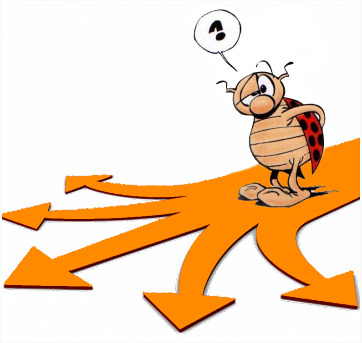
Online Learning with Feedback Graphs

Frédéric Koriche
CRIL - CNRS UMR 8188, Univ. Artois
koriche@cril.fr

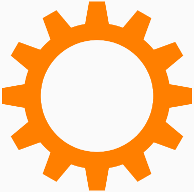
MI'19 - Orléans, France

Outline



Online Learning is the mainstream theoretical framework for making sequential decisions in face of uncertainty.

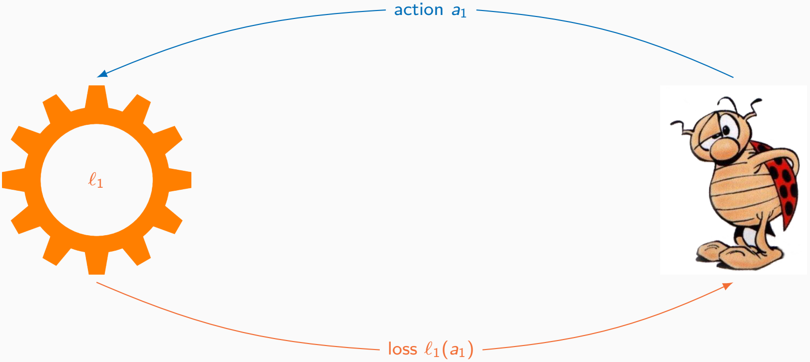
- How should you filter incoming emails?
- Which daily items should you recommender to your customers?
- What move should you consider next when playing Go?



Online learning is a (zero-sum) repeated game between the forecaster and its environment.

For each round $t = 1$ to T

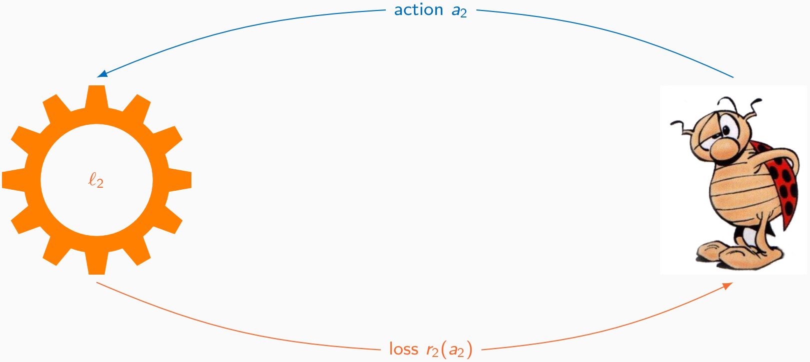
- The forecaster chooses an action $a_t \in \mathcal{A}$ (possibly at random)
- The environment simultaneously chooses a loss function $\ell_t : \mathcal{A} \to \mathbb{R}$
- The forecaster incurs the loss $\ell_t(a_t)$



Online learning is a (zero-sum) repeated game between the forecaster and its environment.

For each round $t = 1$ to T

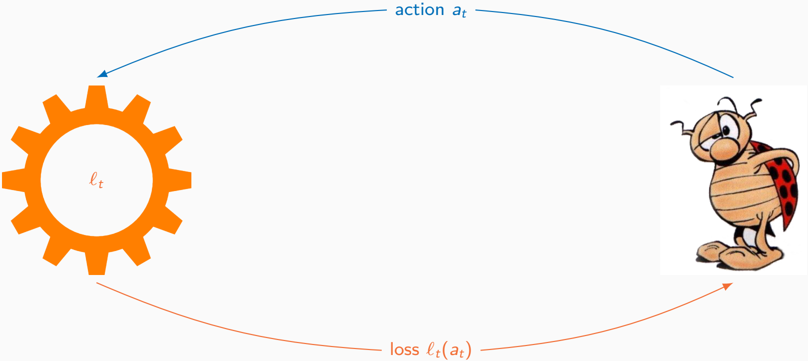
- The forecaster chooses an action $a_t \in \mathcal{A}$ (possibly at random)
- The environment simultaneously chooses a loss function $\ell_t : \mathcal{A} \to \mathbb{R}$
- The forecaster incurs the loss $\ell_t(a_t)$



Online learning is a (zero-sum) repeated game between the forecaster and its environment.

For each round $t = 1$ to T

- The forecaster chooses an action $a_t \in \mathcal{A}$ (possibly at random)
- The environment simultaneously chooses a loss function $\ell_t : \mathcal{A} \to \mathbb{R}$
- The forecaster incurs the loss $\ell_t(a_t)$



Online learning is a (zero-sum) repeated game between the forecaster and its environment.

For each round $t = 1$ to T

- The forecaster chooses an action $a_t \in \mathcal{A}$ (possibly at random)
- The environment simultaneously chooses a loss function $\ell_t : \mathcal{A} \to \mathbb{R}$
- The forecaster incurs the loss $\ell_t(a_t)$

Regret

The goal of the player is to minimize its regret, which is a measure of relative performance the actions taken by the player and the *best* possible action (with benefit of hindsight).

Regret

Formally, the expected regret is defined as

$$\mathrm{Regret}_T = \sum_{t=1}^{T} \mathbb{E}[\ell_t(a_t)] - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell_t(a)$$

Regret

Formally, the expected regret is defined as

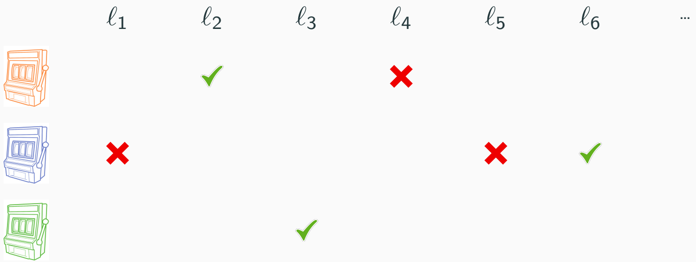
$$\mathrm{Regret}_T = \sum_{t=1}^{T} \mathbb{E}[\ell_t(a_t)] - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell_t(a)$$

Hannan Consistency

A player is Hannan consistent if for any sequence of losses chosen by the environment, the player's regret is always sublinear in T

$$\mathrm{Regret}_T = o(T) \quad \text{i.e.} \quad \lim_{T \to \infty} \frac{\mathrm{Regret}_T}{T} = 0$$

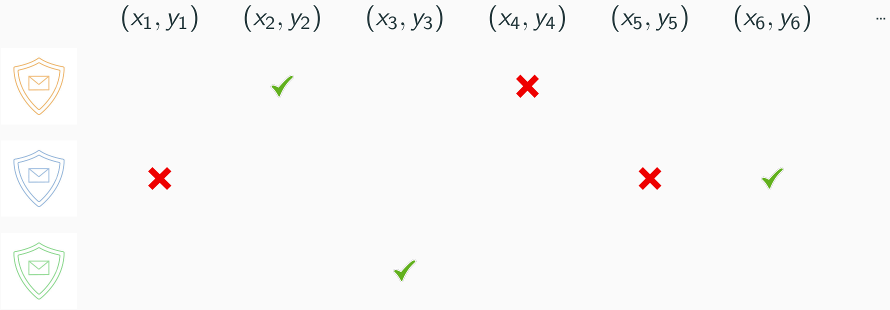
Outline



Multi-Armed Bandits

The forecaster has access to a set $\mathcal{A}$ of *slot machines*. On each round t:

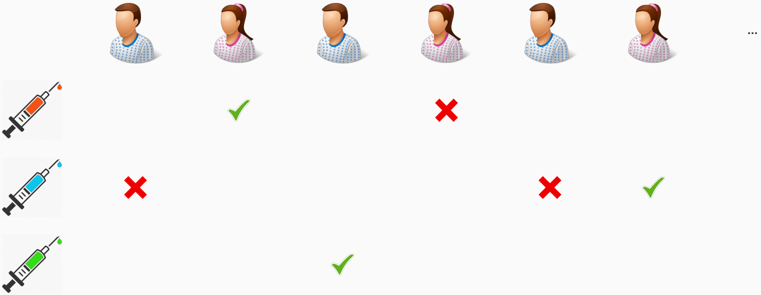
- The forecaster chooses a machine $a_t \in \mathcal{A}$
- Simultaneously, the environment selects a loss function $\ell_t : \mathcal{A} \to \{-1, +1\}$
- The forecaster only observes $\ell_t(a_t)$ and incurs this loss



Antispam Filtering

The forecaster has access to a set $\mathcal{A}$ of *binary classifers*. On each round t:

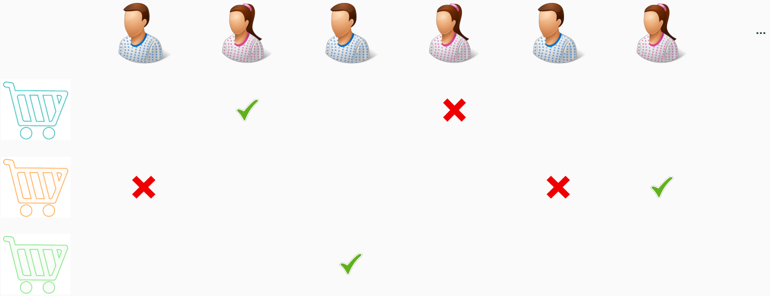
- The forecaster chooses a classifier $a_t \in \mathcal{A}$
- Simultaneously, the environment selects an email (with its label “spam” or “ham”) (x_t, y_t)
- The forecaster observes (x_t, y_t) and incurs the zero-one loss $\mathbb{I}[a_t(x_t) \neq y_t]$



Sequential Treatment Allocation

The forecaster has access to a set $\mathcal{A}$ of *medical treatments*. On each round t:

- A patient arrives with her symptoms x_t
- The forecaster chooses a treatment $a_t \in \mathcal{A}$ according to x_t
- Simultaneously, the environment selects a hidden loss function $\ell_t : \mathcal{A} \to [0, 1]$
- The forecaster observes $\ell_t(a_t)$ and incurs this loss. But can she infer the loss of some other treatments?



Online Advertising

The forecaster has access to a set $\mathcal{A}$ of ads. On each round t:

- A customer arrives with her profile x_t
- The forecaster chooses an ad $a_t \in \mathcal{A}$ according to x_t
- Simultaneously, the environment selects a hidden loss function $\ell_t : \mathcal{A} \to \{0, 1\}$
- The forecaster only observes $\ell_t(a_t)$ and incurs this loss. But again, can she infer the loss of some other ads?

Outline

Prediction Games with Full Information Feedback

The forecaster has access to a set $\mathcal{A} = \{1, \cdots, K\}$ of actions.

The environment has access to a set $\mathcal{L} \subseteq [0,1]^K$ of loss functions.

During each round t

- The forecaster chooses an action $a_t \in \mathcal{A}$
- The environment chooses a loss function $\ell_t \in \mathcal{L}$
- The forecaster observes ℓ_t, and incurs the loss $\ell_t(a_t)$

Prediction Games with Full Information Feedback

The forecaster has access to a set $\mathcal{A} = \{1, \cdots, K\}$ of actions.

The environment has access to a set $\mathcal{L} \subseteq [0, 1]^K$ of loss functions.

During each round t

- The forecaster chooses an action $a_t \in \mathcal{A}$
- The environment chooses a loss function $\ell_t \in \mathcal{L}$
- The forecaster observes ℓ_t, and incurs the loss $\ell_t(a_t)$

This is a basic supervised learning model.

Prediction Games with Bandit Feedback

The forecaster has access to a set $\mathcal{A} = \{1, \cdots, K\}$ of actions.

The environment has access to a set $\mathcal{L} \subseteq [0, 1]^K$ of loss functions.

During each round t

- The forecaster chooses an action $a_t \in \mathcal{A}$
- The environment chooses a loss function $\ell_t \in \mathcal{L}$
- The forecaster only observes $\ell_t(a_t)$, and incurs this loss

Prediction Games with Bandit Feedback

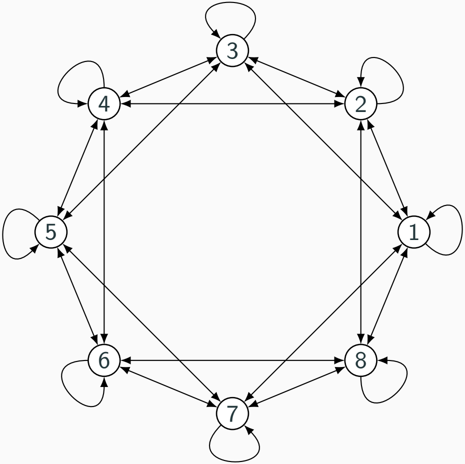
The forecaster has access to a set $\mathcal{A} = \{1, \cdots, K\}$ of actions.

The environment has access to a set $\mathcal{L} \subseteq [0, 1]^K$ of loss functions.

During each round t

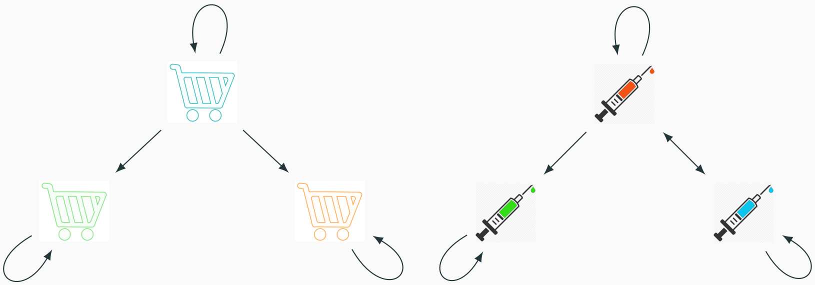
- The forecaster chooses an action $a_t \in \mathcal{A}$
- The environment chooses a loss function $\ell_t \in \mathcal{L}$
- The forecaster only observes $\ell_t(a_t)$, and incurs this loss

This is a basic reinforcement learning model.



Feedback Graphs

For a set $\mathcal{A} = \{1, \cdots, K\}$ of actions, a feedback graph is a digraph $G = (\mathcal{A}, E)$, where each arc $(i, j) \in E$ indicates that if we play action i then we observe the loss on action j.



The interest of feedback graphs stems from the fact that for many applications actions are interdependent. If we play some action a_t at trial t, then we may infer the loss of similar actions.

Prediction Games with Feedback Graphs

The forecaster has access to a set $A = \{1, \cdots, K\}$ of actions.

The environment has access to a set $\mathcal{L} \subseteq [0,1]^K$ of loss functions.

The environment has also access to a class $\mathcal{G} \subseteq \{0,1\}^{K \times K}$ of feedback graphs.

During each round t

- The forecaster chooses an action $a_t \in \mathcal{A}$
- The environment chooses a loss function $\ell_t \in \mathcal{L}$
- The environment chooses a feedback graph $G_t \in \mathcal{G}$
- The forecaster observes G_t and incurs the loss $\ell_t(a_t)$

Prediction Games with Feedback Graphs

The forecaster has access to a set $A = \{1, \cdots, K\}$ of actions.

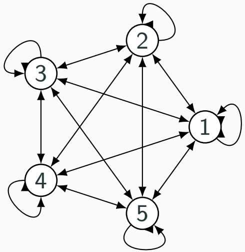
The environment has access to a set $\mathcal{L} \subseteq [0,1]^K$ of loss functions.

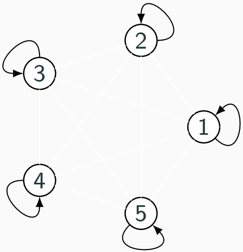
The environment has also access to a class $\mathcal{G} \subseteq \{0,1\}^{K \times K}$ of feedback graphs.

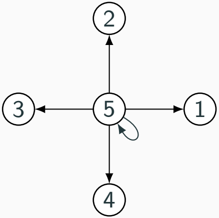
During each round t

- The forecaster chooses an action $a_t \in \mathcal{A}$
- The environment chooses a loss function $\ell_t \in \mathcal{L}$
- The environment chooses a feedback graph $G_t \in \mathcal{G}$
- The forecaster observes G_t and incurs the loss $\ell_t(a_t)$

This covers a wide spectrum of online learning models!



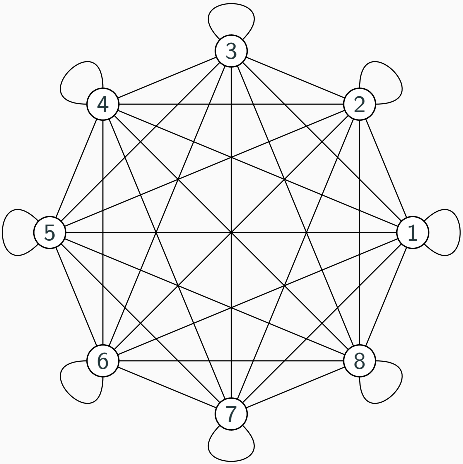




Prediction games with feedback graphs include, among others:

- Full-information games
- Bandit games
- Revealing action games

Outline



Online Learning with Full Information Feedback

- $\mathcal{A}$ is a simple collection $\{1, \cdots, K\}$ of objects.
- $\mathcal{L}$ is a set of (bounded) mappings $\mathcal{A} \to [0, 1]$.
- Loss functions are chosen in an adverarial way.
- The feedback graph is the complete digraph over $\mathcal{A}$.

Hedge (Freund & Schapire, 1997)

Parameter: stepsize η

Initialization: set p_t to the uniform distribution on $\mathcal{A}$

Trials: for $t = 1$ to T

play $A_t \sim p_t$

receive ℓ_t

update $p_{t+1}(i) = \frac{p_t(i)\exp(-\eta\ell_t(i))}{\sum_{j\in\mathcal{A}} p_t(j)\exp(-\eta\ell_t(j))}$

full information

exponential weights

Hedge (Freund & Schapire, 1997)

Parameter: stepsize η

Initialization: set p_t to the uniform distribution on $\mathcal{A}$

Trials: for $t = 1$ to T

play $A_t \sim p_t$

receive ℓ_t

update $p_{t+1}(i) = \frac{p_t(i)\exp(-\eta\ell_t(i))}{\sum_{j\in\mathcal{A}} p_t(j)\exp(-\eta\ell_t(j))}$

full information

exponential weights

Theorem 1

Hedge achieves an optimal regret bound of $O(\sqrt{T \ln K})$

Second-Order Regret Bound (Cesa-Bianchi et al., 2007)

Let i^* be any action in $\mathcal{A}$. Then,

$$\sum_{t=1}^{T} \mathbb{E}[\ell_t(i)] - \sum_{t=1}^{T} \ell_t(i^*) \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \mathbb{E}\left[\ell_t(i)^2\right]$$

Proof

Let $W_t = \sum_{i \in \mathcal{A}} w_t(i)$, where $w_t(i) = \exp\left(-\eta \sum_{s=1}^{t-1} \ell_t(i)\right)$.

Then using $e^x \leq 1 + x + x^2$ for $x \leq 1$,

$$\begin{aligned} \frac{W_{t+1}}{W_t} &= \mathbb{E}[\exp(-\eta \ell_t(i))] \\ &\leq \mathbb{E}\left[1 - \eta \ell_t(i) + \eta^2 \ell_t(i)^2\right] \\ &= 1 - \eta \mathbb{E}[\ell_t(i)] + \eta^2 \mathbb{E}[\ell_t(i)^2] \end{aligned}$$

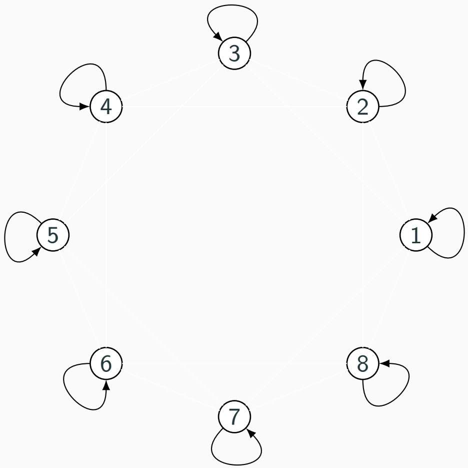
Now, using $\ln(1 - x) \leq -x$ for $x \geq 0$, and summing over T,

$$\ln \frac{W_{T+1}}{W_1} \leq -\eta \sum_t \mathbb{E}[\ell_t(i)] + \eta^2 \sum_t \mathbb{E}[\ell_t(i)^2]$$

Finally, for any fixed action i^*, we also have

$$\ln \frac{W_{T+1}}{W_1} \geq \ln \frac{w_{T+1}(i^*)}{W_1} = -\eta \sum_{t=1}^{T} \ell_t(i^*) - \ln K$$

Combining both inequalities and rearranging gives the result.



Online Learning with Bandit Feedback

- $\mathcal{A}$ is a simple collection $\{1, \cdots, K\}$ of objects.
- $\mathcal{L}$ is a set of (bounded) mappings $\mathcal{A} \to [0, 1]$.
- Loss functions are chosen in an adverarial way.
- The feedback graph is fixed and contains only self-loops.

EXP3 (Auer et al., 2003)

Parameters: stepsize η, exploration γ

Initialization:

let u to the uniform distribution over $\mathcal{A}$

set $q_t = u$

Trials: for $t = 1$ to T

set $p_t = (1-\gamma)q_t + \gamma u$ } exploration-exploitation

play $A_t \sim p_t$

receive $\ell_t(A_t)$ } bandit

estimate $\hat{\ell}_t(i) = \frac{\ell_t(i)}{p_t(i)}\mathbb{I}\{i = A_t\}$ } unbiased estimator

update $q_{t+1}(i) = \frac{q_t(i)\exp(-\eta\hat{\ell}_t(i))}{\sum_{j\in\mathcal{A}} q_t(j)\exp(-\eta\hat{\ell}_t(j))}$ } exponential weights

EXP3 (Auer et al., 2003)

Parameters: stepsize η, exploration γ

Initialization:

let u to the uniform distribution over $\mathcal{A}$

set $q_t = u$

Trials: for $t = 1$ to T

set $p_t = (1-\gamma)q_t + \gamma u$ } exploration-exploitation

play $A_t \sim p_t$

receive $\ell_t(A_t)$ } bandit

estimate $\hat{\ell}_t(i) = \frac{\ell_t(i)}{p_t(i)} \mathbb{I}\{i = A_t\}$ } unbiased estimator

update $q_{t+1}(i) = \frac{q_t(i) \exp(-\eta \hat{\ell}_t(i))}{\sum_{j \in \mathcal{A}} q_t(j) \exp(-\eta \hat{\ell}_t(j))}$ } exponential weights

Theorem 2

EXP3 achieves an optimal regret bound of $\tilde{O}(\sqrt{KT})$

Proof (Sketch)

Decompose the expected regret as follows:

$$\mathbb{E}\left[\sum_t \ell_t(i)\right] - \sum_t \ell_t(i^*) = \mathbb{E}\left[\sum_t \ell_t(i)\right] - \sum_t \sum_i q_t(i)\ell_t(i) \tag{1}$$

$$+ \sum_t \sum_i q_t(i)\ell_t(i) - \sum_t \ell_t(i^*) \tag{2}$$

The term (1) is bounded by γT since

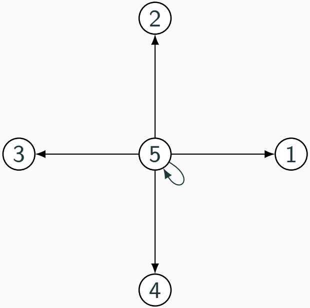
$$\sum_i p_t(i)\ell_t(i) \leq \sum_i q_t(i) + \gamma$$

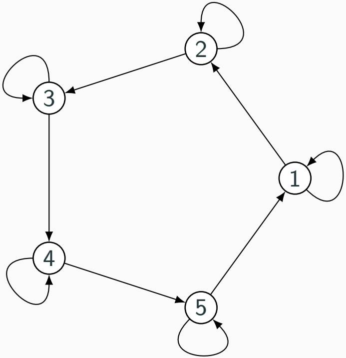
For the term (2), we know that $\hat{\ell}_t(i)$ is an unbiased estimator of $\ell_t(i)$. Therefore,

$$\sum_t \sum_i q_t(i)\ell_t(i) - \sum_t \ell_t(i^*) = \sum_t \sum_i q_t(i)\mathbb{E}[\hat{\ell}_t(i)] - \sum_t \mathbb{E}[\ell_t(i^*)]$$

Applying the second-order regret bound for Hedge, and using $p_t(i) \geq \gamma/K$, we get that

$$\begin{aligned}\sum_t \sum_i q_t(i)\mathbb{E}[\hat{\ell}_t(i)] - \sum_t \mathbb{E}[\ell_t(i^*)] &\leq \frac{\ln K}{\eta} + \eta \sum_t \sum_i \frac{q_t(i)}{p_t(i)} \\ &\leq \frac{\ln K}{\eta} + \frac{\eta K T}{\gamma}\end{aligned}$$



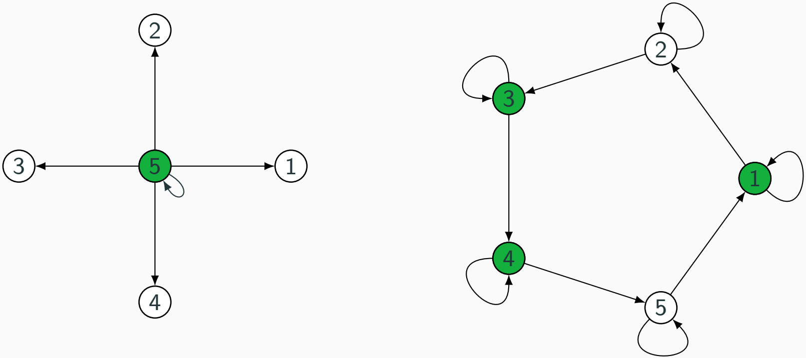


Some Classes of Feedback Graphs

Let G be a directed graph on the action set $\mathcal{A}$.

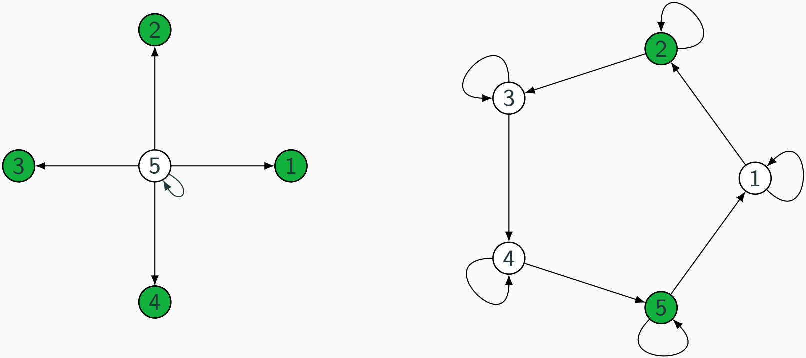
Let $G_{\mathrm{out}}(i)$ be the out-neighborhood of i in G, and $G_{\mathrm{in}}(i)$ be the in-neighborhood of i in G.

- G is weakly observable if $G_{\mathrm{in}}(i) \neq \varnothing$ for each action i,
- G is strongly observable if $i \in G_{\mathrm{in}}(i)$.



Weak Domination Number

A weakly dominating set is a set $D \subseteq \mathcal{A}$ that dominates $\mathcal{A}$, i.e. for every $i \in \mathcal{A}$, there exists $j \in D$ such that $i \in G_{\mathrm{out}}(j)$. The weak domination number δ is the size of any smallest weakly dominating set.



Independence Number

An *independent set* is a set $I \subseteq \mathcal{A}$ of actions that are not connected by any edge. The independence number α is the size of any largest independent set.

EXP3.G (Alon et. al., 2015)

Parameters: feedback graph G, stepsize η, exploration $\gamma \leq 1/2$

Initialization:

let u to the uniform distribution over $\mathcal{A}$

set $q_t = u$

Trials: for $t = 1$ to T

set $p_t = (1-\gamma)q_t + \gamma u$

play $A_t \sim p_t$

receive $\{\ell_t(i) : i \in G_{\mathrm{out}}(A_t)\}$

estimate $\hat{\ell}_t(i) = \frac{\ell_t(i)}{P_t(i)} \mathbb{I}\{i \in G_{\mathrm{out}}(A_t)\}$ $\Big\}\, P_t(i) = \sum_{j \in G_{\mathrm{out}}(a_t)} p_t(j)$

update $q_{t+1}(i) = \frac{q_t(i) \exp(-\eta \hat{\ell}_t(i))}{\sum_{j \in \mathcal{A}} q_t(j) \exp(-\eta \hat{\ell}_t(j))}$

EXP3.G (Alon et. al., 2015)

Parameters: feedback graph G, stepsize η, exploration $\gamma \leq 1/2$

Initialization:

let u to the uniform distribution over $\mathcal{A}$

set $q_t = u$

Trials: for $t = 1$ to T

set $p_t = (1-\gamma)q_t + \gamma u$

play $A_t \sim p_t$

receive $\{\ell_t(i) : i \in G_{\mathrm{out}}(A_t)\}$

estimate $\hat{\ell}_t(i) = \frac{\ell_t(i)}{P_t(i)} \mathbb{I}\{i \in G_{\mathrm{out}}(A_t)\}$ $\Big\}\ P_t(i) = \sum_{j \in G_{\mathrm{out}}(a_t)} p_t(j)$

update $q_{t+1}(i) = \frac{q_t(i)\exp(-\eta\hat{\ell}_t(i))}{\sum_{j\in\mathcal{A}} q_t(j)\exp(-\eta\hat{\ell}_t(j))}$

Theorem 3

EXP3.G achieves an expect regret of

- $\mathcal{O}(\sqrt{\alpha T}\ln(KT))$ for strongly observable feedback graphs
- $\mathcal{O}(\sqrt[3]{\delta \ln KT^2})$ for weakly observable feedback graphs

Main Lemma for Independence Numbers

Let G be a digraph over $\mathcal{A}$ in which each action i is assigned a weight w_i. Assume that $w_i \geq \epsilon$ for $\epsilon \in (0, 1/2)$, and $\sum_i w_i \leq 1$. Then,

$$\sum_i \frac{w_i}{w_i + \sum_{j \in G_{\mathrm{in}}(i)} w_j} \leq 4\alpha \ln \frac{4K}{\alpha\epsilon}$$

Proof (Sketch for strongly observable feedback graphs)

Based on the proof for EXP3, use the above lemma for refining the second-order term.

Using the fact $\hat{\ell}_t(i)$ is again an unbiased estimate of $\ell_t(i)$, together with the fact that $p_t(i) \geq (1-\gamma)q_t(i) \geq 1/2 q_t(i)$, we have

$$\begin{aligned}\sum_t \sum_i q_t(i)\mathbb{E}[\hat{\ell}_t(i)] &\leq \frac{\ln K}{\eta} + \eta \sum_t \sum_i \frac{q_t(i)}{P_t(i)} \\ &\leq \frac{\ln K}{\eta} + 2\eta \sum_t \sum_i \frac{p_t(i)}{P_t(i)}\end{aligned}$$

Since $p_t(i) \geq \gamma/K$, we can use $\epsilon = \gamma/K$, which yields:

$$\sum_t \sum_i q_t(i)\mathbb{E}[\hat{\ell}_t(i)] \leq \frac{\ln K}{\eta} + \eta T \left(8\alpha \ln \frac{4K^2}{\alpha\gamma}\right)$$

Outline

There are many questions related to online learning with feedback graphs.

There are many questions related to online learning with feedback graphs.

About actions

- Is the set of actions A finite or infinite?
- For a finite set A, is it a simple collection of objects, or a combinatorial one?
- For an infinite set A, is it compact? Is it convex?

There are many questions related to online learning with feedback graphs.

About actions

- Is the set of actions A finite or infinite?
- For a finite set A, is it a simple collection of objects, or a combinatorial one?
- For an infinite set A, is it compact? Is it convex?

About losses

- Are the loss functions generated in a stochastic way, or an adversarial way?
- For compact and convex sets A, is $\mathcal{L}$ a set of convex functions?
- For combinatorial sets A, is $\mathcal{L}$ a set of linear, or submodular functions?

There are many questions related to online learning with feedback graphs.

About actions

- Is the set of actions A finite or infinite?
- For a finite set A, is it a simple collection of objects, or a combinatorial one?
- For an infinite set A, is it compact? Is it convex?

About losses

- Are the loss functions generated in a stochastic way, or an adversarial way?
- For compact and convex sets A, is $\mathcal{L}$ a set of convex functions?
- For combinatorial sets A, is $\mathcal{L}$ a set of linear, or submodular functions?

About feedback graphs

- What are the structural properties of the class $\mathcal{G}$?
- Are the feedback graphs fixed in advance, or can they change over time?
- Are the feedback graphs generated in a stochastic way, or an adversarial way?

As an example of recent results in the setting of dynamic feedback graphs ...

Online Learning with Stochastic Feedback Graphs

- $\mathcal{A}$ is a simple collection $\{1, \cdots, K\}$ of objects.
- $\mathcal{L}$ is a set of (bounded) mappings $\mathcal{A} \to [0, 1]$.
- Loss functions are chosen in an adverarial way.
- The feedback graphs are generated accoding to the Erdös-Renyi model, with parameter r

Theorem 4 (Alon et. al., 2017)

The EXP3-G algorithm achieves a regret of

$$\mathcal{O}\left(\sqrt{\frac{T(1-(1-r)^K)\ln K}{r}}\right)$$

References

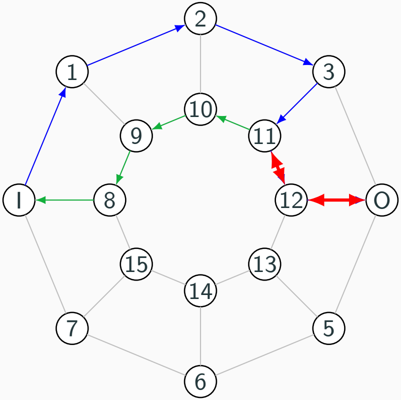
Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 23–35, 2015.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM J. Comput.*, 46(6): 1785–1826, 2017.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2-3):321–352, 2007.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.



Example with Combinatorial Actions: Congestion Games

The forecaster has access a directed graph with a source (I) and a sink (O). Let E be the edge set, and $\mathcal{A} \subseteq \{0,1\}^{|E|}$ be the set of (indicator vectors of) source-sink paths. On each round t:

- The forecaster chooses a path $a_t \in \mathcal{A}$
- The environment chooses a path $\ell_t \in \mathcal{P}$
- The forecaster observes $\langle a_t, \ell_t \rangle$ (number of clashing edges)