Bandits and Exploration
(and a few MDPs)

Tor Lattimore
Contents

• What and why of bandit problems
• A little statistics
• How to solve bandit problems
• Scaling up to RL
Bandits

- Reinforcement learning \( S_1, A_1, R_1, S_2, A_2, R_2, \ldots \)
- Bandits \( A_1, R_1, A_2, R_2, \ldots \)

Learning is important
Balancing exploration/exploitation important
No planning
Bandits

Finite action set $\mathcal{A} = \{1, 2, \ldots, k\}$

For each $a \in \mathcal{A}$ there is an **unknown** distribution $P_a$

Learner chooses $A_t \in \mathcal{A}$ and observes **reward** $R_t \sim P_{A_t}$

Learner wants to maximise $\sum_{t=1}^{n} R_t$
The learning objective

Let \( \mu_a \) be the mean of \( P_a \) and \( \mu^* = \max_{a \in A} \mu_a \)

The optimal action is \( a^* = \arg\max_a \mu_a \)

Our task is to minimise the regret

\[
\mathcal{R}_n = n\mu^* - \mathbb{E} \left[ \sum_{t=1}^{n} R_t \right]
\]

The price paid by the learner for not knowing \( \mu \)
A little step into statistics

Given independent and identically distributed

$X, X_1, X_2, \ldots, X_n$ with mean $\mu$ and variance $\sigma^2$

The empirical mean is

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} X_t$$
A little step into statistics

Given independent and identically distributed $X, X_1, X_2, \ldots, X_n$ with mean $\mu$ and variance $\sigma^2$

The empirical mean is $\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} X_t$

What does the distribution of $\mu$ look like?

We know $\mathbb{E}[\hat{\mu}] = \mu$ and $\text{Var}[\hat{\mu}] = \sigma^2 / n$

Chebyshev’s inequality:

$\mathbb{P} (|\hat{\mu} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$
Subgaussian random variables

The **moment generating function** of $X$ is

$$M_X(\lambda) = \mathbb{E}[\exp(\lambda X)]$$

A random variable is **$\sigma$-subgaussian** if

$$M_X(\lambda) \leq \exp(\sigma^2 \lambda^2/2) \quad \text{for all } \lambda \in \mathbb{R}$$

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$X \sim \mathcal{N}(\mu, \sigma^2)$</th>
<th>$X - \mu$ is $\sigma$-subgaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>$X \sim \mathcal{N}(\mu, \sigma^2)$</td>
<td>$X - \mu$ is $\sigma$-subgaussian</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>$X \sim \mathcal{B}(\mu)$</td>
<td>$X - \mu$ is $\frac{1}{2}$-subgaussian</td>
</tr>
</tbody>
</table>
Tail bound for $\sigma$-subgaussian sums:

$$\mathbb{P}(\hat{\mu} - \mu \geq \varepsilon)$$

$$\exp(\lambda (X - \mu)) \leq \exp(\lambda^2 \sigma^2 / 2)$$
Tail bound for $\sigma$-subgaussian sums:

$$
\mathbb{P}(\hat{\mu} - \mu \geq \varepsilon) = \inf_{\lambda > 0} \mathbb{P} \left( \exp(\lambda(\hat{\mu} - \mu)) \geq \exp(\lambda \varepsilon) \right)
$$

$$
\exp(\lambda(X - \mu)) \leq \exp(\lambda^2\sigma^2/2)
$$
Tail bound for $\sigma$-subgaussian sums:

\[
\mathbb{P}(\hat{\mu} - \mu \geq \varepsilon) = \inf_{\lambda > 0} \mathbb{P}(\exp(\lambda(\hat{\mu} - \mu)) \geq \exp(\lambda \varepsilon))
\]
\[
\leq \inf_{\lambda > 0} \exp(-\lambda \varepsilon) \mathbb{E}[\exp(\lambda(\hat{\mu} - \mu))]
\]

\[
\exp(\lambda(X - \mu)) \leq \exp(\lambda^2 \sigma^2/2)
\]

\[
\mathbb{P}(|Z| \geq c) \leq \mathbb{E}[|Z|]/c
\]
Tail bound for $\sigma$-subgaussian sums:

$$\mathbb{P}(\hat{\mu} - \mu \geq \varepsilon) = \inf_{\lambda > 0} \mathbb{P}(\exp(\lambda(\hat{\mu} - \mu)) \geq \exp(\lambda \varepsilon))$$

$$\leq \inf_{\lambda > 0} \exp(-\lambda \varepsilon) \mathbb{E}[\exp(\lambda(\hat{\mu} - \mu))]$$

$$= \inf_{\lambda > 0} \exp(-\lambda \varepsilon) \prod_{t=1}^{n} \mathbb{E}\left[\exp\left(\frac{\lambda(X_t - \mu)}{n}\right)\right]$$

$$\exp(\lambda(X - \mu)) \leq \exp(\lambda^2 \sigma^2 / 2)$$

$$\lambda(\hat{\mu} - \mu) = \sum_{t=1}^{n} \frac{\lambda(X_t - \mu)}{n}$$
Tail bound for $\sigma$-subgaussian sums:

$$\mathbb{P} (\hat{\mu} - \mu \geq \varepsilon) = \inf_{\lambda > 0} \mathbb{P} (\exp (\lambda (\hat{\mu} - \mu)) \geq \exp (\lambda \varepsilon))$$

$$\leq \inf_{\lambda > 0} \exp (-\lambda \varepsilon) \mathbb{E} [\exp (\lambda (\hat{\mu} - \mu))]$$

$$= \inf_{\lambda > 0} \exp (-\lambda \varepsilon) \prod_{t=1}^{n} \mathbb{E} \left[ \exp \left( \frac{\lambda (X_t - \mu)}{n} \right) \right]$$

$$\leq \inf_{\lambda > 0} \exp (-\lambda \varepsilon) \prod_{t=1}^{n} \exp \left( \frac{\sigma^2 \lambda^2}{2n^2} \right)$$

$$\exp (\lambda (X - \mu)) \leq \exp (\lambda^2 \sigma^2 / 2)$$

$$\lambda (\hat{\mu} - \mu) = \sum_{t=1}^{n} \frac{\lambda (X_t - \mu)}{n}$$
Tail bound for $\sigma$-subgaussian sums:

$$
\mathbb{P}(\hat{\mu} - \mu \geq \varepsilon) = \inf_{\lambda > 0} \mathbb{P}(\exp(\lambda(\hat{\mu} - \mu)) \geq \exp(\lambda \varepsilon))
\leq \inf_{\lambda > 0} \exp(-\lambda \varepsilon) \mathbb{E} \left[ \exp(\lambda(\hat{\mu} - \mu)) \right]
= \inf_{\lambda > 0} \exp(-\lambda \varepsilon) \prod_{t=1}^{n} \mathbb{E} \left[ \exp \left( \frac{\lambda(X_t - \mu)}{n} \right) \right]
\leq \inf_{\lambda > 0} \exp(-\lambda \varepsilon) \prod_{t=1}^{n} \exp \left( \frac{\sigma^2 \lambda^2}{2n^2} \right)
= \inf_{\lambda > 0} \exp \left( \frac{\sigma^2 \lambda^2}{2n} - \lambda \varepsilon \right)
$$

$$
\exp(\lambda(X - \mu)) \leq \exp(\lambda^2 \sigma^2 / 2)
$$

$$
\lambda(\hat{\mu} - \mu) = \sum_{t=1}^{n} \frac{\lambda(X_t - \mu)}{n}
$$
Tail bound for $\sigma$-subgaussian sums:

\[
\mathbb{P}(\hat{\mu} - \mu \geq \varepsilon) = \inf_{\lambda > 0} \mathbb{P}(\exp(\lambda(\hat{\mu} - \mu)) \geq \exp(\lambda \varepsilon))
\]

\[
\leq \inf_{\lambda > 0} \exp(-\lambda \varepsilon) \mathbb{E} \left[ \exp(\lambda(\hat{\mu} - \mu)) \right]
\]

\[
= \inf_{\lambda > 0} \exp(-\lambda \varepsilon) \prod_{t=1}^{n} \mathbb{E} \left[ \exp \left( \frac{\lambda(X_t - \mu)}{n} \right) \right]
\]

\[
\leq \inf_{\lambda > 0} \exp(-\lambda \varepsilon) \prod_{t=1}^{n} \exp \left( \frac{\sigma^2 \lambda^2}{2n^2} \right)
\]

\[
= \inf_{\lambda > 0} \exp \left( \frac{\sigma^2 \lambda^2}{2n} - \lambda \varepsilon \right)
\]

\[
\exp(\lambda(X - \mu)) \leq \exp(\lambda^2 \sigma^2 / 2)
\]

\[
0 = \frac{d}{d\lambda} \left( \frac{\sigma^2 \lambda^2}{2n} - \lambda \varepsilon \right) = \lambda \sigma^2 / n - \varepsilon
\]
Tail bound for $\sigma$-subgaussian sums:

$$
\mathbb{P}(\hat{\mu} - \mu \geq \varepsilon) = \inf_{\lambda > 0} \mathbb{P}(\exp(\lambda(\hat{\mu} - \mu)) \geq \exp(\lambda\varepsilon)) \\
\leq \inf_{\lambda > 0} \exp(-\lambda\varepsilon) \mathbb{E}[\exp(\lambda(\hat{\mu} - \mu))] \\
= \inf_{\lambda > 0} \exp(-\lambda\varepsilon) \prod_{t=1}^{n} \mathbb{E}\left[\exp\left(\frac{\lambda(X_t - \mu)}{n}\right)\right] \\
\leq \inf_{\lambda > 0} \exp(-\lambda\varepsilon) \prod_{t=1}^{n} \exp\left(\frac{\sigma^2\lambda^2}{2n^2}\right) \\
= \inf_{\lambda > 0} \exp\left(\frac{\sigma^2\lambda^2}{2n} - \lambda\varepsilon\right) = \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)
$$

$$
\exp(\lambda(X - \mu)) \leq \exp(\lambda^2\sigma^2/2) \\
0 = \frac{d}{d\lambda}\left(\frac{\sigma^2\lambda^2}{2n} - \lambda\varepsilon\right) = \lambda\sigma^2/n - \varepsilon
$$
Last slide we proved that

$$
\mathbb{P} \left( \hat{\mu} - \mu \geq \varepsilon \right) \leq \exp \left( - \frac{n \varepsilon^2}{2 \sigma^2} \right)
$$

Equating the right-hand side with $\delta$ and rearranging things a little,

$$
\mathbb{P} \left( \hat{\mu} - \mu \geq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} \right) \leq \delta
$$

for any $\delta \in (0, 1)$. Chebyshev’s only gives

$$
\mathbb{P} \left( \hat{\mu} - \mu \geq \sqrt{\frac{\sigma^2}{n\delta}} \right) \leq \delta
$$
Concentration of measure summary

Understanding the **distribution** of the **empirical mean** is important.

Without assumptions **Chebyshev’s** is about the best you can do.

**Subgaussian** assumption leads to much stronger results.

Method is called **Chernoff’s method**.

There are whole books on this topic.
Assumptions

We assume $X - \mu_a$ is 1-subgaussian when $X \sim P_a$ for all actions

Subgaussian bandits
Optimism principle

“You should act as if you are in the **nicest plausible** world possible”
Optimism principle

“You should act as if you are in the nicest plausible world possible”

Guarantees either (a) **optimality** or (b) **exploration**
“Nicest” In bandits, we want the mean to be large

“Plausible” The mean cannot be *much* larger than the empirical mean
“Nicest” In bandits, we want the mean to be large

“Plausible” The mean cannot be much larger than the empirical mean

**Upper Confidence Bound Algorithm**
Choose each arm once and then

$$A_t = \text{argmax}_a \hat{\mu}_a(t - 1) + \sqrt{\frac{2 \log(1/\delta)}{T_a(t - 1)}}$$

$$\hat{\mu}_a(t) = \text{empirical mean of arm } a \text{ after round } t$$

$$T_a(t) = \text{number of plays of arm } a \text{ after round } t$$

$$\delta = \text{confidence level}$$
Regret analysis

Step 1  Decompose the regret over the arms

Step 2  On a “good” event prove that suboptimal arms are not played too often

Step 3  Show the “good” event occurs with high probability
Regret decomposition

\[ R_n = n\mu^* - \mathbb{E}\left[ \sum_{t=1}^{n} R_t \right] \]

\[ \Delta_a = \mu^* - \mu_a \]

\[ T_a(t) = \sum_{s=1}^{t} 1(A_s = a) \]
Regret decomposition

$$\mathcal{R}_n = n\mu^* - \mathbb{E} \left[ \sum_{t=1}^{n} R_t \right]$$

$$= \mathbb{E} \left[ \sum_{t=1}^{n} (\mu^* - R_t) \right]$$

$$\Delta_a = \mu^* - \mu_a$$

$$T_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$$
Regret decomposition

\[ R_n = n\mu^* - \mathbb{E} \left[ \sum_{t=1}^{n} R_t \right] \]

\[ = \mathbb{E} \left[ \sum_{t=1}^{n} (\mu^* - R_t) \right] \]

\[ = \mathbb{E} \left[ \sum_{t=1}^{n} \Delta A_t \right] \]
Regret decomposition

\[ R_n = n\mu^* - \mathbb{E} \left[ \sum_{t=1}^{n} R_t \right] \]

\[ = \mathbb{E} \left[ \sum_{t=1}^{n} (\mu^* - R_t) \right] \]

\[ = \mathbb{E} \left[ \sum_{t=1}^{n} \Delta A_t \right] \]

\[ = \mathbb{E} \left[ \sum_{t=1}^{n} \sum_{a \in A} 1(A_t = a) \Delta_a \right] \]

\[ \Delta_a = \mu^* - \mu_a \]

\[ T_a(t) = \sum_{s=1}^{t} 1(A_s = a) \]
Regret decomposition

\[ \mathcal{R}_n = n \mu^* - \mathbb{E} \left[ \sum_{t=1}^{n} R_t \right] \]

\[ = \mathbb{E} \left[ \sum_{t=1}^{n} (\mu^* - R_t) \right] \]

\[ = \mathbb{E} \left[ \sum_{t=1}^{n} \Delta A_t \right] \]

\[ = \mathbb{E} \left[ \sum_{t=1}^{n} \sum_{a \in A} 1(A_t = a) \Delta_a \right] \]

\[ = \sum_{a \in A} \Delta_a \mathbb{E}[T_a(n)] \]

\[ \Delta_a = \mu^* - \mu_a \]

\[ T_a(t) = \sum_{s=1}^{t} 1(A_s = a) \]
Assume for all $t$ that

$$\hat{\mu}_a^*(t - 1) + \sqrt{\frac{2 \log(1/\delta)}{T_a^*(t - 1)}} \geq \mu^*$$

$$\mu_a + \sqrt{\frac{2 \log(1/\delta)}{T_a(t - 1)}} \geq \hat{\mu}_a(t - 1)$$
Assume for all $t$ that

$$\hat{\mu}_a^*(t - 1) + \sqrt{\frac{2 \log(1/\delta)}{T_a^*(t - 1)}} \geq \mu^*$$

$$\mu_a + \sqrt{\frac{2 \log(1/\delta)}{T_a(t - 1)}} \geq \hat{\mu}_a(t - 1)$$

Now suppose that $A_t = a$ in round $t$

$$\mu_a + 2\sqrt{\frac{2 \log(1/\delta)}{T_a(t - 1)}} \geq \hat{\mu}_a(t - 1) + \sqrt{\frac{2 \log(1/\delta)}{T_a(t - 1)}}$$
Assume for all $t$ that

\[ \hat{\mu}_{a^*}(t - 1) + \sqrt{\frac{2 \log(1/\delta)}{T_{a^*}(t - 1)}} \geq \mu^* \]

Now suppose that $A_t = a$ in round $t$

\[ \mu_a + 2 \sqrt{\frac{2 \log(1/\delta)}{T_a(t - 1)}} \geq \hat{\mu}_a(t - 1) + \sqrt{\frac{2 \log(1/\delta)}{T_a(t - 1)}} \]

\[ \geq \hat{\mu}_{a^*}(t - 1) + \sqrt{\frac{2 \log(1/\delta)}{T_{a^*}(t - 1)}} \geq \mu_a^* \]
Assume for all $t$ that

$$\hat{\mu}_{a^*}(t - 1) + \sqrt{\frac{2 \log(1/\delta)}{T_{a^*}(t - 1)}} \geq \mu^*$$

$$\mu_a + \sqrt{\frac{2 \log(1/\delta)}{T_a(t - 1)}} \geq \hat{\mu}_a(t - 1)$$

Now suppose that $A_t = a$ in round $t$

$$\mu_a + 2\sqrt{\frac{2 \log(1/\delta)}{T_a(t - 1)}} \geq \hat{\mu}_a(t - 1) + \sqrt{\frac{2 \log(1/\delta)}{T_a(t - 1)}}$$

$$\geq \hat{\mu}_{a^*}(t - 1) + \sqrt{\frac{2 \log(1/\delta)}{T_{a^*}(t - 1)}} \geq \mu_{a^*}$$

Hence

$$T_a(t - 1) \leq \frac{8 \log(1/\delta)}{\Delta_a^2} \quad \Longrightarrow \quad T_a(n) \leq 1 + \frac{8 \log(1/\delta)}{\Delta_a^2}$$
Let $\hat{\mu}_{a,s}$ be the empirical mean of arm $a$ after $s$ plays.

The concentration theorem shows that

$$\mathbb{P}\left(\hat{\mu}_{a,s} \geq \mu_a + \sqrt{\frac{2 \log(1/\delta)}{s}}\right) \leq \delta$$

Combining with a union bound,

$$\mathbb{P}\left(\text{exists } s \leq n : \hat{\mu}_{a,s} \geq \mu_a + \sqrt{\frac{2 \log(1/\delta)}{s}}\right) \leq n\delta$$
Putting it together

\[ R_n = \sum_{a \in A} \Delta_a \mathbb{E}[T_a(n)] \leq \sum_{a \in A : \Delta_a > 0} \Delta_a \left( 2\delta n^2 + 1 + \frac{8 \log(1/\delta)}{\Delta_a^2} \right) \leq \sum_{a \in A : \Delta_a > 0} 3\Delta_a + \frac{16 \log(n)}{\Delta_a} \]
Sanity checking our results

We have proven the regret of UCB is at most

$$\mathcal{R}_n \leq \sum_{a \in A: \Delta_a > 0} 3\Delta_a + \frac{16 \log(n)}{\Delta_a}$$

Is this good?
\( R_n = \sum_{a \in A} \Delta_a \mathbb{E}[T_a(n)] \)
Problem independent bound

\[ R_n = \sum_{a \in A} \Delta_a \mathbb{E}[T_a(n)] \]

\[ = \sum_{a \in A: \Delta_a \leq \Delta} \Delta_a \mathbb{E}[T_a(n)] + \sum_{a \in A: \Delta_a > \Delta} \Delta_a \mathbb{E}[T_a(n)] \]
Problem independent bound

\[ R_n = \sum_{a \in A} \Delta_a \mathbb{E}[T_a(n)] \]

\[ = \sum_{a \in A : \Delta_a \leq \Delta} \Delta_a \mathbb{E}[T_a(n)] + \sum_{a \in A : \Delta_a > \Delta} \Delta_a \mathbb{E}[T_a(n)] \]

\[ \leq n \Delta + \sum_{a \in A : \Delta_a > \Delta} 3 \Delta_a + \frac{16 \log(n)}{\Delta_a} \]
Problem independent bound

\[ \mathcal{R}_n = \sum_{a \in A} \Delta_a \mathbb{E}[T_a(n)] \]

\[ = \sum_{a \in A: \Delta_a \leq \Delta} \Delta_a \mathbb{E}[T_a(n)] + \sum_{a \in A: \Delta_a > \Delta} \Delta_a \mathbb{E}[T_a(n)] \]

\[ \leq n\Delta + \sum_{a \in A: \Delta_a > \Delta} 3\Delta_a + \frac{16 \log(n)}{\Delta_a} \]

\[ \leq n\Delta + \frac{16K \log(n)}{\Delta} + 3 \sum_{a \in A} \Delta_a \]
Problem independent bound

\[ R_n = \sum_{a \in A} \Delta_a \mathbb{E}[T_a(n)] \]

\[ = \sum_{a \in A: \Delta_a \leq \Delta} \Delta_a \mathbb{E}[T_a(n)] + \sum_{a \in A: \Delta_a > \Delta} \Delta_a \mathbb{E}[T_a(n)] \]

\[ \leq n \Delta + \sum_{a \in A: \Delta_a > \Delta} 3 \Delta_a + \frac{16 \log(n)}{\Delta_a} \]

\[ \leq n \Delta + \frac{16K \log(n)}{\Delta} + 3 \sum_{a \in A} \Delta_a \]

\[ \leq 8 \sqrt{nk \log(n)} + 3 \sum_{a \in A} \Delta_a \leq 8 \sqrt{nk \log(n)} + 3k \]
There is a lot more..

- Improving constants
- Different noise models
- Linear bandits: $\mathcal{A} \subset \mathbb{R}^d$ and $\mu_a = \langle \mu, a \rangle$
- Other kinds of structure: $\mathcal{A} \subset \mathbb{R}^d$ and $\mu_a = f(a)$ with $f$ ‘smooth’
- Changing action sets
- Delayed rewards
- Non-stationary bandits
- Best arm identification
- Adversarial model

Lots of fun still to be had, but this is an RL workshop
Exploration in reinforcement learning ("We want states")
Episodic MDPs

An episodic MDP is a tuple \((S, A, P, H, r, \mu)\)

- \(S\) is a finite set of states
- \(A\) is a finite set of actions
- \(P\) is the transition kernel
- \(H\) is the episode length
- \(r : S \times A \rightarrow [0, 1]\) is the reward function
- \(\mu\) is the distribution of the initial state

Assumption Only \(P\) is unknown
An episodic MDP is a tuple \((S, A, P, H, r, \mu)\)

- \(S\) is a finite set of states
- \(A\) is a finite set of actions
- \(P\) is the transition kernel
- \(H\) is the episode length
- \(r : S \times A \rightarrow [0, 1]\) is the reward function
- \(\mu\) is the distribution of the initial state

**Assumption** Only \(P\) is unknown
$S = \{1, 2, 3\}$ and $H = 4$
A policy $\pi$ is a function from histories to actions.

The value of a policy $\pi$ is

$$v^\pi = \mathbb{E} \left[ \sum_{h=1}^{H} r(S_h, A_h) \right]$$
Dynamic programming

Think of \( P(s, a) = (P(s, a, 1), \ldots, P(s, a, |S|)) \)

The optimal value function is defined inductively

\[
\begin{align*}
\nu_0(s) &= 0 \\
q_h(s, a) &= r(s, a) + \langle P(s, a), \nu_{h-1} \rangle \\
\nu_h(s) &= \max_{a \in A} q_h(s, a) \\
\pi_h(s) &= \arg\max_{a \in A} q_h(s, a)
\end{align*}
\]

\( P = \{ x \in [0, 1]^{|S|} : \|x\|_1 = 1 \} \)
Learning and regret

In each episode the learner chooses a policy $\pi^t$

Observes a trajectory $S^t_1, A^t_1, S^t_2, A^t_2, \ldots, S^t_H, A^t_H$

Regret over $n$ episodes is

$$R_n = \sum_{t=1}^{n} R^{(t)} = \mathbb{E} \left[ \sum_{t=1}^{n} \langle \mu, \nu^*_H - \nu^{\pi^t}_H \rangle \right]$$
Optimism for RL

Same idea!

**Estimate** the things you don’t know (transitions)

Build **confidence intervals** around the unknowns

Act as if the world is as **nice as plausible**
The empirical transitions are given by

\[ T_{s,a}(t) = \# \text{ plays action } a \text{ in state } s \]

\[ \hat{P}_t(s, a, s') = \# \text{ prop. transitions to } s' \text{ from } s \text{ taking } a \]
Estimation and confidence intervals

The empirical transitions are given by

\[ T_{s,a}(t) = \text{# plays action } a \text{ in state } s \]

\[ \hat{P}_t(s, a, s') = \text{# prop. transitions to } s' \text{ from } s \text{ taking } a \]

The confidence set is \( \ell_1 \)-ball about vector \( \hat{P}_t(s, a) \)

\[ C_t(s, a) = \left\{ p \in \mathcal{P} : \| p - \hat{P}_t(s, a) \|_1 \leq \sqrt{\frac{2|S| \log(2/\delta)}{T_{s,a}(t)}} \right\} \]

\[ \mathcal{P} = \{ x \in [0, 1]^{|S|} : \| x \|_1 = 1 \} \]
Optimistic dynamic programming

At the start of phase $t$,

\[
\tilde{v}_0(s) = 0
\]

\[
\tilde{q}_h(s, a) = r(s, a) + \max_{p \in C_{t-1}(s,a)} \langle p, \tilde{v}_{h-1} \rangle
\]

\[
\tilde{v}_h(s) = \max_{a \in A} \tilde{q}_h(s, a)
\]

\[
\pi^t_h(s) = \arg\max_{a \in A} \tilde{q}_h(s, a)
\]

\[
\tilde{P}_h(s) = \arg\max_{p \in C_{t-1}(s,\pi_h(s))} \langle p, \tilde{v}_{h-1} \rangle
\]
UCB for reinforcement learning

Three steps in each episode

Step 1  Compute empirical estimate of transitions and confidence intervals

Step 2  Use optimistic dynamic programming to find a policy

Step 3  Implement policy for entire episode

Algorithm is called Upper Confidence Bounds for Reinforcement Learning (UCRL)
Analysing UCRL

Use optimism

With high probability $P(s, a) \in C_t(s, a)$ for all $t$ and $s, a$
Analysing UCRL

Use optimism

With high probability $P(s, a) \in C_t(s, a)$ for all $t$ and $s, a$

Assuming this holds, then

$$\langle \mu, v_H - v_H^\pi \rangle = \langle \mu, v_H \rangle - \langle \mu, v_H^\pi \rangle$$

$$\leq \langle \mu, \tilde{v}_H^\pi \rangle - \langle \mu, v_H^\pi \rangle$$

$$= \langle \mu, \tilde{v}_H^\pi - v_H^\pi \rangle$$

Useful because it’s much easier to compare values under the same policy
Value differences

Decompose the value difference:

\[ \langle \mu, \tilde{v}^\pi_H - v^\pi_H \rangle = \mathbb{E} \left[ \sum_{h=1}^{H} \langle \tilde{P}^t_{H-h+1}(S^t_h, A^t_h) - P(S^t_h, A^t_h), \tilde{v}^\pi_{H-h} \rangle \right] \]

We might look at the proof later.
Applying Hölder’s inequality

\[ \mathcal{R}(t) \lesssim \mathbb{E} \left[ \sum_{h=1}^{H} \langle \tilde{P}_{H-h+1}(S_h, A_h) - P(S_h, A_h), \tilde{v}_{H-h}^\pi \rangle \right] \]

Hölder’s inequality: \( \langle x, y \rangle \leq \|x\|_1 \|y\|_\infty \)
Applying Hölder’s inequality

\[ \mathcal{R}^{(t)} \leq \mathbb{E} \left[ \sum_{h=1}^{H} \langle \tilde{P}_{H-h+1}(S_h, A_h) - P(S_h, A_h), \tilde{v}^\pi_{H-h} \rangle \right] \]

\[ \leq \mathbb{E} \left[ \sum_{h=1}^{H} \left\| \tilde{P}_{H-h+1}(S_h, A_h) - P(S_h, A_h) \right\|_1 \left\| \tilde{v}^\pi_{H-h} \right\|_\infty \right] \]

Hölder’s inequality: \( \langle x, y \rangle \leq \|x\|_1 \|y\|_\infty \)
Applying Hölder’s inequality

\[ R(t) \lesssim \mathbb{E} \left[ \sum_{h=1}^{H} \left\langle \tilde{P}_{H-h+1}(S_h, A_h) - P(S_h, A_h), \tilde{\nu}_{H-h}^\pi \right\rangle \right] \]

\[ \leq \mathbb{E} \left[ \sum_{h=1}^{H} \left\| \tilde{P}_{H-h+1}(S_h, A_h) - P(S_h, A_h) \right\|_1 \left\| \tilde{\nu}_{H-h}^\pi \right\|_\infty \right] \]

\[ \lesssim H \mathbb{E} \left[ \sum_{h=1}^{H} \sqrt{|S| \log(1/\delta) \over T_{S_h, A_h}(t-1)} \right] \]

Hölder’s inequality: \[ \langle x, y \rangle \leq \| x \|_1 \| y \|_\infty \]
Applying Hölder’s inequality

\[ R^{(t)} \lesssim \mathbb{E} \left[ \sum_{h=1}^{H} \langle \tilde{P}_{H-h+1}(S_h, A_h) - P(S_h, A_h), \tilde{v}_{H-h}^{\pi} \rangle \right] \]

\[ \lesssim \mathbb{E} \left[ \sum_{h=1}^{H} \left\| \tilde{P}_{H-h+1}(S_h, A_h) - P(S_h, A_h) \right\|_1 \left\| \tilde{v}_{H-h}^{\pi} \right\|_{\infty} \right] \]

\[ \lesssim H \mathbb{E} \left[ \sum_{h=1}^{H} \sqrt{|S| \log(1/\delta)} \right] \]

\[ \lesssim H \mathbb{E} \left[ \sum_{s,a} T_{s,a}(t-1, t) \sqrt{\frac{|S| \log(1/\delta)}{T_{s,a}(t-1)}} \right] \]

Hölder’s inequality: \( \langle x, y \rangle \leq \|x\|_1 \|y\|_{\infty} \)
\[
\sum_{t=1}^{n} R^{(t)} \leq H E \left[ \sum_{s,a} \sum_{t=1}^{n} T_{s,a}(t - 1, t) \sqrt{\frac{|S| \log(1/\delta)}{T_{s,a}(t - 1)}} \right]
\]

\[
\leq H E \left[ \sum_{s,a} \sqrt{|S| T_{s,a}(n) \log(1/\delta)} \right]
\]

\[
\leq H E \left[ \sqrt{|S|^2 |A|} \sum_{s,a} T_{s,a}(n) \log(1/\delta) \right]
\]

\[
= H |S| \sqrt{|A| H n \log(1/\delta)}
\]
At last...

With ‘high probability’ the regret of UCRL is

$$\mathcal{R}_n = O \left(|S|H\sqrt{n|A|\log(1/\delta)}\right)$$

**Lower bound** Any algorithm has regret at least

$$\mathcal{R}_n = \Omega \left(H\sqrt{n|A||S|\log(1/\delta)}\right)$$
Takeaways

- A little concentration of measure
- Optimism as a principle for algorithm design
- Optimism for bandits (UCB) and MDPs (UCRL)
Let us reflect for a moment
Let us reflect for a moment

How big is $H \sqrt{n|A||S| \log(1/\delta)}$?
Let us reflect for a moment

How big is \( H \sqrt{n |A||S| \log(1/\delta)} \)?

\(|S| = 2^{20}\)
Let us reflect for a moment

How big is $H \sqrt{n |A| |S|} \log(1/\delta)$?

$|S| = 2^{20}$

Oh 😞
Big challenges

• Exploring in large unstructured MDPs is hopeless
• Combining exploration with function approximation
• Bringing in bias
• Optimism is not universal
• All known exploration principles are either (a) known to be suboptimal or (b) hopelessly intractible
• Model free exploration

Great time to be in RL (theory and practice!)
“Bandit Algorithms” book

Joint work with Csaba Szepesvári

Free online at http://banditalgs.com
Reading


- UCRL. Auer et al. Near-optimal Regret Bounds for Reinforcement Learning, 2010

Useful keywords  Posterior sampling, information directed sampling, Bellman rank, randomized value functions. Preface with ‘deep’ for more buzz
Categorical concentration

Let $X, X_1, X_2, \ldots, X_n$ be independent and identically distributed with $X_t \in [k]$

Let $p_i = \mathbb{P}(X = i)$ and $\hat{p}_i = \frac{1}{n} \sum_{t=1}^{n} 1(X_t = i)$

You can have fun proving that

$$\mathbb{P}\left( \|p - \hat{p}\|_1 \geq \sqrt{\frac{2k \log(2/\delta)}{n}} \right) \leq \delta$$