Understanding the Role of Feedback in Online Learning with Switching Costs

Duo Cheng [1] **Xingyu Zhou** [2] **Bo Ji** [1]

Abstract

In this paper, we study the role of feedback in online learning with switching costs. It has been shown that the minimax regret is $\widetilde{\Theta}(T^{2/3})$ under bandit feedback and improves to $\widetilde{\Theta}(\sqrt{T})$ under full-information feedback, where T is the length of the time horizon. However, it remains largely unknown how the amount and type of feedback generally impact regret. To this end, we first consider the setting of bandit learning with extra observations; that is, in addition to the typical bandit feedback, the learner can freely make a total of B_{ex} *extra observations*. We fully characterize the minimax regret in this setting, which exhibits an interesting *phase-transition phenomenon*: when $B_{\text{ex}} = O(T^{2/3})$, the regret remains $\widetilde{\Theta}(T^{2/3})$, but when $B_{\text{ex}} = \Omega(T^{2/3})$, it becomes $\widetilde{\Theta}(T/\sqrt{B_{\text{ex}}})$, which improves as the budget B_{ex} increases. To design algorithms that can achieve the minimax regret, it is instructive to consider a more general setting where the learner has a budget of B *total observations*. We fully characterize the minimax regret in this setting as well and show that it is $\widetilde{\Theta}(T/\sqrt{B})$, which scales smoothly with the total budget B. Furthermore, we propose a generic algorithmic framework, which enables us to design different learning algorithms that can achieve matching upper bounds for both settings based on the amount and type of feedback. One interesting finding is that while bandit feedback can still guarantee optimal regret when the budget is relatively limited, it no longer suffices to achieve optimal regret when the budget is relatively large.

1. Introduction

Online learning over a finite set of actions is a classical problem in machine learning research. It can be formulated as a T-round repeated game between a learner and an adversary: at each round, the learner chooses one of the K actions and suffers the loss of this chosen action, where the loss is determined by the adversary. At the end of each round, the learner receives some feedback and uses it to update her policy at the next round. The goal of the learner is to minimize the *regret*, defined as the difference between her cumulative loss and that of the best fixed action in hindsight.

In terms of the type of feedback, two important settings have been extensively studied in the literature: bandit and full information. At each round, if the learner observes only the loss of the chosen action, then it is called *bandit* feedback, and the game is called *adversarial multi-armed bandits* (MAB) or *non-stochastic bandits with adversarial losses* (Auer et al., 2002b). On the other hand, if the losses of all K actions are revealed to the learner, then it is called *full-information* feedback, and the game becomes *prediction with expert advice* (Cesa-Bianchi & Lugosi, 2006).

The regret in these two settings has been well understood. Specifically, the minimax regret is $\Theta(\sqrt{TK})$[1] under bandit feedback (Auer et al., 2002b; Audibert & Bubeck, 2009) and is $\Theta(\sqrt{T \ln K})$ under full information (Cesa-Bianchi & Lugosi, 2006, Theorems 2.2 and 3.7) (Hazan, 2016) (Orabona, 2019, Section 6.8). These results imply that learning under bandit feedback is *slightly harder* than under full information, in the sense that the dependency on K is worse ($\Theta(\sqrt{K})$ vs. $\Theta(\sqrt{\ln K})$). However, the scaling with respect to T remains the same (i.e., $\Theta(\sqrt{T})$).

In the above standard settings, the learner is allowed to *arbitrarily* switch actions at two consecutive rounds. However, in many real-world decision-making problems, switching actions may incur a cost (e.g., due to system reconfiguration and resource reallocation) (Zhang et al., 2005; Kaplan, 2011). Motivated by this practical consideration, a new setting called *online learning with switching costs* has also been extensively studied (Arora et al., 2012; Cesa-Bianchi et al., 2013). In this setting, the learner needs to pay an additional unit loss whenever she switches actions.

Interestingly, it has been shown that in this new setting, learning under bandit feedback is *significantly harder* than

[1]Virginia Tech, Blacksburg, USA [2]Wayne State University, Detroit, USA. Correspondence to: Bo Ji <boji@vt.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]We use standard big O notations (e.g., O, Ω, and Θ); those with tilde (e.g., $\widetilde{O}$, $\widetilde{\Omega}$, and $\widetilde{\Theta}$) hide poly-logarithmic factors.

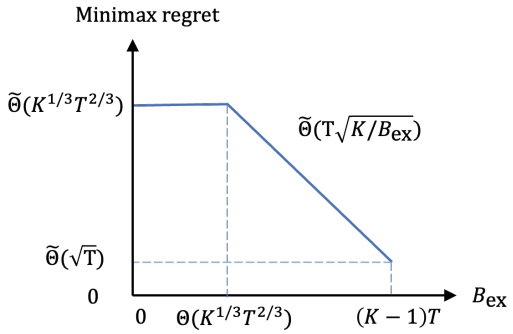
under full information. Under full-information feedback, even with switching costs, the minimax regret remains $\Theta(\sqrt{T \ln K})$, which can be achieved by several algorithms such as Shrinking Dartboard (SD) (Geulen et al., 2010) and Follow-the-Perturbed-Leader (FTPL) (Devroye et al., 2013). On the other hand, Dekel et al. (2013) shows a (worse) lower bound of $\widetilde{\Omega}(K^{1/3}T^{2/3})$ for the bandit setting, which can be matched (up to poly-logarithmic factors) by the batched EXP3 algorithm (Arora et al., 2012). These results reveal that introducing switching costs makes bandit problems *strictly harder* than expert problems due to the worse dependency on T (i.e., $\widetilde{\Theta}(T^{2/3})$ vs. $\widetilde{\Theta}(\sqrt{T})$).

Our Contributions. While these two special cases have been well studied, it remains largely unknown how feedback impacts regret in general. To close this important gap, we aim to fundamentally understand the role of feedback (in terms of both amount and type) in online learning with switching costs. Our main contributions are as follows.

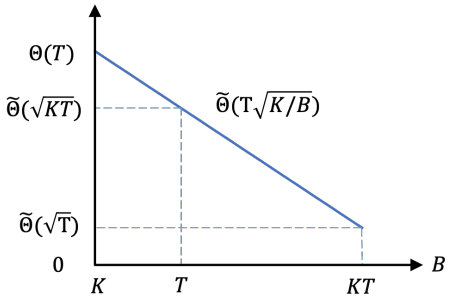
(i) We first consider the setting of bandit learning with extra observations, where in addition to the typical bandit feedback, the learner can freely make a total of B_{ex} *extra observations* in an arbitrary form (Section 3). We present a tight characterization of the minimax regret, which exhibits an interesting *phase-transition phenomenon* (see Fig. 1(a)). Specifically, when $B_{\text{ex}} = O(T^{2/3})$, the regret remains $\widetilde{\Theta}(T^{2/3})$, but when $B_{\text{ex}} = \Omega(T^{2/3})$, it becomes $\widetilde{\Theta}(T/\sqrt{B_{\text{ex}}})$, which improves as the budget B_{ex} increases.

(ii) To understand this phenomenon and design algorithms that can achieve the minimax regret, it is instructive to consider a more general setting where the learner has a budget of B *total observations* (Section 4). We fully characterize the minimax regret in this setting as well and show that it is $\widetilde{\Theta}(T/\sqrt{B})$, which scales smoothly with the total budget B (see Fig. 1(b)). Furthermore, we propose a generic algorithmic framework, which enables us to design different learning algorithms that can achieve matching upper bounds for both settings based on the amount and type of feedback.

(iii) Our findings highlight the crucial impact of feedback type (bandit vs. others) in the second setting (see Table 1). In particular, while both bandit and other types of feedback can achieve optimal regret when the budget is relatively limited, *pure bandit feedback is no longer sufficient to guarantee optimal regret when the budget is relatively large.* However, in the standard setting without switching costs, all three types of feedback we consider can achieve optimal regret in the full range of B. This reveals that the impact of feedback type is (partly) due to switching costs.



(a) Bandit feedback plus B_{ex} extra observations



(b) B total observations

Figure 1. An illustration of the minimax regret vs. observation budget in log-log plots: (a) the learner receives bandit feedback plus no more than B_{ex} extra observations (Theorem 1); (b) the learner can make no more than B total observations (Theorem 2).

Table 1. The minimax regret under different types of feedback in the setting of online learning under a total observation budget B: with (w/) vs. without (w/o) switching costs (SC). A formal description of "Flexible" feedback can be found in Section 4.2.

Feedback Type	**Minimax Regret**	
	w/ SC	w/o SC
Full-information	$\widetilde{\Theta}(T\sqrt{K/B})$	
Flexible		
Bandit ($B = O(K^{1/3}T^{2/3})$)		
Bandit ($B = \Omega(K^{1/3}T^{2/3})$)	$\widetilde{\Theta}(K^{1/3}T^{2/3})$	

2. Problem Setup

In this section, we introduce basic notations and present the problem setup. For any positive integer n, let $[n] := \{1, \ldots, n\}$, and let $\ell_{1:n}$ be the loss sequence $\ell_1, \ldots, \ell_n$. We use $\mathbb{I}_{\{\mathcal{E}\}}$ to denote the indicator function of event $\mathcal{E}$: $\mathbb{I}_{\{\mathcal{E}\}} = 1$ if event $\mathcal{E}$ happens, and $\mathbb{I}_{\{\mathcal{E}\}} = 0$ otherwise.

The learning problem can be viewed as a repeated game between a learner and an adversary. Assume that there are $K > 1$ actions the learner can choose. Let $T \geq K$ be the length of the time horizon, which is fixed at the beginning of the game and is known to the learner. At each round $t \in [T]$, the adversary assigns a loss in $[0, 1]$ to each action in $[K]$; the learner samples an action X_t from a probability

distribution w_t (also determined by the learner) over the action set $[K]$. After taking action X_t, the learner suffers a loss of the chosen action, i.e., $\ell_t[X_t]$. By the end of each round, the learner observes the loss of some actions (specific types of such feedback will be discussed later) and updates probability distribution w_{t+1} that will be used at the next round. Each time when the learner takes an action different from that at the previous round, one unit of switching cost is incurred. The *regret* under a learning algorithm π over a loss sequence $\ell_{1:T}$, denoted by $R_T^\pi(\ell_{1:T})$, is defined as the difference between the cumulative loss (including the switching costs incurred) under algorithm π and that of the optimal (best fixed) action in hindsight:

$$R_T^\pi(\ell_{1:T}) := \sum_{t=1}^{T} \left(\ell_t[X_t] + \mathbb{I}_{\{X_t \neq X_{t-1}\}}\right) - \min_{k\in[K]} \sum_{t=1}^{T} \ell_t[k]. \quad (1)$$

For a randomized algorithm, we consider the *expected regret* (or simply regret), denoted by $\mathbb{E}\left[R_T^\pi(\ell_{1:T})\right]$, where the expectation is taken over the randomness of the algorithm. Without loss of generality, let $\mathbb{I}_{\{X_1 \neq X_0\}} = 0$, i.e., the first action does not incur any switching cost. The adversary is assumed to be *oblivious*, in the sense that the whole loss sequence is determined by the adversary before the game begins. In this paper, for any given algorithm π, we are interested in the *worst-case (expected) regret* over all possible loss sequences (i.e., instance-independent), denoted by R_T^π:

$$R_T^\pi := \sup_{\ell_{1:T}\in[0,1]^{KT}} \mathbb{E}\left[R_T^\pi(\ell_{1:T})\right]. \quad (2)$$

Let Π be the set of all feasible learning algorithms following the specified learning protocol. We define the *minimax (or optimal) regret*, denoted by $R_T^*(\Pi)$, as the minimum worst-case regret under all feasible learning algorithms in Π:

$$R_T^*(\Pi) := \inf_{\pi\in\Pi} R_T^\pi. \quad (3)$$

For notational ease, we may drop Π in $R_T^*(\Pi)$ and simply use R_T^* whenever there is no ambiguity.

To understand the role of feedback in online learning with switching costs, we will consider two different settings with an observation budget: (i) in addition to the typical bandit feedback, the learner can freely make a total of B_{ex} extra observations (Section 3); (ii) the learner can freely make B total observations (Section 4). Due to space limitations, in Appendix A we provide motivating examples for the settings with an observation budget we consider.

3. Bandit Learning with Switching Costs under Extra Observation Budget

Observing the gap in the optimal regret bound under bandit and full-information feedback ($\widetilde{\Theta}(T^{2/3})$ vs. $\widetilde{\Theta}(\sqrt{T})$), it is natural to ask: *How much can one improve upon the $\widetilde{\Theta}(T^{2/3})$ regret if the learner is allowed to make some extra observations in addition to the typical bandit feedback?*

Motivated by this question, we consider the setting of bandit learning with switching costs under an *extra* observation budget. We consider the learning protocol specified in Section 2, and in addition to the typical bandit feedback, the learner is allowed to freely use at most B_{ex} extra observations of the loss of other action(s) throughout the game, where B_{ex} is an integer in $[0, (K-1)T]$. At the two endpoints of 0 and $(K-1)T$, this new setting recovers the bandit and full-information cases, respectively. In this section, by slightly abusing the notation, we also use Π to denote the set of all learning algorithms using typical bandit feedback plus B_{ex} extra observations, and we are interested in the minimax regret R_T^* for $B_{\text{ex}} \in [0, (K-1)T]$.

3.1. Minimax Regret

We first present our main result of the minimax regret R_T^* in this setting, which is formally stated in Theorem 1.

Theorem 1. *In the setting of bandit learning with switching costs under an extra observation budget $B_{\text{ex}} \in [0, (K-1)T]$, the minimax regret is given by*

$$R_T^* = \begin{cases} \widetilde{\Theta}(K^{1/3}T^{2/3}), & B_{\text{ex}} = O(K^{1/3}T^{2/3}), \\ \widetilde{\Theta}(T\sqrt{K/B_{\text{ex}}}), & B_{\text{ex}} = \Omega(K^{1/3}T^{2/3}). \end{cases}$$

Remark 1. Interestingly, this minimax regret exhibits a *phase-transition phenomenon* (see, also, Fig. 1(a)): when the amount of extra observations is relatively small (i.e., $B_{\text{ex}} = O(K^{1/3}T^{2/3})$), they are insufficient for improving the regret, which remains $\widetilde{\Theta}(K^{1/3}T^{2/3})$; however, when the amount is large enough (i.e., $B_{\text{ex}} = \Omega(K^{1/3}T^{2/3})$), the regret decreases smoothly as the budget B_{ex} increases.

3.2. Lower Bound

To establish Theorem 1, we will first show a fundamental lower bound, which is formally stated in Proposition 1.

Proposition 1. *For any learning algorithm π that can use a total of B_{ex} extra observations in addition to the typical bandit feedback, there exists a loss sequence $\ell_{1:T}$ (which may depend on both π and B_{ex}) such that*

$$\mathbb{E}\left[R_T^\pi(\ell_{1:T})\right] = \begin{cases} \widetilde{\Omega}(K^{1/3}T^{2/3}), & B_{\text{ex}} = O(K^{1/3}T^{2/3}), \\ \widetilde{\Omega}(T\sqrt{K/B_{\text{ex}}}), & B_{\text{ex}} = \Omega(K^{1/3}T^{2/3}). \end{cases}$$

We provide detailed proof of the above lower bound in Appendix B. Here, we present a proof sketch that mainly focuses on the key steps of the lower bound analysis with necessary explanations. The proof sketch reveals useful insights that not only help explain the interesting phase-transition phenomenon but also shed light on the design of algorithms that can achieve this lower bound.

Proof Sketch of Proposition 1. We first give an overview of the construction of hard loss sequences in our setting and the main ideas behind the construction.

Generally speaking, the difficulty of bandit problems lies in the *exploitation-exploration* tradeoff. On the one hand, the learner wants to pull empirically good actions in order to enjoy a low instantaneous loss (i.e., exploitation); on the other hand, she may also want to pull other actions and gain useful information to distinguish the optimal (best fixed) action and suboptimal actions (i.e., exploration).

In the presence of switching costs, Dekel et al. (2013) proposes hard instances (i.e., loss sequences) based on a *multi-scale random walk* such that useful information toward distinguishability (between the optimal action and suboptimal actions) *can only be obtained when the learner switches actions*, which, however, incurs switching costs. Using carefully constructed instances, they show that switching costs increase the intrinsic difficulty of bandit learning and result in a regret lower bound of $\widetilde{\Omega}(K^{1/3}T^{2/3})$.

However, the hard instances in Dekel et al. (2013) work for *pure bandit feedback* only. That is, if the learner can obtain full-information feedback at *any* of the T rounds, she would immediately identify the optimal action and suffer no regret in the rest of the game. The reason is that the optimal action has the (unique) lowest loss at *all* T rounds.

To make it still hard to learn even when the learner has some extra feedback, we will borrow an idea from Shi et al. (2022) to modify the original hard instance in Dekel et al. (2013): at each round, an additional layer of action-dependent noise is added to the loss of each action. As a result, the optimal action no longer has the lowest loss at all rounds and therefore cannot be trivially identified even when the learner can make extra observations.

In the rest of the proof sketch, we present three key steps of the proof and provide high-level explanations.

Step 1: Establishing the relationship between two regrets. As in Dekel et al. (2013), each loss value in the initial loss sequence we construct, denoted by $\ell_{1:T}^{\text{init}}$, may not be bounded in $[0,1]$; through truncation, we construct the actual loss sequence $\ell_{1:T}$ by simply projecting each initial loss value onto $[0,1]$. For notational ease, we use R_T^{init} and R_T to denote the regret over loss sequences $\ell_{1:T}^{\text{init}}$ and $\ell_{1:T}$, respectively. Recall that the goal is to obtain a lower bound on $\mathbb{E}[R_T]$, which, however, is hard to analyze directly due to the truncation. Instead, we show that it suffices to obtain a lower bound on $\mathbb{E}\left[R_T^{\text{init}}\right]$ (i.e., the regret under untruncated loss sequence), due to the following relationship:

$$\mathbb{E}[R_T] \geq \mathbb{E}\left[R_T^{\text{init}}\right] - \frac{\epsilon T}{6}, \tag{4}$$

where $\epsilon > 0$ is the gap between the instantaneous losses of the optimal action and a suboptimal action. The value of ϵ will be determined later.

Step 2: Obtaining a lower bound on $\mathbb{E}\left[R_T^{\text{init}}\right]$. Let S be the expected total number of action switches. Through careful information-theoretic analysis, we obtain the following (informal) lower bound on $\mathbb{E}\left[R_T^{\text{init}}\right]$ in terms of the number of switches S and extra observation budget B_{ex}:

$$\mathbb{E}\left[R_T^{\text{init}}\right] \geq \underbrace{\frac{\epsilon T}{2}}_{\mathbf{A}.\mathbf{1}} \underbrace{- C\frac{\epsilon^2 T}{\sqrt{K}}(\sqrt{S}+\sqrt{B_{\text{ex}}})}_{\mathbf{A}.\mathbf{2}} + \underbrace{S}_{\mathbf{A}.\mathbf{3}}, \tag{5}$$

where C is a positive term that contains some constants and poly-logarithmic terms of T.

We now explain each term in Eq. (5). Term $\mathbf{A}.\mathbf{1}$ reflects that without any useful information toward distinguishability, the learner may be stuck with a suboptimal action throughout the game, thus suffering $\Theta(\epsilon T)$ regret. Term $\mathbf{A}.\mathbf{2}$ roughly represents the amount of useful information for gaining distinguishability and thus reducing the regret: better distinguishability leads to a larger $\mathbf{A}.\mathbf{2}$ and thus a lower regret. Term $\mathbf{A}.\mathbf{3}$ is simply the switching costs incurred.

Step 3: Choosing a proper value of ϵ. Note that the lower bound in Eq. (5) is a quadratic function of $\sqrt{S}$. By finding the minimizer of this quadratic function, denoted by S^*, we can further obtain the following lower bound:

$$\mathbb{E}\left[R_T^{\text{init}}\right] \geq \underbrace{\frac{\epsilon T}{2}}_{\mathbf{B}.\mathbf{1}} - \underbrace{\frac{C^2}{4}\cdot\frac{\epsilon^4 T^2}{K}}_{\mathbf{B}.\mathbf{2}} - \underbrace{C\frac{\epsilon^2 T\sqrt{B_{\text{ex}}}}{\sqrt{K}}}_{\mathbf{B}.\mathbf{3}}. \tag{6}$$

It now remains to choose a proper value of ϵ based on B_{ex}. By considering two different cases ($B_{\text{ex}} = \Omega(K^{1/3}T^{2/3})$ and $B_{\text{ex}} = O(K^{1/3}T^{2/3})$) and choosing ϵ accordingly, we show that one of $\mathbf{B}.\mathbf{2}$ and $\mathbf{B}.\mathbf{3}$ dominates the other. Then, we can obtain the desired lower bound by combining these two cases. This completes the proof sketch. □

Remark 2. While we use the same instance construction method in Shi et al. (2022), the problem they study is very different from ours. In particular, their learning protocol and the definition of switching costs are different, and they do not consider an observation budget as we do. We present a detailed discussion about the key difference in Section 5.

3.3. Insights from Lower Bound Analysis

Next, we give some useful observations and important insights that can be obtained from the above proof sketch, in particular, from Eq. (5), which provides a *unified* view of the lower bound in online learning with bandit feedback and flexible extra observations within a budget.

As a warm-up, we begin with the standard bandit case (i.e., $B_{\text{ex}} = 0$), which has been extensively studied (Dekel et al.,

2013). Recall that under the current instance construction, bandit feedback provides useful information *only* when the learner switches actions. From Eq. (5), one can observe that there is a tradeoff between exploration and switching costs: on the one hand, in order to better explore and enjoy a lower regret, the learner has to switch frequently (i.e., a larger S) so as to gain more information (i.e., a larger $\mathbf{A.2}$); on the other hand, however, since the learner has to pay one unit of switching cost for each switch (contributing to $\mathbf{A.3}$), she should not switch too often. To strike the balance between the two, the best the learner can do is to switch $S^* := \Theta(K^{1/3}T^{2/3})$ times; otherwise, the regret can only be worse because S^* is the minimizer of the lower bound in Eq. (5). Finally, choosing ϵ to be $\widetilde{\Theta}(K^{1/3}T^{-1/3})$ in Eq. (6) yields the $\widetilde{\Omega}(K^{1/3}T^{2/3})$ bound for the bandit case.

Remark 3. The above discussion indicates that with switching costs, the worst-case hard instance restrains the learner from obtaining distinguishability from more than $\Theta(K^{1/3}T^{2/3})$ rounds (i.e., rounds associated with action switches) rather than T rounds as in the standard bandit learning setting (without switching costs). This is also the key reason why the minimax regret is worse in bandit learning with switching costs.

Next, we consider the first case: $B_{\text{ex}} = O(K^{1/3}T^{2/3})$. In this case, one might hope to obtain a smaller regret (compared to the bandit case) with the help of additional feedback. However, we will show that unfortunately, the gain from those additional observations is negligible for improving the regret order-wise, and hence, the previous $\widetilde{\Omega}(K^{1/3}T^{2/3})$ bound remains. To see this, let ϵ take the same value as in the bandit case (i.e., $\epsilon = \widetilde{\Theta}(K^{1/3}T^{-1/3})$) in Eq. (6); although $\mathbf{B.3}$ now becomes positive instead of zero (as in the bandit case), it is still dominated by $\mathbf{B.2}$, which results in the same $\widetilde{\Omega}(K^{1/3}T^{2/3})$ bound as in the bandit case.

We now turn to the second case: $B_{\text{ex}} = \Omega(K^{1/3}T^{2/3})$. In contrast to the previous case, due to a relatively large budget, the distinguishability provided by those extra observations (which do not contribute to switching costs) is no longer negligible. This leads to a smaller regret. In particular, by choosing $\epsilon = \widetilde{\Theta}(\sqrt{K/B_{\text{ex}}})$, we have $\mathbf{B.3}$ dominate $\mathbf{B.2}$ and obtain the desired lower bound. In other words, one can reduce the regret through free exploration enabled by such extra observations without incurring switching costs.

3.4. Fundamental Questions about Algorithm Design

The above insights we gain from the lower bound analysis can also shed light on the algorithm design. In fact, these motivate us to ask several fundamental questions, not only about how to achieve optimal regret but also about the role of feedback in online learning with switching costs, in terms of both the amount and type of feedback.

On the one hand, it is straightforward to achieve a matching upper bound when $B_{\text{ex}} = O(K^{1/3}T^{2/3})$. Specifically, one can simply ignore all the extra observations and use bandit feedback only, e.g., batched EXP3 (Arora et al., 2012), which enjoys a $\widetilde{\Theta}(K^{1/3}T^{2/3})$ regret. Although the bounds match, only $\Theta(T^{2/3})$ of the bandit feedback from the T rounds contribute to distinguishability due to the tradeoff introduced by switching costs (see Remark 3). Given this observation, it is natural to ask: ***(Q1)*** *Can one still achieve the same regret of $\widetilde{\Theta}(K^{1/3}T^{2/3})$ while using bandit feedback from $\Theta(K^{1/3}T^{2/3})$ rounds only? Moreover, how would regret scale with the amount of available feedback if the (bandit) feedback is even more limited (e.g., $O(K^{1/3}T^{2/3})$)?*

On the other hand, it remains largely unknown how to match the $\widetilde{\Omega}(T\sqrt{K/B_{\text{ex}}})$ bound when $B_{\text{ex}} = \Omega(K^{1/3}T^{2/3})$. Note that in the derivation of the lower bound, we *optimistically* view that *all* B_{ex} extra observations contribute to useful information toward distinguishability (see term $\mathbf{A.2}$ in Eq. (5)). To achieve this, however, one needs to answer an important question: ***(Q2)*** *How to carefully design a learning algorithm that can properly use these extra observations to indeed gain sufficient useful information toward distinguishability and match the lower bound? Moreover, since B_{ex} now dominates S^* (order-wise), can one still match the lower bound of $\widetilde{\Omega}(T\sqrt{K/B_{\text{ex}}})$ using B_{ex} extra observations only (i.e., not using any bandit feedback)?*

To address these fundamental questions, it turns out that it would be more instructive to consider a general setting where the learner has a budget for total observations (see Section 4) rather than extra observations. We will show that the results obtained for this general setting will naturally answer the aforementioned questions. In particular, we show that there exist learning algorithms that can match the lower bound (up to poly-logarithmic factors), hence concluding the minimax regret stated in Theorem 1.

4. Online Learning with Switching Costs under Total Observation Budget

In this section, we consider a more general setting of online learning with switching costs under a total observation budget. Specifically, at each round, the learner can freely choose to observe the loss of up to K actions (which may not necessarily include the action played), as long as the total number of observations over T rounds does not exceed the budget B, which is an integer in $[K, KT]$. Without loss of generality, we assume $B \geq K$. We aim to understand the role of feedback in this general setting by studying the following fundamental question: ***(Q3)*** *How does the minimax regret scale with the amount of available feedback in general? What is the impact of different types of feedback (bandit, full-information, etc.)?*

To proceed, we need some additional notations for this sec-

tion. Let $\mathcal{O}_t \subseteq [K]$ be the observation set, i.e., the set of actions whose loss the learner chooses to observe at round $t \in [T]$, and let N_{ob} be the total number of observations, i.e., $N_{\text{ob}} := \sum_{t=1}^{T} |\mathcal{O}_t|$. Naturally, we have $N_{\text{ob}} \leq B \leq KT$. For example, bandit feedback is a special case with $\mathcal{O}_t = \{X_t\}, \forall t \in [T]$ and $N_{\text{ob}} = B = T$; full-information feedback is another special case with $\mathcal{O}_t = [K], \forall t \in [T]$ and $N_{\text{ob}} = B = KT$. By slightly abusing the notation in this section, we also use R_T^* to denote the minimax regret over the set of all learning algorithms that satisfy the learning protocol specified in Section 2 and do not exceed the total observation budget B.

4.1. Minimax Regret

We first present the main result of this section and fully characterize the minimax regret for this general setting.

Theorem 2. *In the setting of online learning with switching costs under a total observation budget $B \in [K, KT]$, the minimax regret is given by $R_T^* = \widetilde{\Theta}(T\sqrt{K/B})$.*

Remark 4. This result answers the first part of question **(Q3)**: the minimax regret has a universal $\Theta(1/\sqrt{B})$ scaling across the full range of total budget B (see Fig. 1 (b)), compared to the phase transition in Section 3 (see Fig. 1 (a)).

To establish this result, we need to obtain both a lower bound and a matching upper bound. For the lower bound, it turns out that it suffices to use an existing lower bound, which was originally derived for standard online learning *without* switching costs. We restate this lower bound in Lemma 1.

Lemma 1. *(Seldin et al., 2014, Theorem 2) In the setting of online learning (without switching costs) under a total observation budget $B \in [K, KT]$, the minimax regret is lower bounded by $R_T^* = \Omega(T\sqrt{K/B})$.*

Naturally, this serves as a valid lower bound for the setting with switching costs we consider. In fact, we will show that this lower bound is tight (up to poly-logarithmic factors), which in turn offers the following important message.

Remark 5. If the learner can freely make observations over T rounds within the budget, introducing switching costs *does not increase* the intrinsic difficulty of the online learning problem in terms of the minimax regret.

Now, it only remains to show that there exist algorithms that can achieve a matching upper bound (up to poly-logarithmic factors), which will be the main focus of the next subsection.

4.2. Learning Algorithms and Upper Bounds

In this subsection, we show that there indeed exist algorithms that can achieve the lower bound in Lemma 1, which further implies the tight bound in Theorem 2. Instead of focusing on one particular algorithm, we first propose a generic algorithmic framework, which not only enables us to design various optimal learning algorithms in a unified way but also facilitates a fundamental understanding of the problem by distilling its key components.

Our generic framework builds upon the classic *Online Mirror Descent (OMD)* framework with negative entropy regularizer (also called the *Hedge* algorithm) (Littlestone & Warmuth, 1989) and incorporates the following three key components to tackle both switching costs and observation budget in a synergistic manner.

Batching Technique. The batching technique was originally proposed for addressing adaptive adversaries (Arora et al., 2012), but naturally provides low switching guarantees. We divide T rounds into batches and judiciously distribute the available observations across batches. That is, instead of consuming observations at every round as in standard online learning (which could even be infeasible when observation budget B is relatively small), we use observations only at a *single* round randomly sampled from each batch. One key step to obtain the desired regret guarantee is to feed the (unbiased estimate of) batch-average loss to the learning algorithm at the end of each batch. While this technique is borrowed from Shi et al. (2022), the problem setup we consider is very different (see Section 5).

Shrinking Dartboard (SD). SD is a calibrated technique for controlling the number of action switches in online learning under a lazy version of Hedge. That is, with a carefully crafted probability distribution, the action tends to remain unchanged across two consecutive rounds (Geulen et al., 2010) while preserving the same marginal distribution as in Hedge. In our algorithmic framework, we generalize this idea to the batching case with general feedback: the same action can be played across two consecutive batches (instead of across rounds), and it is no longer required to use only full-information feedback as in Geulen et al. (2010).

Feedback Type. Recall that the learner is allowed to freely request feedback within the total budget. Hence, our last component lies in the feedback type. That is, the learner has the flexibility to choose the observation set $\mathcal{O}_{u_b}$ (not limited to bandit or full-information feedback only). In order to achieve a matching upper bound, however, the choice of the observation set (i.e., the type of feedback) is crucial in some cases. We will elaborate on this in Section 4.3.

Putting these three components together, we arrive at our unified algorithmic framework, which is presented in Algorithm 1. Given the input T, K, and B of the problem, we need to determine the following input of the algorithm: the number of batches N, batch size τ, learning rate η, and indicator I_{SD} (Line 1), along with the initialization of some variables (Line 2). Throughout the game, we maintain a positive weight $W_b[k]$ for each action $k \in [K]$ in each batch $b \in [N]$. Both the weights and the action for each batch may

Algorithm 1 Batched Online Mirror Descent with (Optional) Shrinking Dartboard

1: **Input:** length of time horizon T, number of actions K, and observation budget B; determine the following based on T, K, and B: number of batches N, batch size $\tau = T/N$, learning rate η, and SD indicator I_{SD}
2: **Initialization:** action weight $W_1[k] = 1$ and action sampling distribution $w_1[k] = 1/K, \forall k \in [K]$; $\mathcal{O}_t = \emptyset, \forall t \in [T]$; choose $A_1 \in [K]$ uniformly at random
3: **for** batch $b = 1 : N$ **do**
4: Play action A_b throughout the current batch b, i.e., $X_t = A_b, \forall t = (b-1)\tau + 1, \ldots, b\tau$
5: Sample a round index u_b uniformly at random from integers in $[(b-1)\tau + 1, b\tau]$
6: Choose an observation set $\mathcal{O}_{u_b} \subseteq [K]$ (to be specified later) and observe the loss of each action in $\mathcal{O}_{u_b}$ at round u_b: $\{\ell_{u_b}[i] : i \in \mathcal{O}_{u_b}\}$
7: Construct unbiased estimate $\widehat{\ell}_b$ (to be specified later) of the batch-average loss $\sum_{t=(b-1)\tau+1}^{b\tau} \ell_t/\tau$
8: Run OMD update: update the weight of each action: $W_{b+1}[k] = W_b[k] \cdot \exp(-\eta \cdot \widehat{\ell}_b[k])$ and the sampling probability: $w_{b+1}[k] = \frac{W_{b+1}[k]}{\sum_{i=1}^K W_{b+1}[i]}, \forall k \in [K]$
9: With probability $I_{\text{SD}} \cdot \exp(-\eta \cdot \widehat{\ell}_b)$, keep action $A_{b+1} = A_b$; otherwise, sample action $A_{b+1} \sim w_{b+1}$
10: **end for**

be updated only between two consecutive batches. Hence, in each batch b, we keep playing the chosen action A_b until the end of the batch (Line 4); we sample a round u_b uniformly at random from the current batch (Line 5) and choose an observation set $\mathcal{O}_{u_b}$ in a certain way (to be specified later) such that the loss of each action in $\mathcal{O}_{u_b}$ will be observed at round u_b (Line 6). We then construct an unbiased estimate (Line 7), denoted by $\widehat{\ell}_b = (\widehat{\ell}_b[1], \ldots, \widehat{\ell}_b[K])$, of the batch-average loss $\sum_{t=(b-1)\tau+1}^{b\tau} \ell_t/\tau$ (which depends on the choice of $\mathcal{O}_{u_b}$ and will be specified later) and then update the weight and sampling probability of each action accordingly: $W_{b+1}[k] = W_b[k] \cdot \exp(-\eta \cdot \widehat{\ell}_b[k])$ and $w_{b+1}[k] := W_{b+1}[k] / \sum_{i=1}^K W_{b+1}[i]$ (Line 8). Finally, we determine action A_{b+1} for the next batch (Line 9). Specifically, if the SD indicator $I_{\text{SD}} = 0$, probability $I_{\text{SD}} \cdot \exp(-\eta \cdot \widehat{\ell}_b)$ is always zero, and hence, action A_{b+1} is sampled using fresh randomness with probability proportional to action weights as normally done in Hedge: sample A_{b+1} following distribution $w_{b+1} = (w_{b+1}[1], \ldots, w_{b+1}[K])$. If the SD indicator $I_{\text{SD}} = 1$, with probability $\exp(-\eta \cdot \widehat{\ell}_b)$, we keep the current action for the next batch (i.e., $A_{b+1} = A_b$); otherwise, we sample a new action A_{b+1} following distribution w_{b+1}.

With Algorithm 1 in hand, we are ready to introduce several specific instantiations and study their regret guarantees. In particular, for each instantiation we will specify the choice of the following parameters: number of batches N, batch size τ, learning rate η, SD indicator I_{SD}, and observation set $\mathcal{O}_{u_b}$. In the following, we first demonstrate one simple instantiation that uses full-information feedback only. Then, we show how to generalize this instantiation using more flexible feedback (i.e., not limited to full information only) while achieving the same performance guarantee.

Instantiation via Full-information Feedback. In this instantiation of Algorithm 1, we receive full-information feedback at a randomly selected round u_b in each batch b (i.e., $\mathcal{O}_{u_b} = [K]$ and $\widehat{\ell}_b = \ell_{u_b}$) and SD is turned on (i.e., $I_{\text{SD}} = 1$). At a high level, this can be viewed as a batched generalization of the original SD algorithm (Geulen et al., 2010) with $N = B/K$ batches (since we have K observations in each batch), and hence, the corresponding batch size is $\tau = T/N = KT/B$. For ease of exposition, we assume that N and τ are integers. Specifically, we have $N = B/K$, $\tau = KT/B$, $\eta = \sqrt{\frac{2 \ln K}{3B}}$, $I_{\text{SD}} = 1$, $\mathcal{O}_{u_b} = [K]$, and $\widehat{\ell}_b = \ell_{u_b}$. We use π_{full} to denote this instantiation and present its regret upper bound in Proposition 2. The proof is provided in Appendix C.

Proposition 2. *The worst-case regret under algorithm π_{full} is upper bounded by $R_T^{\pi_{\text{full}}} = O(T\sqrt{K \ln K/B})$.*

Remark 6. This result immediately implies an upper bound of the minimax regret: $R_T^* = O(T\sqrt{K \ln K/B})$, which, along with the lower bound in Lemma 1, further implies the tight bound in Theorem 2. Note that there is an additional $\sqrt{\ln K}$ factor in the upper bound. This shares the same pattern as in the setting even without switching costs (see Seldin et al. (2014, Theorem 1)), where the achieved upper bound also has an additional $\sqrt{\ln K}$ factor.

Remark 7. For the previous setting considered in Section 3, the above result also implies an upper bound of the minimax regret: $\widetilde{O}(T\sqrt{K/B_{\text{ex}}})$, when $B_{\text{ex}} = \Omega(K^{1/3}T^{2/3})$, by simply ignoring all bandit feedback (i.e., $B = B_{\text{ex}}$). On the other hand, as discussed in Section 3.4, when $B_{\text{ex}} = O(K^{1/3}T^{2/3})$, one can simply ignore extra observations and use pure bandit feedback only (e.g., batched EXP3 (Arora et al., 2012)) to achieve a $\widetilde{O}(K^{1/3}T^{2/3})$ regret. Combining these results, along with the lower bound in Proposition 1, implies the tight bound in Theorem 1. Moreover, this also answers question **(Q2)** raised in Section 3.

The result of our first instantiation shows that the optimal regret can indeed be achieved (up to a $\sqrt{\ln K}$ factor) when full-information feedback is employed. However, we can also show that the use of full-information feedback is not essential. In fact, it suffices to have an observation set chosen uniformly at random from all subsets of $[K]$ with the same cardinality, which leads to a more flexible instantiation of Algorithm 1 presented below.

Instantiation via Flexible Feedback. In this instantiation,

instead of having $|\mathcal{O}_{u_b}| = K$ as under full-information feedback, we allow $|\mathcal{O}_{u_b}| = M \leq K$. The key to this flexibility is a careful construction of an unbiased estimate of the batch-average loss (i.e., $\widehat{\ell}_b$). Specifically, let M be any integer that satisfies $M \in [K]$ if $B < T$ and $M \in [\lceil B/T \rceil, K]$ if $B \geq T$.[2] Then, we have $N = B/M$, $\tau = T/N = MT/B$, $\eta = M\sqrt{\frac{2\ln K}{3KB}}$, $I_{\mathrm{SD}} = 1$, $\mathcal{O}_{u_b}$ is chosen uniformly at random from $\{U \in 2^{[K]} : |U| = M\}$, and $\widehat{\ell}_b[k] = \mathbb{I}\{k \in \mathcal{O}_{u_b}\} \cdot \frac{\ell_{u_b}[k]}{M/K}$ for all $k \in [K]$. We use π_{flex} to denote this instantiation and present its regret upper bound in Proposition 3. The proof is provided in Appendix D.

Proposition 3. *The worst-case regret under algorithm π_{flex} is upper bounded by $R_T^{\pi_{\mathrm{flex}}} = O(T\sqrt{K \ln K/B})$.*

An astute reader may already notice that in the above flexible instantiation, while the number of observations can be one (i.e., $|\mathcal{O}_{u_b}| = 1$), it is not the same as standard bandit feedback. This is because here, $\mathcal{O}_{u_b}$ needs to be chosen uniformly at random rather than simply being the action played in that batch (i.e., $\mathcal{O}_{u_b} = \{A_b\}$) as in the standard bandit setting (with a batch size of one). Motivated by this subtle difference, we will devote the next subsection to studying the impact of feedback type.

4.3. Impact of Feedback Type

In this subsection, we study the impact of feedback type by presenting another instantiation of Algorithm 1 via pure bandit feedback only. In this case, we naturally have $B \leq T$.

Instantiation via Bandit Feedback. This instantiation is a generalized version of batched EXP3 (Arora et al., 2012) with *flexible batch size*. Specifically, we have $N = B$, $\tau = T/B$, $\eta = \sqrt{\frac{2\ln K}{BK}}$, $I_{\mathrm{SD}} = 0$, $\mathcal{O}_{u_b} = \{A_b\}$, and $\widehat{\ell}_b[k] = \mathbb{I}\{k \in \mathcal{O}_{u_b}\} \cdot \frac{\ell_{u_b}[k]}{w_b[k]}$ for all $k \in [K]$. We use π_{b} to denote this instantiation. When $B = O(K^{1/3}T^{2/3})$, we obtain a regret upper bound for π_{b} and state it in Proposition 4. The proof is provided in Appendix E.

Proposition 4. *When $B = O(K^{1/3}T^{2/3})$, the worst-case regret under algorithm π_{b} is upper bounded by $R_T^{\pi_{\mathrm{b}}} = O(T\sqrt{K \ln K/B})$.*

Remark 8. This result is encouraging, in the sense that when $B = O(K^{1/3}T^{2/3})$, even using pure bandit feedback can achieve the optimal minimax regret of $\widetilde{\Theta}(T\sqrt{K/B})$. This result also answers question **(Q1)** raised in Section 3. First, it captures the regret scaling with respect to the amount of bandit feedback (i.e., still $\Theta(1/\sqrt{B})$) when B is relatively small. Second, it implies that to achieve a regret of $\widetilde{\Theta}(K^{1/3}T^{2/3})$, it suffices to use bandit feedback from only $B = \Theta(K^{1/3}T^{2/3})$ rounds rather than all T rounds as in the classic algorithms (Arora et al., 2012). The same minimax regret at these two endpoints ($B = \Theta(K^{1/3}T^{2/3})$ and $B = T$) further implies that if only bandit feedback is allowed, the minimax regret is also $\widetilde{\Theta}(K^{1/3}T^{2/3})$ when $B = \Omega(K^{1/3}T^{2/3})$ (i.e., in-between the two endpoints). In this case, bandit feedback is *no longer sufficient* to achieve the optimal minimax regret of $\widetilde{\Theta}(T\sqrt{K/B})$, although full-information and flexible feedback can still achieve this optimal minimax regret (see Propositions 2 and 3). Clearly, this shows the crucial impact of different types of feedback (when the total budget B is large), which answers the second part of question **(Q3)**. On the other hand, however, a straightforward result (Proposition 5 in Appendix F), along with Propositions 2 and 3 and Lemma 1, shows that in the standard setting without switching costs, all three types of feedback can achieve optimal regret in the full range of B. This reveals that the impact of feedback type is partly due to switching costs. We also summarize these results in Table 1.

Remark 9. Under bandit feedback, adopting a different regularizer called *Tsallis entropy* (Audibert & Bubeck, 2009) to the OMD framework could further remove the $\sqrt{\ln K}$ factor in the upper bound from Proposition 4 and exactly match the lower bound (order-wise) presented in Lemma 1.

5. Related Work

In this section, we present detailed discussions on several lines of research that are most relevant to ours. We omit the discussion on bandit and expert problems with switching costs as we have discussed this line of work in Section 1.

Online Learning with Total Observation Budget. In this line of research, the focus is on regret minimization when feedback is not always available and hence "limited" within a total budget. For example, in the so-called "label efficient (bandit) game" (Cesa-Bianchi et al., 2004; Audibert & Bubeck, 2010), the learner can ask for full-information/bandit feedback from no more than $m \in [1, T]$ round(s). It is shown that the tight optimal regrets are $\Theta(T\sqrt{\ln K/m})$ and $\Theta(T\sqrt{K/m})$ under full-information and bandit feedback, respectively. Seldin et al. (2014) also considers a total observation budget in online learning, where the learner can freely request feedback, as long as the total amount of observed losses does not exceed the given total budget B. They establish a tight characterization of the minimax regret in their setting (i.e., $\widetilde{\Theta}(T\sqrt{K/B})$). However, they do not consider switching costs, nor the case when the total observation budget is smaller than T in their algorithm design. Interestingly, we show that introducing switching costs *does not increase* the intrinsic difficulty of online learning in the sense that the minimax regret remains $\widetilde{\Theta}(T\sqrt{K/B})$, but the feedback type becomes crucial.

Bandits with Additional Observations. Yun et al. (2018) considers the bandit setting with additional observations, where the learner can freely make $n \in [0, K-1]$ obser-

[2] To fully use the budget, M cannot be too small when $B \geq T$.

vations at each round in addition to the bandit feedback. Hence, this can be viewed as a special case of online learning with a total observation budget (Seldin et al., 2014). That is, a total of $(n+1)T$ observations are used in a particular way (i.e., bandit plus extra observations). They present a tight characterization of the scaling of the minimax regret with respect to K, T, and n. Similar to Seldin et al. (2014), however, switching costs are not considered.

Online Learning with Switching Costs and Feedback Graphs. Arora et al. (2019) considers online learning with switching costs and feedback graphs, where given a feedback graph G, the learner observes the loss associated with the neighboring action(s) of the chosen action (including itself). However, the feedback graph is given and hence the additional feedback is *not* of the learner's choice. Arora et al. (2019) shows that in this setting, the minimax regret is $\widetilde{\Theta}(\gamma(G)^{1/3}T^{2/3})$, where $\gamma(G)$ is the domination number of the feedback graph G. Hence, the dependency on T remains the same as in the standard bandit setting without additional observations (i.e., $\widetilde{\Theta}(T^{2/3})$). On the contrary, in the setting we consider, the learner can freely decide the loss of which actions to observe, which leads to different (and more interesting) regret bounds.

Online Learning with Limited Switches. Altschuler & Talwar (2018) considers online learning with limited switches. In contrast to the settings with switching costs, here the learner does not pay additional losses for switching actions; instead, the total number of switches allowed is capped at S. Compared to our setting, a key difference is that switching is a constraint rather than a penalty added to the loss/cost function. They show that in the bandit setting, the minimax regret is $\widetilde{\Theta}(T\sqrt{K/S})$, i.e., the regret improves as the switching budget increases; in the expert setting, however, there is a phase-transition phenomenon: while the minimax regret is $\widetilde{\Theta}(T\ln K/S)$ when $S = O(\sqrt{T\ln K})$, it remains $\widetilde{\Theta}(\sqrt{T\ln K})$ when $S = \Omega(\sqrt{T\ln K})$.

Online Learning against Adaptive Adversaries. Online learning with switching costs can also be viewed as a special case of *learning against adaptive adversaries*, where the losses at round t are adapted to actions taken at both rounds t and $t-1$ (in contrast to the oblivious adversaries we consider). Such adversaries have a *bounded memory* (of size one), in the sense that they could adapt only up to the *most recent* action, instead of any history in the earlier rounds (Cesa-Bianchi et al., 2013). Adopting the multi-scale random walk argument in Dekel et al. (2013), it has been shown that against *adaptive adversaries with a memory of size one*, the *minimax policy regret* is $\widetilde{\Theta}(T^{2/3})$ under *both* bandit feedback (Cesa-Bianchi et al., 2013) and full-information feedback (Feng & Loh, 2018). This is fundamentally different from the special case with switching costs, where the minimax regret is different under bandit feedback and full-information feedback ($\widetilde{\Theta}(T^{2/3})$ vs. $\widetilde{\Theta}(\sqrt{T})$).

Stochastic Bandits and the Best of Both Worlds. Note that the above discussions have been focused on the adversarial setting. There is another body of work focused on the stochastic setting (see, e.g., Auer et al. (2002a); Auer (2003); Simchi-Levi & Xu (2019)), where the loss/reward follows some fixed distribution rather than being generated arbitrarily by an adversary. Hence, it is very different from the adversarial setting we consider. An interesting line of work has been focused on designing algorithms that can perform well in both adversarial and stochastic settings, thus achieving *the best of both worlds* (see, e.g., Bubeck & Slivkins (2012); Zimmert et al. (2019)).

Other Related Work. In Shi et al. (2022), a novel bandit setting with switching costs and additional feedback has been considered. Specifically, the learner maintains an "action buffer" for each round, which is a subset of actions with fixed cardinality $m \in [K]$, and the learner can only take an action from this buffer set. Their switching cost can be roughly viewed as how much change is made to this buffer set throughout the game – replacing an action in the buffer set incurs a constant cost. While the learner can observe the losses of all the actions in this buffer set for free, the learner can also choose to receive full-information feedback (i.e., observing the losses of all actions rather than just actions in the buffer set) by paying another (larger) constant cost. Although we draw inspiration from their work for deriving the lower bound and designing algorithms, both their problem setup and regret definition are very different from ours, and more importantly, they do not consider observation budget.

6. Conclusion

Our work is motivated by a well-known gap in the minimax regret under bandit feedback and full-information feedback in online learning with switching costs. We attempted to fundamentally understand the role of feedback by studying two cases of observation budget: (i) bandit feedback plus an extra observation budget and (ii) a total observation budget. Our findings reveal that both the amount and type of feedback play crucial roles when there are switching costs.

One interesting future direction is to consider stronger high-probability regret guarantees (Neu, 2015). Another direction is to achieve *the best of both worlds* guarantees for regrets with switching costs (Rouyer et al., 2021; Amir et al., 2022).

Acknowledgments

We thank the anonymous paper reviewers for their insightful feedback. This work is supported in part by the NSF grants under CNS-2112694 and CNS-2153220.

References

Altschuler, J. M. and Talwar, K. Online learning over a finite action set with limited switching. *ArXiv*, abs/1803.01548, 2018.

Amir, I., Azov, G., Koren, T., and Livni, R. Better best of both worlds bounds for bandits with switching costs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15800–15810, 2022.

Arora, R., Dekel, O., and Tewari, A. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1747–1754, 2012.

Arora, R., Marinov, T. V., and Mohri, M. Bandits with feedback graphs and switching costs. *Advances in Neural Information Processing Systems*, 32, 2019.

Audibert, J.-Y. and Bubeck, S. Minimax policies for adversarial and stochastic bandits. In *Annual Conference Computational Learning Theory*, 2009.

Audibert, J.-Y. and Bubeck, S. Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.*, 11:2785–2836, 2010.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, 2003.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32:48–77, 2002b.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 42–1. JMLR Workshop and Conference Proceedings, 2012.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Cesa-Bianchi, N., Lugosi, G., and Stoltz, G. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51:2152–2162, 2004.

Cesa-Bianchi, N., Dekel, O., and Shamir, O. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Dekel, O., Ding, J., Koren, T., and Peres, Y. Bandits with switching costs: T^{2/3} regret. *arXiv preprint arXiv:1310.2997*, 2013.

Devroye, L., Lugosi, G., and Neu, G. Prediction by random-walk perturbation. In *Annual Conference Computational Learning Theory*, 2013.

Feng, Z. and Loh, P.-L. Online learning with graph-structured feedback against adaptive adversaries. *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 931–935, 2018.

Geulen, S., Vöcking, B., and Winkler, M. Regret minimization for online buffering problems using the weighted majority algorithm. *Electron. Colloquium Comput. Complex.*, TR10, 2010.

Hazan, E. Introduction to online convex optimization. *Found. Trends Optim.*, 2:157–325, 2016.

Kaplan, S. *Power plants: characteristics and costs*. DIANE Publishing, 2011.

Littlestone, N. and Warmuth, M. K. The weighted majority algorithm. *30th Annual Symposium on Foundations of Computer Science*, pp. 256–261, 1989.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *NIPS*, 2015.

Orabona, F. A modern introduction to online learning. *ArXiv*, abs/1912.13213, 2019.

Rouyer, C., Seldin, Y., and Cesa-Bianchi, N. An algorithm for stochastic and adversarial bandits with switching costs. In *International Conference on Machine Learning*, 2021.

Seldin, Y., Bartlett, P. L., Crammer, K., and Abbasi-Yadkori, Y. Prediction with limited advice and multiarmed bandits with paid observations. In *International Conference on Machine Learning*, 2014.

Shi, M., Lin, X., and Jiao, L. Power-of-2-arms for bandit learning with switching costs. In *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 131–140, 2022.

Simchi-Levi, D. and Xu, Y. Phase transitions in bandits with switching constraints. *ERN: Other Econometrics: Mathematical Methods & Programming (Topic)*, 2019.

Yao, A. C.-C. Probabilistic computations: Toward a unified measure of complexity. *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pp. 222–227, 1977.

Yun, D., Proutière, A., Ahn, S., Shin, J., and Yi, Y. Multi-armed bandit with additional observations. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2:1 – 22, 2018.

Zhang, Y., Murata, M., Takagi, H., and Ji, Y. Traffic-based reconfiguration for logical topologies in large-scale wdm optical networks. *Journal of Lightwave Technology*, 23: 2854–2867, 2005.

Zimmert, J., Luo, H., and Wei, C.-Y. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pp. 7683–7692. PMLR, 2019.

Algorithm 2 Loss Sequence Generation Method (Shi et al., 2022)

1: **Input:** suboptimality gap ϵ and noise variance σ^2
2: **Initialization:** choose the identity of optimal action k^* uniformly at random from $[K]$; initialize Gaussian process $G(t) = 0, \forall t \geq 0$; define functions $\delta(t) := \max\{i \geq 0 : 2^i \text{ divides } t\}$ and $\rho(t) := t - 2^{\delta(t)}, \forall t \geq 0$
3: **for** time $t = 1 : T$ **do**
4: $G(t) = G(\rho(t)) + \xi(t)$, where each $\xi(t)$ is an *i.i.d.* sample from $\mathcal{N}(0, \sigma^2)$
5: $\ell_t^{\text{init}}[k] = G(t) - \epsilon \cdot \mathbb{I}_{\{k=k^*\}} + \gamma_k(t), \forall k \in [K]$, where $\gamma_k(t)$ is an *i.i.d.* sample from Gaussian distribution $\mathcal{N}(0, \sigma^2)$
6: $\ell_t[k] = \arg\min_{z \in [0,1]} |z - \ell_t^{\text{init}}[k]|, \forall k \in [K]$
7: **end for**

A. Motivating Examples for Online Learning with Switching Costs and Observation Budget

Consider a retail company that uses online learning to improve its website user interface (UI) design in order to attract more users. In this case, actions correspond to different UI designs. First, switching costs should be taken into account as frequently changing the website interface may become annoying to users. To evaluate other actions (different UI designs), the company can run an A/B test and display different interfaces to separate and relatively small groups of users so that the feedback of other actions is also available (in addition to the one displayed to the main and large population). However, each different website needs additional resources to be deployed and maintained, and hence, one may want to impose a total observation budget.

Another example would be Machine Learning as a Service (MLaaS). Consider a company that uses large ML models for jobs like prediction, chatbots, etc. They may train several different models and dynamically choose the best one via online learning. Changing the deployed ML model is not free: the new model needs to be loaded (which could be costly, especially nowadays when the number of parameters is quite large), and other components in the pipeline may also need to be adjusted accordingly. As a result, it is natural that redeploying or updating model components would incur a cost. While the performance of the deployed model is easily accessible, the company can also run these jobs using other models that are not deployed in the production system, to receive additional feedback. However, running these jobs consumes additional resources (e.g., computing and energy), which is not free either. Therefore, one may want to impose a budget on the number of additional observations (i.e., evaluations).

B. Proof of Proposition 1

Proof of Proposition 1. Our proof is built on Yao's minimax principle (Yao, 1977). That is, we will establish a lower bound on the expected regret for any deterministic algorithm over a class of randomly generated loss sequences, which further implies the same lower bound for any randomized algorithm over some loss sequence in this class.

To begin with, we would like to give some high-level ideas about the proof. Note that while the loss sequence generation method we adopt will be the same as Algorithm 1 in Shi et al. (2022), we need a different analysis to establish the lower bound due to a different setting we consider. Specifically, in the original loss sequence construction based on multi-scale random walk (Dekel et al., 2013), the optimal action k^* has the lowest loss at all T rounds. With bandit feedback, useful information toward distinguishability (between the optimal action and suboptimal actions) is gained only when the learner switches between actions. With full-information feedback, however, the learner can immediately identify the optimal action even at one round only. Therefore, to construct a hard instance (i.e., loss sequence) for the setting where the learner is equipped with additional observations beyond the bandit feedback, Shi et al. (2022) introduced an action-dependent noise in addition to the original loss (which is called the *hidden loss*). Now, the learner's information comes from two parts. On the one hand, the learner still gains distinguishability from switches (which is related to hidden losses). On the other hand, conditioning on hidden losses, the extra observations also provide additional information. Combining two parts together, we obtain a lower bound related to both the number of switches and the number of extra observations. For convenience, we restate this loss sequence generation method in Algorithm 2. Specifically, we first generate the sequence $\{G(t)\}_t$ according to the random walk design (Line 4). Next, we determine the loss before truncation, i.e., $\ell_t^{\text{init}}[k]$ (Line 5). We first add an action-dependent noise $\gamma_k(t)$ (which is an *i.i.d.* Gaussian random variable) to $G(t)$ for each action $k \in [K]$. And then, for the optimal action k^* only, we will further subtract ϵ (which is determined in the very beginning as an input to the algorithm) from the value obtained after adding $\gamma_k(t)$. Finally, we truncate each $\ell_t^{\text{init}}[k]$ onto range $[0, 1]$ and obtain $\ell_t[k]$ (Line 6).

Next, we give some additional notations needed for this proof. For any $k \in [K]$, let $\mathbb{P}_k$ denote the conditional probability

measure given the special (i.e., optimal) action $k^* = k$, i.e., $\mathbb{P}_k(\cdot) := \mathbb{P}(\cdot|k^* = k)$. As a special case, when $k^* = 0$, $\mathbb{P}_0$ denotes the conditional probability measure where all the actions are identical and there is no special action. Let $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot|k^* = k]$ denote the conditional expectation under measure $\mathbb{P}_k$, and let $\mathbb{E}[\cdot] := \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_k[\cdot]$. Let $\ell_{1:T}^{\text{ob}}$ denote the observed loss sequence throughout the game. For two probability distributions $\mathcal{P}$ and $\mathcal{Q}$ on the same space, let $D_{\text{KL}}(\mathcal{P}\|\mathcal{Q}) := \mathbb{E}_{x\sim\mathcal{P}}[\log(d\mathcal{P}(x)/d\mathcal{Q}(x))]$ denote the Kullback-Leibler (KL) divergence (i.e., relative entropy) between $\mathcal{P}$ and $\mathcal{Q}$, and let $D_{\text{TV}}(\mathcal{P}\|\mathcal{Q}) := \sup\{\mathcal{P}(A) - \mathcal{Q}(A) : A \text{ measurable}\}$ denote the total variation distance between $\mathcal{P}$ and $\mathcal{Q}$.

Let $S_t^k := \mathbb{I}_{\{X_{\rho(t)}=k, X_t\neq k\}} + \mathbb{I}_{\{X_{\rho(t)}\neq k, X_t=k\}}$ be the indicator of whether it is switched from *or* to action k between rounds $\rho(t)$ and t, let $\bar{S}^k := \sum_{t=1}^{T} S_t^k$ be the total number of action switches from *or* to action k, let $\bar{S}$ be the total number of action switches, i.e., $\bar{S} := \sum_{t=1}^{T}\mathbb{I}_{\{X_t\neq X_{t-1}\}} = \sum_{k=1}^{K}\bar{S}^k/2$, and let N_{ex}^t be the number of extra observations made at round t in addition to the bandit feedback. Then, we naturally have $N_{\text{ex}}^t \in [0, K-1]$ and $\sum_{t=1}^{T} N_{\text{ex}}^t \leq B_{\text{ex}}$ since the learning algorithm makes no more than B_{ex} extra observations in total. Let R_T be the regret of the deterministic learning algorithm interacting with the loss sequence $\ell_{1:T}$, and let R_T^{init} be the (hypothetical) regret on *the same action sequence* with respect to loss sequence $\ell_{1:T}^{\text{init}}$.

In the following proof, we need Lemmas A.1 and A.4 of Shi et al. (2022) and Lemma 2 of Dekel et al. (2013). We restate these three results as Lemmas 2, 3, and 4, respectively.

Lemma 2. *(Shi et al., 2022, Lemma A.1 (restated)) The KL divergence between $\mathbb{P}_0\left(\ell_{1:T}^{\text{ob}}\right)$ and $\mathbb{P}_k\left(\ell_{1:T}^{\text{ob}}\right)$ can be upper bounded as follows:*

$$
\begin{aligned}
&D_{\text{KL}}\left(\mathbb{P}_0\left(\ell_{1:T}^{\text{ob}}\right)\|\mathbb{P}_k\left(\ell_{1:T}^{\text{ob}}\right)\right)\\
&\leq \sum_{t=1}^{T}\Bigg[\mathbb{P}_0\left(N_{\text{ex}}^t=0, S_t^k=1\right)\cdot\frac{\epsilon^2}{2\sigma^2} + \sum_{j=1}^{K-1}\mathbb{P}_0\left(N_{\text{ex}}^t=j, S_t^k=0, X_t\neq k\right)\cdot\frac{\epsilon^2}{2\sigma^2}\\
&\quad + \sum_{j=1}^{K-1}\mathbb{P}_0\left(N_{\text{ex}}^t=j, S_t^k=0, X_t=k\right)\cdot\frac{j\epsilon^2}{2\sigma^2} + \sum_{j=1}^{K-1}\mathbb{P}_0\left(N_{\text{ex}}^t=j, S_t^k=1, X_t\neq k\right)\cdot\frac{2\epsilon^2}{2\sigma^2}\\
&\quad + \sum_{j=1}^{K-1}\mathbb{P}_0\left(N_{\text{ex}}^t=j, S_t^k=1, X_t=k\right)\cdot\frac{(j+1)\epsilon^2}{2\sigma^2}\Bigg].
\end{aligned}
$$

Lemma 2 is obtained by considering five disjoint cases (which corresponds to the five terms on the Right-Hand-Side (RHS) in terms of different values of N_{ex}^t, S_t^k, and X_t. This lemma reveals the relationship between the KL divergence and the number of switches and the number of extra observations and will be used for deriving Eq. (7).

Lemma 3. *(Shi et al., 2022, Lemma A.4 (restated)) Consider the instance construction in Algorithm 2. Suppose that we have $\epsilon \leq 1/6$ and $\sigma = 1/(9\log_2 T)$. Then, the difference between $\mathbb{E}[R_T]$ and $\mathbb{E}\left[R_T^{\text{init}}\right]$ can be bounded as follows:*

$$
\mathbb{E}\left[R_T^{\text{init}}\right] - \mathbb{E}[R_T] \leq \frac{\epsilon T}{6}.
$$

Although the multi-scale random walk serves as a powerful and convenient tool to help us obtain the desired lower bound, an issue is that the losses could lie out of the range $[0, 1]$, which does not satisfy our problem setup. That is, based on the random walk, we can derive a lower bound on $\mathbb{E}\left[R_T^{\text{init}}\right]$, with respect to a possibly unbounded loss sequence $\ell_{1:T}^{\text{init}}$. However, our goal is to obtain a lower bound with respect to the bounded losses. To get around this issue, Lemma 3 presents a useful result: if ϵ and σ satisfy certain conditions, then the difference between $\mathbb{E}\left[R_T^{\text{init}}\right]$ and $\mathbb{E}[R_T]$ will not be too large, which is sufficient to give us the desired result.

Lemma 4. *(Dekel et al., 2013, Lemma 2 (restated)) Under the instance construction in Algorithm 2, the following is satisfied:*

$$
\sum_{t=1}^{T}\mathbb{P}_0\left(S_t^k=1\right) = \mathbb{E}\left[\sum_{t=1}^{T} S_t^k\right] \leq (\lfloor\log_2 T\rfloor + 1)\cdot\mathbb{E}_0\left[\bar{S}^k\right].
$$

Furthermore, it can be bounded by $2\log_2 T\cdot\mathbb{E}_0\left[\bar{S}^k\right]$ for a sufficiently large T.

Remark 10. Lemma 4 holds regardless of whether the action-dependent is added or not. Therefore, it is true under the instance constructions from both Dekel et al. (2013) and Shi et al. (2022).

Lemma 4 relies on careful design of the random walk. We refer interested readers to (Dekel et al., 2013, Section 3) for technical details. In the following proof, we will first bound the KL divergence in part by $\mathbb{E}[\sum_{t=1}^{T} S_t^k]$. This term is different from the switching costs we consider, as S_t^k roughly denotes the switch between rounds $\rho(t)$ and t, *rather than between two consecutive rounds*. To handle this difference, Lemma 4 builds a connection between the two and will be used for deriving Eq. (7).

With the above three restated lemmas, we are now ready to derive an upper bound on the KL divergence between $\mathbb{P}_0\left(\ell_{1:T}^{\text{ob}}\right)$ and $\mathbb{P}_k\left(\ell_{1:T}^{\text{ob}}\right)$. In particular, for any $k \in [K]$, we have

$$
\begin{aligned}
&D_{\text{KL}}\left(\mathbb{P}_0\left(\ell_{1:T}^{\text{ob}}\right) \| \mathbb{P}_k\left(\ell_{1:T}^{\text{ob}}\right)\right) \\
&\overset{\text{(a)}}{\leq} \sum_{t=1}^{T}\Bigg[\mathbb{P}_0\left(N_{\text{ex}}^t = 0, S_t^k = 1\right)\cdot\frac{\epsilon^2}{2\sigma^2} + \sum_{j=1}^{K-1}\mathbb{P}_0\left(N_{\text{ex}}^t = j, S_t^k = 0, X_t \neq k\right)\cdot\frac{\epsilon^2}{2\sigma^2} \\
&\quad + \sum_{j=1}^{K-1}\mathbb{P}_0\left(N_{\text{ex}}^t = j, S_t^k = 0, X_t = k\right)\cdot\frac{j\epsilon^2}{2\sigma^2} + \sum_{j=1}^{K-1}\mathbb{P}_0\left(N_{\text{ex}}^t = j, S_t^k = 1, X_t \neq k\right)\cdot\frac{2\epsilon^2}{2\sigma^2} \\
&\quad + \sum_{j=1}^{K-1}\mathbb{P}_0\left(N_{\text{ex}}^t = j, S_t^k = 1, X_t = k\right)\cdot\frac{(j+1)\epsilon^2}{2\sigma^2}\Bigg] \\
&\overset{\text{(b)}}{\leq} \sum_{t=1}^{T}\Bigg[\mathbb{P}_0\left(N_{\text{ex}}^t = 0, S_t^k = 1\right)\cdot\frac{\epsilon^2}{2\sigma^2} + \sum_{j=1}^{K-1}\frac{j\epsilon^2}{\sigma^2}\cdot\left(\mathbb{P}_0\left(N_{\text{ex}}^t = j, S_t^k = 0, X_t \neq k\right) + \mathbb{P}_0\left(N_{\text{ex}}^t = j, S_t^k = 0, X_t = k\right)\right. \\
&\quad \left.+\mathbb{P}_0\left(N_{\text{ex}}^t = j, S_t^k = 1, X_t \neq k\right) + \mathbb{P}_0\left(N_{\text{ex}}^t = j, S_t^k = 1, X_t = k\right)\right)\Bigg] \\
&\overset{\text{(c)}}{\leq} \sum_{t=1}^{T}\left[\mathbb{P}_0\left(S_t^k = 1\right)\cdot\frac{\epsilon^2}{2\sigma^2} + \sum_{j=1}^{K-1}\mathbb{P}_0\left(N_{\text{ex}}^t = j\right)\cdot\frac{j\epsilon^2}{\sigma^2}\right] \\
&\overset{\text{(d)}}{\leq} 2\log_2 T\cdot\frac{\epsilon^2}{2\sigma^2}\cdot\left(\mathbb{E}_0\left[\bar{S}^k\right] + 2B_{\text{ex}}\right),
\end{aligned} \tag{7}
$$

where (a) is from Lemma 2, (b) is obtained by enlarging the last four terms using the fact that $2 \leq j+1 \leq 2j, \forall j \geq 1$, (c) is obtained by applying the monotonicity property of probability to the first term and merging the last four disjoint events, and (d) is from Lemma 4 and the fact that $\sum_{t=1}^{T}\sum_{j=1}^{K-1}\mathbb{P}_0(N_{\text{ex}}^t = j)\cdot j = \sum_{t=1}^{T}\mathbb{E}_0\left[N_{\text{ex}}^t\right] \leq B_{\text{ex}}$. Note that Eq. (7) indicates that the KL divergence (which can be viewed as the information obtained by the learner) is related to both the number of switches and the amount of extra feedback. With the upper bound on the KL divergence in Eq. (7), we can also bound the total variation. Specifically, we have

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K} D_{\text{TV}}\left(\mathbb{P}_0\left(\ell_{1:T}^{\text{ob}}\right) \| \mathbb{P}_k\left(\ell_{1:T}^{\text{ob}}\right)\right) \\
&\overset{\text{(a)}}{\leq} \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{\ln 2}{2}}\sqrt{D_{\text{KL}}\left(\mathbb{P}_0\left(\ell_{1:T}^{\text{ob}}\right) \| \mathbb{P}_k\left(\ell_{1:T}^{\text{ob}}\right)\right)} \\
&\overset{\text{(b)}}{\leq} \sqrt{\frac{\ln 2}{2}}\sqrt{2\log_2 T\cdot\frac{\epsilon^2}{2\sigma^2}}\cdot\frac{1}{K}\sum_{k=1}^{K}\sqrt{\mathbb{E}_0\left[\bar{S}^k\right] + 2B_{\text{ex}}} \\
&\overset{\text{(c)}}{\leq} \sqrt{\frac{\ln 2}{2}}\sqrt{2\log_2 T\cdot\frac{\epsilon^2}{2\sigma^2}}\cdot\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}_0\left[\bar{S}^k\right] + 2B_{\text{ex}}\right)} \\
&\overset{\text{(d)}}{=} \frac{\epsilon}{\sigma}\sqrt{\frac{\ln 2\cdot\log_2 T}{K}}\sqrt{\mathbb{E}_0\left[\bar{S}\right] + B_{\text{ex}}},
\end{aligned} \tag{8}
$$

where (a) is from Pinsker's inequality, (b) is from Eq. (7), (c) is from Jensen's inequality, and (d) is from $\sum_{k=1}^{K} \mathbb{E}_0\left[\bar{S}^k\right] = 2\mathbb{E}_0\left[\bar{S}\right]$.

With all the above results, we are ready to derive a lower bound on $\mathbb{E}\left[R_T^{\text{init}}\right]$ after showing two intermediate steps. Let N_k be the number of times action $k \in [K]$ is played up to round T (which is a random variable). We first assume that the deterministic learning algorithm makes at most ϵT switches on *any* loss sequence, which will be used for deriving Eq. (9) below, and we will later relax this assumption. Under this assumption, we have

$$\begin{aligned}
\mathbb{E}_0\left[\bar{S}\right] - \mathbb{E}\left[\bar{S}\right] &\overset{\text{(a)}}{=} \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}_0\left[\bar{S}\right] - \mathbb{E}_k\left[\bar{S}\right]\right) \\
&\overset{\text{(b)}}{=} \frac{1}{K}\sum_{k=1}^{K}\sum_{z=1}^{T}\left(\mathbb{P}_0(\bar{S} \geq z) - \mathbb{P}_k(\bar{S} \geq z)\right) \\
&\overset{\text{(c)}}{=} \frac{1}{K}\sum_{k=1}^{K}\sum_{z=1}^{\epsilon T}\left(\mathbb{P}_0(\bar{S} \geq z) - \mathbb{P}_k(\bar{S} \geq z)\right) \\
&\overset{\text{(d)}}{\leq} \frac{\epsilon T}{K}\sum_{k=1}^{K} D_{\text{TV}}\left(\mathbb{P}_0\left(\ell_{1:T}^{\text{ob}}\right) \| \mathbb{P}_k\left(\ell_{1:T}^{\text{ob}}\right)\right), \qquad (9)
\end{aligned}$$

where (a) is from the definition that $\mathbb{E}[\cdot] := \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_k[\cdot]$, (b) is from rewriting the expectations, (c) is from the assumption of no more than ϵT switches, and (d) is from the definition of the total variation. Also, we have

$$\begin{aligned}
\sum_{k=1}^{K}\mathbb{E}_k[N_k] - T &\overset{\text{(a)}}{=} \sum_{k=1}^{K}\left(\mathbb{E}_k[N_k] - \mathbb{E}_0[N_k]\right) \\
&\overset{\text{(b)}}{=} \sum_{k=1}^{K}\sum_{z=1}^{T}\left(\mathbb{P}_k(N_k \geq z) - \mathbb{P}_0(N_k \geq z)\right) \\
&\overset{\text{(c)}}{\leq} T\sum_{k=1}^{K} D_{\text{TV}}\left(\mathbb{P}_0\left(\ell_{1:T}^{\text{ob}}\right) \| \mathbb{P}_k\left(\ell_{1:T}^{\text{ob}}\right)\right), \qquad (10)
\end{aligned}$$

where (a) is from $\sum_{k=1}^{K}\mathbb{E}_0[N_k] = T$, (b) is from rewriting the expectations, and (c) is from the definition of the total variation.

We now lower bound the expected value of R_T^{init} as follows:

$$\begin{aligned}
\mathbb{E}\left[R_T^{\text{init}}\right] &\overset{\text{(a)}}{=} \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\epsilon(T - N_k) + \bar{S} | k^* = k\right] \\
&= \epsilon T - \frac{\epsilon}{K}\sum_{k=1}^{K}\mathbb{E}_k\left[N_k\right] + \mathbb{E}\left[\bar{S}\right] \\
&\overset{\text{(b)}}{\geq} \epsilon T - \frac{\epsilon T}{K} - \frac{2\epsilon T}{K}\sum_{k=1}^{K} D_{\text{TV}}\left(\mathbb{P}_0\left(\ell_{1:T}^{\text{ob}}\right) \| \mathbb{P}_k\left(\ell_{1:T}^{\text{ob}}\right)\right) + \mathbb{E}_0[\bar{S}] \\
&\overset{\text{(c)}}{\geq} \frac{\epsilon T}{2} - \frac{2\sqrt{\ln 2}\cdot\epsilon^2 T\sqrt{\log_2 T}}{\sigma\sqrt{K}}\sqrt{\mathbb{E}_0\left[\bar{S}\right] + B_{\text{ex}}} + \mathbb{E}_0\left[\bar{S}\right] \\
&\overset{\text{(d)}}{\geq} \frac{\epsilon T}{2} - \frac{2\sqrt{\ln 2}\cdot\epsilon^2 T\sqrt{\log_2 T}}{\sigma\sqrt{K}}\left(\sqrt{\mathbb{E}_0\left[\bar{S}\right]} + \sqrt{B_{\text{ex}}}\right) + \mathbb{E}_0\left[\bar{S}\right] \\
&\overset{\text{(e)}}{\geq} \frac{\epsilon T}{2} - \frac{\ln 2\cdot\epsilon^4 T^2\log_2 T}{K\sigma^2} - \frac{2\sqrt{\ln 2}\cdot\epsilon^2 T\sqrt{\log_2 T}}{\sigma\sqrt{K}}\sqrt{B_{\text{ex}}}, \qquad (11)
\end{aligned}$$

where (a) is from the regret definition (i.e., Eq. (1)), (b) is from Eqs. (9) and (10), (c) is from Eq. (8), (d) is from an elementary inequality: $\sqrt{x} + \sqrt{y} \geq \sqrt{x+y}, \forall x, y \geq 0$, and (e) is obtained by minimizing the quadratic function of $\sqrt{\mathbb{E}_0[\bar{S}]}$.

Now, we turn to lower bound $\mathbb{E}[R_T]$. By Lemma 3, if it holds that $\epsilon \leq 1/6$ and $\sigma = 1/(9\log_2 T)$, then we have $\mathbb{E}\left[R_T^{\text{init}}\right] - \mathbb{E}[R_T] \leq \epsilon T/6$. We first assume $\epsilon \leq 1/6$ and later show that the selected ϵ satisfies this condition. Then, we have

$$\begin{aligned}\mathbb{E}[R_T] &\geq \mathbb{E}\left[R_T^{\text{init}}\right] - \epsilon T/6 \\ &\geq \frac{\epsilon T}{3} - \frac{81\ln 2\cdot\epsilon^4 T^2(\log_2 T)^3}{K} - \frac{18\sqrt{\ln 2}\cdot\epsilon^2 T(\log_2 T)^{3/2}}{\sqrt{K}}\sqrt{B_{\text{ex}}},\end{aligned} \tag{12}$$

where the last step is from Eq. (11) and choosing $\sigma = 1/(9\log_2 T)$.

We now consider two cases for B_{ex}: $\sqrt{B_{\text{ex}}} \leq c_1 K^{1/6}T^{1/3}$ and $\sqrt{B_{\text{ex}}} > c_1 K^{1/6}T^{1/3}$, for some $c_1 > 0$.

In the first case of $\sqrt{B_{\text{ex}}} \leq c_1 K^{1/6}T^{1/3}$, we have

$$\begin{aligned}\mathbb{E}[R_T] &\overset{\text{(a)}}{\geq} \frac{\epsilon T}{3} - \frac{81\ln 2\cdot\epsilon^4 T^2(\log_2 T)^3}{K} - \frac{18\sqrt{\ln 2}\cdot c_1\epsilon^2 T(\log_2 T)^{3/2}}{\sqrt{K}}\cdot K^{1/6}T^{1/3} \\ &\overset{\text{(b)}}{=} \frac{c_2 K^{1/3}T^{2/3}}{3(\log_2 T)^{3/2}} - \frac{81\ln 2\cdot c_2^4 K^{1/3}T^{2/3}}{(\log_2 T)^3} - \frac{18\sqrt{\ln 2}\cdot c_1 c_2^2 K^{1/3}T^{2/3}}{(\log_2 T)^{3/2}} \\ &\overset{\text{(c)}}{\geq} \left(\frac{c_2}{3} - 81\ln 2\cdot c_2^4 - 18\sqrt{\ln 2}\cdot c_1 c_2^2\right)\frac{K^{1/3}T^{2/3}}{18(\log_2 T)^3} \\ &= \widetilde{\Omega}(K^{1/3}T^{2/3}),\end{aligned} \tag{13}$$

where (a) is from $\sqrt{B_{\text{ex}}} \leq c_1 K^{1/6}T^{1/3}$, (b) is obtained by choosing $\epsilon = c_2\frac{K^{1/3}}{T^{1/3}(\log_2 T)^{3/2}}$ (where $c_2 > 0$ satisfies $\frac{1}{3} - 81\ln 2\cdot c_2^3 - 18\sqrt{\ln 2}\cdot c_1 c_2 > 0$), and (c) is simply due to $(\log_2 T)^3 \geq (\log_2 T)^{3/2}$ for a sufficiently large T. Since $K \leq T$, we have $\epsilon \leq 1/6$ when T is sufficiently large.

In the second case of $\sqrt{B_{\text{ex}}} > c_1 K^{1/6}T^{1/3}$, we have

$$\begin{aligned}\mathbb{E}[R_T] &\overset{\text{(a)}}{\geq} \frac{\epsilon T}{3} - \frac{81\ln 2\cdot\epsilon^4 T^2(\log_2 T)^3}{K}\cdot\frac{\sqrt{B_{\text{ex}}}}{c_1 K^{1/6}T^{1/3}} - \frac{18\sqrt{\ln 2}\cdot\epsilon^2 T(\log_2 T)^{3/2}}{\sqrt{K}}\sqrt{B_{\text{ex}}} \\ &= \frac{\epsilon T}{3} - \frac{81\ln 2\cdot\epsilon^4 T^{5/3}(\log_2 T)^3}{c_1 K^{7/6}}\sqrt{B_{\text{ex}}} - \frac{18\sqrt{\ln 2}\cdot\epsilon^2 T(\log_2 T)^{3/2}}{\sqrt{K}}\sqrt{B_{\text{ex}}} \\ &\overset{\text{(b)}}{=} \frac{c_3 T\sqrt{K}}{3(\log_2 T)^{3/2}\cdot\sqrt{B_{\text{ex}}}} - \frac{81\ln 2\cdot c_3^4 K^{5/6}T^{5/3}}{c_1(\log_2 T)^3\cdot(B_{\text{ex}})^{3/2}} - \frac{18\sqrt{\ln 2}\cdot c_3^2\sqrt{K}T}{(\log_2 T)^{3/2}\cdot\sqrt{B_{\text{ex}}}} \\ &\overset{\text{(c)}}{\geq} \frac{c_3 T\sqrt{K}}{3(\log_2 T)^{3/2}\cdot\sqrt{B_{\text{ex}}}} - \frac{81\ln 2\cdot c_3^4\sqrt{K}T}{c_1^3(\log_2 T)^3\cdot\sqrt{B_{\text{ex}}}} - \frac{18\sqrt{\ln 2}\cdot c_3^2\sqrt{K}T}{(\log_2 T)^{3/2}\cdot\sqrt{B_{\text{ex}}}} \\ &\geq \left(\frac{c_3}{3} - \frac{81\ln 2\cdot c_3^4}{c_1^3} - 18\sqrt{\ln 2}\cdot c_3^2\right)\frac{\sqrt{K}T}{(\log_2 T)^3\cdot\sqrt{B_{\text{ex}}}} \\ &= \widetilde{\Omega}\left(T\sqrt{K/B_{\text{ex}}}\right),\end{aligned} \tag{14}$$

where (a) is due to $\frac{\sqrt{B_{\text{ex}}}}{c_1 K^{1/6}T^{1/3}} > 1$, (b) is obtained by choosing $\epsilon = \frac{c_3\sqrt{K}}{(\log_2 T)^{3/2}\sqrt{B_{\text{ex}}}}$ (where $c_3 > 0$ satisfies $1/3 - 81\ln 2\cdot c_3^3/c_1^3 - 18\sqrt{\ln 2}\cdot c_3 > 0$), and (c) again is due to $\frac{\sqrt{B_{\text{ex}}}}{c_1 K^{1/6}T^{1/3}} > 1$ (applied to the second term). Since $K \leq T$, we have $\epsilon = \frac{c_3\sqrt{K}}{(\log_2 T)^{3/2}\sqrt{B_{\text{ex}}}} \leq \frac{c_3 K^{1/3}}{c_1 T^{1/3}(\log_2 T)^{3/2}} \leq 1/6$ when T is sufficiently large.

Now, we want to relax the assumption that the deterministic learning algorithm makes no more than ϵT switches. Similar to the proof of Theorem 2 in Dekel et al. (2013), we consider the following: If the learning algorithm makes more than ϵT switches, then we halt the algorithm at the point when there are exactly ϵT switches and repeat its action after the last switch throughout the rest of the game. We use R_T^{halt} to denote the regret of this halted algorithm over the same loss sequence as in R_T.

We consider two cases for the number of switches made by the original learning algorithm: $\bar{S} \leq \epsilon T$ and $\bar{S} > \epsilon T$.

In the first case of $\bar{S} \leq \epsilon T$, halting does not happen, and trivially, we have $R_T^{\text{halt}} = R_T$.

In the second case of $\bar{S} > \epsilon T$, the original learning algorithm makes more than ϵT switches. We use T' to denote the round index at which the ϵT-th switch happens. Clearly, we have $\epsilon T \leq T'$. As a result, the halted algorithm keeps playing action $X_{T'}$ from round $T'+1$ to the end of the game. We now rewrite R_T^{halt} and R_T as follows:

$$R_T = \sum_{t=1}^{T'} (\ell_t[X_t] - \ell_t[k^*]) + \epsilon T + \sum_{t=T'+1}^{T} (\ell_t[X_t] - \ell_t[k^*]) + (\bar{S} - \epsilon T),$$
$$R_T^{\text{halt}} = \sum_{t=1}^{T'} (\ell_t[X_t] - \ell_t[k^*]) + \epsilon T + \sum_{t=T'+1}^{T} (\ell_t[X_{T'}] - \ell_t[k^*]).$$

By taking the difference between R_T^{halt} and R_T and then taking the expectation with respect to the randomness from loss generation, we have

$$\mathbb{E}\left[R_T^{\text{halt}} - R_T\right] = \underbrace{\mathbb{E}\left[\sum_{t=T'+1}^{T} (\ell_t[X_{T'}] - \ell_t[X_t])\right]}_{\leq \epsilon T} + \underbrace{\mathbb{E}\left[\epsilon T - \bar{S}\right]}_{<0} \leq \epsilon T.$$

To see this, we first observe that at each round, the loss gap between actions is either ϵ or 0 *in expectation* (because the Gaussian noise we add has zero mean). Therefore, the first term $\mathbb{E}\left[\sum_{t=T'+1}^{T} (\ell_t[X_{T'}] - \ell_t[X_t])\right]$ can be bounded by ϵT (i.e., the gap at each round multiplied by the time horizon). Since the original learning algorithm makes more than ϵT switches, we have $\mathbb{E}[R_T] \geq \epsilon T$, which implies $\mathbb{E}\left[R_T^{\text{halt}}\right] \leq 2\mathbb{E}[R_T]$.

Combining the above two cases yields $\mathbb{E}\left[R_T^{\text{halt}}\right] \leq 2\mathbb{E}[R_T]$. Since we already obtain a lower bound for $\mathbb{E}\left[R_T^{\text{halt}}\right]$ (i.e., Eqs. (13) and (14)), the same lower bound also holds for $\mathbb{E}[R_T]$ (within a constant factor of 2).

Finally, we complete the proof by applying Yao's principle. □

We also give two remarks about technical details in the proof of Proposition 1.

Remark 11. To conclude that $\mathbb{E}\left[R_T^{\text{halt}}\right] \leq 2\mathbb{E}[R_T]$, Dekel et al. (2013) shows that $R_T^{\text{halt}} \leq R_T + \epsilon T \leq 2R_T$, which is a stronger result compared to what is needed in an expected sense. This stronger result relies on the fact that the loss gap between actions is either ϵ or 0 at each round. However, this may not be true anymore after introducing an additional action-dependent noise as in Shi et al. (2022). Despite this difference, one can still show $\mathbb{E}\left[R_T^{\text{halt}}\right] \leq \mathbb{E}[R_T] + \epsilon T \leq 2\mathbb{E}[R_T]$, which is used to prove Proposition 1.

Remark 12. Some readers may ask: If the deterministic algorithm switches more than ϵT times, should not the switching costs already imply the desired lower bound on the regret? Why is it necessary to show a reduction from switch-limited algorithms to arbitrary algorithms? To see this, we note that the lower bound of $\widetilde{\Omega}(T^{2/3})$ is obtained after selecting ϵ to be of order $\widetilde{\Theta}(T^{-1/3})$, while such selection is based on the previous analysis, which is further based on the assumption that the algorithm makes *no more than* ϵT switches. Therefore, the reduction from switch-limited algorithms to arbitrary algorithms is necessary.

C. Proof of Proposition 2

Before proving Proposition 2, we first present two lemmas on the properties of SD. These are straightforward extensions of Lemmas 1 and 2 in Geulen et al. (2010) to their batched versions. A key difference is that we consider batches instead of rounds. While the proofs follow the same line of analysis, we provide the proofs below for completeness.

Lemma 5. *For the instantiations π_{full} and π_{flex} of algorithm 1, over any loss sequence, we have*

$$\mathbb{P}(A_b = k) = \mathbb{E}[w_b[k]],\ \forall k \in [K], \forall b \in [N].$$

Proof of Lemma 5. We first note that in these two instantiations of Algorithm 1, the feedback depends on the randomly chosen u_b only, and is thus independent of what actions are taken by the learner throughout the game. As a result, the whole feedback sequence $\widehat{\ell}_{1:N}$ can be fixed even before the learner's actions are determined.

In the following, we will show by induction that conditioning on any random feedback sequence $\widehat{\ell}_{1:N}$, it holds (almost surely) that $\mathbb{P}\left(A_b = k|\widehat{\ell}_{1:N}\right) = w_b[k], \forall k \in [K], \forall b \in [N]$.

The base case of $b = 1$ is trivial due to the algorithm design (specifically, the uniform action weight initialization). In the following, we move forward to the induction step, i.e., we show that for every $b \geq 2$, if it holds that $\mathbb{P}\left(A_{b-1} = k|\widehat{\ell}_{1:N}\right) = w_{b-1}[k],\ \forall k \in [K]$, then it also holds that $\mathbb{P}\left(A_b = k|\widehat{\ell}_{1:N}\right) = w_b[k],\ \forall k \in [K]$, as follows:

$$
\begin{aligned}
\mathbb{P}\left(A_b = k|\widehat{\ell}_{1:N}\right) &\overset{\text{(a)}}{=} W_b[k]/W_{b-1}[k] \cdot \mathbb{P}\left(A_{b-1} = k|\widehat{\ell}_{1:N}\right) \\
&\quad + \sum_{i=1}^{K} (1 - W_b[i]/W_{b-1}[i]) \cdot w_b[k] \cdot \mathbb{P}\left(A_{b-1} = i|\widehat{\ell}_{1:N}\right) \\
&\overset{\text{(b)}}{=} W_b[k]/W_{b-1}[k] \cdot w_{b-1}[k] + w_b[k] \cdot \sum_{i=1}^{K} (1 - W_b[i]/W_{b-1}[i]) \cdot w_{b-1}[i] \\
&= \frac{W_b[k]}{W_{b-1}[k]} \cdot \frac{W_{b-1}[k]}{\sum_{i=1}^{K} W_{b-1}[i]} + w_b[k] \cdot \sum_{i=1}^{K} \left(\frac{W_{b-1}[i] - W_b[i]}{W_{b-1}[i]}\right) \cdot \frac{W_{b-1}[i]}{\sum_{j=1}^{K} W_{b-1}[j]} \\
&= \frac{W_b[k]}{\sum_{i=1}^{K} W_{b-1}[i]} + w_b[k] \cdot \frac{\sum_{i=1}^{K} W_{b-1}[i] - \sum_{i=1}^{K} W_b[i]}{\sum_{i=1}^{K} W_{b-1}[i]} \\
&= w_b[k] \cdot \frac{\sum_{i=1}^{K} W_b[i]}{\sum_{i=1}^{K} W_{b-1}[i]} + w_b[k] \cdot \left(1 - \frac{\sum_{i=1}^{K} W_b[i]}{\sum_{i=1}^{K} W_{b-1}[i]}\right) \\
&= w_b[k],
\end{aligned}
$$

where (a) is due to Line 9 in Algorithm 1 (specifically, there are two disjoint cases for action k to be played in batch b: (i) action k was played in batch $b-1$, and it does not change according to probability $\exp(-\eta \cdot \widehat{\ell}_b[k]) = W_b[k]/W_{b-1}[k]$; (ii) action k is selected based on fresh randomness (i.e., the previous action i does not stay unchanged with probability $(1 - W_b[i]/W_{b-1}[i])$), regardless of which action is played in batch $b-1$), and (b) is from the inductive hypothesis that $\mathbb{P}\left(A_{b-1} = i|\widehat{\ell}_{1:N}\right) = w_b[i], \forall i \in [K]$.

Finally, taking the expectation on both sides of the above completes the proof. □

Lemma 5 implies that SD does not change the marginal distribution of action selection. Hence, one still enjoys the same regret guarantee as that of standard OMD (i.e., without SD).

Lemma 6. *For the instantiations π_{full} and π_{flex} of algorithm 1, over any loss sequence, the expected number of switches satisfies the following:*

$$
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{X_t \neq X_{t-1}\}\right] \leq \sum_{b=2}^{N} \eta \cdot \mathbb{E}\left[\widehat{\ell}_{b-1}[A_{b-1}]\right].
$$

Proof of Lemma 6. Based on the definition of switching costs, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}_{\{X_t\neq X_{t-1}\}}\right] &\overset{\text{(a)}}{=} \mathbb{E}\left[\sum_{b=2}^{N}\mathbb{I}_{\{A_b\neq A_{b-1}\}}\right]\\ &= \sum_{b=2}^{N}\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{\{A_b\neq A_{b-1}\}}|\widehat{\ell}_{1:b-1}, A_{b-1}\right]\right]\\ &\overset{\text{(b)}}{\leq} \sum_{b=2}^{N}\mathbb{E}\left[1-\exp(-\eta\cdot\widehat{\ell}_{b-1}[A_{b-1}])\right]\\ &\overset{\text{(c)}}{\leq} \sum_{b=2}^{N}\eta\cdot\mathbb{E}\left[\widehat{\ell}_{b-1}[A_{b-1}]\right],\end{aligned}$$

where (a) is because switching happens only between two consecutive batches, (b) is due to the action selection rule of Algorithm 1 (Line 9), and (c) is from elementary inequality: $1-\exp(-x)\leq x, \forall x>0$. □

Proof of Proposition 2. Our goal here is to establish an upper bound on the expected regret for algorithm $R_T^{\pi_{\text{full}}}$, which consists of two parts (cf. Eqs. (2) and (1)): (i) standard regret in terms of loss and (ii) switching cost. We will establish bounds on both respectively, hence obtaining the final result in the proposition.

To start with, let us establish an upper bound on the standard regret in terms of loss. To this end, we will build upon the classic analysis of OMD (cf. (Orabona, 2019, Section 6.6)). That is, for any (random) sequence $\widehat{\ell}_{1:N}\in[0,\infty)^K$, learning rate $\eta>0$, and vector u such that its Y_u-th coordinate is 1 and all the others are 0, it holds almost surely that

$$\sum_{b=1}^{N}\left\langle w_b-u,\widehat{\ell}_b\right\rangle\leq\frac{\ln K}{\eta}+\frac{\eta}{2}\sum_{b=1}^{N}\sum_{k=1}^{K}(\widehat{\ell}_b[k])^2\cdot w_b[k],$$

where $w_b[k]=\frac{W_b[k]}{\sum_{i=1}^{K}W_b[i]}$ and $W_b[k]=W_{b-1}[k]\cdot\exp(-\eta\cdot\widehat{\ell}_{b-1}[k]), \forall k\in[K]$, i.e., line 8 in Algorithm 1. Taking the expectation on both sides yields that

$$\begin{aligned}\mathbb{E}\left[\sum_{b=1}^{N}\left\langle w_b-u,\widehat{\ell}_b\right\rangle\right] &\leq \frac{\ln K}{\eta}+\frac{\eta}{2}\sum_{b=1}^{N}\sum_{k=1}^{K}\mathbb{E}\left[(\widehat{\ell}_b[k])^2\cdot w_b[k]\right]\\ &= \frac{\ln K}{\eta}+\frac{\eta}{2}\sum_{b=1}^{N}\sum_{k=1}^{K}\mathbb{E}\left[w_b[k]\cdot\mathbb{E}\left[(\widehat{\ell}_b[k])^2|\widehat{\ell}_{1:b-1}\right]\right]\\ &\overset{\text{(a)}}{=} \frac{\ln K}{\eta}+\frac{\eta}{2}\sum_{b=1}^{N}\sum_{k=1}^{K}\mathbb{E}\left[w_b[k]\cdot(\ell_{u_b}[k])^2\right]\\ &\overset{\text{(b)}}{\leq} \frac{\ln K}{\eta}+\frac{\eta B}{2K},\end{aligned}\tag{15}$$

where (a) follows from the algorithm design of π_{full}, i.e., $\widehat{\ell}_b=\ell_{u_b}$ and u_b is a randomly selected time slot within batch b, and (b) comes from the boundedness of losses, the fact that $\sum_{k=1}^{K}w_b[k]=1, \forall b\in[N]$, and noting that $N=B/K$.

We can further rewrite the left-hand-side (LHS) of the above inequality as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{b=1}^{N}\left\langle w_b - u, \widehat{\ell}_b\right\rangle\right] &= \sum_{b=1}^{N}\mathbb{E}\left[\mathbb{E}\left[\left\langle w_b - u, \widehat{\ell}_b\right\rangle | \widehat{\ell}_{1:b-1}\right]\right] \\
&= \sum_{b=1}^{N}\mathbb{E}\left[\left\langle w_b - u, \mathbb{E}\left[\widehat{\ell}_b | \widehat{\ell}_{1:b-1}\right]\right\rangle\right] \\
&\overset{\text{(a)}}{=} \sum_{b=1}^{N}\mathbb{E}\left[\left\langle w_b - u, \frac{\sum_{t=(b-1)\tau+1}^{b\tau}\ell_t}{\tau}\right\rangle\right] \\
&= \frac{1}{\tau}\sum_{b=1}^{N}\sum_{t=(b-1)\tau+1}^{b\tau}\mathbb{E}\left[\ell_t[A_b] - \ell_t[Y_u]\right] \\
&= \frac{1}{\tau}\sum_{t=1}^{T}\mathbb{E}\left[\ell_t[X_t] - \ell_t[Y_u]\right], \qquad (16)
\end{aligned}$$

where (a) holds since for any $k \in [K]$, we have $\mathbb{E}\left[\widehat{\ell}_b[k]|\widehat{\ell}_{1:b-1}\right] = \sum_{t=(b-1)\tau+1}^{b\tau}\ell_t[k]/\tau$. This is true since under π_{full}, the choice of u_b for constructing $\widehat{\ell}_b$ is a round index chosen uniformly at random from the current batch b.

Now, replacing the LHS of Eq. (15) with Eq. (16), yields

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\ell_t[X_t] - \ell_t[Y_u]\right] &\le \tau\cdot\left(\frac{\ln K}{\eta} + \frac{\eta B}{2K}\right) \\
&= \frac{KT\ln K}{\eta B} + \frac{\eta T}{2}, \qquad (17)
\end{aligned}$$

where the last step is due to $\tau = \frac{KT}{B}$.

After obtaining the above upper bound on the standard regret (i.e., without switching costs), we now turn to bound the switching costs under π_{full}. To this end, we directly leverage Lemma 6 along with the loss estimate $\widehat{\ell}_b = \ell_{u_b}$ to obtain that

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}_{\{X_t \neq X_{t-1}\}}\right] \le \sum_{b=2}^{N}\eta\cdot\mathbb{E}\left[\widehat{\ell}_{b-1}[A_{b-1}]\right] = \sum_{b=2}^{B/K}\eta\cdot\mathbb{E}\left[\widehat{\ell}_{b-1}[A_{b-1}]\right] \le \frac{\eta B}{K}. \qquad (18)$$

Finally, combining Eqs. (17) and (18), we can bound the total regret as follows:

$$\begin{aligned}
R_T^{\pi_{\text{full}}} &\le \frac{KT\ln K}{\eta B} + \frac{\eta T}{2} + \frac{\eta B}{K} \\
&\overset{\text{(a)}}{\le} \frac{KT\ln K}{\eta B} + \frac{3\eta T}{2} \\
&\overset{\text{(b)}}{=} T\sqrt{\frac{6K\ln K}{B}},
\end{aligned}$$

where (a) is from $B \le KT$, and (b) is obtained by choosing $\eta = \sqrt{\frac{2K\ln K}{3B}}$. Hence, we have completed the proof of Proposition 2. □

D. Proof of Proposition 3

Proof of Proposition 3. The organization of this proof is the same as that of Proposition 2, and the only difference lies in the loss estimate in this instantiation.

We first consider the case when $B < T$, i.e., M can be any integer from $[K]$. Recall that B is the total observation budget and M is the number of observations made in each batch. Similarly, we have, for any (random) sequence $\widehat{\ell}_1, \ldots, \widehat{\ell}_N \in [0,\infty)^K$,

learning rate $\eta > 0$, and vector u such that its Y_u-th coordinate is 1 and all the others are 0, it holds almost surely that

$$\sum_{b=1}^{N} \left\langle w_b - u, \widehat{\ell}_b \right\rangle \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{b=1}^{N} \sum_{k=1}^{K} (\widehat{\ell}_b[k])^2 \cdot w_b[k].$$

Taking the expectation on both sides yields that

$$\begin{aligned}
\mathbb{E}\left[\sum_{b=1}^{N} \left\langle w_b - u, \widehat{\ell}_b \right\rangle\right] &\leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{b=1}^{N} \sum_{k=1}^{K} \mathbb{E}\left[(\widehat{\ell}_b[k])^2 \cdot w_b[k]\right] \\
&= \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{b=1}^{N} \sum_{k=1}^{K} \mathbb{E}\left[\mathbb{E}\left[(\widehat{\ell}_b[k])^2 \cdot w_b[k] \middle| \widehat{\ell}_{1:b-1}\right]\right] \\
&\overset{\text{(a)}}{=} \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{b=1}^{N} \sum_{k=1}^{K} \mathbb{E}\left[\left(\frac{\ell_{u_b}[k]}{M/K}\right)^2 \cdot w_b[k] \cdot \frac{M}{K} + 0 \cdot (1 - \frac{M}{K})\right] \\
&\overset{\text{(b)}}{\leq} \frac{\ln K}{\eta} + \frac{\eta K B}{2M^2},
\end{aligned} \tag{19}$$

where (a) follows from the algorithm design of π_{flex}, i.e., the loss estimate $\widehat{\ell}_b[k] = \mathbb{I}\{k \in \mathcal{O}_{u_b}\} \cdot \frac{\ell_{u_b}[k]}{M/K}$ and u_b is a randomly selected time slot within batch b, and (b) comes from the boundedness of losses, the fact that $\sum_{k=1}^{K} w_b[k] = 1, \forall b \in [N]$, and noting that $N = B/K$.

We can further rewrite the LHS of the above inequality as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{b=1}^{N} \left\langle w_b - u, \widehat{\ell}_b \right\rangle\right] &= \sum_{b=1}^{N} \mathbb{E}\left[\mathbb{E}\left[\left\langle w_b - u, \widehat{\ell}_b \right\rangle \middle| \widehat{\ell}_{1:b-1}\right]\right] \\
&= \sum_{b=1}^{N} \mathbb{E}\left[\left\langle w_b - u, \mathbb{E}\left[\widehat{\ell}_b \middle| \widehat{\ell}_{1:b-1}\right]\right\rangle\right] \\
&\overset{\text{(a)}}{=} \sum_{b=1}^{N} \mathbb{E}\left[\left\langle w_b - u, \frac{\sum_{t=(b-1)\tau+1}^{b\tau} \ell_t}{\tau}\right\rangle\right] \\
&= \frac{1}{\tau} \sum_{b=1}^{N} \sum_{t=(b-1)\tau+1}^{b\tau} \mathbb{E}\left[\ell_t[A_b] - \ell_t[Y_u]\right] \\
&= \frac{1}{\tau} \sum_{t=1}^{T} \mathbb{E}\left[\ell_t[X_t] - \ell_t[Y_u]\right],
\end{aligned} \tag{20}$$

where (a) holds since for any $k \in [K]$, we have $\mathbb{E}\left[\widehat{\ell}_b[k] \middle| \widehat{\ell}_{1:b-1}\right] = (1 - \frac{M}{K}) \cdot 0 + \frac{M}{K} \cdot \sum_{t=(b-1)\tau+1}^{b\tau} \frac{1}{\tau} \cdot \frac{\ell_t[k]}{M/K} = \sum_{t=(b-1)\tau+1}^{b\tau} \ell_t[k]/\tau$. This is true since under π_{flex}, the choices of both u_b and $\mathcal{O}_t$ for constructing $\widehat{\ell}_b$ are uniformly random and independent with each other.

Now, replacing the LHS of Eqs. (19) with (20), yields

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\ell_t[X_t] - \ell_t[Y_u]\right] &\leq \tau \cdot \left(\frac{\ln K}{\eta} + \frac{\eta K B}{2M^2}\right) \\
&= \frac{MT \ln K}{\eta B} + \frac{\eta T K}{2M},
\end{aligned} \tag{21}$$

where in the equality, we replace τ with $\frac{MT}{B}$.

After obtaining the above upper bound on the standard regret (i.e., without switching cost), we now turn to bound the switching costs under π_{flex}. To this end, we directly leverage Lemma 6 along with the loss estimate $\widehat{\ell}_b[k] = \mathbb{I}\{k \in$

$\mathcal{O}_{u_b}\} \cdot \frac{\ell_{u_b}[k]}{M/K}$ to obtain that

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}_{\{X_t\neq X_{t-1}\}}\right] \leq \sum_{b=2}^{N}\eta\cdot\mathbb{E}\left[\widehat{\ell}_{b-1}[A_{b-1}]\right] = \sum_{b=2}^{B/M}\eta\cdot\mathbb{E}\left[\widehat{\ell}_{b-1}[A_{b-1}]\right] \leq \eta\frac{B}{M}\cdot\frac{K}{M} = \frac{\eta BK}{M^2}. \tag{22}$$

Finally, combining Eqs. (21) and (22), we can bound the total regret as follows:

$$\begin{aligned} R_T^{\pi_{\text{flex}}} &\leq \frac{MT\ln K}{\eta B} + \frac{\eta TK}{2M} + \frac{\eta BK}{M^2} \\ &\overset{\text{(a)}}{\leq} \frac{MT\ln K}{\eta B} + \frac{3\eta KT}{2M} \\ &\overset{\text{(b)}}{=} T\sqrt{\frac{6K\ln K}{B}}, \end{aligned}$$

where (a) is from $B < MT$ (recall that now we have $B < T$ and $M \geq 1$), and (b) is obtained by choosing $\eta = M\sqrt{\frac{2\ln K}{3BK}}$. Hence, we have completed the proof of Proposition 3.

The proof for the case of $B \geq T$ is exactly the same, except for the (implicit) fact that we need batch size τ to be well-defined, i.e., $\tau \geq 1$. That is why in this case M is less flexible: now M needs to be sufficiently large to fully exploit the total budget. □

E. Proof of Proposition 4

Proof of Proposition 4. The organization of this proof is the same as that of Proposition 2, and the main difference lies in the commonly-used importance-weighted estimator for bandit feedback. In addition, we note that it is now sufficient to directly bounding the switching costs by the number of batches.

We still start with the same fundamental conclusion in OMD analysis. Specifically, for any (random) sequence $\widehat{\ell}_1, \ldots, \widehat{\ell}_N \in [0,\infty)^K$, learning rate $\eta > 0$, and vector u such that its Y_u-th coordinate is 1 and all the others are 0, it holds almost surely that

$$\sum_{b=1}^{N}\left\langle w_b - u, \widehat{\ell}_b\right\rangle \leq \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{b=1}^{N}\sum_{k=1}^{K}(\widehat{\ell}_b[k])^2 w_b[k].$$

Taking the expectation on both sides, we have

$$\begin{aligned} \mathbb{E}\left[\sum_{b=1}^{N}\left\langle w_b - u, \widehat{\ell}_b\right\rangle\right] &\leq \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{b=1}^{N}\mathbb{E}\left[\sum_{k=1}^{K}(\widehat{\ell}_b[k])^2 w_b[k]\right] \\ &= \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{b=1}^{N}\mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{K}(\widehat{\ell}_b[k])^2 w_b[k]\middle|\widehat{\ell}_{1:b-1}\right]\right] \\ &\overset{\text{(a)}}{=} \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{b=1}^{N}\mathbb{E}\left[\sum_{k=1}^{K} w_b[k]\cdot\frac{1}{\tau}\cdot\sum_{t=(b-1)\tau+1}^{b\tau}\frac{(\ell_t[k])^2}{(w_b[k])^2}\cdot w_b[k]\right] \\ &\overset{\text{(b)}}{\leq} \frac{\ln K}{\eta} + \frac{\eta}{2}NK \\ &\overset{\text{(c)}}{=} \sqrt{2NK\ln K}, \end{aligned} \tag{23}$$

where (a) follows from the algorithm design of π_{b}, i.e., the loss estimate $\widehat{\ell}_b[k] = \mathbb{I}\{k \in \mathcal{O}_{u_b}\}\cdot\frac{\ell_{u_b}[k]}{w_b[k]}$ for all $k \in [K]$, (b) follows from the assumption that all losses are bounded by one and the fact that $\sum_{k=1}^{K} w_b[k] = 1, \forall b \in [N]$, and (c) is obtained by choosing $\eta = \sqrt{\frac{2\ln K}{NK}}$.

We can further rewrite the LHS of the above inequality as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{b=1}^{N}\left\langle w_b-u,\widehat{\ell}_b\right\rangle\right] &= \sum_{b=1}^{N}\mathbb{E}\left[\mathbb{E}\left[\left\langle w_b-u,\widehat{\ell}_b\right\rangle\middle|\widehat{\ell}_{1:b-1}\right]\right] \\
&= \sum_{b=1}^{N}\mathbb{E}\left[\left\langle w_b-u,\mathbb{E}\left[\widehat{\ell}_b\middle|\widehat{\ell}_{1:b-1}\right]\right\rangle\right] \\
&\overset{\text{(a)}}{=} \sum_{b=1}^{N}\mathbb{E}\left[\left\langle w_b-u,\frac{\sum_{t=(b-1)\tau+1}^{b\tau}\ell_t}{\tau}\right\rangle\right] \\
&= \frac{1}{\tau}\sum_{b=1}^{N}\sum_{t=(b-1)\tau+1}^{b\tau}\mathbb{E}\left[\ell_t[A_b]-\ell_t(Y_u)\right] \\
&= \frac{1}{\tau}\sum_{t=1}^{T}\mathbb{E}\left[\ell_t[X_t]-\ell_t[Y_u]\right], \qquad (24)
\end{aligned}$$

where (a) holds since for any $k\in[K]$, we have $\mathbb{E}\left[\widehat{\ell}_b[k]\middle|\widehat{\ell}_{1:b-1}\right] = (1-w_b[k])\cdot 0+\sum_{t=(b-1)\tau+1}^{b\tau}\frac{1}{\tau}\cdot w_b[k]\cdot\frac{\ell_t[k]}{w_b[k]} = \sum_{t=(b-1)\tau+1}^{b\tau}\ell_t[k]/\tau$.

Now, replacing the LHS of Eq. (23) with Eq. (24), yields

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\ell_t[X_t]-\ell_t[Y_u]\right] &\le \tau\sqrt{2NK\ln K} \\
&= \frac{T}{B}\cdot\sqrt{2BK\ln K} \\
&= O\left(T\sqrt{\frac{K\ln K}{B}}\right). \qquad (25)
\end{aligned}$$

When $B=O(K^{1/3}T^{2/3})$, we have $\sum_{t=1}^{T}\mathbb{I}_{\{X_t\neq X_{t-1}\}}\le B\le O\left(T\sqrt{\frac{K\ln K}{B}}\right)$. Combining it with Eq. (25), we can conclude that both the standard regret and switching costs are of order $O\left(T\sqrt{\frac{K\ln K}{B}}\right)$, which gives us the desired result. □

F. An Auxiliary Technical Result

In this section, we show that in the standard setting without switching costs, only using bandit feedback (e.g., algorithm π_{b}) can also achieve optimal regret (i.e., matching the lower bound of $\Omega(T\sqrt{K/B})$, up to poly-logarithmic factors) in the full range of $B\in[K,T]$. We state this result in Proposition 5.

Proposition 5. *In the standard setting without switching costs, for any $B\in[K,T]$, the worst-case regret under algorithm π_{b} is upper bounded by $R_T^{\pi_{\mathrm{b}}}=O(T\sqrt{K\ln K/B})$.*

Proof of Proposition 5. The proof follows the same line of analysis as that in the proof of Proposition 4, except that we only require $B<T$ (instead of $B=O(K^{1/3}T^{2/3})$) and do not consider switching costs. Therefore, Eq. (25) implies the following upper bound on the regret:

$$\sum_{t=1}^{T}\mathbb{E}\left[\ell_t[X_t]-\ell_t[Y_u]\right]=O\left(T\sqrt{K\ln K/B}\right).$$

Note that Y_u can be any fixed action, including the best fixed action. This completes the proof. □