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...

\[ ||\hat{Q}_k - Q^*|| \leq \epsilon \quad \text{with probability at least } 1 - \delta \quad \text{if } N \geq f(\epsilon, \delta) \]

\[ ||Q^{\pi_k} - Q^*|| \leq \epsilon \]

not the same thing! $Q^{\pi_k}$ is the true Q-function of policy at iteration $k$

We’ll focus on these types of questions today

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

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

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.

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?

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

“oracle exploration”: for every \((s,a)\), sample \( s' \sim P(s'|s,a) \) \( N \) 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. \(\hat{P}(s'|s, a) = \frac{\#(s, a, s')}{N}\)

2. Given \(\pi\), use \(\hat{P}\) to estimate \(\hat{Q}^\pi\)

how close is \(\hat{Q}^\pi\) to \(Q^\pi\)?

\[ \|Q^\pi(s, a) - \hat{Q}^\pi(s, a)\|_\infty \leq \epsilon \quad \text{with probability at least } 1 - \delta \]
\[ \max_{s, a} |Q^\pi(s, a) - \hat{Q}^\pi(s, a)| \leq \epsilon \]

good to use \(\| \cdot \|_\infty\) if we want worst-case performance

how close is \(\hat{Q}^*\) if we learn it using \(\hat{P}\)?

\[ \|Q^*(s, a) - \hat{Q}^*(s, a)\|_\infty \leq \epsilon \]

optimal Q-function learned under \(\hat{P}\)

how good is the resulting policy?

\[ \|Q^*(s, a) - \hat{Q}^\pi(s, a)\|_\infty \leq \epsilon \]

the arg max \(\pi\) policy corresponding to that Q-function

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. [https://rltheorybook.github.io](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 \(\mu\). Let \(\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i\). 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 \(\mu\) with \(n\) samples the probability we’re off by more than \(\epsilon\) is at most \(2e^{-2n\epsilon^2/(b_+ - b_-)^2}\)
equivalently, if we want this probability to be \(\delta\):

\[
\delta \leq 2e^{-2n\epsilon^2/(b_+ - b_-)^2} \Rightarrow \log \frac{\delta}{2} \leq -2n\epsilon^2/(b_+ - b_-)^2 \Rightarrow \frac{(b_+ - b_-)^2}{2n} \log \frac{2}{\delta} \geq \epsilon^2 \Rightarrow \frac{b_+ - b_-}{\sqrt{2n}} \sqrt{\log \frac{2}{\delta}} \geq \epsilon
\]
or...

\[
n \leq \frac{(b_+ - b_-)^2}{2\epsilon^2} \log \frac{2}{\delta}
\]

error \((\epsilon)\) scales as \(\frac{1}{\sqrt{n}}\)
Concentration inequalities

\[ \hat{P}(s'|s, a) = \frac{\#(s, a, s')}{N} \]  
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 \( \tilde{q} = [\Pr(z = j)]_{j=1}^d \). Assume we have \( N \) i.i.d. samples, and that our empirical estimate of \( \tilde{q} \) is \( \tilde{q} = \sum_{i=1}^N \mathbb{1}[z_i = j]/N \).

We have that \( \forall \epsilon > 0 \):

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

which implies that:

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

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

\[ \Rightarrow \quad N \leq \frac{1}{\epsilon^2} \log \frac{1}{\delta} \quad \Rightarrow \quad \left\| \hat{P}(s'|s, a) - P(s'|s, a) \right\|_1 \leq \sqrt{\frac{|S|}{N}} + \sqrt{\frac{|S| \log 1/\delta}{N}} \leq c \sqrt{\frac{|S| \log 1/\delta}{N}} \]

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. [https://rltheorybook.github.io](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 $Q^\pi$:

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

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

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

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

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

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

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

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. [https://rltheorybook.github.io](https://rltheorybook.github.io) & slides by Aviral Kumar
A few useful lemmas

\[ Q^\pi = (I - \gamma P^\pi)^{-1}r \quad \hat{Q}^\pi = (I - \gamma \hat{P}^\pi)^{-1}r \]

Simulation lemma:

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

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

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

\[ = (I - \gamma \hat{P}^\pi)^{-1}(I - \gamma \hat{P}^\pi)Q^\pi - (I - \gamma \hat{P}^\pi)^{-1}(I - \gamma P^\pi)Q^\pi \]

\[ = (I - \gamma \hat{P}^\pi)^{-1}((I - \gamma \hat{P}^\pi) - (I - \gamma P^\pi))Q^\pi \]

\[ = \gamma(I - \gamma \hat{P}^\pi)^{-1}(P^\pi - \hat{P}^\pi)Q^\pi \]

\[ = \gamma(I - \gamma \hat{P}^\pi)^{-1}(P\Pi - \hat{P}\Pi)Q^\pi \]

\[ = \gamma(I - \gamma \hat{P}^\pi)^{-1}(P - \hat{P})\Pi Q^\pi \]

\[ = \gamma(I - \gamma \hat{P}^\pi)^{-1}(P - \hat{P})V^\pi \]
A few useful lemmas

Another useful lemma: given $P^\pi$ 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

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

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

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

triangle inequality

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

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

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

Putting them together…

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

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

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

$$
\leq \frac{\gamma}{1 - \gamma} \| (P - \hat{P}) V^\pi \|_\infty
$$

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

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

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

Technically need to use the union bound here to account for probabilities.

What does this mean?

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

error grows \textit{quadratically} in the horizon

each backup “accumulates” error

more samples = lower error

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. [https://rltheorybook.github.io](https://rltheorybook.github.io) & slides by Aviral Kumar
Some simple implications...

\[ \| Q^\pi - \hat{Q}^\pi \|_\infty \leq \epsilon \]

\[ \epsilon = \frac{\gamma}{(1 - \gamma)^2} c_2 \sqrt{\frac{|S| \log 1/\delta}{N}} \]

what about \( \| Q^* - \hat{Q}^* \|_\infty \)?

\[ \| Q^* - \hat{Q}^* \|_\infty = \| \sup_\pi Q^\pi - \sup_\pi \hat{Q}^\pi \|_\infty \leq \sup_\pi \| Q^\pi - \hat{Q}^\pi \|_\infty \leq \epsilon \]

what about \( \| Q^* - \hat{Q}^{\hat{\pi}} \|_\infty \)?

\[ \| Q^* - \hat{Q}^{\hat{\pi}} \|_\infty = \| Q^* - \hat{Q}^{\hat{\pi}} + \hat{Q}^{\hat{\pi}} - \hat{Q}^{\hat{\pi}} \|_\infty \leq \| Q^* - \hat{Q}^{\hat{\pi}} \|_\infty + \| \hat{Q}^{\hat{\pi}} - \hat{Q}^{\hat{\pi}} \|_\infty \leq 2\epsilon \]

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. [https://rltheorybook.github.io](https://rltheorybook.github.io) & slides by Aviral Kumar
What About Model-Free RL?
Analyzing fitted Q-iteration

abstract model of exact Q-iteration:

abstract model of approximate fitted Q-iteration:

Question: as \( k \to \infty \), \( \hat{Q}_k \to ? \)

\[
\lim_{k \to \infty} \| \hat{Q}_k - Q^* \|_\infty \leq ?
\]

where do errors come from?

\( T \neq \hat{T} \) “sampling error”

\( \hat{Q}_{k+1} \neq \hat{T} \hat{Q}_k \) “approximation error”

Bellman operator

\[
TQ = r + \gamma \max_a Q
\]

\[
Q_{k+1} \leftarrow TQ_k
\]

approximate Bellman operator

\[
\hat{T}Q = \hat{r} + \gamma \max_a \hat{Q}
\]

\[
\hat{Q}_{k+1} \leftarrow \arg \min_{\hat{Q}} \| \hat{Q} - \hat{T} \hat{Q}_k \|
\]

which norm?

no convergence if \( \| \cdot \|_2 \)

we’ll assume \( \| \cdot \|_\infty \)

\( \hat{r}(s,a) = \frac{1}{N(s,a)} \sum_i \delta((s_i,a_i) = (s,a)) r_i \)

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

Note: these are not models, this is the effect of averaging together transitions in the data!
Let’s analyze sampling error

\[ \hat{Q}_{k+1} \leftarrow \arg \min_{\hat{Q}} \| \hat{Q} - \hat{T}\hat{Q}_k \| \quad T \neq \hat{T} \]

vs. \( T\hat{Q}_k \)

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

estimation error of continuous random variable
just use Hoeffding’s inequality directly!

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

\[
\sum_{s'}(\hat{P}(s'|s, a) - P(s'|s, a))\max_{a'} Q(s', a')
\leq \sum_{s'} |\hat{P}(s'|s, a) - P(s'|s, a)| s', a' \max_{a'} Q(s', a')
= ||\hat{P}(\cdot |s, a) - P(\cdot |s, a)||_1 ||Q||_\infty
\leq c||Q||_\infty \sqrt{\frac{\log 1/\delta}{N}}
\]

Let’s analyze sampling error

\[
\hat{Q}_{k+1} \leftarrow \arg \min_{Q} |Q - \hat{T}Q_k| \quad \text{vs.} \quad T \hat{Q}_k
\]

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

\[
||\hat{T}Q - TQ||_{\infty} \leq 2R_{\max} c_1 \sqrt{\frac{\log |S||A|/\delta}{2N}} + c_2||Q||_{\infty} \sqrt{\frac{\log |S|/\delta}{N}} \quad \text{using union bound}
\]

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. & 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!

\[
||\hat{Q}_k - Q^*||_\infty = ||\hat{Q}_k - T\hat{Q}_{k-1} + T\hat{Q}_{k-1} - Q^*||_\infty 
= ||(\hat{Q}_k - T\hat{Q}_{k-1}) + (T\hat{Q}_{k-1} - TQ^*)||_\infty 
\leq ||\hat{Q}_k - T\hat{Q}_{k-1}||_\infty + ||T\hat{Q}_{k-1} - TQ^*||_\infty 
\leq \epsilon_{k-1} + ||T\hat{Q}_{k-1} - TQ^*||_\infty 
\leq \epsilon_{k-1} + \gamma||\hat{Q}_{k-1} - Q^*||_\infty
\]

using fact that $Q^*$ is fixed point of $T$

using fact that $T$ is a $\gamma$-contraction

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. [https://rltheorybook.github.io](https://rltheorybook.github.io) & slides by Aviral Kumar
Let’s analyze approximation error

\[ ||\hat{Q}_k - Q^*||_\infty \leq \epsilon_{k-1} + \gamma ||\hat{Q}_{k-1} - Q^*||_\infty \]
\[ \leq \epsilon_{k-1} + \gamma \epsilon_{k-2} + \gamma^2 ||\hat{Q}_{k-2} - Q^*||_\infty \]
\[ \leq \epsilon_{k-1} + \gamma \epsilon_{k-2} + \gamma^2 \epsilon_{k-3} + \gamma^3 ||\hat{Q}_{k-2} - Q^*||_\infty \]
\[ \leq \sum_{i=0}^{k-1} \gamma^i \epsilon_{k-i-1} + \gamma^k ||\hat{Q}_0 - Q^*||_\infty \]

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

approximation error scales with “horizon”

Putting it together

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

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

how much \( \hat{Q}_{k+1} \) differs from \( T\hat{Q}_k \)
due to: sampling error (\( T \neq \hat{T} \))
approximation error (\( \hat{Q}_k \neq \hat{T}\hat{Q}_{k-1} \))

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

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

“sampling error”
“approximation error”

Based on RL Theory Textbook. Agarwal, Jiang, Kakade, Sun. [https://rltheorybook.github.io](https://rltheorybook.github.io) & slides by Aviral Kumar
What does it all mean?

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

“sampling error”

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

“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:

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

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