Outline

- Exploration/exploitation tradeoff
- Regret
- Multi-armed bandits
  - Frequentist approaches
    - $\epsilon$-greedy strategies
    - Upper confidence bounds
  - Bayesian bandits
    - Thompson Sampling
Exploration/Exploitation Tradeoff

- Fundamental problem of RL due to the active nature of the learning process
- Consider one-state RL problems known as bandits
Stochastic Bandits

- **Formal definition:**
  - Single state: $S = \{s\}$
  - $A$: set of actions (also known as arms)
  - Space of rewards (often re-scaled to be $[0,1]$)
  - Finite/Infinite horizons
  - Average reward setting ($\gamma = 1$)

- No transition function to be learned since there is a single state

- We simply need to learn the **stochastic** reward function
Origin

- The term bandit comes from gambling where slot machines can be thought as one-armed bandits.

- Problem: which slot machine should we play at each turn when their payoffs are not necessarily the same and initially unknown?
Examples

- Design of experiments (Clinical Trials)
- Online ad placement
- Web page personalization
- Recommender systems
- Networks (packet routing)
Online Ad Placement

Six Ontarians charged in alleged $200-million investment fraud

- WATCH Video: How to protect your bank account from fraud

122 'potential objects' spotted in ocean offer fresh jet lead

- WATCH Sailing the waters where Flight 370 went down

TORONTO: Chow presses Ford to ‘take down the circus tent’ as candidates hammer each other in mayoral debate
Online Ad Optimization

- Problem: which ad should be presented?
- Answer: present ad with highest payoff

\[ \text{payoff} = \text{clickThroughRate} \times \text{payment} \]

- Click through rate: probability that user clicks on ad
- Payment: $$ paid by advertiser
  - Amount determined by an auction
Simplified Problem

- Assume payment is 1 unit for all ads
- Need to estimate click through rate

- Formulate as a bandit problem:
  - Arms: the set of possible ads
  - Rewards: 0 (no click) or 1 (click)

- In what order should ads be presented to maximize revenue?
  - How should we balance exploitation and exploration?
Simple yet Difficult Problem

- Simple: description of the problem is short
- Difficult: no known tractable optimal solution
Simple Heuristics

- **Greedy strategy**: select the arm with the highest average so far
  - May get stuck due to lack of exploration

- **\( \epsilon \)-greedy**: select an arm at random with probability \( \epsilon \) and otherwise do a greedy selection
  - Convergence rate depends on choice of \( \epsilon \)
Regret

- Let $R(a)$ be the **true (unknown) expected reward** of $a$

- Let $r^* = \max_a R(a)$ and $a^* = \arg\max_a R(a)$

- Denote by $loss(a)$ the **expected regret** of $a$

  \[ loss(a) = r^* - R(a) \]

- Denote by $Loss_n$ the **expected cumulative regret** for $n$ time steps

  \[ Loss_n = \sum_{t=1}^{n} loss(a_t) \]
Theoretical Guarantees

- **When $\epsilon$ is constant, then**
  - For large enough $t$: $Pr(a_t \neq a^*) \approx \epsilon$
  - Expected cumulative regret: $Loss_n \approx \sum_{t=1}^{n} \epsilon \times 1 + (1 - \epsilon) \times 0 = \sum_{t=1}^{n} \epsilon = O(n)$
    - Linear regret

- **When $\epsilon_t \propto 1/t$**
  - For large enough $t$: $Pr(a_t \neq a^*) \approx \epsilon_t = O\left(\frac{1}{t}\right)$
  - Expected cumulative regret: $Loss_n \approx \sum_{t=1}^{n} \frac{1}{t} = O(\log n)$
    - Logarithmic regret
Empirical Mean

- Problem: how far is the empirical mean $\tilde{R}(a)$ from the true mean $R(a)$?

- If we knew that $|R(a) - \tilde{R}(a)| \leq \text{bound}$
  - Then we would know that $R(a) \leq \tilde{R}(a) + \text{bound}$
  - And we could select the arm with best $\tilde{R}(a) + \text{bound}$

- Overtime, additional data will allow us to refine $\tilde{R}(a)$ and compute a tighter $\text{bound}$.
Positivism in the Face of Uncertainty

- Suppose that we have an oracle that returns an upper bound $UB_n(a)$ on $R(a)$ for each arm based on $n$ trials of arm $a$.

- Suppose the upper bound returned by this oracle converges to $R(a)$ in the limit:
  - i.e., $\lim_{n \to \infty} UB_n(a) = R(a)$

- Optimistic algorithm
  - At each step, select $\arg \max_a UB_n(a)$
Convergence

- **Theorem:** An optimistic strategy that always selects \( \arg\max_a UB_n(a) \) will converge to \( a^* \)

- **Proof by contradiction:**
  - Suppose that we converge to suboptimal arm \( a \) after infinitely many trials.
  - Then \( R(a) = UB_\infty(a) \geq UB_\infty(a') = R(a') \ \forall a' \)
  - But \( R(a) \geq R(a') \ \forall a' \) contradicts our assumption that \( a \) is suboptimal.
Probabilistic Upper Bound

- Problem: We can’t compute an upper bound with certainty since we are sampling

- However we can obtain measures $f$ that are upper bounds most of the time
  - i.e., $\Pr(R(a) \leq f(a)) \geq 1 - \delta$

  Example: Hoeffding’s inequality
  $$\Pr \left( R(a) \leq \bar{R}(a) + \sqrt{\frac{\log \left( \frac{1}{\delta} \right)}{2n_a}} \right) \geq 1 - \delta$$

  where $n_a$ is the number of trials for arm $a$
Upper Confidence Bound (UCB)

- Set $\delta_n = 1/n^4$ in Hoeffding’s bound
- Choose $a$ with highest Hoeffding bound

\[
\text{UCB}(h)
\]

$V \leftarrow 0$, $n \leftarrow 0$, $n_a \leftarrow 0$ \hspace{1em} \forall a$

Repeat until $n = h$

Execute $\arg\max_a \tilde{R}(a) + \sqrt{\frac{2\log n}{n_a}}$

Receive $r$

$V \leftarrow V + r$

$\tilde{R}(a) \leftarrow \frac{n_a \tilde{R}(a) + r}{n_a + 1}$

$n \leftarrow n + 1$, $n_a \leftarrow n_a + 1$

Return $V$
UCB Convergence

- **Theorem:** Although Hoeffding’s bound is probabilistic, UCB converges.

- **Idea:** As \( n \) increases, the term \( \sqrt{\frac{2\log n}{n_a}} \) increases, ensuring that all arms are tried infinitely often.

- **Expected cumulative regret:** \( \text{Loss}_n = O(\log n) \)
  - Logarithmic regret
Multi-Armed Bandits

- Problem:
  - $N$ bandits with unknown average reward $R(a)$
  - Which arm $a$ should we play at each time step?
  - Exploitation/exploration tradeoff

- Common frequentist approaches:
  - $\epsilon$-greedy
  - Upper confidence bound (UCB)

- Alternative Bayesian approaches
  - Thompson sampling
  - Gittins indices
Bayesian Learning

- **Notation:**
  - $r^a$: random variable for $a$’s rewards
  - $\Pr(r^a; \theta)$: unknown distribution (parameterized by $\theta$)
  - $R(a) = E[r^a]$: unknown average reward

- **Idea:**
  - Express uncertainty about $\theta$ by a prior $\Pr(\theta)$
  - Compute posterior $\Pr(\theta | r_1^a, r_2^a, \ldots, r_n^a)$ based on samples $r_1^a, r_2^a, \ldots, r_n^a$ observed for $a$ so far.

- **Bayes theorem:**
  $$\Pr(\theta | r_1^a, r_2^a, \ldots, r_n^a) \propto \Pr(\theta)\Pr(r_1^a, r_2^a, \ldots, r_n^a | \theta)$$
Distributional Information

- Posterior over \( \theta \) allows us to estimate
  - Distribution over next reward \( r^a \)
    
    \[
    Pr(r_{n+1}^a \mid r_1^a, r_2^a, \ldots, r_n^a) = \int \theta Pr(r_{n+1}^a \mid \theta)Pr(\theta \mid r_1^a, r_2^a, \ldots, r_n^a)d\theta
    \]
  - Distribution over \( R(a) \) when \( \theta \) includes the mean
    
    \[
    Pr\left( R(a) \mid r_1^a, r_2^a, \ldots, r_n^a \right) = Pr\left( \theta \mid r_1^a, r_2^a, \ldots, r_n^a \right) \text{ if } \theta = R(a)
    \]

- To guide exploration:
  - UCB: \( \Pr\left( R(a) \leq \text{bound}(r_1^a, r_2^a, \ldots, r_n^a) \right) \geq 1 - \delta \)
  - Bayesian techniques: \( \Pr\left( R(a) \mid r_1^a, r_2^a, \ldots, r_n^a \right) \)
Coin Example

- Consider two biased coins $C_1$ and $C_2$
  
  $R(C_1) = \Pr(\text{head})$
  
  $R(C_2) = \Pr(\text{head})$

- Problem:
  - Maximize # of heads in $k$ flips
  - Which coin should we choose for each flip?
Bernoulli Variables

- $r_{C_1}, r_{C_2}$ are Bernoulli variables with domain \{0,1\}

- Bernoulli distributions are parameterized by their mean
  - i.e., $\Pr(r_{C_1}; \theta_1) = \theta_1 = R(C_1)$
  - $\Pr(r_{C_2}; \theta_2) = \theta_2 = R(C_2)$
Beta Distribution

- Let the prior $\Pr(\theta)$ be a Beta distribution
  \[ \text{Beta}(\theta; \alpha, \beta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1} \]

  - $\alpha - 1$: # of heads
  - $\beta - 1$: # of tails

- $E[\theta] = \alpha / (\alpha + \beta)$
Belief Update

- Prior: \( \Pr(\theta) = Beta(\theta; \alpha, \beta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1} \)

- Posterior after coin flip:

\[
\Pr(\theta \mid \text{head}) \propto \Pr(\theta) \cdot \Pr(\text{head} \mid \theta) \\
\propto \theta^{\alpha-1}(1 - \theta)^{\beta-1} \cdot \theta \\
= \theta^{(\alpha+1)-1}(1 - \theta)^{\beta-1} \propto Beta(\theta; \alpha + 1, \beta)
\]

\[
\Pr(\theta \mid \text{tail}) \propto \Pr(\theta) \cdot \Pr(\text{tail} \mid \theta) \\
\propto \theta^{\alpha-1}(1 - \theta)^{\beta-1} \cdot (1 - \theta) \\
= \theta^{\alpha-1}(1 - \theta)^{(\beta+1)-1} \propto Beta(\theta; \alpha, \beta + 1)
\]
Thompson Sampling

- Idea:
  - Sample several potential average rewards:
    \[ R_1(a), \ldots, R_k(a) \sim \Pr(R(a) \mid r_1^a, \ldots, r_n^a) \text{ for each } a \]
  - Estimate empirical average
    \[ \hat{R}(a) = \frac{1}{k} \sum_{i=1}^{k} R_i(a) \]
  - Execute \( \text{argmax}_a \hat{R}(a) \)

- Coin example
  - \( \Pr(R(a) \mid r_1^a, \ldots, r_n^a) = \text{Beta}(\theta_a; \alpha_a, \beta_a) \)
    - \( \alpha_a - 1 = \#\text{heads} \) and \( \beta_a - 1 = \#\text{tails} \)
Thompson Sampling Algorithm Bernoulli Rewards

\[
\text{ThompsonSampling}(h)
\]
\[
V \leftarrow 0
\]
For \(n = 1\) to \(h\)
\[
\text{Sample } R_1(a), \ldots, R_k(a) \sim \Pr(R(a)) \quad \forall a
\]
\[
\hat{R}(a) \leftarrow \frac{1}{k} \sum_{i=1}^{k} R_i(a) \quad \forall a
\]
\[
a^* \leftarrow \arg\max_a \hat{R}(a)
\]
Execute \(a^*\) and receive \(r\)
\[
V \leftarrow V + r
\]
Update \(\Pr(R(a^*))\) based on \(r\)
Return \(V\)
Comparison

**Thompson Sampling**
- Action Selection
  \[ a^* = \arg\max_a \hat{R}(a) \]
- Empirical mean
  \[ \hat{R}(a) = \frac{1}{k} \sum_{i=1}^{k} R_i(a) \]
- Samples
  \[ R_j(a) \sim Pr(R(a) | r_1^a, \ldots, r_n^a) \]
  \[ r_i^a \sim Pr(r^a; \theta) \]
- Some exploration

**Greedy Strategy**
- Action Selection
  \[ a^* = \arg\max_a \tilde{R}(a) \]
- Empirical mean
  \[ \tilde{R}(a) = \frac{1}{n} \sum_{i=1}^{n} r_i^a \]
- Samples
  \[ r_i^a \sim Pr(r^a; \theta) \]
- No exploration
Sample Size

- In Thompson sampling, amount of data $n$ and sample size $k$ regulate amount of exploration.

- As $n$ and $k$ increase, $\hat{R}(a)$ becomes less stochastic, which reduces exploration.
  - As $n \uparrow$, $Pr(R(a) | r_1^a, \ldots, r_n^a)$ becomes more peaked.
  - As $k \uparrow$, $\hat{R}(a)$ approaches $\mathbb{E}[R(a) | r_1^a, \ldots, r_n^a]$.

- The stochasticity of $\hat{R}(a)$ ensures that all actions are chosen with some probability.
Analysis

- Thompson sampling converges to best arm

- Theory:
  - Expected cumulative regret: $O(\log n)$
  - On par with UCB and $\epsilon$-greedy

- Practice:
  - Sample size $k$ often set to 1
Summary

- **Stochastic bandits**
  - Exploration/exploitation tradeoff
  - $\epsilon$-greedy and UCB
    - Theory: logarithmic expected cumulative regret

- **In practice:**
  - UCB often performs better than $\epsilon$-greedy
  - Many variants of UCB improve performance

- **Bayesian Bandits**
  - Thompson Sampling