Online Learning Applications

Part 3: Adversarial expert problems and MABs

Adversarial expert problems

Adversarial expert problem

At each time $t = 1, \ldots, T$:

1. The environment chooses a loss function $\ell_t : A \to [0, 1]$
2. The learner chooses an arm $a_t \in A$
3. The learner receives a loss $\ell_t(a_t)$
4. The learner observes the loss $\ell_t(a)$ of **all** arms $a \in A$

Goal

Design an algorithm that achieves sublinear regret ($\lim_{T \to \infty} \frac{R_T}{T} = 0$).

What can we hope to achieve?

Theorem

In the adversarial expert problem, any algorithm suffers regret at least

$$\mathcal{R}_T \geq \Omega\left(\sqrt{\log(K)T}\right).$$

What can we hope to achieve?

Theorem

In the adversarial expert problem, any algorithm suffers regret at least

$$\mathcal{R}_T \geq \Omega\left(\sqrt{\log(K)T}\right).$$

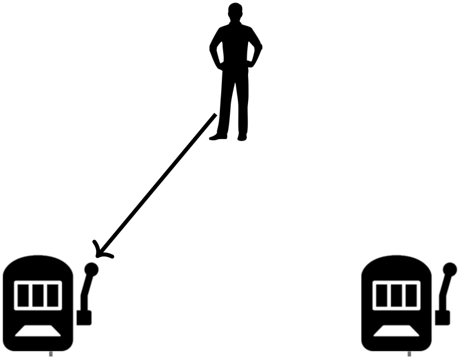
⚠ It doesn't make much sense to talk about instance-dependent bounds (e.g., dependence on Δ_a) since we are not making any assumption about the losses.

Deterministic algorithms fail

If the algorithm is deterministic, the environment "knows" the arm that the learner will choose.

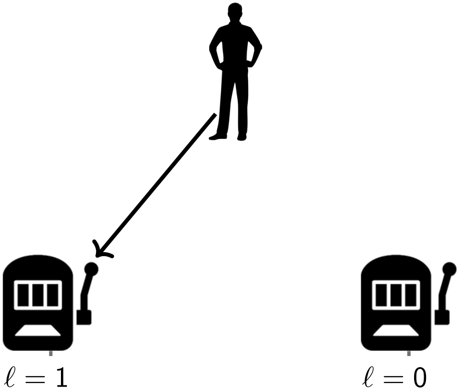
Deterministic algorithms fail

If the algorithm is deterministic, the environment "knows" the arm that the learner will choose.



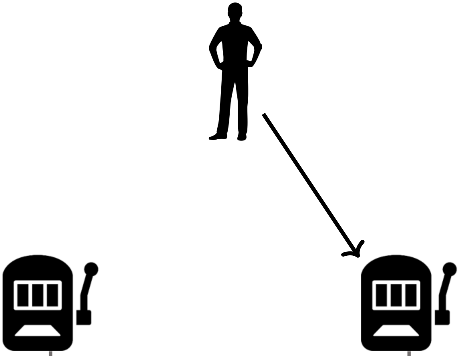
Deterministic algorithms fail

If the algorithm is deterministic, the environment "knows" the arm that the learner will choose.



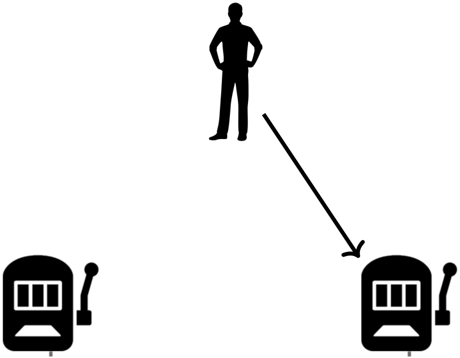
Deterministic algorithms fail

If the algorithm is deterministic, the environment "knows" the arm that the learner will choose.



Deterministic algorithms fail

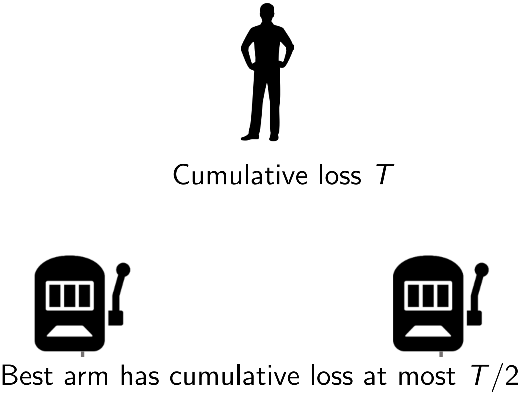
If the algorithm is deterministic, the environment "knows" the arm that the learner will choose.



$\ell = 0$ $\ell = 1$

Deterministic algorithms fail

If the algorithm is deterministic, the environment "knows" the arm that the learner will choose.



Deterministic algorithms fail

If the algorithm is deterministic, the environment "knows" the arm that the learner will choose.

Deterministic Algorithm Regret

Any deterministic algorithm suffers regret $\Omega(T)$.

Randomization

A randomized algorithm chooses a **distribution** over arms.

Idea:

- Assign a weight $w(a)$ to each arm
- Play each arm with probability proportional to the weight

- We should assign a large weight to "good" arms
- We observe only the past losses
- We decrease the weight exponentially in the loss: essential to obtain optimal bounds

Hedge

Algorithm: Hedge

Set of arm A, number of rounds T, learning rate η;
Initialization: $w_1 \leftarrow (1, 1, \ldots, 1)$;
for $t = 1, \ldots, T$ **do**
  $x_t(a) \leftarrow \frac{w_t(a)}{\sum_{a' \in A} w_t(a')}$ for every $a \in A$;
  sample arm $a_t \sim x_t \in \Delta_A$;
  play a_t;
  observe loss vector $\ell_t \in [0, 1]^K$;
  suffer **expected** loss $\langle \ell_t, x_t \rangle$;
  update weights: $w_{t+1}(a) \leftarrow w_t(a) e^{-\eta \ell_t(a)}$ for every $a \in A$;

Hedge

Theorem

Hedge with learning rate $\eta = \sqrt{\frac{\log K}{T}}$ *achieves regret:*

$$R_T \leq O\left(\sqrt{T \log K}\right).$$

Proof.

- Let $\ell_t^2 \in \mathbb{R}^K$ be the vector of squared losses
- Let $z_t = \sum_{a \in A} w_t(a)$ be the sum of the weights at round t

For each $t \in \{1, \dots, T\}$:

$$
\begin{aligned}
z_{t+1} = \sum_{a \in A} w_{t+1}(a) = \sum_{a \in A} w_t(a) e^{-\eta \ell_t(a)} &= z_t \sum_{a \in A} x_t(a) e^{-\eta \ell_t(a)} \\
&\leq z_t \sum_{a \in A} x_t(a) \left(1 - \eta \ell_t(a) + \eta^2 \ell_t(a)^2\right) = z_t \left(1 - \eta \langle \ell_t, x_t \rangle + \eta^2 \langle \ell_t^2, x_t \rangle\right) \\
&\leq z_t e^{-\eta \langle \ell_t, x_t \rangle + \eta^2 \langle \ell_t^2, x_t \rangle},
\end{aligned}
$$

where

- The first inequality comes from $e^{-x} \leq 1 - x + x^2$ for each $x \geq 0$
- The second inequality comes from $1 + x \leq e^x$ for each $x \in \mathbb{R}$

Proof.

By induction, we get

$$z_{T+1} \leq z_1 \prod_{t=1}^{T} e^{-\eta\langle \ell_t, x_t\rangle + \eta^2 \langle \ell_t^2, x_t\rangle} = K e^{-\eta \sum_{t=1}^{T}\langle \ell_t, x_t\rangle + \eta^2 \sum_{t=1}^{T}\langle \ell_t^2, x_t\rangle}.$$

Moreover, by induction we get

$$w_{T+1}(a) = \prod_{t=1}^{T} e^{-\eta \ell_t(a)} = e^{-\eta \sum_{t=1}^{T} \ell_t(a)}.$$

Let $a^* \in A$ be the best arm in hindsight, that is, $a^* \in \arg\min_{a \in A} \sum_{t=1}^{T} \ell_t(a)$. Then,

$$e^{-\eta \sum_{t=1}^{T} \ell_t(a^*)} = w_{T+1}(a^*) \leq z_{T+1} \leq K e^{-\eta \sum_{t=1}^{T}\langle \ell_t, x_t\rangle + \eta^2 \sum_{t=1}^{T}\langle \ell_t^2, x_t\rangle}.$$

Proof.

Taking logs on both sides, we get

$$-\eta \sum_{t=1}^{T} \ell_t(a^*) \leq \log K - \eta \sum_{t=1}^{T} \langle \ell_t, x_t \rangle + \eta^2 \sum_{t=1}^{T} \langle \ell_t^2, x_t \rangle.$$

Hence, we have

$$R_T = \sum_{t=1}^{T} \langle \ell_t, x_t \rangle - \sum_{t=1}^{T} \ell_t(a^*) \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \langle \ell_t^2, x_t \rangle \leq \frac{\log K}{\eta} + \eta T \tag{1}$$

$$= 2\sqrt{T \log K}, \tag{2}$$

where the last inequality comes from $\ell_t^2 \leq \mathbf{1}$ and the last equality from the definition of η. □

Adversarial MABs

Adversarial MABs

At each time $t = 1, \dots, T$:

1. The environment chooses a loss function $\ell_t : A \to [0, 1]$
2. The learner chooses an arm a_t
3. The learner receives a loss $\ell_t(a_t)$
4. ⚠ The learner observes **only** the loss $\ell_t(a_t)$ of arm a_t

Goal

Design an algorithm that achieves sublinear regret ($\lim_{T\to\infty} \frac{\mathcal{R}_T}{T} = 0$).

What can we hope to achieve?

The lower bound is slightly worse that under full information. The dependence on the number of arms is $\sqrt{K}$ instead of $\sqrt{\log(K)}$.

Theorem

In the adversarial MABs, any algorithm suffers regret at least

$$\mathcal{R}_T \geq \Omega\left(\sqrt{KT}\right).$$

What can we hope to achieve?

The lower bound is slightly worse that under full information. The dependence on the number of arms is $\sqrt{K}$ instead of $\sqrt{\log(K)}$.

Theorem

In the adversarial MABs, any algorithm suffers regret at least

$$\mathcal{R}_T \geq \Omega\left(\sqrt{KT}\right).$$

⚠ It doesn't make much sense to talk about instance-dependent bounds (e.g., dependence on Δ_a) since we are not making any assumption about the losses.

Working with limited information

- We cannot directly apply Hedge since we have access only to the loss of the arm we played
- The chosen arm a_t has consequences on the received loss but also on the **information gathered**

How can we adapt Hedge to this feedback model?

Working with limited information

- We cannot directly apply Hedge since we have access only to the loss of the arm we played
- The chosen arm a_t has consequences on the received loss but also on the **information gathered**

How can we adapt Hedge to this feedback model?

Idea:

- Update the weight only for the played arm a_t
- Modify the loss $\ell_t(a_t)$ to $\tilde{\ell}_t(a_t) := \ell_t(a_t)/x_t(a_t)$
- Set all the other $\tilde{\ell}_t(a) = 0$ (i.e., the weights are not updated)
- This is an **unbiased estimator** of the loss ($\mathbb{E}[\tilde{\ell}_t(a)] = \ell_t(a)$)

EXP3

Algorithm: EXP3

set of arms A, number of rounds T, learning rate η;
$w_1 \leftarrow (1, 1, \ldots, 1)$;
for $t = 1, \ldots, T$ **do**
    $x_t(a) \leftarrow \frac{w_t(a)}{\sum_{a' \in A} w_t(a')}$ for every $a \in A$;
    sample arm $a_t \sim x_t \in \Delta_A$;
    play a_t;
    observe loss $\ell_t(a_t)$;
    suffer **expected** loss $\langle \ell_t, x_t \rangle$;
    compute $\tilde{\ell}_t(a) \leftarrow \frac{\ell_t(a_t)}{x_t(a_t)} \mathbb{I}[a_t = a]$ for every $a \in A$;
    update weights: $w_{t+1,i} \leftarrow w_{t,i} e^{-\eta \tilde{\ell}_t(a)}$ for every $a \in A$;

EXP3

Theorem

EXP3 with learning rate $\eta = \sqrt{\frac{\log K}{KT}}$ *achieves regret:*

$$R_T \leq O\left(\sqrt{K \log(K) T}\right).$$

- Regret proportional to $\sqrt{T}$ as in the full-info setting with Hedge
- Worse dependence from the number of arms w.r.t Hedge

EXP3

Theorem

EXP3 with learning rate $\eta = \sqrt{\frac{\log K}{KT}}$ *achieves regret:*

$$R_T \leq O\left(\sqrt{K \log(K) T}\right).$$

- Regret proportional to $\sqrt{T}$ as in the full-info setting with Hedge
- Worse dependence from the number of arms w.r.t Hedge

Relation to Stochastic bandit:

- Same worst-case regret bound under weaker assumptions!
- No instance-dependent regret bounds
- Worse performance in "easy" stochastic instances (no instance-dependent bounds)

Proof.

EXP3 is equivalent to Hedge with losses $\tilde{\ell}_t$. Two challenges to extend the analysis to EXP3:

- $\tilde{\ell}_t \neq \ell_t$
- $\tilde{\ell}_t$ is no more bounded in $[0,1]^K$

We can follow the proof of Hedge to show:

$$\sum_{t=1}^{T}\langle\tilde{\ell}_t, x_t\rangle - \sum_{t=1}^{T}\tilde{\ell}_t(a^*) \leq \frac{\log K}{\eta} + \eta\sum_{t=1}^{T}\langle\tilde{\ell}_t^2, x_t\rangle.$$

Proof.

Taking expectation on both sides:

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle\tilde{\ell}_t, x_t\rangle - \sum_{t=1}^{T}\tilde{\ell}_t(a^*)\right] \leq \frac{\log K}{\eta} + \eta\sum_{t=1}^{T}\sum_{a\in A} x_t(a)\frac{\ell_t(a)^2}{x_t(a)}$$
$$\leq \frac{\log K}{\eta} + \eta\sum_{t=1}^{T}\sum_{a\in A}\ell_t(a)^2 \leq \frac{\log K}{\eta} + \eta TK.$$

Since $\tilde{\ell}_t(a)$ is an unbiased estimator of $\ell_t(a)$:

$$\sum_{t=1}^{T}\langle\ell_t, x_t\rangle - \sum_{t=1}^{T}\ell_t(a^*) = \mathbb{E}\left[\sum_{t=1}^{T}\langle\tilde{\ell}_t, x_t\rangle - \sum_{t=1}^{T}\tilde{\ell}_t(a^*)\right] \leq \frac{\log K}{\eta} + \eta TK = 2\sqrt{K\log(K)T},$$

where the last equality follows from the definition of η.

□

References

Nick Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2): 212–261, 1994.

Yoav Freund and Robert Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44:427–485, 1997.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32:48–77, 2002.