Online learning with feedback graphs and switching costs

Anshuka Rangi
University of California, San Diego

Massimo Franceschetti
University of California, San Diego

Abstract

We study online learning when partial feedback information is provided following every action of the learning process, and the learner incurs switching costs for changing his actions. In this setting, the feedback information system can be represented by a graph, and previous works studied the expected regret of the learner in the case of a clique (Expert setup), or disconnected single loops (Multi-Armed Bandits (MAB)). This work provides a lower bound on the expected regret in the Partial Information (PI) setting, namely for general feedback graphs – excluding the clique. Additionally, it shows that all algorithms that are optimal without switching costs are necessarily sub-optimal in the presence of switching costs, which motivates the need to design new algorithms. We propose two new algorithms: Threshold Based EXP3 and EXP3.SC. For the two special cases of symmetric PI setting and MAB, the expected regret of both of these algorithms is order optimal in the duration of the learning process. Additionally, Threshold Based EXP3 is order optimal in the switching cost, whereas EXP3.SC is not. Finally, empirical evaluations show that Threshold Based EXP3 outperforms the previously proposed order-optimal algorithms EXP3 SET in the presence of switching costs, and Batch EXP3 in the MAB setting with switching costs.

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

1 Introduction

Online learning has a wide variety of applications like classification, estimation, and ranking, and it has been investigated in different areas, including learning theory, control theory, operations research, and statistics. The problem can be viewed as a one player game against an adversary. The game runs for T rounds and at each round the player chooses an action from a given set of K actions. Every action $k \in [K]$ performed at round $t \in [T]$ carries a loss, that is a real number in the interval $[0, 1]$. The losses for all pairs (k, t) are assigned by the adversary before the game starts. The player also incurs a fixed and known Switching Cost (SC) every time he changes his action, that is an arbitrary real number $c > 0$. The expected regret is the expectation of the sum of losses associated to the actions performed by the player plus the SCs minus the losses incurred by the best fixed action in hindsight. The goal of the player is to minimize the expected regret over the duration of the game.

Based on the feedback information received after each action, online learning can be divided into three categories: Multi-Armed Bandit (MAB), Partial Information (PI), and Expert setting. In a MAB setting, at any given round the player only incurs the loss corresponding to the selected action, which implies the player only observes the loss of the selected action. In a PI setting, the player incurs the loss of the selected action $k \in [K]$, as well as observes the losses that he would have incurred in that round by taking actions in a subset of $[K]\backslash\{k\}$. This feedback system can be viewed as a time-varying directed graph G_t with K nodes, where a directed edge $k \to j$ in G_t indicates that performing an action k at round t also reveals the loss that the player would have incurred if action j was taken at round t. In an Expert setting, taking an action reveals the losses that the player would have incurred by taking any of the other actions in that round. In this extremal case, the feedback system G_t corresponds to a time-invariant, undirected clique.

Online learning with PI has been used to design a variety of systems [Gentile and Orabona, 2014, Katariya

Scenarios	Threshold based EXP3	EXP3.SC	Lower Bound
For all t, $G_t = G$	$\tilde{O}(c^{1/3}(\text{mas}(G))^{1/3}T^{2/3})$	$\tilde{O}(c^{4/3}(\text{mas}(G))^{2/3}T^{2/3})$	$\tilde{\Omega}(c^{1/3}\alpha(G)^{1/3}T^{2/3})$
Symmetric PI	$\tilde{O}(c^{1/3}\alpha(G)^{1/3}T^{2/3})$	$\tilde{O}(c^{4/3}\alpha(G)^{2/3}T^{2/3})$	$\tilde{\Omega}(c^{1/3}\alpha(G)^{1/3}T^{2/3})$
MAB	$\tilde{O}(c^{1/3}K^{1/3}T^{2/3})$	$\tilde{O}(c^{4/3}K^{2/3}T^{2/3})$	$\tilde{\Omega}(c^{1/3}K^{1/3}T^{2/3})$
$G_{1:T}$	$\tilde{O}(c\sum_{t=1}^{t^*}\text{mas}(G_{(t)})/\text{mas}(G_{(T)}))$	$\tilde{O}(\sum_{t=1}^{n^*}\text{mas}(G_{(t)})/\text{mas}(G_{(T)}))$	$\tilde{\Omega}(c^{1/3}\beta(G_{1:T})^{1/3}T^{2/3})$
Equi-informational	$\tilde{O}(c^{1/3}\alpha(G_1)^{1/3}T^{2/3})$	$\tilde{O}(c^{4/3}\alpha(G_1)^{2/3}T^{2/3})$	$\tilde{\Omega}(c^{1/3}\beta(G_{1:T})^{1/3}T^{2/3})$

Table 1: Comparison of Threshold based EXP3 and EXP3.SC.

et al., 2016, Zong et al., 2016, Rangi et al., 2018c]. In these applications, feedback captures the idea of side information provided to the player during the learning process. For example, the performance of an employee can provide information about the performance of other employees with similar skills, or the rating of a web page can provide information on ratings of web pages with similar content. In most of these applications, switching between the actions is not free. For example, a company incurs a cost associated to the learning phase while shifting an employee among different tasks, or switching the content of a web page frequently can exasperate users and force them to avoid visiting it. Similarly, re-configuring the production line in a factory is a costly process, and changing the stock allocation in an investment portfolio is subject to certain fees. Despite the many applications where both SC and PI are an integral part of the learning process, the study of online learning with SC has been limited only to the MAB and Expert settings. In the MAB setting, it has been shown that the expected regret of any player is at least $\tilde{\Omega}(c^{1/3}K^{1/3}T^{2/3})$ [Dekel et al., 2014], and that Batch EXP3 is an order optimal algorithm [Arora et al., 2012]. In the Expert setting, it has been shown that the expected regret is at least $\tilde{\Omega}(\sqrt{\log(K)T})$ [Cesa-Bianchi and Lugosi, 2006], and order optimal algorithms have been proposed in [Geulen et al., 2010, Gyorgy and Neu, 2014]. The PI setup has been investigated only in the absence of SC, and for any fixed feedback system $G_t = G$ with independence number $\alpha(G) > 1$, it has been shown that the expected regret is at least $\tilde{\Omega}(\sqrt{\alpha(G)T})$ [Mannor and Shamir, 2011].

1.1 Contributions

We provide a lower bound on the expected regret for any sequence of feedback graphs $G_1, \dots G_T$ in the PI setting with SC. We show that for any sequence of feedback graphs $G_{1:T} = \{G_1, \dots G_T\}$ with independence sequence number $\beta(G_{1:T}) > 1$, the expected regret of any player is at least $\tilde{\Omega}(c^{1/3}\beta(G_{1:T})^{1/3}T^{2/3})$. We then show that for $G_{1:T}$ with $\alpha(G_t) > 1$ for all $t \leq T$, the expected regret of any player is at least $\tilde{\Omega}(c^{1/3}\sum_{G_j \in \mathcal{G}}\alpha(G_j)^{1/3}N(G_j)^{2/3})$, where $\mathcal{G}$ is the set of unique feedback graphs in the sequence $G_{1:T}$, and $N(G_j) = \sum_{t=1}^{T}\mathbf{1}(G_t = G_j)$ is the number of rounds for which the feedback graph G_j is seen in T rounds. These results introduce a new figure of merit $\beta(G_{1:T})$ in the PI setting, which can also be used to generalize the lower bound given in the PI setting without SC [Mannor and Shamir, 2011]. A consequence of these results is that the presence of SC changes the asymptotic regret by at least a factor $T^{1/6}$. Additionally, these results also recover the lower bound on the expected regret in the MAB setting [Dekel et al., 2014].

We also show that in the PI setting for any algorithm that is order optimal without SC, there exists an assignment of losses from the adversary that forces the algorithm to make at least $\tilde{\Omega}(T)$ switches, thus increasing its asymptotic regret by at least a factor $T^{1/2}$. This shows that any algorithm that is order optimal in the PI setting without SC, is necessarily sub-optimal in the presence of SC, and motivates the development of new algorithms in the PI setting and in the presence of SC.

We propose two new algorithms for the PI setting with SC: Threshold-Based EXP3 and EXP3.SC. Threshold-Based EXP3 requires the knowledge of T in advance, whereas EXP3.SC does not. The performance of these algorithms is given for different scenarios in Table 1. The algorithms are order optimal in T and $\beta(G_{1:T})$ for two special cases of feedback information system: symmetric PI setting i.e. the feedback graph $G_t = G$ is fixed and un-directed, and MAB. In these two cases, $\beta(G_{1:T})$ equals $\alpha(G)$ and K respectively. The state-of-art algorithm EXP3 SET in PI setting without SC is known to be order optimal only for these cases as well [Alon et al., 2017]. Threshold Based EXP3 is order optimal in the SC c as well, while EXP3.SC has an additional factor of c in its expected regret. In the time-varying case, for sequence $G_{1:T}$, the expected regret is dependent on the worst t^* and n^* instances of the ratio of $\text{mas}(G_t)$ and $\text{mas}(G_{(T)})$, where $\{\text{mas}(G_{(1)}), \text{mas}(G_{(2)}), \dots, \text{mas}(G_{(T)})\}$ are the sizes of the maximal acyclic subgraphs of $G_{1:T}$ arranged in non-increasing order, $t^* = \lceil T^{2/3}c^{-2/3}\text{mas}^{1/3}(G_{(T)})\rceil$

and $n^* = 0.5\mathrm{mas}^{1/3}(G_{(T)})T^{2/3}c^{1/3}$. Finally, Table 1 also provides the performance in the equi-informational setting, namely when G_t is undirected and all the maximal acyclic subgraphs in $G_{1:T}$ have the same size. The proofs of all these results are available online [Rangi and Franceschetti, 2018b].

Numerical comparison shows that Threshold Based EXP3 outperforms EXP3 SET in the presence of SCs. Threshold Based EXP3 also outperforms Batch EXP3, which is another order optimal algorithm for the MAB setting with SC [Arora et al., 2012].

1.2 Related Work

In the absence of SC, the lower bound on the expected regret is known for all three categories of online learning problems. In the MAB setting, the expected regret is at least $\tilde{\Omega}(\sqrt{KT})$ [Auer et al., 2002, Cesa-Bianchi and Lugosi, 2006, Rangi et al., 2018d]. In the PI setting with fixed feedback graph G, the expected regret is at least $\tilde{\Omega}(\sqrt{\alpha(G)T})$ [Mannor and Shamir, 2011]. In the Expert setting, the expected regret is at least $\tilde{\Omega}(\sqrt{\log(K)T})$ [Cesa-Bianchi and Lugosi, 2006]. All three cases present an asymptotic regret factor $T^{1/2}$. In contrast, in the presence of SC the expected regrets for MAB and Expert settings present different factors, namely $T^{2/3}$ and $T^{1/2}$ respectively. The expected regret is at least $\tilde{\Omega}(c^{1/3}K^{1/3}T^{2/3})$ in the MAB setting and $\tilde{\Omega}(\sqrt{\log(K)T})$ in the Expert setting [Dekel et al., 2014]. This work provides the lower bound on the expected regret $\tilde{\Omega}(c^{1/3}\beta(G_{1:T})^{1/3}T^{2/3})$ for the PI setting in the presence of SC. For the case without SC, this work establishes that the lower bound in PI setting is $\tilde{\Omega}(\sqrt{\beta(G_{1:T})T})$.

The PI setting was first considered in [Alon et al., 2013, Mannor and Shamir, 2011], and many of its variations have been studied without SC [Alon et al., 2015, Alon et al., 2013, Caron et al., 2012, Rangi et al., 2018b, Langford and Zhang, 2008, Kocák et al., 2016, Rangi et al., 2018a, Wu et al., 2015, Rangi and Franceschetti, 2018a]. In the adversarial setting we described, all of these algorithms are order optimal in the MAB and symmetric PI settings, but they also require the player to have knowledge of the graph G_t before performing an action. The algorithm EXP3 SET does not require such knowledge [Alon et al., 2017]. We show that all of these algorithms are sub-optimal in the PI setting with SC, and propose new algorithms that are order optimal in the MAB and symmetric PI settings.

In the expert setting with SC, there are two order optimal algorithms with expected regret $\tilde{O}(\sqrt{\log(K)T})$ [Geulen et al., 2010, Gyorgy and Neu, 2014]. In the MAB setting with SC, Batch EXP3 is an order optimal algorithm with expected regret $\tilde{O}(c^{1/3}K^{1/3}T^{2/3})$ [Arora et al., 2012]. This algorithm has also been used to solve a variant of the MAB setting [Feldman et al., 2016]. In the MAB setting, our algorithm has the same order of expected regret as Batch EXP3 but it numerically outperforms Batch EXP3.

There is a large literature on a continuous variation of the MAB setting, where the number of actions K depends on the number of rounds T. In this setting, the case without the SC was investigated in [Auer et al., 2007, Bubeck et al., 2011, Kleinberg, 2005, Yu and Mannor, 2011]. Recently, the case including SC has also been studied in [Koren et al., 2017a, Koren et al., 2017b]. In [Koren et al., 2017a], the algorithm Slowly Moving Bandits (SMB) has been proposed and in [Koren et al., 2017b], it has been extended to different settings. These algorithms incur an expected regret linear in T when applied in our discrete setting.

2 Problem Formulation

Before the game starts, the adversary fixes a loss sequence $\ell_1, \ldots, \ell_T \in [0,1]^K$, assigning a loss in $[0,1]$ to K actions for T rounds. At round t, the player performs an action $i_t \in [K]$, and incurs the loss $\ell_t(i_t)$ assigned by the adversary. If $i_t \neq i_{t-1}$, then the player also incurs a cost $c > 0$ in addition to the loss $\ell_t(i_t)$.

In the PI setting, the feedback system can be viewed as a time-varying directed graph G_t with K nodes, where a directed edge $k \to j$ indicates that choosing action k at round t also reveals the loss that the player would have incurred if action j were taken at round t. Let $S_t(i) = \{j : i \to j \text{ is a directed edge in } G_t\}$. Following the action i_t, the player observes the losses he would have incurred in round t by performing actions in the subset $S_t(i_t) \subseteq [K]$. Since the player always observes its own loss, $i_t \in S_t(i_t)$. In a MAB setup, the feedback graph G_t has only self loops, i.e. for all $t \leq T$ and $i \in [K]$, $S_t(i) = \{i\}$. In an Expert setup, G_t is a undirected clique i.e. for all $t \leq T$ and $i \in [K]$, $S_t(i) = [K]$. The expected regret of a player's strategy δ is defined as

$$R^\delta(\ell_{1:T}, c) = \mathbf{E}\left[\sum_{t=1}^{T} \ell_t(i_t) + \sum_{t=2}^{T} c \cdot \mathbf{1}(i_{t-1} \neq i_t)\right] - \min_{k\in[K]} \sum_{t=1}^{T} \ell_t(k). \tag{1}$$

In words, the expected regret is the expectation of the sum of losses associated to the actions performed by the player plus the SCs minus the losses incurred by the best fixed action in the hindsight, and the objective of the player is to minimize the expected regret.

Algorithm 1 Adversary's strategy

```
Input:$T > 0$, $G_{1:T}$ with $\beta(G_{1:T}) > 1$;
Set $\epsilon_1 = \epsilon_2 = c^{1/3}\beta(G_{1:T})^{1/3}T^{-1/3}/9\log_2(T)$ and $\sigma = 1/9\log_2(T)$.
Choose an arm $X \in \mathcal{I}(G_{1:T})$ uniformly at random
Draw $T$ variables such that $\forall t \leq T$, $y_t \sim \mathcal{N}(0, \sigma^2)$.
For all $1 \leq t \leq T$ and $i \in [K]$, assign
    $\ell_t(i) = W_t + 0.5 - \epsilon_1 \mathbf{1}(X = i) + \epsilon_2 \mathbf{1}(i \notin \mathcal{I}(G_{1:T}))$,
    $\ell_t(i) = clip(\ell_t(i))$,
where $clip(a) = \min\{\max\{a, 0\}, 1\}$, For all $t \leq T$ $W_t = W_{\rho(t)} + y_t$,, $W_0 = 0$, $\rho(t) = t - 2^{\delta(t)}$ and $\delta(t) = \max\{i \geq 0 : 2^i \text{ divides } t\}$.
Output: loss sequence $\ell_{1:T}$.
```

3 Lower Bound in PI setting with SC

We start by defining the independence sequence number for a sequence of graphs $G_{1:T}$.

Definition 3.1 *Given $G_{1:T}$, let $P(G_t)$ be the set of all the possible independent sets of the graph G_t. The independence sequence number $\beta(G_{1:T})$ is the largest cardinality among all intersections of the independent sets $s_1 \cap s_2 \cap \ldots \cap s_T$, where $s_t \in P(G_t)$. Namely,*

$$\beta(G_{1:T}) = \max_{s_1 \in P(G_1), \ldots s_T \in P(G_T)} |s_1 \cap s_2 \cap \ldots \cap s_T|. \quad (2)$$

Definition 3.2 *The independence sequence set $\mathcal{I}(G_{1:T})$ is the set $s_1 \cap s_2 \cap \ldots s_T$ attaining the maximum in (2).*

We use the notion of $\beta(G_{1:T})$ to provide a lower bound on the expected regret in the PI setting with SC.

Theorem 1 *For any $G_{1:T}$ with $\beta(G_{1:T}) > 1$, there exists a constant $b > 0$ and an adversary's strategy (Algorithm 1) such that for all $T \geq c \cdot \max\{6, \beta(G_{1:T})\}$, and for any player's strategy $\mathcal{A}$, the expected regret of $\mathcal{A}$ is at least $b\, c^{1/3}\beta(G_{1:T})^{1/3}T^{2/3}/\log T$.*

The proof of Theorem 1 relies on Yao's minimax principle [Yao, 1977]. A randomized adversary strategy is constructed such that the expected regret of a player, whose action at any round is a deterministic function of his past observations, is at least $b\, c^{1/3}\beta(G_{1:T})^{1/3}T^{2/3}/\log T$. This adversary strategy is described in Algorithm 1, and is a generalization of the one proposed to establish similar bounds in the MAB setup [Dekel et al., 2014]. The generalization is different than the one proposed for the PI setting without SC [Mannor and Shamir, 2011]. Since $G_{1:T}$ is known to the adversary, it computes the independence sequence set $\mathcal{I}(G_{1:T})$, and the cardinality of this set is $\beta(G_{1:T})$. For all $t \leq T$ and $i, j \in \mathcal{I}(G_{1:T})$, there exists no edge in the graph G_t between the actions i and j. Thus, the selection of any action in $\mathcal{I}(G_{1:T})$ provides no information about the losses of the other actions in $\mathcal{I}(G_{1:T})$. The adversary selects the optimal action uniformly at random from $\mathcal{I}(G_{1:T})$, and assigns an expected loss of $1/2 - \epsilon_1$. The remaining actions in $\mathcal{I}(G_{1:T})$ are assigned an expected loss of $1/2$. On the other hand, since $i \in [K]\backslash\mathcal{I}(G_{1:T})$ provides information about the losses of actions in $\mathcal{I}(G_{1:T})$, action i is assigned an expected loss of $1/2 + \epsilon_2$ to compensate for this additional information. In practice, even a small bias ϵ_2 compensates for the extra information provided by an action in $[K]\backslash\mathcal{I}(G_{1:T})$.

In the PI setup without SC, for a fixed feedback graph $G_t = G$, the expected regret is at least $\tilde{\Omega}(\sqrt{\alpha(G)T})$ [Alon et al., 2017]. The lower bound is provided only for a fixed feedback system, and the lower bound for a general time-varying feedback system $G_{1:T}$ is left as an open question [Alon et al., 2017]. This also motivates the investigation of different graph theoretic measures to study the PI setting [Alon et al., 2017]. Theorem 1 provides a lower bound for a general time-varying feedback system $G_{1:T}$ for the PI setting in presence of SC. The lower bound is dependent on the independence sequence number $\beta(G_{1:T})$ of $G_{1:T}$. Thus, the ideas introduced in Theorem 1 can be extended to close this gap in the literature of PI setting without SC.

Lemma 2 *In the PI setting without SC, for any $G_{1:T}$ with $\beta(G_{1:T}) > 1$, there exists a constant $b > 0$ and an adversary's strategy such that for any player's strategy $\mathcal{A}$, the expected regret of $\mathcal{A}$ is at least $b\sqrt{\beta(G_{1:T})T}$.*

Using Theorem 1 and Lemma 2, it can be concluded that the presence of SC changes the asymptotic regret by at least a factor $T^{1/6}$. In the MAB setup, $\beta(G_{1:T}) = K$, and Theorem 1 recovers the bounds provided in [Dekel et al., 2014].

We now focus on the assumption in Theorem 1, i.e. $\beta(G_{1:T}) > 1$. This is satisfied in many networks of practical interest. For example, networks modeled as p-random graphs where p is the probability of having edge between two nodes. The expected independence number of these graphs is $2\log(Kp)/p$ [Coja-Oghlan and Efthymiou, 2015]. Since the probability of each node being in independent set is same, the expected value of $\beta(G_{1:T})$ is $K(2\log(Kp)/Kp)^T$, and Kp is the expected node degree which is usually a constant as p is inversely proportional to K. This is greater than one for large values of K, and small values of T.

Algorithm 1 depends on the independence sequence set $\mathcal{I}(G_{1:T})$ whose cardinality is non-increasing in T. In such cases, the adversary can split the sequence of feedback graphs $G_{1:T}$ into multiple sub-sequences i.e. say M sub-sequences such that $U_1 = \{G_t : t \in T_1 \subseteq [T]\} \ldots U_M = \{G_t : t \in T_M \subseteq [T]\}$, $[T] = \cup_{m\in[M]} T_m$, and for all $m_1, m_2 \in [M]$, $T_{m_1} \cap T_{m_2}$ is an empty set. For each sub-sequence U_m, compute the independence sequence set and assign losses independently of other sub-sequences according to Algorithm 1. This adversary's strategy, which we call Algorithm 1.1, gives the following bound on the expected regret.

Theorem 3 *For any split of $G_{1:T}$ into disjoint sub-sequences $U_1, \ldots U_M$ with $\beta(U_m) > 1$ and $N(U_m) \geq c \cdot \max\{6, \beta(U_m)\}$ $\forall m \in [M]$, there exists a constant $b > 0$ and an adversary's strategy (Algorithm 1.1) such that for any player's strategy $\mathcal{A}$, the expected regret of $\mathcal{A}$ is at least $b\, c^{1/3} \sum_{m\in[M]} \beta(U_m)^{1/3} N(U_m)^{2/3} / \log T$, where $N(U_m) = \sum_{t=1}^{T} \mathbf{1}(G_t \in U_m)$ is the length of sub-sequence U_m.*

With the insight provided by Theorem 3, the regret can be made large with an appropriate split of $G_{1:T}$ into sub-sequences. This can be formulated as a sub-modular optimization problem where the objective is:

$$\max_{\{U_1,\ldots,U_M\}} c^{1/3} \sum_{m\in[M]} \beta(U_m)^{1/3} N(U_m)^{2/3} / \log T \quad (3)$$

$$\text{subject to} \sum_{m\in[M]} N(U_m) = T, \qquad (4)$$
$$\forall m_1, m_2 \in [M], U_{m_1} \cap U_{m_2} = \phi.$$

This can be solved using greedy algorithms developed in the context of sub-modular maximization.

Until now, we have been focusing on designing an adversary's strategy for maximizing the regret for a given sequence of feedback graphs $G_{1:T}$. Now, we briefly discuss the case when $G_{1:T}$ can also be chosen by the adversary. If the adversary is not constrained about the choice of feedback graphs, then the feedback graph that maximizes the expected regret would be a feedback graph with only self loops, as this reveals the least amount of information. If the adversary is constrained by the choice of independence number, i.e. for all $t \leq T$, $\alpha(G_t) \leq H$, then the optimal value of (3) is achieved for a sequence of fixed feedback graphs i.e. for all $t \leq T$, $\alpha(G_t) = H$, which implies $\beta(G_{1:T}) = H$.

We now discuss the trade-off between the loss incurred and the number of switches performed by the player.

Lemma 4 *If the expected regret computed ignoring the SC of any algorithm $\mathcal{A}$ is $\tilde{O}((\beta(G_{1:T})^{1/2}T)^{\beta})$, then there exists a loss sequence $\ell_{1:T}$ such that $\mathcal{A}$ makes at least $\tilde{\Omega}[(\beta(G_{1:T})^{1/2}T)^{2(1-\beta)}]$ switches.*

Along the same lines of Lemma 4, it can also be shown that if the expected number of switches of $\mathcal{A}$ is $\tilde{O}[(\beta(G_{1:T})^{1/2}T)^{2(1-\beta)}]$, then the expected regret without SC is at least $\tilde{\Omega}((\beta(G_{1:T})^{1/2}T)^{\beta})$. This provides the lower bound on the expected regret given the SC is constrained by a fixed budget. Using Lemma 4, if the expected regret without SC of $\mathcal{A}$ is $\tilde{O}(\sqrt{\beta(G_{1:T})T})$, then there exists a loss sequence that forces $\mathcal{A}$ to make at least $\tilde{\Omega}(T)$ switches. This implies the regret of $\mathcal{A}$ with the SC is linear in T. Thus, any algorithm that is order optimal without SC, is necessarily sub-optimal in the presence of SC, which motivates the design of new algorithms in our setting.

4 Algorithms in PI setting with SC

In this section, we introduce the two algorithms Threshold Based EXP3 and EXP3.SC for an uninformed setting where G_t is only revealed after the action i_t has been performed. This is common in a variety of applications. For instance, a user's selection of some product allows to infer that the user might be interested in similar products. However, no action on the recommended products may mean that user might not be interested in the product, does not need it or did not check the products. Thus, the feedback is revealed only after the action has been performed.

In Threshold Based EXP3 (Algorithm 2), each action $i \in [K]$ is assigned a weight $w_{i,t}$ at round t. When the loss of action i is observed at round t, i.e. $i \in S_t(i_t)$, $w_{i,t}$ is computed by penalizing $w_{i,t-1}$ exponentially by the empirical loss $\ell_t(i)\mathbf{1}(i \in S_t(i_t))/q_{i,t}$. At round t, $p_t = \{p_{1,t}, \ldots, p_{K,t}\}$ is the sampling distribution where $p_{i,t} = w_{i,t}/\sum_{i\in[K]} w_{i,t}$. At round t, action i_t is selected with probability $p_{i,t}$ if the threshold event $E^t = E_1^t \cup E_2^t \cup E_3^t$ is true, where

$$E_1^t = \{t = 1\},$$

$$E_2^t = \{r > \gamma_t, \text{ where } \gamma_t = T^{1/3}c^{2/3}/\text{mas}(G_{(T)})^{1/3}\},$$

$$\begin{aligned} E_3^t = \{&\forall i \in [K]\backslash\{i_t\}, \hat{\ell}_{t-1}(i) + \ell'_{t-1}(i) > \epsilon_t/\eta + 1/q_{i_t,t-1}, \\ &\text{and there exists an } i \in [K]\backslash\{i_t\} \text{ such that} \\ &\hat{\ell}_{t-1}(i) + \ell'_{t-1}(i) - \ell'_{t-1}(i_t) \leq \epsilon_t/\eta + 1/q_{i_t,t-1}\}, \end{aligned} \quad (5)$$

and $\epsilon_t \geq \log(tc^2/\text{mas}(G_{(T)}))/3$. The event E^t contains two threshold conditions, one on the variable r and the other on the empirical losses. The threshold event E^t is critical in balancing the trade-off between the number of switches and the loss incurred by the player. E_1^t corresponds to the first selection of action, and incurs no SC. In E_2^t, the variable r tracks the number of rounds (or time instances) since the event E^t occurred last time. If the choice of a new action has not been considered for past γ_t rounds, then E_2^t forces the player to choose an action according to

Algorithm 2 Threshold based EXP3

```
Initialization: $\eta \in (0,1]$; For all $i \in [K]$, $w_{i,1} = 1$, $\hat{\ell}_0(i) = 0$ and $\ell'_0(i) = 0$; $r = 1$;
for $t = 1, \dots, T$ do
    if $E_1^t$ or $E_2^t$ or $E_3^t$ (see (5)) then
        if $t \neq 1$ then
            $\hat{\ell}_t(i) = \hat{\ell}_{t-1}(i) + \ell'_{t-1}(i)$
            $w_{i,t} = w_{i,t-1} \exp\left(-\eta \ell'_{t-1}(i)\right)$
        end if
        Update $p_{i,t} = w_{i,t} / \sum_{j \in [K]} w_{j,t}$.
        Choose $i_t = i$ with probability $p_{i,t}$.
        Set $r = 1$ and for all $i \in [K]$, set $\ell'_t(i) = 0$
    else
        For all $i \in [K]$, $p_{i,t} = p_{i,t-1}$, $\hat{\ell}_t(i) = \hat{\ell}_{t-1}(i)$
        and $w_{i,t} = w_{i,t-1}$; $i_t = i_{t-1}$;$r = r + 1$
    end if
    For all $i \in S_t(i_t)$, observe the pair $(\ell_t(i), i)$.
    For all $i \in [K]$, $\ell'_t(i) = \ell'_{t-1}(i) + \ell_t(i)\mathbf{1}(i \in S_t(i_t))/q_{i,t}$, where $q_{i,t} = \sum_{j: j \to i} p_{j,t}$
end for
```

the updated sampling distribution p_t at round t. The threshold condition in E_2^t ensures that the regret incurred due to the selection of a sub-optimal action does not grow continuously while trying to save on the SC between the actions. The event E_2^t is independent of the observed losses, and will occur at most $O(T^{2/3})$ times. Unlike event E_2^t, the event E_3^t is dependent on the losses $\hat{\ell}_t(i)$ and $\ell'_t(i)$, for all $i \in [K]$. Each loss $\hat{\ell}_t(i)$ tracks the total empirical loss of action i observed until round $\sigma(t) - 1$, i.e.

$$\hat{\ell}_t(i) = \sum_{k=1}^{\sigma(t)-1} \ell_k(i)\mathbf{1}(i \in S_k(i_k))/q_{i,k},$$

where $\sigma(t) = \max\{k \leq t : E^k \text{ is true }\}$ is the latest round $k^* \leq t$ at which E^{k^*} is true. On the other hand, each loss $\ell'_t(i)$ represents the total empirical loss of action i observed between rounds $\sigma(t)$ and t, i.e.

$$\ell'_t(i) = \sum_{k=\sigma(t)}^{t} \ell_k(i)\mathbf{1}(i \in S_k(i_k))/q_{i,k}.$$

This loss tracks the total empirical loss observed after the selection of an action at time instance $\sigma(t)$. The event E_3^t balances exploration and exploitation while taking into account the SC. In E_3^t, the first condition ensures that the player has sufficient amount of information about the losses of all other actions before exploitation is considered. Given sufficient exploration has been performed, the second condition triggers the exploitation. The selection of a new action is considered when the empirical loss $\ell'_t(i_t)$ incurred by the current action i_t, following its selection at $\sigma(t)$, becomes significant in comparison to the total empirical loss $\hat{\ell}_t(i) + \ell'_t(i)$ incurred by the other actions $i \in [K]\backslash\{i_t\}$. Since the total empirical loss of an action i increases with t, it is desirable that the threshold $\epsilon_t/\eta + 1/q_{i_t,t-1}$ increases with t as well. Since the increment in $\ell'_{t-1}(i_{t-1})$ is bounded above by $1/q_{i,t-1}$ at round t, for all $i \in [K]\backslash\{i_t\}$, E_3^t implies that

$$\hat{\ell}_{t-1}(i) + \ell'_{t-1}(i) - \ell'_{t-1}(i_{t-1}) \geq \epsilon_t/\eta. \tag{6}$$

Thus, E_3^t ensures that the player reconsiders the action selection if the loss incurred due to the current selection becomes significant in comparison to the total empirical loss of other actions. The event also ensures that the loss incurred due to the current selection is sufficiently smaller than the total empirical loss of other actions (see (6)). The event ensures that the sampling distribution p_t has changed significantly from the previous sampling distribution $p_{\sigma(t-1)}$ before selecting the action again. Thus, E_3^t balances exploration and exploitation based on the observed losses.

Batch EXP3, the order optimal algorithm in MAB with SC, is EXP3 performed in batches of $O(T^{1/3})$. A similar strategy to design an algorithm for the PI setting with SC will fail because unlike MAB setting, the feedback graph G_t can change at every round t, and this requires an update of empirical losses based on G_t at every round. In our algorithm, the computation of empirical loss is dependent on G_t via $q_{i,t}$. Additionally, Batch EXP3 does not utilize the information about the observed losses, which is captured in E_3^t. The following theorem presents the performance guarantees of our algorithm.

Theorem 5 *The following statements hold for Threshold Based EXP3:*
(i) *The expected regret without accounting for SC is*

$$\begin{aligned} &\mathbf{E}\left[\sum_{t=1}^{T} \ell_t(i_t) - \min_{k \in [K]} \sum_{t=1}^{T} \ell_t(k)\right] \\ &\leq \frac{\log(K)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{t^*} \frac{T^{2/3} c^{4/3} \mathrm{mas}(G_{(t)})}{(1 - 1/e)\mathrm{mas}^{2/3}(G_{(T)})}, \end{aligned} \tag{7}$$

where $t^* = \lceil T^{2/3} c^{-2/3} \mathrm{mas}^{1/3}(G_{(T)}) \rceil$.
(ii) *The expected number of switches is*

$$\mathbf{E}\left[\sum_{t=2}^{T} \mathbf{1}(i_{t-1} \neq i_t)\right] \leq 2T^{2/3} c^{-2/3} \mathrm{mas}^{1/3}(G_{(T)}). \tag{8}$$

(iii) *Letting* $\eta = \log(K)/T^{2/3} c^{1/3} \mathrm{mas}^{1/3}(G_{(T)})$, *the expected regret (1) is at most*

$$\begin{aligned} &3T^{2/3} c^{1/3} \mathrm{mas}^{1/3}(G_{(T)}) \\ &+ \frac{ec \cdot \log(K)}{2(e-1)\mathrm{mas}(G_{(T)})} \sum_{t=1}^{t^*} \mathrm{mas}(G_{(t)}). \end{aligned} \tag{9}$$

Algorithm 3 EXP3.SC

```
Initialization: For all $i \in [K]$, $\hat{\ell}_1(i) = 0$; $t = 1$, $\epsilon_t = 0.5c^{1/3}\text{mas}^{1/3}(G_{(T)})/t^{1/3}$, $\eta_t = \log(K)/t^{2/3}c^{1/3}\text{mas}^{1/3}(G_{(T)})$
for $t = 1, \ldots, T$ do
    For all $i \in [K]$, update:
        $p_t(i) = \frac{\exp(-\eta_t \hat{L}_{t-1}(i))}{\sum_{j\in[K]}\exp(-\eta_t \hat{L}_{t-1}(j))}$
    Choose $i_t = i_{t-1}$ with probability $1 - \epsilon_t$,
    else, $i_t = i$ with probability $\epsilon_t p_{i,t}$.
    For all $i \in S_t(i_t)$, observe the pair $(\ell_t(i), i)$.
    For all $i \in [K]$, update $\hat{L}_t(i) = \sum_{n=1}^{t}\hat{\ell}_n(i)$,
    where $\hat{\ell}_t(i) = \ell_t(i)\mathbf{1}(i \in S_t(i_t))/q_{i,t}$ and
    $q_{i,t} = \sum_{j:j\rightarrow i} p_{j,t}$.
end for
```

(*iv*) *In a symmetric PI setting i.e. for all $t \leq T$ G_t is un-directed and fixed, the expected regret (1) is at most*

$$4T^{2/3}c^{1/3}\alpha^{1/3}(G_1)\log(K). \tag{10}$$

In the PI setting, $\text{mas}(G_t)$ captures the information provided by the feedback graph G_t. As $\text{mas}(G_t)$ increases, the information provided by G_t about the losses of actions decreases. The regret of the algorithm depends on the $O(T^{2/3})$ instances of $\text{mas}(G_{(t)})$ (see Theorem 5 (*i*)). This is because the algorithm makes a selection of a new action $O(T^{2/3})$ times in expectation (see Theorem 5 (*ii*)), and G_t is not available in advance to influence the selection of the action. Also, the ratio $\text{mas}(G_{(t)})/\text{mas}(G_{(T)})$ is bounded above by K and has no affect on order of T. The bounds of the algorithm on the expected regret are tight in two special cases. In the symmetric PI setting, the expected regret of Threshold Based EXP3 is $\tilde{O}(T^{2/3}c^{1/3}\alpha^{1/3}(G_1))$ (see Theorem 5 (*iii*)), hence, the algorithm is order optimal. In the MAB setting, the expected regret of Threshold Based EXP3 is $\tilde{O}(T^{2/3}c^{1/3}K^{1/3})$, hence, the algorithm is order optimal. The state-of-art algorithm for the case without SCs is known to be order optimal only for these cases as well, and the key challenges for closing this gap are highlighted in the literature [Alon et al., 2017].

EXP3.SC (Algorithm 3) is another algorithm in PI setting with SC. The key differences between Threshold based EXP3 and EXP3.SC are highlighted here. Unlike Threshold based EXP3, EXP3.SC does not require the knowledge of the number of rounds T. Threshold based EXP3 favors the selection of action at regular intervals based on the event E^t. On contrary, EXP3.SC chooses a new action with probability ϵ_t which is decreasing in t. Thus, the algorithm favors exploration in the initial rounds, and favors exploitation as t increases. In Threshold based EXP3, the scaling exponent η is a constant dependent on T. On contrary, in EXP3.SC, the scaling exponent η_t is time-varying, and is decreasing in t. The following theorem provides the performance guarantees of EXP3.SC.

Theorem 6 *The expected regret (1) of EXP3.SC is at most*

$$1.5c^{4/3}\text{mas}^{1/3}(G_{(T)})T^{2/3} + \frac{2\log(K)}{\text{mas}^{2/3}(G_{(T)})}\sum_{j=1}^{n^*}\text{mas}(G_{(j)}),$$

where $n^* = 0.5\text{mas}^{1/3}(G_{(T)})T^{2/3}c^{1/3}$.

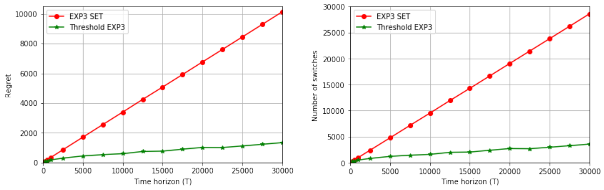
In symmetric PI and MAB settings, the expected regret of EXP3.SC is $\tilde{O}(c^{4/3}\alpha^{2/3}(G_1)T^{2/3})$ and $\tilde{O}(c^{4/3}K^{2/3}T^{2/3})$ respectively. Hence, the algorithm is order optimal in T and $\beta(G_{1:T})$, and has an additional factor of c in the performance guarantees. In EXP3.SC, the dependency on T is removed at the expense of an additional factor of c in its performance.

In an alternative setting where the number of switches are constraint to be $O(T^{2(1-\beta)})$, it can be shown using Lemma 4 that the expected regret without SC is at least $\tilde{\Omega}((\beta(G_{1:T})^{1/2}T)^{\beta})$. The two algorithms in this setting are also simple variations of our two algorithms: Threshold based EXP3 and EXP3.SC. Threshold based EXP3 can be adapted by using threshold $\gamma_t = O(T^{2\beta-1})$, $\epsilon_t = O(\log(t)/2\beta - 1)$ and $\eta = O(T^{-\beta})$. EXP3.SC can be adapted by using $\epsilon_t = O(t^{-(2\beta-1)})$ and $\eta_t = O(t^{-\beta})$. These adapted algorithms would be order optimal in MAB and symmetric PI settings as well.

5 Performance Evaluation

In this section, we numerically compare the performance of Threshold based EXP3 with EXP3 SET and Batch EXP3 in PI and MAB setups with SC respectively. We do not compare the performance of our algorithm with the ones proposed in the Expert setting with SC because in MAB and PI setups, the player needs to balance the exploration-exploitation trade-off, while in the Expert setting the player is only concerned about the exploitation. Hence, there is a fundamental discontinuity in the design of algorithms as we move from the Expert to the PI setting. This gap is also evident from the discontinuity in the lower bounds in these settings, for the Expert setting the expected regret is at least $\tilde{\Omega}(\sqrt{\log(K)T})$, while for the PI setting the expected regret is at least $\tilde{\Omega}(\beta(G_{1:T})^{1/3}T^{2/3})$, for $\beta(G_{1:T}) > 1$ which excludes the clique feedback graph.

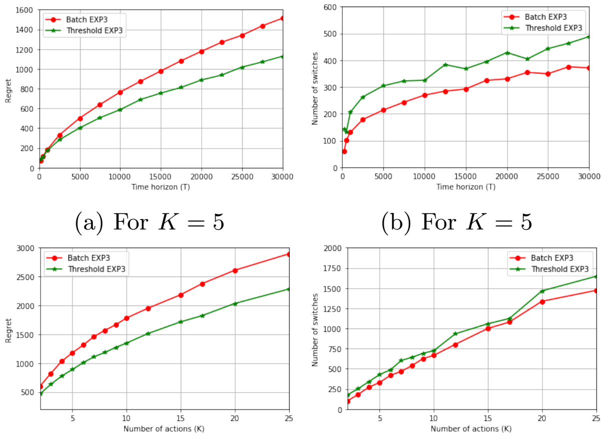
We evaluate these algorithms by simulations because in real data sets, the adversary's strategy is not necessarily unfavorable for the players. Hence, the trends in



(a) For $\alpha(G) = 5$

(b) For $\alpha(G) = 5$

Figure 1: Performance evaluation of EXP3 SET and Threshold based EXP3 for K=25



(a) For $K = 5$

(b) For $K = 5$

(c) For $T = 20000$

(d) For $T = 20000$

Figure 2: Performance evaluation of Batch EXP3 and Threshold based EXP3 in MAB setting

the performance can vary widely across different data sets. For this reason, in the literature only algorithms in stochastic setups rather than adversarial setups are typically evaluated on real data sets [Katariya et al., 2016, Zong et al., 2016]. In our simulations, the adversary uses the Algorithm 1, and $c = 0.35$.

Figure 1 shows that the Threshold based EXP3 outperforms EXP3 SET in the presence of SC. Additionally, the expected regret and the number of switches of EXP3 SET grow linearly with T. These observations are in line with our theoretical results presented in Lemma 4. The results presented here are for $G_t = G$, $\alpha(G) = 5$ and $K = 25$. Similar trends were observed for different value of $\alpha(G)$ and K.

Figure 2 shows that Threshold based EXP3 outperforms Batch EXP3 in MAB setup with SC. The gap in the performance of these algorithm increases with T (Figure 2(a)). Additionally, the number of switches performed by threshold based EXP3 is larger than the number of switches performed by Batch EXP3 (Figure 2(b) and (d)). The former algorithm utilizes the information about the observed losses via E_3^t to balance the trade off between the regret and the number of switches. On contrary, Batch EXP3 does not utilize any information from the observed losses, and switches the action only after playing an action $\tilde{O}(T^{1/3})$ times. Note that MAB setup reveals the least information about the losses, and performance gap due to utilization of this information is significant (Figure 2). This gap in performance grows as $\beta(G_{1:T})$ decreases.

In summary, Threshold Based EXP3 outperforms both EXP3 SET and Batch EXP3 in PI and MAB settings with SC respectively. Threshold Based EXP3 fills a gap in the literature by providing a solution for the PI setting with SC, and improves upon the existing literature in the MAB setup.

6 Conclusion

This work focuses on online learning in the PI setting with SC in the presence of an adversary. The lower bound on the expected regret is presented in the PI setup in terms of independence sequence number. There is a need to design new algorithms in this setting because any algorithm that is order optimal without SC is necessarily sub-optimal in the presence of SC. Two algorithms, Threshold Based EXP3 and EXP3.SC, are proposed and their performance is evaluated in terms of expected regret. These algorithms are order optimal in T in two cases: symmetric PI and MAB setup. Numerical comparisons show that the Threshold Based EXP3 outperforms EXP3 SET and Batch EXP3 in PI setting with SC.

As future work, algorithms can be designed in a partially informed setting and a fully informed setting. In the partially informed setting, the feedback graph G_t at round t is revealed following the action at round $t-1$. Thus, the feedback graphs are revealed one at a time in advance at the beginning of each round. In the fully informed setting, the entire sequence of feedback graphs $G_{1:T}$ is revealed before the game starts. Since the adversary is aware of $G_{1:T}$, these settings are important to study from the player's end as well. Note that without SC, the algorithms in both the partially informed and fully informed settings can exploit the feedback graphs at every round in a greedy manner, and perform an action accordingly. Hence, the algorithm in partially informed setting is also optimal in a fully informed setting in the absence of SC. On the contrary, in the presence of SC, a greedy exploitation of the feedback structure is not possible at every round. Hence, in fully informed setting with SC, the player chooses an action based on $G_{1:T}$ such that the selected action balances the trade off between the regret and the SC. Thus, the partially informed and fully informed settings of PI are of particular interest in the presence of SC, and is an interesting area for further study.

References

[Alon et al., 2015] Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online learning with feedback graphs: Beyond bandits. In *JMLR WORKSHOP AND CONFERENCE PROCEEDINGS*, volume 40. Microtome Publishing.

[Alon et al., 2017] Alon, N., Cesa-Bianchi, N., Gentile, C., Mannor, S., Mansour, Y., and Shamir, O. (2017). Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826.

[Alon et al., 2013] Alon, N., Cesa-Bianchi, N., Gentile, C., and Mansour, Y. (2013). From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems*, pages 1610–1618.

[Arora et al., 2012] Arora, R., Dekel, O., and Tewari, A. (2012). Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400*.

[Auer et al., 2002] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77.

[Auer et al., 2007] Auer, P., Ortner, R., and Szepesvári, C. (2007). Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory*, pages 454–468. Springer.

[Bubeck et al., 2011] Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. (2011). X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695.

[Caron et al., 2012] Caron, S., Kveton, B., Lelarge, M., and Bhagat, S. (2012). Leveraging side observations in stochastic bandits. *arXiv preprint arXiv:1210.4839*.

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

[Coja-Oghlan and Efthymiou, 2015] Coja-Oghlan, A. and Efthymiou, C. (2015). On independent sets in random graphs. *Random Structures & Algorithms*, 47(3):436–486.

[Dekel et al., 2014] Dekel, O., Ding, J., Koren, T., and Peres, Y. (2014). Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 459–467. ACM.

[Feldman et al., 2016] Feldman, M., Koren, T., Livni, R., Mansour, Y., and Zohar, A. (2016). Online pricing with strategic and patient buyers. In *Advances in Neural Information Processing Systems*, pages 3864–3872.

[Gentile and Orabona, 2014] Gentile, C. and Orabona, F. (2014). On multilabel classification and ranking with bandit feedback. *Journal of Machine Learning Research*, 15(1):2451–2487.

[Geulen et al., 2010] Geulen, S., Vöcking, B., and Winkler, M. (2010). Regret minimization for online buffering problems using the weighted majority algorithm. In *COLT*, pages 132–143.

[Gyorgy and Neu, 2014] Gyorgy, A. and Neu, G. (2014). Near-optimal rates for limited-delay universal lossy source coding. *IEEE Transactions on Information Theory*, 60(5):2823–2834.

[Katariya et al., 2016] Katariya, S., Kveton, B., Szepesvari, C., and Wen, Z. (2016). Dcm bandits: Learning to rank with multiple clicks. In *International Conference on Machine Learning*, pages 1215–1224.

[Kleinberg, 2005] Kleinberg, R. D. (2005). Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704.

[Kocák et al., 2016] Kocák, T., Neu, G., and Valko, M. (2016). Online learning with erdős-rényi side-observation graphs. In *Uncertainty in Artificial Intelligence*.

[Koren et al., 2017a] Koren, T., Livni, R., and Mansour, Y. (2017a). Bandits with movement costs and adaptive pricing. *arXiv preprint arXiv:1702.07444*.

[Koren et al., 2017b] Koren, T., Livni, R., and Mansour, Y. (2017b). Multi-armed bandits with metric movement costs. In *Advances in Neural Information Processing Systems*, pages 4122–4131.

[Langford and Zhang, 2008] Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824.

[Mannor and Shamir, 2011] Mannor, S. and Shamir, O. (2011). From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692.

[Rangi and Franceschetti, 2018a] Rangi, A. and Franceschetti, M. (2018a). Multi-armed bandit algorithms for crowdsourcing systems with online estimation of workers' ability. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1345–1352. International Foundation for Autonomous Agents and Multiagent Systems.

[Rangi and Franceschetti, 2018b] Rangi, A. and Franceschetti, M. (2018b). Online learning with feedback graphs and switching costs. *arXiv preprint arXiv:1810.09666*.

[Rangi et al., 2018a] Rangi, A., Franceschetti, M., and Marano, S. (2018a). Consensus-based chernoff test in sensor networks. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6773–6778. IEEE.

[Rangi et al., 2018b] Rangi, A., Franceschetti, M., and Marano, S. (2018b). Decentralized chernoff test in sensor networks. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 501–505. IEEE.

[Rangi et al., 2018c] Rangi, A., Franceschetti, M., and Marano, S. (2018c). Distributed chernoff test: Optimal decision systems over networks. *arXiv preprint arXiv:1809.04587*.

[Rangi et al., 2018d] Rangi, A., Franceschetti, M., and Tran-Thanh, L. (2018d). Unifying the stochastic and the adversarial bandits with knapsack. *arXiv preprint arXiv:1811.12253*.

[Wu et al., 2015] Wu, Y., György, A., and Szepesvári, C. (2015). Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, pages 1360–1368.

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

[Yu and Mannor, 2011] Yu, J. Y. and Mannor, S. (2011). Unimodal bandits. In *ICML*, pages 41–48. Citeseer.

[Zong et al., 2016] Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., and Kveton, B. (2016). Cascading bandits for large-scale recommendation problems. *arXiv preprint arXiv:1603.05359*.