1 Goal: "Earn Almost as Much as the Best Expert"

We consider a contextual/adversarial bandit protocol over rounds $t = 1, \dots, n$ with k arms.

- At each round t, the learner chooses an arm $A_t \in [k]$ according to a distribution P_t over arms.
- The environment assigns rewards $\{x_{t,i}\}_{i=1}^k$ with $x_{t,i} \in [0,1]$.
- The learner observes only the realized reward x_{t,A_t} (bandit feedback).

1.1 Experts

There are M experts indexed by $m \in [M]$. Expert m outputs at each round t a *mixed recommendation* over arms:

$$E^{(t)}_{m,\cdot} \in \Delta_k, \qquad E^{(t)}_{m,i} \geq 0, \quad \sum_{i=1}^k E^{(t)}_{m,i} = 1.$$

1.2 Performance benchmark (regret)

Define the *true expected reward* of expert m at round t:

$$X_{t,m} := \sum_{i=1}^k E^{(t)}_{m,i}\, x_{t,i}. \tag{1}$$

The learner obtains reward x_{t,A_t}. The (random) regret against the best fixed expert in hindsight is

$$R_n := \max_{m \in [M]} \sum_{t=1}^n X_{t,m} \; - \; \sum_{t=1}^n x_{t,A_t}. \tag{2}$$

Key difficulty. In the bandit setting, the learner cannot observe all $\{x_{t,i}\}$, hence cannot directly compute $X_{t,m}$.

2 The Only Honest Substitute Under Bandit Feedback: IPW Estimators

Let $\mathcal{F}_t$ denote the history up to just *before* sampling A_t.

2.1 IPW (importance-weighted) estimator for arm rewards

Since only x_{t,A_t} is observed, define for each arm i the estimator

$$\hat{x}_{t,i} := \frac{x_{t,A_t}\, \mathbf{1}\{A_t = i\}}{P_{t,i}}. \tag{3}$$

Lemma 2.1 (Unbiasedness of IPW). *Conditioned on $\mathcal{F}_t$, for every arm i:*

$$\mathbb{E}[\hat{x}_{t,i} \mid \mathcal{F}_t] = x_{t,i}.$$

Proof.

$$\mathbb{E}[\hat{x}_{t,i} \mid \mathcal{F}_t] = \sum_{a=1}^k P_{t,a} \cdot \frac{x_{t,a}\, \mathbf{1}\{a = i\}}{P_{t,i}} = P_{t,i} \cdot \frac{x_{t,i}}{P_{t,i}} = x_{t,i}.$$

□

2.2 Estimated expert reward

Define the estimated reward of expert m by

$$\hat{X}_{t,m} := \sum_{i=1}^{k} E^{(t)}_{m,i}\, \hat{x}_{t,i}. \tag{4}$$

Lemma 2.2 (Unbiasedness of estimated expert reward). *Conditioned on $\mathcal{F}_t$, for every expert m:*

$$\mathbb{E}[\hat{X}_{t,m} \mid \mathcal{F}_t] = X_{t,m}.$$

Proof.

$$\mathbb{E}[\hat{X}_{t,m} \mid \mathcal{F}_t] = \sum_{i=1}^{k} E^{(t)}_{m,i}\, \mathbb{E}[\hat{x}_{t,i} \mid \mathcal{F}_t] = \sum_{i=1}^{k} E^{(t)}_{m,i}\, x_{t,i} = X_{t,m}.$$

□

3 Why the Exponential Multiplicative Update Appears

Let $w_{t,m} > 0$ be the unnormalized weight of expert m at round t, and define

$$W_t := \sum_{m=1}^{M} w_{t,m}, \qquad Q_{t,m} := \frac{w_{t,m}}{W_t}.$$

The EXP4 update is

$$w_{t+1,m} = w_{t,m} \exp(\eta \hat{X}_{t,m}), \tag{5}$$

where $\eta > 0$ is a learning rate.

3.1 Log-potential and softmax form

Taking logs in (5) yields

$$\log w_{t+1,m} = \log w_{t,m} + \eta \hat{X}_{t,m}, \quad \Rightarrow \quad \log w_{n+1,m} = \log w_{1,m} + \eta \sum_{t=1}^{n} \hat{X}_{t,m}.$$

Also,

$$Q_{t+1,m} = \frac{Q_{t,m} e^{\eta \hat{X}_{t,m}}}{\sum_{j=1}^{M} Q_{t,j} e^{\eta \hat{X}_{t,j}}}.$$

4 Potential Sandwich: Two Inequalities (L1, L2)

4.1 L1: Lower bound via a fixed expert

For any expert $m^\star$,

$$W_{t+1} = \sum_{m=1}^{M} w_{t,m} e^{\eta \hat{X}_{t,m}} \ge w_{t,m^\star} e^{\eta \hat{X}_{t,m^\star}}.$$

Thus,

$$\log W_{n+1} \ge \log w_{1,m^\star} + \eta \sum_{t=1}^{n} \hat{X}_{t,m^\star}. \tag{6}$$

4.2 L2: Upper bound via log-sum-exp / MGF control

We have

$$\frac{W_{t+1}}{W_t} = \sum_{m=1}^{M} Q_{t,m} e^{\eta \hat{X}_{t,m}}, \qquad \log \frac{W_{t+1}}{W_t} = \log\Big(\sum_{m=1}^{M} Q_{t,m} e^{\eta \hat{X}_{t,m}}\Big).$$

A standard exponential-moment bound yields an inequality of the form

$$\log \frac{W_{t+1}}{W_t} \le \eta \sum_{m=1}^{M} Q_{t,m} \hat{X}_{t,m} + \frac{\eta^2}{2} \sum_{m=1}^{M} Q_{t,m} (1 - \hat{X}_{t,m})^2. \tag{7}$$

Summing over $t = 1, \dots, n$ gives

$$\log W_{n+1} - \log W_1 \le \eta \sum_{t=1}^{n} \sum_{m=1}^{M} Q_{t,m} \hat{X}_{t,m} + \frac{\eta^2}{2} \sum_{t=1}^{n} \sum_{m=1}^{M} Q_{t,m} (1 - \hat{X}_{t,m})^2. \tag{8}$$

4.3 Estimated regret skeleton

Combining (6) and (8) yields

$$\sum_{t=1}^{n} \hat{X}_{t,m^\star} - \sum_{t=1}^{n} \sum_{m=1}^{M} Q_{t,m} \hat{X}_{t,m} \le \frac{\log(W_1/w_{1,m^\star})}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \sum_{m=1}^{M} Q_{t,m} (1 - \hat{X}_{t,m})^2. \tag{9}$$

If $w_{1,m} = 1$ for all m, then $W_1 = M$ and $\log(W_1/w_{1,m^\star}) = \log M$.

5 From $\hat{X}$ to True Regret: Unbiasedness Step

Taking expectations and using Lemma 2.2,

$$\mathbb{E}\Big[\sum_{t=1}^{n} \hat{X}_{t,m^\star}\Big] = \mathbb{E}\Big[\sum_{t=1}^{n} X_{t,m^\star}\Big].$$

Also,

$$\sum_{m=1}^{M} Q_{t,m} \hat{X}_{t,m} = \sum_{i=1}^{k} \Big(\sum_{m=1}^{M} Q_{t,m} E^{(t)}_{m,i}\Big) \hat{x}_{t,i}.$$

Define the learner's arm distribution (EXP4 mixture)

$$P_{t,i} := \sum_{m=1}^{M} Q_{t,m} E^{(t)}_{m,i}. \tag{10}$$

Then

$$\sum_{m=1}^{M} Q_{t,m} \hat{X}_{t,m} = \sum_{i=1}^{k} P_{t,i} \hat{x}_{t,i}.$$

By Lemma 2.1,

$$\mathbb{E}\big[\sum_{i=1}^{k} P_{t,i} \hat{x}_{t,i} \mid \mathcal{F}_t\big] = \sum_{i=1}^{k} P_{t,i} x_{t,i} = \mathbb{E}[x_{t,A_t} \mid \mathcal{F}_t].$$

Thus,

$$\mathbb{E}\Big[\sum_{t=1}^{n}\sum_{m} Q_{t,m}\hat{X}_{t,m}\Big] = \mathbb{E}\Big[\sum_{t=1}^{n} x_{t,A_t}\Big].$$

So taking expectation in (9) gives

$$\mathbb{E}\Big[\sum_{t=1}^{n} X_{t,m^\star} - \sum_{t=1}^{n} x_{t,A_t}\Big] \le \frac{\log M}{\eta} + \frac{\eta}{2}\sum_{t=1}^{n}\mathbb{E}\Big[\sum_{m=1}^{M} Q_{t,m}(1-\hat{X}_{t,m})^2\Big]. \tag{11}$$

Choosing $m^\star$ as the best expert in hindsight makes the left-hand side $\mathbb{E}[R_n]$.

6 Miracle Cancellation: Variance Term Collapses to k

Work with losses $y_{t,i} := 1 - x_{t,i} \in [0,1]$. Define

$$\hat{y}_{t,i} := \frac{y_{t,A_t}\,\mathbf{1}\{A_t = i\}}{P_{t,i}}, \qquad \hat{Y}_{t,m} := \sum_{i=1}^{k} E^{(t)}_{m,i}\hat{y}_{t,i}.$$

Since only A_t is observed,

$$\hat{Y}_{t,m} = \frac{E^{(t)}_{m,A_t}\,y_{t,A_t}}{P_{t,A_t}}.$$

Conditioned on $\mathcal{F}_t$,

$$\mathbb{E}[\hat{Y}^2_{t,m} \mid \mathcal{F}_t] = \sum_{i=1}^{k} P_{t,i}\Big(\frac{E^{(t)}_{m,i}y_{t,i}}{P_{t,i}}\Big)^2 = \sum_{i=1}^{k}\frac{(E^{(t)}_{m,i})^2 y^2_{t,i}}{P_{t,i}} \le \sum_{i=1}^{k}\frac{E^{(t)}_{m,i}}{P_{t,i}}. \tag{12}$$

Averaging over $m \sim Q_t$:

$$\sum_{m=1}^{M} Q_{t,m}\mathbb{E}[\hat{Y}^2_{t,m} \mid \mathcal{F}_t] \le \sum_{i=1}^{k}\frac{\sum_{m=1}^{M} Q_{t,m}E^{(t)}_{m,i}}{P_{t,i}} = \sum_{i=1}^{k}\frac{P_{t,i}}{P_{t,i}} = k. \tag{13}$$

7 Final Bound and Optimizing the learning rate

Thus,

$$\mathbb{E}[R_n] \le \frac{\log M}{\eta} + \frac{\eta}{2}nk.$$

Optimizing gives $\eta^\star = \sqrt{\frac{2\log M}{nk}}$ and

$$\mathbb{E}[R_n] \le \sqrt{2nk\log M}.$$