Online Learning with Feedback Graphs

Nicolò Cesa-Bianchi

Università degli Studi di Milano

Joint work with:

Noga Alon (Tel-Aviv University)
Ofer Dekel (Microsoft Research)
Tomer Koren (Technion and Microsoft Research)

Also: Claudio Gentile, Shie Mannor, Yishay Mansour, Ohad Shamir

Theory of repeated games

James Hannan
(1922–2010)

David Blackwell
(1919–2010)

Learning to play a game (1956)

Play a game repeatedly against a possibly suboptimal opponent

Prediction with expert advice

N actions

? ? ? ? ? ? ? ? ?

For $t = 1, 2, \ldots$

1. Loss $\ell_t(i) \in [0, 1]$ is assigned to every action $i = 1, \ldots, N$ (hidden from the player)

Prediction with expert advice

N actions

For $t = 1, 2, \ldots$

1. Loss $\ell_t(i) \in [0, 1]$ is assigned to every action $i = 1, \ldots, N$ (hidden from the player)
2. Player picks an action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

Prediction with expert advice

N actions

.7 .3 .2 .4 .1 .6 .7 .4 .9

For $t = 1, 2, \ldots$

1. Loss $\ell_t(i) \in [0,1]$ is assigned to every action $i = 1, \ldots, N$ (hidden from the player)
2. Player picks an action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$
3. Player gets feedback information: $\boldsymbol{\ell}_t = \big(\ell_t(1), \ldots, \ell_t(N)\big)$

Regret

The loss process $\langle \ell_t \rangle_{t \geqslant 1}$ is deterministic and unknown to the (randomized) player $I_1, I_2, \ldots$

Regret of player $I_1, I_2, \ldots$

$$R_T \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \ell_t(i) \stackrel{\text{want}}{=} o(T)$$

Regret

The loss process $\langle \ell_t \rangle_{t \geqslant 1}$ is deterministic and unknown to the (randomized) player $I_1, I_2, \ldots$

Regret of player $I_1, I_2, \ldots$

$$R_T \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \ell_t(i) \stackrel{\text{want}}{=} o(T)$$

Asymptotic lower bound for experts' game

$$\boxed{R_T = \big(1 - o(1)\big)\sqrt{\frac{T \ln N}{2}}}$$

Proof uses an i.i.d. stochastic loss process

Exponentially weighted forecaster

At time t pick action $I_t = i$ with probability proportional to

$$\exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(i)\right)$$

the sum at the exponent is the total loss of action i up to now

Exponentially weighted forecaster

At time t pick action $I_t = i$ with probability proportional to

$$\exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(i)\right)$$

the sum at the exponent is the total loss of action i up to now

Regret bound

If $\eta = \sqrt{\frac{\ln N}{8T}}$ then $\boxed{R_T \leqslant \sqrt{\frac{T \ln N}{2}}}$

Matching asymptotic lower bound including constants

Dynamic choice $\eta_t = \sqrt{(\ln N)/(8t)}$ only loses small constants

The bandit problem: playing an unknown game

N actions

? ? ? ? ? ? ? ? ?

For $t = 1, 2, \ldots$

1. Loss $\ell_t(i) \in [0,1]$ is assigned to every action $i = 1, \ldots, N$ (hidden from the player)

The bandit problem: playing an unknown game

N actions

For $t = 1, 2, \ldots$

1. Loss $\ell_t(i) \in [0, 1]$ is assigned to every action $i = 1, \ldots, N$ (hidden from the player)
2. Player picks an action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

The bandit problem: playing an unknown game

N actions

For $t = 1, 2, \ldots$

1. Loss $\ell_t(i) \in [0, 1]$ is assigned to every action $i = 1, \ldots, N$ (hidden from the player)
2. Player picks an action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$
3. Player gets feedback information: Only $\ell_t(I_t)$ is revealed

The bandit problem: playing an unknown game

N actions

? .3 ? ? ? ? ? ? ?

For $t = 1, 2, \ldots$

1. Loss $\ell_t(i) \in [0,1]$ is assigned to every action $i = 1, \ldots, N$ (hidden from the player)
2. Player picks an action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$
3. Player gets feedback information: Only $\ell_t(I_t)$ is revealed

Many applications

Ad placement, recommender systems, online auctions, . . .

Bandits as an instance of a general feedback model

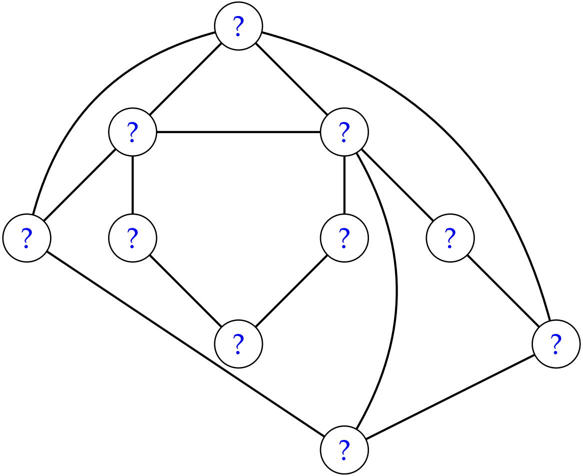
- Besides observing the loss of the played action, the player also observes the loss some other actions
- For example, a recommender system can infer how the user would have reacted had similar products been recommended
- <u>However:</u> we do not insist on assuming that observability between actions implies similarity between losses

Bandits as an instance of a general feedback model

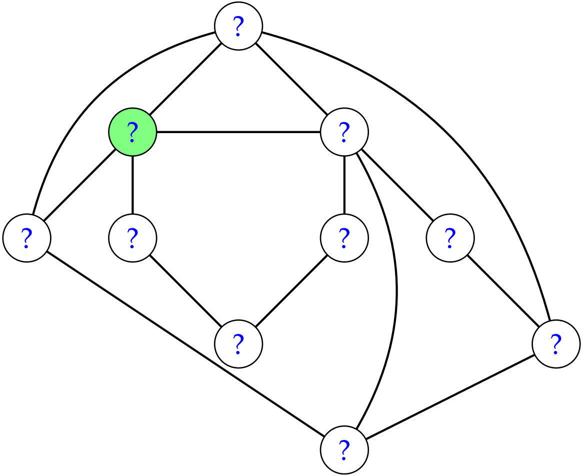
- Besides observing the loss of the played action, the player also observes the loss some other actions
- For example, a recommender system can infer how the user would have reacted had similar products been recommended
- However: we do not insist on assuming that observability between actions implies similarity between losses

How does the observability structure influence regret?

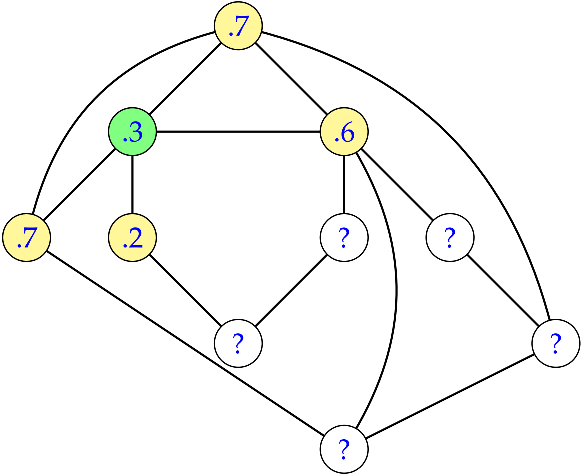
Feedback graph



Feedback graph

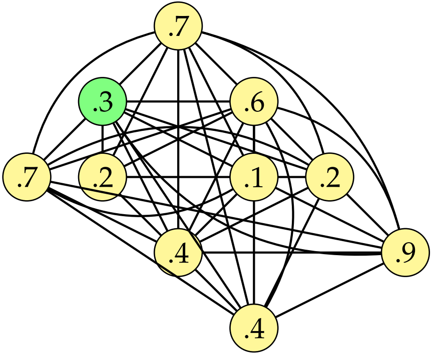


Feedback graph

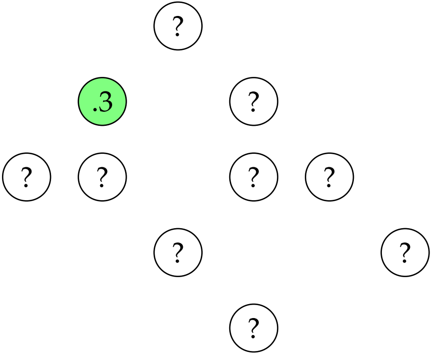


Recovering expert and bandit settings

Experts: clique



Bandits: empty graph



Exponentially weighted forecaster — Reprise

Player's strategy

- $\mathbb{P}_t(I_t = i) \quad \propto \quad \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) \qquad i = 1, \dots, N$
- $\widehat{\ell}_t(i) = \begin{cases} \dfrac{\ell_t(i)}{\mathbb{P}_t\big(\ell_t(i) \text{ observed}\big)} & \text{if } \ell_t(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$

Exponentially weighted forecaster — Reprise

Player's strategy

- $\mathbb{P}_t(I_t = i) \quad \propto \quad \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) \qquad i = 1, \dots, N$
- $\widehat{\ell}_t(i) = \begin{cases} \dfrac{\ell_t(i)}{\mathbb{P}_t\big(\ell_t(i) \text{ observed}\big)} & \text{if } \ell_t(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$

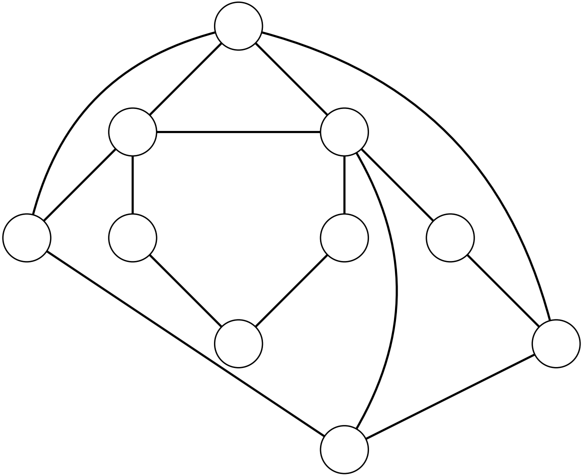
Importance sampling estimator

$\mathbb{E}_t\left[\widehat{\ell}_t(i)\right] = \ell_t(i)$ unbiasedness

$\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right] = \dfrac{\ell_t(i)^2}{\mathbb{P}_t\big(\ell_t(i) \text{ observed}\big)}$ variance control

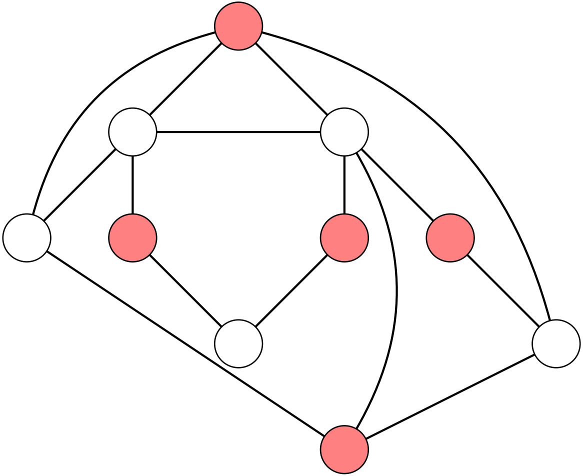
Independence number $\alpha(G)$

The size of the largest independent set



Independence number $\alpha(G)$

The size of the largest independent set



Regret bounds

Analysis (undirected graphs)

$$R_T \leqslant \frac{\ln N}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{\mathbb{P}_t(i \text{ is played})}{\mathbb{P}_t(\ell_t(i) \text{ is observed})}\right]$$

Regret bounds

Analysis (undirected graphs)

$$R_T \leqslant \frac{\ln N}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{\mathbb{P}_t(i \text{ is played})}{\mathbb{P}_t(\ell_t(i) \text{ is observed})}\right]$$

Lemma

For any undirected graph $G = (V, E)$ and for any probability assignment $p_1, \ldots, p_N$ over its vertices

$$\sum_{i=1}^{N}\frac{p_i}{\underbrace{p_i + \sum_{j \in N_G(i)} p_j}_{\mathbb{P}_t(\text{loss of } i \text{ observed})}} \leqslant \alpha(G)$$

Regret bounds

Analysis (undirected graphs)

$$R_T \leqslant \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \alpha(G) = \sqrt{T\alpha(G) \ln N} \qquad \text{by choosing } \eta$$

Regret bounds

Analysis (undirected graphs)

$$R_T \leqslant \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \alpha(G) = \sqrt{T\alpha(G) \ln N} \qquad \text{by choosing } \eta$$

Special cases

Experts (clique): $\alpha(G) = 1$ $\quad R_T \leqslant \sqrt{T \ln N}$

Bandits (empty graph): $\alpha(G) = N$ $\quad R_T \leqslant \sqrt{TN \ln N}$

Regret bounds

Analysis (undirected graphs)

$$R_T \leqslant \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \alpha(G) = \sqrt{T\alpha(G) \ln N} \qquad \text{by choosing } \eta$$

Special cases

Experts (clique):	$\alpha(G) = 1$	$R_T \leqslant \sqrt{T \ln N}$
Bandits (empty graph):	$\alpha(G) = N$	$R_T \leqslant \sqrt{TN \ln N}$

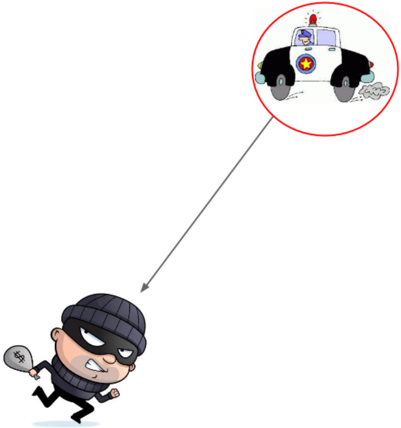
Minimax rate

The general bound is tight: $R_T = \widetilde{\Theta}\left(\sqrt{T\alpha(G) \ln N}\right)$

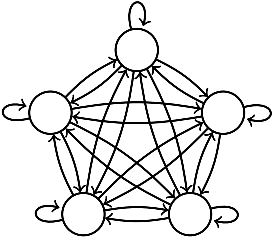
More general feedback models

Directed

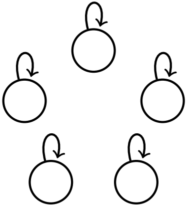
Interventions



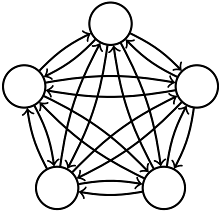
Old and new examples



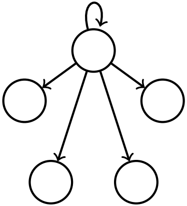
Experts



Bandits



Cops & Robbers



Revealing Action

Exponentially weighted forecaster with exploration

Player's strategy

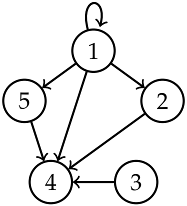
- $$\mathbb{P}_t(I_t = i) = \frac{1-\gamma}{Z_t} \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) + \gamma\, U_G \qquad i = 1, \dots, N$$
- $$\widehat{\ell}_t(i) = \begin{cases} \dfrac{\ell_t(i)}{\mathbb{P}_t\big(\ell_t(i) \text{ observed}\big)} & \text{if } \ell_t(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

U_G is uniform distribution supported on a subset of V

A characterization of feedback graphs

A vertex of G is:

- observable if it has at least one incoming edge (possibly a self-loop)
- strongly observable if it has either a self-loop or incoming edges from all other vertices
- weakly observable if it is observable but not strongly observable



- 3 is not observable
- 2 and 5 are weakly observable
- 1 and 4 are strongly observable

Characterization of minimax rates

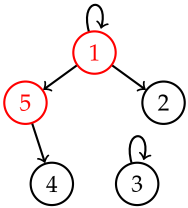
G is strongly observable $\quad R_T = \widetilde{\Theta}\left(\sqrt{\alpha(G)T}\right)$

U_G is uniform on V

G is weakly observable $\quad R_T = \widetilde{\Theta}\left(T^{2/3}\delta(G)\right)$ for $T = \widetilde{\Omega}(N^3)$

U_G is uniform on a weakly dominating set

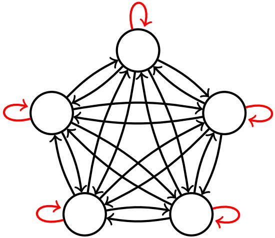
G is not observable $\quad R_T = \Theta(T)$



Weakly dominating set

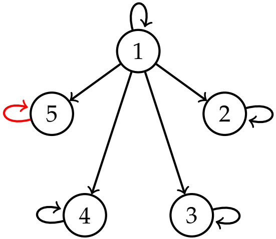
$\delta(G)$ is the size of the smallest set that dominates all weakly observable nodes of G

Some curious cases



Experts vs. Cops & Robbers

Presence of red loops does not affect minimax regret $R_T = \Theta\left(\sqrt{T \ln N}\right)$



Sharp transitions

With red loop: strongly observable with

$\alpha(G) = N - 1 \qquad R_T = \widetilde{\Theta}\left(\sqrt{NT}\right)$

Without red loop: weakly observable with

$\delta(G) = 1 \qquad R_T = \widetilde{\Theta}\left(T^{2/3}\right) \qquad$ for $T = \widetilde{\Omega}(N^3)$

Final remarks

- Theory extends to time-varying feedback graphs
- In the strongly observable case, algorithm can predict without knowing the graph
- Entire framework is a special case of partial monitoring, but our rates exhibit sharp problem-dependent constants

Final remarks

- Theory extends to time-varying feedback graphs
- In the strongly observable case, algorithm can predict without knowing the graph
- Entire framework is a special case of partial monitoring, but our rates exhibit sharp problem-dependent constants

Graph over actions: more interpretations

- Relatedness (rather than observability) structure on loss assignment
- Delay model for loss observations