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

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: $\ell_t = (\ell_t(1), \ldots, \ell_t(N))$
Regret

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

Regret of player $I_1, I_2, \ldots$

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

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

Regret of player $I_1, I_2, \ldots$

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

Asymptotic lower bound for experts' game

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

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

$$R_T \leq \sqrt{T \ln N}$$

Matching asymptotic lower bound including constants

Dynamic choice $\eta_t = \sqrt{\frac{(\ln N)}{8t}}$ 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

Many applications
Ad placement, recommender systems, online auctions, \ldots
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

Many applications

Ad placement, recommender systems, online auctions, \ldots
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.
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
Recovering expert and bandit settings

Experts: clique

Bandits: empty graph
Player’s strategy

\[ P_t(I_t = i) \propto \exp \left( -\eta \sum_{s=1}^{t-1} \hat{\ell}_s(i) \right) \quad i = 1, \ldots, N \]

\[ \hat{\ell}_t(i) = \begin{cases} \frac{\ell_t(i)}{P_t(\ell_t(i) \text{ observed})} & \text{if } \ell_t(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases} \]
Exponentially weighted forecaster — Reprise

Player’s strategy

\[ P_t(I_t = i) \propto \exp \left( -\eta \sum_{s=1}^{t-1} \hat{\ell}_s(i) \right) \quad i = 1, \ldots, N \]

\[ \hat{\ell}_t(i) = \begin{cases} \frac{\ell_t(i)}{P_t(\ell_t(i) \text{ observed})} & \text{if } \ell_t(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases} \]

Importance sampling estimator

\[ \mathbb{E}_t \left[ \hat{\ell}_t(i) \right] = \ell_t(i) \quad \text{unbiasedness} \]

\[ \mathbb{E}_t \left[ \hat{\ell}_t(i)^2 \right] = \frac{\ell_t(i)^2}{P_t(\ell_t(i) \text{ observed})} \quad \text{variance control} \]
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)

\[ R_T \leq \frac{\ln N}{\eta} + \frac{\eta}{2} \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{P_t(i \text{ is played})}{P_t(\ell_t(i) \text{ is observed})} \right] \]
Regret bounds

Analysis (undirected graphs)

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

Lemma

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

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

\( P_t(\text{loss of } i \text{ observed}) \)
Regret bounds

Analysis (undirected graphs)

\[ R_T \leq \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \alpha(G) = \sqrt{T\alpha(G) \ln N} \]

by choosing \( \eta \)

Special cases

Experts (clique):
\[ \alpha(G) = 1 \]

Bandits (empty graph):
\[ \alpha(G) = N \]

Minimax rate

The general bound is tight:
\[ R_T = \tilde{\Theta} \left( \sqrt{T\alpha(G) \ln N} \right) \]
Regret bounds

Analysis (undirected graphs)

\[ R_T \leq \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \alpha(G) = \sqrt{T\alpha(G) \ln N} \]

by choosing \( \eta \)

Special cases

**Experts** (clique): \( \alpha(G) = 1 \) \( R_T \leq \sqrt{T \ln N} \)

**Bandits** (empty graph): \( \alpha(G) = N \) \( R_T \leq \sqrt{TN \ln N} \)
Regret bounds

Analysis (undirected graphs)

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

Special cases

**Experts** (clique):
\[ \alpha(G) = 1 \quad R_T \leq \sqrt{T \ln N} \]

**Bandits** (empty graph):
\[ \alpha(G) = N \quad R_T \leq \sqrt{TN \ln N} \]

Minimax rate

The general bound is tight:
\[ R_T = \tilde{\Theta}(\sqrt{T\alpha(G) \ln N}) \]
More general feedback models

Directed

Interventions
Old and new examples

Experts

Bandits

Cops & Robbers

Revealing Action
Exponentially weighted forecaster with exploration

Player’s strategy

\[ P_t(I_t = i) = \frac{1 - \gamma}{Z_t} \exp \left( -\eta \sum_{s=1}^{t-1} \hat{\ell}_s(i) \right) + \gamma U_G \quad i = 1, \ldots, N \]

\[ \hat{\ell}_t(i) = \begin{cases} 
\ell_t(i) & \text{if } \ell_t(i) \text{ is observed} \\
\frac{P_t(\ell_t(i) \text{ observed})}{\ell_t(i) \text{ observed}} & 0 \text{ otherwise}
\end{cases} \]

\( U_G \) 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

- **1** and **4** are strongly observable
- **2** and **5** are weakly observable
- **3** is not observable
Characterization of minimax rates

- **G is strongly observable**
  \[ R_T = \tilde{\Theta}(\sqrt{\alpha(G)T}) \]
  \[ U_G \text{ is uniform on } V \]

- **G is weakly observable**
  \[ R_T = \tilde{\Theta}(T^{2/3}\delta(G)) \text{ for } T = \tilde{\Omega}(N^3) \]
  \[ U_G \text{ is uniform on a weakly dominating set} \]

- **G is not observable**
  \[ R_T = \Theta(T) \]

**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 $R_T = \Theta(\sqrt{T \ln N})$

Sharp transitions
With red loop: strongly observable with $\alpha(G) = N - 1$ $R_T = \tilde{\Theta}(\sqrt{NT})$
Without red loop: weakly observable with $\delta(G) = 1$ $R_T = \tilde{\Theta}(T^{2/3})$ for $T = \tilde{\Omega}(N^3)$
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