Budget-Constrained Multi-Armed Bandits with Multiple Plays

Datong P. Zhou, Claire J. Tomlin

AAAI-2018

[datong.zhou, tomlin]@berkeley.edu

February 5, 2018

Table of Contents

Table of Contents

Background

Examples

- Product recommendation to maximize sales
- Ad placement to maximize click through rate

Classical Problem and Objective

- Given N arms with unknown reward distribution
- Pull arms sequentially to maximize total expected reward over certain horizon T
- At each round $t \in [T]$:
  - Player selects *exactly one* action a_t
  - Player observes gain $r_{a_t,t}$
- Goal: Minimize cumulative regret

$$\mathcal{R}_{\mathcal{A}}(T) = \left(\max_{i \in [N]} \mathbb{E}\left[\sum_{t=1}^{T} r_{i,t} \right] \right) - \mathbb{E}\left[\sum_{t=1}^{T} r_{a_t,t} \right]$$

- Stochastic generation of rewards vs. adversarial setting
- Recent developments: Cost $c_{i,t}$ and multiple plays K

Background

Examples

- Product recommendation to maximize sales
- Ad placement to maximize click through rate

Classical Problem and Objective

- Given N arms with unknown reward distribution
- Pull arms sequentially to maximize total expected reward over certain horizon T
- At each round $t \in [T]$:
  - Player selects *exactly one* action a_t
  - Player observes gain $r_{a_t,t}$
- Goal: Minimize cumulative regret

$$\mathcal{R}_{\mathcal{A}}(T) = \left(\max_{i \in [N]} \mathbb{E} \left[\sum_{t=1}^{T} r_{i,t} \right] \right) - \mathbb{E} \left[\sum_{t=1}^{T} r_{a_t,t} \right]$$

- Stochastic generation of rewards vs. adversarial setting
- Recent developments: Cost $c_{i,t}$ and multiple plays K

Background

Examples

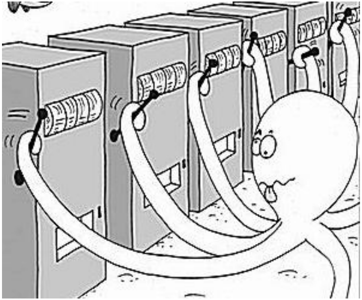
- Product recommendation to maximize sales
- Ad placement to maximize click through rate

Classical Problem and Objective

- Given N arms with unknown reward distribution
- Pull arms sequentially to maximize total expected reward over certain horizon T
- At each round $t \in [T]$:
  - Player selects *exactly one* action a_t
  - Player observes gain $r_{a_t,t}$
- Goal: Minimize cumulative regret

$$\mathcal{R}_{\mathcal{A}}(T) = \left(\max_{i \in [N]} \mathbb{E}\left[\sum_{t=1}^{T} r_{i,t} \right] \right) - \mathbb{E}\left[\sum_{t=1}^{T} r_{a_t,t} \right]$$

- Stochastic generation of rewards vs. adversarial setting
- Recent developments: Cost $c_{i,t}$ and multiple plays K



Background

Examples

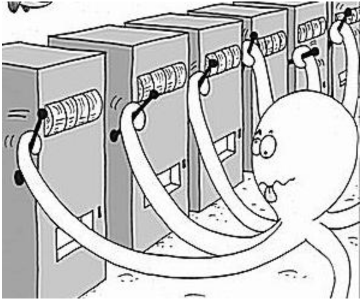
- Product recommendation to maximize sales
- Ad placement to maximize click through rate

Classical Problem and Objective

- Given N arms with unknown reward distribution
- Pull arms sequentially to maximize total expected reward over certain horizon T
- At each round $t \in [T]$:
  - Player selects *exactly one* action a_t
  - Player observes gain $r_{a_t,t}$
- Goal: Minimize cumulative regret

$$\mathcal{R}_{\mathcal{A}}(T) = \left(\max_{i \in [N]} \mathbb{E}\left[\sum_{t=1}^{T} r_{i,t} \right] \right) - \mathbb{E}\left[\sum_{t=1}^{T} r_{a_t,t} \right]$$

- Stochastic generation of rewards vs. adversarial setting
- Recent developments: Cost $c_{i,t}$ and multiple plays K



Background

Examples

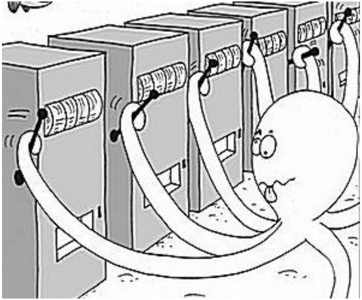
- Product recommendation to maximize sales
- Ad placement to maximize click through rate

Classical Problem and Objective

- Given N arms with unknown reward distribution
- Pull arms sequentially to maximize total expected reward over certain horizon T
- At each round $t \in [T]$:
  - Player selects *exactly one* action a_t
  - Player observes gain $r_{a_t,t}$
- Goal: Minimize cumulative regret

$$\mathcal{R}_{\mathcal{A}}(T) = \left(\max_{i \in [N]} \mathbb{E}\left[\sum_{t=1}^{T} r_{i,t} \right] \right) - \mathbb{E}\left[\sum_{t=1}^{T} r_{a_t,t} \right]$$

- Stochastic generation of rewards vs. adversarial setting
- Recent developments: Cost $c_{i,t}$ and multiple plays K



Background

Examples

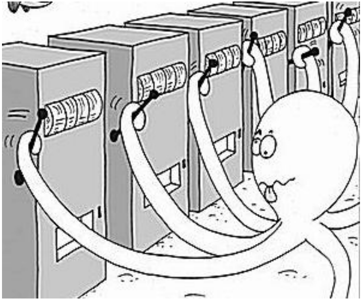
- Product recommendation to maximize sales
- Ad placement to maximize click through rate

Classical Problem and Objective

- Given N arms with unknown reward distribution
- Pull arms sequentially to maximize total expected reward over certain horizon T
- At each round $t \in [T]$:
  - Player selects *exactly one* action a_t
  - Player observes gain $r_{a_t,t}$
- Goal: Minimize cumulative regret

$$\mathcal{R}_{\mathcal{A}}(T) = \left(\max_{i \in [N]} \mathbb{E}\left[\sum_{t=1}^{T} r_{i,t} \right] \right) - \mathbb{E}\left[\sum_{t=1}^{T} r_{a_t,t} \right]$$

- Stochastic generation of rewards vs. adversarial setting
- Recent developments: Cost $c_{i,t}$ and multiple plays K



Background

Examples

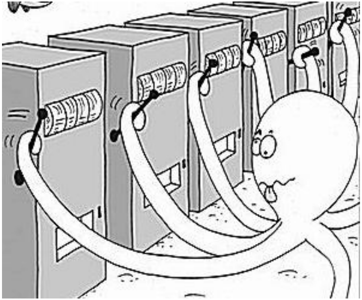
- Product recommendation to maximize sales
- Ad placement to maximize click through rate

Classical Problem and Objective

- Given N arms with unknown reward distribution
- Pull arms sequentially to maximize total expected reward over certain horizon T
- At each round $t \in [T]$:
  - Player selects *exactly one* action a_t
  - Player observes gain $r_{a_t,t}$
- Goal: Minimize cumulative regret

$$\mathcal{R}_{\mathcal{A}}(T) = \left(\max_{i \in [N]} \mathbb{E}\left[\sum_{t=1}^{T} r_{i,t} \right] \right) - \mathbb{E}\left[\sum_{t=1}^{T} r_{a_t,t} \right]$$

- Stochastic generation of rewards vs. adversarial setting
- Recent developments: Cost $c_{i,t}$ and multiple plays K



Contributions

Algorithm	Upper Bound	Lower Bound	Setting	Authors
`Exp3`	$O\left(\sqrt{NT\log N}\right)$	$\Omega\left(\sqrt{NT}\right)$	Fixed T, $K=1$	[1]
`Exp3.M`	$O\left(\sqrt{NTK\log\frac{N}{K}}\right)$	$\Omega\left(\left(1-\frac{K}{N}\right)^2\sqrt{NT}\right)$	Fixed T, $K\geq 1$	[2]
`Exp3.M.B`	$O\left(\sqrt{NB\log\frac{N}{K}}\right)$	$\Omega\left(\left(1-\frac{K}{N}\right)^2\sqrt{NB/K}\right)$	$B>0, K\geq 1$	This paper
`Exp3.P`	$O\left(\sqrt{NT\log\left(NT/\delta\right)}+\log(NT/\delta)\right)$		Fixed T, $K=1$	[3]
`Exp3.P.M`	$O\left(K^2\sqrt{NT\frac{N-K}{N-1}\log\left(NT/\delta\right)}+\frac{N-K}{N-1}\log(NT/\delta)\right)$		Fixed T, $K\geq 1$	This paper
`Exp3.P.M.B`	$O\left(K^2\sqrt{\frac{NB}{K}\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)}+\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)\right)$		$B>0$, $K\geq 1$	This paper
`UCB1`	$O(N\log T)$		Fixed T, $K=1$	[4]
`LLR`	$O(NK^4\log T)$		Fixed T, $K\geq 1$	[5]
`UCB-BV`	$O(N\log B)$		$B>0, K=1$	[6]
`UCB-MB`	$O(NK^4\log B)$		$B>0, K\geq 1$	This paper

[1] P. Auer et al. "The Nonstochastic Multi-Armed Bandit Problem". In: *SIAM Journal on Computing* 32 (2002), pp. 48–77.

[2] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

[3] P. Auer et al. "The Nonstochastic Multi-Armed Bandit Problem". In: *SIAM Journal on Computing* 32 (2002), pp. 48–77.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

[5] Y. Gai, B. Krishnamachari, and R. Jain. "Combinatorial Network Optimization with Unknown Variables: Multi-Armed Bandits with Linear Rewards and Individual Observations". In: *IEEE/ACM Transactions on Networking* 20.5 (2012), pp. 1466–1478.

[6] W. Ding et al. "Multi-Armed Bandit with Budget Constraint and Variable Costs". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013).

Table of Contents

Setup

Problem Description

- Bandit with N distinct arms
- Each arm i has *unknown* cost and reward distributions with means $0 < \mu_r^i \leq 1$ and $0 < c_{\min} \leq \mu_c^i \leq 1$
- Realizations of costs $c_{i,t} \in [c_{\min}, 1]$ and rewards $r_{i,t} \in [0, 1]$ are i.i.d.
- Initial budget $B > 0$ to pay for the materialized costs
- At each round $t = 1, \ldots, \tau_{\mathcal{A}}(B)$:
  - Select exactly $1 \leq K \leq N$ arms into a_t
  - Observe individual costs $\{c_{i,t} \mid i \in a_t\}$ and rewards $\{r_{i,t} \mid i \in a_t\}$
  - Terminate game if $\sum_{i \in a_t} c_{i,t}$ is greater than remaining budget

Goal

- Minimize expected regret

$$\mathcal{R}_{\mathcal{A}}(B) = \mathbb{E}[G_{\mathcal{A}^*}(B)] - \mathbb{E}[G_{\mathcal{A}}(B)]$$

- Utilize modified UCB algorithm with upper confidence bounds:

$$U_{i,t} = \bar{\mu}_t^i + e_{i,t}$$

- At each round, play K arms with K largest $U_{i,t}$

Setup

Problem Description

- Bandit with N distinct arms
- Each arm i has *unknown* cost and reward distributions with means $0 < \mu_r^i \leq 1$ and $0 < c_{\min} \leq \mu_c^i \leq 1$
- Realizations of costs $c_{i,t} \in [c_{\min}, 1]$ and rewards $r_{i,t} \in [0, 1]$ are i.i.d.
- Initial budget $B > 0$ to pay for the materialized costs
- At each round $t = 1, \ldots, \tau_{\mathcal{A}}(B)$:
  - Select exactly $1 \leq K \leq N$ arms into a_t
  - Observe individual costs $\{c_{i,t} \mid i \in a_t\}$ and rewards $\{r_{i,t} \mid i \in a_t\}$
  - Terminate game if $\sum_{i \in a_t} c_{i,t}$ is greater than remaining budget

Goal

- Minimize expected regret

$$\mathcal{R}_{\mathcal{A}}(B) = \mathbb{E}[G_{\mathcal{A}^*}(B)] - \mathbb{E}[G_{\mathcal{A}}(B)]$$

- Utilize modified UCB algorithm with upper confidence bounds:

$$U_{i,t} = \bar{\mu}_t^i + e_{i,t}$$

- At each round, play K arms with K largest $U_{i,t}$

Setup

Problem Description

- Bandit with N distinct arms
- Each arm i has *unknown* cost and reward distributions with means $0 < \mu_r^i \leq 1$ and $0 < c_{\min} \leq \mu_c^i \leq 1$
- Realizations of costs $c_{i,t} \in [c_{\min}, 1]$ and rewards $r_{i,t} \in [0, 1]$ are i.i.d.
- Initial budget $B > 0$ to pay for the materialized costs
- At each round $t = 1, \ldots, \tau_{\mathcal{A}}(B)$:
  - Select exactly $1 \leq K \leq N$ arms into a_t
  - Observe individual costs $\{c_{i,t} \mid i \in a_t\}$ and rewards $\{r_{i,t} \mid i \in a_t\}$
  - Terminate game if $\sum_{i \in a_t} c_{i,t}$ is greater than remaining budget

Goal

- Minimize expected regret

$$\mathcal{R}_{\mathcal{A}}(B) = \mathbb{E}[G_{\mathcal{A}^*}(B)] - \mathbb{E}[G_{\mathcal{A}}(B)]$$

- Utilize modified UCB algorithm with upper confidence bounds:

$$U_{i,t} = \bar{\mu}_t^i + e_{i,t}$$

- At each round, play K arms with K largest $U_{i,t}$

Setup

Problem Description

- Bandit with N distinct arms
- Each arm i has *unknown* cost and reward distributions with means $0 < \mu_r^i \leq 1$ and $0 < c_{\min} \leq \mu_c^i \leq 1$
- Realizations of costs $c_{i,t} \in [c_{\min}, 1]$ and rewards $r_{i,t} \in [0, 1]$ are i.i.d.
- Initial budget $B > 0$ to pay for the materialized costs
- At each round $t = 1, \ldots, \tau_{\mathcal{A}}(B)$:
  - Select exactly $1 \leq K \leq N$ arms into a_t
  - Observe individual costs $\{c_{i,t} \mid i \in a_t\}$ and rewards $\{r_{i,t} \mid i \in a_t\}$
  - Terminate game if $\sum_{i \in a_t} c_{i,t}$ is greater than remaining budget

Goal

- Minimize expected regret

$$\mathcal{R}_{\mathcal{A}}(B) = \mathbb{E}[G_{\mathcal{A}^*}(B)] - \mathbb{E}[G_{\mathcal{A}}(B)]$$

- Utilize modified UCB algorithm with upper confidence bounds:

$$U_{i,t} = \bar{\mu}_t^i + e_{i,t}$$

- At each round, play K arms with K largest $U_{i,t}$

Setup

Problem Description

- Bandit with N distinct arms
- Each arm i has *unknown* cost and reward distributions with means $0 < \mu_r^i \leq 1$ and $0 < c_{\min} \leq \mu_c^i \leq 1$
- Realizations of costs $c_{i,t} \in [c_{\min}, 1]$ and rewards $r_{i,t} \in [0, 1]$ are i.i.d.
- Initial budget $B > 0$ to pay for the materialized costs
- At each round $t = 1, \ldots, \tau_{\mathcal{A}}(B)$:
  - Select exactly $1 \leq K \leq N$ arms into a_t
  - Observe individual costs $\{c_{i,t} \mid i \in a_t\}$ and rewards $\{r_{i,t} \mid i \in a_t\}$
  - Terminate game if $\sum_{i \in a_t} c_{i,t}$ is greater than remaining budget

Goal

- Minimize expected regret

$$\mathcal{R}_{\mathcal{A}}(B) = \mathbb{E}[G_{\mathcal{A}^*}(B)] - \mathbb{E}[G_{\mathcal{A}}(B)]$$

- Utilize modified UCB algorithm with upper confidence bounds:

$$U_{i,t} = \bar{\mu}_t^i + e_{i,t}$$

- At each round, play K arms with K largest $U_{i,t}$

Setup

Problem Description

- Bandit with N distinct arms
- Each arm i has *unknown* cost and reward distributions with means $0 < \mu_r^i \leq 1$ and $0 < c_{\min} \leq \mu_c^i \leq 1$
- Realizations of costs $c_{i,t} \in [c_{\min}, 1]$ and rewards $r_{i,t} \in [0, 1]$ are i.i.d.
- Initial budget $B > 0$ to pay for the materialized costs
- At each round $t = 1, \ldots, \tau_{\mathcal{A}}(B)$:
  - Select exactly $1 \leq K \leq N$ arms into a_t
  - Observe individual costs $\{c_{i,t} \mid i \in a_t\}$ and rewards $\{r_{i,t} \mid i \in a_t\}$
  - Terminate game if $\sum_{i \in a_t} c_{i,t}$ is greater than remaining budget

Goal

- Minimize expected regret

$$\mathcal{R}_{\mathcal{A}}(B) = \mathbb{E}[G_{\mathcal{A}^*}(B)] - \mathbb{E}[G_{\mathcal{A}}(B)]$$

- Utilize modified UCB algorithm with upper confidence bounds:

$$U_{i,t} = \bar{\mu}_t^i + e_{i,t}$$

- At each round, play K arms with K largest $U_{i,t}$

Setup

Problem Description

- Bandit with N distinct arms
- Each arm i has *unknown* cost and reward distributions with means $0 < \mu_r^i \leq 1$ and $0 < c_{\min} \leq \mu_c^i \leq 1$
- Realizations of costs $c_{i,t} \in [c_{\min}, 1]$ and rewards $r_{i,t} \in [0, 1]$ are i.i.d.
- Initial budget $B > 0$ to pay for the materialized costs
- At each round $t = 1, \ldots, \tau_{\mathcal{A}}(B)$:
    - Select exactly $1 \leq K \leq N$ arms into a_t
    - Observe individual costs $\{c_{i,t} \mid i \in a_t\}$ and rewards $\{r_{i,t} \mid i \in a_t\}$
    - Terminate game if $\sum_{i \in a_t} c_{i,t}$ is greater than remaining budget

Goal

- Minimize expected regret

$$\mathcal{R}_{\mathcal{A}}(B) = \mathbb{E}[G_{\mathcal{A}^*}(B)] - \mathbb{E}[G_{\mathcal{A}}(B)]$$

- Utilize modified UCB algorithm with upper confidence bounds:

$$U_{i,t} = \bar{\mu}_t^i + e_{i,t}$$

- At each round, play K arms with K largest $U_{i,t}$

Algorithm UCB-MB

The Algorithm

- Play all arms once to initialize bang-per-buck ratios for each arm
- While B not exhausted, select K arms with K largest $U_{i,t}$

Theorem (Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm UCB-MB)

For the definition of confidence bounds

$$U_{i,t} = \bar{\mu}_t^i + \frac{\sqrt{(K+1)\log t/n_{i,t}}(1+1/c_{\min})}{c_{\min} - \sqrt{(K+1)\log t/n_{i,t}}},$$

Algorithm UCB-MB achieves expected regret $\mathcal{R}_{\mathcal{A}}(B) = O(NK^4 \log B)$.

Proof Idea

- Step 1: Upper bound the number of times a non-optimal selection of arms is made up to a *fixed* stopping time $\tau_{\mathcal{A}}(B)$:

$$\# \text{ suboptimal choices} = O(NK^3 \log \tau_{\mathcal{A}}(B))$$

- Step 2: Relate algorithm UCB-MB and B to stopping time $\tau_{\mathcal{A}}(B)$

Algorithm UCB-MB

The Algorithm

- Play all arms once to initialize bang-per-buck ratios for each arm
- While B not exhausted, select K arms with K largest $U_{i,t}$

Theorem (Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm UCB-MB)

For the definition of confidence bounds

$$U_{i,t} = \bar{\mu}_t^i + \frac{\sqrt{(K+1)\log t/n_{i,t}}(1+1/c_{\min})}{c_{\min} - \sqrt{(K+1)\log t/n_{i,t}}},$$

Algorithm UCB-MB achieves expected regret $\mathcal{R}_{\mathcal{A}}(B) = O(NK^4 \log B)$.

Proof Idea

- Step 1: Upper bound the number of times a non-optimal selection of arms is made up to a *fixed* stopping time $\tau_{\mathcal{A}}(B)$:

$$\#\text{ suboptimal choices} = O(NK^3 \log \tau_{\mathcal{A}}(B))$$

- Step 2: Relate algorithm UCB-MB and B to stopping time $\tau_{\mathcal{A}}(B)$

Algorithm UCB-MB

The Algorithm

- Play all arms once to initialize bang-per-buck ratios for each arm
- While B not exhausted, select K arms with K largest $U_{i,t}$

Theorem (Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm UCB-MB)

For the definition of confidence bounds

$$U_{i,t} = \bar{\mu}_t^i + \frac{\sqrt{(K+1)\log t/n_{i,t}}(1+1/c_{\min})}{c_{\min} - \sqrt{(K+1)\log t/n_{i,t}}},$$

Algorithm UCB-MB achieves expected regret $\mathcal{R}_{\mathcal{A}}(B) = O(NK^4 \log B)$.

Proof Idea

- Step 1: Upper bound the number of times a non-optimal selection of arms is made up to a *fixed* stopping time $\tau_{\mathcal{A}}(B)$:

$$\# \text{ suboptimal choices} = O(NK^3 \log \tau_{\mathcal{A}}(B))$$

- Step 2: Relate algorithm UCB-MB and B to stopping time $\tau_{\mathcal{A}}(B)$

Algorithm UCB-MB

The Algorithm

- Play all arms once to initialize bang-per-buck ratios for each arm
- While B not exhausted, select K arms with K largest $U_{i,t}$

Theorem (Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm UCB-MB)

For the definition of confidence bounds

$$U_{i,t} = \bar{\mu}_t^i + \frac{\sqrt{(K+1)\log t/n_{i,t}}(1+1/c_{\min})}{c_{\min} - \sqrt{(K+1)\log t/n_{i,t}}},$$

Algorithm UCB-MB achieves expected regret $\mathcal{R}_{\mathcal{A}}(B) = O(NK^4 \log B)$.

Proof Idea

- Step 1: Upper bound the number of times a non-optimal selection of arms is made up to a *fixed* stopping time $\tau_{\mathcal{A}}(B)$:

$$\#\text{ suboptimal choices} = O(NK^3 \log \tau_{\mathcal{A}}(B))$$

- Step 2: Relate algorithm UCB-MB and B to stopping time $\tau_{\mathcal{A}}(B)$

Table of Contents

Upper Bound on the Regret

Setup

- Oblivious adversary $\Rightarrow$ no assumptions on reward or cost distributions except boundedness

Algorithm `Exp3.M.B`

- Initialize weights $w_i = 1$ for $i = 1, \dots, N$
- For each round $t = 1, \dots, \tau_{\mathcal{A}}(B)$:
  - Cap weights that are "too large"

$$w_i(t) = v(t) \text{ for } i \in \tilde{S}(t) = \{i \in [N] \mid w_i(t) > v_t\}$$

$$v(t) \leftarrow \left\{ v_t \;\middle|\; \frac{v_t(1-\gamma)}{\sum_{i=1}^{N} v_t \cdot \mathbb{1}(w_i(t) \geq v_t) + w_i(t) \cdot \mathbb{1}(w_i(t) < v_t)} = \frac{1}{K} - \frac{\gamma}{N} \right\}$$

  - Calculate probabilities $p_i(t) = K \left((1-\gamma) \frac{\tilde{w}_i(t)}{\sum_{j=1}^{N} \tilde{w}_j(t)} + \frac{\gamma}{N} \right)$
  - Play arms $a_t \sim p_1, \dots, p_N$
  - Update weights:

$$\hat{r}_i(t) = r_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$
$$\hat{c}_i(t) = c_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$
$$w_i(t+1) = w_i(t) \exp\left[\frac{K\gamma}{N} \left[\hat{r}_i(t) - \hat{c}_i(t) \right] \mathbb{1}_{i \in \tilde{S}(t)} \right]$$

Upper Bound on the Regret

Setup

- Oblivious adversary $\Rightarrow$ no assumptions on reward or cost distributions except boundedness

Algorithm `Exp3.M.B`

- Initialize weights $w_i = 1$ for $i = 1, \ldots, N$
- For each round $t = 1, \ldots, \tau_{\mathcal{A}}(B)$:
  - Cap weights that are "too large"

$$w_i(t) = v(t) \text{ for } i \in \tilde{S}(t) = \{i \in [N] \mid w_i(t) > v_t\}$$

$$v(t) \leftarrow \left\{ v_t \;\middle|\; \frac{v_t(1-\gamma)}{\sum_{i=1}^{N} v_t \cdot \mathbb{1}(w_i(t) \geq v_t) + w_i(t) \cdot \mathbb{1}(w_i(t) < v_t)} = \frac{1}{K} - \frac{\gamma}{N} \right\}$$

  - Calculate probabilities $p_i(t) = K\left((1-\gamma)\frac{\tilde{w}_i(t)}{\sum_{j=1}^{N} \tilde{w}_j(t)} + \frac{\gamma}{N}\right)$
  - Play arms $a_t \sim p_1, \ldots, p_N$
  - Update weights:

$$\hat{r}_i(t) = r_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$
$$\hat{c}_i(t) = c_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$
$$w_i(t+1) = w_i(t) \exp\left[\frac{K\gamma}{N}\left[\hat{r}_i(t) - \hat{c}_i(t)\right] \mathbb{1}_{i \in \tilde{S}(t)}\right]$$

Upper Bound on the Regret

Setup

- Oblivious adversary $\Rightarrow$ no assumptions on reward or cost distributions except boundedness

Algorithm `Exp3.M.B`

- Initialize weights $w_i = 1$ for $i = 1, \dots, N$
- For each round $t = 1, \dots, \tau_{\mathcal{A}}(B)$:
  - Cap weights that are "too large"

$$w_i(t) = v(t) \text{ for } i \in \tilde{S}(t) = \{i \in [N] \mid w_i(t) > v_t\}$$

$$v(t) \leftarrow \left\{ v_t \;\middle|\; \frac{v_t(1-\gamma)}{\sum_{i=1}^N v_t \cdot \mathbb{1}(w_i(t) \geq v_t) + w_i(t) \cdot \mathbb{1}(w_i(t) < v_t)} = \frac{1}{K} - \frac{\gamma}{N} \right\}$$

  - Calculate probabilities $p_i(t) = K\left((1-\gamma)\frac{\tilde{w}_i(t)}{\sum_{j=1}^N \tilde{w}_j(t)} + \frac{\gamma}{N}\right)$
  - Play arms $a_t \sim p_1, \dots, p_N$
  - Update weights:

$$\hat{r}_i(t) = r_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$

$$\hat{c}_i(t) = c_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$

$$w_i(t+1) = w_i(t) \exp\left[\frac{K\gamma}{N}\left[\hat{r}_i(t) - \hat{c}_i(t)\right] \mathbb{1}_{i \in \tilde{S}(t)}\right]$$

Upper Bound on the Regret

Setup

- Oblivious adversary $\Rightarrow$ no assumptions on reward or cost distributions except boundedness

Algorithm `Exp3.M.B`

- Initialize weights $w_i = 1$ for $i = 1, \ldots, N$
- For each round $t = 1, \ldots, \tau_{\mathcal{A}}(B)$:
  - Cap weights that are "too large"

$$w_i(t) = v(t) \text{ for } i \in \tilde{S}(t) = \{i \in [N] \mid w_i(t) > v_t\}$$

$$v(t) \leftarrow \left\{ v_t \;\middle|\; \frac{v_t(1-\gamma)}{\sum_{i=1}^{N} v_t \cdot \mathbb{1}(w_i(t) \geq v_t) + w_i(t) \cdot \mathbb{1}(w_i(t) < v_t)} = \frac{1}{K} - \frac{\gamma}{N} \right\}$$

  - Calculate probabilities $p_i(t) = K\left((1-\gamma)\frac{\tilde{w}_i(t)}{\sum_{j=1}^{N} \tilde{w}_j(t)} + \frac{\gamma}{N}\right)$
  - Play arms $a_t \sim p_1, \ldots, p_N$
  - Update weights:

$$\hat{r}_i(t) = r_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$

$$\hat{c}_i(t) = c_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$

$$w_i(t+1) = w_i(t) \exp\left[\frac{K\gamma}{N}\left[\hat{r}_i(t) - \hat{c}_i(t)\right] \mathbb{1}_{i \in \tilde{S}(t)}\right]$$

Upper Bound on the Regret

Setup

- Oblivious adversary $\Rightarrow$ no assumptions on reward or cost distributions except boundedness

Algorithm `Exp3.M.B`

- Initialize weights $w_i = 1$ for $i = 1, \dots, N$
- For each round $t = 1, \dots, \tau_{\mathcal{A}}(B)$:
  - Cap weights that are "too large"

$$w_i(t) = v(t) \text{ for } i \in \tilde{S}(t) = \{i \in [N] \mid w_i(t) > v_t\}$$

$$v(t) \leftarrow \left\{ v_t \;\middle|\; \frac{v_t(1-\gamma)}{\sum_{i=1}^N v_t \cdot \mathbb{1}(w_i(t) \geq v_t) + w_i(t) \cdot \mathbb{1}(w_i(t) < v_t)} = \frac{1}{K} - \frac{\gamma}{N} \right\}$$

  - Calculate probabilities $p_i(t) = K\left((1-\gamma)\frac{\tilde{w}_i(t)}{\sum_{j=1}^N \tilde{w}_j(t)} + \frac{\gamma}{N}\right)$
  - Play arms $a_t \sim p_1, \dots, p_N$
  - Update weights:

$$\hat{r}_i(t) = r_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$
$$\hat{c}_i(t) = c_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$
$$w_i(t+1) = w_i(t) \exp\left[\frac{K\gamma}{N}\left[\hat{r}_i(t) - \hat{c}_i(t)\right]\mathbb{1}_{i \in \tilde{S}(t)}\right]$$

Upper Bound on the Regret

Setup

- Oblivious adversary $\Rightarrow$ no assumptions on reward or cost distributions except boundedness

Algorithm `Exp3.M.B`

- Initialize weights $w_i = 1$ for $i = 1, \ldots, N$
- For each round $t = 1, \ldots, \tau_{\mathcal{A}}(B)$:
  - Cap weights that are "too large"

$$w_i(t) = v(t) \text{ for } i \in \tilde{S}(t) = \{i \in [N] \mid w_i(t) > v_t\}$$

$$v(t) \leftarrow \left\{ v_t \;\middle|\; \frac{v_t(1-\gamma)}{\sum_{i=1}^N v_t \cdot \mathbb{1}(w_i(t) \geq v_t) + w_i(t) \cdot \mathbb{1}(w_i(t) < v_t)} = \frac{1}{K} - \frac{\gamma}{N} \right\}$$

  - Calculate probabilities $p_i(t) = K\left((1-\gamma)\frac{\tilde{w}_i(t)}{\sum_{j=1}^N \tilde{w}_j(t)} + \frac{\gamma}{N}\right)$
  - Play arms $a_t \sim p_1, \ldots, p_N$
  - Update weights:

$$\hat{r}_i(t) = r_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$
$$\hat{c}_i(t) = c_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$
$$w_i(t+1) = w_i(t) \exp\left[\frac{K\gamma}{N}\left[\hat{r}_i(t) - \hat{c}_i(t)\right] \mathbb{1}_{i \in \tilde{S}(t)}\right]$$

Upper Bound on the Regret (cont'd.)

Analysis of Algorithm `Exp3.M.B`

Theorem (Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.M.B`)

Algorithm `Exp3.M.B` *achieves cumulative regret* $\mathcal{R}_{\mathcal{A}}(B) = O(\sqrt{BN \log(N/K)})$.

Proof Idea

- Modify existing proof techniques[7]:
  - Step 1: Assume fixed time horizon $T = \max(\tau_{\mathcal{A}}(B), \tau_{\mathcal{A}^*}(B))$
  - Step 2: Relate T to budget B

Remarks

- Our bound recovers previous findings for the following special cases with fixed T:
  - Recovers $O(\sqrt{BN \log N})$ bound for $K = 1$[8]
  - Recovers $O(\sqrt{TN \log N})$ bound from for $K = 1$, no costs / budget[9]

[7] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389; P. Auer et al. "The Nonstochastic Multi-Armed Bandit Problem". In: *SIAM Journal on Computing* 32 (2002), pp. 48–77.

[8] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

Upper Bound on the Regret (cont'd.)

Analysis of Algorithm `Exp3.M.B`

Theorem (Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.M.B`)

Algorithm `Exp3.M.B` achieves cumulative regret $\mathcal{R}_{\mathcal{A}}(B) = O(\sqrt{BN \log(N/K)})$.

Proof Idea

- Modify existing proof techniques[7]:
  - Step 1: Assume fixed time horizon $T = \max(\tau_{\mathcal{A}}(B), \tau_{\mathcal{A}^*}(B))$
  - Step 2: Relate T to budget B

Remarks

- Our bound recovers previous findings for the following special cases with fixed T:
  - Recovers $O(\sqrt{BN \log N})$ bound for $K = 1$[8]
  - Recovers $O(\sqrt{TN \log N})$ bound from for $K = 1$, no costs / budget[9]

[7] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389; P. Auer et al. "The Nonstochastic Multi-Armed Bandit Problem". In: *SIAM Journal on Computing* 32 (2002), pp. 48–77.

[8] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

Upper Bound on the Regret (cont'd.)

Analysis of Algorithm `Exp3.M.B`

Theorem (Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.M.B`)

Algorithm `Exp3.M.B` *achieves cumulative regret* $\mathcal{R}_{\mathcal{A}}(B) = O(\sqrt{BN\log(N/K)})$.

Proof Idea

- Modify existing proof techniques[7]:
  - Step 1: Assume fixed time horizon $T = \max(\tau_{\mathcal{A}}(B), \tau_{\mathcal{A}^*}(B))$
  - Step 2: Relate T to budget B

Remarks

- Our bound recovers previous findings for the following special cases with fixed T:
  - Recovers $O(\sqrt{BN\log N})$ bound for $K = 1$[8]
  - Recovers $O(\sqrt{TN\log N})$ bound from for $K = 1$, no costs / budget[9]

---

[7] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389; P. Auer et al. "The Nonstochastic Multi-Armed Bandit Problem". In: *SIAM Journal on Computing* 32 (2002), pp. 48–77.

[8] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

Upper Bound on the Regret (cont'd.)

Analysis of Algorithm `Exp3.M.B`

Theorem (Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.M.B`)

Algorithm `Exp3.M.B` achieves cumulative regret $\mathcal{R}_{\mathcal{A}}(B) = O(\sqrt{BN \log(N/K)})$.

Proof Idea

- Modify existing proof techniques[7]:
  - Step 1: Assume fixed time horizon $T = \max(\tau_{\mathcal{A}}(B), \tau_{\mathcal{A}^*}(B))$
  - Step 2: Relate T to budget B

Remarks

- Our bound recovers previous findings for the following special cases with fixed T:
  - Recovers $O(\sqrt{BN \log N})$ bound for $K = 1$[8]
  - Recovers $O(\sqrt{TN \log N})$ bound from for $K = 1$, no costs / budget[9]

[7] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389; P. Auer et al. "The Nonstochastic Multi-Armed Bandit Problem". In: *SIAM Journal on Computing* 32 (2002), pp. 48–77.

[8] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

Upper Bound on the Regret (cont'd.)

Analysis of Algorithm `Exp3.M.B`

Theorem (Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.M.B`)

Algorithm `Exp3.M.B` *achieves cumulative regret* $\mathcal{R}_{\mathcal{A}}(B) = O(\sqrt{BN \log(N/K)})$.

Proof Idea

- Modify existing proof techniques[7]:
  - Step 1: Assume fixed time horizon $T = \max(\tau_{\mathcal{A}}(B), \tau_{\mathcal{A}^*}(B))$
  - Step 2: Relate T to budget B

Remarks

- Our bound recovers previous findings for the following special cases with fixed T:
  - Recovers $O(\sqrt{BN \log N})$ bound for $K = 1$[8]
  - Recovers $O(\sqrt{TN \log N})$ bound from for $K = 1$, no costs / budget[9]

[7] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389; P. Auer et al. "The Nonstochastic Multi-Armed Bandit Problem". In: *SIAM Journal on Computing* 32 (2002), pp. 48–77.

[8] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

Upper Bound on the Regret (cont'd.)

Analysis of Algorithm `Exp3.M.B`

Theorem (Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.M.B`)

Algorithm `Exp3.M.B` *achieves cumulative regret* $\mathcal{R}_{\mathcal{A}}(B) = O(\sqrt{BN \log(N/K)})$.

Proof Idea

- Modify existing proof techniques[7]:
  - Step 1: Assume fixed time horizon $T = \max(\tau_{\mathcal{A}}(B), \tau_{\mathcal{A}^*}(B))$
  - Step 2: Relate T to budget B

Remarks

- Our bound recovers previous findings for the following special cases with fixed T:
  - Recovers $O(\sqrt{BN \log N})$ bound for $K = 1$[8]
  - Recovers $O(\sqrt{TN \log N})$ bound from for $K = 1$, no costs / budget[9]

[7] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389; P. Auer et al. "The Nonstochastic Multi-Armed Bandit Problem". In: *SIAM Journal on Computing* 32 (2002), pp. 48–77.

[8] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

Upper Bound on the Regret (cont'd.)

Analysis of Algorithm `Exp3.M.B`

Theorem (Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.M.B`)

Algorithm `Exp3.M.B` achieves cumulative regret $\mathcal{R}_{\mathcal{A}}(B) = O(\sqrt{BN \log(N/K)})$.

Proof Idea

- Modify existing proof techniques[7]:
  - Step 1: Assume fixed time horizon $T = \max(\tau_{\mathcal{A}}(B), \tau_{\mathcal{A}^*}(B))$
  - Step 2: Relate T to budget B

Remarks

- Our bound recovers previous findings for the following special cases with fixed T:
  - Recovers $O(\sqrt{BN \log N})$ bound for $K = 1$[8]
  - Recovers $O(\sqrt{TN \log N})$ bound from for $K = 1$, no costs / budget[9]

[7] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389; P. Auer et al. "The Nonstochastic Multi-Armed Bandit Problem". In: *SIAM Journal on Computing* 32 (2002), pp. 48–77.

[8] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

Lower Bound on the Regret

Theorem (Lower Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.M.B`)

The weak regret of Algorithm Exp3.M.B is at least

$$\mathcal{R} \geq \min\left(\frac{c_{\min}^{3/2}(1-K/N)^2}{8\sqrt{\log(4/3)}} \sqrt{\frac{NB}{K}},\ \frac{B(1-K/N)}{8} \right).$$

This bound is of order $\Omega((1-K/N)^2\sqrt{NB/K})$.

Proof Idea

- Step 1: Derive auxiliary lemma. Select K out of N arms at random to be "good" arms with $r_i(t) \sim \text{Bern}(1/2+\varepsilon)$; $c_i(t) = c_{\min}$ w.p. $1/2+\varepsilon$, $c_i(t) = 1$ w.p. $1/2-\varepsilon$

Lemma

Let $f : \{\{0,1\},\{c_{\min},1\}\}^{\tau_{\max}} \to [0,M]$ be any function defined on reward and cost sequences $\{\mathbf{r},\mathbf{c}\}$ of length less than or equal $\tau_{\max} = \frac{B}{Kc_{\min}}$. Then for the best action set a^*:

$$\mathbb{E}_{a^*}[f(\mathbf{r},\mathbf{c})] \leq \mathbb{E}_u[f(\mathbf{r},\mathbf{c})] + \frac{Bc_{\min}^{-3/2}}{2}\sqrt{-\mathbb{E}_u[N_{a^*}]\log(1-4\varepsilon^2)},$$

Lower Bound on the Regret

Theorem (Lower Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm Exp3.M.B)

The weak regret of Algorithm Exp3.M.B is at least

$$\mathcal{R} \geq \min\left(\frac{c_{\min}^{3/2}(1-K/N)^2}{8\sqrt{\log(4/3)}} \sqrt{\frac{NB}{K}},\ \frac{B(1-K/N)}{8} \right).$$

This bound is of order $\Omega((1-K/N)^2\sqrt{NB/K})$.

Proof Idea

- Step 1: Derive auxiliary lemma. Select K out of N arms at random to be "good" arms with $r_i(t) \sim \mathrm{Bern}(1/2+\varepsilon)$; $c_i(t) = c_{\min}$ w.p. $1/2+\varepsilon$, $c_i(t) = 1$ w.p. $1/2-\varepsilon$

Lemma

Let $f : \{\{0,1\},\{c_{\min},1\}\}^{\tau_{\max}} \to [0,M]$ be any function defined on reward and cost sequences $\{\mathbf{r},\mathbf{c}\}$ of length less than or equal $\tau_{\max} = \frac{B}{Kc_{\min}}$. Then for the best action set a^*:

$$\mathbb{E}_{a^*}[f(\mathbf{r},\mathbf{c})] \leq \mathbb{E}_u[f(\mathbf{r},\mathbf{c})] + \frac{Bc_{\min}^{-3/2}}{2}\sqrt{-\mathbb{E}_u[N_{a^*}]\log(1-4\varepsilon^2)},$$

Lower Bound on the Regret

Theorem (Lower Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm Exp3.M.B)

The weak regret of Algorithm Exp3.M.B is at least

$$\mathcal{R} \geq \min\left(\frac{c_{\min}^{3/2}(1-K/N)^2}{8\sqrt{\log(4/3)}}\sqrt{\frac{NB}{K}},\ \frac{B(1-K/N)}{8}\right).$$

This bound is of order $\Omega((1-K/N)^2\sqrt{NB/K})$.

Proof Idea

- Step 1: Derive auxiliary lemma. Select K out of N arms at random to be "good" arms with $r_i(t) \sim \mathrm{Bern}(1/2+\varepsilon)$; $c_i(t) = c_{\min}$ w.p. $1/2+\varepsilon$, $c_i(t) = 1$ w.p. $1/2-\varepsilon$

Lemma

Let $f : \{\{0,1\},\{c_{\min},1\}\}^{\tau_{\max}} \to [0, M]$ *be any function defined on reward and cost sequences* $\{\mathbf{r}, \mathbf{c}\}$ *of length less than or equal* $\tau_{\max} = \frac{B}{Kc_{\min}}$. *Then for the best action set* a^*:

$$\mathbb{E}_{a^*}[f(\mathbf{r},\mathbf{c})] \leq \mathbb{E}_u[f(\mathbf{r},\mathbf{c})] + \frac{Bc_{\min}^{-3/2}}{2}\sqrt{-\mathbb{E}_u[N_{a^*}]\log(1-4\varepsilon^2)},$$

Lower Bound on the Regret (cont'd.)

Proof Idea (cont'd.)

- Step 2: Notice there exist $\binom{N}{K}$ unique combinations of K tuples.
  - Let $\mathbf{C}([N], K)$ denote the set of all such subsets[10]
  - Let $\mathbb{E}_*[\cdot]$ denote the expected value w.r.t. uniform assignment of "good" arms.

$$\mathbb{E}_*[G_{\max}] = \left(\frac{1}{2} + \varepsilon\right) K\mathbb{E}_*[\tau_{\mathcal{A}}(B),]$$

$$\mathbb{E}_{a^*}[G_{\mathcal{A}}] = \frac{1}{2}K\mathbb{E}_{a^*}[\tau_{\mathcal{A}}(B)] + \varepsilon\mathbb{E}_{a^*}[N_{a^*}],$$

$$\mathbb{E}_*[G_{\mathcal{A}}] = \frac{1}{\binom{N}{K}} \sum_{a^* \in \mathbf{C}([N],K)} \mathbb{E}_{a^*}[G_{\mathcal{A}}] = \frac{1}{2}K\mathbb{E}_*[\tau_{\mathcal{A}}(B)] + \frac{\varepsilon}{\binom{N}{K}} \sum_{a^* \in \mathbf{C}([N],K)} \mathbb{E}_{a^*}[N_{a^*}].$$

- Step 3: Use previous lemma to bound $\mathbb{E}_*[G_{\max} - G_{\mathcal{A}}]$:

$$\mathbb{E}_*[G_{\max} - G_{\mathcal{A}}] \geq \varepsilon B\left(1 - \frac{K}{N}\right) - \frac{2\varepsilon B}{c_{\min}^{3/2}}\sqrt{\frac{BK}{N}\log(4/3)}.$$

- Step 4: Tune ε to optimize bound:

$$\varepsilon = \min\left(\frac{1}{4}, \frac{c_{\min}^{3/2}}{4\log(4/3)}(1 - K/N)\sqrt{\frac{N}{BK}}\right).$$

[10] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

Lower Bound on the Regret (cont'd.)

Proof Idea (cont'd.)

- Step 2: Notice there exist $\binom{N}{K}$ unique combinations of K tuples.
  - Let $\mathbf{C}([N], K)$ denote the set of all such subsets[10]
  - Let $\mathbb{E}_*[\cdot]$ denote the expected value w.r.t. uniform assignment of "good" arms.

$$\mathbb{E}_*[G_{\max}] = \left(\frac{1}{2} + \varepsilon\right) K\mathbb{E}_*[\tau_{\mathcal{A}}(B),]$$

$$\mathbb{E}_{a^*}[G_{\mathcal{A}}] = \frac{1}{2}K\mathbb{E}_{a^*}[\tau_{\mathcal{A}}(B)] + \varepsilon\mathbb{E}_{a^*}[N_{a^*}],$$

$$\mathbb{E}_*[G_{\mathcal{A}}] = \frac{1}{\binom{N}{K}} \sum_{a^* \in \mathbf{C}([N],K)} \mathbb{E}_{a^*}[G_{\mathcal{A}}] = \frac{1}{2}K\mathbb{E}_*[\tau_{\mathcal{A}}(B)] + \frac{\varepsilon}{\binom{N}{K}} \sum_{a^* \in \mathbf{C}([N],K)} \mathbb{E}_{a^*}[N_{a^*}].$$

- Step 3: Use previous lemma to bound $\mathbb{E}_*[G_{\max} - G_{\mathcal{A}}]$:

$$\mathbb{E}_*[G_{\max} - G_{\mathcal{A}}] \geq \varepsilon B\left(1 - \frac{K}{N}\right) - \frac{2\varepsilon B}{c_{\min}^{3/2}}\sqrt{\frac{BK}{N}\log(4/3)}.$$

- Step 4: Tune ε to optimize bound:

$$\varepsilon = \min\left(\frac{1}{4}, \frac{c_{\min}^{3/2}}{4\log(4/3)}(1 - K/N)\sqrt{\frac{N}{BK}}\right).$$

[10] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

Lower Bound on the Regret (cont'd.)

Proof Idea (cont'd.)

- Step 2: Notice there exist $\binom{N}{K}$ unique combinations of K tuples.
  - Let $\mathbf{C}([N], K)$ denote the set of all such subsets[10]
  - Let $\mathbb{E}_*[\cdot]$ denote the expected value w.r.t. uniform assignment of "good" arms.

$$\mathbb{E}_*[G_{\max}] = \left(\frac{1}{2} + \varepsilon\right) K\mathbb{E}_*[\tau_{\mathcal{A}}(B),]$$

$$\mathbb{E}_{a^*}[G_{\mathcal{A}}] = \frac{1}{2} K\mathbb{E}_{a^*}[\tau_{\mathcal{A}}(B)] + \varepsilon\mathbb{E}_{a^*}[N_{a^*}],$$

$$\mathbb{E}_*[G_{\mathcal{A}}] = \frac{1}{\binom{N}{K}} \sum_{a^* \in \mathbf{C}([N],K)} \mathbb{E}_{a^*}[G_{\mathcal{A}}] = \frac{1}{2} K\mathbb{E}_*[\tau_{\mathcal{A}}(B)] + \frac{\varepsilon}{\binom{N}{K}} \sum_{a^* \in \mathbf{C}([N],K)} \mathbb{E}_{a^*}[N_{a^*}].$$

- Step 3: Use previous lemma to bound $\mathbb{E}_*[G_{\max} - G_{\mathcal{A}}]$:

$$\mathbb{E}_*[G_{\max} - G_{\mathcal{A}}] \geq \varepsilon B \left(1 - \frac{K}{N}\right) - \frac{2\varepsilon B}{c_{\min}^{3/2}} \sqrt{\frac{BK}{N} \log(4/3)}.$$

- Step 4: Tune ε to optimize bound:

$$\varepsilon = \min\left(\frac{1}{4}, \frac{c_{\min}^{3/2}}{4\log(4/3)} (1 - K/N) \sqrt{\frac{N}{BK}}\right).$$

[10] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

Lower Bound on the Regret (cont'd.)

Proof Idea (cont'd.)

- Step 2: Notice there exist $\binom{N}{K}$ unique combinations of K tuples.
  - Let $\mathbf{C}([N], K)$ denote the set of all such subsets[10]
  - Let $\mathbb{E}_*[\cdot]$ denote the expected value w.r.t. uniform assignment of "good" arms.

$$\mathbb{E}_*[G_{\max}] = \left(\frac{1}{2} + \varepsilon\right) K\mathbb{E}_*[\tau_{\mathcal{A}}(B),]$$

$$\mathbb{E}_{a^*}[G_{\mathcal{A}}] = \frac{1}{2}K\mathbb{E}_{a^*}[\tau_{\mathcal{A}}(B)] + \varepsilon\mathbb{E}_{a^*}[N_{a^*}],$$

$$\mathbb{E}_*[G_{\mathcal{A}}] = \frac{1}{\binom{N}{K}} \sum_{a^* \in \mathbf{C}([N],K)} \mathbb{E}_{a^*}[G_{\mathcal{A}}] = \frac{1}{2}K\mathbb{E}_*[\tau_{\mathcal{A}}(B)] + \frac{\varepsilon}{\binom{N}{K}} \sum_{a^* \in \mathbf{C}([N],K)} \mathbb{E}_{a^*}[N_{a^*}].$$

- Step 3: Use previous lemma to bound $\mathbb{E}_*[G_{\max} - G_{\mathcal{A}}]$:

$$\mathbb{E}_*[G_{\max} - G_{\mathcal{A}}] \geq \varepsilon B\left(1 - \frac{K}{N}\right) - \frac{2\varepsilon B}{c_{\min}^{3/2}}\sqrt{\frac{BK}{N}\log(4/3)}.$$

- Step 4: Tune ε to optimize bound:

$$\varepsilon = \min\left(\frac{1}{4}, \frac{c_{\min}^{3/2}}{4\log(4/3)}(1 - K/N)\sqrt{\frac{N}{BK}}\right).$$

[10] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

Lower Bound on the Regret (cont'd.)

Proof Idea (cont'd.)

- Step 2: Notice there exist $\binom{N}{K}$ unique combinations of K tuples.
  - Let $\mathbf{C}([N], K)$ denote the set of all such subsets[10]
  - Let $\mathbb{E}_*[\cdot]$ denote the expected value w.r.t. uniform assignment of "good" arms.

$$\mathbb{E}_*[G_{\max}] = \left(\frac{1}{2} + \varepsilon\right) K\mathbb{E}_*[\tau_{\mathcal{A}}(B),]$$

$$\mathbb{E}_{a^*}[G_{\mathcal{A}}] = \frac{1}{2}K\mathbb{E}_{a^*}[\tau_{\mathcal{A}}(B)] + \varepsilon\mathbb{E}_{a^*}[N_{a^*}],$$

$$\mathbb{E}_*[G_{\mathcal{A}}] = \frac{1}{\binom{N}{K}} \sum_{a^* \in \mathbf{C}([N],K)} \mathbb{E}_{a^*}[G_{\mathcal{A}}] = \frac{1}{2}K\mathbb{E}_*[\tau_{\mathcal{A}}(B)] + \frac{\varepsilon}{\binom{N}{K}} \sum_{a^* \in \mathbf{C}([N],K)} \mathbb{E}_{a^*}[N_{a^*}].$$

- Step 3: Use previous lemma to bound $\mathbb{E}_*[G_{\max} - G_{\mathcal{A}}]$:

$$\mathbb{E}_*[G_{\max} - G_{\mathcal{A}}] \geq \varepsilon B\left(1 - \frac{K}{N}\right) - \frac{2\varepsilon B}{c_{\min}^{3/2}}\sqrt{\frac{BK}{N}\log(4/3)}.$$

- Step 4: Tune ε to optimize bound:

$$\varepsilon = \min\left(\frac{1}{4}, \frac{c_{\min}^{3/2}}{4\log(4/3)}(1 - K/N)\sqrt{\frac{N}{BK}}\right).$$

[10] T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

High Probability Bound on the Regret

Algorithm `Exp3.P.M.B`

- Modification of Algorithm `Exp3.M.B`:
  - Initialize parameter $\alpha = 2\sqrt{6}\sqrt{(N-K)/(N-1)\log\left(NB/(Kc_{\min}\delta)\right)}$.
  - Initialize weights w_i for $i \in [N]$: $w_i(1) = \exp\left(\alpha\gamma K^2\sqrt{B/(NKc_{\min})}/3\right)$.
  - Update weights for $i \in [N]$ as follows:

$$w_i(t+1) = w_i(t)\exp\left[\mathbb{1}_{i \notin \tilde{S}(t)}\frac{\gamma K}{3N}\left(\hat{r}_i(t) - \hat{c}_i(t) + \frac{\alpha\sqrt{Kc_{\min}}}{p_i(t)\sqrt{NB}}\right)\right].$$

Theorem (High Probability Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.P.M.B`)

For the multiple play algorithm ($1 \le K \le N$) and the budget $B > 0$, the following bound on the regret holds with probability at least $1-\delta$:

$$\begin{aligned}\mathcal{R} &\le 2\sqrt{3}\sqrt{\frac{NB(1-c_{\min})}{c_{\min}}\log\frac{N}{K}} + 4\sqrt{6}\frac{N-K}{N-1}\log\left(\frac{NB}{Kc_{\min}\delta}\right)\\ &\quad + 2\sqrt{6}(1+K^2)\sqrt{\frac{N-K}{N-1}\frac{NB}{Kc_{\min}}\log\left(\frac{NB}{Kc_{\min}\delta}\right)}\\ &= O\left(K^2\sqrt{\frac{NB}{K}\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)} + \frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)\right)\end{aligned}$$

High Probability Bound on the Regret

Algorithm `Exp3.P.M.B`

- Modification of Algorithm `Exp3.M.B`:
  - Initialize parameter $\alpha = 2\sqrt{6}\sqrt{(N-K)/(N-1)\log(NB/(Kc_{\min}\delta))}$.
  - Initialize weights w_i for $i \in [N]$: $w_i(1) = \exp\left(\alpha\gamma K^2\sqrt{B/(NKc_{\min})}/3\right)$.
  - Update weights for $i \in [N]$ as follows:

$$w_i(t+1) = w_i(t)\exp\left[\mathbb{1}_{i \notin \tilde{S}(t)}\frac{\gamma K}{3N}\left(\hat{r}_i(t) - \hat{c}_i(t) + \frac{\alpha\sqrt{Kc_{\min}}}{p_i(t)\sqrt{NB}}\right)\right].$$

Theorem (High Probability Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.P.M.B`)

For the multiple play algorithm ($1 \le K \le N$) and the budget $B > 0$, the following bound on the regret holds with probability at least $1-\delta$:

$$\begin{aligned}\mathcal{R} &\le 2\sqrt{3}\sqrt{\frac{NB(1-c_{\min})}{c_{\min}}\log\frac{N}{K}} + 4\sqrt{6}\frac{N-K}{N-1}\log\left(\frac{NB}{Kc_{\min}\delta}\right) \\ &\quad + 2\sqrt{6}(1+K^2)\sqrt{\frac{N-K}{N-1}\frac{NB}{Kc_{\min}}\log\left(\frac{NB}{Kc_{\min}\delta}\right)} \\ &= O\left(K^2\sqrt{\frac{NB}{K}\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)} + \frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)\right)\end{aligned}$$

High Probability Bound on the Regret

Algorithm `Exp3.P.M.B`

- Modification of Algorithm `Exp3.M.B`:
  - Initialize parameter $\alpha = 2\sqrt{6}\sqrt{(N-K)/(N-1)\log\left(NB/(Kc_{\min}\delta)\right)}$.
  - Initialize weights w_i for $i \in [N]$: $w_i(1) = \exp\left(\alpha\gamma K^2\sqrt{B/(NKc_{\min})}/3\right)$.
  - Update weights for $i \in [N]$ as follows:

$$w_i(t+1) = w_i(t)\exp\left[\mathbb{1}_{i \notin \tilde{S}(t)}\frac{\gamma K}{3N}\left(\hat{r}_i(t) - \hat{c}_i(t) + \frac{\alpha\sqrt{Kc_{\min}}}{p_i(t)\sqrt{NB}}\right)\right].$$

Theorem (High Probability Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.P.M.B`)

For the multiple play algorithm ($1 \le K \le N$) and the budget $B > 0$, the following bound on the regret holds with probability at least $1-\delta$:

$$\begin{aligned}\mathcal{R} &\le 2\sqrt{3}\sqrt{\frac{NB(1-c_{\min})}{c_{\min}}\log\frac{N}{K}} + 4\sqrt{6}\frac{N-K}{N-1}\log\left(\frac{NB}{Kc_{\min}\delta}\right) \\ &\quad + 2\sqrt{6}(1+K^2)\sqrt{\frac{N-K}{N-1}\frac{NB}{Kc_{\min}}\log\left(\frac{NB}{Kc_{\min}\delta}\right)} \\ &= O\left(K^2\sqrt{\frac{NB}{K}\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)} + \frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)\right)\end{aligned}$$

High Probability Bound on the Regret

Algorithm `Exp3.P.M.B`

- Modification of Algorithm `Exp3.M.B`:
  - Initialize parameter $\alpha = 2\sqrt{6}\sqrt{(N-K)/(N-1)\log\left(NB/(Kc_{\min}\delta)\right)}$.
  - Initialize weights w_i for $i \in [N]$: $w_i(1) = \exp\left(\alpha\gamma K^2\sqrt{B/(NKc_{\min})}/3\right)$.
  - Update weights for $i \in [N]$ as follows:

$$w_i(t+1) = w_i(t)\exp\left[\mathbb{1}_{i \notin \tilde{S}(t)}\frac{\gamma K}{3N}\left(\hat{r}_i(t) - \hat{c}_i(t) + \frac{\alpha\sqrt{Kc_{\min}}}{p_i(t)\sqrt{NB}}\right)\right].$$

Theorem (High Probability Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.P.M.B`)

For the multiple play algorithm ($1 \leq K \leq N$) and the budget $B > 0$, the following bound on the regret holds with probability at least $1-\delta$:

$$\begin{aligned}\mathcal{R} &\leq 2\sqrt{3}\sqrt{\frac{NB(1-c_{\min})}{c_{\min}}\log\frac{N}{K}} + 4\sqrt{6}\frac{N-K}{N-1}\log\left(\frac{NB}{Kc_{\min}\delta}\right) \\ &\quad + 2\sqrt{6}(1+K^2)\sqrt{\frac{N-K}{N-1}\frac{NB}{Kc_{\min}}\log\left(\frac{NB}{Kc_{\min}\delta}\right)} \\ &= O\left(K^2\sqrt{\frac{NB}{K}\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)} + \frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)\right)\end{aligned}$$

High Probability Bound on the Regret

Algorithm `Exp3.P.M.B`

- Modification of Algorithm `Exp3.M.B`:
  - Initialize parameter $\alpha = 2\sqrt{6}\sqrt{(N-K)/(N-1)\log\left(NB/(Kc_{\min}\delta)\right)}$.
  - Initialize weights w_i for $i \in [N]$: $w_i(1) = \exp\left(\alpha\gamma K^2\sqrt{B/(NKc_{\min})}/3\right)$.
  - Update weights for $i \in [N]$ as follows:

$$w_i(t+1) = w_i(t)\exp\left[\mathbb{1}_{i \notin \tilde{S}(t)} \frac{\gamma K}{3N}\left(\hat{r}_i(t) - \hat{c}_i(t) + \frac{\alpha\sqrt{Kc_{\min}}}{p_i(t)\sqrt{NB}}\right)\right].$$

Theorem (High Probability Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.P.M.B`)

For the multiple play algorithm ($1 \leq K \leq N$) and the budget $B > 0$, the following bound on the regret holds with probability at least $1-\delta$:

$$\begin{aligned}\mathcal{R} &\leq 2\sqrt{3}\sqrt{\frac{NB(1-c_{\min})}{c_{\min}}\log\frac{N}{K}} + 4\sqrt{6}\frac{N-K}{N-1}\log\left(\frac{NB}{Kc_{\min}\delta}\right) \\ &\quad + 2\sqrt{6}(1+K^2)\sqrt{\frac{N-K}{N-1}\frac{NB}{Kc_{\min}}\log\left(\frac{NB}{Kc_{\min}\delta}\right)} \\ &= O\left(K^2\sqrt{\frac{NB}{K}\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)} + \frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)\right)\end{aligned}$$

High Probability Bound on the Regret (cont'd.)

Theorem (High Probability Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm `Exp3.P.M.B`)

For the multiple play algorithm ($1 \leq K \leq N$) and the budget $B > 0$, the following bound on the regret holds with probability at least $1 - \delta$:

$$\mathcal{R} = O\left(K^2 \sqrt{\frac{NB}{K} \frac{N-K}{N-1} \log\left(\frac{NB}{K\delta}\right)} + \frac{N-K}{N-1} \log\left(\frac{NB}{K\delta}\right)\right)$$

Remark

- Recovers $O(\sqrt{NT \log(NT/\delta)} + \log(NT/\delta))$ bound[11] for $K = 1$, no costs

Proof Idea

- Step 1: Derive upper confidence bound $\hat{U}$ on $G_{\max}$ that holds *w.h.p.*
- Step 2: Lower bound $G_{\texttt{Exp3.P.M.B}}$ as function of $\hat{U}$
- Step 3: Combine to obtain upper bound on $G_{\max} - G_{\texttt{Exp3.P.M.B}}$

[11] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

High Probability Bound on the Regret (cont'd.)

Theorem (High Probability Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm Exp3.P.M.B)

For the multiple play algorithm ($1 \leq K \leq N$) and the budget $B > 0$, the following bound on the regret holds with probability at least $1 - \delta$:

$$\mathcal{R} = O\left(K^2\sqrt{\frac{NB}{K}\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)} + \frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)\right)$$

Remark

- Recovers $O(\sqrt{NT\log(NT/\delta)} + \log(NT/\delta))$ bound[11] for $K = 1$, no costs

Proof Idea

- Step 1: Derive upper confidence bound $\hat{U}$ on $G_{\max}$ that holds *w.h.p.*
- Step 2: Lower bound $G_{\mathtt{Exp3.P.M.B}}$ as function of $\hat{U}$
- Step 3: Combine to obtain upper bound on $G_{\max} - G_{\mathtt{Exp3.P.M.B}}$

[11] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

High Probability Bound on the Regret (cont'd.)

Theorem (High Probability Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm Exp3.P.M.B)

For the multiple play algorithm ($1 \leq K \leq N$) and the budget $B > 0$, the following bound on the regret holds with probability at least $1 - \delta$:

$$\mathcal{R} = O\left(K^2 \sqrt{\frac{NB}{K} \frac{N-K}{N-1} \log\left(\frac{NB}{K\delta}\right)} + \frac{N-K}{N-1} \log\left(\frac{NB}{K\delta}\right)\right)$$

Remark

- Recovers $O(\sqrt{NT \log(NT/\delta)} + \log(NT/\delta))$ bound[11] for $K = 1$, no costs

Proof Idea

- Step 1: Derive upper confidence bound $\hat{U}$ on $G_{\max}$ that holds *w.h.p.*
- Step 2: Lower bound $G_{\texttt{Exp3.P.M.B}}$ as function of $\hat{U}$
- Step 3: Combine to obtain upper bound on $G_{\max} - G_{\texttt{Exp3.P.M.B}}$

[11] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

High Probability Bound on the Regret (cont'd.)

Theorem (High Probability Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm Exp3.P.M.B)

For the multiple play algorithm ($1 \le K \le N$) and the budget $B > 0$, the following bound on the regret holds with probability at least $1 - \delta$:

$$\mathcal{R} = O\left(K^2 \sqrt{\frac{NB}{K}\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)} + \frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)\right)$$

Remark

- Recovers $O(\sqrt{NT\log(NT/\delta)} + \log(NT/\delta))$ bound[11] for $K = 1$, no costs

Proof Idea

- Step 1: Derive upper confidence bound $\hat{U}$ on $G_{\max}$ that holds *w.h.p.*
- Step 2: Lower bound $G_{\texttt{Exp3.P.M.B}}$ as function of $\hat{U}$
- Step 3: Combine to obtain upper bound on $G_{\max} - G_{\texttt{Exp3.P.M.B}}$

[11] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

High Probability Bound on the Regret (cont'd.)

Theorem (High Probability Upper Bound on $\mathcal{R}_{\mathcal{A}}(B)$ for Algorithm Exp3.P.M.B)

For the multiple play algorithm ($1 \leq K \leq N$) and the budget $B > 0$, the following bound on the regret holds with probability at least $1-\delta$:

$$\mathcal{R} = O\left(K^2\sqrt{\frac{NB}{K}\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)} + \frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)\right)$$

Remark

- Recovers $O(\sqrt{NT\log(NT/\delta)} + \log(NT/\delta))$ bound[11] for $K=1$, no costs

Proof Idea

- Step 1: Derive upper confidence bound $\hat{U}$ on $G_{\max}$ that holds *w.h.p.*
- Step 2: Lower bound $G_{\texttt{Exp3.P.M.B}}$ as function of $\hat{U}$
- Step 3: Combine to obtain upper bound on $G_{\max} - G_{\texttt{Exp3.P.M.B}}$

[11] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

High Probability on the Regret (cont'd.)

Proof Idea (cont'd.)

- Step 1: Derive upper confidence bound $\hat{U}$ on $G_{\max}$ that holds *w.h.p.*:
  - Define upper confidence bound:

$$\hat{U}^* = \sum_{i \in a^*} \left(\alpha \hat{\sigma}_i + \sum_{t=1}^{\tau_{a^*}(B)} \left(\hat{r}_i(t) - \hat{c}_i(t) \right) \right)$$

  - For $2\sqrt{6}\sqrt{\frac{N-K}{N-1} \log \frac{NB}{Kc_{\min}\delta}} \leq \alpha \leq 12\sqrt{\frac{NB}{Kc_{\min}}}$, we can show $\mathbb{P}\left(\hat{U}^* > G_{\max} - B\right) \geq 1 - \delta$
- Step 2: Lower bound $G_{\mathtt{Exp3.P.M.B}}$ as function of $\hat{U}$:
  - For $\alpha \leq 2\sqrt{\frac{NB}{Kc_{\min}}}$, the gain of Algorithm `Exp3.P.M.B` is bounded below as follows:

$$G_{\mathtt{Exp3.P.M.B}} \geq \left(1 - \gamma - \frac{2\gamma}{3} \frac{1 - c_{\min}}{c_{\min}}\right) \hat{U}^* - \frac{3N}{\gamma} \log \frac{N}{K} - 2\alpha^2 - \alpha(1 + K^2) \frac{BN}{Kc_{\min}}.$$

- Step 3: Combine to obtain upper bound on $G_{\max} - G_{\mathtt{Exp3.P.M.B}}$, tune γ:

$$\gamma = \min\left(\left(1 + \frac{2}{3} \frac{1 - c_{\min}}{c_{\min}}\right)^{-1}, \left(\frac{3N \log(N/K)}{(G_{\max} - B)\left(1 + 2(1 - c_{\min})/(3c_{\min})\right)}, \right)^{1/2} \right)$$

High Probability on the Regret (cont'd.)

Proof Idea (cont'd.)

- Step 1: Derive upper confidence bound $\hat{U}$ on $G_{\max}$ that holds *w.h.p.*:
  - Define upper confidence bound:

$$\hat{U}^* = \sum_{i \in a^*} \left(\alpha \hat{\sigma}_i + \sum_{t=1}^{\tau_{a^*}(B)} (\hat{r}_i(t) - \hat{c}_i(t)) \right)$$

  - For $2\sqrt{6}\sqrt{\frac{N-K}{N-1} \log \frac{NB}{Kc_{\min}\delta}} \leq \alpha \leq 12\sqrt{\frac{NB}{Kc_{\min}}}$, we can show $\mathbb{P}\left(\hat{U}^* > G_{\max} - B\right) \geq 1 - \delta$
- Step 2: Lower bound $G_{\mathtt{Exp3.P.M.B}}$ as function of $\hat{U}$:
  - For $\alpha \leq 2\sqrt{\frac{NB}{Kc_{\min}}}$, the gain of Algorithm `Exp3.P.M.B` is bounded below as follows:

$$G_{\mathtt{Exp3.P.M.B}} \geq \left(1 - \gamma - \frac{2\gamma}{3} \frac{1 - c_{\min}}{c_{\min}}\right) \hat{U}^* - \frac{3N}{\gamma} \log \frac{N}{K} - 2\alpha^2 - \alpha(1 + K^2) \frac{BN}{Kc_{\min}}.$$

- Step 3: Combine to obtain upper bound on $G_{\max} - G_{\mathtt{Exp3.P.M.B}}$, tune γ:

$$\gamma = \min\left(\left(1 + \frac{2}{3}\frac{1 - c_{\min}}{c_{\min}}\right)^{-1}, \left(\frac{3N \log(N/K)}{(G_{\max} - B)\left(1 + 2(1 - c_{\min})/(3c_{\min})\right)},\right)^{1/2}\right)$$

High Probability on the Regret (cont'd.)

Proof Idea (cont'd.)

- Step 1: Derive upper confidence bound $\hat{U}$ on $G_{\max}$ that holds *w.h.p.*:
  - Define upper confidence bound:

$$\hat{U}^* = \sum_{i \in a^*} \left(\alpha \hat{\sigma}_i + \sum_{t=1}^{\tau_{a^*}(B)} \left(\hat{r}_i(t) - \hat{c}_i(t) \right) \right)$$

  - For $2\sqrt{6}\sqrt{\frac{N-K}{N-1} \log \frac{NB}{Kc_{\min}\delta}} \leq \alpha \leq 12\sqrt{\frac{NB}{Kc_{\min}}}$, we can show $\mathbb{P}\left(\hat{U}^* > G_{\max} - B\right) \geq 1 - \delta$
- Step 2: Lower bound $G_{\mathtt{Exp3.P.M.B}}$ as function of $\hat{U}$:
  - For $\alpha \leq 2\sqrt{\frac{NB}{Kc_{\min}}}$, the gain of Algorithm Exp3.P.M.B is bounded below as follows:

$$G_{\mathtt{Exp3.P.M.B}} \geq \left(1 - \gamma - \frac{2\gamma}{3} \frac{1 - c_{\min}}{c_{\min}}\right) \hat{U}^* - \frac{3N}{\gamma} \log \frac{N}{K} - 2\alpha^2 - \alpha(1 + K^2) \frac{BN}{Kc_{\min}}.$$

- Step 3: Combine to obtain upper bound on $G_{\max} - G_{\mathtt{Exp3.P.M.B}}$, tune γ:

$$\gamma = \min\left(\left(1 + \frac{2}{3} \frac{1 - c_{\min}}{c_{\min}}\right)^{-1}, \left(\frac{3N \log(N/K)}{(G_{\max} - B)\left(1 + 2(1 - c_{\min})/(3c_{\min})\right)}, \right)^{1/2} \right)$$

High Probability on the Regret (cont'd.)

Proof Idea (cont'd.)

- Step 1: Derive upper confidence bound $\hat{U}$ on $G_{\max}$ that holds *w.h.p.*:
  - Define upper confidence bound:

$$\hat{U}^* = \sum_{i \in a^*} \left(\alpha \hat{\sigma}_i + \sum_{t=1}^{\tau_{a^*}(B)} (\hat{r}_i(t) - \hat{c}_i(t)) \right)$$

  - For $2\sqrt{6}\sqrt{\frac{N-K}{N-1} \log \frac{NB}{Kc_{\min}\delta}} \leq \alpha \leq 12\sqrt{\frac{NB}{Kc_{\min}}}$, we can show $\mathbb{P}\left(\hat{U}^* > G_{\max} - B\right) \geq 1 - \delta$
- Step 2: Lower bound $G_{\texttt{Exp3.P.M.B}}$ as function of $\hat{U}$:
  - For $\alpha \leq 2\sqrt{\frac{NB}{Kc_{\min}}}$, the gain of Algorithm `Exp3.P.M.B` is bounded below as follows:

$$G_{\texttt{Exp3.P.M.B}} \geq \left(1 - \gamma - \frac{2\gamma}{3}\frac{1 - c_{\min}}{c_{\min}}\right)\hat{U}^* - \frac{3N}{\gamma}\log\frac{N}{K} - 2\alpha^2 - \alpha(1 + K^2)\frac{BN}{Kc_{\min}}.$$

- Step 3: Combine to obtain upper bound on $G_{\max} - G_{\texttt{Exp3.P.M.B}}$, tune γ:

$$\gamma = \min\left(\left(1 + \frac{2}{3}\frac{1 - c_{\min}}{c_{\min}}\right)^{-1}, \left(\frac{3N\log(N/K)}{(G_{\max} - B)(1 + 2(1 - c_{\min})/(3c_{\min}))}\right)^{1/2}\right)$$

High Probability on the Regret (cont'd.)

Proof Idea (cont'd.)

- Step 1: Derive upper confidence bound $\hat{U}$ on $G_{\max}$ that holds *w.h.p.*:
  - Define upper confidence bound:

$$\hat{U}^* = \sum_{i \in a^*} \left(\alpha \hat{\sigma}_i + \sum_{t=1}^{\tau_{a^*}(B)} \left(\hat{r}_i(t) - \hat{c}_i(t) \right) \right)$$

  - For $2\sqrt{6}\sqrt{\frac{N-K}{N-1} \log \frac{NB}{Kc_{\min}\delta}} \leq \alpha \leq 12\sqrt{\frac{NB}{Kc_{\min}}}$, we can show $\mathbb{P}\left(\hat{U}^* > G_{\max} - B\right) \geq 1 - \delta$
- Step 2: Lower bound $G_{\mathtt{Exp3.P.M.B}}$ as function of $\hat{U}$:
  - For $\alpha \leq 2\sqrt{\frac{NB}{Kc_{\min}}}$, the gain of Algorithm Exp3.P.M.B is bounded below as follows:

$$G_{\mathtt{Exp3.P.M.B}} \geq \left(1 - \gamma - \frac{2\gamma}{3}\frac{1 - c_{\min}}{c_{\min}}\right)\hat{U}^* - \frac{3N}{\gamma}\log\frac{N}{K} - 2\alpha^2 - \alpha(1 + K^2)\frac{BN}{Kc_{\min}}.$$

- Step 3: Combine to obtain upper bound on $G_{\max} - G_{\mathtt{Exp3.P.M.B}}$, tune γ:

$$\gamma = \min\left(\left(1 + \frac{2}{3}\frac{1 - c_{\min}}{c_{\min}}\right)^{-1}, \left(\frac{3N\log(N/K)}{(G_{\max} - B)\left(1 + 2(1 - c_{\min})/(3c_{\min})\right)},\right)^{1/2}\right)$$

Table of Contents

Conclusion and Outlook

Summary of Contributions

- Analysis of *budget-constrained MABs* with *multiple plays*
- Stochastic Setting
    - Algorithm `UCB-MB` based on `UCB1`[12] and `UCB-BV`[13] to upper-bound regret
- Adversarial Setting
    - Algorithm `Exp3.M.B` to upper-bound regret
    - Modification of weight updates in `Exp3.M.B` to lower-bound regret
    - Algorithm `Exp3.P.M.B` to for high probability upper bound

Future Work

- Simulations with data
- Connect to Badanidiyuru et al.[14] to recover $O(\sqrt{B})$ adversarial bound with null arm
- Explore connection to e-commerce (mechanism design)
    - Strategic considerations of agents in repeated interactions between buyer and seller
    - Pool of buyers vs. single buyer

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

[13] W. Ding et al. "Multi-Armed Bandit with Budget Constraint and Variable Costs". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013).

[14] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. "Bandits with Knapsacks". In: *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science* (2013), pp. 207–216.

Conclusion and Outlook

Summary of Contributions

- Analysis of *budget-constrained MABs* with *multiple plays*
- Stochastic Setting
  - Algorithm `UCB-MB` based on `UCB1`[12] and `UCB-BV`[13] to upper-bound regret
- Adversarial Setting
  - Algorithm `Exp3.M.B` to upper-bound regret
  - Modification of weight updates in `Exp3.M.B` to lower-bound regret
  - Algorithm `Exp3.P.M.B` to for high probability upper bound

Future Work

- Simulations with data
- Connect to Badanidiyuru et al.[14] to recover $O(\sqrt{B})$ adversarial bound with null arm
- Explore connection to e-commerce (mechanism design)
  - Strategic considerations of agents in repeated interactions between buyer and seller
  - Pool of buyers vs. single buyer

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

[13] W. Ding et al. "Multi-Armed Bandit with Budget Constraint and Variable Costs". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013).

[14] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. "Bandits with Knapsacks". In: *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science* (2013), pp. 207–216.

Conclusion and Outlook

Summary of Contributions

- Analysis of *budget-constrained MABs* with *multiple plays*
- Stochastic Setting
  - Algorithm `UCB-MB` based on `UCB1`[12] and `UCB-BV`[13] to upper-bound regret
- Adversarial Setting
  - Algorithm `Exp3.M.B` to upper-bound regret
  - Modification of weight updates in `Exp3.M.B` to lower-bound regret
  - Algorithm `Exp3.P.M.B` to for high probability upper bound

Future Work

- Simulations with data
- Connect to Badanidiyuru et al.[14] to recover $O(\sqrt{B})$ adversarial bound with null arm
- Explore connection to e-commerce (mechanism design)
  - Strategic considerations of agents in repeated interactions between buyer and seller
  - Pool of buyers vs. single buyer

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

[13] W. Ding et al. "Multi-Armed Bandit with Budget Constraint and Variable Costs". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013).

[14] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. "Bandits with Knapsacks". In: *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science* (2013), pp. 207–216.

Conclusion and Outlook

Summary of Contributions

- Analysis of *budget-constrained MABs* with *multiple plays*
- Stochastic Setting
  - Algorithm `UCB-MB` based on `UCB1`[12] and `UCB-BV`[13] to upper-bound regret
- Adversarial Setting
  - Algorithm `Exp3.M.B` to upper-bound regret
  - Modification of weight updates in `Exp3.M.B` to lower-bound regret
  - Algorithm `Exp3.P.M.B` to for high probability upper bound

Future Work

- Simulations with data
- Connect to Badanidiyuru et al.[14] to recover $O(\sqrt{B})$ adversarial bound with null arm
- Explore connection to e-commerce (mechanism design)
  - Strategic considerations of agents in repeated interactions between buyer and seller
  - Pool of buyers vs. single buyer

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

[13] W. Ding et al. "Multi-Armed Bandit with Budget Constraint and Variable Costs". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013).

[14] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. "Bandits with Knapsacks". In: *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science* (2013), pp. 207–216.

Conclusion and Outlook

Summary of Contributions

- Analysis of *budget-constrained MABs* with *multiple plays*
- Stochastic Setting
  - Algorithm `UCB-MB` based on `UCB1`[12] and `UCB-BV`[13] to upper-bound regret
- Adversarial Setting
  - Algorithm `Exp3.M.B` to upper-bound regret
  - Modification of weight updates in `Exp3.M.B` to lower-bound regret
  - Algorithm `Exp3.P.M.B` to for high probability upper bound

Future Work

- Simulations with data
- Connect to Badanidiyuru et al.[14] to recover $O(\sqrt{B})$ adversarial bound with null arm
- Explore connection to e-commerce (mechanism design)
  - Strategic considerations of agents in repeated interactions between buyer and seller
  - Pool of buyers vs. single buyer

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

[13] W. Ding et al. "Multi-Armed Bandit with Budget Constraint and Variable Costs". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013).

[14] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. "Bandits with Knapsacks". In: *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science* (2013), pp. 207–216.

Conclusion and Outlook

Summary of Contributions

- Analysis of *budget-constrained MABs* with *multiple plays*
- Stochastic Setting
  - Algorithm `UCB-MB` based on `UCB1`[12] and `UCB-BV`[13] to upper-bound regret
- Adversarial Setting
  - Algorithm `Exp3.M.B` to upper-bound regret
  - Modification of weight updates in `Exp3.M.B` to lower-bound regret
  - Algorithm `Exp3.P.M.B` to for high probability upper bound

Future Work

- Simulations with data
- Connect to Badanidiyuru et al.[14] to recover $O(\sqrt{B})$ adversarial bound with null arm
- Explore connection to e-commerce (mechanism design)
  - Strategic considerations of agents in repeated interactions between buyer and seller
  - Pool of buyers vs. single buyer

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

[13] W. Ding et al. "Multi-Armed Bandit with Budget Constraint and Variable Costs". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013).

[14] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. "Bandits with Knapsacks". In: *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science* (2013), pp. 207–216.

Table of Contents

References I

P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

R. Agrawal, M. V. Hegde, and D. Teneketzis. "Multi-Armed Bandits with Multiple Plays and Switching Cost". In: *Stochastics and Stochastic Reports* 29 (1990), pp. 437–459.

V. Anantharam, P. Varaiya, and J. Walrand. "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem - Part I: IID Rewards". In: *IEEE Transactions on Automatic Control* 32 (1986), pp. 968–976.

P. Auer et al. "The Nonstochastic Multi-Armed Bandit Problem". In: *SIAM Journal on Computing* 32 (2002), pp. 48–77.

A. Badanidiyuru, R. Kleinberg, and A. Slivkins. "Bandits with Knapsacks". In: *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science* (2013), pp. 207–216.

W. Ding et al. "Multi-Armed Bandit with Budget Constraint and Variable Costs". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013).

Y. Gai, B. Krishnamachari, and R. Jain. "Combinatorial Network Optimization with Unknown Variables: Multi-Armed Bandits with Linear Rewards and Individual Observations". In: *IEEE/ACM Transactions on Networking* 20.5 (2012), pp. 1466–1478.

References II

L. Tran-Thanh et al. "Epsilon-First Policies for Budget-Limited Multi-Armed Bandits". In: *Twenty-Fourth AAAI Conference on Artificial Intelligence* (2010), pp. 1211–1216.

L. Tran-Thanh et al. "Knapsack Based Optimal Policies for Budget–Limited Multi–Armed Bandits". In: *Twenty-Sixth AAAI Conference on Artificial Intelligence* (2012), pp. 1134–1140.

T. Uchiya, A. Nakamura, and M. Kudo. "Algorithms for Adversarial Bandit Problems with Multiple Plays". In: *International Conference on Algorithmic Learning Theory* (2010), pp. 375–389.

Y. Xia et al. "Budgeted Multi-Armed Bandits with Multiple Plays". In: *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (2016), pp. 2210 –2216.

THANK YOU!
QUESTIONS?