Bandits Games

Sébastien Bubeck

Introduction

Bandits games are a framework for sequential decision making under various scenarios:

- Continuous or discrete set of actions,
- Adversarial or stochastic environment,
- different objectives: cumulative regret or simple regret,

... and many more extensions, with additional rules, new regret notions, different feedback assumptions, etc ...

Real applications include:

- ads placement on webpages,
- computer Go,
- cognitive radio,
- packets routing.

Introduction

Bandits games are a framework for sequential decision making under various scenarios:

- Continuous or discrete set of actions,
- Adversarial or stochastic environment,
- different objectives: cumulative regret or simple regret,

... and many more extensions, with additional rules, new regret notions, different feedback assumptions, etc ...

Real applications include:

- ads placement on webpages,
- computer Go,
- cognitive radio,
- packets routing.

Introduction

Bandits games are a framework for sequential decision making under various scenarios:

- Continuous or discrete set of actions,
- Adversarial or stochastic environment,
- different objectives: cumulative regret or simple regret,

... and many more extensions, with additional rules, new regret notions, different feedback assumptions, etc ...

Real applications include:

- ads placement on webpages,
- computer Go,
- cognitive radio,
- packets routing.

Introduction

Bandits games are a framework for sequential decision making under various scenarios:

- Continuous or discrete set of actions,
- Adversarial or stochastic environment,
- different objectives: cumulative regret or simple regret,

... and many more extensions, with additional rules, new regret notions, different feedback assumptions, etc ...

Real applications include:

- ads placement on webpages,
- computer Go,
- cognitive radio,
- packets routing.

Introduction

Bandits games are a framework for sequential decision making under various scenarios:

- Continuous or discrete set of actions,
- Adversarial or stochastic environment,
- different objectives: cumulative regret or simple regret,

... and many more extensions, with additional rules, new regret notions, different feedback assumptions, etc ...

Real applications include:

- ads placement on webpages,
- computer Go,
- cognitive radio,
- packets routing.

Introduction

Bandits games are a framework for sequential decision making under various scenarios:

- Continuous or discrete set of actions,
- Adversarial or stochastic environment,
- different objectives: cumulative regret or simple regret,

... and many more extensions, with additional rules, new regret notions, different feedback assumptions, etc ...

Real applications include:

- ads placement on webpages,
- computer Go,
- cognitive radio,
- packets routing.

Introduction

Bandits games are a framework for sequential decision making under various scenarios:

- Continuous or discrete set of actions,
- Adversarial or stochastic environment,
- different objectives: cumulative regret or simple regret,

... and many more extensions, with additional rules, new regret notions, different feedback assumptions, etc ...

Real applications include:

- ads placement on webpages,
- computer Go,
- cognitive radio,
- packets routing.

Introduction

Bandits games are a framework for sequential decision making under various scenarios:

- Continuous or discrete set of actions,
- Adversarial or stochastic environment,
- different objectives: cumulative regret or simple regret,

... and many more extensions, with additional rules, new regret notions, different feedback assumptions, etc ...

Real applications include:

- ads placement on webpages,
- computer Go,
- cognitive radio,
- packets routing.

Introduction

Bandits games are a framework for sequential decision making under various scenarios:

- Continuous or discrete set of actions,
- Adversarial or stochastic environment,
- different objectives: cumulative regret or simple regret,

... and many more extensions, with additional rules, new regret notions, different feedback assumptions, etc ...

Real applications include:

- ads placement on webpages,
- computer Go,
- cognitive radio,
- packets routing.

Introduction

Bandits games are a framework for sequential decision making under various scenarios:

- Continuous or discrete set of actions,
- Adversarial or stochastic environment,
- different objectives: cumulative regret or simple regret,

... and many more extensions, with additional rules, new regret notions, different feedback assumptions, etc ...

Real applications include:

- ads placement on webpages,
- computer Go,
- cognitive radio,
- packets routing.

Classical bandit game, Robbins (1952)

Parameters available to the player: the number of rounds n and the number of arms K.

Parameters unknown to the player: the reward distributions (over $[0,1]$) $\nu_1, \dots, \nu_K$ of the arms (with respective means $\mu_1, \dots, \mu_K$). Notations: $\mu^* = \max_{i=1,\dots,K} \mu_i$, $\Delta_i = \mu^* - \mu_i$, $\Delta = \min_{i:\Delta_i>0} \Delta_i$, c denotes an absolute numerical constant.

For each round $t = 1, 2, \dots, n$;

1. The player chooses an arm $I_t \in \{1, \dots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

Goal: Maximize (in expectation) the cumulative rewards. Equivalently we want to minimize the cumulative regret:

$$R_n = n\mu^* - \mathbb{E} \sum_{t=1}^{n} Y_t.$$

Classical bandit game, Robbins (1952)

Parameters available to the player: the number of rounds n and the number of arms K.

Parameters unknown to the player: the reward distributions (over $[0, 1]$) $\nu_1, \ldots, \nu_K$ of the arms (with respective means $\mu_1, \ldots, \mu_K$). Notations: $\mu^* = \max_{i=1,\ldots,K} \mu_i$, $\Delta_i = \mu^* - \mu_i$, $\Delta = \min_{i:\Delta_i>0} \Delta_i$, c denotes an absolute numerical constant.

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

1. The player chooses an arm $I_t \in \{1, \ldots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

Goal: Maximize (in expectation) the cumulative rewards. Equivalently we want to minimize the cumulative regret:

$$R_n = n\mu^* - \mathbb{E} \sum_{t=1}^{n} Y_t.$$

Classical bandit game, Robbins (1952)

Parameters available to the player: the number of rounds n and the number of arms K.

Parameters unknown to the player: the reward distributions (over $[0,1]$) $\nu_1, \dots, \nu_K$ of the arms (with respective means $\mu_1, \dots, \mu_K$). Notations: $\mu^* = \max_{i=1,\dots,K} \mu_i$, $\Delta_i = \mu^* - \mu_i$, $\Delta = \min_{i:\Delta_i>0} \Delta_i$, c denotes an absolute numerical constant.

For each round $t = 1, 2, \dots, n$;

1. The player chooses an arm $I_t \in \{1, \dots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

Goal: Maximize (in expectation) the cumulative rewards. Equivalently we want to minimize the cumulative regret:

$$R_n = n\mu^* - \mathbb{E}\sum_{t=1}^{n} Y_t.$$

Classical bandit game, Robbins (1952)

Parameters available to the player: the number of rounds n and the number of arms K.

Parameters unknown to the player: the reward distributions (over $[0,1]$) $\nu_1, \dots, \nu_K$ of the arms (with respective means $\mu_1, \dots, \mu_K$). Notations: $\mu^* = \max_{i=1,\dots,K} \mu_i$, $\Delta_i = \mu^* - \mu_i$, $\Delta = \min_{i:\Delta_i>0} \Delta_i$, c denotes an absolute numerical constant.

For each round $t = 1, 2, \dots, n$;

1. The player chooses an arm $I_t \in \{1, \dots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

Goal: Maximize (in expectation) the cumulative rewards. Equivalently we want to minimize the cumulative regret:

$$R_n = n\mu^* - \mathbb{E} \sum_{t=1}^{n} Y_t.$$

Classical bandit game, Robbins (1952)

Parameters available to the player: the number of rounds n and the number of arms K.

Parameters unknown to the player: the reward distributions (over $[0, 1]$) $\nu_1, \ldots, \nu_K$ of the arms (with respective means $\mu_1, \ldots, \mu_K$). Notations: $\mu^* = \max_{i=1,\ldots,K} \mu_i$, $\Delta_i = \mu^* - \mu_i$, $\Delta = \min_{i:\Delta_i>0} \Delta_i$, c denotes an absolute numerical constant.

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

1. The player chooses an arm $I_t \in \{1, \ldots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

Goal: Maximize (in expectation) the cumulative rewards. Equivalently we want to minimize the cumulative regret:

$$R_n = n\mu^* - \mathbb{E} \sum_{t=1}^{n} Y_t.$$

Classical bandit game, Robbins (1952)

Parameters available to the player: the number of rounds n and the number of arms K.

Parameters unknown to the player: the reward distributions (over $[0,1]$) $\nu_1, \dots, \nu_K$ of the arms (with respective means $\mu_1, \dots, \mu_K$). Notations: $\mu^* = \max_{i=1,\dots,K} \mu_i$, $\Delta_i = \mu^* - \mu_i$, $\Delta = \min_{i:\Delta_i>0} \Delta_i$, c denotes an absolute numerical constant.

For each round $t = 1, 2, \dots, n$;

1. The player chooses an arm $I_t \in \{1, \dots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

Goal: Maximize (in expectation) the cumulative rewards. Equivalently we want to minimize the cumulative regret:

$$R_n = n\mu^* - \mathbb{E} \sum_{t=1}^{n} Y_t.$$

Classical bandit game, Robbins (1952)

Parameters available to the player: the number of rounds n and the number of arms K.

Parameters unknown to the player: the reward distributions (over $[0,1]$) $\nu_1, \dots, \nu_K$ of the arms (with respective means $\mu_1, \dots, \mu_K$). Notations: $\mu^* = \max_{i=1,\dots,K} \mu_i$, $\Delta_i = \mu^* - \mu_i$, $\Delta = \min_{i:\Delta_i>0} \Delta_i$, c denotes an absolute numerical constant.

For each round $t = 1, 2, \dots, n$;

1. The player chooses an arm $I_t \in \{1, \dots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

Goal: Maximize (in expectation) the cumulative rewards.

Equivalently we want to minimize the cumulative regret:

$$R_n = n\mu^* - \mathbb{E} \sum_{t=1}^{n} Y_t.$$

Strategies based on optimism in face of uncertainty

- Let $T_i(t)$ be the number of times arm i has been selected up to time t.
- Let $\widehat{X}_{i,t}$ be the empirical mean of arm i at time t (that is based on $T_i(t)$ rewards).
- UCB (Upper Confidence Bound), Auer, Cesa-Bianchi, and Fischer (2002):

$$I_{t+1} = \arg\max_{i \in \{1,\dots,K\}} \widehat{X}_{i,t} + \sqrt{\frac{\alpha \log t}{T_i(t)}}.$$

- MOSS (Minimax Optimal Stochastic Strategy), Audibert and Bubeck (2009):

$$I_{t+1} = \arg\max_{i \in \{1,\dots,K\}} \widehat{X}_{i,t} + \sqrt{\frac{\max\left(\log\left(\frac{n}{K T_i(t)}\right), 0\right)}{T_i(t)}}.$$

Strategies based on optimism in face of uncertainty

- Let $T_i(t)$ be the number of times arm i has been selected up to time t.
- Let $\widehat{X}_{i,t}$ be the empirical mean of arm i at time t (that is based on $T_i(t)$ rewards).
- UCB (Upper Confidence Bound), Auer, Cesa-Bianchi, and Fischer (2002):

$$I_{t+1} = \arg \max_{i \in \{1,\dots,K\}} \widehat{X}_{i,t} + \sqrt{\frac{\alpha \log t}{T_i(t)}} .$$

- MOSS (Minimax Optimal Stochastic Strategy), Audibert and Bubeck (2009):

$$I_{t+1} = \arg \max_{i \in \{1,\dots,K\}} \widehat{X}_{i,t} + \sqrt{\frac{\max\left(\log\left(\frac{n}{KT_i(t)}\right),0\right)}{T_i(t)}} .$$

Strategies based on optimism in face of uncertainty

- Let $T_i(t)$ be the number of times arm i has been selected up to time t.
- Let $\widehat{X}_{i,t}$ be the empirical mean of arm i at time t (that is based on $T_i(t)$ rewards).
- UCB (Upper Confidence Bound), Auer, Cesa-Bianchi, and Fischer (2002):

$$I_{t+1} = \arg \max_{i \in \{1,\dots,K\}} \widehat{X}_{i,t} + \sqrt{\frac{\alpha \log t}{T_i(t)}} \; .$$

- MOSS (Minimax Optimal Stochastic Strategy), Audibert and Bubeck (2009):

$$I_{t+1} = \arg \max_{i \in \{1,\dots,K\}} \widehat{X}_{i,t} + \sqrt{\frac{\max\left(\log\left(\frac{n}{K T_i(t)}\right), 0 \right)}{T_i(t)}} \; .$$

Strategies based on optimism in face of uncertainty

- Let $T_i(t)$ be the number of times arm i has been selected up to time t.
- Let $\widehat{X}_{i,t}$ be the empirical mean of arm i at time t (that is based on $T_i(t)$ rewards).
- UCB (Upper Confidence Bound), Auer, Cesa-Bianchi, and Fischer (2002):

$$I_{t+1} = \arg \max_{i \in \{1,\dots,K\}} \widehat{X}_{i,t} + \sqrt{\frac{\alpha \log t}{T_i(t)}} .$$

- MOSS (Minimax Optimal Stochastic Strategy), Audibert and Bubeck (2009):

$$I_{t+1} = \arg \max_{i \in \{1,\dots,K\}} \widehat{X}_{i,t} + \sqrt{\frac{\max\left(\log \left(\frac{n}{K T_i(t)} \right) , 0 \right)}{T_i(t)}} .$$

Regret bounds for UCB and MOSS

Theorem (Auer, Cesa-Bianchi, and Fischer (2002), Audibert, Munos, and Szepesvári (2009), Bubeck (2010))

There exists $f : (1/2, +\infty) \to \mathbb{R}$ *such that UCB with* $\alpha > 1/2$ *satisfies for any* $n \geq K \geq 2$*:*

$$R_n \leq \sum_{i:\Delta_i > 0} \frac{4\alpha}{\Delta_i} \log(n) + Kf(\alpha), \text{ and } R_n \leq \sqrt{nK(4\alpha \log(n) + f(\alpha))}.$$

Theorem

MOSS satisfies:

$$R_n \leq \frac{cK}{\Delta} \log(n), \text{ and } R_n \leq c\sqrt{nK}.$$

Regret bounds for UCB and MOSS

Theorem (Auer, Cesa-Bianchi, and Fischer (2002), Audibert, Munos, and Szepesvári (2009), Bubeck (2010))

There exists $f : (1/2, +\infty) \to \mathbb{R}$ *such that* UCB *with* $\alpha > 1/2$ *satisfies for any* $n \geq K \geq 2$*:*

$$R_n \leq \sum_{i:\Delta_i>0} \frac{4\alpha}{\Delta_i} \log(n) + Kf(\alpha), \text{ and } R_n \leq \sqrt{nK(4\alpha \log(n) + f(\alpha))}.$$

Theorem

MOSS *satisfies:*

$$R_n \leq \frac{cK}{\Delta} \log(n), \text{ and } R_n \leq c\sqrt{nK}.$$

Pure exploration bandit game, joint work with Jean-Yves Audibert, Rémi Munos and Gilles Stoltz

Classical bandit game for n rounds. Then the player outputs a recommendation $J_n \in \{1, \dots, K\}$.

Goal: Maximize the expected reward of the recommended arm.
We consider the regret $r_n = \mu^* - \mathbb{E}\mu_{J_n}$.

Theorem

$$\inf_{\text{player's strategy}} \sup_{\nu} r_n = \Theta\left(\sqrt{\frac{K}{n}}\right).$$

Here we focus on the **speed of convergence** (to 0) of r_n as a function of ν.

Pure exploration bandit game, joint work with Jean-Yves Audibert, Rémi Munos and Gilles Stoltz

Classical bandit game for n rounds. Then the player outputs a recommendation $J_n \in \{1, \dots, K\}$.

Goal: Maximize the expected reward of the recommended arm.

We consider the regret $r_n = \mu^* - \mathbb{E}\mu_{J_n}$.

Theorem

$$\inf_{\textit{player's strategy}} \sup_{\nu} r_n = \Theta\left(\sqrt{\frac{K}{n}}\right).$$

Here we focus on the **speed of convergence** (to 0) of r_n as a function of ν.

Pure exploration bandit game, joint work with Jean-Yves Audibert, Rémi Munos and Gilles Stoltz

Classical bandit game for n rounds. Then the player outputs a recommendation $J_n \in \{1, \dots, K\}$.

Goal: Maximize the expected reward of the recommended arm.

We consider the regret $r_n = \mu^* - \mathbb{E}\mu_{J_n}$.

Theorem

$$\inf_{\text{player's strategy}} \sup_{\nu} r_n = \Theta\left(\sqrt{\frac{K}{n}}\right).$$

Here we focus on the **speed of convergence** (to 0) of r_n as a function of ν.

Pure exploration bandit game, joint work with Jean-Yves Audibert, Rémi Munos and Gilles Stoltz

Classical bandit game for n rounds. Then the player outputs a recommendation $J_n \in \{1, \dots, K\}$.

Goal: Maximize the expected reward of the recommended arm.

We consider the regret $r_n = \mu^* - \mathbb{E}\mu_{J_n}$.

Theorem

$$\inf_{\textit{player's strategy}} \sup_{\nu} r_n = \Theta\left(\sqrt{\frac{K}{n}}\right).$$

Here we focus on the **speed of convergence** (to 0) of r_n as a function of ν.

Uniform strategy

For each $i \in \{1, \dots, K\}$, select arm i during $\lfloor n/K \rfloor$ rounds.
Recommend the arm with highest empirical mean.

Theorem

The uniform strategy satisfies:

$$r_n \leq K \exp\left(-c\frac{n\Delta^2}{K}\right).$$

Informally, the uniform strategy needs (of order of) K/Δ^2 rounds to have a small regret. Can we do better ?
Assume that there exists a unique optimal arm i^*, then we have strategies that require only $H = \sum_{i \neq i^*} 1/\Delta_i^2$ rounds to have a small regret.

Uniform strategy

For each $i \in \{1, \ldots, K\}$, select arm i during $\lfloor n/K \rfloor$ rounds.
Recommend the arm with highest empirical mean.

Theorem

The uniform strategy satisfies:

$$r_n \leq K \exp\left(-c\frac{n\Delta^2}{K}\right).$$

Informally, the uniform strategy needs (of order of) K/Δ^2 rounds to have a small regret. Can we do better ?
Assume that there exists a unique optimal arm i^*, then we have strategies that require only $H = \sum_{i \neq i^*} 1/\Delta_i^2$ rounds to have a small regret.

Uniform strategy

For each $i \in \{1, \dots, K\}$, select arm i during $\lfloor n/K \rfloor$ rounds.
Recommend the arm with highest empirical mean.

Theorem

The uniform strategy satisfies:

$$r_n \leq K \exp\left(-c\frac{n\Delta^2}{K}\right).$$

Informally, the uniform strategy needs (of order of) K/Δ^2 rounds to have a small regret. Can we do better ?

Assume that there exists a unique optimal arm i^*, then we have strategies that require only $H = \sum_{i \neq i^*} 1/\Delta_i^2$ rounds to have a small regret.

Uniform strategy

For each $i \in \{1, \ldots, K\}$, select arm i during $\lfloor n/K \rfloor$ rounds.
Recommend the arm with highest empirical mean.

Theorem

The uniform strategy satisfies:

$$r_n \leq K \exp\left(-c\frac{n\Delta^2}{K}\right).$$

Informally, the uniform strategy needs (of order of) K/Δ^2 rounds to have a small regret. Can we do better ?
Assume that there exists a unique optimal arm i^*, then we have strategies that require only $H = \sum_{i \neq i^*} 1/\Delta_i^2$ rounds to have a small regret.

The smaller R_n the larger r_n !

Theorem

Consider any strategy and let $\epsilon : \mathbb{N} \to \mathbb{R}$ be such that for all (Bernoulli) distributions $\nu_1, \ldots, \nu_K$ on the rewards, we have

$$R_n \leq c\epsilon(n),$$

then for all sets of $K \geq 3$ (distinct, Bernoulli) distributions on the rewards, all different from a Dirac distribution at 1, up to a permutation of the arms we have,

$$r_n \geq \Delta \exp(-c\epsilon(n)) \ .$$

Successive Rejects (SR)

Let $A_1 = \{1, \dots, K\}$.

For each phase $k = 1, 2, \dots, K-1$:

(1) For each $i \in A_k$, select arm i during n_k rounds.

(2) Let $A_{k+1} = A_k \setminus \{j\}$, where j is the arm in A_k with the smallest empirical mean.

Let J_n be the unique element of A_K.

Theorem

SR satisfies (for well chosen (n_k)):

$$r_n \leq K^2 \exp\left(-c\frac{n}{\log(K)H}\right).$$

Successive Rejects (SR)

Let $A_1 = \{1, \ldots, K\}$.

For each phase $k = 1, 2, \ldots, K-1$:

(1) For each $i \in A_k$, select arm i during n_k rounds.

(2) Let $A_{k+1} = A_k \setminus \{j\}$, where j is the arm in A_k with the smallest empirical mean.

Let J_n be the unique element of A_K.

Theorem

SR satisfies (for well chosen (n_k)):

$$r_n \leq K^2 \exp\left(-c \frac{n}{\log(K) H}\right).$$

Successive Rejects (SR)

Let $A_1 = \{1, \dots, K\}$.

For each phase $k = 1, 2, \dots, K-1$:

(1) For each $i \in A_k$, select arm i during n_k rounds.

(2) Let $A_{k+1} = A_k \setminus \{j\}$, where j is the arm in A_k with the smallest empirical mean.

Let J_n be the unique element of A_K.

Theorem

SR satisfies (for well chosen (n_k)):

$$r_n \leq K^2 \exp\left(-c \frac{n}{\log(K) H}\right).$$

Successive Rejects (SR)

Let $A_1 = \{1, \dots, K\}$.

For each phase $k = 1, 2, \dots, K-1$:

(1) For each $i \in A_k$, select arm i during n_k rounds.

(2) Let $A_{k+1} = A_k \setminus \{j\}$, where j is the arm in A_k with the smallest empirical mean.

Let J_n be the unique element of A_K.

Theorem

SR satisfies (for well chosen (n_k)):

$$r_n \leq K^2 \exp\left(-c\frac{n}{\log(K)H}\right).$$

Successive Rejects (SR)

Let $A_1 = \{1, \dots, K\}$.

For each phase $k = 1, 2, \dots, K-1$:

(1) For each $i \in A_k$, select arm i during n_k rounds.

(2) Let $A_{k+1} = A_k \setminus \{j\}$, where j is the arm in A_k with the smallest empirical mean.

Let J_n be the unique element of A_K.

Theorem

SR satisfies (for well chosen (n_k)):

$$r_n \leq K^2 \exp\left(-c \frac{n}{\log(K) H}\right).$$

Successive Rejects (SR)

Let $A_1 = \{1, \dots, K\}$.

For each phase $k = 1, 2, \dots, K-1$:

(1) For each $i \in A_k$, select arm i during n_k rounds.

(2) Let $A_{k+1} = A_k \setminus \{j\}$, where j is the arm in A_k with the smallest empirical mean.

Let J_n be the unique element of A_K.

Theorem

SR satisfies (for well chosen (n_k)):

$$r_n \leq K^2 \exp\left(-c\frac{n}{\log(K)H}\right).$$

Successive Rejects (SR)

Let $A_1 = \{1, \dots, K\}$.

For each phase $k = 1, 2, \dots, K-1$:

(1) For each $i \in A_k$, select arm i during n_k rounds.

(2) Let $A_{k+1} = A_k \setminus \{j\}$, where j is the arm in A_k with the smallest empirical mean.

Let J_n be the unique element of A_K.

Theorem

SR satisfies (for well chosen (n_k)):

$$r_n \leq K^2 \exp\left(-c \frac{n}{\log(K) H}\right).$$

Lower bound

Theorem

Let $\nu_1, \dots, \nu_K$ be Bernoulli distributions with parameters in $[1/3, 2/3]$ (and a unique optimal arm). Then, for any strategy, up to a permutation of the arms,

$$r_n \geq \Delta \exp\left(-c \frac{n \log(K)}{H}\right).$$

Informally, any algorithm requires at least (of order of) $H/\log(K)$ rounds to have a small regret (and recall that SR has a small regret with $\log(K) H$ rounds).

Lower bound

Theorem

Let $\nu_1, \ldots, \nu_K$ be Bernoulli distributions with parameters in $[1/3, 2/3]$ (and a unique optimal arm). Then, for any strategy, up to a permutation of the arms,

$$r_n \geq \Delta \exp\left(-c \frac{n \log(K)}{H}\right).$$

Informally, any algorithm requires at least (of order of) $H/\log(K)$ rounds to have a small regret (and recall that SR has a small regret with $\log(K)H$ rounds).

$\mathcal{X}$-armed bandit game, joint work with Rémi Munos, Gilles Stoltz and Csaba Szepesvari

Classical bandit game where the set of arms $\{1, \dots, K\}$ is replaced by an arbitrary set $\mathcal{X}$.

Theorem

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^D$ and $\mathcal{F}$ be the set of bandits problems such that the mean-payoff function is 1-Lipschitz (with respect to some norm). Then we have

$$\inf_{\text{player's strategy}} \sup_{\mathcal{F}} R_n = \tilde{\Theta}\left(n^{\frac{D+1}{D+2}}\right).$$

Can we avoid the exponential dependence on the dimension ?

$\mathcal{X}$-armed bandit game, joint work with Rémi Munos, Gilles Stoltz and Csaba Szepesvari

Classical bandit game where the set of arms $\{1, \ldots, K\}$ is replaced by an arbitrary set $\mathcal{X}$.

Theorem

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^D$ and $\mathcal{F}$ be the set of bandits problems such that the mean-payoff function is 1-Lipschitz (with respect to some norm). Then we have

$$\inf_{\textit{player's strategy}} \sup_{\mathcal{F}} R_n = \tilde{\Theta}\left(n^{\frac{D+1}{D+2}}\right).$$

Can we avoid the exponential dependence on the dimension ?

$\mathcal{X}$-armed bandit game, joint work with Rémi Munos, Gilles Stoltz and Csaba Szepesvari

Classical bandit game where the set of arms $\{1, \ldots, K\}$ is replaced by an arbitrary set $\mathcal{X}$.

Theorem

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^D$ and $\mathcal{F}$ be the set of bandits problems such that the mean-payoff function is 1-Lipschitz (with respect to some norm). Then we have

$$\inf_{\text{player's strategy}} \sup_{\mathcal{F}} R_n = \tilde{\Theta}\left(n^{\frac{D+1}{D+2}}\right).$$

Can we avoid the exponential dependence on the dimension ?

Near-optimality dimension

Let ℓ be a *dissimilarity* measure, that is, a non-negative mapping $\ell : \mathcal{X}^2 \to \mathbb{R}$ satisfying $\ell(x, x) = 0$.

Definition

Let $f : \mathcal{X} \to [0, 1]$, $\mathcal{X}_\epsilon = \{x \in \mathcal{X}, \sup f - f(x) \le \epsilon\}$ and $\mathcal{P}(\mathcal{X}_\epsilon, \ell, \epsilon)$ be the packing number of $\mathcal{X}$ with ℓ-open balls of radius ϵ. The near-optimality dimension of f is defined as $d(f) = \limsup_{\epsilon \to 0} \frac{\log \mathcal{P}(\mathcal{X}_\epsilon, \ell, \epsilon)}{\log \epsilon^{-1}}$.

Example

Let $\mathcal{X} = [0, 1]^D$ and ℓ be some norm $|| \cdot ||$. Then $f(x) = ||x||$ satisfies $d(f) = 0$ and $g(x) = ||x||^2$ satisfies $d(g) = D/2$.

Near-optimality dimension

Let ℓ be a *dissimilarity* measure, that is, a non-negative mapping $\ell : \mathcal{X}^2 \to \mathbb{R}$ satisfying $\ell(x, x) = 0$.

Definition

Let $f : \mathcal{X} \to [0, 1]$, $\mathcal{X}_\epsilon = \{x \in \mathcal{X}, \sup f - f(x) \leq \epsilon\}$ and $\mathcal{P}(\mathcal{X}_\epsilon, \ell, \epsilon)$ be the packing number of $\mathcal{X}$ with ℓ-open balls of radius ϵ. The near-optimality dimension of f is defined as $d(f) = \limsup_{\epsilon \to 0} \frac{\log \mathcal{P}(\mathcal{X}_\epsilon, \ell, \epsilon)}{\log \epsilon^{-1}}$.

Example

Let $\mathcal{X} = [0, 1]^D$ and ℓ be some norm $||\cdot||$. Then $f(x) = ||x||$ satisfies $d(f) = 0$ and $g(x) = ||x||^2$ satisfies $d(g) = D/2$.

Near-optimality dimension

Let ℓ be a *dissimilarity* measure, that is, a non-negative mapping $\ell : \mathcal{X}^2 \to \mathbb{R}$ satisfying $\ell(x, x) = 0$.

Definition

Let $f : \mathcal{X} \to [0, 1]$, $\mathcal{X}_\epsilon = \{x \in \mathcal{X}, \sup f - f(x) \leq \epsilon\}$ and $\mathcal{P}(\mathcal{X}_\epsilon, \ell, \epsilon)$ be the packing number of $\mathcal{X}$ with ℓ-open balls of radius ϵ. The near-optimality dimension of f is defined as $d(f) = \limsup_{\epsilon \to 0} \frac{\log \mathcal{P}(\mathcal{X}_\epsilon, \ell, \epsilon)}{\log \epsilon^{-1}}$.

Example

Let $\mathcal{X} = [0, 1]^D$ and ℓ be some norm $|| \cdot ||$. Then $f(x) = ||x||$ satisfies $d(f) = 0$ and $g(x) = ||x||^2$ satisfies $d(g) = D/2$.

Regret bounds with near-optimality dimension

Theorem (Kleinberg, Slivkins, and Upfal (2008))

Let $\mathcal{X}$ be a compact metric space (with **metric** *ℓ). Consider a bandit problem such that the mean-payoff is 1-***Lipschitz** *and has a near-optimality dimension $d \geq 0$ (with respect to ℓ). Then the Zooming algorithm satisfies $R_n = \tilde{O}\left(n^{\frac{d+1}{d+2}}\right)$.*

Theorem

Let ℓ be **any dissimilarity** *and consider a bandit problem such that the mean-payoff is* **weakly-Lipschitz** *and has a near-optimality dimension $d \geq 0$ (with respect to ℓ). Then HOO satisfies (under mild 'compactness' assumption on $\mathcal{X}$) $R_n = \tilde{O}\left(n^{\frac{d+1}{d+2}}\right)$.*

Regret bounds with near-optimality dimension

Theorem (Kleinberg, Slivkins, and Upfal (2008))

Let $\mathcal{X}$ be a compact metric space (with **metric** *ℓ). Consider a bandit problem such that the mean-payoff is 1-***Lipschitz** *and has a near-optimality dimension $d \geq 0$ (with respect to ℓ). Then the Zooming algorithm satisfies $R_n = \tilde{O}\left(n^{\frac{d+1}{d+2}}\right)$.*

Theorem

Let ℓ be **any dissimilarity** *and consider a bandit problem such that the mean-payoff is* **weakly-Lipschitz** *and has a near-optimality dimension $d \geq 0$ (with respect to ℓ). Then HOO satisfies (under mild 'compactness' assumption on $\mathcal{X}$) $R_n = \tilde{O}\left(n^{\frac{d+1}{d+2}}\right)$.*

Example

$\mathcal{X} = [0,1]^D$, $\alpha \geq 0$ and mean-payoff function f locally "α-smooth" around (any of) its maximum x^* (in finite number):

$$f(x^*) - f(x) = \Theta(||x - x^*||^\alpha) \text{ as } x \to x^*.$$

Theorem

Assume that we run HOO using $\ell(x, y) = ||x - y||^\beta$.

- **Known smoothness:** $\beta = \alpha$. $R_n = \tilde{O}(\sqrt{n})$, **i.e., the rate is independent of the dimension** D.
- **Smoothness underestimated:** $\beta < \alpha$. $R_n = \tilde{O}(n^{(d+1)/(d+2)})$ *where* $d = D\left(\frac{1}{\beta} - \frac{1}{\alpha}\right)$.
- **Smoothness overestimated:** $\beta > \alpha$. *No guarantee. Note: UCT corresponds to* $\beta = +\infty$.

Example

$\mathcal{X} = [0,1]^D$, $\alpha \geq 0$ and mean-payoff function f locally "α-smooth" around (any of) its maximum x^* (in finite number):

$$f(x^*) - f(x) = \Theta(||x - x^*||^\alpha) \text{ as } x \to x^*.$$

Theorem

Assume that we run HOO using $\ell(x, y) = ||x - y||^\beta$.

- **Known smoothness:** $\beta = \alpha$. $R_n = \tilde{O}(\sqrt{n})$, i.e., the rate is independent of the dimension D.
- **Smoothness underestimated:** $\beta < \alpha$. $R_n = \tilde{O}(n^{(d+1)/(d+2)})$ *where* $d = D\left(\frac{1}{\beta} - \frac{1}{\alpha}\right)$.
- **Smoothness overestimated:** $\beta > \alpha$. *No guarantee. Note: UCT corresponds to* $\beta = +\infty$.

Example

$\mathcal{X} = [0,1]^D$, $\alpha \geq 0$ and mean-payoff function f locally "α-smooth" around (any of) its maximum x^* (in finite number):

$$f(x^*) - f(x) = \Theta(||x - x^*||^\alpha) \text{ as } x \to x^*.$$

Theorem

Assume that we run HOO using $\ell(x,y) = ||x-y||^\beta$.

- **Known smoothness:** $\beta = \alpha$. $R_n = \tilde{O}(\sqrt{n})$, **i.e., the rate is independent of the dimension** D.
- **Smoothness underestimated:** $\beta < \alpha$. $R_n = \tilde{O}(n^{(d+1)/(d+2)})$ *where* $d = D\left(\frac{1}{\beta} - \frac{1}{\alpha}\right)$.
- **Smoothness overestimated:** $\beta > \alpha$. *No guarantee. Note: UCT corresponds to* $\beta = +\infty$.

Example

$\mathcal{X} = [0,1]^D$, $\alpha \geq 0$ and mean-payoff function f locally "α-smooth" around (any of) its maximum x^* (in finite number):

$$f(x^*) - f(x) = \Theta(||x - x^*||^\alpha) \text{ as } x \to x^*.$$

Theorem

Assume that we run HOO using $\ell(x,y) = ||x-y||^\beta$.

- **Known smoothness:** $\beta = \alpha$. $R_n = \tilde{O}(\sqrt{n})$, **i.e., the rate is independent of the dimension** D.
- **Smoothness underestimated:** $\beta < \alpha$. $R_n = \tilde{O}(n^{(d+1)/(d+2)})$ *where* $d = D\left(\frac{1}{\beta} - \frac{1}{\alpha}\right)$.
- **Smoothness overestimated:** $\beta > \alpha$. *No guarantee. Note: UCT corresponds to* $\beta = +\infty$.

Example

$\mathcal{X} = [0,1]^D$, $\alpha \geq 0$ and mean-payoff function f locally "α-smooth" around (any of) its maximum x^* (in finite number):

$$f(x^*) - f(x) = \Theta(||x - x^*||^\alpha) \text{ as } x \to x^*.$$

Theorem

Assume that we run HOO using $\ell(x,y) = ||x-y||^\beta$.

- **Known smoothness:** $\beta = \alpha$. $R_n = \tilde{O}(\sqrt{n})$, **i.e., the rate is independent of the dimension** D.
- **Smoothness underestimated:** $\beta < \alpha$. $R_n = \tilde{O}(n^{(d+1)/(d+2)})$ *where* $d = D\left(\frac{1}{\beta} - \frac{1}{\alpha}\right)$.
- **Smoothness overestimated:** $\beta > \alpha$. *No guarantee. Note: UCT corresponds to* $\beta = +\infty$.

Adversarial multi-armed bandit game, joint work with Jean-Yves Audibert

For each round $t = 1, 2, \dots, n$;

1. The player chooses an arm $I_t \in \{1, \dots, K\}$, possibly with the help of an external randomization.
2. Simultaneously the adversary chooses a gain vector $g_t = (g_{1,t}, \dots, g_{K,t}) \in [0,1]^K$.
3. The player receives (and observes) the gain $g_{I_t,t}$.

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

$$R_n = \max_{i=1,\dots,K} \mathbb{E} \sum_{t=1}^{n} g_{i,t} - \mathbb{E} \sum_{t=1}^{n} g_{I_t,t}.$$

Adversarial multi-armed bandit game, joint work with Jean-Yves Audibert

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

1. The player chooses an arm $I_t \in \{1, \ldots, K\}$, possibly with the help of an external randomization.
2. Simultaneously the adversary chooses a gain vector $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0, 1]^K$.
3. The player receives (and observes) the gain $g_{I_t,t}$.

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

$$R_n = \max_{i=1,\ldots,K} \mathbb{E} \sum_{t=1}^{n} g_{i,t} - \mathbb{E} \sum_{t=1}^{n} g_{I_t,t}.$$

Adversarial multi-armed bandit game, joint work with Jean-Yves Audibert

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

1. The player chooses an arm $I_t \in \{1, \ldots, K\}$, possibly with the help of an external randomization.
2. Simultaneously the adversary chooses a gain vector $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0,1]^K$.
3. The player receives (and observes) the gain $g_{I_t,t}$.

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

$$R_n = \max_{i=1,\ldots,K} \mathbb{E} \sum_{t=1}^{n} g_{i,t} - \mathbb{E} \sum_{t=1}^{n} g_{I_t,t}.$$

Adversarial multi-armed bandit game, joint work with Jean-Yves Audibert

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

1. The player chooses an arm $I_t \in \{1, \ldots, K\}$, possibly with the help of an external randomization.
2. Simultaneously the adversary chooses a gain vector $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0,1]^K$.
3. The player receives (and observes) the gain $g_{I_t,t}$.

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

$$R_n = \max_{i=1,\ldots,K} \mathbb{E} \sum_{t=1}^{n} g_{i,t} - \mathbb{E} \sum_{t=1}^{n} g_{I_t,t}.$$

Adversarial multi-armed bandit game, joint work with Jean-Yves Audibert

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

1. The player chooses an arm $I_t \in \{1, \ldots, K\}$, possibly with the help of an external randomization.
2. Simultaneously the adversary chooses a gain vector $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0, 1]^K$.
3. The player receives (and observes) the gain $g_{I_t,t}$.

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

$$R_n = \max_{i=1,\ldots,K} \mathbb{E} \sum_{t=1}^{n} g_{i,t} - \mathbb{E} \sum_{t=1}^{n} g_{I_t,t}.$$

Known results

Theorem (Auer, Cesa-Bianchi, Freund, and Schapire (1995))

For any strategy,

$$\sup R_n \geq \frac{1}{20}\sqrt{nK}.$$

Moreover Exp3 satisfies:

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

We propose a new strategy, INF, which satisfies $R_n \leq 8\sqrt{nK}$.

Due to time constraints, we skip all the interesting extensions: label efficient games, high probability bounds, tracking the best expert bounds, bounds that scale with the optimal arm rewards.

Known results

Theorem (Auer, Cesa-Bianchi, Freund, and Schapire (1995))

For any strategy,

$$\sup R_n \geq \frac{1}{20}\sqrt{nK}.$$

Moreover Exp3 satisfies:

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

We propose a new strategy, INF, which satisfies $R_n \leq 8\sqrt{nK}$.

Due to time constraints, we skip all the interesting extensions: label efficient games, high probability bounds, tracking the best expert bounds, bounds that scale with the optimal arm rewards.

Known results

Theorem (Auer, Cesa-Bianchi, Freund, and Schapire (1995))

For any strategy,

$$\sup R_n \geq \frac{1}{20}\sqrt{nK}.$$

Moreover Exp3 satisfies:

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

We propose a new strategy, INF, which satisfies $R_n \leq 8\sqrt{nK}$.

Due to time constraints, we skip all the interesting extensions: label efficient games, high probability bounds, tracking the best expert bounds, bounds that scale with the optimal arm rewards.

INF (Implicitly Normalized Forecaster)

Parameter: function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ increasing, convex, twice continuously differentiable, and such that $(0, 1] \subset \psi(\mathbb{R}_-^*)$.

Let p_1 be the uniform distribution over $\{1, \ldots, K\}$.

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

1. $I_t \sim p_t$.
2. Compute $\tilde{g}_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t = i}$ and $\tilde{G}_{i,t} = \sum_{s=1}^{t} \tilde{g}_{i,s}$.
3. Compute the new probability distribution:

$$p_{i,t+1} = \psi(\tilde{G}_{i,t} - C_t)$$

where C_t is the unique real number such that $\sum_{i=1}^{K} p_{i,t+1} = 1$.

INF (Implicitly Normalized Forecaster)

Parameter: function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ increasing, convex, twice continuously differentiable, and such that $(0, 1] \subset \psi(\mathbb{R}_-^*)$.

Let p_1 be the uniform distribution over $\{1, \ldots, K\}$.

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

1. $I_t \sim p_t$.
2. Compute $\tilde{g}_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t = i}$ and $\tilde{G}_{i,t} = \sum_{s=1}^{t} \tilde{g}_{i,s}$.
3. Compute the new probability distribution:

$$p_{i,t+1} = \psi(\tilde{G}_{i,t} - C_t)$$

where C_t is the unique real number such that $\sum_{i=1}^{K} p_{i,t+1} = 1$.

INF (Implicitly Normalized Forecaster)

Parameter: function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ increasing, convex, twice continuously differentiable, and such that $(0, 1] \subset \psi(\mathbb{R}_-^*)$.

Let p_1 be the uniform distribution over $\{1, \ldots, K\}$.

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

1. $I_t \sim p_t$.
2. Compute $\tilde{g}_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t = i}$ and $\tilde{G}_{i,t} = \sum_{s=1}^{t} \tilde{g}_{i,s}$.
3. Compute the new probability distribution:

$$p_{i,t+1} = \psi(\tilde{G}_{i,t} - C_t)$$

where C_t is the unique real number such that $\sum_{i=1}^{K} p_{i,t+1} = 1$.

INF (Implicitly Normalized Forecaster)

Parameter: function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ increasing, convex, twice continuously differentiable, and such that $(0, 1] \subset \psi(\mathbb{R}_-^*)$.

Let p_1 be the uniform distribution over $\{1, \ldots, K\}$.

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

1. $I_t \sim p_t$.
2. Compute $\tilde{g}_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t = i}$ and $\tilde{G}_{i,t} = \sum_{s=1}^{t} \tilde{g}_{i,s}$.
3. Compute the new probability distribution:

$$p_{i,t+1} = \psi(\tilde{G}_{i,t} - C_t)$$

where C_t is the unique real number such that $\sum_{i=1}^{K} p_{i,t+1} = 1$.

INF (Implicitly Normalized Forecaster)

Parameter: function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ increasing, convex, twice continuously differentiable, and such that $(0, 1] \subset \psi(\mathbb{R}_-^*)$.

Let p_1 be the uniform distribution over $\{1, \dots, K\}$.

For each round $t = 1, 2, \dots, n$;

1. $I_t \sim p_t$.
2. Compute $\tilde{g}_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t = i}$ and $\tilde{G}_{i,t} = \sum_{s=1}^{t} \tilde{g}_{i,s}$.
3. Compute the new probability distribution:

$$p_{i,t+1} = \psi(\tilde{G}_{i,t} - C_t)$$

where C_t is the unique real number such that $\sum_{i=1}^{K} p_{i,t+1} = 1$.

INF (Implicitly Normalized Forecaster)

Parameter: function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ increasing, convex, twice continuously differentiable, and such that $(0, 1] \subset \psi(\mathbb{R}_-^*)$.

Let p_1 be the uniform distribution over $\{1, \dots, K\}$.

For each round $t = 1, 2, \dots, n$;

1. $I_t \sim p_t$.
2. Compute $\tilde{g}_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t = i}$ and $\tilde{G}_{i,t} = \sum_{s=1}^{t} \tilde{g}_{i,s}$.
3. Compute the new probability distribution:

$$p_{i,t+1} = \psi(\tilde{G}_{i,t} - C_t)$$

where C_t is the unique real number such that $\sum_{i=1}^{K} p_{i,t+1} = 1$.

Examples

1. $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\eta > 0$ and $\gamma \in [0, 1)$; this corresponds exactly to the Exp3 strategy.
2. $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $q > 1$, $\eta > 0$ and $\gamma \in [0, 1)$; this is a new strategy which will be proved to be minimax optimal for appropriate parameters.

Examples

1. $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\eta > 0$ and $\gamma \in [0, 1)$; this corresponds exactly to the Exp3 strategy.
2. $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $q > 1$, $\eta > 0$ and $\gamma \in [0, 1)$; this is a new strategy which will be proved to be minimax optimal for appropriate parameters.

Regret bound for Poly INF

Theorem

Consider $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ *with* $\gamma = \min\left(\frac{1}{2}, \sqrt{\frac{3K}{n}}\right)$, $\eta = \sqrt{5n}$ *and* $q = 2$. *Then INF satisfies:*

$$R_n \leq 8\sqrt{nK}.$$

Proof

By an Abel transform we shift the focus from:

$$\sum_{t=1}^{n} g_{I_t,t} = \sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t}(\tilde{G}_{i,t} - \tilde{G}_{i,t-1})$$

to

$$\sum_{t=1}^{n-1} \sum_{i=1}^{K} \tilde{G}_{i,t}(p_{i,t+1} - p_{i,t}) = \sum_{i=1}^{K} \sum_{t=1}^{n-1} \psi^{-1}(p_{i,t+1})(p_{i,t+1} - p_{i,t}).$$

Then a Taylor expansion gives us:

$$(p_{i,t+1} - p_{i,t})\psi^{-1}(p_{i,t+1}) = -\int_{p_{i,t+1}}^{p_{i,t}} \psi^{-1}(u)du + \frac{(p_{i,t} - p_{i,t+1})^2}{2\psi'(\psi^{-1}(\tilde{p}_{i,t+1}))}.$$

The first resulting term: $-\sum_{i=1}^{K} \int_{p_{i,n+1}}^{1/K} \psi^{-1}(u)du$ is easy to control. On the other hand for the second term we need to do a multivariate Taylor expansion on

Proof

By an Abel transform we shift the focus from:

$$\sum_{t=1}^{n} g_{I_t,t} = \sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t}(\tilde{G}_{i,t} - \tilde{G}_{i,t-1})$$

to

$$\sum_{t=1}^{n-1} \sum_{i=1}^{K} \tilde{G}_{i,t}(p_{i,t+1} - p_{i,t}) = \sum_{i=1}^{K} \sum_{t=1}^{n-1} \psi^{-1}(p_{i,t+1})(p_{i,t+1} - p_{i,t}).$$

Then a Taylor expansion gives us:

$$(p_{i,t+1} - p_{i,t})\psi^{-1}(p_{i,t+1}) = -\int_{p_{i,t+1}}^{p_{i,t}} \psi^{-1}(u)du + \frac{(p_{i,t} - p_{i,t+1})^2}{2\psi'(\psi^{-1}(\tilde{p}_{i,t+1}))}.$$

The first resulting term: $-\sum_{i=1}^{K} \int_{p_{i,n+1}}^{1/K} \psi^{-1}(u)du$ is easy to control. On the other hand for the second term we need to do a multivariate Taylor expansion on

Proof

By an Abel transform we shift the focus from:

$$\sum_{t=1}^{n} g_{I_t,t} = \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}(\tilde{G}_{i,t} - \tilde{G}_{i,t-1})$$

to

$$\sum_{t=1}^{n-1}\sum_{i=1}^{K} \tilde{G}_{i,t}(p_{i,t+1} - p_{i,t}) = \sum_{i=1}^{K}\sum_{t=1}^{n-1} \psi^{-1}(p_{i,t+1})(p_{i,t+1} - p_{i,t}).$$

Then a Taylor expansion gives us:

$$(p_{i,t+1} - p_{i,t})\psi^{-1}(p_{i,t+1}) = -\int_{p_{i,t+1}}^{p_{i,t}} \psi^{-1}(u)du + \frac{(p_{i,t} - p_{i,t+1})^2}{2\psi'(\psi^{-1}(\tilde{p}_{i,t+1}))}.$$

The first resulting term: $-\sum_{i=1}^{K}\int_{p_{i,n+1}}^{1/K}\psi^{-1}(u)du$ is easy to control. On the other hand for the second term we need to do a multivariate Taylor expansion on

$$p_{i,t} - p_{i,t+1} = \psi(\tilde{G}_{i,t} - C_t) - \psi(\tilde{G}_{i,t+1} - C_{t+1})$$

Proof

By an Abel transform we shift the focus from:

$$\sum_{t=1}^{n} g_{I_t,t} = \sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t}(\tilde{G}_{i,t} - \tilde{G}_{i,t-1})$$

to

$$\sum_{t=1}^{n-1} \sum_{i=1}^{K} \tilde{G}_{i,t}(p_{i,t+1} - p_{i,t}) = \sum_{i=1}^{K} \sum_{t=1}^{n-1} \psi^{-1}(p_{i,t+1})(p_{i,t+1} - p_{i,t}).$$

Then a Taylor expansion gives us:

$$(p_{i,t+1} - p_{i,t})\psi^{-1}(p_{i,t+1}) = -\int_{p_{i,t+1}}^{p_{i,t}} \psi^{-1}(u)du + \frac{(p_{i,t} - p_{i,t+1})^2}{2\psi'(\psi^{-1}(\tilde{p}_{i,t+1}))}.$$

The first resulting term: $-\sum_{i=1}^{K} \int_{p_{i,n+1}}^{1/K} \psi^{-1}(u)du$ is easy to control. On the other hand for the second term we need to do a multivariate Taylor expansion on

$$p_{i,t} - p_{i,t+1} = \psi(\tilde{G}_{i,t} - C_t) - \psi(\tilde{G}_{i,t+1} - C_{t+1})$$

as well as a careful treatment of the "shift" introduced by $\tilde{p}_{i,t+1}$.

Proof

By an Abel transform we shift the focus from:

$$\sum_{t=1}^{n} g_{I_t,t} = \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}(\tilde{G}_{i,t} - \tilde{G}_{i,t-1})$$

to

$$\sum_{t=1}^{n-1}\sum_{i=1}^{K} \tilde{G}_{i,t}(p_{i,t+1} - p_{i,t}) = \sum_{i=1}^{K}\sum_{t=1}^{n-1} \psi^{-1}(p_{i,t+1})(p_{i,t+1} - p_{i,t}).$$

Then a Taylor expansion gives us:

$$(p_{i,t+1} - p_{i,t})\psi^{-1}(p_{i,t+1}) = -\int_{p_{i,t+1}}^{p_{i,t}} \psi^{-1}(u)du + \frac{(p_{i,t} - p_{i,t+1})^2}{2\psi'(\psi^{-1}(\tilde{p}_{i,t+1}))}.$$

The first resulting term: $-\sum_{i=1}^{K}\int_{p_{i,n+1}}^{1/K} \psi^{-1}(u)du$ is easy to control. On the other hand for the second term we need to do a multivariate Taylor expansion on

$$p_{i,t} - p_{i,t+1} = \psi(\tilde{G}_{i,t} - C_t) - \psi(\tilde{G}_{i,t+1} - C_{t+1})$$

as well as a careful treatment of the "shift" introduced by $\tilde{p}_{i,t+1}$.

Perspectives

The possible extensions of classical bandits games are almost unlimited. The following cases are of special interest (to me).

- Exploiting the combinatorial structure in linear bandits.
- Specific forms of dependency between the arms for stochastic bandits.
- Mortal bandits: set of arms varying over time.

Perspectives

The possible extensions of classical bandits games are almost unlimited. The following cases are of special interest (to me).

- Exploiting the combinatorial structure in linear bandits.
- Specific forms of dependency between the arms for stochastic bandits.
- Mortal bandits: set of arms varying over time.

Perspectives

The possible extensions of classical bandits games are almost unlimited. The following cases are of special interest (to me).

- Exploiting the combinatorial structure in linear bandits.
- Specific forms of dependency between the arms for stochastic bandits.
- Mortal bandits: set of arms varying over time.

Perspectives

The possible extensions of classical bandits games are almost unlimited. The following cases are of special interest (to me).

- Exploiting the combinatorial structure in linear bandits.
- Specific forms of dependency between the arms for stochastic bandits.
- Mortal bandits: set of arms varying over time.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 1952.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 1995.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning Journal*, 2002.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th ACM Symposium on Theory of Computing*, 2008.

J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 2009.

S. Bubeck, R. Munos, G. Stoltz, and Cs. Szepesvari. Online optimization in $\mathcal{X}$-armed bandits. In *Advances in Neural Information Processing Systems (NIPS) 22*, 2009.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proc. of the 22nd annual conference on learning theory (COLT)*, 2009.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Proc. of the 20th International Conference on Algorithmic Learning Theory (ALT)*, 2009.

J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Proc. of the 23rd annual conference on learning theory (COLT)*, 2010.

S. Bubeck and R. Munos. Open loop optimistic planning. In *23rd annual conference on learning theory (COLT)*, 2010.