Bandits and Exploration

(and a few MDPs)

Tor Lattimore

DeepMind

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

S_1 S_2 S_3 S_4 S_5 S_6

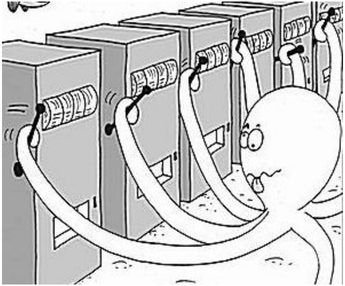
Bandits

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 μ_a be the mean of P_a and $\mu^* = \max_{a \in \mathcal{A}} \mu_a$

The **optimal action** is $a^* = \operatorname{argmax}_a \mu_a$

Our task is to minimise the **regret**

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

The price paid by the learner for not knowing μ

A little step into statistics

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

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

A little step into statistics

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

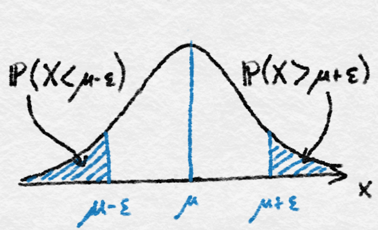
The **empirical mean** is $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$

What does the distribution of μ look like?

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

Chebyshev's inequality:

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



Subgaussian random variables

The **moment generating function** of X is

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

A random variable is $\boldsymbol{\sigma}$**-subgaussian** if

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

Gaussian	$X \sim \mathcal{N}(\mu, \sigma^2)$	$X - \mu$ is σ-subgaussian
Bernoulli	$X \sim \mathcal{B}(\mu)$	$X - \mu$ is $\frac{1}{2}$-subgaussian

Tail bound for σ-subgaussian sums:

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

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

Tail bound for σ-subgaussian sums:

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

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

Tail bound for σ-subgaussian sums:

$$\begin{aligned}\mathbb{P}\left(\hat{\mu} - \mu \geq \varepsilon\right) &= \inf_{\lambda > 0} \mathbb{P}\left(\exp\left(\lambda(\hat{\mu} - \mu)\right) \geq \exp(\lambda\varepsilon)\right) \\ &\leq \inf_{\lambda > 0} \exp(-\lambda\varepsilon)\mathbb{E}\left[\exp\left(\lambda(\hat{\mu} - \mu)\right)\right]\end{aligned}$$

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

$$\mathbb{P}\left(|Z| \geq c\right) \leq \mathbb{E}[|Z|]/c$$

Tail bound for σ-subgaussian sums:

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

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

$$\lambda(\hat{\mu} - \mu) = \sum_{t=1}^{n} \frac{\lambda(X_t - \mu)}{n}$$

Tail bound for σ-subgaussian sums:

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

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

$\lambda(\hat{\mu} - \mu) = \sum_{t=1}^{n} \frac{\lambda(X_t - \mu)}{n}$

Tail bound for σ-subgaussian sums:

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

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

$$\lambda(\hat{\mu} - \mu) = \sum_{t=1}^{n} \frac{\lambda(X_t - \mu)}{n}$$

Tail bound for σ-subgaussian sums:

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

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

$$0 = \frac{d}{d\lambda}\left(\frac{\sigma^2\lambda^2}{2n} - \lambda\varepsilon\right) = \lambda\sigma^2/n - \varepsilon$$

Tail bound for σ-subgaussian sums:

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

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

$$0 = \frac{d}{d\lambda}\left(\frac{\sigma^2\lambda^2}{2n} - \lambda\varepsilon\right) = \lambda\sigma^2/n - \varepsilon$$

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 δ 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

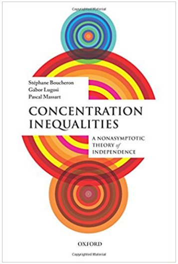
Understanding the **distribution** of the **empirical mean** is important

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

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

$\hat{\mu}_a(t) =$ empirical mean of arm a after round t
$T_a(t) =$ number of plays of arm a after round t
$\delta =$ 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

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

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

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

Regret decomposition

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

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

$$\begin{aligned}
\mathfrak{R}_n &= n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} R_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} (\mu^* - R_t)\right]
\end{aligned}$$

Regret decomposition

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

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

$$\begin{aligned}
\mathfrak{R}_n &= n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} R_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} (\mu^* - R_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}\right]
\end{aligned}$$

Regret decomposition

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

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

$$\begin{aligned}
\mathfrak{R}_n &= n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} R_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n}(\mu^* - R_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} \mathbb{1}(A_t = a)\Delta_a\right]
\end{aligned}$$

Regret decomposition

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

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

$$\begin{aligned}
\mathfrak{R}_n &= n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} R_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} (\mu^* - R_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \mathbb{1}(A_t = a)\Delta_a\right] \\
&= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)]
\end{aligned}$$

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

Now suppose that $A_t = a$ in round t

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

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

$$T_a(t-1) \leq \frac{8\log(1/\delta)}{\Delta_a^2} \implies T_a(n) \leq 1 + \frac{8\log(1/\delta)}{\Delta_a^2}$$

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,

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

$$\mathbb{P}\left(\cup_i B_i\right) \leq \sum_i \mathbb{P}\left(B_i\right)$$

Putting it together

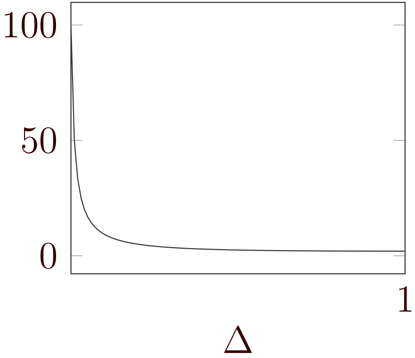
$$\begin{aligned}
\mathfrak{R}_n &= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \\
&\leq \sum_{a \in \mathcal{A}: \Delta_a > 0} \Delta_a \left(2\delta n^2 + 1 + \frac{8 \log(1/\delta)}{\Delta_a^2} \right) \\
&\leq \sum_{a \in \mathcal{A}: \Delta_a > 0} 3\Delta_a + \frac{16 \log(n)}{\Delta_a}
\end{aligned}$$

Sanity checking our results

We have proven the regret of UCB is at most

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

Is this good?



Problem independent bound

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

Problem independent bound

$$\begin{aligned}\mathfrak{R}_n &= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \\ &= \sum_{a \in \mathcal{A}: \Delta_a \leq \Delta} \Delta_a \mathbb{E}[T_a(n)] + \sum_{a \in \mathcal{A} \Delta_a > \Delta} \Delta_a \mathbb{E}[T_a(n)]\end{aligned}$$

Problem independent bound

$$\begin{aligned}
\mathfrak{R}_n &= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \\
&= \sum_{a \in \mathcal{A} : \Delta_a \leq \Delta} \Delta_a \mathbb{E}[T_a(n)] + \sum_{a \in \mathcal{A} \Delta_a > \Delta} \Delta_a \mathbb{E}[T_a(n)] \\
&\leq n\Delta + \sum_{a \in \mathcal{A} : \Delta_a > \Delta} 3\Delta_a + \frac{16 \log(n)}{\Delta_a}
\end{aligned}$$

Problem independent bound

$$\begin{aligned}
\mathfrak{R}_n &= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \\
&= \sum_{a \in \mathcal{A}: \Delta_a \leq \Delta} \Delta_a \mathbb{E}[T_a(n)] + \sum_{a \in \mathcal{A} \Delta_a > \Delta} \Delta_a \mathbb{E}[T_a(n)] \\
&\leq n\Delta + \sum_{a \in \mathcal{A}: \Delta_a > \Delta} 3\Delta_a + \frac{16 \log(n)}{\Delta_a} \\
&\leq n\Delta + \frac{16K \log(n)}{\Delta} + 3 \sum_{a \in \mathcal{A}} \Delta_a
\end{aligned}$$

Problem independent bound

$$
\begin{aligned}
\mathfrak{R}_n &= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \\
&= \sum_{a \in \mathcal{A} : \Delta_a \leq \Delta} \Delta_a \mathbb{E}[T_a(n)] + \sum_{a \in \mathcal{A} \Delta_a > \Delta} \Delta_a \mathbb{E}[T_a(n)] \\
&\leq n\Delta + \sum_{a \in \mathcal{A} : \Delta_a > \Delta} 3\Delta_a + \frac{16 \log(n)}{\Delta_a} \\
&\leq n\Delta + \frac{16K \log(n)}{\Delta} + 3 \sum_{a \in \mathcal{A}} \Delta_a \\
&\leq 8\sqrt{nk \log(n)} + 3 \sum_{a \in \mathcal{A}} \Delta_a \leq 8\sqrt{nk \log(n)} + 3k
\end{aligned}
$$

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 $(\mathcal{S}, \mathcal{A}, P, H, r, \mu)$

- $\mathcal{S}$ is a finite set of **states**
- $\mathcal{A}$ is a finite set of **actions**
- P is the **transition kernel**
- H is the **episode length**
- $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the **reward function**
- μ is the distribution of the initial state

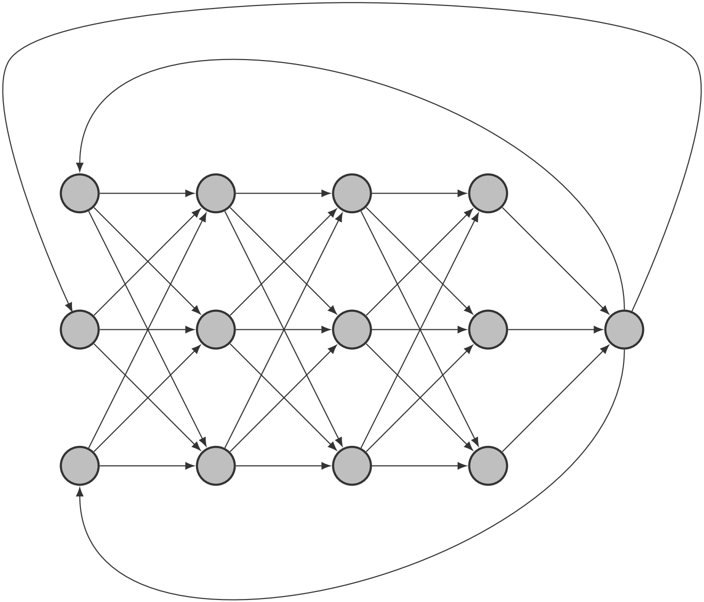
Episodic MDPs

An **episodic MDP** is a tuple $(\mathcal{S}, \mathcal{A}, P, H, r, \mu)$

- $\mathcal{S}$ is a finite set of **states**
- $\mathcal{A}$ is a finite set of **actions**
- P is the **transition kernel**
- H is the **episode length**
- $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the **reward function**
- μ is the distribution of the initial state

Assumption Only P is unknown

$\mathcal{S} = \{1, 2, 3\}$ and $H = 4$



Policies and values

A **policy** π is a function from histories to actions

The **value** of a policy π is

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

Dynamic programming

Think of $P(s,a) = (P(s,a,1), \ldots, P(s,a,|\mathcal{S}|))$

The optimal value function is defined inductively

$$v_0(s) = 0$$

$$q_h(s,a) = r(s,a) + \langle P(s,a), v_{h-1} \rangle$$

$$v_h(s) = \max_{a \in \mathcal{A}} q_h(s,a)$$

$$\pi_h(s) = \operatorname{argmax}_{a \in \mathcal{A}} q_h(s,a)$$

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

Learning and regret

In each episode the learner chooses a policy π^t

Observes a trajectory $S_1^t, A_1^t, S_2^t, A_2^t, \dots, S_H^t, A_H^t$

Regret over n episodes is

$$\mathfrak{R}_n = \sum_{t=1}^n \mathfrak{R}^{(t)} = \mathbb{E}\left[\sum_{t=1}^n \langle \mu, v_H^* - v_H^{\pi^t} \rangle\right]$$

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**

Estimation and confidence intervals

The empirical transitions are given by

$$T_{s,a}(t) = \text{\# plays action } a \text{ in state } s$$
$$\hat{P}_t(s,a,s') = \text{\# prop. transitions to } s' \text{ from } s \text{ taking } a$$

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 ℓ_1-ball about vector $\hat{P}_t(s,a)$

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

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

Optimistic dynamic programming

At the start of phase t,

$$\tilde{v}_0(s) = 0$$

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

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

$$\pi_h^t(s) = \operatorname{argmax}_{a \in \mathcal{A}} \tilde{q}_h(s,a)$$

$$\tilde{P}_h(s) = \operatorname{argmax}_{p \in \mathcal{C}_{t-1}(s, \pi_h(s))} \langle p, \tilde{v}_{h-1} \rangle$$

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 $P(s,a) \in \mathcal{C}_t(s,a)$ for all t and s,a

Analysing UCRL

Use optimism

With high probability $P(s,a) \in \mathcal{C}_t(s,a)$ for all t and s, a

Assuming this holds, then

$$\begin{aligned}\langle \mu, v_H - v_H^{\pi^t} \rangle &= \langle \mu, v_H \rangle - \langle \mu, v_H^{\pi^t} \rangle \\ &\leq \langle \mu, \tilde{v}_H^{\pi^t} \rangle - \langle \mu, v_H^{\pi^t} \rangle \\ &= \langle \mu, \tilde{v}_H^{\pi^t} - v_H^{\pi^t} \rangle\end{aligned}$$

Useful because it's **much** easier to compare values under the same policy

Value differences

Decompose the value difference:

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

We might look at the proof later

Applying Hölder's inequality

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

Hölder's inequality: $\langle x, y \rangle \leq \|x\|_1 \|y\|_\infty$

Applying Hölder's inequality

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

Hölder's inequality: $\langle x, y \rangle \leq \|x\|_1 \|y\|_{\infty}$

Applying Hölder's inequality

$$\begin{aligned}\mathfrak{R}^{(t)} &\lesssim \mathbb{E}\left[\sum_{h=1}^{H}\langle \tilde{P}_{H-h+1}(S_h, A_h) - P(S_h, A_h), \tilde{v}^{\pi}_{H-h}\rangle\right] \\ &\leq \mathbb{E}\left[\sum_{h=1}^{H}\left\|\tilde{P}_{H-h+1}(S_h, A_h) - P(S_h, A_h)\right\|_1 \|\tilde{v}^{\pi}_{H-h}\|_\infty\right] \\ &\lesssim H\mathbb{E}\left[\sum_{h=1}^{H}\sqrt{\frac{|\mathcal{S}|\log(1/\delta)}{T_{S_h,A_h}(t-1)}}\right]\end{aligned}$$

Hölder's inequality: $\langle x, y\rangle \leq \|x\|_1 \|y\|_\infty$

Applying Hölder's inequality

$$
\begin{aligned}
\mathfrak{R}^{(t)} &\lesssim \mathbb{E}\left[\sum_{h=1}^{H} \langle \tilde{P}_{H-h+1}(S_h, A_h) - P(S_h, A_h), \tilde{v}^{\pi}_{H-h} \rangle \right] \\
&\leq \mathbb{E}\left[\sum_{h=1}^{H} \left\| \tilde{P}_{H-h+1}(S_h, A_h) - P(S_h, A_h) \right\|_1 \|\tilde{v}^{\pi}_{H-h}\|_\infty \right] \\
&\lesssim H\mathbb{E}\left[\sum_{h=1}^{H} \sqrt{\frac{|\mathcal{S}| \log(1/\delta)}{T_{S_h, A_h}(t-1)}} \right] \\
&\lesssim H\mathbb{E}\left[\sum_{s,a} T_{s,a}(t-1, t) \sqrt{\frac{|\mathcal{S}| \log(1/\delta)}{T_{s,a}(t-1)}} \right]
\end{aligned}
$$

Hölder's inequality: $\langle x, y \rangle \leq \|x\|_1 \|y\|_\infty$

$$
\begin{aligned}
\sum_{t=1}^{n} \mathfrak{R}^{(t)} &\leq H\mathbb{E}\left[\sum_{s,a}\sum_{t=1}^{n} T_{s,a}(t-1,t)\sqrt{\frac{|\mathcal{S}|\log(1/\delta)}{T_{s,a}(t-1)}}\right] \\
&\lesssim H\mathbb{E}\left[\sum_{s,a}\sqrt{|\mathcal{S}|T_{s,a}(n)\log(1/\delta)}\right] \\
&\leq H\mathbb{E}\left[\sqrt{|\mathcal{S}|^2|\mathcal{A}|\sum_{s,a}T_{s,a}(n)\log(1/\delta)}\right] \\
&= H|\mathcal{S}|\sqrt{|\mathcal{A}|Hn\log(1/\delta)}
\end{aligned}
$$

$$\int \frac{f'(x)}{\sqrt{f(x)}}dx = 2\sqrt{f(x)}$$

At last...

With ‘high probability’ the regret of UCRL is

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

Lower bound Any algorithm has regret at least

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

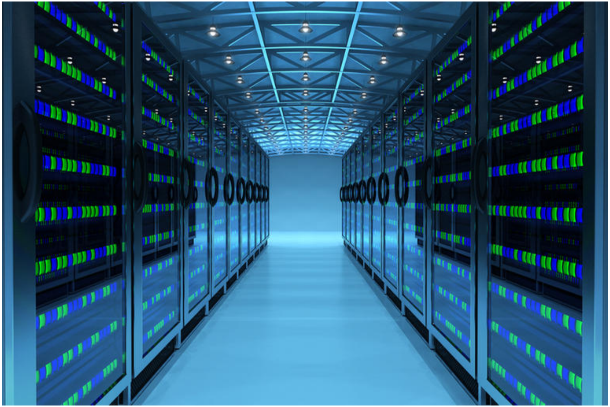
Takeaways

- 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

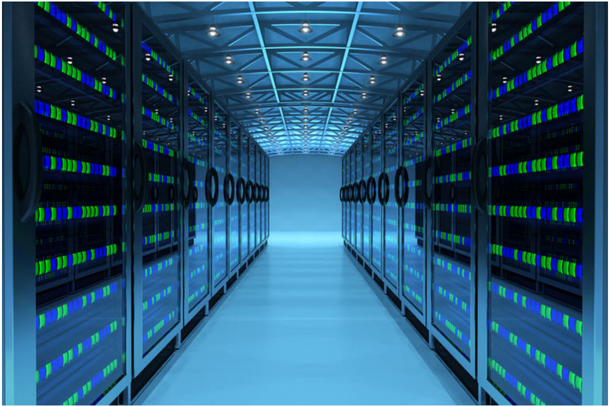
Let us reflect for a moment

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



Let us reflect for a moment

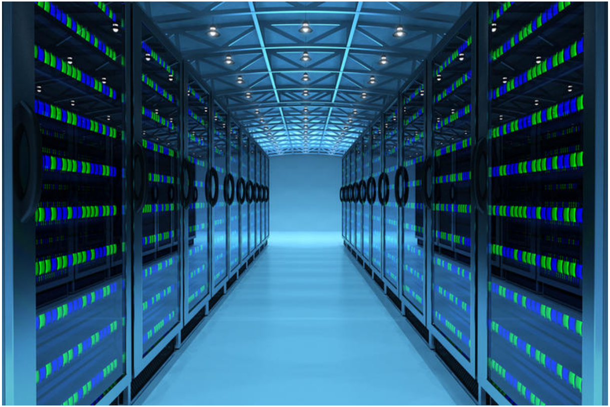
How big is $H\sqrt{n|\mathcal{A}||\mathcal{S}|\log(1/\delta)}$?



$|\mathcal{S}| = 2^{20}$

Let us reflect for a moment

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



$|\mathcal{S}| = 2^{20}$

Oh ☹

Big challenges

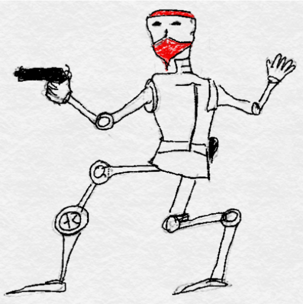
- Exploring in large unstructured MDPs is hopeless
- Combining exploration with function approximation
- Bringing in bias
- Optimism is **not** universal
- All known exploration principles are either (a) known to be **suboptimal** or (b) **hopelessly intractible**
- Model free exploration

Great time to be in RL (theory and practice!)

“Bandit Algorithms” book

Joint work with Csaba Szepesvári

Free online at `http://banditalgs.com`



Reading

- UCB. Tze Leung Lai. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem, 1987
- 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 $\hat{p}_i = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}(X_t = i)$

You can have fun proving that

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