Online Learning with Feedback Graphs

Claudio Gentile
INRIA and Google NY
cla.gentile@gmail.com

NYC
March 6th, 2018
Content of this lecture

Regret analysis of sequential prediction problems lying between full and bandit information regimes:

- Motivation

- Nonstochastic setting:
  - Brief review of background
  - Feedback graphs

- Stochastic setting:
  - Brief review of background
  - Feedback graphs

- Examples (nonstochastic)
Motivation

Sequential prediction problems with partial information where items in action space have semantic connections turning into observability dependencies of associated losses/gains
Background/1: Nonstochastic experts

\( K \) actions for Learner

For \( t = 1, 2, \ldots \) :

1. Losses \( \ell_t(i) \in [0,1] \) are assigned deterministically by Nature to every action \( i = 1 \ldots K \) (hidded to Learner)

2. Learner picks action \( I_t \) (possibly using randomization) and incurs loss \( \ell_t(I_t) \)

3. Learner gets feedback information: \( \ell_t(1), \ldots, \ell_t(I_t), \ldots, \ell_t(K) \)
Background/1: Nonstochastic experts

\( K \) actions for Learner

For \( t = 1, 2, \ldots \):

1. Losses \( \ell_t(i) \in [0, 1] \) are assigned deterministically by opponent to every action \( i = 1 \ldots K \) (hidded to Learner)

2. Learner picks action \( I_t \) (possibly using randomization) and incurs loss \( \ell_t(I_t) \)

3. Learner gets feedback information: \( \ell_t(1), \ldots, \ell_t(I_t), \ldots, \ell_t(K) \)
Background: Nonstochastic experts

For $t = 1, 2, \ldots$:

1. Losses $\ell_t(i) \in [0, 1]$ are assigned deterministically by opponent to every action $i = 1 \ldots K$ (hidded to Learner)

2. Learner picks action $I_t$ (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Learner gets feedback information: $\ell_t(1), \ldots, \ell_t(I_t), \ldots, \ell_t(K)$
No (External, Pseudo) Regret

Goal: Given $T$ rounds, Learner’s total loss

$$\sum_{t=1}^{T} \ell_t(I_t)$$

must be close to that of single best action in hindsight for Learner

Regret of Learner for $T$ rounds:

$$R_T = \mathbb{E} \left[ \sum_{t=1}^{T} \ell_t(I_t) \right] - \min_{i=1\ldots K} \sum_{t=1}^{T} \ell_t(i)$$

Want: $R_T = o(T)$ as $T$ grows large ("no regret")

Notice: No stochastic assumptions on losses, but assume for simplicity Nature is deterministic and oblivious

Lower bound:

$$R_T \geq (1 - o(1)) \sqrt{\frac{T \ln K}{2}}$$

[CB+97]

as $T, K \to \infty$

($\ell_t(i)$ random coin flips + simple probabilistic argument)
Exponentially-weighted Algorithm [CB+97]

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

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

- if $\eta = \sqrt{\frac{\ln K}{8T}}$ \quad \implies \quad R_T \leq \sqrt{\frac{T \ln K}{2}}$
- Dynamic $\eta = \sqrt{\frac{\ln K}{t}}$ \quad \implies \quad $R_T$ looses constant factors
Nonstochastic bandit problem/1

$K$ actions for Learner

\[ \cdots \cdot \cdots \cdot \cdots \cdot \cdots \cdot \cdots \cdot \cdots \cdot \circ \]

For $t = 1, 2, \ldots$:

1. Losses $\ell_t(i) \in [0, 1]$ are assigned deterministically by Nature to every action $i = 1 \ldots K$ (hidded to Learner)

2. Learner picks action $I_t$ (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Learner gets feedback information: $\ell_t(I_t)$
Nonstochastic bandit problem

For $t = 1, 2, \ldots$:

1. Losses $\ell_t(i) \in [0, 1]$ are assigned deterministically by Nature to every action $i = 1 \ldots K$ (hidden to Learner)

2. Learner picks action $I_t$ (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Learner gets feedback information: $\ell_t(I_t)$
Nonstochastic bandit problem/1

For $t = 1, 2, \ldots$:

1. Losses $\ell_t(i) \in [0, 1]$ are assigned deterministically by Nature to every action $i = 1 \ldots K$ (hidded to Learner)

2. Learner picks action $I_t$ (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Learner gets feedback information: $\ell_t(I_t)$
Nonstochastic bandit problem/2

Goal: same as before

Regret of Learner for $T$ rounds:

$$R_T = \mathbb{E} \left[ \sum_{t=1}^{T} \ell_t(I_t) \right] - \min_{i=1\ldots K} \sum_{t=1}^{T} \ell_t(i)$$

Want: $R_T = o(T)$ as $T$ grows large ("no regret")

Tradeoff exploration vs. exploitation
Nonstochastic bandit problem/3: Exp3 Alg./1 [Auer+ 02]

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

$$\exp \left( -\eta \sum_{s=1}^{t-1} \hat{\ell}_s(i) \right), \quad i = 1 \ldots K$$

$$\hat{\ell}_s(i) = \begin{cases} \frac{\ell_s(i)}{\Pr_s(\ell_s(i) \text{ is observed in round } s)} & \text{if } \ell_s(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

- Only one nonzero component in $\hat{\ell}_t$

- Exponentially-weighted alg with (importance sampling) loss estimates

$$\hat{\ell}_t(i) \approx \ell_t(i)$$
Nonstochastic bandit problem/3: Exp3 Alg./2 [Auer+ 02]

Properties of loss estimates:

- $\mathbb{E}_t[\hat{\ell}_t(i)] = \ell_t(i)$ unbiasedness
- $\mathbb{E}_t[\hat{\ell}_t(i)^2] \leq \frac{1}{\Pr_t(\ell_t(i) \text{ is observed in round } t)}$ variance control

Regret analysis:

- Set $p_t(i) = \Pr_t(I_t = i)$
- Approximate $\exp(x)$ up to 2nd order, sum over rounds $t$ and overapprox.:

$$
\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i) \hat{\ell}_t(i) - \min_{i=1,\ldots,K} \sum_{t=1}^{T} \hat{\ell}_t(i) \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i) \hat{\ell}_t(i)^2
$$

- Take expectations (tower rule), and optimize over $\eta$:

$$
R_T \leq \frac{\ln K}{\eta} + \frac{\eta}{2} TK = \sqrt{2TK \ln K}
$$

- Lower bound $\Omega(\sqrt{TK})$ (improved upper bound by the INF alg. [AB09])
Contrasting expert to nonstochastic bandit problem

Experts:

- Learner observes all losses $\ell_t(1), \ldots, \ell_t(K)$
- $\Pr_t(\ell_t(i) \text{ is observed in round } t) = 1$
- Regret $R_T = O(\sqrt{T \ln K})$

Nonstochastic bandits:

- Learner only observes loss $\ell_t(I_t)$ of chosen action
- $\Pr_t(\ell_t(i) \text{ is observed in round } t) = \Pr_t(I_t = i)$
  - Note: Exp3 collapses to Exponentially-weighted alg.
- Regret $R_T = O(\sqrt{TK})$

Exponential gap $\ln K$ vs. $K$: relevant when actions are many
Nonstochastic bandits with Feedback Graphs

*K* actions for Learner

For *t* = 1, 2, ... :

1. Losses \( \ell_t(i) \in [0, 1] \) are assigned deterministically by Nature to every action \( i = 1 \ldots K \) (hidden to Learner)

2. Feedback graph \( G_t = (V, E_t) \), \( V = \{1, \ldots, K\} \) generated by exogenous process (hidden to Learner) – all self-loops included

3. Learner picks action \( I_t \) (possibly using randomization) and incurs loss \( \ell_t(I_t) \)

4. Learner gets feedback information: \( \{\ell_t(j) : (I_t, j) \in E_t\} + G_t \)
Nonstochastic bandits with Feedback Graphs/1 [MS11,A+13,K+14]

*K* actions for Learner

For *t* = 1, 2, ...:

1. Losses \( \ell_t(i) \in [0, 1] \) are assigned deterministically by Nature to every action \( i = 1, \ldots, K \) (hidden to Learner)

2. Feedback graph \( G_t = (V, E_t) \), \( V = \{1, \ldots, K\} \)
generated by exogenous process (hidden to Learner) – all self-loops included

3. Learner picks action \( I_t \) (possibly using randomization) and incurs loss \( \ell_t(I_t) \)

4. Learner gets feedback information: \( \{\ell_t(j) : (I_t, j) \in E_t\} + G_t \)
Nonstochastic bandits with Feedback Graphs

$K$ actions for Learner

For $t = 1, 2, \ldots$ :

1. Losses $\ell_t(i) \in [0, 1]$ are assigned deterministically by Nature to every action $i = 1 \ldots K$ (hidden to Learner)

2. Feedback graph $G_t = (V, E_t)$, $V = \{1, \ldots, K\}$ generated by exogenous process (hidden to Learner) – all self-loops included

3. Learner picks action $I_t$ (possibly using randomization) and incurs loss $\ell_t(I_t)$

4. Learner gets feedback information: $\{\ell_t(j) : (I_t, j) \in E_t\} + G_t$
Nonstochastic bandits with Feedback Graphs/1 [MS11, A+13, K+14]

K actions for Learner

For t = 1, 2, ... :

1. Losses \( \ell_t(i) \in [0, 1] \) are assigned deterministically by Nature to every action \( i = 1 \ldots K \) (hidden to Learner)

2. Feedback graph \( G_t = (V, E_t) \), \( V = \{1, \ldots, K\} \) generated by exogenous process (hidden to Learner) – all self-loops included

3. Learner picks action \( I_t \) (possibly using randomization) and incurs loss \( \ell_t(I_t) \)

4. Learner gets feedback information: \( \{\ell_t(j) : (I_t, j) \in E_t\} \ + \ G_t \)
Nonstochastic bandits with Feedback Graphs/2: Exp3-IX Alg.  

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

$$\exp \left( -\eta \sum_{s=1}^{t-1} \hat{\ell}_s(i) \right), \quad i = 1 \ldots K$$

$$\hat{\ell}_s(i) = \begin{cases} \frac{\ell_s(i)}{\gamma_t + \Pr_s(\ell_s(i) \text{ is observed in round } s)} & \text{if } \ell_s(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

- **Note:** prob. of observing loss of action $\neq$ prob. of playing action

- Exponentially-weighted alg with $\gamma_t$-biased (importance sampling) loss estimates

$$\hat{\ell}_t(i) \approx \ell_t(i)$$

- Bias is controlled by $\gamma_t = 1/\sqrt{t}$
Nonstochastic bandits with Feedback Graphs \cite{A+13,K+14}

Independence number $\alpha(G_t)$: disregard edge orientation

\[
\begin{align*}
\text{clique: expert problem} & \quad \quad \leq \quad \alpha(G_t) \quad \leq \quad \text{edgeless: bandit problem}
\end{align*}
\]

Regret analysis:

- If $G_t = G \quad \forall t$:
  \[
  R_T = \tilde{O}\left(\sqrt{T\alpha(G)}\right)
  \]
  (also lower bound up to logs)

- In general:
  \[
  R_T = O\left(\ln(TK)\sqrt{\sum_{t=1}^{T} \alpha(G_t)}\right)
  \]
Nonstochastic bandits with Feedback Graphs

Properties of loss estimates:

- \( p_t(i) = \Pr_t(I_t = i) \) (prob. of playing)
- \( Q_t(i) = \Pr_t(\ell_t(i) \text{ is observed in round } t) \) (prob. of observing)
- \( \hat{\ell}_t(i) = \frac{\ell_t(i) \{\ell_t(i) \text{ is observed in round } t\}}{\gamma_t + Q_t(i)} \)
- \( \mathbb{E}_t[\hat{\ell}_t(i)] = \ell_t(i) \) unbiasedness
- \( \mathbb{E}_t[\hat{\ell}_t(i)^2] \leq \frac{1}{Q_t(i)} \) variance control

Some details of regret analysis:

- From
  \[
  \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i) \hat{\ell}_t(i) - \min_{i=1,\ldots,K} \sum_{t=1}^{T} \hat{\ell}_t(i) \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i) \hat{\ell}_t(i)^2
  \]
- Take expectations:
  \[
  R_T \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E} \left[ \sum_{i=1}^{K} \frac{p_t(i)}{Q_t(i)} \right] \leftarrow \text{variance}
  \]
Nonstochastic bandits with Feedback Graphs

Relating variance to $\alpha(G)$:

- Suppose $G$ is undirected (with self-loops)

$$\Sigma = \sum_{i=1}^{K} \frac{p(i)}{Q^G(i)} = \sum_{i=1}^{K} \sum_{j : j \rightarrow i} \frac{p(i)}{p(j)} \leq |S|$$

where $S \subseteq V$ is an independent set for $G = (V, E)$

- Init: $S = \emptyset$, $G_1 = G$, $V_1 = V$
  - Pick $i_1 = \arg\min_{i \in V_1} Q^{G_1}(i)$
  - Augment $S \leftarrow S \cup \{i_1\}$
  - Remove $i_1$ from $V_1$, all its neighbors (and incident edges in $G_1$):

$$\Sigma \leftarrow \Sigma - \sum_{j : j \rightarrow i_1} \frac{p(j)}{Q^{G_1}(j)} \geq \Sigma - \sum_{j : j \rightarrow i_1} \frac{p(j)}{Q^{G_1}(i_1)} = \Sigma - \frac{Q^{G_1}(i_1)}{Q^{G_1}(i_1)} = \Sigma - 1$$

- . . . get smaller graph $G_2 = (V_2, E_2)$ and iterate
Nonstochastic bandits with Feedback Graphs

Relating variance to $\alpha(G)$:

- Suppose $G$ is undirected (with self-loops)

$$\Sigma = \sum_{i=1}^{K} \frac{p(i)}{Q^{G}(i)} = \sum_{i=1}^{K} \sum_{j: j \xrightarrow{G} i} \frac{p(j)}{p(i)} \leq |S|$$

where $S \subseteq V$ is an independent set for $G = (V, E)$

- Init: $S = \emptyset$, $G_1 = G$, $V_1 = V$
  - Pick $i_2 = \arg \min_{i \in V_2} Q^{G_2}(i)$
  - Augment $S \leftarrow S \cup \{i_2\}$
  - Remove $i_2$ from $V_2$, all its neighbors (and incident edges in $G_2$):

$$\Sigma \leftarrow \Sigma - \sum_{j: j \xrightarrow{G_2} i_2} \frac{p(j)}{Q^{G_1}(j)} \geq \Sigma - \sum_{j: j \xrightarrow{G_2} i_2} \frac{p(j)}{Q^{G_2}(i_2)} = \Sigma - \frac{Q^{G_2}(i_1)}{Q^{G_2}(i_2)} = \Sigma - 1$$

- ... get smaller graph $G_3 = (V_3, E_3)$ and iterate
Nonstochastic bandits with Feedback Graphs

Relating variance to $\alpha(G)$:

- Suppose $G$ is undirected (with self-loops)

$$
\Sigma = \sum_{i=1}^{K} \frac{p(i)}{Q^G(i)} = \sum_{i=1}^{K} \frac{p(i)}{\sum_{j : j \xrightarrow{G} i} p(j)} \leq |S|
$$

where $S \subseteq V$ is an independent set for $G = (V, E)$

- Init: $S = \emptyset$; $G_1 = G$, $V_1 = V$
  - Pick $i_3 = \arg\min_{i \in V_3} Q^{G_3}(i)$
  - Augment $S \leftarrow S \cup \{i_3\}$
  - Remove $i_3$ from $V_3$, all its neighbors (and incident edges in $G_3$):

$$
\Sigma \leftarrow \Sigma - \sum_{j : j \xrightarrow{G_3} i_3} \frac{p(j)}{Q^{G_1}(j)} \geq \Sigma - \sum_{j : j \xrightarrow{G_3} i_3} \frac{p(j)}{Q^{G_3}(i_3)} = \Sigma - \frac{Q^{G_3}(i_3)}{Q^{G_3}(i_3)} = \Sigma - 1
$$

- . . . get smaller graph $G_4 = (V_4, E_4)$ and iterate
Nonstochastic bandits with Feedback Graphs

Relating variance to $\alpha(G)$:

- Suppose $G$ is undirected (with self-loops)

\[
\Sigma = \sum_{i=1}^{K} \frac{p(i)}{Q^G(i)} = \sum_{i=1}^{K} \sum_{j: j \rightarrow_{G} i} p(j) \leq |S|
\]

where $S \subseteq V$ is an independent set for $G = (V, E)$

- Init: $S = \emptyset$, $G_1 = G$, $V_1 = V$
  - Pick $i_4 = \arg\min_{i \in V_4} Q^{G_4}(i)$
  - Augment $S \leftarrow S \cup \{i_4\}$
  - Remove $i_4$ from $V_4$, all its neighbors (and incident edges in $G_4$):

\[
\Sigma \leftarrow \Sigma - \sum_{j: j \rightarrow_{G_4} i_4} \frac{p(j)}{Q_{G_1}(j)} \geq \Sigma - \sum_{j: j \rightarrow_{G_4} i_4} \frac{p(j)}{Q_{G_4}(i_4)} = \Sigma - \frac{Q^{G_4}(i_4)}{Q^{G_4}(i_4)} = \Sigma - 1
\]

- . . . get smaller graph $G_4 = (V_4, E_4)$ and iterate
Nonstochastic bandits with Feedback Graphs

Hence:

- $\Sigma$ decreases by at most 1
- $|S|$ increases by 1
- Potential $|S| + \Sigma$ increases over iterations:
  - has minimal value at the beginning ($S = \emptyset$)
  - reaches maximal value is when $G$ becomes empty ($\Sigma = 0$)
- $S$ is independent set by construction
- $|S| \leq \alpha(G)$

When $G$ directed analysis gets more complicated (needs lower bound on $p_t(i)$) and adds a log $T$ factor in bound

Have obtained:

$$R_T \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \alpha(G_t) = \mathcal{O} \left( \sqrt{\ln K \sum_{t=1}^{T} \alpha(G_t)} \right)$$
Stochastic bandit problem

- $K$ actions for Learner
- When picking action $i$ at time $t$, Learner receives as reward independent realization of random variable $X_i : \mathbb{E}[X_i] = \mu_i, \quad X_i \in [0, 1]$
- The $\mu_i$s are hidden to Learner

For $t = 1, 2, \ldots$ :

1. Learner picks action $I_t$ (possibly using random.) and gathers reward $X_{I_t,t}$
2. Learner gets feedback information: $X_{I_t,t}$

Goal: Optimize (pseudo)regret

$$R_T = \max_{i=1 \ldots K} \mathbb{E} \left[ \sum_{t=1}^{T} X_{i,t} \right] - \mathbb{E} \left[ \sum_{t=1}^{T} X_{I_t,t} \right] = \mu^* T - \mathbb{E} \left[ \sum_{t=1}^{T} X_{I_t,t} \right] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(T)]$$
Stochastic bandit problem/1

- \( K \) actions for Learner
- When picking action \( i \) at time \( t \), Learner receives as \textit{reward} independent realization of random variable \( X_i : \mathbb{E}[X_i] = \mu_i, \quad X_i \in [0, 1] \)
- The \( \mu_i \)s are hidden to Learner

For \( t = 1, 2, \ldots \):

1. Learner picks action \( I_t \) (possibly using random.) and gathers reward \( X_{I_t,t} \)

2. Learner gets feedback information: \( X_{I_t,t} \)

Goal: Optimize \text{(pseudo)}regret

\[
R_T = \max_{i=1 \ldots K} \mathbb{E} \left[ \sum_{t=1}^{T} X_{i,t} \right] - \mathbb{E} \left[ \sum_{t=1}^{T} X_{I_t,t} \right] = \mu^* T - \mathbb{E} \left[ \sum_{t=1}^{T} X_{I_t,t} \right] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(T)]
\]
Stochastic bandit problem

- \( K \) actions for Learner
- When picking action \( i \) at time \( t \), Learner receives as reward independent realization of random variable \( X_i : \mathbb{E}[X_i] = \mu_i, \quad X_i \in [0, 1] \)
- The \( \mu_i \)'s are hidden to Learner

For \( t = 1, 2, \ldots \) :

1. Learner picks action \( I_t \) (possibly using random.) and gathers reward \( X_{I_t,t} \)
2. Learner gets feedback information: \( X_{I_t,t} \)

Goal: Optimize (pseudo)regret

\[
R_T = \max_{i=1 \ldots K} \mathbb{E} \left[ \sum_{t=1}^{T} X_{i,t} \right] - \mathbb{E} \left[ \sum_{t=1}^{T} X_{I_t,t} \right] = \mu^* T - \mathbb{E} \left[ \sum_{t=1}^{T} X_{I_t,t} \right] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(T)]
\]
Stochastic bandit problem

- $K$ actions for Learner
- When picking action $i$ at time $t$, Learner receives as reward independent realization of random variable $X_i: \mathbb{E}[X_i] = \mu_i, \quad X_i \in [0, 1]$
- The $\mu_i$s are hidden to Learner

For $t = 1, 2, \ldots$ :

1. Learner picks action $I_t$ (possibly using random.) and gathers reward $X_{I_t,t}$
2. Learner gets feedback information: $X_{I_t,t}$

Goal: Optimize (pseudo)regret

$$R_T = \max_{i=1 \ldots K} \mathbb{E} \left[ \sum_{t=1}^{T} X_{i,t} \right] - \mathbb{E} \left[ \sum_{t=1}^{T} X_{I_t,t} \right] = \mu^* T - \mathbb{E} \left[ \sum_{t=1}^{T} X_{I_t,t} \right] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(T)]$$
Stochastic bandit problem/2: UCB alg

At round $t$ pick action

$$I_t = \arg\max_{i=1...K} \left( \bar{X}_{i,t-1} + \sqrt{\frac{\ln t}{T_{i,t-1}}} \right)$$

- $T_{i,t-1} =$ no. of times reward of action $i$ has been observed so far

- $\bar{X}_{i,t-1} = \frac{1}{T_{i,t-1}} \sum_{s \leq t-1: I_s = i} X_{i,s} =$ average reward of action $i$ observed so far

(Pseudo)Regret:

$$R_T = O \left( \left( \sum_{i=1}^{K} \frac{1}{\Delta_i} \right) \ln T + K \right)$$
Stochastic bandits with feedback graphs

- $K$ actions for Learner, arranged into a fixed graph $G = (V,E)$
- When picking action $i$ at time $t$, Learner receives as reward independent realization of random variable $X_i : \mathbb{E}[X_i] = \mu_i$, but also reward of nearby actions in $G$
- The $\mu_i$s are hidden to Learner

For $t = 1, 2, \ldots$

1. Learner picks action $I_t$ (possibly using random.) and gathers reward $X_{I_t,t}$
2. Learner gets feedback information: $\{X_{j,t} : (I_t, j) \in E\}$

Goal: Optimize (pseudo)regret

$$R_T = \max_{i=1 \ldots K} \mathbb{E} \left[ \sum_{t=1}^{T} X_{i,t} \right] - \mathbb{E} \left[ \sum_{t=1}^{T} X_{I_t,t} \right] = \mu^* T - \mathbb{E} \left[ \sum_{t=1}^{T} X_{I_t,t} \right] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(T)]$$
Stochastic bandits with feedback graphs/2: UCB-N [Ca+12]

At round $t$ pick action

$$I_t = \arg\max_{i=1 \ldots K} \left( \bar{X}_{i,t-1} + \sqrt{\frac{\ln t}{O_{i,t-1}}} \right)$$

- $O_{i,t-1}$ = no. of times reward of action $i$ has been observed so far
- $\bar{X}_{i,t-1} = \frac{1}{O_{i,t-1}} \sum_{s\leq t-1: I_s \rightarrow i} X_{i,s}$ = average reward of action $i$ observed so far
Stochastic bandits with feedback graphs/3

Clique covering number $\chi(G)$ : assume $G$ is undirected

$c(G) = 4$

1 clique: expert problem $\leq \alpha(G) \leq \chi(G) \leq \sqrt{\sum_{C \in \mathcal{C}} \max_{i \in C} \Delta_i \min_{i \in C} \Delta_i^2} \ln T + K$

edgeless: bandit problem

Regret analysis:

- Given any partition $\mathcal{C}$ of $V$ into cliques: $\mathcal{C} = \{C_1, C_2, \ldots, C_{|\mathcal{C}|}\}$

- $R_T = \mathcal{O}\left(\sum_{C \in \mathcal{C}} \max_{i \in C} \Delta_i \min_{i \in C} \Delta_i^2 \ln T + K\right)$

- Sum over $\leq \chi(G)$ regret terms (but can be improved to "$\leq \alpha(G)$")

- Term $K$ replaced by $\chi(G)$ by modified alg.

- No tight lower bounds available
**Simple examples/1: Auctions (nonstoc.)**

- Second-price auction with reserve (seller side)
  - highest bid revealed to seller (e.g. AppNexus)

- Auctioneer is third party

- After seller plays reserve price $I_t$, both seller’s revenue and highest bid revealed to him/her

- Seller/Player in a position to observe all revenues for prices $j \geq I_t$

- $\alpha(G) = 1$: $R_T = O\left(\ln(TK)\sqrt{T}\right)$ (expert problem up to logs) [CB+17]
Simple examples/2: “Contextual” bandits (nonstoc.) [Auer+02]

$K$ predictors

$$f_i : \{1 \ldots T\} \rightarrow \{1 \ldots N\}, \quad i = 1 \ldots K,$$

each one having the same $N << K$ actions

Learner's “action space” is the set of $K$ predictors

For $t = 1, 2, \ldots$:

1. $\ell_t(j) \in [0, 1]$ are assigned deterministically by Nature to every action $j = 1 \ldots N$ (hidded to Learner)

2. Learner observes $f_1(t) \ f_2(t) \ \ldots \ f_K(t)$

3. Learner picks predictor $f_{I_t}$ (possibly using randomization) and incurs loss $\ell_t(f_{I_t}(t))$

4. Learner gets feedback information: $\ell_t(f_{I_t}(t))$

Feedback graph $G_t$ on $K$ predictors made up of $\leq N$ cliques

$$\{i : f_i(t) = 1\} \ \{i : f_i(t) = 2\} \ \ldots \ \{i : f_i(t) = N\}$$

Independence number: $\alpha(G_t) \leq N \ \forall t$
References


Ne15: G. Neu, Explore no more: Improved high-probability regret bounds for non-stochastic bandits, NIPS 2015.