Multi-Armed Bandit model: general setting

\( K \) arms:

for \( a \in \{1, \ldots, K\} \), \((X_{a,t})_{t \in \mathbb{N}}\) is a stochastic process.

\((\text{unknown distributions})\)

\textbf{Bandit game:} a each round \( t \), an agent

- chooses an arm \( A_t \in \{1, \ldots, K\} \)
- receives a reward \( X_t = X_{A_t,t} \)

\textbf{Goal:} Build a sequential strategy

\[ A_t = F_t(A_1, X_1, \ldots, A_{t-1}, X_{t-1}) \]

maximizing

\[ \mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha_t X_t \right], \]

where \((\alpha_t)_{t \in \mathbb{N}}\) is a discount sequence. [Berry and Fristedt, 1985]
Multi-Armed Bandit model: the i.i.d. case

K independent arms:

for \( a \in \{1, \ldots, K\} \), \((X_{a,t})_{t \in \mathbb{N}}\) is i.i.d. \( \sim \nu_a \)

(unknown distributions)

Bandit game: a each round \( t \), an agent

- chooses an arm \( A_t \in \{1, \ldots, K\} \)
- receives a reward \( X_t \sim \nu_{A_t} \)

Goal: Build a sequential strategy

\[ A_t = F_t(A_1, X_1, \ldots, A_{t-1}, X_{t-1}) \]

maximizing

<table>
<thead>
<tr>
<th>Finite-horizon MAB</th>
<th>Discounted MAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbb{E} \left[ \sum_{t=1}^{T} X_t \right] )</td>
<td>( \mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha^{t-1} X_t \right] )</td>
</tr>
</tbody>
</table>
Why MABs?

**Goal:** maximize ones’ gains in a casino? (HOPELESS)
Why MABs? Real motivations

Clinical trials:

\[ \mathcal{B}(\mu_1) \quad \mathcal{B}(\mu_2) \quad \mathcal{B}(\mu_3) \quad \mathcal{B}(\mu_4) \quad \mathcal{B}(\mu_5) \]

- choose a treatment \( A_t \) for patient \( t \)
- observe a response \( X_t \in \{0, 1\} : \mathbb{P}(X_t = 1) = \mu_{A_t} \)
- **Goal**: maximize the number of patient healed

Recommendation tasks:

- recommend a movie \( A_t \) for visitor \( t \)
- observe a rating \( X_t \sim \nu_{A_t} \) (e.g. \( X_t \in \{1, \ldots, 5\} \))
- **Goal**: maximize the sum of ratings
Bernoulli bandit models

$K$ independent arms:

for $a \in \{1, \ldots, K\}$, $(X_{a,t})_{t \in \mathbb{N}}$ is i.i.d. $\sim \mathcal{B}(\mu_a)$

Bandit game: at each round $t$, an agent
- chooses an arm $A_t \in \{1, \ldots, K\}$
- receives a reward $X_t \sim \mathcal{B}(\mu_{A_t}) \in \{0, 1\}$

Goal: maximize

<table>
<thead>
<tr>
<th>Finite-horizon MAB</th>
<th>Discounted MAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{E} \sum_{t=1}^{T} X_t$</td>
<td>$\mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha^{t-1} X_t \right]$</td>
</tr>
</tbody>
</table>
Bernoulli bandit models

\( K \) independent arms:

\[ \text{for } a \in \{1, \ldots, K\}, \quad (X_{a,t})_{t \in \mathbb{N}} \text{ is i.i.d } \sim \mathcal{B}(\mu_a) \]

**Bandit game**: at each round \( t \), an agent
- chooses an arm \( A_t \in \{1, \ldots, K\} \)
- receives a reward \( X_t \sim \mathcal{B}(\mu_{A_t}) \in \{0, 1\} \)

**Goal**: maximize

<table>
<thead>
<tr>
<th>Finite-horizon MAB</th>
<th>Discounted MAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbb{E} \left[ \sum_{t=1}^{T} X_t \right] )</td>
<td>( \mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha^{t-1} X_t \right] )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Frequentist model</strong></th>
<th><strong>Bayesian model</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_1, \ldots, \mu_K ) unknown parameters</td>
<td>( \mu_1, \ldots, \mu_K ) drawn from a prior distribution: ( \mu_a \sim \pi_a )</td>
</tr>
<tr>
<td>arm ( a ): ( (X_{a,t})_t \text{ i.i.d. } \sim \mathcal{B}(\mu_a) )</td>
<td>arm ( a ): ( (X_{a,t})_t</td>
</tr>
</tbody>
</table>
Outline

1. Bayesian bandits: a planning problem
2. Frequentist bandits: asymptotically optimal algorithms
3. Non stochastic bandits: minimax algorithms
Outline

1 Bayesian bandits: a planning problem
2 Frequentist bandits: asymptotically optimal algorithms
3 Non stochastic bandits: minimax algorithms
A Markov Decision Process

Bandit model ($B(\mu_1), \ldots, B(\mu_K)$)

- Prior distribution: $\mu_a \sim_{\text{i.i.d}} \mathcal{U}([0, 1])$
- Posterior distribution: $\pi^t_a := \mathcal{L}(\mu_a | X_1, \ldots, X_t)$

$$\pi^t_a = \text{Beta}\left(S_a(t) + 1, N_a(t) - S_a(t) + 1\right)$$

$S_a(t)$: sum of the rewards gathered from arm $a$ up to time $t$
$N_a(t)$: number of draws of arm $a$ up to time $t$

State $\Pi^t = (\pi^t_a)_{a=1}^K$ that evolves in a MDP.
A Markov Decision Process

An example of transition:

\[
\begin{pmatrix}
1 & 2 \\
5 & 1 \\
0 & 2 \\
\end{pmatrix}
\xrightarrow{A_t=2}
\begin{pmatrix}
1 & 2 \\
6 & 1 \\
0 & 2 \\
\end{pmatrix}
\text{if } X_t = 1
\]

Solving a planning problem: there exists an exact solution to

- The finite-horizon MAB:
  \[
  \arg \max_{(A_t)} \mathbb{E}_{\mu \sim \pi} \left[ \sum_{t=1}^{T} X_t \right]
  \]

- The discounted MAB:
  \[
  \arg \max_{(A_t)} \mathbb{E}_{\mu \sim \pi} \left[ \sum_{t=1}^{\infty} \alpha^{t-1} X_t \right]
  \]

Optimal policy = solution to dynamic programming equations.
A Markov Decision Process

An example of transition:

\[
\begin{pmatrix}
1 & 2 \\
5 & 1 \\
0 & 2
\end{pmatrix}
\rightarrow
\begin{pmatrix}
1 & 2 \\
6 & 1 \\
0 & 2
\end{pmatrix}
\text{if } X_t = 1
\]

Solving a planning problem: there exists an exact solution to

- The finite-horizon MAB:

\[
\arg \max_{(A_t)} \mathbb{E}_{\mu \sim \pi} \left[ \sum_{t=1}^{T} X_t \right]
\]

- The discounted MAB:

\[
\arg \max_{(A_t)} \mathbb{E}_{\mu \sim \pi} \left[ \sum_{t=1}^{\infty} \alpha^{t-1} X_t \right]
\]

Optimal policy = solution to dynamic programming equations.

Problem: The state space is too large!
A reduction of the dimension

[Gittins 79]: the solution of the discounted MAB reduces to an index policy:

\[ A_{t+1} = \arg\max_{a=1 \ldots K} G_\alpha(\pi_a^t). \]

- The Gittins indices:

\[
G_\alpha(p) = \sup_{\text{stopping times } \tau>0} \frac{\mathbb{E}_{Y_t \sim B(\mu)} \left[ \sum_{t=1}^{\tau} \alpha^{t-1} Y_t \right]}{\mathbb{E}_{Y_t \sim B(\mu)} \left[ \sum_{t=1}^{\tau} \alpha^{t-1} \right]}
\]

"instantaneous rewards when committing to arm \( \mu \sim p \), when rewards are discounted by \( \alpha \)"
An alternative formulation:

\[ G_\alpha(p) = \inf\{ \lambda \in \mathbb{R} : V_\alpha^*(p, \lambda) = 0 \}, \]

with

\[ V_\alpha^*(p, \lambda) = \sup_{\text{stopping times } \tau > 0} \mathbb{E}_{Y_t \sim B(\mu) \mu \sim p} \left[ \sum_{t=1}^{\tau} \alpha^{t-1} (Y_t - \lambda) \right]. \]

“price worth paying for committing to arm \( \mu \sim p \) when rewards are discounted by \( \alpha \)”
Gittins indices for finite horizon

The Finite-Horizon Gittins indices: depend on the remaining time to play $r$

$$G(p, r) = \inf\{ \lambda \in \mathbb{R} : V_r^*(p, \lambda) = 0 \},$$

with

$$V_r^*(p, \lambda) = \sup_{0 < \tau \leq r} \mathbb{E}_{\substack{\tau \text{ stopping times} \\ \mu \sim p}} \left[ \sum_{t=1}^{\tau} (Y_t - \lambda) \right].$$

“price worth paying for playing arm $\mu \sim p$ for at most $r$ rounds”

The Finite-Horizon Gittins algorithm

$$A_{t+1} = \arg\max_{a=1 \ldots K} G(\pi_a^t, T - t)$$

does NOT coincide with the optimal solution [Berry and Fristedt 85]... but is conjectured to be a good approximation!
Outline

1. Bayesian bandits: a planning problem
2. Frequentist bandits: asymptotically optimal algorithms
3. Non stochastic bandits: minimax algorithms
Regret minimization

\[ \mu = (\mu_1, \ldots, \mu_K) \] unknown parameters, \( \mu^* = \max_a \mu_a. \)

- The **regret** of a strategy \( A = (A_t) \) is defined as

\[
R_\mu(A, T) = \mathbb{E}_\mu \left[ \mu^* T - \sum_{t=1}^T X_t \right]
\]

and can be rewritten

\[
R_\mu(A, T) = \sum_{a=1}^K (\mu^* - \mu_a) \mathbb{E}_\mu [N_a(T)].
\]

\( N_a(t) \) : number of draws of arm \( a \) up to time \( t \)

**Maximizing rewards \( \Leftrightarrow \) Minimizing regret**

**Goal:** Design strategies that have small regret for all \( \mu. \)
Optimal algorithms for regret minimization

All the arms should be drawn infinitely often!

- [Lai and Robbins, 1985]: a uniformly efficient strategy $(\forall \mu, \forall \alpha \in [0, 1[, R_\mu(A, T) = o(T^\alpha))$ satisfies

$$\mu_a < \mu^* \Rightarrow \liminf_{T \to \infty} \frac{\mathbb{E}_\mu[N_a(T)]}{\log T} \geq \frac{1}{d(\mu_a, \mu^*)},$$

where

$$d(\mu, \mu') = \text{KL}(B(\mu), B(\mu'))$$

$$= \mu \log \frac{\mu}{\mu'} + (1 - \mu) \log \frac{1 - \mu}{1 - \mu'}.$$

Definition

A bandit algorithm is **asymptotically optimal** if, for every $\mu$,

$$\mu_a < \mu^* \Rightarrow \limsup_{T \to \infty} \frac{\mathbb{E}_\mu[N_a(T)]}{\log T} \leq \frac{1}{d(\mu_a, \mu^*)}.$$
First algorithms

- **Idea 1**: Draw each arm $T/K$ times

  $\Rightarrow$ EXPLORATION
First algorithms

- **Idea 1**: Draw each arm $T/K$ times
  ⇒ **EXPLORATION**

- **Idea 2**: Always choose the empirical best arm:
  $$A_{t+1} = \arg\max_a \hat{\mu}_a(t)$$
  ⇒ **EXPLOITATION**
First algorithms

- **Idea 1**: Draw each arm \( T/K \) times
  \[ \Rightarrow \text{EXPLORATION} \]

- **Idea 2**: Always choose the empirical best arm:
  \[ A_{t+1} = \arg\max_a \hat{\mu}_a(t) \]
  \[ \Rightarrow \text{EXPLOITATION} \]

- **Idea 3**: Draw the arms uniformly during \( T/2 \) rounds, then draw the empirical best until the end
  \[ \Rightarrow \text{EXPLORATION followed EXPLOITATION} \]
Idea 1: Draw each arm $T/K$ times

⇒ EXPLORATION

Idea 2: Always choose the empirical best arm:

$$A_{t+1} = \arg\max_a \hat{\mu}_a(t)$$

⇒ EXPLOITATION

Idea 3: Draw the arms uniformly during $T/2$ rounds, then draw the empirical best until the end

⇒ EXPLORATION followed EXPLOITATION

Linear regret...
For each arm $a$, build a confidence interval on $\mu_a$:

$$\mu_a \leq \text{UCB}_a(t) \quad \text{w.h.p}$$

Figure: Confidence intervals on the arms at round $t$

Optimism principle:

“act as if the best possible model were the true model”

$$A_{t+1} = \arg \max_a \text{UCB}_a(t)$$
A UCB algorithm in action!
The UCB1 algorithm

UCB1 [Auer et al. 02] is based on the index

\[ UCB_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{\alpha \log(t)}{2N_a(t)}} \]

- Hoeffding’s inequality:
  \[
  \mathbb{P}\left( \hat{\mu}_{a,s} + \sqrt{\frac{\alpha \log(t)}{2s}} \leq \mu_a \right) \leq \exp\left( -2s \left( \frac{\alpha \log(t)}{2s} \right) \right) = \frac{1}{t^\alpha}. 
  \]

- Union bound:
  \[
  \mathbb{P}(UCB_a(t) \leq \mu_a) \leq \mathbb{P}\left( \exists s \leq t : \hat{\mu}_{a,s} + \sqrt{\frac{\alpha \log(t)}{2s}} \leq \mu_a \right) \leq \sum_{s=1}^{t} \frac{1}{t^\alpha} = \frac{1}{t^{\alpha-1}}. 
  \]
Theorem

For every $\alpha > 2$ and every sub-optimal arm $a$, there exists a constant $C_\alpha > 0$ such that

$$\mathbb{E}_\mu [N_a(T)] \leq \frac{2\alpha}{(\mu^* - \mu_a)^2} \log(T) + C_\alpha.$$
The UCB1 algorithm

**Theorem**

For every $\alpha > 2$ and every sub-optimal arm $a$, there exists a constant $C_\alpha > 0$ such that

$$\mathbb{E}_\mu[N_a(T)] \leq \frac{2\alpha}{(\mu^* - \mu_a)^2} \log(T) + C_\alpha.$$ 

**Remark:**

$$\frac{2\alpha}{(\mu^* - \mu_a)^2} > 4\alpha \frac{1}{d(\mu_a, \mu^*)}$$

(UCB1 not asymptotically optimal)
Assume $\mu^* = \mu_1$ and $\mu_2 < \mu_1$.

$$N_2(T) = \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1}=2)$$

$$= \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1}=2) \cap (\text{UCB}_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1}=2) \cap (\text{UCB}_1(t) > \mu_1)$$

$$\leq \sum_{t=0}^{T-1} \mathbb{1}(\text{UCB}_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1}=2) \cap (\text{UCB}_2(t) > \mu_1)$$
Proof: 1/3

Assume $\mu^* = \mu_1$ and $\mu_2 < \mu_1$.

\[
N_2(T) = \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = 2)
\]

\[
= \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = 2) \cap (\text{UCB}_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = 2) \cap (\text{UCB}_1(t) > \mu_1)
\]

\[
\leq \sum_{t=0}^{T-1} \mathbb{1}(\text{UCB}_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = 2) \cap (\text{UCB}_2(t) > \mu_1)
\]

\[
\mathbb{E}[N_2(T)] \leq \sum_{t=0}^{T-1} \mathbb{P}(\text{UCB}_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \text{UCB}_2(t) > \mu_1)
\]

\begin{align*}
&\underbrace{\sum_{t=0}^{T-1} \mathbb{P}(\text{UCB}_1(t) \leq \mu_1)}_{A} + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \text{UCB}_2(t) > \mu_1)}_{B}
\end{align*}
\[ \mathbb{E}[N_2(T)] \leq \sum_{t=0}^{T-1} \mathbb{P}(\text{UCB}_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \text{UCB}_2(t) > \mu_1) \]

- **Term A:** if \( \alpha > 2 \),

\[
\sum_{t=0}^{T-1} \mathbb{P}(\text{UCB}_1(t) \leq \mu_1) \leq 1 + \sum_{t=1}^{T-1} \frac{1}{t^{\alpha-1}} \\
\leq 1 + \zeta(\alpha - 1) := C_\alpha / 2.
\]
Term B:

\[
(B) = \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \text{UCB}_2(t) > \mu_1)
\]

\[
\leq \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \text{UCB}_2(t) > \mu_1, \text{LCB}_2(t) \leq \mu_2) + C_{\alpha}/2
\]

with

\[
\text{LCB}_2(t) = \hat{\mu}_2(t) - \sqrt{\frac{\alpha \log t}{2N_2(t)}}.
\]

(LCB_2(t) < \mu_2 < \mu_1 \leq \text{UCB}_2(t))

\[
\Rightarrow (\mu_1 - \mu_2) \leq 2 \sqrt{\frac{\alpha \log(T)}{2N_2(t)}}
\]

\[
\Rightarrow N_2(t) \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2} \log(T)
\]
\textbf{Term B:} (continued)

\[
(B) \leq \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \text{UCB}_2(t) > \mu_1, \text{LCB}_2(t) \leq \mu_2) + C_{\alpha}/2
\]
\[
\leq \sum_{t=0}^{T-1} \mathbb{P}\left(A_{t+1} = 2, N_2(t) \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2 \log(T)}\right) + C_{\alpha}/2
\]
\[
\leq \frac{2\alpha}{(\mu_1 - \mu_2)^2 \log(T)} + C_{\alpha}/2
\]

\textbf{Conclusion:}

\[
\mathbb{E}[N_2(T)] \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2 \log(T)} + C_{\alpha}.
\]
The KL-UCB algorithm

- A UCB-type algorithm: \( A_{t+1} = \arg \max_a u_a(t) \)
- ... associated to the right upper confidence bounds:

\[
u_a(t) = \max \left\{ q \geq \hat{\mu}_a(t) : d(\hat{\mu}_a(t), x) \leq \frac{\log(t)}{N_a(t)} \right\},\]

\( \hat{\mu}_a(t) \): empirical mean of rewards from arm \( a \) up to time \( t \).

[Cappé et al. 13]: KL-UCB satisfies

\[
E_{\mu} [N_a(T)] \leq \frac{1}{d(\mu_a, \mu^*)} \log T + O(\sqrt{\log(T)}).
\]
Bayesian algorithms for regret minimization?

Algorithms based on Bayesian tools can be good to solve (frequentist) regret minimization.

**Ideas:**

- use the Finite-Horizon Gittins
- use posterior quantiles
- use posterior samples
Thompson Sampling

\((\pi_a^t, \ldots, \pi_K^t)\) posterior distribution on \((\mu_1, \ldots, \mu_K)\) at round \(t\).

**Algorithm: Thompson Sampling**

**Thompson Sampling** is a randomized Bayesian algorithm:

\[
\forall a \in \{1..K\}, \quad \theta_a(t) \sim \pi_a^t \\
A_{t+1} = \arg\max_a \theta_a(t)
\]

“Draw each arm according to its posterior probability of being optimal” [Thompson 1933]

Thompson Sampling is asymptotically optimal. [K., Korda, Munos 2012]
Outline

1. Bayesian bandits: a planning problem
2. Frequentist bandits: asymptotically optimal algorithms
3. Non stochastic bandits: minimax algorithms
Minimax regret

In stochastic (Bernoulli) bandits, we exhibited algorithm satisfying

\[ \forall \mu, \ R_\mu(A, T) = \left( \sum_{a=1}^{K} \frac{(\mu^* - \mu_a)}{d(\mu_a, \mu^*)} \right) \log(T) + o(\log(T)). \]

problem-dependent term, can be large

For those algorithms, one can also prove that, for some constant \( C \),

\[ \forall \mu, \ R_\mu(A, T) \leq C \sqrt{KT \log(T)} \]

problem-independent bound

Minimax rate of the regret

\[ \inf_{A} \sup_{\mu} R_\mu(A, T) = O\left( \sqrt{KT} \right) \]
A new bandit game: at round $t$

- the player chooses arm $A_t$
- simultaneously, an adversary chooses the vector of rewards $(x_{t,1}, \ldots, x_{t,K})$
- the player receives the reward $x_t = x_{A_t,t}$

Goal: maximize rewards, or minimize regret

$$R(A, T) = \max_a \mathbb{E}\left[\sum_{t=1}^T x_{a,t}\right] - \mathbb{E}\left[\sum_{t=1}^T x_t\right].$$
Full information: Exponential Weighted Forecaster

The full-information game: at round $t$

- the player chooses arm $A_t$
- simultaneously, an adversary chooses the vector of rewards
  $$(x_{t,1}, \ldots, x_{t,K})$$
- the player receives the reward $x_t = x_{A_t,t}$
- and he observes the reward vector $(x_{t,1}, \ldots, x_{t,K})$

The EWF algorithm [Littelstone, Warmuth 1994]

With $\hat{p}_t$ the probability distribution

$$\hat{p}_{a,t} \propto e^{\eta \left( \sum_{s=1}^{t-1} x_{a,s} \right)}$$

at round $t$, choose

$$A_t \sim \hat{p}_t$$
We don’t have access to the \((x_{a,t})\) for all \(a\)... 

\[
\hat{x}_{a,t} = \frac{x_{a,t}}{\hat{p}_{a,t}} 1(A_t=a)
\]

satisfies \(\mathbb{E}[\hat{x}_{a,t}] = x_{a,t}\).

**The EXP3 strategy [Auer et al. 2003]**

With \(\hat{p}_t\) the probability distribution

\[
\hat{p}_{a,t} \propto e^{\eta \left( \sum_{s=1}^{t-1} \hat{x}_{a,s} \right)}
\]

at round \(t\), choose

\[A_t \sim \hat{p}_t\]
Theoretical results

The EXP3 strategy [Auer et al. 2003]

With $\hat{p}_t$ the probability distribution

$$\hat{p}_{a,t} \propto e^{\eta \left( \sum_{s=1}^{t-1} \hat{x}_{a,s} \right)}$$

at round $t$, choose

$$A_t \sim \hat{p}_t$$

[Bubeck and Cesa-Bianchi 12] EXP3 with

$$\eta = \sqrt{\frac{1}{K} \log(K/T)}$$

satisfies

$$R(\text{EXP3}, T) \leq \sqrt{2 \log K \sqrt{KT}}$$

Remarks:

- almost the same guarantees for $\eta_t = \sqrt{\frac{1}{Kt} \log(K)}$
- extra exploration is needed to have high probability results
Under different assumptions, different types of strategies to achieve an exploration-exploitation tradeoff in bandit models:

**Index policies:**
- Gittins indices
- UCB-type algorithms

**Randomized algorithms:**
- Thompson Sampling
- Exponential weights

More complex bandit models not covered today: restless bandits, contextual bandits, combinatorial bandits...
Books and surveys

Bayesian bandits:

Stochastic and non-stochastic bandits: