Jeux de bandits

Sébastien Bubeck,

Centre de Recerca Matemàtica, Barcelone

Standard prediction game

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

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 $\mathcal{A} = \{1, \dots, K\}$.
For each round $t = 1, 2, \dots, n$;

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 $\mathcal{A} = \{1, \dots, K\}$.
For each round $t = 1, 2, \dots, n$;

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 $\mathcal{A} = \{1, \dots, K\}$.
For each round $t = 1, 2, \dots, n$;

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:

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

Standard prediction game

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 $\mathcal{A}$ 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 $\mathcal{A}$ 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,*

$$\sup_{adversaries} R_n \geq \sqrt{\frac{n \log K}{2}} + o(\sqrt{n \log K}).$$

Standard prediction game

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,

$$\sup_{\text{adversaries}} R_n \geq \sqrt{\frac{n \log K}{2}} + o(\sqrt{n \log K}).$$

Bandit information

- In the bandit game, the player only observes the gain $g_{A_t,t}$ of the choosen action.
- This type of feedback raises an exploration versus exploitation tradeoff.
- Bandit information is suited to many real-world applications.

Bandit information

- In the bandit game, the player only observes the gain $g_{A_t,t}$ of the choosen action.
- This type of feedback raises an exploration versus exploitation tradeoff.
- Bandit information is suited to many real-world applications.

Bandit information

- In the bandit game, the player only observes the gain $g_{A_t,t}$ of the choosen 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,

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

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,

$$\sup_{\textit{adversaries}} R_n \geq \frac{1}{4}\sqrt{nK} + o(\sqrt{nK}).$$

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,

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

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

Poly INF satisfies:

$$R_n \leq 2\sqrt{2nK}.$$

The stochastic bandit game, Robbins [1952]

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).

The stochastic bandit game, Robbins [1952]

Here we assume that for any action $a \in \mathcal{A}$, the sequence $g_{a,1}, \dots, g_{a,n}$ 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).

The stochastic bandit game, Robbins [1952]

Here we assume that for any action $a \in \mathcal{A}$, the sequence $g_{a,1}, \dots, g_{a,n}$ 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).

The pure exploration game, Bubeck et al. [2009, 2010, 2011]

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.

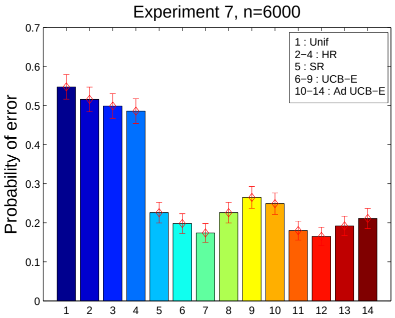

Figure: Three groups of bad actions, $K = 30$, $1 \times Ber(0.5)$, $5 \times Ber(0.45)$, $14 \times Ber(0.43)$, $10 \times Ber(0.38)$.

Conclusion

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 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: $\mathcal{A}$ 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 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: $\mathcal{A}$ 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 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: $\mathcal{A}$ 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 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: $\mathcal{A}$ 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 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.