Mistake Bounds on the Noise-Free Multi-Armed Bandit Game[☆]

Atsuyoshi Nakamura[a,*], David P. Helmbold[b], Manfred K. Warmuth[b]

[a] *Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo 060-0814, Japan*
[b] *University of California at Santa Cruz, USA*

Abstract

We study the $\{0,1\}$-loss version of adaptive adversarial multi-armed bandit problems with $\alpha(\geq 1)$ lossless arms. For the problem, we show a tight bound $K - \alpha - \Theta(1/T)$ on the minimax expected number of mistakes (1-losses), where K is the number of arms and T is the number of rounds.

Keywords: computational learning theory, online learning, bandit problem, mistake bound

1. Introduction

We study a game, which we call the *noise-free multi-armed bandit game*. This game is the $\{0,1\}$-loss version of an adversarial multi-armed bandit problem with $\alpha(\geq 1)$ lossless arms. It is a T-round game. In each round t, the adversary sets losses $\ell_{t,i}(\in \{0,1\})$ for each arm i, then the player chooses an arm i_t from the K arms, and finally, the loss ℓ_{t,i_t} of the chosen arm is revealed to the player. The player is said to make a mistake if $\ell_{t,i_t} = 1$. The player's objective is to minimize the expected number of mistakes over the T rounds while the adversary's objective is to maximize the expected number of mistakes. If the adversary were unconstrained, it could set each $\ell_{t,i}$ to 1, forcing T mistakes. However, here the adversary must ensure that at least α arms remain loss free (i.e. at the end of the game there are some α js such that $\sum_{t=1}^{T} \ell_{t,j} = 0$). We call the situation with this constraint on

[☆] The preliminary version of this paper has appeared in [1].

[*] Corresponding author (E-mail:atsu@ist.hokudai.ac.jp, Phone:+81-11-706-6806, Fax:+81-11-706-7832)

the adversary the *noise-free setting*. A trivial upper bound on the minimax number of mistakes is $K - \alpha$ because any clever player will not re-choose an arm that has been observed to have loss.

There have been many studies on the adversarial multi-armed bandit problem [2, 3, 4]. To the best of our knowledge, however, the problem in the noise-free setting has not been studied yet. Many people might wonder if the noise-free setting is interesting.

In fact, after a little consideration, we can find the minimax strategy in the case of the full-information (all losses revealed at each trial) and oblivious adversary settings: in the case with $\alpha = 1$, for $T \geq K - 1$, the minimax expected number of mistakes in the noise-free setting is $\sum_{i=2}^{K}(1/i) = \Theta(\log K)$ for the full-information version of the problem and $(K-1)/2$ for the oblivious adversary version of the problem. (See the detailed analyses in Sec. 5.1 and Sec. 5.2). Note that a nice adversary's algorithm [2] for the (noisy) multi-armed bandit problem is known and can be modified for the noise-free problem with one lossless arm, but it is an oblivious adversary algorithm, so the expected number of mistakes forced by that adversary cannot be lower bounded by any value larger than $(K-1)/2$. The lower bound of $(K-1)/2$ for an oblivious adversary leaves a large uncertainty on the minimax expected number of mistakes in the K-armed bandit game with one lossless arm: it is far from the trivial upper bound of $K - 1$. Our analysis significantly closes this gap, yielding both upper and lower bounds of $K - 1 - \Theta(1/T)$, and exactly matching bounds for the two-arm case.

Determining the minimax adaptive adversary strategy in the noise-free setting is non-trivial. An adaptive adversary does not decide which arms will be lossless at the beginning, but needs only ensure that at least α arms remain lossless throughout. Setting 1-losses to some arms restricts the choice of eventual lossless arm, and there are cases where it is better for the adversary to maintain flexibility by keeping the number of remaining lossless arms. For example, in the noise-free 2-armed bandit game with one lossless arm, consider a player who chooses each arm i with equal probability at $t = 1$, and continues to choose the same arm i until it makes a mistake. The player then switches to choosing the other arm, which must be a lossless arm by the constraint on the adversary. This player has an expected number of mistakes of $1/2$ if the adversary sets any loss to 1 at the first $t = 1$ trial. But If the adversary sets both losses to zero at trial $t = 1$, it can see which arm is chosen by the player and can force a mistake in trial $t = 2$.

A natural strategy for the player is choosing an arm uniformly at random

from the arms with no observed 1-loss so far, but this is not the best strategy. Its expected number of mistakes is $1-1/2^T$ in a T-round 2-armed game with one lossless arm when the adversary selects one arm to have loss 1 at every trial (and the other arm always has loss 0). To reduce the expected number of mistakes the player can instead adopt a gradually sticking strategy: choose one arm uniformly at random when $t = 1$, and at each time $t > 1$, choose the same arm as the previous trial with probability $t/(t+1)$ and choose the other arm with probability $1/(t+1)$. This gradually sticking strategy reduces the expected number of mistakes to $1-1/(T+1)$ against the above non-adaptive adversary. Lemma 2 implies that this player's expected number of mistakes is also bounded by $1 - 1/(T+1)$ when matched against any adversary for the 2-armed game.

We analyze expected mistake bounds for the T-round K-armed bandit game with an adaptive adversary in the case with $\alpha(\geq 1)$ lossless arms. We design both the player's and adversary's algorithms and prove a tight bound $K-\alpha-\Theta(1/T)$ on the minimax expected number of mistakes by analyzing the expected number of mistakes for those algorithms. Our algorithms repeatedly call algorithms for a novel *survival game* as subroutines. The survival game $G(T, K, k)$ is a simpler version of a noise-free K-armed bandit game with T rounds and a more restricted adversary. The adversary starts with the all-zero loss vector, and at any round can set the arms' losses to any 0-1 vector with exactly k ones. However, once the adversary changes to a non-zero vector of losses, it must keep using that loss setting for the remainder of the trials. The goal of the player is to maximize the probability of never making a mistake over the T trials, while the adversary's goal is to maximize the probability that at least one mistake is made. By analyzing algorithms for the survival game, we prove an upper bound $\binom{K-1}{k} / \left(T + \binom{K}{k} - 1\right)$ and a lower bound $(K-k)/(K+(T-1)k)$ on the minimax no-mistake probability for the problem, where K is the number of arms and $\binom{K}{k} = \frac{K!}{k!(K-k)!}$.

This paper is organized as follows. In Sec. 2, we explain necessary notions and notations for describing our algorithms and their analyses and define the noise-free multi-armed bandit game and the survival game. We first analyze the no-mistake probability bounds for the survival game in Sec. 3, then analyze mistake bounds for the noise-free multi-armed bandit game in Sec. 4. In Sec. 5, we analyze the mistake bounds for the full-information and oblivious adversary version of the problem, and compare the bounds for our

setting (adaptive adversary version of the problem) with those. We conclude the paper with an open problem in Sec. 6.

2. Problem Setting

For any natural numbers i, j with $i \leq j$, $[i..j]$ denotes the set $\{i, \ldots, j\}$ and we let $[j]$ denote $[1..j]$. For any sequence $x_1, \ldots, x_n$, we let $\boldsymbol{x}[b..e]$ denote its contiguous subsequence $x_b, \ldots, x_e$. We use $\boldsymbol{x}[b..b-1]$ for the *null* sequence, that is, the sequence with no element.

The *noise-free multi-armed bandit problem* we consider here is the $\{0,1\}$-loss version of an adversarial multi-armed bandit problem with $\alpha(\geq 1)$ lossless arms. It is a T-round game between a player and an adversary. There are K arms (of slot machines): arm $1, \ldots,$ arm K. At each time $t = 1, \ldots, T$, the adversary picks a loss $\ell_{t,i} \in \{0,1\}$ for each arm $i \in [K]$. Let $\boldsymbol{\ell}_t \in \{0,1\}^K$ denote the K-dimensional vector $(\ell_{t,1}, \ldots, \ell_{t,K})$. The player, who does not know $\boldsymbol{\ell}_t$, chooses arm I_t and suffers loss ℓ_{t,I_t}. We say that the player makes a mistake at time t when $\ell_{t,I_t} = 1$.

We allow the player to use a randomized strategy, so at each time t the player's choice I_t is a random variable. Let i_t denote a realization of random variable I_t. We call (i_t, ℓ_{t,i_t}) a player's *observation* at time t and denote it by o_t. Each player's choice I_t can depend only on his/her past observations $\boldsymbol{o}[1..t-1]$. The adversary is allowed to behave adaptively: the adversary's decision $\boldsymbol{\ell}_t$ can depend on both the player's past choices $\boldsymbol{i}[1..t-1]$ and the adversary's past decisions $\boldsymbol{\ell}[1..t-1]$. We also assume that the adversary has sufficient power to analyze the algorithm and determine the probabilities of its possible choices.

The player's and the adversary's objectives are minimization and maximization, respectively, of the player's expected number of mistakes, $\mathbb{E}(\sum_{t=1}^{T} \ell_{t,I_t})$. We evaluate the minimax expected number of mistakes for player's minimizing and adversary's maximizing strategy.

We further introduce the following notions and notations for description of algorithms and analyses. For any set $S \subseteq [K]$, define $\mathbb{1}_S$ to be the K-dimensional $\{0,1\}$-vector whose ith component is 1 if and only if $i \in S$, and let $\mathbb{0}$ denote $\mathbb{1}_\emptyset$. Then, any loss vector $\boldsymbol{\ell}_t$ can be represented as $\mathbb{1}_S$ for $S = \{i \mid \ell_{t,i} = 1\}$. We say that arm i is *polluted* at time t if $\ell_{s,i} = 1$ for some $s \in [1..t-1]$, and that arm i is *clean* at time t otherwise. A polluted arm is said to be *dirty* if the loss of the arm has been already revealed to the player.

Note that the adversary can distinguish clean arms from polluted arms but the player cannot.

We also consider a following simpler variant of the noise-free multi-armed bandit problem. A *survival game* $G(T, K, k)$ is a noise-free K-armed bandit problem with T rounds in which the adversary can change his/her loss vectors to some $\mathbb{1}_S$ with $|S| = k$ only once and must pick zero vectors until then, where $|\cdot|$ is the number of elements in set '$\cdot$'. The game is over when the player makes a mistake. In this problem, we evaluate the minimax no-mistake probability, which coincides with one minus the expected number of mistakes in this case (since the survival game stops when the algorithm makes a mistake).

We first analyze mistake bounds for the survival game, then prove mistake bounds for the noise-free multi-armed bandit game making use of the results on the survival game.

3. No-Mistake Probability Bounds for the Survival Game

In this section, we show upper and lower bounds on the no-mistake probability for survival game $G(T, K, k)$.

First, we analyze the player's algorithm GradSticking(T, K, k) (Algorithm 1) to prove a lower bound for the $G(T, K, k)$ survival game. Choosing an initial arm according to uniform distribution, algorithm GradSticking gradually increases its probability of repeating the previously chosen arm; the probability that the arm chosen at time $t-1$ is also selected at time t, becomes larger as t becomes larger.

We state the following general facts about the algorithm's mistake probabilities, which are used not only for analysis of the survival game but also for that of the noise-free bandit game in the next section. Note that the lemma holds for a more general adversary than the adversary in the survival game: the survival game adversary corresponds to using $m = 1$ and $k_1 = k$ in the lemma.

Lemma 1. *Let* $0 = t_0 < t_1 < t_2 < \cdots < t_m < t_{m+1} = T + 1$. *Assume the adversary's loss vectors* $\mathbb{1}_{S_t}$ *at time* t *satisfy:*

1. $S_0 = \emptyset$
2. *if* t *is not one of the* t_i*'s then* $S_t = S_{t-1}$
3. *if* t *is one of the* t_i*'s, then* $S_t \supset S_{t-1}$. *Let* $k_i = |S_t| - |S_{t-1}|$.

Algorithm 1: Player GradSticking(T, K, k)

input : T:the number of rounds, K: number of arms,
k: number of 1s in a changed loss vector

initialize: $i_0 \leftarrow 1$

for *time* $t = 1, \ldots, T$ **do**

Select $i_t \in [K]$ as

$$i_t = \begin{cases} i_{t-1} & \text{with probability } \dfrac{1+(t-1)k}{K+(t-1)k} \text{ and} \\ j & \text{with probability } \dfrac{1}{K+(t-1)k} \text{ for } j \neq i_{t-1}. \end{cases}$$

Receive $\ell_{t,i_t} \in \{0, 1\}$.

if $\ell_{t,i_t} = 1$ **then**

return `/* Game over with a mistake */`

end

end

return `/* Game over without a mistake */`

Then, for each time $t = 1, \ldots, T$*, if algorithm GradSticking*(T, K, k) *makes no mistake at times* $1, \ldots, t-1$ *(i.e. conditioning on* $\ell_{t',I_{t'}} = 0$ *for all* $t' < t$*), we have for time* t

$$\textit{its no-mistake probability} \begin{cases} \geq \dfrac{K - \sum_{j=1}^{i} k_j}{K+(t-1)k} & \textit{when } t = t_i \textit{ and} \\ = \dfrac{(t-1)k + K - \sum_{j=1}^{i} k_j}{K+(t-1)k} & \textit{when } t_i < t < t_{i+1}. \end{cases}$$

(Proof) For $t = t_i$, the adversary can set his/her loss vector to $\mathbb{1}_{S_t}$ so that S_t contains the previously chosen arm i_{t-1}. Since the no-mistake probability is minimized in this case, it is at least

$$\frac{K - \sum_{j=1}^{i} k_j}{K+(t-1)k}.$$

For $t_i < t < t_{i+1}$, the assumption that GradSticking(T, K, k) makes no mistake at time $t-1$, means $i_{t-1} \notin S_{t-1}$. Furthermore, $S_{t-1} = S_t$ in this case.

Thus the algorithm's no-mistake probability is

$$\frac{(t-1)k + K - \sum_{j=1}^{i} k_j}{K + (t-1)k}.$$

□

For this algorithm, we obtain a lower bound on the no-mistake probability shown in Theorem 1 below.

Theorem 1. *In the survival game $G(T, K, k)$, algorithm GradSticking(T, K, k) makes no mistake with probability at least* $\dfrac{K-k}{K+(T-1)k}$.

(Proof) Assume that the adversary changes his/her loss vector to $\mathbb{1}_S$ at some time t_0. Then, any k-sized set S that contains i_{t_0-1} maximizes the probability that algorithm GradSticking makes a mistake. For such S, the probability that algorithm GradSticking makes no mistake is (using Lemma 1)

$$\frac{K-k}{K+(t_0-1)k} \cdot \frac{K+(t_0-1)k}{K+t_0k} \cdots \frac{K+(T-2)k}{K+(T-1)k} = \frac{K-k}{K+(T-1)k}.$$

□

Since the no-mistake probability is $((K-k)/K)^T$ when arms are always chosen according to the uniform distribution and the adversary uses $\mathbb{1}_S$ from the beginning, we know that no-mistake probability is improved significantly by gradually increasing the probability of repeating the previously chosen arm. Does Algorithm GradSticking(T, K, k) increase these probabilities in the best possible way?

Consider a player algorithm that chooses arm j with probability $p_{j|i}(t)$ at time t right after arm i is chosen at time $t-1$. The arm-selection probabilities of algorithm GradSticking(T, K, k) satisfies the following two conditions C1 and C2.

C1 The probability $p_{j|i}(t)$ depends on time t and the arm i that is chosen at time $t-1$, but does not depend on the choices before time $t-1$.

C2 $p_{1|1}(t) = p_{2|2}(t) = \cdots = p_{K|K}(t) \geq 1/K$ for all $t \in \{1, 2, \ldots, T\}$.

The following theorem says that GradSticking(T, K, k) is optimal among the algorithms using arm-selection probabilities satisfying conditions C1 and C2, so there is no better schedule for increasing the probability of repeating the previously chosen arm under the constraint of those conditions.

Theorem 2. *For survival game $G(T, K, k)$, any algorithm using arm-selection probabilities that satisfy conditions C1 and C2, makes no mistake with probability at most $\dfrac{K-k}{K+(T-1)k}$ in the worst case.*

(Proof) As in the proof of Theorem 1, consider the adversary's strategy that changes his/her loss vector to $\mathbb{1}_S$ at time t_0. The adversary wants to use set S that minimizes the probability that the player makes no mistake, which is

$$\begin{aligned}&\sum_{i_{t_0},\ldots,i_T \notin S} p_{i_{t_0}|i_{t_0-1}}(t_0) \times \cdots \times p_{i_T|i_{T-1}}(T) \\ &\quad = \sum_{i_{t_0},\ldots,i_{T-1} \notin S} p_{i_{t_0}|i_{t_0-1}}(t_0) \times \cdots \times p_{i_{T-1}|i_{T-2}}(T-1) \times \sum_{i_T \notin S} p_{i_T|i_{T-1}}(T). \quad (1)\end{aligned}$$

Under Condition C1, the minimum of probability $\sum_{i_T \notin S} p_{i_T|i_{T-1}}(T)$ for each $i_{T-1} \notin S$ can be maximized using a probability distribution

$$p_{j|i}(T) = \frac{1-p(T)}{K-1} \text{ for } j \neq i,$$

where $p(t)$ is the probability of sticking, $p(t) \equiv p_{1|1}(t) = \cdots = p_{K|K}(t)$ for $t = 1, \ldots, T$. Then, the right-hand side of Eq. (1) is upper bounded by

$$\begin{aligned}&\sum_{i_{t_0},\ldots,i_{T-1} \notin S} p_{i_{t_0}|i_{t_0-1}}(t) \times \cdots \times p_{i_{T-1}|i_{T-2}}(T-1) \times \left(p(T) + (K-k-1)\frac{1-p(T)}{K-1} \right) \\ &\quad = \sum_{i_{t_0},\ldots,i_{T-1} \notin S} p_{i_{t_0}|i_{t_0-1}}(t) \times \cdots \times p_{i_{T-1}|i_{T-2}}(T-1) \times \frac{kp(T)+(K-k-1)}{K-1} \\ &\quad = \sum_{i_{t_0},\ldots,i_{T-2} \notin S} p_{i_{t_0}|i_{t_0-1}}(t) \times \cdots \times p_{i_{T-2}|i_{T-3}}(T-2) \sum_{i_{T-1} \notin S} p_{i_{T-1}|i_{T-2}}(T-1) \\ &\qquad \times \frac{kp(T)+(K-k-1)}{K-1} \qquad (2)\end{aligned}$$

The minimum of probability $\sum_{i_{T-1} \notin S} p_{i_{T-1}|i_{T-2}}(T-1)$ for each $i_{T-2} \notin S$ can be maximized similarly, and applying the same argument repeatedly, the right-hand side of Eq. (2) (and thus the probability of no mistake) is upper bounded by

$$(K-k)\frac{1-p(t_0)}{K-1}\prod_{t=t_0+1}^{T}\frac{kp(t)+(K-k-1)}{K-1} \equiv q(t_0). \tag{3}$$

Note that $\sum_{i_{t_0} \notin S} p_{i_{t_0}|i_{t_0-1}} \leq (K-k)((1-p(t_0))/(K-1))$ even for the best player's strategy because the adversary can set S so as to include arm i_{t_0-1}. Let $q(t_0)$ denote Expression (3). The player wants to maximize $\min_{t_0=1,\ldots,T} q(t_0)$. Only $q(1)$ depends on probability $p(1)$, and $q(1)$ is maximized as a function of $p(1) \in [1/K, 1]$ by setting $p(1) = 1/K$. Assume that $\min\{q(1), \ldots, q(t')\}$ is maximized by setting $p(t) = (1+(t-1)k)/(K+(t-1)k)$ for $t = 1, \ldots, t'$. Consider maximization problem of $\min\{q(1), \ldots, q(t'+1)\}$. This problem is equivalent to the maximization problem of

$$\min\{\min\{q(1), \ldots, q(t')\}, q(t'+1)\}.$$

Probability $q(t'+1)$ does not depend on $p(t)$ for $t = 1, \ldots, t'$, so the maximization in terms of $p(t)$ for $t = 1, \ldots, t'$ affects only the maximization of $\min\{q(1), \ldots, q(t')\}$, which is done by setting $p(t) = (1+(t-1)k)/(K+(t-1)k)$ for $t = 1, \ldots, t'$ from the assumption. Note that, in such setting,

$$\begin{aligned}\min\{q(1), \ldots, q(t')\} =& q(1) = \cdots = q(t') \\ =& \frac{K-k}{K+(t'-1)k}\prod_{t=t'+1}^{T}\frac{kp(t)+(K-k-1)}{K-1}\end{aligned}$$

holds. As a function of $p(t'+1)$, $\min\{q(1), \ldots, q(t')\}$ is increasing but $q(t'+1)$ is decreasing, so $\min\{\min\{q(1), \ldots, q(t')\}, q(t'+1)\}$ is maximized when $\min\{q(1), \ldots, q(t')\} = q(t'+1)$, that is, in the case with

$$\frac{1}{K+(t'-1)k} \cdot \frac{kp(t'+1)+(K-k-1)}{K-1} = \frac{1-p(t'+1)}{K-1}.$$

By solving this equation, we get $p(t'+1) = (1+t'k)/(K+t'k)$. Therefore, by mathematical induction, we know that $\min\{q(1), \ldots, q(T)\}$ is maximized

by setting $p(t) = (1 + (t-1)k)/(K + (t-1)k)$ for $t = 1, \ldots, T$, and in that case,

$$\min\{q(1), \ldots, q(T)\} = \frac{K-k}{K+(T-1)k}$$

holds. □

Next, we prove an upper bound on the no-mistake probability for the survival game $G(T, K, k)$ by analyzing an adversary's algorithm that generates loss vectors $\boldsymbol{\ell}_t$ adaptively. Algorithm Wait&Sticking(T, K, k) (Algorithm 2) is an adversary algorithm for survival game $G(T, K, k)$. At each time $t \in [T]$, if the previous loss vector was the waiting loss vector, $\boldsymbol{\ell}_{t-1} = \mathbb{0}$, then it examines the algorithm to calculate the best time $c^* \in [t, T]$ to change the loss vector. If this best time is the current time ($c^* = t$), then it sets $\boldsymbol{\ell}_t$ to the loss vector $\boldsymbol{\ell}_{c^*}$ minimizing the no-mistake probability. Here, the best time c^* to change the loss vector is that time $c \in [t, T]$ where the no-mistake probability can be minimized by changing to the best (for the adversary) loss vector at time c and sticking to it from then on.

We use the following function in the statement of the theorem: for three natural numbers T, K, k, define $F(T, K, k)$ as

$$F(T, K, k) \equiv \frac{\binom{K-1}{k}}{T + \binom{K}{k} - 1}.$$

Theorem 3. *In the survival game $G(T, K, k)$, the Wait&Sticking(T, K, k) adversary forces the no-mistake probability of any player algorithm to be at most $F(T, K, k)$.*

(Proof) Consider the probability of finishing without a mistake after we fix the set S of lossy arms during time b. Let $R = T - b + 1$ (so $T = b + R - 1$) denote the remaining time, including time b, and

$$p_{b,S}(\boldsymbol{o}[1..b-1], R) = P\left\{ \sum_{t=b}^{b+R-1} \ell_{t,I_t} = 0 \,\middle|\, \begin{array}{l} \boldsymbol{O}[1..b-1] = \boldsymbol{o}[1..b-1] \\ \boldsymbol{\ell}_b = \cdots = \boldsymbol{\ell}_{b+R-1} = \mathbb{1}_S \end{array} \right\}$$

for any natural number b, observation sequence $\boldsymbol{o}[1..b-1] \in ([K] \times \{0\})^{b-1}$ and $S \subseteq [K]$. For $c \in [b..b+R-1]$, let $p_c(\boldsymbol{o}[1..b-1], R)$ denote the value p_c

Algorithm 2: Adversary Wait&Sticking(T, K, k)

recall that $\boldsymbol{O}$ is the random variable for the loss observations, $\boldsymbol{o}$ is its realization, $\boldsymbol{\ell}_t$ is the loss vector at time t, and I is the sequence of arm choices.

input : T: number of rounds, K: number of arms,
k: number of 1s in a changed loss vector

$S_* \leftarrow \emptyset$; `/* best arms to assign loss, initially undecided */`
for *time* $t = 1, \ldots, T$ **do**
  if $S_* = \emptyset$ **then** `/* still waiting */`
    $p_{\min} = 2$
    for $c = t, \ldots, T$ **do**
      $p_c \leftarrow$

$$E_{I[t..T]}\left[\min_{\substack{S \subseteq [K] \\ |S| = k}} P\left\{\sum_{s=c}^{T} \ell_{s,I_s} = 0 \;\middle|\; \begin{array}{l} \boldsymbol{O}[1..t-1] = \boldsymbol{o}[1..t-1], \\ I[t..c-1], \\ \boldsymbol{\ell}_t = \cdots = \boldsymbol{\ell}_{c-1} = \mathbb{0}, \\ \boldsymbol{\ell}_c = \cdots = \boldsymbol{\ell}_T = \mathbb{1}_S \end{array}\right\}\right]$$

      if $p_c < p_{min}$ **then**
        $c^* \leftarrow c$, $p_{\min} \leftarrow p_c$
      end
    end
    if $c^* = t$ **then** `/* stick now */`
      $S_* \leftarrow$

$$\arg\min_{S \subseteq [K], |S| = k} P\left\{\sum_{s=t}^{T} \ell_{s,I_s} = 0 \;\middle|\; \begin{array}{l} \boldsymbol{O}[1..t-1] = \boldsymbol{o}[1..t-1], \\ \boldsymbol{\ell}_t = \cdots = \boldsymbol{\ell}_T = \mathbb{1}_S \end{array}\right\}$$

    end
  end
  $\boldsymbol{\ell}_t \leftarrow \mathbb{1}_{S_*}$
  Observe the player's choice i_t
  if $\ell_{t,i_t} = 1$ **then**
    return `/* Game over` ***with*** `a mistake */`
  end
end
return `/* Game over` ***without*** `a mistake */`

that is set in Wait&Sticking$(b + R - 1, K, k)$ at time b, namely,

$p_c(\boldsymbol{o}[1..,b-1], R)$

$$=E_{I[b..b+R-1]}\left[\min_{S\subseteq[K],|S|=k} P\left\{\sum_{t=c}^{b+R-1}\ell_{t,I_t}=0 \;\middle|\; \begin{array}{l}\boldsymbol{O}[1..b-1]=\boldsymbol{o}[1..b-1],\\ I[b..c-1],\\ \boldsymbol{\ell}_b=\cdots=\boldsymbol{\ell}_{c-1}=\mathbb{0},\\ \boldsymbol{\ell}_c=\cdots=\boldsymbol{\ell}_{b+R-1}=\mathbb{1}_S\end{array}\right\}\right].$$

Then,

$$\begin{aligned}&p_{c^*}(\boldsymbol{o}[1..0],T)=\min_{c\in[T]}p_c(\boldsymbol{o}[1..0],T)\\ \geq&P\left\{\sum_{t=1}^{T}\ell_{t,I_t}=0\middle|\,\text{Wait\&Sticking}(T,K,k)\text{ chooses }\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T\right\}\end{aligned}$$

holds. We prove that

$$p_{c^*}(\boldsymbol{o}[1..b-1],R)\leq F(R,K,k)$$

holds for any natural number b, observation sequence $\boldsymbol{o}[1..b-1]\in([K]\times\{0\})^{b-1}$ and for any natural number R of remaining time. It suffices to prove the inequality

$$\sum_{S\subseteq[K],|S|=k}p_{b,S}(\boldsymbol{o}[1..b-1],R)+\sum_{c=b+1}^{b+R-1}p_c(\boldsymbol{o}[1..b-1],R)\leq\binom{K-1}{k}\qquad(4)$$

because $p_{c^*}(\boldsymbol{o}[1..b-1],R)$ is the minimum of the terms on the left hand side of the inequality, and the average of these terms is at most

$$\binom{K-1}{k}\Big/\left(R+\binom{K}{k}-1\right)$$

when Ineq. (4) holds.

We prove Ineq. (4) by mathematical induction on R for any fixed b. When $R=1$, for any natural number b and any $\boldsymbol{o}[1..b-1]\in([K]\times\{0\})^{b-1}$,

$$\begin{aligned}p_{b,S}(\boldsymbol{o}[1..b-1],1)=&P\left\{\ell_{b,I_b}=0\;\middle|\;\begin{array}{l}\boldsymbol{O}[1..b-1]=\boldsymbol{o}[1..b-1]\\ \boldsymbol{\ell}_b=\mathbb{1}_S\end{array}\right\}\\ =&P\left\{I_b\notin S\mid\boldsymbol{O}[1..b-1]=\boldsymbol{o}[1..b-1]\right\}\end{aligned}$$

holds. Thus, Ineq. (4) holds because

$$\begin{aligned}&\sum_{S\subseteq[K],|S|=k}p_{b,S}(\boldsymbol{o}[1..b-1],1)+\sum_{c=b+1}^{b}p_c(\boldsymbol{o}[1..b-1],1)\\ =&\sum_{S\subseteq[K],|S|=k}P\left\{I_b\notin S\mid\boldsymbol{O}[1..b-1]=\boldsymbol{o}[1..b-1]\right\}=\binom{K-1}{k}.\end{aligned}$$

Here, the last equality holds because there are just $\binom{K-1}{k}$ size-k subsets of $[K] \setminus \{i\}$ for each arm $i \in [K]$. Assume that Ineq. (4) holds for any natural number b when there is R time remaining. When there is $R+1$ time remaining we have:

$$
\begin{aligned}
&\sum_{S\subseteq[K],|S|=k} p_{b,S}(\boldsymbol{o}[1..b-1], R+1) + \sum_{c=b+1}^{b+R} p_c(\boldsymbol{o}[1..b-1], R+1) \\
&= \sum_{S\subseteq[K],|S|=k} \sum_{i\notin S} P\{I_b = i\} p_{b+1,S}((\boldsymbol{o}[1..b-1], (i,0)), R) \\
&\quad + \sum_{c=b+1}^{b+R} \sum_{i\in[K]} P\{I_b = i\} p_c((\boldsymbol{o}[1..b-1], (i,0)), R) \\
&= \sum_{i\in[K]} P\{I_b = i\} \times \\
&\quad \left(\sum_{i\notin S\subseteq[K],|S|=k} p_{b+1,S}((\boldsymbol{o}[1..b-1], (i,0)), R) \right. \\
&\qquad \left. + \min_{S\subseteq[K]} p_{b+1,S}((\boldsymbol{o}[1..b-1], (i,0)), R) + \sum_{c=b+2}^{b+R} p_c((\boldsymbol{o}[1..b-1], (i,0)), R) \right) \\
&\leq \sum_{i\in[K]} P\{I_b = i\} \times \\
&\quad \left(\sum_{S\subseteq[K],|S|=k} p_{b+1,S}((\boldsymbol{o}[1..b-1], (i,0)), R) + \sum_{c=b+2}^{b+R} p_c((\boldsymbol{o}[1..b-1], (i,0)), R) \right) \\
&\leq \sum_{i\in[K]} P\{I_b = i\} \binom{K-1}{k} = \binom{K-1}{k}
\end{aligned}
$$

holds, where the last inequality is due to the assumption that Ineq. (4) holds when there is R time remaining. Therefore, Ineq. (4) holds for all natural numbers b and R. □

Remark 1. *By Theorem 1 and 3, the minimax no-mistake probability* $P^*(T, K, k)$

for survival game $G(T, K, k)$ can be bounded as

$$\frac{1}{1+\dfrac{kT}{K-k}} \leq P^*(T,K,k) \leq \frac{1}{1+\dfrac{T-1}{\binom{K-1}{k}}+\dfrac{k}{K-k}}.$$

The upper bound coincides with the lower bound when $k = 1$; both the values are $(K-1)/(T+K-1)$. However, the difference between the coefficients of T in the denominators can be significant when $K \gg k \geq 2$:

$$\frac{k}{K-k} \approx \frac{k}{K} \text{ while } \frac{1}{\binom{K-1}{k}} \approx \frac{k!}{K^k}.$$

4. Mistake Bounds for the Noise-Free Bandit Game

In the survival game $G(T, K, k)$, the adversary is restricted to change his/her loss vector just once, from the zero vector to a vector with k 1's. How much does the no-mistake probability increase when this restriction is removed? We answer this question for player GradSticking$(T, K, 1)$ in Lemma 2, which will be used to prove an upper bound on the expected number of mistakes for the noise-free multi-armed bandit game. The following proposition (proven in Appendix A) is necessary for proving the lemma. Here m represents the number of times that the adversary augments the loss vector.

Proposition 1. *For any integers $m \geq 1$, $K > m$, and $T \geq m$,*

$$\begin{aligned}&\prod_{i=1}^{m}\left[\left(\frac{K-\sum_{j=1}^{i}k_j}{K+t_i-1}\right)\prod_{t_i<t<t_{i+1}}\left(\frac{K-\sum_{j=1}^{i}k_j+t-1}{K+t-1}\right)\right]\\ =&\frac{\prod_{i=1}^{m}\left[\left(K-\sum_{j=1}^{i}k_j\right)\prod_{h=1}^{k_i-1}\left(K-\sum_{j=1}^{i}k_j+t_i-1+h\right)\right]}{\prod_{h=K-\beta}^{K-1}(T+h)}\end{aligned} \tag{5}$$

holds for any integers $k_1, ..., k_m \geq 1$ with $\sum_{i=1}^{m} k_i = \beta < K$ and any integers $1 \leq t_1 < \cdots < t_m < t_{m+1} \equiv T+1$.

For notational convenience, define $Q(K,T,\alpha)$ as

$$Q(K,T,\alpha) = \prod_{h=\alpha}^{K-1} \frac{h}{T+h} = \prod_{h=1}^{T} \frac{\alpha-1+h}{K-1+h}$$

for any natural numbers K, T and $\alpha < K$. Note that asymptotic behavior of $Q(K,T,\alpha)$ with respect to T is $\Theta(1/T^{K-\alpha})$ for fixed K, and that with respect to K is $\Theta(1/K^T)$ for fixed T.

Lemma 2. *Algorithm GradSticking$(T,K,1)$ makes no mistakes over T trials with probability at least $Q(K,T,\alpha)$ in the K arm, T round noise-free multi-armed bandit game with α lossless arms.*

(Proof) Until it makes a mistake, Algorithm GradSticking$(T,K,1)$ is symmetric in its choice of arms: at each time t the previously chosen arm has the same probability of being re-pulled regardless of which arm was previously chosen, and each other arm has the same lesser probability of being chosen. We exploit these symmetries in the following argument.

Consider an arbitrary adversary. Without loss of generality, we assume that the adversary always sets the loss of polluted arms to 1. Let $t_1 < t_2 < \ldots < t_m$ be the m times that the adversary pollutes arms (sets an arms loss to 1 for the first time), and for $1 \leq i \leq m$ let k_i be the number of newly-polluted arms at time t_i. Although technically m and the t_i and k_i values may be random variables depending on the adversaries randomization or the the particular arms chosen previously by the algorithm, we will bound the probability of no mistake for each realization, and thus the average of any distribution induced by a particular adversary.

Since the number of clean arms at time $t_i + 1$ is $K - \sum_{j=1}^{i} k_j \geq \alpha$, by Lemma 1 the probability that the algorithm makes no mistake at time t_i, given that it has not previously made a mistake, is at least $\frac{K-\sum_{j=1}^{i} k_j}{K+t_i-1}$. Similarly, the probability that the algorithm makes no mistake at some time t between t_i and t_{i+1} (when no arms become polluted) is at least $\frac{K-\sum_{j=1}^{i} k_j+t-1}{K+t-1}$. Furthermore, the probability that the algorithm makes no mistake at times before t_1 is 1. Note that we define $t_{m+1} \equiv T+1$ for notational convenience. Therefore, the probability the algorithm never makes a mistake over all T trials is at least

$$\prod_{i=1}^{m} \left\{ \left(\frac{K - \sum_{j=1}^{i} k_j}{K + t_i - 1} \right) \prod_{t_i < t < t_{i+1}} \left(\frac{K - \sum_{j=1}^{i} k_j + t - 1}{K + t - 1} \right) \right\}. \qquad (6)$$

Algorithm 3: Player GradStickingSub(b, e, A_D)

input : b: beginning time, e: ending time, A_D: dirty arm set
output : t: game-over time, i_t: 1-loss arm
initialize: $i_{b-1} \leftarrow$ 1st arm in $[K] \setminus A_D$, $K_{\overline{D}} \leftarrow K - |A_D|$

for *time* $t = b, \ldots, e$ **do**
  Select $i_t \in [K] \setminus A_D$ as

$$i_t = \begin{cases} i_{t-1} & \text{with probability } \dfrac{1+t-b}{K_{\overline{D}}+t-b} \text{ and} \\ j & \text{with probability } \dfrac{1}{K_{\overline{D}}+t-b} \text{ for } j \neq i_{t-1}. \end{cases}$$

  Receive $\ell_{t,i_t} \in \{0, 1\}$.
  if $\ell_{t,i_t} = 1$ **then**
    return (t, i_t) `/* Game over with a mistake */`
  end
end
return (e, i_e) `/* Game over without a mistake */`

If $\sum_{j=1}^{m} k_j < K-\alpha$, the value of Expression (6) can be decreased by increasing any of $k_1, \ldots, k_m$. Thus, the value is minimized when $\sum_{j=1}^{m} k_j = K - \alpha$. In that case, by Proposition 1, it is equal to

$$\frac{\prod_{i=1}^{m} \left(\left(K - \sum_{j=1}^{i} k_j \right) \prod_{h=1}^{k_i - 1} \left(K - \sum_{j=1}^{i} k_j + t_i - 1 + h \right) \right)}{\prod_{h=\alpha}^{K-1} (T+h)}.$$

Since $t_i \geq 1$ for all $i = 1, \ldots, m$, this is lower bounded by

$$\prod_{h=\alpha}^{K-1} \frac{h}{T+h} = Q(K, T, \alpha).$$

□

Player Algorithm GradStickingSub(b, e, A_D) (Algorithm 3) is a version of GradSticking$(T, K, 1)$ in which the interface is modified so as to be usable as a subroutine. In GradStickingSub(b, e, A_D), the set of dirty arms A_D is given as input, and the algorithm prevents those dirty arms from being chosen. The following corollary of Lemma 2 holds trivially.

Algorithm 4: Player RepGradSticking(T, K)

parameter: T: number of trials, K: number of arms
initialize : $t \leftarrow 0$, $A_D \leftarrow \emptyset$

repeat
  $(t, i) \leftarrow$ GradStickingSub$(t+1, T, A_D)$
  $A_D \leftarrow A_D \cup \{i\}$
until $t = T$;

Corollary 1. *In the noise-free K-armed bandit game with α noiseless arms, let A_D be the set of dirty arms at the beginning of trial b. Then, from time b to e, Algorithm GradStickingSub(b, e, A_D) makes no mistakes with probability at least $Q(K_{\overline{D}}, T', \alpha)$, where $K_{\overline{D}} = K - |A_D|$ is the number of non-dirty arms and $T' = e - b + 1$ is the number of rounds.*

Player algorithm RepGradSticking (Algorithm 4) repeatedly calls GradStickingSub. Each call to Algorithm GradStickingSub returns at the end time T or the first time $t(< T)$ when a mistake is made. In the latter case, the arm i that became dirty due to the mistake is added to the dirty arm set A_D, and GradStickingSub is re-called with the beginning time $t+1$ and the new dirty arm set.

The following theorem shows an upper bound on the expected number of mistakes for the noise-free multi-armed bandit game.

Theorem 4. *The expected number of mistakes made by player algorithm RepGradSticking(T, K) is at most*

$$\begin{aligned} &\sum_{j=1}^{\min\{T, K-\alpha\}} \prod_{i=1}^{j} (1 - Q(K-i+1, T-i+1, \alpha)) \\ \leq &\min\{T, K-\alpha\} - \sum_{j=1}^{\min\{T, K-\alpha\}} Q(K-j+1, T-j+1, \alpha) \end{aligned} \tag{7}$$

in noise-free multi-armed bandit game with α lossless arms.

(Proof) By Corollary 1, algorithm RepGradSticking(T, K) makes a mistake at least once with probability at most $1 - Q(K, T, \alpha)$ by calling GradStickingSub$(1, T, \emptyset)$.

With the same argument, it further makes a mistake at least once more with probability at most $(1-Q(K,T,\alpha))(1-Q(K-1,T-1,\alpha))$ by calling GradStickingSub the second time. Continuing the same argument, algorithm RepGradSticking(T,K) makes at most

$$\sum_{j=1}^{\min\{T,K-\alpha\}} \prod_{i=1}^{j}(1-Q(K-i+1,T-i+1,\alpha)) \leq T - \sum_{j=1}^{\min\{T,K-\alpha\}} Q(K-j+1,T-j+1,\alpha)$$

mistakes. □

Remark 2. *Note that the dominant non-constant minus-term of the right-hand side of Ineq. (7) is $Q(\alpha+1,T-K+\alpha+1,\alpha)=\alpha/(T-K+2\alpha+1)$ when $T \geq K-\alpha$, and $Q(K-T+1,1,\alpha)=\alpha/(K-T+1)$ otherwise. Thus, the righthand side of Ineq. (7) is asymptotically $K-\alpha-\Omega(\alpha/T)$ for fixed K and $T-\Omega(\alpha/K)$ for fixed T.*

Remark 3. *From the way of the proof of Theorem 4, we know that player algorithm RepGradSticking(T,K) suffers at most k mistakes with probability at least $1-\prod_{i=1}^{k+1}(1-Q(K-i+1,T-i+1,\alpha))$ for $0 \leq k \leq \max\{T-1,K-\alpha-1\}$. Note that the dominant term of this probability lower bound is $Q(K-k,T-k,\alpha)=\prod_{h=\alpha}^{K-k-1}\frac{h}{T-k+h}=\Theta(1/T^{K-k-\alpha})$, which rapidly converges to 0 when $T\to\infty$ for any $k<K-\alpha$.*

Next we give our adversary algorithm for noise-free multi-armed bandit problem. It is based on the adversary algorithm for the survival game and uses information about the player through expectations over its arm choices.

Adversary Algorithm Wait&StickingSub(b,e,A_P,k) (Algorithm 5) is a version of Wait&Sticking(T,K,k) that is modified so as to be usable during time period $[b..e] \subseteq [T]$ in the noise-free K-armed bandit game with T rounds. During time period $[b..e]$, Wait&StickingSub changes its loss vector only once, from $\mathbb{1}_{A_P}$ to $\mathbb{1}_{A_P \cup S_*}$ where S_* is a k-sized set of clean arms at time b. For the player, this game is more difficult than the survival game $G(e-b+1,K-|A_P|,k)$ when there is at least one polluted but non-dirty arm. However, we obtain the following corollary of Theorem 3 without exploiting this added difficulty.

Algorithm 5: Adversary Wait&StickingSub(b, e, A_P, k)

input : b: beginning time, e: ending time, A_P: polluted arm set, k: number of 1s in a changed loss vector

output : t: game-over time, S_*: set of arms whose loss is set to 1

initialize: $A_C \leftarrow [K] \setminus A_P$

$S_* \leftarrow \emptyset$

for *time* $t = b, \ldots, e$ **do**

  if $S_* = \emptyset$ **then**

    $p_{\min} = 2$

    for $c = t, \ldots, T$ **do**

      $p_c \leftarrow$

      $$E_{I[t..T]}\left[\min_{\substack{S \subseteq A_C \\ |S| = k}} P\left\{\sum_{s=c}^{T} \ell_{s,I_s} = 0 \;\middle|\; \begin{array}{l} \boldsymbol{O}[1..t-1] = \boldsymbol{o}[1..t-1], \\ I[t..c-1], \\ \boldsymbol{\ell}_t = \cdots = \boldsymbol{\ell}_{c-1} = \mathbb{1}_{A_P}, \\ \boldsymbol{\ell}_c = \cdots = \boldsymbol{\ell}_T = \mathbb{1}_{A_P \cup S} \end{array}\right\}\right]$$

      if $p_c < p_{min}$ **then**

        $c^* \leftarrow c$, $p_{\min} \leftarrow p_c$

      end

    end

    if $c^* = t$ **then**

      $S_* \leftarrow$

      $$\arg\min_{S \subseteq A_C, |S| = k} P\left\{\sum_{s=t}^{T} \ell_{s,I_s} = 0 \;\middle|\; \begin{array}{l} \boldsymbol{O}[1..t-1] = \boldsymbol{o}[1..t-1], \\ \boldsymbol{\ell}_t = \cdots = \boldsymbol{\ell}_T = \mathbb{1}_{A_P \cup S} \end{array}\right\}$$

    end

  end

  $\boldsymbol{\ell}_t \leftarrow \mathbb{1}_{A_P \cup S_*}$

  Observe the player's choice i_t

  if $\ell_{t,i_t} = 1$ **then**

    return (t, S_*) `/* Game over with a mistake */`

  end

end

return (e, S_*) `/* Game over without a mistake */`

Corollary 2. *In the the noise-free K-armed bandit game with T rounds, adversary algorithm Wait&StickingSub(b, e, A_P, k) forces any player algorithm to have a no-mistake probability during time period $[b..e]$ upper bounded by*

$F(e-b+1, K-|A_P|, k)$.

Algorithm RepeatW&S(K,T) (Algorithm 6) is an adversary algorithm for the noise-free multi-armed bandit game. First, the algorithm partitions the whole range of times $[1..T]$ into m time periods $[1..t_1-1], [t_1..t_2-1], \ldots, [t_{m-1}..T]$. To simplify the indexing, we define $t_0 \equiv 1$ and $t_m \equiv T+1$. The algorithm also allocates a number of arms to each time period by picking numbers $k_1, \ldots, k_m$ with $\sum_{i=1}^{m} k_i = K - \alpha$. Then, for each pair $([t_{i-1}..t_i - 1], k_i)$, the algorithm calls Wait&StickingSub$(t_{i-1}, t_i - 1, A_P, k_i)$ where A_P is the set of arms polluted so far. When Wait&StickingSub returns back before time $t_i - 1$, RepeatW&S updates the set A_P of polluted arms and uses the loss vector $\boldsymbol{\ell}_t = \mathbb{1}_{A_P}$ for the remainder of the period.

The solution to the following optimization problem is used by the adversary to obtain a partition of the whole time period $[1..T]$ and a division of $K-1$ that are difficult for any algorithm.

Problem 1. *Given two integers $T \geq 1$ and $K \geq 2$, find two natural number sequences $t_0, \ldots, t_m$ and $k_1, \ldots, k_m$ that minimize*

$$\sum_{i=1}^{m} F(t_i - t_{i-1}, K - \sum_{j=1}^{i-1} k_j, k_i)$$

subject to

$$1 \leq m \leq K - \alpha, \tag{8}$$

$$1 = t_0 < t_1 < \cdots < t_m = T + 1 \text{ and} \tag{9}$$

$$k_1 + \cdots + k_m = K - \alpha. \tag{10}$$

The following theorem gives a lower bound on the expected number of mistakes in the nose-free multi-armed bandit game.

Theorem 5. *The adversary algorithm RepeatW&S(K,T) forces the expected number of mistakes made by any player algorithm to be at least*

$$m - \sum_{i=1}^{m} F(t_i - t_{i-1}, K - \sum_{j=1}^{i-1} k_j, k_i)$$

for any positive integers m, $t_0, \ldots, t_m$, $k_1, \ldots, k_m$ satisfying (8),(9) and (10).

Algorithm 6: Adversary RepeatW&S(K, T)

parameter: K: number of arms, T: number of trials
initialize : $t_0, \dots, t_m, k_1, \dots, k_m \leftarrow$ the solution of Problem 1,
$A_P \leftarrow \emptyset$

for $i = 1, \dots, m$ **do**
  $(t', S) \leftarrow$ Wait&StickingSub$(t_{i-1}, t_i - 1, A_P, k_i)$
  $A_P \leftarrow A_P \cup S$
  for $t = t', \dots, t_i - 1$ **do**
    $\boldsymbol{\ell}_t \leftarrow \mathbb{1}_{A_P}$
    Observe the player's choice i_t
  end
end

(Proof) By Corollary 2, within each call of Wait&StickingSub$(t_{i-1}, t_i - 1, A_P, k_i)$, the expected number of mistakes made by any player algorithm is at least $1 - F(t_i - t_{i-1}, K - \sum_{j=1}^{i-1} k_j, k_i)$. Thus the total expected number of mistakes is at least $m - \sum_{i=1}^{m} F(t_i - t_{i-1}, K - \sum_{j=1}^{i-1} k_j, k_i)$. □

The following corollary says that the upper and lower bounds shown in this section are equal when $\alpha = K - 1$.

Corollary 3. *The minimax expected number of mistakes for the noise-free K-armed bandit game with T rounds and $\alpha = K - 1$ lossless arms is*

$$1 - \frac{K-1}{T+K-1}.$$

(Proof) By Theorem 4, the minimax expected number of mistakes is upper bounded by

$$1 - Q(K, T, \alpha) = 1 - \frac{K-1}{T+K-1}$$

for any $T \geq 1$ and $K \geq 2$ in the $(K-1)$-lossless-arm case ($\alpha = K - 1$). For $\alpha = K - 1$, natural numbers $m, t_0, \dots, t_m, k_1, \dots, k_m$ that satisfy (8),(9) and (10) are uniquely determined: $m = 1$, $t_0 = 1$, $t_1 = T + 1$, $k_1 = 1$. So by Theorem 5, the minimax expected number of mistakes is lower bounded by

$$1 - F(T, K, 1) = 1 - \frac{K-1}{T+K-1}.$$

□

For general K larger than 2, concrete lower bounds on the expected number of mistakes for the noise-free K-armed bandit game are shown in the following two corollaries which can be derived from Theorem 5.

Corollary 4. *RepeatW&S[K, T] forces any player algorithm to make an expected number of mistakes at least*

$$T\left(1-\left(\frac{\alpha}{K}\right)^{1/T}\right)-\frac{\left(\frac{K}{\alpha}\right)^{(T-1)/T}-1}{K\left(\left(\frac{K}{\alpha}\right)^{1/T}-1\right)} \tag{11}$$

when $T \leq K-\alpha-1$, and, for each $1 \leq h \leq K-\alpha$, at least

$$H_K - H_{\alpha+h} + h - \frac{A^2(h)(B(h)+4h)}{2B^2(h)} \tag{12}$$

when $T \geq \frac{h(h-1)}{2} + K - \alpha$. Here H_n is the nth harmonic number and

$$\begin{aligned} A(h) =& 2\sum_{j=1}^{h}\sqrt{\alpha+j-1} \\ B(h) =& 2T-2(K-\alpha)+(2\alpha+h+1)h. \end{aligned}$$

(Proof) For the first bound we have $T \leq K-\alpha-1$. Let $m=T$, and consider positive integers $t_0,\ldots,t_m$ satisfying

$$t_0=1, t_1=2, t_2=3, ..., t_{m-1}=T, t_m=T+1.$$

Then, starting from the expected mistake bound in Theorem 5,

$$\begin{aligned} m-\sum_{i=1}^{m}F(t_i-t_{i-1},K-\sum_{j=1}^{i-1}k_j,k_i) =& T-\sum_{i=1}^{T}F(1,K-\sum_{j=1}^{i-1}k_j,k_i) \\ =& T-\sum_{i=1}^{T}\frac{\binom{K-\sum_{j=1}^{i-1}k_j-1}{k_i}}{\binom{K-\sum_{j=1}^{i-1}k_j}{k_i}} \\ =& T-\sum_{i=1}^{T}\frac{K-\sum_{j=1}^{i}k_j}{K-\sum_{j=1}^{i-1}k_j} \end{aligned}$$

expected mistakes are forced. By the inequality of arithmetic and geometric means, we have

$$T - \sum_{i=1}^{T} \frac{K - \sum_{j=1}^{i} k_j}{K - \sum_{j=1}^{i-1} k_j} \leq T - T \left(\frac{\alpha}{K}\right)^{1/T}$$

with equality if and only if

$$\frac{K - \sum_{j=1}^{i} k_j}{K - \sum_{j=1}^{i-1} k_j} = \left(\frac{\alpha}{K}\right)^{1/T} \tag{13}$$

for all $i = 1, \ldots, T$. Unfortunately, $k_1, \ldots, k_T$ that satisfy (13) are not integers. As an approximate solution, use $k_1, \ldots, k_T$ that satisfy

$$K - \sum_{j=1}^{i} k_j = \left\lceil \alpha^{i/T} K^{(T-i)/T} \right\rceil,$$

then

$$\begin{aligned}
T - \sum_{i=1}^{T} \frac{K - \sum_{j=1}^{i} k_j}{K - \sum_{j=1}^{i-1} k_j} =& T - \sum_{i=1}^{T} \frac{\left\lceil \alpha^{i/T} K^{(T-i)/T} \right\rceil}{\left\lceil \alpha^{(i-1)/T} K^{(T-i+1)/T} \right\rceil} \\
\geq& T - \sum_{i=1}^{T-1} \frac{\alpha^{i/T} K^{(T-i)/T} + 1}{\alpha^{(i-1)/T} K^{(T-i+1)/T}} - \frac{\alpha}{\alpha^{(T-1)/T} K^{1/T}} \\
=& T \left(1 - \left(\frac{\alpha}{K}\right)^{1/T}\right) - \frac{\left(\frac{K}{\alpha}\right)^{(T-1)/T} - 1}{K \left(\left(\frac{K}{\alpha}\right)^{1/T} - 1\right)}
\end{aligned}$$

completing the proof of the first bound.

For the second bound, consider an arbitrary integer h in $[1, K - \alpha]$ and assume that $T \geq \frac{h(h-1)}{2} + K - \alpha$. We will partition the times into $m = K - \alpha$ periods where one arm will become polluted in each period and the first $K - h - \alpha$ periods are length 1. More precisely, let the t_i period boundaries be

$$t_i = \begin{cases} t_{i-1} + 1 & (i = 1, \ldots, K - \alpha - h) \\ t_{i-1} + T_i + 1 & (i = K - \alpha - (h-1), \ldots, K - \alpha) \end{cases}$$

where the T_i's are non-negative integers to be optimized later subject to

$$\sum_{i=1}^{h} T_{K-\alpha-(i-1)} = T - (K - \alpha). \tag{14}$$

We define $t_0 \equiv 1$ for convenience, and set the arm budgets for each period to 1, i.e. $k_i = 1$ for $i \in [m]$.

Then, from the theorem, the adversary forces every player to have an expected number of mistakes at least

$$\begin{aligned}
&m - \sum_{i=1}^{m} F(t_i - t_{i-1}, K - \sum_{j=1}^{i-1} k_j, k_i) \\
=& \sum_{i=1}^{K-\alpha-h} (1 - F(1, K-i+1, 1)) + h - \sum_{i=K-\alpha-(h-1)}^{K-\alpha} F(T_i + 1, K-i+1, 1) \\
=& \sum_{i=1}^{K-\alpha-h} \frac{1}{K-i+1} + h - \sum_{i=K-\alpha-(h-1)}^{K-\alpha} \frac{K-i}{K-i+1+T_i} \\
=& H_K - H_{\alpha+h} + h - \sum_{i=1}^{h} \frac{\alpha+i-1}{\alpha+i+T_{K-\alpha-(i-1)}} \qquad (15)
\end{aligned}$$

holds. Let

$$f(T_{K-\alpha-(h-1)}, \ldots, T_{K-\alpha}) = \sum_{i=1}^{h} \frac{\alpha+i-1}{\alpha+i+T_{K-\alpha-(i-1)}}$$

By solving the problem of maximizing $f(T_{K-\alpha-(h-1)}, \ldots, T_{K-\alpha})$ subject to Constraint (14) using the method of Lagrange multipliers, we obtain

$$T_{K-\alpha-(i-1)} = \frac{B(h)\sqrt{\alpha+i-1}}{A(h)} - (\alpha+i) \text{ for } i = 1, \ldots, h. \qquad (16)$$

All the $T_{K-\alpha-(i-1)}$ are non-negative because

$$\begin{aligned}
&\frac{B(h)\sqrt{\alpha+i-1}}{A(h)} - (\alpha+i) \\
=& \frac{2T - 2(K-\alpha) + (2\alpha+h+1)h}{2\sum_{j=1}^{h}\sqrt{\alpha+j-1}}\sqrt{\alpha+i-1} - (\alpha+i) \\
\geq& \frac{2(\alpha+h)h}{2\sum_{j=1}^{h}\sqrt{\alpha+j-1}}\sqrt{\alpha+i-1} - (\alpha+i) \\
\geq& \frac{2(\alpha+h)h}{h\sqrt{2(2\alpha+h-1)}}\sqrt{\alpha+i-1} - (\alpha+i)
\end{aligned}$$

$$
\begin{aligned}
=&\sqrt{\frac{2(\alpha+h)^2(\alpha+i-1)}{2\alpha+h-1}}-(\alpha+i)\\
=&\sqrt{(\alpha+i)^2+\frac{(h-i)\left\{(\alpha+i-2)(3\alpha+2h+i)+2(\alpha+h)\right\}+(\alpha+i)^2(i-1)}{2\alpha+h-1}}\\
&-(\alpha+i)\geq 0
\end{aligned}
$$

holds for $i=1,\ldots,h$. Here, the first inequality holds because $T\geq\frac{h(h-1)}{2}+K-\alpha$ and the second inequality holds by inequality $\sum_{j=1}^{h}\sqrt{\alpha+j-1}\leq h\sqrt{\frac{2\alpha+h-1}{2}}$. Due to the integrality constraints, instead of the (real-valued) $T_{K-\alpha-(i-1)}$ defined by Eq. (16), we use a rounded version $T_{K-\alpha-(i-1)}$ defined as follows:

$$
i+1+T_{K-\alpha-(i-1)}=\left\lfloor\sum_{j=1}^{i}\frac{B(h)\sqrt{\alpha+j-1}}{A(h)}\right\rfloor-\left\lfloor\sum_{j=1}^{i-1}\frac{B(h)\sqrt{\alpha+j-1}}{A(h)}\right\rfloor.
$$

Then,

$$
\begin{aligned}
\sum_{i=1}^{h}\frac{\alpha+i-1}{i+1+T_{K-\alpha-(i-1)}}<&\sum_{i=1}^{h}\frac{\alpha+i-1}{\frac{B(h)}{A(h)}\sqrt{\alpha+i-1}-1}\\
=&A(h)\sum_{i=1}^{h}\frac{\alpha+i-1}{B(h)\sqrt{\alpha+i-1}-A(h)}\\
=&\frac{A^2(h)}{2B(h)}+\frac{A^2(h)}{B^2(h)}h+\frac{A^3(h)}{B^2(h)}\sum_{i=1}^{h}\frac{1}{B(h)\sqrt{\alpha+i-1}-A(h)}\\
\leq&\frac{A^2(h)}{2B(h)}+\frac{A^2(h)}{B^2(h)}h+\frac{A^2(h)}{B^2(h)}h\\
=&\frac{A^2(h)(B(h)+4h)}{2B^2(h)}
\end{aligned}
\tag{17}
$$

holds. Here, the first inequality uses

$$
\left\lfloor\sum_{j=1}^{i}\frac{B(h)\sqrt{\alpha+j-1}}{A(h)}\right\rfloor-\left\lfloor\sum_{j=1}^{i-1}\frac{B(h)\sqrt{\alpha+j-1}}{A(h)}\right\rfloor>\frac{B(h)\sqrt{\alpha+i-1}}{A(h)}-1
$$

and the second inequality uses the fact that

$$
B(h)\sqrt{\alpha+i-1}-A(h)\geq B(h)-A(h)\geq A(h),
$$

which can be implied from the inequalities

$$A(h) \leq h\sqrt{2(2\alpha + h - 1)}$$

and

$$\begin{aligned} B(h) =& 2\left\{T - (K-\alpha) - \frac{h(h-1)}{2}\right\} + 2(\alpha + h)h \\ \geq& 2(\alpha + h)h \\ =& h\sqrt{2(2\alpha + h - 1)} \cdot \left(\frac{\sqrt{2(2\alpha + h - 1)}}{2} + \frac{h+1}{\sqrt{2(2\alpha + h - 1)}} \right) \\ \geq& h\sqrt{2(2\alpha + h - 1)} \cdot 2\sqrt{\frac{h+1}{2}} \geq 2h\sqrt{2(2\alpha + h - 1)}. \end{aligned}$$

By Eq. (15) and Ineq. (17), Bound (12) holds in this case. □

Corollary 5. *RepeatW&S[K, T] forces the expected number of mistakes made by any player algorithm to be at least*

$$K - \alpha - \frac{(K + \alpha - 1)(K - \alpha)^2(2T + (K + \alpha + 3)(K - \alpha))}{(2T + (K + \alpha - 1)(K - \alpha))^2}$$

for $T \geq (K + \alpha - 1)(K - \alpha)/2$.

(Proof) This corollary can be derived from Bound (12) of Corollary 4 with $h = K - \alpha$ and the fact that

$$\begin{aligned} A^2(K-\alpha) = 4\left(\sum_{j=1}^{K-\alpha} \sqrt{\alpha + j - 1}\right)^2 \leq& 4(K-\alpha)\sum_{j=1}^{K-\alpha}(\alpha + j - 1) \\ =& 2(K + \alpha - 1)(K - \alpha)^2. \end{aligned}$$

□

Remark 4. *By Theorem 4 and Corollary 5, the minimax expected number of mistakes for the noise-free K-armed bandit problem with T rounds is*

$$K - \alpha - \Theta\left(\frac{1}{T}\right).$$

5. Comparison with variations on the model

In this section we first examine two variations of the noise-free multi-armed bandit game and prove minimax expected mistake bounds for them. The first variation is the full-information setting, where the entire loss vector $\boldsymbol{\ell}_t$ is revealed to the algorithm each time, so this variation is not a bandit setting. The second variation is when the adversary is oblivious, and thus must assign losses to arms independent of the player's strategy. We close the section by contrasting these minimax bounds with those from previous section, and making some concluding remarks.

5.1. Full-information Setting

In the full-information setting the whole $\boldsymbol{\ell}_t$ is revealed at every time t regardless of which arm is selected by the player. This is not a bandit setting, so we call it the "noise-free multi-armed full-information game" and call both bandit and full-information games *noise-free multi-armed games*. In this case, the minimax number $L^*(K, T, \alpha)$ of mistakes satisfies

$$L^*(K,T,\alpha) = \max_{\substack{0 < k_1,\ldots,k_m \\ m = \min\{T, K-\alpha\} \\ \sum_{i=1}^m k_i = K-\alpha}} \sum_{i=1}^{m} \frac{k_i}{K - \sum_{j=1}^{i-1} k_j}, \tag{18}$$

which can be proved by induction on T. For $T = 1$ and any $K > 0$, the adversary's best strategy is to set $\boldsymbol{\ell}_1$ to $\mathbb{1}_S$ for the arm set S consisting of the $K-\alpha$ arms having the largest probabilities of being selected by the algorithm. The player's corresponding minimization strategy is to choose each arm with equal probability, so the minimax number of mistakes is $(K-\alpha)/K$, satisfying Eq. (18).

Assume that Eq. (18) holds for $T = T_0$ and any $K > 0$. Consider the game with $T = T_0 + 1$. When the losses of k arms are set to 1 at time 1, the minimax expected number of mistakes from time 2 on is $L^*(K-k, T_0, \alpha)$, and by the inductive assumption,

$$L^*(K-k,T_0,\alpha) = \max_{\substack{0 < k_1,\ldots,k_m \\ m = \min\{T_0, K-k-\alpha\} \\ \sum_{i=1}^m k_i = K-k-\alpha}} \sum_{i=1}^{m} \frac{k_i}{K - k - \sum_{j=1}^{i-1} k_j}.$$

The minimax number $L^*(K-k, T_0, \alpha)$ depends on k, but not on which set of k-arms is assigned loss at time 1, and the adversary's best strategy is to

assign loss at time 1 to the k arms with highest probability of being selected at that time by the player. The best counter for this adversary's strategy is for the player to select an arm uniformly at random, so

$$\begin{aligned} L^*(K, T_0+1, \alpha) &= \max_{0<k<K} \left(\frac{k}{K} + L^*(K-k, T_0, \alpha) \right) \\ &= \max_{\substack{0 < k, k_1, \ldots, k_m \\ m = \min\{T_0, K-k-\alpha\} \\ k + \sum_{i=1}^m k_i = K-\alpha}} \left(\frac{k}{K} + \sum_{i=1}^m \frac{k_i}{K - k - \sum_{j=1}^{i-1} k_j} \right) \\ &= \max_{\substack{0 < k_1, \ldots, k_m \\ m = \min\{T_0+1, K-k_1-\alpha+1\} \\ \sum_{i=1}^m k_i = K-\alpha}} \sum_{i=1}^m \frac{k_i}{K - \sum_{j=1}^{i-1} k_j}. \end{aligned}$$

We can show

$$\max_{\substack{0 < k_1, \ldots, k_m \\ m = \min\{T_0+1, K-k_1-\alpha+1\} \\ \sum_{i=1}^m k_i = K-\alpha}} \sum_{i=1}^m \frac{k_i}{K - \sum_{j=1}^{i-1} k_j} = \max_{\substack{0 < k_1, \ldots, k_m \\ m = \min\{T_0+1, K-\alpha\} \\ \sum_{i=1}^m k_i = K-\alpha}} \sum_{i=1}^m \frac{k_i}{K - \sum_{j=1}^{i-1} k_j}$$

because $\min\{T_0+1, K-k_1-\alpha+1\} \leq \min\{T_0+1, K-\alpha\}$ for $k_1 > 0$ and for $m < \min\{T_0+1, K-\alpha\}$, any $k_1, \ldots, k_m > 0$ with $\sum_{i=1}^m k_i = K-\alpha$, there are $k'_1, \ldots, k'_{m+1} > 0$ with $\sum_{i=1}^{m+1} k'_i = K-\alpha$ such that $m \leq \min\{T_0+1, K-\alpha\}$ and

$$\sum_{i=1}^m \frac{k_i}{K - \sum_{j=1}^{i-1} k_j} < \sum_{i=1}^{m+1} \frac{k'_i}{K - \sum_{j=1}^{i-1} k'_j} \tag{19}$$

holds. Such $k'_1, \ldots, k'_{m+1}$ can be constructed as follows. Since $m < K-\alpha$, there is at least one k_i that is larger than 1. Divide such k_i into $k'_i > 0$ and $k'_{i+1} > 0$, that is, $k'_i + k'_{i+1} = k_i$. Then, consider sequence $k'_1, \ldots, k'_{m+1}$ defined as $k'_1 = k_1, \ldots, k'_{i-1} = k_{i-1}, k'_i, k'_{i+1}, k'_{i+2} = k_{i+1}, \ldots, k'_{m+1} = k_m$. For sequences $k_1, \ldots, k_m$ and $k'_1, \ldots, k'_{m+1}$, Ineq. (19) holds because

$$\frac{k_i}{K - K_0} < \frac{k'_i}{K - K_0} + \frac{k_i - k'_i}{K - K_0 - k'_i}$$

holds, where $K_0 = \sum_{j=1}^{i-1} k_i = \sum_{j=1}^{i-1} k'_i$. Thus Eq. (18) holds for $T = T_0 + 1$ and any $K > 0$. Therefore, minimax number $L^*(K, T, \alpha)$ of mistakes becomes

$$\max_{\substack{0 < k_1, \dots, k_m \\ m = K - \alpha \\ \sum_{i=1}^m k_i = K - \alpha}} \sum_{i=1}^{m} \frac{k_i}{K - \sum_{j=1}^{i-1} k_i} = \sum_{i=\alpha+1}^{K} \frac{1}{i} = \Theta\left(\log \frac{K}{\alpha}\right)$$

for $T \geq K - \alpha$, and

$$\begin{aligned}\max_{\substack{k_1, \dots, k_m \in [K-\alpha] \\ m = T \\ \sum_{i=1}^m k_i = K - \alpha}} \sum_{i=1}^{m} \frac{k_i}{K - \sum_{j=1}^{i-1} k_j} \leq & \max_{\substack{k_1, \dots, k_m \in (0, K-\alpha) \\ m = T \\ \sum_{i=1}^m k_i = K - \alpha}} \sum_{i=1}^{m} \frac{k_i}{K - \sum_{j=1}^{i-1} k_j} \\ = & T\left(1 - \left(\frac{\alpha}{K}\right)^{1/T}\right)\end{aligned}$$

for $T < K - \alpha$, where the right-hand side of the inequality is maximized among $(k_1, \dots, k_m)$ of m-dimensional space of real numbers instead of natural numbers. (The last equality holds when $k_i = K(1-(\alpha/K)^{1/T})(\alpha/K)^{(i-1)/T}$.)

5.2. Oblivious Adversary Game

Oblivious adversary strategies cannot depend on the past choices of a randomized player. In effect, this means that the adversary might as well (randomly) pick which arms will remain lossless at the start of the game. The safest strategy is to select α lossless arms, which we call $a_1^*, \dots, a_\alpha^*$, from the uniform distribution over the K arms. The adversary can maximize the loss of the best players by always setting the loss of each other arm to 1, so at each time $t \in [T]$, $\boldsymbol{\ell}_t = \mathbb{1}_S$ where $S = [K] \setminus \{a_1^*, \dots, a_\alpha^*\}$. For this best adversary, the player's best strategy is to repeatedly choose a non-dirty arm uniformly at random, and keep playing that arm until the player makes a mistake. Analyzing the interaction of these best strategies, the probability of making exactly m mistakes is

$$\frac{K-\alpha}{K} \times \frac{K-\alpha-1}{K-1} \times \cdots \times \frac{K-\alpha-m+1}{K-m+1} \times \frac{\alpha}{K-m} = \frac{\alpha}{K} \prod_{i=1}^{\alpha-1} \frac{K-m-i}{K-i}$$

when $m < \min\{T, K-\alpha\}$ and

$$\frac{K-\alpha}{K} \times \frac{K-\alpha-1}{K-1} \times \cdots \times \frac{K-\alpha-m+1}{K-m+1} = \prod_{i=0}^{\alpha-1} \frac{K-m-i}{K-i}$$

when $m = \min\{T, K-\alpha\}$. Let $m_0 = \min\{T, K-\alpha\}$. Then, the minimax expected number of mistakes, $L^*(K,T,\alpha)$, for the oblivious adversary is

$$
\begin{aligned}
&\frac{\alpha \sum_{m=1}^{m_0-1} m \prod_{i=1}^{\alpha-1}(K-m-i) + m_0 \prod_{i=0}^{\alpha-1}(K-m_0-i)}{\prod_{i=0}^{\alpha-1}(K-i)} \\
=&\frac{\sum_{m=1}^{m_0-1} m \left(\prod_{i=0}^{\alpha-1}(K-m-i) - \prod_{i=1}^{\alpha}(K-m-i)\right) + m_0 \prod_{i=0}^{\alpha-1}(K-m_0-i)}{\prod_{i=0}^{\alpha-1}(K-i)} \\
=&\frac{\displaystyle\sum_{m=1}^{m_0-1}\prod_{i=0}^{\alpha-1}(K-m-i) - (m_0-1)\prod_{i=1}^{\alpha}(K-m_0+1-i) + m_0\prod_{i=0}^{\alpha-1}(K-m_0-i)}{\prod_{i=0}^{\alpha-1}(K-i)} \\
=&\frac{\frac{1}{\alpha+1}\sum_{m=1}^{m_0-1}\left(\prod_{i=-1}^{\alpha-1}(K-m-i) - \prod_{i=0}^{\alpha}(K-m-i)\right) + \prod_{i=0}^{\alpha-1}(K-m_0-i)}{\prod_{i=0}^{\alpha-1}(K-i)} \\
=&\frac{\frac{1}{\alpha+1}\left(\prod_{i=-1}^{\alpha-1}(K-1-i) - \prod_{i=0}^{\alpha}(K-m_0+1-i)\right) + \prod_{i=0}^{\alpha-1}(K-m_0-i)}{\prod_{i=0}^{\alpha-1}(K-i)} \\
=&\frac{\frac{K-\alpha}{\alpha+1}\prod_{i=0}^{\alpha-1}(K-i) - \frac{K-m_0+1}{\alpha+1}\prod_{i=0}^{\alpha-1}(K-m_0-i) + \prod_{i=0}^{\alpha-1}(K-m_0-i)}{\prod_{i=0}^{\alpha-1}(K-i)} \\
=&\frac{\frac{K-\alpha}{\alpha+1}\prod_{i=0}^{\alpha-1}(K-i) - \frac{K-m_0-\alpha}{\alpha+1}\prod_{i=0}^{\alpha-1}(K-m_0-i)}{\prod_{i=0}^{\alpha-1}(K-i)} \\
=&\frac{K-\alpha}{\alpha+1} - \frac{K-\alpha}{\alpha+1}\prod_{i=0}^{\alpha}\frac{K-m_0-i}{K-i} \\
=&\frac{K-\alpha}{\alpha+1}\left(1 - \prod_{i=0}^{\alpha}\frac{K-m_0-i}{K-i}\right) \\
=&\begin{cases} \dfrac{K-\alpha}{\alpha+1}\left(1-\displaystyle\prod_{i=0}^{\alpha}\frac{K-T-i}{K-i}\right) & (\text{when } T < K-\alpha) \\ \dfrac{K-\alpha}{\alpha+1} & (\text{when } T \geq K-\alpha). \end{cases}
\end{aligned}
$$

5.3. Comparison

In the case with a large T, the above shows that the minimax expected number of mistakes is $\Theta(\log(K/\alpha))$ for the full-information setting and $(K-\alpha)/(\alpha+1)$ for the oblivious adversary setting in the K-armed game.

Table 1: Comparison of $L^*(K,T,\alpha)$, the minimax expected number of mistakes , for variants of the noise-free multi-armed game with K arms and T rounds

Full-information

$$L^*(K,T,\alpha)\begin{cases} \leq T\left(1-\left(\frac{\alpha}{K}\right)^{1/T}\right) & (T<K-\alpha) \\ \\ =\sum_{i=\alpha+1}^{K}\frac{1}{i}=\Theta\left(\log\frac{K}{\alpha}\right) & (T\geq K-\alpha) \end{cases}$$

Bandit with an oblivious adversary

$$L^*(K,T,\alpha)=\begin{cases} \frac{K-\alpha}{\alpha+1}\left(1-\prod_{i=0}^{\alpha}\frac{K-T-i}{K-i}\right) & (T<K-\alpha) \\ \\ \frac{K-\alpha}{\alpha+1} & (T\geq K-\alpha). \end{cases}$$

Bandit with an adaptive adversary

$$\left.\begin{array}{ll} (T<K-\alpha) & T\left(1-\left(\frac{\alpha}{K}\right)^{1/T}\right)-\frac{\left(\frac{K}{\alpha}\right)^{(T-1)/T}-1}{K\left(\left(\frac{K}{\alpha}\right)^{1/T}-1\right)} \\ \\ \left(T\geq\frac{(K+\alpha-1)(K-\alpha)}{2}\right) & K-\alpha-\frac{(K+\alpha-1)(K-\alpha)^2(2T+(K+\alpha+3)(K-\alpha))}{(2T+(K+\alpha-1)(K-\alpha))^2} \end{array}\right\}\leq$$

$$L^*(K,T,\alpha)\leq\min\{T,K-\alpha\}-\sum_{j=1}^{\min\{T,K-\alpha\}}Q(K-j+1,T-j+1,\alpha),$$

$$\left(L^*(K,T,\alpha)=K-\alpha-\Theta\left(\frac{1}{T}\right)\right)$$

where $Q(K,T,\alpha)=\prod_{h=\alpha}^{K-1}\frac{h}{T+h}=\prod_{h=1}^{T}\frac{\alpha-1+h}{K-1+h}$. For $\alpha=K-1$,

$$L^*(K,T,\alpha)=1-\frac{K-1}{T+K-1}.$$

Therefore the $K-\alpha-\Theta(1/T)$ expected number of mistakes forced by the adaptive adversary case is large in comparison, which indicates the power of an adaptive adversary. Table 1 gives a more detailed comparison of the bounds for the three different settings and includes the $T<K-\alpha$ case.

5.4. Bounds from the Noisy Bandit Problem

In the general (noisy) bandit problem, the pseudo-regret [4]

$$\min_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \sum_{t=1}^{T} \ell_{t,i}\right]$$

is the most popular evaluation measure. In the noise-free case, the pseudo-regret coincides with the expected loss $\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t}\right]$, hence it coincides with the expected number of mistakes in the case of $\{0,1\}$-loss. For the noisy bandit problem, a pseudo-regret lower bound $\min\{\sqrt{KT}, T\}/20$ can be proved for the $[0,1]$-loss-version of the multi-armed bandit problem by a slight modification of the proof [2] for its $[0,1]$-reward-version. This lower bound is known to be optimal unless computational efficiency is not required [5].

The adversary with the pseudo-regret $\sqrt{KT}/20$ used in the proof [2] is an oblivious adversary who generates losses $\ell_{t,i}$ for $i \in K$ according to a Bernoulli distribution with parameter $1/2$, except a best arm i_* selected according to the uniform distribution whose loss is generated according to a Bernoulli distribution with parameter $1/2 - \epsilon$ for some $\epsilon > 0$. To achieve the pseudo-regret $\sqrt{KT}/20$, ϵ must be set to $\sqrt{K/T}/4$, which means that the adversary is very noisy; the parameter of the Bernoulli distribution for the best arm is almost $1/2$ when T is large enough compared to K. Setting ϵ to $1/2$ corresponds to ensuring that there is a noise-free arm. Unfortunately, using this setting of ϵ in the ϵ-dependent bound of [2] leads to a trivial (negative) lower bound on the pseudo-regret. On the other hand, the optimal oblivious adversary for the ($\alpha = 1$) noise-free setting forces any player to have pseudo-regret at least $(K-1)/2$.

6. Concluding Remarks

The simple oblivious adversary analysis provides a $(K-\alpha)/(\alpha+1)$ lower bound on the minimax expected number of mistakes in the K-armed bandit game with an adaptive adversary in the case with $\alpha(\geq 1)$ lossless arms. Conversely, any sensible player will make at most one mistake on each arm, giving a trivial $K-\alpha$ upper bound on the minimax expected number of mistakes. Our analysis of the noise-free multi-armed bandit game goes through a simpler "survival game", where the goals of the adversary/player are to maximize/minimize the probability of making any mistake (rather than the

expected number of mistakes). This analysis provides nearly matching upper and lower bounds on the minimax expected number of mistakes, showing that it is $K-\alpha-\Theta(1/T)$ for the noise-free multi-armed bandit game, and we obtain the exact minimax value of $1-\frac{K-1}{T+K-1}$ when $\alpha = K-1$. However, for larger K there is still a small gap between the upper and the lower bounds. This gap on our bounds exists even in our bounds on the simpler survival game. We conjecture that our lower bound of the survival game is tight, but the correctness of this conjecture remains an open problem. Our player algorithm does not make use of the number α of lossless arms. Whether α can be used by the player to reduce the expected number of mistakes, is also an open problem.

Acknowledgments

We would like to thank anonymous reviewers for their valuable comments that have improved the quality of this paper.

[1] A. Nakamura, D. P. Helmbold, M. K. Warmuth, Noise free multi-armed bandit game, in: 10th International Conference on Language and Automata Theory and Applications (LATA 2016), 2016, pp. 412–423.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, The nonstochastic multiarmed bandit problem, SIAM J. Comput. 32 (1) (2003) 48–77.

[3] N. Cesa-Bianchi, G. Lugosi, Prediction, Learning, and Games, Cambridge University Press, 2006.

[4] S. Bubeck, N. Cesa-Bianchi, Regret analysis of stochastic and nonstochastic multi-armed bandit problems, Foundations and Trends in Machine Learning 5 (1) (2012) 1–122.

[5] J. Audibert, S. Bubeck, Minimax policies for adversarial and stochastic bandits, in: Proc. of the 22nd Conference on Learning Theory (COLT 2009), 2009.

Appendix A. Proof of Proposition 1

Recall that Proposition 1 states:

For any integers $m \geq 1$, $K > m$, and $T \geq m$,

$$\begin{aligned}
&\prod_{i=1}^{m}\left[\left(\frac{K-\sum_{j=1}^{i}k_j}{K+t_i-1}\right)\prod_{t_i<t<t_{i+1}}\left(\frac{K-\sum_{j=1}^{i}k_j+t-1}{K+t-1}\right)\right] \\
=&\frac{\prod_{i=1}^{m}\left[\left(K-\sum_{j=1}^{i}k_j\right)\prod_{h=1}^{k_i-1}\left(K-\sum_{j=1}^{i}k_j+t_i-1+h\right)\right]}{\prod_{h=K-\beta}^{K-1}(T+h)}
\end{aligned} \tag{A.1}$$

holds for any integers $k_1, ..., k_m \geq 1$ with $\sum_{i=1}^{m} k_i = \beta < K$ and any integers $1 \leq t_1 < \cdots < t_m < t_{m+1} \equiv T+1$.

(Proof) We prove Eq. (A.1) by mathematical induction on m. When $m = 1$, then $k_1 = \beta$, so the lefthand-side of Eq. (A.1) is equal to

$$\begin{aligned}
&\frac{K-\beta}{K+t_1-1}\times\frac{K-\beta+t_1}{K+t_1}\times\cdots\times\frac{K-\beta+T-1}{K+T-1} \\
=&\frac{(K-\beta)(K-\beta+T-1)!/(K-\beta+t_1-1)!}{(K+T-1)!/(K+t_1-2)!} \\
=&\frac{(K-\beta)(K+t_1-2)!/(K-\beta+t_1-1)!}{(K+T-1)!/(K-\beta+T-1)!} \\
=&\frac{(K-\beta)\prod_{h=1}^{\beta-1}(K-\beta+t_1-1+h)}{\prod_{h=K-\beta}^{K-1}(T+h)},
\end{aligned}$$

which is the righthand-side of Eq. (A.1) for $m = 1$ and $k_1 = \beta$. Thus, Eq. (A.1) holds for any $1 \leq k_1 = \beta < K$ and any $1 \leq t_1 < t_2 = T+1$ when $m = 1$.

Assume now that the proposition holds when $m = m_0 \geq 1$, and consider the case when $m = m_0 + 1 \geq 2$. Fix arbitrary K, T, $k_1, ..., k_{m_0+1}$, and $1 \leq t_1 < \cdots < t_{m_0+1} < t_{m_0+2} = T+1$ satisfying the conditions of the proposition. The lefthand-side of Eq. (A.1) can be factored into three terms:

$$\begin{aligned}
&\prod_{i=1}^{m}\left(\frac{K-\sum_{j=1}^{i}k_j}{K+t_i-1}\right)\prod_{t_i<t<t_{i+1}}\left(\frac{K-\sum_{j=1}^{i}k_j+t-1}{K+t-1}\right) \\
=&\left(\frac{K-k_1}{K+t_1-1}\right)\prod_{t_1<t<t_2}\left(\frac{K-k_1+t-1}{K+t-1}\right)
\end{aligned} \tag{A.2}$$

$$\times \frac{\prod_{t=t_2}^{T}((K-k_1)+t-1)}{\prod_{t=t_2}^{T}(K+t-1)} \tag{A.3}$$

$$\times \prod_{i=2}^{m_0+1}\left(\frac{(K-k_1)-\sum_{j=2}^{i}k_j}{(K-k_1)+t_i-1}\right)\prod_{t_i<t<t_{i+1}}\left(\frac{(K-k_1)-\sum_{j=2}^{i}k_j+t-1}{(K-k_1)+t-1}\right). \tag{A.4}$$

Then, each term can be calculated as follows.

$$\begin{aligned}
\text{(A.2)} =& \frac{(K-k_1)(K-k_1+t_2-2)!/(K-k_1+t_1-1)!}{(K+t_2-2)!/(K+t_1-2)!} \\
=& \frac{(K-k_1)(K+t_1-2)!/(K-k_1+t_1-1)!}{(K+t_2-2)!/(K-k_1+t_2-2)!} \\
=& \frac{(K-k_1)\prod_{h=1}^{k_1-1}(K-k_1+t_1-1+h)}{(K+t_2-2)!/(K-k_1+t_2-2)!} \\
\text{(A.3)} =& \frac{(K-k_1+T-1)!/(K-k_1+t_2-2)!}{(K+T-1)!/(K+t_2-2)!} \\
=& \frac{(K+t_2-2)!/(K-k_1+t_2-2)!}{(K+T-1)!/(K-k_1+T-1)!} \\
=& \frac{(K+t_2-2)!/(K-k_1+t_2-2)!}{\prod_{h=K-k_1}^{K-1}(T+h)} \\
\text{(A.4)} =& \frac{\prod_{i=2}^{m_0+1}\left((K-k_1)-\sum_{j=2}^{i}k_j\right)\prod_{h=1}^{k_i-1}\left((K-k_1)-\sum_{j=2}^{i}k_j+t_i-1+h\right)}{\prod_{h=(K-k_1)-(\beta-k_1)}^{(K-k_1)-1}(T+h)}
\end{aligned}$$

Note that the last equality uses the $m = m_0$ inductive assumption. Thus,

$$\text{(A.2)} \times \text{(A.3)} \times \text{(A.4)} = \frac{\prod_{i=1}^{m_0+1}\left(K-\sum_{j=1}^{i}k_j\right)\prod_{h=1}^{k_i-1}\left(K-\sum_{j=1}^{i}k_j+t_i-1+h\right)}{\prod_{h=K-\beta}^{K-1}(T+h)},$$

proving the proposition when $m = m_0 + 1$ and completing the induction. □