From Bandits to Experts: A Tale of Domination and Independence

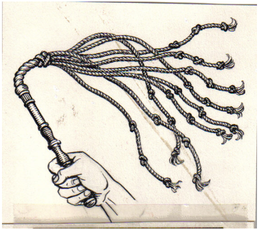
Nicolò Cesa-Bianchi

Università degli Studi di Milano

From Bandits to Experts:
A Tale of Domination and Independence

Nicolò Cesa-Bianchi

Università degli Studi di Milano



Joint work with:

Noga Alon
Ofer Dekel
Tomer Koren

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

Zero-sum 2-person games played more than once

	1	2	$\dots$	M
1	$\ell(1,1)$	$\ell(1,2)$	$\dots$	
2	$\ell(2,1)$	$\ell(2,2)$	$\dots$	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
N				

$N \times M$ known loss matrix over $\mathbb{R}$

- Row player (player) has N actions
- Column player (opponent) has M actions

For each game round $t = 1, 2, \dots$

- Player chooses action i_t and opponent chooses action y_t
- The player suffers loss $\ell(i_t, y_t)$ (= gain of opponent)

Player can learn from opponent's history of past choices $y_1, \dots, y_{t-1}$

Prediction with expert advice

Volodya Vovk

Manfred Warmuth

	$t=1$	$t=2$	$\dots$
1	$\ell_1(1)$	$\ell_2(1)$	$\dots$
2	$\ell_1(2)$	$\ell_2(2)$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$
N	$\ell_1(N)$	$\ell_2(N)$	

Play an unknown loss matrix

Opponent's moves $y_1, y_2, \dots$ define a sequential prediction problem with a time-varying loss function $\ell(i_t, y_t) = \ell_t(i_t)$

Playing the experts 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)

Playing the experts 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 $\ell_t(I_t)$

Playing the experts game

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: $\boldsymbol{\ell}_t = \big(\ell_t(1), \ldots, \ell_t(N)\big)$

Oblivious opponents

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

Oblivious regret minimization

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

Bounds on regret [How to use expert advice, 1997]

Lower bound using random losses

- Losses $\ell_t(i)$ are independent random coin flips $L_t(i) \in \{0,1\}$
- For any player strategy $\mathbb{E}\left[\sum_{t=1}^{T} L_t(I_t)\right] = \frac{T}{2}$
- Then the expected regret is

$$\mathbb{E}\left[\max_{i=1,\dots,N} \sum_{t=1}^{T} \left(\frac{1}{2} - L_t(i)\right)\right] = \big(1 - o(1)\big)\sqrt{\frac{T \ln N}{2}}$$

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 [How to use expert advice, 1997]

- If $\eta = \sqrt{(\ln N)/(8T)}$ then $\boxed{R_T \leqslant \sqrt{\frac{T \ln N}{2}}}$
- Matching 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

? 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

The bandit problem: playing an unknown game

N actions

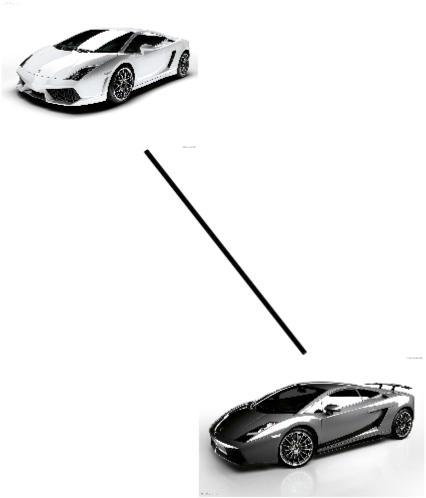
? 3 ? ? ? ? ? ? ?

For $t = 1, 2, \ldots$

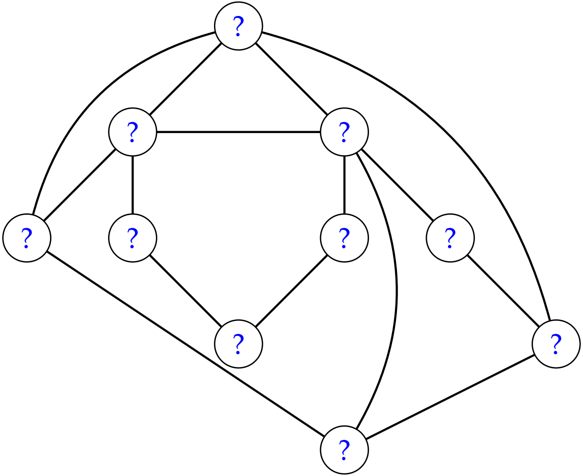
1. Loss $\ell_t(i) \in [0, 1]$ is assigned to every action $i = 1, \ldots, N$ (hidden from the player)
2. Player picks an action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$
3. Player gets feedback information: Only $\ell_t(I_t)$ is revealed

Many applications

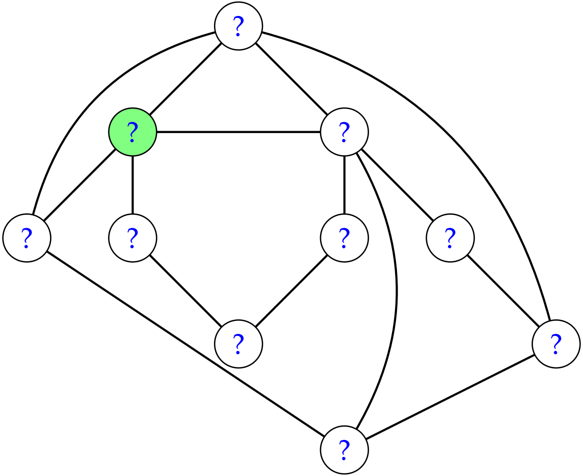
Ad placement, dynamic content adaptation, routing, online auctions



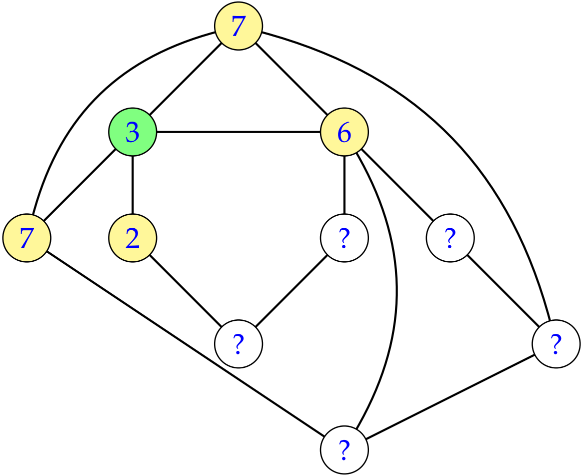
A graph of relationships over actions



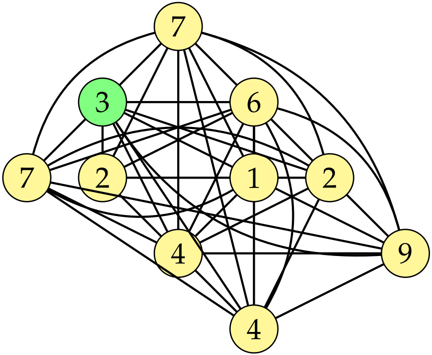
A graph of relationships over actions



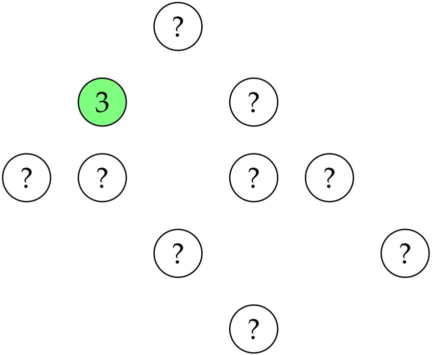
A graph of relationships over actions



Experts: clique



Bandits: empty graph



Exponentially weighted forecaster — Reprise

Player's strategy [Alon, C-B, Gentile, Mannor, Mansour and Shamir, 2013]

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

Importance sampling estimator

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

$$\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right] \leqslant \frac{1}{\mathbb{P}_t\big(\ell_t(i) \text{ observed}\big)} \qquad \text{variance control}$$

Regret bounds

Analysis (undirected graphs)

$$R_T \leqslant \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{\mathbb{P}_t(I_t = i)}{\mathbb{P}_t(I_t = i) + \sum_{j \in N_G(i)} \mathbb{P}_t(I_t = j)}$$

Lemma

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

$$\sum_{i=1}^{N} \frac{p_i}{p_i + \sum_{j \in N_G(i)} p_j} \leqslant \alpha(G)$$

$\alpha(G)$ is the independence number of G (largest subset of V such that no two distinct vertices in it are adjacent in G)

Regret bounds

Analysis (undirected graphs)

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

Special cases

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

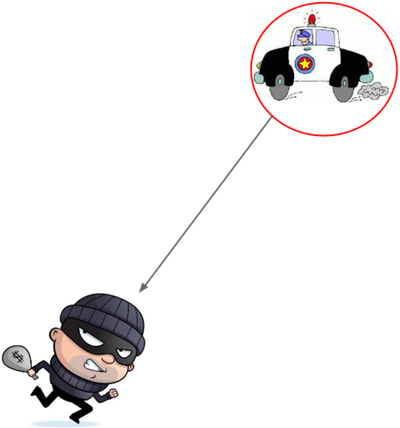
Minimax rate

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

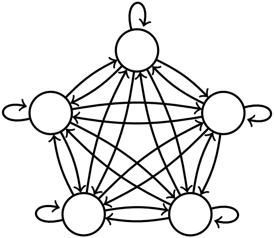
More general feedback models

Directed

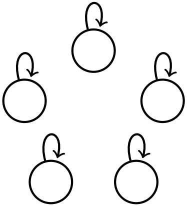
Interventions



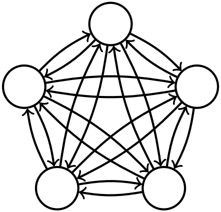
Old and new examples



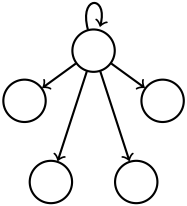
Experts



Bandits



Cops & Robbers



Revealing Action

Exponentially weighted forecaster with exploration

Player's strategy [Alon, C-B, Dekel and Koren, 2015]

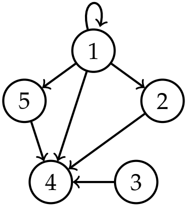
- $$\mathbb{P}_t(I_t = i) \propto \frac{1-\gamma}{Z_t} \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right) + \gamma\, U_G \qquad i = 1, \dots, N$$
- $$\widehat{\ell}_t(i) = \begin{cases} \dfrac{\ell_t(i)}{\mathbb{P}_t\big(\ell_t(i) \text{ observed}\big)} & \text{if } \ell_t(i) \text{ is observed}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

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

Minimax rates

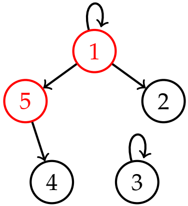
G is strongly observable $\quad \boxed{R_T = \widetilde{\Theta}\left(\sqrt{\alpha(G)T}\right)}$

U_G is uniform on V

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

U_G is uniform on a weakly dominating set

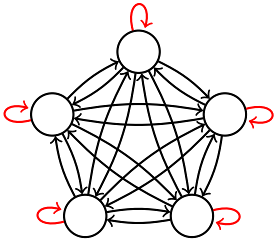
G is not observable $\quad R_T = \Theta(T)$

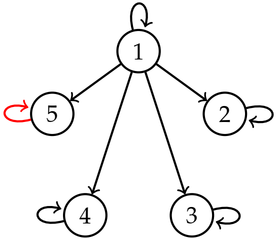


Weakly dominating set

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

Minimax regret





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

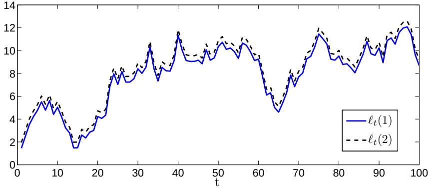
With red loop: strongly observable with $\alpha(G) = N - 1$
$R_T = \widetilde{\Theta}\left(\sqrt{NT}\right)$

Without red loop: weakly observable with $\delta(G) = 1$
$R_T = \widetilde{\Theta}\left(T^{2/3}\right)$

The loss of action i at time t depends on the player's past m actions
$\ell_t(i) \to L_t(I_{t-m}, \ldots, I_{t-1}, i)$

Adaptive regret

$$R_T^{\mathrm{ada}} = \mathbb{E}\left[\sum_{t=1}^{T} L_t(I_{t-m}, \ldots, I_{t-1}, I_t) - \min_{i=1,\ldots,N} \sum_{t=1}^{T} L_t(\underbrace{i, \ldots, i}_{m \text{ times}}, i)\right]$$



Minimax rate ($m > 0$)

$$R_T^{\mathrm{ada}} = \Theta\left(T^{2/3}\right)$$

Conclusions

- An abstract, game-theoretic framework for studying a variety of sequential decisions problems
- Applicable to machine learning (e.g., binary classification) and online convex optimization settings
- Exponential weights can be replaced by polynomial weights (cfr. Mirror Descent for convex optimization)
- Connections to gambling, portfolio management, competitive analysis of algorithms