Pure Exploration with Feedback Graphs

Alessio Russo, Yichen Song and Aldo Pacchiano

AISTATS 2025

Pacchiano's Lab for Adaptive and Intelligent Algorithms (PLAIA)

Best Arm Identification with Fixed Confidence

K arms with reward distributions ν_a with $a \in \{1, \ldots, K\}$. Assume $(\nu_a)_a$ belong to the family of single-parameter exponential distributions, with $\mu_a = \mathbb{E}_{r \sim \nu_a}[r]$.

μ_1

μ_2

μ_3

μ_4

μ_5

- **Sequential**: In round t the learner pulls arm $a_t \in [K]$ and receives the reward $r_t \sim \nu_{a_t}$.
- **Best Arm Identification objective**: quickly find the optimal arm $a^\star = \arg\max_a \mu_a$ with confidence $\delta \in (0, 1/2) \Rightarrow$ minimize sample complexity $\mathbb{E}[\tau]$ subject to $\mathbb{P}(\hat{a}_\tau \neq a^\star) \leq \delta$.
  - τ is a random stopping time and $\hat{a}_\tau$ is the estimated best arm at τ.

Let's introduce a graph structure

Best Arm Identification with Fixed Confidence

K arms with reward distributions ν_a with $a \in \{1, \dots, K\}$. Assume $(\nu_a)_a$ belong to the family of single-parameter exponential distributions, with $\mu_a = \mathbb{E}_{r \sim \nu_a}[r]$.


μ_1


μ_2


μ_3


μ_4

μ_5

- **Sequential**: In round t the learner pulls arm $a_t \in [K]$ and receives the reward $r_t \sim \nu_{a_t}$.
- **Best Arm Identification objective**: quickly find the optimal arm $a^\star = \arg\max_a \mu_a$ with confidence $\delta \in (0, 1/2) \Rightarrow$ minimize sample complexity $\mathbb{E}[\tau]$ subject to $\mathbb{P}(\hat{a}_\tau \neq a^\star) \leq \delta$.
  - τ is a random stopping time and $\hat{a}_\tau$ is the estimated best arm at τ.

Let's introduce a graph structure

Best Arm Identification with Fixed Confidence

K arms with reward distributions ν_a with $a \in \{1, \ldots, K\}$. Assume $(\nu_a)_a$ belong to the family of single-parameter exponential distributions, with $\mu_a = \mathbb{E}_{r \sim \nu_a}[r]$.

μ_1 μ_2 μ_3 μ_4 μ_5

- **Sequential**: In round t the learner pulls arm $a_t \in [K]$ and receives the reward $r_t \sim \nu_{a_t}$.
- **Best Arm Identification objective**: quickly find the optimal arm $a^\star = \arg\max_a \mu_a$ with confidence $\delta \in (0, 1/2) \Rightarrow$ minimize sample complexity $\mathbb{E}[\tau]$ subject to $\mathbb{P}(\hat{a}_\tau \neq a^\star) \leq \delta$.
  - τ is a random stopping time and $\hat{a}_\tau$ is the estimated best arm at τ.

Let's introduce a graph structure

Best Arm Identification with Fixed Confidence

K arms with reward distributions ν_a with $a \in \{1, \dots, K\}$. Assume $(\nu_a)_a$ belong to the family of single-parameter exponential distributions, with $\mu_a = \mathbb{E}_{r \sim \nu_a}[r]$.

μ_1

μ_2

μ_3

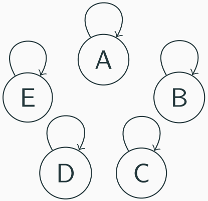
μ_4

μ_5

- **Sequential**: In round t the learner pulls arm $a_t \in [K]$ and receives the reward $r_t \sim \nu_{a_t}$.
- **Best Arm Identification objective**: quickly find the optimal arm $a^\star = \arg\max_a \mu_a$ with confidence $\delta \in (0, 1/2) \Rightarrow$ minimize sample complexity $\mathbb{E}[\tau]$ subject to $\mathbb{P}(\hat{a}_\tau \neq a^\star) \leq \delta$.
  - τ is a random stopping time and $\hat{a}_\tau$ is the estimated best arm at τ.

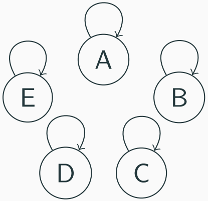
Let's introduce a graph structure

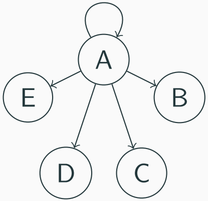
Graph Structure



1. Bandit model: when selecting an action, you observe the reward of that action.
2. Revealing action: when selecting A, you observe the reward of all other nodes.
3. Ring graph: you observe only the reward of two neighboring nodes.
4. Loopless clique: all connected.

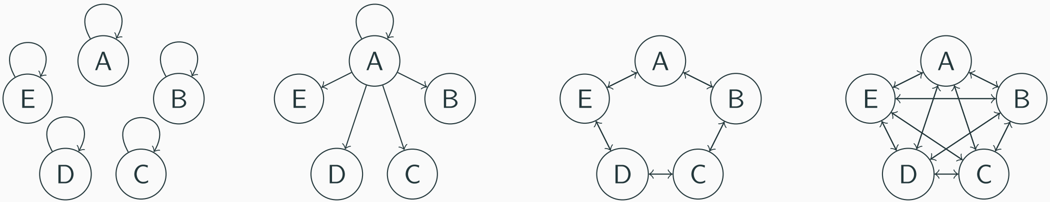
Graph Structure





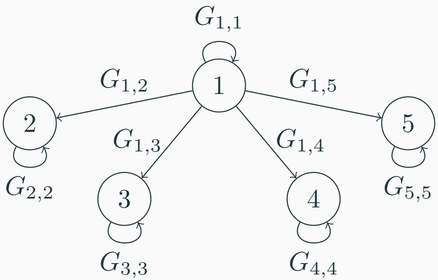
1. Bandit model: when selecting an action, you observe the reward of that action.
2. Revealing action: when selecting A, you observe the reward of all other nodes.
3. Ring graph: you observe only the reward of two neighboring nodes.
4. Loopless clique: all connected.

Graph Structure



1. Bandit model: when selecting an action, you observe the reward of that action.
2. Revealing action: when selecting A, you observe the reward of all other nodes.
3. Ring graph: you observe only the reward of two neighboring nodes.
4. Loopless clique: all connected.

BAI with a Graph

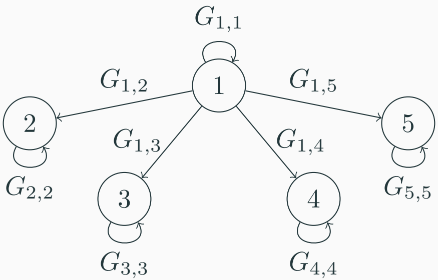


BAI Problem

Goal: Estimate $a^\star$ as quickly as possible subject to $\mathbb{P}(\hat{a}_\tau \neq a^\star) \leq \delta$.

- Graph characterized by the adjacency matrix $G \in [0,1]^{K \times K}$.
- When selecting a the agent observes $(Z_{a,u})_{u \in [K]}$, where $Z_{a,u} = Y_{a,u} R_u$ for all nodes u, with $Y_{a,u} \sim \mathrm{Ber}(G_{a,u})$ and $R_u \sim \nu_u$.
- What is the sample complexity lower bound? Can we use the graph to speed-up learning?

BAI with a Graph

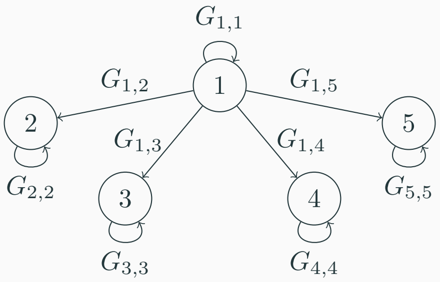


BAI Problem

Goal: Estimate $a^\star$ as quickly as possible subject to $\mathbb{P}(\hat{a}_\tau \neq a^\star) \leq \delta$.

- Graph characterized by the adjacency matrix $G \in [0,1]^{K \times K}$.
- When selecting a the agent observes $(Z_{a,u})_{u \in [K]}$, where $Z_{a,u} = Y_{a,u} R_u$ for all nodes u, with $Y_{a,u} \sim \mathrm{Ber}(G_{a,u})$ and $R_u \sim \nu_u$.
- What is the sample complexity lower bound? Can we use the graph to speed-up learning?

BAI with a Graph

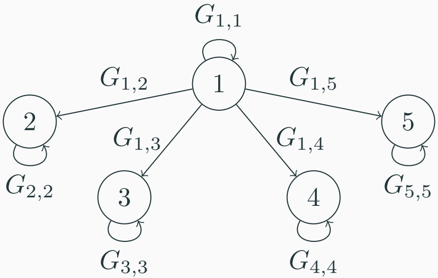


BAI Problem

Goal: Estimate $a^\star$ as quickly as possible subject to $\mathbb{P}(\hat{a}_\tau \neq a^\star) \leq \delta$.

- Graph characterized by the adjacency matrix $G \in [0,1]^{K\times K}$.
- When selecting a the agent observes $(Z_{a,u})_{u\in[K]}$, where $Z_{a,u} = Y_{a,u}R_u$ for all nodes u, with $Y_{a,u} \sim \mathrm{Ber}(G_{a,u})$ and $R_u \sim \nu_u$.
- What is the sample complexity lower bound? Can we use the graph to speed-up learning?

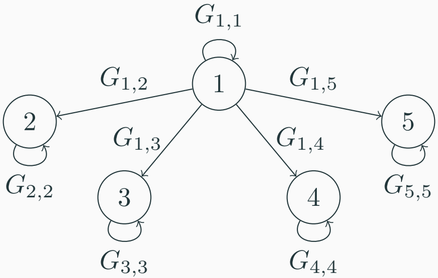
BAI with a Graph



However, the agent may/may not know the graph. **Two settings**:

- Uninformed setting: The learner does not know G nor which edge is activated at each time-step t.
- Informed setting: The learner either knows G or which edge was activated after choosing a node.

BAI with a Graph



However, the agent may/may not know the graph. **Two settings**:

- Uninformed setting: The learner does not know G nor which edge is activated at each time-step t.
- Informed setting: The learner either knows G or which edge was activated after choosing a node.

Sample Complexity Lower Bounds

Sample Complexity Lower Bounds - Uninformed Setting

Theorem

For any δ-PC algorithm and any model ν with reward distributions $\{\nu_u\}_{u\in V}$ with continuous support, in the **uninformed setting**[1] *we have that*

$$\mathbb{E}_\nu[\tau] \geq T^\star(\nu) \log \frac{1}{2.4\delta}, \tag{1}$$

$$\underbrace{(T^\star(\nu))^{-1}}_{\text{information rate}} = \underbrace{\sup_{\omega\in\Delta(V)}}_{\text{sampling policy}} \min_{u\neq a^\star} (m_u + m_{a^\star}) \underbrace{I_{\frac{m_{a^\star}}{m_u+m_{a^\star}}}(\nu_{a^\star}, \nu_u)}_{\text{Generalized Jensen-Shannon divergence}}$$

$$\text{s.t.} \quad \underbrace{m_u}_{\text{observation rate}} = \sum_{v\in N_{in}(u)} \omega_v G_{v,u} \quad \forall u \in V.$$

Concretely, for Gaussian rewards $\mathcal{N}(\mu_a, \lambda^2)$

$$T^\star(\nu) = \inf_{\omega\in\Delta(V)} \max_{u\neq a^\star} \left(m_u^{-1} + m_{a^\star}^{-1}\right) \frac{2\lambda^2}{\Delta_u^2} \text{ s.t. } m = G^\top \omega \quad (\text{where } \Delta_u = \mu_{a^\star} - \mu_u).$$

[1] Uninformed setting: The learner does not know G nor which edge is activated at each time-step t

Sample Complexity Lower Bounds - Uninformed Setting

Theorem

For any δ-PC algorithm and any model ν with reward distributions $\{\nu_u\}_{u\in V}$ with continuous support, in the **uninformed setting**[1] *we have that*

$$\mathbb{E}_\nu[\tau] \geq T^\star(\nu) \log \frac{1}{2.4\delta}, \tag{1}$$

$$\underbrace{(T^\star(\nu))^{-1}}_{\text{information rate}} = \underbrace{\sup_{\omega\in\Delta(V)}}_{\text{sampling policy}} \min_{u\neq a^\star} (m_u + m_{a^\star}) \underbrace{I_{\frac{m_{a^\star}}{m_u+m_{a^\star}}}(\nu_{a^\star}, \nu_u)}_{\text{Generalized Jensen-Shannon divergence}}$$

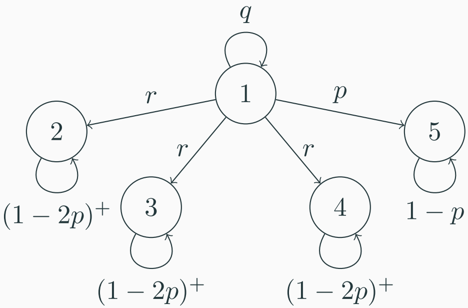
$$\text{s.t.} \quad \underbrace{m_u}_{\text{observation rate}} = \sum_{v\in N_{in}(u)} \omega_v G_{v,u} \quad \forall u \in V.$$

Concretely, for Gaussian rewards $\mathcal{N}(\mu_a, \lambda^2)$

$$T^\star(\nu) = \inf_{\omega\in\Delta(V)} \max_{u\neq a^\star} \left(m_u^{-1} + m_{a^\star}^{-1}\right) \frac{2\lambda^2}{\Delta_u^2} \text{ s.t. } m = G^\top \omega \quad (\text{where } \Delta_u = \mu_{a^\star} - \mu_u).$$

[1] Uninformed setting: The learner does not know G nor which edge is activated at each time-step t

An Example: The Loopy Star

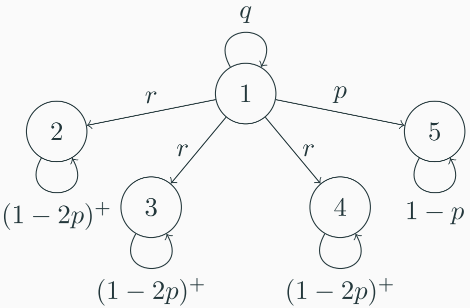


Loopy star graph. To each edge is associated an activation probability (obs. that $(x)^+ = \max(x, 0)$).

We consider Gaussian rewards, with $\lambda = 1$, $\mu_5 = 1$ and $\mu_u = 0.5, u \in \{1, \dots, 4\}$.

- This graph is the union of a bandit feedback graph and revealing action graph. Removing any self-loop changes the minimax regret from $\tilde{\Theta}(\sqrt{\alpha(G)T})$ to $\tilde{\Theta}(T^{2/3})$ [ACBDK15].

An Example: The Loopy Star

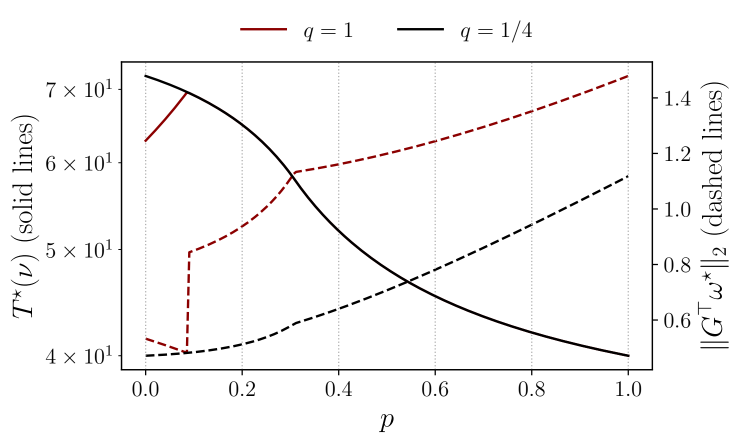


Loopy star graph. To each edge is associated an activation probability (obs. that $(x)^+ = \max(x, 0)$).

We consider Gaussian rewards, with $\lambda = 1$, $\mu_5 = 1$ and $\mu_u = 0.5, u \in \{1, \dots, 4\}$.

- This graph is the union of a bandit feedback graph and revealing action graph. Removing any self-loop changes the minimax regret from $\tilde{\Theta}(\sqrt{\alpha(G)T})$ to $\tilde{\Theta}(T^{2/3})$ [ACBDK15].

The Loopy Star



Loopy star example with $r = 1/4$. The solid lines depict $T^\star(\nu)$ for $q = 1$ and $q = 1/4$ for different values of p. Similarly, on the right axis, the dashed lines show $\|G^\top \omega^\star\|_2$, which indicates the amount of information gathered per time-step.

Sample Complexity Lower Bounds - Uninformed Setting [Proof 1/3]

Overall proof idea: take the log-likelihood ratio (LLR) of the observed data up to time τ between the true model ν and an alternative model ν' that admits a different optimal vertex.

- Selecting the model ν' that minimizes the LLR yields a lower bound on the sample complexity.

Step 1 (LLR): Consider two bandit models $\nu = \{G, (\nu_u)_u\}, \nu' = \{G', (\nu'_u)_u\}$. For each v, ν_v and ν'_v have, respectively, densities f_v and f'_v. $Z_{v,u} = Y_{v,u}R_u$ has density $f_{v,u}$ (sim. $f'_{v,u}$)

$$L_t = \ln \frac{d\mathbb{P}_\nu(V_1, Z_1, \ldots, V_t, Z_t)}{d\mathbb{P}_{\nu'}(V_1, Z_1, \ldots, V_t, Z_t)}, \qquad \big(V_t \text{ is the chosen vertex; } Z_t \text{ are the observed } \{Z_{v,u}\}_{v,u} \text{ at time } t\big)$$

$$= \sum_{u \in V} \sum_{v \in N_{in}(u)} \sum_{j=1}^{N_v(t)} \ln\left(\frac{f_{v,u}(W_{j,(v,u)})}{f'_{v,u}(W_{j,(v,u)})}\right), \qquad \big(W_{j,(v,u)} \text{ is the } j\text{-th obs. of } Z_{v,u}\big)$$

$$\Longrightarrow \mathbb{E}_\nu[L_t] = \sum_{u \in V} \sum_{v \in N_{in}(u)} \mathbb{E}_\nu[N_v(t)]\mathrm{KL}(\nu_{v,u}, \nu'_{v,u}).$$

Sample Complexity Lower Bounds - Uninformed Setting [Proof 1/3]

Overall proof idea: take the log-likelihood ratio (LLR) of the observed data up to time τ between the true model ν and an alternative model ν' that admits a different optimal vertex.

- Selecting the model ν' that minimizes the LLR yields a lower bound on the sample complexity.

Step 1 (LLR): Consider two bandit models $\nu = \{G, (\nu_u)_u\}, \nu' = \{G', (\nu'_u)_u\}$. For each v, ν_v and ν'_v have, respectively, densities f_v and f'_v. $Z_{v,u} = Y_{v,u} R_u$ has density $f_{v,u}$ (sim. $f'_{v,u}$)

$$L_t = \ln \frac{d\mathbb{P}_\nu(V_1, Z_1, \ldots, V_t, Z_t)}{d\mathbb{P}_{\nu'}(V_1, Z_1, \ldots, V_t, Z_t)}, \qquad (V_t \text{ is the chosen vertex; } Z_t \text{ are the observed } \{Z_{v,u}\}_{v,u} \text{ at time } t)$$

$$= \sum_{u \in V} \sum_{v \in N_{in}(u)} \sum_{j=1}^{N_v(t)} \ln \left(\frac{f_{v,u}(W_{j,(v,u)})}{f'_{v,u}(W_{j,(v,u)})} \right), \qquad (W_{j,(v,u)} \text{ is the } j\text{-th obs. of } Z_{v,u})$$

$$\implies \mathbb{E}_\nu[L_t] = \sum_{u \in V} \sum_{v \in N_{in}(u)} \mathbb{E}_\nu[N_v(t)] \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}).$$

Sample Complexity Lower Bounds - Uninformed Setting [Proof 1/3]

Overall proof idea: take the log-likelihood ratio (LLR) of the observed data up to time τ between the true model ν and an alternative model ν' that admits a different optimal vertex.

- Selecting the model ν' that minimizes the LLR yields a lower bound on the sample complexity.

Step 1 (LLR): Consider two bandit models $\nu = \{G, (\nu_u)_u\}, \nu' = \{G', (\nu'_u)_u\}$. For each v, ν_v and ν'_v have, respectively, densities f_v and f'_v. $Z_{v,u} = Y_{v,u} R_u$ has density $f_{v,u}$ (sim. $f'_{v,u}$)

$$L_t = \ln \frac{\mathrm{d}\mathbb{P}_\nu(V_1, Z_1, \ldots, V_t, Z_t)}{\mathrm{d}\mathbb{P}_{\nu'}(V_1, Z_1, \ldots, V_t, Z_t)}, \qquad \big(V_t \text{ is the chosen vertex; } Z_t \text{ are the observed } \{Z_{v,u}\}_{v,u} \text{ at time } t\big)$$

$$= \sum_{u \in V} \sum_{v \in N_{in}(u)} \sum_{j=1}^{N_v(t)} \ln \left(\frac{f_{v,u}(W_{j,(v,u)})}{f'_{v,u}(W_{j,(v,u)}))} \right), \qquad \big(W_{j,(v,u)} \text{ is the } j\text{-th obs. of } Z_{v,u}\big)$$

$$\Longrightarrow \mathbb{E}_\nu[L_t] = \sum_{u \in V} \sum_{v \in N_{in}(u)} \mathbb{E}_\nu[N_v(t)] \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}).$$

Sample Complexity Lower Bounds - Uninformed Setting [Proof 2/3]

Using an information processing inequality[KCG16], we can lower bound the expected LLR at τ as

$$\mathbb{E}_\nu[L_\tau] \geq \log(1/(2.4\delta)),$$

and by letting $\omega_v = \mathbb{E}_\nu[N_v(\tau)]/\mathbb{E}_\nu[\tau]$, we obtain

$$\underbrace{\mathbb{E}_\nu[\tau]}_{\text{sample complexity}} \sum_{u\in V} \sum_{v \in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) \geq \log(1/(2.4\delta)).$$

Lastly, because of the nature of the problem, we can prove that for continuous rewards we have[2]

$$\mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) = \underbrace{\mathrm{kl}(G_{v,u}, G'_{v,u})}_{\text{Bernoulli KL divergence}} + G_{v,u} \underbrace{\mathrm{KL}(\nu_u, \nu'_u)}_{\text{KL divergence of the rewards}}.$$

[2] Recall that $G_{v,u}$ is the edge activation probability (probability of observing u when selecting v).

Sample Complexity Lower Bounds - Uninformed Setting [Proof 2/3]

Using an information processing inequality[KCG16], we can lower bound the expected LLR at τ as

$$\mathbb{E}_\nu[L_\tau] \geq \log(1/(2.4\delta)),$$

and by letting $\omega_v = \mathbb{E}_\nu[N_v(\tau)]/\mathbb{E}_\nu[\tau]$, we obtain

$$\underbrace{\mathbb{E}_\nu[\tau]}_{\text{sample complexity}} \sum_{u\in V} \sum_{v\in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) \geq \log(1/(2.4\delta)).$$

Lastly, because of the nature of the problem, we can prove that for continuous rewards we have[2]

$$\mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) = \underbrace{\mathrm{kl}(G_{v,u}, G'_{v,u})}_{\text{Bernoulli KL divergence}} + G_{v,u} \underbrace{\mathrm{KL}(\nu_u, \nu'_u)}_{\text{KL divergence of the rewards}}.$$

[2] Recall that $G_{v,u}$ is the edge activation probability (probability of observing u when selecting v).

Sample Complexity Lower Bounds - Uninformed Setting [Proof 2/3]

Using an information processing inequality[KCG16], we can lower bound the expected LLR at τ as

$$\mathbb{E}_\nu[L_\tau] \geq \log(1/(2.4\delta)),$$

and by letting $\omega_v = \mathbb{E}_\nu[N_v(\tau)]/\mathbb{E}_\nu[\tau]$, we obtain

$$\underbrace{\mathbb{E}_\nu[\tau]}_{\text{sample complexity}} \sum_{u \in V} \sum_{v \in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) \geq \log(1/(2.4\delta)).$$

Lastly, because of the nature of the problem, we can prove that for continuous rewards we have[2]

$$\mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) = \underbrace{\mathrm{kl}(G_{v,u}, G'_{v,u})}_{\text{Bernoulli KL divergence}} + G_{v,u} \underbrace{\mathrm{KL}(\nu_u, \nu'_u)}_{\text{KL divergence of the rewards}}.$$

[2] Recall that $G_{v,u}$ is the edge activation probability (probability of observing u when selecting v).

Sample Complexity Lower Bounds - Uninformed Setting [Proof 3/3]

Step 2 (Optimizing ν'): We focus on alternative models ν' that admit a different optimal vertex

$$\mathrm{Alt}(\nu) = \cup_{v \neq a^\star} \mathrm{Alt}_v(\nu), \quad \mathrm{Alt}_v(\nu) = \{\nu' \mid \mu'_v > \mu'_{a^\star}\}.$$

Choose ν' as to minimize the LLR!

$$\begin{aligned}
&\inf_{\nu' \in \mathrm{Alt}(\nu)} \sum_{u \in V} \sum_{v \in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) \\
&= \min_{u \neq a^\star} \inf_{\nu' : \mu'_u \geq \mu'_{a^\star}} \sum_{v \in N_{in}(u)} \omega_v G_{v,u} \mathrm{KL}(\nu_u, \nu'_u) + \sum_{w \in N_{in}(a^\star)} \omega_w G_{w,a^\star} \mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}), \\
&= \min_{u \neq a^\star} \inf_{\nu' : \mu'_u \geq \mu'_{a^\star}} m_u \mathrm{KL}(\nu_u, \nu'_u) + m_{a^\star} \mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}). \qquad (m_u := \textstyle\sum_{v \in N_{in}(u)} \omega_v G_{v,u})
\end{aligned}$$

Therefore, by optimizing over ν' as in [GK16, Lemma 3] we obtain

$$\min_{u \neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u + m_{a^\star}}}(\nu_{a^\star}, \nu_u) \geq \log(1/(2.4\delta)).$$

Sample Complexity Lower Bounds - Uninformed Setting [Proof 3/3]

Step 2 (Optimizing ν'): We focus on alternative models ν' that admit a different optimal vertex

$$\mathrm{Alt}(\nu) = \cup_{v\neq a^\star}\mathrm{Alt}_v(\nu), \quad \mathrm{Alt}_v(\nu) = \{\nu' \mid \mu'_v > \mu'_{a^\star}\}.$$

Choose ν' as to minimize the LLR!

$$\begin{aligned}
&\inf_{\nu'\in\mathrm{Alt}(\nu)} \sum_{u\in V}\sum_{v\in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) \\
&= \min_{u\neq a^\star} \inf_{\nu':\mu'_u\geq\mu'_{a^\star}} \sum_{v\in N_{in}(u)} \omega_v G_{v,u}\mathrm{KL}(\nu_u,\nu'_u) + \sum_{w\in N_{in}(a^\star)} \omega_w G_{w,a^\star}\mathrm{KL}(\nu_{a^\star},\nu'_{a^\star}), \\
&= \min_{u\neq a^\star} \inf_{\nu':\mu'_u\geq\mu'_{a^\star}} m_u \mathrm{KL}(\nu_u,\nu'_u) + m_{a^\star}\mathrm{KL}(\nu_{a^\star},\nu'_{a^\star}). \qquad (m_u := \textstyle\sum_{v\in N_{in}(u)} \omega_v G_{v,u})
\end{aligned}$$

Therefore, by optimizing over ν' as in [GK16, Lemma 3] we obtain

$$\min_{u\neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u+m_{a^\star}}}(\nu_{a^\star},\nu_u) \geq \log(1/(2.4\delta)).$$

Sample Complexity Lower Bounds - Uninformed Setting [Proof 3/3]

Step 2 (Optimizing ν'): We focus on alternative models ν' that admit a different optimal vertex

$$\mathrm{Alt}(\nu) = \cup_{v \neq a^\star} \mathrm{Alt}_v(\nu), \quad \mathrm{Alt}_v(\nu) = \{\nu' \mid \mu'_v > \mu'_{a^\star}\}.$$

Choose ν' as to minimize the LLR!

$$\begin{aligned}
&\inf_{\nu' \in \mathrm{Alt}(\nu)} \sum_{u \in V} \sum_{v \in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) \\
&= \min_{u \neq a^\star} \inf_{\nu' : \mu'_u \geq \mu'_{a^\star}} \sum_{v \in N_{in}(u)} \omega_v G_{v,u} \mathrm{KL}(\nu_u, \nu'_u) + \sum_{w \in N_{in}(a^\star)} \omega_w G_{w,a^\star} \mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}), \\
&= \min_{u \neq a^\star} \inf_{\nu' : \mu'_u \geq \mu'_{a^\star}} m_u \mathrm{KL}(\nu_u, \nu'_u) + m_{a^\star} \mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}). \qquad (m_u := \textstyle\sum_{v \in N_{in}(u)} \omega_v G_{v,u})
\end{aligned}$$

Therefore, by optimizing over ν' as in [GK16, Lemma 3] we obtain

$$\min_{u \neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u + m_{a^\star}}}(\nu_{a^\star}, \nu_u) \geq \log(1/(2.4\delta)).$$

Sample Complexity Lower Bounds - Uninformed Setting [Proof 3/3]

Step 2 (Optimizing ν'): We focus on alternative models ν' that admit a different optimal vertex

$$\mathrm{Alt}(\nu) = \cup_{v \neq a^\star} \mathrm{Alt}_v(\nu), \quad \mathrm{Alt}_v(\nu) = \{\nu' \mid \mu'_v > \mu'_{a^\star}\}.$$

Choose ν' as to minimize the LLR!

$$\begin{aligned}
&\inf_{\nu' \in \mathrm{Alt}(\nu)} \sum_{u \in V} \sum_{v \in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) \\
&= \min_{u \neq a^\star} \inf_{\nu' : \mu'_u \geq \mu'_{a^\star}} \sum_{v \in N_{in}(u)} \omega_v G_{v,u} \mathrm{KL}(\nu_u, \nu'_u) + \sum_{w \in N_{in}(a^\star)} \omega_w G_{w,a^\star} \mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}), \\
&= \min_{u \neq a^\star} \inf_{\nu' : \mu'_u \geq \mu'_{a^\star}} m_u \mathrm{KL}(\nu_u, \nu'_u) + m_{a^\star} \mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}). \qquad (m_u := \textstyle\sum_{v \in N_{in}(u)} \omega_v G_{v,u})
\end{aligned}$$

Therefore, by optimizing over ν' as in [GK16, Lemma 3] we obtain

$$\min_{u \neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u + m_{a^\star}}}(\nu_{a^\star}, \nu_u) \geq \log(1/(2.4\delta)).$$

Sample Complexity Lower Bounds - Uninformed Setting [Proof 3/3]

Step 2 (Optimizing ν'): We focus on alternative models ν' that admit a different optimal vertex

$$\mathrm{Alt}(\nu) = \cup_{v \neq a^\star} \mathrm{Alt}_v(\nu), \quad \mathrm{Alt}_v(\nu) = \{\nu' \mid \mu'_v > \mu'_{a^\star}\}.$$

Choose ν' as to minimize the LLR!

$$\begin{aligned}
&\inf_{\nu' \in \mathrm{Alt}(\nu)} \sum_{u \in V} \sum_{v \in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) \\
&= \min_{u \neq a^\star} \inf_{\nu' : \mu'_u \geq \mu'_{a^\star}} \sum_{v \in N_{in}(u)} \omega_v G_{v,u} \mathrm{KL}(\nu_u, \nu'_u) + \sum_{w \in N_{in}(a^\star)} \omega_w G_{w,a^\star} \mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}), \\
&= \min_{u \neq a^\star} \inf_{\nu' : \mu'_u \geq \mu'_{a^\star}} m_u \mathrm{KL}(\nu_u, \nu'_u) + m_{a^\star} \mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}). \qquad (m_u := \textstyle\sum_{v \in N_{in}(u)} \omega_v G_{v,u})
\end{aligned}$$

Therefore, by optimizing over ν' as in [GK16, Lemma 3] we obtain

$$\min_{u \neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u + m_{a^\star}}}(\nu_{a^\star}, \nu_u) \geq \log(1/(2.4\delta)).$$

Sample Complexity Lower Bounds - Uninformed Setting - Bernoulli

- For Bernoulli rewards something funny happens in the **uninformed setting**[3]...
- Observing $Z_{a,u} = 0$ can mean either "edge did not fire" or "reward was 0,"

$$P(Z = 0) = 1 - G_{v,u}\, \mu_u.$$

Because the learner never sees which edge fired, it is possible to construct an alternative model resemblingly perfectly the true model, under which an alternative arm is optimal!

Proposition

If $(\nu_u)_{u\in V}$ are Bernoulli distributions with parameters $(\mu_u)_{u\in V}$, then $a^\star$ is unidentifiable, in the sense that $(T^\star(\nu))^{-1} = 0$.

[3] Uninformed setting: The learner does not know G nor which edge is activated at each time-step t.

Sample Complexity Lower Bounds - Uninformed Setting - Bernoulli

- For Bernoulli rewards something funny happens in the **uninformed setting**[3]...
- Observing $Z_{a,u} = 0$ can mean either "edge did not fire" or "reward was 0,"

$$P(Z = 0) = 1 - G_{v,u}\,\mu_u.$$

Because the learner never sees which edge fired, it is possible to construct an alternative model resemblingly perfectly the true model, under which an alternative arm is optimal!

Proposition

If $(\nu_u)_{u \in V}$ are Bernoulli distributions with parameters $(\mu_u)_{u \in V}$, then $a^\star$ is unidentifiable, in the sense that $(T^\star(\nu))^{-1} = 0$.

[3] Uninformed setting: The learner does not know G nor which edge is activated at each time-step t.

Sample Complexity Lower Bounds - Uninformed Setting - Bernoulli

- For Bernoulli rewards something funny happens in the **uninformed setting**[3]...
- Observing $Z_{a,u} = 0$ can mean either "edge did not fire" or "reward was 0,"

$$P(Z = 0) = 1 - G_{v,u}\, \mu_u .$$

Because the learner never sees which edge fired, it is possible to construct an alternative model resemblingly perfectly the true model, under which an alternative arm is optimal!

Proposition

If $(\nu_u)_{u \in V}$ are Bernoulli distributions with parameters $(\mu_u)_{u \in V}$, then $a^\star$ is unidentifiable, in the sense that $(T^\star(\nu))^{-1} = 0$.

[3] Uninformed setting: The learner does not know G nor which edge is activated at each time-step t.

Sample Complexity Lower Bounds - Informed Setting - Bernoulli

What about the informed setting[4]?

- All good here, and the original sample complexity holds also for Bernoulli rewards.

$$\mathbb{E}_\nu[\tau] \geq T^\star(\nu) \log \frac{1}{2.4\delta} \tag{2}$$

where

$$(T^\star(\nu))^{-1} = \sup_{\omega \in \Delta(V)} \min_{u \neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u + m_{a^\star}}}(\nu_{a^\star}, \nu_u) \text{ s.t. } m = G^\top \omega$$

[4] Informed setting: The learner either knows G or which edge was activated after choosing a node.

Sample Complexity Lower Bounds - Informed Setting - Bernoulli

What about the informed setting[4]?

- All good here, and the original sample complexity holds also for Bernoulli rewards.

$$\mathbb{E}_\nu[\tau] \geq T^\star(\nu) \log \frac{1}{2.4\delta} \tag{2}$$

where

$$(T^\star(\nu))^{-1} = \sup_{\omega \in \Delta(V)} \min_{u \neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u + m_{a^\star}}}(\nu_{a^\star}, \nu_u) \text{ s.t. } m = G^\top \omega$$

[4] Informed setting: The learner either knows G or which edge was activated after choosing a node.

TaS-FG: Track And Stop for Feedback Graphs

Components of a Strategy

A strategy is defined by

- Sampling rule
- Stopping rule
- Recommendation rule (we use the MLE)

TaS-FG: Sampling Rule

How do we design an algorithm that approaches the optimal sample complexity?

$$T(\omega;\nu)^{-1} = \min_{u\neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u+m_{a^\star}}}(\nu_{a^\star}, \nu_u) \text{ s.t. } m = G^\top \omega$$

The solution $\omega^\star \in \arg\inf_{\omega\in\Delta(V)} T(\omega;\nu)$ provides the best proportion of draws.

Design

- Ensure that N_t/t (average selection frequency) tracks $\omega^\star(t)$ (computed w.r.t. $\hat{\nu}(t)$, the estimated model), where N_t is the visitation vector $N(t) := \begin{bmatrix} N_1(t) & \dots & N_K(t) \end{bmatrix}^\top$.
- Sampling rule:

$$A_t \in \begin{cases} \arg\min_{u\in S_t} N_u(t) & \exists u : N_u(t) < \sqrt{t} - K/2 \\ \arg\min_{u\in V} N_u(t) - \sum_{n=1}^t \omega_u^\star(n) & \text{otherwise} \end{cases}, \tag{3}$$

ensures $\lim_{t\to\infty} \inf_{\omega\in C^\star(\nu)} \|N(t)/t - \omega\|_\infty \to 0$ ($C^\star$ is the set of optimal allocations)[5].

[5] Tracking a convex combination of all past solutions guarantees convergence to a unique point in $C^\star$.

TaS-FG: Sampling Rule

How do we design an algorithm that approaches the optimal sample complexity?

$$T(\omega;\nu)^{-1} = \min_{u \neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u + m_{a^\star}}}(\nu_{a^\star}, \nu_u) \text{ s.t. } m = G^\top \omega$$

The solution $\omega^\star \in \arg\inf_{\omega \in \Delta(V)} T(\omega;\nu)$ provides the best proportion of draws.

Design

- Ensure that N_t/t (average selection frequency) tracks $\omega^\star(t)$ (computed w.r.t. $\hat{\nu}(t)$, the estimated model), where N_t is the visitation vector $N(t) := \begin{bmatrix} N_1(t) & \dots & N_K(t) \end{bmatrix}^\top$.
- Sampling rule:

$$A_t \in \begin{cases} \arg\min_{u \in S_t} N_u(t) & \exists u : N_u(t) < \sqrt{t} - K/2 \\ \arg\min_{u \in V} N_u(t) - \sum_{n=1}^{t} \omega_u^\star(n) & \text{otherwise} \end{cases}, \quad (3)$$

ensures $\lim_{t\to\infty} \inf_{\omega \in C^\star(\nu)} \|N(t)/t - \omega\|_\infty \to 0$ ($C^\star$ is the set of optimal allocations)[5].

[5] Tracking a convex combination of all past solutions guarantees convergence to a unique point in $C^\star$.

TaS-FG: Sampling Rule

How do we design an algorithm that approaches the optimal sample complexity?

$$T(\omega;\nu)^{-1} = \min_{u \neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u + m_{a^\star}}}(\nu_{a^\star}, \nu_u) \text{ s.t. } m = G^\top \omega$$

The solution $\omega^\star \in \arg\inf_{\omega \in \Delta(V)} T(\omega;\nu)$ provides the best proportion of draws.

Design

- Ensure that N_t/t (average selection frequency) tracks $\omega^\star(t)$ (computed w.r.t. $\hat{\nu}(t)$, the estimated model), where N_t is the visitation vector $N(t) := \begin{bmatrix} N_1(t) & \dots & N_K(t) \end{bmatrix}^\top$.
- Sampling rule:

$$A_t \in \begin{cases} \arg\min_{u \in S_t} N_u(t) & \exists u : N_u(t) < \sqrt{t} - K/2 \\ \arg\min_{u \in V} N_u(t) - \sum_{n=1}^{t} \omega_u^\star(n) & \text{otherwise} \end{cases}, \qquad (3)$$

ensures $\lim_{t\to\infty} \inf_{\omega \in C^\star(\nu)} \|N(t)/t - \omega\|_\infty \to 0$ ($C^\star$ is the set of optimal allocations)[5].

[5] Tracking a convex combination of all past solutions guarantees convergence to a unique point in $C^\star$.

TaS-FG: Stopping Rule

When do we stop?

$$T(\omega;\nu)^{-1} = \min_{u \neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u + m_{a^\star}}}(\nu_{a^\star}, \nu_u) \text{ s.t. } m = G^\top \omega$$

Stopping Rule

- The lower bound tells us that $\tau \sim T^\star(\nu)\log(1/\delta)$. But we don't know the model! Additional price to pay $O(\log\log(t))$.
- Stopping as soon as

$$t \approx T(N(t)/t; \hat{\nu}(t))^6 \left[\log\left(\frac{K-1}{\delta}\right) + O(\log\log(t))\right]$$

guarantees correctness

$$\mathbb{P}_\nu(\tau < \infty, \hat{a}_\tau \neq a^\star(\mu)) \leq \delta.$$

- With the previous sampling rule, we can guarantee $\limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau]}{\ln(1/\delta)} \leq T^\star(\nu)$.

[6] Empirical characteristic time based on the MLE

TaS-FG: Stopping Rule

When do we stop?

$$T(\omega;\nu)^{-1} = \min_{u \neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u + m_{a^\star}}}(\nu_{a^\star}, \nu_u) \text{ s.t. } m = G^\top \omega$$

Stopping Rule

- The lower bound tells us that $\tau \sim T^\star(\nu)\log(1/\delta)$. But we don't know the model! Additional price to pay $O(\log\log(t))$.
- Stopping as soon as

$$t \approx T(N(t)/t; \hat{\nu}(t))^{6} \left[\log\left(\frac{K-1}{\delta}\right) + O(\log\log(t))\right]$$

guarantees correctness

$$\mathbb{P}_\nu(\tau < \infty, \hat{a}_\tau \neq a^\star(\mu)) \leq \delta.$$

- With the previous sampling rule, we can guarantee $\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\ln(1/\delta)} \leq T^\star(\nu)$.

[6] Empirical characteristic time based on the MLE

TaS-FG: Stopping Rule

When do we stop?

$$T(\omega; \nu)^{-1} = \min_{u \neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u + m_{a^\star}}}(\nu_{a^\star}, \nu_u) \text{ s.t. } m = G^\top \omega$$

Stopping Rule

- The lower bound tells us that $\tau \sim T^\star(\nu) \log(1/\delta)$. But we don't know the model! Additional price to pay $O(\log\log(t))$.
- Stopping as soon as

$$t \approx T(N(t)/t; \hat{\nu}(t))^6 \left[\log\left(\frac{K-1}{\delta}\right) + O(\log\log(t))\right]$$

guarantees correctness

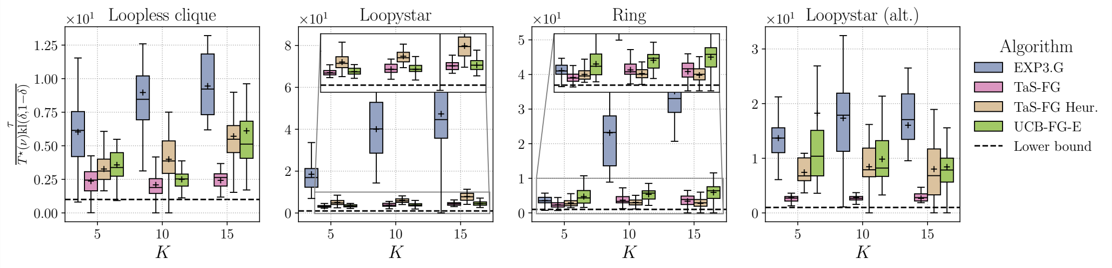
$$\mathbb{P}_\nu(\tau < \infty, \hat{a}_\tau \neq a^\star(\mu)) \leq \delta.$$

- With the previous sampling rule, we can guarantee $\limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau]}{\ln(1/\delta)} \leq T^\star(\nu)$.

[6] Empirical characteristic time based on the MLE

Numerical Results

Numerical Results



Box plots of the normalized sample complexity $\frac{\tau}{T^\star(\nu)\mathrm{kl}(\delta,1-\delta)}$ for $\delta = e^{-7}$ over 100 seeds. Boxes indicate the interquartile range, while the median and mean values are, respectively, the solid line and the $+$ sign in black.

Conclusions

Thank you for listening!

- Github repo:
  `https://github.com/rssalessio/Pure-Exploration-with-Feedback-Graphs`

References i

- Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren, *Online learning with feedback graphs: Beyond bandits*, Conference on Learning Theory, PMLR, 2015, pp. 23–35.
- Aurélien Garivier and Emilie Kaufmann, *Optimal best arm identification with fixed confidence*, Conference on Learning Theory, PMLR, 2016, pp. 998–1027.
- Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier, *On the complexity of best-arm identification in multi-armed bandit models*, The Journal of Machine Learning Research **17** (2016), no. 1, 1–42.