# A Tight Lower Bound for Non-stochastic Multi-armed Bandits with Expert Advice

**Zachary Chase** ZACHMAN99323@GMAIL.COM
*Kent State University*

**Shinji Ito** SHINJI@MIST.I.U-TOKYO.AC.JP
*The University of Tokyo and RIKEN*

**Idan Mehalel** IDANMEHALEL@GMAIL.COM
*The Hebrew University*

**Editors:** Steve Hanneke and Tor Lattimore

## Abstract

We determine the minimax optimal expected regret in the classic *non-stochastic multi-armed bandit with expert advice* problem, by proving a lower bound that matches the upper bound of Kale (2014). The two bounds determine the minimax optimal expected regret to be $\Theta\left(\sqrt{TK\log\frac{N}{K}}\right)$, where $K$ is the number of arms, $N$ is the number of experts, and $T$ is the time horizon.

**Keywords:** Multi-armed bandits, Adversarial bandits, Online learning.

## 1. Introduction

The seminal work Auer et al. (2002) presented the fundamental *non-stochastic multi-armed bandit with expert advice* problem (BwE, for short), which is defined as follows. Let $T, N, K \in \mathbb{N}$ where $K$ is the number of arms, $N$ is the number of experts, and $T$ is the time horizon. On each round $t \in [T]$, each expert $j \in [N]$ selects an advice $e_t(j) \in [K]$ which is presented to the learner. Then, the learner pulls an arm $I_t \in [K]$ and the adversary simultaneously assigns a loss $\ell_t(r) \in \{0, 1\}$ for each arm $r \in [K]$. The learner observes only the loss it suffers, namely $\ell_t(I_t)$, and the other losses remain unrevealed. Generally, the expert advice and losses are allowed to be fractional, that is, $e_t(j)$ is a probability distribution over $[K]$, and $\ell_t(r) \in [0, 1]$. However, for the sake of proving the lower bound, the simpler discrete version of the problem suffices. Note that the adversary is *adaptive*: the assignment of expert advice and losses to arms in each round may (and will, in our construction) depend on the learner's past choices. The learner's goal is to minimize the expected *regret*: the difference between its cumulative loss and the cumulative loss of the best expert, which is the expert with minimal cumulative loss.

### 1.1. Previous results

Throughout the paper, we assume that $T \geq \Omega(K' \log(2N/K'))$, where $K' = \min\{K, N\}$; in the complementing case, the lower bounds proved in this work and in Auer et al. (2002) imply that the minimax optimal expected regret is $\Theta(T)$. We further assume that $N \geq K$; in the complementing case, it is known that the minimax optimal expected regret is $\Theta\left(\sqrt{TN}\right)$ Auer et al. (2002); Kale (2014).

The seminal work Auer et al. (2002) came up with the Exp4 algorithm, which has been used as a building block or inspiration for many partial feedback algorithms. A partial list of examples

includes McMahan and Streeter (2009); Seldin et al. (2011); Daniely and Helbertal (2013); Raman et al. (2024). The analysis of Auer et al. (2002) bounds the *pseudo-regret* (see Section 3 for a formal definition) of Exp4 by $O\left(\sqrt{TK\log N}\right)$. This upper bound remained the best known for this problem, until the work Kale (2014) which considered a generalization of the problem where the expert advice is only partially observed by the learner, which, as a by-product, improved the upper bound in the general case as well to $O\left(\sqrt{TK\log(N/K)}\right)$. Furthermore, this quantity upper bounds the optimal expected regret of the problem, not just the weaker pseudo-regret.

As for lower bounds, until the work Seldin and Lugosi (2016), the best known lower bound was $\Omega\left(\sqrt{T}\left(\sqrt{K}+\sqrt{\log N}\right)\right)$, where the left summand is from the lower bound for the multi-armed bandit (without experts) problem given in Auer et al. (2002), and the right summand is from the lower bound on the full-info prediction with expert advice problem, given e.g. in Cesa-Bianchi and Lugosi (2006). The work Seldin and Lugosi (2016) significantly improved the lower bound to $\Omega\left(\sqrt{TK\frac{\log N}{\log K}}\right)$, only off by a $\log K$ factor from the upper bound. They conjectured that this lower bound is tight, which is refuted in this work.

The works Mannor and Tsitsiklis (2004); Chen et al. (2024) identified a useful connection between minimizing regret problems (such as BwE), and identification of best arm/expert problems. Such identification problems are often referred to as the *pure-exploration* variations of the regret minimization problems Even-Dar et al. (2002); Mannor and Tsitsiklis (2004). Intuitively, the idea is that in settings where a single entity (arm/expert) is significantly better than all other entities throughout the game, a learner who minimizes regret must identify this entity in an early stage of the game. This idea allows us to consider identification problems instead of regret minimization problems. Armed with this approach, and inspired by a problem instance presented in Chen et al. (2024), the two recent works Ito (2024); Cesa-Bianchi et al. (2025) have managed to prove a tight lower bound of $\Omega\left(\sqrt{TK\log(N/K)}\right)$ against a restricted learner, having the property that its choice in round $t$ is independent of the expert advice of round $t$ (the result of Cesa-Bianchi et al. (2025) is somewhat weaker, as they consider a slightly even more restricted learner). This setting is sometimes referred to as *proper* online learning [1] Hanneke et al. (2021), while the general setting is known as *improper* learning. The connection between regret minimization and best arm identification was also used in the book of Slivkins Slivkins (2019) to prove the known $\Omega\left(\sqrt{KT}\right)$ lower bound for multi-armed bandits (without experts).

### 1.2. Our contribution

Building on the ideas and the hard problem instance used in Ito (2024), we manage to prove the same $\Omega\left(\sqrt{TK\log(N/K)}\right)$ lower bound for the classic problem, without placing any restrictions on the learner. The formal statement of the bound appears in Theorem 5. A summary of the bounds proved for this problem may be found in Table 1. In the following subsection, we describe the proof sketch.

1. In Hanneke et al. (2021), a learner is called proper if, before observing the current instance, it must output a hypothesis from a reference class. In the BwE setting, the analogous restriction is that the learner must effectively commit on an expert before seeing the current-round advice.

| Setup | Reference | Bound |
|---|---|---|
| Standard (pr) | Auer et al. (2002) | $O\left(\sqrt{TK\log N}\right)$ |
| Standard | Kale (2014) | $O\left(\sqrt{TK\log\frac{N}{K}}\right)$ |
| Standard | Seldin and Lugosi (2016) | $\Omega\left(\sqrt{TK\frac{\log N}{\log K}}\right)$ |
| Proper* | Cesa-Bianchi et al. (2025) | $\Omega\left(\sqrt{TK\log\frac{N}{K}}\right)$ |
| Proper | Ito (2024) | $\Omega\left(\sqrt{TK\log\frac{N}{K}}\right)$ |
| Standard | **[This work]** | $\Omega\left(\sqrt{TK\log\frac{N}{K}}\right)$ |

Table 1: Upper ($O(\cdot)$) and lower ($\Omega(\cdot)$) bounds on the expected regret for the multi-armed bandit with expert advice (BwE) problem. The *Standard (pr)* setup refers to the standard setup considered in this paper, but where the bounded quantity is the weaker pseudo-regret, and not the expected regret. The *proper* setup refers to a restricted learner whose choice of arm in round $t$ does not depend on the advice of round $t$. The *proper** setting refers to the setting considered in Cesa-Bianchi et al. (2025), in which the learner is slightly more restricted than in the proper setting.

### 1.3. Proof Sketch

Our proof proceeds in four steps, described in high-level in the following subsections. Many of the ideas we build on were discovered by Ito (2024), but there are a few key differences between the proofs, which we discuss in Section 1.4.

#### 1.3.1. STEP 1: REDUCTION FROM *Special Batch Identification* (SBI) TO BWE

The first idea used in the proof, originated from Chen et al. (2024) and further used in Ito (2024) is to reduce some sort of identification problem to BwE. We call the specific identification problem we use *Special Batch Identification* or SBI, for short. This problem is well defined in the setting we consider, in which we partition the $N$ experts to roughly[2] $k = K/2$ many batches, where each batch contains roughly $n = N/k$ experts. Every batch $u \in [k]$ is a distinct "world" in the sense that the experts in batch $u \in [k]$ can only advise one of two arms associated with their batch: $(u, 0)$ or $(u, 1)$. Furthermore, in every round, precisely one arm of each batch is "correct" (has loss 0) and the other is incorrect (has loss 1). We stress that the correct label of every batch is freshly chosen in each round. This creates a situation of $k$ many distinct binary "worlds". The adversary chooses in advance precisely one of those batches to be "special". The special batch behaves slightly different from the other batches. In the standard batches, all experts have probability $1/2$ to choose the correct label in every round. The special batch, however, contains a special expert $e^\star$, for which the adversary makes sure that in every round, the arm advised by $e^\star$ is chosen to be correct with probability $1/2 + \epsilon$,

2. In the formal proof we use one more "dummy" batch, which we add for technical reasons.

where $\epsilon$ is some small parameter. The point here is that a learner who wishes to minimize regret essentially competes with this special expert. Therefore, it must identify the special expert with high probability, and then repeatedly copy its advice. We show that the bottleneck of this process is the identification of the special batch. This gives the reduction from SBI to BwE in our setting.

#### 1.3.2. STEP 2: THE ONE-BATCH SBI GAME

Intuitively, since every batch functions as a separate "world", in order to find the special batch, the learner has no choice but to go through the batches one by one and query them for some time, until it finds the special batch. Therefore, the learner essentially must solve $\Omega(k)$ one-batch SBI instances in expectation, until it finds the special batch (In a one-batch SBI game, the learner should decide for a given batch, if it is special or not).

Since in every batch, and in every round, precisely one arm is correct, the special case of a single batch is a full-information problem: the learner receives all relevant information in every round, no matter which of the two arms it picks. This makes the problem reasonable to analyze: we need to compare the distributions on the input generated by the adversary when the batch is special, and when it is not special. This was done by Ito (2024), by bounding the KL-divergence between the two distributions. We use the same bound in our proof as well. This bound shows that as long as the learner sees the information produced by the adversary for fewer than $c\frac{\log n}{\epsilon^2}$ many rounds (for some constant $c$), the cases where the batch is special or not are essentially indistinguishable.

#### 1.3.3. STEP 3: FROM THE ONE-BATCH SBI GAME TO THE GENERAL SBI GAME

As mentioned before, the general idea is to show that the adversary can force the learner to solve $\Omega(k)$ many one-batch instances until it identifies the special batch. Having that in hand, the learner indeed must run for $\Omega\left(k\frac{\log n}{\epsilon^2}\right)$ many rounds in order to identify the special batch with high probability. However, there is a subtle issue here that needs to be addressed. Since we study the case where the learner is allowed to devise its prediction only after seeing the current round's advice, we cannot immediately deduce that the learner cannot use this information to choose in each round a specific batch for which the expert advice is most informative. The role of our adaptive adversary is to handle this exact issue: it draws fresh advice for some batch only after rounds where the learner has queried it.

#### 1.3.4. STEP 4: FROM THE GENERAL SBI GAME LOWER BOUND TO A BWE LOWER BOUND

Fix $K, N, T$. We assume that for the setting we have defined in Section 1.3.1, there exists a learner $A$ with expected regret at most $O(\epsilon T)$. By Step 1, this intuitively means that $A$ must have (with high probability) identified the special batch at some round $T' < T$, and it is imitating the predictions of the special expert since then. However, from Step 3, we know that to identify the special batch, $A$ must run for at least $T' = \Omega\left(k\frac{\log n}{\epsilon^2}\right)$ many rounds. Therefore, for small enough $\epsilon$, we get $T' > T$, which means that the special batch is not identified, and every learner has expected regret at least $\Omega(\epsilon T)$. Choosing $\epsilon = \Theta\left(\sqrt{\frac{k\log n}{T}}\right)$ is small enough to get $T' > T$, and the lower bound $\Omega(\epsilon T) = \Omega\left(\sqrt{TK\log\frac{N}{K}}\right)$ is implied.

### 1.4. Key differences from the proof of Ito (2024)

We prove the same lower bound that Ito (2024) proved against a restricted learner whose predictions in round $t$ are independent of the expert advice of round $t$ (that is, against a proper learner). While our proof is based on the ideas of Ito (2024), there are a few key differences, allowing us to prove the same lower bound against a non-restricted learner. Below, we briefly describe the main differences between our proof and the proof of Ito (2024).

1. To make Step 3 possible, we have to make sure that the $k$ "one-batch" problems are "identical but independent" even when the learner is improper (not restricted as in Ito (2024)). This forces the learner to solve from scratch at least $\Omega(k)$ one-batch problems in order to identify the special batch. To implement it, we adapt the expert advice to the learner's choices of arms. This is where we use the adaptiveness of the adversary, which is not used in Ito (2024). As explained in Section 1.3.3, the idea is to draw fresh advice for some batch only after rounds where this batch was queried. The formal implementation of the adversary is described in Section 2.

2. In Ito (2024), the reduction to BwE (Step 1) is proved from the more standard problem of *Best Expert Identification* (BEI), in which a specific special expert needs to be identified, not only the batch containing it, as in SBI. However, the adaptive adversary we have defined to make Step 3 feasible, makes such a reduction infeasible, and therefore we prove the reduction from the seemingly easier SBI problem, and for an improper learner. Nevertheless, we observe that SBI is still difficult enough, and the lower bound proved in Ito (2024) for BEI holds for SBI as well.

### 1.5. Rest of the paper organization

The rest of the paper (except for Section 7 which discusses future work), is dedicated to the formal proof. To avoid any confusion, we stress here that in the formal proof, we add another "dummy batch", which is not mentioned in the proof sketch. Therefore, in the formal proof, the "one-batch" instance of the problem mentioned in the above sketch is referred to as the "two-batch" instance (including the dummy batch). The exact construction is provided in the following section.

## 2. A reduced setting of BwE (Step 0)

To prove the lower bound, we consider only the following set of instances of the problem. We assume that $k = (K-1)/2 \in \mathbb{N}$, and that $n = (N-1)/k \in \mathbb{N}$. Note that for the lower bound that we prove, this assumption is without loss of generality: for other values of $N, K$, we may simply use only the maximal number of arms $K'$ and experts $N'$ that fits the above requirements, and let the other arms and experts always suffer loss 1. This will result in the same bound as if $N, K$ fit the above requirements, up to a multiplicative constant.

Under the assumption above, we may partition $N-1$ of the experts to $k$ disjoint batches, each of size $n$. The experts in the $u$'th batch (where $u \in [k]$) will only set their advice to one of two arms: $(u,0)$ or $(u,1)$. Each of the $N-1$ experts in the $k$ batches is identified by a pair $(u,v)$ where $u \in [k]$ identifies the batch, and $v \in [n]$ identifies the index of the expert inside the batch. Note that we have one unused expert and one unused arm. We identify this expert with 0 and this arm also with 0. Expert 0 can only set its advice to 0.

Within this reduced setting, we only consider the following possible strategies of the adaptive adversary.

**Expert advice assignment.** We denote the advice in round $t$ (of all experts other than 0) by a vector $A_t = (a_t^{(1)}, \ldots, a_t^{(k)})$, where $a_t^{(u)}$ is a binary vector of length $n$ setting the advice of batch $u$ in the natural way: the $v$'th index of $a_t^{(u)}$, denoted $a_t^{(u,v)}$ sets the right value of $e_t(u,v)$. In all the adversarial strategies that we consider, the expert advice $A_t$ of each round $t$ is chosen as follows.

1. Initialize $A_t$ by choosing every entry iid from $\mathrm{Ber}(1/2)$.
2. For round $t = 1, 2, \ldots$:
   (a) Use the advice $A_t$ for the course of this round.
   (b) Upon the choice of $I_t$ by the learner, for each $u \in [k]$:
      i. If $I_t \in \{(u,0),(u,1)\}$, draw a binary vector $b$ of length $n$, each entry iid from $\mathrm{Ber}(1/2)$, and set $a_{t+1}^{(u)} = b$.
      ii. Otherwise, set $a_{t+1}^{(u)} = a_t^{(u)}$.

In simple words, the advice of every batch $u \in [k]$ is initialized by a vector of entries drawn iid from $\mathrm{Ber}(1/2)$. Only after rounds where an arm of batch $u$ is pulled, new expert advice is picked for this batch $u$, again by a vector of entries drawn iid from $\mathrm{Ber}(1/2)$. After rounds in which an arm of $u$ is not pulled, the same advice is reused for the next round.

**Loss assignments.** For any fixed $0 < \epsilon \leq 0.1$, we define $N$ different randomized loss assignment strategies that depend on $\epsilon$. Formally, a strategy is a randomized algorithm that the adversary uses to draw the expert advice and the loss of each arm. In the reduced setting that we consider, the adversary may choose precisely one of those strategies to be used throughout the entire game. The learner does not know which strategy is chosen by the adversary. If an adversary follows strategy $S$, it is called an $S$-adversary. Fix $0 < \epsilon \leq 0.1$. In all strategies we define, the loss of all arms is always taken from $\{0,1\}$. Furthermore, for any batch $u \in [k]$ and any round $t$, $\ell_t(r)$ equals 0 for precisely one $r \in \{(u,0),(u,1)\}$. We refer to this arm as the *correct arm* in batch $u$ and round $t$. Let us define the strategies. First, any of the strategies we define draws $\ell_t(0) \sim \mathrm{Ber}(1/2 - \epsilon/2)$. We describe the differences between the strategies below.

- The first strategy is called $S_0$. An $S_0$-adversary draws $\ell_t(u,0) \sim \mathrm{Ber}(1/2)$ for all $u \in [k], t \in [T]$. Note that this determines $\ell_t(u,1)$ for all $u \in [k]$ and all $t \in [T]$.
- The other $N-1$ strategies are $\{S_{(u,v)}\}_{u\in[k],v\in[n]}$. An $S_{(u^\star,v^\star)}$-adversary draws $\ell_t(u,0) \sim \mathrm{Ber}(1/2)$ for all $t$, and for all $u \in [k]\backslash\{u^\star\}$. It draws $\ell_t(u^\star, a_t^{(u^\star,v^\star)}) \sim \mathrm{Ber}(1/2 - \epsilon)$.

In simple words, in $S_0$ all experts are equally handled, and $S_{(u^\star,v^\star)}$ gives a slight advantage to the predictions of $(u^\star, v^\star)$.

## 3. A reduction from *Special Batch Identification* (SBI) to BwE (Step 1)

We say that $u \in [k]$ is the *special batch* if there exists $v \in [n]$ such that the adversarial strategy is $S_{(u,v)}$. If there is no such $u$, then 0 is the special batch. Note that in our reduced setting, there is always precisely one special batch.

In this paper, we lower bound the *pseudo-regret* of the learner (which can only be smaller than the expected regret):

$$R_T = \max_{j^\star \in [N]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t) - \sum_{t=1}^{T} \ell_t(e_t(j^\star))\right].$$

Further discussion on standard measures of regret in bandit problems may be found in Bubeck and Cesa-Bianchi (2012). Note that from linearity of expectation, the pseudo-regret against any fixed expert $j \in [N]$ can be seen as a sum of $T$ *per-round* pseudo-regret values, the $t^{th}$ of which is $\mathbb{E}[\ell_t(I_t) - \ell_t(e_t(j))]$.

In this section, we prove that in our reduced setting, any algorithm achieving pseudo-regret at most $\epsilon T/1000$, must identify the special batch with high probability. Note that this rather is intuitive: a learner who does not know which batch is the special one, and thus keeps pulling arms not from the special batch in at least, say, half of the rounds, will suffer at least $\epsilon/2$ per-round pseudo-regret against the best expert in at least half of the rounds, resulting in at least $\epsilon T/4$ cumulative pseudo-regret.

We call this problem of identifying the special batch *Special Batch Identification (SBI)*. Formally, an algorithm $A'$ for SBI is an algorithm which is being executed in the same setting as BwE. The difference is that the goal of $A'$ is different from the goal of an algorithm for BwE. The goal of $A'$ is to identify the special batch with high certainty, after running for the least possible number of rounds. That is, in SBI, the algorithm may stop the game at any point, and it is required to output a prediction of the special batch's identity that is correct with high probability. In particular, it does not make any predictions or suffer any loss; it only chooses an arm to pull in each round, in a way that maximizes output accuracy, in the least possible number of rounds. Therefore, note that:

1. Expert $0$ and arm $0$ behave the same under all strategies, and thus its per-round expected loss is known to always be $1/2 - \epsilon/2$. That is, pulling $0$ does not provide any information required to estimate if $0$ is the special batch or not. We can thus assume w.l.o.g that $A'$ never pulls $0$.

2. We can say that $A'$ "pulls a batch" $u$ instead of a specific arm $(u, 0)$ or $(u, 1)$, as pulling $(u, 0)$ or $(u, 1)$ gives $A'$ the exact same information. Indeed, $\ell_t(u, 0) = 1 - \ell_t(u, 1)$.

We say that $A'$ is *good* if for every possible strategy from our pool, $A'$ outputs the correct special batch with probability at least $0.95$. We denote by $T(A', S)$ the expected number of rounds $A'$ is executed before outputting its estimate, when the adversarial strategy is $S$. The expectation is taken over the randomness of $A'$, as well as the randomness of $S$.

**Lemma 1** *Suppose that $A$ is an algorithm for BwE, such that for any $S$ from our pool of strategies, the pseudo-regret of $A$ is bounded by $R_T \leq r(T)$ for all $T$. Let $T^\star \geq 1000r(T^\star)/\epsilon$. Then, there exists a good algorithm $A'$ for SBI such that for any strategy $S$ from our pool, $T(A', S) \leq T^\star$.*

**Proof** Let us define a good algorithm $A'$ with $T(A', S) \leq T^\star$ for all $S$. $A'$ executes $A$ for $T^\star$ many rounds. For any $u \in [k] \cup \{0\}$, let $T_u$ be the number of rounds where $A$ pulls an arm of batch $u$. $A'$ outputs $\arg\max_{u \in [k] \cup \{0\}} T_u$. Clearly, $T(A', S) \leq T^\star$ for all $S$. It remains to show that $A'$ is good. Let $u^\star$ be the special batch. We show that $T_{u^\star}$ is likely to be very large compared to other $T_u$'s:

$$\begin{aligned} \Pr[T_{u^\star} < 0.75T^\star] &= \Pr[T^\star - T_{u^\star} \geq 0.25T^\star] \\ &\leq \frac{4\mathbb{E}[T^\star - T_{u^\star}]}{T^\star} \qquad \text{(Markov)} \end{aligned}$$

$$\begin{aligned}
&\leq \frac{4}{T^\star} \cdot \frac{2R_{T^\star}}{\epsilon} && (R_{T^\star} \geq \tfrac{\epsilon}{2}\mathbb{E}[T^\star - T_{u^\star}]) \\
&\leq \frac{4}{T^\star} \cdot \frac{2r(T^\star)}{\epsilon} \\
&\leq \frac{4}{T^\star} \cdot \frac{2\frac{T^\star\epsilon}{1000}}{\epsilon} \\
&\leq 0.01.
\end{aligned}$$

To see that $R_{T^\star} \geq \frac{\epsilon}{2}\mathbb{E}[T^\star - T_{u^\star}]$, note that if the special batch is 0, then the per-round pseudo-regret of any arm other than 0 against expert 0 is exactly $\epsilon/2$. If the special batch is some $u \in [k]$, then when pulling an arm of some other $u' \in [k]$ the per-round pseudo-regret against the special expert is exactly $\epsilon$, and when pulling 0, the per-round pseudo-regret against the special expert is exactly $\epsilon/2$.

Note that $T_{u^\star} \geq 0.75T^\star$ implies $u^\star = \arg\max_{u \in [k]\cup\{0\}} T_u$. Therefore, $A'$ outputs $u^\star$ with probability at least 0.99, as required. ■

## 4. Lower bound for the two-batch game (Step 2)

In this section we consider the case $K = 3$, in which $k = 1$, and we only have the batch containing only 0, and another single batch. We prove a lower bound on $T(A', S_0)$, for any good algorithm $A'$ for SBI. First, observe that in the case $k = 1$, the learner is fixed to always choose batch 1. Therefore, for any adversarial strategy, the learner is completely passive throughout the game. Its only choice is for how many rounds to play, and which batch, 0 or 1, to output at the end of the game. Therefore, for any fixed strategy, the adversary draws the expert advice and arm losses from the same distribution in all rounds. For a strategy $S_\star$, let $P_\star$ be the distribution from which the adversary pulls the advice and losses in every round when the strategy is $S_\star$. For any distribution $P$, let $P^T$ be $T$ iid draws from $P$. So, if the adversary uses strategy $S_\star$ and it is given that the learner runs for $T$ many rounds, then the distribution over all expert advice and losses throughout the game is $P_\star^T$. Since the learner depends only on the observed advice and losses, $P_\star^T$ induces a distribution on the learner's choice of special batch, when running for $T$ many rounds against an $S_\star$-adversary. Let $S_{mix^T}$ be the adversarial strategy where the adversary draws in advance $v$ from $[n]$ uniformly at random, and then uses the strategy $S_{(1,v)}$ for $T$ many rounds. Likewise, denote $P_{mix^T} = \frac{1}{n}\sum_{v\in[n]} P_{(1,v)}^T$. In all notation defined above, when $T$ is replaced with $\infty$, this is interpreted as a product of unbounded length. We denote the KL-divergence between two distributions $P, Q$ by $D_{\mathrm{KL}}(P||Q)$.

We begin with the following lemma, which was proved in Ito (2024) for equivalent distributions that were defined slightly differently therein. For completeness, we provide the proof below.

**Lemma 2 (Ito (2024))** *If $0 < \epsilon \leq 0.1$, then for any $T \geq 1, n \geq 1$:*

$$D_{\mathrm{KL}}(P_{mix^T}||P_0^T) \leq \frac{(1+4\epsilon^2)^T - 1}{n}.$$

**Proof** By definitions of $P_{mix^T}, P_0^T$, we have

$$D_{\mathrm{KL}}(P_{mix^T}||P_0^T) = D_{\mathrm{KL}}\left(\frac{1}{n}\sum_{v\in[n]} P_{(1,v)}^T||P_0^T\right)$$

Let $p_v$ and $p_0$ be the probability mass functions for $P^T_{(1,v)}$ and $P^T_0$, respectively. By definition of the KL-divergence and linearity of expectation, the above is equal to

$$\frac{1}{n}\sum_{v^\star\in[n]}\mathbb{E}_{g\sim P^T_{(1,v^\star)}}\left[\ln\left(\frac{1}{n}\sum_{v\in[n]}\frac{p_v(g)}{p_0(g)}\right)\right]\le\frac{1}{n}\sum_{v^\star\in[n]}\ln\left(\frac{1}{n}\sum_{v\in[n]}\mathbb{E}_{g\sim P^T_{(1,v^\star)}}\left[\frac{p_v(g)}{p_0(g)}\right]\right), \quad (1)$$

where the inequality is due to Jensen's inequality combined with concavity of $\ln$, and linearity of expectation. It remains to calculate the expectation $\mathbb{E}_{g\sim P^T_{(1,v^\star)}}\left[\frac{p_v(g)}{p_0(g)}\right]$ for every $v, v^\star$. Fix $g$, and for every round $t$, let $c_t\in\{0,1\}$ such that in round $t$, the correct arm is $(1,c_t)$. For every round $t$ and $v\in[n]$, define $c_{t,v}=1[e_t(v)=(1,c_t)]$. Now, note that:

$$\frac{p_v(g)}{p_0(g)}=\prod_{t\in[T]}\frac{(1/2)^n(1/2+(2c_{t,v}-1)\epsilon)}{(1/2)^{n+1}}=\prod_{t\in[T]}(1+(2c_{t,v}-1)2\epsilon).$$

Therefore, if $v\neq v^\star$ then:

$$\mathbb{E}_{g\sim P^T_{(1,v^\star)}}\left[\frac{p_v(g)}{p_0(g)}\right]=\prod_{t\in[T]}\mathbb{E}_{c_{t,v}\sim\mathrm{Ber}(1/2)}[1+(2c_{t,v}-1)2\epsilon]=1.$$

Otherwise, if $v=v^\star$ then:

$$\mathbb{E}_{g\sim P^T_{(1,v^\star)}}\left[\frac{p_v(g)}{p_0(g)}\right]=\prod_{t\in[T]}\mathbb{E}_{c_{t,v}\sim\mathrm{Ber}(1/2+\epsilon)}[1+(2c_{t,v}-1)2\epsilon]=(1+4\epsilon^2)^T.$$

Plugging this into the RHS of (1) which upper bounds $D_{\mathrm{KL}}(P_{mix^T}||P^T_0)$, we obtain:

$$\begin{aligned}D_{\mathrm{KL}}(P_{mix^T}||P^T_0)&\le\frac{1}{n}\sum_{v^\star\in[n]}\ln\left(\frac{n-1+(1+4\epsilon^2)^T}{n}\right)\\&=\ln\left(1+\frac{(1+4\epsilon^2)^T-1}{n}\right)\\&\le\frac{(1+4\epsilon^2)^T-1}{n},\end{aligned}$$

as claimed. ■

We now prove a lower bound on $T(A',S_0)$ for the two-batch case ($k=1$) that holds under the assumption that $A'$ is good for any strategy $S$ from our pool of strategies.

**Lemma 3** *Suppose that $k=1$, $0<\epsilon\le 0.1$, and $n\ge 10$. Let $A'$ be good. Then $T(A',S_0)\ge T^\star/2:=\frac{\ln(n/10)}{8\epsilon^2}$.*

**Proof** Let $T$ be a random variable counting the rounds for which $A'$ is being executed. We will show that $P^\infty_0[T>T^\star]\ge 1/2$, which implies the claim. Let $\hat{J}\in\{0,1\}$ be the output of $A'$. Let $E$ be the

event where $A'$ stops after $T \leq T^\star$ many rounds and outputs $\hat{J} = 0$. By Pinsker's inequality, we have that:

$$\begin{aligned}\left|P_0^{T^\star}[E] - P_{mix^{T^\star}}[E]\right| &\leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_{mix^{T^\star}} \| P_0^{T^\star})} \\ &\leq \sqrt{\frac{1}{2}\frac{(1+4\epsilon^2)^{T^\star} - 1}{n}} \qquad \text{(Lemma 2)} \\ &\leq \sqrt{\frac{1}{2}\frac{e^{4\epsilon^2 T^\star} - 1}{n}} \\ &\leq 1/4.\end{aligned}$$

Note that for $P_\star$ induced by $S_\star$ from our pool, we have $P_\star^\infty(A) = P_\star^{T^\star}(A)$ for any event $A$ contained in the event $T \leq T^\star$. Indeed, determining whether $A$ occurs or not can be done after $T^\star$ draws from $P_\star$. Since $A'$ is good, we also have

$$P_{mix^{T^\star}}[E] = P_{mix^\infty}[E] \leq 0.05.$$

Again since $A'$ is good, we have:

$$\begin{aligned}1 - P_0^{T^\star}[E] &= P_0^{T^\star}[T > T^\star] + P_0^{T^\star}[T \leq T^\star \wedge \hat{J} \neq 0] \\ &= P_0^{T^\star}[T > T^\star] + P_0^\infty[T \leq T^\star \wedge \hat{J} \neq 0] \\ &\leq P_0^{T^\star}[T > T^\star] + 0.05.\end{aligned}$$

Now, note that $P_0^\infty[T > T^\star] = P_0^{T^\star}[T > T^\star]$, as the occurrence of the event $T > T^\star$ is already determined after $T^\star$ many rounds. Combining this identity with the three inequalities above, we obtain:

$$\begin{aligned}P_0^\infty[T > T^\star] &= P_0^{T^\star}[T > T^\star] \\ &\geq 0.95 - P_0^{T^\star}[E] \\ &\geq 0.95 - 1/4 - P_{mix^{T^\star}}[E] \\ &\geq 0.95 - 1/4 - 0.05 \\ &= 0.65.\end{aligned}$$

This concludes the proof. ■

## 5. The general case (Step 3)

We now bootstrap the result to the general case, where there are $k + 1 \geq 2$ many batches.

**Lemma 4** *Let $A'$ be a good algorithm for SBI. Then $T(A', S_0) \geq k\frac{\ln(n/10)}{20\epsilon^2}$.*

**Proof** Let $T_u(A', S_0)$ be the expected number of rounds where $A'$ pulls $u \in [k]$ against an $S_0$-adversary. That is, $T(A', S_0) = \sum_{u \in [k]} T_u(A', S_0)$ (from linearity of expectation, and since we

assume that $A'$ never pulls 0). Suppose towards contradiction that $T(A', S_0) \leq k\frac{\ln(n/10)}{20\epsilon^2}$. Therefore, there exists $u$ so that $T_u(A', S_0) \leq \frac{\ln(n/10)}{20\epsilon^2}$. Based on the queries of $A'$ on batch $u$, we will construct a good algorithm $B'$ for SBI for the case $k = 1$, which we call the "small instance". The instance of the problem that $A'$ solves is called the "big instance". We stress that the adversary is unaware of $A'$, and produces input for the small instance, handled by $B'$. However, $B'$ will simulate input for $A'$, and use its decisions after adapting them to the small instance. The non-0 experts in the small instance will be treated as the experts of batch $u$ in the big instance. Expert 0 in the small instance remains expert 0 in the big instance. $B'$ operates as follows.

1. In rounds $t = 1, \ldots$
   (a) $B'$ receives advice for the small instance from the adversary.
   (b) $B'$ generates the advice of all batches of the big instance, except from $0, u$, according to the policy described in Section 2. It passes to $A'$ the advice of $0, u$ generated by the adversary, and also the self-generated advice for the other batches.
   (c) $B'$ checks what $A'$ wishes to do.
   (d) If $A'$ queries a batch $u' \neq u$:
      i. $B'$ draws the correct arm of $u'$ according to $\mathrm{Ber}(1/2)$ and passes the outcome to $A'$.
      ii. $B'$ generates advice for all batches of the big instance according to the policy described in Section 2, and passes it to $A'$.
      iii. Return to Item 1c.
   (e) Else, if $A'$ queries $u$: pass the correct arm received from the adversary to $A'$, as the correct arm for batch $u$.
   (f) If $A'$ stops, then:
      i. If $A'$ outputs $u$ or 0, $B'$ outputs the same output as $A'$ (where $u$ refers to the non-0 batch in the small instance).
      ii. Else, if $A'$ outputs some batch other than $u$ or 0, $B'$ outputs batch 0.

Since the adversary must choose a strategy from the pool for the small instance, the big problem instance generated for $A'$ is also from the pool, and note that the special batch in the small and big instances is the same batch: it is one of the two batches of the small instance. When $A'$ stops and outputs a batch, $B'$ stops as well. If $A'$ outputs a valid batch for the small instance ($u$ or 0), $B'$ outputs the same batch, and otherwise it outputs 0. Since $A'$ is good, it outputs the correct batch with probability at least 0.95, and therefore $B'$ also outputs the correct batch with probability at least 0.95. Thus, $B'$ is good for the small instance. However, we also know that $T_u(A', S_0) < \frac{\ln(n/10)}{20\epsilon^2}$, and $T_u(A', S_0)$ is the expected number of rounds that $B'$ is executed for against an $S_0$-adversary. This contradicts Lemma 3. Therefore, $T(A', S_0) \geq k\frac{\ln(n/10)}{20\epsilon^2}$. ■

## 6. Concluding the lower bound for BwE (Step 4)

**Theorem 5** *Let $N, K$ be positive natural numbers such that $N \geq 2K$. Then, for any $T \geq K\ln(N/K)$, and for any BwE algorithm, there exists an adversarial strategy for which $R_T = \Omega\left(\sqrt{TK\log(N/K)}\right)$.*

**Proof** Assume without loss of generality that $N, K$ match our reduced setting ($k := (K-1)/2 \in \mathbb{N}$ and $n := (N-1)/k \in \mathbb{N}$), and that $n > 10$. Let $T \geq K \ln(N/K)$ and fix $\epsilon := \sqrt{\frac{k \ln(n/10)}{100T}}$. Therefore, $k\frac{\ln(n/10)}{20\epsilon^2} = 5T$ and $\epsilon \leq 0.1$. Let $A$ be an algorithm for BwE, and suppose it has $R_T \leq \sqrt{Tk \ln(n/10)}/100000 := r(T)$. Thus:

$$\frac{1000r(T)}{\epsilon} = \frac{\sqrt{Tk \ln(n/10)}/100}{\sqrt{\frac{k \ln(n/10)}{100T}}} = T/10 < T.$$

Then, by Lemma 1, there exists a good SBI algorithm $A'$ with $T(A', S) \leq T = k\frac{\ln(n/10)}{100\epsilon^2}$ for any strategy $S$ from the pool. This contradicts Lemma 4, thus

$$R_T > \sqrt{Tk \ln(n/10)}/100000 = \Omega(\sqrt{TK \log(N/K)}),$$

as desired. ■

## 7. Future work

While in its full generality, the BwE problem allows the adversary to be adaptive, previous lower bounds hold also against an *oblivious* adversary who sets the advice and losses for all rounds in advance. We conjecture that the tight lower bound proved in this work holds against an oblivious adversary as well. More specifically, we conjecture that the oblivious adversary used in Ito (2024) achieves the desired lower bound even against an improper learner.

## Acknowledgments

We thank Emmanuel Esposito for insightful discussions on the works Cesa-Bianchi et al. (2025); Ito (2024).

IM is supported by the European Research Council (ERC) under the European Union's Horizon 2022 research and innovation program (grant agreement No. 101041711), the Israel Science Foundation (grant number 2258/19), and the Simons Foundation (as part of the Collaboration on the Mathematical and Scientific Foundations of Deep Learning). SI is supported by JSPS KAKENHI Grant Number JP25K03184 and by JST PRESTO, Japan, Grant Number JPMJPR2511.

## References

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolò Cesa-Bianchi, Khaled Eldowa, Emmanuel Esposito, and Julia Olkhovskaya. Improved regret bounds for bandits with expert advice. *Journal of Artificial Intelligence Research*, 83, 2025.

Houshuang Chen, Yuchen He, and Chihao Zhang. On interpolating experts and multi-armed bandits. In *International Conference on Machine Learning*, pages 6776–6802. PMLR, 2024.

Amit Daniely and Tom Helbertal. The price of bandit information in multiclass online classification. In *Conference on Learning Theory*, pages 93–104. PMLR, 2013.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

Steve Hanneke, Roi Livni, and Shay Moran. Online learning with simple predictors and a combinatorial characterization of minimax in 0/1 games. In *Conference on Learning Theory*, pages 2289–2314. PMLR, 2021.

Shinji Ito. On the minimax regret for contextual linear bandits and multi-armed bandits with expert advice. *Advances in Neural Information Processing Systems*, 37:61793–61812, 2024.

Satyen Kale. Multiarmed bandits with limited expert advice. In *Conference on Learning Theory*, pages 107–122. PMLR, 2014.

Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

H Brendan McMahan and Matthew J Streeter. Tighter bounds for multi-armed bandits with expert advice. In *COLT*, 2009.

Ananth Raman, Vinod Raman, Unique Subedi, Idan Mehalel, and Ambuj Tewari. Multiclass online learnability under bandit feedback. In *International Conference on Algorithmic Learning Theory*, pages 997–1012. PMLR, 2024.

Yevgeny Seldin and Gábor Lugosi. A lower bound for multi-armed bandits with expert advice. In *13th European Workshop on Reinforcement Learning (EWRL)*, volume 2, page 7, 2016.

Yevgeny Seldin, Peter Auer, John Shawe-taylor, Ronald Ortner, and François Laviolette. Pac-bayesian analysis of contextual bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.