Nonstochastic Bandit Problems on Graphs

Nicolò Cesa-Bianchi

Università degli Studi di Milano

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	...	M
1	$\ell(1,1)$	$\ell(1,2)$	...	
2	$\ell(2,1)$	$\ell(2,2)$	...	
⋮	⋮	⋮	⋱	
K				

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

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

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

- 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 $y_1, \ldots, y_{t-1}$

Prediction with expert advice

Volodya Vovk

Manfred Warmuth

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

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

A sequential decision problem

- K actions
- Unknown deterministic assignment of losses to actions $\ell_t = \big(\ell_t(1), \dots, \ell_t(K)\big) \in [0,1]^K$ for $t = 1, 2, \dots$

For $t = 1, 2, \dots$

Playing the experts game

A sequential decision problem

- K actions
- Unknown deterministic assignment of losses to actions $\ell_t = \big(\ell_t(1), \ldots, \ell_t(K)\big) \in [0,1]^K$ for $t = 1, 2, \ldots$

For $t = 1, 2, \ldots$

1. Player picks an action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

Playing the experts game

A sequential decision problem

- K actions
- Unknown deterministic assignment of losses to actions $\ell_t = (\ell_t(1), \ldots, \ell_t(K)) \in [0,1]^K$ for $t = 1, 2, \ldots$

For $t = 1, 2, \ldots$

1. Player picks an action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$
2. Player gets feedback information: $\ell_t(1), \ldots, \ell_t(K)$

Regret analysis

Regret

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

Regret analysis

Regret

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

Lower bound using random losses [Experts' paper, 1997]

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

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

for $K, T \to \infty$

Exponentially weighted forecaster (Hedge)

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

[Experts' paper, 1997]

- If $\eta = \sqrt{\frac{\ln K}{8T}}$ then $R_T \leqslant \sqrt{\frac{T \ln K}{2}}$
- This matches the asymptotic lower bound, including constants

The bandit problem: playing an unknown game

- K actions
- Unknown deterministic assignment of losses to actions
  $\ell_t = \big(\ell_t(1), \ldots, \ell_t(K)\big) \in [0,1]^K$ for $t = 1, 2, \ldots$

For $t = 1, 2, \ldots$

The bandit problem: playing an unknown game

- K actions
- Unknown deterministic assignment of losses to actions $\ell_t = (\ell_t(1), \ldots, \ell_t(K)) \in [0,1]^K$ for $t = 1, 2, \ldots$

For $t = 1, 2, \ldots$

1. Player picks an action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

The bandit problem: playing an unknown game

- K actions
- Unknown deterministic assignment of losses to actions $\ell_t = (\ell_t(1), \ldots, \ell_t(K)) \in [0,1]^K$ for $t = 1, 2, \ldots$

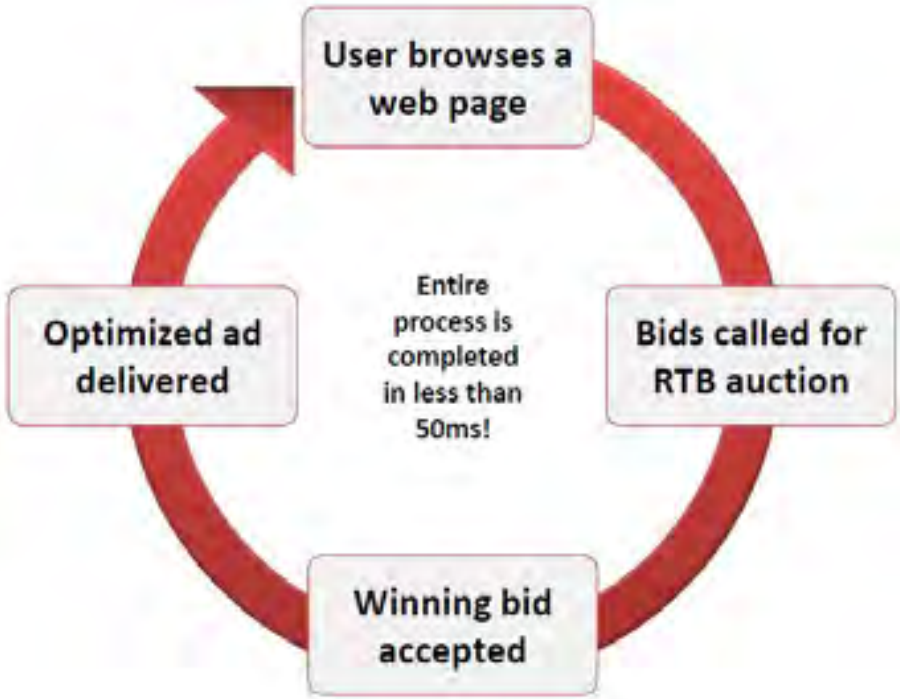
For $t = 1, 2, \ldots$

1. Player picks an action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$
2. Player gets feedback information: Only $\ell_t(I_t)$ is revealed

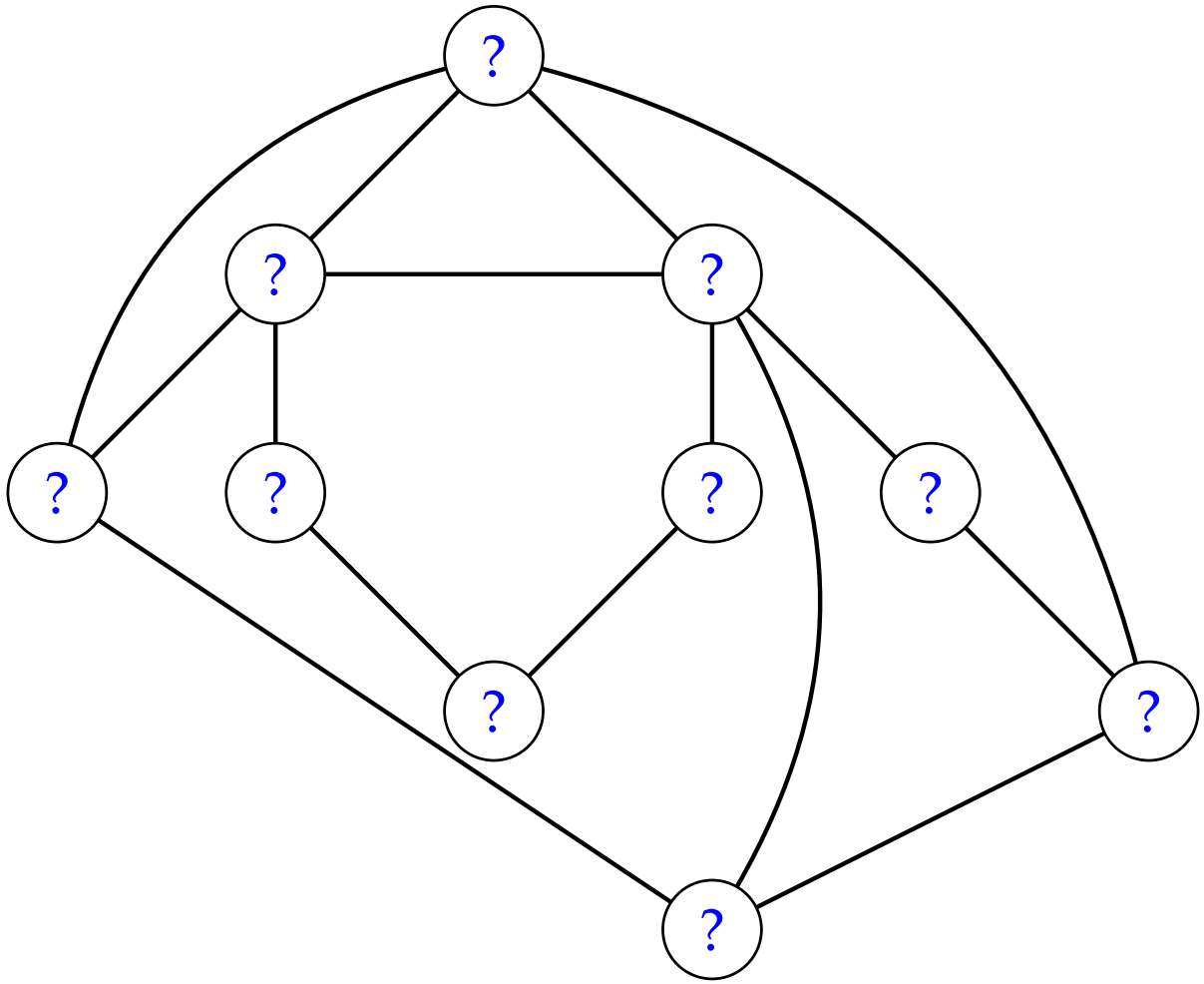
A growing range of applications

- Ad placement
- Dynamic content/layout optimization
- Recommender systems
- Clinical trials
- Network protocol optimization

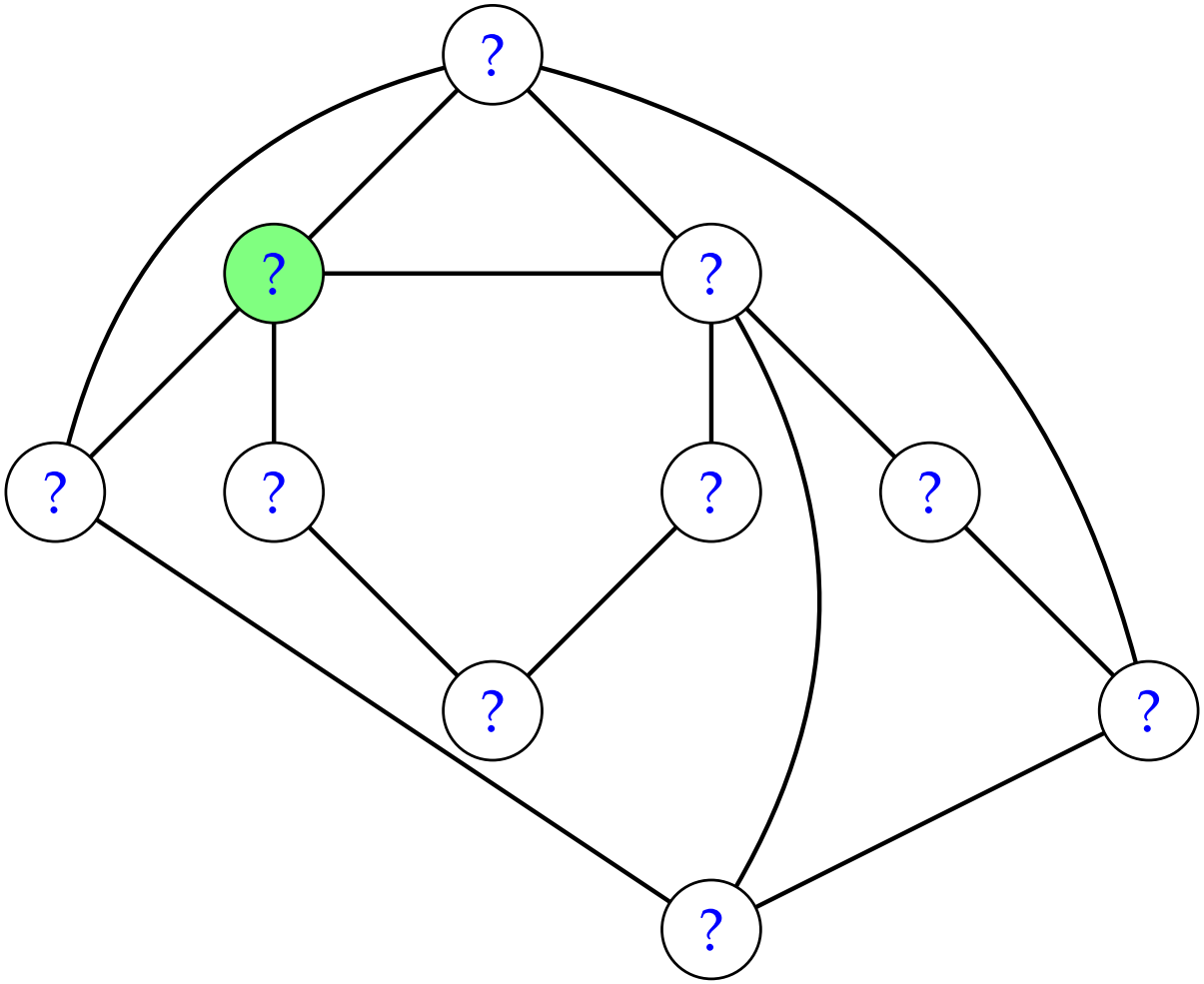
Relationships between actions [Mannor and Shamir, 2011]



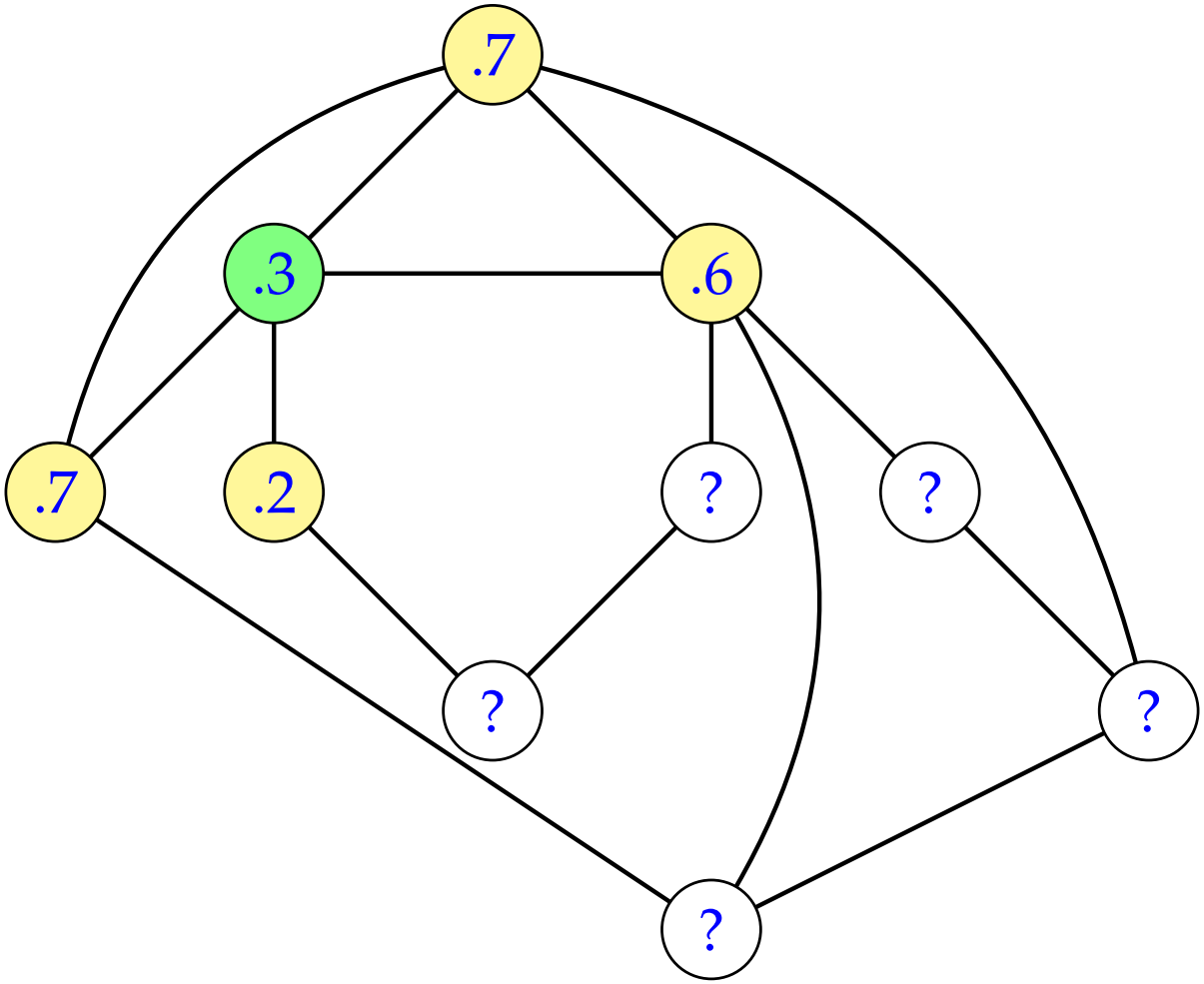
A graph of relationships over actions



A graph of relationships over actions

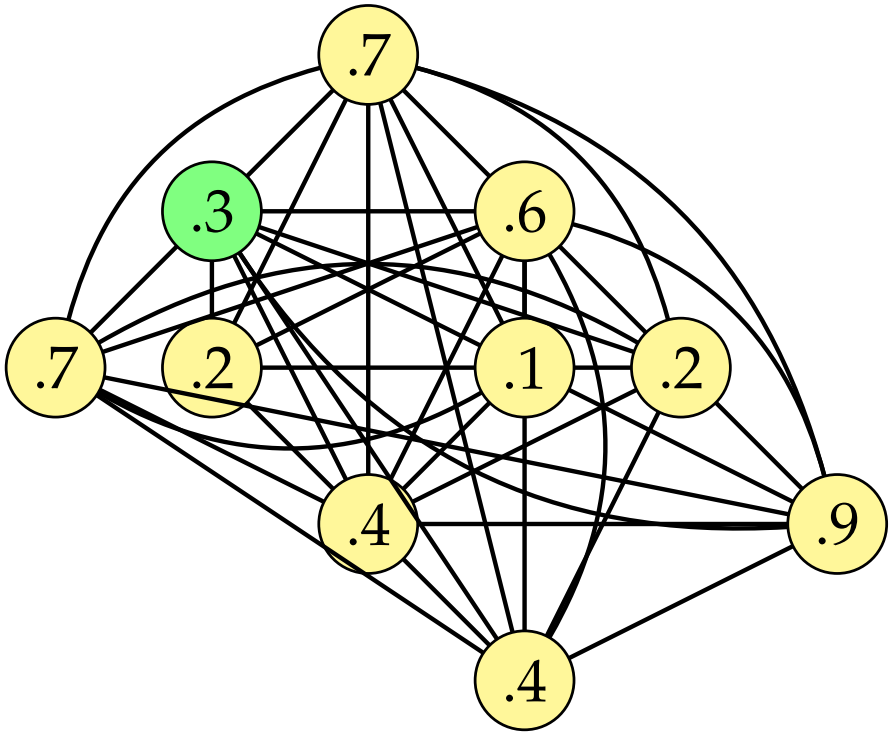


A graph of relationships over actions

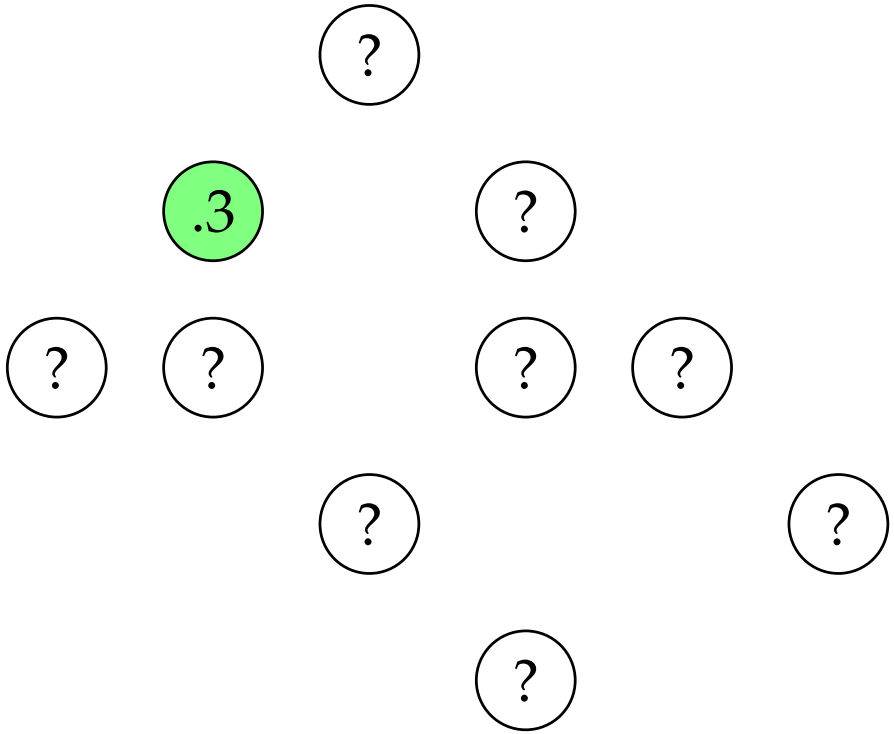


Recovering expert and bandit settings

Experts: clique



Bandits: empty graph



Hedge revisited

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

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

Hedge revisited

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

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

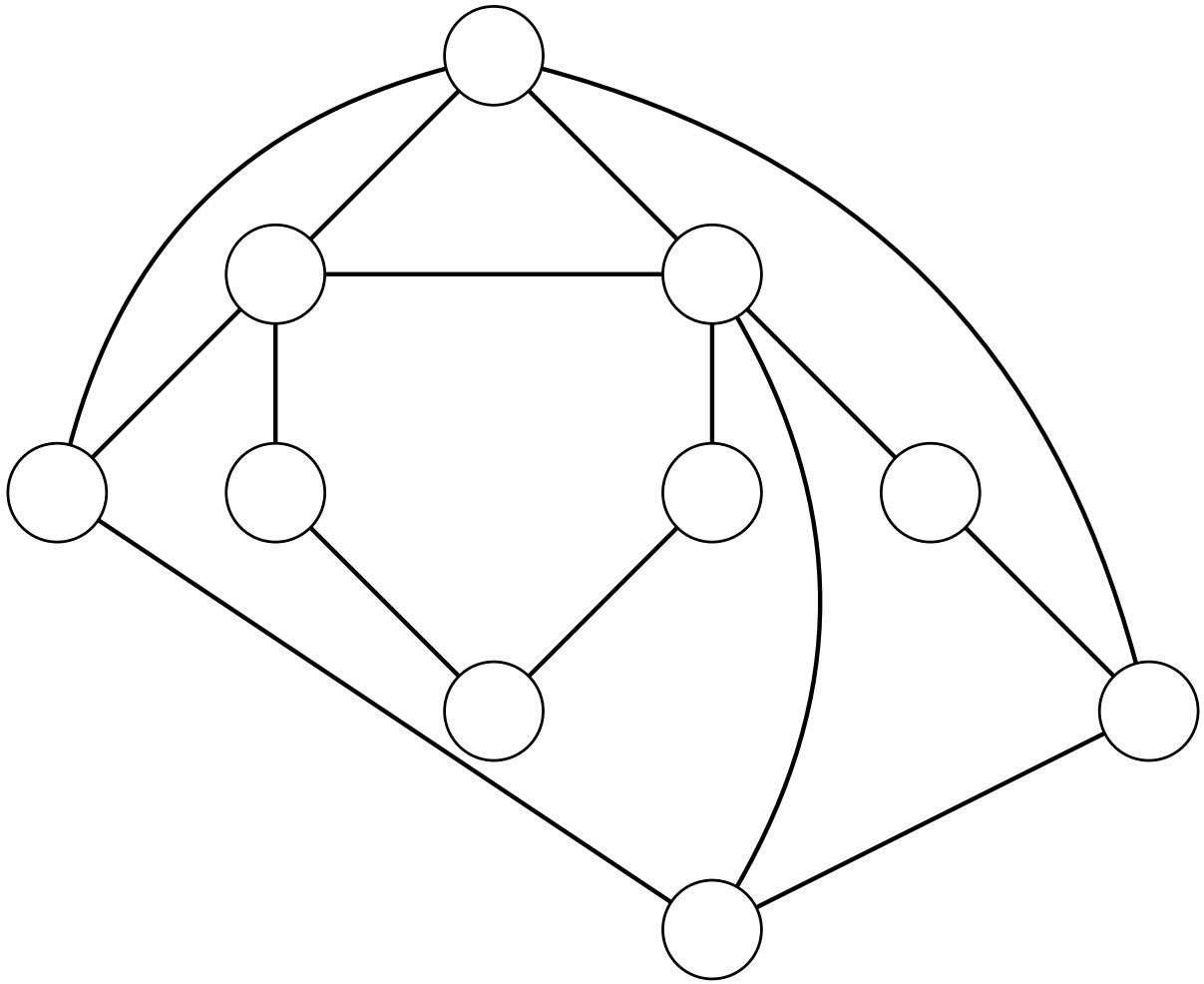
Importance sampling estimator

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

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

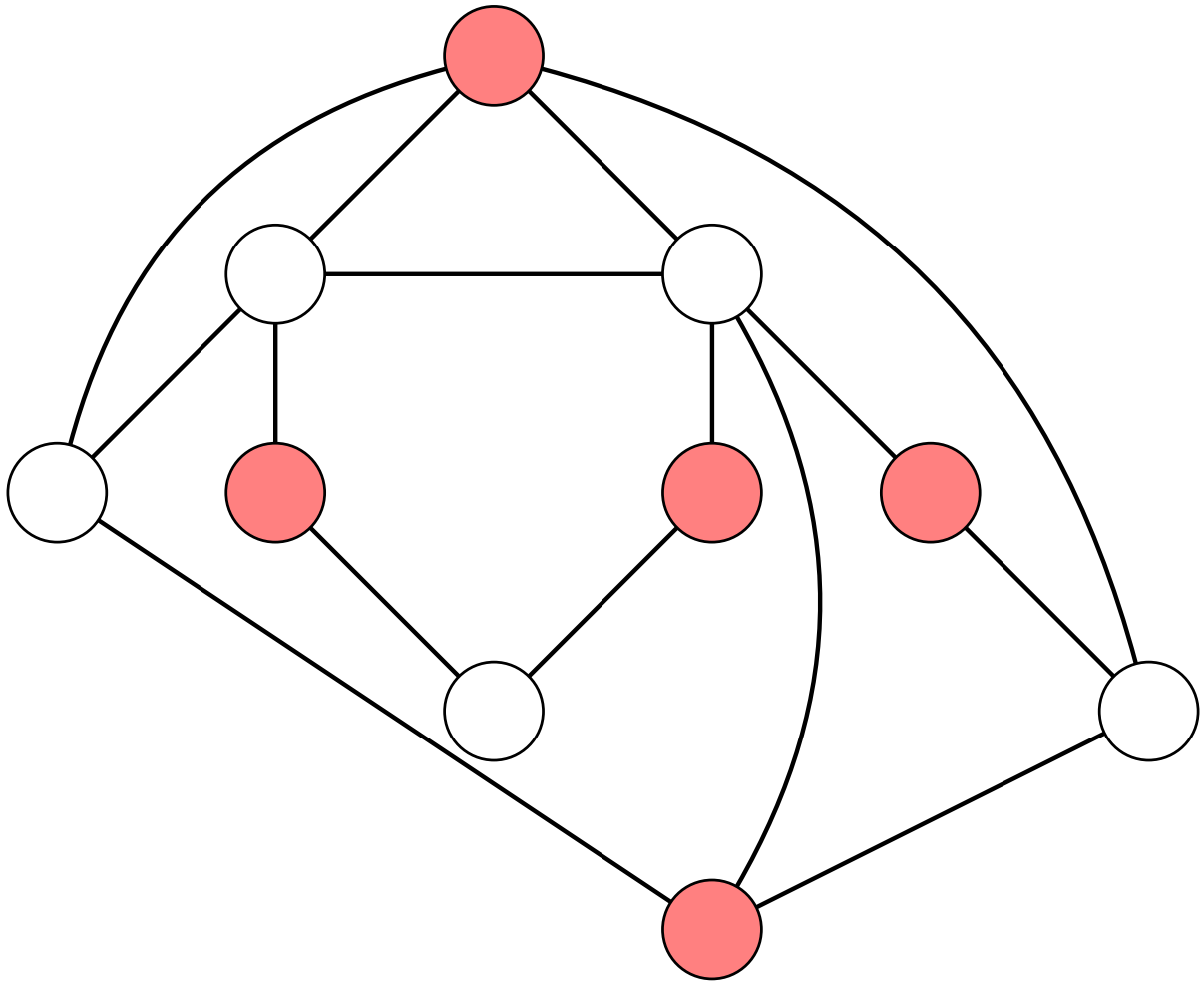
Independence number $\alpha(G)$

The size of the largest independent set



Independence number $\alpha(G)$

The size of the largest independent set



Regret bound

$$
\begin{aligned}
R_T &\leqslant \frac{\ln K}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} \mathbb{P}_t(I_t = i)\,\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right]\right] \\
&\leqslant \frac{\ln K}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} \frac{\mathbb{P}_t(I_t = i)}{\mathbb{P}_t\big(\ell_t(i) \text{ is observed}\big)}\right] \qquad \text{variance control} \\
&\leqslant \frac{\ln K}{\eta} + \frac{\eta}{2} T\alpha(G) \quad = \quad \sqrt{T\alpha(G)\ln K}
\end{aligned}
$$

$\alpha(G)$ is the independence number of G

Regret bound

$$R_T \leqslant \frac{\ln K}{\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} \mathbb{P}_t(I_t = i)\,\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right]\right]$$

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

$$\leqslant \frac{\ln K}{\eta} + \frac{\eta}{2} T\alpha(G) \quad = \quad \sqrt{T\alpha(G)\ln K}$$

$\alpha(G)$ is the independence number of G

Special cases

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

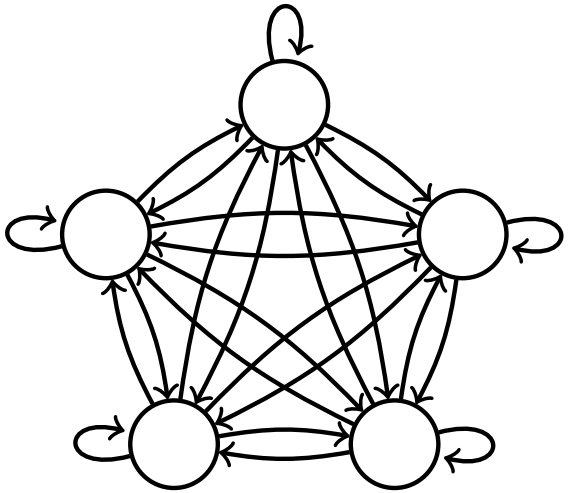
Bandits (empty graph): $\alpha(G) = K$ $\quad R_T \leqslant \sqrt{TK \ln K}$

More general feedback models [C-B, Dekel, and Koren, 2015]

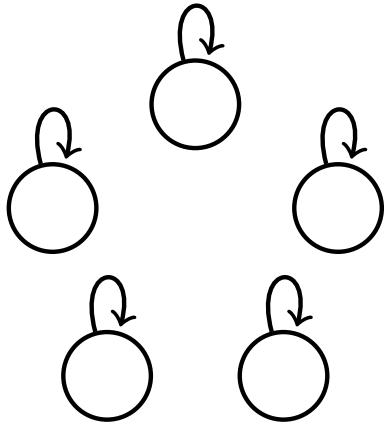
Directed

Interventions

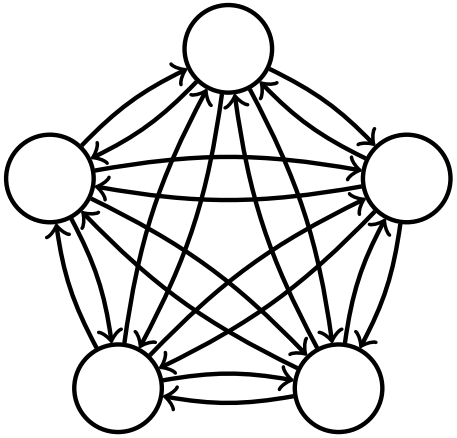
Old and new examples



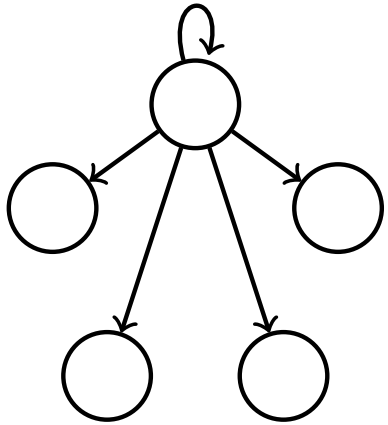
Experts



Bandits



Cops & Robbers

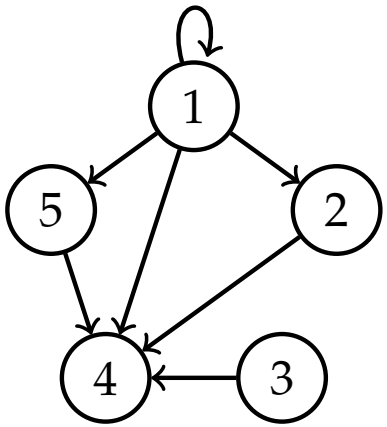


Revealing Action

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

Hedge uses additional exploration due to reduced feedback

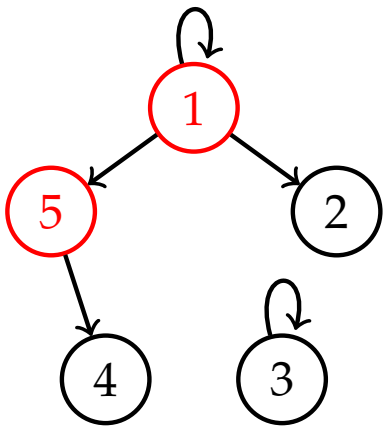
G is strongly observable $\quad R_T = \widetilde{\Theta}\left(\sqrt{\alpha(G)T}\right)$

Hedge mixed with uniform distribution

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

Hedge mixed with uniform distribution on a weakly dominating set

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



Weakly dominating set

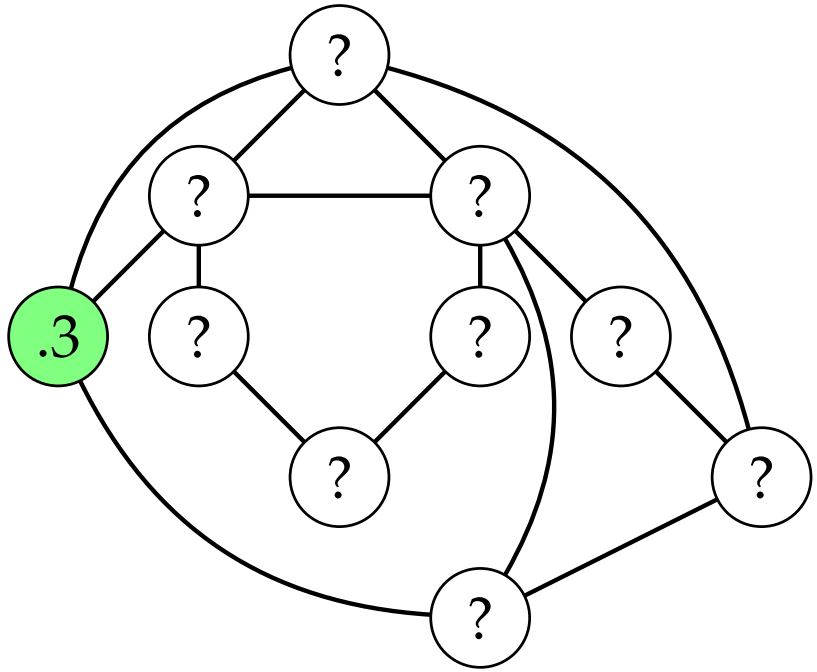
$\delta(G)$ is the size of the smallest set that dominates all weakly observable nodes of G

Smooth losses

[C-B and Shamir, 2017]

- Back in the bandit model: we only observe the loss of the chosen action
- But we want to take advantage of scenarios where ℓ_t is smooth over the graph:

$$\sum_{(i,j)\in E} \big(\ell_t(i) - \ell_t(j)\big)^2 \text{ is small}$$



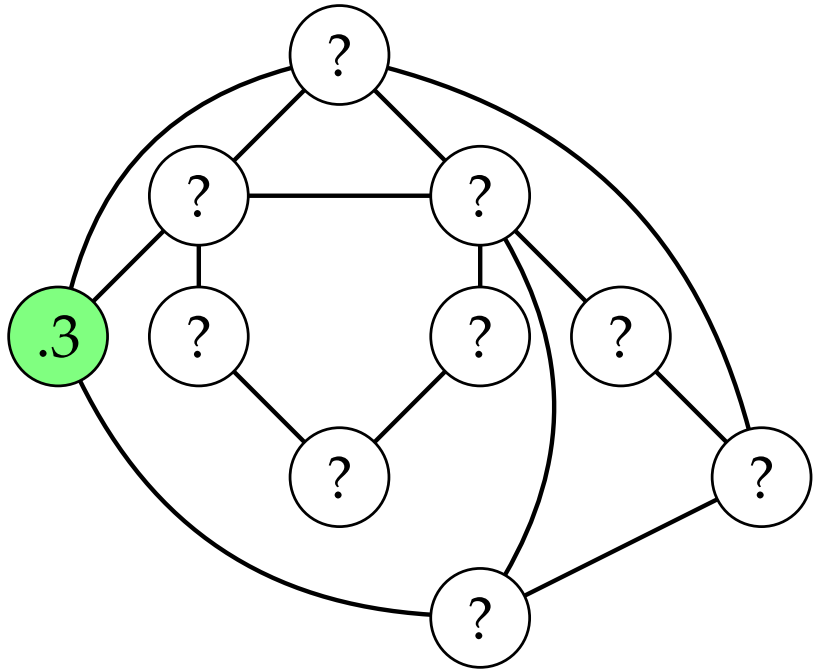
Smooth losses

[C-B and Shamir, 2017]

- Back in the bandit model: we only observe the loss of the chosen action
- But we want to take advantage of scenarios where ℓ_t is smooth over the graph:

$$\sum_{(i,j)\in E} \big(\ell_t(i) - \ell_t(j)\big)^2 \text{ is small}$$

- We also need to know the smallest loss $\min_{i=1,\dots,K} \ell_t(i)$
  (no need to know the action achieving it)



Regret with smooth losses

$$R_T \leqslant \sqrt{(\ln K)\frac{C^2}{\lambda_2(L)}T}$$

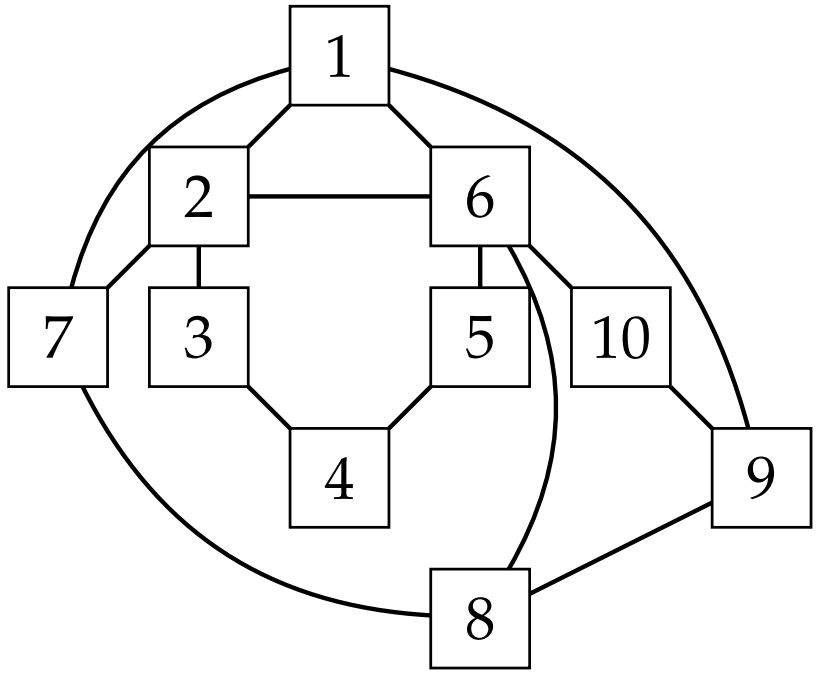
for the Laplacian L of any simple and connected graph such that

$$\underbrace{\sum_{(i,j)\in E} \big(\ell_t(i) - \ell_t(j)\big)^2}_{\ell_t^\top L \ell_t} \leqslant C^2 \qquad \text{for all } t = 1, \dots, T$$

- $\lambda_2(K) \in [0, K]$ is the smallest nonzero eigenvalue of L measuring connectivity of the graph
- This regret bound vanishes when $\ell_t(1) = \cdots = \ell_t(K)$ for all t

Cooperation with delay [C-B, Gentile, Mansour and Minora, 2016]

- N agents sitting on the vertices of an unknown communication graph
- Agents cooperate to solve a common bandit problem
- Each agent runs an instance of the same bandit algorithm



The communication protocol with delay d

For each $t = 1, \ldots, T$ each agent $v \in V$ does the following:

1. Plays an action $I_t(v)$ drawn according to his private distribution $\mathbf{p}_t(v)$ observing loss $\ell_t\big(I_t(v)\big)$ (same loss vector for all agents)
2. Sends to his neighbors the message
$$m_t(v) = \Big\langle t, v, I_t(v), \ell_t\big(I_t(v)\big), \mathbf{p}_t(v) \Big\rangle$$
3. Receives messages from his neighbors, forwarding those that are not older than d

The communication protocol with delay d

For each $t = 1, \ldots, T$ each agent $v \in V$ does the following:

1. Plays an action $I_t(v)$ drawn according to his private distribution $\mathbf{p}_t(v)$ observing loss $\ell_t\big(I_t(v)\big)$ (same loss vector for all agents)
2. Sends to his neighbors the message
$$m_t(v) = \Big\langle t, v, I_t(v), \ell_t\big(I_t(v)\big), \mathbf{p}_t(v) \Big\rangle$$
3. Receives messages from his neighbors, forwarding those that are not older than d

- An agent receives a message from another agent with a delay equal to the shortest path between them
- A message sent by some agent v at time t will be received by all agents whose shortest-path distance from v is at most d

Average welfare regret

$$R_T^{\text{coop}} = \frac{1}{N} \sum_{v \in V} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t\left(I_t(v)\right)\right] - \min_{i=1,\dots,K} \sum_{t=1}^{T} \ell_t(i)$$

Remarks

- Clearly, $R_T^{\text{coop}} \leqslant \sqrt{TK \ln K}$ when agents run vanilla Exp3 (no cooperation)
- By using other agent's plays, each agent may estimate ℓ_t better (thus learning nearly at full info rate)
- In general, d trades off between quality and quantity of information

Key inequality

$$\mathbb{E}\left[\sum_{v\in V}\sum_{i=1}^{K}\frac{\mathbb{P}_t\big(I_t(v)=i\big)}{\mathbb{P}_{t-d}\big(\ell_{t-d}(i)\text{ is observed by }v\big)}\right] \leqslant \frac{e}{1+e^{-1}}\big(K\,\alpha_d+N\big)$$

α_d is the independence number of the graph obtained from G by connecting any two vertices whose shortest path distance is at most d

Average welfare regret bound

$$R_T^{\text{coop}} \leqslant \sqrt{\underbrace{\left((d+1) + \frac{K}{N}\,\alpha_d\right)}_{\text{main term}} \underbrace{T \ln K}_{\text{unavoidable}}}$$

Average welfare regret bound

$$R_T^{\text{coop}} \leqslant \sqrt{\underbrace{\left((d+1) + \frac{K}{N}\,\alpha_d\right)}_{\text{main term}} \underbrace{T \ln K}_{\text{unavoidable}}}$$

$d = K^{1/2}$ with any connected graph G

- Then $\alpha_d \leqslant \frac{2N}{d+2}$ and

$$R_T^{\text{coop}} \leqslant \sqrt{K^{1/2} T \ln K}$$

- This is better than $\sqrt{KT}$ (minimax for non-cooperating bandits)

Final remarks

- Feedback graphs provide a unified model for studying experts, bandits, and more general settings

Final remarks

- Feedback graphs provide a unified model for studying experts, bandits, and more general settings
- We can characterize the minimax regret achievable on any directed feedback graph in terms of its structure

Final remarks

- Feedback graphs provide a unified model for studying experts, bandits, and more general settings
- We can characterize the minimax regret achievable on any directed feedback graph in terms of its structure
- Besides observability, graphs can be used to define bandit problems with smooth losses

Final remarks

- Feedback graphs provide a unified model for studying experts, bandits, and more general settings
- We can characterize the minimax regret achievable on any directed feedback graph in terms of its structure
- Besides observability, graphs can be used to define bandit problems with smooth losses
- Communication graphs provide a natural setting for the study of interacting bandits using tools similar to those developed for the analysis of feedback graphs