without divorcing us (too much) from reality

Exploration: Performance of RL methods is greatly complicated by exploration – how likely are we to find (potentially sparse) rewards?

Theoretical guarantees typically address worst case performance, and worst case exploration is extremely hard

$$\text{Goal: show that exploration method (e.g., counts) is } \text{Poly}(|S|, |A|, 1/(1-\gamma))$$

Learning: If we somehow “abstract away” exploration, how many samples do we need to effectively learn a model or value function that results in good performance?

$$\text{“generative model” assumption: assume we can sample from } P(s'|s,a) \text{ for } \textit{any } (s,a)$$

$$\text{“oracle exploration”: for every } (s,a)\text{, sample } s' \sim P(s'|s,a) \; N \text{ times}$$

What’s the **point**?

1. Prove that our RL algorithms will work perfectly every time

Usually not possible with current deep RL methods, which are often not even guaranteed to converge

2. Understand how errors are affected by problem parameters

Do larger discounts work better than smaller ones?

If we want half the error, do we need 2x the samples? 4x? something else?

Usually we use precise theory to get imprecise **qualitative** conclusions about how various factors influence the performance of RL algorithms under **strong assumptions**, and try to make the assumptions reasonable enough that these conclusions are likely to apply to real problems (but they are not guaranteed to apply to real problems)

Don’t take someone seriously if they say their RL algorithm has “provable guarantees” – the assumptions are always unrealistic, and theory is at best a rough guide to what might happen

Some basic sample complexity analysis

"oracle exploration": for every (s, a), sample $s' \sim P(s'|s, a)$ N times

simple "model based" algorithm:

1. $\hat{P}(s'|s, a) = \dfrac{\#(s, a, s')}{N}$
2. Given π, use $\hat{P}$ to estimate $\hat{Q}^\pi$

how close is $\hat{Q}^\pi$ to Q^π? ⟵ $||Q^\pi(s, a) - \hat{Q}^\pi(s, a)||_\infty \leq \epsilon$ with probability at least $1 - \delta$ if $N \geq f(\epsilon, \delta)$

$$\max_{s,a} |Q^\pi(s, a) - \hat{Q}^\pi(s, a)| \leq \epsilon$$

good to use $|| \cdot ||_\infty$ if we want *worst-case* performance

how close is $\hat{Q}^\star$ if we learn it using $\hat{P}$? ⟵ $||Q^\star(s, a) - \hat{Q}^\star(s, a)||_\infty \leq \epsilon$

optimal Q-function learned under $\hat{P}$

how good is the resulting policy? ⟵ $||Q^\star(s, a) - Q^{\hat{\pi}}(s, a)||_\infty \leq \epsilon$

the arg max policy corresponding to that Q-function

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

Concentration inequalities

whenever we need to answer questions about how close a learned function is to the true function, in terms of # of samples

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,

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

interpretation:

if we estimate μ with n samples the probability we're off by more than ϵ is at most $2e^{-2n\epsilon^2/(b_+ - b_-)^2}$

equivalently, if we want this probability to be δ:

$$\delta \leq 2e^{-2n\epsilon^2/(b_+ - b_-)^2} \Rightarrow \log \frac{\delta}{2} \leq -2n\epsilon^2/(b_+ - b_-)^2 \Rightarrow \frac{(b_+ - b_-)^2}{2n} \log \frac{2}{\delta} \geq \epsilon^2 \quad \Rightarrow \quad \frac{b_+ - b_-}{\sqrt{2n}} \sqrt{\log \frac{2}{\delta}} \geq \epsilon$$

or... $\quad n \leq \dfrac{(b_+ - b_-)^2}{2\epsilon^2} \log \dfrac{2}{\delta}$

error (ϵ) scales as $\frac{1}{\sqrt{n}}$

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

Concentration inequalities

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

Proposition A.8. *(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 $[\hat{q}]_j = \sum_{i=1}^{N} \mathbf{1}[z_i = j]/N$.*

We have that $\forall \epsilon > 0$:

$$\Pr\left(\|\hat{q} - \vec{q}\|_2 \geq 1/\sqrt{N} + \epsilon\right) \leq e^{-N\epsilon^2}.$$

which implies that:

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

$$\delta \leq e^{-N\epsilon^2} \;\Rightarrow\; \epsilon \leq \frac{1}{\sqrt{N}}\sqrt{\log\frac{1}{\delta}} \qquad ||\hat{P}(s'|s,a) - P(s'|s,a)||_1 \leq \sqrt{|S|}(1/\sqrt{N} + \epsilon) \qquad \text{with prob } 1-\delta$$

$$\Rightarrow\; N \leq \frac{1}{\epsilon^2}\log\frac{1}{\delta} \qquad ||\hat{P}(s'|s,a) - P(s'|s,a)||_1 \leq \sqrt{\frac{|S|}{N}} + \sqrt{\frac{|S|\log 1/\delta}{N}} \leq c\sqrt{\frac{|S|\log 1/\delta}{N}}$$

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

A few useful lemmas

Next goal: relate error in $\hat{P}$ to error in $\hat{Q}^\pi$

Relating P to Q^π:

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

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^\pi(s')$$

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

Q^π: $|S||A|$

r: $|S||A|$

P: $|S||A| \times |S|$

V^π: $|S|$

$$V^\pi = \Pi Q^\pi$$

Π: $|S| \times |S||A|$

$$Q^\pi = r + \gamma P^\pi Q^\pi$$

$$P^\pi = P\Pi$$

$$Q^\pi = r + \gamma P^\pi Q^\pi$$

$$Q^\pi - \gamma P^\pi Q^\pi = r$$

$$(I - \gamma P^\pi) Q^\pi = r$$

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

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

A few useful lemmas

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

Simulation lemma: $Q^\pi - \hat{Q}^\pi = \gamma \underbrace{(I - \gamma \hat{P}^\pi)^{-1}}_{\text{evaluation operator}} \underbrace{(P - \hat{P})}_{\text{difference in probabilities}} V^\pi$

true value

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

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

A few useful lemmas

Another useful lemma: given P^π and *any* vector $v \in \mathbb{R}^{|S||A|}$, we have:

$$||(I - \gamma P^\pi)^{-1} v||_\infty \leq ||v||_\infty / (1 - \gamma)$$

"Q-function" corresponding to "reward" v is at most $1/(1-\gamma)$ times larger

$$\sum_{t=0}^{\infty} \gamma^t c = \frac{c}{1-\gamma}$$

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

$$||v||_\infty = ||(I - \gamma P^\pi) w||_\infty \geq ||w||_\infty - \gamma ||P^\pi w||_\infty \geq ||w||_\infty - \gamma ||w||_\infty = (1-\gamma)||w||_\infty$$

triangle inequality

$$||a - b|| \geq ||a|| - ||b||$$

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

$$||v||_\infty / (1 - \gamma) \geq ||w||_\infty$$

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

Putting them together...

$$||(I - \gamma P^\pi)^{-1} v||_\infty \leq ||v||_\infty / (1 - \gamma)$$

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

can we bound $||V^\pi||_\infty$?

assume $R_{\max} = 1$

$$\sum_{t=0}^{\infty} \gamma^t r_t \leq \sum_{t=0}^{\infty} \gamma^t R_{\max} = \frac{1}{1-\gamma} R_{\max}$$

$$||\hat{P}(s'|s,a) - P(s'|s,a)||_1 \leq c \sqrt{\frac{|S| \log 1/\delta}{N}}$$

$$\begin{aligned}
||Q^\pi - \hat{Q}^\pi||_\infty &= ||\gamma (I - \gamma \hat{P}^\pi)^{-1} (P - \hat{P}) V^\pi||_\infty \\
&\leq \frac{\gamma}{1-\gamma} ||(P - \hat{P}) V^\pi||_\infty \\
&\leq \frac{\gamma}{1-\gamma} \left(\max_{s,a} ||P(\cdot|s,a) - \hat{P}(\cdot|s,a)||_1 \right) ||V^\pi||_\infty \\
&\leq \frac{\gamma}{(1-\gamma)^2} \left(\max_{s,a} ||P(\cdot|s,a) - \hat{P}(\cdot|s,a)||_1 \right) \\
&\leq \frac{\gamma}{(1-\gamma)^2} c_2 \sqrt{\frac{|S| \log 1/\delta}{N}}
\end{aligned}$$

technically need to use the union bound here to account for probabilities

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

What does this mean?

$$||Q^\pi - \hat{Q}^\pi||_\infty \leq \frac{\gamma}{(1-\gamma)^2} c_2 \sqrt{\frac{|S| \log 1/\delta}{N}}$$

more samples = lower error

error grows **quadratically** in the horizon

each backup "accumulates" error

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

Some simple implications...

$$||Q^\pi - \hat{Q}^\pi||_\infty \leq \epsilon \qquad \epsilon = \frac{\gamma}{(1-\gamma)^2} c_2 \sqrt{\frac{|S| \log 1/\delta}{N}}$$

what about $||Q^\star - \hat{Q}^\star||_\infty$?

$$|\sup_x f(x) - \sup_x g(x)| \leq \sup_x |f(x) - g(x)|$$

$$||Q^\star - \hat{Q}^\star||_\infty = ||\sup_\pi Q^\pi - \sup_\pi \hat{Q}^\pi||_\infty \leq \sup_\pi ||Q^\pi - \hat{Q}^\pi||_\infty \leq \epsilon$$

what about $||Q^\star - Q^{\hat{\pi}^\star}||_\infty$?

$$||Q^\star - Q^{\hat{\pi}^\star}||_\infty = ||Q^\star - \hat{Q}^{\hat{\pi}^\star} + \hat{Q}^{\hat{\pi}^\star} - Q^{\hat{\pi}^\star}||_\infty \leq ||Q^\star - \underbrace{\hat{Q}^{\hat{\pi}^\star}}_{\hat{Q}^\star}||_\infty + ||Q^{\hat{\pi}^\star} - \hat{Q}^{\hat{\pi}^\star}||_\infty \leq 2\epsilon$$

same policy

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

What About Model-Free RL?

Analyzing fitted Q-iteration

Bellman operator

$$TQ = r + \gamma P \max_a Q$$

abstract model of exact Q-iteration: $Q_{k+1} \leftarrow TQ_k$

abstract model of approximate fitted Q-iteration: $\hat{Q}_{k+1} \leftarrow \arg\min_{\hat{Q}} ||\hat{Q} - \hat{T}\hat{Q}_k||$

which norm?

no convergence if $||\cdot||_2$

we'll assume $||\cdot||_\infty$

Question: as $k \rightarrow \infty$, $\hat{Q}_k \rightarrow ?$

$$\lim_{k\rightarrow\infty} ||\hat{Q}_k - Q^\star||_\infty \leq ?$$

approximate Bellman operator

$$\hat{T}Q = \hat{r} + \gamma \hat{P} \max_a Q$$

$$\hat{r}(s,a) = \frac{1}{N(s,a)} \sum_i \delta((s_i,a_i) = (s,a)) r_i \qquad \hat{P}(s'|s,a) = \frac{N(s,a,s')}{N(s,a)}$$

where do errors come from?

$T \neq \hat{T}$ "sampling error"

$\hat{Q}_{k+1} \neq \hat{T}\hat{Q}_k$ "approximation error"

Note: these are **not** models, this is the effect of averaging together transitions in the data!

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

Let’s analyze sampling error

$$\hat{Q}_{k+1} \leftarrow \arg\min_{\hat{Q}} ||\hat{Q} - \hat{T}\hat{Q}_k|| \qquad T \neq \hat{T}$$

$$\text{vs. } T\hat{Q}_k$$

$$|\hat{T}Q(s,a) - TQ(s,a)| = |\hat{r}(s,a) - r(s,a) + \gamma(E_{\hat{P}(s'|s,a)}[\max_{a'} Q(s',a')] - E_{P(s'|s,a)}[\max_{a'} Q(s',a')])|$$

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

estimation error of continuous random variable

just use Hoeffding’s inequality directly!

$$|\hat{r}(s,a) - r(s,a)| \leq 2R_{\max}\sqrt{\frac{\log 1/\delta}{2N}}$$

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

$$\leq \sum_{s'} |\hat{P}(s'|s,a) - P(s'|s,a)| \max_{s',a'} Q(s',a')$$

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

$$\leq c||Q||_\infty \sqrt{\frac{\log 1/\delta}{N}}$$

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

Let’s analyze sampling error

$$\hat{Q}_{k+1} \leftarrow \arg\min_{\hat{Q}} ||\hat{Q} - \hat{T}\hat{Q}_k| \qquad T \neq \hat{T}$$

$$\text{vs. } T\hat{Q}_k$$

$$|\hat{T}Q(s,a) - TQ(s,a)| \leq 2R_{\max}\sqrt{\frac{\log 1/\delta}{2N}} + c||Q||_\infty\sqrt{\frac{\log 1/\delta}{N}}$$

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

using union bound

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

Let's analyze approximation error

approximation error assumption: $||\hat{Q}_{k+1} - T\hat{Q}_k||_\infty \leq \epsilon_k$

This is a strong assumption!

we'll analyze the exact backup operator for now,
but we'll come back to approximate backups later!

using fact that $Q^\star$ is fixed point of T

$$\begin{aligned}
||\hat{Q}_k - Q^\star||_\infty &= ||\hat{Q}_k - T\hat{Q}_{k-1} + T\hat{Q}_{k-1} - Q^\star||_\infty \\
&= ||(\hat{Q}_k - T\hat{Q}_{k-1}) + (T\hat{Q}_{k-1} - TQ^\star)||_\infty \\
&\leq ||\hat{Q}_k - T\hat{Q}_{k-1}||_\infty + ||T\hat{Q}_{k-1} - TQ^\star||_\infty \\
&\leq \epsilon_{k-1} + ||T\hat{Q}_{k-1} - TQ^\star||_\infty \\
&\leq \epsilon_{k-1} + \gamma||\hat{Q}_{k-1} - Q^\star||_\infty
\end{aligned}$$

using fact that T is a γ-contraction

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

Let's analyze approximation error

$$
\begin{aligned}
||\hat{Q}_k - Q^\star||_\infty &\leq \epsilon_{k-1} + \gamma||\hat{Q}_{k-1} - Q^\star||_\infty \\
&\leq \epsilon_{k-1} + \gamma\epsilon_{k-2} + \gamma^2||\hat{Q}_{k-2} - Q^\star||_\infty \\
&\leq \epsilon_{k-1} + \gamma\epsilon_{k-2} + \gamma^2\epsilon_{k-3} + \gamma^3||\hat{Q}_{k-2} - Q^\star||_\infty \\
&\leq \sum_{i=0}^{k-1} \gamma^i \epsilon_{k-i-1} + \gamma^k||\hat{Q}_0 - Q^\star||_\infty
\end{aligned}
$$

$$
\lim_{k\to\infty} ||\hat{Q}_k - Q^\star||_\infty \leq \sum_{i=0}^{\infty} \gamma^i \max_k \epsilon_k \; = \frac{1}{1-\gamma}||\epsilon||_\infty
$$

approximation error scales with "horizon"

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

Putting it together

$$||\hat{T}Q - TQ||_\infty \leq 2R_{\max}c_1\sqrt{\frac{\log|S||A|/\delta}{2N}} + c_2||Q||_\infty\sqrt{\frac{\log|S|/\delta}{N}}$$ "sampling error"

$$\lim_{k\to\infty}||\hat{Q}_k - Q^\star||_\infty \leq \frac{1}{1-\gamma}\max_k \epsilon_k = \frac{1}{1-\gamma}\max_k ||\hat{Q}_k - T\hat{Q}_{k-1}||_\infty$$ "approximation error"

how much $\hat{Q}_{k+1}$ differs from $T\hat{Q}_k$

due to: sampling error $(T \neq \hat{T})$

approximation error $(\hat{Q}_k \neq \hat{T}\hat{Q}_{k-1})$

$$||\hat{Q}_k - T\hat{Q}_{k-1}||_\infty = ||\hat{Q}_k - \hat{T}\hat{Q}_{k-1} + \hat{T}\hat{Q}_{k-1} - T\hat{Q}_{k-1}||_\infty$$

$$\leq \underbrace{||\hat{Q}_k - \hat{T}\hat{Q}_{k-1}||_\infty}_{\text{approximation}} + \underbrace{||\hat{T}\hat{Q}_{k-1} - T\hat{Q}_{k-1}||_\infty}_{\text{sampling}}$$

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar

What does it all mean?

$$\mathcal{O}\left(R_{\max}\frac{1}{1-\gamma}\right)$$

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

"sampling error"

$$\lim_{k\to\infty}||\hat{Q}_k - Q^\star||_\infty \leq \frac{1}{1-\gamma}\max_k \epsilon_k = \frac{1}{1-\gamma}\max_k||\hat{Q}_k - T\hat{Q}_{k-1}||_\infty$$

"approximation error"

$$||\hat{Q}_k - T\hat{Q}_{k-1}||_\infty \leq ||\hat{Q}_k - \hat{T}\hat{Q}_{k-1}||_\infty + ||\hat{T}\hat{Q}_{k-1} - T\hat{Q}_{k-1}||_\infty$$

error "compounds" with horizon, over iterations and due to sampling

so far we needed strong (infinity norm) assumptions

more advanced results can be derived with p-norms under some distribution:

$$||\hat{Q}_k - Q^\star||_{p,\mu} = \left(E_{(s,a)\sim\mu(s,a)}[|\hat{Q}_k(s,a) - Q^\star(s,a)|^p]\right)^{1/p}$$

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. https://rltheorybook.github.io & slides by Aviral Kumar