Multi-Armed Bandits With Best-Action Queries

Francesco Bacchiocchi*
Politecnico di Milano
francesco.bacchiocchi@polimi.it

Matteo Castiglioni*
Politecnico di Milano
matteo.castiglioni@polimi.it

Alberto Marchesi*
Politecnico di Milano
alberto.marchesi@polimi.it

Francesco Emanuele Stradi*
Politecnico di Milano
francescoemanuele.stradi@polimi.it

Abstract

We study *multi-armed bandits* (MABs) augmented with *best-action queries*, in which the learner may additionally query an oracle that reveals the best arm in the current round. This setting was recently characterized by Russo et al. [2024] in the *full-feedback* model, where the learner observes the rewards of all arms after each round. They show that, in both *stochastic* and *adversarial* environments, k best-action queries reduce the optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret to $\widetilde{\mathcal{O}}(\min\{T/k, \sqrt{T}\})$. Whether this improvement extends to the more realistic *bandit-feedback* model—where the learner observes only the reward of the played arm—was left as an open problem. We address this question. When rewards are stochastic but correlated among arms, we show that the full-feedback result does not carry over: any algorithm must incur regret at least $\Omega(\sqrt{T-k})$. This lower bound directly extends to adversarial environments. On the positive side, we show that $\widetilde{\mathcal{O}}(\min\{T/k, \sqrt{T-k}\})$ regret is still achievable when rewards are stochastic and independent across arms, and establish a matching lower bound, up to logarithmic factors. Together, these results provide a complete characterization of the benefits of *best-action queries* in the *bandit-feedback* model.

1 Introduction

Multi-armed bandits (MABs) constitute a canonical framework for sequential decision-making under uncertainty [Auer et al., 2002a, Cesa-Bianchi and Lugosi, 2006, Lattimore and Szepesvári, 2020]. In a MAB problem, the learner repeatedly selects one among n actions (arms) over a horizon of T rounds, observing only the reward of the played arm. The goal of the learner is to maximize the cumulative reward collected during learning, or equivalently, to minimize the *regret* with respect to the best fixed arm in expectation. MABs capture a fundamental exploration-exploitation tradeoff and have thus become a standard model for applications such as recommendation systems [Li et al., 2010], online experimentation [Villar et al., 2015], and digital advertising [Chapelle and Li, 2011].

Motivated by the recent growing literature on algorithms with machine learning-based predictions (see, *e.g.*, the survey [Mitzenmacher and Vassilvitskii, 2022]), we study MABs augmented with *best-action queries* [Russo et al., 2024], in which the learner may query an oracle in at most k rounds, where the oracle reveals the best arm in the current round. This models situations in which highly reliable information about the instantaneous optimal choice is available, but only through an expensive channel, such as expert intervention, high-fidelity simulation, offline evaluation, or a slower and more accurate decision pipeline. These high-quality interventions are typically too costly to invoke at every round because of monetary, computational, or latency constraints, and must therefore

*Authors are listed in alphabetical order.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

be used sparingly. From this perspective, best-action queries provide a clean abstraction for studying the value of *rare but highly informative feedback* in online learning. A concrete example is an online platform that continuously moderates user-generated content (*e.g.*, Meta or Google). When a new post is created, the platform must decide whether to flag the content as harmful. The platform has two options: (i) relying on reviews by human experts (*i.e.*, best-action queries), or (ii) invoking a learning algorithm that makes automated decisions. Clearly, due to budget constraints, human review is a scarce resource, and thus the platform can choose option (i) only a limited number of times. Russo et al. [2024] fully characterize the impact of best-action queries in *full-feedback* settings, where after each round the learner observes the rewards of all arms. They show that k best-action queries improve the optimal regret rate to $\widetilde{\mathcal{O}}(\min\{T/k, \sqrt{(T-k)}\})$ in both stochastic and adversarial settings.[2] It is quite surprising that a sublinear number of T^α queries, $\alpha \in (1/2, 1)$, is sufficient to drop the regret to $\widetilde{\mathcal{O}}(T^{1-\alpha})$. For instance, $T^{2/3}$ queries are sufficient to drop the regret to $\widetilde{\mathcal{O}}(T^{1/3})$. Whether similar results can be recovered under the more realistic *bandit-feedback* settings—where only the reward of the played arm is observed—was left open by Russo et al. [2024]. Understanding whether the same improvements can still be achieved under bandit feedback is substantially more challenging, since the learner can no longer rely on observing the outcomes of unplayed actions, and the informational value of a query must be assessed in a partial-feedback regime. In this work, we address the following research question:

Can best-action queries improve the performance of an online learner when only bandit feedback is available?

Original Contribution We fully resolve this question. First, we establish a negative result: any algorithm with access to k best-action queries must incur regret of $\Omega(\sqrt{T-k})$ when rewards are stochastic but *correlated* across arms. This lower bound naturally extends to *adversarial* environments. Thus, in such settings, best-action queries cannot substantially improve the learner's performance. To prove this, we construct two correlated instances that are not only difficult to distinguish but also limit the "query power" by simultaneously bounding the negative regret and the information gained about the optimal action. We also note that our result is tight: a simple $\widetilde{\mathcal{O}}(\sqrt{n(T-k)})$ upper bound is achievable by combining any standard adversarial regret minimizer with k arbitrary oracle queries.

We then turn to the *stochastic* setting with i.i.d. rewards. Here, the landscape changes significantly: we design an algorithm achieving regret $\widetilde{\mathcal{O}}(\min\{Tn/k, \sqrt{n(T-k)}\})$. This result stems from a key observation: the previous lower bound breaks in this regime because making two instances hard to distinguish requires the arm rewards to have large variance. This high variance, in turn, allows each query to yield a large negative per-round regret.

Our analysis formalizes this observation by lower bounding the expected negative regret of a query as a function of the variance of the underlying distributions. This highlights the trade-off between two extremes: (i) the variance of the distributions is small, and hence a variance-aware learner can achieve regret significantly better than the standard $\mathcal{O}(\sqrt{T})$, (ii) the variance of the distributions is large, and hence, while the optimal regret in the rounds without query is the standard $\mathcal{O}(\sqrt{T})$, the queries compensate by providing a large negative regret. Although our analysis is fairly involved, the resulting algorithm is remarkably simple: we query the optimal arm in the first k rounds (ignoring the feedback), and then deploy a variance-aware regret minimizer, such as UCB-V [Audibert et al., 2009], for the remaining rounds. Finally, we provide a matching lower bound of $\Omega(\min\{Tn/k, \sqrt{n(T-k)}\})$. Notice that the analogous lower bound for full-feedback [Russo et al., 2024] is inapplicable here as it relies on correlated rewards. Overall, our results, summarized in Table 1, provide a complete characterization of the value of best-action queries under *bandit feedback*, resolving the problem left open by Russo et al. [2024].

1.1 Related Works

In this section, we provide additional discussion on related works.

Hints and query-based online learning Best-action queries are related to the broader literature on online learning with hints. Dekel et al. [2017] study online linear optimization with hints that are

[2] In this work, we use $\widetilde{\mathcal{O}}$ to hide logarithmic factors.

Setting	Upper bound	Lower bound
Stoc. *independent*	$\widetilde{\mathcal{O}}\left(\min\left\{\frac{T}{k}, \sqrt{T-k}\right\}\right)$	$\Omega\left(\min\left\{\frac{T}{k}, \sqrt{T-k}\right\}\right)$
Stoc. *correlated* (Adv.)	$\widetilde{\mathcal{O}}\left(\sqrt{T-k}\right)$	$\Omega\left(\sqrt{T-k}\right)$

Table 1: Regret guarantees in MABs with best-action queries under bandit feedback

correlated with the loss vectors, showing that such hints can lead to logarithmic regret over strongly convex domains. This line was extended to imperfect hints by Bhaskara et al. [2020], to sublinear hint budgets by Bhaskara et al. [2021], and to bandit online linear optimization with hints and queries by Bhaskara et al. [2023a]. These works differ from ours in both the feedback structure and the nature of the side information: hints are typically vector-valued and correlated with the loss, whereas our queries return only ordinal information, namely the identity of the best arm in a single round. A closely related query model is studied by Bhaskara et al. [2023b], where the learner may compare a small subset of actions and observe which action in the queried subset has the smallest loss before playing. Their results show that ordinal queries can lead to strong improvements in online learning and stochastic bandits. Our model is different in two key ways. First, a best-action query compares all arms rather than a small queried subset, and is therefore more informative at the single-round level. Second, and crucially, we impose a global query budget k, so the learner must decide how to allocate a scarce and expensive source of information over time.

Algorithms with predictions Our work also fits within the growing literature on algorithms with predictions, where machine-learned advice is used to improve performance beyond worst-case guarantees while retaining robustness when predictions are inaccurate [Mitzenmacher and Vassilvitskii, 2022]. Examples include online algorithms for ski rental and scheduling [Purohit et al., 2018], caching with learned advice [Lykouris and Vassilvitskii, 2021], rent-or-buy problems with expert advice [Gollapudi and Panigrahi, 2019], and scheduling with learned weights [Lattanzi et al., 2020]. In contrast to much of this literature, our side information is not an exogenous prediction available at every round, but an actively requested oracle response subject to a hard budget. Thus, the main challenge is not only how to exploit side information when it is available, but also whether and when such scarce information can improve regret under partial feedback.

Bandits with preference feedback Our work is also connected to bandit models with ordinal, preference, or ranking feedback. In the dueling-bandit framework, the learner selects a pair of arms and observes only a noisy preference between them, rather than cardinal rewards [Yue and Joachims, 2009, Yue et al., 2012]. This line of work has led to algorithms based on pairwise preference estimates, such as Beat-the-Mean [Yue and Joachims, 2011], RUCB [Zoghi et al., 2014], and variants that target notions of optimality beyond Condorcet winners, such as Copeland winners [Zoghi et al., 2015]. More general ranking-feedback models have also been studied in learning-to-rank and recommendation settings, including ranked bandits [Radlinski et al., 2008], cascading bandits [Kveton et al., 2015], top-k feedback [Chaudhuri and Tewari, 2017], and subset-wise ranking feedback under Plackett–Luce models [Saha and Gopalan, 2019]. Most closely related to this perspective, Maran et al. [2024] introduce bandits with ranking feedback, where the learner receives ordinal information that allows arms to be ranked based on previous pulls, without observing numerical reward differences. These works differ from ours in the type and timing of the ordinal information. In dueling-bandit and ranking-feedback models, ordinal feedback is typically the learner's primary feedback signal and is obtained repeatedly from queried pairs, subsets, ranked lists, or rankings induced by the history of past pulls. By contrast, in our setting the learner normally receives standard bandit rewards, and ordinal information is available only through a limited number of best-action queries.

2 Multi-Armed Bandits With Best-Action Queries

We consider a *multi-armed bandit* (MAB) problem with $n \in \mathbb{N}$ arms and horizon $T \in \mathbb{N}$. At each round $t \in [T]$,[3] the environment specifies a reward vector $X_t = (X_{t,1}, \ldots, X_{t,n}) \in [0,1]^n$,

[3]Throughout the paper, we write $[a] := \{1, \ldots, a\}$ for the set of the first $a \in \mathbb{N}$ natural numbers.

the learner selects an arm (or action) $I_t \in [n]$, possibly at random, and observes the reward X_{t,I_t} associated with the chosen arm. We study two different *stochastic* settings. In the simpler independent setting, each reward $X_{t,i}$ is sampled independently across rounds and arms from a distribution D_i. In the more general *correlated* setting, the rewards of all the arms are sampled from a joint distribution D at each round t, *i.e.*, $X_t \sim D$, independently across rounds. Notice that this is a special case of the *adversarial* setting.

In addition to the standard MAB interaction, the learner is endowed with a budget of $k \in [T]$ *best-action queries*. At the beginning of each round $t \in [T]$, the learner may query the oracle. If a query is issued, the oracle reveals a reward-maximizing arm for that round, namely an arm $i_t^\star \in \arg\max_{i\in[n]} X_{t,i}$.[4] In this case, the learner plays $I_t = i_t^\star$ and observes the corresponding reward $X_{t,i_t^\star}$. Otherwise, the interaction proceeds as in a standard MAB: the learner selects an arm $I_t \in [n]$ and then observes the reward X_{t,I_t}. The learner may issue at most k queries in total over the T rounds. Formally, we let $Q_t \in \{0,1\}$ denote whether a query is issued at round t, and define $\mathcal{Q} := \{t \in [T] \mid Q_t = 1\}$ as the set of rounds in which the learner queries the oracle. Notice that $|\mathcal{Q}| \le k$ must hold deterministically.

Protocol 1 Learner-Environment Interaction

Require: Number of available queries $k \in [T]$

for $t = 1, \dots, T$ **do**
  The environment selects $X_t = (X_{t,1}, \dots, X_{t,n}) \in [0,1]^n$
  The learner decides $Q_t \in \{0,1\}$
  $Q_t \leftarrow Q_t \cdot \mathbb{I}\{\sum_{\tau=1}^{t} Q_\tau \le k\}$
  if $Q_t = 1$ **then**
    The oracle reveals $i_t^\star$
    The learner plays $I_t = i_t^\star$
  else
    The learner plays $I_t \in [n]$
  end if
  The learner observes X_{t,I_t}
end for

Protocol 1 formalizes the interaction between the learner and the environment in a MAB problem with best-action queries.

The goal of the learner is to minimize the *regret*. We denote the (cumulative) regret suffered by a learning algorithm $\mathfrak{A}$ as follows:

$$R_{T,k}(\mathfrak{A}) := \max_{i\in[n]} \sum_{t\in[T]} \mathbb{E}[X_{t,i}] - \mathbb{E}\left[\sum_{t\in[T]} X_{t,I_t}\right],$$

where the expectation is taken with respect to the randomness of the algorithm and the environment.

3 When Best-Action Queries Do Not Help: The Stochastic Correlated Setting

We start by providing an impossibility result for the case in which the rewards are stochastic and correlated across arms. This clearly implies that the same impossibility result holds for the more general adversarial setting. Specifically, we prove that, in contrast to the full-feedback setting [Russo et al., 2024], best-action queries do not significantly improve the regret guarantees attainable, that is, any algorithm is guaranteed to suffer $\Omega(\sqrt{T-k})$ regret. Interestingly, this shows that correlation is sufficient to make best-action queries essentially useless.

Theorem 3.1. *For every $T \ge 1$ and $k \in [T-1]$, and every randomized learning algorithm $\mathfrak{A}$ that uses at most k best-action queries, there exists a MAB problem instance with two arms and stochastic correlated rewards such that the regret of $\mathfrak{A}$ satisfies:*

$$R_{T,k}(\mathfrak{A}) \ge \Omega(\sqrt{T-k}).$$

Our impossibility result is based on building a family of two hard instances ν_1, ν_2. In both instances, we use a uniform random variable $W_t \sim \text{Unif}\{1,2\}$ to correlate the rewards of the two possible arms. Then, the two instances ν_1, ν_2 are defined as follows:

$$\textcircled{\nu_1} \;:\; (X_{t,1}, X_{t,2}) = \begin{cases} (Z_t^+,\, Y_t), & \text{if } W_t = 1, \\ (Z_t^-,\, Z_t^+), & \text{if } W_t = 2. \end{cases}$$

[4] In this paper, we assume that the oracle breaks ties according to a fixed priority order over the arms, chosen in advance and used in every round. This order may be known to the learner.

$$\textcircled{\nu_2} \;:\; (X_{t,1}, X_{t,2}) = \begin{cases} (Z_t^+,\, Z_t^-), & \text{if } W_t = 1, \\ (Y_t,\, Z_t^+), & \text{if } W_t = 2. \end{cases}$$

In both instances, $Y_t \sim \text{Unif}[0,1]$ and Z_t^-, Z_t^+ are chosen as a function of Y_t in order to guarantee that $Y_t < Z_t^- < Z_t^+$ holds almost surely (we defer the formal construction to the proof).

It is easy to see that the two instances differ only in which arm gets the smaller reward Z_t^-. Indeed, under ν_1 this creates a slight advantage for arm 1 in expectation, whereas under ν_2 the same advantage is given to arm 2. In the following, we denote by Δ the gap between the expected rewards of the optimal arm and the suboptimal arms.

Our construction deals with the tension between two opposing objectives: keeping the KL between the two instances small and controlling the negative regret obtained by best-action queries. The main challenge is dealing with the tradeoff between the suffered (positive) regret and the negative regret accumulated by using best-action queries. To ensure that the (negative) regret is small in magnitude, we would like the best fixed arm, *i.e.*, our baseline $i^\star$, to be optimal at every round. This idea has been used in previous work such as [Gerchinovitz and Lattimore, 2016]. However, this is not suitable in our setting, since it would allow the learner to identify the optimal arm with a single query.

We use correlated rewards to deal with this tension. Indeed, in our construction a query reveals the arm that is optimal at the current round and allows the learner to collect the top reward Z_t^+, but this feedback has the same distribution under ν_1 and ν_2, and therefore does not help identify which arm is optimal in expectation. In contrast, on rounds without queries, the learner observes only the reward of the selected arm, and these observations are drawn from two nearby distributions whose KL divergence is of order Δ^2. Hence, up to the m-th non-query round, where $m := T - k$, the total information available for distinguishing ν_1 from ν_2 is only of order $m\Delta^2$. Choosing $\Delta = \Theta(m^{-1/2})$ keeps the two environments statistically close, so any learner must frequently pull the wrong fixed arm on non-queried rounds, yielding regret of order $m \cdot \Delta = \Theta(\sqrt{T-k})$.

We finally remark that any adversarial no-regret algorithm which attains a regret upper bound $\widetilde{\mathcal{O}}(\sqrt{T})$ (*e.g.*, [Auer et al., 2002b]) provides $\widetilde{\mathcal{O}}(\sqrt{T-k})$ regret when endowed with best-action queries. This is because the regret incurred by any algorithm during the best-action query rounds is at most zero.

4 Best-Action Queries in the Stochastic Independent Setting

In this section, we show that when the rewards are independent across arms and i.i.d. over time, the benefit of best-action queries improves dramatically. Specifically, the results of this section will be presented in the following order:

- In Section 4.1, we study the gap between the value of the optimal static policy and the optimal dynamic one, highlighting a clear connection between this gap and the variance of the instance.
- In Section 4.2, we present our algorithm and the main intuition behind it.
- In Section 4.3, we provide the theoretical guarantees achieved by our algorithm.
- In Section 4.4, we prove a matching lower bound, showing that the regret guarantee achieved by our algorithm is optimal up to logarithmic factors in the stated regime.

4.1 The Gap Between Dynamic and Static Policies and Its Relation to Variance

As a first step, in this section we establish a relationship between the gap separating the optimal dynamic and static policies and the variance of the rewards. To this end, we first need to introduce some additional useful notation for the stochastic independent setting.

For ease of presentation, we let $X_i \sim D_i$ denote a generic random variable drawn from the reward distribution D_i of arm $i \in [n]$. Then, we define $\mu_i := \mathbb{E}[X_i]$ and $\sigma_i^2 := \text{Var}(X_i)$ as the mean and variance of arm i's reward distribution. Additionally, we denote with $i^\star \in \arg\max_{i\in[n]} \mu_i$ an optimal arm and with $\mu^\star = \mu_{i^\star}$ its mean reward. As usual, we denote by $\Delta_i := \mu^\star - \mu_i$ the suboptimality gap of arm $i \in [n]$ and we denote with $\mathcal{S} := \{i \in [n] \mid \Delta_i > 0\}$ the set of suboptimal arms. We also define two quantities corresponding to the per-round optimal value of the optimal *static* and *dynamic*

policies, respectively, as follows:

$$\mathrm{OPT_s} := \max_{i\in[n]} \mathbb{E}\left[X_i\right] = \mu^\star, \qquad \mathrm{OPT_d} := \mathbb{E}\left[\max_{i\in[n]} X_i\right].$$

Intuitively, the first policy just knows the distribution, while the second one picks an arm *after* observing a sample from the distribution. Now, we show how to relate the difference between the dynamic and static optimal values to the suboptimality gaps and the variances of the arms. To do that, we observe that:

$$\mathrm{OPT_d} - \mathrm{OPT_s} = \mathbb{E}\left[\max_{j\in[n]} X_j\right] - \mu_{i^\star} \geq \frac{1}{n}\sum_{i\in[n]} \Big(\mathbb{E}[\max\{X_{i^\star}, X_i\}] - \mu_{i^\star}\Big).$$

The right-hand side of the inequality is related to the variance of the variables. Specifically,

$$\mathbb{E}[\max\{X_{i^\star}, X_i\}] - \mu_{i^\star} = \frac{1}{2}\left(\mathbb{E}[|X_{i^\star} - X_i|] - \Delta_i\right) \quad \forall i \in \mathcal{S},$$

which follows from employing $\max\{a, b\} = (a + b + |a - b|)/2$. As a final observation, notice that:

$$\mathbb{E}\left[|X_{i^\star} - X_i|\right] \geq \mathbb{E}\left[(X_{i^\star} - X_i)^2\right] = \mathrm{Var}\left(X_{i^\star} - X_i\right) + \left(\mathbb{E}\left[X_{i^\star} - X_i\right]\right)^2 \geq \sigma_{i^\star}^2 + \sigma_i^2 \geq \sigma_i^2,$$

since the rewards are independent. Define $(\,\cdot\,)_+ := \max\{\,\cdot\,, 0\}$. We can now combine these results to get the following lemma.

Lemma 4.1. *Given a MAB instance with stochastic independent rewards, the following holds:*

$$\mathrm{OPT_d} - \mathrm{OPT_s} \geq \frac{1}{2n}\sum_{i\in\mathcal{S}} \left(\sigma_i^2 - \Delta_i\right)_+.$$

Lemma 4.1 relates the difference between optimal *dynamic* and *static* policies, showing that this gap admits a lower bound that grows with the variance of suboptimal arms whenever the latter is sufficiently large compared to their suboptimality gap.

We now show how the gap between the dynamic and static optimal values affects the learner's regret. Intuitively, whenever the learner queries the oracle, it receives the dynamic optimal value, while the benchmark against which regret is measured—namely, the static optimum—remains unchanged. Therefore, each query contributes exactly minus the gap between the dynamic and static optimal values, which yields the following decomposition of the learner's regret.

Lemma 4.2. *Consider any algorithm $\mathfrak{A}$ that, given as input a time horizon $T \in \mathbb{N}$ and a number $k \in [T]$ of allowed best-action queries, issues a query to the oracle for exactly k rounds and, in the remaining $T - k$ rounds, ignores the oracle feedback and employs an algorithm $\mathfrak{A}_0$ for standard stochastic independent MAB problems. The base algorithm uses only its own observations, private randomness, and local non-query round counter. Then, the following holds:*

$$R_{T,k}(\mathfrak{A}) = R_{T-k,0}(\mathfrak{A}_0) - k(\mathrm{OPT_d} - \mathrm{OPT_s}).$$

We use the convention $R_{0,0}(\mathfrak{A}_0) = 0$.

Proof. Let $Q \subseteq [T]$ be the set of query rounds. By assumption, $|Q| = k$. On every queried round $t \in Q$, the learner plays an arm:

$$i_t^\star \in \arg\max_{i\in[n]} X_{t,i},$$

and thus the reward is $X_{t,I_t} = \max_{i\in[n]} X_{t,i}$. The query decision Q_t is made before observing the current reward vector and is independent of it, conditional on the past and the learner's private randomness. Since reward vectors are i.i.d. over time, $\mathbb{E}[Q_t X_{t,I_t}] = \mathbb{E}[Q_t]\mathrm{OPT_d}$. Therefore,

$$\mathbb{E}\left[\sum_{t\in Q} X_{t,I_t}\right] = k\,\mathrm{OPT}_d. \tag{1}$$

Now consider the $T - k$ rounds without queries. By assumption, on those rounds the algorithm ignores the oracle feedback and runs the standard stochastic bandit algorithm $\mathfrak{A}_0$. Conditionally on

the past, the learner's initial random seed, and the decision not to query, the current reward vector still has its original distribution. Since rewards are i.i.d. across rounds, the sequence of observations collected on the non-query rounds has the same law as the sequence of observations that $\mathfrak{A}_0$ would receive in a fresh stochastic bandit instance of horizon $T-k$. Thus, we have:

$$\mathbb{E}\left[\sum_{t\notin Q} X_{t,I_t}\right] = (T-k)\,\mathrm{OPT}_s - R_{T-k,0}(\mathfrak{A}_0). \tag{2}$$

Summing Equation (1) and Equation (2) we get $\mathbb{E}\left[\sum_{t=1}^{T} X_{t,I_t}\right] = (T-k)\,\mathrm{OPT}_s - R_{T-k,0}(\mathfrak{A}_0) + k\,\mathrm{OPT}_d$. Thus,

$$\begin{aligned} R_{T,k}(\mathfrak{A}) &= T\,\mathrm{OPT_s} - \mathbb{E}\left[\sum_{t=1}^{T} X_{t,I_t}\right] \\ &= T\,\mathrm{OPT_s} - \Big((T-k)\,\mathrm{OPT_s} - R_{T-k,0}(\mathfrak{A}_0) + k\,\mathrm{OPT}_d\Big) \\ &= R_{T-k,0}(\mathfrak{A}_0) - k(\mathrm{OPT}_d - \mathrm{OPT}_s). \end{aligned}$$

This concludes the proof. □

Lemma 4.2 shows that endowing an algorithm with k best-action queries reduces its regret by an amount proportional to both k and the gap $\mathrm{OPT_d} - \mathrm{OPT_s}$. We can now combine Lemma 4.1 and Lemma 4.2. This allows us to lower bound the gap between dynamic and static optimal values $\mathrm{OPT_d} - \mathrm{OPT_s}$, in order to get the following corollary.

Corollary 4.3. *For every $k \in [T]$, under the same assumptions as Lemma 4.2, it holds:*

$$R_{T,k}(\mathfrak{A}) \le R_{T-k,0}(\mathfrak{A}_0) - \frac{k}{2n}\sum_{i\in\mathcal{S}}(\sigma_i^2 - \Delta_i)_+.$$

The above corollary has an important implication. Consider hard instances in which the regret of a standard stochastic MAB algorithm is of order $\Theta(\sqrt{T})$. For the usual fixed-n hard instances, the arm variances remain constant, while the suboptimality gaps satisfy $\Delta_i = \Theta(1/\sqrt{T})$. Hence, for sufficiently large T, each term $(\sigma_i^2 - \Delta_i)_+$ is bounded away from zero, and therefore the second term in the corollary is of order $\Theta(k)$. In other words, in these instances, employing k best-action queries leads to a $\Theta(k)$ improvement in the final regret guarantees. For instance, if we set $k = T^{2/3}$ and choose UCB1 [Auer et al., 2002a] as the underlying algorithm, then the negative term induced by the queries eventually dominates the usual regret term of UCB1, so that the resulting regret becomes negative for sufficiently large T. This suggests that access to best-action queries can dramatically improve performance.

4.2 A Variance-Aware Regret Minimizer

In the previous section, we showed how best-action queries can significantly improve the regret guarantees when the variance-related quantities $(\sigma_i^2 - \Delta_i)_+$ are large. In this section, we deal with the settings in which such quantities are small, so this lower bound on the query gain may be small. In this case, we show that using a variance-aware regret minimizer in the rounds without queries leads to a small regret. Accordingly, we employ the UCB-V algorithm [Audibert et al., 2009] as a regret minimizer. For the sake of clarity, we restate here the regret guarantees of UCB-V.

Theorem 4.4 (UCB-V guarantees [Audibert et al., 2009])**.** *Consider a MAB instance with stochastic independent rewards in $[0,1]$, and let $\mathcal{S} := \{i \in [n] \mid \Delta_i > 0\}$ denote the set of strictly suboptimal arms. For every horizon $T > 1$, let $N_i(T)$ denote the number of times arm $i \in [n]$ has been selected up to round T. Then, for every $i \in \mathcal{S}$, the UCB-V algorithm guarantees:*

$$\Delta_i \mathbb{E}[N_i(T)] \le C\left(\frac{\sigma_i^2}{\Delta_i} + 1\right)\log T,$$

for an absolute constant $C > 0$.

Intuitively, Theorem 4.4 shows how, by employing the UCB-V algorithm, it is possible to attain regret guarantees that strongly benefit from instances characterized by low variance. For instance, when $\sigma_i^2 = \Theta(\Delta_i)$, a regret upper bound of order $\log T$ can be recovered even for $\Delta_i = \Theta(1/\sqrt{T})$.

Algorithm 1 UCB-V with Best-Action Queries

Require: Number of available queries $k \in [T]$

1: **for** $t = 1, \dots, k$ **do**
2: The oracle reveals $i_t^\star$
3: The learner plays $I_t = i_t^\star$
4: **end for**
5: Initialize UCB-V from scratch
6: Run UCB-V for the remaining $T - k$ rounds

In Algorithm 1, we show how to instantiate UCB-V into our setting in which the learner is endowed with best-action queries. Specifically, for the first k rounds the learner exploits the queries to the oracle. Notice that, since the rewards are stochastic, spending the entire query budget at the beginning of the learning process yields the same expected reward as any other way of allocating the same budget over time. Thus, for simplicity, we deplete the query budget in the first rounds. Then, UCB-V is initialized from scratch, *i.e.*, we discard all the information collected during the probing phase, since it is not needed for the worst-case regret guarantee proved below. Finally, we run the UCB-V algorithm for the remaining $T - k$ rounds.

4.3 The Final Regret Bound

In this section, we show how to combine the results of Section 4.1 and Section 4.2 to deal with the tension between the variance and the gap $\mathrm{OPT_d} - \mathrm{OPT_s}$. Specifically, we show that when the variance is large then $\mathrm{OPT_d} - \mathrm{OPT_s}$ is large, making the queries extremely useful. On the other hand, when the variance is small, a variance-aware regret minimizer such as UCB-V provides a much better regret rate than the standard $\widetilde{\mathcal{O}}(\sqrt{T})$.

Formally, we prove the following theorem.

Theorem 4.5. *Fix $k \in [T]$ and let $\mathfrak{A}$ be Algorithm 1 with input k. Then,*

$$R_{T,k}(\mathfrak{A}) \leq \mathcal{O}\left(\min\left\{\frac{nT\log T}{k}, \sqrt{n(T-k)\log T}\right\}\right).$$

Proof. The claim is trivial when $k = T$, since the learner queries in every round and thus: $R_{T,T}(\mathfrak{A}) = T \cdot (\mathrm{OPT_s} - \mathrm{OPT_d}) \leq 0$. Hence, in the rest of the proof, we assume $k \leq T - 1$. We define $M := T - k$. If $M = 1$, the regret is at most 1, which satisfies the stated bound since $T \geq 2$. Hence we may assume $M \geq 2$. For each $i \in \mathcal{S}$, let $N_i(M)$ denote the number of times arm $i \in [n]$ is pulled by the UCB-V algorithm after the initial querying phase. By Lemma 4.2,

$$R_{T,k}(\mathfrak{A}) = R_{M,0}(\text{UCB-V}) + k(\mathrm{OPT_s} - \mathrm{OPT_d}).$$

Since only strictly suboptimal arms contribute to the regret of the UCB-V phase, it holds:

$$R_{M,0}(\text{UCB-V}) = \sum_{i \in \mathcal{S}} \Delta_i \mathbb{E}[N_i(M)],$$

where $\mathbb{E}[N_i(M)]$ is the expected number of times UCB-V played arm i during these M rounds.

Notice that we can employ Lemma 4.2 since our algorithm performs best-action queries in the first k rounds and then runs the UCB-V algorithm after these rounds. Moreover, by Lemma 4.1,

$$\mathrm{OPT_d} - \mathrm{OPT_s} \geq \frac{1}{2n}\sum_{i \in \mathcal{S}} (\sigma_i^2 - \Delta_i)_+.$$

Therefore, it holds:

$$R_{T,k}(\mathfrak{A}) \leq \sum_{i \in \mathcal{S}} \left[\Delta_i \mathbb{E}[N_i(M)] - \frac{k}{2n}(\sigma_i^2 - \Delta_i)_+\right]. \tag{3}$$

For each $i \in \mathcal{S}$, we define $R_i := \Delta_i \mathbb{E}[N_i(M)]$. By Theorem 4.4, we have that:

$$R_i \leq C\left(\frac{\sigma_i^2}{\Delta_i} + 1\right)\log M = 2C\log M + C\left(\frac{\sigma_i^2}{\Delta_i} - 1\right)\log M.$$

Let $x_i := (\sigma_i^2 - \Delta_i)_+$, then we have:

$$\begin{aligned} R_i &= \min\left\{2C\log M + C\left(\frac{\sigma_i^2}{\Delta_i} - 1\right)\log M, \Delta_i \mathbb{E}[N_i(M)]\right\} \\ &\leq 2C\log M + \min\left\{C\left(\frac{\sigma_i^2}{\Delta_i} - 1\right)\log M, \Delta_i \mathbb{E}[N_i(M)]\right\} \\ &\leq 2C\log M + \sqrt{C\left(\frac{\sigma_i^2}{\Delta_i} - 1\right)_+ \Delta_i \mathbb{E}[N_i(M)]\log M} \\ &\leq 2C\log T + \sqrt{C x_i \mathbb{E}[N_i(M)]\log T}, \end{aligned}$$

where in the second inequality we used $\min\{a, b\} \leq \sqrt{ab_+}$ that holds for every $a \geq 0$ and $b \in \mathbb{R}$. Then,

$$R_i - \frac{k}{2n}x_i \leq 2C\log T + \sqrt{C x_i \mathbb{E}[N_i(M)]\log T} - \frac{k}{2n}x_i. \tag{4}$$

Now apply the inequality $a\sqrt{x} - bx \leq a^2/4b$ that holds for every $a \geq 0$, $b > 0$ and $x \geq 0$, where we choose $a := \sqrt{C\mathbb{E}[N_i(M)]\log T}$ and $b := k/2n$, which gives the following bound:

$$\sqrt{C\, x_i\, \mathbb{E}[N_i(M)]\, \log T} - \frac{k}{2n}x_i \leq \frac{Cn}{2k}\mathbb{E}[N_i(M)]\log T. \tag{5}$$

Therefore, plugging Equation (5) into Equation (4), for every $i \in \mathcal{S}$, we get:

$$R_i - \frac{k}{2n}x_i \leq 2C\log T + \frac{Cn}{2k}\mathbb{E}[N_i(M)]\log T.$$

Summing over all $i \in \mathcal{S}$, and recalling Equation (3) and the definitions of R_i and x_i, we obtain:

$$R_{T,k}(\mathfrak{A}) \leq 2Cn\log T + \frac{Cn\log T}{2k}\sum_{i\in\mathcal{S}}\mathbb{E}[N_i(M)].$$

The number of pulls of suboptimal arms during the UCB-V phase is at most the total number of pulls, *i.e.*, $\sum_{i\in\mathcal{S}}\mathbb{E}[N_i(M)] \leq M \leq T$. Thus, it holds:

$$R_{T,k}(\mathfrak{A}) \leq 2Cn\log T + \frac{CnT\log T}{2k} \leq \frac{5C}{2}\frac{nT\log T}{k},$$

since $k \leq T$. To conclude the proof, we observe that UCB-V attains the same instance-independent regret upper bound as UCB1 [Auer et al., 2002a], which is of order $\mathcal{O}(\sqrt{nM\log M})$. Indeed, $\sigma_i^2 \leq 1$ and $\Delta_i \leq 1$ give $R_i \leq 2C\log M/\Delta_i$, and therefore $R_i^2 \leq 2C\log M\,\mathbb{E}[N_i(M)]$. By Cauchy-Schwarz,

$$R_{T,k}(\mathfrak{A}) \leq \sum_{i\in\mathcal{S}} R_i \leq \sqrt{2C|\mathcal{S}|\log M \sum_{i\in\mathcal{S}}\mathbb{E}[N_i(M)]} \leq \sqrt{2CnM\log M}. \qquad \square$$

4.4 A Matching Lower Bound

In this section, we provide a matching lower bound for the stochastic independent setting. For clarity, we describe the core intuition using a two-arm setting, though our theorem and formal proof generalize to an arbitrary number of arms n. Our construction stems from the observation that the hardest instances arise when the problem parameters satisfy the following relationship:

$$\sigma^2 = \Theta\left(\Delta\frac{T}{k}\right) \quad \text{and} \quad \Delta = \Theta\left(\frac{1}{k}\right).$$

We consider two problem instances, ν_1 and ν_2. In ν_1, the first arm follows a Bernoulli distribution with mean $1-p$, and the second arm follows a Bernoulli with mean $1-p+\Delta$. In ν_2, these reward distributions are swapped. A distinctive feature of our approach is the use of Bernoulli distributions where the means are close to 1. While standard lower bounds often use distributions centered around $1/2$, as we will see in the following, our construction leverages the fact that the mean approaches

the boundary. In principle, this makes the instance easier to learn, since the variance is smaller and the KL divergence scales as $\mathcal{O}(\Delta^2/p)$ rather than the typical $\mathcal{O}(\Delta^2)$. Nonetheless, the motivation behind this choice is rooted in Lemma 4.1, which shows that high-variance distributions (like those centered at $1/2$) incur large negative regret during the querying rounds. To see this, consider two Bernoulli distributions with means close to $1/2$. In this case, the gap between the value of optimal dynamic policy $\mathrm{OPT}_{\mathrm{d}} \simeq 3/4$ and the optimal static one $\mathrm{OPT}_{\mathrm{s}} = 1/2$ is constant. However, by shifting the distributions such that means are at least $1 - p$, this gap is compressed to at most p, as $\mathrm{OPT}_{\mathrm{s}} \geq 1 - p$ and $\mathrm{OPT}_{\mathrm{d}} \leq 1$. This compression is essential for matching the upper bound.

Formally we prove the following.

Theorem 4.6. *There exists an absolute constant C such that, for $2 \leq n \leq T$ and any randomized algorithm $\mathfrak{A}$ which issues at most k best-action queries with $1 \leq k \leq T/C$, there exists a stochastic independent instance for which:*

$$R_{T,k}(\mathfrak{A}) \geq \Omega\left(\min\left\{\frac{Tn}{k}, \sqrt{n(T-k)}\right\}\right).$$

Theorem 4.6 shows that Algorithm 1 is optimal up to logarithmic factors in the stochastic independent setting for $k \leq T/100$ and $2 \leq n \leq T$.

Acknowledgments

This publication was funded with the contribution of Ministero dell'Università e della ricerca pursuant to D.D. n. 18010 of 12 November 2025 – BANDO FIS 3. Project FIS-2024-05736 (Starting Grant), title: "Towards a trustworthy strategic use of data in machine learning pipelines" (STRATDATA). CUP: D53C25002380001.

References

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002a.

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

Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Online learning with imperfect hints. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 822–831. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bhaskara20a.html.

Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Logarithmic regret from sublinear hints. *Advances in Neural Information Processing Systems*, 34:28222–28232, 2021.

Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Bandit online linear optimization with hints and queries. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2313–2336. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/bhaskara23a.html.

Aditya Bhaskara, Sreenivas Gollapudi, Sungjin Im, Kostas Kollias, and Kamesh Munagala. Online learning and bandits with queried hints. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 16:1–16:24. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023b.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. *Advances in Neural Information Processing Systems*, 24, 2011.

Sougata Chaudhuri and Ambuj Tewari. Online learning to rank with top-k feedback. *Journal of Machine Learning Research*, 18(103):1–50, 2017. URL http://jmlr.org/papers/v18/16-285.html.

Ofer Dekel, Arthur Flajolet, Nika Haghtalab, and Patrick Jaillet. Online learning with a hint. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/22b1f2e0983160db6f7bb9f62f4dbb39-Paper.pdf.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. *Advances in Neural Information Processing Systems*, 29, 2016.

Sreenivas Gollapudi and Debmalya Panigrahi. Online algorithms for rent-or-buy with expert advice. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2319–2327. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/gollapudi19a.html.

Branislav Kveton, Csaba Szepesvári, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 767–776, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/kveton15.html.

Silvio Lattanzi, Thomas Lavastida, Benjamin Moseley, and Sergei Vassilvitskii. Online scheduling via learned weights. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 1859–1877. SIAM, 2020. doi: 10.1137/1.9781611975994.114. URL https://doi.org/10.1137/1.9781611975994.114.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. URL https://tor-lattimore.com/downloads/book/book.pdf.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010.

Thodoris Lykouris and Sergei Vassilvitskii. Competitive caching with machine learned advice. *J. ACM*, 68(4):24:1–24:25, July 2021. ISSN 0004-5411. doi: 10.1145/3447579. URL https://doi.org/10.1145/3447579.

Davide Maran, Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Castiglioni, Nicola Gatti, and Marcello Restelli. Bandits with ranking feedback. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 80567–80608. Curran Associates, Inc., 2024. doi: 10.52202/079017-2562. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/936ce22b767cf1a1496083e4725d3b21-Paper-Conference.pdf.

Michael Mitzenmacher and Sergei Vassilvitskii. Algorithms with predictions. *Communications of the ACM*, 65(7):33–35, 2022.

Manish Purohit, Zoya Svitkina, and Ravi Kumar. Improving online algorithms via ML predictions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/73a427badebe0e32caa2e1fc7530b7f3-Paper.pdf.

Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, ACM International Conference Proceeding Series, pages 784–791. ACM, 2008. doi: 10.1145/1390156.1390255. URL https://doi.org/10.1145/1390156.1390255.

Matteo Russo, Andrea Celli, Riccardo Colini-Baldeschi, Federico Fusco, Daniel Haimovich, Dima Karamshuk, Stefano Leonardi, and Niek Tax. Online learning with sublinear best-action queries. *Advances in Neural Information Processing Systems*, 37:40407–40433, 2024.

Aadirupa Saha and Aditya Gopalan. PAC battling bandits in the Plackett-Luce model. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 700–737. PMLR, 22–24 Mar 2019. URL https://proceedings.mlr.press/v98/saha19a.html.

Sofía S. Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical Science*, 30(2):199–215, 2015.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1201–1208, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553527. URL https://doi.org/10.1145/1553374.1553527.

Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning*, ICML'11, pages 241–248, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78(5):1538–1556, September 2012. ISSN 0022-0000. doi: 10.1016/j.jcss.2011.12.028. URL https://doi.org/10.1016/j.jcss.2011.12.028.

Masrour Zoghi, Shimon Whiteson, Rémi Munos, and Maarten de Rijke. Relative upper confidence bound for the K-armed dueling bandit problem. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32(2) of *Proceedings of Machine Learning Research*, pages 10–18, Beijing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/zoghi14.html.

Masrour Zoghi, Zohar S. Karnin, Shimon Whiteson, and Maarten de Rijke. Copeland dueling bandits. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/9872ed9fc22fc182d371c3e9ed316094-Paper.pdf.

Appendix

Contents

1 Introduction **1**

1.1 Related Works 2

2 Multi-Armed Bandits With Best-Action Queries **3**

3 When Best-Action Queries Do Not Help: The Stochastic Correlated Setting **4**

4 Best-Action Queries in the Stochastic Independent Setting **5**

4.1 The Gap Between Dynamic and Static Policies and Its Relation to Variance 5

4.2 A Variance-Aware Regret Minimizer 7

4.3 The Final Regret Bound 8

4.4 A Matching Lower Bound 9

A Omitted Proofs of Section 3 **13**

B Omitted Proofs of Section 4 **16**

B.1 Lower Bound 16

A Omitted Proofs of Section 3

Theorem 3.1. *For every $T \geq 1$ and $k \in [T-1]$, and every randomized learning algorithm $\mathfrak{A}$ that uses at most k best-action queries, there exists a MAB problem instance with two arms and stochastic correlated rewards such that the regret of $\mathfrak{A}$ satisfies:*

$$R_{T,k}(\mathfrak{A}) \geq \Omega(\sqrt{T-k}).$$

Proof. Fix an arbitrary randomized algorithm $\mathfrak{A}$ that uses at most k best-action queries, and construct two instances on at least one of which its expected regret is $\Omega(\sqrt{T-k})$. All expectations include the algorithm's internal randomness. Let $m := T - k$. Fix parameters:

$$0 < a < \frac{1}{4}, \qquad 0 < \eta < \frac{1}{4} - a, \qquad b := a + \eta.$$

For every $c \in [0, 1/2]$, we define:

$$H_c(x) := x - cx(1-x), \qquad x \in [0,1].$$

Since,

$$H_c'(x) = 1 + c(2x-1) \geq \frac{1}{2},$$

the map H_c is strictly increasing and therefore invertible on $[0,1]$. Let $(U_t)_{t=1}^T$ be i.i.d. $\mathrm{Unif}[0,1]$, let $(W_t)_{t=1}^T$ be i.i.d. $\mathrm{Unif}\{1,2\}$, with both sequences mutually independent and independent of the learner's randomness, and define:

$$Y_t := U_t, \qquad Z_t^- := H_a^{-1}(U_t), \qquad Z_t^+ := H_b^{-1}(U_t).$$

Because $a < b$, we have $H_b(x) \leq H_a(x) \leq x$ for all $x \in [0,1]$, and therefore $Y_t \leq Z_t^- \leq Z_t^+$. We now define two stochastic correlated instances ν_1, ν_2:

$$\nu_1: \quad (X_{t,1}, X_{t,2}) = \begin{cases} (Z_t^+, Y_t), & \text{if } W_t = 1, \\ (Z_t^-, Z_t^+), & \text{if } W_t = 2, \end{cases}$$

and,

$$\nu_2: \quad (X_{t,1}, X_{t,2}) = \begin{cases} (Z_t^+, Z_t^-), & \text{if } W_t = 1, \\ (Y_t, Z_t^+), & \text{if } W_t = 2. \end{cases}$$

In both instances, the arm indexed by W_t is almost surely the unique best arm at round t. For every $c \in [0, 1/2]$ and every $y \in [0, 1]$,

$$\mathbb{P}\big(H_c^{-1}(U_t) \le y\big) = \mathbb{P}(U_t \le H_c(y)) = H_c(y),$$

hence $H_c^{-1}(U_t)$ has density:

$$f_c(x) = H_c'(x) = 1 + c(2x - 1), \qquad x \in [0, 1].$$

Therefore,

$$\mathbb{E}[H_c^{-1}(U_t)] = \int_0^1 x\big(1 + c(2x-1)\big)\, dx = \frac{1}{2} + \frac{c}{6}.$$

In particular, $\mathbb{E}[Y_t] = 1/2$, $\mathbb{E}[Z_t^-] = 1/2 + a/6$, $\mathbb{E}[Z_t^+] = 1/2 + b/6$. Under ν_1, arm 1 has mean:

$$\mu_1^{(1)} = \frac{1}{2}\mathbb{E}[Z_t^+] + \frac{1}{2}\mathbb{E}[Z_t^-] = \frac{1}{2} + \frac{a+b}{12},$$

whereas arm 2 has mean:

$$\mu_2^{(1)} = \frac{1}{2}\mathbb{E}[Y_t] + \frac{1}{2}\mathbb{E}[Z_t^+] = \frac{1}{2} + \frac{b}{12}.$$

Thus, under ν_1, the best fixed arm is arm 1, with gap $\Delta = \mu_1^{(1)} - \mu_2^{(1)} = \frac{a}{12}$. By symmetry, under ν_2 the best fixed arm is arm 2, with the same gap $\Delta = a/12$.

Let $R_T^{(1)}(\mathfrak{A})$ and $R_T^{(2)}(\mathfrak{A})$ be the expected regrets of $\mathfrak{A}$ under ν_1 and ν_2, respectively.[5] On a queried round, the oracle reveals the best arm index W_t, so the learner obtains reward Z_t^+. Since the best fixed arm has mean $1/2 + (a+b)/12$, the expected regret of a queried round is

$$\left(\frac{1}{2} + \frac{a+b}{12}\right) - \left(\frac{1}{2} + \frac{b}{6}\right) = -\frac{b-a}{12} = -\frac{\eta}{12}.$$

Moreover, the queried-round feedback is exactly (W_t, Z_t^+) under both ν_1 and ν_2, so queried rounds give no information for distinguishing the two instances.

Let N_1 and N_2 be the numbers of times the learner pulls arms 1 and 2, respectively, over the first $m = T - k$ non-queried rounds, which exist because at most k queries are issued. Let $\tau \le T$ be the round of the m-th non-query action. Then $N_1 + N_2 = m$. On a non-queried round, pulling the wrong fixed arm incurs conditional expected regret exactly $\Delta = a/12$. Discarding the nonnegative expected regret of non-query rounds after τ and using the query budget, we obtain:

$$R_T^{(1)}(\mathfrak{A}) \ge \Delta\, \mathbb{E}_1[N_2] - \frac{\eta k}{12}, \qquad R_T^{(2)}(\mathfrak{A}) \ge \Delta\, \mathbb{E}_2[N_1] - \frac{\eta k}{12},$$

where $\mathbb{E}_\sigma$ denotes expectation under ν_σ. Now define the event $E := \{N_1 > m/2\}$. Since E^c implies $N_2 \ge m/2$, we obtain:

$$\mathbb{E}_1[N_2] \ge \frac{m}{2}\mathbb{P}_1(E^c), \qquad \mathbb{E}_2[N_1] \ge \frac{m}{2}\mathbb{P}_2(E).$$

Therefore,

$$R_T^{(1)}(\mathfrak{A}) + R_T^{(2)}(\mathfrak{A}) \ge \frac{\Delta m}{2}\Big(\mathbb{P}_1(E^c) + \mathbb{P}_2(E)\Big) - \frac{\eta k}{6}.$$

By the Bretagnolle–Huber inequality,

$$\mathbb{P}_1(E^c) + \mathbb{P}_2(E) \ge \frac{1}{2}\exp\big(-D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2)\big),$$

where $\mathbb{P}_1, \mathbb{P}_2$ are the laws of the feedback histories under ν_1, ν_2 , including the learner's random seed and decisions, stopped at τ. Hence,

$$R_T^{(1)}(\mathfrak{A}) + R_T^{(2)}(\mathfrak{A}) \ge \frac{\Delta m}{4}\exp\big(-D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2)\big) - \frac{\eta k}{6}. \tag{6}$$

[5] We drop the dependence on k from the definition of regret to ease the notation.

It remains to bound the KL divergence. On a non-queried round, if the learner plays arm 1, then the observation law is:

$$p_+(x) = \frac{1}{2} f_b(x) + \frac{1}{2} f_a(x) = 1 + \frac{a+b}{2}(2x-1)$$

under ν_1, and,

$$p_-(x) = \frac{1}{2} \cdot 1 + \frac{1}{2} f_b(x) = 1 + \frac{b}{2}(2x-1)$$

under ν_2. If the learner plays arm 2, the two laws are swapped. Since $a + b < 1/2$, both densities are uniformly lower bounded by $3/4$ on $[0, 1]$. Moreover,

$$p_+(x) - p_-(x) = a\left(x - \frac{1}{2}\right).$$

Using $\log u \le u - 1$, we get:

$$D_{\mathrm{KL}}(p_+ \| p_-) \le \int_0^1 \frac{(p_+(x) - p_-(x))^2}{p_-(x)}\, dx \le \frac{4}{3}a^2 \int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \frac{a^2}{9}.$$

By the same argument,

$$D_{\mathrm{KL}}(p_- \| p_+) \le \frac{a^2}{9}.$$

Since queried rounds contribute zero KL and there are exactly m non-queried rounds up to and including τ, the chain rule for histories stopped at τ gives:

$$D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2) \le m \max\{D_{\mathrm{KL}}(p_+ \| p_-), D_{\mathrm{KL}}(p_- \| p_+)\} \le \frac{ma^2}{9}. \tag{7}$$

Now, we define:

$$a := \min\left\{\frac{1}{8}, \frac{1}{\sqrt{m}}\right\}, \qquad \eta := \frac{a}{24(k+1)}, \qquad b := a + \eta.$$

Then $0 < a < 1/4$, $\eta > 0$, and $a + b = 2a + \eta < 1/2$, so all previous bounds apply. Also,

$$\frac{ma^2}{9} \le \frac{1}{9}, \qquad \frac{\eta k}{6} = \frac{ak}{144(k+1)} \le \frac{a}{144}.$$

Plugging these estimates and $\Delta = a/12$ into Equation (6), we obtain:

$$R_T^{(1)}(\mathfrak{A}) + R_T^{(2)}(\mathfrak{A}) \ge \frac{am}{48} e^{-1/9} - \frac{a}{144}. \tag{8}$$

Now recall that $a = \min\{1/8, 1/\sqrt{m}\}$. For every $m \ge 1$, we have $am \ge \sqrt{m}/8$ and $a \le 1/8$. Therefore, using Equation (8), we have:

$$R_T^{(1)}(\mathfrak{A}) + R_T^{(2)}(\mathfrak{A}) \ge \frac{e^{-1/9}}{384}\sqrt{m} - \frac{1}{1152}.$$

Since $m \ge 1$ and $e^{-1/9} \ge 8/9$, it follows that:

$$R_T^{(1)}(\mathfrak{A}) + R_T^{(2)}(\mathfrak{A}) \ge \left(\frac{1}{432} - \frac{1}{1152}\right)\sqrt{m} = \frac{5}{3456}\sqrt{m}.$$

Recalling that $m = T - k$, we conclude that:

$$R_T^{(1)}(\mathfrak{A}) + R_T^{(2)}(\mathfrak{A}) \ge \frac{5}{3456}\sqrt{T-k}.$$

Therefore, under the uniform prior on $\{\nu_1, \nu_2\}$,

$$\frac{R_T^{(1)}(\mathfrak{A}) + R_T^{(2)}(\mathfrak{A})}{2} \ge \frac{5}{6912}\sqrt{T-k}.$$

Consequently, there exists a stochastic correlated two-armed instance on which $\mathfrak{A}$ suffers regret:

$$R_{T,k}(\mathfrak{A}) \ge \Omega(\sqrt{T-k}),$$

as claimed. □

B Omitted Proofs of Section 4

Lemma 4.1. *Given a MAB instance with stochastic independent rewards, the following holds:*

$$\mathrm{OPT_d} - \mathrm{OPT_s} \geq \frac{1}{2n} \sum_{i \in \mathcal{S}} \left(\sigma_i^2 - \Delta_i\right)_+ .$$

Proof. Since $\max_{j \in [n]} X_j \geq \max\{X_{i^\star}, X_i\}$ for every $i \in [n]$, averaging over $i \in [n]$ yields:

$$\max_{j \in [n]} X_j \geq \frac{1}{n} \sum_{i=1}^{n} \max\{X_{i^\star}, X_i\}. \tag{9}$$

Then, we have:

$$\begin{aligned}
\mathrm{OPT_d} - \mathrm{OPT_s} = \mathbb{E}\left[\max_{j \in [n]} X_j\right] - \mu_{i^\star} &\geq \frac{1}{n} \sum_{i=1}^{n} \Big(\mathbb{E}[\max\{X_{i^\star}, X_i\}] - \mu_{i^\star}\Big) \\
&= \frac{1}{n} \sum_{i \neq i^\star} \Big(\mathbb{E}[\max\{X_{i^\star}, X_i\}] - \mu_{i^\star}\Big) \\
&\geq \frac{1}{2n} \sum_{i \neq i^\star} \Big(\mathbb{E}[|X_{i^\star} - X_i|] - \Delta_i\Big)_+ .
\end{aligned} \tag{10}$$

The first inequality above holds by taking expectations and subtracting $\mu_{i^\star}$ from both sides of Equation (9). The first equality holds since $\mathbb{E}[\max\{X_{i^\star}, X_{i^\star}\}] = \mu_{i^\star}$. The second inequality holds since $\max\{X_{i^\star}, X_i\} \geq X_{i^\star}$ point-wise, so $\mathbb{E}[\max\{X_{i^\star}, X_i\}] - \mu_{i^\star} \geq 0$ and using:

$$\max\{a, b\} = \frac{a+b}{2} + \frac{|a-b|}{2}.$$

Since $X_{i^\star}, X_i \in [0,1]$, we have $|X_{i^\star} - X_i| \in [0,1]$. Thus, using independence across arms, we get:

$$\mathbb{E}[|X_{i^\star} - X_i|] \geq \mathbb{E}[(X_{i^\star} - X_i)^2] = \mathrm{Var}(X_{i^\star} - X_i) + \big(\mathbb{E}[X_{i^\star} - X_i]\big)^2 \geq \sigma_{i^\star}^2 + \sigma_i^2 \geq \sigma_i^2. \tag{11}$$

Plugging Equation (11) into Equation (10), we get:

$$\mathrm{OPT_d} - \mathrm{OPT_s} \geq \frac{1}{2n} \sum_{i \neq i^\star} (\sigma_i^2 - \Delta_i)_+ \geq \frac{1}{2n} \sum_{i \in \mathcal{S}} (\sigma_i^2 - \Delta_i)_+,$$

concluding the proof. □

B.1 Lower Bound

Theorem 4.6. *There exists an absolute constant C such that, for $2 \leq n \leq T$ and any randomized algorithm $\mathfrak{A}$ which issues at most k best-action queries with $1 \leq k \leq T/C$, there exists a stochastic independent instance for which:*

$$R_{T,k}(\mathfrak{A}) \geq \Omega\left(\min\left\{\frac{Tn}{k}, \sqrt{n(T-k)}\right\}\right).$$

Proof. Take $C = 100$ and let $M := T - k$. Since $k \leq T/100$, we have $M \geq 99T/100$ and, in particular, $M \geq 99k$. Fix an arbitrary randomized algorithm $\mathfrak{A}$. We construct an instance on which its expected regret is large, with all expectations also taken over its internal randomness.

Set $K := \max\{k, \sqrt{nT}\}$. We show a lower bound of order Tn/K. Since $M = \Theta(T)$, it is enough to prove a lower bound of order Mn/K. Without loss of generality, index the arms according to the fixed tie-breaking order, so that lower-index arms have higher priority. Let J be the set of the $\lfloor n/2 \rfloor$ arms with lowest tie-breaking priority. Since the learner pulls one arm in each non-query round, $\sum_{i \in J} \mathbb{E}_1[N_i] \leq T$. Therefore, there exists an arm $j \in J$ such that $\mathbb{E}_1[N_j] \leq 3T/n$. Here N_i denotes the number of times arm i is pulled in the non-query rounds, and $\mathbb{E}_1$ is expectation under the first instance defined below.

We set the following parameters:

$$\Delta := \frac{n}{1000K}, \qquad \epsilon := \frac{M\Delta}{50K}, \qquad p := 2\Delta + \epsilon.$$

Since $K \geq \sqrt{nT}$ and $2 \leq n \leq T$, the above parameters satisfy $0 < \Delta \leq 1/1000$, $0 < \epsilon \leq 1/50000$, $p \leq 1/4$, and $1 - p + 2\Delta < 1$.

Consider the base instance ν_1 in which arm 1 has reward distribution $\mathrm{Be}(1 - p + \Delta)$, while every arm $i \neq 1$ has reward distribution $\mathrm{Be}(1 - p)$. Thus arm 1 is the unique optimal arm in ν_1, with gap Δ. Now define a second instance ν_2, identical to ν_1 except that arm j has reward distribution $\mathrm{Be}(1 - p + 2\Delta)$. Thus arm j is the unique optimal arm in ν_2, again with gap Δ over arm 1.

Let $R_T^{(1)}(\mathfrak{A})$ and $R_T^{(2)}(\mathfrak{A})$ be the regret of $\mathfrak{A}$ under ν_1 and ν_2, respectively. Let $\mathbb{P}_1, \mathbb{P}_2$ denote the laws of the complete interaction histories, including the algorithm's internal randomness. On query rounds, the learner receives the dynamic optimum. Under ν_1, the static optimum is $1 - p + \Delta$, so each query contributes regret at least $-(p - \Delta)$. Under ν_2, the static optimum is $1 - p + 2\Delta$, so each query contributes regret at least $-(p - 2\Delta)$.

Since the algorithm issues at most k best-action queries, it plays at least $M = T - k$ rounds without querying the oracle. Let $E := \{N_1 \leq M/2\}$. On ν_1, if E occurs, the learner pulls suboptimal arms at least $M/2$ times in the non-query rounds. Hence:

$$R_T^{(1)}(\mathfrak{A}) \geq -k(p - \Delta) + \frac{M\Delta}{2}\mathbb{P}_1(E). \tag{12}$$

On ν_2, if E^c occurs, the learner pulls arm 1 at least $M/2$ times, and arm 1 is suboptimal by gap Δ. Thus,

$$R_T^{(2)}(\mathfrak{A}) \geq -k(p - 2\Delta) + \frac{M\Delta}{2}\mathbb{P}_2(E^c). \tag{13}$$

Summing Equations (12)–(13) and applying Bretagnolle–Huber, we get:

$$\begin{aligned} R_T^{(1)}(\mathfrak{A}) + R_T^{(2)}(\mathfrak{A}) &\geq -k(\Delta + 2\epsilon) + \frac{M\Delta}{2}\big(\mathbb{P}_1(E) + \mathbb{P}_2(E^c)\big) \\ &\geq -k(\Delta + 2\epsilon) + \frac{M\Delta}{4}\exp\big(-D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2)\big). \end{aligned} \tag{14}$$

It remains to bound the KL divergence between the two feedback histories. Let $N_Q := \sum_{t=1}^{T} Q_t \leq k$ be the number of queries. By the chain rule for KL divergence, since the algorithm uses the same randomized decision rules conditional on the observed history under both instances, we have:

$$D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2) = \mathbb{E}_1[N_Q] D_{\mathrm{KL}}(Q_1 \| Q_2) + \mathbb{E}_1[N_j]\, D_{\mathrm{KL}}\big(\mathrm{Be}(1 - p), \mathrm{Be}(1 - p + 2\Delta)\big),$$

where Q_1, Q_2 are the distributions of the feedback received in one query round under ν_1, ν_2. Thus, the query term above is at most $kD_{\mathrm{KL}}(Q_1 \| Q_2)$.

First, for the bandit feedback term, since:

$$D_{\mathrm{KL}}(\mathrm{Be}(1 - p), \mathrm{Be}(1 - p + 2\Delta)) = D_{\mathrm{KL}}(\mathrm{Be}(p), \mathrm{Be}(p - 2\Delta)) = D_{\mathrm{KL}}(\mathrm{Be}(p), \mathrm{Be}(\epsilon)),$$

and $p - \epsilon = 2\Delta$, the standard Bernoulli KL bound gives:

$$D_{\mathrm{KL}}(\mathrm{Be}(p), \mathrm{Be}(\epsilon)) \leq \frac{8\Delta^2}{\epsilon} = \frac{400K\Delta}{M} = \frac{2}{5}\frac{n}{M}.$$

Here we used $\epsilon = M\Delta/(50K)$. Therefore, it holds

$$\mathbb{E}_1[N_j]\, D_{\mathrm{KL}}\big(\mathrm{Be}(1 - p), \mathrm{Be}(1 - p + 2\Delta)\big) \leq \frac{3T}{n} \cdot \frac{2}{5}\frac{n}{M} = \frac{6T}{5M} \leq \frac{40}{33}. \tag{15}$$

We now bound the query term. Since j belongs to the lower half of the tie-breaking order, there are at least $n/2 - 1$ arms with higher priority than j. If at least one of these arms obtains reward 1, then the oracle never returns j, independently of whether the instance is ν_1 or ν_2. Therefore, the query feedback can differ under ν_1 and ν_2 only on the event that all these higher-priority arms obtain reward 0, while on its complement the conditional query-feedback laws coincide. This event, which we denote by G, has the same probability in both instances, at most $p^{n/2-1}$.

Conditioned on G, the only remaining difference between ν_1 and ν_2 is the distribution of arm j, which changes from $\mathrm{Be}(1-p)$ to $\mathrm{Be}(1-p+2\Delta)$. Thus,

$$\begin{aligned} D_{\mathrm{KL}}\big(Q_1(\cdot \mid G) \| Q_2(\cdot \mid G)\big) &\le D_{\mathrm{KL}}(P_1^G \| P_2^G) \\ &= D_{\mathrm{KL}}\big(\mathrm{Be}(1-p), \mathrm{Be}(1-p+2\Delta)\big). \end{aligned}$$

Here, P_1^G and P_2^G are the distributions of the full reward vector conditioned on G under the two instances. The inequality follows from data processing, since the query feedback is the same deterministic function of this vector. The equality follows from independence. Indeed, conditioning on G leaves the reward of arm j unchanged and independent of the remaining rewards, whose conditional joint distribution is identical under both instances. Let $\alpha := \mathbb{P}_1(G) = \mathbb{P}_2(G)$. By the law of total probability, for $s \in \{1, 2\}$,

$$Q_s = \alpha Q_s(\cdot \mid G) + (1-\alpha) Q_s(\cdot \mid G^c).$$

Joint convexity of KL divergence and the equality of the conditional query-feedback laws on G^c give

$$\begin{aligned} D_{\mathrm{KL}}(Q_1 \| Q_2) \le \alpha D_{\mathrm{KL}}\big(Q_1(\cdot \mid G) \| Q_2(\cdot \mid G)\big) \\ + (1-\alpha) \underbrace{D_{\mathrm{KL}}\big(Q_1(\cdot \mid G^c) \| Q_2(\cdot \mid G^c)\big)}_{=0}. \end{aligned}$$

Using $\alpha \le p^{n/2-1}$ and the preceding conditional bound, we get:

$$D_{\mathrm{KL}}(Q_1 \| Q_2) \le p^{n/2-1} D_{\mathrm{KL}}\big(Q_1(\cdot \mid G) \| Q_2(\cdot \mid G)\big) \le p^{n/2-1}\, D_{\mathrm{KL}}\big(\mathrm{Be}(1-p), \mathrm{Be}(1-p+2\Delta)\big).$$

Using the previous KL bound,

$$D_{\mathrm{KL}}(Q_1 \| Q_2) \le p^{n/2-1} \frac{2}{5} \frac{n}{M}.$$

Thus, since $M \ge 99k$ and $p \le 1/4$, we get the final bound:

$$k D_{\mathrm{KL}}(Q_1 \| Q_2) \le \frac{2n}{5} \frac{k}{M} p^{n/2-1} \le \frac{2n}{495} p^{n/2-1} \le \frac{1}{100}, \tag{16}$$

where the last inequality uses $np^{n/2-1} \le 2$ for all $n \ge 2$ and $p \le 1/4$. Combining Equation (15) and Equation (16), we get:

$$D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2) \le \frac{40}{33} + \frac{1}{100} \le \frac{5}{4}.$$

Therefore, $\exp\big(-D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2)\big) \ge e^{-5/4}$. Plugging this into Equation (14) gives:

$$\begin{aligned} R_T^{(1)}(\mathfrak{A}) + R_T^{(2)}(\mathfrak{A}) &\ge -k(\Delta + 2\epsilon) + \frac{M\Delta}{4e^{5/4}} \\ &\ge -k\Delta - \frac{M\Delta}{25} + \frac{M\Delta}{4e^{5/4}} \\ &\ge -\frac{M\Delta}{99} - \frac{M\Delta}{25} + \frac{M\Delta}{4e^{5/4}} \\ &\ge \frac{M\Delta}{50}, \end{aligned}$$

where we used $M \ge 99k$ and $k \le K$. Since $\Delta = n/(1000K)$ and $M = \Theta(T)$, we obtain:

$$R_T^{(1)}(\mathfrak{A}) + R_T^{(2)}(\mathfrak{A}) = \Omega\left(\frac{Tn}{K}\right).$$

Thus, there exists one of the two instances on which $R_{T,k}(\mathfrak{A}) \ge \frac{M\Delta}{100} = \Omega\left(\frac{Tn}{K}\right)$. Since $K = \max\{k, \sqrt{nT}\}$ and $M = \Theta(T)$, it holds:

$$R_{T,k}(\mathfrak{A}) \ge \Omega\left(\min\left\{\frac{Tn}{k}, \sqrt{n(T-k)}\right\}\right),$$

which concludes the proof. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state all the main contributions made by the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: All the assumptions are clearly stated in Section 2.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All the theoretical results clearly state their assumptions, while all their proofs are provided in Section 3, Section 4, and in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does *not* include experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed, since the work is mainly theoretical.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The paper does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.