ONLINE LEARNING / ADAPTIVE DECISION MAKING

Set of actions X (e.g. $X = \{1, \dots, k\}$, $X \subseteq \mathbb{R}^d$)

For $t = 1 : T$
- <ins>Adversary</ins> picks loss $\ell_t \in \mathcal{F} : X \to \mathbb{R}$ <ins>unknown</ins>
- DM picks $x_t \in X$
- DM observes <ins>something</ins> about ℓ_t

Goal: $\sum_{t=1}^{T} \ell_t(x_t) \to \min!$

What does DM observe?

- <ins>Full information setting</ins>: ℓ_t
- <ins>Bandit setting</ins>: $\ell_t(x_t)$

<ins>Performance metric</ins>

Regret $R_T = \sum_{t=1}^{T} \ell_t(x_t) - \min_{x \in X} \sum_t \ell_t(x)$

Want $R_T / T \to 0$

"average excess loss compared to best fixed action in hindsight vanishes"

<ins>Ex</ins> X = set of experts / classifiers

$h_1, \dots, h_k$

At each round, receive data point, and predictions $y_t(1) \dots y_t(k)$ for the k experts

$y_t(x) \in \{0, 1\}$

We predict y_t, <ins>then</ins> observe y_t^*

$$\ell_t(x) = \begin{cases} 0 & \text{if } y_t(x) = y_t^* \\ 1 & \text{o/w.} \end{cases}$$

"full information" <ins>setting</ins>

FULL INFORMATION

Warmup: Assume $X = \{1..k\}$

$\ell_t(x) \in \{0,1\} \quad \forall x, t$

$\exists x : \ell_t(x) = 0 \quad \forall t$

Halving alg. maintain weights per expert

$w_1(x) = 1 \quad \forall x \in X$

At time t: let

$\forall y \in \{+1,-1\}: W_t(y) := \sum_{x : y_t(x) = y} w_t(x)$ total weight of all experts voting for y

Predict weighted majority

$\hat{y}_t := \arg\max_{y \in \{+1,-1\}} W_t(y)$

Pick some x_t who predicted $\hat{y}_t$

Recieve y_t^*

$\forall x$: set $w_{t+1}(x) = \begin{cases} 0 & \text{if } \hat{y}_t \neq y_t^* \\ w_t(x) \end{cases}$

Claim: $R_T \leq \lceil \log_2 k \rceil$

Proof: Each round either: no mistake $(\ell_t(x_t) = 0)$

or $\sum_{x=1}^{k} w_{t+1}(x) \leq \frac{1}{2} \sum_{x=1}^{k} w_t(x)$

$\square$

More generally

What if <u>no</u> expert is always correct?

What if $\ell_t(x) \in [0,1]$?

Multiplicative weights / Hedge:

Init $w_1(x) = 1 \quad \forall x$

for $t = 1 : T$

$$p_t(x) = \frac{w_t(x)}{\sum_{x'} w_t(x')}$$

pick $x_t \sim p_t$

Incur loss $\ell_t(x_t)$

$$w_{t+1}(x) = w_t(x) \cdot \exp(-\varepsilon\, \ell_t(x)) \quad \forall x$$

If $\ell_t \in \{0,1\}$ for "incorrect" experts, reduce weight $\exp(-\varepsilon) \approx (1-\varepsilon)$

<u>Theorem</u>

$$\mathbb{E}[R_T] \leq \frac{\log k}{\varepsilon} + \underbrace{\varepsilon \sum_{t=1}^{T} \underbrace{p_t^T \ell_t^2}_{\sum_{x=1}^{k} p_t(x)\, \ell_t(x)^2}}_{\leq\, \varepsilon \cdot T \text{ if } \ell_t(x) \in [0,1]}$$

for $\varepsilon := \frac{\sqrt{\log k}}{\sqrt{T}}$ $\overset{\text{Assuming } \ell_t \in [0,1]}{\Longrightarrow}$ $\mathbb{E}[R_T] \leq 2\sqrt{T \log k}$

$$\Rightarrow \mathbb{E}\left[\frac{R_T}{T}\right] \leq \frac{2\sqrt{\log k}}{\sqrt{T}}$$

<u>**Proof**</u> $\Phi_t = \sum_{x=1}^{k} w_t(x)$. Then $\Phi_1 = k$

$$
\begin{aligned}
\Phi_{t+1} &= \sum_{x=1}^{k} w_t(x) \cdot \exp(-\varepsilon \ell_t(x)) \\
&= \Phi_t \sum_{x=1}^{k} p_t(x) \exp(-\varepsilon \ell_t(x)) \\
&\le \Phi_t \sum_{x} p_t(x) \left(1 - \varepsilon \ell_t(x) + \varepsilon^2 \ell_t^2(x)\right) \\
&= \Phi_t \left(1 - \varepsilon p_t^T \ell_t + \varepsilon^2 p_t^T \ell_t^2\right) \\
&\le \Phi_t \exp\left(-\varepsilon p_t^T \ell_t + \varepsilon^2 p_t^T \ell_t^2\right) \\
&\le \underbrace{\Phi_1}_{k} \exp\left(-\varepsilon \sum_{t=1}^{T} p_t^T \ell_t + \varepsilon^2 \sum_{t=1}^{T} p_t^T \ell_t^2\right)
\end{aligned}
$$

Note: $p_t(x) = \frac{w_t(x)}{\Phi_t}$

$e^{-x} \le 1 - x + x^2$ for $x \ge 0$

$e^x \ge 1 + x \quad \forall x$

For best expert x^*, $w_T(x^*) = \exp\left(-\varepsilon \sum_{t=1}^{T} \ell_t(x^*)\right)$

Thus: $w_T(x^*) \le \Phi_T \le k \cdot \exp\left(-\varepsilon \sum_{t=1}^{T} p_t^T \ell_t + \varepsilon^2 \sum_{t=1}^{T} p_t^T \ell_t^2\right)$

$$\overset{\log}{\Rightarrow} -\varepsilon \sum_{t=1}^{T} \ell_t(x^*) \le \log k - \varepsilon \sum_{t=1}^{T} p_t^T \ell_t + \varepsilon^2 \sum_{t=1}^{T} p_t^T \ell_t^2$$

$$\Rightarrow \mathbb{E}[R_T] = \sum_{t=1}^{T} p_t^T \ell_t - \sum_{t=1}^{T} \ell_t(x^*) \le \frac{\log k}{\varepsilon} + \varepsilon \sum_{t=1}^{T} p_t^T \ell_t^2$$

$\square$

Bandit Setting

DM only observes $\ell_t(x_t)$

$\hookrightarrow$ exploration - exploitation

Key idea: Reduction to full information setting by constructing estimate of $\ell_t(x)$

Ex. Recommender systems

X = set of k possible recommend.

At time t, user arrives

Pick x_t, if user interested in x_t, user watches/clicks/..

$$\ell_t(x_t) = \begin{cases} 0 & \text{if user interested} \\ 1 & \text{otw.} \end{cases}$$

Sps we play $x_t \sim p_t$ (as in Hedge)

Define
$$\tilde{\ell}_t(x) := \begin{cases} \frac{1}{p_t(x_t)} \ell_t(x_t) & \text{if } x = x_t \\ 0 & \text{otw.} \end{cases}$$

Then $\forall x$:
$$\mathbb{E}_{x_t \sim p_t}\left[\hat{\ell}_t(x)\right] = \sum_{x'} p_t(x') \cdot \tilde{\ell}_t(x') = p_t(x_t) \cdot \tilde{\ell}(x_t) + 0$$
$$= \cancel{p_t(x_t)} \cdot \frac{\ell_t(x_t)}{\cancel{p_t(x_t)}} = \ell_t(x_t)$$

$$\forall x: \mathbb{E}_{x_t \sim p_t}\left[\tilde{\ell}_t^2(x)\right] = p_t(x_t) \cdot \tilde{\ell}_t^2(x_t) = p_t(x_t) \frac{\ell_t^2(x_t)}{p_t^2(x_t)} = \frac{\ell_t^2(x_t)}{p_t(x_t)}$$

Algorithm: EXP3 : play Hedge on $\tilde{\ell}_t$

Theorem: For $\ell_t \in [0,1]$ $\forall t, x$: Sps we run EXP3 w. $\varepsilon = \sqrt{\frac{\log k}{Tk}}$

Then it holds: $\mathbb{E}[R_T] \leq 2\sqrt{T k \log k}$ ← due to partial inf.

Proof:

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(x^*)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} p_t^T \ell_t - \sum_{t=1}^{T} \ell_t(x^*)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} p_t^T \tilde{\ell}_t - \sum_{t=1}^{T} \tilde{\ell}_t(x^*)\right]$$

lin. of expect. x^* independent of $\tilde{\ell}_{1:T}$; $\mathbb{E}[\tilde{\ell}_t] = \ell_t$

$$\overset{\text{Hedge on } \tilde{\ell}_{1:T}}{\leq} \mathbb{E}\left[\varepsilon \cdot \sum_{t=1}^{T} p_t^T \tilde{\ell}_t^2 + \frac{\log k}{\varepsilon}\right]$$

Note: $\mathbb{E}[p_t^T \tilde{\ell}_t^2] = \sum_{x=1}^{k} p_t(x) \cdot \mathbb{E}_{x_t \sim p_t}[\tilde{\ell}_t^2(x)]$

$$= \sum_{x=1}^{k} \cancel{p_t(x)} \frac{\ell_t^2(x)}{\cancel{p_t(x)}} = \sum_{x=1}^{k} \ell_t^2(x) \leq k$$

$$\Rightarrow \mathbb{E}[R_T] \leq \varepsilon \cdot k \cdot T + \frac{\log k}{\varepsilon}$$

$$\overset{\text{Def. of } \varepsilon}{\leq} 2\sqrt{T \cdot k \cdot \log k}$$

LEARNING IN REPEATED GAMES

For $t = 1:T$

We pick $x_t \in X$

Opp. picks $y_t \in Y$ fixed

We obtain $\ell_t^X(x_t) = f(x_t, y_t)$

Opp. obtains $\ell_t^Y(y_t) = 1 - f(x_t, y_t)$

Can directly apply Hedge (or EXP3) with sublinear regret $O(\sqrt{T \log k})$ (or $O(\sqrt{T k \log k})$)

<u>Fact</u> if both players play no-regret, the "average actions" $P_T^X = \text{Unif}\{x_1 \ldots x_T\}$

$$P_T^Y = \text{Unif}\{y_1 \ldots y_T\}$$

Converge to a Nash equilibrium in game f

OUTLOOK

Exponential gap (in k) between full info and bandit setting.

In hindsight, often can also observe y_t (other player's action)

Can show:

$$\mathbb{E}[R_T] = O\left(\sqrt{T \log k} + \gamma_T \sqrt{T}\right)$$

regret under full info

depends on structure of game

E.g. $f(x,y) = w^T \phi(x,y)$, for $w \in \mathbb{R}^d$

$\Rightarrow \gamma_T = d \cdot \log T$