Learning Periodic Strategies in Blocking Bandits Is as Hard as Bandits with Switching Costs

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Università degli Studi di Milano, Milan, Italy

Junya Honda HONDA@I.KYOTO-U.AC.JP
Kyoto University, Kyoto, Japan
RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

Yuko Kuroki YUKO.MIYAUCHI@INTESASANPAOLO.COM
Intesa Sanpaolo AI Research, Turin, Italy

Atsushi Miyauchi ATSUSHI.MIYAUCHI@INTESASANPAOLO.COM
Intesa Sanpaolo, Turin, Italy

Lukas Zierahn LUKAS.ZIERAHN@CWI.NL
Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands
Booking.com, Amsterdam, The Netherlands

Editors: Steve Hanneke and Tor Lattimore

Abstract

In blocking K-armed bandits, playing an arm renders it unavailable for a fixed number of future rounds. While this model is relatively well understood in the stochastic regime, much less is known when rewards are generated adversarially. Via a novel reduction, we first show that computing the total reward of the best dynamic policy is NP-hard, even when the blocking time $d > 1$ is identical across arms. We therefore turn to tractable comparators and study the class of d-periodic policies, proving that the optimal periodic policy is efficiently computable and always obtains at least a $\frac{1}{K}$ fraction of the dynamic optimum. We also show that this $\frac{1}{K}$ factor is information-theoretically tight: no algorithm can achieve sublinear α-regret with respect to the offline optimal dynamic policy for any $\alpha > \frac{1}{K}$. Our main result shows that $T^{2/3}$ is the minimax rate for the regret (against periodic policies) for adversarial blocking bandits with identical blocking times, and that this rate is achievable by an efficient algorithm. Our main technical contribution is the lower bound, which establishes that blocking bandits are at least as hard as bandits with switching costs. The matching upper bound instead follows from a reduction to combinatorial semi-bandits over bipartite matchings. Finally, we show that $\sqrt{T}$ regret rates are efficiently achievable in the full information setting, and more generally via α-regret with $\alpha = \frac{1}{2}$.

Keywords: Regret minimization, combinatorial bandits, bipartite matchings.

1. Introduction

Multiarmed bandits are an established mathematical framework for the study of sequential decision-making problems (Lattimore and Szepesvári, 2020). A variety of adaptations of the basic K-armed bandit model have been proposed to address different features of real-world applications. In this work we focus on blocking bandits (Basu et al., 2019), a simple variant of the K-armed model where selecting an arm makes it unavailable to the player for a number of future timesteps. This constraint arises in concrete decision-making scenarios. For example, when pulling an arm corresponds to allocating an incoming task to one of K servers, a server may not accept new tasks until

the current one is completed (a similar phenomenon occurs in crowdsourcing, where workers can refuse new tasks when busy terminating the one previously assigned). Further examples include sequential experimental design or automated laboratories, where experiments occupy instruments, and financial settings, where trades and position changes temporarily consume inventory, margin, or risk budget, limiting how quickly the same action can be repeated.

Basu et al. (2019) established the computational hardness (under complexity-theoretic assumptions) of maximizing the reward over T rounds when each arm has a constant reward and a constant blocking time (possibly different for each arm). The hardness arises because computing the optimal dynamic policy (whose total reward is conventionally denoted by OPT) reduces to an intractable scheduling problem.

While blocking durations can in principle depend on the chosen arm, operational constraints that are imposed at the system level are naturally modeled by a single global cool-down parameter d shared across arms (e.g., standardized reset windows in experimental platforms, or frequency-capping rules in recommendation/advertising). Moreover, for such systems, rewards may be non-stationary—due to changing demand, drifting user preferences, or strategic effects—so the stochastic i.i.d. assumption becomes unrealistic. This motivates the study of blocking bandits in the harder adversarial regime, but under the simplifying assumption that the blocking time d is identical across arms. Identical blocking times in the stochastic regime are studied as a special case in Basu et al. (2019). They observe that the optimal dynamic policy becomes d-periodic, and that competing against the best periodic class is equivalent to competing against the best d-subset in combinatorial bandits with semi-bandit feedback (Audibert et al., 2014).

However, this benign structure is specific to stochastic rewards. In the adversarial setting, the offline dynamic optimum remains problematic in two distinct senses even when all blocking times are identical. First, computing the best dynamic policy is NP-hard. Second, the offline dynamic optimum is too strong as a comparator of online learning: for every $\alpha > \frac{1}{K}$, no algorithm can achieve sublinear α-regret with respect to OPT, even disregarding efficiency and restricting to oblivious adversaries. To focus on a tractable comparator, we thus consider the class of periodic strategies. We show that the optimal d-periodic policy is efficiently offline-computable (i.e., when all rewards are known in advance) and the best d-periodic policy always obtains at least a $\frac{1}{K}$ fraction of OPT, and that this factor is tight in the worst case.

Despite the simple structure of d-periodic policies, competing against the best such policy is not straightforward under the blocking constraint. A generic reduction to adversarial deterministic MDPs, using results of Dekel and Hazan (2013), gives a $\widetilde{\mathcal{O}}(T^{2/3})$ regret bound against the best periodic strategy. However, the resulting dependence on the problem parameters is exponential in the blocking time d (see Appendix A).

Our main contribution establishes that the rate $T^{2/3}$ is indeed minimax: irrespective of computational efficiency, any algorithm for blocking bandits must suffer a regret $\Omega(T^{2/3})$ against the best d-periodic strategy for any $d > 1$. In addition to that, we also prove that the rate $\widetilde{\mathcal{O}}(T^{2/3})$—matching the dependence on the time horizon T in the lower bound up to log factors—can be achieved efficiently, with a polynomial dependence on the relevant parameters K and d.

Note that our model is a special case of the one studied by Bishop et al. (2020), who allow both rewards and blocking durations to be generated adversarially for each arm and time step. As computing the optimal policy is NP-hard, they investigate upper bounds on the α-regret, where they consider an α that depends on the range of the blocking durations and on the total path variation of the rewards. In the adversarial regime, their α-regret also scales with the total path variation and

can be $\Theta(T)$ even for fixed d, making the choice of a meaningful comparator non-trivial. Choosing d-periodic comparators, as we do, ensures a near-optimal regret with an efficient algorithm and without any constraints on the adversary.

Due to space constraints, we defer a more detailed discussion of related work—including connections to recharging (Kleinberg and Immorlica, 2018) and sleeping bandits (Kleinberg et al., 2010)—to Appendix B.

Technical challenges. Our lower bound construction shows that blocking bandits are at least as hard as bandits with switching costs (BSC), which have a minimax regret of $\Theta(T^{2/3})$ (Dekel et al., 2014). The main idea is to force the blocking strategy to pay a large loss to explore while respecting the blocking constraint. We show that for any blocking time $1 < d < K$ and for any $d/2$ instances of BSC, each over two arms, there exists an instance of (K, d)-blocking bandits such that the total regret over the $d/2$ instances of BSC, including the switching costs, is bounded from above by the total regret over the blocking bandit instance. This immediately implies that any lower bound for BSC also applies to blocking bandits. Our construction assigns each BSC strategy two unique arms in $[K]$ (call these the legal arms for that strategy). Then we consider T/d blocks of d consecutive time steps, where each BSC instance is assigned two consecutive timesteps. If the blocking strategy plays a legal arm, then it will incur a corresponding loss from the BSC instance. On the other hand, if the blocking strategy plays an illegal arm, it will suffer a maximum loss. Crucially, due to the blocking constraint, any blocking strategy that switches arms must play an illegal arm at least once, thus simulating a switching cost. We then prove a lower bound for combinatorial bandits with semi-bandit feedback over (K, d)-bipartite matchings, and use it to strengthen our result by exploiting the equivalence between bipartite matchings and periodic strategies.

This equivalence is also at the core of our efficient algorithm for blocking bandits, which runs using a combinatorial bandit over bipartite matchings. However, to account for the blocking constraint, we implement the combinatorial bandit algorithm over batches of size $T^{1/3}$ following the standard template for adversaries with bounded memory (Arora et al., 2012). Whenever the matching changes across consecutive batches, we use the initial part of the new batch to make a transition that respects the blocking constraint. The length of this transition is a variable that depends on how much K is bigger than d.

Finally, to go beyond the $T^{2/3}$ barrier, we discuss how stronger feedback (full information) or relaxed benchmarks (α-regret)[1] can recover the $T^{1/2}$ regret rate. Our full information result relies on the low-switch algorithm for combinatorial bandits of Devroye et al. (2015), which is efficiently implementable over bipartite matchings.

2. Problem Setting

In the adversarial (K, d)-blocking bandit, a player faces a standard adversarial K-armed bandit with the additional constraint that the arm played at any time step t cannot be played again for the subsequent $d - 1$ steps, where $1 \leq d < K$. Formally, for any step t, if $I_t \in [K]$ is played at time t, then the blocking constraint requires that $I_{t+1} \neq I_t, \ldots, I_{t+d-1} \neq I_t$. Note that the learner must select *exactly one* arm from the set of available arms in every round (there is no "do-nothing"

1. Following the standard convention in the adversarial bandit literature, we define our setting in terms of loss minimization. This is equivalent to reward maximization for regret. However, we revert to reward maximization when considering α-regret, because multiplicative benchmarks become vacuous in loss form, when the optimal total loss can be arbitrarily close to zero.

action). Note also that $(K, 1)$-blocking bandits correspond to standard K-armed bandits, as the blocking constraint disappears for $d = 1$. We use $\ell_t(i) \in [0, 1]$ to denote the loss of arm $i \in [K]$ at time t and assume losses are generated by an oblivious adversary. A nonoblivious adversary can set all losses equal to 1 and then strategically place losses of 0 in actions the learner is not able to play due to the blocking constraint, leading to a scenario where the learner is deterministically not able to observe any information. We use $L_T(\pi)$ to denote a total loss $\sum_{t=1}^{T} \ell_t(I_t)$.

The comparator in our setting must obey the same blocking constraints as the player. However, as computing the best feasible dynamic comparator is NP-hard even with identical blocking times (Theorem 2 in Section 3), we are led to define regret with respect to the class $\mathcal{A}_d$ of d-periodic comparator sequences, on which the minimizer can be computed in polynomial time (Theorem 1 in Section 3). Note also that the d-periodic comparator reduces to the standard fixed-action comparator for $d = 1$.

Each $\boldsymbol{a} \in \mathcal{A}_d$ is a sequence $\boldsymbol{a} = (a_1, a_2, \ldots)$ where $a_t \in [K]$ and $a_{t+d} = a_t$ for all t. Moreover, $\boldsymbol{a}$ must satisfy the blocking constraint: for all t, arms $a_t, \ldots, a_{t+d-1}$ are all distinct. Formally, for any horizon T, the regret of a (possibly randomized) player's strategy π playing arm I_t at time t is defined by

$$R_T := \mathbb{E}\big[L_T(\pi)\big] - \min_{\boldsymbol{a} \in \mathcal{A}_d} L_T(\boldsymbol{a}) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{\boldsymbol{a} \in \mathcal{A}_d} \sum_{t=1}^{T} \ell_t(a_t)\,.$$

The *best d-periodic comparator sequence* for horizon T is any minimizer $\boldsymbol{a}_T^\star \in \arg\min_{\boldsymbol{a} \in \mathcal{A}_d} L_T(\boldsymbol{a})$.

3. Offline Optimization and Approximation Barriers

We first prove that the offline problem of minimizing the cumulative loss over the d-periodic comparator class is polynomial-time solvable.

Theorem 1 *Given an integer $1 \le d < K$ and a loss matrix $(\ell_t(k))_{t \in [T],\, k \in [K]}$ with $\ell_t(k) \in [0, 1]$, the d-periodic comparator sequence $\boldsymbol{a} \in \mathcal{A}_d$ that minimizes the total loss $L_T(\boldsymbol{a}) = \sum_{t=1}^{T} \ell_t(a_t)$ can be computed in polynomial time.*

Proof First, we transform the original problem into a maximization problem by converting losses $\ell_t(k)$ into rewards $1 - \ell_t(k)$. From the d-periodicity, in each period of length d every arm is either used exactly once at a fixed position or not used at all. Aggregating rewards over periods to define $w(k, s) = \sum_{t \le T : t \equiv s \,(\mathrm{mod}\ d)} (1 - \ell_t(k))$ for $k \in [K], s \in [d]$, the problem reduces to selecting exactly one arm at each of the d positions, which can be formulated as a maximum-weight matching problem on a bipartite graph with vertex sets $[K]$ (arms) and $[d]$ (time slots in a period). This is polynomial-time solvable—e.g., by the Hungarian method (Korte and Vygen, 2012). ■

Next, we show that computing the best dynamic policy with identical blocking times is NP-hard. Unlike previous results, which relied on reductions from 3-SAT or scheduling problems, ours is a novel reduction from a graph coloring problem called DISTANCE PRECOLORING EXTENSION (DPE), which was introduced and shown to be NP-complete in Das et al. (2025). In DPE, given a path graph $P = (V, E)$, where $V = \{v_1, \ldots, v_n\}$ and $E = \{\{v_i, v_{i+1}\} \mid i = 1, \ldots, n-1\}$, a set of colors $C = \{c_1, \ldots, c_k\}$, a nonnegative integer q, and a pre-coloring $\gamma' \colon S \to C$ for some $S \subseteq V$, the goal is to decide whether there exists a coloring $\gamma \colon V \to C$ such that γ extends γ' (i.e.,

$\gamma(v) = \gamma'(v)$ for any $v \in S$) and any pair of nodes at distance at most q on P receive different colors.

Theorem 2 *Let $\mathcal{F}_{K,d,T} \subseteq [K]^T$ denote the set of all length-T action sequences satisfying the blocking constraint. Given an integer $1 < d < K$ and a loss matrix $(\ell_t(k))_{t\in[T],\, k\in[K]}$ with $\ell_t(k) \in [0,1]$, computing the optimal (not necessarily d-periodic) comparator sequence $\boldsymbol{a} \in \mathcal{F}_{K,d,T}$ that minimizes the total loss $L_T(\boldsymbol{a}) = \sum_{t=1}^{T} \ell_t(a_t)$ is NP-hard.*

Proof We construct a polynomial-time reduction from DPE. Given an instance of DPE, we construct an instance of the problem of computing the optimal comparator $\boldsymbol{a} \in \mathcal{F}_{K,d,T}$ as follows. We set $T = |V|$, $K = |C|$, and $d = q + 1$. We also make the loss matrix $(\ell_t(k))_{t\in[T],\, k\in[K]}$ in the way that for $v_i \in S$ and $c_j \in C$, $\ell_i(j) = 0$ if $\gamma'(v_i) = c_j$ and $\ell_i(j) = 1$ otherwise, and for $v_i \in V \setminus S$ and $c_j \in C$, $\ell_i(j) = 0$. Then, based on the one-to-one correspondence between colors in C and arms in $[K]$ and the equivalence between the coloring constraint and the blocking constraint, we see that there exists a desired coloring γ for the DPE instance if and only if the total loss of the optimal comparator $\boldsymbol{a}$ is equal to zero. ■

Remark 3 *The above reduction also shows that the offline loss-minimization problem is NP-hard to approximate within any finite multiplicative factor. Indeed, the constructed instance has optimum loss zero if and only if the underlying DPE instance is feasible, so any finite-factor approximation algorithm would decide DPE feasibility. This offline approximation hardness is orthogonal to the online regret question addressed below.*

We next quantify the relationship between the best d-periodic policy and the best dynamic policy in the reward-maximization formulation. The best d-periodic policy always obtains at least a $1/K$ fraction of the dynamic optimum, and this factor is tight in the worst case.

Theorem 4 *Given an integer $1 < d < K$ and a payoff matrix $(g_t(k))_{t\in[T],\, k\in[K]}$ with $g_t(k) \in [0,1]$, let* OPT *be the total payoff $G_T(\boldsymbol{a}^\star) = \sum_{t=1}^{T} g_t(a_t^\star)$ of the optimal (not necessarily d-periodic) comparator sequence $\boldsymbol{a}^\star \in \mathcal{F}_{K,d,T}$. Then*

$$\max_{\boldsymbol{a}\in\mathcal{A}_d} G_T(\boldsymbol{a}) \geq \frac{\mathrm{OPT}}{K} .$$

Moreover, there exists a payoff matrix $(g_t(k))_{t\in[T],\, k\in[K]}$ such that

$$\max_{\boldsymbol{a}\in\mathcal{A}_d} G_T(\boldsymbol{a}) \leq \frac{\mathrm{OPT}}{K} .$$

Proof Let $\boldsymbol{a}^\star$ be the comparator sequence that achieves OPT. Draw a d-periodic strategy $\boldsymbol{a} \in \mathcal{A}_d$ uniformly at random. For any $t \in [T]$, $\mathbb{P}(a_t = a_t^\star) = \frac{1}{K}$ by symmetry. Therefore, $\mathbb{E}\big[g_t(a_t)\big] \geq \mathbb{P}(a_t = a_t^\star) g_t(a_t^\star) = \frac{g_t(a_t^\star)}{K}$, which implies the first statement because t was chosen arbitrarily.

To prove the second statement, take d, K, T such that $\gcd(K, d) = 1$ and Kd divides T, and consider the binary payoff matrix $(g_t(k))_{t\in[T],\, k\in[K]}$ in which arms get a non-zero reward in a round-robin fashion: $g_1(1) = 1$ and $g_1(i) = 0$ for all $i \neq 1$, $g_2(2) = 1$ and $g_2(i) = 0$ for all $i \neq 2$, and so on until the time step K when we start over. Formally, $g_t(i) = 1$ if and only if $i = ((t-1)$

$\bmod K) + 1$. Clearly, $\mathrm{OPT} = T$ on this instance. Now pick any d-periodic policy $\boldsymbol{a} \in \mathcal{A}_d$ and consider any position $r \in [d]$. Define the time steps $\mathcal{T}_r = \{r + nd : n \in \mathbb{N}\}$ when a_r is played. Because $\gcd(K, d) = 1$, the values in $\mathcal{T}_r$ modulo K cycle through all $\{0, \dots, K-1\}$ according to some permutation. As a_r is played exactly $\frac{T}{d}$ times, $g_t(a_r) = 1$ occurs exactly $\frac{T}{Kd}$ times (i.e., when $a_r = ((t-1) \bmod K) + 1$), which summed over the d positions gives a total payoff of $\frac{T}{K}$. ■

We show that no algorithm can guarantee a multiplicative factor strictly larger than $1/K$ against the dynamic optimum up to sublinear regret.

Theorem 5 *Fix integers $1 \le d < K$ and a horizon T divisible by d. For every possibly randomized learning algorithm π and every $\alpha > 1/K$, there exists a deterministic oblivious payoff matrix $(g_t(k))_{t\in[T],\, k\in[K]}$ with $g_t(k) \in \{0, 1\}$ such that*

$$\alpha\mathrm{OPT} - \mathbb{E}\left[\sum_{t=1}^{T} g_t(I_t)\right] \ge \left(\alpha - \frac{1}{K}\right)\frac{T}{d},$$

where $\mathrm{OPT} := \max_{\boldsymbol{a}\in\mathcal{F}_{K,d,T}} \sum_{t=1}^{T} g_t(a_t)$. Consequently, for fixed d, no algorithm can achieve sublinear α-regret with respect to the best dynamic policy for any $\alpha > 1/K$.

Proof Let $B = T/d$. For each block $b \in [B]$, draw A_b independently and uniformly from $[K]$, and set

$$g_t(i) = \begin{cases} \mathbb{I}\{i = A_b\}, & t = bd \text{ for some } b \in [B], \\ 0, & \text{otherwise.} \end{cases}$$

This defines an oblivious random payoff matrix.

For every realization of $(A_b)_{b=1}^{B}$, we have $\mathrm{OPT} = B$. Indeed, rewards are available only at the block endpoints, so $\mathrm{OPT} \le B$. Conversely, in each block b, a dynamic comparator can play A_b at time bd and fill the preceding $d-1$ rounds arbitrarily while avoiding both A_b and the currently blocked arms. At each such preceding round, at most d arms are forbidden, so feasibility follows from $K > d$.

Now fix any learning algorithm π. Since the only nonzero rewards occur at block endpoints,

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(I_t)\right] = \sum_{b=1}^{B} \mathbb{E}\big[\mathbb{I}\{I_{bd} = A_b\}\big].$$

For each b, the arm A_b is independent of the learner's action I_{bd} before observing g_{bd} and is uniform on $[K]$. Hence

$$\mathbb{E}\big[\mathbb{I}\{I_{bd} = A_b\}\big] = \frac{1}{K},$$

and therefore

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(I_t)\right] = \frac{B}{K}.$$

Thus

$$\alpha\mathrm{OPT} - \mathbb{E}\left[\sum_{t=1}^{T} g_t(I_t)\right] = \left(\alpha - \frac{1}{K}\right)B = \left(\alpha - \frac{1}{K}\right)\frac{T}{d}.$$

Finally, since the above equality holds in expectation over the random payoff matrix, there exists a deterministic realization of the payoff matrix for which the desired inequality holds. ■

Together with Theorem 4, this shows that the factor $1/K$ is the tight worst-case multiplicative threshold for comparison with the offline dynamic optimum. One may ask whether allowing comparators with a longer period $p > d$ improves the worst-case approximation to OPT. In general, this does not remove the $1/K$ barrier. For example, let $p = mK$ for some integer m. Partition timesteps into blocks of length p. Within each block, let the unique rewarding arm cycle as $1, 2, \ldots, K$ repeated m times, and shift this entire pattern by one arm from one block to the next. The dynamic comparator can follow the rewarding arm at every round, so $\mathrm{OPT} = T$. However, for each residue class modulo p, the rewarding arm cycles uniformly through all K arms across blocks. Any fixed p-periodic comparator therefore matches the rewarding arm only a $1/K$ fraction of the time, and obtains reward at most T/K. Thus even natural longer periods such as $p = mK$ do not improve the worst-case approximation factor. A complete characterization for arbitrary $p > d$ remains open.

4. Lower Bound

In this section we present two separate lower bounds. The first one is based on a reduction to bandits with switching costs and yields a lower bound of $\Omega(d^{1/3}T^{2/3})$. The second one is based on bipartite matchings and yields a bound of $\Omega(\sqrt{dKT})$. While the result shows that a dependence of order $\Omega(T^{2/3})$ is unavoidable, the second one hints at what the true dependence on other parameters might be, including that a dependence on K is likely necessary.

In bandits with switching costs (BSC) (Arora et al., 2012), the learner pays a unit cost whenever the action of the current timestep is different from the one in the previous timestep, modeling the cost a learner may incur for switching their financial strategy or by reconfiguring a system. In a K-armed BSC, the regret of a strategy playing arms $I_1, \ldots, I_T$ is measured against the best arm and defined by

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t) + \sum_{t=2}^{T} \mathbb{I}\{I_t \neq I_{t-1}\}\right] - \min_{a\in[K]} \sum_{t=1}^{T} \ell_t(a)\,.$$

A breakthrough result by Dekel et al. (2014) shows that the minimax regret of K-armed BSC is $\Theta\big(K^{1/3}T^{2/3}\big)$. To explore the mechanism we use to prove our first lower bound, we restrict ourselves to $K = 3$ actions and a blocking time of $d = 2$. The losses will be as follows:

$$\text{At } t \text{ odd: } \ell_t(a) = \begin{cases} \ell'_{\frac{t-1}{2}+1}(a), & \text{if } a_1 \text{ or } a_2, \\ 2, & \text{if } a_3 \end{cases} \quad \text{and at } t \text{ even: } \ell_t(a) = \begin{cases} 0, & \text{if } a_1 \text{ or } a_2, \\ 2, & \text{if } a_3\,. \end{cases}$$

We define $\ell'_t(a)$ as the losses of an arbitrary BSC with two actions, ran for $T/2$ timesteps. Due to the blocking constraint, if we played action a_1 on an odd timestep and we wish to play a_2 on future odd timesteps instead, we will have to play action a_3 at least once in between, which captures the switching cost. Thus, the cumulative loss plus the switching cost of the BSC lower bounds the loss of the blocking bandit. We now extend this intuition to arbitrary K in Theorem 6. In the rest of this section, we make our argument more transparent by augmenting the notation for regret as follows: $R_T(\mathcal{I}, \pi)$ denotes the regret of strategy π on problem instance $\mathcal{I}$ over the horizon T.

Theorem 6 *For any $K > d > 1$ where d is even and for any T such that d divides T, consider arbitrary BSC instances $\mathcal{I}_{\mathrm{sw}}(1), \dots, \mathcal{I}_{\mathrm{sw}}(d/2)$ of over two arms, $[0,1]$-valued losses, and horizon $T' = T/d$. Then there is an instance $\mathcal{I}_{\mathrm{bl}}$ of (K,d)-blocking bandits, $[0,2]$-valued losses, and horizon T, such that for any blocking bandits strategy π_{bl} there exist BSC strategies $\pi_{\mathrm{sw}}(1), \dots, \pi_{\mathrm{sw}}(d/2)$ satisfying*

$$\sum_{j=1}^{d/2} R_{T'}\big(\mathcal{I}_{\mathrm{sw}}(j), \pi_{\mathrm{sw}}(j)\big) \le R_T\big(\mathcal{I}_{\mathrm{bl}}, \pi_{\mathrm{bl}}\big) \,.$$

Proof We divide the timesteps into $T' = T/d$ blocks of length d each. Each BSC instance $\mathcal{I}_{\mathrm{sw}}(j)$ will be active once for each block s at time step $t(s,j) = (s-1)d + 2j - 1$. Each instance $\mathcal{I}_{\mathrm{sw}}(j)$ has action set $\mathcal{A}(j) = \{2j-1, 2j\}$ and we let $\ell'_s(j,i)$ be the loss of action $i \in \mathcal{A}(j)$ in instance $\mathcal{I}_{\mathrm{sw}}(j)$ at block $s \in [T']$. Then the loss of π_{bl} in block s is defined as follows:

$$\ell_t(i) = \begin{cases} \ell'_s(j,i) & \text{if } i \in \mathcal{A}(j) \text{ and } t = t(s,j), \\ 0 & \text{if } i \in \mathcal{A}(j) \text{ and } t = t(s,j)+1, \\ 2 & \text{otherwise.} \end{cases}$$

Let $B_t \in [K]$ be the action chosen by π_{bl} at any odd time $t = t(s,j)$. If possible, the BSC policy $\pi_{\mathrm{sw}}(j)$ will follow the action of the blocking strategy $I_s(j) = B_t$ but if $B_t \notin \mathcal{A}(j)$ then $I_s(j) = I_{s-1}(j)$. The cumulative loss of π_{bl} is then computed as follows:

$$\begin{aligned}
&\sum_{t=1}^{T} \ell_t(B_t) \\
&\quad = \sum_{j=1}^{d/2} \sum_{s=1}^{T'} \Big(\ell'_s\big(j, I_s(j)\big) \mathbb{I}\big\{B_{t(s,j)} \in \mathcal{A}(j)\big\} + 2\,\mathbb{I}\big\{B_{t(s,j)} \notin \mathcal{A}(j)\big\} + 2\,\mathbb{I}\big\{B_{t(s,j)+1} \notin \mathcal{A}(j)\big\} \Big) \\
&\quad \ge \sum_{j=1}^{d/2} \sum_{s=1}^{T'} \Big(\ell'_s\big(j, I_s(j)\big) + \mathbb{I}\big\{B_{t(s,j)} \notin \mathcal{A}(j)\big\} + \mathbb{I}\big\{B_{t(s,j)+1} \notin \mathcal{A}(j)\big\} \Big) \qquad (1) \\
&\quad \ge \sum_{j=1}^{d/2} \sum_{s=1}^{T'} \ell'_s\big(j, I_s(j)\big) + \sum_{j=1}^{d/2} \sum_{s=2}^{T'} \mathbb{I}\big\{I_{s-1}(j) \ne I_s(j)\big\} \qquad (2)
\end{aligned}$$

where in (1) we used $\mathbb{I}\big\{B_{t(s,j)} \notin \mathcal{A}(j)\big\} \ge \ell'_s\big(j, I_s(j)\big)\mathbb{I}\big\{B_{t(s,j)} \notin \mathcal{A}(j)\big\}$. For (2), consider $t = t(s,j)$ such that $I_s(j) \ne I_{s-1}(j)$. Then $B_t \in \mathcal{A}(j)$ and $B_t \ne B_{t-d}$, and so it must be that $B_{t-d} \notin \mathcal{A}(j)$ or $B_{t-d+1} \notin \mathcal{A}(j)$ because of the blocking constraint.

Without loss of generality, we assume that the best fixed action in each BSC instance $\mathcal{I}_{\mathrm{sw}}(j)$ is the first action so that $\sum_{s=1}^{T'} \ell'_s(j, 2j-1) \le \sum_{s=1}^{T'} \ell'_s(j, 2j)$. We then find the optimal d-periodic policy as $(a_1, a_2, \dots, a_{d-1}, a_d)$ and by construction, the sum of losses obtained by the optimal actions for the switching cases is equal to the losses incurred by this periodic policy

$$\sum_{j=1}^{d/2} \min_{i \in \mathcal{A}(j)} \sum_{s=1}^{T'} \ell'_s(j,i) = \min_{\boldsymbol{a} \in \mathcal{A}_d} \sum_{t=1}^{T} \ell_t(a_t) \,.$$

Therefore,

$$\begin{aligned}
&\sum_{j=1}^{d/2} R_{T'}\big(\mathcal{I}_{\mathrm{sw}}(j), \pi_{\mathrm{sw}}(j)\big) \\
&\quad = \sum_{j=1}^{d/2} \left(\sum_{s=1}^{T'} \ell'_s\big(j, I_s(j)\big) + \sum_{s=2}^{T'} \mathbb{I}\big\{I_{s-1}(j) \neq I_s(j)\big\} - \min_{i \in \mathcal{A}(j)} \sum_{s=1}^{T'} \ell'_s(j,i) \right) \\
&\quad \leq \sum_{t=1}^{T} \ell_t(B_t) - \min_{\boldsymbol{a} \in \mathcal{A}_d} \sum_{t=1}^{T} \ell_t(a_t) = R_T\big(\mathcal{I}_{\mathrm{bl}}, \pi_{\mathrm{bl}}\big)
\end{aligned}$$

concluding the proof. ■

Combining Theorem 6 with the minimax lower bound for two-armed BSC requires the classical Yao's argument. Let $\mathcal{D}_{\mathrm{sw}}$ be a hard distribution for two-armed BSC over horizon $T' = T/d$. Draw $I_{\mathrm{sw}}(1), \ldots, I_{\mathrm{sw}}(d/2)$ independently from $\mathcal{D}_{\mathrm{sw}}$, and construct I_{bl} as in Theorem 6. Then, for any deterministic blocking strategy π_{bl}, Theorem 6 and the BSC lower bound give

$$\mathbb{E}\left[R_T(I_{\mathrm{bl}}, \pi_{\mathrm{bl}})\right] \geq \sum_{j=1}^{d/2} \mathbb{E}\left[R_{T'}(I_{\mathrm{sw}}(j), \pi_{\mathrm{sw}}(j))\right] = \Omega\Big(d(T/d)^{2/3}\Big) = \Omega\Big(d^{1/3}T^{2/3}\Big)\,.$$

Yao's principle extends this to randomized blocking strategies. Rescaling the $[0,2]$-valued losses by $1/2$ gives the claimed lower bound for losses in $[0,1]$.

This gives our first lower bound.[2]

Corollary 7 *Pick K, d such that $K > d > 1$. In (K,d)-blocking bandits, any player strategy incurs a regret against the best d-periodic policy of at least $\Omega\big(d^{1/3}T^{2/3}\big)$.*

For the second lower bound, we consider the (K,d)-bipartite matching problem, with $d < K$ and where the action set is the set $\mathcal{M}$ of all injective mappings (matchings) $V : [d] \to [K]$. The adversary chooses losses $\ell_t(r,i) \in [0,1]$ for all $(r,i) \in [d] \times [K]$. The loss of a matching V at time t is

$$L_t(V) = \sum_{r \in [d]} \ell_t\big(r, V(r)\big)\,.$$

Bipartite matchings work for blocking bandits because a matching V corresponds to an action sequence of length d which respects the blocking constraint, and vice versa. Moreover, in blocking bandits at the end of each d-sequence of actions the player has observed $\big\{\ell_t\big(r, V(r)\big) : r \in [d]\big\}$, which corresponds to the semi-bandit feedback model. Our analysis first extends the standard lower bound $\Omega\big(d\sqrt{KT}\big)$ of Audibert et al. (2014)—originally proven for the set of all d-sized subsets of K base actions—to bipartite matchings, and then relates it back to blocking bandits. Note that this view allows any arbitrary matchings to be played consecutively without any switching cost in between, which is likely the reason why we cannot force the optimal $T^{2/3}$ dependence.

2. We can drop the requirement that d be even as the minimax regret of (K,d)-blocking bandits for $d > 1$ odd is at least as big as the minimax regret of $(K, d-1)$-blocking bandits, and this does not affect the asymptotic rate.

Theorem 8 *Pick K, d such that $K > d > 1$. In combinatorial bandits with semi-bandit feedback over (K, d)-bipartite matchings, any player strategy incurs a regret of at least $\Omega\big(d\sqrt{KT}\big)$.*

Proof Without loss of generality, we may assume $K > 2d$. If $K \le 2d$ then one can simply set the losses of edges $\lfloor d/2 \rfloor + 1, \dots, d$ to zero and thus the best possible regret for (K, d)-bipartite matching is not smaller than the one for $\big(K, \lfloor d/2 \rfloor\big)$. By Yao's principle, it suffices to exhibit a distribution over loss sequences such that every deterministic algorithm has large expected regret. Let $K' = K - d + 1$. Choose

$$\varepsilon = \frac{1}{4}\sqrt{\frac{K'}{T}} \le \frac{1}{4}.$$

We select the optimal matching $V^\star : [d] \to [K]$ at uniform random and assign i.i.d. losses

$$\ell_t(r, j) \sim \begin{cases} \text{Bernoulli}\big(\frac{1}{2} - \varepsilon\big), & j = V^\star(r), \\ \text{Bernoulli}\big(\frac{1}{2}\big), & j \neq V^\star(r). \end{cases}$$

Now, the expected regret of any deterministic learner that at time t plays V_t is

$$\mathbb{E}\big[R_T\big] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{r=1}^{d}\Big(\ell_t(r, V_t(r)) - \ell_t(r, V^\star(r))\Big)\right] \ge \varepsilon \sum_{r=1}^{d}\Big(T - \mathbb{E}\big[N_T\big(r, V^\star(r)\big)\big]\Big)$$

where $N_T(r, j) = \sum_{t=1}^{T} \mathbb{I}\big\{V_t(r) = j\big\}$. Now, fix an arbitrary deterministic player and some $r \in [d]$, and condition on the realization $\mathcal{F}_r$ of $\big\{V^\star(u) \,:\, u \neq r\big\}$ and of $\ell_t(r, j)$ for all $u \neq r$ and $j \in [K]$. Note that conditioned on $\mathcal{F}_r$, $V^\star(r)$ is uniform over the K' nodes that are not taken by $V^\star(u)$ for $u \in [d] \setminus \{r\}$. Similarly, the losses $\ell_t(r, j)$ for $t \in [T]$ and $j \in [K']$ are all i.i.d. Bernoulli of parameter $\frac{1}{2}$ except $\ell_t\big(r, V^\star(r)\big)$, which is i.i.d. Bernoulli of parameter $\frac{1}{2} - \varepsilon$. Since the player observes $\ell_t(r, V_t(r))$ after each round t, The player's behavior on r is that of a (deterministic) K'-armed bandit algorithm whose expected regret is $\varepsilon\big(T - \mathbb{E}\big[N_T\big(r, V^\star(r)\big) \,\big|\, \mathcal{F}_r\big]\big)$. We can then write

$$\begin{aligned}\mathbb{E}\big[R_T\big] &\ge \varepsilon \sum_{r=1}^{d}\Big(T - \mathbb{E}\big[N_T\big(r, V^\star(v)\big)\big]\Big) \\ &= \sum_{r=1}^{d}\mathbb{E}\Big[\varepsilon\Big(T - \mathbb{E}\big[N_T\big(r, V^\star(r)\big) \,\big|\, \mathcal{F}_r\big]\Big)\Big] = \Omega\big(d\sqrt{K'T}\big)\end{aligned}$$

where in the last transition we applied the standard $\Omega\big(\sqrt{K'T}\big)$ lower bound for K'-armed bandits with our choice of ε. Since $K \ge 2d$, $K' \ge K/2$ implying the desired lower bound $\Omega\big(d\sqrt{KT}\big)$ for the (K, d)-bipartite matching problem. ■

Equipped with this result, it is now easy to prove our second lower bound.

Corollary 9 *Pick K, d such that $K > d > 1$. In (K, d)-blocking bandits, any player strategy incurs a regret against the best d-periodic policy of at least $\Omega\big(\sqrt{dKT}\big)$.*

Proof To relate Theorem 8 to blocking bandits, assume d divides K. Define an instance $\mathcal{I}_{\text{bip}}$ of combinatorial bandits over bipartite matchings with left nodes $[d]$, right nodes $[K]$, and losses ℓ_τ for $\tau \in [T']$ and $T' = T/d$. We now construct the instance $\mathcal{I}_{\text{bl}}$ over T steps with losses ℓ'_t simply

Algorithm 1 Reduction from combinatorial semi-bandits

Input: Arms $[K]$, blocking time d, batch size B, combinatorial semi-bandit algorithm $\mathbb{A}$ over $\mathcal{M}$.

for $b = 1, 2, \ldots, N = \lfloor \frac{T}{dB} \rfloor$ **do**
  Obtain a matching $V_b \in \mathcal{M}$ from $\mathbb{A}$.
  if $b = 1$ ***or*** $V_b = V_{b-1}$ **then**
    Roll out the d-step schedule V_b for B times and observe the corresponding losses.
  else
    Compute $V^{(0)}, \ldots, V^{(n)} \in \mathcal{M}$ to transition from $V_{b-1} = V^{(0)}$ to $V_b = V^{(n)}$.
    Roll out each d-step schedule $V^{(1)}, \ldots, V^{(n)}$ and observe the corresponding losses.
    Roll out the d-step schedule V_b for the remaining $B - n$ segments and observe the corresponding losses.
  end
  Feed to $\mathbb{A}$ the observed losses aggregated over the corresponding segments.
end

by setting $\ell'_t(i) = \ell_\tau(v, i)$ for all $v \in [d]$, where $\tau = \lfloor t/d \rfloor + 1$. Clearly, if π_{bl} is a policy for $\mathcal{I}_{\mathrm{bl}}$ playing actions $I_t \in [K]$, then μ_τ with $\mu_\tau(v) = I_{(\tau-1)d+v}$ is a bipartite matching for all $\tau \in [T']$. Hence,

$$\sum_{t=1}^{T} \ell'_t(I_t) = \sum_{\tau=1}^{T'} \sum_{r\in[d]} \ell_\tau\big(r, I_{(\tau-1)d+r}\big) \quad \text{and} \quad \min_{a\in\mathcal{A}_d} \sum_{t=1}^{T} \ell'_t(a_t) = \min_{V\in\mathcal{M}} \sum_{\tau=1}^{T'} \sum_{r\in[d]} \ell_\tau\big(r, V(r)\big)\,.$$

This implies $R_{T'}\big(\mathcal{I}_{\mathrm{bip}}, \pi_{\mathrm{bip}}\big) \le R_T\big(\mathcal{I}_{\mathrm{bl}}, \pi_{\mathrm{bl}}\big)$ where π_{bip} is the policy for $\mathcal{I}_{\mathrm{bip}}$ that at time τ plays $V_t : [d] \to [K]$ such that $V_t(r) = I_{(\tau-1)d+r}$ for $r \in [d]$. By invoking Theorem 8, we obtain that the regret of (K, d)-blocking bandits is at least $\Omega\big(d\sqrt{K(T/d)}\big) = \Omega\big(\sqrt{dKT}\big)$. ■

5. Upper Bound via Reduction to Combinatorial Semi-Bandits

We establish the upper bound by reducing the (K, d)-blocking bandit problem to combinatorial bandits over matchings in bipartite graphs (with semi-bandit feedback). Once more, the core observation is the bijection between d-periodic strategies $\boldsymbol{a} \in \mathcal{A}_d$ and bipartite matchings $V \in \mathcal{M}$. The key difficulty is the blocking constraint, which prevents us from directly applying the upper bound for combinatorial bandits.

We partition the time horizon T into $N := \lfloor \frac{T}{dB} \rfloor$ meta-rounds, indexed by $b \in [N]$. Each meta-round consists of a batch of B segments, where each segment has length d. Within a batch b, we denote the segment index by $r \in [B]$ and the slot index within a segment by $s \in [d]$. These map to the original time step $\tau \in [T]$ via $t = \tau(b, r, s) := (b-1)dB + (r-1)d + s$. For simplicity, we assume that dB divides T as we can ignore the last incomplete batch, which contributes at most $\mathcal{O}(dB)$ additional loss. We denote by $\ell_t \in [0, 1]^K$ the loss vector at time t and by $\ell_t(i) \in [0, 1]$ the loss of arm i at time t.

Algorithm 1 details the reduction procedure. The learner interacts with a combinatorial semi-bandit algorithm that selects a matching $V_b \in \mathcal{M}$ at the start of each meta-round $b \in [N]$ (recall that $\mathcal{M}$ is the set of size-d matchings $V : [d] \to [K]$ in the bipartite graph). We view a matching $V \in \mathcal{M}$ as a d-step schedule of actions $V(1), \ldots, V(d) \in [K]$ for the blocking bandit algorithm.

When the combinatorial semi-bandit algorithm does not switch action between two adjacent batches, $V_b = V_{b-1}$, the learner plays V_b repeatedly for all B segments. When the matching changes, $V_b \neq V_{b-1}$, we insert a short transition phase of length at most n (i.e., dn time steps) to clear the blocking memory and align the schedule from V_{b-1} to V_b, where $n \leq \left\lceil \frac{d}{\min\{d,\, K-d\}} \right\rceil + 1$. In particular, the overhead becomes constant when K is sufficiently larger than d (e.g., $n \leq 2$ when $K \geq 2d$), and it degrades gracefully as K approaches d (e.g., $n \leq d+1$) when $K = d+1$. This transition ensures that the blocking constraints are satisfied globally, at the cost of a small overhead. The semi-bandit algorithm then receives the aggregated loss of the chosen matching V_b over the segments where it was played. The constructive proof of the next lemma is provided in Appendix C.1.

Lemma 10 (Transition between matchings) *Let $V', V \in \mathcal{M}$ be size-d matchings. Then there exists a sequence of $n+1$ matchings $\big(V^{(0)}, V^{(1)}, \ldots, V^{(n)}\big)$ such that $V' = V^{(0)}$, $V = V^{(n)}$, and the dn-step schedule $V^{(0)}, V^{(1)}, \ldots, V^{(n)}$ satisfies the blocking constraint. Moreover, we can take*

$$n \leq \left\lceil \frac{d}{\min\{d,\ K-d\}} \right\rceil + 1.$$

In particular, if $d+1 \leq K < 2d$, then $n \leq 2\left\lceil \frac{d}{K-d} \right\rceil$, and if $K \geq 2d$, then $n \leq 2$.

The above structure allows us to decompose the total regret into (i) the semi-bandit regret on the aggregated meta-losses and (ii) an overhead incurred during transition phases. Balancing the semi-bandit regret term with this transition overhead yields our final regret bound.

Theorem 11 *For any $1 < d < K$, the regret of Algorithm 1 run with a minimax optimal combinatorial semi-bandit algorithm $\mathbb{A}$ satisfies*

$$R_T = \mathcal{O}\left(\left(\frac{d^2 K}{\min\{d,\ K-d\}} \right)^{1/3} T^{2/3} \right),$$

with an appropriate choice of batch size B. Specifically, $R_T = \mathcal{O}\big((dK)^{1/3} T^{2/3}\big)$, if $K \geq 2d$.

Proof A d-periodic comparator in blocking bandits is uniquely defined with an action $V \in \mathcal{M}$ in combinatorial bandits. Let $V^\star \in \arg\min_{V \in \mathcal{M}} \sum_{t=1}^{T} \ell_t(V(((t-1) \bmod d) + 1))$ be the optimal size-d matching in $\mathcal{M}$ corresponding to best d-periodic comparator.

In each meta-round b, our procedure plays V_b throughout segments $r = n+1, \ldots, B$. For action $V \in \mathcal{M}$, define the meta-loss, i.e., the total loss incurred on these segments as:

$$L_b(V) := \sum_{r=n+1}^{B} \sum_{s=1}^{d} \ell_{\tau(b,r,s)}\big(V(s)\big)\,.$$

Let $R_N(\mathbb{A})$ denote the standard static regret of $\mathbb{A}$ in combinatorial bandits on the sequence $\{L_b\}_{b=1}^{N}$, namely

$$R_N(\mathbb{A}) := \sum_{b=1}^{N} L_b(V_b) - \min_{V \in \mathcal{M}} \sum_{b=1}^{N} L_b(V) \geq \sum_{b=1}^{N} \big(L_b(V_b) - L_b(V^\star)\big)\,,$$

where V_b is the action selected by $\mathbb{A}$ at meta-round b.

Using Lemma 10, the additional regret incurred by the algorithm at the beginning of each new block is bounded by dn, where $n \leq \left\lceil \frac{d}{\min\{d,\, K-d\}} \right\rceil + 1$. Therefore, the regret of Algorithm 1 against the best d-periodic comparator is bounded as $R_T \leq R_N(\mathbb{A}) + dnN$.

The combinatorial semi-bandit algorithm of Audibert et al. (2014) guarantees the minimax regret bound $\Theta\big(\sqrt{mDN}\big)$ where m is the largest 1-norm of $V \in \mathcal{M}$ when $\mathcal{M}$ is represented as a subset of $\{0,1\}^D$ such that ℓ_t is function linear in V. In our case, we can represent bipartite matchings using incident vectors on the set of dK edges between $[d]$ and $[K]$, which gives $D = dK$ and $m = d$. Recalling that $N = \frac{T}{dB}$, we get that $R_N(\mathbb{A})$ is of order $\sqrt{d(dK)\frac{T}{dB}}$. Finally, using $R_T \leq R_N(\mathbb{A}) + dnN$ and scaling the bound by B to express regret in terms of times steps instead of batches we get

$$R_T = \mathcal{O}\left(B\sqrt{dK\frac{T}{B}} + dn\frac{T}{dB}\right).$$

Recalling that $n \leq \left\lceil \frac{d}{\min\{d,\, K-d\}} \right\rceil + 1$ and tuning B accordingly, concludes the proof. ■

6. Discussion: Breaking the $T^{2/3}$ Barrier

A natural question raised by our results is under what conditions the better rate $\sqrt{T}$ can be recovered. In this section, we discuss two ways of ensuring this rate: either by improving the feedback quality (full information instead of bandit feedback) or by relaxing the performance benchmark (α-regret with $\alpha < 1$).

Full information feedback. With full information feedback, the learner observes the entire loss vector $\ell_t \in [0,1]^K$ at the end of each round t. This stronger feedback allows for $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bounds using Algorithm 1 where the semi-bandit combinatorial algorithm $\mathbb{A}$ is replaced by a full information variant. For example, under full information feedback we can use the low-switch combinatorial algorithm of Devroye et al. (2015), which has regret $\widetilde{\mathcal{O}}\big(d^{3/2}\sqrt{T}\big)$ over (K,d)-bipartite matchings with a $\mathcal{O}\big(d\ln(dK)\sqrt{T}\big)$ bound on the expected number of switches. Hence the reduction gives $R_T \leq R_N(\mathbb{A}) + dn\sum_{b=1}^{N}\mathbb{I}\{V_{b+1} \neq V_b\}$. As the algorithm is an instance of Follow-the-Perturbed-Leader (FPL), it only requires to run a static linear optimization oracle at each round, which is efficiently implementable over (K,d)-bipartite matchings—see Theorem 1. We have thus the following corollary from Theorem 11 (proof in Appendix C.2).

Corollary 12 *For any $1 < d < K$, the regret of Algorithm 1 run using the algorithm of Devroye et al. (2015, Theorem 4) as base algorithm $\mathbb{A}$ under full information feedback satisfies*

$$R_T = \widetilde{\mathcal{O}}\left(\frac{d^{7/4}}{\sqrt{\min\{d,\ K-d\}}}\sqrt{T}\right).$$

Thus full information recovers the $\sqrt{T}$ rate and removes the polynomial dependence on K (leaving only a $\log(dK)$ dependence). However, balancing $R_N(\mathbb{A})$ and the transition cost via the batch size B negatively impacts the polynomial dependence on d. We suspect the current exponent $\frac{7}{4}$ be an artifact of our analysis.

Bounds on α-regret. We say that a policy π for (K, d)-blocking bandits achieves α-regret if the total reward $G_T(\pi)$ of the policy satisfies $G_T(\pi) \geq \alpha \max_{\boldsymbol{a} \in \mathcal{A}_d} G_T(\boldsymbol{a}) - o(T)$ where $G_T(\boldsymbol{a})$ is the total reward of the d-periodic policy $\boldsymbol{a}$. Note that we define α-regret for reward instead of losses, as losses and rewards are not equivalent when we look at multiplicative approximations of the best comparator.

We can prove a $\frac{1}{2}$-regret bound of $\mathcal{O}(\sqrt{T})$ using the following simple observation: if we partition time into batches of length d, then the total reward of the best d-periodic policy is at most the sum of the total rewards of the best d-periodic policies restricted to the odd and the even batches. Therefore, at least one parity must account for at least half of the optimal total reward. This suggests a simple strategy: randomly draw one of the two parities (odd or even) and run a standard adversarial combinatorial semi-bandit algorithm for (K, d)-bipartite matching only on the batches with the drawn (active) parity. The batches of the other (inactive) parity are used as buffers to satisfy the blocking constraint. Indeed, when $K \geq 2d$, Lemma 10 guarantees the following: let V_t be the sequence (or matching) of d-actions to be played next in the active parity. Then we can find a d-sequence V' such that V_{t-1}, V', V_t is a sequence of $3d$ actions satisfying the blocking constraints, where V_{t-1} is the d-sequence previously played in the active batch.

For example, using the OSMD algorithm from Audibert et al. (2014) over the class of (K, d)-bipartite matchings, we get that the regret over any parity is at most $\mathcal{O}\big(d\sqrt{KT}\big)$ and we obtain the following result, bounding the regret of our strategy compared to half of the optimal d-periodic reward.

Theorem 13 *Assume $K \geq 2d$. Then there exists an efficient randomized strategy π for (K, d)-blocking bandits satisfying*

$$G_T(\pi) \geq \frac{1}{2} \max_{\boldsymbol{a} \in \mathcal{A}_d} G_T(\boldsymbol{a}) - \mathcal{O}\big(d\sqrt{KT}\big) \,.$$

7. Conclusion

Our results leave a gap concerning the optimal dependence of the minimax regret on K and d. Our reduction to bipartite matchings (Algorithm 1) becomes costly for blocking bandits whenever the current matching changes. A different, and potentially better approach would be to embed the blocking constraints in the states of an adversarial deterministic MDP, and then compete against the best periodic trajectory on this MDP. This would require improving the analysis of Dekel and Hazan (2013) by leveraging the structure of the MDP implementing the blocking constraints. On the other hand, improving the dependence on K in the lower bound seems challenging: the $T^{2/3}$ rate in the lower bound is due to the cost of exploring arms that are currently blocked, and when K is large compared to d, most arms are available.

Another direction is to identify additional structural assumptions or restricted adversarial models under which one can use a comparator achieving a better approximation to OPT than the $1/K$ worst-case barrier of periodic strategies, while still ensuring tractability and sublinear regret.

Acknowledgments

The authors are grateful to Gergely Neu for helpful discussions. NCB acknowledges the financial support from the EU Horizon CL4-2022-HUMAN-02 research and innovation action under grant agreement 101120237, project ELIAS (European Lighthouse of AI for Sustainability). JH is supported by JSPS KAKENHI Grant Number JP25K03184. YK is partially supported by Japan Science and Technology Agency (JST) Strategic Basic Research Programs PRESTO "R&D Process Innovation by AI and Robotics: Technical Foundations and Practical Applications" grant number JPMJPR24T.

References

Ahsan Alvi, Binxin Ru, Jan-Peter Calliess, Stephen Roberts, and Michael A. Osborne. Asynchronous batch Bayesian optimisation with improved local penalisation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 253–262, 2019.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1747–1754, 2012.

Alexia Atsidakou, Orestis Papadigenopoulos, Soumya Basu, Constantine Caramanis, and Sanjay Shakkottai. Combinatorial blocking bandits with stochastic delays. In *Proceedings of the 38th International Conference on Machine Learning*, pages 404–413, 2021.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM*, 65(3), 2018.

Amotz Bar-Noy, Randeep Bhatia, Joseph Naor, and Baruch Schieber. Minimizing service and operation costs of periodic scheduling. *Mathematics of Operations Research*, 27(3):518–544, 2002.

Soumya Basu, Rajat Sen, Sujay Sanghavi, and Sanjay Shakkottai. Blocking bandits. In *Advances in Neural Information Processing Systems*, pages 4785–4794, 2019.

Soumya Basu, Orestis Papadigenopoulos, Constantine Caramanis, and Sanjay Shakkottai. Contextual blocking bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 271–279, 2021.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. Bandits with replenishable knapsacks: the best of both worlds. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

Nicholas Bishop, Hau Chan, Debmalya Mandal, and Long Tran-Thanh. Adversarial blocking bandits. In *Advances in Neural Information Processing Systems*, 2020.

Leonardo Cella and Nicoló Cesa-Bianchi. Stochastic bandits with delay-dependent payoffs. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 1168–1177, 2020.

Nicolò Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Sapana Chaudhary and Dileep Kalathil. Safe online convex optimization with unknown linear safety constraints. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 6175–6182, 2022.

Arun Kumar Das, Michal Opler, and Tomáš Valla. Precoloring extension with demands on paths. In *Proceedings of the 36th International Symposium on Algorithms and Computation*, pages 23:1–23:15, 2025.

Ofer Dekel and Elad Hazan. Better rates for any adversarial deterministic MDP. In *Proceedings of the 30th International Conference on Machine Learning*, pages 675–683, 2013.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: $T^{2/3}$ regret. In *Proceedings of the 46th ACM Symposium on Theory of Computing*, page 459–467, 2014.

Luc Devroye, Gábor Lugosi, and Gergely Neu. Random-walk perturbations for online combinatorial optimization. *IEEE Transactions on Information Theory*, 61(7):4099–4106, 2015.

Yanyan Dong and Vincent Y. F. Tan. Adversarial combinatorial bandits with switching costs. *IEEE Transactions on Information Theory*, 70(7):5213–5227, 2024.

Ayoub Foussoul, Vineet Goyal, Orestis Papadigenopoulos, and Assaf Zeevi. Last switch dependent bandits with monotone payoff functions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10265–10284, 2023.

Dan Garber and Ben Kretzu. Projection-free online convex optimization with time-varying constraints. In *Proceedings of the 41st International Conference on Machine Learning*, pages 14988–15005, 2024.

Robert Holte, Aloysius Mok, Louis Rosier, Igor Tulchinsky, and Donald Varvel. Pinwheel: A real-time scheduling problem. In *Proceedings of the 22nd Hawaii International Conference on System Science*, pages 693–702, 1989.

Spencer Hutchinson, Tianyi Chen, and Mahnoosh Alizadeh. Optimistic safety for online convex optimization with unknown linear constraints. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, pages 2809–2817, 2025.

Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Bandit task assignment with unknown processing time. In *Advances in Neural Information Processing Systems*, 2023.

Rodolphe Jenatton, Jim Huang, and Cedric Archambeau. Adaptive algorithms for online convex optimization with long-term constraints. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 402–411, 2016.

Robert Kleinberg and Nicole Immorlica. Recharging bandits. In *Proceedings of the 59th IEEE Symposium on Foundations of Computer Science*, pages 309–319, 2018.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine Learning*, 80(2):245–272, 2010.

Bernhard Korte and Jens Vygen. *Combinatorial Optimization: Theory and Algorithms*, volume 21 of *Algorithms and Combinatorics*. Springer, 2012.

Raunak Kumar and Robert Kleinberg. Non-monotonic resource utilization in the bandits with knapsacks problem. In *Advances in Neural Information Processing Systems*, 2022.

Pierre Laforgue, Giulia Clerici, Nicolò Cesa-Bianchi, and Ran Gilad-Bachrach. A last switch dependent analysis of satiation and seasonality in bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 971–990, 2022.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: online convex optimization with long term constraints. *The Journal of Machine Learning Research*, 13(1): 2503–2528, 2012.

Gergely Neu and Michal Valko. Online combinatorial optimization with stochastic decision sets and adversarial losses. In *Advances in Neural Information Processing Systems*, pages 2780–2788, 2014.

Quan M. Nguyen and Nishant Mehta. Near-optimal per-action regret bounds for sleeping bandits. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pages 2827–2835, 2024.

Shunhao Oh, Anuja Meetoo Appavoo, and Seth Gilbert. Periodic bandits and wireless network selection. In *Proceedings of the 46th International Colloquium on Automata, Languages, and Programming*, pages 149:1–149:15, 2019.

Orestis Papadigenopoulos and Constantine Caramanis. Recurrent submodular welfare and matroid blocking semi-bandits. In *Advances in Neural Information Processing Systems*, pages 23334–23346, 2021.

Aadirupa Saha, Pierre Gaillard, and Michal Valko. Improved sleeping bandits with stochastic actions sets and adversarial rewards. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8357–8366, 2020.

Hao Yu, Michael J. Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*, page 1427–1437, 2017.

Jingxuan Zhu and Bin Liu. Thompson sampling for bandits with cool-down periods. *Transactions on Machine Learning Research*, 2025.

Appendix A. Reduction to Adversarial Deterministic MDP

In this appendix, we show that although the adversarial deterministic Markov Decision Processes (ADMDPs) reduction achieves the optimal dependence on the horizon T, it suffers from a prohibitive dependence on the number of arms K and blocking length d.

Let $G = (\mathcal{V}, \mathcal{E})$ be the state transition graph defined as follows. The state space $\mathcal{V}$ consists of all valid sequences of the last $d-1$ actions. Since no action can be repeated within this window, each state corresponds to a permutation of length $d-1$ from the set of K arms. Formally, a state $v \in \mathcal{V}$ is defined as:

$$v = (a_1, a_2, \ldots, a_{d-1}),$$

where $a_k \in [K]$ represents the arm played $d-k$ steps ago. The size of the state space is:

$$|\mathcal{V}| = P(K, d-1) = \frac{K!}{(K-d+1)!}.$$

The set of edges $\mathcal{E}$ defines the transitions. From a current state $v = (a_1, a_2, \ldots, a_{d-1})$, the player chooses a new action a_{new}. The blocking constraint requires that a_{new} is not currently blocked, meaning $a_{\text{new}} \notin \{a_1, a_2, \ldots, a_{d-1}\}$. The number of available actions (and thus outgoing edges) for any state is exactly $K-(d-1)$. Upon choosing a_{new}, the system transitions to the new state v' by shifting the window:

$$v' = (a_2, \ldots, a_{d-1}, a_{\text{new}}).$$

The oldest action a_1 leaves the history window and becomes unblocked for the next step. The total number of edges is:

$$|\mathcal{E}| = |\mathcal{V}| \times (K-d+1).$$

In the setting of Dekel and Hazan (2013), the adversary defines a sequence of T loss functions, $f_1, \ldots, f_T$, where each $f_t : \mathcal{E} \to [0,1]$. They compare the player's loss to the loss of the best deterministic policy in hindsight. Formally, a deterministic policy $\pi \in \Pi$ is a mapping from $\mathcal{V}$ to itself mapping each state v to one of its outgoing neighbors. With the initial state v_0, $\pi^t(v)$ denotes the next state at t by the fixed policy π, and the loss incurred by π at time t is given by $f_t(\pi^{t-1}(v_0), \pi^t(v_0))$. Using this notation, now the player's regret $R_{\text{MDP}}(T)$ is defined as:

$$R_{\text{MDP}}(T) = \mathbb{E}\left[\sum_{t=1}^{T} f_t(V_{t-1}, V_t)\right] - \min_{\pi \in \Pi} \sum_{t=1}^{T} f_t(\pi^{t-1}(v_0), \pi^t(v_0)).$$

Proposition 14 *There exists an algorithm that achieves the regret against the best d-periodic comparator of order*

$$R_T = \mathcal{O}\left(K^{3d-2} T^{2/3}\right).$$

Proof By Corollary 4.1 of Dekel and Hazan (2013), there exists an algorithm that achieves

$$R_{\text{MDP}}(T) = \mathcal{O}(|\mathcal{V}|^2 |\mathcal{E}| T^{2/3}).$$

Using $|\mathcal{V}| = P(K, d-1) = \frac{K!}{(K-d+1)!}$ and $|\mathcal{E}| = |\mathcal{V}|(K-d+1)$ when applying the above regret bound, we obtain

$$R_{\text{MDP}}(T) = \mathcal{O}\left(\left(\frac{K!}{(K-d+1)!}\right)^3 (K-d+1) \cdot T^{2/3}\right) = \mathcal{O}\left(K^{3d-2} \cdot T^{2/3}\right).$$

When we regard the adversarial blocking bandit as a special case of the AMDP problem, a deterministic policy class $\Pi = \{\pi : \mathcal{V} \to \mathcal{V}\}$ clearly includes a d-periodic policy class $\Pi_d = \{\pi : [d] \to [K]\}$. Therefore, the best d-periodic comparator is weaker than the optimal deterministic policy, which implies that this upper bound also applies to the regret against the best d-periodic comparator:

$$R_T \leq R_{\mathrm{MDP}}(T) = \mathcal{O}\left(K^{3d-2}T^{2/3}\right).$$

■

Appendix B. Additional Related Work

Extensions of stochastic blocking bandits have been developed for contextual bandits (Basu et al., 2021), combinatorial bandits with stochastic delays (Atsidakou et al., 2021), and matroid-structured semi-bandits (Papadigenopoulos and Caramanis, 2021). Our model is also related to variants of non-stationary bandits where the payoff of an arm depends on the time elapsed since the last play of that arm, such as recharging bandits (Kleinberg and Immorlica, 2018), bandits with delay-dependent payoffs (Cella and Cesa-Bianchi, 2020), and last-switch dependent bandits (Laforgue et al., 2022; Foussoul et al., 2023).

It also connects to bandits with availability constraints, most notably sleeping bandits (Kleinberg et al., 2010; Neu and Valko, 2014; Saha et al., 2020; Nguyen and Mehta, 2024). However, unlike sleeping bandits, blocking bandits enforce action-induced hard constraints where arms temporarily disappear. Related action-induced constraints also appear in asynchronous Bayesian optimization (Alvi et al., 2019) and in bandit task assignment with unknown processing times (Ito et al., 2023), though these works focus on continuous action space or stochastic reward settings.

Regarding periodicity, Oh et al. (2019) investigates periodic strategies in adversarial settings to model scenarios like network traffic. However, their periodicity models patterns inherent to the environment. In contrast, the periodicity in our setting is structurally induced by the blocking constraints. Motivated by applications in cognitive radio networks, the notion of cool-down periods is recently studied in Zhu and Liu (2025). They focused on the stochastic regime and their regret is defined against a dynamic oracle that greedily selects the best available arm at each step. This benchmark is ill-suited for the adversarial setting, where such a myopic policy can be arbitrarily suboptimal compared to the global offline optimum.

While our lower-bound construction embeds an instance of BSC, the existing analysis of Cesa-Bianchi et al. (2013); Dekel et al. (2014) does not directly transfer to the blocking setting. In their analysis, it is crucial to relate the information gain to the number of switches, while blocking models impose no switching penalty for exploration, but enforce history-dependent feasibility. Additionally, the comparator differs: switching-cost regret is defined in relation to the best fixed action. Furthermore, while we identify d-periodic strategies with size-d matchings in the bipartite graph $[d] \times [K]$, algorithms for combinatorial bandits with switching costs, such as Dong and Tan (2024), are insufficient. A sequence of matchings output by such algorithms may violate the blocking constraints, necessitating the insertion of additional transition schedules as done in our upper bound analysis.

A related line of work studies Online Convex Optimization (OCO) under constraints. In particular, long-term constraints require constraints to be satisfied only in aggregate over a horizon T (Mahdavi et al., 2012; Jenatton et al., 2016), and this has been extended to stochastic constraints where

constraint functions are revealed after actions (Yu et al., 2017). More recent work considers time-varying constraints (Garber and Kretzu, 2024). In contrast, safe OCO imposes per-round safety, and has been studied under unknown linear safety constraints (Chaudhary and Kalathil, 2022), with subsequent refinements for unknown linear constraints and bandit constraint feedback (Hutchinson et al., 2025). Another relevant framework is Bandits with Knapsacks (BwK), which introduces resource consumption constraints, where actions consume a global budget (Badanidiyuru et al., 2018). Generalizations to renewable or replenishing resources have also been explored (Bernasconi et al., 2024; Kumar and Kleinberg, 2022). Our setting differs fundamentally from constrained OCO and BwK: blocking is a discrete, action-induced feasibility constraint, thereby creating temporal dependencies, distinct from global budget limits or aggregate and safety constraints.

Finally, prior hardness results for periodic policies in stochastic blocking bandits often rely on heterogeneous blocking times via reductions from periodic scheduling problems (Bar-Noy et al., 2002; Holte et al., 1989); these do not apply to our identical-d setting, where we instead use a reduction from Distance Precoloring Extension on paths (Das et al., 2025).

Appendix C. Omitted Proofs

C.1. Proof of Lemma 10

Lemma C.1 (Restatement of Lemma 10) *Let $V', V \in \mathcal{M}$ be size-d matchings. Then there exists a sequence of $n+1$ matchings $\left(V^{(0)}, V^{(1)}, \dots, V^{(n)}\right)$ such that $V' = V^{(0)}$, $V = V^{(n)}$, and the dn-step schedule $V^{(0)}, V^{(1)}, \dots, V^{(n)}$ satisfies the blocking constraint. Moreover, we can take*

$$n \le \left\lceil \frac{d}{\min\{d,\ K-d\}} \right\rceil + 1.$$

In particular, if $d+1 \le K < 2d$, then $n \le 2\left\lceil \frac{d}{K-d} \right\rceil$, and if $K \ge 2d$, then $n \le 2$.

Proof

We proceed by constructing the transition sequence inductively. For any target matching V, let $\sigma_V = (v_1^*, \dots, v_d^*) \in [K]^d$ be the corresponding d-periodic schedule. Set $m := \min\{d, K-d\}$ and $n := \lceil d/m \rceil + 1$. For each $\ell = 1, \dots, n-1$, define the ℓ-th segment of offsets $(\ell-1)m < j \le \min(\ell m, d)$, and let

$$\mathcal{U}_\ell \ := \ \{\, v_j^* : (\ell-1)m < j \le \min(\ell m, d) \,\}.$$

So $\mathcal{U}_\ell$ is simply the set of target arms we newly introduce in segment ℓ, and it satisfies $|\mathcal{U}_\ell| \le m \le K-d$.

We construct matchings $V^{(1)}, \dots, V^{(n)}$ after the given preceding $V^{(0)} = V'$. The construction maintains the invariant that, for $\ell = 1, \dots, n-2$, the matching $V^{(\ell+1)}$ avoids $\mathcal{U}_{\ell+1}$ everywhere (this is what makes the next segment available).

First, construct $V^{(1)}$ as any valid size-d matching that avoids $\mathcal{U}_1$ entirely. This is feasible by a greedy fill: at any step, at most $d-1$ arms are blocked by the last $d-1$ plays, and additionally we forbid $|\mathcal{U}_1| \le K-d$ arms, so at least one arm (edge) is available.

For each $\ell = 1, \dots, n-1$, we build block $V^{(\ell+1)}$ so that: (i) it plays v_j^* at every offset $j \le \min(\ell m, d)$ (i.e., once an offset is fixed, it stays fixed thereafter), and (ii) if $\ell \le n-2$, it avoids $\mathcal{U}_{\ell+1}$ everywhere in the block. By construction, the previous block $V^{(\ell)}$ avoids $\mathcal{U}_\ell$, hence none of the arms in $\mathcal{U}_\ell$ appeared in the previous d steps; therefore, every newly introduced target arm $v_j^* \in \mathcal{U}_\ell$

is unblocked when it is first placed. Already-fixed target arms can be repeated at the same offset in the next block, which is allowed since repeats occur exactly d steps apart. All remaining offsets are filled greedily while respecting blocking (and additionally avoiding $\mathcal{U}_{\ell+1}$ when $\ell \le n-2$); the same counting as in $V^{(1)}$ guarantees a feasible choice at each step. For $\ell = n-1$, we have $\min(\ell m, d) = d$, so the entire block $V^{(n)}$ equals $(v_1^*, \dots, v_d^*)$ and thus realizes V.

■

C.2. Proof of Corollary 12

Corollary C.2 (Restatement of Corollary 12) *There exists an efficient algorithm for the full-information (K, d)-blocking bandit problem that achieves an expected regret of: For any $1 < d < K$, the regret of Algorithm 1 run with a full-information algorithm of Devroye et al. (2015) satisfies*

$$R_T = \widetilde{\mathcal{O}}\left(\frac{d^{7/4}}{\sqrt{\min\{d,\ K-d\}}} T^{1/2}\right).$$

Proof We employ the FPL algorithm for bipartite matchings on the meta-game over $N = \lfloor T/dB \rfloor$ rounds as an online combinatorial optimization algorithm $\mathbb{A}$ in Algorithm 1. Let dn be the length of transition phase to clear the blocking constraint, where n is upper bounded by Lemma 10. We follow the same step as Theorem 11, and the regret against the best d-periodic comparator is bounded by $R_T \le R_N(\mathbb{A}) + dn \sum_{b=1}^{N} \mathbb{I}\{V_{b+1} \neq V_b\}$.

According to Devroye et al. (2015) in the weighted matching case on the complete bipartite graph of $\mathrm{Bip}_{K,d}$ and since the meta-losses are scaled by B, the expected regret $R_N(\mathbb{A})$ and the expected number of switches $S(N) = \mathbb{E}[\sum_{b=1}^{N} \mathbb{I}\{V_{b+1} \neq V_b\}]$ of the FPL algorithm satisfy:

$$R_N(\mathbb{A}) = 4Bd^{3/2}\sqrt{N \log(dK)} + B\frac{d \log N}{2}, \quad S(N) = d\sqrt{N}\log(dK).$$

Then induced regret satisfies for the parameter B,

$$R_T \le 4Bd^{3/2}\sqrt{N \log(dK)} + B\frac{d \log N}{2} + d^2 n\sqrt{N}\log(dK)$$

Hence

$$R_T = \mathcal{O}\left(d\sqrt{T \log(dK)}\,\sqrt{B} \ + \ d^{3/2} n \ \log(dK)\,\frac{\sqrt{T}}{\sqrt{B}} \ + \ dB \log\frac{T}{dB}\right).$$

Balancing the first and the second terms yields the tuning $B = \Theta\Big(n\sqrt{d \log(dK)}\Big)$. With this choice,

$$R_T = \mathcal{O}\left(d^{5/4}\big(\log(dK)\big)^{3/4}\sqrt{n\,T} \ + \ d^{3/2}\ n\sqrt{\log(dK)}\log\frac{T}{d^{3/2} n\sqrt{\log(dK)}}\right).$$

Ignoring the logarithmic term and assuming T is sufficiently large, the main statement is

$$R_T = \widetilde{\mathcal{O}}\Big(d^{5/4} n^{1/2} T^{1/2}\Big).$$

■