From Bandits to Experts: On the Value of Side-Observations

Shie Mannor
Department of Electrical Engineering
Technion, Israel
shie@ee.technion.ac.il

Ohad Shamir
Microsoft Research New England
USA
ohadsh@microsoft.com

Abstract

We consider an adversarial online learning setting where a decision maker can choose an action in every stage of the game. In addition to observing the reward of the chosen action, the decision maker gets side observations on the reward he would have obtained had he chosen some of the other actions. The observation structure is encoded as a graph, where node i is linked to node j if sampling i provides information on the reward of j. This setting naturally interpolates between the well-known "experts" setting, where the decision maker can view all rewards, and the multi-armed bandits setting, where the decision maker can only view the reward of the chosen action. We develop practical algorithms with provable regret guarantees, which depend on non-trivial graph-theoretic properties of the information feedback structure. We also provide partially-matching lower bounds.

1 Introduction

One of the most basic learning settings studied in the online learning framework is learning from experts. In its simplest form, we assume that each round t, the learning algorithm must choose one of k possible actions, which can be interpreted as following the advice of one of k "experts"[1]. At the end of the round, the performance of all actions, measured here in terms of some reward, is revealed. This process is iterated for T rounds, and our goal is to minimize the *regret*, namely the difference between the total reward of the single best action in hindsight, and our own accumulated reward. We follow the standard online learning framework, in which nothing whatsoever can be assumed on the process generating the rewards, and they might even be chosen by an adversary who has full knowledge of our learning algorithm.

A crucial assumption in this setting is that we get to see the rewards of all actions at the end of each round. However, in many real-world scenarios, this assumption is unrealistic. A canonical example is web advertising, where at any timepoint one may choose only a single ad (or small number of ads) to display, and observe whether it was clicked, but not whether other ads would have been clicked or not if presented to the user. This partial information constraint has led to a flourishing literature on multi-armed bandits problems, which model the setting where we can only observe the reward of the action we chose. While this setting has been long studied under stochastic assumptions, the landmark paper [4] showed that this setting can also be dealt with under adversarial conditions, making the setting comparable to the experts setting discussed above. The price in terms of the provable regret is usually an extra $\sqrt{k}$ multiplicative factor in the bound. The intuition for this factor has long been that in the bandit setting, we only get "$1/k$ of the information" obtained in the expert setting (as we observe just a single reward rather than k). While the bandits setting received much theoretical interest, it has also been criticized for not capturing additional side-information we often

[1]The more general setup, which is beyond the scope of this paper, considers k experts providing advice for choosing among n actions, where in general $n \neq k$ [4].

have on the rewards of the different actions. This has led to studying richer settings, which make various assumptions on the relationship between the rewards; see below for more details.

In this paper, we formalize and initiate a study on a range of settings that interpolates between the bandits setting and the experts setting. Intuitively, we assume that after choosing some action i, and obtaining the action's reward, we observe not just action i's reward (as in the bandit setting), and not the rewards of all actions (as in the experts setting), but rather some (possibly noisy) information on a *subset* of the other actions. This subset may depend on action i in an arbitrary way, and may change from round to round. This information feedback structure can be modeled as a sequence of directed graphs $G_1, \ldots, G_T$ (one per round t), so that an edge from action i to action j implies that by choosing action i, "sufficiently good" information is revealed on the reward of action j as well. The case of G_t being the complete graph corresponds to the experts setting. The case of G_t being the empty graph corresponds to the bandit setting. The broad scenario of arbitrary graphs in between the two is the focus of our study.

As a motivating example, consider the problem of web advertising mentioned earlier. In the standard multi-armed bandits setting, we assume that we have no information whatsoever on whether undisplayed ads would have been clicked on. However, in many cases, we do have some side-information. For instance, if two ads i, j are for similar vacation packages in Hawaii, and ad i was displayed and clicked on by some user, it is likely that the other ad j would have been clicked on as well. In contrast, if ad i is for running shoes, and ad j is for wheelchair accessories, then a user who clicked on one ad is unlikely to clique on the other. This sort of side-information can be better captured in our setting.

As another motivating example, consider a sensor network where each sensor collects data from a certain geographic location. Each sensor covers an area that may overlap the area covered by other sensors. At every stage a centralized controller activates one of the sensors and receives input from it. The value of this input is modeled as the integral of some "information" in the covered area. Since the area covered by each of the sensors overlaps the area covered by other sensors, the reward obtained when choosing sensor i provides an indication of the reward that would have been obtained when sampling sensor j. A related example comes from ultra wideband communication networks, where every agent can select which channel to use for transmission. When using a channel, the agent senses if the transmission was successful, and also receives some indication of the noise level in other channels that are in adjacent frequency bands [2].

Our results portray an interesting picture, with the attainable regret depending on non-trivial properties of these graphs. We provide two practical algorithms with regret guarantees: the ExpBan algorithm that is based on a combination of existing methods, and the more fundamentally novel ELP algorithm that has superior guarantees. We also study lower bounds for our setting. In the case of undirected graphs, we show that the information-theoretically attainable regret is precisely characterized by the average *independence number* (or stability number) of the graph, namely the size of its largest independent set. For the case of directed graphs, we obtain a weaker regret which depends on the average *clique-partition number* of the graphs. More specifically, our contributions are as follows:

- We formally define and initiate a study of the setting that interpolates between learning with expert advice (with $\mathcal{O}(\sqrt{\log(k)T})$ regret) that assumes that all rewards are revealed and the multi-armed bandits setting (with $\tilde{\mathcal{O}}(\sqrt{kT})$ regret) that assumes that only the reward of the action selected is revealed. We provide an answer to a range of models in between.
- The framework we consider assumes that by choosing each action, other than just obtaining that action's reward, we can also observe some side-information about the rewards of other actions. We formalize this as a graph G_t over the actions, where an edge between two actions means that by choosing one action, we can also get a "sufficiently good" estimate of the reward of the other action. We consider both the case where G_t changes at each round t, as well as the case that $G_t = G$ is fixed throughout all rounds.
- We establish upper and lower bounds on the achievable regret, which depends on two combinatorial properties of G_t: Its independence number $\alpha(G_t)$ (namely, the largest number of nodes without edges between them), and its clique-partition number $\bar{\chi}(G_t)$ (namely, the smallest number of cliques into which the nodes can be partitioned).

- We present two practical algorithms to deal with this setting. The first algorithm, called ExpBan, combines existing algorithms in a natural way, and applies only when $G_t = G$ is fixed at all T rounds. Ignoring computational constraints, the algorithm achieves a regret bound of $\mathcal{O}(\sqrt{\bar{\chi}(G)\log(k)T})$. With computational constraints, its regret bound is $\mathcal{O}(\sqrt{c\log(k)T})$, where c is the size of the minimal clique partition one can efficiently find for G. However, note that for general graphs, it is NP-hard to find a clique partition for which $c = \mathcal{O}(k^{1-\epsilon})$ for any $\epsilon > 0$.
- The second algorithm, called ELP, is an improved algorithm, which can handle graphs which change between rounds. For undirected graphs, where sampling i gives an observation on j and vice versa, it achieves a regret bound of $\mathcal{O}(\sqrt{\log(k)\sum_{t=1}^{T}\alpha(G_t)})$. For directed graphs (where the observation structure is not symmetric), our regret bound is at most $\mathcal{O}(\sqrt{\log(k)\sum_{t=1}^{T}\bar{\chi}(G_t)})$. Moreover, the algorithm is computationally efficient. This is in contrast to the ExpBan algorithm, which in the worst case, cannot efficiently achieve regret significantly better than $\mathcal{O}(\sqrt{k\log(k)T})$.
- For the case of a fixed graph $G_t = G$, we present an information-theoretic $\Omega\left(\sqrt{\alpha(G)T}\right)$ lower bound on the regret, which holds regardless of computational efficiency.
- We present some simple synthetic experiments, which demonstrate that the potential advantage of the ELP algorithm over other approaches is real, and not just an artifact of our analysis.

1.1 Related Work

The standard multi-armed bandits problem assumes no relationship between the actions. Quite a few papers studied alternative models, where the actions are endowed with a richer structure. However, in the large majority of such papers, the feedback structure is the same as in the standard multi-armed bandits. Examples include [11], where the actions' rewards are assumed to be drawn from a statistical distribution, with correlations between the actions; and [1, 8], where the actions reward's are assumed to satisfy some Lipschitz continuity property with respect to a distance measure between the actions.

In terms of other approaches, the combinatorial bandits framework [7] considers a setting slightly similar to ours, in that one chooses and observes the rewards of some subset of actions. However, it is crucially assumed that the reward obtained is the sum of the rewards of all actions in the subset. In other words, there is no separation between earning a reward and obtaining information on its value. Another relevant approach is partial monitoring, which is a very general framework for online learning under partial feedback. However, this generality comes at the price of tractability for all but specific cases, which do not include our model.

Our work is also somewhat related to the contextual bandit problem (e.g., [9, 10]), where the standard multi-armed bandits setting is augmented with some side-information provided in each round, which can be used to determine which action to pick. While we also consider additional side-information, it is in a more specific sense. Moreover, our goal is still to compete against the best single action, rather than some set of policies which use this side-information.

2 Problem Setting

Let $[k] = \{1, \ldots, k\}$ and $[T] = \{1, \ldots, T\}$. We consider a set of actions $1, 2, \ldots, k$. Choosing an action i at round t results in receiving a reward $g_i(t)$, which we shall assume without loss of generality to be bounded in $[0, 1]$. Following the standard adversarial framework, we make no assumptions whatsoever about how the rewards are selected, and they might even be chosen by an adversary. We denote our choice of action at round t as i_t. Our goal is to minimize regret with respect to the best single action in hindsight, namely

$$\max_i \sum_{t=1}^{T} g_i(t) - \sum_{t=1}^{T} g_{i_t}(t).$$

Algorithm 1 The ExpBan Algorithm

Input: neighborhood sets $\{N_i(t)\}_{i\in[k]}$.
Split the graph induced by the neighborhood sets into c cliques ($c \leq k$ as small as possible)
For each clique, define a "meta-action" to be a standard experts algorithm over the actions in the clique
Run a multi-armed-bandits algorithm over the c meta-actions

For simplicity, we will focus on a finite-horizon setting (where the number of rounds T is known in advance), on regret bounds which hold in expectation, and on oblivious adversaries, namely that the reward sequence $g_i(t)$ is unknown but fixed in advance (see Sec. 8 for more on this issue).

Each round t, the learning algorithm chooses a single action i_t. In the standard multi-armed bandits setting, this results in $g_{i_t}(t)$ being revealed to the algorithm, while $g_j(t)$ remains unknown for any $j \neq i_t$. In our setting, we assume that by choosing an action i, other than getting $g_i(t)$, we also get some side-observations about the rewards of the other actions. Formally, we assume that one receives $g_i(t)$, and for some fixed parameter b is able to construct unbiased estimates $\hat{g}_j(t)$ for all actions j in some subset of $[k]$, such that $\mathbb{E}[\hat{g}_j(t)|\text{action i chosen}] = g_j(t)$ and $\Pr(|\hat{g}_j(t)| \leq b) = 1$. For any action j, we let $N_j(t)$ be the set of actions, for which we can get such an estimate $\hat{g}_j(t)$ on the reward of action j. This is essentially the "neighborhood" of action j, which receives sufficiently good information (as parameterized by b) on the reward of action j. We note that j is always a member of N_j, and moreover, N_j may be larger or smaller depending on the value of b we choose. We assume that $N_j(t)$ for all j, t are known to the learner in advance.

Intuitively, one can think of this setting as a sequence of graphs, one graph per round t, which captures the information feedback structure between the actions. Formally, we define G_t to be a graph on the k nodes $1, \ldots, k$, with an edge from node i to node j if and only if $j \in N_i(t)$. In the case that $j \in N_i(t)$ if and only if $i \in N_j(t)$, for all i, j, we say that G_t is undirected. We will use this graph viewpoint extensively in the remainder of the paper.

3 The ExpBan Algorithm

We begin by presenting the ExpBan algorithm (see Algorithm 1 above), which builds on existing algorithms to deal with our setting, in the special case where the graph structure remains fixed throughout the rounds - namely, $G_t = G$ for all t. The idea of the algorithm is to split the actions into c cliques, such that choosing an action in a clique reveals unbiased estimates of the rewards of all the other actions in the clique. By running a standard experts algorithm (such as the exponentially weighted forecaster - see [6, Chapter 2]), we can get low regret with respect to any action in that clique. We then treat each such expert algorithm as a meta-action, and run a standard bandits algorithm (such as the EXP3 [4]) over these c meta-actions. We denote this algorithm as ExpBan, since it combines an experts algorithm with a bandit algorithm.

The following result provides a bound on the expected regret of the algorithm. The proof appears in the appendix.

Theorem 1. *Suppose $G_t = G$ is fixed for all T rounds. If we run ExpBan using the exponentially weighted forecaster and the EXP3 algorithm, then the expected regret is bounded as follows:*[2]

$$\sum_{t=1}^{T} g_j(t) - \mathbb{E}\left[\sum_{t=1}^{T} g_{i_t}(t)\right] \leq 4b\sqrt{c\log(k)T}. \tag{1}$$

For the optimal clique partition, we have $c = \bar{\chi}(G)$, the clique-partition number of G.

It is easily seen that $\bar{\chi}(G)$ is a number between 1 and k. The case $\bar{\chi}(G) = 1$ corresponds to G being a clique, namely, that choosing any action allows us to estimate the rewards of all other actions. This corresponds to the standard experts setting, in which case the algorithm attains the optimal $\mathcal{O}(\sqrt{\log(k)T})$ regret. At the other extreme, $\bar{\chi}(G) = k$ corresponds to G being the empty graph,

[2] Using more sophisticated methods, it is now known that the $\log(k)$ factor can be removed (e.g., [3]). However, we will stick with this slightly less tight analysis for simplicity.

namely, that choosing any action only reveals the reward of that action. This corresponds to the standard bandit setting, in which case the algorithm attains the standard $\mathcal{O}(\sqrt{\log(k)kT})$ regret. For general graphs, our algorithm interpolates between these regimes, in a way which depends on $\bar{\chi}(G)$.

While being simple and using off-the-shelf components, the ExpBan algorithm has some disadvantages. First of all, for a general graph G, it is NP-hard to find $c \leq \mathcal{O}(k^{1-\epsilon})$ for any $\epsilon > 0$. (This follows from [12] and the fact that the clique-partition number of G equals the chromatic number of its complement.) Thus, with computational constraints, one cannot hope to obtain a bound better than $\tilde{\mathcal{O}}(\sqrt{kT})$. That being said, we note that this is only a worst-case result, and in practice or for specific classes of graphs, computing a good clique partition might be relatively easy. A second disadvantage of the algorithm is that it is not applicable for an observation structure that changes with time.

4 The ELP Algorithm

We now turn to present the ELP algorithm (which stands for "Exponentially-weighted algorithm with Linear Programming"). Like all multi-armed bandits algorithms, it is based on a tradeoff between exploration and exploitation. However, unlike standard algorithms, the exploration component is not uniform over the actions, but is chosen carefully to reflect the graph structure at each round. In fact, the optimal choice of the exploration requires us to solve a simple linear program, hence the name of the algorithm. Below, we present the pseudo-code as well as a couple of theorems that bound the expected regret of the algorithm under appropriate parameter choices. The proofs of the theorems appear in the appendix. The first theorem concerns the symmetric observation case, where if choosing action i gives information on action j, then choosing action j must also give information on i. The second theorem concerns the general case. We note that in both cases the graph G_t may change arbitrarily in time.

Algorithm 2 The ELP Algorithm

Input: $\beta, \{\gamma(t)\}_{t\in[T]}, \{s_i(t)\}_{i\in[k],t\in[T]}$, neighborhood sets $\{N_i(t)\}_{i\in[k],t\in[T]}$.
$\forall\, j \in [k] \quad w_j(1) := 1/k$.
for $t = 1, \ldots, T$ **do**
  $\forall\, i \in [k] \quad p_i(t) := (1-\gamma(t))\frac{w_i(t)}{\sum_{l=1}^{k} w_l(k)} + \gamma(t)s_i(t)$
  Choose action i_t with probability $p_{i_t}(t)$, and receive reward $g_{i_t}(t)$
  Compute $\hat{g}_j(t)$ for all $j \in N_{i_t}(t)$
  For all $j \in [k]$, let $\tilde{g}_j(t) = \frac{\hat{g}_j(t)}{\sum_{l\in N_j(t)} p_l(t)}$ if $i_t \in N_j(t)$, and $\tilde{g}_j(t) = 0$ otherwise.
  $\forall\, j \in [k] \quad w_j(t+1) = w_j(t)\exp(\beta\tilde{g}_j(t))$
end for

4.1 Undirected Graphs

The following theorem provides a regret bound for the algorithm, as well as appropriate parameter choices, in the case of undirected graphs. Later on, we will discuss the case of directed graphs. In a nutshell, the theorem shows that the regret bound depends on the average independence number $\alpha(G_t)$ of each graph G_t - namely, the size of its largest independent set.

Theorem 2. *Suppose that for all t, G_t is an undirected graph. Suppose we run Algorithm 2 using some $\beta \in (0, 1/2bk)$, and choosing*

$$\{s_i(t)\}_{i\in[k]} = \underset{\forall i\ s_i(t)\geq 0, \sum_i s_i(t)=1}{\operatorname{argmax}} \ \min_{j\in[k]} \sum_{l\in N_j(t)} s_l(t),$$

(which can be easily done via linear programming) and $\gamma(t) = \beta b / \min_{j\in[k]} \sum_{l\in N_j(t)} s_l(t)$. Then it holds for any fixed action j that

$$\sum_{t=1}^{T} g_j(t) - \mathbb{E}\left[\sum_{t=1}^{T} g_{i_t}(t)\right] \ \leq\ 3\beta b^2 \sum_{t=1}^{T} \alpha(G_t) + \frac{\log(k)}{\beta}. \tag{2}$$

If we choose $\beta = \sqrt{\log(k)/3b^2 \sum_t \alpha(G_t)}$*, then the bound equals*

$$b\sqrt{3\log(k)\sum_{t=1}^{T}\alpha(G_t)}. \tag{3}$$

Comparing Thm. 2 with Thm. 1, we note that for any graph G_t, its independence number $\alpha(G_t)$ lower bounds its clique-partition number $\bar{\chi}(G_t)$. In fact, the gap between them can be very large (see Sec. 6). Thus, the attainable regret using the ELP algorithm is better than the one attained by the ExpBan algorithm. Moreover, the ELP algorithm is able to deal with time-changing graphs, unlike the ExpBan algorithm.

If we take worst-case computational efficiency into account, things are slightly more involved. For the ELP algorithm, the optimal value of β, needed to obtain Eq. (3), requires knowledge of $\sum_{t=1}^{T}\alpha(G_t)$, but computing or approximating the $\alpha(G_t)$ is NP-hard in the worst case. However, there is a simple fix: we create $\lceil\log(k)\rceil$ copies of the ELP algorithm, where copy i assumes that $\sum_{t=1}^{T}\alpha(G_t)$ equals 2^{i-1}. Note that one of these values must be wrong by a factor of at most 2, so the regret of the algorithm using that value would be larger by a factor of at most 2. Of course, the problem is that we don't know in advance which of those $\lceil\log(k)\rceil$ copies is the best one. But this can be easily solved by treating each such copy as a "meta-action", and running a standard multi-armed bandits algorithm (such as EXP3) over these $\lceil\log(k)\rceil$ actions. Note that the same idea was used in the construction of the ExpBan algorithm. Since there are $\lceil\log(k)\rceil$ meta-actions, the additional regret incurred is $\mathcal{O}(\sqrt{\log^2(k)T})$. So up to logarithmic factors in k, we get the same regret as if we could actually compute the optimal value of β.

4.2 Directed Graphs

So far, we assumed that the graphs we are dealing with are all undirected. However, a natural extension of this setting is to assume a directed graph, where choosing an action i may give us information on the reward of action j, but not vice-versa. It is readily seen that the ExpBan algorithm would still work in this setting, with the same guarantee. For the ELP algorithm, we can provide the following guarantee:

Theorem 3. *Under the conditions of Thm. 2 (with the relaxation that the graphs G_t may be directed), it holds for any fixed action j that*

$$\sum_{t=1}^{T} g_j(t) - \mathbb{E}\left[\sum_{t=1}^{T} g_{i_t}(t)\right] \;\leq\; 3\beta b^2 \sum_{t=1}^{T}\bar{\chi}(G_t), + \frac{\log(k)}{\beta}. \tag{4}$$

where $\bar{\chi}(G_t)$ is the clique-partition number of G_t. If we choose $\beta = \sqrt{log(k)/3b^2 \sum_t \bar{\chi}(G_t)}$, then the bound equals

$$b\sqrt{3\log(k)\sum_{t=1}^{T}\bar{\chi}(G_t)}. \tag{5}$$

Note that this bound is weaker than the one of Thm. 2, since $\alpha(G_t) \leq \bar{\chi}(G_t)$ as discussed earlier. We do not know whether this bound (relying on the clique-partition number) is tight, but we conjecture that the independence number, which appears to be the key quantity in undirected graphs, is not the correct combinatorial measure for the case of directed graphs[3]. In any case, we note that even with the weaker bound above, the ELP algorithm still seems superior to the ExpBan algorithm, in the sense that it allows us to deal with time-changing graphs, and that an explicit clique decomposition of the graph is not required. Also, we again have the issue of β which is determined by a quantity which is NP-hard to compute, i.e. $\bar{\chi}(G_t)$. However, this can be circumvented using the same trick discussed in the context of undirected graphs.

[3]It is possible to construct examples where the analysis of the ELP algorithm necessarily leads to an $\mathcal{O}(\sqrt{k\log(k)T})$ bound, even when the independence number is 1

5 Lower Bound

The following theorem provides a lower bound on the regret in terms of the independence number $\alpha(G)$, for a constant graph $G_t = G$.

Theorem 4. *Suppose $G_t = G$ for all t, and that actions which are not linked in G get no side-observations whatsoever between them. Then there exists a (randomized) adversary strategy, such that for every $T \geq 374\alpha(G)^3$ and any learning strategy, the expected regret is at least $0.06\sqrt{\alpha(G)T}$.*

A proof is provided in the appendix. The intuition of the proof is that if the graph G has $\alpha(G)$ independent vertices, then an adversary can make this problem as hard as a standard multi-armed bandits problem, played on $\alpha(G)$ actions. Using a known lower bound of $\Omega(\sqrt{nT})$ for multi-armed bandits on n actions, our result follows[4].

For constant undirected graphs, this lower bound matches the regret upper bound for the ELP algorithm (Thm. 2) up to logarithmic factors. For directed graphs, the difference between them boils down to the difference between $\bar{\chi}(G)$ and $\alpha(G)$. For many well-behaved graphs, this gap is rather small. However, for general graphs, the difference can be huge - see the next section for details.

6 Examples

Here, we briefly discuss some concrete examples of graphs G, and show how the regret performance of our algorithms depend on their structure. An interesting issue to notice is the potential gap between the performance of our algorithms, through the graph's independence number $\alpha(G)$ and clique-partition number $\bar{\chi}(G)$.

First, consider the case where there exists a single action, such that choosing it reveals the rewards of all the other actions. In contrast, choosing the other actions only reveal their own reward. At first blush, it may seem that having such a "super-action", which reveals everything that happens in the current round, should help us improve our regret. However, the independence number $\alpha(G)$ of such a graph is easily seen to be $k-1$. Based on our lower bound, we see that this "super-action" is actually not helpful at all (up to negligible factors).

Second, consider the case where the actions are endowed with some metric distance function, and edge (i,j) is in G if and only if the distance between i,j is at most some fixed constant r. We can think of each action i as being in the center of a sphere of radius r, such that the reward of action i is propagated to every other action in that sphere. In this case, $\alpha(G)$ is essentially the number of non-overlapping spheres we can *pack* in G. In contrast, $\bar{\chi}(G)$ is essentially the number of spheres we need to *cover* G. Both numbers shrink rapidly as r increases, improving the regret of our algorithms. However, the sphere covering size can be much larger than the sphere packing size. For example, if the actions are placed as the elements in $\{0, 1/2, 1\}^n$, we use the l_∞ metric, and $r \in (1/2, 1)$, it is easily seen that the sphere packing number is just 1. In contrast, the sphere covering number is at least $2^n = k^{\log_3(2)} \approx k^{0.63}$, since we need a separate sphere to cover every element in $\{0,1\}^n$.

Third, consider the random Erdös - Rényi graph $G = G(k,p)$, which is formed by linking every action i to every action j with probability p independently. It is well known that when p is a constant, the independence number $\alpha(G)$ of this graph is only $\mathcal{O}(\log(k))$, whereas the clique-partition number $\bar{\chi}(G)$ is at least $\Omega(k/\log(k))$. This translates to a regret bound of $\mathcal{O}(\sqrt{kT})$ for the ExpBan algorithm, and only $\mathcal{O}(\sqrt{\log^2(k)T})$ for the ELP algorithm. Such a gap would also hold for a directed random graph.

7 Empirical Performance Gap between ExpBan and ELP

In this section, we show that the gap between the performance of the ExpBan algorithm and the ELP algorithm can be real, and is not just an artifact of our analysis.

[4] We note that if the maximal degree of every node is bounded by d, it is possible to get the lower bound for $T \geq \Omega(d^2\alpha(G))$ (as opposed to $T \geq \Omega(\alpha(G)^3)$); see the proof for details.

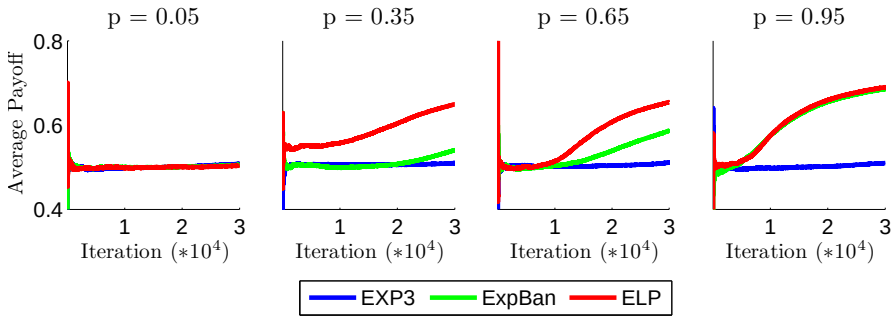

Figure 1: Experiments on random graphs.

To show this, we performed the following simple experiment: we created a random Erdös - Rényi graph over 300 nodes, where each pair of nodes were linked independently with probability p. Choosing any action results in observing the rewards of neighboring actions in the graph. The reward of each action at each round was chosen randomly and independently to be 1 with probability $1/2$ and 0 with probability $1/2$, except for a single node, whose reward equals 1 with a higher probability of $3/4$. We then implemented the ExpBan and ELP algorithms in this setting, for $T = 30,000$. For comparison, we also implemented the standard EXP3 multi-armed bandits algorithm [4], which doesn't use any side-observations. All the parameters were set to their theoretically optimal values. The experiment was repeated for varying p and over 10 independent runs.

The results are displayed in Figure 1. The X-axis is the iteration number, and the Y-axis is the mean payoff obtained so far, averaged over the 10 runs (the variance in the numbers was minuscule, and therefore we do not report confidence intervals). For $p = 0.05$, the graph is rather empty, and the advantage of using side observations is not large. As a result, all 3 algorithms perform roughly the same for this choice of T. As p increases, the value of side-obervations increase, and the the performance of our two algorithms, which utilize side-observations, improves over the standard multi-armed bandits algorithm. Moreover, for intermediate values of p, there is a noticeable gap between the performance of ExpBan and ELP. This is exactly the regime where the gap between the clique-partition number (governing the regret bound of ExpBan) and the independence number (governing the regret bound for the ELP algorithm) tends to be larger as well[5]. Finally, for large p, the graph is almost complete, and the advantage of ELP over ExpBan becomes small again (since most actions give information on most other actions).

8 Discussion

In this paper, we initiated a study of a large family of online learning problems with side observations. In particular, we studied the broad regime which interpolates between the experts setting and the bandits setting of online learning. We provided algorithms, as well as upper and lower bounds on the attainable regret, with a non-trivial dependence on the information feedback structure.

There are many open questions that warrant further study. First, the upper and lower bounds essentially match only in particular settings (i.e., in undirected graphs, where no side-observations whatsoever, other than those dictated by the graph are allowed). Can this gap be narrowed or closed? Second, our lower bounds depend on a reduction which essentially assumes that the graph is constant over time. We do not have a lower bound for changing graphs. Third, it remains to be seen whether other online learning results can be generalized to our setting, such as learning with respect to policies (as in EXP4 [4]) and obtaining bounds which hold with high probability. Fourth, the model we have studied assumed that the observation structure is known. In many practical cases, the observation structure may be known just partially or approximately. Is it possible to devise algorithms for such cases?

Acknowledgements. This research was supported in part by the Google Inter-university center for Electronic Markets and Auctions.

[5]Intuitively, this can be seen by considering the extreme cases - for a complete graph over k nodes, both numbers equal 1, and for an empty graph over k nodes, both numbers equal k. For constant $p \in (0, 1)$, there is a real gap between the two, as discussed in Sec. 6

References

[1] R. Agrawal. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33:1926–1951, 1995.

[2] H. Arslan, Z. N. Chen, and M. G. Di Benedetto. *Ultra Wideband Wireless Communication*. Wiley - Interscience, 2006.

[3] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, 2009.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

[5] V. Baston. Some cyclic inequalities. *Proceedings of the Edinburgh Mathematical Society (Series 2)*, 19:115–118, 1974.

[6] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[7] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. In *COLT*, 2009.

[8] R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *STOC*, pages 681–690, 2008.

[9] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, 2007.

[10] L. Li, W. Chu, J. Langford, and R. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[11] P. Rusmevichientong and J. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35(2):395–411, 2010.

[12] D. Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, 3(1):103–128, 2007.