Exp 3 Algorithm

Source: Bandit Algorithms, L. & S., Chap. 11

Adversarial Setting

Bandit system evolves over discrete time steps

Two actors: Adversary chooses arm rewards
Player chooses arm to play

k arms, $j = \{1, 2, \dots, k\}$, no stochastic assumption on arm rewards.

At each time $t = 1, 2, \dots, n$:

① Adversary chooses reward for each arm j:

$$x_{tj} \in [0, 1], \quad j = 1, 2, \dots, n$$

arbitrary / no assumptions on the way these are chosen

equivalently chooses loss for each arm j:

$$y_{tj} \in [0, 1], \quad j = 1, 2, \dots, n$$

$$y_{tj} = (1 - x_{tj})$$

② Player chooses action $A_t \in \{1, 2, \cdots, k\}$.
A_t chosen in a two step process:

ⓐ A policy $\pi = (\pi_1, \pi_2, \cdots, \pi_n)$ (time) is fixed before start of play, where

$\pi_t : (A_1, X_1, A_2, X_2, \cdots, A_{t-1}, X_{t-1}) \longrightarrow P_t$

(history of action, rewards)

A_1: action at time 1; X_1: reward of selected arm at time 1

$P_t \in \mathcal{P}_{k-1}$

P_t: a probability dist. over $\{1, 2, \cdots, k\}$. i.e., $q \in \mathcal{P}_{k-1} \Rightarrow q = (q^{(1)}, \cdots, q^{(k)})$ with $q^{(j)} \geq 0$, $\sum_{j=1}^{k} q^{(j)} = 1$.

$\mathcal{P}_{k-1}$: probability simplex over k dimensions

$$x_s = \sum_{j=1}^{k} x_{sj} \, \chi_{\{A_s = j\}}$$

- P_t is a function of all observations of actions and reward / loss up to and including time $(t-1)$.

- The adversary knows the selection function, i.e., adversary knows P_t before choosing $(x_{t1}, \cdots, x_{tk})$ at time t.

(b) Player has access to secret random bits not known to adversary. Player uses these random bits to choose action according to policy $P_t = (P_{t1}\ P_{t2}\ \cdots\ P_{tk})$, i.e.,

$$P(A_t = j \mid A_1, X_1, \cdots, A_{t-1}, X_{t-1}) = P_{tj}$$

Thus adversary knows player policy but does not know player action, when deciding arm rewards.

(3) Bandit feedback: Player gets to observe the reward/loss according to the specific action at time t, i.e.,
if $A_t = j$, then player gets to only observe x_{tj} (or y_{tj}).

(Other arms' rewards are NOT revealed).

Regret: In the adversarial setting, regret is with respect to the best fixed arm policy in hindsight.

policy chosen by player

$$R_n(\pi, x) = \left(\max_{1 \le j \le k} \sum_{t=1}^{n} x_{tj} \right) - E\left[\sum_{t=1}^{n} x_{t A_t} \right]$$

reward

environment x

highest fixed arm reward in hindsight

players reward with $E[\cdot]$ over randomization by player.

$= \text{sequence } \{ x_{tj},\ j = 1, \dots, k,\ t = 1, \dots n \}$

Chosen by adversary

Equivalently,

$$R_n(\pi, y) = E\left[\sum_{t=1}^{n} y_{t A_t} \right] - \left(\min_{1 \le j \le k} \sum_{t=1}^{n} y_{tj} \right)$$

loss environment y

player's loss

fixed arm with smallest loss in hindsight.

Aside: Why compare to best fixed arm in hindsight? One motivation is in the setting of classification, where the goal is to use n labeled samples, and return a weight vector characterizing a classifier.

Consider an algorithm that optimizes for the ERM (Empirical Risk Minimization) criterion.

It can be shown that developing an online learning algorithm that optimizes with the "best fixed arm in hindsight" criterion automatically leads to tight generalization bounds.

We will discuss this formally later when studying OCO (Online Convex Optimization).

Ref: Chap 5 — Online-to-Batch Conversions from the monograph: Online Learning and Online Convex Optimization, by Shai Shalev-Shwartz, NOW Publishers, Vol. 4, No. 2, 2011.

Exp 3 Algorithm:

→ Exponential weight algorithm for Explore and Exploit.

Idea 1: Suppose player had access to the rewards for all arms. Then, exponentially boost the probability of that arm

that has been best in hindsight at time t.

i.e., suppose all $\{y_{sj},\ s=1,2,\dots,t,\ j=1,\dots,k\}$ were revealed. Then, cumulative loss from arm j is $L_{t-1,j} = \sum_{s=1}^{t-1} y_{sj}$

Thus, choose arm j at time t

$$\propto e^{-\eta L_{t-1,j}}$$

(or equivalently, $P_{tj} \propto e^{+\eta S_{t-1,j}}$) → cumulative reward until time $(t-1)$

Idea 2: We do not know $\{y_{tj},\ j=1,\dots,k\}$ because of bandit feedback.

Thus, use Idea 1 (exp boost best in hindsight) using conditionally unbiased estimators the unknown y_{tj}.

(Seems strange, as we have no model for y_{tj}, so how can we create an estimator?)

Importance Sampling Estimators for $\{x_{tj}, y_{tj}\}$.

reward of arm A_1 $\quad x_s = \sum_{j=1}^{k} x_{sj} \chi_{\{A_s=j\}}$

Let

$$I_{t-1} = (A_1, X_1, A_2, X_2, \cdots, A_{t-1}, X_{t-1})$$

or

$$= (A_1, Y_1, A_2, Y_2, \cdots, A_{t-1}, Y_{t-1})$$

$y_s = \sum_{j=1}^{k} y_{sj} \chi_{\{A_s=j\}}$.

(we are misusing the same notation for both as $Y_t = (1 - X_t)$, so conditioning on either is equivalent, and is usually clear from context which we are using).

$$P_{tj} = P(A_t = j \mid I_{t-1}) \quad \text{r.v.}$$

(the algo. we consider satisfies $P_{tj} > 0$ a.s., $\forall j = 1, 2, \cdots, k$, $t = 1, 2, \cdots, n$)

The estimators:

$$\hat{X}_{tj} = \frac{\chi_{\{A_t=j\}} \cdot X_t}{P_{tj}} \qquad \hat{Y}_{tj} = \frac{\chi_{\{A_t=j\}} Y_t}{P_{tj}}$$

Now, $E[\cdot]$ w.r.t. to the randomness of the player's actions

$$E\left[\hat{x}_{tj} \mid I_{t-1}\right] = E\left[\frac{\chi_{\{A_t=j\}}\, x_t}{p_{tj}} \,\middle|\, I_{t-1}\right]$$

because p_{tj} is a known fn of I_{t-1}

$$= \frac{1}{p_{tj}} E\left[\chi_{\{A_t=j\}}\, x_t \mid I_{t-1}\right] \Leftarrow$$

Next, observe that given I_{t-1}, the action A_t and reward vector x_t are independent. This is because the adversary has access only to the player's policy (a deterministic function of I_{t-1}), but not the random bits the player uses to choose the action.

$$\therefore E\left[\hat{x}_{tj} \mid I_{t-1}\right] = \frac{1}{p_{tj}} \underbrace{E\left[\chi_{\{A_t=j\}} \mid I_{t-1}\right]}_{= p_{tj}} x_{tj}$$

x_{tj}: arbit. reward chosen by adv.

$$= x_{tj}$$

Similarly, $\boxed{E\left[\hat{y}_{tj} \mid I_{t-1}\right] = y_{tj}}$

Aside: Consider the conditional variances of these estimators.

$$V_t\left[\hat{X}_{tj}^2\right] = E\left[\hat{X}_{tj}^2 \,\middle|\, I_{t-1}\right] - x_{tj}^2$$

$$= E\left[\frac{x_{\{A_t=j\}} X_t^2}{P_{tj}^2} \,\middle|\, I_{t-1}\right] - x_{tj}^2$$

$$= \frac{x_{tj}^2 (1 - P_{tj})}{P_{tj}}$$

→ good if rewards are small

Similarly,

$$V_t\left[\hat{Y}_{tj}^2\right] = \frac{y_{tj}^2 (1 - P_{tj})}{P_{tj}}$$

→ good if losses are small

we will use this in the proofs

The Exp 3 Algorithm (sec. 11.3 in text)

The discussion below is based on losses. Please read text for the one based on rewards.

Let $\hat{L}_{t,j} = \sum_{s=1}^{t} \hat{Y}_{sj}$, $j = 1, 2, \ldots, k$
$t = 1, 2, 3, \ldots, n$

and $\hat{L}_{0,j} = 0$

$$\hat{Y}_{sj} = \frac{\chi_{\{A_s = j\}} Y_s}{P_{sj}} = \frac{\chi_{\{A_s = j\}} (1 - X_s)}{P_{sj}}$$

Algorithm input: $\eta > 0$

For each time $t = 1, 2, \ldots, n$:

- $P_{tj} = \left(\frac{e^{-\eta \hat{L}_{t-1,j}}}{\sum_{r=1}^{k} e^{-\eta \hat{L}_{t-1,r}}} \right)$, $j = 1, 2, \ldots, k$ (arms.)
- Sample $A_t \sim P_{tj}$ (using indep/secret randomness)
- Update $\hat{L}_{t,j} = \hat{L}_{t-1,j} + \hat{Y}_{tj}$

Thm (11.1 in text for rewards; below mirrors the proof for loss).

Setting as above. Then for $\eta = \sqrt{\frac{\ln k}{nk}}$,

$$R_n(\pi, y) \leq 2\sqrt{nk \ln(k)}$$

Proof: Fix any arm i, and we will analyze the regret with respect to this fixed arm. As i is arbitrary, the analysis bounds the regret wrt any fixed arm (note this is diff. from regret decomposition we have been studying for stochastic bandits).

$$\rightarrow \quad R_{ni} = E\left[\sum_{t=1}^{n} Y_t\right] - \sum_{t=1}^{n} y_{ti}$$

Proof has 3 main steps, marked (A), (B), (C) below.

---------- (A): Express Regret wrt Estimates ----------

$$E\left[\hat{L}_{ni}\right] = E\left[\sum_{s=1}^{n} \hat{Y}_{si}\right]$$

$$= E\left[\sum_{s=1}^{n} E\left[\hat{Y}_{si} \mid I_{s-1}\right]\right]$$

$$E[\hat{\Sigma}_{ni}] = E\left[\sum_{s=1}^{n} y_{si}\right] = \sum_{s=1}^{n} y_{si} \quad \rightarrow (1)$$

$$E\left[Y_t \mid I_{t-1}\right] = \sum_{j=1}^{k} p_{tj}\, y_{tj}$$

$$= \sum_{j=1}^{k} p_{tj}\, E\left[\hat{Y}_{tj} \mid I_{t-1}\right]$$

$\rightarrow (2)$

Next, let

$$\hat{\Sigma}_n = \sum_{t=1}^{n} \sum_{j=1}^{k} p_{tj}\, \hat{Y}_{tj}$$

Then $E[\hat{\Sigma}_n] = E\left[\sum_{t=1}^{n} E\left[\sum_{j=1}^{k} p_{tj}\hat{Y}_{tj} \mid I_{t-1}\right]\right]$

$$\overset{(2)}{=} E\left[\sum_{t=1}^{n} E\left[Y_t \mid I_{t-1}\right]\right]$$

(Tower rule used in several steps) $\rightarrow$

$$= E\left[\sum_{t=1}^{n} Y_t\right] \quad \rightarrow (3)$$

Thus, $R_{ni} = E\left[\sum_{t=1}^{n} Y_t\right] - \sum_{t=1}^{n} y_{ti}$

$$\therefore \quad \boxed{R_{ni} \overset{(3)(1)}{=} E\left[\hat{L}_n\right] - E\left[\hat{L}_{ni}\right]} \quad \rightarrow (A), \rightarrow (4)$$

(Note: We will use $\hat{L}_{n,i} = \hat{L}_{ni}$ interchangeably)

---------- Step (B): Track $(\hat{L}_n - \hat{L}_{ni})$ ----------

Let $W_t = e^{\eta t} \cdot \sum_{j=1}^{k} e^{-\eta \hat{L}_{tj}}$, $t = 0, 1, 2, \cdots, n$

with $W_0 = k$ $\left(\because \hat{L}_{0j} = 0 \text{ by definition}\right)$

selected fixed arm

$$e^{\eta n} e^{-\eta \hat{L}_{ni}} \leq \sum_{j=1}^{k} e^{-\eta \hat{L}_{nj}} e^{\eta n}$$

$$= W_n = W_0 \left(\frac{W_1}{W_0}\right) \cdots \left(\frac{W_n}{W_{n-1}}\right)$$

$\rightarrow (5)$

$$= k \prod_{t=1}^{n} \left(\frac{W_t}{W_{t-1}}\right)$$

Now,

$$\left(\frac{W_t}{W_{t-1}}\right) = \frac{e^{\eta t} \sum_{j=1}^{k} e^{-\eta \hat{L}_{tj}}}{e^{\eta(t-1)} \sum_{j=1}^{k} e^{-\eta \hat{L}_{t-1,j}}} = e^{\eta} \sum_{j=1}^{k} \frac{e^{-\eta \hat{L}_{t-1,j}}}{\sum_{r=1}^{k} e^{-\eta \hat{L}_{t-1,r}}} \cdot e^{-\eta \hat{Y}_{tj}}$$

$$= \sum_{j=1}^{k} p_{tj} \, e^{\eta(1 - \hat{Y}_{tj})}$$

Now, observe that:

(a) For $x \leq 1$, $(1+x) \overset{(i)}{\leq} e^x \overset{(ii)}{\leq} 1 + x + x^2$.

(b) $\hat{Y}_{tj} = \dfrac{\mathbb{1}_{\{A_t = j\}} Y_t}{P_{tj}} = \begin{cases} 0 & \text{if } A_t \neq j \\ \dfrac{y_{tj}}{P_{tj}} & \text{if } A_t = j \end{cases}$

$Y_t \in [0,1]$, $0 < P_{tj} < 1$

$\Rightarrow (1 - \hat{Y}_{tj}) \leq 1$

$$\therefore \frac{W_t}{W_{t-1}} = \sum_{j=1}^{k} P_{tj} \cdot e^{\eta(1-\hat{Y}_{tj})}$$

$$\overset{(ii)}{\leq} \sum_{j=1}^{k} P_{tj} \cdot \left(1 + \eta(1-\hat{Y}_{tj}) + \eta^2 (1-\hat{Y}_{tj})^2\right)$$

$$= \left(1 + \eta \sum_{j=1}^{k} P_{tj}(1-\hat{Y}_{tj}) + \eta^2 \sum_{j=1}^{k} P_{tj} \cdot (1-\hat{Y}_{tj})^2\right)$$

$$\overset{(i)}{\leq} e^{\eta \sum_{j=1}^{k} P_{tj}(1-\hat{Y}_{tj}) + \eta^2 \sum_{j=1}^{k} P_{tj}(1-\hat{Y}_{tj})^2}$$

$$\therefore \quad e^{n\eta} \cdot e^{-\eta \hat{L}_{ni}} \leq K \prod_{t=1}^{n} \left(\frac{W_t}{W_{t-1}} \right) \qquad \text{(from (5))}$$

$$\leq e^{\ln K + \eta \sum_{t=1}^{n} \sum_{j=1}^{K} p_{tj} (1 - \hat{y}_{tj}) + \eta^2 \sum_{t=1}^{n} \sum_{j=1}^{K} p_{tj} (1 - \hat{y}_{tj})^2}$$

$n - \hat{L}_n$ (see below)

$$\therefore \quad (n - \hat{L}_{ni}) \leq \frac{\ln K}{\eta} + \sum_{t=1}^{n} \sum_{j=1}^{K} p_{tj} (1 - \hat{y}_{tj}) + \eta \sum_{t=1}^{n} \sum_{j=1}^{K} p_{tj} (1 - \hat{y}_{tj})^2$$

$\log(\cdot)$ and divide by η.

Recall: $\hat{L}_n = \sum_{t=1}^{n} \sum_{j=1}^{K} p_{tj} \hat{y}_{tj}$, and note that $\sum_{t=1}^{n} \sum_{j=1}^{K} p_{tj} = n$ → (B)

$$\Rightarrow \boxed{(\hat{L}_n - \hat{L}_{ni}) \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{n} \sum_{j=1}^{K} p_{tj} (1 - \hat{y}_{tj})^2}$$

$E[\cdot] = R_{ni}$

→ (6)

⑥: Bound the expectation of RHS in ⑥:

$$E\left[\sum_{t=1}^{n}\sum_{j=1}^{k} p_{tj}\left(1-\hat{Y}_{tj}\right)^2\right]$$

$$= E\left[\sum_{t=1}^{n}\sum_{j=1}^{k} p_{tj}\left(1-\frac{\chi_{\{A_t=j\}}\, Y_t}{p_{tj}}\right)^2\right]$$

$$= \sum_{t=1}^{n} E\left[\sum_{j=1}^{k} p_{tj}\left(1+\frac{\chi_{\{A_t=j\}}\, Y_t^2}{p_{tj}^2}-\frac{2\chi_{\{A_t=j\}}\, Y_t}{p_{tj}}\right)\right]$$

$$= \sum_{t=1}^{n} E\left[1+\sum_{j=1}^{k} E\left[\frac{\chi_{\{A_t=j\}}\, Y_t^2}{p_{tj}}-2\chi_{\{A_t=j\}}\, Y_t \,\middle|\, I_{t-1}\right]\right]$$

Note: $\chi_{\{A_t=j\}}\, Y_t = \chi_{\{A_t=j\}}\, Y_{tj}$

$$= \sum_{t=1}^{n} E\left[1+E\left[\sum_{j=1}^{k}\frac{\chi_{\{A_j=j\}}\, Y_t^2}{p_{tj}}\,\middle|\, I_{t-1}\right]\right.$$

$$\left.-2E\left[\sum_{j=1}^{k}\chi_{\{A_t=j\}}\, Y_t \,\middle|\, I_{t-1}\right]\right]$$

$\because\ E\left[\chi_{\{A_t=j\}}\,\middle|\, I_{t-1}\right]=p_{tj}$

$$\sum_{j=1}^{k}\chi_{\{A_t=j\}}\, Y_t = Y_t$$

$$= \sum_{t=1}^{n} E\left[1 - 2\gamma_t + E\left[\sum_{j=1}^{k} \frac{\chi_{\{A_t=j\}}\, y_{tj}^2}{p_{tj}} \,\middle|\, I_{t-1}\right]\right]$$

$$= \sum_{t=1}^{n} E\left[1 - 2\gamma_t + \sum_{j=1}^{k} y_{tj}^2\right]$$

$$= \sum_{t=1}^{n} E\left[1 - 2\gamma_t + \gamma_t^2 - \gamma_t^2 + \sum_{j=1}^{k} y_{tj}^2\right]$$

$$= \sum_{t=1}^{n} E\left[\underbrace{(1-\gamma_t)^2}_{\leq 1} + \underbrace{\sum_{\substack{j=1 \\ j \neq A_t}}^{k} y_{tj}^2}_{\leq (k-1)}\right] \leq nk$$

$$\therefore \quad \boxed{E\left[\sum_{t=1}^{n} \sum_{j=1}^{k} p_{tj}\left(1-\hat{\gamma}_{tj}\right)^2\right] \leq nk} \rightarrow (c)$$

Substituting in (6) and taking expectations,

$$R_{ni} \leq \frac{\ln k}{\eta} + \eta nk = 2\sqrt{nk \ln(k)}$$

$$\eta = \sqrt{\ln k / nk}$$

QED

Comment: We can tighten the constant by $\sqrt{2}$,

i.e., $R_n \leq \sqrt{2nk\ln(k)}$

Please read text, Thm 11.2 for the proof.

Note: Instead of bandit feedback, suppose we had full feedback, i.e., the player gets to see rewards/losses for all arms (but incurs the one that the player chose to play).

$\longrightarrow$ Hedge algorithm $\longrightarrow$ A decision-theoretic generalization of online learning and an application to boosting, Freund and Schapire, 1997

Then regret $\leq \sqrt{\frac{n\ln(k)}{2}}$

* No $\sqrt{k}$ term! $\longrightarrow$ The gain due to full information

(See Note 3, Chap 11 in text for discussion.)

To see this scaling, recall (5):

$$\left(\hat{L}_n - \hat{L}_{ni}\right) \le \frac{\ln k}{\eta} + \eta \sum_{t=1}^{n} \sum_{j=1}^{k} p_{tj} \left(1 - \hat{y}_{tj}\right)^2$$

$E[\cdot] = R_{ni}$

Suppose $\hat{y}_{tj} = y_{tj}$ (i.e., no noise in the estimator because of full information)

Then

$$\sum_{t=1}^{n} \sum_{j=1}^{k} p_{tj} \left(1 - \hat{y}_{tj}\right)^2 = \sum_{t=1}^{n} \sum_{j=1}^{k} p_{tk} \underbrace{\left(1 - y_{tj}\right)^2}_{\le 1}$$

$$\le n \qquad \le 1$$

i.e., bound is $\left(\frac{\ln k}{\eta} + \eta n\right)$. Choose

$\eta = \sqrt{\frac{\ln k}{n}}$ to balance terms $\Rightarrow$

$$R_n \le 2\sqrt{n \ln k}$$

(Analysis can be sharpened to provide tighter constant)