Globally Optimal Algorithms for Fixed-Budged Best Arm Identification

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider the fixed-budget best arm identification problem where the goal is to find the arm of the largest mean with a fixed number of samples. It is known that the probability of misidentifying the best arm is exponentially small to the number of rounds. However, limited characterizations have been discussed on the rate (exponent) of this value. In this paper, we characterize the optimal rate as a result of global optimization over all possible parameters. We introduce two rates, R^{go} and R^{go}_{∞}, corresponding to lower bounds on the misidentification probability, each of which is associated with a proposed algorithm. The rate R^{go} is associated with R^{go}-tracking, which can be efficiently implemented by a neural network and is shown to outperform existing algorithms. However, this rate requires a nontrivial condition to be achievable. To deal with this issue, we introduce the second rate R^{go}_{∞}. We show that this rate is indeed achievable by introducing a conceptual algorithm called delayed optimal tracking (DOT).

1 Introduction

We consider K-armed best arm identification problem with T samples. In this problem, each arm $i \in [K] = \{1, 2, \dots, K\}$ is associated with (unknown) distribution $P_i \in \mathcal{P}$ for some class of distributions $\mathcal{P}$. Upon choosing arm i, the forecaster observes reward $X(t)$, which is independently drawn from P_i. The forecaster then tries to identify (one of) the best arm[1] $\mathcal{I}^* = \arg\max_i \mu_i$ with the largest mean $\mu^* = \max_i \mu_i$ for $\mu_i = \mathbb{E}_{X \sim P_i}[X]$. The problem[2] is called the best arm identification (BAI, Audibert et al. (2010)), or the ranking and selection (R&S, Hong et al. (2021)).

To this aim, the forecaster uses some algorithm that would adaptively choose an arm based on its history of rewards. At each round t, the algorithm chooses one of the arms $I(t) \in [K]$ and receives the corresponding reward $X(t)$. After the T-th round, the algorithm outputs a recommendation arm $J(T) \in [K]$, which corresponds to an estimator of the best arm. The misidentification probability is expressed by $\mathbb{P}[J(T) \notin \mathcal{I}^*]$, which will be referred to as the probability of the error (PoE) throughout the paper. Best arm identification has two settings. In the fixed confidence setting, the forecaster minimizes the number of draws T until the confidence level on the PoE reaches a given value $\delta \in (0, 1)$. In this case, T is a stopping time that can be chosen adaptively. In the fixed-budget setting, the forecaster tries to minimize the PoE given a constant T. In this paper, we shall focus on the fixed-budget setting. In

[1] We use $\mathcal{I}^* = \mathcal{I}^*(\boldsymbol{P}) \subset [K]$ as the set of best arms and $i^* = i^*(\boldsymbol{P}) \in \mathcal{I}^*(\boldsymbol{P})$ as one of them (ties are broken in an arbitrary way). These differences does not matter much in this paper.

[2] See Section 1.3 regarding the related work on BAI and R&S.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

general, a good algorithm for the fixed-confidence setting is very different from that for the fixed-budget setting. To be more specific, an algorithm for the fixed-confidence setting can be instance-wise optimal[3]. Namely, several algorithms exist (Garivier and Kaufmann, 2016) that can be optimized for each instance of distributions $\boldsymbol{P} = (P_1, P_2, \ldots, P_K)$ as far as we consider algorithms called δ-PAC. By contrast, an algorithm for the fixed-budget setting requires consideration of the possibility that improving the PoE for an instance $\boldsymbol{P}$ worsens the PoE for another instance $\boldsymbol{P}'$. Thus, we must consider a kind of a global optimization of the performance over all possible $\boldsymbol{P} \in \mathcal{P}^K$.

1.1 Global optimality in the fixed-budget setting

In the fixed-budget setting, the PoE decays exponentially to T as $\exp(-RT)$ for some *rate* $R > 0$. The instance-wise optimality given above is no longer available here. To demonstrate this, assume that we make an estimate of $\boldsymbol{P}$ based on the initial $o(T)$ rounds, say, $\sqrt{T}$ rounds. In this case, we can obtain the estimate of $\boldsymbol{P}$ that is ϵ-correct with probability $\exp(-\epsilon^2 O(\sqrt{T})) = \exp(-o(T))$. However, this estimation does not help to improve the rate of exponential convergence. In other words, estimating $\boldsymbol{P}$ requires non-negligible (i.e. $O(T)$) cost for exploration. As a result, we cannot optimize the PoE for each instance $\boldsymbol{P}$ unlike the fixed-confidence setting. Instead, to discuss optimality in the fixed-budget setting, we must choose a complexity function $H(\boldsymbol{P})$, and the performance of an algorithm must be evaluated on the rate normalized by the complexity.

In literature, little is known about the optimal rate of the exponent. Audibert et al. (2010) proposed the successive rejects (SR) algorithm, which has the rate of $1/(\log K)$ with the complexity function $H_2(\boldsymbol{P}) := \max_{i \in [K]} \frac{i}{\Delta_i^2}$ for $\Delta_i = \max_j \mu_j - \mu_i$ satisfying $\Delta_1 \leq \Delta_2 \leq \cdots \leq \Delta_K$. Carpentier and Locatelli (2016) showed a particular set of instances such that this rate matches the lower bound up to a constant factor. However, the constant used there is by far loose[4], and there is limited discussion on the actual rate of such algorithms.

1.2 Contributions

This paper tightly characterizes the optimal minimax rate of the PoE as a result of a global optimization given $\boldsymbol{P}$. Let $H = H(\boldsymbol{P}) : \mathcal{P}^K \to \mathbb{R}^+$ be any continuous complexity measure. We then discuss the best possible rate $R > 0$ such that the PoE is bounded by $\exp(-RT/H(\boldsymbol{P}) + o(T))$ for all $\boldsymbol{P} \in \mathcal{P}^K$ and make the following contributions.

- We derive an upper bound on R (corresponding to a lower bound of the PoE), denoted by R^{go}, which we obtain by considering a class of oracle algorithms that can determine the allocation of trials to each arm knowing the final empirical distribution after T rounds (Theorem 1).

- We propose an algorithm (R^{go}-tracking) that greedily tracks this oracle allocation based on the current empirical distribution (Section 2.1). Though this oracle allocation is expressed by a complicated minimax optimization, we propose a technique to learn this by a neural network and empirically confirm that the PoE of the learned algorithm is close to the lower bound (Sections 3 and 4). We also discuss that the algorithm is unlikely to provably achieve the bound even when the minimax problem is perfectly solved because of the impossibility of the tracking.

- We tighten the PoE lower bound by weakening the oracle algorithms. Based on this refined bound, we propose the delayed optimal tracking (DOT) algorithm that asymptotically achieves the tightened lower bound for Bernoulli and Gaussian arms, though the algorithm is computationally almost infeasible (Sections 2.2 and 2.3).

In summary, we propose a nearly-tight PoE lower bound with a computationally feasible algorithm that is empirically close to this bound. We also propose a provably tight lower bound and matching algorithm in a computationally infeasible form. Notation is listed in the supplementary material.

[3] A more complete discussion on this topic can be found in Section D of supplementary material.
[4] Theorem 1 therein includes a large constant 400.

1.3 Related work

Compared with the works of the fixed-confidence BAI, less is known about the fixed-budget BAI. For example, a book on this subject (Lattimore and Szepesvári, 2020) spends only two pages on the fixed-budget BAI.[5] Many algorithms designed for the fixed-confidence BAI, such as D-tracking (Kaufmann et al., 2016), do not have a finite-time PoE guarantee when we apply them to the fixed-budget setting. Nevertheless, there are two well-known fixed-budget BAI algorithms: Successive rejection (SR, Audibert et al. (2010)) and successive halving (SH, Shahrampour et al. (2017)). Both SR and SH progressively narrow the candidate of the best arm at the end of each segment. While SR discards one arm after each segment, SH discards half of the remaining arms after each segment. SR and SH have the guarantee on PoE of the rate $\exp(-RT/H_2(\boldsymbol{P}))$ for some constant $R > 0$. Other fixed-budget BAI algorithms, such as UCB-E (Audibert et al., 2010) and UGapE (Gabillon et al., 2012), require the knowledge of minimum gap $\min_i \Delta_i$, and thus are not universal to all best arm identification instances.

Another literature on this topic is the ranking and selection (R&S) problems (Powell and Ryzhov, 2018; Hong et al., 2021). Although the goal of R&S problems is to identify the best arm, many R&S papers do not consider the estimation error of $\boldsymbol{P}$ in a finite time. As a result, algorithms therein do not have the guarantee on the PoE in the best arm identification setting. The optimal computing budget allocation (OCBA Chen et al. (2000); Glynn and Juneja (2004)) algorithm tries to minimize the PoE assuming the plug-in estimator matches the true parameter. Bayesian R&S algorithms try to solve the dynamic programming of minimizing the PoE given a prior, which is computationally prohibitive, and thus approximated solutions have been sought (Frazier et al., 2008; Powell and Ryzhov, 2018).

2 Globally optimal algorithm

In this section, we derive several lower bounds on the PoE and propose algorithms to empirically or theoretically achieve these bounds.

First, we formalize the problem. Let $\mathcal{P}$ be a known class of reward distributions. We consider the case where $\mathcal{P}$ is the set of Bernoulli distributions with mean $\Theta \subset [0,1]$ (including the case $\Theta = [0,1]$), or Gaussian distributions with mean in $\Theta \subset \mathbb{R}$ (including the case $\Theta = \mathbb{R}$) and known variance $\sigma^2 > 0$. It should be noted that many parts of results in this paper can be generalized to much wider classes of distributions, but it makes the notation much longer and is discussed in Appendix E.

When we derive lower bounds and construct algorithms, we introduce $\mathcal{Q}$ as a class of distributions corresponding to the estimated distributions of the arms. Namely, we set $\mathcal{Q}$ as the set of all Bernoulli (resp. Gaussian) distributions with mean in $[0,1]$ (resp. $\mathbb{R}$) when $\mathcal{P}$ is the set of Bernoulli (resp. Gaussian) distributions with mean in Θ. As such, we take $\mathcal{Q} \supset \mathcal{P}$ so that the estimator of P_i is always in $\mathcal{Q}$. In these models, we identify the distribution $P_i \in \mathcal{P}$ with its mean parameter in $\Theta \subset \mathbb{R}$.

Our interest lies in the rate $\lim_{T\to\infty} \frac{1}{T} \log(1/\mathbb{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})])$ of convergence of the PoE. Since we are interested in lower and upper bounds of the rate of algorithms including those requiring the knowledge of T, we define the rate for a sequence of algorithms $\{\pi_T\}$ by

$$R(\{\pi_T\}) = \inf_{\boldsymbol{P}\in\mathcal{P}^K} H(\boldsymbol{P}) \liminf_{T\to\infty} \frac{1}{T} \log(1/\mathbb{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})]). \tag{1}$$

Here, a larger $R(\{\pi_T\})$ corresponds to a faster convergence of the PoE.

2.1 PoE for oracle algorithms

First, we derive a lower bound on the PoE that is unlikely to be achievable but strongly related to an optimal algorithm. Let $D(P\|Q) = \mathbb{E}_{X\sim P}[\frac{\mathrm{d}P}{\mathrm{d}Q}(X)]$ be the Kullback–Leibler (KL) divergence between P and Q. Then we have the following bound.

[5] Section 33.3 therein.

Algorithm 1: R^{go}-Tracking

input : (ϵ-)optimal solution $(\boldsymbol{r}^*(\cdot), J^*(\cdot))$ of (2).

Draw each arm once.

for $t = K+1, 2, \ldots, T$ **do**

  Draw arm $\arg\max_{i\in[K]} \{r_i^*(\boldsymbol{Q}(t-1)) - N_i(t-1)/(t-1)\}$.

return $J(T) = J^*(\boldsymbol{Q})$.

Theorem 1. Under any sequence of algorithm $\{\pi_T\}$ it holds that

$$R(\{\pi_T\}) \leq \sup_{\boldsymbol{r}(\cdot)\in\Delta^K,\, J(\cdot)\in[K]} \inf_{\boldsymbol{Q}\in\mathcal{Q}^K} \inf_{\boldsymbol{P}\in\mathcal{P}^K : J(\boldsymbol{Q})\notin\mathcal{I}^*(\boldsymbol{P})} H(\boldsymbol{P}) \sum_{i\in[K]} r_i(\boldsymbol{Q}) D(Q_i\|P_i) =: R^{\mathrm{go}}, \quad (2)$$

where the outer supremum is taken over all functions $\boldsymbol{r}(\cdot) : \mathcal{Q}^K \to \Delta^K$, $J(\cdot) : \mathcal{Q}^K \to [K]$.

All proofs are provided in the supplementary material owing to page limitation. This theorem states that under any algorithm there exists an instance $\boldsymbol{P}$ such that the PoE is at least $\exp(-TR^{\mathrm{go}}/H(\boldsymbol{P}) + o(T))$. Intuitively speaking, the bound in Theorem 1 corresponds to the best possible rate of oracle algorithms that can determine the allocation as $\boldsymbol{r} = \boldsymbol{r}^*(\boldsymbol{Q}) \in \Delta^K$ knowing the final empirical mean $\boldsymbol{Q} = \boldsymbol{Q}(T)$, where $\boldsymbol{r}^*(\cdot)$ is the (ϵ-)optimal[6] solution of (2).

From the technical viewpoint, the main difference from the lower bound on the fixed-confidence setting is that we also have to consider candidates of empirical distributions $\boldsymbol{Q}$ as well as the true distributions $\boldsymbol{P}$. This makes the analysis much more difficult, because a slight difference of the empirical distribution might (possibly discontinuously) affect the allocation unlike the difference of the true distribution $\boldsymbol{P}$ unknown to the algorithm. A naive analysis just depending on the empirical distribution fails because of this discontinuity of the allocation. To overcome this difficulty, we adopt a technique inspired by the *typical set analysis* often used in the information theory (Cover and Thomas, 2006). We define the *typical allocation* for each candidate of empirical distribution $\boldsymbol{Q}$ and prove the theorem by evaluating the error probability based on the typical allocation.

Remark 1. We can take arbitrary $H(\boldsymbol{P}) > 0$ as a complexity measure, but R^{go} might become zero if $H(\boldsymbol{P})$ is not taken reasonably. When $R^{\mathrm{go}} = 0$ any algorithm trivially satisfies PoE $\leq \exp(-TR^{\mathrm{go}}/H(\boldsymbol{P}) + o(T))$. This means that any algorithm is minimax-optimal in terms of $H(\boldsymbol{P})$, that is, such choice of $H(\boldsymbol{P})$ gives meaningless results.

In the actual trial, the algorithm can only know the empirical mean $\boldsymbol{Q}(t-1)$ at the beginning of the current round t and we cannot ensure the achievability of the bound for oracle algorithms. Despite this, one reasonable choice of the algorithm would be to keep tracking this optimal allocation $\boldsymbol{r}^*(\boldsymbol{Q}(t-1))$, expecting that the current empirical mean $\boldsymbol{Q}(t-1)$ is close to $\boldsymbol{Q}(T)$. R^{go}-tracking in Algorithm 1 is the algorithm based on this idea. Here, $N_i(t-1)$ is the number of times that the arm i is drawn at the beginning of the t-th round, and it draws the arm such that the current fraction of the allocation $N_i(t-1)/(t-1)$ is the most insufficient compared with the estimated optimal allocation $\boldsymbol{r}^*(\boldsymbol{Q}(t-1))$.

As we will see in Section 4, the empirical performance of Algorithm 1 is very close to the PoE lower bound stated above. However, it is difficult to expect that this algorithm provably achieves this bound in general because of the following: We could prove that R^{go}-tracking is optimal if the fraction of allocation always satisfies $N_i(t)/t = \boldsymbol{r}(\boldsymbol{Q}(t)) + o(1)$, that is, the algorithm can track the ideal allocation $\boldsymbol{r}(\boldsymbol{Q}(t))$. However, this does not generally hold. For example, the empirical mean $\boldsymbol{Q}(t)$ sometimes changes rapidly in the Gaussian case. Whilst this event occurs with exponentially small probability, the PoE itself is also an exponentially small probability and it is highly nontrivial to specify in which case the tracking failure probability becomes negligible.

Remark 2. Eq. (2) also involves the optimization of the recommendation arm $J(\boldsymbol{Q})$ as well as $\boldsymbol{r}(\boldsymbol{Q})$. We can easily see that it is optimal to set $J(\boldsymbol{Q}) = i^*(\boldsymbol{Q})$, that is, taking

[6] This paper uses $\epsilon > 0$ as an arbitrarily small gap to the optimal solution. An asterisk is used to denote optimality.

the empirical best arm as the recommendation arm when $\mathcal{P} = \mathcal{Q}$ since $R(\pi)$ becomes zero otherwise. However, $J(\boldsymbol{Q}) = i^*(\boldsymbol{Q})$ might not hold for $\boldsymbol{Q} \notin \mathcal{P}$ when $\mathcal{P} \subsetneq \mathcal{Q}$.

2.2 PoE considering trackability

To construct an algorithm that is provably optimal, we begin with refining the PoE lower bound by weakening the "strength" of the oracle algorithm.

We consider splitting T rounds into B batches of size $\lfloor T/B \rfloor$ or $\lfloor T/B \rfloor + 1$. Let

$$\boldsymbol{r}^B = (\boldsymbol{r}_1(\boldsymbol{Q}_1),\, \boldsymbol{r}_2(\boldsymbol{Q}_1, \boldsymbol{Q}_2),\, \boldsymbol{r}_3(\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3), \dots, \boldsymbol{r}_B(\boldsymbol{Q}_1, \dots, \boldsymbol{Q}_B))$$

be a sequence of B functions, where $\boldsymbol{r}_b : \mathcal{Q}^{Kb} \to \Delta^K$ corresponds the allocation in the b-th batch when the empirical means of the first b batches are $\boldsymbol{Q}^b = (\boldsymbol{Q}_1, \boldsymbol{Q}_2, \dots, \boldsymbol{Q}_b)$. Based on this class of allocation rule, we have the following PoE lower bound.

Theorem 2. (PoE Bound for batch-oracle algorithms) Under any sequence of algorithms π_T and $B \in \mathbb{N}$,

$$R(\{\pi_T\}) \leq \sup_{\boldsymbol{r}^B(\cdot), J(\cdot)} \inf_{\boldsymbol{Q}^B \in \mathcal{Q}^{KB}} \inf_{\boldsymbol{P}: J(\boldsymbol{Q}^B) \notin \mathcal{I}^*(\boldsymbol{P})} \frac{H(\boldsymbol{P})}{B} \sum_{i \in [K], b \in [B]} r_{b,i} D(Q_{b,i} \| P_i) =: R_B^{\mathrm{go}}. \quad (3)$$

Here, the outer supremum is taken over all functions $\boldsymbol{r}^B(\cdot) = (\boldsymbol{r}_1(\cdot), \boldsymbol{r}_2(\cdot), \dots, \boldsymbol{r}_B(\cdot))$ for $\boldsymbol{r}_b(\cdot) : \mathcal{Q}^{Kb} \to \Delta^K$ and $J(\cdot) : \mathcal{Q}^{KB} \to [K]$.

Theorem 1 is the special case of this theorem with $B = 1$. This bound corresponds to the best bound of oracle algorithms that can determine the allocation of the b-th batch knowing the empirical distribution of this batch. It is tighter than that given in Theorem 1, as the oracle considered here cannot know the empirical distribution of the later batches $b+1, b+2, \dots, B$. It follows that we can obtain the following result.

Corollary 3. We have $R_B^{\mathrm{go}} \leq R^{\mathrm{go}}$ for any $B \in \mathbb{N}$.

We will show that $R_\infty^{\mathrm{go}} := \lim_{B \to \infty} R_B^{\mathrm{go}}$ exists and is the best possible rate.

2.3 Matching algorithm

In this section, we derive an algorithm that has a rate that almost matches R_B^{go}. For any $\epsilon > 0$, let an ϵ-optimal solution of Eq. (3) be $(\boldsymbol{r}^{B,*}(\cdot), J^*(\cdot)) = (\boldsymbol{r}_1^*(\cdot), \boldsymbol{r}_2^*(\cdot), \boldsymbol{r}_3^*(\cdot), \dots, \boldsymbol{r}_B^*(\cdot), J^*(\cdot))$ with its objective at least

$$\inf_{\boldsymbol{Q}^B \in \mathcal{Q}^{KB}} \inf_{\boldsymbol{P}: J^*(\boldsymbol{Q}^B) \notin \mathcal{I}^*(\boldsymbol{P})} \frac{H(\boldsymbol{P})}{B} \sum_{i \in [K], b \in [B]} r_{b,i}^*(\boldsymbol{Q}^b) D(Q_{b,i} \| P_i) \geq R_B^{\mathrm{go}} - \epsilon.$$

We cannot naively follow the allocation $r_{b,i}^*(\boldsymbol{Q}^b)$ because it requires the empirical mean of the current batch $\boldsymbol{Q}_b$, which is not fully available until the end of the current batch. The delayed optimal tracking algorithm (DOT, Algorithm 2) addresses this issue. This algorithm divides T rounds into $B + K - 1$ batches, where the b-th batch corresponds to $(bT_B + 1, bT_B + 2, \dots, (b+1)T_B)$-th rounds for $T_B = T/(B + K - 1)$. Here, for simplicity, we assume that T is a multiple of $B + K - 1$. In the other case, we can reach almost the same result by just ignoring the last $T - (B + K - 1)\lfloor T/(B + K - 1) \rfloor$ rounds.

The crux of Algorithm 2 is to determine allocation $\boldsymbol{r}_b$ by using the *stored* empirical mean $\boldsymbol{Q}'_1, \boldsymbol{Q}'_2, \dots \boldsymbol{Q}'_B$ rather than the true empirical mean $\boldsymbol{Q}_1, \boldsymbol{Q}_2, \dots \boldsymbol{Q}_{B+K-1}$: The first K batches are devoted to uniform exploration and the samples are stored in a queue (though this explanation is not strict, in that the actual procedure is done after taking the mean of the stored samples). At the b-th batch for $b \geq K+1$, we draw each arm i for $n_{b,i} \approx T_B \boldsymbol{r}_b$ times[7], where $\boldsymbol{r}_b$ is determined based on the stored samples in the queue. When drawing arm i for $n_{b,i}$ times, we dequeue and open $n_{b,i}$ stored samples instead of opening the actual $n_{b,i}$ samples, the latter of which are enqueued and kept unopened.

By the nature of this algorithm we can ensure the following property.

[7] The $-K$ in Line 6 of Algorithm 2 is for the ceiling fractional values. This is reflected in the term T' in Theorem 5. If T is large compared to B, K, the difference between T and T' does not matter.

Algorithm 2: Delayed optimal tracking (DOT)

input : ϵ-optimal solution $\boldsymbol{r}^{B,*}(\cdot) = (\boldsymbol{r}_1^*(\cdot), \boldsymbol{r}_2^*(\cdot), \dots, \boldsymbol{r}_B^*(\cdot), J^*(\cdot))$ of (3).

for $b = 1, 2, \dots, K$ **do**

  Set $r_{b,i} = \mathbf{1}[i = b]$ for $i \in [K]$ and draw arm b for T_B times.

Set $\boldsymbol{Q}'_1 := \boldsymbol{Q}_K$ for the empirical mean $\boldsymbol{Q}_K$.

for $b = K+1, K+2, \dots, B+K-1$ **do**

  Compute $\boldsymbol{r}_b = (r_{b,1}, r_{b,2}, \dots, r_{b,K}) = \boldsymbol{r}^*_{b-K}(\boldsymbol{Q}'_1, \boldsymbol{Q}'_2, \dots, \boldsymbol{Q}'_{b-K})$.

  Draw each arm i for $n_{b,i}$ times, where $n_{b,i} \geq r_{b,i}(T_B - K)$ is taken so that $\sum_{i\in[K]} n_{b,i} = T_B$.

  Observe empirical mean $\boldsymbol{Q}_b$ of the batch.

  Update the *stored* empirical average as

$$\boldsymbol{Q}'_{b-K+1} = \boldsymbol{Q}'_{b-K} + \boldsymbol{r}_b(\boldsymbol{Q}_b - \boldsymbol{Q}'_{b-K}),$$

  where $\boldsymbol{r}_b\boldsymbol{Q}$ denotes the element-wise product.

Recommend $J(T) = J^*(\boldsymbol{Q}'_1, \boldsymbol{Q}'_2, \dots, \boldsymbol{Q}'_B)$.

Lemma 4. Assume that we run Algorithm 2. Then, the following inequality always holds:

$$\frac{1}{B+K-1} \sum_{i\in[K], b\in[B+K-1]} r_{b,i} D(Q_{b,i}||P_i) \geq \frac{B}{B+K-1} \frac{R^{\mathrm{go}}_B - \epsilon}{H(\boldsymbol{P})}. \tag{4}$$

Lemma 4 states that the empirical divergence of DOT given in the LHS of (4) almost matches the upper bound $R^{\mathrm{go}}_B/H(\boldsymbol{P})$ for sufficiently large B despite the delayed allocation. Using this property we obtain the following achievability bound.

Theorem 5. (Performance bound of Algorithm 2) The PoE of the DOT algorithm satisfies

$$\mathbb{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})] \leq \exp\left(-\frac{BT'}{B+K-1}\frac{R^{\mathrm{go}}_B - \epsilon}{H(\boldsymbol{P})} + f(K,B,T)\right),$$

where $T' = T - (B+K-1)K$ and $f(K,B,T) = 2BK\log(2T)$.

The following corollary is immediate since $f(K,B,T) = o(T)$ holds for fixed K, B.

Corollary 6. The worst-case rate of the DOT algorithm $\pi_{\mathrm{DOT},T}$ satisfies

$$R(\{\pi_{\mathrm{DOT},T}\}) \geq \frac{B}{B+K-1}(R^{\mathrm{go}}_B - \epsilon).$$

2.4 Optimality

In this section, we show the rate $R(\pi_{\mathrm{DOT}})$ of DOT becomes arbitrarily close to optimal when we take a sufficiently large number of batches B.

Theorem 7. (Optimality of DOT) Assume $H(\boldsymbol{P})$ be such that $R^{\mathrm{go}} < \infty$. Then, the limit

$$R^{\mathrm{go}}_\infty := \lim_{B\to\infty} R^{\mathrm{go}}_B \tag{5}$$

exists. Moreover, for any $\eta > 0$, there exist parameters B, ϵ such that the following holds on the performance of the DOT algorithm:

$$R(\{\pi_{\mathrm{DOT},T}\}) = \inf_{\boldsymbol{P}\in\mathcal{P}} H(\boldsymbol{P}) \liminf_{T\to\infty} \frac{\log(1/\mathbb{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})])}{T} \geq R^{\mathrm{go}}_\infty - \eta. \tag{6}$$

Remark 3. (η-optimality) Since $R(\{\pi_T\}) \leq R^{\mathrm{go}}_B$ holds for any sequence of algorithms $\{\pi_T\}$ and $b \in \mathbb{N}$, we have

$$R(\{\pi_T\}) \leq \inf_{B\in\mathbb{N}} R^{\mathrm{go}}_B \leq \liminf_{B\to\infty} R^{\mathrm{go}}_B = R^{\mathrm{go}}_\infty$$

from (5). Therefore the rate of the DOT algorithm given in (6) is optimal up to η for arbitrary small $\eta > 0$. This essentially states that, no algorithm can be η-better than DOT in terms of the rate against the worst-case instance $\boldsymbol{P}$.

Algorithm 3: Gradient descent method for $\boldsymbol{\theta}$

input : learning rate η

while *not converged* **do**

  `/* Compute` $\boldsymbol{P}^{\min}$ `and` $\boldsymbol{Q}^{\min}$ `which minimizes the negative exponent */`

  Set $E^{\min} \leftarrow \infty$.

  for $n_1 = 1, 2, \ldots, N^{\mathrm{true}}$ **do**

    Sample true parameters $\boldsymbol{P}$ from $\mathcal{P}$ uniformly at random.

    for $n_2 = 1, 2, \ldots, N^{\mathrm{emp}}$ **do**

      Sample $\boldsymbol{Q}$ from $\{\boldsymbol{Q} \in \mathcal{Q}^K : \mathcal{I}^*(\boldsymbol{Q}) \cap \mathcal{I}^*(\boldsymbol{P}) = \emptyset\}$.

      if $E(\boldsymbol{P}, \boldsymbol{Q}; \boldsymbol{\theta}) < E^{\min}$ **then**

        $\boldsymbol{P}^{\min} \leftarrow \boldsymbol{P}, \quad \boldsymbol{Q}^{\min} \leftarrow \boldsymbol{Q}, \quad E^{\min} \leftarrow E(\boldsymbol{P}^{\min}, \boldsymbol{Q}^{\min}; \boldsymbol{\theta})$

  Update parameters $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} E(\boldsymbol{P}^{\min}, \boldsymbol{Q}^{\min}; \boldsymbol{\theta})$.

Algorithm 4: R^{go}-Tracking by Neural Network (TNN)

Draw each arm once.

for $t = K+1, 2, \ldots, T$ **do**

  Draw arm $\arg\max\left(r_{\boldsymbol{\theta},i}(\boldsymbol{Q}(t-1)) - N_i(t-1)/(t-1)\right)$.

return $J(T) = \arg\max Q_i(T)$ (empirical best arm).

Remark 4. (Utility of the DOT algorithm) Although DOT (Algorithm 2) has an asymptotically optimal rate R^{go}_{∞}, it is difficult to calculate, or to even approximate, the optimal solution of (3) since it is not an optimization of a finite-dimensional vector but an optimization of function $\boldsymbol{r}^B$, which has high input dimension proportional to B. In this sense, DOT algorithm as well as Theorem 7 is purely theoretical thus far, and the existence of a computationally tractable and provably optimal algorithm is an important open question.

3 Learning

In this section, we propose a method to learn $\boldsymbol{r}(\boldsymbol{Q})$ of Eq. (2) by utilizing a neural network to practically realize R^{go}-tracking in Algorithm 1. Throughout this section, we assume a class of algorithms satisfying $J(\boldsymbol{Q}) \in \mathcal{I}^*(\boldsymbol{Q})$, that is, algorithms that recommend the empirical best arm, which is guaranteed to be optimal when $\mathcal{P} = \mathcal{Q}$ (see Remark 2).

3.1 Learning allocation

Let $\boldsymbol{r}_{\boldsymbol{\theta}}(\boldsymbol{Q}) : \mathcal{Q}^K \to \Delta^K$ be a neural network with a set of parameters $\boldsymbol{\theta}$. We consider alternately optimizing $r_{\boldsymbol{\theta}}(\cdot)$ and $(\boldsymbol{P}, \boldsymbol{Q})$, and we update $\boldsymbol{\theta}$ via mini-batch gradient descent. Given a complexity function $H(\boldsymbol{P})$, Eq. (2) is defined as the minimum over all $(\boldsymbol{P}, \boldsymbol{Q})$ such that the best arm is different. Our learning method (Algorithm 3) uses L mini-batches[8]. Let

$$E(\boldsymbol{P}, \boldsymbol{Q}; \boldsymbol{\theta}) \coloneqq H(\boldsymbol{P}) \sum_{i=1}^{K} r_{\boldsymbol{\theta},i}(\boldsymbol{Q}) D\left(Q_i \| P_i\right). \tag{7}$$

Given allocation $\boldsymbol{r}_{\boldsymbol{\theta}}$, Eq. (7) is the negative log-likelihood (rate) of the bandit instance $\boldsymbol{P}$ given the empirical means $\boldsymbol{Q}$. At each batch, it obtains the pair $\boldsymbol{P}^{\min}, \boldsymbol{Q}^{\min}$ such that Eq. (7) is minimized. Specifically, for each iteration, we sample N^{true} candidates of true means $\boldsymbol{P}$ uniformly from $\mathcal{P}^K$, then for each $\boldsymbol{P}$, we sample N^{emp} values of empirical means $\boldsymbol{Q} \in \mathcal{Q}^K$ such that $\mathcal{I}^*(\boldsymbol{Q}) \cap \mathcal{I}^*(\boldsymbol{P}) = \emptyset$ uniformly at random.

[8] Algorithm 3 is a conceptual explanation and the actual implementation used a momentum method. See Section 4.1 for implementation details.

3.2 Tracking by neural network

Having trained $\boldsymbol{r}_{\boldsymbol{\theta}}$, we propose the R^{go}-Tracking by Neural Network (TNN) algorithm (Algorithm 4), which is an implementation of R^{go}-Tracking by the trained neural network. This algorithm draws the arm such that the current fraction of samples $N_i(t-1)/(t-1)$ is the most insufficient compared with the learned allocation $\boldsymbol{r}_{\boldsymbol{\theta}}(\boldsymbol{Q}(t-1))$.

4 Simulation

This section tests numerically the performance of TNN algorithm. We compared the performance of TNN (Algorithm 4) with two algorithms: Uniform algorithm, which samples each arm in a round-robin fashion, and Successive Rejects (SR, Audibert et al., 2010), where the entire trial is divided into segments before the game starts, and one arm with the smallest estimated mean reward is removed for each segment.

We consider Bernoulli bandits with $K = 3$ arms, where each mean parameter is in $[0, 1]$. In particular, we consider the three sets of true parameters: (instance 1) $\boldsymbol{P} = (0.5, 0.45, 0.3)$, (instance 2) $\boldsymbol{P} = (0.5, 0.45, 0.05)$, and (instance 3) $\boldsymbol{P} = (0.5, 0.45, 0.45)$. The number of the rounds T is fixed to 2000, and we repeated the experiments for 10^5 times.

4.1 Training neural networks

Here, we show experimental details for training neural networks for the TNN algorithm discussed in Section 3.2.

We used the complexity measure $H_1(\boldsymbol{P}) = \sum_{i \neq i^*(\boldsymbol{P})} (P^* - P_i)^{-2}$ as a standard choice of $H(\boldsymbol{P})$. We used the neural network with four layers (including the input layer and output layer), where we used the ReLU for the activation functions and introduced the skip-connection (He et al., 2016) between each hidden layer to make training the network easier. To obtain the map to Δ^K, we adopted the softmax function. The number of nodes in the hidden layers was fixed to $K \times 3$. We used AdamW (Loshchilov and Hutter, 2019) with a learning rate 10^{-3} and weight decay 10^{-7} to update the parameters.

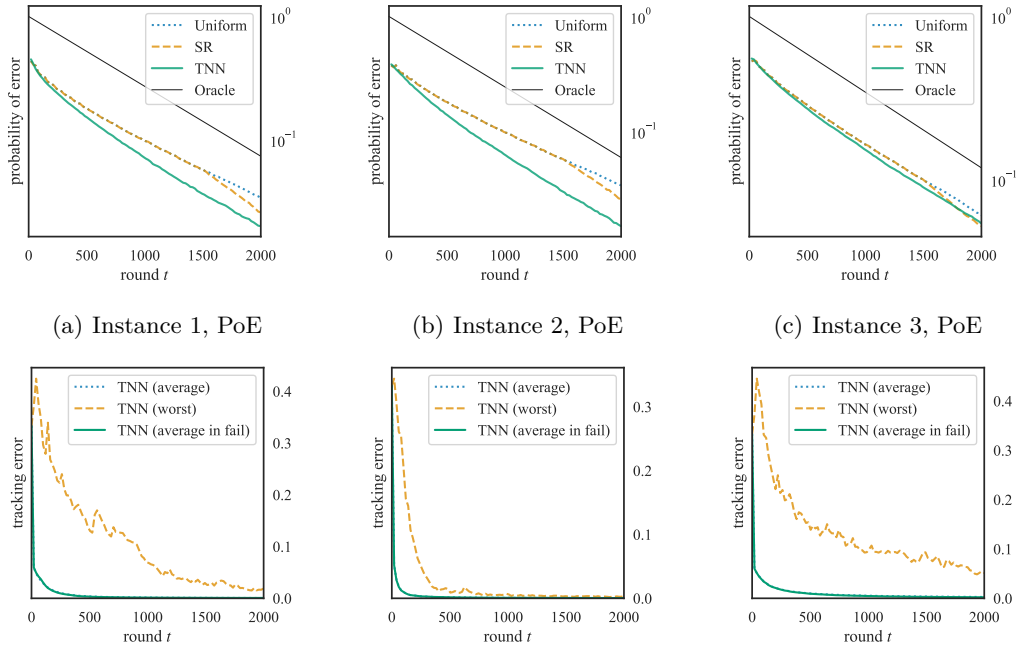
For training the neural network, we ran Algorithm 3 with $N^{\mathrm{true}} = 32$ and $N^{\mathrm{emp}} = 90$. Additionally, to allow the neural network to easily learn $\boldsymbol{r}$, the elements of $\boldsymbol{P} = (P_1, P_2, \ldots, P_K)$ were sorted beforehand. Other details of the implementation is given in Appendix C.

4.2 Experimental results

Figure 1 illustrates the results of our simulations. Each column corresponds to the result for each instance.

The first row ((a)–(c)) shows the PoE of the compared methods when the arm with the largest empirical mean is regarded as the estimated best arm $J(t)$ at each round t. Here, the black line represents $\exp(-t \inf_{\boldsymbol{Q}} \sum_i r_{\boldsymbol{\theta},i}(\boldsymbol{Q}) D(Q_i \| P_i))$, which corresponds to the exponent of the oracle algorithm that can perfectly track the allocation $r_{i,\boldsymbol{\theta}}(\boldsymbol{Q})$. Therefore, the asymptotic slope of TNN cannot be better than that of the black line. We can see from the figures that the slope of the TNN is close to the oracle algorithm and performs better than or comparable to the other algorithms. Note that this is the result for fixed time horizon T. Though the final slope of SR may look outperforming TNN, it just comes from the fact that SR is not anytime and is an algorithm that divides T rounds into several segments.

The second row ((d)–(f)) shows the tracking error of the TNN algorithm, which is defined as $\mathrm{disc}(t) = \max_{i \in [K]} |r_i(\boldsymbol{Q}(t)) - N_i(t)/t|$, which measures the discrepancy between the ideal allocation $r_i(\boldsymbol{Q}(t))$ and the actual allocation $N_i(t)/t$. If this quantity is $o(T)$ in almost all trials (including the ones where the algorithm failed to recommend the best arm) and all instances, then we can guarantee $R^{\mathrm{go}} = R^{\mathrm{go}}_{\infty}$. The labels TNN (average), TNN (worst), and TNN (average in fail) corresponds to the average tracking error of all trials, the worst-case tracking error and the average tracking error of all failed trials, respectively. The fact that 'TNN (worst)' is small at $T = 2{,}000$ implies that the gap between R^{go} and R^{go}_{∞} is small, which supports the reasonableness of algorithms based on R^{go}.



(a) Instance 1, PoE (b) Instance 2, PoE (c) Instance 3, PoE

(d) Instance 1, tracking error (e) Instance 2, tracking error (f) Instance 3, tracking error

Figure 1: Bernoulli bandits, $K = 3, T = 2000$, average over 10^5 trials.

5 Conclusion

This paper considered the fixed-budget best arm identification problem. We identified the minimax rate R^{go}_{∞} on the exponent of the probability of error by introducing a matching algorithm (DOT algorithm). Optimization on the rate R^{go}_{∞} is very challenging to implement, and we considered the learning of a simpler optimization problem of rate R^{go} by using a neural network (TNN algorithm). The TNN algorithm ourperformed existing algorithms. A number of possible lines of future work include the following points.

- A more scalable learning of $\boldsymbol{r}(\boldsymbol{Q})$: TNN adopted a neural network to obtain the oracle allocation $\boldsymbol{r}(\boldsymbol{Q})$ associated with the rate bound. While its empirical results are promising and support our theoretical findings, the current experiment is limited to the case of $K = 3$ arms because the learning is very costly even for small K. A more sophisticated learning algorithm is desired to realize R^{go}-tracking for larger K.

- Identifying the existence (or non-existence) of the gap: though the empirical results suggest that R^{go} is very close (or maybe equal) to the optimal rate R^{go}_{∞} for the Bernoulli case, a formal analysis of this gap for general cases is demanded since the DOT algorithm to achieve R^{go}_{∞} is computationally almost infeasible.

- A bound for another rate measure: we defined the worst-case rate of convergence by (1), which first takes the limit of T and then takes the worst-case instance $\boldsymbol{P}$. Another natural choice of the rate would be to exchange them, that is, to consider

$$R'(\{\pi_T\}) = \liminf_{T\to\infty} \inf_{\boldsymbol{P}\in\mathcal{P}^K} \frac{H(\boldsymbol{P})}{T} \log(1/\mathbb{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})]) \le R(\{\pi_T\}).$$

  Whereas Theorems 1 and 2 on the upper bounds of $R(\pi)$ are still valid for $R'(\{\pi_T\}) \le R(\{\pi_T\})$, the current achievability analysis does not apply and analyzing the tightness of R^{go}_{∞} for $R'(\{\pi_T\})$ is an open problem.

References

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT 2010 - The 23rd Conference on Learning Theory*, pages 41–53. Omnipress, 2010.

L. Jeff Hong, Weiwei Fan, and Jun Luo. Review on ranking and selection: A new perspective. *Frontiers of Engineering Management*, 8(3):321–343, Sep 2021.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, Proceedings of Machine Learning Research, 2016.

Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17 (1):1–42, 2016.

Shahin Shahrampour, Mohammad Noshad, and Vahid Tarokh. On sequential elimination algorithms for best-arm identification in multi-armed bandits. *IEEE Transactions on Signal Processing*, 65(16):4281–4292, 2017.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, 2012.

Warren Buckler Powell and Ilya O. Ryzhov. *Optimal Learning*. Second edition. Unpublished Manuscript, March 2018. URL `https://castlelab.princeton.edu/wp-content/uploads/2019/02/Powell-OptimalLearningWileyMarch112018.pdf`.

Chun-Hung Chen, Jianwu Lin, Enver Yücesan, and Stephen E. Chick. Simulation budget allocation for further enhancing theefficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270, July 2000.

Peter Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *Winter Simulation Conference*, volume 1. IEEE, 2004.

Peter I. Frazier, Warren B. Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM J. Control Optim.*, 47(5):2410–2439, sep 2008.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, second edition, July 2006. ISBN 0471241954.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.2
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] This is a methodology paper and ethical concerns do not directly apply here.

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendices
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We will publish the source code upon acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Table 1: Major notation

symbol	definition
K	number of the arms
T	number of the rounds
B	number of the batches
T_B	$= T/(B+K-1)$
T'	$= T-(B+K-1)K$
$I(t)$	arm selected at round t
$X(t)$	reward at round t
$J(T)$	recommendation arm at the end of round T
$\boldsymbol{P} \in \mathcal{P}^K$	true parameters
$P_i \in \mathcal{P}$	i-th component of $\boldsymbol{P}$
$\mathcal{I}^* = \mathcal{I}^*(\boldsymbol{P})$	Set of best arms under parameter $\boldsymbol{P}$
$i^*(\boldsymbol{P})$	one arm in $\mathcal{I}^*(\boldsymbol{P})$ (taken arbitrary in a deterministic way)
$\boldsymbol{Q} \in \mathcal{Q}^K$	estimated parameters of $\boldsymbol{P}$
$Q_i \in \mathcal{Q}$	i-th component of $\boldsymbol{Q}$
$\boldsymbol{Q}_b \in \mathcal{Q}^K$	estimated parameters of b-th batch
$Q_{b,i} \in \mathcal{Q}$	i-th component of $\boldsymbol{Q}_b$
$\boldsymbol{Q}^b$	$= (\boldsymbol{Q}_1, \boldsymbol{Q}_2, \dots, \boldsymbol{Q}_b)$
$\boldsymbol{Q}'_b$	stored parameters (in Algorithm 2)
$Q'_{b,i} \in \mathcal{Q}$	i-th component of $\boldsymbol{Q}'_b$
$\mathcal{P}$	hypothesis class of $\boldsymbol{P}$
$\mathcal{Q}$	distribution of estimated parameter of $\boldsymbol{Q}$
$D(Q\|P)$	KL divergence between Q and P
Δ^K	probability simplex in K dimensions
$\boldsymbol{r} \in \Delta^K$	allocation (proportion of arm draws)
r_i	i-th component of $\boldsymbol{r}$
$\boldsymbol{r}_b \in \Delta^K$	allocation at b-th batch
$r_{b,i}$	i-th component of $\boldsymbol{r}_b$
$\boldsymbol{r}^b$	$= (\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_b)$
$\boldsymbol{n}_b$	Number of draws of Algorithm 2 at b-th batch
$n_{b,i}$	i-th component of $\boldsymbol{n}_b$. Note that $n_{b,i} \geq r_{b,i}(T_B - K)$ holds.
$J(\boldsymbol{Q}^B)$	recommendation arm given $\boldsymbol{Q}^B$
$(\boldsymbol{r}^{B,*}, J^*)$	ϵ-optimal allocation
$H(\cdot)$	complexity measure of instances
$R(\{\pi_T\})$	worst-case rate of PoE of sequence of algorithms $\{\pi_T\}$ in (1)
R^{go}	best possible $R(\{\pi_T\})$ for oracle algorithms in (2)
R^{go}_B	best possible $R(\{\pi_T\})$ for B-batch oracle algorithms in (3)
R^{go}_∞	$\lim_{B\to\infty} R^{\mathrm{go}}_B$. Limit exists (Theorem 7)
$\boldsymbol{\theta}$	model parameter of the neural network
$\boldsymbol{r}_{\boldsymbol{\theta}}$	allocation by a neural network with model parameters $\boldsymbol{\theta}$
$r_{\boldsymbol{\theta},i}$	i-th component of $\boldsymbol{r}_{\boldsymbol{\theta}}$

A Notation table

Table 1 summarizes our notation.

B Computational resources

We used a modern laptop (Macbook Pro) for learning $\boldsymbol{\theta}$. It took less than one hour to learn $\boldsymbol{\theta}$. For conducting a large number of simulations (i.e., Run TNN and existing algorithms for

10^5 times), we used a 2-CPU Xeon server of sixteen cores. It took less than twelve hours to complete simulations. We did not use a GPU for computation.

C Implementation details

To speed up computation, the same $\boldsymbol{Q}$ was used for each $\boldsymbol{P}$ with the same optimal arm $i^*(\boldsymbol{P})$ in the mini-batches.

The final model $\boldsymbol{\theta}$ of the neural network is chosen as follows. We stored sequence of models $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$ during training (Algorithm 3). Among these models, we chose the one with the maximum objective function $\arg\max_l \min_{(\boldsymbol{P},\boldsymbol{Q})\in(\mathcal{P}^{\mathrm{emp}},\mathcal{Q}^{\mathrm{emp}})} E(\boldsymbol{P},\boldsymbol{Q};\boldsymbol{\theta}^{(l)})$. Here, the minimum is taken over a finite dataset of size $|\mathcal{P}^{\mathrm{emp}}| = 32$ and $|\mathcal{Q}^{\mathrm{emp}}| = 10^5$.

The black lines in Figure 1 (a)–(c) representing $\exp(-t \inf_{\boldsymbol{Q}} \sum_i r_{\boldsymbol{\theta},i}(\boldsymbol{Q}) D(Q_i \| P_i))$ are computed by the grid search of $\boldsymbol{Q}$ with each Q_i separated by intervals of 5.0×10^{-3}.

D Instance optimality in the fixed-confidence setting

For sufficiently small $\delta > 0$, the asymptotic sample complexity for fixed-confidence setting is known. Namely, any fixed-confidence algorithm is required to draw at least

$$\liminf_{\delta \to +0} \frac{T}{\log(\delta^{-1})} \geq C^{\mathrm{conf}}(\boldsymbol{P}) \tag{8}$$

times, where

$$C^{\mathrm{conf}}(\boldsymbol{P}) = \left(\sup_{\boldsymbol{r}(\boldsymbol{P}) \in \Delta^K} \inf_{\boldsymbol{P}' : i^*(\boldsymbol{P}') \notin \mathcal{I}^*(\boldsymbol{P})} \sum_{i=1}^K r_i D(P_i \| P_i') \right)^{-1}.$$

Garivier and Kaufmann (2016) proposed C-Tracking and D-Tracking algorithms that have a sample complexity bound that matches Eq. (8). This bound implies that an algorithm adapts the true parameter $\boldsymbol{P}$ without paying essential cost of exploration. In fact, building an optimal algorithm such that Eq. (8) holds is not very difficult.

Roughly speaking, a $o(\log(1/\delta))$ cost, say, uniform exploration of $\sqrt{\log(1/\delta)}$ rounds, enables us to obtain enough accuracy the bound of

$$|\hat{\boldsymbol{P}} - \boldsymbol{P}| \sim (\log(1/\delta))^{-1/4} = o(1) \tag{9}$$

with probability $1 - o(1)$. The expected value of the stopping time is bounded as:

$$\underbrace{\sqrt{\log(1/\delta)}}_{\text{uniform exploration}} + \underbrace{(C^{\mathrm{conf}}(\boldsymbol{P}) + o(1)) \log(\delta^{-1})}_{\text{stopping time bound under Eq. (9)}} + \underbrace{o(1)}_{\text{probability of Eq. (9) does not hold}} \times O\left(\log(\delta^{-1})\right).$$

The first and the third terms does not hurt the optimal rate, and thus the bound of Eq. (8) is derived.

E Extension to wider models

In the main body of the paper, we assumed that $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ are Bernoulli or Gaussian distributions. Many parts of the results of the paper can be extended to exponential families or distributions over a support set $\mathcal{S} \subset \mathbb{R}$.

Let us consider an exponential family of form

$$\mathrm{d}P(x|\theta) = \exp(\theta^\top Y(x) - A(\theta)) \mathrm{d}F(x),$$

where F is a base measure and $\theta \in \Theta \subset \mathbb{R}^d$ is a natural parameter. We assume that $A'(\theta) = \mathbb{E}_{X \sim F(\cdot|\theta)}[Y(X)]$ has the inverse $(A')^{-1} : \mathrm{im}(Y) \to \Theta$, where $\mathrm{im}(Y)$ is the image of Y.

Let $\mathcal{P}$ be a class of reward distributions. $\mathcal{P}$ can be the family of distributions over a known support $\mathcal{S} \subset \mathbb{R}$. We can also consider the case where $\mathcal{P}$ is the above exponential family with possibly restricted parameter set $\Theta' \subset \Theta$. For example, $\mathcal{P}$ can be the set of Gaussian distributions with mean parameters in $[0, 1]$ and variances in $(0, \infty)$.

When we derive lower bounds and construct algorithms, we introduce $\mathcal{Q}$ as a class of distributions corresponding to the estimated reward distributions of the arms. We set $\mathcal{Q} = \mathcal{P}$ when $\mathcal{P}$ is a family of distributions over a known support $\mathcal{S} \subset \mathbb{R}$. When we consider a natural exponential family with parameter set $\Theta' \subset \Theta$, we set $\mathcal{Q}$ as this exponential family with parameter set Θ, so that the estimator of P_i is always within $\mathcal{Q}$. For example, if we consider $\mathcal{P}$ as a class of Gaussians with means in $[0, 1]$ and variances in $(0, \infty)$, $\mathcal{Q}$ is the class of all Gaussians with means in $(-\infty, \infty)$ and variances in $(0, \infty)$.

In Algorithm 2, we use a convex combination of distributions Q and Q'. The key property used in the analysis is the convexity of KL divergence between distributions. When we consider the family $\mathcal{P}$ of distributions over support set $\mathcal{S}$, the convexity

$$D(\alpha Q + (1-\alpha)Q' \,\|P) \le \alpha D(Q\|P) + (1-\alpha)D(Q'\|P)$$

holds for any $P, Q, Q' \in \mathcal{Q}$ when we define $\alpha Q + (1-\alpha)Q'$ as the mixture of Q and Q' with weight $(\alpha, 1-\alpha)$. When $\mathcal{P}$ is the exponential family, the convexity of the KL divergence holds when $\alpha Q + (1-\alpha)Q'$ is defined as the distribution in this family such that the expectation of the sufficient statistics $Y(X)$ is equal to $\alpha\mathbb{E}_{X\sim Q}[Y(X)] + (1-\alpha)\mathbb{E}_{X\sim Q'}[Y(X)]$. Note that this corresponds to taking the convex combination of the empirical means when we consider Bernoulli distributions or Gaussian distributions with a known variance.

By the convexity of the KL divergence, most parts of the analysis apply to $\mathcal{P}$ in this section and we straightforwardly obtain the following result.

Proposition 8. Theorems 1 and 2, Corollary 3, and Lemma 4 hold under the models $\mathcal{P}$ with the definition of the convex combination in this section.

The only part where the analysis is limited to Bernoulli or Gaussian is Theorem 5 on the PoE upper bound of the DOT algorithm. The subsequent results immediately follow if Theorem 5 is extended to the models in this section. Since the key property of the DOT algorithm in Lemma 4 on the trackability of the empirical divergence is still valid for these models, we expect that Theorem 5 can also be extended though it remains as an open question.

F Proofs

F.1 Proofs of Theorems 1 and 2

We only give the proof of Theorem 2 since Theorem 1 is a special case of this theorem with $B = 1$.

In this proof, we consider many candidates of the true distributions $\boldsymbol{P} = (P_1, P_2, \dots, P_K)$ and we write $\boldsymbol{P}[A]$ to denote the probability of the event A when the reward of each arm i follows P_i. We divide T rounds into B batches, and the b-th batch corresponds to $(t_b, t_b+1, \dots, t_{b+1}-1)$-th rounds for $b \in [B]$ and $t_b = \lfloor (b-1)T/B \rfloor + 1$. We define the history of the b-th batch by $\mathcal{H}_b = ((I(t_b), X(t_b)), (I(t_b+1), X(t_b+2)), \dots, I(t_{b+1}-1), X(t_{b+1}-1))$. The entire history is denoted by $\mathcal{H}^B = (\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_B)$.

By slight abuse of notation, we interchangeably write

$$\mathcal{H}_b = ((X_{b,1,1}, X_{b,1,2}, \dots, X_{b,1,N_{b,1}}), (X_{b,2,1}, X_{b,2,2}, \dots, X_{b,2,N_{b,2}}), \dots, (X_{b,K,1}, X_{b,K,2}, \dots, X_{b,K,N_{b,K}})),$$

where $X_{b,k,n}$ is the reward of the n-th draw of arm k in the b-th batch and $N_{b,k}$ is the number of draws of arm k in the b-th batch.

We adopt the formulation of the random rewards such that every $X_{b,k,m}$, the m-th reward of arm k in the b-th batch, is randomly generated before the game begins, and if an arm is drawn then this reward is revealed to the player. Then $Y_{b,k,m}$ is well-defined even if arm k is not drawn m times in the b-th batch.

Fix an arbitrary $\epsilon > 0$. We define sets of "typical" rewards under $\boldsymbol{Q}^B$: we write $\mathcal{T}_\epsilon(\boldsymbol{Q}^B)$ to denote the event such that rewards (a part of which might be unrevealed as noted above) satisfy

$$\sum_{k=1}^{K} \left| \left(n_{b,k} D(Q_{b,k} \| P_k) - \sum_{m=1}^{n_{b,k}} \log \frac{\mathrm{d}Q_{b,k}}{\mathrm{d}P_k}(X_{b,k,m}) \right) \right| \leq \epsilon T / B \tag{10}$$

for any $b \in [B]$ and $\boldsymbol{n}_b = (n_{b,1}, n_{b,2}, \ldots, n_{b,K})$ such that $\sum_{k \in [K]} n_{b,k} = t_{b+1} - t_b$. By the strong law of large numbers, $\lim_{T \to \infty} \boldsymbol{Q}^B[\mathcal{T}_\epsilon^B(\boldsymbol{Q}^B)] = 1$, where $\boldsymbol{Q}^B[\cdot]$ denotes the probability under which $X_k(t)$ follows distribution $Q_{b,k}$ for $t \in \{t_b, t_b + 1, \ldots, t_{b+1} - 1\}$.

We define $\boldsymbol{r}^B = \boldsymbol{r}^B(\mathcal{H}^B) = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_B)$ for $\boldsymbol{r}_b = \boldsymbol{n}_b / (t_{b+1} - t_b)$, where $\boldsymbol{n}_b = (n_{b,1}, n_{b,2}, \ldots, n_{b,K})$. In other words, $\boldsymbol{r}_b$ is the fractions of arm-draws in the b-th batch under history $\mathcal{H}_b$.

Let $\mathcal{R}_{T,B} \subset (\Delta^K)^B$ be the set of all possible $\boldsymbol{r}^B(\mathcal{H}^B)$. Since $n_{b,k} \in \{0, 1, \ldots, t_{b+1} - t_b\}$ and $t_{b+1} - t_b \leq T/B + 1$, we see that

$$|\mathcal{R}_{T,B}| \leq (T/B + 2)^{KB},$$

which is polynomial in T.

Consider an arbitrary algorithm π and define the "typical" allocation $\boldsymbol{r}^b(\boldsymbol{Q}^b; \pi, \epsilon)$ and decision $J(\boldsymbol{Q}^B; \pi, \epsilon)$ of the algorithm for distributions $\boldsymbol{Q}^b = (\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_b)$ as

$$\begin{aligned}
\boldsymbol{r}_1(\boldsymbol{Q}^1; \pi, \epsilon) &= \operatorname*{arg\,max}_{\boldsymbol{r} \in \mathcal{R}_{T,1}} \boldsymbol{Q}^1 \left[\boldsymbol{r}_1(\mathcal{H}_1) = \boldsymbol{r} \middle| \mathcal{T}_\epsilon(\boldsymbol{Q}^B) \right], \\
\boldsymbol{r}_b(\boldsymbol{Q}^b; \pi, \epsilon) &= \operatorname*{arg\,max}_{\boldsymbol{r} \in \mathcal{R}_{T,b}} \boldsymbol{Q}^b \left[\boldsymbol{r}_b(\mathcal{H}_b) = \boldsymbol{r} \middle| \boldsymbol{r}^{b-1}(\mathcal{H}^{b-1}) = \boldsymbol{r}^{b-1}(\boldsymbol{Q}^{b-1}; \pi, \epsilon),\ \mathcal{T}_\epsilon(\boldsymbol{Q}^B) \right], \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad b = 2, 3, \ldots, B, \\
J(\boldsymbol{Q}^B; \pi, \epsilon) &= \operatorname*{arg\,max}_{i \in [K]} \boldsymbol{Q}^B \left[J(T) = i \middle| \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B; \pi, \epsilon),\ \mathcal{T}_\epsilon(\boldsymbol{Q}^B) \right].
\end{aligned}$$

Then we have

$$\boldsymbol{Q}^B \left[\boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B; \pi, \epsilon) \middle| \mathcal{T}_\epsilon(\boldsymbol{Q}^B) \right] \geq \frac{1}{|\mathcal{R}_{T,B}|}, \tag{11}$$

$$\boldsymbol{Q}^B \left[J(T) = J(\boldsymbol{Q}^B; \pi, \epsilon) \middle| \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B; \pi, \epsilon),\ \mathcal{T}_\epsilon(\boldsymbol{Q}^B) \right] \geq \frac{1}{K}. \tag{12}$$

Lemma 9. Let $\epsilon > 0$ and algorithm π be arbitrary. Then, for any $\boldsymbol{P}, \boldsymbol{Q}^B$ be such that $J(\boldsymbol{Q}^B; \pi, \epsilon) \neq \mathcal{I}^*(\boldsymbol{P})$ it holds that

$$\frac{1}{T} \log \boldsymbol{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})] \geq -\frac{1}{B} \sum_{b=1}^{B} \sum_{k=1}^{K} \boldsymbol{r}_{b,k}(\boldsymbol{Q}^b; \pi, \epsilon) D(Q_{b,k} \| P_k) - \epsilon - \delta_{\boldsymbol{P}, \boldsymbol{Q}^B, \epsilon}(T)$$

for a function $\delta_{\boldsymbol{P}, \boldsymbol{Q}^B, \epsilon}(T)$ satisfying $\lim_{T \to \infty} \delta_{\boldsymbol{P}, \boldsymbol{Q}^B, \epsilon}(T) = 0$.

Proof. For arbitrary $\boldsymbol{Q}^B$ we obtain by a standard argument of a change of measures that

$$\begin{aligned}
&\boldsymbol{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})] \\
&\geq \boldsymbol{P}[\mathcal{T}_\epsilon(\boldsymbol{Q}^B),\ \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B; \pi, \epsilon),\ J(T) = J(\boldsymbol{Q}^B; \pi, \epsilon)] \\
&= \boldsymbol{P}[\mathcal{T}_\epsilon(\boldsymbol{Q}^B),\ \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B; \pi, \epsilon)] \\
&\qquad \times \boldsymbol{P}[J(T) = J(\boldsymbol{Q}^B; \pi, \epsilon) \mid \mathcal{H}^B \in \mathcal{T}_\epsilon(\boldsymbol{Q}^B),\ \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B; \pi, \epsilon)] \\
&= \boldsymbol{P}[\mathcal{T}_\epsilon(\boldsymbol{Q}^B),\ \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B; \pi, \epsilon)] \\
&\qquad \times \boldsymbol{Q}^B[J(T) = J(\boldsymbol{Q}^B; \pi, \epsilon) \mid \mathcal{H}^B \in \mathcal{T}_\epsilon(\boldsymbol{Q}^B),\ \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B; \pi, \epsilon)] \qquad (13) \\
&\geq \frac{1}{K} \boldsymbol{P}[\mathcal{T}_\epsilon(\boldsymbol{Q}^B),\ \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B; \pi, \epsilon)] \qquad \text{(by (12))}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{K}\mathbb{E}_{\boldsymbol{P}}\left[\mathbf{1}\left[\mathcal{H}^B \in \mathcal{T}_\epsilon(\boldsymbol{Q}^B),\, \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B;\pi,\epsilon)\right]\right] \\
&= \frac{1}{K}\mathbb{E}_{\boldsymbol{Q}^B}\left[\mathbf{1}\left[\mathcal{T}_\epsilon(\boldsymbol{Q}^B),\, \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B;\pi,\epsilon)\right]\prod_{b=1}^{B}\prod_{t=t_b}^{t_{b+1}-1}\frac{\mathrm{d}P_{I(t)}}{\mathrm{d}Q_{b,I(t)}}(X(t))\right] \\
&\geq \frac{1}{K}\mathbb{E}_{\boldsymbol{Q}^B}\left[\mathbf{1}\left[\mathcal{H}^B \in \mathcal{T}_\epsilon(\boldsymbol{Q}^B),\, \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B;\pi,\epsilon)\right]\right] \\
&\qquad \times \exp\left(-\frac{T}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} r_{b,k}(\boldsymbol{Q}^b;\pi,\epsilon)D(Q_{b,k}\|P_k) - \epsilon T\right) \qquad \text{(by (10))} \\
&= \frac{1}{K}\boldsymbol{Q}^B\left[\mathcal{T}_\epsilon(\boldsymbol{Q}^B),\, \boldsymbol{r}^B(\mathcal{H}^B) = \boldsymbol{r}^B(\boldsymbol{Q}^B;\pi,\epsilon)\right] \\
&\qquad \times \exp\left(-\frac{T}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} r_{b,k}(\boldsymbol{Q}^b;\pi,\epsilon)D(Q_{b,k}\|P_k) - \epsilon T\right) \\
&\geq \frac{\boldsymbol{Q}^B[\mathcal{H}^B \in \mathcal{T}_\epsilon(\boldsymbol{Q}^B)]}{K|\mathcal{R}_{T,B}|}\exp\left(-\frac{T}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} r_{b,k}(\boldsymbol{Q}^b;\pi,\epsilon)D(Q_{b,k}\|P_k) - \epsilon T\right), \qquad \text{(by (11))}
\end{aligned}
$$

where (13) holds since $J(T)$ does not depend on the true distribution $\boldsymbol{P}$ given the history $\mathcal{H}^B$. The proof is completed by letting $\delta_{\boldsymbol{P},\boldsymbol{Q}^B,\epsilon} = \log\frac{\boldsymbol{Q}^B[\mathcal{H}^B\in\mathcal{T}_\epsilon(\boldsymbol{Q}^B)]}{K|\mathcal{R}_{T,B}|}$. □

Proof of Theorem 2. For each $\boldsymbol{Q}^B$, let $\boldsymbol{r}^B(\boldsymbol{Q}^B;\{\pi_T\},\epsilon)$, $J(\boldsymbol{Q}^B;\{\pi_T\},\epsilon)$ be such that there exists a subsequence $\{T_n\}_n \subset \mathbb{N}$ satisfying

$$
\begin{aligned}
\lim_{n\to\infty}\boldsymbol{r}^B(\boldsymbol{Q}^B;\pi^{T_n},\epsilon) &= \boldsymbol{r}^B(\boldsymbol{Q}^B;\{\pi_T\},\epsilon), \\
J(\boldsymbol{Q}^B;\pi^{T_n},\epsilon) &= J(\boldsymbol{Q}^B;\{\pi_T\},\epsilon), \quad \forall n.
\end{aligned}
$$

Such $\boldsymbol{r}^B(\boldsymbol{Q}^B;\{\pi_T\},\epsilon) \in (\Delta^K)^B$ and $J(\boldsymbol{Q}^B;\{\pi_T\},\epsilon) \in [K]$ exist since $(\Delta^K)^B$ and $[K]$ are compact. By Lemma 9, for any $J(\boldsymbol{Q}^B;\{\pi_T\},\epsilon) \notin \mathcal{I}^*(\boldsymbol{P})$ we have

$$
\begin{aligned}
\liminf_{T\to\infty}\frac{1}{T}\log 1/\boldsymbol{P}[J(T)\notin\mathcal{I}^*(\boldsymbol{P})] &\leq \liminf_{n\to\infty}\frac{1}{T_n}\log 1/\boldsymbol{P}[J(T_n)\notin\mathcal{I}^*(\boldsymbol{P})] \\
&\leq \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} r_{b,k}(\boldsymbol{Q}^b;\{\pi_T\},\epsilon)D(Q_{b,k}\|P_k) + \epsilon. \qquad (14)
\end{aligned}
$$

By taking the worst case we have

$$
\begin{aligned}
R(\{\pi_T\}) &= \inf_{\boldsymbol{P}} H(\boldsymbol{P})\liminf_{T\to\infty}\frac{1}{T}\log 1/\boldsymbol{P}[J(T)\notin\mathcal{I}^*(\boldsymbol{P})] \\
&\leq \inf_{\boldsymbol{P}\in\mathcal{P}^K,\boldsymbol{Q}^B\in\mathcal{Q}^{KB}:J(\boldsymbol{Q}^B;\{\pi_T\},\epsilon)\notin\mathcal{I}^*(\boldsymbol{P})}\frac{H(\boldsymbol{P})}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} r_{b,k}(\boldsymbol{Q}^b;\{\pi_T\},\epsilon)D(Q_{b,k}\|P_k) + \epsilon.
\end{aligned}
$$

By optimizing $\{\pi^T\}$ we have

$$
\begin{aligned}
R(\{\pi_T\}) &\leq \sup_{\{\pi_T\}}\inf_{\boldsymbol{P}\in\mathcal{P}^K} H(\boldsymbol{P})\liminf_{T\to\infty}\frac{1}{T}\log 1/\boldsymbol{P}[J(T)\notin\mathcal{I}^*(\boldsymbol{P})] \\
&= \sup_{\boldsymbol{r}^B(\cdot),J(\cdot)}\sup_{\{\pi_T\}:\boldsymbol{r}^B(\cdot;\{\pi_T\},\epsilon)=\boldsymbol{r}^B(\cdot)}\inf_{\boldsymbol{P}\in\mathcal{P}^K}\frac{H(\boldsymbol{P})}{B}\liminf_{T\to\infty}\frac{1}{T}\log 1/\boldsymbol{P}[J(T)\notin\mathcal{I}^*(\boldsymbol{P})] \\
&\leq \sup_{\boldsymbol{r}^B(\cdot),J(\cdot)}\sup_{\{\pi_T\}:\boldsymbol{r}^B(\cdot;\{\pi_T\},\epsilon)=\boldsymbol{r}^B(\cdot)}\inf_{\boldsymbol{P}\in\mathcal{P}^K,\boldsymbol{Q}^B\in\mathcal{Q}^{KB}:J(\boldsymbol{Q}^B)\notin\mathcal{I}^*(\boldsymbol{P})}\frac{H(\boldsymbol{P})}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} r_{b,k}(\boldsymbol{Q}^b)D(Q_{b,k}\|P_k) + \epsilon \\
&\qquad\qquad \text{(by (14))} \\
&\leq \sup_{\boldsymbol{r}^B(\cdot),J(\cdot)}\inf_{\boldsymbol{P}\in\mathcal{P}^K,\boldsymbol{Q}^B\in\mathcal{Q}^{KB}:J(\boldsymbol{Q}^B)\notin\mathcal{I}^*(\boldsymbol{P})}\frac{H(\boldsymbol{P})}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} r_{b,k}(\boldsymbol{Q}^b)D(Q_{b,k}\|P_k) + \epsilon.
\end{aligned}
$$

We obtain the desired result since $\epsilon > 0$ is arbitrary. □

F.2 Proof of Corollary 3

Proof of Corollary 3. We have

$$
\begin{aligned}
&R_B^{\mathrm{go}} \\
&:= \sup_{\boldsymbol{r}^B(\boldsymbol{Q}^B), J(\boldsymbol{Q}^B)} \inf_{\boldsymbol{Q}^B} \inf_{\boldsymbol{P}: J(\boldsymbol{Q}^B) \notin \mathcal{I}^*(\boldsymbol{P})} \frac{H(\boldsymbol{P})}{B} \sum_{i \in [K], b \in [B]} r_{b,i} D(Q_{b,i} || P_i) \\
&\leq \sup_{\boldsymbol{r}^B(\boldsymbol{Q}^B), J(\boldsymbol{Q}^B)} \inf_{\boldsymbol{Q}^B : \boldsymbol{Q}_1 = \boldsymbol{Q}_2 = \cdots = \boldsymbol{Q}_B} \inf_{\boldsymbol{P}: J(\boldsymbol{Q}^B) \notin \mathcal{I}^*(\boldsymbol{P})} \frac{H(\boldsymbol{P})}{B} \sum_{i \in [K], b \in [B]} r_{b,i} D(Q_{b,i} || P_i) \quad \text{(inf over a subset).} \\
&= \sup_{\boldsymbol{r}^B(\boldsymbol{Q}), J(\boldsymbol{Q})} \inf_{\boldsymbol{Q}} \inf_{\boldsymbol{P}: J(\boldsymbol{Q}) \notin \mathcal{I}^*(\boldsymbol{P})} H(\boldsymbol{P}) \sum_{i \in [K]} \left(\frac{1}{B} \sum_{b \in [B]} r_{b,i} \right) D(Q_i || P_i) \\
&\quad \text{(by denoting } \boldsymbol{Q} = \boldsymbol{Q}_1 = \boldsymbol{Q}_2 = \ldots \boldsymbol{Q}_B) \\
&= \sup_{\boldsymbol{r}(\boldsymbol{Q}), J(\boldsymbol{Q})} \inf_{\boldsymbol{Q}} \inf_{\boldsymbol{P}: J(\boldsymbol{Q}) \notin \mathcal{I}^*(\boldsymbol{P})} H(\boldsymbol{P}) \sum_{i \in [K]} r_i D(Q_i || P_i) \\
&\quad \text{(by letting } r_i = (1/B) \sum_b r_{b,i}) \\
&= R^{\mathrm{go}} \quad \text{(by definition).}
\end{aligned}
$$

□

F.3 Additional Lemma

The following lemma is used to derive the regret bound.

Lemma 10. Assume that we run Algorithm 2. Then, for any $B_C \in K, K+1, \ldots, B$, it follows that

$$
\sum_{i,b \in [B_C]} r_{b,i} D(Q_{b,i} || P_i) \quad \geq \quad \sum_{i,a \in [B_C - K]} r^*_{a,i} D(Q'_{a,i} || P_i) + \sum_{i \in [K]} D(Q'_{B_C - K + 1, i} || P_i). \tag{15}
$$

Proof of Lemma 10. We use induction over $B_C = K, K+1, \ldots, B$. (i) It is trivial to derive Eq. (15) for $B_C = K$. (ii) Assume that Eq. (15) holds for B_C. In batch $B_C + 1$, the algorithm draws arms in accordance with allocation $\boldsymbol{r}_{B_C+1} = \boldsymbol{r}^*_{B_C - K + 1}$. We have,

$$
\begin{aligned}
&\sum_{i \in [K], b \in [B_C + 1]} r_{b,i} D(Q_{b,i} || P_i) \\
&\geq \sum_{i \in [K], a \in [B_C - K]} r^*_{a,i} D(Q'_{a,i} || P_i) + \sum_{i \in [K]} D(Q'_{B_C - K + 1, i} || P_i) + \underbrace{\sum_i r_{B_C+1,i} D(Q_{B_C+1,i} || P_i)}_{\text{Batch } B_C + 1} \\
&\quad \text{(by the assumption of the induction)} \\
&= \sum_i \left(\sum_{a \in [B_C - K]} r^*_{a,i} D(Q'_{a,i} || P_i) + r^*_{B_C - K + 1, i} D(Q'_{B_C - K + 1, i} || P_i) \right) + \sum_i \left(1 - r^*_{B_C - K + 1, i}\right) D(Q'_{B_C - K + 1, i} || P_i) \\
&\qquad + \sum_i r_{B_C+1,i} D(Q_{B_C+1,i} || P_i) \\
&= \sum_i \left(\sum_{a \in [B_C - K]} r^*_{a,i} D(Q'_{a,i} || P_i) + r^*_{B_C - K + 1, i} D(Q'_{B_C - K + 1, i} || P_i) \right) + \sum_i \left(1 - r_{B_C+1,i}\right) D(Q'_{B_C - K + 1, i} || P_i)
\end{aligned}
$$

$$+\sum_i r_{B_C+1,i} D(Q_{B_C+1,i}||P_i)$$

(by definition)

$$= \sum_i \left(\sum_{a\in[B_C-K]} r^*_{a,i} D(Q'_{a,i}||P_i) + r^*_{B_C-K+1,i} D(Q'_{B_C-K+1,i}||P_i) \right) + \sum_i D(Q'_{B_C-K+2,i}||P_i)$$

(by Jensen's inequality and $Q'_{B_C-K+2,i} = r_{B_C+1,i} Q_{B_C+1,i} + (1 - r_{B_C+1,i}) Q'_{B_C-K+1,i}$)

$$= \sum_i \sum_{a\in[B_C-K+1]} r^*_{a,i} D(Q'_{a,i}||P_i) + \sum_i D(Q'_{B_C-K+2,i}||P_i).$$

□

F.4 Proof of Lemma 4

Proof of Lemma 4.

$$\begin{aligned}
\sum_{i,b\in[B+K-1]} r_{b,i} D(Q_{b,i}||P_i) &\geq \sum_{i,b\in[B-1]} r^*_{b,i} D(Q'_{b,i}||P_i) + \sum_i D(Q'_{B,i}||P_i). \qquad \text{(by (15))} \\
&\geq \sum_{i,b\in[B]} r^*_{b,i} D(Q'_{b,i}||P_i) \\
&\geq \frac{B(R^{\mathrm{go}}_B - \epsilon)}{H(\boldsymbol{P})} \qquad \text{(by definition of } \epsilon\text{-optimal solution).}
\end{aligned}$$

□

F.5 Proof of Theorem 5

Proof of Theorem 5, Bernoulli rewards. Since the reward is binary, the possible values that $Q_{b,i}$ take lies in a finite set

$$\mathcal{V} = \left\{ \frac{l}{m} : l \in \mathbb{N}, m \in \mathbb{N}^+ \right\},$$

where it is easy to prove $|\mathcal{V}| \leq (T/(B+K-1)+2)^2 \leq (T/B+2)^2$. We have

$$\begin{aligned}
\mathbb{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})] &= \sum_{\boldsymbol{V}_1,\ldots,\boldsymbol{V}_B \in \mathcal{V}^K} \mathbb{P}\left[J(T) \notin \mathcal{I}^*(\boldsymbol{P}), \bigcap_b \{\boldsymbol{Q}_b = \boldsymbol{V}_b\} \right] \\
&= \sum_{\boldsymbol{V}_1,\ldots,\boldsymbol{V}_B \in \mathcal{V}^K : J^*(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_B) \notin \mathcal{I}^*(\boldsymbol{P})} \mathbb{P}\left[\bigcap_b \{\boldsymbol{Q}_b = \boldsymbol{V}_b\} \right].
\end{aligned}$$

By using the Chernoff bound, we have

$$\mathbb{P}\left[Q_{b,i} = V_{b,i} \middle| \bigcap_{b'\in[b-1]} \{\boldsymbol{Q}_{b'} = \boldsymbol{V}_{b'}\} \right] \leq e^{-\frac{T'}{B+K-1} r_{b,i} D(V_{b,i}||P_i)}, \tag{16}$$

and thus

$$\begin{aligned}
&\mathbb{P}\left[\bigcap_b \{\boldsymbol{Q}_b = \boldsymbol{V}_b\} \right] \\
&= \prod_b \mathbb{P}\left[\boldsymbol{Q}_b = \boldsymbol{V}_b \middle| \bigcap_{b'=1}^{b-1} \{\boldsymbol{Q}_{b'} = \boldsymbol{V}_{b'}\} \right]
\end{aligned}$$

$$\leq \prod_b e^{-\frac{T'}{B+K-1}\sum_i r_{b,i} D(V_{b,i}||P_i)} \quad \text{(by Eq. (16))}$$
$$= e^{-\frac{T'}{B+K-1}\sum_{b,i} r_{b,i} D(V_{b,i}||P_i)}. \tag{17}$$

Furthermore,

$$\begin{aligned}
&\mathbb{P}\left[\bigcap_b \{\boldsymbol{Q}_b = \boldsymbol{V}_b\}\right] \\
&= \mathbb{P}\left[\bigcap_b \{\boldsymbol{Q}_b = \boldsymbol{V}_b\}, \sum_{i,b\in[B+K-1]} r_{b,i} D(Q_{b,i}||P_i) \geq \frac{B(R_B^{\mathrm{go}} - \epsilon)}{H(\boldsymbol{P})}\right] \\
&\quad \text{(by Lemma 4).} \\
&= \mathbb{P}\left[\bigcap_b \{\boldsymbol{Q}_b = \boldsymbol{V}_b\}\right] \mathbb{P}\left[\sum_{i,b\in[B+K-1]} r_{b,i} D(Q_{b,i}||P_i) \geq \frac{B(R_B^{\mathrm{go}} - \epsilon)}{H(\boldsymbol{P})} \,\middle|\, \bigcap_b \{\boldsymbol{Q}_b = \boldsymbol{V}_b\}\right] \\
&= \mathbb{P}\left[\bigcap_b \{\boldsymbol{Q}_b = \boldsymbol{V}_b\}\right] \mathbb{P}\left[\sum_{i,b\in[B+K-1]} r_{b,i} D(V_{b,i}||P_i) \geq \frac{B(R_B^{\mathrm{go}} - \epsilon)}{H(\boldsymbol{P})}\right] \\
&= \mathbb{P}\left[\bigcap_b \{\boldsymbol{Q}_b = \boldsymbol{V}_b\}\right] \mathbb{E}\left[\mathbf{1}\left[\sum_{i,b\in[B+K-1]} r_{b,i} D(V_{b,i}||P_i) \geq \frac{B(R_B^{\mathrm{go}} - \epsilon)}{H(\boldsymbol{P})}\right]\right] \\
&\leq e^{-\frac{T'}{B+K-1}\sum_{b,i} r_{b,i} D(V_{b,i}||P_i)} \mathbb{E}\left[\mathbf{1}\left[\sum_{i,b\in[B+K-1]} r_{b,i} D(V_{b,i}||P_i) \geq \frac{B(R_B^{\mathrm{go}} - \epsilon)}{H(\boldsymbol{P})}\right]\right] \\
&\quad \text{(by Eq. (17))} \\
&= \mathbb{E}\left[e^{-\frac{T'}{B+K-1}\sum_{b,i} r_{b,i} D(V_{b,i}||P_i)} \mathbf{1}\left[\sum_{i,b\in[B+K-1]} r_{b,i} D(V_{b,i}||P_i) \geq \frac{B(R_B^{\mathrm{go}} - \epsilon)}{H(\boldsymbol{P})}\right]\right] \\
&\leq \mathbb{E}\left[e^{-\frac{T'}{B+K-1}\frac{B(R_B^{\mathrm{go}} - \epsilon)}{H(\boldsymbol{P})}}\right] \\
&= e^{-\frac{T'}{B+K-1}\frac{B(R_B^{\mathrm{go}} - \epsilon)}{H(\boldsymbol{P})}}.
\end{aligned} \tag{18}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})] \\
&\leq \sum_{\boldsymbol{V}_1,\ldots,\boldsymbol{V}_B \in \mathcal{V}^K} e^{-\frac{B}{B+K-1}\frac{(R_B^{\mathrm{go}} - \epsilon)T'}{H(\boldsymbol{P})}} \\
&\quad \text{(by Eq. (18))} \\
&\leq (T/B+2)^{2KB} e^{-\frac{B}{B+K-1}\frac{(R_B^{\mathrm{go}} - \epsilon)T'}{H(\boldsymbol{P})}}.
\end{aligned}$$

Here, $\log((T/B+2)^{2KB}) = o(T)$ to T when we consider K, B as constants.

□

Proof of Theorem 5, Normal rewards. Let

$$\mathcal{B} = \bigcup_{i,b} \{|Q_{b,i}| \geq T\}.$$

Then, it is easy to see

$$\mathbb{P}[\mathcal{B}] = T^{2KB} O(e^{-T^2/2}),$$

which is negligible because $\log(1/\mathbb{P}[\mathcal{B}])/T$ diverges.

The PoE is bounded as

$$\mathbb{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})] = \mathbb{P}\left[J(T) \notin \mathcal{I}^*(\boldsymbol{P}), \mathcal{B}^c\right] + \mathbb{P}[\mathcal{B}]$$

We have,

$$\begin{aligned}
&\mathbb{P}\left[J(T) \notin \mathcal{I}^*(\boldsymbol{P}), \mathcal{B}^c\right] \\
&= \int_{-T}^{T} \cdots \int_{-T}^{T} \mathbf{1}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})] p(\boldsymbol{Q}_B | \boldsymbol{Q}_{B-1} \dots \boldsymbol{Q}_1) d\boldsymbol{Q}_B \dots p(\boldsymbol{Q}_B | \boldsymbol{Q}_{B-1} \dots \boldsymbol{Q}_1) d\boldsymbol{Q}_b \dots p(\boldsymbol{Q}_1) d\boldsymbol{Q}_1.
\end{aligned} \tag{19}$$

Here,

$$\begin{aligned}
p(\boldsymbol{Q}_B | \boldsymbol{Q}_{B-1} \dots \boldsymbol{Q}_1) &= \prod_{i \in [K]} \frac{n_{b,i}}{\sqrt{2\pi}} \exp\left(-\frac{n_{b,i}(Q_{b,i} - P_i)^2}{2} \right) \\
&= \prod_{i \in [K]} \frac{n_{b,i}}{\sqrt{2\pi}} \exp\left(-n_{b,i} D(Q_{b,i} || P_i) \right) \\
&\leq \prod_{i \in [K]} T \exp\left(-n_{b,i} D(Q_{b,i} || P_i) \right).
\end{aligned}$$

Finally, we have

$$\begin{aligned}
(19) &\leq T^{BK} \int_{-T}^{T} \cdots \int_{-T}^{T} \mathbf{1}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})] \prod_{b \in [B]} \prod_{i \in [K]} \exp\left(-n_{b,i} D(Q_{b,i} || P_i) \right) d\boldsymbol{Q}_B \dots d\boldsymbol{Q}_1 \\
&\leq T^{BK} \int_{-T}^{T} \cdots \int_{-T}^{T} \mathbf{1}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})] \prod_{b \in [B]} \prod_{i \in [K]} \exp\left(-\frac{T' r^{(b,i)}}{B+K-1} D(Q_{b,i} || P_i) \right) d\boldsymbol{Q}_B \dots d\boldsymbol{Q}_1 \\
&\leq T^{BK} \int_{-T}^{T} \cdots \int_{-T}^{T} \mathbf{1}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})] \exp\left(-\frac{B}{B+K-1} \frac{(R_B^{\mathrm{go}} - \epsilon) T'}{H(\boldsymbol{P})} \right) d\boldsymbol{Q}_B \dots d\boldsymbol{Q}_1 \quad \text{(by Lemma 4)} \\
&\leq T^{BK} \int_{-T}^{T} \cdots \int_{-T}^{T} \exp\left(-\frac{B}{B+K-1} \frac{(R_B^{\mathrm{go}} - \epsilon) T'}{H(\boldsymbol{P})} \right) d\boldsymbol{Q}_B \dots d\boldsymbol{Q}_1 \\
&\leq T^{BK} (2T)^{BK} \exp\left(-\frac{B}{B+K-1} \frac{(R_B^{\mathrm{go}} - \epsilon) T'}{H(\boldsymbol{P})} \right).
\end{aligned}$$

□

F.6 Proof of Theorem 7

Proof of Theorem 7. We first show that the limit

$$R_\infty^{\mathrm{go}} = \lim_{B \to \infty} R_B^{\mathrm{go}}$$

exists. Namely, for any $\eta > 0$ there exists $B_0 \in \mathbb{N}$ such that for any $B_1 > B_0$ we have

$$|R_{B_0}^{\mathrm{go}} - R_{B_1}^{\mathrm{go}}| \leq \eta.$$

Theorem 5 implies that Algorithm 2 with $B = B_0$ and $\epsilon = \eta/2$ satisfies[9]

$$\liminf_{T \to \infty} \frac{\log(1/\mathbb{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})])}{T} \geq \frac{B_0}{B_0 + K - 1} \frac{R_{B_0}^{\mathrm{go}} - \eta/2}{H(\boldsymbol{P})},$$

[9]Strictly speaking, Algorithm 2 depends on T, and we take sequence of the algorithm $(\pi_{\mathrm{DOT},T})_{T=1,2,\dots}$.

and thus

$$\inf H(\boldsymbol{P}) \liminf_{T\to\infty} \frac{\log(1/\mathbb{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})])}{T} \geq \frac{B_0}{B_0+K-1}\left(R^{\mathrm{go}}_{B_0} - \frac{\eta}{2}\right). \tag{20}$$

Moreover, Theorem 2 implies that any algorithm satisfies

$$\inf H(\boldsymbol{P}) \limsup_{T\to\infty} \frac{\log(1/\mathbb{P}[J(T) \notin \mathcal{I}^*(\boldsymbol{P})])}{T} \leq R^{\mathrm{go}}_{B_1}. \tag{21}$$

Combining Eq. (20) and Eq. (21), we have

$$\frac{B_0}{B_0+K-1}\left(R^{\mathrm{go}}_{B_0} - \eta/2\right) \leq R^{\mathrm{go}}_{B_1}$$

and thus

$$\begin{aligned}
R^{\mathrm{go}}_{B_0} &\leq R^{\mathrm{go}}_{B_1} + \frac{\eta}{2} + \frac{K-1}{B_0+K-1} R^{\mathrm{go}}_{B_0} \\
&\leq R^{\mathrm{go}}_{B_1} + \frac{\eta}{2} + \frac{K-1}{B_0+K-1} R^{\mathrm{go}} \quad \text{(by Corollary 3)} \\
&\leq R^{\mathrm{go}}_{B_1} + \frac{\eta}{2} + \frac{\eta}{2} \quad \text{(by } K \geq 2\text{, by taking } B_0 \geq 2KR^{\mathrm{go}}/\eta\text{)} \\
&\leq R^{\mathrm{go}}_{B_1} + \eta.
\end{aligned}$$

By swapping B_0, B_1, it is easy to show that

$$R^{\mathrm{go}}_{B_1} \leq R^{\mathrm{go}}_{B_0} + \eta,$$

and thus

$$|R^{\mathrm{go}}_{B_0} - R^{\mathrm{go}}_{B_1}| \leq \eta,$$

which implies that the limit exists. It is easy to confirm that the performance of Algorithm 2 with any $B \geq 2KR^{\mathrm{go}}_1/\eta$ and $\epsilon = \eta/2$ satisfies Eq. (6). □