Pure Exploration in Multi-Armed Bandits Problems

Sébastien Bubeck[1]

joint work with Rémi Munos[1] & Gilles Stoltz[2,3]

[1] INRIA Lille, SequeL team
[2] Ecole normale supérieure, CNRS, Paris, France
[3] HEC Paris, CNRS, Jouy-en-Josas, France

Outline

- Mathematical description of the problem
- Concrete examples
- Analysis

Outline

- Mathematical description of the problem
- Concrete examples
- Analysis

Outline

- Mathematical description of the problem
- Concrete examples
- Analysis

Pure exploration bandit game

Parameters: K probability distributions $\nu_1, \ldots, \nu_K$ on $[0,1]$ (with respective means $\mu_1, \ldots, \mu_K$). Notation: $\mu^* = \max_{i=1,\ldots,K} \mu_i$.

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

1. The forecaster 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).
3. The forecaster outputs a recommendation $J_t \in \{1, \ldots, K\}$.

Goal: Maximize the expected reward of the recommended arm. We consider the regret at time n:

$$r_n = \mu^* - \mu_{J_n}.$$

Remark: The classical regret is $R_n = \sum_{t=1}^{n} \mu^* - \mu_{I_t}$.

Pure exploration bandit game

Parameters: K probability distributions $\nu_1, \ldots, \nu_K$ on $[0,1]$ (with respective means $\mu_1, \ldots, \mu_K$). Notation: $\mu^* = \max_{i=1,\ldots,K} \mu_i$.

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

1. The forecaster 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).
3. The forecaster outputs a recommendation $J_t \in \{1, \ldots, K\}$.

Goal: Maximize the expected reward of the recommended arm. We consider the regret at time n:

$$r_n = \mu^* - \mu_{J_n}.$$

Remark: The classical regret is $R_n = \sum_{t=1}^{n} \mu^* - \mu_{I_t}$.

Pure exploration bandit game

Parameters: K probability distributions $\nu_1, \dots, \nu_K$ on $[0,1]$ (with respective means $\mu_1, \dots, \mu_K$). Notation: $\mu^* = \max_{i=1,\dots,K} \mu_i$.

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

1. The forecaster 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).
3. The forecaster outputs a recommendation $J_t \in \{1, \dots, K\}$.

Goal: Maximize the expected reward of the recommended arm. We consider the regret at time n:

$$r_n = \mu^* - \mu_{J_n}.$$

Remark: The classical regret is $R_n = \sum_{t=1}^{n} \mu^* - \mu_{I_t}$.

Pure exploration bandit game

Parameters: K probability distributions $\nu_1, \ldots, \nu_K$ on $[0,1]$ (with respective means $\mu_1, \ldots, \mu_K$). Notation: $\mu^* = \max_{i=1,\ldots,K} \mu_i$.

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

1. The forecaster 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).
3. The forecaster outputs a recommendation $J_t \in \{1, \ldots, K\}$.

Goal: Maximize the expected reward of the recommended arm.
We consider the regret at time n:

$$r_n = \mu^* - \mu_{J_n}.$$

Remark: The classical regret is $R_n = \sum_{t=1}^{n} \mu^* - \mu_{I_t}$.

Pure exploration bandit game

Parameters: K probability distributions $\nu_1, \dots, \nu_K$ on $[0,1]$ (with respective means $\mu_1, \dots, \mu_K$). Notation: $\mu^* = \max_{i=1,\dots,K} \mu_i$.

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

1. The forecaster 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).
3. The forecaster outputs a recommendation $J_t \in \{1, \dots, K\}$.

Goal: Maximize the expected reward of the recommended arm. We consider the regret at time n:

$$r_n = \mu^* - \mu_{J_n}.$$

Remark: The classical regret is $R_n = \sum_{t=1}^{n} \mu^* - \mu_{I_t}$.

Pure exploration bandit game

Parameters: K probability distributions $\nu_1, \dots, \nu_K$ on $[0,1]$ (with respective means $\mu_1, \dots, \mu_K$). Notation: $\mu^* = \max_{i=1,\dots,K} \mu_i$.

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

1. The forecaster 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).
3. The forecaster outputs a recommendation $J_t \in \{1, \dots, K\}$.

Goal: Maximize the expected reward of the recommended arm.
We consider the regret at time n:

$$r_n = \mu^* - \mu_{J_n}.$$

Remark: The classical regret is $R_n = \sum_{t=1}^{n} \mu^* - \mu_{I_t}$.

Pure exploration bandit game

1. Note that in general $R_n \neq r_1 + \ldots + r_n$ and we even expect $\mathbb{E} r_1 + \ldots r_n << \mathbb{E} R_n$.
2. Allocation strategy (I_t) to minimize $\mathbb{E} R_n$: tradeoff between exploration and exploitation.
3. Recommendation strategy J_n to minimize $\mathbb{E} r_n$: pure exploitation of the results obtained so far.
4. Allocation strategy (I_t) to minimize $\mathbb{E} r_n$: pure exploration (to make the recommendation strategy more efficient!).

Pure exploration bandit game

1. Note that in general $R_n \neq r_1 + \ldots + r_n$ and we even expect $\mathbb{E} r_1 + \ldots r_n << \mathbb{E} R_n$.
2. Allocation strategy (I_t) to minimize $\mathbb{E} R_n$: tradeoff between exploration and exploitation.
3. Recommendation strategy J_n to minimize $\mathbb{E} r_n$: pure exploitation of the results obtained so far.
4. Allocation strategy (I_t) to minimize $\mathbb{E} r_n$: pure exploration (to make the recommendation strategy more efficient!).

Pure exploration bandit game

1. Note that in general $R_n \neq r_1 + \ldots + r_n$ and we even expect $\mathbb{E} r_1 + \ldots r_n << \mathbb{E} R_n$.
2. Allocation strategy (I_t) to minimize $\mathbb{E} R_n$: tradeoff between exploration and exploitation.
3. Recommendation strategy J_n to minimize $\mathbb{E} r_n$: pure exploitation of the results obtained so far.
4. Allocation strategy (I_t) to minimize $\mathbb{E} r_n$: pure exploration (to make the recommendation strategy more efficient!).

Pure exploration bandit game

1. Note that in general $R_n \neq r_1 + \ldots + r_n$ and we even expect $\mathbb{E} r_1 + \ldots r_n << \mathbb{E} R_n$.
2. Allocation strategy (I_t) to minimize $\mathbb{E} R_n$: tradeoff between exploration and exploitation.
3. Recommendation strategy J_n to minimize $\mathbb{E} r_n$: pure exploitation of the results obtained so far.
4. Allocation strategy (I_t) to minimize $\mathbb{E} r_n$: pure exploration (to make the recommendation strategy more efficient!).

Examples

- **Test phase of a treatment:** sequentially test different treatments for a given period of time and then select the one to be commercialized.
  Goal: minimize the regret of the commercialized product (i.e the simple regret) and not the regret of the test phase (i.e. the cumulative regret)
- **Computer Go:** Given a limited CPU time and a goban position, output the next action to play.
  Idea: This problem is a bandit game where actions = arms and round = evaluation (costly in CPU time) of an action. One wants to minimize the simple regret of the selected action once the budget is exhausted.

Examples

- **Test phase of a treatment:** sequentially test different treatments for a given period of time and then select the one to be commercialized.
  Goal: minimize the regret of the commercialized product (i.e the simple regret) and not the regret of the test phase (i.e. the cumulative regret)
- **Computer Go:** Given a limited CPU time and a goban position, output the next action to play.
  Idea: This problem is a bandit game where actions = arms and round = evaluation (costly in CPU time) of an action. One wants to minimize the simple regret of the selected action once the budget is exhausted.

Strategies

Notation: $T_i(t)$ is the number of times we pulled arm i up to time t and $\widehat{\mu}_{i,s}$ is the empirical average of rewards for arm i after s pulls of this arm.

- **Uniform forecaster:** pulls each arm one after an other and recommend the arm with the highest empirical mean.
- **UCB(p), [Auer et al 02]:** pulls at round t the arm with the highest upper confidence bound

$$\widehat{\mu}_{i,T_i(t-1)} + \sqrt{\frac{p \log(t)}{T_i(t-1)}}$$

and recommend the most played arm.

Strategies

Notation: $T_i(t)$ is the number of times we pulled arm i up to time t and $\widehat{\mu}_{i,s}$ is the empirical average of rewards for arm i after s pulls of this arm.

- **Uniform forecaster:** pulls each arm one after an other and recommend the arm with the highest empirical mean.
- **UCB(p), [Auer et al 02]:** pulls at round t the arm with the highest upper confidence bound

$$\widehat{\mu}_{i,T_i(t-1)} + \sqrt{\frac{p \log(t)}{T_i(t-1)}}$$

and recommend the most played arm.

Main Result: the smaller R_n the larger r_n !

Theorem (Main result)

Let $\epsilon : \{1, 2, \ldots\} \to \mathbb{R}$ *be such that for all (Bernoulli) distributions* $\nu_1, \ldots, \nu_K$ *on the rewards, there exists a constant* $C \geq 0$ *with*

$$\mathbb{E} R_n \leq C \epsilon(n),$$

then for all sets of $K \geq 3$ *(distinct, Bernoulli) distributions on the rewards, all different from a Dirac distribution at* 1*, there exists two constants* $D, D' > 0$ *and an ordering* $\nu_1, \ldots, \nu_K$ *of the considered distributions with*

$$\mathbb{E} r_n \geq D \, e^{-D' \epsilon(n)} \ .$$

Consequences of the main result

Corollary

For all sets of $K \geq 3$ (distinct, Bernoulli) distributions on the rewards, all different from a Dirac distribution at 1, there exists two constants $D, D' > 0$ and an ordering $\nu_1, \ldots, \nu_K$ of the considered distributions with

$$\mathbb{E} r_n \geq D\, e^{-D' n} \, .$$

Remark: As we will see shortly, the basic uniform forecaster has an exponential rate of decrease for the simple regret.

Consequences of the main result

Corollary

For all sets of $K \geq 3$ *(distinct, Bernoulli) distributions on the rewards, all different from a Dirac distribution at* 1*, there exists two constants* $D, D' > 0$ *and an ordering* $\nu_1, \ldots, \nu_K$ *of the considered distributions with*

$$\mathbb{E}r_n \geq D\, e^{-D' n} \, .$$

Remark: As we will see shortly, the basic uniform forecaster has an exponential rate of decrease for the simple regret.

Consequences of the main result

Reminder: the optimal rate of growth for the cumulative regret is logarithmic. For instance UCB(p) satisfies for any distributions a regret bound of the form $\mathbb{E}R_n \leq C \log(n)$.

Corollary

For all sets of $K \geq 3$ (distinct, Bernoulli) distributions on the rewards, all different from a Dirac distribution at 1, there exists two constants $D, D' > 0$ and an ordering $\nu_1, \ldots, \nu_K$ of the considered distributions such that the simple regret of UCB(p) satisfies

$$\mathbb{E}r_n \geq D\, n^{-D'} .$$

Consequences of the main result

Reminder: the optimal rate of growth for the cumulative regret is logarithmic. For instance UCB(p) satisfies for any distributions a regret bound of the form $\mathbb{E}R_n \leq C \log(n)$.

Corollary

For all sets of $K \geq 3$ (distinct, Bernoulli) distributions on the rewards, all different from a Dirac distribution at 1, there exists two constants $D, D' > 0$ and an ordering $\nu_1, \ldots, \nu_K$ of the considered distributions such that the simple regret of UCB(p) satisfies

$$\mathbb{E}r_n \geq D\, n^{-D'} \ .$$

Conclusion at this point

- Optimal forecasters for the cumulative regret (*i.e.*, $\mathbb{E}R_n \sim \log n$) are suboptimal for the simple regret (*i.e.*, $\mathbb{E}r_n \sim n^{-D}$).
- Basic forecasters outperform famous strategies like UCB (since for the uniform strategy $\mathbb{E}r_n \sim \exp(-Dn)$).
- Is this conclusion still valid in "finite time" ? More precisely, what is the form of the distribution-dependent constants D ?

Conclusion at this point

- Optimal forecasters for the cumulative regret (*i.e.*, $\mathbb{E}R_n \sim \log n$) are suboptimal for the simple regret (*i.e.*, $\mathbb{E}r_n \sim n^{-D}$).
- Basic forecasters outperform famous strategies like UCB (since for the uniform strategy $\mathbb{E}r_n \sim \exp(-Dn)$).
- Is this conclusion still valid in "finite time" ? More precisely, what is the form of the distribution-dependent constants D ?

Conclusion at this point

- Optimal forecasters for the cumulative regret (*i.e.*, $\mathbb{E}R_n \sim \log n$) are suboptimal for the simple regret (*i.e.*, $\mathbb{E}r_n \sim n^{-D}$).
- Basic forecasters outperform famous strategies like UCB (since for the uniform strategy $\mathbb{E}r_n \sim \exp(-Dn)$).
- Is this conclusion still valid in "finite time" ? More precisely, what is the form of the distribution-dependent constants D ?

Precise distribution-dependent rate

Notation: $\Delta_i = \mu^* - \mu_i$.

Theorem

The uniform exploration satisfies:

$$\mathbb{E} r_n \leq \sum_{j:\Delta_j>0} \Delta_j \, e^{-\frac{n\Delta_j^2}{2K}} .$$

Theorem

For $p > 1$, UCB(p) satisfies:

$$\mathbb{E} r_n \leq \frac{K^{2p-1}}{p-1} \left(\frac{1}{n}\right)^{2(p-1)}$$

for all $n \geq \max\left(K + \frac{4Kp\ln n}{\Delta^2}, K(K+2)\right)$.

Precise distribution-dependent rate

Notation: $\Delta_i = \mu^* - \mu_i$.

Theorem

The uniform exploration satisfies:

$$\mathbb{E} r_n \leq \sum_{j:\Delta_j>0} \Delta_j \, e^{-\frac{n\Delta_j^2}{2K}} .$$

Theorem

For $p > 1$, UCB(p) satisfies:

$$\mathbb{E} r_n \leq \frac{K^{2p-1}}{p-1} \left(\frac{1}{n}\right)^{2(p-1)}$$

for all $n \geq \max\left(K + \frac{4Kp\ln n}{\Delta^2}, K(K+2)\right)$.

Precise distribution-dependent rate

Notation: $\Delta_i = \mu^* - \mu_i$.

Theorem

The uniform exploration satisfies:

$$\mathbb{E} r_n \leq \sum_{j:\Delta_j > 0} \Delta_j \, e^{-\frac{n\Delta_j^2}{2K}}.$$

Theorem

For $p > 1$, UCB(p) satisfies:

$$\mathbb{E} r_n \leq \frac{K^{2p-1}}{p-1} \left(\frac{1}{n}\right)^{2(p-1)}$$

for all $n \geq \max\left(K + \frac{4Kp \ln n}{\Delta^2}, K(K+2)\right)$.

Distribution-free analysis

Theorem

The uniform exploration satisfies: $\mathbb{E} r_n \leq 2\sqrt{\frac{2K\log(K)}{n}}$.

Theorem

For $p > 1$, *UCB*(p) *satisfies:* $\mathbb{E} r_n \leq \sqrt{\frac{4pK\log(n) + \left(\frac{3}{2} + \frac{1}{2(p-1)}\right)}{n}}$.

Theorem

For any forecaster and any time n *there exists a set of distributions such that* $\mathbb{E} r_n \geq \frac{1}{20}\sqrt{\frac{K}{n}}$.

Distribution-free analysis

Theorem

The uniform exploration satisfies: $\mathbb{E} r_n \leq 2\sqrt{\frac{2K\log(K)}{n}}$.

Theorem

For $p > 1$, *UCB*(p) *satisfies:* $\mathbb{E} r_n \leq \sqrt{\frac{4pK\log(n) + \left(\frac{3}{2} + \frac{1}{2(p-1)}\right)}{n}}$.

Theorem

For any forecaster and any time n *there exists a set of distributions such that* $\mathbb{E} r_n \geq \frac{1}{20}\sqrt{\frac{K}{n}}$.

Distribution-free analysis

Theorem

The uniform exploration satisfies: $\mathbb{E} r_n \leq 2\sqrt{\frac{2K\log(K)}{n}}$.

Theorem

For $p > 1$, *UCB*(p) *satisfies:* $\mathbb{E} r_n \leq \sqrt{\frac{4pK\log(n) + \left(\frac{3}{2} + \frac{1}{2(p-1)}\right)}{n}}$.

Theorem

For any forecaster and any time n *there exists a set of distributions such that* $\mathbb{E} r_n \geq \frac{1}{20}\sqrt{\frac{K}{n}}$.

Conclusion and ongoing work

- Different regimes:
  - Asymptotically $\mathbb{E}r_n(\text{uniform}) << \mathbb{E}r_n(\text{strategy with low } \mathbb{E}R_n)$.
  - Finite time $\mathbb{E}r_n(UCB(p)) << \mathbb{E}r_n(\text{uniform})$ (for some distributions).
- New algorithms using the insights gained from the present analysis.
- Optimal distribution-dependent rate for the simple regret.

Conclusion and ongoing work

- Different regimes:
  - Asymptotically $\mathbb{E}r_n(\text{uniform}) << \mathbb{E}r_n(\text{strategy with low } \mathbb{E}R_n)$.
  - Finite time $\mathbb{E}r_n(UCB(p)) << \mathbb{E}r_n(\text{uniform})$ (for some distributions).
- New algorithms using the insights gained from the present analysis.
- Optimal distribution-dependent rate for the simple regret.

Conclusion and ongoing work

- Different regimes:
  - Asymptotically $\mathbb{E}r_n(\text{uniform}) << \mathbb{E}r_n(\text{strategy with low } \mathbb{E}R_n)$.
  - Finite time $\mathbb{E}r_n(UCB(p)) << \mathbb{E}r_n(\text{uniform})$ (for some distributions).
- New algorithms using the insights gained from the present analysis.
- Optimal distribution-dependent rate for the simple regret.