Online Learning with Uncertain Feedback Graphs

Pouya M. Ghari, and Yanning Shen, *Member, IEEE*

***Abstract*—Online learning with expert advice is widely used in various machine learning tasks. It considers the problem where a learner chooses one from a set of experts to take advice and make a decision. In many learning problems, experts may be related, henceforth the learner can observe the losses associated with a subset of experts that are related to the chosen one. In this context, the relationship among experts can be captured by a feedback graph, which can be used to assist the learner's decision-making. However, in practice, the nominal feedback graph often entails uncertainties, which renders it impossible to reveal the actual relationship among experts. To cope with this challenge, the present work studies various cases of potential uncertainties and develops novel online learning algorithms to deal with uncertainties while making use of the uncertain feedback graph. The proposed algorithms are proved to enjoy sublinear regret under mild conditions. Experiments on real datasets are presented to demonstrate the effectiveness of the novel algorithms.**

***Index Terms*—Online Learning, Graphs, Expert Advice, Uncertainty.**

I. Introduction

Online learning with expert advice considers the case where there exists a learner and a set of experts, where the learner interacts with the experts to make a decision [2]. At each time instant, the learner chooses one of the experts and it takes the action advised by the chosen expert, then incurs the loss associated with the taken action. Such framework can be used to model different learning tasks such as online multi-kernel learning see e.g., [3], [4]. Conventional online learning literature mostly focuses on two settings, *full information* setting [5]–[8] or *bandit* setting [8]–[11]. In the full information setting, at each time instant, the learner can observe the loss associated with all experts. By contrast, in the bandit setting, the learner can only observe the loss associated with the chosen expert. However, in some applications such as the web advertising problem, where a user clicks on an ad and information about other related ads is revealed, the learner can make partial observations of losses associated with a subset of experts. In cases where querying for advice from expert incurs cost, the learner may choose to observe the loss of subset of experts, see e.g. [12], [13]. To cope with this scenario, *online learning with feedback graphs* was developed in [14], where partial observations of losses are modeled using a directed *feedback graph*. Each node represents an expert, and an edge from node i to node j exists if the learner can observe the loss associated with expert j while choosing expert i. The observations of losses associated with other experts are called side observations. The full information and the bandit settings are both special cases of online learning with either a fully connected feedback graph or a feedback graph with only self loops. Given the feedback graph either before or after decision making, [15] has proposed algorithms with sub-linear regret bounds. Online learning with feedback graphs and sleeping experts has been studied in [16] where at each time instant, a subset of experts may not be available. [17] has studied the case where there is a dependency between the feedback graph and expert losses. Moreover, [10] has proposed an algorithm for bandit setting which obtains sub-linear regret with respect to the best switching expert selection strategy.

Most of existing works rely on the assumption that the feedback graph in known *perfectly* before decision making [15], [16], [18]–[20], or after decision making [15], [17], [21]–[23]. However, such information may not be available in practice. In addition, due to possible uncertainty of the environment, the feedback graph may be uncertain. As an example, consider an online clothing store that offers discount on an item for new customers. Suppose there are two brands A and B producing similar shirts at comparable price. The store has small and medium sizes of brand A and medium and large sizes shirts of brand B in stock. Assuming that the store offers discount on brand B. If the user accepts the offer, and buys a medium size shirt of brand B, it implies the user is also interested in shirts of brand A. Moreover, if the user buys a large size of shirt B, this indicates no interest in shirts of brand A. Otherwise, if the user declines the offer of brand B, it only shows the user is not interested in shirts of brand B but no information is available about the preference of the user on the shirts of brand A. Considering the case where the exact feedback graph may not be available, [24] shows that not knowing the entire feedback graph can make the side observations useless and the learner may simply ignore them. [25] studies the case where the exact feedback graph is unknown but is known to be generated from the Erdös-Rényi model. However, such assumption may not be valid in practice. In addition, both [24] and [25] assume that the loss associated with the chosen expert is guaranteed to be observed. Moreover, the probabilistic feedback graph in stochastic setting has been studied in [26] where the loss of each expert randomly generated using a certain probability distribution.

The present paper extensively studies the case where the learner only has access to a feedback graph that may contain uncertainties, namely *nominal feedback graph*, and the learner may not be able to observe the loss associated with the chosen expert. Moreover, the present paper studies non-stochastic adversarial online learning problems where at each time instant, the environment privately selects a loss function. The learner relies on the nominal feedback graph to choose among experts, and then incurs a loss associated with the chosen expert. At the same time, it observes the loss associated with a subset

P. M. Ghari and Y. Shen are with the Department of Electrical Engineering and Computer Science, University of California, Irvine, CA, USA. Emails: pmollaeb@uci.edu and yannings@uci.edu. Preliminary results of this work were presented in part at the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [1]. The work in this paper is supported by NSF ECCS 2207457. Corresponding author: Yanning Shen.

of experts resulting from the unknown actual feedback graph. Furthermore, different from [24] and [25], the present work does not assume that it is guaranteed that the learner observes the loss associated with the chosen expert. This is true in, e.g., apple tasting problem [27]. The present work studies various cases of potential uncertainties, and develops novel online learning algorithms to cope with different uncertainties in the nominal feedback graph. Regret analysis is provided to prove that our novel algorithms can achieve sublinear regret under mild conditions. Experiments on a number of real datasets are presented to showcase the effectiveness of our novel algorithms.

II. Problem Statement

Consider the case where there exist K experts and the learner chooses to take the advice of one of the experts at each time instant t. Let $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$ represent the directed nominal feedback graph at time t with a set of vertices $\mathcal{V}$, where the vertex $v_i \in \mathcal{V}$ represents the i-th expert, and there exist an edge from v_i to v_j (i.e. $(i,j) \in \mathcal{E}_t$), if the learner observes the loss associated with the j-th expert (i.e. $\ell_t(v_j)$) with probability p_{ij} while choosing the i-th expert. Let $\mathcal{N}_{i,t}^{\text{in}}$ and $\mathcal{N}_{i,t}^{\text{out}}$ represent in-neighborhood and out-neighborhood of v_i in $\mathcal{G}_t$, respectively. Thus, $v_j \in \mathcal{N}_{i,t}^{\text{out}}$ if there is an edge from v_i to v_j at time t (i.e. $(i,j) \in \mathcal{E}_t$). Similarly, $v_j \in \mathcal{N}_{i,t}^{\text{in}}$ if there is an edge from v_j to v_i at time t (i.e. $(j,i) \in \mathcal{E}_t$). The present paper considers non-stochastic adversarial online learning problems. At each time instant t, the environment privately selects a loss function $\ell_t(.)$ with $\ell_t(.) : \mathcal{V} \to [0,1]$, and the nominal feedback graph $\mathcal{G}_t$ is revealed to the learner before decision making. The learner then chooses one of the experts to take its advice. Then, the learner will incur the loss associated with the chosen expert. Let I_t denote the index of the chosen expert. Note that the learner observes $\ell_t(v_{I_t})$ with probability of $p_{I_t I_t}$, hence the loss remains unknown with the probability of $1 - p_{I_t I_t}$.

The present paper discusses different potential uncertainties in the feedback graphs, and develops novel algorithms for online learning with uncertain feedback graph. Specifically, two cases are discussed: i) *online learning with informative probabilistic feedback graph*: where the probability p_{ij} associated with each edge is given along with the nominal feedback graph $\mathcal{G}_t$; and ii) *online learning with uninformative probabilistic feedback graph*: where only the nominal feedback graph $\mathcal{G}_t$ is revealed, but not the probabilities.

III. Online Learning with Informative Probabilistic Feedback Graphs

First consider the case where $\{p_{ij}\}$ are given along with the $\mathcal{G}_t$. This can be the case in various applications. For instance, consider a network of agents in a wireless sensor network that cooperate with each other on certain tasks such as environmental monitoring. Online learning algorithms distributed over spatial locations have been employed in climate informatics field [28], [29]. Assume that each agent in the network keeps updating its local model, and there is a central unit (learner) wishes to perform a learning task based on models and data samples distributed among agents. In this case, the agents in the network can be viewed as experts. Consider the case where the learner chooses one of the experts and sends a request for the corresponding expert advice through a wireless link. Subset of experts which receive the request, send their advice to the learner. However, due to uncertainty in the environment or power limitation, some of the agents in the network including the chosen one may not detect the request. Therefore, the learner can only observe the advice of subset of agents in the network which detect its request. In this case, the learner can model probable advice that it can receive from experts with a nominal feedback graph. If learner knows the characteristics of the environment which is true in many wireless communication applications, the probabilities associated with edges in the nominal feedback graph is revealed.

Algorithm 1 Exp3-IP: Online learning with informative probabilistic feedback graph

Input: learning rate $\eta > 0$.
Initialize: $w_{i,1} = 1$, $\forall i \in [K]$.
for $t = 1, \ldots, T$ **do**
  Observe $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$ and choose one of the experts according to the PMF π_t in (3).
  Observe $\{\ell_t(v_i)\}_{v_i \in \mathcal{S}_t}$ and calculate loss estimate $\hat{\ell}_t(v_i)$, $\forall i \in [K]$ via (2).
  Update $w_{i,t+1}$, $\forall i \in [K]$ via (1).
end for

At each time instant t, upon selecting an expert and observing the losses of a subset of experts, the weights $\{w_{i,t}\}_{i=1}^K$ which indicate the reliability of experts can be updated as follows

$$w_{i,t+1} = w_{i,t} \exp\left(-\eta \hat{\ell}_t(v_i)\right), \quad \forall i \in [K] \tag{1}$$

where $[K] := \{1, \ldots, K\}$ and η is the learning rate. Function $\hat{\ell}_t(v_i)$ denotes the importance sampling loss estimate which can be obtained as

$$\hat{\ell}_t(v_i) = \frac{\ell_t(v_i)}{q_{i,t}} \mathcal{I}(v_i \in \mathcal{S}_t) \tag{2}$$

where $\mathcal{S}_t$ represent the set of vertices associated with experts whose losses are observed by the learner at time instant t. The indicator function is denoted by $\mathcal{I}(.)$ and $q_{i,t}$ is the probability that the loss $\ell_t(v_i)$ is observed. Its value depends on the algorithm, and will be specified later.

Let A_t denote the adjacency matrix of the nominal feedback graph $\mathcal{G}_t$ with $A_t(i,j)$ denoting the (i,j)th entry of A_t. Let X_{ij} be a Bernoulli random process with random variables $X_{ij}(t) = 1$ with probability p_{ij}. When the learner chooses the i-th expert at time t, the learner observes $\ell_t(v_j)$ only if $v_j \in \mathcal{N}_{i,t}^{\text{out}}$ and $X_{ij}(t) = 1$. Let F_t denote the number of losses observed by the learner. Due to the stochastic nature of the observations available to the learner, F_t is a random variable. Furthermore, let $F_{i,t}$ denote the expected number of observed losses if the learner chooses the i-th expert at time t. Thus, we can write

$$F_{i,t} = \mathbb{E}_t[F_t | I_t = i, A_t] = \sum_{\forall j : v_j \in \mathcal{N}_{i,t}^{\text{out}}} \mathbb{E}[X_{ij}(t)] = \sum_{\forall j : v_j \in \mathcal{N}_{i,t}^{\text{out}}} p_{ij}.$$

The learner then chooses one expert according to the probability mass function (PMF) $\pi_t := (\pi_{1,t}, \ldots, \pi_{K,t})$ with

$$\pi_{i,t} = (1-\eta)\frac{w_{i,t}}{W_t} + \eta\frac{F_{i,t}}{\sum_{j\in\mathcal{D}_t} F_{j,t}}\mathcal{I}(v_i \in \mathcal{D}_t) \tag{3}$$

where $W_t := \sum_{i=1}^{K} w_{i,t}$, and $\mathcal{D}_t$ denotes the dominating set of graph $\mathcal{G}_t$. Note that a dominating set $\mathcal{D}$ of a graph is a subset of vertices such that there is an edge from at least one vertex in $\mathcal{D}$ to any vertex not in $\mathcal{D}$. It can be observed from (3) that η controls the trade-off between exploitation and exploration. With a smaller η, more emphasis is placed on the first term which promotes exploitation, and the learner tends to choose the expert with larger $w_{i,t}$. The second term allows the learner to select experts in the dominating set $\mathcal{D}_t$ with certain probability independent of their performance in previous rounds. Based on (3), $q_{i,t}$ in (2) can be computed as

$$q_{i,t} = \sum_{\forall j: v_j \in \mathcal{N}_{i,t}^{\text{in}}} \pi_{j,t} p_{ji}. \tag{4}$$

The overall algorithm for online learning with uncertain feedback graph in the informative probabilistic setting, termed Exp3-IP, is summarized in Algorithm 1. In order to analyze the performance of Algorithm 1, as well as the ensuing algorithms, we first preset two assumptions needed:
(a1) $0 \le \ell_t(v_i) \le 1$, $\forall t : t \in \{1, \ldots, T\}, \forall i : i \in \{1, \ldots, K\}$.
(a2) If $(i,j) \in \mathcal{E}_t$, the learner can observe the loss associated with the j-th expert with probability at least $\epsilon > 0$ when it chooses the i-th expert, and $(i,i) \in \mathcal{E}_t,\ \forall i$.

Note that (a1) is a general assumption in online learning literature e.g., [18]. And (a2) assumes a nonzero probability of observing (but not guaranteed observation of) the loss associated with the chosen expert $\ell_t(v_{I_t})$. The following theorem presents the regret bound for Exp3-IP.

Theorem 1. *Under (a1), the expected regret of Exp3-IP can be bounded by*

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i\in\mathcal{V}}\sum_{t=1}^{T}\ell_t(v_i)\\ \le&\frac{\ln K}{\eta} + \eta(1-\frac{\eta}{2})T + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}}.\end{aligned} \tag{5}$$

Proof of Theorem 1 is included in Appendix A. It can be seen from Theorem 1 that the value of $\pi_{i,t}/q_{i,t}$ plays an important role in regret bound. Choosing an expert using (3), it is ensured that every vertex in $\mathcal{D}_t$ is chosen by the learner with non-zero probability. Moreover, since there is at least one edge from a node in $\mathcal{D}_t$ to any node not in $\mathcal{D}_t$, under (a2), the probability $q_{i,t}$, $\forall i$ is non-zero. Lower bounding $q_{i,t}$, (a2) enables Exp3-IP to achieve sub-linear regret. Building upon Theorem 1, the ensuing lemma further explores under which circumstances Exp3-IP can achieve sub-linear regret bound.

Lemma 2. *Let the doubling trick (see e.g. [15]) be employed to determine the value of η and greedy set cover algorithm (see e.g. [30]) is exploited to derive a dominating set $\mathcal{D}_t$ for the nominal feedback graph $\mathcal{G}_t$. Under (a1) and (a2), the expected regret of Exp3-IP satisfies*

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i\in\mathcal{V}}\sum_{t=1}^{T}\ell_t(v_i)\\ \le&\mathcal{O}\left(\sqrt{\ln K \ln(\frac{K}{\epsilon}T)\sum_{t=1}^{T}\frac{\alpha(\mathcal{G}_t)}{\epsilon}} + \ln(\frac{K}{\epsilon}T)\right)\end{aligned} \tag{6}$$

where $\alpha(\mathcal{G}_t)$ denotes the independence number of the nominal feedback graph $\mathcal{G}_t$.

Proof of Lemma 2 is included in Appendix B. As it is proved in Appendix B, the assumption $(i,i) \in \mathcal{E}_t,\ \forall i$ in (a2) guarantees that $\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}} \le \mathcal{O}\left(\frac{\alpha(\mathcal{G}_t)}{\epsilon}\ln(\frac{KT}{\epsilon})\right)$ (see Lemma 7 and (51)–(54) in Appendix B). In order to guarantee the regret bound in (6), it is required that $\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}} \le \mathcal{O}\left(\frac{\alpha(\mathcal{G}_t)}{\epsilon}\ln(\frac{KT}{\epsilon})\right)$ holds true. Therefore, without (a2), the regret bound in (6) cannot be satisfied. Furthermore, if the learner does not know the time horizon T before start decision making, doubling trick can be exploited to determine η. In particular, using the doubling trick, Exp3-IP adjusts the learning rate η 'on the fly' without knowing the time horizon T. At time instant t, as long as

$$\sum_{\tau=1}^{t}(1+\frac{1}{2}\sum_{i=1}^{K}\frac{\pi_{i,\tau}}{q_{i,\tau}}) \le 2^{r_t} \tag{7}$$

holds true, Exp3-IP employs learning rate $\eta = \sqrt{\frac{\ln K}{2^{r_t+1}}}$, where $r_t \ge 0$ is the smallest integer that can satisfy the inequality in (7). According to (6), Exp3-IP can achieve sub-linear regret. Furthermore, (6) shows that the regret bound of Exp3-IP depends on $\frac{1}{\epsilon}$. Larger ϵ indicates that the learner is less uncertain about the nominal feedback graph. In other words higher confidence of the nominal feedback graph leads to a tighter regret bound.

Comparison with [15]. Exp3-DOM of [15] deals with the cases that the feedback graph is certain and revealed to the learner before decision making at each time instant. In this case, Exp3-DOM achieves regret of $\mathcal{O}\left(\ln(K)\sqrt{\ln(KT)\sum_{t=1}^{T}\alpha(\mathcal{G}_t)} + \ln(K)\ln(KT)\right)$ (see Theorem 8 in [15]). When the graph is certain such that $p_{ij} = 1$ for all $(i,j) \in \mathcal{E}$, then $\epsilon = 1$. Therefore, when the graph is certain and given to the learner, the proposed Exp3-IP achieves regret of $\mathcal{O}\left(\sqrt{\ln K \ln(KT)\sum_{t=1}^{T}\alpha(\mathcal{G}_t)} + \ln(KT)\right)$.

IV. Online Learning with Uninformative Probabilistic Feedback Graphs

The previous section deals with the case where the nominal feedback graph $\mathcal{G}_t$ can be time-variant and probabilities associated with edges of $\mathcal{G}_t$ are revealed. In this section, we will study the scenario where the nominal feedback graph $\mathcal{G}_t$ is static and is revealed to the learner while the probabilities $\{p_{ij}\}$ associated with edges are not given, which is called *uninformative probabilistic feedback graph*. In this section the nominal feedback graph is denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. In this case, estimates of probabilities $\{p_{ij}\}$ can be updated and

employed to assist the learner with future decision making. For example, consider the problem of online advertisement, where a website is trying to decide which product to be advertised via online survey with a multiple choice question. Specifically, users are asked whether they are interested in certain product along with possible reasons (cost, color, etc). Note that the answer to certain product may also indicate the participant's potential interest in other products with similar cost or color. For instance, if the participant indicates that he or she is interested in the product because of its affordable cost, this implies *potential* interest in other products with the same or lower price. In this case, the relationship among products can be modeled by a nominal feedback graph, where an edge exists between two nodes (products) if they share same or similar attributes (cost, color), which implies that users *may be* interested in both products. Such nominal feedback graph can then be used to assist the website to make a decision on which product to advertise . However, the actual relationship between the the user's interests in the products remains uncertain, which leads to uncertainty in the nominal feedback graph. Since attributes (cost, color, etc) of products do not change over time, the nominal feedback graph is static, while the probabilities associated with edges in the nominal feedback graph are unknown. Faced with this practical challenge, two approaches will be developed in this section, to estimate either the unknown probability or the importance sampling loss in (2), which will then be employed to assist the learner's decision making.

A. Estimation-based Approach

In the present subsection, we will further explore the general scenario where the value of p_{ij} may vary across edges, while the nominal feedback graph $\mathcal{G}_t$ is static. Since X_{ij} defined under (2) is a mean ergodic random process [31] in this scenario, the sample mean of $\{X_{ij}(t)\}$ converges to p_{ij}, i.e., the expected value of $X_{ij}(t)$. Let $\mathcal{T}_{ij,t}$ represent a set collecting time instants before t when the learner chooses to take the advice of the i-th expert and there is an edge between v_i and v_j in the nominal feedback graph $\mathcal{G}$. In other word, $\mathcal{T}_{ij,t}$ can be defined as

$$\mathcal{T}_{ij,t} = \{\tau | A_\tau(i,j) = 1, I_\tau = i, 0 < \tau < t\}. \tag{8}$$

Based on the above discussion, p_{ij} can be estimated as

$$\hat{p}_{ij,t} = \frac{1}{C_{ij,t}} \sum_{\tau \in \mathcal{T}_{ij,t}} X_{ij}(\tau) \tag{9}$$

where $C_{ij,t} := |\mathcal{T}_{ij,t}|$ is the cardinality of $\mathcal{T}_{ij,t}$. Since X_{ij} is a mean ergodic Bernoulli random process, $\hat{p}_{ij,t}$ is an unbiased maximum likelihood (ML) estimator of p_{ij}.

Note that a sufficient number of observations of the random process X_{ij} is needed, in order to provide a reliable estimation in (9). To this end, the learner performs exploration in the first KM time instants to ensure that $C_{ij,t} \geq M$, $\forall (i,j) \in \mathcal{E}_t$, where the value of M is determined by the learner. Specifically, in the first KM time instants, the learner chooses all experts in $\mathcal{V}$, one by one M times, i.e. the learner selects expert v_k, with $k = t - \lfloor \frac{t}{K} \rfloor K$ when $t \leq KM$. For $t > KM$, the learner draws one of the experts according to the following PMF

$$\pi_{i,t} = (1-\eta)\frac{w_{i,t}}{W_t} + \frac{\eta}{|\mathcal{D}|}\mathcal{I}(v_i \in \mathcal{D}), \forall i \in [K] \tag{10}$$

where $\mathcal{D}$ denotes a dominating set for the nominal feedback graph $\mathcal{G}$. In order to obtain a reliable loss estimate to assist the learner's decision making, we will approximate the importance sampling loss estimate in (2) using the estimated probability $\hat{p}_{ij,t}$. In this context, the probability of observing $\ell_t(v_i)$ can be approximated as

$$\hat{q}_{i,t} = \sum_{\forall j: v_j \in \mathcal{N}_{i,t}^{\text{in}}} \pi_{j,t}(\hat{p}_{ji,t} + \frac{\xi}{\sqrt{M}}) \tag{11}$$

where $\xi \geq 1$ is a parameter selected by the learner. Then the importance sampling loss estimates can be obtained as

$$\tilde{\ell}_t(v_i) = \frac{\ell_t(v_i)}{\hat{q}_{i,t}} \mathcal{I}(v_i \in \mathcal{S}_t). \tag{12}$$

With the estimates in hand, the weights $\{w_{i,t}\}_{i=1}^K$ can be updated as follows

$$w_{i,t+1} = w_{i,t} \exp\left(-\eta \tilde{\ell}_t(v_i)\right), \quad \forall i \in [K]. \tag{13}$$

The procedure that the learner chooses among experts when the probabilities are unknown is presented in Algorithm 2, named Exp3-UP. The following theorem establishes the regret bound of Exp3-UP.

Algorithm 2 Exp3-UP: Online learning with uninformative probabilistic feedback graphs

Input: learning rate $\eta > 0$, the minimum number of observations M, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
Initialize: $w_{i,1} = 1$, $\forall i \in [K]$, $\hat{p}_{ij,1} = 0$, $\forall (i,j) \in \mathcal{E}$.
for $t = 1, \ldots, T$ **do**
  if $t \leq KM$ **then**
    Set $k = t - \lfloor \frac{t}{K} \rfloor K$ and draw the expert node v_k.
  else
    Select one of the experts according to the PMF $\pi_t = (\pi_{1,t}, \ldots, \pi_{K,t})$, with $\pi_{i,t}$ in (10).
  end if
  Observe $\{(i, \ell_t(v_i)) : v_i \in \mathcal{S}_t\}$ and compute $\tilde{\ell}_t(v_i)$, $\forall i \in [K]$ as in (12).
  Update $\hat{p}_{ij,t+1}$, $\forall (i,j) \in \mathcal{E}_t$ via (9).
  Update $w_{i,t+1}$, $\forall i \in [K]$ via (13).
end for

Theorem 3. *Under (a1), the expected regret of Exp3-UP satisfies*

$$\begin{aligned} &\sum_{t=1}^T \mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i \in \mathcal{V}} \sum_{t=1}^T \ell_t(v_i) \\ \leq &\frac{\ln K}{\eta} + (K-1)M + \eta(1 - \frac{\eta}{2})(T - KM) \\ &+ \sum_{t=KM+1}^T \sum_{i=1}^K \frac{\pi_{i,t}}{q_{i,t}}(\frac{2\xi}{\sqrt{M}} + \frac{\eta}{2}) \end{aligned} \tag{14}$$

with probability at least

$$\delta_\xi := \prod_{t=KM+1}^{T} \prod_{(i,j)\in\mathcal{E}_t} \left(1 - 2\exp(-\frac{2\xi^2 C_{ij,t}}{M + 4\xi\sqrt{M}})\right).$$

See proof of Theorem 3 in Appendix C. The following Corollary states conditions under which the regret bound in (14) holds with high probability, i.e., $\delta_\xi = 1 - \mathcal{O}(\frac{1}{T})$ and the proof can be found in Appendix D.

Corollary 3.1. *If* $M \geq \left(\frac{4\xi \ln(KT)}{\xi^2 - \ln(KT)}\right)^2$ *and* $\xi > \sqrt{\ln(KT)}$*, under (a1) and (a2) the expected regret of Exp3-UP satisfies*

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i\in\mathcal{V}} \sum_{t=1}^{T} \ell_t(v_i)\\ \leq&\mathcal{O}\left(\frac{\alpha(\mathcal{G})}{\epsilon} \ln(KT)\sqrt{K\ln(KT)}T^{\frac{2}{3}}\right)\end{aligned} \tag{15}$$

with probability at least $1 - \mathcal{O}(\frac{1}{T})$.

Note that according to Algorithm 2 and Corollary 3.1, knowing the value of the time horizon T is required so that the learner can choose the values for M and ξ to achieve the sublinear regret bound in (15), which may not be feasible, and can be resolved by resorting to doubling trick. In this case, if $2^b < t \leq 2^{b+1}$ where $b \in \mathbb{N}$, the learner performs the Exp3-UP with parameters

$$\eta = \sqrt{\frac{\ln K}{2^{b+1}}} \tag{16a}$$

$$M = \left\lceil 2^{\frac{2(b+1)}{3}} \frac{1}{\sqrt{K}} + \ln 4K \right\rceil \tag{16b}$$

$$\xi = \left(2K^{\frac{1}{4}} + \sqrt{4\sqrt{K} + 1}\right)\sqrt{\ln(K2^{b+3})}. \tag{16c}$$

When the learner realizes that the value of M needs to be increased, it then performs exploration to guarantee that at least M samples of the mean ergodic random process X_{ij} are observed. The following lemma shows that when doubling trick is employed, Exp3-UP can achieve sub-linear regret without knowing the time horizon beforehand, the proof of which is in Appendix E.

Lemma 4. *Assuming that the doubling trick is employed to determine the value of* η*,* M *and* ξ *at each time instant and the greedy set cover algorithm is utilized to obtain a dominating set* $\mathcal{D}$ *of the nominal feedback graph. If* $T > K$*, the regret of Exp3-UP satisfies*

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i\in\mathcal{V}} \sum_{t=1}^{T} \ell_t(v_i)\\ \leq&\mathcal{O}\left(\frac{\alpha(\mathcal{G})}{\epsilon} \ln(T)\ln(KT)\sqrt{K\ln(KT)}T^{\frac{2}{3}} + \ln T\right)\end{aligned} \tag{17}$$

with probability at least $1 - \mathcal{O}(\frac{1}{K})$.

B. Geometric Resampling-based Approach

Another approach to obtain a reliable loss estimate is to employ geometric resampling. Similar to Exp3-UP, if $t \leq KM$ the learner chooses the k-th expert at time instant t where $k = t - \lfloor t/K \rfloor K$. In this way, it is guaranteed that at least M samples of the mean ergodic random process X_{ij} are observed. Based on these observations, a loss estimate is obtained whose expected value is an approximation of the loss $\ell_t(v_i)$, $\forall i \in [K]$. At $t > KM$, the learner draws one of the experts according to the following PMF

$$\pi_{i,t} = (1-\eta)\frac{w_{i,t}}{W_t} + \frac{\eta}{|\mathcal{D}|}\mathcal{I}(v_i \in \mathcal{D}), \quad \forall i \in [K] \tag{18}$$

where $\mathcal{D}$ represents a dominating set for $\mathcal{G}$. Furthermore, at each time instant $t > KM$, let $\tau_{i,1}^{(t)}, \ldots, \tau_{i,M}^{(t)}$ denote the last M time instants before t at which the i-th expert was chosen by the learner. If $(i,j) \in \mathcal{E}$, the learner observes $X_{ij}(\tau_{i,1}^{(t)}), \ldots, X_{ij}(\tau_{i,M}^{(t)})$ which are samples of the random process X_{ij} at $\tau_{i,1}^{(t)}, \ldots, \tau_{i,M}^{(t)}$. Let $Y_{ij,1}(t), \ldots, Y_{ij,M}(t)$ denote a random permutation of $X_{ij}(\tau_{i,1}^{(t)}), \ldots, X_{ij}(\tau_{i,M}^{(t)})$. At each time instant t, the learner draws with replacement M experts according to PMF $\{\pi_{i,t}\}$ in (18) in M independent trials. Let $P_{i,1}(t), \ldots, P_{i,M}(t)$ be a sequence of random variables associated with v_i at time instant t where $P_{i,u}(t) = 1$ if the learner draws the i-th expert at the u-th trial and $P_{i,u}(t) = 0$ otherwise. Let

$$Z_{i,u}(t) = \sum_{\forall j: v_j \in \mathcal{N}_{i,t}^{\text{in}}} P_{j,u}(t) Y_{ji,u}(t) \tag{19}$$

for all $1 \leq u \leq M$. An under-estimate of loss can then be obtained as

$$\tilde{\ell}_t(v_i) = Q_{i,t}\ell_t(v_i)\mathcal{I}(v_i \in \mathcal{S}_t). \tag{20}$$

where $Q_{i,t} := \min\{\{u \mid 1 \leq u \leq M, Z_{i,u}(t) = 1\} \cup \{M\}\}$, and the expected value of $\tilde{\ell}_t(v_i)$ can be written as

$$\mathbb{E}_t[\tilde{\ell}_t(v_i)] = \left(1 - (1 - q_{i,t})^M\right)\ell_t(v_i), \tag{21}$$

see (101) – (104) in Appendix F for detailed derivation. Then, the weights $\{w_{i,t}\}_{i=1}^K$ are updated as in (13) using the loss estimate $\tilde{\ell}_t(v_i)$ in (20). The geometric resampling based online expert learning framework (Exp3-GR) is summarized in Algorithm 3, and Theorem 5 presents its regret bound.

Theorem 5. *Under (a1) and (a2), the expected regret of Exp3-GR is bounded by*

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i\in\mathcal{V}} \sum_{t=1}^{T} \ell_t(v_i)\\ \leq&\frac{\ln K}{\eta} + (K-1)M + \sum_{t=KM+1}^{T} (1-q_{i,t})^M\\ &+ \eta(1-\eta)(T - KM) + \eta \sum_{t=KM+1}^{T} \sum_{i=1}^{K} \frac{\pi_{i,t}}{q_{i,t}}.\end{aligned} \tag{22}$$

The proof of Theorem 5 is presented in Appendix F. Building upon Theorem 5, the following Corollary presents the conditions under which Exp3-GR can obtain sub-linear regret.

Corollary 5.1. *Assume that greedy set cover algorithm is employed to find a dominating set of the nominal feedback*

Algorithm 3 Exp3-GR: Exp3 with geometric resampling

Input: learning rate $\eta > 0$, the minimum number of observations M, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
Initialize: $w_{i,1} = 1$, $\forall i \in [K]$.
for $t = 1, \dots, T$ **do**
  if $t \leq KM$ **then**
    Set $k = t - \lfloor \frac{t}{K} \rfloor K$ and draw the expert node v_k.
  else
    Select one expert according to PMF π_t in (18).
    Observe $\{\ell_t(v_i) : v_i \in \mathcal{S}_t\}$ and compute $\tilde{\ell}_t(v_i)$, $\forall i \in [K]$ via (20).
    Update $w_{i,t+1}$, $\forall i \in [K]$ via (13).
  end if
end for

graph $\mathcal{G}$. If $M \geq \frac{|\mathcal{D}| \ln T}{2\eta\epsilon}$, under (a1) and (a2), Exp3-GR satisfies

$$\sum_{t=1}^{T} \mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i \in \mathcal{V}} \sum_{t=1}^{T} \ell_t(v_i) \leq \mathcal{O}\left(\frac{\alpha(\mathcal{G})}{\epsilon} \ln(KT) \sqrt{KT \ln K} \right). \quad (23)$$

Proof. According to (a2), if $(i,j) \in \mathcal{E}$, the learner observes the loss of the j-th expert when it chooses the i-th expert with probability at least ϵ. Recalling (18) it can be inferred that $\pi_{i,t} > \eta/|\mathcal{D}|$, $\forall i \in \mathcal{D}$. Combining (4) with the fact that for each $v_i \in \mathcal{V}$ there is at least one edge from $\mathcal{D}$ to v_i, $\forall i \in [K]$, $q_{i,t}$ can be bounded below as

$$q_{i,t} > \frac{\eta\epsilon}{|\mathcal{D}|}. \quad (24)$$

Combining the condition $M \geq \frac{|\mathcal{D}| \ln T}{2\eta\epsilon}$ with (24), we have $M q_{i,t} \geq \frac{1}{2} \ln T$ which leads to $e^{-M q_{i,t}} \leq \frac{1}{\sqrt{T}}$. Thus, using the fact $1 + x \leq e^x$, we have

$$(1 - q_{i,t})^M \leq e^{-M q_{i,t}} \leq \frac{1}{\sqrt{T}}. \quad (25)$$

Hence, the third term in (22), i.e., $\sum_{t=t'}^{T} (1 - q_{i,t})^M$ can be bounded by $\mathcal{O}(\sqrt{T})$.

Furthermore, consider the case where we have $\eta = \mathcal{O}(\sqrt{\frac{K \ln K}{T}})$. Therefore, taking into account that greedy set cover algorithm is used to determine the dominating set, it can be inferred that $|\mathcal{D}| = \mathcal{O}(\alpha(\mathcal{G}) \ln K)$ (see e.g. [15]) based on which it can be obtained that $M = \mathcal{O}(\frac{\alpha(\mathcal{G})}{\epsilon\sqrt{K}} \ln T \sqrt{T \ln K})$, satisfies the condition $M \geq \frac{|\mathcal{D}| \ln T}{2\eta\epsilon}$. Hence, the expected regret of Exp3-GR satisfies (23), and the Corollary 5.1 is proved. □

Achieving the sub-linear regret in (23) requires that the learner knows the time horizon T, beforehand which may not be possible in some cases. When the learner does not know T, doubling trick can be utilized to achieve sub-linear regret. The following Lemma is proved in Appendix G, shows the regret bound for Exp3-GR when doubling trick is employed to find values of η and M without knowing the time horizon T. In this case, at time instant t, when $2^b < t \leq 2^{b+1}$, parameters η and M can be chosen as $\eta = \sqrt{\frac{K \ln K}{2^{b+1}}}$, $M = \left\lceil \frac{(b+1)\sqrt{2^{b-1}} |\mathcal{D}| \ln 2}{\epsilon \sqrt{K \ln K}} \right\rceil$.

When the learner realizes that M needs to be increased, it performs exploration to guarantee that at least M samples of the mean ergodic random process X_{ij} are observed.

Lemma 6. *Employing doubling trick to select η and M at each time instant, and supposing that a dominating set for the nominal feedback graph $\mathcal{G}$ is obtained using greedy set cover algorithm, the expected regret of Exp3-GR satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i \in \mathcal{V}} \sum_{t=1}^{T} \ell_t(v_i) \leq \mathcal{O}\left(\frac{\alpha(\mathcal{G}) \ln T}{\epsilon} \ln(KT) \sqrt{KT \ln K} \right). \quad (26)$$

Comparing Lemma 4 with Lemma 6, it can be observed that Exp3-GR achieves a tighter regret bound *with probability 1*. However, note that choosing an appropriate M for Exp3-GR requires knowing ϵ or a lower bound of ϵ, which may not be feasible in general, while such information is not required for Exp3-UP in order to guarantee the regret bound in (17).

Comparison with [25]. Note that while Exp3-GR and Exp3-Res proposed in [25] both employ the geometric resampling technique, there exist two major differences: i) Exp3-Res assumes the actual feedback graph is generated from Erdös-Rényi model, and the probabilities of the presence of edges are equal across all edges, while Exp3-GR considers the unequally probable case and does not assume that the probabilities of existence of all edges are equal; and ii) unlike Exp3-Res, Exp3-GR does not assume that the learner is guaranteed to observe the loss associated with the chosen expert. Furthermore, it is useful to compare the regret bound of Exp3-GR with that of Exp3-Res when the actual feedback graph is generated from the Erdös-Rényi model with $p_{ij} = p$, $\forall i,j \in [K]$. In this case, according to Corollary 5.1, Exp3-GR achieves regret of $\mathcal{O}\left(\frac{\ln(KT)}{p}\sqrt{KT \ln K}\right)$. On the other hand, under the assumption that $p \geq \frac{\ln T}{2K-2}$ and knowing that probabilities associated with all edges are equal, Exp3-Res obtains regret of $\mathcal{O}\left(\sqrt{K^2 \ln K + \frac{T \ln K}{p}}\right)$. Hence, having access to knowledge that the probabilities associated with all edges are equal enables Exp3-Res to achieve tighter regret bound than Exp3-GR in this special case.

Dependence of loss and feedback graph. The proposed algorithms Exp3-IP, Exp3-UP and Exp3-GR can also deal with cases where there is dependence between actual feedback graphs and losses. Consider the case that the environment generates $(\mathbf{x}_t, y_t)$ stochastically following certain time-invariant distribution. The i-th expert obtains the input $\mathbf{x}_t$ and outputs the prediction $\hat{y}_{i,t}$. In this case, the loss $\ell_t(v_i)$ can measure the discrepancy between $\hat{y}_{i,t}$ and y_t using some metrics such as squared loss. Furthermore, assume that the actual feedback graph at time t denoted by $\mathcal{H}_t$ depends on $\mathbf{x}_t$. In this case, if the learner knows the possible relations among experts, the learner can construct the nominal feedback graph $\mathcal{G}$ where the existence of each edge depends on $\mathbf{x}_t$. Therefore, the edge between two vertices exist with some time-invariant probability. Before decision making the learner is uncertain about $\mathbf{x}_t$ and as a result the learner can utilize one of the proposed algorithms Exp3-

Table I
CUMULATIVE REGRET ON VARIOUS DATASETS AND FULLY CONNECTED NOMINAL FEEDBACK GRAPH IN EQUALLY PROBABLE SETTING.

	Air	CCPP	Twitter	Tom's
Exp3	47.56	152.34	71.03	45.39
Exp3.G	39.16	122.93	59.72	40.63
Exp3-Res	33.36	100.22	52.64	38.46
Exp3-SET	38.21	122.62	59.28	39.49
Exp3-DOM	39.04	122.16	61.30	41.10
Exp3-IP	**33.33**	**98.22**	**52.47**	**37.63**
Exp3-UP	35.97	109.87	56.86	38.68
Exp3-GR	33.60	**99.91**	53.33	**38.25**

Table II
CUMULATIVE REGRET ON VARIOUS DATASETS AND PARTIALLY-CONNECTED NOMINAL FEEDBACK GRAPH $\mathcal{P}$ IN UNEQUALLY PROBABLE SETTING.

	Air	CCPP	Twitter	Tom's
Exp3	47.56	152.97	71.04	45.39
Exp3.G	41.19	128.74	62.21	41.53
Exp3-Res	35.83	109.62	56.55	39.51
Exp3-SET	40.21	129.32	62.22	40.77
Exp3-DOM	41.19	127.84	63.90	41.98
Exp3-IP	**32.95**	**97.37**	**52.23**	**37.19**
Exp3-UP	38.77	120.97	61.06	40.28
Exp3-GR	**33.60**	**99.59**	**53.04**	**38.01**

IP, Exp3-UP and Exp3-GR to decide based on the nominal feedback graph.

V. EXPERIMENTS

Performance of the proposed algorithms Exp3-IP, Exp3-UP and Exp3-GR are compared with online learning algorithms Exp3 [9], Exp3.G [18], Exp3-Res [25], Exp3-SET [15] and Exp3-DOM [15]. Exp3 considers bandit setting, and Exp3-Res assumes Erdös-Rényi model for the feedback graph. Furthermore, Exp3.G and Exp3-DOM treats the nominal feedback graph $\mathcal{G}_t$ as the actual one without considering uncertainties. Exp3-SET observes the nominal feedback graph $\mathcal{G}_t$ and the loss of out-neighbors of the chosen expert after decision making. Exp3-SET treats connectivity information given by $\mathcal{G}_t$ associated with nodes other than the chosen one as certain information without considering the uncertainty. Note that Exp3-SET observers the actual feedback graph partially after decision making since Exp3-SET observes the loss of chosen experts' out-neighbors. Performance is tested for regression task on several real datasets downloaded from the UCI Machine Learning Repository [32]:

Air Quality: This dataset contains $9,358$ responses from sensors in a polluted area, each with 13 features. The goal is to predict polluting chemical concentration in the air [33].

CCPP: The dataset has $9,568$ samples, with 4 features such as temperature, collected from a combined cycle power plant. The goal is predicting hourly electrical energy output [34].

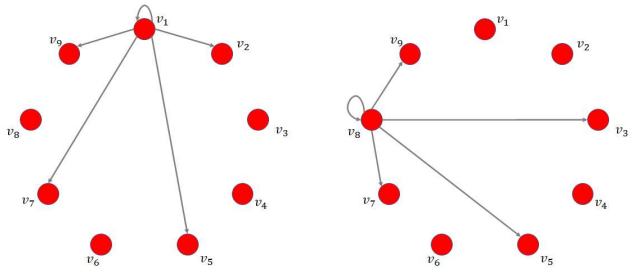
Twitter: This dataset contains $14,000$ samples with 77 features including e.g., the length of discussion on a given topic and the number of new interactive authors. The goal is to predict average number of active discussion on a certain topic [35].

Tom's Hardware: The dataset contains $10,000$ samples from a technology forum with 96 features. The goal is to predict the average number of display about a certain topic on Tom's hardware [35].

Let $(\mathbf{x}_i, y_i)$ and $(\bar{\mathbf{x}}_i, \bar{y}_i)$ be the i-th data sample and the normalized one, respectively. The data is normalized as $\bar{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\max_j \|\mathbf{x}_j\|}$, $\bar{y}_i = \frac{y_i - \min_j y_j}{\max_j y_j - \min_j y_j}$. Therefore, $\|\bar{\mathbf{x}}_i\| \leq 1$, $0 \leq \bar{y}_i \leq 1$, $\forall i$. In the experiments, there are 9 experts such that each expert is a trained model. In particular, each expert is trained on 10% of each dataset before the start online learning task associated with the corresponding dataset. Among them, 8 experts are trained via kernel ridge regression such that 5 experts exploit RBF kernels with bandwidth of $10^{-2}, 10^{-1}, 1, 10, 100$ while 3 experts employ Laplacian kernels with bandwidth of $10^{-2}, 1, 100$. Moreover, one expert is a trained linear regression model. Performance of algorithms are evaluated based on cumulative regret averaged over 20 independent runs. Recall that cumulative regret of an algorithm is the cumulative difference between the loss of the algorithm and that of the best expert in hindsight over time. In experiments, squared loss function is employed to measure the loss of experts. The learning rate η is set to $\frac{0.5}{\sqrt{T}}$ for all algorithms. Note that online learning algorithms may achieve better regret experimentally with carefully tuned learning rate. However, for fair comparison, the learning rates of all online learning algorithms are set to be the same. Parameter M is set as 25 for both Exp3-UP and Exp3-GR and $\xi = 1$ for Exp3-UP.

We first tested the equally probable setting where the nominal graph $\mathcal{G}_t$ is fully connected and probabilities $p_{ij} = 0.5$, $\forall i, j$. Table I lists the regret performance for various datasets. It can be observed that, knowing the exact probability enables Exp3-IP to achieve the lowest regret. Moreover, the proposed Exp3-UP and Exp3-GR obtain lower regret than Exp3.G, Exp3-SET and Exp3-DOM which treat the nominal feedback graph as actual one. Note that in this case, the actual feedback graph is indeed generated from the Erdös-Rényi model. The regret of the proposed Exp3-GR is comparable to that of Exp3-Res while Exp3-Res makes decision under the assumption that the actual feedback graph is generated from the Erdös-Rényi model.

We further tested the unequally probable case, when the graph is partially connected. In particular, $v_j \in \mathcal{N}_{i,t}^{\text{out}}$ if j is either the remainder of $i-1$, i, $i+1$, $i+4$ and $i+6$ to 9. Note that if the remainder is zero, it is considered to be 9. The resulting nominal feedback graph in this case is represented by $\mathcal{P}$. Therefore, in the nominal feedback graph $\mathcal{P}$, each node has 5 out-neighbors. As an example, out-neighbors of v_1 and v_8 are illustrated in Figure 1. The probability associated with each edge is drawn from uniform distribution $\mathcal{U}[0.25, 0.5]$. Table II lists the cumulative regret of all algorithms for Air Quality, CCPP, Twitter and Tom's Hardware datasets. It can be observed that Exp3-IP obtains the lowest regret. This shows that knowing the probabilities can indeed help obtain better performance. Furthermore, it can be observed that Exp3-UP and Exp3-GR can achieve lower regret in comparison with Exp3 which shows the effectiveness of using the information given by the uncertain graph. In addition, lower regret of Exp3-UP and Exp3-GR



(a) Out-neighbors of v_1 in $\mathcal{P}$. (b) Out-neighbors of v_8 in $\mathcal{P}$.

Figure 1. Out-neighbors of v_1 and v_8 are illustrated in partially-connected nominal feedback graph $\mathcal{P}$.

Table III
CUMULATIVE REGRET ON VARIOUS DATASETS AND PARTIALLY-CONNECTED CERTAIN NOMINAL FEEDBACK GRAPH $\mathcal{P}$.

	Air	CCPP	Twitter	Tom's
Exp3	46.46	150.23	70.01	45.05
Exp3.G	32.86	96.80	50.52	36.87
Exp3-Res	32.10	98.47	50.68	37.22
Exp3-SET	31.93	97.31	50.24	36.29
Exp3-DOM	32.84	96.39	52.07	37.22
Exp3-IP	33.19	97.37	52.44	36.88
Exp3-UP	35.92	109.93	56.25	38.44
Exp3-GR	33.34	99.08	52.87	37.60

compared to Exp3.G, Expe-SET and Exp3-DOM indicates that considering the uncertain graph $\mathcal{G}_t = \mathcal{P}$ as a certain graph can increase regret. Moreover, it can be observed Exp3-GR outperforms Exp3-Res when the actual feedback graph is not generated by Erdös-Rényi model. It can be observed Exp3-IP achieves lower regret than Exp3-GR and Exp3-UP, since the learner has access to the probabilities, while Exp3-UP and Exp3-GR do not rely on such prior information.

In addition, we tested the performance of algorithms when the nominal feedback graph $\mathcal{P}$ is partially-connected, and the probability associated with each edge is 1. As it can be seen from Table III, Exp3.G, Exp3-SET, Exp3-DOM and the proposed Exp3-IP which utilize the certain feedback graph obtain lower regret than those of Exp3-UP and Exp3-GR which treat the certain feedback graph as uncertain one. In fact, Exp3-UP and Exp3-GR do not know the probability associated with edges. Furthermore, the regret of Exp3-IP is comparable to Exp3.G, Exp3-SET and Exp3-DOM.

VI. CONCLUSION

The present paper studied the problem of online learning with *uncertain* feedback graphs, where potential uncertainties in the feedback graphs were modeled using probabilistic models. Novel algorithms were developed to exploit information revealed by the nominal feedback graph and different scenarios were discussed. Specifically, in the informative case, where the probabilities associated with edges are also revealed, Exp3-IP was developed. It is proved that Exp3-IP can achieve sublinear regret bound. Furthermore, Exp3-UP and Exp3-GR were developed for the uninformative case. It is proved that Exp3-GR can achieve tighter sublinear regret bound than that of Exp3-UP when the number of experts is negligible compared to time horizon, while Exp3-UP requires less prior information than Exp3-GR. Experiments on a number of real datasets were carried out to demonstrate that our novel algorithms can effectively address uncertainties in the feedback graph, and help enhance the learning ability of the learner.

REFERENCES

[1] P. M. Ghari and Y. Shen, "Online learning with probabilistic feedback," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4183–4187, May 2022.

[2] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. USA: Cambridge University Press, 2006.

[3] Y. Shen, T. Chen, and G. B. Giannakis, "Random feature-based online multi-kernel learning in environments with unknown dynamics," *Journal of Machine Learning Research*, vol. 20, pp. 773–808, Jan 2019.

[4] P. M. Ghari and Y. Shen, "Online multi-kernel learning with graph-structured feedback," in *Proceedings of the International Conference on Machine Learning*, vol. 119, pp. 3474–3483, Jul. 2020.

[5] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Information and Computation*, vol. 108, no. 2, pp. 212 – 261, 1994.

[6] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth, "How to use expert advice," *Journal of the ACM*, vol. 44, p. 427–485, May 1997.

[7] E. Hazan and N. Megiddo, "Online learning with prior knowledge," in *Proceedings of Annual Conference on Learning Theory*, p. 499–513, Jun 2007.

[8] A. Resler and Y. Mansour, "Adversarial online learning with noise," in *Proceedings of International Conference on Machine Learning*, pp. 5429–5437, Jun 2019.

[9] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, p. 48–77, Jan 2003.

[10] K. Gokcesu and S. S. Kozat, "An online minimax optimal algorithm for adversarial multiarmed bandit problem," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 5565–5580, Mar. 2018.

[11] S. Yang and Y. Gao, "An optimal algorithm for the stochastic bandits while knowing the near-optimal mean reward," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 2285–2291, Jun. 2021.

[12] K. Amin, S. Kale, G. Tesauro, and D. Turaga, "Budgeted prediction with expert advice," in *AAAI Conference on Artificial Intelligence*, (Austin, Texas, USA), Feb 2015.

[13] P. M. Ghari and Y. Shen, "Graph-aided online learning with expert advice," in *Asilomar Conference on Signals, Systems, and Computers*, pp. 470–474, 2020.

[14] S. Mannor and O. Shamir, "From bandits to experts: On the value of side-observations," in *Proc. of International Conference on Neural Information Processing Systems*, pp. 684–692, 2011.

[15] N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour, and O. Shamir, "Nonstochastic multi-armed bandits with graph-structured feedback," *SIAM Journal on Computing*, vol. 46, no. 6, pp. 1785–1826, 2017.

[16] C. Cortes, G. Desalvo, C. Gentile, M. Mohri, and S. Yang, "Online learning with sleeping experts and feedback graphs," in *Proceedings of International Conference on Machine Learning*, pp. 1370–1378, Jun 2019.

[17] C. Cortes, G. DeSalvo, C. Gentile, M. Mohri, and N. Zhang, "Online learning with dependent stochastic feedback graphs," in *Proceedings of International Conference on Machine Learning*, Jul 2020.

[18] N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren, "Online learning with feedback graphs: Beyond bandits," in *Proceedings of Conference on Learning Theory*, vol. 40, (Paris, France), pp. 23–35, Jul 2015.

[19] F. Liu, S. Buccapatnam, and N. B. Shroff, "Information directed sampling for stochastic bandits with graph feedback," in *Proceedings of AAAI Conference on Artificial Intelligence*, Feb 2018.

[20] R. Arora, T. V. Marinov, and M. Mohri, "Bandits with feedback graphs and switching costs," in *Advances in Neural Information Processing Systems*, pp. 10397–10407, Dec 2019.

[21] T. Kocák, G. Neu, M. Valko, and R. Munos, "Efficient learning by implicit exploration in bandit problems with side observations," in *Proceedings of International Conference on Neural Information Processing Systems*, p. 613–621, Dec 2014.

[22] T. Kocák, G. Neu, and M. Valko, "Online learning with noisy side observations," in *Proceedings of International Conference on Artificial Intelligence and Statistics*, (Cadiz, Spain), pp. 1186–1194, May 2016.

[23] A. Rangi and M. Franceschetti, "Online learning with feedback graphs and switching costs," in *Proceedings of International Conference on Artificial Intelligence and Statistics*, pp. 2435–2444, Apr 2019.

[24] A. Cohen, T. Hazan, and T. Koren, "Online learning with feedback graphs without the graphs," in *Proceedings of International Conference on Machine Learning*, p. 811–819, Jun 2016.

[25] T. Kocák, G. Neu, and M. Valko, "Online learning with Erdös-Rényi side-observation graphs," in *Proceedings of Conference on Uncertainty in Artificial Intelligence*, p. 339–346, Jun 2016.

[26] S. Li, W. Chen, Z. Wen, and K.-S. Leung, "Stochastic online learning with probabilistic graph feedback," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 4675–4682, Apr. 2020.

[27] D. P. Helmbold, N. Littlestone, and P. M. Long, "Apple tasting," *Information and Computation*, vol. 161, p. 85–139, Sep 2000.

[28] N. Cesa-Bianchi, T. Cesari, and C. Monteleoni, "Cooperative online learning: Keeping your neighbors updated," in *Proceedings of the International Conference on Algorithmic Learning Theory*, vol. 117, pp. 234–250, Feb 2020.

[29] S. McQuade and C. Monteleoni, "Global climate model tracking using geospatial neighborhoods," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, pp. 335–341, Jul 2012.

[30] V. Chvatal, "A greedy heuristic for the set-covering problem," *Mathematics of Operations Research*, vol. 4, pp. 233–235, Aug 1979.

[31] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. McGraw-Hill, 4th ed., 2002.

[32] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[33] S. D. Vito, E. Massera, M. Piga, L. Martinotto, and G. D. Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, vol. 129, no. 2, pp. 750 – 757, 2008.

[34] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *International Journal of Electrical Power and Energy Systems*, vol. 60, pp. 126 – 140, 2014.

[35] F. Kawala, A. Douzal-Chouakria, E. Gaussier, and E. Dimert, "Prédictions d'activité dans les réseaux sociaux en ligne," in *4ième conférence sur les modèles et l'analyse des réseaux : Approches mathématiques et informatiques*, (France), p. 16, Oct. 2013.

[36] V. Yurinskiĭ, "Exponential inequalities for sums of random vectors," *Journal of Multivariate Analysis*, vol. 6, pp. 473 – 499, Dec 1976.

APPENDIX

A. Proof of Theorem 1

Recall that $W_t = \sum_{i=1}^K w_{i,t}$ (below (3)), we have

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^K \frac{w_{i,t+1}}{W_t} = \sum_{i=1}^K \frac{w_{i,t}}{W_t} \exp\Big(-\eta\hat{\ell}_t(v_i)\Big). \tag{27}$$

According to (3), we can write

$$\frac{w_{i,t}}{W_t} = \frac{\pi_{i,t} - \eta\bar{F}_{i,t}}{1-\eta} \tag{28}$$

where $\bar{F}_{i,t} = \frac{F_{i,t}}{\sum_{j\in\mathcal{D}_t} F_{j,t}}\mathcal{I}(v_i \in \mathcal{D}_t)$. Substituting (28) into (27) obtains

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^K \frac{\pi_{i,t} - \eta\bar{F}_{i,t}}{1-\eta} \exp\Big(-\eta\hat{\ell}_t(v_i)\Big). \tag{29}$$

Using the inequality $e^{-x} \le 1 - x + \frac{1}{2}x^2, \forall x \ge 0$, the following inequality holds

$$\frac{W_{t+1}}{W_t} \le \sum_{i=1}^K \frac{\pi_{i,t} - \eta\bar{F}_{i,t}}{1-\eta} \left(1 - \eta\hat{\ell}_t(v_i) + \frac{1}{2}(\eta\hat{\ell}_t(v_i))^2\right). \tag{30}$$

Taking logarithm of both sides of (30) and using the fact that $1 + x \le e^x$, we have

$$\ln\frac{W_{t+1}}{W_t} \le \sum_{i=1}^K \frac{\pi_{i,t} - \eta\bar{F}_{i,t}}{1-\eta} \left(-\eta\hat{\ell}_t(v_i) + \frac{1}{2}(\eta\hat{\ell}_t(v_i))^2\right). \tag{31}$$

Summing (31) over time obtains

$$\ln\frac{W_{T+1}}{W_1} \le \sum_{t=1}^T\sum_{i=1}^K \frac{\pi_{i,t} - \eta\bar{F}_{i,t}}{1-\eta} \left(-\eta\hat{\ell}_t(v_i) + \frac{1}{2}(\eta\hat{\ell}_t(v_i))^2\right). \tag{32}$$

Furthermore, the left hand side of (31) can be bounded from below as

$$\ln\frac{W_{T+1}}{W_1} \ge \ln\frac{w_{i,T+1}}{W_1} = -\eta\sum_{t=1}^T \hat{\ell}_t(v_i) - \ln K \tag{33}$$

where the equality holds due to the fact that $W_1 = \sum_{j=1}^K w_{j,1} = K$. Then, (32) and (33) lead to

$$\begin{aligned}
&\sum_{t=1}^T\sum_{i=1}^K \frac{\eta\pi_{i,t}}{(1-\eta)}\hat{\ell}_t(v_i) - \eta\sum_{t=1}^T \hat{\ell}_t(v_i) \\
\le& \ln K + \sum_{t=1}^T\sum_{i\in\mathcal{D}_t} \frac{\eta^2\bar{F}_{i,t}}{(1-\eta)}\hat{\ell}_t(v_i) \\
&+ \sum_{t=1}^T\sum_{i=1}^K \eta^2\frac{\pi_{i,t} - \eta\bar{F}_{i,t}}{2(1-\eta)}\hat{\ell}_t(v_i)^2.
\end{aligned} \tag{34}$$

Multiplying both sides of (34) by $\frac{(1-\eta)}{\eta}$

$$\begin{aligned}
&\sum_{t=1}^T\sum_{i=1}^K \pi_{i,t}\hat{\ell}_t(v_i) - \sum_{t=1}^T \hat{\ell}_t(v_i) \\
\le& \frac{\ln K}{\eta} + \sum_{t=1}^T\sum_{i\in\mathcal{D}_t} \eta\bar{F}_{i,t}\hat{\ell}_t(v_i) \\
&+ \sum_{t=1}^T\sum_{i=1}^K \frac{\eta}{2}(\pi_{i,t} - \eta\bar{F}_{i,t})\hat{\ell}_t(v_i)^2.
\end{aligned} \tag{35}$$

Furthermore, the expected values of $\hat{\ell}_t(v_i)$ and $\hat{\ell}_t(v_i)^2$ can be written as

$$\mathbb{E}_t[\hat{\ell}_t(v_i)] = \sum_{j=1}^K \pi_{j,t}p_{ji,t}\frac{\ell_t(v_i)}{q_{i,t}} = \ell_t(v_i) \tag{36a}$$

$$\mathbb{E}_t[\hat{\ell}_t(v_i)^2] = \sum_{j=1}^K \pi_{j,t}p_{ji,t}\frac{\ell_t(v_i)^2}{q_{i,t}^2} = \frac{\ell_t(v_i)^2}{q_{i,t}} \le \frac{1}{q_{i,t}} \tag{36b}$$

where the inequality in (36b) holds because of (a1) which implies $\ell_t(v_i) \le 1$. Taking the expectation of both sides of (35), we arrive at

$$\begin{aligned}
&\sum_{t=1}^T\sum_{i=1}^K \pi_{i,t}\ell_t(v_i) - \sum_{t=1}^T \ell_t(v_i) \\
\le& \frac{\ln K}{\eta} + \sum_{t=1}^T\sum_{i=1}^K \eta\bar{F}_{i,t}\ell_t(v_i)
\end{aligned}$$

$$+\sum_{t=1}^{T}\sum_{i=1}^{K}\frac{\eta}{2}(\pi_{i,t}-\eta\bar{F}_{i,t})\frac{1}{q_{i,t}}. \quad (37)$$

Moreover, using the fact that $q_{i,t}\leq 1$ we have

$$\frac{\eta^2}{2}\sum_{t=1}^{T}\sum_{i=1}^{K}\frac{\bar{F}_{i,t}}{q_{i,t}}\geq\frac{\eta^2}{2}\sum_{t=1}^{T}\sum_{i=1}^{K}\bar{F}_{i,t}=\frac{\eta^2}{2}\sum_{t=1}^{T}1=\frac{\eta^2 T}{2}. \quad (38)$$

Furthermore, since based on (a1) $\ell_t(v_i)\leq 1$, the second term on the RHS of (37) can be bounded by

$$\eta\sum_{t=1}^{T}\sum_{i=1}^{K}\bar{F}_{i,t}\ell_t(v_i)\leq\eta\sum_{t=1}^{T}\sum_{i=1}^{K}\bar{F}_{i,t}=\eta\sum_{t=1}^{T}1=\eta T. \quad (39)$$

Combining (38), (39) with (37) we have

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{i=1}^{K}\pi_{i,t}\ell_t(v_i)-\sum_{t=1}^{T}\ell_t(v_i)\\ \leq&\frac{\ln K}{\eta}+\eta T-\frac{\eta^2 T}{2}+\frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}}. \quad (40)\end{aligned}$$

By definition, the first term on the RHS of (40) equals to $\mathbb{E}_t[\ell_t(v_{I_t})]$. In addition, note that (40) holds for all $v_i\in\mathcal{V}$, hence the following inequality holds

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})]-\min_{v_i\in\mathcal{V}}\sum_{t=1}^{T}\ell_t(v_i)\\ \leq&\frac{\ln K}{\eta}+\eta(1-\frac{\eta}{2})T+\frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}} \quad (41)\end{aligned}$$

which completes the proof of Theorem 1.

B. Proof of Lemma 2

Based on Theorem 1, the upper bound of the expected regret of Exp3-IP is

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})]-\min_{v_i\in\mathcal{V}}\sum_{t=1}^{T}\ell_t(v_i)\\ \leq&\frac{\ln K}{\eta}+\eta(1-\frac{\eta}{2})T+\frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}}. \quad (42)\end{aligned}$$

Let at each time instant t, Q_t is defined as

$$Q_t=1+\frac{1}{2}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}}. \quad (43)$$

Furthermore, let τ_r be the largest time instant satisfying $\sum_{t=1}^{\tau_r}Q_t\leq 2^r$. According to the doubling trick, at $\tau_{r-1}+1$, such that $\sum_{t=1}^{\tau_{r-1}+1}Q_t>2^{r-1}$, the algorithm restarts with

$$\eta_r=\sqrt{\frac{\ln K}{2^r}}. \quad (44)$$

Also, the algorithm starts with $r=0$. Therefore, based on (42) and (44), it can be concluded that

$$\sum_{t=1}^{\tau_r}\pi_{i,t}\ell_t(v_i)-\min_{v_i\in\mathcal{V}}\sum_{t=1}^{\tau_r}\ell_t(v_i)\leq 2\sqrt{2^r\ln K}-\frac{\ln K}{2^{r+1}}\tau_r \quad (45)$$

when $2^{r-1}<\sum_{t=1}^{\tau_r}Q_t\leq 2^r$. The maximum number of restarts required is $\left\lceil\log_2\sum_{t=1}^{T}Q_t\right\rceil$. Moreover, it can be written that

$$\sum_{r=0}^{\lceil\log_2\sum_{t=1}^{T}Q_t\rceil}2\sqrt{2^r\ln K}<\frac{4\sqrt{\ln K}}{\sqrt{2}-1}\sqrt{\sum_{t=1}^{T}Q_t}. \quad (46)$$

Therefore, based on (42) and considering the fact that the maximum possible value for incurred loss at each restart is 1, combining (45) with (46) leads to

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})]-\min_{v_i\in\mathcal{V}}\sum_{t=1}^{T}\ell_t(v_i)\\ \leq&\mathcal{O}\left(\sqrt{(\ln K)\sum_{t=1}^{T}Q_t}+\left\lceil\log_2\sum_{t=1}^{T}Q_t\right\rceil\right)\\ =&\mathcal{O}\left(\sqrt{\ln K\sum_{t=1}^{T}(1+\frac{1}{2}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}})}+\left\lceil\log_2\sum_{t=1}^{T}Q_t\right\rceil\right) \quad (47)\end{aligned}$$

Based on (a2), we can write $p_{ij}\geq\epsilon>0$ if $(i,j)\in\mathcal{E}_t$. According to (4) and the fact that the i-th expert is chosen by the learner with probability of $\pi_{i,t}$, based on (a2) the inequality $q_{i,t}\geq\pi_{i,t}\epsilon$ holds. Thus, we have

$$\left\lceil\log_2\sum_{t=1}^{T}Q_t\right\rceil=\mathcal{O}\left(\ln\left(\frac{K}{\epsilon}T\right)\right). \quad (48)$$

Combining (47) with (48) obtains

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})]-\min_{v_i\in\mathcal{V}}\sum_{t=1}^{T}\ell_t(v_i)\\ \leq&\mathcal{O}\left(\sqrt{\ln K\sum_{t=1}^{T}(1+\frac{1}{2}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}})}+\ln(\frac{K}{\epsilon}T)\right) \quad (49)\end{aligned}$$

In addition, the following Lemma is used as a step stone [15].

Lemma 7. *Let $\mathcal{G}=(\mathcal{V},\mathcal{E})$ be a directed graph with a set of vertices $\mathcal{V}$ and a set of edges $\mathcal{E}$ such that each vertex in $\mathcal{V}$ has a self-loop. Let $\mathcal{D}\subseteq\mathcal{V}$ be a dominating set for $\mathcal{G}$ and $p_1,\ldots,p_K$ be a probability distribution defined over $\mathcal{V}$, such that $p_i\geq\beta>0$, for $i\in\mathcal{D}$. Then*

$$\sum_{i=1}^{K}\frac{p_i}{\sum_{j:j\to i}p_j}\leq 2\alpha(\mathcal{G})\ln(1+\frac{\left\lceil\frac{K^2}{\beta|\mathcal{D}|}\right\rceil+K}{\alpha(\mathcal{G})})+2|\mathcal{D}| \quad (50)$$

where $\alpha(\mathcal{G})$ represents independence number for the graph $\mathcal{G}$.

Based on Lemma 7 and (a2), we get

$$\begin{aligned}&\sum_{i=1}^{K}\frac{\pi_{i,t}}{\sum_{\forall j:j\in\mathcal{N}_{i,t}^{\text{in}}}\pi_{j,t}}\\ <&2\alpha(\mathcal{G}_t)\ln(1+\frac{\left\lceil\frac{K^3}{\eta\epsilon}\right\rceil+K}{\alpha(\mathcal{G}_t)})+2|\mathcal{D}_t|. \quad (51)\end{aligned}$$

Considering the fact that $q_{i,t}\geq\epsilon\sum_{\forall j:j\in\mathcal{N}_{i,t}^{\text{in}}}\pi_{j,t}$ which is induced by (a2), from (51), it can be inferred that

$$\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}}<\frac{2\alpha(\mathcal{G}_t)}{\epsilon}\ln(1+\frac{\left\lceil\frac{K^3}{\eta\epsilon}\right\rceil+K}{\alpha(\mathcal{G}_t)})+\frac{2|\mathcal{D}_t|}{\epsilon}. \quad (52)$$

Furthermore, if greedy set cover algorithm by [30] is employed to obtain the dominating set $|\mathcal{D}_t|$, it can be written that [15]

$$|\mathcal{D}_t| = \mathcal{O}(\alpha(\mathcal{G}_t)\ln K). \tag{53}$$

Therefore, from (52) we can conclude that

$$\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}} \leq \mathcal{O}\left(\frac{\alpha(\mathcal{G}_t)}{\epsilon}\ln(\frac{KT}{\epsilon})\right) \tag{54}$$

Combining (49) with (53) and (54), we arrive at

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i\in\mathcal{V}}\sum_{t=1}^{T}\ell_t(v_i)\\ \leq&\mathcal{O}\left(\sqrt{\ln K\ln(\frac{K}{\epsilon}T)\sum_{t=1}^{T}\frac{\alpha(\mathcal{G}_t)}{\epsilon}} + \ln(\frac{K}{\epsilon}T)\right)\end{aligned} \tag{55}$$

which completes the proof of Lemma 2.

C. Proof of Theorem 3

In order to prove Theorem 3, let's first consider when $t \leq KM$. When the learner chooses among experts in a deterministic fashion. The (expected) loss can be written as $\mathbb{E}_t[\ell_t(v_i)] = \ell_t(v_k)$. Since $\ell_t(v_i) \leq 1$, we have

$$\sum_{t=1}^{KM}\mathbb{E}_t[\ell_t(v_i)] - \sum_{t=1}^{KM}\ell_t(v_i) \leq (K-1)M. \tag{56}$$

On the other hand, for any $t > KM$, the following equality holds

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^{K}\frac{w_{i,t+1}}{W_t} = \sum_{i=1}^{K}\frac{w_{i,t}}{W_t}\exp\left(-\eta\tilde{\ell}_t(v_i)\right). \tag{57}$$

Recall (10), we have

$$\frac{w_{i,t}}{W_t} = \frac{\pi_{i,t} - \eta\hat{\bar{F}}_{i,t}}{1-\eta} \tag{58}$$

where $\hat{\bar{F}}_{i,t} = \frac{\eta}{|\mathcal{D}|}\mathcal{I}(v_i \in \mathcal{D}_t)$. Following similar steps from (29) to (34), and from (57) and (58) we obtain

$$\begin{aligned}&\sum_{t=t'}^{T}\sum_{i=1}^{K}\pi_{i,t}\tilde{\ell}_t(v_i) - \sum_{t=t'}^{T}\tilde{\ell}_t(v_i)\\ \leq&\frac{\ln K}{\eta} + \sum_{t=t'}^{T}\sum_{i=1}^{K}\eta\hat{\bar{F}}_{i,t}\tilde{\ell}_t(v_i)\\ &+\sum_{t=t'}^{T}\sum_{i=1}^{K}\frac{\eta}{2}(\pi_{i,t}-\eta\hat{\bar{F}}_{i,t})\tilde{\ell}_t(v_i)^2\end{aligned} \tag{59}$$

where $t' = KM+1$. In addition, the expected value of $\tilde{\ell}_t(v_i)$ and $\tilde{\ell}_t(v_i)^2$ at time instant t can be written as

$$\mathbb{E}_t[\tilde{\ell}_t(v_i)] = \sum_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\pi_{j,t}p_{ji}\frac{1}{\hat{q}_{i,t}}\ell_t(v_i) = \frac{q_{i,t}}{\hat{q}_{i,t}}\ell_t(v_i) \tag{60a}$$

$$\begin{aligned}\mathbb{E}_t[\tilde{\ell}_t(v_i)^2] &= \sum_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\pi_{j,t}p_{ji}\frac{1}{\hat{q}_{i,t}^2}\ell_t(v_i)^2\\ &= \frac{q_{i,t}}{\hat{q}_{i,t}^2}\ell_t(v_i)^2 \leq \frac{q_{i,t}}{\hat{q}_{i,t}^2}.\end{aligned} \tag{60b}$$

Let $e_{ij,t} := |\hat{p}_{ij,t} - p_{ij}|$. According to (11), the probability that $\hat{q}_{i,t} \geq q_{i,t}$ is at least $\prod_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\Pr(e_{ij,t} \leq \xi/\sqrt{M})$ since the incidents $\{e_{ij,t} \leq \xi/\sqrt{M},\ \forall(i,j)\in\mathcal{E}\}$ are independent from each other. Let ε denote $\xi/\sqrt{M}$ and $\mu_{i,t} := \frac{1}{\hat{q}_{i,t}} - \frac{1}{q_{i,t}}$, we have

$$\begin{aligned}\mu_{i,t} &= \frac{q_{i,t}-\hat{q}_{i,t}}{\hat{q}_{i,t}q_{i,t}} = \frac{\sum_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\pi_{j,t}(p_{ji}-\hat{p}_{ji,t}-\varepsilon)}{\hat{q}_{i,t}q_{i,t}}\\ &\geq -\frac{\sum_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}2\pi_{j,t}\varepsilon}{q_{i,t}^2}\end{aligned} \tag{61}$$

where the last inequality holds with probability $\prod_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\Pr(e_{ij,t}\leq\varepsilon)$. Therefore, the following inequalities hold with the probability $\prod_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\Pr(e_{ij,t}\leq\varepsilon)$

$$\begin{aligned}&\ell_t(v_i) - \sum_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\frac{2\pi_{j,t}\varepsilon}{q_{i,t}}\ell_t(v_i) \leq \mathbb{E}_t[\tilde{\ell}_t(v_i)]\\ =&\ell_t(v_i) + q_{i,t}\mu_{i,t}\ell_t(v_i) \leq \ell_t(v_i)\end{aligned} \tag{62a}$$

$$\mathbb{E}_t[\tilde{\ell}_t(v_i)^2] \leq \frac{1}{q_{i,t}}. \tag{62b}$$

Taking expectation of both sides of (59) and combining with (62), we obtain the following inequality

$$\begin{aligned}&\sum_{t=t'}^{T}\sum_{i=1}^{K}\pi_{i,t}\ell_t(v_i) - \sum_{t=t'}^{T}\ell_t(v_i)\\ &-\sum_{t=t'}^{T}\sum_{i=1}^{K}\pi_{i,t}\sum_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\frac{2\pi_{j,t}\varepsilon}{q_{i,t}}\ell_t(v_i)\\ \leq&\frac{\ln K}{\eta} + \sum_{t=t'}^{T}\sum_{i=1}^{K}\eta\hat{\bar{F}}_{i,t}\ell_t(v_i) + \sum_{t=t'}^{T}\sum_{i=1}^{K}\frac{\eta}{2}(\pi_{i,t}-\eta\hat{\bar{F}}_{i,t})\frac{1}{q_{i,t}}\\ \leq&\frac{\ln K}{\eta} + \sum_{t=t'}^{T}\sum_{i=1}^{K}\eta\hat{\bar{F}}_{i,t} + \sum_{t=t'}^{T}\sum_{i=1}^{K}\frac{\eta}{2}(\pi_{i,t}-\eta\hat{\bar{F}}_{i,t})\frac{1}{q_{i,t}}\end{aligned} \tag{63}$$

which holds with probability at least $\prod_{(i,j)\in\mathcal{E}_t}\Pr(e_{ij,t'}\leq\varepsilon,\ldots,e_{ij,T}\leq\varepsilon)$. Applying the chain rule for one term in the product, we have

$$\begin{aligned}&\Pr(e_{ij,t'}\leq\varepsilon,\ldots,e_{ij,T}\leq\varepsilon)\\ =&\Pr(e_{ij,t'}\leq\varepsilon)\prod_{t=t'+1}^{T}\Pr(e_{ij,t}\leq\varepsilon \mid e_{ij,t-1}\leq\varepsilon,\ldots,e_{ij,t'}\leq\varepsilon)\\ \geq&\prod_{t=t'}^{T}\Pr(e_{ij,t}\leq\varepsilon).\end{aligned} \tag{64}$$

In order to obtain the lower bound of the probability $\Pr(e_{ij,t}\leq\varepsilon_{ij,t})$, the Bernstein inequality is employed. To this end consider the following lemma [36].

Lemma 8. *Let $\zeta_1\ldots\zeta_n$ be independent random variables with*

$$\mathbb{E}[\zeta_i] = 0, \forall i: 1\leq i\leq n \tag{65a}$$

$$|\mathbb{E}[\zeta_i^m]| \leq \frac{m!}{2}b_i^2H^{m-2}, m=2,3,\ldots,\forall i:1\leq i\leq n. \tag{65b}$$

Then for $x \geq 0$, we have

$$\Pr(|\zeta_1 + \ldots + \zeta_n| \geq xB_n) \leq 2\exp(-\frac{\frac{x^2}{2}}{1+\frac{xH}{B_n}}) \quad (66)$$

where $B_n^2 = b_1^2 + \ldots + b_n^2$.

Let $\theta_{ij}(t) := X_{ij}(t) - p_{ij}, \forall (i,j) \in \mathcal{E}_t$. Since $X_{ij}(t)$ follows Bernoulli distribution with the parameter p_{ij}, it can be readily obtained that $\mathbb{E}[\theta_{ij}(t)] = 0$. Furthermore, for the moment generating function of $\theta_{ij}(t)$, we have

$$M_{\theta_{ij}(t)}(z) = (1-p_{ij})e^{-p_{ij}z} + p_{ij}e^{(1-p_{ij})z}. \quad (67)$$

Therefore, the expected value of $\theta_{ij}^m(t)$, $m = 2, 3, \ldots$ satisfies

$$\begin{aligned}\mathbb{E}[\theta_{ij}^m(t)] &= \frac{d^m M_{\theta_{ij}(t)}(z)}{dz^m}\,|_{z=0} \\ &= (-p_{ij})^m(1-p_{ij}) + (1-p_{ij})^m p_{ij}. \end{aligned} \quad (68)$$

From (68), we can conclude that

$$\begin{aligned}|\mathbb{E}[\theta_{ij}^m(t)]| &\leq p_{ij}(1-p_{ij}) \leq \frac{1}{4} \\ &\leq \frac{m!}{8} = \frac{m!}{2}\left(\frac{1}{2}\right)^2 \times 1^{m-2}, m = 2, 3, \ldots \end{aligned} \quad (69)$$

Thus, letting $b_i = \frac{1}{2}$, $H = 1$ in Lemma 8 and combining with (69), the following inequality can be obtained

$$\begin{aligned}&\Pr\left(|\sum_{\tau \in \mathcal{T}_{ij,t}} \theta_{ij}(\tau)| \geq \frac{\xi C_{ij,t}}{\sqrt{M}}\right) \\ =&\Pr\left(|\sum_{\tau \in \mathcal{T}_{ij,t}} X_{ij}(\tau) - p_{ij}| \geq \frac{\xi C_{ij,t}}{\sqrt{M}}\right) \\ \leq& 2\exp\left(-\frac{2\xi^2 \frac{C_{ij,t}}{M}}{1+\frac{4\xi}{\sqrt{M}}}\right) = 2\exp\left(-\frac{2\xi^2 C_{ij,t}}{M+4\xi\sqrt{M}}\right) \end{aligned} \quad (70)$$

which leads to

$$\Pr(e_{ij,t} \geq \varepsilon) \leq 2\exp\left(-\frac{2\xi^2 C_{ij,t}}{M+4\xi\sqrt{M}}\right) \quad (71)$$

Therefore, (63) holds with probability at least

$$\delta_\xi = \prod_{t=t'}^{T} \prod_{(i,j)\in\mathcal{E}_t} \left(1 - 2\exp(-\frac{2\xi^2 C_{ij,t}}{M+4\xi\sqrt{M}})\right). \quad (72)$$

Since $\sum_{\forall j: v_j \in \mathcal{N}_{i,t}^{\text{in}}} \pi_{j,t} \leq 1$ and $\varepsilon = \frac{\xi}{\sqrt{M}}$, the following inequality holds

$$\begin{aligned}\sum_{t=t'}^{T}\sum_{i=1}^{K}\pi_{i,t} \sum_{\forall j: v_j \in \mathcal{N}_{i,t}^{\text{in}}} \frac{2\pi_{j,t}\varepsilon}{q_{i,t}} &= \sum_{t=t'}^{T}\sum_{i=1}^{K}\pi_{i,t} \sum_{\forall j: v_j \in \mathcal{N}_{i,t}^{\text{in}}} \frac{2\pi_{j,t}\frac{\xi}{\sqrt{M}}}{q_{i,t}} \\ &\leq \sum_{t=t'}^{T}\sum_{i=1}^{K}\frac{2\pi_{i,t}\xi}{q_{i,t}\sqrt{M}}. \end{aligned} \quad (73)$$

Using (73) and the fact that $\frac{1}{q_{i,t}} \geq 1$, (63) can be rewritten as

$$\begin{aligned}&\sum_{t=t'}^{T}\sum_{i=1}^{K}\pi_{i,t}\ell_t(v_i) - \sum_{t=t'}^{T}\ell_t(v_i) \\ \leq& \frac{\ln K}{\eta} + \sum_{t=t'}^{T}\eta(1-\frac{\eta}{2}) + \sum_{t=t'}^{T}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}}(\frac{2\xi}{\sqrt{M}} + \frac{\eta}{2}) \end{aligned} \quad (74)$$

Combining (74) with (56) results in following inequality

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})] - \sum_{t=1}^{T}\ell_t(v_i) \\ \leq& \frac{\ln K}{\eta} + (K-1)M + \eta(1-\frac{\eta}{2})(T-KM) \\ &+ \sum_{t=t'}^{T}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}}(\frac{2\xi}{\sqrt{M}} + \frac{\eta}{2}) \end{aligned} \quad (75)$$

which holds with probability at least δ_ξ and the proof of Theorem 3 is completed.

D. Proof of Corollary 3.1

The proof of Corollary 3.1 will be built upon the following Lemma.

Lemma 9. *Let $\zeta_1, \ldots, \zeta_N$ ($N > 1$) be a sequence of real positive numbers such that $\forall i : 1 \leq i \leq N$, $0 < \zeta_i < 1$ and $\forall n : 1 \leq n \leq N$, $\sum_{i=1}^{n}\zeta_i < 1$. Then, it can be written that*

$$\prod_{i=1}^{N}(1-\zeta_i) > 1 - \sum_{i=1}^{N}\zeta_i \quad (76)$$

Proof. We prove this Lemma using mathematical induction. Firstly, Consider (76) for $N = 2$

$$(1-\zeta_1)(1-\zeta_2) = 1 - \zeta_1 - \zeta_2 + \zeta_1\zeta_2 > 1 - \zeta_1 - \zeta_2. \quad (77)$$

Assuming that (76) holds for $N = n$. Then, based on (77) we have for $N = n+1$

$$\begin{aligned}&\prod_{i=1}^{n+1}(1-\zeta_i) \\ =&\left(\prod_{i=1}^{n}(1-\zeta_i)\right) \times (1-\zeta_{n+1}) > \quad (1-\sum_{i=1}^{n}\zeta_i)(1-\zeta_{n+1}) \\ >&1 - \sum_{i=1}^{n+1}\zeta_i. \end{aligned} \quad (78)$$

Hence, (76) also holds for $N = n+1$, and Lemma 9 is proved by induction. □

Assuming M satisfies

$$M \geq \left(\frac{4\xi\ln(KT)}{\xi^2 - \ln(KT)}\right)^2. \quad (79)$$

Hence, (79) can be re-written as

$$\frac{1}{K^2T^2} \geq \exp(-\frac{2\xi^2\sqrt{M}}{\sqrt{M}+4\xi}) \geq \exp(-\frac{2\xi^2 C_{ij,t}}{M+4\xi\sqrt{M}}) \quad (80)$$

where the second inequality holds since $C_{ij,t} \geq M$. Let $t' = KM + 1$. Note that the regret bound in (14) holds with probability at least δ_ξ in (72). According to Lemma 9, we can obtain the following inequality

$$\delta_\xi = \prod_{t=t'}^{T} \prod_{(i,j)\in\mathcal{E}_t} \left(1 - 2\exp(-\frac{2\xi^2 C_{ij,t}}{M+4\xi\sqrt{M}})\right)$$

$$>1-\sum_{(i,j)\in\mathcal{E}_t}\sum_{t=t'}^{T}2\exp(-\frac{2\xi^2C_{ij,t}}{M+4\xi\sqrt{M}}). \tag{81}$$

Combining (80) with (81) obtains

$$\delta_\xi \geq 1-\frac{2(T-KM)|\mathcal{E}|}{K^2T^2} \tag{82}$$

where $|\mathcal{E}|$ denotes the cardinality of the $\mathcal{E}$. Since $\mathcal{G}$ does not change over time, $|\mathcal{E}|$ is a constant. According to (82), it can be readily obtained that when (79) holds, the regret bound in (14) holds with probability at least of order $1-\mathcal{O}(\frac{1}{T})$. Consider the case where the learner sets η, M and ξ as follows

$$\eta=\mathcal{O}(\sqrt{\frac{\ln K}{T}}) \tag{83a}$$

$$M=\mathcal{O}(\frac{1}{\sqrt{K}}T^{\frac{2}{3}}) \tag{83b}$$

$$\xi=\mathcal{O}(K^{\frac{1}{4}}\sqrt{\ln(KT)}). \tag{83c}$$

Putting η, M and ξ in (83) into (14) and based on Lemma 7, it can be concluded that the expected regret of Exp3-UP satisfies

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})]-\sum_{t=1}^{T}\ell_t(v_i)\\ \leq&\mathcal{O}\left(\frac{\alpha(\mathcal{G})}{\epsilon}\ln(KT)(\sqrt{T\ln K}+\sqrt{K\ln(KT)}T^{\frac{2}{3}})\right)\end{aligned} \tag{84}$$

with probability at least $1-\mathcal{O}(\frac{1}{T})$.

E. Proof of Lemma 4

In this section, doubling trick technique is employed such that Exp3-UP can achieve sub-linear regret. If $2^b<t\leq 2^{b+1}$, the value of the learning rate η_b, M_b and ξ_b are

$$\eta_b=\sqrt{\frac{\ln K}{2^{b+1}}} \tag{85a}$$

$$M_b=\left\lceil 2^{\frac{2(b+1)}{3}}\frac{1}{\sqrt{K}}+\ln 4K\right\rceil \tag{85b}$$

$$\xi_b=\left(2K^{\frac{1}{4}}+\sqrt{4\sqrt{K}+1}\right)\sqrt{\ln(K2^{b+3})} \tag{85c}$$

When the learner realizes that $t>2^{b+1}$, the algorithm restarts with η_{b+1}, M_{b+1} and ξ_{b+1}. The algorithm starts with $b=\lceil\log_2 K\rceil$. Therefore, when $t<2^{\lceil\log_2 K\rceil}$, the value of η_b, M_b and ξ_b are set with respect to $b=\lceil\log_2 K\rceil$. Let $\mathcal{M}_i$ denotes a set which includes the time instants when the learner chooses the i-th expert in a deterministic fashion for exploration. Specifically, when at time instant τ, the learner chooses the i-th expert for exploration without using the PMF in (10), the time instant τ is appended to $\mathcal{M}_i$. At each restart the learner chooses the experts one by one for the exploration until the condition $|\mathcal{M}_i|\geq M_b$, $\forall i\in[K]$ is satisfied. Then, the learner chooses among experts according to PMF in (10) using the learning rate η_b. Therefore, based on Theorem 3, for each b, the algorithm satisfies

$$\begin{aligned}&\sum_{t=2^b+1}^{T_b}\mathbb{E}_t[\ell_t(v_{I_t})]-\sum_{t=2^b+1}^{T_b}\ell_t(v_i)\\ \leq&\frac{\ln K}{\eta_b}+(K-1)(M_b-M_{b-1})\\ &+\eta_b(1-\frac{\eta_b}{2})(T_b-2^b-K(M_b-M_{b-1}))\\ &+\sum_{t=2^b+1}^{T_b}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}}(\frac{2\xi_b}{\sqrt{M_b}}+\frac{\eta_b}{2})\end{aligned} \tag{86}$$

with probability at least δ_b where it can be expressed as

$$\delta_b=\prod_{t=t'_b}^{T_b}\prod_{(i,j)\in\mathcal{E}_t}\left(1-2\exp(-\frac{2\xi_b^2C_{ij,t}}{M_b+4\xi_b\sqrt{M_b}})\right) \tag{87}$$

where T_b denote the greatest time instant which satisfies $2^b<T_b\leq 2^{b+1}$ and t'_b can be written as

$$t'_b=\min(T_{b-1}+K(M_b-M_{b-1})+1,T_b). \tag{88}$$

Note that $M_{b-1}=0$ when $b=\lceil\log_2 K\rceil$. Since for each b, M_b and ξ_b in (85) meet the following condition

$$M_b\geq\left(\frac{4\xi_b\ln(4KT_b)}{\xi_b^2-\ln(4KT_b)}\right)^2, \tag{89}$$

it can be concluded that the following inequality holds true

$$\begin{aligned}\frac{1}{16K^2T_b^2}\geq&\exp(-\frac{2\xi_b^2\sqrt{M_b}}{\sqrt{M_b}+4\xi_b})\\ \geq&\exp(-\frac{2\xi_b^2C_{ij,t}}{M_b+4\xi_b\sqrt{M_b}}),\end{aligned} \tag{90}$$

and as a result according to Lemma 9 we can write

$$\delta_b>1-\sum_{(i,j)\in\mathcal{E}_t}\sum_{t=t'_b}^{T_b}2\exp(-\frac{2\xi_b^2C_{ij,t}}{M_b+4\xi_b\sqrt{M_b}}). \tag{91}$$

Combining (90) with (91), it can be concluded that

$$\delta_b>1-\max(0,\frac{(T_b-t'_b)|\mathcal{E}_{T_b}|}{8K^2T_b^2}). \tag{92}$$

Therefore, for each b from (86), (92) and Lemma 7 it can be inferred that

$$\sum_{t=2^b+1}^{T_b}\mathbb{E}_t[\ell_t(v_{I_t})]-\sum_{t=2^b+1}^{T_b}\ell_t(v_i)\leq\mathcal{O}(\omega) \tag{93}$$

where $\omega:=\frac{\alpha(\mathcal{G})}{\epsilon}\ln(KT_b)(\sqrt{T_b\ln K}+\sqrt{K\ln(KT_b)}T_b^{\frac{2}{3}})$ holds with probability at least $1-\mathcal{O}(\frac{1}{T_b})$. Summing (93) over all possible values of b, from $b:=\lceil\log_2 K\rceil$ to $\lceil\log_2 T\rceil$ and taking into account that the maximum value of the loss at each restart is 1, we arrive at

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})]-\sum_{t=1}^{T}\ell_t(v_i)\\ \leq&\sum_{b=\lceil\log_2 K\rceil}^{\lceil\log_2 T\rceil}\mathcal{O}(\omega)+\lceil\log_2 T\rceil-\lceil\log_2 K\rceil\\ \leq&\mathcal{O}(\omega\ln T)+\mathcal{O}(\ln T)\end{aligned} \tag{94}$$

which holds with probability at least

$$\Delta=\prod_{b=\lceil\log_2 K\rceil}^{\lceil\log_2 T\rceil}\left(1-\max(0,\frac{(T_b-t'_b)|\mathcal{E}_{T_b}|}{8K^2T_b^2})\right). \tag{95}$$

When $b=\lceil\log_2 K\rceil$, we have $T_b\geq 2K$. Furthermore, when $\lceil\log_2 K\rceil<b\leq\lfloor\log_2 T\rfloor$, it can be concluded that $T_b=2T_{b-1}$. Therefore, we can write

$$\begin{aligned}&\sum_{b=\lceil\log_2 K\rceil}^{\lfloor\log_2 T\rfloor}\max(0,\frac{(T_b-t_b')|\mathcal{E}_{T_b}|}{8K^2T_b^2})\\<&\sum_{b=\lceil\log_2 K\rceil}^{\lfloor\log_2 T\rfloor}\frac{1}{8T_b}\leq\frac{1}{8K}(\sum_{b=\lceil\log_2 K\rceil}^{\lfloor\log_2 T\rfloor}(\frac{1}{2})^{b-\lceil\log_2 K\rceil})\\=&\frac{1}{8K}(2-(\frac{1}{2})^{\lfloor\log_2 T\rfloor-\lceil\log_2 K\rceil}).\end{aligned}\tag{96}$$

Based on (96) and under the assumption that $T>K$, we find

$$\begin{aligned}&\sum_{b=\lceil\log_2 K\rceil}^{\lceil\log_2 T\rceil}\max(0,\frac{(T_b-t_b')|\mathcal{E}_{T_b}|}{8K^2T_b^2})\\<&\frac{1}{8K}(2-(\frac{1}{2})^{\lfloor\log_2 T\rfloor-\lceil\log_2 K\rceil})+\frac{1}{8T}<\frac{3}{8K}.\end{aligned}\tag{97}$$

Thus, Δ meet the conditions in the Lemma 9 and it can be inferred that

$$\Delta>1-\sum_{b=\lceil\log_2 K\rceil}^{\lceil\log_2 T\rceil}\max(0,\frac{(T_b-t_b')|\mathcal{E}_{T_b}|}{8K^2T_b^2})\geq 1-\mathcal{O}(\frac{1}{K}).$$

Therefore, in this case, Exp3-UP satisfies

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})]-\sum_{t=1}^{T}\ell_t(v_i)\\\leq&\mathcal{O}\left(\frac{\alpha(\mathcal{G})}{\epsilon}\ln(T)\ln(KT)(\sqrt{T\ln K}+\sqrt{K\ln(KT)}T^{\frac{2}{3}})\right)\end{aligned}$$

with probability at least $1-\mathcal{O}(\frac{1}{K})$. This completes the proof of Lemma 4.

F. Proof of Theorem 5

Since Exp3-GR chooses the experts one by one for the exploration at the first KM time instants, $\mathbb{E}_t[\ell_t(v_i)]=\ell_t(v_k)$ and (56) hold true. In addition, for $t>KM$ we have

$$\frac{W_{t+1}}{W_t}=\sum_{i=1}^{K}\frac{w_{i,t+1}}{W_t}=\sum_{i=1}^{K}\frac{w_{i,t}}{W_t}\exp\left(-\eta\tilde{\ell}_t(v_i)\right).\tag{98}$$

According to (18), $\frac{w_{i,t}}{W_t}$ can be expressed as

$$\frac{w_{i,t}}{W_t}=\frac{\pi_{i,t}-\frac{\eta}{|\mathcal{D}|}\mathcal{I}(v_i\in\mathcal{D})}{1-\eta}.\tag{99}$$

Following similar steps performed to obtain (35) from (27) and (28), given (98) and (99) we get

$$\begin{aligned}&\sum_{t=t'}^{T}\sum_{i=1}^{K}\pi_{i,t}\tilde{\ell}_t(v_i)-\sum_{t=t'}^{T}\tilde{\ell}_t(v_i)\\\leq&\frac{\ln K}{\eta}+\sum_{t=t'}^{T}\sum_{i\in\mathcal{D}}\frac{\eta}{|\mathcal{D}|}\tilde{\ell}_t(v_i)\\&+\sum_{t=t'}^{T}\sum_{i=1}^{K}\frac{\eta}{2}(\pi_{i,t}-\frac{\eta}{|\mathcal{D}|}\mathcal{I}(v_i\in\mathcal{D}))\tilde{\ell}_t(v_i)^2\end{aligned}\tag{100}$$

where $t'=KM+1$. According to (20), expected value of loss estimate $\tilde{\ell}_t(v_i)$ can be expressed as

$$\begin{aligned}\mathbb{E}_t[\tilde{\ell}_t(v_i)]=&\sum_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\pi_{j,t}p_{ji}\mathbb{E}_t[Q_{i,t}]\ell_t(v_i)\\=&q_{i,t}\mathbb{E}_t[Q_{i,t}]\ell_t(v_i)\end{aligned}\tag{101a}$$

$$\begin{aligned}\mathbb{E}_t[\tilde{\ell}_t(v_i)^2]=&\sum_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\pi_{j,t}p_{ji}\mathbb{E}_t[Q_{i,t}^2]\ell_t(v_i)^2\\=&q_{i,t}\mathbb{E}_t[Q_{i,t}^2]\ell_t(v_i)^2.\end{aligned}\tag{101b}$$

Note that the expected values depend on random variable $\{Z_{i,u}(t)\}_{u=1}^{M}$ in (19), where $P_{i,u}(t)$ and $Y_{ij,u}(t)$, $\forall i\in[K]$, $\forall(i,j)\in\mathcal{E}_t$ are independent Bernoulli random variables with parameters $\pi_{i,t}$ and p_{ij}, respectively. Therefore, $\{Z_{i,u}(t)\}_{u=1}^{M}$ are also Bernoulli random variables with expected value

$$\begin{aligned}\mathbb{E}_t[Z_{i,u}(t)]=&\mathbb{E}_t\left[\sum_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}P_{j,u}(t)Y_{ji,u}(t)\right]\\=&\sum_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\mathbb{E}_t[P_{j,u}(t)]\mathbb{E}_t[Y_{ji,u}(t)]\\=&\sum_{\forall j:v_j\in\mathcal{N}_{i,t}^{\text{in}}}\pi_{j,t}p_{ji}=q_{i,t}.\end{aligned}\tag{102}$$

In other words, $Z_{i,u}(t)$ is a Bernoulli random variable whose value is 1 with probability $q_{i,t}$. The expected value of $Q_{i,t}$ and $Q_{i,t}^2$ can henceforth be written as

$$\begin{aligned}\mathbb{E}_t[Q_{i,t}]=&\sum_{u=1}^{M}uq_{i,t}(1-q_{i,t})^{u-1}+M(1-q_{i,t})^M\\=&\frac{1-(Mq_{i,t}+1)(1-q_{i,t})^M}{q_{i,t}}+M(1-q_{i,t})^M\\=&\frac{1-(1-q_{i,t})^M}{q_{i,t}}\end{aligned}\tag{103a}$$

$$\begin{aligned}\mathbb{E}_t[Q_{i,t}^2]=&\sum_{u=1}^{M}u^2q_{i,t}(1-q_{i,t})^{u-1}+M^2(1-q_{i,t})^M\\=&\frac{2-2(1-q_{i,t}^{M+2})}{q_{i,t}^2}-\frac{1+(2M+3)(1-q_{i,t})^{M+1}}{q_{i,t}}\\&-(M+1)^2(1-q_{i,t})^M+M^2(1-q_{i,t})^M\\=&\frac{2-2(1-q_{i,t}^{M+2})}{q_{i,t}^2}-\frac{1+(2M+3)(1-q_{i,t})^{M+1}}{q_{i,t}}\\&-(2M+1)(1-q_{i,t})^M.\end{aligned}\tag{103b}$$

Combining (101) with (103), we arrive at

$$\begin{aligned}\mathbb{E}_t[\tilde{\ell}_t(v_i)]=&q_{i,t}\frac{1-(1-q_{i,t})^M}{q_{i,t}}\ell_t(v_i)\\=&\left(1-(1-q_{i,t})^M\right)\ell_t(v_i)\leq\ell_t(v_i)\end{aligned}\tag{104a}$$

$$\begin{aligned}\mathbb{E}_t[\tilde{\ell}_t(v_i)^2]=&\left(\frac{2-2(1-q_{i,t}^{M+2})}{q_{i,t}}-1\right)\ell_t(v_i)^2\\&+(2M+3)(1-q_{i,t})^{M+1}\ell_t(v_i)^2\\&-q_{i,t}(2M+1)(1-q_{i,t})^M\ell_t(v_i)^2\leq\frac{2}{q_{i,t}}.\end{aligned}\tag{104b}$$

Combining (100) and (104), it can be concluded that

$$\begin{aligned}
&\sum_{t=t'}^{T}\sum_{i=1}^{K}\pi_{i,t}\ell_t(v_i) - \sum_{t=t'}^{T}\ell_t(v_i) - \sum_{t=t'}^{T}\sum_{i=1}^{K}\pi_{i,t}(1-q_{i,t})^M\ell_t(v_i)\\
\leq&\frac{\ln K}{\eta} + \sum_{t=t'}^{T}\sum_{i=1}^{K}\frac{\eta}{2}(\pi_{i,t} - \frac{\eta}{|\mathcal{D}|}\mathcal{I}(v_i\in\mathcal{D}))\frac{2}{q_{i,t}}\\
&+\sum_{t=t'}^{T}\sum_{i\in\mathcal{D}}\frac{\eta}{|\mathcal{D}|}\ell_t(v_i). \qquad (105)
\end{aligned}$$

According to (a1) $\ell_t(v_i)\leq 1$ and using the fact that $\frac{2}{q_{i,t}}\geq 2$, (105) can be further bounded by

$$\begin{aligned}
&\sum_{t=t'}^{T}\sum_{i=1}^{K}\pi_{i,t}\ell_t(v_i) - \sum_{t=t'}^{T}\ell_t(v_i) \qquad (106)\\
\leq&\frac{\ln K}{\eta} + \sum_{t=t'}^{T}(1-q_{i,t})^M + \sum_{t=t'}^{T}\sum_{i\in\mathcal{D}}\frac{\eta-\eta^2}{|\mathcal{D}|} + \sum_{t=t'}^{T}\sum_{i=1}^{K}\eta\frac{\pi_{i,t}}{q_{i,t}}\\
=&\frac{\ln K}{\eta} + \sum_{t=t'}^{T}(1-q_{i,t})^M + \eta(1-\eta)(T-KM) + \eta\sum_{t=t'}^{T}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}}.
\end{aligned}$$

Note that when $t>t'$, we have $\mathbb{E}_t[\ell_t(v_{I_t})] = \sum_{i=1}^{K}\pi_{i,t}\ell_t(v_i)$. Combining (56) with (106) leads to

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})] - \sum_{t=1}^{T}\ell_t(v_i)\\
\leq&\frac{\ln K}{\eta} + (K-1)M + \sum_{t=t'}^{T}(1-q_{i,t})^M\\
&+\eta(1-\eta)(T-KM) + \eta\sum_{t=t'}^{T}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}} \qquad (107)
\end{aligned}$$

which completes the proof of Theorem 5.

G. Proof of Lemma 6

In this section, the doubling trick is employed to choose η and M when the learner does not know the time horizon T beforehand. At time instant t, when $2^b < t \leq 2^{b+1}$, η_b and M_b the are chosen as

$$\eta_b = \sqrt{\frac{K\ln K}{2^{b+1}}}, M_b = \left\lceil\frac{(b+1)\sqrt{2^{b-1}}|\mathcal{D}|\ln 2}{\epsilon\sqrt{K\ln K}}\right\rceil. \qquad (108)$$

When $t > 2^{b+1}$ holds true, the algorithm restarts with η_{b+1} and M_{b+1}. The algorithm starts with $b=0$. At each restart, the algorithm chooses the experts one by one for exploration until the condition that each expert is chosen at least M_b times is met. Then, the learner uses the last M_b observed samples from each expert to perform geometric resampling. In this case, for each b, Exp3-GR satisfies

$$\begin{aligned}
&\sum_{t=2^b+1}^{T_b}\mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i\in\mathcal{V}}\sum_{t=2^b+1}^{T_b}\ell_t(v_i) \qquad (109)\\
\leq&\frac{\ln K}{\eta_b} + (K-1)(M_b - M_{b-1}) + \sum_{t=t'_b}^{T_b}(1-q_{i,t})^{M_b}\\
&+\eta_b(1-\eta_b)(T_b - 2^b - K(M_b - M_{b-1})) + \eta_b\sum_{t=t'_b}^{T_b}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}}
\end{aligned}$$

where T_b denote the greatest time instant which satisfies $2^b < T_b \leq 2^{b+1}$ and t'_b can be expressed as in (88). Note that when $b=0$, we have $M_{b-1}=0$. Taking into account that the maximum loss at each restart is 1, summing (109) over all possible values for b obtains

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i\in\mathcal{V}}\sum_{t=1}^{T}\ell_t(v_i)\\
\leq&\lceil\log_2 T\rceil + \sum_{b=0}^{\lfloor\log_2 T\rfloor}\frac{\ln K}{\eta_b} + (K-1)M\\
&+\sum_{b=0}^{\lfloor\log_2 T\rfloor}\eta_b(1-\eta_b)(T_b - 2^b - K(M_b - M_{b-1}))\\
&+\sum_{b=0}^{\lfloor\log_2 T\rfloor}\eta_b\sum_{t=t'_b}^{T_b}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}} + \sum_{b=0}^{\lfloor\log_2 T\rfloor}\sum_{t=t'_b}^{T_b}(1-q_{i,t})^{M_b} \qquad (110)
\end{aligned}$$

where M is the number of samples for each expert when $b = \lfloor\log_2 T\rfloor$ which are used for geometric resampling. According to (108) and the fact that $\mathcal{D}$ is obtained using the greedy set cover algorithm, we have

$$M = \mathcal{O}\left(\frac{\alpha(\mathcal{G})}{\epsilon\sqrt{K}}\ln T\sqrt{T\ln K}\right). \qquad (111)$$

Furthermore, for each b, the inequality $q_{i,t} > \frac{\eta_b\epsilon}{|\mathcal{D}|}$ holds. Therefore, according to (108), we can write $M_b q_{i,t} > \frac{b+1}{2}\ln 2$. Thus, it can be concluded that $(1-q_{i,t})^{M_b} \leq e^{-M_b q_{i,t}} < \frac{1}{\sqrt{2^{b+1}}}$ from which we obtain

$$\begin{aligned}
\sum_{b=0}^{\lfloor\log_2 T\rfloor}\sum_{t=t'_b}^{T_b}(1-q_{i,t})^{M_b} &< \sum_{b=0}^{\lfloor\log_2 T\rfloor}\frac{T_b - 2^b}{\sqrt{2^{b+1}}} \leq \sum_{b=0}^{\lfloor\log_2 T\rfloor}\sqrt{2^{b-1}}\\
&\leq \frac{\sqrt{2T}-1}{2-\sqrt{2}}. \qquad (112)
\end{aligned}$$

In addition, based on the Lemma 7, it can be written that

$$\begin{aligned}
&\sum_{b=0}^{\lfloor\log_2 T\rfloor}\eta_b\sum_{t=t'_b}^{T_b}\sum_{i=1}^{K}\frac{\pi_{i,t}}{q_{i,t}}\\
\leq&\sum_{b=0}^{\lfloor\log_2 T\rfloor}\sqrt{\frac{K\ln K}{2^{b+1}}}(T_b - 2^b)\mathcal{O}\left(\frac{\alpha(\mathcal{G})}{\epsilon}\ln(KT)\right)\\
\leq&\lceil\log_2 T\rceil\sqrt{2^{b-1}K\ln K}\mathcal{O}\left(\frac{\alpha(\mathcal{G})}{\epsilon}\ln(KT)\right)\\
=&\mathcal{O}\left(\frac{\alpha(\mathcal{G})}{\epsilon}(\ln T)\ln(KT)\sqrt{KT\ln K}\right). \qquad (113)
\end{aligned}$$

Therefore, combining (110) with (111), (112) and (113), it can be inferred that Exp3-GR satisfies

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_t[\ell_t(v_{I_t})] - \min_{v_i\in\mathcal{V}}\sum_{t=1}^{T}\ell_t(v_i)\\
\leq&\mathcal{O}\left(\frac{\alpha(\mathcal{G})\ln T}{\epsilon}\ln(KT)\sqrt{KT\ln K}\right) \qquad (114)
\end{aligned}$$

which completes the proof of Lemma 6.