Online Learning with Dependent Stochastic Feedback Graphs

Corinna Cortes 1  Giulia DeSalvo 1  Claudio Gentile 1  Mehryar Mohri 1  Ningshan Zhang 2

Abstract
A general framework for online learning with partial information is one where feedback graphs specify which losses can be observed by the learner. We study a challenging scenario where feedback graphs vary stochastically with time and, more importantly, where graphs and losses are dependent. This scenario appears in several real-world applications that we describe where the outcome of actions are correlated. We devise a new algorithm for this setting that exploits the stochastic properties of the graphs and that benefits from favorable regret guarantees. We present a detailed theoretical analysis of this algorithm, and also report the results of a series of experiments on real-world datasets, which show that our algorithm outperforms standard baselines for online learning with feedback graphs.

1. Introduction
Prediction with expert advice is a classical framework for modeling the repeated interactions between a learner and the environment (Littlestone & Warmuth, 1994; Cesa-Bianchi & Lugosi, 2006). In this framework, the learner maintains performance estimates for a set of experts (actions), based on their past behavior. At each round, she uses those estimates to select an action, which incurs a loss determined by the environment, and subsequently updates her estimates. The learner’s objective is to minimize her regret, that is, the difference between her cumulative loss over a finite number of interactions and that of the best expert in hindsight.

The two most standard settings in this framework are the full information setting, where the learner is informed of the loss incurred by all actions, and the bandit setting, where she can only observe the loss incurred by the selected action. These settings are both special instances of a general model of online learning introduced by Mannor & Shamir (2011), where loss observability is specified by a feedback graph. In a directed feedback graph, each vertex represents an action and an edge from vertex $i$ to vertex $j$ indicates that the loss of action $j$ is observed when action $i$ is selected. The bandit setting corresponds to a feedback graph reduced to only self-loops at each vertex, the full information setting corresponds to a fully connected graph. Online learning with feedback graphs has been further extensively analyzed by Alon et al. (2013; 2017) and by a series of other recent publications (Caron et al., 2012; Alon et al., 2015; Kocák et al., 2014; Neu, 2015; Cohen et al., 2016; Buccapatnam et al., 2014b; Wu et al., 2015a; Tossou et al., 2017; Liu et al., 2018; Yun et al., 2018; Cortes et al., 2019; Li et al., 2020).

In the scenarios of online learning with feedback graphs studied by some authors, the feedback graph $G$ is assumed to be fixed over time (Caron et al., 2012; Buccapatnam et al., 2014b; Neu, 2015), while others allow a time-varying feedback graph $G_t$, chosen adversarially at each round $t$ (Alon et al., 2013; 2017; 2015; Kocák et al., 2014; Cohen et al., 2016; Kocák et al., 2016). This paper considers a scenario commonly appearing in practice where feedback graphs are time-varying but are stochastic, that is, at each round $t$, the feedback graph $G_t$ is drawn i.i.d. according to some unknown distribution. Crucially, we will not assume that the losses and feedback graphs are statistically independent, a key assumption adopted by many authors, including Cohen et al. (2016); Tossou et al. (2017); Liu et al. (2018); Li et al. (2020), in their analysis of (partially accessible) feedback graphs. The novelty of the setting presented in this paper is thus not just to study stochastic feedback graphs, but more importantly, to allow for an underlying dependence between the feedback graphs and the losses, which as we will see makes the analysis intrinsically harder. We call this new setting online learning with dependent stochastic feedback graphs highlighting that losses and graphs are dependent stochastic variables.

Our scenario of online learning with dependent stochastic graphs is relevant to many real-world problems. As an example, consider the problem of a doctor selecting an appropriate treatment (action) for a patient, based on his symptoms. The patient can be modeled by a feature vector $x_t$ representing his health history and other characteristics and we assume a fixed distribution over the patients $x_t$. 

---

*Equal contribution 1Google Research, New York, NY 2Hudson River Trading, New York, NY. Correspondence to: Giulia DeSalvo <giuliad@google.com>.

visiting that doctor. Each \( x_t \) induces a feedback graph over treatments, since the success of treatment \( i \) for curing a collection of symptoms that the patient displays could also reveal the effectiveness of an alternate treatment \( j \). Thus, just as \( x_t \), the feedback graph is stochastic and since both the loss and the feedback graph depend on \( x_t \), they are not independent. More generally, the scenario of online learning with stochastic feedback graphs is pertinent to a variety of problems where a stochastic context or feature vector \( x_t \) induces a relationship between experts or actions, as the outcome of one expert for \( x_t \) also reveals information about that of some other experts. We will discuss in more detail several examples of such structured experts.

Our study admits some connection with that of Kocák et al. (2016), who examine a scenario of online learning with an imperfect feedback, modeled by a weighted graph, or the closely related one of Wu et al. (2015b); however, the definition of the edge weights in those studies is entirely distinct from ours. Related to feedback graph problems, but in settings quite different from ours, Yun et al. (2018) consider stochastic bandits where the extra information admits a cost to be traded off against regret, while Cortes et al. (2019) analyze feedback graphs in the context of the sleeping experts.

Our setting naturally raises the following question: how can we leverage the stochastic properties of the graphs while at the same time dealing with the dependency between losses and graphs? One can prove that simply averaging loss observations of an expert based on the feedback graph \( G_t \) may result in an arbitrarily biased empirical estimate of the expected loss of the expert, since there is a dependency between losses and graphs and since we make no assumptions on the nature of this dependency. In order to avoid this pitfall, we introduce an estimated feedback graph \( \hat{G}_t \) based on past observations up to time \( t−1 \) instead of relying on the feedback graph \( G_t \) at the time \( t \). The design of our algorithms then relies on the observability specified by \( \hat{G}_t \) so that the empirical estimates defined by averaging loss observations based on the estimated graphs will be unbiased. Specifically, the empirical estimate of the performance of action \( j \) is updated whenever action \( I_t \) is selected by the algorithm and there is an edge from \( I_t \) to \( j \) in \( \hat{G}_t \).

Using estimated graphs, however, leads to another complication: there are times \( t \) when even though there is an edge from \( I_t \) to \( j \) in \( \hat{G}_t \), the loss of action \( j \) is not revealed, which is when graph \( G_t \) does not contain an edge from \( I_t \) to \( j \). In order for these events not to occur too often, we introduce a threshold \( \theta_t \) to remove edges with low probability in the estimated graph. That is, depending only on information up to time \( t−1 \), we estimate the probability of an edge from vertex \( i \) to vertex \( j \) and define, at each round \( t \), the graph \( \hat{G}_t \) whose edges admit a probability above this threshold \( \theta_t \). Furthermore, over these events, since the loss of action \( j \) is not revealed, we also introduce an estimated loss used to update the empirical estimate of the performance of expert \( j \). These estimated losses effectively estimate the expectation of the loss of expert \( j \) when there is no edge from \( i \) to \( j \) in graph \( G_t \).

Intuitively, denser feedback graphs imply that more loss observations are revealed and hence, the algorithm should converge faster since the empirical estimates more quickly converge to their respective means. The novelty of our approach is thus based on carefully leveraging the interplay between the estimated graphs, thresholds, and estimated losses in order to reliably update the empirical estimates of the expected loss of an expert via the densest graph possible, thereby circumventing the dependency of losses and graphs while also exploiting the presence of edges in the feedback graph in order to attain favorable theoretical guarantees.

We start with a formal description of the scenario of online learning with dependent stochastic feedback graphs and introduce the relevant notation (Section 2). In Section 3, we present a novel algorithm that exploits the stochasticity of the feedback graphs and carefully deals with the dependency between losses and graphs. In Section 4, we prove that our algorithm benefits from favorable pseudo-regret guarantees that are logarithmic in the number of rounds \( T \) and are expressed in terms of the expected feedback graph, \( \mathbb{E}[G_t] \).

In that section, we also illustrate, through a lower bounding argument, that the dependency structure between losses and graphs may render the extra side information delivered by feedback graphs mostly useless. In Section 5, we discuss a natural scenario with stochastic graphs and report the results of several experiments on real-world datasets suggesting that our algorithm outperforms standard baselines.

### 2. Learning Scenario

We consider the familiar sequential learning problem over \( T \) rounds, where the algorithm has access to a set of \( K \) experts (or actions) \( \mathcal{E} = \{\xi_1, \ldots, \xi_K\} \), which we abusively identify as \( \mathcal{E} = \{1, \ldots, K\} \) in some cases. At each round \( t \in [T] \), each expert \( \xi_t \in \mathcal{E} \) is assigned a loss \( \ell_t(\xi_t) \in [0,1] \). The algorithm selects an expert \( \hat{\xi}_t \), and incurs the corresponding loss \( \ell_t(\hat{\xi}_t) \). Additionally, at round \( t \in [T] \), the algorithm is supplied with a feedback graph \( G_t = (\mathcal{E}, E_t) \), where each vertex corresponds to an expert, and where the presence of an edge from vertex \( i \) to vertex \( j \) indicates that the loss of expert \( \xi_j \) is revealed to the algorithm if it selects expert \( \xi_i \) at time \( t \). In what follows, we denote by \( N_t(i) \) the out-neighborhood of vertex \( i \) in \( G_t \), that is the set of vertices \( j \) that can be reached from \( i \) via an edge in \( E_t \). Note, the feedback graph \( G_t \) only specifies the observability of expert losses and not loss values.
We consider the so-called action to play. The standard measure of the performance of the algorithm includes at least the loss value \( \ell_t(\xi_{t}) \) of the action \( \xi_{t} \), it selects, that is the information available in the bandit setting.

We consider the setting where expert losses and feedback graphs are stochastic and arbitrarily dependent. Let \( \mathcal{G} \) denote the family of all directed graphs with \( K \) vertices (and self-loops on all of them), and let \( \mathcal{D} \) be a distribution over \( \mathcal{G} \times \{0, 1\}^K \) unknown to the algorithm. At each round \( t \in [T] \), the environment generates the pair \((G_t, \xi_t) \in \mathcal{G} \times \{0, 1\}^K \) i.i.d. according to \( \mathcal{D} \), where \( \xi_t = (\ell_t(\xi_1), \ldots, \ell_t(\xi_K)) \) is the vector of all \( K \) expert losses.

In particular, this setting includes the important learning scenario of statistical learning, so far not studied in the bandits literature, where pairs \((x_t, y_t)\) are drawn i.i.d. according to some distribution over \( X \times Y \), where \( X \) is the input space and \( Y \) is the output space. Expert \( \xi_t \) is a mapping from \( X \) to \( Y \) and its loss at time \( t \) is given by \( \ell_t(\xi_t) = L(\xi_t(x_t), y_t) \), for some loss function \( L \). The feedback graph \( G_t \) is then typically a function of \( x_t \). As an example, consider the following sequence of graphs: at each time \( t \), graph \( G_t \) admits (bi-directional) edges \((i, j)\) whenever \( \xi_t(x_t) = \xi_j(x_t) \). The probability \( p_{t,ij} \) of an edge in this graph is then given by \( p_{t,ij} = \mathbb{P}(\{x: \xi_t(x) = \xi_j(x)\}) \), where \( p_{t,ii} = 1 \) if the graph admits self-loops. Experts may be, for example, decision trees with the same function value in a given region of the space, which thereby implies that these trees will admit the same loss in this region, and the graph \( G_t \) will then account for the shared loss observability between these trees whenever an \( x_t \) falls in this region. In such scenarios, \( G_t \) and the losses \( \ell_{t} \) are dependent as they are both functions of \( x_t \). Note that the full vector of losses \( \ell_{t} \) is generated i.i.d. at each time \( t \), but its components \( \ell_{t}(\xi_t) \) can be dependent. Likewise, the graphs \( G_t \) are i.i.d. across time but, for any given \( t \), their edges \((i, j)\) may be correlated.

Another instance of this scenario is learning with abstenion, where an expert can elect to either make a prediction or abstain, at a cost typically less than that of misclassification (Cortes et al., 2016; 2018). In the stochastic case, both the loss vector and the feedback graph at round \( t \) are functions of \( x_t \), with the feedback graph \( G_t \) defined by the following: if the expert selected by the algorithm elects to predict on input \( x_t \), then the label \( y_t \) is revealed, and the loss \( \ell_t(\xi_t) = L(\xi_t(x_t), y_t) \) of every expert \( \xi_t \) is revealed; otherwise, \( \xi_t \), abstains, the label \( y_t \) is not revealed and only the loss of \( \xi_t \), along with that of other abstaining experts is revealed, which is some fixed abstention cost \( c \).

We consider the so-called uninformed case, where graph \( G_t \) is revealed to the algorithm only after it has selected an action to play. The standard measure of the performance of the algorithm is the pseudo-regret, \( R_T \), defined as follows:

\[
R_T = \max_{\xi \in \mathcal{E}} \mathbb{E} \left[ \sum_{t=1}^{T} \ell_t(\xi_{t}) - \ell_t(\xi) \right],
\]

where the expectation is taken over the i.i.d. sequence of pairs \((G_t, \xi_t)\) drawn from \( \mathcal{D} \).

3. Algorithm

In this section, we present our new algorithm UCB-DSG, or UCB with Dependent Stochastic Graphs, for the stochastic and loss-dependent feedback graph scenario just discussed. In Section 4, we prove that this algorithm admits favorable pseudo-regret guarantees. The design of UCB-DSG builds on the classical UCB algorithm of Auer et al. (2002a). The pseudocode of UCB-DSG is given in Algorithm 1.

At a high level, UCB-DSG maintains an empirical estimate \( \hat{\mu}_{i,t} \) for the average loss \( \mu_i = \mathbb{E}[\ell_t(\xi_i)] \) of each expert \( \xi_i \in \mathcal{E} \). At each round, it selects the expert with the smallest lower confidence estimate, and uses the loss observations from a feedback graph, defined below, to update its estimates.

Ideally, the algorithm would update its empirical means at round \( t \) according to graph \( G_t \). However, since \( G_t \) may not be a complete graph and may depend on the losses \( \ell_t(\xi_i) \), simply taking an average over the losses observed could lead to biased empirical estimates. As a straightforward example, consider a graph \( G_t \) that, when selecting expert \( \xi_t \), also reveals the loss of expert \( \xi_j \), but only when \( \ell_t(\xi_j) \geq 0.5 \). For a detailed discussion of this technical bias issue in the context of the online learning with abstention, see (Cortes et al., 2018). Section 4 below contains an illustrative example in the form of a lower bound that shows that even if the (marginal) probabilities \( p_{ij} \) of edges \((i,j)\) are all close to 1, we cannot in general take advantage of the presence of these edges when graphs and losses are statistically dependent.

The first key idea behind the design of our algorithm is to use at time \( t \) a surrogate graph \( \hat{G}_t \), derived from an average of past observed graphs \( G_1, \ldots, G_{t-1} \), which bypasses this technical bias issue. However, the loss of an expert \( \xi_j \) in the out-neighborhood \( \hat{N}_j(I_t) \) of graph \( \hat{G}_t \) may not be actually observed at round \( t \). This is because, in general, graphs \( G_t \) and \( \hat{G}_t \) do not coincide and \( N_j(I_t) \) may not contain \( \xi_j \). In order to control the estimation error from such cases, we introduce a threshold that is used to trim the estimated graph of any low-probability edges. Moreover, in these cases, UCB-DSG resorts to an estimated loss \( \hat{\ell}_t(\xi_j) \) for expert \( \xi_j \) since the true loss is not revealed. The threshold and the estimated losses are carefully crafted to guarantee that the true mean of each expert is, with high probability, within a confidence interval around the empirical mean. Specifically, there is a critical interplay between the estimated losses and thresholds to favor a denser graph while at the same time...
ALGORITHM 1: UCB-DSG

Parameters: Experts $E = \{\xi_1, \ldots, \xi_K\}$, # of rounds $T$; Init: $U_{i,0} = Q_{i,0} = 1$ for all $i \in [K]$;
for $t \in [T]$ do
$S_{i,t-1} \leftarrow \frac{O(\log(Kt))}{\sqrt{Q_{i,t-1}}}$; {Constants in App.B}
$I_t \leftarrow \arg\min_{i \in [K]} (p_{i,t-1} - S_{i,t-1});$
$\tilde{N}_t(I_t) \leftarrow \{j \in [K] : \hat{\rho}_{i,j,t} \geq \theta_{j,t-1}\};$ {Create $\tilde{G}_t$}
for $j \in [K]$ do
if $j \in \tilde{N}_t(I_t)$ then
$Q_{j,t} \leftarrow Q_{j,t-1} + 1;
$ if $j \in N_t(I_t)$ then
$\tilde{\mu}_{j,t} \leftarrow \frac{t(j)}{Q_{j,t}} + (1 - \frac{1}{Q_{j,t}})\tilde{\mu}_{j,t-1};$
{Use true loss}
else
$\nu_{j,t} \leftarrow \frac{\nu(j)}{Q_{j,t}} + (1 - \frac{1}{Q_{j,t}})\nu_{j,t-1};$
{Use estimated loss}
else
$Q_{j,t} \leftarrow Q_{j,t-1}; S_{j,t} \leftarrow S_{j,t-1}; \nu_{j,t} \leftarrow \nu_{j,t-1};$
{No update}
if $\hat{\rho}^{t-1}_{j,s} \geq 1 - \frac{O(\log(Kt))}{\sqrt{Q_{j,t-1}}}$ + 1 then
$U_{j,t} \leftarrow U_{j,t-1} + 1;
$ for $i \in [K]$ do
$\nu_{j,i,t} \leftarrow \frac{\nu(i)}{Q_{j,t}} \mathbb{1}\{j \notin N_t(i) \} \mathbb{1}\{j \in N_t(I_t)\}$
+ $(1 - \frac{1}{U_{j,t}}) \nu_{j,i,t-1};$
else
$U_{j,t} \leftarrow U_{j,t-1}; \nu_{j,i,t} \leftarrow \nu_{j,i,t-1}, \forall i \in [K];$
{No update}
end if
end for
end if
end for
end for

reliably updating the empirical estimates, $\tilde{\mu}_{i,t}$.

To devise our surrogate graph $\tilde{G}_t$, for $t \geq 2$ and for all $i, j \in [K]$, we use an unbiased estimate of the probability of a directed edge from $i$ to $j$, $\hat{p}_{i,j,t}^{-1}$, based on the information available up to time $t - 1$:

$$\hat{p}_{i,j,t}^{-1} = \frac{\sum_{s=1}^{t-1} \mathbb{1}\{j \in N_t(i)\}}{t - 1},$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function of the argument, and $\hat{p}_{i,j,0} = 0$. We define $\tilde{G}_t$ by allowing an edge from $i$ to $j$ only if $\hat{p}_{i,j,t}^{-1}$ is above the aforementioned threshold. That is, the graph $\tilde{G}_t$ is determined by the out-neighborhoods $\tilde{N}_t(i) = \{j \in [K] : \hat{p}_{i,j,t}^{-1} \geq \theta_{j,t-1}\} \cup \{i\}$, with the threshold $\theta_{j,t-1}$ specified by

$$\theta_{j,t-1} = \min \left\{ 2 - \frac{U_{j,t-1} + 1}{Q_{j,t-1} + 1} - \frac{O(\log(Kt))}{\sqrt{Q_{j,t-1} + 1}}, \right.$$

$$\left. 1 - \frac{O(\log(Kt))}{\sqrt{Q_{j,t-1} + 1}} \right\}. \quad (1)$$

In the above, $Q_{j,t-1}$ is the number of times expert $\xi_j$ was updated up to time $t - 1$ and similarly, $U_{j,t-1}$ is the number of times the estimated loss, $\tilde{\ell}_t(\xi_j)$, has been updated up to time $t - 1$. The exact constants in front of the log terms hidden in the big-O notation are given in Appendix B. Note that, by construction, $\tilde{G}_t$ admits self-loops at all vertices.

Since the algorithm uses the estimated loss of expert $\xi_j$ when the loss of expert $\xi_j$ is not revealed, that is when $j \notin N_t(I_t)$, our estimated loss aims at estimating the following conditional expected loss:

$$E[\ell_t(\xi_j)|j \notin N_t(I_t)] = \frac{E[\ell_t(\xi_j)|j \notin N_t(I_t)]}{E[\mathbb{1}\{j \notin N_t(I_t)\}]}.$$

The denominator in the above fraction can be readily estimated by $1 - \tilde{p}_{1,t}^{-1}$, which converges to its (conditional) mean, $1 - p_{1,t}$. Concretely, for each pair $i, j \in K \times K$ and time $t \in [T]$, we maintain estimates $\hat{p}_{i,j,t}^{-1}$ and, at time $t$, the probability estimate $\hat{p}_{i,j,t}^{-1}$ corresponding to $i = I_t$ is used.

On the other hand for the numerator $E[\ell_t(\xi_j)|j \notin N_t(i)]$, we need to introduce a new estimator and so we define

$$\tilde{\nu}_{j,i,t} = \frac{1}{U_{j,t}} \sum_{s=1}^{t} \ell_s(\xi_j) \mathbb{1}\{j \notin N_s(i)\} \mathbb{1}\{j \in N_s(I_s)\} C_{j,s}$$

with

$$C_{j,s} = 1 \left\{ \frac{\hat{p}_{1,s}^{-1} - 1}{\tilde{p}_{1,s}^{-1} - 1} - \frac{O(\log(Ks))}{\sqrt{U_{j,s-1} + 1}} \right\} \text{ and } U_{j,t} = \sum_{s=1}^{t} C_{j,s}$$

for all $j, i, t$. Note that the sum in the definition of $\tilde{\nu}_{j,i,t}$ must be over the times $s$ such that $j \notin N_s(I_s)$ since the loss of expert $\xi_j$ is revealed only when playing action $I_s$. A careful reader would notice that this injects a bias in the estimator $\tilde{\nu}_{j,i,t}$, but the condition defining $C_{j,s}$, which holds whenever the (estimated) probability of the edge from $I_s$ to $j$ is large enough, is chosen so as to guarantee that $\tilde{\nu}_{j,i,t}$ converges to $E[\ell_t(\xi_j)|j \notin N_t(i)]$. One may, in fact, adopt a condition like $C_{j,s}$ in order to define the estimated graph $\tilde{G}_s$ directly, but this would result in a worse regret guarantee, since $C_{j,s}$ is provably more conservative than the condition $\tilde{p}_{1,s}^{-1} \geq \theta_{j,s-1}$ used in $\tilde{G}_s$. Putting the above together, our estimated loss is defined as

$$\tilde{\ell}_t(\xi_j) = \min \left\{ \tilde{\nu}_{j,I_t,t-1}, 1 \right\} \left(\tilde{p}_{1,t}^{-1}, 1 \right).$$

Thresholds $\theta_{j,t}$ and estimated losses $\tilde{\ell}_t(\xi_j)$ together play a key role in the regret guarantee of our algorithm. As we will prove in our analysis, a small threshold $\theta_{j,t}$ (implying that the algorithm will resort to estimated losses more often) results in a more favorable regret bound. The value of $\theta_{j,t}$ in (1) chiefly depends on the interplay between the two quantities $U_{j,t}$ and $Q_{j,t}$. One can show by induction
(see Appendix B) that \( U_{j,t} \leq Q_{j,t} + 1 \) for all \( j \) and \( t \). Hence, if \( U_{j,t} \) is as large as \( Q_{j,t} \) (which implies that \( \hat{\ell}_i(\xi_j) \) is a very accurate estimator of its own expectation), then \( \theta_{j,t} \) will be small. On the other hand, when \( U_{j,t} \) is much smaller than \( Q_{j,t} \), then the estimated loss \( \hat{\ell}_i(\xi_j) \) is a less reliable estimator. In this case, the threshold is \( \theta_{j,t} = 1 - 2O(\log(KT)\sqrt{Q_{j,t} + 1}) \), which will be large if action \( j \) has already undergone many updates. In particular, if the algorithm has already resorted to \( \hat{\ell}_i(\xi_j) \) very often and \( \hat{\ell}_i(\xi_j) \) is not dependable because of a small \( U_{j,t} \), a large \( \theta_{j,t} \) will ensure that the algorithm will rely on \( \hat{\ell}_i(\xi_j) \) less frequently in the future. Finally, \( \theta_{j,t} \) also mildly decreases with time, due to its logarithmic dependence on \( t \) through \( O(\log(KT)) \).

In summary, at each round \( t \in [T] \), UCB-DSG selects the expert with the smallest optimistic empirical loss, that is, \( I_t = \arg\min_{i \in [K]} \hat{\mu}_{i,t-1} - S_{i,t-1} \), and operates with the surrogate graph \( \hat{G}_t \) to update all estimates \( \hat{\mu}_{j,t} \), for \( j \in \hat{N}(I_t) \), by either using a true loss \( \ell_i(\xi_j) \), when available \((j \in N(I_t))\), or an estimated loss \( \hat{\ell}_i(\xi_j) \) otherwise. Notice that, since the definition of \( \hat{G}_t \) only depends on graphs \( G_{t,1}, \ldots, G_{t-1} \) and not on \( G_t \), Algorithm 1 need not receive graph \( G_t \) before playing action \( I_t \) (uninformed setting).

### 4. Theoretical Guarantees

In this section, we analyze the UCB-DSG algorithm just described and prove that it benefits from favorable logarithmic pseudo-regret guarantees in terms of the independent sets of certain graphs derived from the matrix of probabilities \([p_{i,j}]_{i,j=1}^{K \times K}\). Our regret upper bound is given below, a proof sketch is provided in Section 4.1 and detailed in Appendix B. In Section 4.2, we complement our upper bound with a lower bound showing that in the case of strong dependency between graphs and losses, even when \( p_{i,j} \) is close to 1 for all \( i \) and \( j \), we cannot in general achieve better regret performance than \( O(K) \).

As in standard UCB pseudo-regret bounds, our guarantees are expressed in terms of the loss gaps \( \Delta_j = \mu_j - \mu_{i^*} \), \( j \in [K] \), where \( \mu_j \) is the expected loss of \( \xi_j \) and \( \mu_{i^*} = \min_{j \in [K]} \mu_j \) that of the best expert, \( \xi_{i^*} \). Naturally, our results are also expressed in terms of graph properties. For any \( i, j \in [K] \), let \( p_{i,j} \) denote the marginal probability of an edge from \( i \) to \( j \). We define \( G \) as the weighted directed graph that admits an edge from \( i \) to \( j \) with weight \( p_{i,j} \), for any \( i, j \in [K] \). \( G \) can then be viewed as the process graph from which graphs \( G_t \) are drawn i.i.d., or their expectation, \( E[G_t] \). For any vector of thresholds \( \theta = \{\theta_1, \ldots, \theta_K\} \in [0,1]^K \), we define \( G_\theta = \{E, E_\theta\} \) as the unweighted and undirected graph derived from \( G \) by keeping only those edges in \( G \) that satisfy both \( p_{i,j} \geq \theta_i \) and \( p_{j,i} \geq \theta_j \). In our bound, we only consider strictly positive loss gaps, \( \Delta_j > 0 \) and omit this condition to avoid a notational burden. Recalling that an independent set, \( I(G) \), of an undirected graph \( G \) is the set of vertices such that no two vertices in \( I(G) \) are adjacent, we denote by \( \bar{I}(G) = \{I^1(G), I^2(G), \ldots\} \) the set of all independent sets of graph \( G \). The bound in the theorem that follows is in terms of the independent sets of \( G_\theta \), for suitably chosen \( \theta \).

**Theorem 1.** Let

\[
K^* = \{ j \in [K] : p_{i^*,j} \geq 1 - O\left( \left( \frac{\log(KT)}{\Delta_j} + 1 \right)^{-1/2} \right) \}
\]

and for \( \gamma \in [0,1) \), let \( \bar{\theta}(\gamma) = (\theta_1(\gamma), \ldots, \theta_K(\gamma)) \) be defined as

\[
\theta_i(\gamma) = \begin{cases} 1 - \frac{\gamma}{1 - \frac{\log(KT)}{\sqrt{\theta}}} & \text{for } i \in K^* \\ 1 - \frac{\log(KT)\sqrt{T}}{\gamma} & \text{for } i \notin K^*. \end{cases}
\]

The pseudo-regret \( R_T \) of UCB-DSG over \( T \) rounds is bounded by:

\[
O \left( \min_{\gamma \in [0,1]} \left( \max_{I \in \bar{I}(G_{\bar{\theta}(\gamma)})} \sum_{j \in I} \frac{\ln(KT)\log(T)}{\Delta_j} + \frac{f(D,K)}{1-(1+\gamma)^2/4} \right) \right).
\]

In the above, \( f(D,K) \) is a function linear in the number of experts \( K \) and inversely proportional to each gap \( \Delta_j \in D = \{\Delta_j : j \in [K], j \neq i^*\} \), but independent of \( T \).

The denser the graph \( G_{\bar{\theta}(\gamma)} \) the smaller the number of the independent sets, and thus the more favorable the bound is. Recalling that the independence number of a graph is the size of the maximum independent set of the graph, the number of terms in the above sum over any independent set, \( I \), is upper bounded by the independence number, \( \alpha(\gamma) \), of \( G_{\bar{\theta}(\gamma)} \). Thus, in the case that \( \Delta_j = \Delta \) for all \( j \neq i^* \), this sum can be bounded by \( \frac{\alpha(\gamma)}{\Delta} \ln(KT) \log(T) \). Moreover, \( \gamma \in [0,1) \) is a free parameter that allows for a limited tradeoff between the two terms in the sum. As \( \gamma \) increases, the number of independent sets decreases while the second term \( f(D,K)/(1-(1+\gamma)^2/4) \) increases, and vice-versa.

Figure 1 contains an example of the thresholded process graph \( G_{\bar{\theta}(\gamma)} \) defined in the theorem. Due to how the thresholds \( \bar{\theta}(\gamma) \) are defined, more edges connecting nodes within \( K^* \) are included in the graph \( G_{\bar{\theta}(\gamma)} \), as compared to the others, thereby implying that the larger the set \( K^* \) is, the better the resulting regret bound. More concretely, if \( T \) is large and \( \gamma \) is a constant (say, \( \gamma = 1/2 \)), then \( 1 - \gamma \) is large, so that \( K^* \) is more likely to form a clique in \( G_{\bar{\theta}(\gamma)} \). In this latter case, \( \alpha(\gamma) \leq K - |K^*| + 1 \) and hence, the bigger \( K^* \) the more favorable the bound is. Notice that the size of \( K^* \) is also influenced by the gaps \( \Delta_j \). In particular, for a given
The standard UCB bound corresponds to a sum over all experts and since $G$ admits self-loops at all vertices with weight one, our guarantee is always at least as favorable. In the setting of fixed undirected graphs, Buccapatnam et al. (2014a) present an interesting action elimination algorithm that solves at every round a linear program based on the feedback graph. The authors prove favorable guarantees in terms of the minimum dominating set of the graph. Unfortunately, several difficulties arise when adapting (Buccapatnam et al., 2014a) to our setting of dependent stochastic graphs. From a computational standpoint, we would have to solve a linear program at every round based on the feedback graph’s edge probabilities $p^t_{i,j}$. Moreover, we would need to know the time horizon $T$ in advance (a key step of our analysis of UCB-DSG heavily depends on this property). One might instead estimate the $p_{i,j}$’s in an initial stage where no actions are eliminated and then, say, adopt the strategy in (Li et al., 2020) that works in the case when the stochastic graph is independent of the losses and the $p_{i,j}$ are known. However, even this solution looks problematic in our setting. This is because the duration of this initial stage should be at least $\sqrt{T}$ in order to insure accurate estimation and, if we don’t rely on feedback graphs during this stage we would suffer anyway a regret of the form $\frac{K}{T} \log T$. More importantly, observe that even if we know the marginal probabilities $p_{i,j}$ exactly, we still have to face the loss bias issue intrinsic to our dependency assumptions.

4.1. Sketch of the Pseudo-regret Analysis

Here, we sketch the analysis that proves Theorem 1. Please see Appendix B for the details. At a high level, the proof proceeds via two steps: 1) we prove that $\mu_{i,t}$ converges to its mean at a rate of $\tilde{O}(1/T)$ and then bound the regret using a UCB-type analysis; 2) we relate the resulting bound to the independent sets of the process graph, $G_{\tilde{e}(\gamma)}$.

For step 1), we first quantify the accuracy of our loss estimates $\tilde{e}_t(x)$ (Lemma 1 and Lemma 2 in Appendix B). Specifically, Lemma 1 shows that even though $\mu_{i,t}$ is a biased estimator, it still concentrates around $\tilde{e}_t(x)$, converging to its expectation at a rate of $O(1/\sqrt{T})$. By using the definition of threshold $\theta_{i,j}$ and then, a regret of the form $\frac{K}{T} \log T$.

The independence number of this graph is $\alpha_G$.

The independence number of this graph is $\alpha_G$.

$\alpha(\gamma)$ is highly connected to experts with large expected losses, the pseudo-regret bound is favorable even if their connecting edge weights are comparatively small.

The standard UCB bound corresponds to a sum over all experts and since $G$ admits self-loops at all vertices with weight one, our guarantee is always at least as favorable. In the special case where the probability matrix $[p_{i,j}]_{i,j=1}^K$ is binary and symmetric, then graphs $G_i$ are deterministic, and all coincide with $G$. In this case, our bound coincides with that of Lykouris et al. (2020) [Theorem 3.2], which holds for a UCB-type algorithm, called UCB-N, that averages all available loss observations. In Appendix B (see Theorem 3 therein), we also prove a regret guarantee in terms of the clique partition number of $G_{\tilde{e}(\gamma)}$ which, in the case of deterministic graphs, matches the clique-based bound for UCB-N in (Caron et al., 2012) [Theorem 2]. Even though summing over nodes of an independent set is smaller than summing over a clique cover, the bound based on cliques shaves off a $\log(T)$ factor compared to that in Theorem 1.

In the setting of fixed undirected graphs, Buccapatnam et al. (2014a) present an interesting action elimination algorithm that solves at every round a linear program based on the feedback graph. The authors prove favorable guarantees in terms of the minimum dominating set of the graph. Unfortunately, several difficulties arise when adapting (Buccapatnam et al., 2014a) to our setting of dependent stochastic graphs. From a computational standpoint, we would have to solve a linear program at every round based on the feedback graph’s edge probabilities $p^t_{i,j}$. Moreover, we would need to know the time horizon $T$ in advance (a key step of our analysis of UCB-DSG heavily depends on this property). One might instead estimate the $p_{i,j}$’s in an initial stage where no actions are eliminated and then, say, adopt the strategy in (Li et al., 2020) that works in the case when the stochastic graph is independent of the losses and the $p_{i,j}$ are known. However, even this solution looks problematic in our setting. This is because the duration of this initial stage should be at least $\sqrt{T}$ in order to insure accurate estimation and, if we don’t rely on feedback graphs during this stage we would suffer anyway a regret of the form $\frac{K}{T} \log T$. More importantly, observe that even if we know the marginal probabilities $p_{i,j}$ exactly, we still have to face the loss bias issue intrinsic to our dependency assumptions.

4.1. Sketch of the Pseudo-regret Analysis

Here, we sketch the analysis that proves Theorem 1. Please see Appendix B for the details. At a high level, the proof proceeds via two steps: 1) we prove that $\hat{\mu}_{i,t}$ converges to its mean at a rate of $\tilde{O}(1/\sqrt{T})$ and then bound the regret using a UCB-type analysis; 2) we relate the resulting bound to the independent sets of the process graph, $G_{\tilde{e}(\gamma)}$.

For step 1), we first quantify the accuracy of our loss estimates $\tilde{e}_t(x)$ (Lemma 1 and Lemma 2 in Appendix B). Specifically, Lemma 1 shows that even though $\hat{\mu}_{i,t}$ is a biased estimator, it still concentrates around $\tilde{e}_t(x)$, converging to its expectation at a rate of $O(1/\sqrt{T})$. By using the definition of threshold $\theta_{i,j}$ and then, a regret of the form $\frac{K}{T} \log T$.

The independence number of this graph is $\alpha_G$.
We now show that the intrinsic difficulty of learning with dependent stochastic feedback graphs does not derive from the stochasticity of the graphs (or the lack of prior knowledge of the probabilities $p_{i,j}$) but, rather, from the arbitrariness of the dependency structure of graphs and losses.

**Theorem 2.** For any number of arms $K \geq 2$, any gap value $\Delta \in (0, 1/4)$, any edge probability $p \in [0, 1)$, and any strongly consistent policy\footnote{A strongly consistent policy is one such that $R_T = o(T^\alpha)$ for all $\alpha \in (0, 1)$.} there exists a dependent stochastic feedback graph problem with process graph $G$ with $p_{i,j} = p$ for all $i \neq j$, $p_{i,i} = 1$ for all $i$, and expected losses $\mu_1, \ldots, \mu_K$, with $\mu_i - \mu_\star = (1-p)\Delta$ for all $i \neq \star$, such that $R_T = \Omega(\frac{K}{\Delta} \log T)$.

In the above lower bound, $p$ can be arbitrarily close to 1, but is assumed to be constant (independent of $\Delta$, $K$, and $T$). Notice that this lower bound holds even if the algorithm knows $p$. Yet, this result does not violate the upper bound in Theorem 1, since that theorem refers to the independence number of an undirected graph $G_\theta(\gamma)$ that retains enough edges $(i,j)$ only when $|K^\ast|$ is big, which in turn happens only if the probabilities $p_{i,j}$ defining $K^\ast$ increase to 1 as $T \to \infty$ and since $f(D, K)$ increases as the gaps decrease.

The lower bound also illustrates the stark difference between dependent and independent feedback graph problems. For instance, if $G_t$ is an undirected Erdos-Renyi stochastic graph with probability $p$ of independently generating each edge, then the gap-dependent pseudo-regret upper bound is of the form $\frac{\log K}{p} \log T$, i.e., logarithmic in $K$ rather than linear. This conclusion can be extracted from both (Buccapatnam et al., 2014a) [Remark 5] and (Li et al., 2020) [Theorem 3], and essentially follows from the fact that the independence number of an Erdos-Renyi graph with $K$ nodes and edge probability $p$ is $\approx (2/p) \log K$ (Frieze, 1990).

### 4.2. Lower Bound

We now show that the intrinsic difficulty of learning with dependent stochastic feedback graphs does not derive from the stochasticity of the graphs (or the lack of prior knowledge of the probabilities $p_{i,j}$) but, rather, from the arbitrariness of the dependency structure of graphs and losses.

We present results for the CIFAR dataset, as well as ten datasets from the UCI repository: letter, pendigits, poker, satimage, shuttle, segment, covtype, acoustic, HIGGS and sensorless (see Appendix D for dataset statistics). The features of each of the UCI datasets were scaled to $[-1, 1]$. For the CIFAR dataset, we extracted via PCA the first 25 components and used them as features. The experiment was set up as follows. We randomly draw four times a set of hyperplanes, $\{w_i^1, \ldots, w_i^K : i \in [K]\}$, and for each set of hyperplanes, we randomly shuffle the data six times. Our results are averages over these 24 runs.

For all algorithms, since the constants in front of the log terms are artifacts of the analysis, we introduced a parameter $\lambda \in \{0.1, 0.5, 1, 10, 100\}$, and tuned each algorithm over this parameter as is standard practice. Specifically, the $\lambda$ multiplies the slack terms in the confidence intervals. For UCB-DSG, we tune the algorithm over two $\lambda$s: one for the slack, $S_{i,t}$, and threshold, $\theta_{i,t}$, as they both contain terms of the form $1/\sqrt{Q_{i,t}}$, and one for the estimated loss’ conditions $C_{i,t}$, as it contains a term of the form $1/\sqrt{U_{i,t}}$. For each algorithm, we ran the above experiment for the different values of $\lambda$ and report the results of the value of $\lambda$ that admits the smallest regret at the last round.

Figure 2 (top row) shows the averaged regret results for some of the datasets and Appendix D contains the plots for all datasets. These figures show that UCB-DSG outperforms UCB on all eleven datasets except for two, for which it admits a comparable performance. UCB-DSG attains a performance that in a few cases is even close to FS.
We then compared UCB-DSG against two other baselines that we designed based on the concepts analyzed in this paper, ALG-1 and ALG-2. Both algorithms pick the expert with the smallest lower confidence bound $I_t = \arg\min_{i \in [K]} \hat{\mu}_{i,t-1} - S_{i,t-1}$, but the way they update the experts’ empirical estimates differs.

The first baseline, ALG-1, is in the full information setting and is not relying on feedback graphs. ALG-1 updates the chosen expert by using its true loss, and uses $\ell'_t(\xi_i) = L(\xi_i(x_i), \xi_t(x_i))$ to update all other experts $i \neq I_t$. Note that, whenever $\xi_{I_t}$ admits zero loss, all $\ell'_t(\xi_i) = \ell_t(\xi_i)$; however, ALG-1 could suffer a potentially linear regret, as the algorithm may add noisy labels at each round that corrupt the empirical estimates.

ALG-2 updates any expert that predicts the same way as $\xi_I$, by using the true loss, that is, it updates all $i \in [K]$ that satisfy $\xi_i(x_i) = \xi_I(x_i)$ by using the loss $L(\xi_I(x_i), y_i)$. This algorithm is the natural extension of UCB-N of Caron et al. (2012) to this setting but, again, it may suffer linear regret since it relies on biased empirical estimates. To see why, suppose that ALG-2 selects $\xi_1$, and that $\xi_1(x) = \xi_2(x)$ only on those $x$ such that $\xi_1(x) = \xi_2(x) = m$, for some $m \in [c]$. Then, the empirical estimate for $\xi_2$ may be biased, because it is averaged only over rounds $t$ when $\xi_2(x_t) = m$.

Despite neither of these two algorithms are theoretically motivated, they are natural baselines. In our experiments (see the bottom row of Figure 2 as well as Appendix D), we found that UCB-DSG is in fact outperforming both baselines on all datasets except for three, letter, HIGGS and CIFAR. For letter and CIFAR datasets, UCB, ALG-1, ALG-2 and UCB-DSG all admit a similar performance, which seems to suggest that feedback graphs are not beneficial for these two datasets. For the HIGGS dataset, ALG-1, ALG-2 and UCB-DSG algorithms admit a performance close to FS, thereby indicating that given the current experimental setup, there is no opportunities for improvements for this dataset. In Appendix D, we further discuss how the density of the graphs affects the algorithms’ performance.

Altogether, these experiments show that, in a natural scenario where feedback graph and losses are dependent, our algorithm UCB-DSG achieves a performance that is substantially more favorable than that of readily available baselines.

6. Conclusion

We initiated the analysis of online learning with stochastic feedback graphs in a setting often emerging in applications where graphs and losses are statistically dependent.

Our algorithm estimates the probability of an edge via past realizations of the stochastic graphs and, based on carefully chosen thresholds, uses estimated losses to update its empirical estimates. Our pseudo-regret bound is in terms of thresholded distributional properties of the process generating the feedback graphs. In this setting, our algorithm benefits from strong theoretical guarantees that become more favorable in cases where playing the best action tends to reveal more losses of the other actions. We have complemented our upper bound with a regret lower bound that illustrates the inherent hardness of the general dependent feedback graph setting. Finally, we report a series of experiments showing the favorable empirical performance of UCB-DSG, as compared to the UCB algorithm, as well as to readily available UCB-based variants of UCB-DSG that propagate information across actions in a way not supported by our theory.

It can be shown that, under some assumption, our analysis can be extended to the more general setting where feedback graphs may not admit self-loops at all vertices. This is an important setting that includes naturally active learning scenarios, as discussed in Appendix C.
References


A. Feedback Graphs in Multi-class Classification

Figure 3 and Figure 4 give examples of graphs $G_t$ and $\hat{G}_t$ for the special case of multiclass classification considered in our experiments.

Figure 3. Example of graphs $G_t$ in the case for multiclass classification setting with 3 classes and $K = 4$ experts. If two experts predict the same way at time $t$, then they admit the same loss and hence, there is an edge between them in graph $G_t$.

Figure 4. On the left: An illustrations of graphs $\hat{G}_1, \ldots, \hat{G}_t$ used by UCB-DSG for the setting of multiclass classification with $K = 4$ experts. If the empirical probability that there is an edge in $G_t$ is higher than a carefully chosen threshold, then $\hat{G}_t$ contains this edge. See Figure 3 for the graphs $G_t$ in this setting. On the right: An illustration of graph $\hat{G}_{\theta(\gamma)}$. Graph $\hat{G}_{\theta(\gamma)}$ is similar to $\hat{G}_t$, in that instead of empirical probabilities, we use true probabilities along with thresholds (where in this specific example, the same threshold $\theta$ is used for all edges).
B. Proofs for Pseudo-regret Bounds of UCB-DSG Algorithm

In this appendix, we present the proofs for the pseudo-regret bounds of UCB-DSG algorithm (Theorem 1 in the main body of the paper). The analysis of the UCB-DSG algorithm proceeds as follows. First, we present Proposition 1 that bounds the correction term $\rho_{i,t}$ due to resorting to estimated losses $\tilde{\ell}_s(\xi_i)$ when constructing estimator $\hat{\mu}_{i,t}$:

$$\rho_{i,t} = \frac{1}{Q_{i,t}} \sum_{s=1}^{t} (\ell_s(\xi_i) - \tilde{\ell}_s(\xi_i) \mathbb{I}\{i \in \hat{N}_s(I_s)\} \mathbb{I}\{i \notin N_s(I_s)\}) .$$

To prove Proposition 1, we leverage Lemma 1 and Lemma 2, which quantify the accuracy of our estimators. Then, we use this proposition to show that $\hat{\mu}_{i,t} + \rho_{i,t}$ is an unbiased estimate that concentrates around the mean $\mu_i$. By using the fact that, with high probability, the mean lies inside the confidence interval of this estimate, the pseudo-regret can be bounded by a term that is logarithmic in $T$ and a constant term that does not depend on $T$. We then show in Proposition 2 that the out-neighborhoods of $\hat{G}_t$ are contained in the out-neighborhoods of the thresholded process graph $\hat{G}_{\hat{\theta}(\gamma)}$. To prove Proposition 2, we use Lemma 3 and Lemma 4, which present properties of $U_{i,t}$ and $Q_{i,t}$ as well as a high probability bound on the number of times an expert is pulled. Lastly, in Proposition 3, we derive a bound that depends on the independence sets of $\hat{G}_{\hat{\theta}(\gamma)}$.

In what follows, we denote by $\mathbb{E}_t[\cdot]$ the expectation of the random variable at argument, conditioned on history up to time $t - 1$, that is, including losses $\ell_1(\cdot), \ldots, \ell_{t-1}(\cdot)$, and graphs $G_1, \ldots, G_{t-1}$. We also need a few ancillary definitions. For brevity, we set throughout

$$\theta_{j,t} = \min \left\{ 2 - 2 \frac{U_{j,t} + 1}{Q_{j,t} + 1} - \frac{a_t(\delta_t/4)}{\sqrt{Q_{j,t} + 1}}, 1 - 2 \frac{a_t(\delta_t/4)}{Q_{j,t} + 1} \right\} ,$$

$$a_t(\delta) = 4 \sqrt{\log \left( \frac{2K^2(t+1)(1+\log(4t+1))}{\delta} \right) ,}$$

$$\delta_t = \frac{1}{K^3 \tau} ,$$

$$S_{i,t} = \frac{88 a_t(\delta_t) + 34 + 2 \sqrt{\log(K(t+1))}}{Q_{i,t}} ,$$

$$T_{i,t} = \sum_{s=1}^{t} \mathbb{I}\{I_s = i\} ,$$

$$C_{j,t} = \mathbb{I}\{\hat{p}_{i,t-1}^{j} \geq 1 - \epsilon_{i,t-1}\} \text{ where } \epsilon_{i,t} = \frac{a_t(\delta_t/4) + 1}{\sqrt{U_{i,t} + 1}} .$$

In all proofs that follow culminating into the proof of Theorem 1, the confidence level $\delta$ will be set $\delta_t = \frac{1}{K^3 \tau^2}$ so as to turn high probability statements throughout to expectations statements.

B.1. Bounding the Correction Term $\rho_{i,t}$

The following two lemmas allow us quantify the accuracy of our estimators, which are in turn used to bound the correction term, $\rho_{i,t}$, in Proposition 1. The first lemma shows that $\hat{\mu}_{i,t}$ converges to $\mathbb{E}[\ell(\xi_i) \mathbb{I}\{i \notin N(j)\}]$ at a rate of $1/\sqrt{U_{i,t}}$ and the second lemma uses this fact to show that estimated loss converges to its mean at a similar rate.

**Lemma 1.** For any $\delta > 0$, with probability at least $1 - \delta$,

$$|\hat{\mu}_{i,t} - \mathbb{E}[\ell(\xi_i) \mathbb{I}\{i \notin N(j)\}]| \leq (6.3a_t(\delta) + 3) \sqrt{\frac{1}{U_{i,t}}}$$

holds simultaneously over $i, j \in [K]$, and all $t \geq 1$. 


Proof. The proof follows by giving upper and lower bounds on

\[
\tilde{v}_{i,j,t} U_{i,t} = \sum_{s=1}^{t} \sum_{r=1}^{K} \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} \mathbb{I} \{ I_s = r \} \mathbb{I} \{ \hat{P}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1} \}
\]

in terms of \( \mathbb{E} [ \ell(\xi_i) \mathbb{I} \{ i \not\in N_j(j) \} ] U_{i,t} \).

Recall that, while \( \ell_s(\xi_i) \) and \( N_s \) are statistically dependent (they both depend on graph \( G_s \)), the variables \( I_s, \hat{P}_{r,i}^{s-1}, \ell_s(\xi_i) \), and \( \epsilon_{i,s-1} \) in (2) are determined by the losses and graphs up to \( s - 1 \). Hence, for \( s \in [t] \), the random variable

\[
\sum_{r=1}^{K} \left( \mathbb{E} [ \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} \mathbb{I} \{ i \in N_s(r) \} ] \mathbb{I} \{ I_s = r \} \mathbb{I} \{ \hat{P}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1} \} - \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} \mathbb{I} \{ i \in N_s(r) \} \mathbb{I} \{ I_s = r \} \mathbb{I} \{ \hat{P}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1} \} \right)
\]

has zero conditional mean. Then, via a concentration inequality for martingale difference sequences, specifically the one in (van de Geer, 2002) (Lemma 2.4 and Theorem 2.5 therein) combined with a standard peeling argument, one can see that with probability at least \( 1 - \delta/2 \) over the generation of graphs and losses, we have both

\[
\tilde{v}_{i,j,t} U_{i,t} \leq \sum_{s=1}^{t} \sum_{r=1}^{K} \mathbb{E} [ \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} \mathbb{I} \{ i \in N_s(r) \} ] \mathbb{I} \{ I_s = r \} \mathbb{I} \{ \hat{P}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1} \} + a_t(\delta) \sqrt{1 + U_{i,t}} \tag{3}
\]

and

\[
\tilde{v}_{i,j,t} U_{i,t} \geq \sum_{s=1}^{t} \sum_{r=1}^{K} \mathbb{E} [ \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} \mathbb{I} \{ i \in N_s(r) \} ] \mathbb{I} \{ I_s = r \} \mathbb{I} \{ \hat{P}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1} \} - a_t(\delta) \sqrt{1 + U_{i,t}} \tag{4}
\]

simultaneously over \( i, j \in [K] \), and \( t = 1, 2, \ldots \).

For the upper bound (3), we observe that

\[
\mathbb{E} [ \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} \mathbb{I} \{ i \in N_s(r) \} ] \leq \mathbb{E} [ \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} ] = \mathbb{E} [ \ell(\xi_i) \mathbb{I} \{ i \not\in N(j) \} ] ,
\]

the latter equality deriving from the fact that the process jointly generating graphs and losses is i.i.d. over time. Plugging back into (3), and recalling the definition of \( U_{i,t} \), we can write

\[
\tilde{v}_{i,j,t} U_{i,t} \leq \mathbb{E} [ \ell(\xi_i) \mathbb{I} \{ i \not\in N(j) \} ] U_{i,t} + a_t(\delta) \sqrt{1 + U_{i,t}} .
\]

For the lower bound (4), it holds that

\[
\mathbb{E} [ \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} \mathbb{I} \{ i \in N_s(r) \} ] = \mathbb{E} [ \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} ] - \mathbb{E} [ \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} \mathbb{I} \{ i \not\in N_s(r) \} ] \\
\geq \mathbb{E} [ \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} ] - \mathbb{E} [ \mathbb{I} \{ i \not\in N_s(r) \} ] \\
(\text{since } \ell_s(\xi_i) \mathbb{I} \{ i \not\in N_s(j) \} \leq 1) \\
= \mathbb{E} [ \ell(\xi_i) \mathbb{I} \{ i \not\in N(j) \} ] - (1 - p_{r,i}) \\
(\text{since the process is i.i.d. and, by definition, } \mathbb{E} [ \mathbb{I} \{ i \not\in N_s(r) \} ] = 1 - p_{r,i}) \\
\geq \mathbb{E} [ \ell(\xi_i) \mathbb{I} \{ i \not\in N(j) \} ] - (1 - \hat{P}_{r,i}^{s-1}) - \frac{a_t(\delta)}{\sqrt{s-1}} ,
\]

the last step being a consequence of the standard Hoeffding inequality \( \hat{P}_{r,i}^{s-1} \leq p_{r,i} + \frac{a_{s-1}(\delta)}{\sqrt{s-1}} \) applied to i.i.d. random variables, holding with probability at least \( 1 - \delta/2 \), simultaneously over \( r, i \) and \( s \).
Thus, from (4) we can write

\[ \tilde{\nu}_{i,j,t} U_{i,t} \geq \mathbb{E}[\ell(\xi_i) I\{i \not\in N(j)\}] U_{i,t} - a_t(\delta) \sqrt{1 + U_{i,t}} \]

\[ \quad - \sum_{s=1}^{t} \sum_{r=1}^{K} (1 - \hat{p}_{r,i}^{s-1}) I\{s = r\} I\{\hat{p}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1}\} \]

\[ \quad - \sum_{s=1}^{t} \sum_{r=1}^{K} \sum_{s'=1}^{s} \tilde{\nu}_{i,j,t} I\{s = r\} I\{\hat{p}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1}\} . \]

We then handle in turn the two terms (I) and (II).

Focusing on the first term, it holds that

\[ (I) \leq \sum_{s=1}^{t} \sum_{r=1}^{K} \epsilon_{i,s-1} I\{s = r\} I\{\hat{p}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1}\} \]

\[ = \sum_{s=1}^{t} \sum_{r=1}^{K} \epsilon_{i,s-1} \frac{1}{\sqrt{U_{i,s-1} + 1}} I\{s = r\} I\{\hat{p}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1}\} \]

\[ \leq (a_t(\delta) + 1) \sum_{s=1}^{t} \frac{1}{\sqrt{U_{i,s}}} \sum_{s=1}^{K} I\{s = r\} I\{\hat{p}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1}\} . \]

At this point we use the following inequality (see, for example, (Auer et al., 2002b) [Lemma 3.5]), holding for \( x_1, \ldots, x_t \geq 0 \) with \( 0/\sqrt{0} = 0 \):

\[ \sum_{s=1}^{t} \frac{x_s}{\sqrt{\sum_{s'=1}^{t} x_{s'}}} \leq 2 \sqrt{\sum_{s=1}^{t} x_s} . \] (5)

Setting \( x_s = \sum_{r=1}^{K} I\{s = r\} I\{\hat{p}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1}\} \) therein allows us to conclude that

\[ (I) \leq 2(a_t(\delta) + 1) \sqrt{U_{i,t}} . \]

A bound on the second term follows by a similar argument:

\[ (II) \leq 1 + a_t(\delta) \sum_{s=1}^{t} \sum_{r=1}^{K} \sqrt{\frac{x}{s}} I\{s = r\} I\{\hat{p}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1}\} \]

\[ \leq 1 + a_t(\delta) \sum_{s=1}^{t} \sum_{r=1}^{K} \sqrt{\frac{1}{U_{i,s}}} \sum_{r=1}^{K} I\{s = r\} I\{\hat{p}_{r,i}^{s-1} \geq 1 - \epsilon_{i,s-1}\} \]

\[ \leq 1 + 2a_t(\delta) \sqrt{2U_{i,t}} . \]

Putting together as in (4) yields

\[ \tilde{\nu}_{i,j,t} U_{i,t} \geq \mathbb{E}[\ell(\xi_i) I\{i \not\in N(j)\}] U_{i,t} - a_t(\delta) \sqrt{1 + U_{i,t}} - 2(a_t(\delta) + 1) \sqrt{U_{i,t}} - 2a_t(\delta) \sqrt{2U_{i,t}} - 1 . \]

We further upper bound and divide by \( U_{i,t} \) to obtain the claimed bound on \( \tilde{\nu}_{i,j,t} = \mathbb{E}[\ell(\xi_i) I\{i \not\in N(j)\}] \).

Using the above lemma, we can then find the rate at which the estimated loss, \( \tilde{\ell}_i(\xi) \), converges to its mean since the estimated loss is defined in terms of \( \tilde{\nu}_{i,j,t} \).
Lemma 2. For any $\delta > 0$, with probability at least $1 - \delta$,
\[
|\mathbb{E}_s [\ell_s(\xi_i)I\{i \not\in N_s(I_s)\}] - \tilde{\ell}_s(\xi_i)\mathbb{E}_s [I\{i \not\in N_s(I_s)\}]| \leq 2 \left( a_{s-1}(\delta/3) \sqrt{\frac{1}{s-1}} + b_{s-1}(\delta/3) \sqrt{\frac{1}{U_{i,s-1}}} \right)
\]
holds simultaneously over $i$ and $s$, where $b_{s-1}(\delta) = 6.3a_{s-1}(\delta) + 3$.

Proof. First suppose that $1 - p_{I_s,i} \leq \frac{2a_{s-1}(\delta/3)}{\sqrt{s-1}}$. Since $\ell_s(\xi) \leq 1$ and $\tilde{\ell}_s(\xi) \leq 1$ by definition, we have
\[
|\mathbb{E}_s [\ell_s(\xi_i)I\{i \not\in N_s(I_s)\}] - \tilde{\ell}_s(\xi_i)\mathbb{E}_s [I\{i \not\in N_s(I_s)\}]| \leq \mathbb{E}_s [I\{i \not\in N_s(I_s)\}] = 1 - p_{I_s,i}
\]
which clearly implies the desired bound.

Hence, we continue by assuming $1 - p_{I_s,i} > \frac{2a_{s-1}(\delta/3)}{\sqrt{s-1}}$. We establish upper and lower bounds on
\[
\mathbb{E}_s [\ell_s(\xi_i)I\{i \not\in N_s(I_s)\}] - \tilde{\ell}_s(\xi_i)\mathbb{E}_s [I\{i \not\in N_s(I_s)\}] . \tag{6}
\]
As for the upper bound, whenever round $s$ is such that $\tilde{\ell}_s(\xi) = 1$, we have
\[
(6) \leq \mathbb{E}_s [I\{i \not\in N_s(I_s)\}] - \mathbb{E}_s [I\{i \not\in N_s(I_s)\}] = 0 .
\]
On the other hand, when $\tilde{\ell}_s(\xi) = \frac{\nu_{i,s-1}}{1 - \tilde{p}_{I_s,i}}$ we can use Lemma 1. Specifically, with probability at least $1 - \delta/3$ it holds that
\[
\tilde{\nu}_{i,s-1} \geq \mathbb{E}_s [\ell_s(\xi_i)I\{i \not\in N_s(I_s)\}] - \frac{b_{s-1}(\delta/3)}{\sqrt{U_{i,s-1}}} .
\]
Moreover, the standard Hoeffding concentration inequality also gives
\[
\tilde{\nu}_{I_s,i} \geq p_{I_s,i} - \frac{a_{s-1}(\delta/3)}{\sqrt{s-1}}, \text{ with probability at least } 1 - \frac{2}{3}\delta .
\]
Thus, with the same probability,
\[
\frac{\tilde{\nu}_{i,s-1}}{1 - \tilde{p}_{I_s,i}} \geq \frac{\mathbb{E}_s [\ell_s(\xi_i)I\{i \not\in N_s(I_s)\}] - \frac{b_{s-1}(\delta/3)}{\sqrt{U_{i,s-1}}}}{1 - p_{I_s,i} + \frac{a_{s-1}(\delta/3)}{\sqrt{s-1}}} .
\]
Combining with $\mathbb{E}_s [\ell_s(\xi_i)I\{i \not\in N_s(I_s)\}] \leq p_{I_s,i}$, allows us to bound (6) as follows:
\[
(6) \leq \mathbb{E}_s [\ell_s(\xi_i)I\{i \not\in N_s(I_s)\}] - \frac{\mathbb{E}_s [\ell_s(\xi_i)I\{i \not\in N_s(I_s)\}] - \frac{b_{s-1}(\delta/3)}{\sqrt{U_{i,s-1}}}}{1 - p_{I_s,i} + \frac{a_{s-1}(\delta/3)}{\sqrt{s-1}}}(1 - p_{I_s,i})
\]
\[
= \mathbb{E}_s [\ell_s(\xi_i)I\{i \not\in N_s(I_s)\}] \left[ 1 - p_{I_s,i} + \frac{a_{s-1}(\delta/3)}{\sqrt{s-1}} \right] \cdot \frac{1}{1 - p_{I_s,i} + \frac{a_{s-1}(\delta/3)}{\sqrt{s-1}}}
\]
\[
\leq \frac{a_{s-1}(\delta/3)}{\sqrt{s-1}} + \frac{b_{s-1}(\delta/3)}{\sqrt{U_{i,s-1}}} \left( 1 - p_{I_s,i} \right)
\]
\[
\leq \frac{a_{s-1}(\delta/3)}{\sqrt{s-1}} + \frac{b_{s-1}(\delta/3)}{\sqrt{U_{i,s-1}}} .
\]
As for the lower bound on (6), we observe that, by definition, $\tilde{\ell}_s(\xi_i) \leq \frac{\nu_{i,s-1}}{1 - \tilde{p}_{I_s,i}}$ so that
\[
(6) \geq \mathbb{E}_s [\ell_s(\xi_i)I\{i \not\in N_s(I_s)\}] - \frac{\tilde{\nu}_{i,s-1}}{1 - \tilde{p}_{I_s,i}}(1 - p_{I_s,i}) .
\]
As a consequence, we are ready to prove the proposition we announced at the beginning of this section. Using Lemma 2.4 and Theorem 2.5 in (van de Geer, 2002) (along with peeling) twice, once to the martingale difference $G$ on $R_{s} - s \xi_{i}$, the variable $\hat{\text{DSG}}_{s}$ and $\ell_{s}$, the following inequality holds for the UCB-DSG algorithm:

$$
|\rho_{t,i}| \leq \left( 88 a_{t}(\delta/3) + 34 \right) \sqrt{\frac{1}{Q_{i,t}}}.
$$

**Proof.** The proof proceeds by bounding the following quantity:

$$
Q_{i,t} \left| \rho_{t,i} \right| = \left| \sum_{s=1}^{t} (\ell_{s}(\xi_{i}) - \hat{\ell}_{s}(\xi_{i})) \mathbb{I}\{i \in \tilde{N}_{s}(I_{s})\} \mathbb{I}\{i \notin N_{s}(I_{s})\} \right|.
$$

Putting together the above upper and lower bounds on (6) yields the claimed result.

Armed with the above two lemmas, we are ready to prove the proposition we announced at the beginning of this section.

**Proposition 1.** For any $\delta > 0$, with probability at least $1 - \delta$, for all $t > 0$ and $i \in [K]$, the following inequality holds for the UCB-DSG algorithm:

$$
|\rho_{t,i}| \leq \left( 88 a_{t}(\delta/3) + 34 \right) \sqrt{\frac{1}{Q_{i,t}}}.
$$
Dependent Stochastic Feedback Graphs

we conclude that with probability at least $1 - \delta/3$ over the generation of graphs and losses, Eq. (7) is upper bounded as

$$\tag{7} \Pr \left\{ \sum_{s=1}^{t} I\{ i \in \hat{N}_s(I_s) \} D_s + 2a_t(\delta/3) \sqrt{1 + \sum_{s=1}^{t} I\{ i \in \hat{N}_s(I_s) \}} \right\} ,$$

simultaneously over all $i \in [K]$ and $t$, where

$$D_s = \mathbb{E}_s \left[ \ell_s(\xi_i)I\{ i \not\in N_s(I_s) \} \right] - \mathbb{E}_s \left[ I\{ i \not\in N_s(I_s) \} \right] .$$

We now focus on bounding the first sum in the above inequality and split it according to the possible values of $\theta_{i,s}$. We have

$$\sum_{s=2}^{t} I\{ i \in \hat{N}_s(I_s) \} D_s = \sum_{s=2}^{t} I\{ \theta_{i,s-1} = 1 - 2 \frac{a_{s-1}(\delta/3)}{\sqrt{Q_{i,s-1}} + 1} \} I\{ i \in \hat{N}_s(I_s) \} \sum_{s=2}^{t} I\{ i \in \hat{N}_s(I_s) \} \left(1 - p_{I,s,i} \right) ,$$

and analyze the resulting terms separately. Since $\bar{\ell}_s(\xi_i) \leq 1$ and $\ell_s(\xi_i) \leq 1$, we have

$$D_s \leq \mathbb{E}_s \left[ I\{ i \not\in N_s(I_s) \} \right] = 1 - p_{I,s,i} .$$

Thus,

$$\tag{8} \sum_{s=2}^{t} I\{ \theta_{i,s-1} = 1 - 2 \frac{a_{s-1}(\delta/3)}{\sqrt{Q_{i,s-1}} + 1} \} I\{ i \in \hat{N}_s(I_s) \} \left(1 - p_{I,s,i} \right) ,$$

where $I\{ \theta_{i,s-1} = 1 - 2 \frac{a_{s-1}(\delta/3)}{\sqrt{Q_{i,s-1}} + 1} \} I\{ i \in \hat{N}_s(I_s) \} = I\{ p_{I,s,i}^{-1} \geq 1 - 2 \frac{a_{s-1}(\delta/3)}{\sqrt{Q_{i,s-1}} + 1} \}$ holds by definition of neighborhood $\hat{N}_s(I_s)$. The standard Hoeffding bound gives, with probability at least $1 - \delta/3$, $p_{I,s,i}^{-1} \leq p_{I,s,i} + \frac{a_{s-1}(\delta/3)}{\sqrt{Q_{i,s-1}} + 1} \leq 1 - p_{I,s,i} ,$

and we can bound (8) as follows:

$$\tag{8} \sum_{s=2}^{t} I\{ i \in \hat{N}_s(I_s) \} \left( \frac{a_{s-1}(\delta/3)}{\sqrt{s - 1}} + 2 \frac{a_{s-1}(\delta/3)}{\sqrt{Q_{i,s-1}} + 1} \right) \left(1 - p_{I,s,i} \right) ,$$

where in the last step we used again (5).
Next, we analyze (9). For brevity, we set below $\theta_{t,s-1}' = 2 - 2\sqrt{\frac{U_{i,s-1} + 1}{Q_{i,s-1} + 1}} - \frac{a_{s-1}(\delta/3)}{\sqrt{Q_{i,s-1} + 1}}$. From Lemma 2 we have $D_s \leq \frac{2a_{s-1}(\delta/3)}{\sqrt{Q_{i,s-1} + 1}} + \frac{2b_{s-1}(\delta/3)}{\sqrt{U_{i,s-1} + 1}}$, and since $\ell_s(\xi_i) \leq 1$ and $\ell_s(\xi_i) \leq 1$, it also holds that $D_s \leq 1$. Thus we can write

$$(9) \leq \sum_{s=2}^{t} \mathbb{I}\{\theta_{i,s-1}' = \theta_{t,s-1}'\} \mathbb{I}\{i \in \tilde{N}_s(I_s)\} \min\left\{\frac{2a_{s-1}(\delta/3)}{\sqrt{s-1}} + \frac{2b_{s-1}(\delta/3)}{\sqrt{U_{i,s-1} + 1}}, 1\right\}.$$

By a similar reasoning as in the previous case, it holds that

$$\sum_{s=2}^{t} \mathbb{I}\{\theta_{i,s-1}' = \theta_{t,s-1}'\} \mathbb{I}\{i \in \tilde{N}_s(I_s)\} \frac{2a_{s-1}(\delta/3)}{\sqrt{s-1}} \leq 4a_t(\delta/3) \sqrt{Q_{i,t}},$$

so we are left to bound

$$\sum_{s=2}^{t} \mathbb{I}\{\theta_{i,s-1}' = \theta_{t,s-1}'\} \mathbb{I}\{i \in \tilde{N}_s(I_s)\} \min\left\{\frac{2b_{s-1}(\delta/3)}{\sqrt{U_{i,s-1} + 1}}, 1\right\}.$$

Since, by definition $\hat{p}_{i,t,s} \leq 1$, we have

$$\mathbb{I}\{\theta_{i,s-1}' = \theta_{t,s-1}'\} \mathbb{I}\{i \in \tilde{N}_s(I_s)\} = \mathbb{I}\left\{\hat{p}_{i,s}^{-1} \geq \theta_{t,s-1}'\right\} \leq \max\{1 - \theta_{t,s-1}' + \hat{p}_{i,s}^{-2}, 0\} \leq 2 \sqrt{\frac{U_{i,s-1} + 1}{Q_{i,s-1} + 1}} + \frac{a_{s-1}(\delta/3)}{Q_{i,s-1} + 1}.$$

Using this fact along with some algebra,

$$(10) \leq \sum_{s=2}^{t} \mathbb{I}\{i \in \tilde{N}_s(I_s)\} \left(2 \sqrt{\frac{U_{i,s-1} + 1}{Q_{i,s-1} + 1}} + \frac{a_{s-1}(\delta/3)}{Q_{i,s-1} + 1}\right) \min\left\{\frac{2b_{s-1}(\delta/3)}{\sqrt{U_{i,s-1} + 1}}, 1\right\} \leq 4 \sum_{s=2}^{t} \mathbb{I}\{i \in \tilde{N}_s(I_s)\} \frac{U_{i,s-1} + 1}{Q_{i,s-1} + 1} \frac{b_{s-1}(\delta/3)}{\sqrt{U_{i,s-1} + 1}} + \frac{t}{s=2} \mathbb{I}\{i \in \tilde{N}_s(I_s)\} \frac{a_{s-1}(\delta/3)}{Q_{i,s-1} + 1} \leq 4\sqrt{2b_t(\delta/3)} \sum_{s=2}^{t} \mathbb{I}\{i \in \tilde{N}_s(I_s)\} \frac{1}{Q_{i,s}} + a_t(\delta/3) \sum_{s=2}^{t} \mathbb{I}\{i \in \tilde{N}_s(I_s)\} \frac{1}{Q_{i,s}} \leq (8\sqrt{2b_t(\delta/3)} + 2a_t(\delta/3)) \sqrt{Q_{i,t}},$$

where we used the simple facts $\sqrt{\frac{U_{i,s-1} + 1}{U_{i,t-1}}} \leq \sqrt{2}$, $Q_{i,s} \leq Q_{i,s-1} + 1$, and again Inequality (5). Thus,

$$(9) \leq 4a_t(\delta/3) \sqrt{2Q_{i,t}} + (8\sqrt{2b_t(\delta/3)} + 2a_t(\delta/3)) \sqrt{Q_{i,t}}.$$

Combining the bounds for (8) and (9), plugging back, and dividing by $Q_{i,t}$ gives the claimed result.

\[\Box\]

### B.2. Connection to the Thresholded Process Graph $G_{\hat{\delta}(\gamma)}$

The next two lemmas, which show properties of the number of times an arm is updated and chosen, are used to prove Proposition 2, which connects the sequence of graphs $\tilde{G}_t$ to the thresholded process graph $G_{\hat{\delta}(\gamma)}$. 
Lemma 3. For all $t$, with probability at least $1 - 2/t^3$, it holds that

$$
\sum_{j \neq i^*} T_{j,t} \leq \sum_{j \neq i^*} \left\lceil \frac{4d_{t-1}(\delta_{t-1})^2}{\Delta_j^2} \right\rceil,
$$

where $d_i(\delta)$ and $\delta_t$ are as in Proposition 2.

Proof. A straightforward adaptation of Theorem 2 by Audibert et al. (2009) (Eq. (12) therein) to our setting reads as follows: for any $\tau \in \mathbb{R}$ and any integer $u > 1$, it holds that

$$
P(T_{j,t} \geq \tau) \leq \sum_{s=u+1}^{2} \mathbb{P} \left( \frac{d_{s-1}(\delta_{s-1})}{\sqrt{u}} < \tau \right) + \mathbb{P} \left( \exists 1 \leq s \leq t - u \text{ s.t. } \mu_{i,s} - d_{u+s-1}(\delta_{u+s-1}) \geq \tau \right),
$$

where $\mu_{j,u,s}$ denotes the average empirical estimate of expert $j$ that was updated $u$ times up to time $s - 1$ by playing that expert.

Then, by setting $\tau = \mu_i - 2d_{t-1}(\delta_{t-1})/\sqrt{u}$ and $u = \frac{4d_{t-1}(\delta_{t-1})^2}{\Delta_i^2}$ and by an argument which is similar to the one culminating in the bounds of (18) and (19) in the proof of Theorem 1 where $Q_{t,i}$ is replaced by $T_{t,i}$ respectively, it follows that the right hand side of (11) is less than $\frac{2}{t^3}$. A union bound over $j$ completes the proof.

Lemma 4. For all $t \geq 1$ and all $i \in [K]$ it holds that $Q_{i,t} + 1 \geq U_{i,t}$.

Proof. The proof proceeds by induction. For the base case, recalling that we set $U_{i,0} = 1$, $Q_{i,0} = 1$ in the algorithm, it holds that at time $t = 1$, $U_{i,1} = 2$ since $1 - \frac{a_i(\delta_i/4) + 1}{\sqrt{2}} \leq 0 \leq \hat{p}_{i,1}$ and that $Q_{i,1} = 2$ since $\sqrt{2} - 2\frac{a_i(\delta_i/4)}{\sqrt{2}} \leq 0 \leq \hat{p}_{i,1}$. Thus it holds that $Q_{i,t+1} + 1 \geq U_{i,t}$.

By induction hypothesis, suppose that $Q_{i,t-1} + 1 \geq U_{i,t-1}$ holds. In the case that $Q_{i,t-1} + 1 > U_{i,t-1}$, then it follow directly that $Q_{i,t} + 1 \geq U_{i,t}$ since $U_{i,t}$ can increase by at most 1 and even if $Q_{i,t-1}$ remains the same, the inequality stills holds. In the case that $Q_{i,t-1} + 1 = U_{i,t-1}$, then at time $t$, it must be that $Q_{i,t} = Q_{i,t-1} + 1$. To see this, recall that $Q_{i,t} = Q_{i,t-1} + 1$ only if $\hat{p}_{i,t}^{-1} \geq \theta_{i,t-1}$ and since $0 \geq 2 - 2\sqrt{\frac{Q_{i,t-1} + 1}{Q_{i,t-1}}} - \frac{a_i(\delta_{i,t-1}/4)}{\sqrt{Q_{i,t-1} + 1}} = 2 - 2\sqrt{\frac{U_{i,t-1} + 1}{Q_{i,t-1} + 1}} - \frac{a_i(\delta_{i,t-1}/4)}{\sqrt{Q_{i,t-1} + 1}} \geq \theta_{i,t-1}$ by definition of the threshold, the condition $\hat{p}_{i,t}^{-1} \geq \theta_{i,t-1}$ is satisfied. The fact that $Q_{i,t} = Q_{i,t-1} + 1$ in turn implies that $Q_{i,t} + 1 \geq U_{i,t}$ since $U_{i,t}$ can increase by at most 1.

Proposition 2. Let $K^* = \left\{ j \in [K] : p_{i,j} \geq 1 - \left( \frac{4d_{t}(\delta_{t})^2}{\Delta_j^2} + 2 \right)^{-1/2} \right\}$ where for all $t \in [T]$ and $i \in (0,1]$, $d_i(\delta) = 88a_i(\delta) + 34 + \sqrt{2} \log(K(t+1))$ and where $\delta_t = \frac{1}{KT^2}$. For $\gamma \in (0,1)$, let $\theta(\gamma)(\gamma, \ldots, \theta_K(\gamma))$ be defined as

$$
\theta_i(\gamma) = \begin{cases} 
1 - \gamma, & \text{for } i \in K^*, \\
1 - \frac{a_i(\delta_t/4)}{\sqrt{T}}, & \text{for } i \notin K^*,
\end{cases}
$$

set $Q_{i,t}(\gamma) = \sum_{s=1}^{t} \mathbb{I}\{p_{i,s} \geq \theta_i(\gamma)\}$. Then, the following holds

$$
\sum_{t=1}^{T} \sum_{i=1}^{K} \Delta_i \mathbb{E} \left[ \mathbb{I}\{I_t = i\} \mathbb{I}\{Q_{i,t-1} \leq \left\lceil \frac{4d_{t}(\delta_t)^2}{\Delta_j^2} \right\rceil\} \right] = O \left( \min_{\gamma \in (0,1)} \left( \sum_{t=1}^{T} \sum_{i=1}^{K} \Delta_i \mathbb{E} \left[ \mathbb{I}\{I_t = i\} \mathbb{I}\{Q_{i,t-1}^0 \leq \left\lceil \frac{4d_{t}(\delta_t)^2}{\Delta_j^2} \right\rceil\} \right] + \frac{f(D,K)}{1 - (1 + \gamma)^2/4} \right) \right),
$$

where $f(D,K) = \log\left( \sum_{j \neq i^*} \frac{1}{\Delta_j} \right) \sum_{j \neq i^*} \frac{1}{\Delta_j}$.

Proof. Coarsely speaking, we want to prove that $Q_{i,t}(\gamma) \leq Q_{i,t-1}$, as this directly implies the bound of the proposition. Equivalently, our goal will be to prove that if $p_{j,t} \geq \theta_i(\gamma)$, then $\hat{p}_{j,t}^{-1} \geq \theta_{i,t-1}$ for all $j \in [K]$. In order to achieve this,
we split the proof into two steps. In the first step, we show that with a carefully chosen \( \hat{\theta}_{i,t-1} \), if \( p_{j,i} \geq \hat{\theta}_{i,t-1} \), then \( p_{j,i}^{*-1} \geq \theta_{i,t-1} \). Specifically in this step, we replace the estimated probabilities \( \hat{p}_{i,j}^{*-1} \) by exact probabilities \( p_{i,j} \), and find a threshold \( \hat{\theta}_{i,t-1} \) that is independent of any algorithmic quantities. In the second step, we show that if \( p_{j,i} \geq \hat{\theta}_{i,t-1} \), then \( p_{j,i}^{*-1} \geq \theta_{i,t-1} \), that for each \( i \in [K] \), there exist a threshold that is independent of time \( t \) bounding \( \theta_{i,t-1} \). By combining these two steps, we attain the desired bound.

For the first step, we want to show that if \( p_{j,i} \geq \hat{\theta}_{i,t-1} \), then with high probability \( p_{j,i}^{*-1} \geq \theta_{i,t-1} \). By Hoeffding inequality, when \( t \geq 2 \), with probability at least \( 1 - 4/4 \) we have \( p_{j,i}^{*-1} - p_{j,i} \leq \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{t}} \). Since \( \delta_t \) is of the form \( 1/K_t^3 \), the low probability event gives a constant contribution the overall expectation. For the high probability event, if edge \( j \rightarrow i \) is such that \( p_{j,i} \geq \hat{\theta}_{i,t-1} \), then \( p_{j,i}^{*-1} \geq \hat{\theta}_{i,t-1} - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{t}} \). Thus, in order to prove that \( p_{j,i}^{*-1} \geq \theta_{i,t-1} \), it suffices to show that \( \hat{\theta}_{i,t-1} - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{t}} \geq \theta_{i,t-1} \).

As we will see, the threshold \( \hat{\theta}_{i,t-1} \) will be of the form \( \hat{\theta}_{i,t-1} = \min\{1 - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{t}}, c_{i,t-1}\} \) where \( c_{i,t-1} \) is defined below. If \( \hat{\theta}_{i,t-1} = 1 - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{t}} \), then \( \hat{\theta}_{i,t-1} - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{t}} = 1 - 2 \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{t}} \geq 1 - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{t}} + 1 \geq \theta_{i,t-1} \) and hence it follows that \( p_{j,i}^{*-1} \geq \theta_{i,t-1} \). Thus, for the rest of the proof, we will focus on the case when \( \hat{\theta}_{i,t-1} = c_{i,t-1} \).

We proceed by splitting into three cases: \( i \notin K^* \), \( i = i^* \), and \( i \in K^* \setminus \{i^*\} \).

- \( i \notin K^* \). We define \( \hat{\theta}_{i,t-1} = \min\{1 - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{t}}, 2 - 2\sqrt{\frac{1}{t}}\} \).

Since \( \hat{\theta}_{i,t-1} = 2 - 2\sqrt{\frac{1}{t}} \), we have
\[
\hat{\theta}_{i,t-1} - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{t}} = 2 - (2 + a_{t-1}(\delta_{t-1}/4))\sqrt{\frac{1}{t}} \geq 2 - 2\sqrt{\frac{U_{i,t-1} + 1}{Q_{i,t-1} + 1}} \geq \theta_{i,t-1}.
\]

as, by definition, \( \frac{U_{i,t-1} + 1}{Q_{i,t-1} + 1} \geq \frac{1}{t} \) and \( \frac{1}{Q_{i,t-1} + 1} \geq \frac{1}{t} \).

- \( i = i^* \). Since we want an upper bound on \( \theta_{i^*,t-1} \), we seek a lower bound on \( \sqrt{\frac{U_{i^*,t-1} + 1}{Q_{i^*,t-1} + 1}} \) and \( \sqrt{\frac{1}{Q_{i^*,t-1} + 1}} \). By definition of \( U_{i^*,t-1} \), it holds that \( U_{i^*,t-1} \geq T_{i^*,t-1} \) and hence,
\[
\frac{U_{i^*,t-1} + 1}{Q_{i^*,t-1} + 1} \geq \frac{T_{i^*,t-1} + 1}{t} = 1 - \sum_{j \not= i^*} T_{j,t-1} \frac{t}{t} = 1 - \sum_{j \not= i^*} T_{j,t-1}.
\]

By Lemma 3, for all \( t \), with probability at least \( 1 - 2/(t - 1)^3 \), it holds that
\[
\sum_{j \not= i^*} T_{j,t-1} \leq \sum_{j \not= i^*} \left[ \frac{4d_{j-1}(\delta_{t-1})^2}{\Delta_j^2} \right].
\]

Thus, by letting
\[
\hat{\theta}_{i^*,t-1} = \min\left(1 - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{t}}, 2 - 2\sqrt{\left(1 - \frac{\sum_{j \not= i^*} \left[ \frac{4d_{j-1}(\delta_{t-1})^2}{\Delta_j^2} \right]}{t}\right)}\right),
\]

and following a similar reasoning as in the previous case yields \( \hat{\theta}_{i^*,t-1} \geq \theta_{i^*,t-1} \) and hence \( p_{j,i^*}^{*-1} \geq \theta_{i^*,t-1} \).
Dependent Stochastic Feedback Graphs

- $i \in K^* - i^*$. We want to show that $U_{i,t-1} \geq T_{i^*,t-1}$ as this implies that

$$\frac{U_{i,t-1} + 1}{Q_{i,t-1} + 1} \geq \frac{T_{i^*,t-1} + 1}{Q_{i^*,t-1} + 1},$$

and from here, by the same reasoning as the case $i = i^*$ via Eq. (12), we let

$$\hat{\theta}_{i,t-1} = \min \left(1 - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{T}}, 2 - 2 \left(1 - \frac{\sum_{j \neq i^*} \left[\frac{4d_{i^*}(\delta_{t-1})^2}{\Delta_i^2}\right]}{t}\right)^{1/2}\right)$$

in order to conclude that $\hat{\theta}_{i,t-1} - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{T}} \geq \theta_{i,t-1}$ for all $i \in [K]$. In order to show that $U_{i,t-1} \geq T_{i^*,t-1}$, we will use the definition of $K^*$ to prove that whenever arm $i^*$ is pulled, $U_{i,t-1}$ increases by 1. Equivalently, by definition of $U_{i,t-1}$, we need to show that $\hat{p}_{i^*,i,t-1} \geq 1 - \epsilon_{i,t-1}$. By definition of $K^*$, it holds that $p_{i^*,i} \geq 1 - \left(\frac{4d_{i^*}^2(\delta_{t-1})^2}{\Delta_i^2}\right)^{-1/2}$.

Then with probability at least $1 - \delta_{t-1}/4$,

$$\hat{p}_{i^*,i,t-1} \geq 1 - \left(\frac{4d_{i^*}^2(\delta_{t-1})^2}{\Delta_i^2}\right)^{1/2} - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{T}}$$

(13)

also holds by the standard Hoeffding concentration inequality. Since we are analyzing $t$ such that $Q_{i,t-1} \leq \left[\frac{4d_{i^*}^2(\delta_{t-1})^2}{\Delta_i^2}\right]$ and since by Lemma 4, $U_{i,t-1} \leq Q_{i,t-1} + 1$, it follows that

$$1 - \epsilon_{i,t-1} = 1 - \frac{1 + a_{t-1}(\delta_{t-1}/4)}{\sqrt{U_{i,t-1} + 1}}\right)$$

$$\leq 1 - \left(\frac{4d_{i^*}^2(\delta_{t-1})^2}{\Delta_i^2}\right)^{-1/2} - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{U_{i,t-1} + 1}}\right)$$

$$\leq 1 - \left(\frac{4d_{i^*}^2(\delta_{t-1})^2}{\Delta_i^2}\right)^{-1/2} - \frac{a_{t-1}(\delta_{t-1}/4)}{\sqrt{T}}\right).$$

Combining the above inequality with (13), we conclude that $\hat{p}_{i^*,i,t-1} \geq 1 - \epsilon_{i,t-1}$, which directly implies that $U_{i,t-1} \geq T_{i^*,t-1}$.

For the second step of the proof, in order to show that if $p_{j,i} \geq \theta_i(\gamma)$, then $p_{j,i} \geq \hat{\theta}_{i,t-1}$, we find a threshold such that $\theta_i(\gamma) \geq \hat{\theta}_{i,t-1}$ for each $i \in [K]$. For $i \in K^*$, by choosing $\theta_i(\gamma) = \min \left(1 - \frac{2\tau_{i-1}(\delta_{t-1}/4)}{\sqrt{\Delta_i}}, 2 - 2\sqrt{T/\Delta_i}\right) = 1 - \frac{2\tau_{i-1}(\delta_{t-1}/4)}{\sqrt{\Delta_i}}$ for $T > 2$, it holds that $\theta_i(\gamma) \geq \hat{\theta}_{i,t-1}$. For $i \in K^*$, we find $\gamma \in [0, 1)$ such that

$$2 \left(1 - \frac{\sum_{j \neq i^*} \left[\frac{4d_{i^*}(\delta_{t-1})^2}{\Delta_i^2}\right]}{t}\right)^{1/2} \geq 1 + \gamma.$$

The above holds whenever $t \geq t_0 = \Omega\left(\frac{\sum_{i \in K^*} \Delta_i^{-2} \log(\sum_{j \neq i^*} \Delta_i^{-2})}{4 + (1 + \gamma)^2/4}\right)$ and thus, over these time steps, $\theta_i(\gamma) = 1 - \gamma \geq \hat{\theta}_{i,t-1}$ for $i \in K^*$. Hence,

$$\sum_{i=1}^{K} \sum_{t=1}^{T} \Delta_i E[I_t = i] I\{Q_{i,t-1} \leq s_i\}$$

$$= \sum_{i \leq \tau_0} \sum_{t=1}^{K} \Delta_i E[I_t = i] I\{Q_{i,t-1} \leq s_i\} + \sum_{t > \tau_0} \sum_{i=1}^{K} \Delta_i E[I_t = i] I\{Q_{i,t-1} \leq s_i\}$$

$$= O\left(\min_{\gamma \in [0, 1]} \frac{1}{1 - (1 + \gamma)^2/4} \log \left(\sum_{j \neq i^*} \frac{1}{\Delta_j^2}\right) + \sum_{t=1}^{T} \sum_{i=1}^{K} \Delta_i E\left[I_t = i \right] I\{Q_{i,t-1} \leq s_i\}\right).$$
and this concludes the proof.

\[ \square \]

### B.3. Connection to Independent Sets

The following proposition connects the expert observations from thresholded process graph \( G_{\bar{\theta}(\gamma)} \) to its independent sets by directly using ideas from (Lykouris et al., 2020).

**Proposition 3.** With the same notation as in Proposition 2, let \( G_{\bar{\theta}(\gamma)} = \{ E, F \} \) be the undirected graph where each vertex corresponds to an expert, and where there is an edge from vertex \( j \) to vertex \( i \) iff both \( p_{j,i} \geq \theta_i(\gamma) \) and \( p_{i,j} \geq \theta_j(\gamma) \) hold. Then the following holds

\[
\sum_{t=1}^{T} \sum_{i=1}^{K} \Delta_i \mathbb{E} \left[ \mathbb{I}\{ I_t = i \} \mathbb{I}\left\{ Q^\theta_{i,t-1} \leq \frac{4d_T(\delta_T)^2}{\Delta_i} \right\} \right] \leq O \left( \ln(KT) \log(T) \max_{I \in \mathcal{I}(G_{\bar{\theta}(\gamma)})} \sum_{j \in I} \frac{1}{\Delta_j} \right).
\]

**Proof.** We will follow an analysis of (Lykouris et al., 2020), which holds for undirected graphs and thus, we consider the undirected graph, \( G_{\bar{\theta}(\gamma)} \). We let \( \tilde{Q}^\theta_{i,t-1} \) be the observations defined by graph \( G_{\bar{\theta}(\gamma)} \), which implies that by definition, \( \tilde{Q}^\theta_{i,t-1} \leq Q^\theta_{i,t-1} \) and thus,

\[
\sum_{t=1}^{T} \sum_{i=1}^{K} \Delta_i \mathbb{E} \left[ \mathbb{I}\{ I_t = i \} \mathbb{I}\left\{ \tilde{Q}^\theta_{i,t-1} \leq \frac{4d_T(\delta_T)^2}{\Delta_i} \right\} \right] \leq \sum_{t=1}^{T} \sum_{i=1}^{K} \Delta_i \mathbb{E} \left[ \mathbb{I}\{ I_t = i \} \mathbb{I}\left\{ Q^\theta_{i,t-1} \leq \frac{4d_T(\delta_T)^2}{\Delta_i} \right\} \right].
\]

The proof proceeds by following the analysis from (Lykouris et al., 2020), that is, we divide the run of the algorithm into layers, but the definition of the layers is based on the loss observability induced by the graph \( G_{\bar{\theta}(\gamma)} \). Specifically, when an arm \( i \) is pulled and when \( \tilde{Q}^\theta_{i,t-1} \leq \frac{4d_T(\delta_T)^2}{\Delta_i} \), arm \( i \) along with all its neighbors is placed into the earliest layer \( l \) in which arm \( i \) has not yet been observed according to the graph \( G_{\bar{\theta}(\gamma)} \). Since we are considering the times \( t \) such that \( \tilde{Q}^\theta_{i,t-1} \leq \frac{4d_T(\delta_T)^2}{\Delta_i} \), any arm where \( \tilde{Q}^\theta_{i,t-1} > \frac{4d_T(\delta_T)^2}{\Delta_i} \) can be eliminated. This implies that any layers \( l \geq \frac{4d_T(\delta_T)^2}{\Delta_i} \) cannot contain arm \( i \) as a selected arm. The rest of analysis follows directly from Theorem 3.2 in (Lykouris et al., 2020).

\[ \square \]

### B.4. Proof of Theorem 1

Leveraging Proposition 1, Proposition 2, and Proposition 3, we can now prove Theorem 1, which bounds the regret of UCB-DSG in terms of the graph \( G_{\bar{\theta}(\gamma)} \).

**Theorem 1.** Let

\[
K^* = \{ j \in [K] : p_{i^*,j} \geq 1 - O \left( \frac{\log(KT)}{\Delta_j} + 1 \right)^{-1/2} \}
\]

and for \( \gamma \in [0, 1) \), let \( \bar{\theta}(\gamma) = (\theta_1(\gamma), \ldots, \theta_K(\gamma)) \) be defined as

\[
\theta_i(\gamma) = \begin{cases} 1 - \gamma & \text{for } i \in K^* \\ 1 - \frac{O(\log(KT))}{\sqrt{T}} & \text{for } i \notin K^* \end{cases}.
\]

The pseudo-regret \( R_T \) of UCB-DSG over \( T \) rounds is bounded by:

\[
O \left( \min_{\gamma \in [0, 1)} \left( \max_{I \in \mathcal{I}(G_{\bar{\theta}(\gamma)})} \sum_{j \in I} \frac{\ln(KT) \log(T)}{\Delta_j} + \frac{f(D, K)}{1 - (1 + \gamma)^2/T} \right) \right).
\]

In the above, \( f(D, K) \) is a function linear in the number of experts \( K \) and inversely proportional to each gap \( \Delta \in D = \{ \Delta_j : j \in [K], j \neq i^* \} \), but independent of \( T \).
Proof. Let index $*$ be a short-hand for $\arg\min_{i \in [K]} \mu_i$. The pseudo-regret of the algorithm can be written as:

$$E \left[ \sum_{t=1}^{T} \ell_t(\xi_{I_t}) - \ell_t(\xi_*) \right] = \sum_{t=1}^{T} E[\ell_t(\xi_{I_t}) - \ell_t(\xi_*)]$$

$$= \sum_{t=1}^{T} E \left[ E_t[\ell_t(\xi_{I_t}) - \ell_t(\xi_*)] \right]$$

$$= \sum_{t=1}^{T} E \left[ E_t \left[ \sum_{i=1}^{K} \mathbb{I}\{I_t = i\} \left( \ell_t(\xi_{i}) - \ell_t(\xi_*) \right) \right] \right]$$

$$= \sum_{i=1}^{K} \sum_{t=1}^{T} E[\ell_t(\xi_{i}) - \ell_t(\xi_*) \mathbb{I}\{I_t = i\}]$$

$$= \sum_{i=1}^{K} (\mu_i - \mu_*) E[\mathbb{I}\{I_t = i\}]$$

$$= \sum_{i=1}^{K} (\mu_i - \mu_*) E \left[ \sum_{t=1}^{T} \mathbb{I}\{I_t = i\} \right].$$  \hspace{1cm} (14)

We focus on bounding the term $E \left[ \sum_{t=1}^{T} \mathbb{I}\{I_t = i\} \right]$ for each $i$. For any fixed $s_t$, the events $Q_{s_t-1} > s_t$ and $Q_{s_t-1} \leq s_t$ can be used to split the expectation:

$$E \left[ \sum_{t=1}^{T} \mathbb{I}\{I_t = i\} \right] = E \left[ \sum_{t=1}^{T} \mathbb{I}\{I_t = i\} \mathbb{I}\{Q_{s_t-1} < s_t\} \right] + E \left[ \sum_{t=1}^{T} \mathbb{I}\{I_t = i\} \mathbb{I}\{Q_{s_t-1} \geq s_t\} \right].  \hspace{1cm} (15)$$

Since the chosen expert $I_t = i$ equals arm $i$ only if the upper confidence bound of $i$ is smaller than that of any other expert, the last sum in (15) can be bounded as follows:

$$E \left[ \sum_{t=1}^{T} \mathbb{I}\{I_t = i\} \mathbb{I}\{Q_{s_t-1} \geq s_t\} \right] = \sum_{t=1}^{T} P(I_t = i, Q_{s_t-1} \geq s_t)$$

$$\leq \sum_{t=1}^{T} P \left( \mu_{i,t-1} - S_{1,t-1} \leq \mu_{s,t-1} - S_{s,t-1}, Q_{s_t-1} \geq s_t \right).  \hspace{1cm} (16)$$

The empirical estimate $\hat{\mu}_{i,t-1}$ includes the estimated losses, thus we must use the correction term from Proposition 1, where we set $\delta = \frac{1}{\sqrt{Kt}}$, therein, to turn it into an empirical estimate that concentrates around its mean. More concretely, for any given round $s < t$, and expert $\xi_i$ we can write

$$\bar{\ell}_s(\xi_i) \mathbb{I}\{i \in \hat{N}_s(I_s)\} + (\ell_s(\xi_i) - \bar{\ell}_s(\xi_i)) \mathbb{I}\{i \in \hat{N}_s(I_s)\} \mathbb{I}\{i \notin N_s(I_s)\}$$

$$= \ell_s(\xi_i) \mathbb{I}\{i \in \hat{N}_s(I_s)\} \mathbb{I}\{i \notin N_s(I_s)\} + \ell_s(\xi_i) \mathbb{I}\{i \in \hat{N}_s(I_s)\} \mathbb{I}\{i \notin N_s(I_s)\}$$

$$= \ell_s(\xi_i) \mathbb{I}\{i \in \hat{N}_s(I_s)\} \mathbb{I}\{i \notin N_s(I_s)\}$$

$$= \ell_s(\xi_i) \mathbb{I}\{i \in \hat{N}_s(I_s)\} ,$$

so that summing over $s = 1, \ldots, t - 1$, and dividing by $Q_{s_t-1}$ yields

$$\hat{\mu}_{i,t-1} + \rho_{i,t-1} = \frac{1}{Q_{s_t-1}} \sum_{s=1}^{t-1} \ell_s(\xi_i) \mathbb{I}\{i \in \hat{N}_s(I_s)\} : = \bar{\mu}_{i,t-1}.$$  

Then, the estimate $\bar{\mu}_{i,t-1}$ is an average of i.i.d. realizations of the random variable $\ell(\xi_i)$ because the out-neighborhood $\hat{N}_s(I_s)$ of the chosen expert $I_s$ on graph $\hat{G}_s$ only depends on previous observations. In what follows, we use Proposition 1,
where we set $\delta = \frac{1}{Kt^3}$, so that the extra pseudo-regret deriving from this proposition being false only contributes a constant term overall.

Turning back to Equation (16), we introduce the terms $\mu_s, \mu_i, \rho_s, t-1, \rho_i, t-1$, and reorder as follows:

$$\tilde{\mu}_{i,t-1} - S_{i,t-1} \leq \mu_* - \rho_s, t-1 - S_{s,t-1}$$

which implies

$$\begin{array}{c}
0 \leq \tilde{\mu}_{s, t-1} + \rho_s, t-1 - \rho_s, t-1 - S_{s,t-1} - \tilde{\mu}_{i, t-1} - \rho_i, t-1 + \rho_i, t-1 + S_{i,t-1} \\

0 \leq (\tilde{\mu}_{s, t-1} - \mu_* - \rho_s, t-1 - S_{s,t-1}) + (\mu_i - \tilde{\mu}_i, t-1 + \rho_i, t-1 + S_{i,t-1} - 2S_{i,t-1}) \\

+ (\mu_* - \mu_i + 2S_{i,t-1}) .
\end{array}$$

(17)

Then, by setting $s_i = \left[ \frac{4d_i^2(\delta_T)}{\Delta_s} \right]$ where for all $t \in [T]$ and $\delta \in (0, 1]$, $d_t(\delta) = 88a_t(\delta) + 34 + \sqrt{2 \log(K(t + 1))}$ and where $\delta_T = \frac{1}{Kt^3}$, we ensure that the event $Q_{i, t-1} > s_i$ implies that the term $\mu_* - \mu_i + 2S_{i,t-1}$ in (17) is negative. In turn, this implies that either the first or the second term of (17) must be non-negative since, otherwise, the sum of the three terms therein cannot be non-negative. We then show that the first two terms are non-negative with an extremely low probability. More concretely, we have

$$\begin{align}
(16) & \leq \sum_{t=1}^{T} \mathbb{P} \left( \tilde{\mu}_{s, t-1} - \mu_* - \rho_s, t-1 - S_{s,t-1} \geq 0 \right) \\
& + \sum_{t=1}^{T} \mathbb{P} \left( \mu_i - \tilde{\mu}_i, t-1 + \rho_i, t-1 + S_{i,t-1} - 2S_{i,t-1} \geq 0 \right) .
\end{align}$$

Using Proposition 1, the probabilities in the first sum are upper bounded as follows:

$$\begin{align}
\mathbb{P} \left( \tilde{\mu}_{s, t-1} - \mu_* - \rho_s, t-1 - S_{s,t-1} \geq 0 \right) & \leq \mathbb{P} \left( \tilde{\mu}_{s, t-1} - \mu_* \geq S_{s,t-1} - |\rho_s, t-1| \right) \\
& \leq \mathbb{P} \left( \tilde{\mu}_{s, t-1} - \mu_* \geq \sqrt{\frac{2 \log(K)}{Q_{s, t-1}}} \right) .
\end{align}$$

(18)

Note that the low probability events from Proposition 1 sum to a constant independent of $T$. Then, a standard stratification over the value of $Q_{s, t-1}$, together with Hoeffding’s bound allows us to conclude that the above is upper bounded by a term of the form $1/(Kt^3)$ which, further summed over $t = 1, \ldots, T$ gives rise to a constant term, independent of $T$. Similarly, for the probabilities in the second sum we have

$$\begin{align}
\mathbb{P} \left( \mu_i - \tilde{\mu}_i, t-1 + \rho_i, t-1 + S_{i,t-1} - 2S_{i,t-1} \geq 0 \right) \\
& \leq \mathbb{P} \left( \mu_i - \tilde{\mu}_i, t-1 \geq \sqrt{\frac{2 \log(K)}{Q_{i, t-1}}} \right) .
\end{align}$$

(19)

Again, a standard stratification over the value of $Q_{i, t-1}$, together with Hoeffding’s bound yields an upper bound of the form $1/(Kt^3)$ which contributes a constant term overall.

Now from Equations (15) and (14), and the argument above, all we have left to do is bounding the following quantity

$$\sum_{i=1}^{K} \sum_{t=1}^{T} \Delta_i \mathbb{E} \left[ I_k = i \right] \mathbb{E} \left[ Q_{i, t-1} \leq s_i \right] .$$

(20)

We can bound this term via Proposition 2 and then applying Proposition 3 directly implies the claimed bound.

\[ \square \]
We now denote by $\mathcal{P}(\tilde{\theta}(\gamma)) = \{P_1, P_2, \ldots\}$ a partition of the vertices $\mathcal{E}$ of $G_{\tilde{\theta}(\gamma)}$ in cliques $P_k$: that is, for any $i, j \in P_k$, there is an edge from $i$ to $j$ and an edge from $j$ to $i$ in $G_{\tilde{\theta}(\gamma)}$. Such partitions are not unique and our bounds depend on the cliques (or clusters) in each partition. Thus, they are given in terms of the minimum over $\mathcal{P}(\tilde{\theta}(\gamma))$ of some quantities.

**Theorem 3.** With the notation of Theorem 1, the pseudo-regret of UCB-DSG run over $T$ rounds is bounded as follows:

$$R_T = O\left(\min_{\gamma \in [0,1]} \left(\min_{P \in \mathcal{P}(\tilde{\theta}(\gamma))} \sum_{i \in P} \frac{\max_{i,t} P_{i,t} \Delta_i}{\min_{i \in P} \Delta_i} \log(KT) + \frac{f(D, K)}{1 - (1 + \gamma)^2/4}\right)\right).$$

**Proof.** Recalling that for $t \in [T]$ and any $\delta \in (0, 1]$, $d_t(\delta) = 88a_t(\delta) + 34 + \sqrt{2\log(K(t + 1)}$ and $\delta_t = \frac{1}{K_t^2}$. Following the analysis of Theorem 1 and applying Proposition 3, we are left to bound

$$\sum_{i=1}^{T-1} \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[\mathbb{1}\{I_t = i\}\mathbb{1}\left\{Q_{i,t-1}^{\theta_i(\gamma)} < \left\lceil \frac{4d_t(\delta_t)^2}{\Delta_i^2} \right\rceil\right\}\right]$$

in terms of clique partition.

We then fix a partition $\mathcal{P}(\tilde{\theta}(\gamma))$ of the graph $G_{\tilde{\theta}(\gamma)}$ and bound the right-hand side of the last inequality based on this partitioning. We can write

$$(21) \leq \sum_{t=1}^{T} \sum_{P \in \mathcal{P}(\tilde{\theta}(\gamma))} \sum_{i \in P} \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[\mathbb{1}\{I_t = i\}\mathbb{1}\left\{Q_{i,t-1}^{\theta_i(\gamma)} < \left\lceil \frac{4d_t(\delta_t)^2}{\Delta_i^2} \right\rceil\right\}\right].$$

We now denote by $Q_{P,t-1}^{\theta_i(\gamma)} = \sum_{s=1}^{t-1} \mathbb{1}\{I_t \in P\}$ the number of times any expert in cluster $P$ for graph $G_{\tilde{\theta}(\gamma)}$ has been played by the algorithm up to time $t - 1$. Since experts in a cluster are observed together, we have $Q_{P,t-1}^{\theta_i(\gamma)} \leq Q_{i,t-1}^{\theta_i(\gamma)}$ for any $i \in P$, the inequality originating from cases where expert $i$ is observed while in the out-neighborhood of some expert not in $P$. This allows us to write

$$(21) \leq \sum_{t=1}^{T} \sum_{P \in \mathcal{P}(\tilde{\theta}(\gamma))} \sum_{i \in P} \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[\mathbb{1}\{I_t = i\}\mathbb{1}\left\{Q_{P,t-1}^{\theta_i(\gamma)} < \max_{i \in P} \left\lceil \frac{4d_t(\delta_t)^2}{\Delta_i^2} \right\rceil\right\}\right].$$

where in the last inequality we used the fact that, for all $s \geq 0$, we have

$$\sum_{t=1}^{T} \mathbb{1}\{I_t \in P\}\mathbb{1}\left\{Q_{P,t-1}^{\theta_i(\gamma)} < s\right\} \leq s.$$

Since the above holds for any fixed partition $\mathcal{P}(\tilde{\theta}(\gamma))$, we can take the minimum over all such partitions. Then, putting the above together, we obtain the bound of this theorem.

**B.6. Proof of Theorem 2**

In this section, we prove the lower bound on the regret, which reveals the intrinsic difficulty of learning with dependent stochastic graphs.
Theorem 2. For any number of arms $K \geq 2$, any gap value $\Delta \in (0, 1/4)$, any edge probability $p \in [0, 1)$, and any strongly consistent policy\(^3\) there exists a dependent stochastic feedback graph problem with process graph $G$ with $p_{i,j} = p$ for all $i \neq j$, $p_{i,i} = 1$ for all $i$, and expected losses $\mu_1, \ldots, \mu_K$, with $\mu_i - \mu_{i^*} = (1 - p)\Delta$ for all $i \neq i^*$, such that $R_T = \Omega(\frac{K}{\Delta} \log T)$.

Proof. Consider the following process generating the sequence $(G_t, \ell_t) \in \mathcal{G} \times [0, 1]^K$, for $t = 1, \ldots, T$. The loss realizations $\ell_t(\xi_i)$ are independent across $i$ (as well as across time) and are generated according to the following distributions. For $i \neq i^*$, we have

$$
\ell_t(\xi_i) = \begin{cases} 
0 & \text{with prob. } p \\
1/2 & \text{with prob. } (1 - p)/2 \\
1 & \text{with prob. } (1 - p)/2 ,
\end{cases}
$$

so that

$$
E[\ell_t(\xi_i)] = \frac{3}{4}(1 - p) .
$$

On the other hand, for $i = i^*$ we have

$$
\ell_t(\xi_i) = \begin{cases} 
0 & \text{with prob. } p \\
1/2 & \text{with prob. } (1/2 + 2\Delta)(1 - p) \\
1 & \text{with prob. } (1/2 - 2\Delta)(1 - p) ,
\end{cases}
$$

so that

$$
E[\ell_t(\xi_{i^*})] = (1 - p)(\frac{3}{4} - \Delta) .
$$

Moreover, the graph $G_t$ has self loops and is such that for $i \neq j$ there is an edge from $i$ to $j$ iff $\ell_t(\xi_j) = 0$. As a consequence, $p_{i,j} = p$ for all $i \neq j$, and $p_{i,i} = 1$. Finally, observe that the above loss distributions have been constructed in such a way that the (unconditional) gap $\mu_i - \mu_{i^*}$ satisfies

$$
\mu_i - \mu_{i^*} = (1 - p)\Delta ,
$$

hence the conditional gap, conditioned on the event that the loss value is not equal to zero satisfies

$$
E[\ell_t(\xi_i) \mid \ell_t(\xi_i) > 0] - E[\ell_t(\xi_{i^*}) \mid \ell_t(\xi_{i^*}) > 0] = \frac{(E[\ell_t(\xi_i)] - E[\ell_t(\xi_{i^*})])}{1 - p} = \Delta .
$$

Now, every time we pull action $I_t$ we observe the loss $\ell_t(\xi_{I_t})$, as well as the losses of all experts $i \neq I_t$ whose loss is 0. Since the probability of the loss being zero is the same for all arms, this extra observations cannot help identify $i^*$. It is only when $\ell_t(\xi_i) > 0$ that this loss value is useful to discriminate between $i$ and $i^*$, but such a value can only be observed when we do actually play action $I_t = i$. Hence, despite the large (marginal) probability $p_{i,j} = p$ on the edges, this problem is indeed equivalent to a standard stochastic multiarmed bandit problem with no feedback graph (i.e., with self-loops only), whose arm losses $\ell'_t(\xi_i)$ are distributed according to the conditional probabilities (given $\ell_t(\xi_i) > 0$) of the corresponding $\ell_t(\xi_i)$:

$$
\ell'_t(\xi_i) = \begin{cases} 
1/2 & \text{with prob. } 1/2 \\
1 & \text{with prob. } 1/2 ,
\end{cases}
$$

for $i \neq i^*$, and

$$
\ell'_t(\xi_{i^*}) = \begin{cases} 
1/2 & \text{with prob. } 1/2 + 2\Delta \\
1 & \text{with prob. } 1/2 - 2\Delta .
\end{cases}
$$

The resulting gaps $\mu'_i - \mu'_{i^*}$ are clearly equal to $\Delta$. For this problem it is easy to construct a lower bound via standard techniques (e.g., (Lai & Robbins, 1985; Burnetas & Katehakis, 1996; Salomon et al., 2013)) showing that for all strongly consistent policies, i.e., such that $R_T = o(T^\alpha)$ for all $\alpha \in (0, 1)$, we must have, for all $i \neq i^*$, $\lim_{T \to \infty} \frac{E[T_{i,T}]}{\ln T} \geq \frac{1}{KL_i}$, where $T_{i,T}$ is the number of times we play (suboptimal) action $i$ during the $T$ round, and $KL_i$ is the KL divergence between the distribution of any suboptimal arm and the optimal one, which in this case is $O(\Delta^2)$. In turn, this implies that the pseudo-regret $R_T$ must be $\Omega(\frac{K}{\Delta} \log T)$.

\(\square\)

\(\text{\textsuperscript{3}A strongly consistent policy is one such that } R_T = o(T^\alpha) \text{ for all } \alpha \in (0, 1).\)
C. Connections to Active Learning

As we mentioned in the main body of the paper, the absence of self-loops on the feedback graphs lends our setting to be instantiated to active learning scenarios with experts that can choose to either request labels or return predictions without seeing labels.

To see why, consider the following setting. Let $\mathcal{X}$ denote the input space and let $\mathcal{Y} = \{-1, 1\}$ denote the (binary) output space. Consider a set of training points of the form $z = (x, y)$ drawn i.i.d. according to some fixed yet unknown distribution over $\mathcal{X} \times \mathcal{Y}$, where the labels $y \in \mathcal{Y}$ of each point is not revealed unless explicitly requested. Let the set of experts $\mathcal{E}$ be of the form $\mathcal{E} = \{(h, r) : h \in \mathcal{H}, r \in \mathcal{R}\}$, defined by a pair of functions $\xi = (h, r)$, where $h : \mathcal{X} \to \mathbb{R}$ is a classifier function in a family of functions $\mathcal{H}$, and $r : \mathcal{X} \to \mathbb{R}$ is a requester function in a family of functions $\mathcal{R}$. The sign of $h(x)$ is used to predict the label of $x$, while the sign of $r(x)$ decides whether to request that label.

In practice, there is an intrinsic cost associated with requesting labels, e.g., monetary cost paid to a human to label each point individually. In order to capture this property, consider the following active learning loss function for expert $\xi = (h, r)$ on pairing $z = (x, y)$:

$$L(\xi, z) = c(x)I\{r(x) > 0\} + I\{yh(x) \leq 0\}I\{r(x) \leq 0\}.$$

This is the sum of the cost, $c(x) > 0$, of labelling a point $x$ when $\xi$ is requesting the associated label $y$ (i.e., $r(x) > 0$), and the cost of misclassifying the point when $\xi$ is not requesting the label ($r(x) \leq 0$). This framework fully captures the goal of active learning since an expert with a small expected active learning loss would correctly classify points while requesting labels a small amount of times. An ideal expert would in fact request fewer labels over the set of points where its prediction is incorrect and correctly classify the rest. These type of paired experts along with the active learning loss function was introduced in (Cortes et al., 2016), and further investigated by (Cortes et al., 2018), where the authors analyze the abstention setting. Yet, despite the syntactic similarity, the loss in the abstention setting has very different properties than the one in active learning defined above, because in the abstention setting the corresponding feedback graphs do always admit self-loops at all experts (see Section 2), while this is not the case for active learning.
Table 1. Table of dataset statistics.

<table>
<thead>
<tr>
<th>Name</th>
<th>No. Features $d$</th>
<th>No. Classes $c$</th>
<th>No. Edges in $\mathbb{E}[G_t]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>pendigits</td>
<td>16</td>
<td>10</td>
<td>101,023</td>
</tr>
<tr>
<td>covtype</td>
<td>54</td>
<td>7</td>
<td>143,554</td>
</tr>
<tr>
<td>poker</td>
<td>10</td>
<td>10</td>
<td>100,833</td>
</tr>
<tr>
<td>satimage</td>
<td>36</td>
<td>6</td>
<td>167,473</td>
</tr>
<tr>
<td>shuttle</td>
<td>9</td>
<td>7</td>
<td>143,783</td>
</tr>
<tr>
<td>CIFAR</td>
<td>25</td>
<td>10</td>
<td>100,877</td>
</tr>
<tr>
<td>sensorless</td>
<td>48</td>
<td>11</td>
<td>91,542</td>
</tr>
<tr>
<td>HIGGS</td>
<td>28</td>
<td>2</td>
<td>500,405</td>
</tr>
<tr>
<td>segment</td>
<td>19</td>
<td>7</td>
<td>143,762</td>
</tr>
<tr>
<td>letter</td>
<td>17</td>
<td>26</td>
<td>39,393</td>
</tr>
<tr>
<td>acoustic</td>
<td>50</td>
<td>3</td>
<td>333,933</td>
</tr>
</tbody>
</table>

D. Additional Experimental Results

In this section, we present further experimental results that compare our algorithm against several baselines.

We first show in Table 1 dataset statistics including number of features, number of classes, and expected number of edges in graph $\mathbb{E}[G_t]$, which was calculated via edge probabilities on the last round, $\hat{p}^T_{i,j}$. The number of classes affects the density of the resulting feedback graphs and, in fact, the table shows that the number of edges in graph $G_t$ decreases as the number of classes $c$ increases. For comparison, recall that for all datasets $K = 1,000$, thus the total number of edges is at most $10^6$.

Then, we given Figure 5 average regret comparing UCB, FS, and UCB-DSG algorithms for all datasets occurring in Table 1. We find that UCB-DSG algorithm outperforms UCB on almost all datasets and attains a performance close to that of FS on some of the datasets. Next, Figure 6 shows the average regret for ALG-1, ALG-2, and UCB-DSG algorithm. We find that the UCB-DSG algorithm outperforms these two baselines on most datasets.

In general, we expect a better performance of UCB-DSG relative to that of UCB whenever the percentage of edges of $\hat{G}_t$ is large since more feedback information is used. This is corroborated by the regret results of Figure 5 together with Figure 7. Figure 7 contain plots of the percentage of the number edges in graph $\hat{G}_t$ divided by the number of edges in graph $\mathbb{E}[G_t]$ for two datasets, sensorless and letter. For the letter dataset, the UCB-DSG admits a similar performance to that of UCB and Figure 7 shows, in fact, that the percentage of edges of $\hat{G}_t$ for the letter dataset is relatively small. On the other hand, the majority of edges are retained in graph $\hat{G}_t$ for the sensorless dataset and we find that UCB-DSG outperforms UCB on this dataset. This general pattern holds throughout and so we omit all other edge plots.
Figure 5. We show graphs of the average regret $R_t(\cdot)/t$ as a function of $t$ (log-scale) for UCB, UCB-DSG, and FS.
Figure 6. We show graphs of the average regret $R_t(\cdot)/t$ as a function of $t$ (log-scale) for $\text{ALG-1}$, $\text{UCB-DSG}$, and $\text{ALG-2}$. 
Figure 7. Plots of percentage of edges in the graph $\hat{G}_t$ divided by the edges of graph $E[G_t]$ as a function of $t$ (log-scale) for the letter and sensorless datasets.