Deep Reinforcement Learning and Control

Bandit Algorithms

Recitation 2

Spring 2022, CMU 10-403

Robin Schmucker

Overview

Focus: Provide an overview of some important bandit algorithms

- Stochastic bandits
- Contextual bandits
- Bayesian bandits

Some references

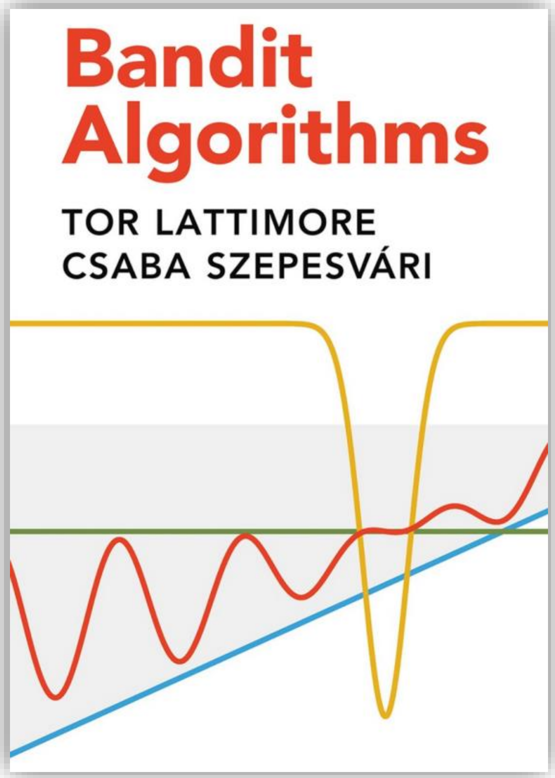
- Sutton & Barto, Chapter 2

Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.

- A comprehensive reference

Lattimore, Tor, and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

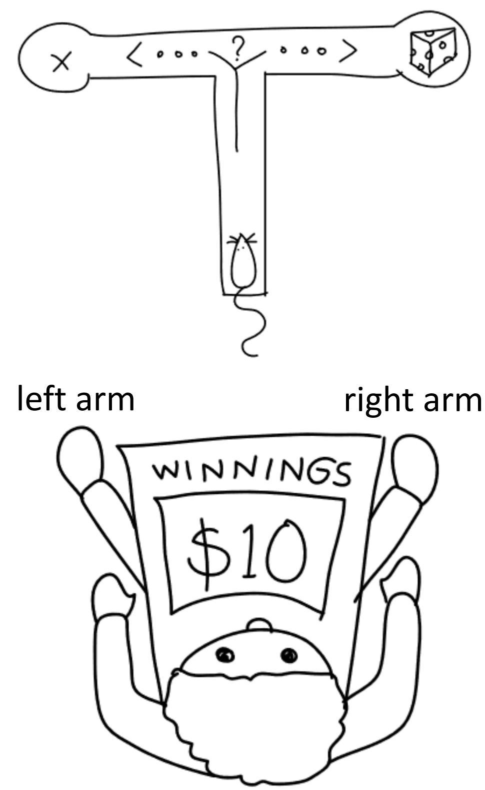
- The bandit framework allows to analyze diverse repeated 1-step interaction problems



What is a Bandit problem?

Sequential game between an *agent* and an *environment*

Round	1	2	3	4	5	6	7	8	9	10
LEFT	0		10	0		0				10
RIGHT		10			0		0	0	0	



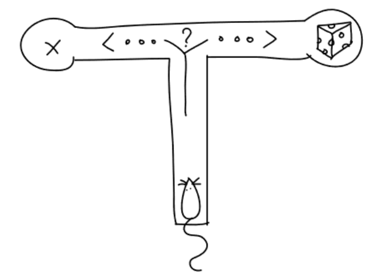
Source: [Lattimore & Szepesvári]

What is a Bandit problem?

Sequential game between a *agent* and an *environment*

In each round $t = 1, ..., n$:

- Agent chooses *action* $A_t \in \mathcal{A}$
- Environment reveals *reward* $X_t \in \mathbb{R}$



History up to time t: $H_t = (A_1, X_1, \ldots, A_{t-1}, X_{t-1})$

Policy: mapping from *history to action*

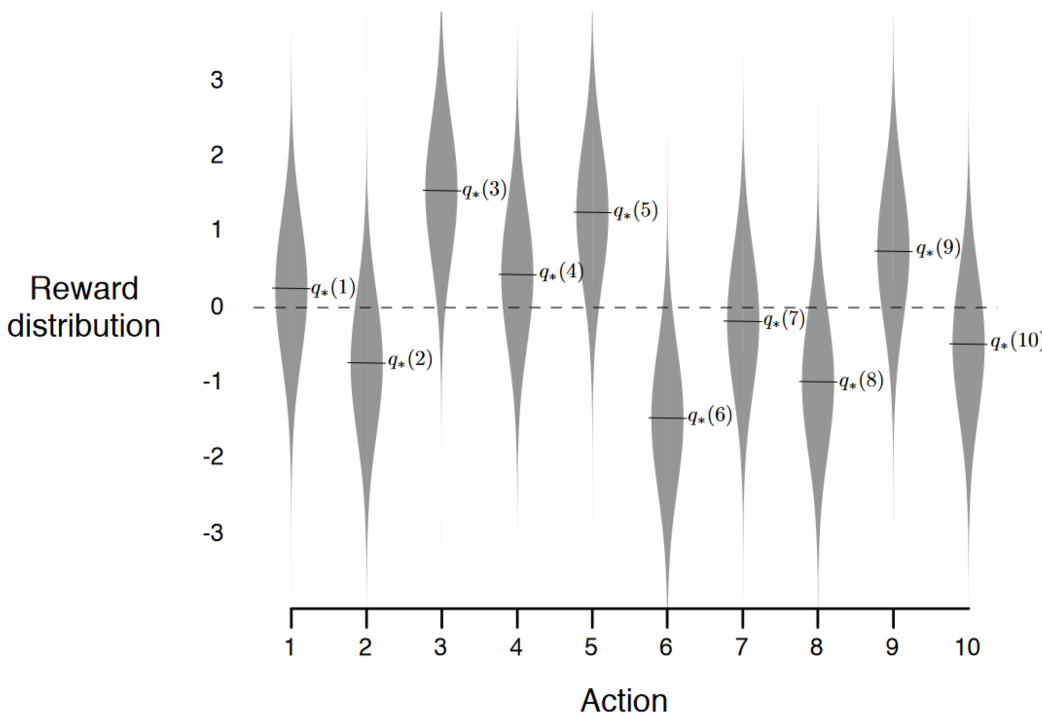
Environment: mapping from *history & action* to *reward*

Env. class: $\mathcal{E}$ describes a family of similar *environments*

Source: [Lattimore & Szepesvári]

Example: Gaussian Bandits

- Each action $A \in \{1, \ldots, k\}$ returns a reward $X \sim \mathcal{N}(\mu_A, \sigma_A^2)$
- At each step, the reward distribution is identical



Source: Sutton

How to evaluate bandit algo.?

DEFINITION 1.1. The regret of the learner relative to a policy π (not necessarily that followed by the learner) is the difference between the total expected reward using policy π for n rounds and the total expected reward collected by the learner over n rounds. The regret relative to a set of policies Π is the maximum regret relative to any policy $\pi \in \Pi$ in the set.

- Total expected reward: $S_n = \mathbb{E}[\sum_{t=1}^{n} X_t]$
- Our total expected reward depends on the policy induced by the bandit algo. and randomness in the environment
- Competitor class Π: a set of policies to benchmark against

Source: Lattimore & Szepesvári

How to evaluate a bandit algo.?

- Regret: $R_n = \max_{\pi \in \Pi} \mathbb{E}_\pi [\sum_{t=1}^{n} X_t] - \mathbb{E}[\sum_{t=1}^{n} X_t]$

- Worst-case regret: Max. regret over all environments in $\mathcal{E}$

- A good bandit algorithm achieves sublinear regret:

$$\lim_{n \to \infty} R_n / n = 0 \quad => \quad R_n = o(n)$$

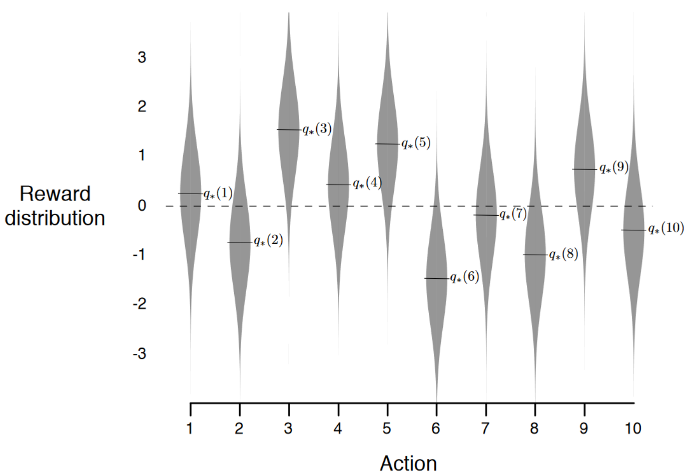
Can we do better ($R_n = O(\sqrt{n}), R_n = O(\log(n))$, ...)?

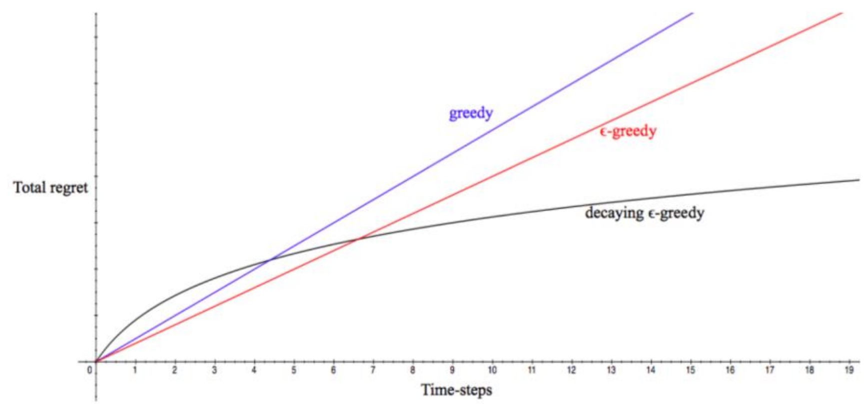
Source: Lattimore & Szepesvári

Example: Gaussian Bandits

- Want to do as well as the optimal action ($\Pi = \{1, ..., k\}$)
- Regret:

$$R_n = n \max_{a \in \mathcal{A}} \mu_a - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$$





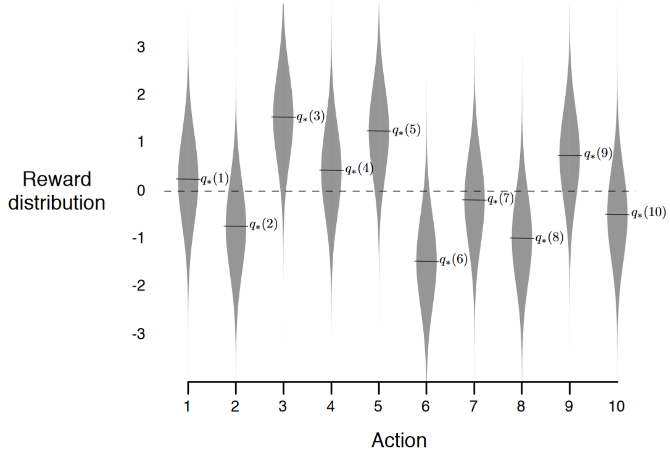
- Question: Why has ε-Greedy linear regret?

Source: [Sutton]

Stochastic Finite Bandits

- Defined via a set of distributions $v = (P_A \,:\, a \in \mathcal{A})$
- Each action $A \in \{1, \ldots, k\}$ returns a reward $X \sim P_A$
- At each step, the reward distribution is identical
- Want to do as well as the best action ($\Pi = \{1, \ldots, k\}$)
- Regret:

$$R_n = n \max_{a \in \mathcal{A}} \mu_a - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$$



Gaussian Bandit

Source: [Lattimore & Szepesvári]

Upper Confidence Bound (UCB)

- Idea: Optimism in the face of uncertainty ☺
- $T_i(t)$: number of times arm i has been sampled
- $\hat{\mu}_i(t)$: sample mean $\hat{\mu}_i(t) = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Assign each i a value which is likely to be an overestimate

$$\mathrm{UCB}_i(t-1,\delta) = \begin{cases} \infty & \text{if } T_i(t-1) = 0 \\ \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} & \text{otherwise}\,. \end{cases}$$

exploration parameter

Source: Lattimore & Szepesvári

Upper Confidence Bound (UCB)

Input k and δ
for $t \in 1, \ldots, n$ **do**
    Choose action $A_t = \operatorname{argmax}_i \mathrm{UCB}_i(t-1, \delta)$
    Observe reward X_t and update upper confidence bounds
end for

Algorithm 3: UCB(δ).

- δ: confidence level that controls degree of exploration

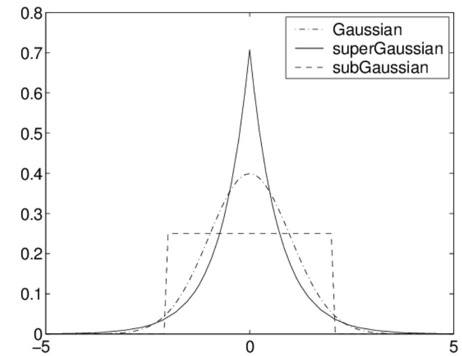
$$\mathrm{UCB}_i(t-1,\delta) = \begin{cases} \infty & \text{if } T_i(t-1) = 0 \\ \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} & \text{otherwise}\,. \end{cases}$$

exploration parameter

Source: [Lattimore & Szepesvári]

UCB: Regret bound

DEFINITION 5.2 (Subgaussianity). A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$, it holds that $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2\sigma^2/2)$.



Source: Sarvotham

-> Tails decay at least as fast as a Gaussian

Useful concentration inequality:

COROLLARY 5.5. *Assume that $X_i - \mu$ are independent, σ-subgaussian random variables. Then for any $\varepsilon \geq 0$,*

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

where $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$.

Source: Lattimore & Szepesvári

UCB: Regret bound

- Suboptimality gap: $\Delta_a = \mu^* - \mu_a$

LEMMA 4.5 (Regret decomposition lemma). *For any policy π and stochastic bandit environment ν with $\mathcal{A}$ finite or countable and horizon $n \in \mathbb{N}$, the regret R_n of policy π in ν satisfies*

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}\left[T_a(n)\right] . \tag{4.5}$$

- Environment class $\mathcal{E}$ of interest: subgaussian distributions

Source: Lattimore & Szepesvári

UCB: Regret bound

THEOREM 7.1. *Consider UCB as shown in Algorithm 3 on a stochastic k-armed 1-subgaussian bandit problem. For any horizon n, if $\delta = 1/n^2$, then*

$$R_n \leq 3 \sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i > 0} \frac{16 \log(n)}{\Delta_i} .$$

- Regret decomposition: $R_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}\left[T_i(n)\right]$
- For each arm i prove that $\mathbb{E}[T_i(t)]$ is small
- Question: When does UCB select arm i?

Source: Lattimore & Szepesvári

UCB: Regret bound

- For i to be selected at least one of the following must hold:

(a) The index of action i is larger than the true mean of a specific optimal arm.
(b) The index of a specific optimal arm is smaller than its true mean.

- Without loss of generality assume $\mu_1 = \mu^*$
- G_i describes an event in which we select A_1 over A_i

(a) (b)

$$G_i = \left\{\mu_1 < \min_{t\in[n]} \text{UCB}_1(t,\delta)\right\} \cap \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\log\left(\frac{1}{\delta}\right)} < \mu_1\right\},$$

Average of observed rewards for arm i

We will choose an $u_i \in \{1,\ldots,n\}$ later

Source: Lattimore & Szepesvári

UCB: Regret bound

Game Plan: We will show two things

1 If G_i occurs, then arm i will be played at most u_i times: $T_i(n) \leq u_i$.

2 The complement event G_i^c occurs with low probability (governed in some way yet to be discovered by u_i).

Because $T_i(n) \leq n$ no matter what, this will mean that

$$\mathbb{E}[T_i(n)] = \mathbb{E}[\mathbb{I}\{G_i\} T_i(n)] + \mathbb{E}[\mathbb{I}\{G_i^c\} T_i(n)] \leq u_i + \mathbb{P}(G_i^c)\, n\,. \tag{7.5}$$

Source: Lattimore & Szepesvári

UCB: Regret bound

1 If G_i occurs, then arm i will be played at most u_i times: $T_i(n) \leq u_i$.

By contradiction: Assume G_i holds and $T_i(n) > u_i$

$$\implies \exists t \in \{1, \ldots, n\} : T_i(t-1) = u_i \wedge A_t = i$$

It follows:

$$\begin{aligned}
\text{UCB}_i(t-1, \delta) &= \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} && \text{(definition of } \text{UCB}_i(t-1,\delta)) \\
&= \hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} && \text{(since } T_i(t-1) = u_i) \\
&< \mu_1 && \text{(definition of } G_i) \\
&< \text{UCB}_1(t-1, \delta)\,. && \text{(definition of } G_i)
\end{aligned}$$

Hence $A_t = \text{argmax}_j \, \text{UCB}_j(t-1, \delta) \neq i$, which is a contradiction. Therefore if G_i occurs, then $T_i(n) \leq u_i$.

Source: Lattimore & Szepesvári

UCB: Regret bound

2 The complement event G_i^c occurs with low probability (governed in some way yet to be discovered by u_i).

(a) (b)

By definition: $$G_i^c = \left\{\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right\} \cup \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1\right\}.$$

Analyze term (a):

$$\mathbb{P}\left(\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right) \leq \mathbb{P}\left(\bigcup_{s\in[n]}\left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\}\right)$$
$$\leq \sum_{s=1}^{n} \mathbb{P}\left(\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right) \leq n\delta\,. \quad (7.7)$$

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

Source: [Lattimore & Szepesvári]

UCB: Regret bound

Analyze term (b): Select u_i large enough s.t. for some $c \in (0,1)$

$$\Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq c\Delta_i$$

With $\mu_i = \mu_i + \Delta_i$:

$$\mathbb{P}\left(\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1\right) = \mathbb{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}}\right)$$

$$\leq \mathbb{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i\right) \leq \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$$

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

Combining (a) and (b): $\mathbb{P}\left(G_i^c\right) \leq n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$

Source: [Lattimore & Szepesvári]

UCB: Regret bound

We showed: $\mathbb{E}\left[T_i(n)\right] = \mathbb{E}\left[\mathbb{I}\left\{G_i\right\} T_i(n)\right] + \mathbb{E}\left[\mathbb{I}\left\{G_i^c\right\} T_i(n)\right] \leq u_i + \mathbb{P}\left(G_i^c\right) n$

$$T_i(n) \leq u_i \qquad \mathbb{P}\left(G_i^c\right) \leq n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$$

It follows: $\mathbb{E}\left[T_i(n)\right] \leq u_i + n\left(n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)\right)$

Set $u_i = \left\lceil \frac{2\log(1/\delta)}{(1-c)^2 \Delta_i^2} \right\rceil$ and $c = 1/2$, it follows:

$$\mathbb{E}[T_i(n)] \leq u_i + 1 + n^{1-2c^2/(1-c)^2} = \left\lceil \frac{2\log(n^2)}{(1-c)^2 \Delta_i^2} \right\rceil + 1 + n^{1-2c^2/(1-c)^2}$$

$$\mathbb{E}\left[T_i(n)\right] \leq 3 + \frac{16\log(n)}{\Delta_i^2}$$

Source: Lattimore & Szepesvári

UCB: Regret bound

Completing the proof by substitution

$$\mathbb{E}[T_i(n)] \leq 3 + \frac{16\log(n)}{\Delta_i^2} \qquad R_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}[T_i(n)]$$

$\square$

THEOREM 7.1. *Consider UCB as shown in Algorithm 3 on a stochastic k-armed 1-subgaussian bandit problem. For any horizon n, if $\delta = 1/n^2$, then*

$$R_n \leq 3\sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i>0} \frac{16\log(n)}{\Delta_i}\,.$$

Problem: Bound is meaningless for small gaps $\Delta_a = \mu^* - \mu_a$

Source: Lattimore & Szepesvári

UCB: Regret bound

THEOREM 7.2. *If $\delta = 1/n^2$, then the regret of UCB, as defined in Algorithm 3, on any $\nu \in \mathcal{E}^k_{\mathrm{SG}}(1)$ environment, is bounded by*

1-subgaussian

$$R_n \leq 8\sqrt{nk\log(n)} + 3\sum_{i=1}^{k} \Delta_i \,.$$

- No inverse relationship to suboptimality gap ☺
- Optimal algorithm for 1-subgaussian up to $\log(n)$ factor

Source: Lattimore & Szepesvári

UCB: Regret bound

Proof Let $\Delta > 0$ be some value to be tuned subsequently, and recall from the proof of Theorem 7.1 that for each suboptimal arm i, we can bound

$$\mathbb{E}[T_i(n)] \leq 3 + \frac{16\log(n)}{\Delta_i^2}\,.$$

Again, relying on the regret decomposition

$$\begin{aligned} R_n &= \sum_{i=1}^{k} \Delta_i \mathbb{E}\left[T_i(n)\right] = \sum_{i:\Delta_i<\Delta} \Delta_i \mathbb{E}\left[T_i(n)\right] + \sum_{i:\Delta_i\geq\Delta} \Delta_i \mathbb{E}\left[T_i(n)\right] \\ &\leq n\Delta + \sum_{i:\Delta_i\geq\Delta}\left(3\Delta_i + \frac{16\log(n)}{\Delta_i}\right) \leq n\Delta + \frac{16k\log(n)}{\Delta} + 3\sum_i \Delta_i \\ &\leq 8\sqrt{nk\log(n)} + 3\sum_{i=1}^{k}\Delta_i\,, \end{aligned}$$

where the first inequality follows because $\sum_{i:\Delta_i<\Delta} T_i(n) \leq n$ and the last line by choosing $\Delta = \sqrt{16k\log(n)/n}$. □

Source: Lattimore & Szepesvári

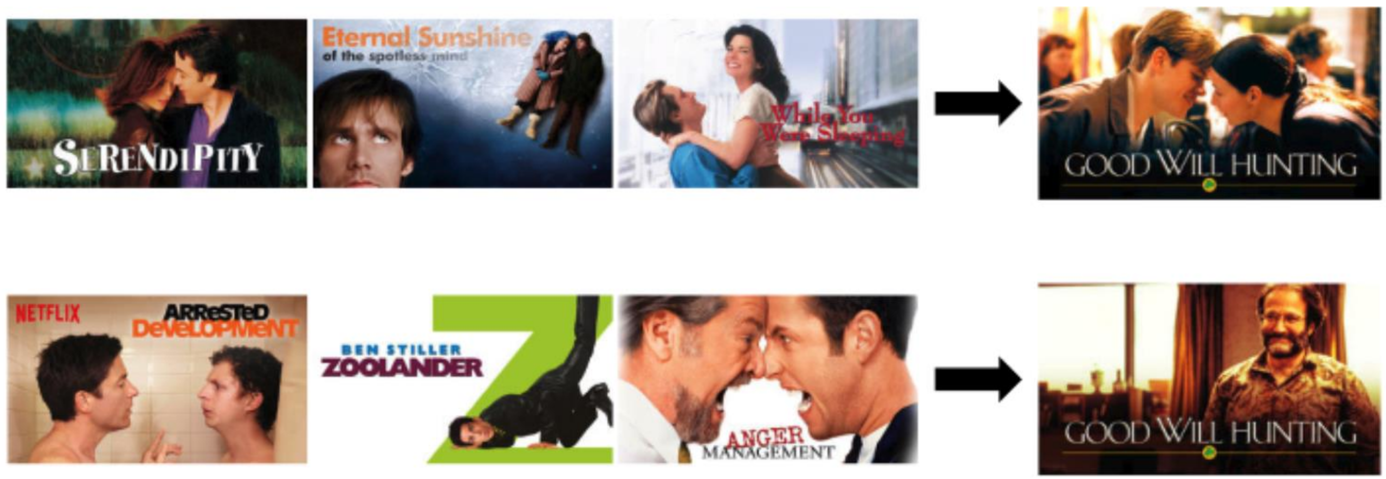
Boltzmann Exploration

- Not covered in class, but similar to UCB
  - Control degree of exploration using temperature param. $\tau \in \mathbb{R}_{\geq 0}$
  - Resembles a "softmax" over action values
  - Stochastic policy

$$p(A_t = a \mid H_t) = \frac{\exp(\tau \hat{q}_{a,t})}{\sum_{a' \in \mathcal{A}} \exp(\tau \hat{q}_{a',t})}$$

- As $\tau \to 0$, converges to uniform random policy
- As $\tau \to \infty$, converges to pure greedy policy
- Recent analysis of convergence properties: Cesa-Bianchi et al.

Example: artwork selection



For a particular title and a particular user, we can use the contextual bandit framework to decide what image to show.

- Context: user attributes, language preferences, previously watched movies, time and day of week, …

Stochastic Contextual Bandits

Context: $C \in \mathcal{C}$ information observed by the agent

Reward function: $r : \mathcal{C} \times \mathcal{A} \to \mathbb{R}$ Noise: $\eta_t \sim P_\eta$

In each round $t = 1, ..., n$:

- Environment determines $C_t \in \mathcal{C}$
- Agent chooses *action* $A_t \in \mathcal{A}$
- Agent receives *reward* $X_t = r(C_t, A_t) + \eta_t$

we will need additional assumptions on reward function for analysis

History up to time t: $H_t = (C_1, A_1, X_1, \ldots, C_{t-1}, A_{t-1}, X_{t-1})$

Regret:

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} \max_{a \in [k]} r(C_t, a) - \sum_{t=1}^{n} X_t\right]$$

Optimal action depends on context

Source: Lattimore & Szepesvári

Bayesian Bandits

- Assume a prior Θ on the parametric reward distribution

  $X_i \sim P(X \mid \theta_i)$ and $\theta_i \sim \Theta$

- Use observed history $D = (A_1, X_1, \ldots, A_{t-1}, X_{t-1})$ to compute posterior using Bayes rule

$$p_t(\theta \mid D) = \frac{p(D \mid \theta) p(\theta)}{p(D)} \propto p(D \mid \theta) \cdot p(\theta)$$

posterior likelihood prior

- Idea: Use posterior to guide the exploration

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370-418

Conjugate Priors

- A *prior* and *model* are called a <u>conjugate pair</u> if the *posterior* has the same parametric form as the prior distribution
- This allows a closed-form expression of posterior

same form

$P(\theta)$ $P(\theta \;\; D)$

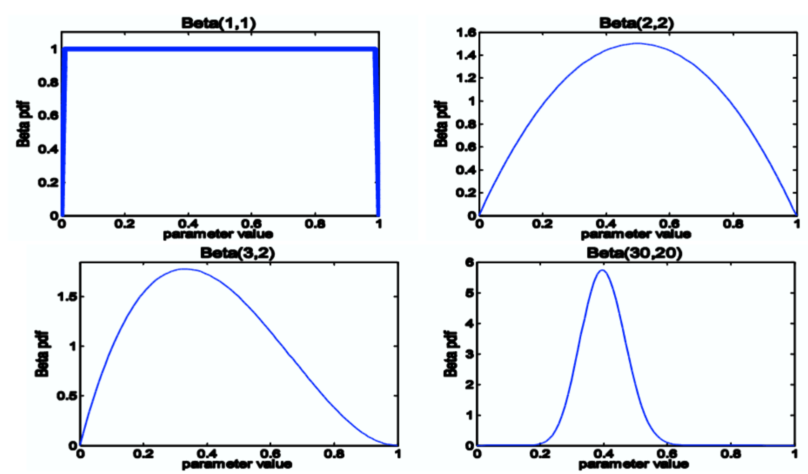
prior posterior

- Example: The beta distribution is a <u>conjugate prior</u> for the Bernoulli distribution

Assume $\theta \sim \mathrm{Beta}(\beta_H, \beta_T)$

I.e., $P(\theta) = \dfrac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)}$

More concentrated as values of β_H, β_T increase



Source: Balcan

θ biased coin -- each arm can be though of as different coin

$$a_H = \sum X_i$$

$$a_T = \sum (1 - X_i)$$

Conjugate Priors

Assume $\theta \sim \mathrm{Beta}(\beta_H, \beta_T)$ I.e., $P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-}}{B(\beta_H,\beta_T)}$

Likelihood function $P(D|\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$ (Binomial)

Posterior: $P(\theta|D) \propto P(D|\theta)P(\theta)$

$$P(\theta|D) \propto \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1} \sim \mathrm{Beta}(\alpha_H+\beta_H, \alpha_T+\beta_T)$$

Interpretation: like MLE, but *hallucinating* $\beta_H - 1$ additional heads & $\beta_T - 1$ additional tails

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{(\alpha_T+\beta_T-1) + (\alpha_H+\beta_H-1)}$$

Note: as we get more sample effect of prior washed out.

Source: Balcan

Thompson Sampling

Explores based on posterior reward distribution

In each round $t = 1, ..., n$:

- For $A \in \mathcal{A}$ agent samples $\theta_{A,t} \sim P(\theta_A \mid D_t)$
- Agent selects $A_t \in \arg\max_{A \in \mathcal{A}} \mathbb{E}_{\theta_{A,t}}[X_A] = \arg\max_{A \in \mathcal{A}} \mu_{\theta_{A,t}}$
- Agent observes reward
- Agent updates posterior distribution

Regret analysis: Agrawal & Goyal

Questions?