Jeux de bandits

Sébastien Bubeck,

Centre de Recerca Matemàtica, Barcelone
Standard prediction game

**Parameters:** number of rounds $n$, set of actions $\mathcal{A} = \{1, \ldots, K\}$. For each round $t = 1, 2, \ldots, n$;

1. The player chooses an action $A_t \in \mathcal{A}$.
2. Simultaneously a gain $g_{a,t} \in [0, 1]$ is assigned to each action $a \in \mathcal{A}$.
3. The player receives the gain $g_{A_t,t}$. He observes the gain $g_{a,t}$ of every action $a \in \mathcal{A}$.

**Goal:** Maximize the cumulative gains obtained. We consider the regret:

$$R_n = \max_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} g_{a,t} - \mathbb{E} \sum_{t=1}^{n} g_{A_t,t}.$$
**Parameters:** number of rounds $n$, set of actions $\mathcal{A} = \{1, \ldots, K\}$.  
For each round $t = 1, 2, \ldots, n$;

1. The player chooses an action $A_t \in \mathcal{A}$.
2. Simultaneously a gain $g_{a,t} \in [0, 1]$ is assigned to each action $a \in \mathcal{A}$.
3. The player receives the gain $g_{A_t, t}$. He observes the gain $g_{a,t}$ of every action $a \in \mathcal{A}$.

**Goal:** Maximize the cumulative gains obtained. We consider the regret:

$$R_n = \max_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} g_{a,t} - \mathbb{E} \sum_{t=1}^{n} g_{A_t, t}.$$
Standard prediction game

**Parameters:** number of rounds \( n \), set of actions \( \mathcal{A} = \{1, \ldots, K\} \).

For each round \( t = 1, 2, \ldots, n \);

1. The player chooses an action \( A_t \in \mathcal{A} \).
2. Simultaneously a gain \( g_{a,t} \in [0, 1] \) is assigned to each action \( a \in \mathcal{A} \).
3. The player receives the gain \( g_{A_t,t} \). He observes the gain \( g_{a,t} \) of every action \( a \in \mathcal{A} \).

**Goal:** Maximize the cumulative gains obtained. We consider the regret:

\[
R_n = \max_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} g_{a,t} - \mathbb{E} \sum_{t=1}^{n} g_{A_t,t}.
\]
Parameters: number of rounds $n$, set of actions $A = \{1, \ldots , K\}$. For each round $t = 1, 2, \ldots , n$;

1. The player chooses an action $A_t \in A$.

2. Simultaneously a gain $g_{a,t} \in [0, 1]$ is assigned to each action $a \in A$.

3. The player receives the gain $g_{A_t,t}$. He observes the gain $g_{a,t}$ of every action $a \in A$.

Goal: Maximize the cumulative gains obtained. We consider the regret:

$$R_n = \max_{a \in A} \mathbb{E} \sum_{t=1}^{n} g_{a,t} - \mathbb{E} \sum_{t=1}^{n} g_{A_t,t}.$$
Parameters: number of rounds \( n \), set of actions \( \mathcal{A} = \{1, \ldots, K\} \).

For each round \( t = 1, 2, \ldots, n \);

1. The player chooses an action \( A_t \in \mathcal{A} \).

2. Simultaneously a gain \( g_{a,t} \in [0, 1] \) is assigned to each action \( a \in \mathcal{A} \).

3. The player receives the gain \( g_{A_t,t} \). He observes the gain \( g_{a,t} \) of every action \( a \in \mathcal{A} \).

Goal: Maximize the cumulative gains obtained. We consider the regret:

\[
R_n = \max_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} g_{a,t} - \mathbb{E} \sum_{t=1}^{n} g_{A_t,t}.
\]
### Example (Prediction with expert advice)

Here $A$ is a set of experts trying to predict some sequence, and $g_{a,t}$ is the quality of the prediction of expert $a$ for the $t^{th}$ element of the sequence.

<table>
<thead>
<tr>
<th>Theorem (Hannan [1957])</th>
</tr>
</thead>
<tbody>
<tr>
<td>There exists a strategy such that $R_n = o(n)$.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theorem (Cesa-Bianchi et al. [1997])</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Exp}$ satisfies $R_n \leq \sqrt{\frac{n \log K}{2}}$. Moreover for any strategy, $\sup_{\text{adversaries}} R_n \geq \sqrt{\frac{n \log K}{2}} + o(\sqrt{n \log K})$.</td>
</tr>
</tbody>
</table>
Example (Prediction with expert advice)

Here $\mathcal{A}$ is a set of experts trying to predict some sequence, and $g_{a,t}$ is the quality of the prediction of expert $a$ for the $t^{th}$ element of the sequence.

Theorem (Hannan [1957])

There exists a strategy such that $R_n = o(n)$.

Theorem (Cesa-Bianchi et al. [1997])

Exp satisfies $R_n \leq \sqrt{\frac{n \log K}{2}}$. Moreover for any strategy, $\sup_{\text{adversaries}} R_n \geq \sqrt{\frac{n \log K}{2}} + o(\sqrt{n \log K})$. 
Standard prediction game

Example (Prediction with expert advice)
Here $A$ is a set of experts trying to predict some sequence, and $g_{a,t}$ is the quality of the prediction of expert $a$ for the $t^{th}$ element of the sequence.

Theorem (Hannan [1957])

There exists a strategy such that $R_n = o(n)$.

Theorem (Cesa-Bianchi et al. [1997])

$\text{Exp}$ satisfies $R_n \leq \sqrt{\frac{n \log K}{2}}$. Moreover for any strategy, \[ \sup_{\text{adversaries}} R_n \geq \sqrt{\frac{n \log K}{2}} + o\left(\sqrt{n \log K}\right). \]
Example (Prediction with expert advice)

Here $A$ is a set of experts trying to predict some sequence, and $g_{a,t}$ is the quality of the prediction of expert $a$ for the $t^{th}$ element of the sequence.

Theorem (Hannan [1957])

There exists a strategy such that $R_n = o(n)$.

Theorem (Cesa-Bianchi et al. [1997])

$Exp$ satisfies $R_n \leq \sqrt{\frac{n \log K}{2}}$. Moreover for any strategy,

$$\sup_{\text{adversaries}} R_n \geq \sqrt{\frac{n \log K}{2}} + o\left(\sqrt{n \log K}\right).$$
In the **bandit game**, the player only observes the gain $g_{A_t,t}$ of the chosen action.

This type of feedback raises an exploration versus exploitation tradeoff.

**Bandit information** is suited to many real-world applications.
• In the bandit game, the player only observes the gain $g_{A_t,t}$ of the chosen action.

• This type of feedback raises an exploration versus exploitation tradeoff.

• Bandit information is suited to many real-world applications.
In the **bandit game**, the player only observes the gain $gA_{t,t}$ of the chosen action.

This type of feedback raises an **exploration versus exploitation tradeoff**.

**Bandit information** is suited to many real-world applications.
Minimax regret for the bandit game

**Theorem (Auer et al. [1995])**

*Exp3 satisfies:*

\[ R_n \leq \sqrt{2nK \log K}. \]

Moreover for any strategy,

\[ \sup_{\text{adversaries}} R_n \geq \frac{1}{4} \sqrt{nK} + o(\sqrt{nK}). \]

**Theorem (Audibert and Bubeck [2009], Audibert and Bubeck [2010], Audibert, Bubeck, Lugosi [2011])**

*Poly INF satisfies:*

\[ R_n \leq 2\sqrt{2nK}. \]
Minimax regret for the bandit game

**Theorem (Auer et al. [1995])**

*Exp3 satisfies:*
\[
R_n \leq \sqrt{2nK \log K}.
\]

*Moreover for any strategy,*
\[
\sup_{\text{adversaries}} R_n \geq \frac{1}{4} \sqrt{nK} + o(\sqrt{nK}).
\]

**Theorem (Audibert and Bubeck [2009], Audibert and Bubeck [2010], Audibert, Bubeck, Lugosi [2011])**

*Poly INF satisfies:*
\[
R_n \leq 2\sqrt{2nK}.
\]
**Theorem (Auer et al. [1995])**

*Exp3 satisfies:*

\[ R_n \leq \sqrt{2nK \log K}. \]

*Moreover for any strategy,*

\[ \sup_{\text{adversaries}} R_n \geq \frac{1}{4} \sqrt{nK} + o(\sqrt{nK}). \]

**Theorem (Audibert and Bubeck [2009], Audibert and Bubeck [2010], Audibert, Bubeck, Lugosi [2011])**

*Poly INF satisfies:*

\[ R_n \leq 2\sqrt{2nK}. \]
Here we assume that for any action \( a \in \mathcal{A} \), the sequence \( g_{a,1}, \ldots, g_{a,n} \) is an i.i.d sequence of random variables (and independent of each other).

- This assumption allows for powerful new strategies that exploit concentration properties of sums of independent random variables.
- For instance there exists strategies with \( R_n = O(\log n) \) (instead of \( R_n = O(\sqrt{n}) \) in the general case).
Here we assume that for any action $a \in A$, the sequence $g_{a,1}, \ldots, g_{a,n}$ is an i.i.d sequence of random variables (and independent of each other).

- This assumption allows for powerful new strategies that exploit concentration properties of sums of independent random variables.

- For instance there exists strategies with $R_n = O(\log n)$ (instead of $R_n = O(\sqrt{n})$ in the general case).
Here we assume that for any action $a \in \mathcal{A}$, the sequence $g_{a,1}, \ldots, g_{a,n}$ is an i.i.d sequence of random variables (and independent of each other).

- This assumption allows for powerful new strategies that exploit concentration properties of sums of independent random variables.
- For instance there exists strategies with $R_n = O(\log n)$ (instead of $R_n = O(\sqrt{n})$ in the general case).
Here we consider the stochastic bandit game with a new objective: the player seeks to maximize the gain $g_{A_n,n}$ of the last round. This new objective changes dramatically the optimal strategies.

Figure: Three groups of bad actions, $K = 30$, $1 \times Ber(0.5)$, $5 \times Ber(0.45)$, $14 \times Ber(0.43)$, $10 \times Ber(0.38)$. 
There exists many more extensions of the bandit game:

- **Linear bandits**: $A$ is a vector space and the gain is a linear function of the action taken,
- **Lipschitz bandits**: $A$ is a metric space and the gain is a Lipschitz function,
- **Contextual bandits**: a side information is given at each round,
- **Specific forms of dependency** between the actions for stochastic bandits,
- **Mortal bandits**: set of actions varying over time.
There exists many more extensions of the bandit game:

- **Linear bandits**: $\mathcal{A}$ is a vector space and the gain is a linear function of the action taken,
- **Lipschitz bandits**: $\mathcal{A}$ is a metric space and the gain is a Lipschitz function,
- **Contextual bandits**: a side information is given at each round,
- **Specific forms of dependency** between the actions for stochastic bandits,
- **Mortal bandits**: set of actions varying over time.
Conclusion

There exists many more extensions of the bandit game:

- **Linear bandits**: $A$ is a vector space and the gain is a linear function of the action taken,
- **Lipschitz bandits**: $A$ is a metric space and the gain is a Lipschitz function,
- **Contextual bandits**: a side information is given at each round,
- Specific forms of dependency between the actions for stochastic bandits,
- **Mortal bandits**: set of actions varying over time.
There exists many more extensions of the bandit game:

- **Linear bandits**: $\mathcal{A}$ is a vector space and the gain is a linear function of the action taken,
- **Lipschitz bandits**: $\mathcal{A}$ is a metric space and the gain is a Lipschitz function,
- **Contextual bandits**: a side information is given at each round,
- **Specific forms of dependency** between the actions for stochastic bandits,
- **Mortal bandits**: set of actions varying over time.
There exists many more extensions of the bandit game:

- **Linear bandits:** $\mathcal{A}$ is a vector space and the gain is a linear function of the action taken,
- **Lipschitz bandits:** $\mathcal{A}$ is a metric space and the gain is a Lipschitz function,
- **Contextual bandits:** a side information is given at each round,
- **Specific forms of dependency** between the actions for stochastic bandits,
- **Mortal bandits:** set of actions varying over time.