Stochastic bandits with vector losses: Minimizing ℓ^∞-norm of relative losses

Xuedong Shang, Han Shao, Jian Qian

Stochastic bandits with vector losses: Minimizing ℓ^∞-norm of relative losses

Xuedong Shang
Inria Lille, SequeL Team

Han Shao
Toyota Technological Institute at Chicago

Jian Qian
MIT

Abstract

Multi-armed bandits are widely applied in scenarios like recommender systems, for which the goal is to maximize the click rate. However, more factors should be considered, e.g., user stickiness, user growth rate, user experience assessment, etc. In this paper, we model this situation as a problem of K-armed bandit with multiple losses. We define relative loss vector of an arm where the i-th entry compares the arm and the optimal arm with respect to the i-th loss. We study two goals: (a) finding the arm with the minimum ℓ^∞-norm of relative losses with a given confidence level (which refers to fixed-confidence best-arm identification); (b) minimizing the ℓ^∞-norm of cumulative relative losses (which refers to regret minimization). For goal (a), we derive a problem-dependent sample complexity lower bound and discuss how to achieve matching algorithms. For goal (b), we provide a regret lower bound of $\Omega(T^{2/3})$ and provide a matching algorithm.

1 Introduction

Multi-armed bandit is a classical sequential decision-making problem, where an agent/learner sequentially chooses actions (also called "arms") and observes a stochastic scalar loss of the chosen arm for T rounds (Thompson, 1933). The two classical goals are to identify the best arm (which is the arm with the minimum expected loss) and to minimize the cumulative losses. Practical applications of multi-armed bandit, among many others, range from recommendation systems (Zeng et al., 2016), clinical trials (Durand et al., 2018) to portfolio management (Huo and Fu, 2017). For example, when a user comes to an e-commerce website, traditional recommender systems choose a product (an arm) to recommend and observe whether the user clicks it or not (the loss). However, in addition to click rates, other factors like user stickiness should be considered as well in practice. Another example is fairness in public policy making, where each policy (an arm) can have drastic impacts over different gender/race groups (vector losses). These problems can be modeled as multi-armed bandits with *vector/multi-dimensional* losses. In each dimension i, we measure the performance of an arm k by comparing its i-th loss with the minimum loss among the i-th dimension, which we call relative loss[1].

We provide a simple problem instance shown in Table 1 to explain the intuition on how our setting differs from the usual one: the bandit model contains 3 arms, each row corresponds to the loss vector incurred by playing each arm and each column corresponds to the vector of absolute i-th losses for each arm. In this example, the optimal arm with respect to each arm has zero-loss, thus the relative losses coincide with the absolute losses. The optimal arm would be arm 3 since it minimizes the maximum of each row. We formalize the intuition later in Section 2.

Table 1: An instance.

arms	$\boldsymbol{\ell}^{(1)}$	$\boldsymbol{\ell}^{(2)}$
arm 1	1	0
arm 2	0	1
arm 3	1/2	1/2

In this paper, we study both the two classical goals under the vector-loss setting: best-arm identification and regret minimization.

Best-arm identification, as a particular type of *pure exploration*, only cares about identifying the optimal arm given some stopping criterion. Two kinds of stopping criterion exist: (a) *fixed-budget* for which the algorithm stops when a given budget is exhausted (Bubeck et al., 2009; Audibert and Bubeck, 2010; Gabillon et al., 2012; Karnin et al., 2013; Carpentier and Locatelli, 2016); (b) *fixed-confidence* for which the algorithm stops when we are able to spot the best arm with a high confidence level (Even-dar et al., 2003; Kalyanakrishnan et al., 2012; Gabillon et al., 2012; Jamieson et al., 2014; Garivier and Kaufmann, 2016; Qin et al., 2017; Yu et al., 2018; Degenne et al., 2019; Ménard, 2019; Shang et al., 2020). In this paper, we focus on the second type and the detailed setting is described in Section 2.2.

[1] One may notice that the relative loss coincides with the traditional definition of regret in the scalar case.

Contrary to best-arm identification, the objective of regret minimization, as indicated by its name, is to minimize the *regret*: the gap between the total reward gathered by the agent and the cumulative reward obtained by optimal strategy. Regret minimization naturally balances between exploration and exploitation. An asymptotic lower bound on the regret is given by Lai and Robbins (1985). Since then the problem has been extensively studied. Typical solutions include optimistic algorithms (Auer et al., 2002; Cappé et al., 2013; Honda and Takemura, 2015), their Bayesian competitor Thompson sampling (Thompson, 1933; Kaufmann et al., 2012; Agrawal and Goyal, 2013; Korda et al., 2013), and non-parametric methods (Baransi et al., 2014; Chan, 2020; Baudry et al., 2020). In our paper, the objective is somehow different. We aim to minimize the ℓ^∞-norm of cumulative relative loss, which requires a more specific definition of regret that we give in Section 2.3.

Related work. Vector payoffs/losses, as a core ingredient of this work, mostly finds its popularity among literature of online learning, in particular in a game theory point of view. The problem is closely related to multi-objective optimization where the use of *Blackwell approachability* has been thoroughly investigated both for the *full information* setting (Perchet, 2014) and the *partial monitoring* setting (Kwon and Perchet, 2017; Perchet, 2011). The very same problem is less studied for multi-armed bandit. To the best of our knowledge, minimizing the ℓ^∞-norm of (cumulative) relative loss has never been looked into in the bandit literature. For best-arm identification, a related setting refers to to Katz-Samuels and Scott (2019), where the feedback is also multi-dimensional, but the goal is constrained maximization. For regret minimization, the most similar setting to ours is the multi-objective multi-armed bandit that considers conflicting sub-objectives. It is first proposed by Drugan and Nowe (2013) and Zuluaga et al. (2013), and is followed by a series of extensions (Auer et al., 2016; Drugan and Nowe, 2014; Lu et al., 2019). Multi-objective multi-armed bandit aims to find the *Pareto frontier* of different sub-objectives, while our setting only cares about the maximum. For example, arm 1 and arm 2 in Table 1 are on the Pareto frontier as well, but do not achieve optimality in our definition.

Contributions. The contributions of this paper are the following:

- We describe a novel multi-armed bandit setting with d-dimensional vector losses and we study the problem in both best-arm identification and regret minimization. We design the performance measure as minimizing the ℓ^∞-norm of relative loss over all single dimensions.
- We first investigate best-arm identification. We derive a *problem-dependent* lower bound on the sample complexity and discuss how to achieve matching algorithms for *fixed-confidence* best-arm identification.
- We then study regret minimization. We show that any algorithm suffers a *worst-case* regret of order $\Omega(T^{2/3})$ under our setting. We provide an algorithm based on *two-player game* with matching upper regret bound up to a log factor.

Outline. The rest of the paper is organized as follows. We start by the problem formulation in Section 2 where we specify both best-arm identification and regret minimization under our setting. We first study best-arm identification in Section 3 for which we focus on the sample complexity. It then follows regret minimization in Section 4 where we provide the worst-case lower bound along with a simple matching algorithm before we conclude.

2 Problem formulation

Our model $\boldsymbol{\nu}$ for the environment is a K-armed bandit with *unknown* vector payoffs, i.e., vector loss distributions $(\nu_k^{(1)}, \dots, \nu_k^{(d)})_{k\in[K]}$ where $\nu_k^{(i)}$ is the i-th sub-(scalar) loss distribution for the k-th arm. Each distribution $\nu_k^{(i)}$ is from a known sub-Gaussian canonical exponential family with one parameter (the mean of the distribution) for all i and $k \in [K]$.

We consider two mainstream multi-armed bandit frameworks in this paper (see Kaufmann and Garivier 2017 for a survey), namely best-arm identification and regret minimization. In both settings, a learning algorithm $\mathcal{A}$ selects an arm $\mathcal{A}_t \in [K]$ at each round $t = 1, \dots, T$, and then observes a loss vector of arm k: $y_{\mathcal{A}_t,t} \sim (\nu_{\mathcal{A}_t}^{(1)}, \dots, \nu_{\mathcal{A}_t}^{(d)})$. Let $\mathcal{F}_t = \sigma(\mathcal{A}_1, y_{\mathcal{A}_1,1}, \dots, \mathcal{A}_t, y_{\mathcal{A}_t,t,})$ be the information available to the algorithm after t rounds. We specify the two frameworks under our setting in this section.

2.1 Some notations

Let $\Sigma_n \triangleq \{\boldsymbol{\omega} \in [0,1]^n : \sum_{i=1}^n \omega_i = 1\}$ with $n \in \mathbb{N}$ denote the n-dimensional probability simplex. Let $\mathbf{1}$ denote the all-one vector whose dimension can be known from the context. We let $d(x, y)$ denote the Kullback-Leibler divergence from the distribution parameterized by x to that parameterized by y for $x, y \in [0, 1]$. We let $d^+(x, y) = d(x, y)\mathbb{1}_{\{x>y\}}$. For simplicity, we abuse $\arg\min$ (resp. $\arg\max$) to represent an arbitrary element that achieves the minimum (resp. maximum) and fix this element thereafter[2]. We introduce several notions of loss for problem formulation. Note that all the following loss definitions depend on the bandit model $\boldsymbol{\nu}$. For simplicity, we omit it in the notations whenever there is no ambiguity.

[2]It is not hard to check that the choice does not affect the results in this paper.

For each $i \in [d]$, we define the i-th *expected loss* as

$$\boldsymbol{\ell}^{(i)} \triangleq (\ell_1^{(i)}, \dots, \ell_K^{(i)}),$$

where $\ell_k^{(i)} = \mathbb{E}[\nu_k^{(i)}] \in [0,1]$ for $k \in [K]$. Similarly, we denote by

$$\boldsymbol{\ell}_k \triangleq (\ell_k^{(1)}, \dots, \ell_k^{(d)})$$

the *expected loss* vector of arm k. A bandit model in this paper can thus be interchangeably represented by $\boldsymbol{\nu}$ or $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_K) \in [0,1]^{d \times K}$.

Let $\star_i \triangleq \arg\min_{k \in [K]} \ell_k^{(i)}$ denote the index of the arm with the lowest i-th expected loss. We further define the i-th *expected relative loss* for $i \in [d]$ as

$$\boldsymbol{\ell}_{|\star}^{(i)} = (\ell_{1|\star}^{(i)}, \dots, \ell_{K|\star}^{(i)}),$$

and the *expected relative loss* of arm k as

$$\boldsymbol{\ell}_{k|\star} \triangleq (\ell_{k|\star}^{(1)}, \dots, \ell_{k|\star}^{(d)}),$$

where $\ell_{k|\star}^{(i)} \triangleq \ell_k^{(i)} - \ell_{\star_i}^{(i)}$. And we denote the matrix of the expected relative losses by

$$\boldsymbol{\ell}_{|\star} = (\boldsymbol{\ell}_{1|\star}, \dots, \boldsymbol{\ell}_{K|\star}) \in [0,1]^{d \times K}.$$

We define the i-th *expected loss of weight* $\boldsymbol{\omega} \in \Sigma_K$ as

$$\ell_{\boldsymbol{\omega}}^{(i)} \triangleq \boldsymbol{\omega}^\top \boldsymbol{\ell}^{(i)},$$

and the i-th *expected relative loss of the weight* $\boldsymbol{\omega}$ as

$$\ell_{\boldsymbol{\omega}|\star}^{(i)} \triangleq \boldsymbol{\omega}^\top \boldsymbol{\ell}_{|\star}^{(i)}.$$

Finally, we denote by

$$\ell_{\boldsymbol{\omega}|\star} \triangleq \left\| (\ell_{\boldsymbol{\omega}|\star}^{(1)}, \dots, \ell_{\boldsymbol{\omega}|\star}^{(d)}) \right\|_\infty$$

the *ℓ^∞-norm of the expected relative loss of the weight* $\boldsymbol{\omega}$.

2.2 Best-arm identification

We first detail the framework of fixed-confidence best-arm identification in our case: the objective is to identify the arm with the minimum relative loss in terms of infinite norm. That is, for each bandit model $\boldsymbol{\ell} \in [0,1]^{d \times K}$, the *unique* correct answer is given by

$$i^\star(\boldsymbol{\ell}) \triangleq \arg\min_{k \in [K]} \|\boldsymbol{\ell}_{k|\star}\|_\infty = \arg\min_{k \in [K]} \max_{i \in [d]} \ell_{k|\star}^{(i)}$$

among the set of possible correct answers $\mathcal{I} = [K]$.

Motivation. In general, the vector-loss/payoff settings considered by previous works mainly focus on the Pareto frontier of different sub-objectives. This notion of optimality is unreasonable in some cases, where some dimensional losses suffer extremely high scalar regrets. To avoid the risk of incredibly high scalar regrets for any single dimension, we target at minimizing the infinite norm of the relative losses (which are scalar regrets) and thus, we can bound the scalar regrets for all dimensions at the same time.

Algorithm. A deterministic pure-exploration algorithm under the fixed-confidence setting is given by three components: (1) a *sampling rule* $(\mathcal{A}_t)_{t \geq 1}$, where $\mathcal{A}_t \in [K]$ is $\mathcal{F}_{t-1}$-measurable. (2) a *stopping rule* τ_δ, a stopping time for the filtration $(\mathcal{F}_t)_{t \geq 1}$, and (3) a *decision rule* $\hat{\imath} \in \mathcal{I}$ which is $\mathcal{F}_{\tau_\delta}$-measurable. Non-deterministic algorithms could also be considered by allowing the rules to depend on additional internal randomization. The algorithms we present are deterministic.

δ-correctness and fixed-confidence objective. An algorithm is δ-correct if it predicts the correct answer with probability at least $1-\delta$, precisely if $\mathbb{P}_{\boldsymbol{\ell}}\big(\hat{\imath} \neq i^\star(\boldsymbol{\ell})\big) \leq \delta$ and $\tau_\delta < +\infty$ almost surely for all $\boldsymbol{\ell} \in [0,1]^{d \times K}$. The goal is to find a δ-correct algorithm that minimizes the *sample complexity*, that is, the expected number of samples $\mathbb{E}_{\boldsymbol{\ell}}[\tau_\delta]$ needed to predict an answer.

2.3 Regret minimization

We now detail the setting for regret minimization. Let $\boldsymbol{L}_{\mathcal{A}} \triangleq \sum_{t=1}^T \boldsymbol{\ell}_{\mathcal{A}_t}$ denote the expected cumulative loss of algorithm $\mathcal{A}$ where $\boldsymbol{\ell}_{\mathcal{A}_t} = (\ell_{\mathcal{A}_t}^{(1)}, \dots, \ell_{\mathcal{A}_t}^{(d)})$ and $L_{\mathcal{A}}^{(i)} \triangleq \sum_{t=1}^T \ell_{\mathcal{A}_t}^{(i)}$ be the expected cumulative losses. The traditional regret (which we call relative loss) w.r.t. the scalar loss $\ell_{\star_i}^{(i)}$ is defined as

$$L_{\mathcal{A}|\star}^{(i)} \triangleq L_{\mathcal{A}}^{(i)} - T\ell_{\star_i}^{(i)}$$

for $i \in [d]$. The goal is to minimize the ℓ^∞-norm of cumulative relative loss, which differs from the goal of classical stochastic multi-armed bandits with scalar payoffs. However, comparing the ℓ^∞-norm of cumulative relative loss of an algorithm with a single optimal arm may be unreasonable. For example, a bandit problem with three arms $(1,0)$, $(0,1)$ and $(3/4, 3/4)$ has the optimal arm $(3/4, 3/4)$. But we can achieve ℓ^∞-norm of cumulative relative loss $T/2$ by pulling arm 1 and arm 2 for $T/2$ rounds respectively while always pulling the single optimal arm can only achieve $3T/4$. Therefore, it is more reasonable to look into the optimal proportion of arm pulls instead of only considering the single optimal arm under the context of vector losses. We call the optimal proportion of arm pulls the *optimal weight*.

Definition 1 (optimal weight)**.** *We define*

$$\boldsymbol{\omega}^\star \triangleq \arg\min_{\boldsymbol{\omega} \in \Sigma_K} \ell_{\boldsymbol{\omega}|\star}$$

the optimal weight of arms.

Consequently, it is also natural to measure the performance by comparing with the optimal weight. Therefore, we introduce the following regret in terms of the relative losses defined w.r.t. the optimal weight.

Definition 2 (regret)**.** *The expected regret of algorithm $\mathcal{A}$ is defined as*

$$\mathbb{E}[R_{\mathcal{A}}(T)] \triangleq \mathbb{E}\left[\left\|\boldsymbol{L}_{\mathcal{A}|\star}\right\|_\infty\right] - \ell_{\boldsymbol{\omega}^\star|\star} T, \tag{1}$$

where $\boldsymbol{L}_{\mathcal{A}|\star} = (L^{(1)}_{\mathcal{A}|\star}, \dots, L^{(d)}_{\mathcal{A}|\star})$.

3 Best-arm identification

We first study best-arm identification for our setting in a fixed-confidence context. We are thus interested in the sample complexity. We begin with particularizing the general problem-dependent lower bound by Garivier and Kaufmann (2016) to our setting. Then we discuss how to design asymptotically optimal algorithms that we precise the definition.

3.1 Lower bound on the sample complexity

We first derive a problem-dependent lower bound as stated in the following theorem.

Theorem 3. *Let* $S_y(\eta) \triangleq \{i | \eta_i \leq y_i\}$ *and* $C_\gamma(z) \triangleq \{i | z_i \leq \gamma\}$ *for* $\eta, y \in \mathbb{R}^d$, $z \in \mathbb{R}^K$ *and* $\gamma \in \mathbb{R}$. *For any* δ*-correct strategy and any* $\boldsymbol{\ell} \in [0,1]^{d\times K}$, *we have*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\ell}}[\tau_\delta]}{\log(1/\delta)} \geq T^\star(\boldsymbol{\ell})\,,$$

where $T^\star(\boldsymbol{\ell})$ *is a characteristic time defined by*

$$\begin{aligned} &T^\star(\boldsymbol{\ell})^{-1} \\ &= \max_{\boldsymbol{\omega}\in\Sigma_K} \min_{\substack{k^\star\in[K]\\ j\in[d]}} \inf_{\substack{x\in[0,1], y\in[0,1]^d:\\ y\leq(1-x)\mathbf{1}}} \omega_{i^\star(\boldsymbol{\ell})} d(\ell^{(j)}_{i^\star(\boldsymbol{\ell})}, x + y_j) \\ &\quad + \omega_{k^\star} \left(\sum_{i \notin S_y(\boldsymbol{\ell}_{k^\star})} d^+(\ell^{(i)}_{k^\star}, x + y_i) \right) \\ &\quad + \sum_{i\neq j} \sum_{k\in C_{y_i}(\boldsymbol{\ell}^{(i)})} \omega_i d(\ell^{(i)}_k, y_i) \\ &\quad + \sum_{k\in C_{y_j}(\boldsymbol{\ell}^{(j)})/\{i^\star(\boldsymbol{\ell})\}} \omega_j d(\ell^{(j)}_k, y_j)\,. \end{aligned} \tag{2}$$

Proof. By Theorem 1 of Garivier and Kaufmann (2016), we have

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\ell}}[\tau_\delta]}{\log(1/\delta)} \geq T^\star(\boldsymbol{\ell})\,,$$

where

$$T^\star(\boldsymbol{\ell})^{-1} = \max_{\boldsymbol{\omega}\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\ell})} \left(\sum_{k\in[K]} \sum_{i\in[d]} \omega_k d(\ell^{(i)}_k, \lambda^{(i)}_k) \right)\,,$$

where $\mathrm{Alt}(\boldsymbol{\ell})$ is an alternative bandit problem with different optimal arm, i.e.,

$$\mathrm{Alt}(\boldsymbol{\ell}) \triangleq \{\boldsymbol{\lambda} \in [0,1]^{d\times K} : i^\star(\boldsymbol{\lambda}) \neq i^\star(\boldsymbol{\ell})\}\,.$$

Then we just need to calculate $T^\star(\boldsymbol{\ell})^{-1}$ to complete the proof. For an alternative $\boldsymbol{\lambda}$ with $i^\star(\boldsymbol{\lambda}) = k^\star$, we let $y = (\lambda^{(1)}_{\star_1}, \dots, \lambda^{(d)}_{\star_d}) \in [0,1]^d$ and $x = \max_i \lambda^{(i)}_{k^\star} - y_i \in [0,1]$. Then we have $\exists j \in [d], \lambda^{(j)}_{i^\star(\boldsymbol{\ell})} - y_j \geq x$. For any $\boldsymbol{\omega} \in \Sigma_K$, we have

$$\begin{aligned} &\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\ell})} \left(\sum_{k\in[K]} \sum_{i\in[d]} \omega_k d(\ell^{(i)}_k, \lambda^{(i)}_k) \right) \\ &= \min_{k^\star\in[K]} \inf_{\substack{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\ell}):\\ i^\star(\boldsymbol{\lambda})=k^\star}} \left(\sum_{k\in[K]} \sum_{i\in[d]} \omega_k d(\ell^{(i)}_k, \lambda^{(i)}_k) \right) \\ &= \min_{k^\star\in[K]} \inf_{\substack{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\ell}):\\ i^\star(\boldsymbol{\lambda})=k^\star}} \left(\sum_{\substack{k\in\cup_{i\in[d]} C_{y_i}(\boldsymbol{\ell}^{(i)})\\ \cup\{i^\star(\boldsymbol{\ell}),k^\star\}}} \sum_{i\in[d]} \omega_k d(\ell^{(i)}_k, \lambda^{(i)}_k) \right) \\ &= \min_{\substack{k^\star\in[K]\\ j\in[d]}} \inf_{\substack{x\in[0,1], y\in[0,1]^d:\\ y\leq(1-x)\mathbf{1}}} \omega_{i^\star(\boldsymbol{\ell})} d(\ell^{(j)}_{i^\star(\boldsymbol{\ell})}, x + y_j) \\ &\quad + \omega_{k^\star} \left(\sum_{i\notin S_y(\ell_{k^\star})} d^+(\ell^{(i)}_{k^\star}, x + y_i) \right) \\ &\quad + \sum_{i\neq j} \sum_{k\in C_{y_i}(\boldsymbol{\ell}^{(i)})} \omega_i d(\ell^{(i)}_k, y_i) \\ &\quad + \sum_{k\in C_{y_j}(\boldsymbol{\ell}^{(j)})/\{i^\star(\boldsymbol{\ell})\}} \omega_j d(\ell^{(j)}_k, y_j)\,, \end{aligned}$$

which completes the proof. □

3.2 Asymptotically optimal algorithms

To design an algorithm for fixed-confidence best-arm identification, one needs to specify three components as previously mentioned: a stopping rule, a decision rule and a sampling rule. The *Track-and-Stop* strategy proposed by Garivier and Kaufmann (2016) can be adopted in our setting with optimal sample complexity. For completeness, we describe the algorithm briefly below.

In the next, we use the empirical average $\hat{\boldsymbol{\ell}}_t$ to estimate the expected losses $\boldsymbol{\ell}$ at time t, that is

$$\forall k \in [K], i \in [d], \hat{\ell}^{(i)}_{t,k} \triangleq \frac{1}{t} \sum_{\tau=1}^{t} y^{(i)}_{\tau,k}\,.$$

Decision rule. Let $f(\cdot)$ be a function of time-dependent exploration bonus (e.g. $\log(t)$) for $t \in \mathbb{N}$. Let $[c_{t,k}, d_{t,k}] \triangleq \{\xi : \sum_{i\in[d]} N_{t-1,k} d(\hat{\ell}^{(i)}_{t-1,k}, \xi) \leq f(t)\}$. Now, let

$$\tilde{\boldsymbol{\ell}}_{t-1} \triangleq \operatorname*{arg\,min}_{\substack{\boldsymbol{\lambda}\in[0,1]^{d\times K}\cap\\ \prod_{k=1}^K [c_{t,k}, d_{t,k}]^d}} \left(\sum_{k\in[K]} \sum_{j\in[d]} N_{t-1,k} d(\hat{\ell}^{(j)}_{t-1,k}, \lambda^{(j)}_k) \right)\,.$$

Then for the decision rule, we choose to recommend $\hat{\imath} = i^\star(\tilde{\boldsymbol{\ell}}_{t-1})$. Note that if the empirical loss matrix $\hat{\boldsymbol{\ell}}_{t-1} \in [0,1]^{d\times K}$, then $\tilde{\boldsymbol{\ell}}_{t-1}$ coincides with $\hat{\boldsymbol{\ell}}_{t-1}$ and the decision is simply the empirical best arm.

Stopping rule. In this paper, we choose to use the classical Chernoff stopping rule (see e.g. Chernoff 1959; Garivier and Kaufmann 2016) that can be concretized (for exponential family bandit models) to the following form:

$$\tau_\delta \triangleq \inf\left\{t \in \mathbb{N} : \mathrm{GLR}_t(\mathrm{Alt}(\hat{\boldsymbol{\ell}}_t)) > \beta(t,\delta)\right\},$$

where $\beta(t,\delta)$ is a threshold function to be chosen carefully and

$$\mathrm{GLR}_t(\mathrm{Alt}(\hat{\boldsymbol{\ell}}_t)) = \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\hat{\boldsymbol{\ell}}_t)} \left(\sum_{k\in[K]}\sum_{i\in[d]} N_{t,k} d(\hat{\ell}_{t,k}^{(i)}, \lambda_k^{(i)})\right)$$

is the *generalized log-likelihood ratio* between the alternative set $\mathrm{Alt}(\hat{\boldsymbol{\ell}}_t)$ and the whole parameter space.

Using the same reasoning as Shang et al. (2020), one can show that the Chernoff stopping rule coupled with the threshold

$$\beta(t,\delta) \triangleq 4\log(4+\log(t)) + 2\mathcal{C}\left(\frac{\log((Kd-1)/\delta)}{2}\right)$$

leads to the δ-correctness, i.e. $\mathbb{P}\left[\tau_\delta < \infty \wedge \hat{\imath} \neq i^\star(\boldsymbol{\ell})\right] \leq \delta$ for any sampling rule. The function $\mathcal{C}$ is given by Kaufmann and Koolen (2018) that satisfies $\mathcal{C}(x) \simeq x + \log(x)$. Note that in practice, one can simply choose to set $\beta(t,\delta) = \log((1+\log(t))/\delta)$.

Sampling rule and the whole picture. We aim to design algorithms that match the lower bound derived in Theorem 3. We call such algorithms asymptotically optimal. Formally, a fixed-confidence algorithm is asymptotically optimal if

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\ell}}[\tau_\delta]}{\log(1/\delta)} \leq T^\star(\boldsymbol{\ell}).$$

To achieve this property, the learner needs to allocate her pulls according to the optimal weight vector given by the characteristic time (Garivier and Kaufmann, 2016; Russo, 2016), that is

$$\begin{aligned}
\boldsymbol{\omega}^\star(\boldsymbol{\ell}) &= \arg\max_{\boldsymbol{\omega}\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\ell})} \left(\sum_{k\in[K]}\sum_{i\in[d]} \omega_k d(\ell_k^{(i)}, \lambda_k^{(i)})\right) \\
&= \arg\max_{\boldsymbol{\omega}\in\Sigma_K} \min_{\substack{k^\star\in[K]\\ j\in[d]}} \inf_{\substack{x\in[0,1], y\in[0,1]^d:\\ y\leq(1-x)\mathbf{1}}} \omega_{i^\star(\boldsymbol{\ell})} d(\ell_{i^\star(\boldsymbol{\ell})}^{(j)}, x+y_j) \\
&\quad + \omega_{k^\star}\left(\sum_{i\notin S_y(\boldsymbol{\ell}_{k^\star})} d^+(\ell_{k^\star}^{(i)}, x+y_i)\right) \\
&\quad + \sum_{i\neq j}\sum_{k\in C_{y_i}(\boldsymbol{\ell}^{(i)})} \omega_i d(\ell_k^{(i)}, y_i) \\
&\quad + \sum_{k\in C_{y_j}(\boldsymbol{\ell}^{(j)})/\{i^\star(\boldsymbol{\ell})\}} \omega_j d(\ell_k^{(j)}, y_j), \qquad (3)
\end{aligned}$$

which can be considered as solving a minimax saddle-point problem. Although the inf part is non-convex, it is computable by calculating the infimum over x, y for each $k^\star$ and j. To calculate the infimum over x, y, for each $i \in [d]$, we consider the case that y_i is larger than the i-th losses of m_i arms with $m_i = 0, 1, \ldots, K$ separately. In each case of fixed $k^*, j, \{m_i\}_{i\in[d]}$, the infimum part of (3) is convex and solvable. However, this incurs a computational complexity of $\Theta(dK^{d+1})$.

The aforementioned problem requires the knowledge of the true means, one simple way to overcome this is to adopt the `D-Tracking` rule (Garivier and Kaufmann, 2016), where we choose to sample

$$\mathcal{A}_{t+1} \in \arg\max_{k\in[K]} \omega_k^\star(\hat{\boldsymbol{\ell}}_t) - N_{t,k}/t$$

using 'plug-in' estimates of the optimal weight. `D-Tracking` is proved to be asymptotically optimal (Garivier and Kaufmann, 2016), with a known drawback as its computational liability due to the optimization problem (3) that has to be treated once at each step, since there is no known closed form expression or even no computationally feasible approximation approach in general.

An improved algorithm without solving the optimization problem every round by solving a two-player game derived from Degenne et al. (2019) is given in Appendix A.

4 Regret minimization

We turn our attention to regret minimization. We first derive a worst-case lower bound. Then we present an efficient algorithm that matches the lower bound. For simplicity, we omit the t in the subscripts in this section, e.g., we denote $\hat{\ell}_k^{(i)}$ instead of $\hat{\ell}_{t,k}^{(i)}$.

4.1 Worst-case lower bound

Theorem 4. *For $T > 27$, let* sup *be the supremum over all distributions of losses and* inf *be the infimum over all algorithms. Then we have,*

$$\inf_{\mathcal{A}} \sup_{\boldsymbol{\nu}} \mathbb{E}\left[R_{\mathcal{A}}(T)\right] \geq \frac{1}{2304} T^{\frac{2}{3}}.$$

Proof. Let $\epsilon \in [0, 1/6]$ be a constant, we consider a bandit model $\boldsymbol{\nu}$ with the following 2-dimensional loss vectors:

$$\boldsymbol{\ell}_1 = \left(\frac{1}{4}, \frac{3}{4}\right), \boldsymbol{\ell}_2 = \left(\frac{3}{4}, \frac{1}{4}\right),$$
$$\boldsymbol{\ell}_3 = \left(\frac{3-\epsilon}{8}, \frac{3+\epsilon}{8}\right), \boldsymbol{\ell}_4 = \left(\frac{3+\epsilon}{8}, \frac{3-\epsilon}{8}\right),$$

where the losses are Gaussian distributions with variance 1 and expectation of the indicated value.

Denote N_1, N_2, N_3 and N_4 the number each arm is pulled. Since there is a symmetry between arm 3 and arm 4. Without loss of generality, we assume for the given algorithm $\mathcal{A}$ under consideration, that $\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}}[N_3] \le \mathbb{E}_{\mathcal{A},\boldsymbol{\nu}}[N_4]$. Then according to the assumption between N_3 and N_4, we consider an alternative bandit model $\boldsymbol{\nu}'$ with the following losses,

$$\boldsymbol{\ell}'_1 = \left(\frac{1-\epsilon}{4}, \frac{3}{4}\right), \boldsymbol{\ell}'_2 = \left(\frac{3}{4}, \frac{1}{4}\right),$$
$$\boldsymbol{\ell}'_3 = \left(\frac{3-\epsilon}{8}, \frac{3+\epsilon}{8}\right), \boldsymbol{\ell}'_4 = \left(\frac{3+\epsilon}{8}, \frac{3-\epsilon}{8}\right).$$

For $\epsilon < 1/6$: The optimal arms for each loss are $\star_1 = 1$, $\star_2 = 2$.

$$\begin{aligned}
\omega^\star_{\boldsymbol{\nu}'} &\triangleq \arg\min_{\boldsymbol{\omega}\in\Sigma_K} \max_{i\in[d]} \left\{\boldsymbol{\omega}^\top (\boldsymbol{\ell}')^{(i)} - \ell^{(i)}_{\star_i,\boldsymbol{\nu}'}\right\} \\
&= \arg\min_{\boldsymbol{\omega}\in\Sigma_K} \max \Bigg\{ \frac{(2+\epsilon)\omega_2}{4} + \frac{(1+\epsilon)\omega_3}{8} \\
&\quad + \frac{(1+3\epsilon)\omega_4}{8}, \frac{\omega_1}{2} + \frac{(1+\epsilon)\omega_3}{8} + \frac{(1-\epsilon)\omega_4}{8} \Bigg\} \\
&= (0,0,1,0).
\end{aligned}$$

Thus the regret is lower bounded as follows,

$$\begin{aligned}
R_{\boldsymbol{\nu}'}(T) &\triangleq \max_{i\in[d]} \left(L^{(i)} - \ell^{(i)}_{\boldsymbol{\omega}^\star_{\boldsymbol{\nu}'}} T\right) \\
&= \max \Bigg\{ \frac{(1-\epsilon)N_1}{4} + \frac{3N_2}{4} + \frac{(3-\epsilon)N_3}{8} \\
&\quad + \frac{(3+\epsilon)N_4}{8} - \frac{3-\epsilon}{8}T, \frac{3N_1}{4} + \frac{N_2}{4} \\
&\quad + \frac{(3+\epsilon)N_3}{8} + \frac{(3-\epsilon)N_4}{8} - \frac{3+\epsilon}{8}T \Bigg\} \\
&= \max \Bigg\{ -\frac{(1+\epsilon)N_1}{8} + \frac{(3+\epsilon)N_2}{8} + \frac{\epsilon N_4}{4}, \\
&\qquad \frac{(3-\epsilon)N_1}{8} - \frac{(1+\epsilon)N_2}{8} - \frac{\epsilon N_4}{4} \Bigg\} \\
&\ge \frac{2}{3}\left(-\frac{(1+\epsilon)N_1}{8} + \frac{(3+\epsilon)N_2}{8} + \frac{\epsilon N_4}{4}\right) \\
&\quad + \frac{1}{3}\left(\frac{(3-\epsilon)N_1}{8} - \frac{(1+\epsilon)N_2}{8} - \frac{\epsilon N_4}{4}\right) \\
&\ge \frac{1}{48}N_1 + \frac{5}{24}N_2 + \frac{\epsilon}{12}N_4 \\
&= (\frac{1}{48} - \frac{\epsilon}{12})N_1 + (\frac{5}{24} - \frac{\epsilon}{12})N_2 + \frac{\epsilon}{12}(T - N_3) \\
&\ge \frac{1}{144}N_1 + \frac{1}{6}N_2 + \frac{\epsilon}{12}(T - N_3).
\end{aligned}$$

So we have the following regret for the bandit model $\boldsymbol{\nu}'$,

$$\mathbb{E}[R_{\mathcal{A},\boldsymbol{\nu}'}(T)] \ge \frac{1}{144}\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_1] + \frac{\epsilon}{12}\left(T - \mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_3]\right). \tag{4}$$

According to the inequality (6) by Garivier et al. (2018), we have,

$$\begin{aligned}
\frac{\epsilon^2}{32}\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_1] &\ge \mathrm{kl}\left(\frac{\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_3]}{T}, \frac{\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}}[N_3]}{T}\right) \\
&\ge \frac{1}{2}\left(\frac{\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}}[N_3]}{T} - \frac{\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_3]}{T}\right)^2.
\end{aligned}$$

Therefore,

$$\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_3] \le \frac{\epsilon}{4}T\sqrt{\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_1]} + \mathbb{E}_{\mathcal{A},\boldsymbol{\nu}}[N_3]$$

Furthermore with $\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}}[N_3] \le T/2$, and according to our assumption,

$$\begin{aligned}
\mathbb{E}[R_{\mathcal{A},\boldsymbol{\nu}'}(T)] &\ge \frac{1}{144}\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_1] + \frac{\epsilon}{12}\left(T - \mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_3]\right) \\
&\ge \frac{1}{144}\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_1] + \frac{\epsilon}{12}T - \frac{\epsilon^2}{48}T\sqrt{\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_1]} \\
&\quad - \frac{\epsilon}{12}\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}}[N_3] \\
&\ge \frac{1}{144}\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_1] + \frac{\epsilon}{24}T - \frac{\epsilon^2}{48}T\sqrt{\mathbb{E}_{\mathcal{A},\boldsymbol{\nu}'}[N_1]}
\end{aligned}$$

Take $\epsilon = T^{-1/3}/2 < 1/6$, we have,

$$\mathbb{E}[R_{\mathcal{A},\boldsymbol{\nu}'}(T)] \ge \frac{1}{144}\mathbb{E}_{\boldsymbol{\nu}'}[N_1] + \frac{1}{48}T^{\frac{2}{3}} - \frac{1}{192}T^{\frac{1}{3}}\sqrt{\mathbb{E}_{\boldsymbol{\nu}'}[N_1]}$$

If $\mathbb{E}_{\boldsymbol{\nu}'}[N_1] \ge T^{2/3}/16$, then by (4), we have,

$$\mathbb{E}[R_{\mathcal{A},\boldsymbol{\nu}'}(T)] \ge \frac{1}{144}\mathbb{E}_{\boldsymbol{\nu}'}[N_1] \ge \frac{1}{2304}T^{\frac{2}{3}}$$

Else, we have,

$$\begin{aligned}
\mathbb{E}[R_{\mathcal{A},\boldsymbol{\nu}'}(T)] &\ge \frac{1}{144}\mathbb{E}_{\boldsymbol{\nu}'}[N_1] + \frac{1}{48}T^{\frac{2}{3}} - \frac{1}{192}T^{\frac{1}{3}}\sqrt{\mathbb{E}_{\boldsymbol{\nu}'}[N_1]} \\
&\ge \frac{1}{48}T^{\frac{2}{3}} - \frac{1}{192}T^{\frac{1}{3}}\frac{1}{4}T^{\frac{1}{3}} \\
&\ge \frac{1}{2304}T^{\frac{2}{3}}
\end{aligned}$$

□

A simple method derived from the lower bound proof. The lower bound proof actually indicates that the minimum losses for each dimension are crucial in order to achieve optimality. To this regard, following a simple scheme of forced exploration, then exploit, we could easily derive an algorithm matching the lower bound for the minimax regret. Detailed description of the algorithm and analysis can be found in Appendix B. Despite its simplicity, the computation complexity scales exponentially with d. To cope with this issue, we develop a second algorithm with the two-player game scheme.

4.2 A minimax game

We propose an algorithm called Combinatorial Game (CG), whose pseudo-code is displayed in Algorithm 1.

The idea is to introduce a two-player game scheme as recently studied by Degenne et al. (2020b), where one tries to identify the best allocation of probability across the arms while the opponent always replies with a best response. More specifically, at each round t we request from the first learner its probability allocation, and pull arms accordingly. When the losses are revealed, we calculate the fictitious losses the learner would have suffered if it had played the arm, and feed the fictitious losses to the learner, as displayed in Algorithm 1. The learner is supposed to have regret bounds similar to AdaHedge (de Rooij et al., 2014).

Concretely, the arm with the smallest empirical i-th loss is denoted by $\hat{\star}_i \triangleq \arg\min_{k\in[K]} \hat{\ell}_k^{(i)}$ for $i \in [d]$. Let $\mathrm{LCB}(\boldsymbol{\ell}_{|\star})$ be the lower confidence bound of $\boldsymbol{\ell}_{|\star}$, calculated as,

$$\mathrm{LCB}(\boldsymbol{\ell}_{|\star})_{i,k} = \hat{\ell}_{k|\hat{\star}}^{(i)} - \sqrt{\frac{2\log(T)}{N_{t,k}}} - \sqrt{\frac{2\log(T)}{N}}\,, \quad (5)$$

where $\hat{\ell}_{k|\hat{\star}}^{(i)} = \hat{\ell}_k^{(i)} - \hat{\ell}_{\hat{\star}_i}^{(i)}$. Then we can define the best response in an optimistic fashion: $\mathbf{x}_t = \arg\max_{\mathbf{x}\in\Sigma_d} \mathbf{x}^\top \mathrm{LCB}(\boldsymbol{\ell}_{|\star})\boldsymbol{\omega}_t$, and feed the optimistic loss $\mathrm{LCB}(\boldsymbol{\ell}_{|\star})^\top \mathbf{x}_t$ back to $\mathcal{L}$.

Algorithm 1 The algorithm of CG

Input: Time horizon T, number of forced exploration rounds N, learner $\mathcal{L}$ for linear losses on the simplex
Pull each arm for N rounds
Start an instance of $\mathcal{L}$ and set $N_{1,k} = 0$ for all $k \in [K]$
for $t = 1, \cdots, T - KN$ **do**
  $\hat{\star}_i = \arg\min_{k\in[K]} \hat{\ell}_k^{(i)}$ for $i \in [d]$
  Get $\boldsymbol{\omega}_t$ from $\mathcal{L}$
  // Track the weights
  Play arm $\mathcal{A}_t = \arg\min_{k\in[K]} \left(N_{t,k} - \sum_{\tau=0}^{t-1} \omega_{\tau,k}\right)$
  $N_{t+1,\mathcal{A}_t} = N_{t,\mathcal{A}_t} + 1$ and $N_{t+1,k} = N_{t,k}$ for $k \neq \mathcal{A}_t$
  $\mathbf{x}_t = \arg\max_{\mathbf{x}\in\Sigma_d} \mathbf{x}^\top \mathrm{LCB}(\boldsymbol{\ell}_{|\star})\boldsymbol{\omega}_t$,
  where $\mathrm{LCB}(\boldsymbol{\ell}_{|\star})$ is calculated as in (5)
  // Feed optimistic loss
  Feed loss $\mathrm{LCB}(\boldsymbol{\ell}_{|\star})^\top \mathbf{x}_t$ to $\mathcal{L}$ and update $\mathrm{LCB}(\boldsymbol{\ell}_{|\star})$
end for

4.3 Analysis of CG

We show that CG achieves a matching upper bound for the regret. We first show that the empirical estimation is valid with high probability at each round. Specifically, we have the following lemma.

Lemma 5. *Define the following event:*

$$E_{1,t} \triangleq \left\{ \forall k \in [K], i \in [d] : \left|\hat{\ell}_k^{(i)} - \ell_k^{(i)}\right| \leq \sqrt{\frac{2\log(t)}{N_k}} \right\},$$

where N_k denotes the number of pulls of arm k. This event happens with probability at least $1 - dK/t^2$:

$$\mathbb{P}\left[E_{1,t}\right] \geq 1 - \frac{dK}{t^2}$$

Proof. This is a direct application of the Hoeffding's Inequality with the union bound. □

With the Lemma above, we proceed to show that $\mathrm{LCB}(\boldsymbol{\ell}_{|\star})$ is a valid approximation for $\boldsymbol{\ell}_{|\star}$. Concretely, we have the following lemma.

Lemma 6. *Assume that $E_{1,t}$ holds, we have,*

$$\mathrm{LCB}_t(\boldsymbol{\ell}_{|\star})_{i,k} \leq \ell_{k|\star}^{(i)}\,,$$

$$\ell_{k|\star}^{(i)} \leq \mathrm{LCB}_t(\boldsymbol{\ell}_{|\star})_{i,k} + 2\sqrt{\frac{2\log(t)}{N_{t,k}}} + 2\sqrt{\frac{2\log(t)}{N}}.$$

Proof. This is a easy deduction of Lemma 5. □

Theorem 7. *For $T \geq dK$, CG achieves a $\widetilde{\mathcal{O}}(T^{2/3})$ regret.*

Proof. Recall the event $E_{1,t}$ defined in Lemma 5, we can decompose the regret as follows.

$$\mathbb{E}\left[R_{\mathrm{CG}}(T)\right] = \mathbb{E}\left[\max_{i\in[d]} \sum_{t=1}^{T} \ell_{\mathcal{A}_t|\star}^{(i)}\right] - \ell_{\boldsymbol{\omega}^\star|\star} T$$

$$\leq KN + \mathbb{E}\left[\max_{i\in[d]} \sum_{t\geq KN} \ell_{\mathcal{A}_t|\star}^{(i)} \mathbb{1}_{\{\neg E_{1,t}\}}\right]$$

$$+ \mathbb{E}\left[\max_{i\in[d]} \sum_{t\geq KN} \ell_{\mathcal{A}_t|\star}^{(i)} \mathbb{1}_{\{E_{1,t}\}}\right] - \ell_{\boldsymbol{\omega}^\star|\star} T\,.$$

For the second term, due to Lemma 5, we have

$$\mathbb{E}\left[\max_{i\in[d]} \sum_{t\geq KN} \ell_{\mathcal{A}_t|\star}^{(i)} \mathbb{1}_{\{\neg E_{1,t}\}}\right]$$

$$\leq \mathbb{E}\left[\sum_{t\geq KN} \mathbb{1}_{\{\neg E_{1,t}\}}\right] \leq \sum_{t\geq KN} \frac{dK}{t^2} \leq \frac{\pi^2 dK}{6}\,.$$

For the third term, we first decompose the regret into a term related to strategies of both players and the tracking error.

$$\mathbb{E}\left[\max_{i\in[d]} \sum_{t\geq KN} \ell_{\mathcal{A}_t|\star}^{(i)} \mathbb{1}_{\{E_{1,t}\}}\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\max_{i\in[d]}\sum_{k=1}^{K}\sum_{t=KN+1}^{T}\ell^{(i)}_{k|\star}\mathbb{1}_{\{E_{1,t},\mathcal{A}_t=k\}}\right]\\
&\leq \mathbb{E}\left[\max_{i\in[d]}\sum_{k=1}^{K}\sum_{t=KN+1}^{T}\omega_{t,k}\ell^{(i)}_{k|\star}\mathbb{1}_{\{E_{1,t}\}}\right]\\
&\quad+\mathbb{E}\left[\max_{i\in[d]}\sum_{k=1}^{K}\sum_{t=1+KN}^{T}\left(\mathbb{1}_{\{\mathcal{A}_t=k\}}-\omega_{t,k}\right)\mathbb{1}_{\{E_{1,t}\}}\ell^{(i)}_{k|\star}\right]\\
&\leq \mathbb{E}\left[\max_{i\in[d]}\sum_{k=1}^{K}\sum_{t=KN+1}^{T}\omega_{t,k}\ell^{(i)}_{k|\star}\mathbb{1}_{\{E_{1,t}\}}\right]\\
&\quad+\mathbb{E}\left[\max_{i\in[d]}\sum_{k=1}^{K}\sum_{t=1+KN}^{T}\left(\mathbb{1}_{\{\mathcal{A}_t=k\}}-\omega_{t,k}\right)\ell^{(i)}_{k|\star}\right]\\
&\quad+\mathbb{E}\left[\max_{i\in[d]}\sum_{k=1}^{K}\sum_{t=1+KN}^{T}\left(\omega_{t,k}-\mathbb{1}_{\{\mathcal{A}_t=k\}}\right)\ell^{(i)}_{k|\star}\mathbb{1}_{\{\neg E_{1,t}\}}\right]\\
&\leq \mathbb{E}\left[\sum_{t=KN+1}^{T}\max_{\mathbf{x}\in\Sigma^d}\mathbf{x}^\top\boldsymbol{\ell}_{|\star}\boldsymbol{\omega}_t\mathbb{1}_{\{E_{1,t}\}}\right]+K+\frac{\pi^2 dK}{6}\,,
\end{aligned}
\tag{6}
$$

where (6) adopts Lemma 15 by Garivier and Kaufmann (2016). The first term in (6) can be further estimated.

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=KN+1}^{T}\max_{\mathbf{x}\in\Sigma^d}\mathbf{x}^\top\boldsymbol{\ell}_{|\star}\boldsymbol{\omega}_t\mathbb{1}_{\{E_{1,t}\}}\right]\\
&\leq \mathbb{E}\left[\sum_{t=KN+1}^{T}\max_{\mathbf{x}\in\Sigma^d}\mathbf{x}^\top\mathrm{LCB}(\boldsymbol{\ell}_{|\star})\boldsymbol{\omega}_t\mathbb{1}_{\{E_{1,t}\}}\right]\\
&\quad+\mathbb{E}\left[2\sum_{t=KN+1}^{T}\sum_{k=1}^{K}\left(\sqrt{\frac{2\log(T)}{N_{t,k}}}\right)\omega_{t,k}\right]\\
&\quad+2T\sqrt{\frac{2\log(T)}{N}}\\
&\leq \mathbb{E}\left[\sum_{t=KN+1}^{T}\max_{\mathbf{x}\in\Sigma^d}\mathbf{x}^\top\mathrm{LCB}(\boldsymbol{\ell}_{|\star})\boldsymbol{\omega}_t\mathbb{1}_{\{E_{1,t}\}}\right]\\
&\quad+2(K^2+\sqrt{2KT})\sqrt{2\log(T)}+2T\sqrt{\frac{2\log(T)}{N}}
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\sum_{t=KN+1}^{T}\max_{\mathbf{x}\in\Sigma^d}\mathbf{x}^\top\mathrm{LCB}(\boldsymbol{\ell}_{|\star})\boldsymbol{\omega}_t\right]\\
&\quad+2(K^2+\sqrt{2KT})\sqrt{2\log(T)}+2T\sqrt{\frac{2\log(T)}{N}}+\frac{\pi^2}{6}dK\\
&\leq \mathbb{E}\left[\sum_{t=KN+1}^{T}\mathbf{x}_t^\top\mathrm{LCB}(\boldsymbol{\ell}_{|\star})\boldsymbol{\omega}^\star\right]+\sqrt{T}\\
&\quad+2(K^2+\sqrt{2KT})\sqrt{2\log(T)}+2T\sqrt{\frac{2\log(T)}{N}}+\frac{\pi^2}{6}dK\,,
\end{aligned}
\tag{8}
$$

where (7) adopts Lemma 9 by Degenne et al. (2019) and (8) uses the fact that $\mathcal{L}$ has regret $\sqrt{T}$. Now we are only left to bound the first term in (8).

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=KN+1}^{T}\mathbf{x}_t^\top\mathrm{LCB}(\boldsymbol{\ell}_{|\star})\boldsymbol{\omega}^\star\right]\\
&\leq \mathbb{E}\left[\sum_{t=KN+1}^{T}\mathbf{x}_t^\top\boldsymbol{\ell}_{|\star}\boldsymbol{\omega}^\star\right]+\frac{\pi^2}{6}dK\\
&\leq (T-KN)\max_{\mathbf{x}\in\Sigma^d}\mathbf{x}^\top\boldsymbol{\ell}_{|\star}\boldsymbol{\omega}^\star+\frac{\pi^2}{6}dK\\
&= (T-KN)\ell_{\boldsymbol{\omega}^\star|\star}+\frac{\pi^2}{6}dK\,,
\end{aligned}
$$

Therefore, aggregating all the terms above we have the regret is upper bounded by $\mathcal{O}(KN+T\sqrt{\log(T)/N}) = \widetilde{\mathcal{O}}(T^{2/3})$ by setting $N=(K^2T^2\log(T))^{1/3}$. □

Adaptive algorithm: The term $T^{2/3}$ comes from the trade-off between the exploration of N rounds and the confidence bonus $\sqrt{2\log(T)/N}$. In fact, `CG` does not need time horizon T and forced exploration rounds N as inputs. `CG` can be easily refined by keeping each arm pulled for at least $t^{2/3}$ rounds at time t and using a learner which is also adaptive, e.g., `AdaHedge` (de Rooij et al., 2014).

5 Discussion

We studied a new setup of multi-armed bandit with vector losses. The main purpose of the paper was to investigate a framework for which we carefully constructed appropriate performance measures. We derived a problem-dependent lower bound of the sample complexity for best-arm identification and discussed how to design asymptotically optimal matching algorithms. We also derived a worst-case lower bound for regret minimization and designed a minimax game algorithm that achieves matching upper bound.

We are mainly interested in the maximum of different losses in this work. One possible future direction is to study how can we extend to a more general objective function instead of taking the maximum. Another interesting problem is to investigate whether we can derive a problem-dependent lower bound for regret minimization, for which the alternative bandit problem has a different optimal weight instead of a different single optimal arm.

References

Agrawal, S. and Goyal, N. (2013). Further optimal regret bounds for Thompson sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AIStats)*, pages 99–107.

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Annual Conference on Learning Theory (CoLT)*.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multi-armed bandit problem. *Machine Learning Journal*, 47(2-3):235–256.

Auer, P., Chiang, C. K., Ortner, R., and Drugan, M. M. (2016). Pareto front identification from stochastic bandit feedback. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AIStats)*, pages 939–947.

Baransi, A., Maillard, O.-a., and Mannor, S. (2014). Sub-sampling for multi-armed bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases 2014 (ECML-PKDD)*.

Baudry, D., Kaufmann, E., and Maillard, O.-A. (2020). Sub-sampling for efficient non-parametric bandit exploration. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT)*, pages 23–37.

Cappé, O., Garivier, A., Maillard, O. A., Munos, R., and Stoltz, G. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541.

Carpentier, A. and Locatelli, A. (2016). Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Proceedings of the 29th Annual Conference on Learning Theory (CoLT)*.

Chan, H. P. (2020). The multi-armed bandit problem: An efficient nonparametric solution. *Annals of Statistics*, 48(1):346–373.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.

de Rooij, S., Van Erven, T., Grünwald, P. D., and Koolen, W. M. (2014). Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316.

Degenne, R., Koolen, W., and Ménard, P. (2019). Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*.

Degenne, R. and Koolen, W. M. (2019). Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*.

Degenne, R., Ménard, P., Shang, X., and Valko, M. (2020a). Gamification of pure exploration for linear bandits. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Degenne, R., Shao, H., and Koolen, W. M. (2020b). Structure Adaptive Algorithms for Stochastic Bandits. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Drugan, M. M. and Nowe, A. (2013). Designing multi-objective multi-armed bandits algorithms: A study. In *Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 2358–2365.

Drugan, M. M. and Nowe, A. (2014). Scalarization based Pareto optimal set of arms identification algorithms. In *Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN)*, pages 2690–2697.

Durand, A., Achilleos, C., Iacovides, D., Strati, K., and Pineau, J. (2018). Contextual bandits for adapting treatment in a mouse model of de Novo Carcinogenesis. In *Proceedings of the 3rd Machine Learning for Health Care Conference (MLHC)*.

Even-dar, E., Mannor, S., and Mansour, Y. (2003). Action elimination and stopping conditions for reinforcement learning. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 162–169.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 3212–3220.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Annual Conference on Learning Theory (CoLT)*.

Garivier, A., Ménard, P., and Stoltz, G. (2018). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399.

Honda, J. and Takemura, A. (2015). Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16:3721–3756.

Huo, X. and Fu, F. (2017). Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society Open Science*, 4(11).

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the 27th Annual Conference on Learning Theory (CoLT)*, pages 423–439.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 655–662.

Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 1238–1246.

Katz-Samuels, J. and Scott, C. (2019). Top feasible arm identification. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTats)*.

Kaufmann, E. and Garivier, A. (2017). Learning the distribution with largest mean: two bandit frameworks. *ESAIM: Proceedings and Surveys*, 60:114–131.

Kaufmann, E. and Koolen, W. (2018). Mixture martingales revisited with applications to sequential tests and confidence intervals. *arXiv preprint arXiv:1811.11419*.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory (ALT)*.

Korda, N., Kaufmann, E., and Munos, R. (2013). Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 1448–1456.

Kwon, J. and Perchet, V. (2017). Online learning and Blackwell approachability with partial monitoring: Optimal convergence rates. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AIStats)*, volume 54.

Lai, T.-L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Lu, S., Wang, G., Hu, Y., and Zhang, L. (2019). Multi-objective generalized linear bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3080–3086.

Ménard, P. (2019). Gradient ascent for active exploration in bandit problems. *arXiv preprint arXiv:1905.08165*.

Perchet, V. (2011). Approachability of convex sets in games with partial monitoring. *Journal of Optimization Theory and Applications*, 149(3):665–677.

Perchet, V. (2014). Approachability, regret and calibration: Implications and equivalences. *Journal of Dynamics and Games*, 1(2):181–254.

Qin, C., Klabjan, D., and Russo, D. (2017). Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 5381–5391.

Russo, D. (2016). Simple Bayesian algorithms for best arm identification. In *Proceedings of the 29th Annual Conference on Learning Theory (CoLT)*.

Shang, X., de Heide, R., Kaufmann, E., Ménard, P., and Valko, M. (2020). Fixed-confidence guarantees for Bayesian best-arm identification. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AIStats)*.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285.

Yu, X., Shao, H., Lyu, M. R., and King, I. (2018). Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Zeng, W., Fang, M., Shao, J., and Shang, M. (2016). Uncovering the essential links in online commercial networks. *Scientific Reports*, 6.

Zuluaga, M., Krause, A., Sergent, G., and Puschel, M. (2013). Active learning for multi-criterion optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*.

A More details on best-arm identification algorithm

We provide an improved algorithm for best-arm identification. The idea is to view the problem again as a minimax game as for regret minimization, which is also a natural observation from the lower bound: given a bandit model $\boldsymbol{\ell}$, at each time step a learner plays an arm, and a fictive opponent, tries to fool the learner by playing an alternative bandit model $\boldsymbol{\lambda}$ with a different correct answer. Such framework allows to obtain algorithms that adapt to any structure with asymptotic optimality guarantees, and is extensively studied recently for best-arm identification (Degenne et al., 2019; Degenne and Koolen, 2019; Ménard, 2019; Degenne et al., 2020a).

In Algorithm 2, we show one instance of such gamified sampling rule by Degenne and Koolen (2019), adopted to our setting, along with the decision rule and the stopping rule we described in Section 3.

By applying the game scheme, we can actually approach the minimax saddle-point

$$\boldsymbol{\omega}^{\star}(\boldsymbol{\ell}) = \arg\max_{\boldsymbol{\omega}\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\ell})} \left(\sum_{k\in[K]} \sum_{i\in[d]} \omega_k d(\ell_k^{(i)}, \lambda_k^{(i)}) \right)$$

step by step by leveraging an iterative algorithm for both the real learner and the fictive opponent.

We first need to implement a learner $\mathcal{L}_{\boldsymbol{\omega}}^{i}$ for each possible answer $i \in \mathcal{I}$, for which we can apply choose to use `AdaHedge` again (as for regret minimization). Not to enter into details, `AdaHedge` is a regret minimizing algorithm of the exponential weights family, that achieves a $\mathcal{O}(\sqrt{T})$ regret for bounded losses (de Rooij et al., 2014). At each time step, we get a weight vector $\boldsymbol{\omega}_t$ from $\mathcal{L}_{\boldsymbol{\omega}}^{i_t}$ where i_t is the empirical best answer. However, a bandit algorithm cannot play a fraction vector. We can incorporate a tracking procedure (see line 20 in Algorithm 2) to circumvent this difficulty. For the opponent learner, we choose to use the best response, that is the most confusing model as formalized in Line 14 of the pseudo-code.

The present procedure involves also an optimization problem, but simpler than that of `D-Tracking`, and is computationally more feasible in practice (note that other combination of sub-algorithms for both opponents are possible, see e.g. Degenne et al. 2019; Ménard 2019).

Sample complexity. In this paper, since we assume that our model $\boldsymbol{\ell} \in [0,1]^{d\times K}$, there exists constants $\gamma_{\boldsymbol{\ell}}$ and $D_{\boldsymbol{\ell}}$, that only depend on the model $\boldsymbol{\ell}$, such that for all $y \in [0,1]$, the function $x \mapsto d(x,y)$ is $\gamma_{\boldsymbol{\ell}}$-Lipschitz on $[0,1]$ and $d(x,y) \leq D_{\boldsymbol{\ell}}$ (see Appendix F of Degenne et al. 2019 for detailed discussions).

According to Theorem 2 of Degenne et al. (2019), the sample complexity of Algorithm 2 at the stopping time τ_δ is bounded by T_δ as defined in Theorem 8, which is a non-asymptotic bound that depends on regrets incurred by both the `AdaHedge` learner and the best response learner. For `AdaHedge`, the regret incurred is $R^k(t) = \sqrt{t\log(dK)}\log(t)$, and the best-response learner has zero-regret: $R^{\boldsymbol{\lambda}}(t) \leq 0$. And with $\beta(t,\delta) \approx \log(1/\delta) + o(t)$, the asymptotic optimality of Algorithm 2 is also retained. We do not intend to reproduce the proof here since it can be (almost) adopted directly from the proof of Theorem 2 by Degenne et al. (2019) (up to a factor of d).

Theorem 8. *The sample complexity of Algorithm 2 on model $\boldsymbol{\ell}$ is*

$$\mathbb{E}_{\boldsymbol{\ell}}[\tau_\delta] \leq T_\delta + \mathtt{CST}.$$

The quantity T_δ is defined as

$$T_\delta \triangleq \max\left\{ t \in \mathbb{N} : t \leq \beta(t,\delta)/D_{\boldsymbol{\ell}} + C_{\boldsymbol{\ell}}(R^{\boldsymbol{\lambda}}(t) + R^k(t) + \mathcal{O}(t\log(t))) \right\}$$

where $C_{\boldsymbol{\ell}}$[3] depends on the model $\boldsymbol{\ell}$.

B A simple algorithm for regret minimization and its analysis

B.1 A combinatorial lemma

We now describe in detail an algorithm based on the proof idea of Theorem 4. As we stressed, it is more viable to consider the proportion (weight) of arm pulls, in particular the optimal weight of arm pulls for regret minimization. To simplify the problem, we first show that the optimal distribution is a linear combination over d arms.

[3] See Appendix D of Degenne et al. (2019) for an exact definition.

Algorithm 2 A gamified algorithm for best-arm identification

Input: Learners for each possible answer $(\mathcal{L}^i_{\boldsymbol{\omega}})_{i\in\mathcal{I}}$, threshold function $\beta(\cdot,\delta)$, exploration bonus $f(\cdot)$, number of forced exploration rounds N
pull each arm N rounds
for $t > KN$ **do**
  // Stopping rule
  if at time t, we have

$$\max_{k\in[K]} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\hat{\boldsymbol{\ell}}_{t-1})} \left(\sum_{k\in[K]} \sum_{j\in[d]} N_{t-1,k} d(\hat{\ell}^{(j)}_{t-1,k}, \lambda^{(j)}_k) \right) > \beta(t-1,\delta)$$

  then
    stop and **return**
    $\hat{\imath} = i^\star(\tilde{\boldsymbol{\ell}}_{t-1})$
  end if
  // Best answer
  $i_t = i^\star(\tilde{\boldsymbol{\ell}}_{t-1})$
  // The learner plays
  Get $\boldsymbol{\omega}_t$ from $\mathcal{L}^{i_t}_{\boldsymbol{\omega}}$ and update $\mathbf{W}_t = \mathbf{W}_{t-1} + \boldsymbol{\omega}_t$
  // Best response from the nature

$$\boldsymbol{\lambda}_t \in \operatorname*{arg\,min}_{\boldsymbol{\lambda}\in \mathrm{Alt}(\hat{\boldsymbol{\ell}}_{t-1})} \left(\sum_{k\in[K]} \sum_{j\in[d]} \omega_{t,k} d(\hat{\ell}^{(j)}_{t-1,k}, \lambda^{(j)}_k) \right)$$

  // Feed optimistic losses
  For $k \in [K]$, let

$$U_{t,k} = \max\left(\frac{f(t-1)}{N_{t-1,k}}, \max_{\xi\in\{c_{t,k},d_{t,k}\}} \sum_{j\in[d]} d(\xi, \lambda^{(j)}_{t,k}) \right)$$

  Feed $-\sum_{k\in[K]} \omega_k U_{t,k}$ to learner $\mathcal{L}^{i_t}_{\boldsymbol{\omega}}$
  // Track the weights
  Pull $\mathcal{A}_t \in \arg\min_{k\in[K]} N_{t-1,k} - W_{t,k}$
end for

Lemma 9. *In the case of d losses, there exists a $\boldsymbol{\omega}^\star$ such that it has at most d non-zero elements.*

Proof. We first define the quadrant $H^+ \triangleq \{\mathbf{x} | x_i \geq 0\}$ and we define an addition operation of two sets A and B as $A + B \triangleq \{a + b | a \in A, b \in B\}$. For any compact set X we note that

$$\inf_{\mathbf{x}\in X} \max(x_1, ..., x_d) = \inf_{\mathbf{y}\in(X+H^+)\bigcap \mathrm{Diag}} y_1 \,,$$

where $\mathrm{Diag} \triangleq \{\mathbf{x} | x_1 = x_2 = ... = x_d\}$. Let $A \triangleq \mathrm{Conv}\left(\left\{ \boldsymbol{\ell}_{k|\star} \right\}_{k\in[K]} \right)$ be the convex hull over the relative losses of all K arms. Then we have

$$\ell_{\boldsymbol{\omega}^\star|\star} = \inf_{\mathbf{y}\in(A+H^+)\bigcap \mathrm{Diag}} y_1 \,.$$

Therefore, there exists $\boldsymbol{\omega} \in A$ and $h \in H^+$ such that $\ell_{\boldsymbol{\omega}^\star|\star} = \boldsymbol{\omega}^\top \boldsymbol{\ell}^{(i)}_{|\star} + h_i$ for all $i \in [d]$. Moreover, it is obvious that there is at least one $i \in [d]$ such that $h_i = 0$. Thus, here $\boldsymbol{\omega}$ is an optimal weight, i.e., $\ell_{\boldsymbol{\omega}^\star|\star} = \ell_{\boldsymbol{\omega}|\star}$. Furthermore, the vector $\boldsymbol{\omega}^\top \boldsymbol{\ell}^{(i)}_{|\star}$ is not an interior point of A, since this would enable a $(-1, ..., -1)$ direction translation, thus $\boldsymbol{\omega}^\top \boldsymbol{\ell}^{(i)}_{|\star}$ is on a surface of A, a convex hull of finite points in a $d-1$ dimension space, we conclude that there exists such an $\boldsymbol{\omega}$ with at most d non-zero elements.

□

B.2 A straightforward algorithm

We assume that $K \geq d$. According to Lemma 9, we only need to consider linear combinations of d arms. We define a combinatorial arm $\mathbf{c} \in C$ as a set of d arms where $C = \{\{c_1, ..., c_d\} \subseteq [K] | c_1 < c_2 < \ldots < c_d\}$. For all $\mathbf{c} \in C$, we are interested in the quantity $\ell^{(i)}_{\mathbf{c}|\star} \triangleq \min_{\boldsymbol{\alpha} \in \Sigma_d} \boldsymbol{\alpha}^\top \boldsymbol{\ell}^{(i)}_{\mathbf{c}|\star}$, where $\boldsymbol{\ell}^{(i)}_{\mathbf{c}|\star} = (\ell^{(i)}_{c_1|\star}, \ldots, \ell^{(i)}_{c_d|\star})$ is the vector of the i-th relative losses of arm set $\mathbf{c}$. We further denote by $\ell_{\mathbf{c}|\star} = \max_{i \in [d]} \ell^{(i)}_{\mathbf{c}|\star}$ the ℓ^∞-norm of the relative loss and $\ell^{(i)}_{\mathbf{c}} = \ell^{(i)}_{\mathbf{c}|\star} + \ell^{(i)}_{\star_i}$ the absolute loss of the combinatorial arm.

A straightforward idea of algorithm for regret minimization is to track the values of $\ell_{\mathbf{c}|\star}$ for every $\mathbf{c} \in C$. We thus need to have a good estimate of the relative loss of all d combinations of arms. We propose Combinatorial Play (CP) as shown in Algorithm 3. The empirical relative loss of arm $\mathbf{c} \in C$ w.r.t. $\hat{\star}_i$ is defined as

$$\hat{\ell}_{\mathbf{c}|\hat{\star}} = \max_{i \in [d]} \min_{\boldsymbol{\alpha} \in \Sigma_d} \boldsymbol{\alpha}^\top \hat{\boldsymbol{\ell}}^{(i)}_{\mathbf{c}|\hat{\star}} . \tag{9}$$

Let $\hat{\boldsymbol{\alpha}}_{\mathbf{c}}$ denote the value of $\boldsymbol{\alpha}$.

Our algorithm thus chooses among $\mathbf{c} \in C$ and calculates the empirical optimal allocation $\hat{\boldsymbol{\alpha}}_{\mathbf{c}} \in \Sigma_d$ among $\mathbf{c}$. Then we use the tracking procedure from the literature (see e.g. Garivier and Kaufmann 2016) to decide which real arm to pull.

Algorithm 3 The algorithm of CP

Input: time horizon T and number of forced exploration rounds N
pull each arm N rounds
$\hat{\star}_i = \arg\min_{k \in [K]} \hat{\ell}^{(i)}_k$ for $i \in [d]$
for all $\mathbf{c} \in C$, we calculate its estimate $\hat{\ell}_{\mathbf{c}|\hat{\star}}$ and its optimal allocation $\hat{\boldsymbol{\alpha}}_{\mathbf{c}}$ based on Eq. (9)
$\hat{\mathbf{c}} \in \arg\min_{\mathbf{c} \in C} \hat{\ell}_{\mathbf{c}|\hat{\star}}$ and the corresponding optimal allocation $\hat{\boldsymbol{\alpha}}_{\hat{\mathbf{c}}}$.
for $t = KN + 1, \cdots, T$ **do**
    Pull arm $\mathcal{A}_t$ according to probability distribution $\hat{\boldsymbol{\alpha}}_{\hat{\mathbf{c}}}$ over $\hat{\mathbf{c}}$.
end for

B.3 Analysis of CP

We analyze CP in this section. Our main result is stated below.

Theorem 10. *Assume that $\hat{\ell}^{(i)}_k \in [0,1]$ for all $k \in [K]$ and $i \in [d]$, CP achieves a $\widetilde{\mathcal{O}}(T^{2/3})$ regret.*

Proof. First, according to Lemma 5, we have $\mathbb{P}[E_{1,t}] \geq 1 - dK/t^2$. When $E_{1,t}$ holds, we have for all $\mathbf{c} \in C$

$$\hat{\ell}_{\mathbf{c}|\hat{\star}} = \max_{i \in [d]} \min_{\boldsymbol{\alpha} \in \Sigma_d} \boldsymbol{\alpha}^\top \hat{\boldsymbol{\ell}}^{(i)}_{\mathbf{c}|\hat{\star}} \geq \max_{i \in [d]} \min_{\alpha \in \Sigma_d} \left(\boldsymbol{\alpha}^\top \boldsymbol{\ell}^{(i)}_{\mathbf{c}|\star} - 2\sqrt{\frac{2\log(t)}{N}} \right) = \ell_{\mathbf{c}|\star} - 2\sqrt{\frac{2\log(t)}{N}} .$$

And similarly, we have

$$\hat{\ell}_{\mathbf{c}|\hat{\star}} \leq \ell_{\mathbf{c}|\star} + 2\sqrt{\frac{2\log(t)}{N}} .$$

Then the regret is

$$\begin{aligned}
\mathbb{E}[R_{\text{CP}}(T)] &\leq KN + \sum_{t=KN+1}^{T} \mathbb{E}\left[\mathbb{1}_{\{\neg E_{1,t}\}}\right] + (\ell_{\mathbf{c}|\star} - \ell_{\boldsymbol{\omega}^\star|\star})(T - KN)\mathbb{E}\left[\mathbb{1}_{\{\forall t, E_{1,t}\}}\right] \\
&\leq KN + \sum_{t=KN+1}^{T} \frac{dK}{t^2} + (\hat{\ell}_{\mathbf{c}|\hat{\star}} - \min_{\mathbf{c}} \ell_{\mathbf{c}|\star})(T - KN)\mathbb{E}\left[\mathbb{1}_{\{\forall t, E_{1,t}\}}\right]
\end{aligned}$$

$$
\begin{aligned}
&+ 2\sqrt{\frac{2\log(t)}{N}}(T-KN)\mathbb{E}\left[\mathbb{1}_{\{\forall t, E_{1,t}\}}\right] \\
&\leq KN + \frac{d}{N} + \left(\hat{\ell}_{\mathbf{c}|\hat{\star}} - \min_{\mathbf{c}}\left(\hat{\ell}_{\mathbf{c}|\hat{\star}} - 2\sqrt{\frac{2\log(t)}{N}}\right)\right)(T-KN)\mathbb{E}\left[\mathbb{1}_{\{\forall t, E_{1,t}\}}\right] \\
&+ 2\sqrt{\frac{2\log(t)}{N}}(T-KN)\mathbb{E}\left[\mathbb{1}_{\{\forall t, E_{1,t}\}}\right] \\
&\leq KN + \frac{d}{N} + 4\sqrt{\frac{2\log(T)}{N}}(T-KN)\,,
\end{aligned}
$$

which completes the proof by setting $N = (32T^2\log(T)/K^2)^{\frac{1}{3}}$. □