

Advanced Machine Learning

Bandit Problems

MEHRYAR MOHRI MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

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 $i \in \{1, \dots, K\}$,
 - reward distribution P_i .
 - reward mean μ_i .
 - gap to best: $\Delta_i = \mu^* - \mu_i$, where $\mu^* = \max_{i \in [1, K]} \mu_i$.

Bandit Setting

- For $t = 1$ to T do
 - player selects action $I_t \in \{1, \dots, K\}$ (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

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

■ Pseudo-regret

$$\begin{aligned} \bar{R}_T &= \max_{i \in [1, K]} \mathbb{E} \left[\sum_{t=1}^T X_{i,t} - \sum_{t=1}^T X_{I_t,t} \right]. \\ &= \mu^* T - \mathbb{E} \left[\sum_{t=1}^T X_{I_t,t} \right]. \end{aligned}$$

■ By Jensen's inequality, $\bar{R}_T \leq \mathbb{E}[R_T]$.

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 Δ_i s:

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

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:
$$\begin{aligned} \bar{R}_T &= \mu^* T - \mathbb{E} \left[\sum_{t=1}^T X_{I_t,t} \right] = \mathbb{E} \left[\sum_{t=1}^T (\mu^* - X_{I_t,t}) \right] \\ &= \mathbb{E} \left[\sum_{t=1}^T \sum_{i=1}^K (\mu^* - X_{i,t}) 1_{I_t=i} \right] = \sum_{t=1}^T \sum_{i=1}^K \mathbb{E}[(\mu^* - X_{i,t})] \mathbb{E}[1_{I_t=i}] \\ &= \sum_{i=1}^K (\mu^* - \mu_i) \mathbb{E} \left[\sum_{t=1}^T 1_{I_t=i} \right] = \sum_{i=1}^K \mathbb{E}[T_i(T)]\Delta_i. \end{aligned}$$

ϵ -Greedy Strategy

(Auer et al. 2002a)

- At time t ,
 - with probability $1 - \epsilon_t$, select arm i with best emp. mean.
 - with probability ϵ_t , select random arm.
- For $\epsilon_t = \min\left(\frac{6K}{\Delta^2 t}, 1\right)$, 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 Δ .
 - sub-optimal arms treated similarly (naive search).

UCB Strategy

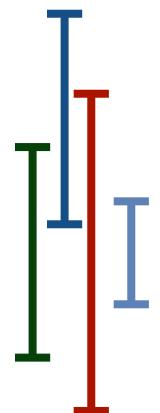
(Lai and Robbins, 1985; Agrawal 1995; Auer et al. 2002a)

■ 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 too long.

- by definition, $\mu_i \leq \mu^* \leq \text{UCB}_i$.
- pulling i often → UCB closer to μ_i .



Note on Concentration Ineqs

- Let X be a random variable such that for all $t \geq 0$,

$$\log \mathbb{E} [e^{t(X - \mathbb{E}[X])}] \leq \Psi(t),$$

where Ψ 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

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

$$I_t \in \operatorname{argmax}_{i \in [1, K]} \left[\hat{\mu}_{i,t-1} + \psi^{*-1} \left(\frac{\alpha \log t}{T_i(t-1)} \right) \right].$$

(α, ψ) -UCB Guarantee

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

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

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

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

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

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

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

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

- 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 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,

$$\begin{aligned} E[T_i(T)] &= E \left[\sum_{t=1}^T 1_{I_t=i} \right] \\ &\leq s + E \left[\sum_{t=s+1}^T 1_{I_t=i} 1_{T_i(t-1) \geq s} \right] \\ &\leq s + \sum_{t=s+1}^T \Pr[\hat{\mu}_{i,t-1} - \mu_{i,t-1} - \eta_{i,t-1} \geq 0] + \Pr[\mu^* - \hat{\mu}_{i^*,t-1} - \eta_{i^*,t-1} \geq 0]. \end{aligned}$$

Proof

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

$$\begin{aligned} & \Pr[\mu^* - \hat{\mu}_{i^*, t-1} - \eta_{i^*, t-1} \geq 0] \\ & \leq \Pr \left[\exists s \in [1, t]: \mu^* - \hat{\mu}_{i^*, s} - \psi^{*-1} \left(\frac{\alpha \log t}{s} \right) \geq 0 \right] \\ & \leq \sum_{s=1}^t \frac{1}{t^\alpha} = \frac{1}{t^{\alpha-1}}. \end{aligned}$$

- Final constant of the bound obtained by further simple calculations.

Lower Bound

(Lai and Robbins, 1985)

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

$$\liminf_{T \rightarrow +\infty} \frac{\bar{R}_T}{\log T} \geq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{D(\mu_i \parallel \mu^*)}.$$

- a more general result holds for general distributions.

Notes

- Observe that

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

$$\begin{aligned} \text{since } D(\mu_i \| \mu^*) &= \mu_i \log \frac{\mu_i}{\mu^*} + (1 - \mu_i) \log \frac{1 - \mu_i}{1 - \mu^*} \\ &\leq \mu_i \frac{\mu_i - \mu^*}{\mu^*} + (1 - \mu_i) \frac{\mu^* - \mu_i}{1 - \mu^*} \\ &= \frac{(\mu_i - \mu^*)^2}{\mu^*(1 - \mu^*)} = \frac{\Delta_i^2}{\mu^*(1 - \mu^*)}. \end{aligned}$$

Outline

- Stochastic bandits
- Adversarial bandits

Adversarial Model

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

Bandit Setting

- For $t=1$ to T do
 - player selects action $I_t \in \{1, \dots, K\}$ (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, \dots, 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, \dots, 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:

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

- By Jensen's inequality, $\bar{R}_T \leq \mathbb{E}[R_T]$.

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:

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}} \mathbf{1}_{I_t=i},$$

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

- Unbiased estimate: for any i ,

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

EXP3

EXP3(K)

(Auer et al. 2002b)

```
1    $\mathbf{p}_1 \leftarrow (\frac{1}{K}, \dots, \frac{1}{K})$ 
2    $(\tilde{L}_{1,0}, \dots, \tilde{L}_{K,0}) \leftarrow (0, \dots, 0)$ 
3   for  $t \leftarrow 1$  to  $T$  do
4       SAMPLE( $I_t \sim \mathbf{p}_t$ )
5       RECEIVE( $\ell_{I_t, t}$ )
6       for  $i \leftarrow 1$  to  $K$  do
7            $\tilde{\ell}_{i,t} \leftarrow \frac{\ell_{i,t}}{p_{i,t}} 1_{I_t=i}$ 
8            $\tilde{L}_{i,t} \leftarrow \tilde{L}_{i,t-1} + \tilde{\ell}_{i,t}$ 
9       for  $i \leftarrow 1$  to  $K$  do
10           $p_{i,t+1} \leftarrow \frac{e^{-\eta \tilde{L}_{i,t}}}{\sum_{j=1}^K e^{-\eta \tilde{L}_{j,t}}}$ 
11  return  $\mathbf{p}_{T+1}$ 
```

EXP3 (Exponential weights for Exploration and Exploitation)

EXP3 Guarantee

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

$$\bar{R}_T \leq \frac{\log K}{\eta} + \frac{\eta KT}{2}.$$

Choosing η to minimize the bound gives

$$\boxed{\bar{R}_T \leq \sqrt{2KT \log K}.}$$

- **Proof:** similar to that of EG, but we cannot use Hoeffding's inequality since $\tilde{\ell}_{i,t}$ is unbounded.

Proof

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

■ Upper bound:

$$\begin{aligned}
 \Phi_t - \Phi_{t-1} &= \log \frac{\sum_{i=1}^K e^{-\eta \tilde{L}_{i,t}}}{\sum_{i=1}^N e^{-\eta \tilde{L}_{i,t-1}}} = \log \frac{\sum_{i=1}^K e^{-\eta \tilde{L}_{i,t-1}} e^{-\eta \tilde{\ell}_{i,t}}}{\sum_{i=1}^N e^{-\eta \tilde{L}_{i,t-1}}} \\
 &= \log \left[\mathbb{E}_{i \sim \mathbf{p}_t} [e^{-\eta \tilde{\ell}_{i,t}}] \right] \\
 &\leq \mathbb{E}_{i \sim \mathbf{p}_t} [e^{-\eta \tilde{\ell}_{i,t}}] - 1 \quad (\log x \leq x - 1) \\
 &\leq \mathbb{E}_{i \sim \mathbf{p}_t} \left[-\eta \tilde{\ell}_{i,t} + \frac{\eta^2}{2} \tilde{\ell}_{i,t}^2 \right] \quad (e^{-x} \leq 1 - x + \frac{x^2}{2}) \\
 &= -\eta \mathbb{E}_{i \sim \mathbf{p}_t} [\tilde{\ell}_{i,t}] + \frac{\eta^2}{2} \mathbb{E}_{i \sim \mathbf{p}_t} \left[\frac{l_{i,t}^2 \mathbf{1}_{I_t=i}}{p_{i,t}^2} \right] \\
 &= -\eta \ell_{I_t,t} + \frac{\eta^2}{2} \frac{l_{I_t,t}^2}{p_{I_t,t}} \leq -\eta \ell_{I_t,t} + \frac{\eta^2}{2} \frac{1}{p_{I_t,t}}.
 \end{aligned}$$

Proof

■ Upper bound: summing up the inequalities yields

$$\mathbb{E}[\Phi_T - \Phi_0] \leq -\eta \mathbb{E}_{I_t \sim \mathbf{p}_t} \left[\sum_{t=1}^T \ell_{I_t, t} \right] + \mathbb{E}_{I_t \sim \mathbf{p}_t} \left[\sum_{t=1}^T \frac{\eta^2}{2p_{I_t, t}} \right] = -\eta \mathbb{E} \left[\sum_{t=1}^T \ell_{I_t, t} \right] + \frac{\eta^2 KT}{2}.$$

■ Lower bound: for all $j \in [1, K]$,

$$\begin{aligned} \mathbb{E}[\Phi_T - \Phi_0] &= \mathbb{E}_{I_t \sim \mathbf{p}_t} \left[\log \left[\sum_{i=1}^K e^{-\eta \tilde{L}_{i, T}} \right] - \log K \right] \\ &\geq -\eta \mathbb{E}_{I_t \sim \mathbf{p}_t} [\tilde{L}_{j, T}] - \log K = -\eta \mathbb{E}_{I_t \sim \mathbf{p}_t} [L_{j, T}] - \log K. \end{aligned}$$

■ Comparison:

$$\begin{aligned} \forall j \in [1, K], \quad \eta \mathbb{E} \left[\sum_{t=1}^T \ell_{I_t, t} \right] - \eta \mathbb{E}[L_{j, T}] &\leq \log K + \frac{\eta^2}{2} KT \\ \Rightarrow \bar{R}_T &\leq \frac{\log K}{\eta} + \frac{\eta KT}{2}. \end{aligned}$$

Notes

When T is not known:

- standard doubling trick.
- or, use $\eta_t = \sqrt{\frac{\log K}{Kt}}$, then $\bar{R}_T \leq 2\sqrt{KT \log K}$.

High probability bounds:

- importance weighting problem: unbounded second moment (see [\(Cortes, Mansour, MM, 2010\)](#)), $E_{i \sim p_t} [\tilde{\ell}_{i,t}^2] = \frac{\ell_{I_{t,t}}^2}{p_{I_{t,t}}}$.
- [\(Auer et al., 2002b\)](#): 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 $\tilde{\ell}_{i,t} = \frac{\ell_{i,t} 1_{I_t=i} + \beta}{p_{i,t}}$ with $\beta > 0$ a parameter to tune.

Lower Bound

(Bubeck 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 ψ 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

- R. Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Mathematics*, vol. 27, pp. 1054–1078, 1995.
- Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, vol. 11, pp. 2635– 2686, 2010.
- Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi- armed bandit problem, *Machine Learning Journal*, vol. 47, no. 2–3, pp. 235– 256, 2002a.
- Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002b.
- Sébastien Bubeck, Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* 5, 1–122, 2012.

References

- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In NIPS, 2010.
- T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.
- Gilles Stoltz. Incomplete information and internal regret in prediction of individual sequences. *Ph.D. thesis, Universite Paris-Sud*, 2005.
- R. Tyrrell Rockafellar. Convex Analysis. *Princeton University Press*, 1970.
- W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Bulletin of the American Mathematics Society*, vol. 25, pp. 285–294, 1933.