Advanced Machine Learning
Bandit Problems
Multi-Armed Bandit Problem

Problem: which arm of a $K$-slot machine should a gambler pull to maximize his cumulative reward over a sequence of trials?

- stochastic setting.
- adversarial setting.
Motivation

- Clinical trials: potential treatments for a disease to select from, new patient or category at each round (Thompson, 1933).

- Ads placement: selection of ad to display out of a finite set (which could vary with time though) for each new web page visitor.

- Adaptive routing: alternative paths for routing packets through a “series of tubes” or alternative roads for driving from a source to a destination.

- Games: different moves at each round of a game such as chess, or Go.
Key Problem

Exploration vs exploitation dilemma (or trade-off):
- inspect new arms with possibly better rewards.
- use existing information to select best arm.
Outline

- Stochastic bandits
- Adversarial bandits
Stochastic Model

- \( K \) arms: for each arm \( i \in \{1, \ldots, K\} \),
  - reward distribution \( P_i \).
  - reward mean \( \mu_i \).
  - gap to best: \( \Delta_i = \mu^* - \mu_i \), where \( \mu^* = \arg\max_{i \in [1,K]} \mu_i \).
Bandit Setting

For $t = 1$ to $T$ do

- player selects action $I_t \in \{1, \ldots, K\}$ (randomized).
- player receives reward $X_{I_t,t} \sim P_{I_t}$.

Equivalent descriptions:

- on-line learning with partial information ($\neq$ full).
- one-state MDPs (Markov Decision Processes).
Objectives

- **Expected regret**

\[
E[R_T] = E \left[ \max_{i \in [1,K]} \sum_{t=1}^{T} X_{i,t} - \sum_{t=1}^{T} X_{I_t,t} \right].
\]

- **Pseudo-regret**

\[
\overline{R}_T = \max_{i \in [1,K]} E \left[ \sum_{t=1}^{T} X_{i,t} - \sum_{t=1}^{T} X_{I_t,t} \right].
\]

\[
= \mu^* T - E \left[ \sum_{t=1}^{T} X_{I_t,t} \right].
\]

- By Jensen’s inequality, \( \overline{R}_T \leq E[R_T] \).
**Expected Regret**

- If \((X_{i,t} - \mu_i)\)s take values in \([-r, +r]\), then
  \[
  \mathbb{E} \left[ \max_{i \in [1,K]} \sum_{t=1}^{T} (X_{i,t} - \mu^*) \right] \leq r \sqrt{2T \log K}.
  \]

- The \(O(\sqrt{T})\) dependency cannot be improved;
  better guarantees can be achieved for pseudo-regret.
Pseudo-Regret

Expression in terms of $\Delta_i$s:

$$\overline{R}_T = \sum_{i=1}^{K} E[T_i(T)] \Delta_i,$$

where $T_i(t)$ denotes the number of times arm $i$ was pulled up to time $t$, $T_i(t) = \sum_{s=1}^{t} 1_{I_s=i}$.

Proof: $\overline{R}_T = \mu^* T - E \left[ \sum_{t=1}^{T} X_{I_t,t} \right] = E \left[ \sum_{t=1}^{T} (\mu^* - X_{I_t,t}) \right]$

$$= \sum_{t=1}^{T} \sum_{i=1}^{K} (\mu^* - X_{i,t}) 1_{I_t=i} = \sum_{t=1}^{T} \sum_{i=1}^{K} E[(\mu^* - X_{i,t})] E[1_{I_t=i}]$$

$$= \sum_{i=1}^{K} (\mu^* - \mu_i) E \left[ \sum_{t=1}^{T} 1_{I_t=i} \right] = \sum_{i=1}^{K} E[T_i(T)] \Delta_i.$$
\(\varepsilon\)-Greedy Strategy

- At time \(t\),
  - with probability \(1 - \varepsilon_t\), select arm \(i\) with best emp. mean.
  - with probability \(\varepsilon_t\), select random arm.

- For \(\varepsilon_t = \min\left(\frac{6K}{\Delta_i^2 t}, 1\right)\), with \(\Delta = \min_{i: \Delta_i > 0} \Delta_i\),
  - for \(t \geq \frac{6K}{\Delta^2}\), \(\Pr[I_t \neq i^*] \leq \frac{C}{\Delta^2 t}\) for some \(C > 0\).
  - thus, \(E[T_i(T)] \leq \frac{C}{\Delta^2} \log T\) and \(\bar{R}_T \leq \sum_{i: \Delta_i > 0} \frac{C \Delta_i}{\Delta^2} \log T\).

- Logarithmic regret but,
  - requires knowledge of \(\Delta\).
  - sub-optimal arms treated similarly (naive search).
Note on Concentration Ineqs

Let $X$ be a random variable such that for all $t \geq 0$,
\[ \log \mathbb{E} \left[ e^{t(X-\mathbb{E}[X])} \right] \leq \Psi(t), \]
where $\Psi$ is a convex function. For Hoeffding’s inequality and $X \in [a, b], \Psi(t) = \frac{t^2(b-a)^2}{8}$.

Then, $\Pr[X - \mathbb{E}[X] > \epsilon] = \Pr[e^{t(X-\mathbb{E}[X])} > e^{t\epsilon}]$
\[ \leq \inf_{t>0} e^{-t\epsilon} \mathbb{E}[e^{t(X-\mathbb{E}[X])}] \]
\[ \leq \inf_{t>0} e^{-t\epsilon} e^{\Psi(t)} \]
\[ = e^{-\sup_{t>0}(t\epsilon - \Psi(t))} \]
\[ = e^{-\Psi^*(\epsilon)}. \]
UCB Strategy

(Lai and Robbins, 1985; Agrawal 1995; Auer et al. 2002a)

- **Optimism in face of uncertainty:**
  - at each time $t \in [1, T]$ compute upper confidence bound (UCB) on the expected reward of each arm $i \in [1, K]$.
  - select arm with largest UCB.

- **Idea:** wrong arm $i$ cannot be selected for long.
  - by definition, $\mu_i \leq \mu^* \leq \text{UCB}_i$.
  - pulling $i$ often $\rightarrow$ UCB closer to $\mu_i$. 

UCB Strategy

- Average reward estimate for arm $i$ by time $t$:
  \[
  \hat{\mu}_{i,t} = \frac{1}{T_i(t)} \sum_{s=1}^{t} X_{i,s} 1_{I_s=i}.
  \]

- Concentration inequality (e.g., Hoeffding’s ineq.):
  \[
  \Pr[\mu_i - \frac{1}{t} \sum_{s=1}^{t} X_{i,s} > \epsilon] \leq e^{-t\psi^*(\epsilon)}.
  \]

- Thus, for any $\delta > 0$, with probability at least $1 - \delta$,
  \[
  \mu_i < \frac{1}{t} \sum_{s=1}^{t} X_{i,s} + \psi^{-1}\left(\frac{1}{t} \log \frac{1}{\delta}\right).
  \]
(α, ψ)-UCB Strategy

Parameter $\alpha > 0$; $(\alpha, \psi)$-UCB strategy consists of selecting at time $t$

$$I_t \in \arg\max_{i \in [1, K]} \left[ \hat{\mu}_{i,t-1} + \psi^{-1}(\frac{\alpha \log t}{T_i(t-1)}) \right].$$
(α, ψ)-UCB Guarantee

- **Theorem**: for $\alpha > 2$, the pseudo-regret of $(\alpha, \psi)$-UCB satisfies

$$
\overline{R}_T \leq \sum_{i: \Delta_i > 0} \left( \frac{\alpha \Delta_i}{\psi^*(\frac{\Delta_i}{2})} \log T + \frac{\alpha}{\alpha - 2} \right).
$$

- for Hoeffding's lemma, $\alpha$-UCB, $\psi^*(\epsilon) = 2\epsilon^2$ (Auer et al. 2002a),

$$
\overline{R}_T \leq \sum_{i: \Delta_i > 0} \left( \frac{2\alpha}{\Delta_i} \log T + \frac{\alpha}{\alpha - 2} \right).
$$
Proof

Lemma: for any $s \geq 0$,

$$\sum_{t=1}^{T} 1_{I_t=i} \leq s + \sum_{t=s+1}^{T} 1_{I_t=i} 1_{T_i(t-1) \geq s}.$$  

Proof: observe that

$$\sum_{t=1}^{T} 1_{I_t=i} = \sum_{t=1}^{T} 1_{I_t=i} 1_{T_i(t-1) \leq s} + \sum_{t=1}^{T} 1_{I_t=i} 1_{T_i(t-1) \geq s}.$$  

- Now, for $t^* = \arg\max_{t \leq T} 1_{T_i(t-1) \leq s} \neq 0$

  $$\sum_{t=1}^{T} 1_{I_t=i} 1_{T_i(t-1) \leq s} = \sum_{t=1}^{t^*} 1_{I_t=i} 1_{T_i(t-1) \leq s},$$

- By definition of $t^*$, the number of non-zero terms in the sum is at most $s$. 
Proof

- For any $i$ and $t$ define $\eta_{i,t-1} = \psi^*-1\left(\frac{\alpha \log t}{T_i(t-1)}\right)$. At time $t$, if $i$ is selected, then

$$(\hat{\mu}_{i,t-1} + \eta_{i,t-1}) - (\hat{\mu}_{i^*,t} + \eta_{i^*,t-1}) \geq 0$$

$$\Leftrightarrow [\hat{\mu}_{i,t-1} - \mu_{i,t-1} - \eta_{i,t-1}] + [2\eta_{i,t-1} - \Delta_i] + [\mu^* - \hat{\mu}_{i^*,t-1} - \eta_{i^*,t-1}] \geq 0.$$ 

Thus, at least of one of these three terms is non-negative. Also, if one is non-positive, at least one of the other two is non-negative.
Proof

To bound the pseudo-regret, we bound $E[T_i(T)]$. But, observe first that

$$T_i(t - 1) \geq s = \left\lceil \frac{\alpha \log T}{\psi^*(\frac{\Delta_i}{2})} \right\rceil \geq \frac{\alpha \log t}{\psi^*(\frac{\Delta_i}{2})} \Rightarrow \Delta_i - 2\eta_{i,t-1} \geq 0.$$

Thus,

$$E[T_i(T)] = E\left[\sum_{t=1}^{T} 1_{I_t=i}\right]$$

$$\leq s + E\left[\sum_{t=s+1}^{T} 1_{I_t=i} 1_{T_i(t-1) \geq s}\right]$$

$$\leq s + \sum_{t=s+1} \Pr[\hat{\mu}_{i,t-1} - \mu_{i,t-1} - \eta_{i,t-1} \geq 0] + \Pr[\mu^* - \hat{\mu}_{i^*,t-1} - \eta_{i^*,t-1} \geq 0].$$
Proof

Each of the two probability terms can be bounded as follows using the union bound:

\[ \Pr[\mu^* - \hat{\mu}_{i^*},t-1 - \eta_{i^*},t-1 \geq 0] \]

\[ \leq \Pr \left[ \exists s \in [1,t]: \mu^* - \hat{\mu}_{i^*},s - \psi^{-1} \left( \frac{\alpha \log t}{s} \right) \geq 0 \right] \]

\[ \leq \sum_{s=1}^{t} \frac{1}{t^\alpha} = \frac{1}{t^{\alpha-1}}. \]

Final constant of the bound obtained by further simple calculations.
Lower Bound

Theorem: for any strategy such that \( \mathbb{E}[T_i(T)] = o(T^\beta) \) for any arm \( i \) and any \( \beta > 0 \) for any set of Bernoulli reward distributions, the following holds for all Bernoulli reward distributions:

\[
\liminf_{T \to +\infty} \frac{R_T}{\log T} \geq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{D(\mu_i \parallel \mu^*)}.
\]

- a more general result holds for general distributions.
Observe that

\[
    \sum_{i: \Delta_i > 0} \frac{\Delta_i}{D(\mu_i \parallel \mu^*)} \geq \mu^*(1 - \mu^*) \sum_{i: \Delta_i > 0} \frac{1}{\Delta_i},
\]

since

\[
    D(\mu_i \parallel \mu^*) = \mu_i \log \frac{\mu_i}{\mu^*} + (1 - \mu_i) \log \frac{1 - \mu_i}{1 - \mu^*}
\]

\[
\leq \frac{\mu_i (\mu_i - \mu^*)}{\mu^*} + (1 - \mu_i) \frac{\mu^* - \mu_i}{1 - \mu^*}
\]

\[
= \frac{(\mu - \mu^*)^2}{\mu^*(1 - \mu^*)} = \frac{\Delta_i^2}{\mu^*(1 - \mu^*)}. \]
Outline

- Stochastic bandits
- Adversarial bandits
Adversarial Model

- $K$ arms: for each arm $i \in \{1, \ldots, K\}$,
  - no stochastic assumption.
  - rewards in $[0, 1]$. 
Bandit Setting

For $t = 1$ to $T$ do

• player selects action $I_t \in \{1, \ldots, K\}$ (randomized).
• player receives reward $x_{I_t, t}$.

Notes:

• rewards $x_{i, t}$ for all arms determined by adversary simultaneously with the selection $I_t$ of an arm by player.
• adversary oblivious or nonoblivious (or adaptive).
• strategies: deterministic, regret of at least $\frac{T}{2}$ for some (bad) sequences, thus must consider randomized.
Scenarios

- Oblivious case:
  - adversary rewards selected independently of the player’s actions; thus, reward vector at time $t$ only a function of $t$.

- Non-oblivious case:
  - adversary rewards at time $t$ function of the player’s past actions $I_1, \ldots, I_{t-1}$.
  - notion of regret problematic: cumulative reward compared to a quantity that depends on the player’s actions! (single best action in hindsight function of actions $I_1, \ldots, I_T$ played; playing that single “best” action could have resulted in different rewards.)
Objectives

- Minimize regret \((\ell_{i,t} = 1 - x_{i,t})\), expectation or high prob.:

\[
R_T = \max_{i \in [1,K]} \sum_{t=1}^{T} x_{i,t} - \sum_{t=1}^{T} x_{I_t,t} = \sum_{t=1}^{T} \ell_{I_t,t} - \min_{i \in [1,K]} \sum_{t=1}^{T} \ell_{i,t}.
\]

- Pseudo-regret:

\[
\overline{R}_T = E\left[\sum_{t=1}^{T} \ell_{I_t,t}\right] - \min_{i \in [1,K]} E\left[\sum_{t=1}^{T} \ell_{i,t}\right].
\]

- By Jensen’s inequality, \(\overline{R}_T \leq E[R_T]\).
Importance Weighting

- In the bandit setting, the cumulative loss of each arm is not observed, so how should we update the probabilities?

- Estimates via surrogate loss:

\[ \tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}} 1_{I_t = i}, \]

where \( p_t = (p_{1,t}, \ldots, p_{K,t}) \) is the probability distribution the player uses at time \( t \) to draw an arm (\( p_{i,t} > 0 \)).

- Unbiased estimate: for any \( i \),

\[
\mathbb{E}_{I_t \sim p_t} [\tilde{\ell}_{i,t}] = \sum_{j=1}^{K} p_{j,t} \frac{\ell_{i,t}}{p_{i,t}} 1_{j = i} = \ell_{i,t}.
\]
EXP3

EXP3($K$)
1 p₁ ← (\frac{1}{K}, \ldots, \frac{1}{K})
2 (\tilde{L}_{1,0}, \ldots, \tilde{L}_{K,0}) ← (0, \ldots, 0)
3 for $t$ ← 1 to $T$ do
4 \hspace{1em} \text{SAMPLE}(I_t \sim p_t)
5 \hspace{1em} \text{RECEIVE}(\ell_{I_t,t})
6 \hspace{2em} \text{for } i ← 1 \text{ to } K \text{ do}
7 \hspace{3em} \ell_{i,t} ← \frac{\ell_{i,t}}{p_i,t} 1_{I_t=i}
8 \hspace{3em} \tilde{L}_{i,t} ← \tilde{L}_{i,t-1} + \ell_{i,s}
9 \hspace{2em} \text{for } i ← 1 \text{ to } K \text{ do}
10 \hspace{3em} p_{i,t+1} ← \frac{e^{-\eta \tilde{L}_{i,t}}}{\sum_{j=1}^{K} e^{-\eta \tilde{L}_{j,t}}}
11 \hspace{1em} \text{return } p_{T+1}

EXP3 (Exponential weights for Exploration and Exploitation)
EXP3 Guarantee

- **Theorem**: the pseudo-regret of EXP3 can be bounded as follows:

\[ \overline{R}_T \leq \frac{\log K}{\eta} + \frac{\eta KT}{2}. \]

Choosing \( \eta \) to minimize the bound gives

\[ \overline{R}_T \leq \sqrt{2KT \log K}. \]

- **Proof**: similar to that of EG, but we cannot use Hoeffding's inequality since \( \ell_{i,t} \) is unbounded.
Proof

Potential: \( \Phi_t = \log \sum_{i=1}^{K} e^{-\eta \tilde{L}_{i,t}} \).

Upper bound:

\[
\Phi_t - \Phi_{t-1} = \log \frac{\sum_{i=1}^{K} e^{-\eta \tilde{L}_{i,t}}}{\sum_{i=1}^{N} e^{-\eta \tilde{L}_{i,t-1}}} = \log \frac{\sum_{i=1}^{K} e^{-\eta \tilde{L}_{i,t-1}} e^{-\eta \tilde{\ell}_{i,t}}}{\sum_{i=1}^{N} e^{-\eta \tilde{L}_{i,t-1}}}
\]

\[
= \log \left[ \mathbb{E}_{i \sim p_t} \left[ e^{-\eta \tilde{\ell}_{i,t}} \right] \right]
\]

\[
\leq \mathbb{E}_{i \sim p_t} \left[ e^{-\eta \tilde{\ell}_{i,t}} \right] - 1 \quad \text{(log } x \leq x - 1) \]

\[
\leq \mathbb{E}_{i \sim p_t} \left[ -\eta \tilde{\ell}_{i,t} + \frac{\eta^2}{2} \tilde{\ell}_{i,t}^2 \right] \quad \text{(} e^{-x} \leq 1 - x + \frac{x^2}{2} \text{)}
\]

\[
= -\eta \mathbb{E}_{i \sim p_t} \tilde{\ell}_{i,t} + \frac{\eta^2}{2} \mathbb{E}_{i \sim p_t} \left[ \frac{l_{i,t}^2 1_{I_t=i}}{\pi_{i,t}} \right]
\]

\[
= -\eta \ell_{I_t,t} + \frac{\eta^2}{2} \frac{l_{I_t,t}^2}{\pi_{I_t,t}} \leq -\eta \ell_{I_t,t} + \frac{\eta^2}{2} \frac{1}{\pi_{I_t,t}}.
\]
Proof

- **Upper bound**: summing up the inequalities yields

\[
E[\Phi_T - \Phi_0] \leq -\eta \sum_{t=1}^{T} \ell_{I_t, t} + \frac{\eta^2}{2} \sum_{t=1}^{T} \frac{p_{I_t, t}}{2p_{I_t, t}} = -\eta \sum_{t=1}^{T} \ell_{I_t, t} + \frac{\eta^2 KT}{2}.
\]

- **Lower bound**: for all \( j \in [1, K] \),

\[
E[\Phi_T - \Phi_0] = E \sum_{t=1}^{K} e^{-\eta \tilde{L}_{i, t}} - \log K \\
\geq -\eta \sum_{i=1}^{K} E \tilde{L}_{i, t} - \log K = -\eta \sum_{i=1}^{K} L_{i, t} - \log K.
\]

- **Comparison**:

\[
\forall j \in [1, K], \quad \eta \sum_{t=1}^{T} \ell_{I_t, t} - \eta E[L_{j, T}] \leq \log K + \frac{\eta^2}{2} KT \\
\Rightarrow R_T \leq \frac{\log K}{\eta} + \frac{\eta KT}{2}.
\]
Notes

- When $T$ is not known:
  - standard doubling trick.
  - or, use $\eta_t = \sqrt{\log K / K_t}$, then $\bar{R}_T \leq 2\sqrt{KT\log K}$.

- High probability bounds:
  - importance weighting problem: unbounded second moment (see (Cortes, Mansour, MM, 2010)), $E_{i \sim p_t}[\tilde{l}_{i,t}^2] = \frac{\ell_{I_t,t}^2}{p_{I_t,t}}$.
  - (Auer et al., 2002b): mixing probability with a uniform distribution to ensure a lower bound on $p_{i,t}$; but not sufficient for high probability bound.
  - solution: biased estimate $\tilde{l}_{i,t} = \frac{\ell_{i,t}1_{I_t=i}+\beta}{p_{i,t}p_{I_t,t}}$ with $\beta > 0$ a parameter to tune.
Lower Bound

(Bubek and Cesa-Bianchi, 2012)

- Sufficient lower bound in a stochastic setting for the pseudo-regret (and therefore for the expected regret).

- **Theorem:** for any $T \geq 1$ and any player strategy, there exists a distribution of losses in $\{0, 1\}$ for which

  $$\overline{R}_T \geq \frac{1}{20} \sqrt{KT}.$$
Notes

- Bound of EXP3 matching lower bound modulo Log term.

- Log-free bound: $p_{i,t+1} = \psi(C_t - \tilde{L}_{i,t})$ where $C_t$ is a constant ensuring $\sum_{i=1}^{K} p_{i,t+1} = 1$ and $\psi$ increasing, convex, twice differentiable over $\mathbb{R}^*$ (Audibert and Bubeck, 2010).
  
  - EXP3 coincides with $\psi(x) = e^{\eta x}$.
  - log-free bound with $\psi(x) = (-\eta x)^{-q}$ and $q = 2$.
  - formulation as mirror descent.
  - only in oblivious case.
References


References


