Towards Fundamental Limits of Multi-armed Bandits with Random Walk Feedback

Tianyu Wang* Lin F. Yang† Zizhuo Wang‡

Abstract

Despite the ubiquitous applications of bandit learning algorithms in recommendation systems, social network, or online advertisement, where user behaviors can be modeled as a random walk over a network, few studies have utilized the network structure to improve learning efficiency. In this paper, we address this issue by providing a novel bandit learning formulation, where each arm is the starting node of a random walk in a network and the reward is the length of walk. This formulation not only captures a large number of applications in practice but also provides a framework to actively reduce learning complexity by utilizing graph structure in the random walk feedback. We provide a comprehensive understanding of this formulation by establishing matching learning complexity upper and lower bounds, in both the stochastic and the adversarial setting. In the stochastic setting, by utilizing the feedback structure, our learning method can achieve constant regret, whereas regular bandit algorithms' regret grows with T. In the adversarial setting, we establish novel algorithms that achieve regret bound of order $\widetilde{\mathcal{O}}\left(\kappa\sqrt{T}\right)$, where κ is a constant that depends on the structure of the graph, instead of number of arms (nodes). This bounds significantly improves regular bandit algorithms, whose complexity depends on number of arms (nodes).

1 Introduction

Bandit problems simultaneously call for exploitation of good options and exploration of the decision space. Algorithms for this problem find applications in various domains, from medical trials (Robbins, 1952) to online advertisement (Li et al., 2010). Many authors have studied bandit problems from different perspectives.

In this paper, we study a new bandit learning problem where the feedback is depicted by a random walk over the arms. That is, each time a node i is played, one observes a random walk over the arms(nodes) from i to an absorbing node, and the reward is the length of this random walk. Such feedback structure may show up in different scenarios. One concrete motivation is the browsing behavior of internet users within a certain web domain. Specifically, one may view a user's browsing record as a random walk. Web pages within a web domain are viewed as graph nodes. The user transits to another node when she opens another page in this domain. This random walk ends (hits an absorbing node) when the user exits the domain. In this learning setting, we want to carefully select entrances to this web domain (e.g. providing a recommendation link to a user), so that the profit (generated by browsing) is maximized. We therefore ask the following question:

In a graph with an absorbing node, if we can select the initial node to seed a random walk and observe the random walk trajectory, how should we select the initial nodes, so that the random walks are as long-lasting as possible? **(P)**

We study this problem from an online learning perspective. To be more precise, we consider the following model. The environment is modeled by a graph $G = (V, E)$, where V consists of transient nodes $[K] := \{0, 1, \ldots, K-1\}$ and an absorbing node $*$. Each edge ij ($i \in [K], j \in [K] \cup \{*\}$) can encode two quantities, a transition probability from i to j and the distance from i to j. For $t = 1, 2, \ldots, T$, we pick a node to start a random walk, and observe the random walk trajectory from the selected node to the absorbing node. For each random walk, we use its hitting time (to the absorbing node $*$) to model how long-lasting it is. With this formulation, we can define a bandit learning problem for the question (**P**). Each time, the agent picks a node in G to start a random walk, observes the trajectory, and receives the hitting time of the random walk as reward. In this setting, the performance of learning algorithms is measured by *regret*, which is the difference between the rewards of an optimal node and the rewards of the nodes played by the algorithm. Unlike standard multi-armed bandit problems, the feedback is random walk trajectories and thus reveals information not only about the node played,

*tianyu@cs.duke.edu, Duke University.
†linyang@ee.ucla.edu, UCLA.
‡wangzizhuo@cuhk.edu.cn, CUHK-SZ.

but the environment (transitions/distances among nodes) as well. This new feedback structure calls for new insights on learning with graph random walk feedback.

We start with a stochastic version. In the stochastic version, the graph G is fixed and unknown. In studying the stochastic formulation, we first formulate the problem, and introduce some concepts. In fact, by a variant of the the UCB principle (Lai and Robbins, 1985; Auer et al., 2002a), we solve this setting optimally. Then we study an adversarial version, where the edge lengths in graph G change adversarially over time. This setting takes care of changing environments, which can model the potential change of users' preference. We develop a variant of the exponential weight algorithm (Littlestone and Warmuth, 1994; Auer et al., 2002b) for the adversarial formulation. In the adversarial formulation, a high probability regret of order $\widetilde{\mathcal{O}}\left(\sqrt{\kappa T}\right)$ can be achieved, where κ depends on the graph structure. Also, our algorithm for the adversarial case uses a new odd-even trick. This trick decouples the randomness induced by decision made by the algorithm, the environment, and the estimation of the environment. Our odd-even trick might be useful in other bandit learning settings as well. We also provide lower bounds that match our upper bounds.

The construction of the lower bound introduces several novel insights compared to regular bandit lower bounds, where the events are from a simpler sample space. If we execute a policy π for T epochs on a problem instance, the sample space is then $\left(\cup_{h=1}^{\infty}\mathcal{B}^h\right)^T$, where $\mathcal{B}$ is the space of all events that a single step on a trajectory can generate. For example, if all edge length are fixed, then $\mathcal{B}=[K]$, since a single step on a trajectory might be any node. The difficulty in the sample space is: the trajectory can be arbitrarily long, which is reflected by the union up to infinity. To handle this issue, we make use of the Markov property: conditioning on the k-th node being j, the subsequent sample space generated is identical to the sample space conditioning on the first node being j. By using this observation, one can "disentangle" the randomness induced by possible infinite length of random walks. Also, using this observation, one may be able to derive other forms of lower bounds under this random walk feedback structure.

In terms of applications, many other scenarios also fit in our model. For example, our model can also describe the browsing over items (e.g., videos, news articles, commodities) in mobile apps. In this case, items are modeled as nodes in a graph, in which tapping a node leads to a transition to another node, and closing the mobile app means hitting an absorbing node. We may also want to prolong the browsing activity in this case. Many other recommendation system applications follow a similar structure.

In summary, we study a novel bandit learning problem motivated by user's browsing behavior. We provide understandings of the fundamental statistical limits of this problem by developing algorithms for both the stochastic and adversarial settings, and establishing matching lower bounds. Notably, the problem (**P**) we ask is generic and we hope our study could inspire further investigations in this broad problem.

1.1 Related Works

Bandit problems date its history back to at least Thompson (1933), and have been studied extensively in the literature. One of the the most popular approaches to the stochastic bandit problem is the Upper Confidence Bound (UCB) algorithms (Robbins, 1952; Lai and Robbins, 1985; Auer, 2002). Various extensions of UCB algorithms have been studied (Srinivas et al., 2010a; Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2012; Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014). Specifically, some works use KL-divergence to construct the confidence bound (Lai and Robbins, 1985; Garivier and Cappé, 2011; Maillard et al., 2011), or include variance estimates within the confidence bound (Audibert et al., 2009; Auer and Ortner, 2010). UCB is also used in the contextual learning setting (e.g., Li et al., 2010; Krause and Ong, 2011; Slivkins, 2014). Parallel to the stochastic setting, studies on the adversarial bandit problem form another line of literature. Since randomized weighted majorities (Littlestone and Warmuth, 1994), exponential weights remains a top strategy for adversarial bandits (Auer et al., 1995; Cesa-Bianchi et al., 1997; Auer et al., 2002b). Many efforts have been made to improve/extend exponential algorithms. For example, Kocák et al. (2014) target at implicit variance reduction. Mannor and Shamir (2011); Alon et al. (2013) study a partially observable setting. Despite the large body of literature, no previous work has, to the best of our knowledge, explicitly focused the question (**P**). Specifically, if one applies vanilla bandit algorithms without using the graph structure. The regret bound would depend on the number of options (arms/nodes). This may be much worse than a environment-dependent bound.

For both stochastic bandits and adversarial bandits, lower bounds in different scenarios have been derived, since the $\mathcal{O}(\log T)$ asymptotic lower bounds for consistent policies (Lai and Robbins, 1985). Worst case bound of order $\mathcal{O}(\sqrt{T})$ have also been derived (Auer et al., 1995) for the stochastic setting. In addition to the classic stochastic setting, lower bounds in other stochastic (or stochastic-like) settings have also been considered, including PAC-learning complexity (Mannor and Tsitsiklis, 2004), best arm identification complexity (Kaufmann et al., 2016; Chen et al., 2017), and lower bounds in continuous spaces (Kleinberg et al., 2008). Lower bound problems for adversarial bandits may be converted to lower bound problems for stochastic bandits (Auer et al.,

1995) in many cases. An intriguing lower bound beyond the expected regret is the high probability lower bound of order $\widetilde{\mathcal{O}}\left(\sqrt{T}\right)$ by Gerchinovitz and Lattimore (2016).

In terms of the problem setting, a related problem is the *Stochastic Shortest Path* problem (Bertsekas and Tsitsiklis, 1991), where the agent selects actions to travel to absorbing states as quickly as possible. In stochastic shortest path problems, the agent makes decisions at every state (node). In contrast, in our setting, the agent only selects the initial node, and has no further control of the random walk trajectory.

Another related setting is bandit with side information (Mannor and Shamir, 2011; Alon et al., 2015, 2017). In such problems, playing a node reveals information about other nodes, where the feedback structure is governed by an observation graph. Such problem assumes that the observation graph is revealed, either before or after the player has made a decision. In our setting, however, and the observation model is unknown and needs to be learned. Very importantly, our setting has much more randomness than that for bandit with side information and is harder in a statistical sense; See Section 4 for more discussion.

While bandit problems have been studied in different settings using various techniques, no prior works, to the best of our knowledge, focus on answering the important question (**P**) to obtain near optimal bounds. Our paper provides a comprehensive answer to this important class of problems (**P**) for propagation over graphs.

2 Problem Setting

In this section, we formulate the problem and put forward notations and definitions that will be used throughout the rest of the paper. The learning process repeats for T epochs and the learning environment is described by graphs $G_1, G_2, \ldots, G_T$ for epochs $t = 1, 2, \ldots, T$. The graph G_t is defined on K transient nodes $[K] = \{0, 1, \ldots, K-1\}$ and one absorbing node denoted by $*$. We will use $V = [K]$ to denote the set of transient nodes, and use $\widetilde{V} := [K] \cup \{*\}$ to denote the transient nodes together with the absorbing node. On this node set $\widetilde{V}$, graph G_t encodes transition probabilities and edge lengths: $G_t := \left(\{m_{ij}\}_{i\in V, j\in\widetilde{V}}, \{l_{ij}^{(t)}\}_{i\in V, j\in\widetilde{V}}\right)$, where m_{ij} is the probability of transiting from i to j and $l_{ij}^{(t)} \in [0,1]$ is the length from i to j (at epoch t). We gather the transition probabilities among transient nodes to form a transition matrix $M = [m_{ij}]_{i,j\in[K]}$. We make the following assumption about M.

Assumption 1. *The transition matrix $M = [m_{ij}]_{i,j\in[K]}$ among transient nodes is primitive.* [1] *In addition, there is a constant ρ, such that $\|M\|_\infty \le \rho < 1$, where $\|M\|_\infty = \max_{i\in[K]} \sum_{j\in[K]} |m_{ij}|$ is the maximum absolute row sum.*

In Assumption 1, the primitivity assumption ensures that we can get to any transient node v from any other node state u. The infinite norm of M being strictly less than 1 means that the random walk will transit to the absorbing node starting from any node (eventually with probability 1). This describes the absorptiveness of the environment. Note that this infinite norm assumption can be replaced by other notions of matrix norms.

Playing node j at epoch t generates a random walk trajectory $\mathcal{P}_{t,j} := \left(X_{t,0}^{(j)}, L_{t,1}^{(j)}, X_{t,1}^{(j)}, L_{t,2}^{(j)}, X_{t,2}^{(j)}, \ldots, L_{t,H_{t,j}}^{(j)}, X_{t,H_{t,j}}^{(j)}\right)$, where $X_{t,0}^{(j)} = j$ is the starting nodes, $X_{t,H_{t,j}}^{(j)} = *$ is the absorbing node, $X_{t,i}^{(j)}$ is the i-th node in the random walk trajectory, $L_{t,i}^{(j)}$ is the edge length from $X_{t,i-1}^{(j)}$ to $X_{t,i}^{(j)}$, and $H_{t,j}$ is the number of edges in trajectory $\mathcal{P}_{t,j}$. For simplicity, we write $X_{t,i}^{(j)}$ (resp. $L_{t,i}^{(j)}$) as $X_{t,i}$ (resp. $L_{t,i}$) when it is clear from context.

For the random trajectory $\mathcal{P}_{t,j} := \left(X_{t,0}, L_{t,1}, X_{t,1}, L_{t,2}, X_{t,2}, \ldots, L_{t,H_{t,j}}, X_{t,H_{t,j}}\right)$, the length of the trajectory (or ***hitting time*** of node j at epoch t) is defined as

$$\mathcal{L}\left(\mathcal{P}_{t,j}\right) := \sum_{i=1}^{H_{t,j}} L_{t,i}. \tag{1}$$

Here we use the edge length to represent the reward of the trajectory. In practice, the edge lengths may have real-world meanings. For example, the out-going edge from a node may represent utility (e.g., profit) of visiting this node. At epoch t, the agent selects a node $J_t \in [K]$ to initiate a random walk, and observe trajectory $\mathcal{P}_{t,J_t}$. In stochastic environments, the environment does not change across epochs. Thus for any fixed node $v \in [K]$, the random trajectories $\mathcal{P}_{1,v}, \mathcal{P}_{2,v}, \mathcal{P}_{3,v}, \ldots$ are independently identically distributed. Because of this, we write $\mu_v := \mathbb{E}\left[\mathcal{L}\left(\mathcal{P}_{t,v}\right)\right]$ for simplicity. In this case, the regret is defined as

$$\mathrm{Reg}(T) = \max_{i\in[K]} \sum_{t=1}^{T} \mu_i - \sum_{t=1}^{T} \mu_{J_t}, \tag{2}$$

[1] A matrix M is primitive if there exists a positive integer k, such that all entries in M^k is positive.

where J_t is the node played by the algorithm (at epoch t). In adversarial problems, the environment can change across epochs, and the regret against a fixed node j is defined as

$$\mathrm{Reg}_j^{\mathrm{adv}}(T) = \sum_{t=1}^{T} \mathcal{L}\left(\mathcal{P}_{t,j}\right) - \sum_{t=1}^{T} \mathcal{L}\left(\mathcal{P}_{t,J_t}\right). \tag{3}$$

In the adversarial setting, it suffices to bound $\mathrm{Reg}_j^{\mathrm{adv}}(T)$ for all $j \in [K]$. We also define a notion of centrality that will be used later.

Definition 1. *Let $X_0, X_1, X_2, \ldots, X_\tau = *$ be nodes on a random trajectory. Under Assumption 1, we define, for node $v \in [K]$,*

$$\alpha_v := \min_{u \in [K],\, u \neq v} \mathbb{P}\left(v \in \{X_1, X_2, \ldots, X_\tau\} | X_0 = u\right)$$

*to be the **hitting centrality** of node v. We also define $\alpha = \min_v \alpha_v$*

Hitting centrality of a node v is how likely it is visited by a trajectory starting from another node. In non-absorptive (and ergodic) Markov chains, the hitting centrality of any node is 1. This quantity is less than 1 for networks with absorbing nodes. As we will show in the analysis, the hitting centrality will be a factor in the regret bound as a problem intrinsic parameter.

3 Stochastic Setting

In the stochastic setting, the graphs G_t do not change across epochs. To solve this problem, we estimate the expected hitting times μ_j for all nodes $j \in [K]$ (and maintain a confidence interval of the estimations). We use the mean estimates (and the confidence intervals) to construct decision indices. In each epoch, we play a node with a node with the highest decision index.

To formally introduce our strategies, we first define the following quantities. We define number of times a node is played, and number of times a node is covered by a trajectory. For a node $v \in [K]$, define

$$N_t(v) := 1 \vee \sum_{s<t} \mathbb{I}_{[J_s = v]}, \qquad N_t^+(v) := 1 \vee \sum_{s<t} \mathbb{I}_{[v \in \mathcal{P}_{s,J_s}]}, \tag{4}$$

where $a \vee b := \max\{a, b\}$.

In words, $N_t(v)$ is the number of times we play a node v up to epoch t, and $N_t^+(v)$ is number of trajectories that cover node v up to epoch t.

Recall $\mathcal{P}_{t,J_t} = \left(X_{t,0}, L_{t,1}, X_{t,1}, \ldots, L_{t,H_{t,J_t}}, X_{t,H_{t,J_t}}\right)$ is the trajectory at epoch t. For a node v and the trajectories $\mathcal{P}_{1,J_1}, \mathcal{P}_{2,J_2}, \mathcal{P}_{3,J_3}, \ldots$, let $k_{v,i}$ be the index (epoch) of the i-th trajectory that covers node v. Let $Y_{v,k_{v,i}}$ be the sum of edge lengths between the first occurrence of v and the absorbing node $*$ in trajectory $k_{v,i}$. For a transient node $v \in [K]$ and n trajectories (at epochs $k_{v,1}, k_{v,2}, \ldots, k_{v,n}$) that cover node v, the hitting time estimator of v is computed as

$$\widetilde{Z}_{v,n} := \frac{1}{n} \sum_{i=1}^{n} Y_{v,k_{v,i}}. \tag{5}$$

Since $v \in \mathcal{P}_{t,v}$, $Y_{v,k_{v,i}}$ is an identical copy of the hitting time $\mathcal{L}\left(\mathcal{P}_{t,v}\right)$ (Proposition 1). For (5), one can also use robust mean estimators (Bubeck et al., 2013), but a plain mean estimator is sufficient for the purpose of this paper.

Proposition 1. *In the stochastic setting, for any nodes $v \in [K]$, we have, for $\forall t, i \in \mathbb{N}_+, \forall r \in \mathbb{R}$*

$$\mathbb{P}\left(Y_{v,k_{v,i}} = r\right) = \mathbb{P}\left(\mathcal{L}\left(\mathcal{P}_{t,v}\right) = r\right). \tag{6}$$

Proof. In a trajectory $\mathcal{P}_{t,J_t} = \left(X_{t,0}, L_{t,1}, X_{t,1}, \ldots, L_{t,H_{t,J_t}}, X_{t,H_{t,J_t}}\right)$, conditioning on $X_{t,i} = j$ being known (and no future information is revealed), the randomness generated by $L_{t,i+1}, X_{t,i+1}, L_{t,i+2}, X_{t,i+2}, \ldots$ is identical to the randomness generated by $L_{t,1}, X_{t,1}, L_{t,2}, X_{t,2}, \ldots$ conditioning on $X_{t,0} = j$ being fixed. □

Proposition 1 is a simple and clean consequence of the Markov property. Note that even if each trajectory can visit a node multiple times, only one hitting time sample can be used. This is because extracting multiple sample would break Markovianity, by revealing that the random walk will visit a same node again. In fact, the virtue of this proposition will be used more than once as we expand the paper.

The simplest strategy is to play the node with the largest estimated hitting time. In each epoch t, we play the node that maximizes the empirical estimates of the hitting times, $\widetilde{Z}_{v,N_t^+(v)}$. This strategy is formally stated in Algorithm 1.

Algorithm 1

1: **Input:** A set of nodes $[K]$ (and an absorbing node $*$).
2: **Warm up:** Play each node once to initialize. Observe trajectories.
3: **for** $t = 1, 2, 3, \ldots$ **do**
4: Select J_t to start a random walk, such that

$$J_t \in \arg\max_{v \in V} \widetilde{Z}_{v,N_t^+(v)},$$

where with ties broken arbitrarily.
5: Observe the trajectory $\mathcal{P}_{t,v_t} := \left\{X_{t,0}, L_{t,1}, X_{t,1}, L_{t,2}, X_{t,2}, \cdots, L_{t,H_{t,v_t}} X_{t,H_{t,v_t}}\right\}$. Update $N_t^+(v)$ and estimates $\widetilde{Z}_{t,N_t(v)}$ for all $v \in [K]$.
6: **end for**

Before proceeding to analysis of Algorithm 1 and further sections, we define the notion of optimality gap. In a stochastic environment, the graph does not change across epochs. We define the optimality gap Δ_v of a node v to be $\Delta_v := \max_{i \in [K]} \mu_i - \mu_v$.

Analysis of Algorithm 1: Gap-Dependent Constant Regret

The above simple algorithm, which does not require any prior knowledge of the hitting time distribution or the network structure, achieves a constant regret as stated in Theorem 1.

Theorem 1. *Suppose Assumption 1 holds. Algorithm 1 achieves a constant regret that only depends on* $\alpha_v, \alpha_{v^*}, \rho$, *and* Δ_v: $\mathrm{Reg}(T) \leq \widetilde{\mathcal{O}}\left(\sum_{v:\Delta_v>0}\left(\frac{1}{\min\{\alpha_v,\alpha_{v^*}\}(1-\rho)^2\Delta_v} + \Delta_v\right)\right)$, *where* v^* *is the optimal node (the node with maximum hitting time), and* $\widetilde{\mathcal{O}}$ *omits absolute constants and logarithmic dependence on problem intrinsics.*

Unlike the standard bandit problem, playing any node in our problem reveals information about the environment (other nodes). In this stochastic settings, this means: After a certain time, all hitting time estimates are close enough to the true hitting time, and few mistakes will be made. More details about Theorem 1 can be found in Appendix B.3.

3.1 Optimistic Strategies

For optimistic strategies, we estimate $\mathcal{L}(\mathcal{P}_{t,i})$ from past observations, and compute confidence intervals of our estimations. Based on the estimations and confidence intervals, we play a node that maximizes the UCB index.

Given $N_t^+(v)$ trajectories covering v, the confidence terms (at epoch t) are

$$\widetilde{C}_{N_t^+(v),t} := \sqrt{\frac{8\xi_t \log t}{N_t^+(v)}},$$

where $\xi_t = \max\left\{1 + \frac{\rho}{(1-\rho)^2}, \frac{\log(1-\rho)}{\log\rho} + \frac{5\log t}{\log 1/\rho}\right\}$.

At each time t, we play a node J_t that maximizes the UCB index, $\widetilde{Z}_{v,N_t^+(v)} + \widetilde{C}_{N_t^+(v),t}$. This strategy is described in Algorithm 2.

Analysis of Algorithm 2: Gap-Dependent Constant Regret

Similar to Algorithm 1, optimistic strategies achieve constant instance-dependent regret.

Theorem 2. *On a problem instance that satisfies Assumption 1, Algorithm 2 achieves constant regret of order* $\widetilde{\mathcal{O}}\left(\sum_{v:\Delta_v>0}\left(\Delta_v + \frac{1}{(1-\rho)^2\Delta_v}\right)\right)$, *where* $\widetilde{\mathcal{O}}$ *omits absolute constants and logarithmic dependence on problem intrinsics.*

Algorithm 2

Input: A set of nodes $[K]$. Parameters: a constant ρ that bounds the spectral radius of P.
Warm up: Play each node once to initialize. Observe trajectories.
For any $v \in [K]$, define the decision index

$$I_{v,N_t^+(v),t} = \widetilde{Z}_{v,N_t^+(v)} + \widetilde{C}_{N_t^+(v),t}. \tag{7}$$

for $t = 1, 2, 3, \ldots$ **do**
  Select J_t to start a random walk, such that

$$J_t \in \arg\max_{v \in V} I_{v,N_t^+(v),t}, \tag{8}$$

  with ties broken arbitrarily.
  Observe the trajectory $\mathcal{P}_{t,v_t} := \left\{X_{t,0}, L_{t,1}, X_{t,1}, L_{t,2}, X_{t,2}, \cdots, L_{t,H_{t,v_t}} X_{t,H_{t,J_t}}\right\}$. Update $N_t^+(\cdot)$ and decision indices for all $v \in [K]$.
end for

Similar to that for Algorithm 1, all hitting time estimates are close enough to the true hitting times, since playing one nodes reveals information about the environment (other nodes). In the stochastic setting, this means few mistakes will be made after a certain epoch T. Due to UCB exploration, Algorithm 2 learns the hitting times quicker than Algorithm 1, especially when the hitting centralities are small. This dependence on hitting centralities (α) is reflected in the theorem statements: in Theorem 1, the regret rate scales with $\frac{1}{\min\{\alpha_v, \alpha_{v^*}\}}$, while in Theorem 2, the regret rate scales with $\log \frac{1}{\alpha_v}$. This is also illustrated empirically in Section 5.

Analysis of Algorithm 2: Risk Analysis

In this section, we study the risk of the Algorithm 2. As suggested by the pseudo-regret measure by Audibert et al. (2009), the risk can be measured by the distribution of number of times a sub-optimal arm is played.

For any sub-optimal node v, above its mean, the tail decay of $N_t(v)$ by Algorithm 2 follows a power-law tail decay. In particular, the following theorem extends previous results for standard bandits (Theorem 8 & 10, Audibert et al., 2009) to our setting.

Theorem 3. *For any suboptimal node v, and T and x such that*

$$x, T \geq \min\left\{t : \frac{\xi_t \log t}{\alpha_v t - t^{3/4}} \leq \frac{1}{32}\Delta_v^2\right\},$$

Algorithm 2 satisfies

$$\mathbb{P}\left(N_T(v) \geq x\right) \leq \widetilde{\mathcal{O}}\left(\frac{1}{(1+\alpha_v)x^3}\right), \tag{9}$$

where $\widetilde{\mathcal{O}}$ omits problem-dependent constants and terms of (exponentially) smaller order.

An interesting observation from Theorem 3 is: when playing one node reveals information about the environment and other nodes (or equivalently, $\alpha_v > 0$), the algorithm exhibits a power-law tail decay rate on $\mathbb{P}\left(N_T(v) \geq x\right) \lesssim \frac{1}{x^3}$, no matter how small α_v is. A proof of Theorem 3 can be found in Appendix B.5.

Analysis of Algorithm 2: Instance-independent Regret Bound

While in Section 3.1 we derived a constant regret for Algorithm 2, the regret bound depends on problem intrinsics. In particular, the bounds in Theorems 1 and 2 scale with $\frac{1}{\Delta_v}$. In this section, we have the following gap-independent regret bound stated in Theorem 4.

Theorem 4. *Let T be any positive integer. Under Assumption 1, Algorithm 2 admits a regret of order*

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\min\left\{\frac{1}{(1-\rho)^2}\sqrt{\frac{T}{\alpha}}, \frac{1}{(1-\rho)^2}\sqrt{KT}\right\}\right),$$

where $\alpha = \min_{v \in V} \alpha_v$ (α_v defined in Definition 1).

The regret rate in Theorem 4 is similar to that for standard bandit setting. The proof (Appendix B.6) uses concentration of estimators at each epoch, followed by a Cauchy-Schwarz inequality. As shown in Theorem 5 later, $\mathcal{O}\left(\sqrt{T}\right)$ is the best instance-independent regret one can achieve.

3.2 Worst Case Lower Bound for Stochastic Setting

In this section, we show for any given T, one can construct a problem instance such that no algorithm achieves better regret than $\Omega\left(T^{1/2}\right)$. In this worst case analysis, we allow the adversary to choose a hard instance with access to all randomness, including when we stop the algorithm run.

Theorem 5. *For any given T and any policy π, there exists a problem instance $\mathfrak{I}$ satisfying Assumption 1 such that the T step regret of π on instance $\mathfrak{I}$ is lower bounded by $\Omega\left(T^{1/2}\right)$.*

The proof idea of Theorem 5 is as follows. We construct two similar problem instances, $\mathfrak{J}$ and $\mathfrak{J}'$, so that no policy can quickly distinguish them from each other. To make this precise, we study the probability space of playing a policy π on $\mathfrak{J}$ and $\mathfrak{J}'$. Let $\mathbb{P}_{\mathfrak{J},\pi}$ (resp. $\mathbb{P}_{\mathfrak{J}',\pi}$) be the distribution of executing π on $\mathfrak{J}$ (resp. $\mathfrak{J}'$). We show that the total variation distance between $\mathbb{P}_{\mathfrak{J},\pi}$ and $\mathbb{P}_{\mathfrak{J}',\pi}$ is small. If this distance between $\mathbb{P}_{\mathfrak{J},\pi}$ and $\mathbb{P}_{\mathfrak{J}',\pi}$ is small, the policy π must play node 1 frequently (or infrequently) in both $\mathfrak{J}$ and $\mathfrak{J}'$. This means π will make enough mistakes in either $\mathfrak{J}$ (if π plays node 1 infrequently in both $\mathfrak{J}$ and $\mathfrak{J}'$) or $\mathfrak{J}'$ (if π plays node 1 frequently in both $\mathfrak{J}$ and $\mathfrak{J}'$).

A tricky side of our problem is that that sample space (on which both $\mathbb{P}_{\mathfrak{J},\pi}$ and $\mathbb{P}_{\mathfrak{J}',\pi}$ are defined) is different from a regular bandit problem. If we execute π for T epochs, the sample space $\mathfrak{S}$ is then $\left(\cup_{h=1}^{\infty}[K]^h\right)^T$. This is because each trajectory can be arbitrarily long, and the nodes on the trajectory can be any of $[K]$.

More discussions on this sample space and the lower bound proof can be found in Section 4 following Theorem 7. A proof of Theorem 5 can be found in Appendix C.2.

4 Adversarial Setting

In this section, we consider the case in which the network structure G_t changes over time. Here we study a version of this problem in which the adversary alters edge length across epochs: In each epoch, the adversary can arbitrarily pick edge lengths $l_{ij}^{(t)}$ from $[0,1]$. Recall, in this case, the performance is measured by the regret against playing any fixed node $j \in [K]$:

$$\mathrm{Reg}_j^{\mathrm{adv}}(T) = \sum_{t=1}^{T} \mathcal{L}\left(\mathcal{P}_{t,j}\right) - \sum_{t=1}^{T} \mathcal{L}\left(\mathcal{P}_{t,J_t}\right),$$

where J_t is the node played in epoch t, and $\mathcal{L}\left(\mathcal{P}_{t,j}\right)$ is defined in (1).

We will use an extension of the exponential weight algorithm to solve this adversarial problem. Intuitively, our algorithm maintains a probability distribution over the nodes. This probability distribution gives higher weights to historically more rewarding nodes. In each epoch, we sample a node from this probability distribution, play this node, and record down information from this action. To symbolically describe the strategy, we first define some notations. We first extract a sample of $\mathcal{L}\left(\mathcal{P}_{t,j}\right)$ from the trajectory $\mathcal{P}_{t,J_t}$, where J_t is the node played in epoch t.

Given the trajectory for epoch t $\mathcal{P}_{t,J_t} = \left(X_{t,0}, L_{t,1}, X_{t,1}, \ldots, L_{t,H_{t,J_t}}, X_{t,H_{t,J_t}}\right)$, we define, for $v \in [K]$,

$$Y_v(\mathcal{P}_{t,J_t}) = \max_{i:0\le i<H_{t,J_t}} \mathbb{I}_{[X_{t,i}=v]} \cdot \mathcal{L}_i(\mathcal{P}_{t,J_t}), \tag{10}$$

where $\mathcal{L}_i(\mathcal{P}_{t,J_t}) := \sum_{k=i+1}^{H_{t,J_t}} L_{t,k}$. In words, $\mathcal{L}_i(\mathcal{P}_{t,J_t})$ is the distance (sum of edge lengths) from the first occurrence of v to the absorbing node.

By the principle of Proposition 1, if node i is covered by trajectory $\mathcal{P}_{t,J_t}$, $Y_v(\mathcal{P}_{t,J_t})$ is a sample of $\mathcal{L}\left(\mathcal{P}_{t,i}\right)$. We define, for the trajectory $\mathcal{P}_{t,J_t} = \{X_{t,0}, L_{t,1}, X_{t,1}, L_{t,2}, \ldots, L_{t,H_{t,J_t}}, X_{t,H_{t,J_t}}\}$,

$$Z_{t,v} := Y_v(\mathcal{P}_{t,J_t}), \quad \forall v \in [K], \tag{11}$$

where $Y_v(\mathcal{P}_{t,J_t})$ is defined above in (10).

Define $\mathbb{I}_{t,ij} := \mathbb{I}_{\left[i\in\mathcal{P}_{t,J_t} \text{ and } j\in\mathcal{P}_{t,J_t}, Y_i(\mathcal{P}_{t,J_t})>Y_j(\mathcal{P}_{t,J_t})\right]}$. This indicator random variable is 1 iff i and j both show up in $\mathcal{P}_{t,J_t}$ and the first occurrence of j is after the first occurence of i. We then define

$$\widehat{q}_{t,ij} := \frac{\sum_{s=1}^{t-1} \mathbb{I}_{s,ij}}{N_t^+(i)}, \tag{12}$$

which is an estimator of how likely j is visited via a trajectory starting at i.

Using the above defined $\widehat{q}_{t,ij}$ and sample $Z_{t,i}$, we define an estimator for $\mathbb{E}\left[\mathcal{L}\left(\mathcal{P}_{t,j}\right)\right] - B$ as

$$\widehat{Z}_{t,i} := \frac{\left(Z_{t,i} - B\right)\mathbb{I}_{\left[i \in \mathcal{P}_{t,J_t}\right]}}{p_{ti} + \sum_{j \neq i} \widehat{q}_{t,ji} p_{tj}}, \quad \forall i \in [K], \tag{13}$$

where B is an algorithm parameter. With the estimators $\widehat{Z}_{t,i}$, we define

$$\widehat{S}_{t,j} = \begin{cases} \sum_{s:s \text{ is even, and } 0 \leq s \leq t} \widehat{Z}_{s,i}, & \text{if } t \text{ is even,} \\ \sum_{s:s \text{ is odd, and } 1 \leq s \leq t} \widehat{Z}_{s,i}, & \text{if } t \text{ is odd.} \end{cases} \tag{14}$$

By convention, we set $\widehat{S}_{0,i} = 0$ and $\widehat{S}_{1,i} = 0$ for all $i \in [K]$.

Now we define the probability of playing i in epoch t as

$$p_{ti} := \begin{cases} \frac{1}{K}, & \text{if } t = 1, 2, \\ \frac{\exp\left(\eta \widehat{S}_{t-2,i}\right)}{\sum_{j=1}^{K} \exp\left(\eta \widehat{S}_{t-2,j}\right)}, & \text{if } t \geq 3, \end{cases} \tag{15}$$

Note that by our algorithm design, the probability of playing a node in epoch t depends only on randomness up to epoch $t-2$. However, the estimator $\widehat{q}_{t,ij}$ depends on randomness in epoch $t-1$. This odd-even trick disentangles the interleaved randomness, which is critical in our analysis.

Against any arm $j \in [K]$, following the sampling rule (15) can guarantee an $\widetilde{\mathcal{O}}\left(\sqrt{T}\right)$ regret bound. We now summarize our strategy in Algorithm 3, and state the performance guarantee in Theorem 6.

Algorithm 3

1: **Input:** A set of nodes $[K]$, transition matrix M, total number of epochs T, probability parameter $\epsilon \in \left(0, \frac{1}{(1-\rho)KT}\right)$. Algorithm parameters: $B = \frac{\log \frac{(1-\rho)\epsilon}{KT}}{\log \rho}$, $\eta = \frac{1}{B\sqrt{T}} \cdot \frac{\sqrt{\log K}}{\sqrt{1+2\sum_{j\in[K]} \frac{1-\alpha_j}{1+\alpha_j}}}$.

2: **for** $t = 1, 2, 3, \ldots, T$ **do**

3: Randomly play node $J_t \in [K]$, such that

$$\mathbb{P}\left(J_t = i\right) = p_{ti}, \forall i \in [K]. \tag{16}$$

where p_{ti} is defined in (15).

4: Observe the trajectory $\mathcal{P}_{t,J_t}$. Update estimates $\widehat{Z}_{t,j}$ according to (13).

5: **end for**

With high probability, Algorithm 3 admits a regret rate bounded as in Theorem 6.

Theorem 6. *Fix any T and $i \in [K]$. Algorithm 3 satisfies*

$$\mathbb{E}\left[\mathrm{Reg}_i^{\mathrm{adv}}(T)\right] \lesssim \sqrt{\left(1 + \sum_{j\in[K]} \frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right) T}.$$

In addition, if the estimators $\widehat{q}_{t,ij}$ are exact, then with probability at least $1 - 7\epsilon$ (for any $\epsilon \in \left(0, \frac{1}{(1-\rho)KT}\right)$), Algorithm 3 achieves, for $\forall i \in [K]$,

$$\mathrm{Reg}_i^{\mathrm{adv}}(T) \lesssim \sqrt{\left(1 + 2\sum_{j\in[K]} \frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right) T \log(KT/\delta)}.$$

In Figure 1, we provide a plot of $f(x) = \frac{1-\sqrt{x}}{1+\sqrt{x}}$ with $x \in [0, 1]$. This shows that the regret dependence on the connectivity of the graph is very nonlinear. To prove this theorem, we need to handle variance of estimators $\widehat{q}_{t,ij}$ and disentangle the randomness by using the odd-even trick.

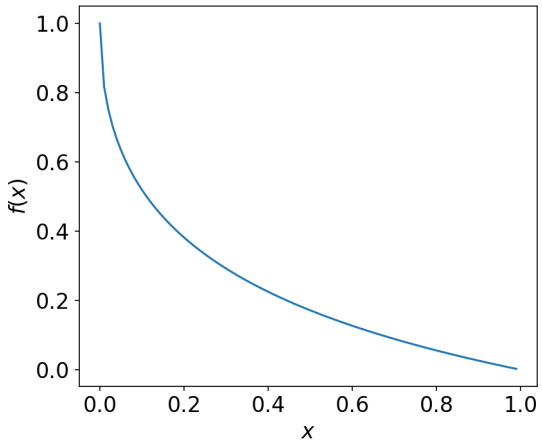

Figure 1: A plot of function $f(x) = \frac{1-\sqrt{x}}{1+\sqrt{x}}$, $x \in [0,1]$. This shows that in Theorem 6, the dependence on graph connectivity is highly non-linear.

Details can be found in Appendix C.

4.1 Lower Bound for the Adversarial Setting

We also provide a lower bound for the adversarial case, which is summarized in Theorem 7. This lower bound shows no algorithm can always perform better than $\Omega\left(\sqrt{T}\right)$ with high probability.

Theorem 7. *Fix any $T > \sqrt{128 \log 8}$ and $\sigma < \frac{1}{7}$. On a graph of K transient nodes where any pair of transient nodes are connected with probability p, there exists a sequence of edge lengths and a node $i \in [K]$, such that for any policy, the regret incurred by any π against i satisfies*

$$\mathbb{P}\left(\mathbb{E}\left[\mathrm{Reg}_i^{\mathrm{adv}}(T)\right] \geq \frac{1}{2}\sqrt{\frac{(1-Kp)\,\sigma^2 T}{\left(1+\frac{2p}{1-Kp}\right)}}\right) \geq \frac{3}{16}.$$

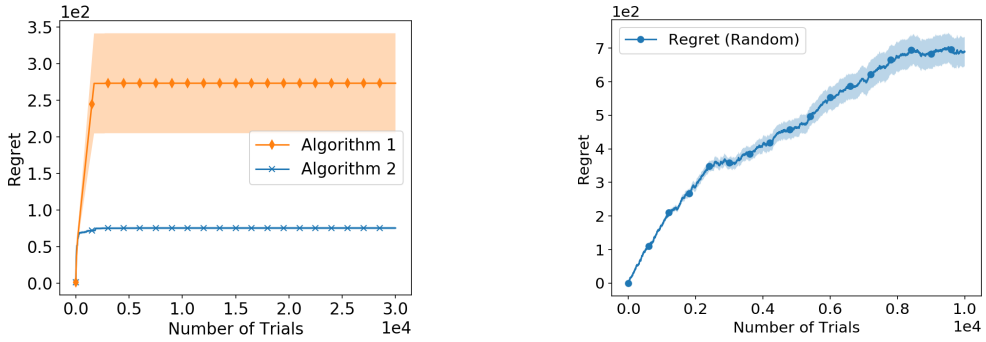

Figure 2: *Left:* Results for stochastic setting. *Right:* Results for the adversarial setting. Each line is averaged over 10 runs. The shaded areas (around the solid lines) indicate one standard deviation below and above the average. For the adversarial case, the regret is measured against arm 0, which, by construction, has largest hitting time in expectation.

To prove this theorem, we first tweak the problem so that the adversary chooses distribution over edge lengths (Proposition 4 in Appendix C.3), and give a bound under this randomization.

Similar to the proof for Theorem 5, we construct two problem instances $\mathfrak{J}$ and $\mathfrak{J}'$ such that no policy can quickly tell the difference between them. We again use $\mathbb{P}_{\mathfrak{J},\pi}$ (resp. $\mathbb{P}_{\mathfrak{J}',\pi}$) to denote the probability measure generated by playing π on $\mathfrak{J}$ (resp. $\mathfrak{J}'$). Similar to the proof for Theorem 5, the sample space (on which

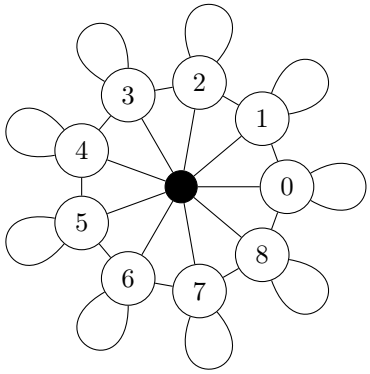

Figure 3: The network structure for experiments. The dark node at the center is the absorbing node $*$, and nodes labelled with numbers are transient nodes. Nodes without edges connecting them visits each other with zero probability.

both $\mathbb{P}_{\mathfrak{J},\pi}$ and $\mathbb{P}_{\mathfrak{J}',\pi}$ is defined) is different from a regular bandit problem. If we execute π for T epochs, the sample space is then $\left(\cup_{h=1}^{\infty}\left([0,1]\times[K]\right)^h\right)^T$ (with the σ-algebra generated by singletons in $[K]$ and Borel sets in $[0,1]$). This is because (1) each trajectory can be arbitrarily long, (2) the nodes on the trajectory can be any of $[K]$, and (3) each edge on the trajectory can take values from $[0,1]$. Specifically, for each trajectory $\mathcal{P}_{t,J_t}=\left(X_{t,0},L_{t,1},X_{t,1},\ldots,L_{t,H_{t,J_t}},X_{t,H_{t,J_t}}\right)$, H_{t,J_t} can be any positive integer, $X_{t,i}$ can be any integer from $[K]$, and each $L_{t,i}$ can be any number from $[0,1]$. This sample space is fundamentally different from the space generated by interaction with a standard K-armed bandit problem for T rounds, which is $[0,1]^{KT}$. We use the Markov property of random walks to handle this difficulty. Specifically, for any fixed i and j, conditioning on $X_{t,i}=j$ being known, the space generated by $L_{t,i+1},X_{t,i+1},L_{t,i+2},X_{t,i+2},\ldots$ is identical to the space generated by $L_{t,1},X_{t,1},L_{t,2},X_{t,2},\ldots$ conditioning on $X_{t,0}=j$ being fixed. A proof of Theorem 7 can be found in Appendix C.2.

5 Experiments

We deploy our algorithms on a problem with 9 transient nodes. For stochastic setting (left subfigure in Figure 2), we have $l_{ij}=1$ for $i\in[9]$ and $j\in[9]\cup\{*\}$. The transition probabilities among transient nodes are

$$m_{ij}=\begin{cases}0.6,\ \text{if } i=j=1,\\ 0.4,\ \text{if } i=j \text{ and } i\neq 1,\\ 0.1,\ \text{if } i=j\pm 1 \mod 9,\\ 0,\ \text{otherwise.}\end{cases}$$

For adversarial setting (right subfigure in Figure 2), the transition probabilities among transient nodes are

$$m_{ij}=\begin{cases}0.3,\ \text{if } i=j,\\ 0.1,\ \text{if } i=j\pm 1 \mod 9,\\ 0,\ \text{otherwise.}\end{cases}$$

The edge lengths are sampled from Gaussian distributions and truncated to between 0 and 1. Specifically, for all $t=1,2,\cdots,T$,

$$l_{ij}^{(t)}=\begin{cases}\text{clip}_{[0,1]}\left(W_t+0.5\right),\ \text{if } i=0 \text{ and } j=*\\ \text{clip}_{[0,1]}\left(W_t\right),\ \text{if } i\neq 0 \text{ and } j=*\\ 1,\ \text{otherwise,}\end{cases}$$

where $\text{clip}_{[0,1]}(z)$ takes a number z and clips it to $[0,1]$, and $W_t\overset{i.i.d.}{\sim}\mathcal{N}(0.5,0.1)$. The results (Figure 2) empirically consolidate our theorems:

6 Conclusion

In this paper, we propose the problem (**P**) motivated by online advertisement, and study it from a bandit learning perspective. We study both the stochastic setting and the adversarial setting for this problem. We

provide algorithms for both settings, and derive matching lower bounds. Our paper provides a comprehensive study of this important problem (**P**) from a bandit learning perspective.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1.

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pages 23–35. PMLR.

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

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

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331. IEEE.

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

Auer, P. and Ortner, R. (2010). UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65.

Bertsekas, D. P. and Tsitsiklis, J. N. (1991). An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

Bubeck, S. and Slivkins, A. (2012). The best of both worlds: stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1.

Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., and Warmuth, M. K. (1997). How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485.

Chen, L., Li, J., and Qiao, M. (2017). Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, pages 101–110.

Csiszár, I. and Shields, P. C. (2004). *Information theory and statistics: A tutorial*. Now Publishers Inc.

Garivier, A. and Cappé, O. (2011). The KL–UCB algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory*, pages 359–376.

Gerchinovitz, S. and Lattimore, T. (2016). Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems*, pages 1198–1206.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42.

Kleinberg, R., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *ACM Symposium on Theory of Computing*, pages 681–690. ACM.

Kocák, T., Neu, G., Valko, M., and Munos, R. (2014). Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, pages 613–621.

Krause, A. and Ong, C. S. (2011). Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pages 2447–2455.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

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

Littlestone, N. and Warmuth, M. K. (1994). The weighted majority algorithm. *Information and computation*, 108(2):212–261.

Maillard, O.-A., Munos, R., and Stoltz, G. (2011). A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Conference On Learning Theory*, pages 497–514.

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

Mannor, S. and Tsitsiklis, J. N. (2004). The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Seldin, Y. and Slivkins, A. (2014). One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295.

Slivkins, A. (2014). Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010a). Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010b). Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022.

Tao, T. and Vu, V. (2015). Random matrices: universality of local spectral statistics of non-Hermitian matrices. *The Annals of Probability*, 43(2):782–874.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

A Notations

Table 1: Table of common notations.

Symbol	Definition
$[K]$	Transient nodes $[K] = \{0, 1, \cdots, K-1\}$.
$*$	Absorbing node.
M	Transition matrix among transient nodes.
m_{ij}	The probability of visiting j from i.
ρ	Upper bound on the ∞-norm of matrix M.
α_v, α	Hitting centrality of node v. $\alpha = \min_v \alpha_v$
$l_{ij}^{(t)}$	Edge length from i to j at epoch t.
J_t	Node played at time t.
$\mathcal{P}_{t,j}$	Trajectory of playing node j at time t.
$\mathcal{L}(\mathcal{P}_{t,j})$	Length of trajectory $\mathcal{P}_{t,j}$. This is also the hitting time of playing node j at epoch t.
$l_{t,j}$	$l_{t,j} := \mathbb{E}\left[\mathcal{L}(\mathcal{P}_{t,j})\right]$.
$H_{t,j}$	Number of edges in the trajectory of playing node j at epoch t.
$N_t(v)$	Number of times v is played up to epoch t.
$N_t^+(v)$	Number of times v is covered by some trajectory up to epoch t.
$\widetilde{Z}_{v,n}$	Nontruncated estimation for $\mathcal{L}(\mathcal{P}_{t,v})$ (in stochastic setting). See (5).
$Z_{t,j}$	Sum of edge lengths from the first occurrence of node j to $*$ in $\mathcal{P}_{t,J_t}$ (for adversarial setting).
$\widehat{Z}_{t,j}$	Estimator for $Z_{t,j}$ (in adversarial setting).
p_{tj}	Probability of playing j in epoch t (in adversarial setting).

B Proofs for the Stochastic Setting

The proofs (for Theorems 1 and 2) will use machinery from UCB analysis for standard multi-armed bandit problems Auer et al. (2002a).

B.1 Concentrations of Estimators

Note. In the stochastic setting, the graph is time-invariant. For easier reference, we define random variables Z_v such that Z_v has the same distribution as $\mathcal{L}(\mathcal{P}_{t,v})$ for all t. Then, in the stochastic setting,

$$\Delta_v = \mathbb{E}[Z_{v^*}] - \mathbb{E}[Z_v], \quad \text{where} \quad v^* \in \arg\max_v \mathbb{E}[Z_v]. \tag{17}$$

In this part, we derive concentration bounds for hitting time estimators $\widetilde{Z}_{v,n}$. To start with, we show that the hitting times are sub-exponential.

Lemma 1. *For any $v \in [K]$, $i = 1, 2, \cdots$ and any integer $x > 0$, we have*

$$\mathbb{P}\left(Y_{v,k_{v,i}} \geq x\right) \leq \frac{\rho^x}{1-\rho}.$$

where ρ is defined in Assumption 1.

Proof. Let M be the transition matrix among transient nodes.

Since for any (v, i, x),

$$\begin{aligned}&\left\{Y_{v,k_{v,i}} \geq x\right\}\\ =&\{\text{a random walk starting from } v \text{ does not reach the absorbing node in } x \text{ steps}\},\end{aligned}$$

we have,

$$\begin{aligned}\mathbb{P}\left(Y_{v,k_{v,i}} \geq x\right) &\leq \mathbb{P}(\{\text{random walk starting from } v \text{ does not terminate in } x \text{ steps}\})\\ &= \sum_{h=x}^{\infty} \mathbb{P}(\{\text{random walk starting from } v \text{ terminates at step } h\}).\end{aligned} \tag{18}$$

Writing out the probability in (18) gives

$$\mathbb{P}\left(Y_{v,k_{v,i}} \geq x\right) \leq \sum_{l=x}^{\infty}\sum_{j=1}^{K}\left[M^l\right]_{ij} \leq \sum_{l=x}^{\infty}\|M^l\|_{\infty} \leq \sum_{l=x}^{\infty}\rho^l \leq \frac{\rho^x}{1-\rho}.$$

□

For the concentration results, we use Lemmas 2 and 3.

Lemma 2 (Proposition 34 by Tao and Vu (2015))**.** *Consider a martingale sequence $X_1, X_2, \cdots$ adapted to filtration $\mathcal{F}_1, \mathcal{F}_2 \cdots$. For constants $c_1, c_2, \cdots < \infty$, we have*

$$\mathbb{P}\left(|X_n - X_0| > \lambda\sqrt{\sum_{i=1}^{n} c_i^2}\right) \leq 2\exp\left(-\frac{\lambda^2}{2}\right) + \sum_{i=1}^{n}\mathbb{P}(|X_i - X_{i-1}| > c_i). \tag{19}$$

Lemma 2 is an extension of the Azuma's inequality with an extra term bounding the probability of any term in the martingale difference sequence being unbounded.

Lemma 3. *For any transient node $v \in [K]$, if $N_t^+(v) > 0$, we have*

$$\mathbb{P}\left(\left|\widetilde{Z}_{v,N_t^+(v)} - \mathbb{E}\left[\mathcal{L}\left(\mathcal{P}_{s,v}\right)\right]\right| \geq \sqrt{\frac{8\xi_t \log t}{N_t^+(v)}}\right) \leq 3t^{-4}, \quad \forall s, t \in \mathbb{N}_+ \tag{20}$$

where $\xi_t = \max\left\{1 + \frac{\rho}{(1-\rho)^2}, \frac{\log(1-\rho)}{\log\rho} + \frac{5\log t}{\log 1/\rho}\right\}$.

Proof. By Lemma 1, we have, when $x \geq \mathbb{E}\left[Y_{v,k_{v,i}}\right]$,

$$\begin{aligned}\mathbb{P}\left(\left|Y_{v,k_{v,i}} - \mathbb{E}\left[Y_{v,k_{v,i}}\right]\right| \geq x\right) \leq & \mathbb{P}\left(Y_{v,k_{v,i}} - \mathbb{E}\left[Y_{v,k_{v,i}}\right] \geq x\right) + \mathbb{P}\left(-Y_{v,k_{v,i}} + \mathbb{E}\left[Y_{v,k_{v,i}}\right] \geq x\right) \\ \leq & \mathbb{P}\left(Y_{v,k_{v,i}} \geq x\right) + \mathbb{P}\left(Y_{v,k_{v,i}} \leq \mathbb{E}\left[Y_{v,k_{v,i}}\right] - x\right) \qquad \text{(the second term is zero.)} \\ \leq & \frac{\rho^x}{1-\rho}. \end{aligned} \tag{21}$$

Also by Lemma 1, $\mathbb{E}\left[Y_{v,k_{v,i}}\right] = \sum_{x=0}^{\infty} \mathbb{P}\left(Y_{v,k_{v,i}} \geq x\right) \leq 1 + \sum_{x=1}^{\infty} \frac{\rho^x}{1-\rho} \leq 1 + \frac{\rho}{(1-\rho)^2}$. Since (21) is true only for $x \geq \mathbb{E}\left[Y_{v,k_{v,i}}\right]$, we have, by setting

$$\xi_t = \max\left\{1 + \frac{\rho}{(1-\rho)^2}, \frac{\log(1-\rho)}{\log \rho} + \frac{5\log t}{\log 1/\rho}\right\},$$

$$\mathbb{P}\left(\left|Y_{v,k_{v,i}} - \mathbb{E}\left[Y_{v,k_{v,i}}\right]\right| \geq \xi_t\right) \leq t^{-5}. \tag{22}$$

Applying Lemma 2 gives

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{i=1}^{N_t^+(v)} \left[Y_{v,k_{v,i}} - \mathbb{E}\left[Y_{v,k_{v,i}}\right]\right]\right| \geq \lambda\sqrt{\sum_{i=1}^{N_t^+(v)} \xi_t}\right) \\ \leq & 2\exp\left(\frac{-\lambda^2}{2}\right) + \sum_{i=1}^{N_t^+(v)} \mathbb{P}\left(\left|Y_{v,k_{v,i}} - \mathbb{E}\left[Y_{v,k_{v,i}}\right]\right| \geq \xi_t\right) \\ \leq & 2\exp\left(\frac{-\lambda^2}{2}\right) + t^{-4}. \end{aligned} \tag{23}$$

At time t, we set $\lambda = \sqrt{8\log t}$, and from (23) we get

$$\begin{aligned}&\mathbb{P}\left(\left|\widetilde{Z}_{v,N_t^+(v)} - \mathbb{E}\left[\widetilde{Z}_{v,N_t^+(v)}\right]\right| \geq \sqrt{\frac{8\xi_t \log t}{N_t^+(v)}}\right) \\ = & \mathbb{P}\left(\left|\sum_{i=1}^{N_t^+(v)} \left[Y_{v,k_{v,i}} - \mathbb{E}\left[Y_{v,k_{v,i}}\right]\right]\right| \geq \sqrt{8\log t}\sqrt{\sum_{i=1}^{N_t^+(v)} \xi_t}\right) \leq 3t^{-4}. \end{aligned}$$

We now conclude the proof by using $\mathbb{E}\left[\widetilde{Z}_{v,N_t^+(v)}\right] = \mathbb{E}\left[\mathcal{L}\left(\mathcal{P}_{s,v}\right)\right]$ for all $s, t \in \mathbb{N}_+$. □

B.2 A Node is Visited from Other Nodes with High Probability

In this part, we show that with high probability, playing one node will reveal some information about other nodes.

Lemma 4. *For any $v \in V$ and a positive integer t*

$$\mathbb{P}\Big(N_t^+(v) - N_t(v) - \alpha_v\left(t - N_t(v)\right) \geq \lambda\Big) \leq \exp\left(-\frac{\lambda^2}{2t}\right). \tag{24}$$

Proof. Recall $\mathcal{P}_{t,J_t} = \left(X_{t,0}, L_{t,1}, X_{t,1}, L_{t,2}, X_{t,2}, \ldots, L_{t,H_{t,J_t}}, X_{t,H_{t,J_t}}\right)$ is the trajectory for epoch t and $X_{i,0}$ is the node played at epoch i. For a fixed node $v \in V$, consider the random variables $\left\{\mathbb{I}_{\left[v \in \mathcal{P}_{t,J_t} \setminus \{X_{t,0}\}\right]}\right\}_t$, which is the indicator that takes value 1 when v is covered in path $\mathcal{P}_{t,J_t}$ but is not played at t. From this definition, we have

$$\sum_{k=1}^{t} \mathbb{I}_{\left[v \in \mathcal{P}_{k,J_k} \setminus \{X_{k,0}\}\right]} = N_t^+(v) - N_t(v).$$

From definition of α_v, we have

$$\mathbb{E}\left[N_t^+(v) - N_t(v)\right] = \mathbb{E}\left[\sum_{k=1}^{t} \mathbb{I}_{[v \in \mathcal{P}_i \setminus \{X_{k,0}\}]}\right] \geq \alpha_v\left(t - \mathbb{E}\left[N_t(v)\right]\right).$$

Thus by one-sided Azuma's inequality, we have for any $\lambda > 0$,

$$\mathbb{P}\Big(N_t^+(v) - N_t(v) - \alpha_v\left(t - N_t(v)\right) \geq \lambda\Big) \leq \exp\left(-\frac{\lambda^2}{2t}\right). \tag{25}$$

□

B.3 Proof of Theorem 1

Theorem 1. *Suppose Assumption 1 holds. Algorithm 1 achieves a constant regret that only depends on* $\alpha_v, \alpha_{v^*}, \rho$, *and* Δ_v: $\mathrm{Reg}(T) \leq \widetilde{\mathcal{O}}\left(\sum_{v:\Delta_v>0}\left(\frac{1}{\min\{\alpha_v,\alpha_{v^*}\}(1-\rho)^2\Delta_v} + \Delta_v\right)\right)$, *where* v^* *is the optimal node (the node with maximum hitting time), and* $\widetilde{\mathcal{O}}$ *omits absolute constants and logarithmic dependence on problem intrinsics.*

Proof. First we define

$$T^{(1)}_{\min,v} := \min\left\{t \in \mathbb{N} : \left(\sqrt{\frac{8\xi_t \log t}{\alpha_v t - \sqrt{t\log t}}} \leq \frac{\Delta_v}{2}\right) \wedge \left(\sqrt{\frac{8\xi_t \log t}{\alpha_{v^*} t - \sqrt{t\log t}}} \leq \frac{\Delta_v}{2}\right)\right\}, \tag{26}$$

For a sub-optimal node v and a time $t \in \mathbb{N}$, we consider

$$\mathcal{E}_{v,t} = \left\{N_t^+(v) - N_t(v) - (t - N_t(v))\alpha_v \geq -\sqrt{4t\log t}\right\}. \tag{27}$$

By Lemma 4, $\mathcal{E}_{v,t}$ is true with probability at least $1 - \frac{1}{t^2}$.
For simplicity, in (and only in) this part of proof, we write

$$B_{n,t} = \sqrt{\frac{8\xi_t \log t}{n}},$$

where $\xi_t := \max\left\{1 + \frac{\rho}{(1-\rho)^2}, \frac{\log(1-\rho)}{\log\rho} + \frac{5\log t}{\log 1/\rho}\right\}$.

We have, for any sub-optimal node v, the probability of v being played at time t satisfies

$$\begin{aligned}
\mathbb{P}\left(J_t = v\right) \leq& \mathbb{P}\left(\widetilde{Z}_{v,N_t^+(v)} \geq \widetilde{Z}_{v^*,N_t^+(v^*)}\right) && (\ v^* \in \arg\max_{v\in V}\mathbb{E}\left[Z_v\right]) \\
=& \mathbb{P}\left(\widetilde{Z}_{v,N_t^+(v)} - \mathbb{E}\left[Z_v\right] - \left(\widetilde{Z}_{v^*,N_t^+(v^*)} - \mathbb{E}\left[Z_{v^*}\right]\right) \geq \Delta_v\right) && (\ \Delta_v = \mathbb{E}\left[Z_{v^*}\right] - \mathbb{E}\left[Z_v\right]) \\
\leq& \Big[\mathbb{P}\left(2\max\{B_{N_t^+(v),t}, B_{N_t^+(v^*),t}\} > \Delta_v\right) + \mathbb{P}\left(\widetilde{Z}_{v,N_t^+(v)} \geq \mathbb{E}\left[Z_v\right] + B_{N_t^+(v),t}\right) && (28) \\
&+ \mathbb{P}\left(\widetilde{Z}_{v^*,N_t^+(v^*)} \leq \mathbb{E}\left[Z_{v^*}\right] - B_{N_t^+(v^*),t}\right)\Big],
\end{aligned}$$

where in (28) we use

$$\begin{aligned}
&\left\{\widetilde{Z}_{v,N_t^+(v)} - \mathbb{E}\left[Z_v\right] - \left(\widetilde{Z}_{v^*,N_t^+(v^*)} - \mathbb{E}\left[Z_{v^*}\right]\right) \geq \Delta_v\right\} \\
\Longrightarrow &\left\{2\max\{B_{N_t^+(v),t}, B_{N_t^+(v^*),t}\} > \Delta_v\right\} \\
&\quad\bigcup\left\{\widetilde{Z}_{v^*,N_t^+(v^*)} \geq \mathbb{E}\left[Z_{v^*}\right] + B_{N_t^+(v^*),t}\right\} \\
&\quad\bigcup\left\{\widetilde{Z}_{v,N_t^+(v)} \leq \mathbb{E}\left[Z_v\right] - B_{N_t^+(v),t}\right\},
\end{aligned}$$

which can be verified by checking its contrapositive statement.

When (1) event $\mathcal{E}_{v,t}$ is true, (2) $\sqrt{\frac{8\xi_t\log t}{\alpha_v t - \sqrt{t\log t}}} \leq \frac{\Delta_v}{2}$ and (3) $\sqrt{\frac{8\xi_t\log t}{\alpha_{v^*} t - \sqrt{t\log t}}} \leq \frac{\Delta_v}{2}$, we have

$$2\max\{B_{N_t^+(v),t}, B_{N_t^+(v^*),t}\} \leq \Delta_v.$$

Thus when $t \geq T^{(1)}_{\min,v}$, we have

$$\begin{aligned}
&\mathbb{P}\left(2\max\{B_{N_t^+(v),t}, B_{N_t^+(v^*),t}\} \geq \Delta_v\right) \\
=&\mathbb{P}\left(2\max\{B_{N_t^+(v),t}, B_{N_t^+(v^*),t}\} \geq \Delta_v \middle| \mathcal{E}_{v,t}\right)\mathbb{P}\left(\mathcal{E}_{v,t}\right) \\
&+\mathbb{P}\left(2\max\{B_{N_t^+(v),t}, B_{N_t^+(v^*),t}\} \geq \Delta_v \middle| \overline{\mathcal{E}_{v,t}}\right)\left[1-\mathbb{P}\left(\mathcal{E}_{v,t}\right)\right] \\
\leq& 1-\mathbb{P}\left(\mathcal{E}_{v,t}\right) \leq \frac{1}{t^2}.
\end{aligned} \tag{29}$$

Also by Lemma 3, we have

$$\mathbb{P}\left(\widetilde{Z}_{v^*,N_t^+(v^*)} \geq \mathbb{E}\left[Z_{v^*}\right] + B_{N_t^+(v^*),t}\right) \leq \frac{2}{t^2}, \quad \mathbb{P}\left(\widetilde{Z}_{v,N_t^+(v)} \leq \mathbb{E}\left[Z_v\right] - B_{N_t^+(v),t}\right) \leq \frac{2}{t^2}. \tag{30}$$

We can now combine (28), (29) and (30) to get

$$\begin{aligned}
\mathbb{E}\left[N_T(v)\right] =& \sum_{t=1}^{T}\mathbb{P}\left(J_t = v\right) \\
\leq& T^{(1)}_{\min,v} + \sum_{t=T^{(1)}_{\min,v}}^{T}\mathbb{P}\left(v_t = v\right) \\
\leq& T^{(1)}_{\min,v} + \sum_{t=T^{(1)}_{\min,v}}^{T}\Bigg[\mathbb{P}\left(2\max\{B_{N_t^+(v),t}, B_{N_t^+(v^*),t}\} \geq \Delta_v\right) \\
&+\mathbb{P}\left(\widetilde{H}_{v,N_t^+(v)} \leq \mathbb{E}\left[H_v\right] - B_{N_t^+(v),t}\right) \\
&+\mathbb{P}\left(\widetilde{H}_{v^*,N_t^+(v^*)} \geq \mathbb{E}\left[H_{v^*}\right] + B_{N_t^+(v^*),t}\right)\Bigg] \\
\leq& T^{(1)}_{\min,v} + \sum_{t=1}^{\infty}\left[\frac{1}{t^2}+\frac{2}{t^2}+\frac{2}{t^2}\right] \leq T^{(1)}_{\min,v} + \frac{5\pi^2}{6},
\end{aligned} \tag{31}$$

where on the last line we use (29) and (30).

Finally, we use the Wald's equation to get

$$\begin{aligned}
\mathrm{Reg}(T) =& \sum_{t=1}^{T}\left(\mathbb{E}\left[\mathcal{L}\left(\mathcal{P}_{t,v^*}\right)\right] - \mathbb{E}\left[\mathcal{L}\left(\mathcal{P}_{t,J_t}\right)\right]\right) \\
=& \sum_{t=1}^{T}\sum_{v\in[K]}\left(\mathbb{E}\left[Z_{v^*}\right] - \mathbb{E}\left[Z_v\right]\right)\mathbb{P}\left(J_t = v\right) = \sum_{v\in[K],\Delta_v>0}\Delta_v\mathbb{E}\left[N_T(v)\right].
\end{aligned}$$

From here we use $T^{(1)}_{\min,v} = \widetilde{\mathcal{O}}\left(\frac{1}{\min\{\alpha_v,\alpha_{v^*}\}(1-\rho)^2\Delta_v^2}\right)$ to conclude the proof. □

B.4 Proof of Theorem 2

Theorem 2. *On a problem instance that satisfies Assumption 1, Algorithm 2 achieves constant regret of order* $\widetilde{\mathcal{O}}\left(\sum_{v:\Delta_v>0}\left(\Delta_v + \frac{1}{(1-\rho)^2\Delta_v}\right)\right)$*, where* $\widetilde{\mathcal{O}}$ *omits absolute constants and logarithmic dependence on problem intrinsics.*

Proof. First we define

$$T^{(2)}_{\min,v} := \min\left\{t \in \mathbb{N} : t\alpha_v - \sqrt{t\log t} \geq \frac{32\xi_t\log t}{\Delta_v^2}\right\}. \tag{32}$$

The proof of this theorem is developed in three steps.

Step 1: When $t \geq T^{(2)}_{\min,v}$, $\widetilde{C}_{N_t^+(v),t} \leq 2\Delta_v$ with high probability.

For any $t \in \mathbb{N}$ and node $v \in V$, consider the following event.

$$\mathcal{E}_{v,t} := \left\{ N_t^+(v) - N_t(v) - (t - N_t(v))\alpha_v \geq -\sqrt{t \log t} \right\}. \tag{33}$$

By Lemma 4, we have for any $v \in V$, with probability at least $1 - \frac{1}{t^2}$,

$$N_t^+(v) - N_t(v) - (t - N_t(v))\alpha_v \geq -\sqrt{t \log t}.$$

For a sub-optimal node v, when

$$N_t^+(v) \geq \frac{32\xi_t \log t}{\Delta_v^2}, \tag{34}$$

we have

$$\widetilde{C}_{N_t^+(v),t} = \sqrt{\frac{8\xi_t \log t}{N_t^+(v)}} \leq \frac{1}{2}\Delta_v. \tag{35}$$

For $t \geq T^{(2)}_{\min,v}$ where $T^{(2)}_{\min,v}$ is defined in (32), under event $\mathcal{E}_{v,t}$, we have

$$\begin{aligned} N_t^+(v) &\geq \alpha_v t + (1 - \alpha_v) N_t(v) - \sqrt{t \log t} && \text{(under event } \mathcal{E}_{v,t}) \\ &\geq \alpha_v t - \sqrt{t \log t} && \text{(since } \alpha_v \leq 1) \\ &\geq \frac{32\xi_t \log t}{\Delta_v^2}. && (36) \end{aligned}$$

From above, we know (34) is true as long as $\mathcal{E}_{v,t}$ is true. Thus for any $t \geq T^{(2)}_{\min,v}$,

$$\mathbb{P}\left(\widetilde{C}_{N_t^+(v),t} \leq \frac{1}{2}\Delta_v \right) = \mathbb{P}\left(N_t^+(v) \geq \frac{32\xi_t \log t}{\Delta_v^2} \right) \geq \mathbb{P}\left(\mathcal{E}_{v,t} \right) \geq 1 - \frac{1}{t^2}.$$

In words, $\widetilde{C}_{N_t^+(v),t} \leq \frac{1}{2}\Delta_v$ as long as (1) $t \geq T^{(2)}_{\min,v}$ and (2) $\mathcal{E}_{v,t}$ is true.

Step 2: After time $T^{(2)}_{\min,v}$, a sub-optimal node is played constant number of times (in expectation).

For easier reference, for any $t \in \mathbb{N}$ and any $v \in V$, we write

$$w_{v,t} := \frac{32\xi_t \log t}{\Delta_v^2}. \tag{37}$$

In Step 1 (Eq. 36), we have shown that for any $t \geq$, event $\mathcal{E}_{v,t}$ implies

$$N_t^+(v) \geq w_{v,t}. \tag{38}$$

After time $T^{(2)}_{\min,v}$, we can bound the number of times a sub-optimal arm is played (in expectation) by

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=T^{(2)}_{\min,v}}^{T} \mathbb{I}_{[J_t = v]} \right] \\ &= \sum_{t=T^{(2)}_{\min,v}}^{T} \left\{ \mathbb{E}\left[\mathbb{I}_{[J_t=v]} \middle| \mathcal{E}_{v,t} \right] \mathbb{P}\left(\mathcal{E}_{v,t} \right) + \mathbb{E}\left[\mathbb{I}_{[J_t=v]} \middle| \overline{\mathcal{E}_{v,t}} \right] \left(1 - \mathbb{P}\left(\mathcal{E}_{v,t} \right) \right) \right\} \\ &\leq \left(\sum_{t=T^{(2)}_{\min,v}}^{T} \mathbb{P}\left(J_t = v, t \geq T^{(2)}_{\min,v} \middle| \mathcal{E}_{v,t} \right) \right) + \sum_{t=1}^{\infty} \frac{1}{t^2} \\ &\leq \sum_{t=T^{(2)}_{\min,v}}^{T} \mathbb{P}\left(J_t = v, t \geq T^{(2)}_{\min,v} \middle| \mathcal{E}_{v,t} \right) + \frac{\pi^2}{6} \\ &\leq \sum_{t=T^{(2)}_{\min,v}}^{T} \mathbb{P}\left(\widetilde{Z}_{v,N_t^+(v)} + \widetilde{C}_{N_t^+(v),t} \geq \widetilde{Z}_{v^*,N_t^+(v^*)} + \widetilde{C}_{N_t^+(v^*),t},\ t \geq T^{(2)}_{\min,v} \middle| \mathcal{E}_{v,t} \right) + \frac{\pi^2}{6} \end{aligned} \tag{39}$$

where (39) uses

$$\left\{ \text{ a node } v \text{ is played at } t \right\} \Rightarrow \left\{ \widetilde{Z}_{v,N_t^+(v)} + \widetilde{C}_{N_t^+(v),t} \geq \widetilde{Z}_{v^*,N_t^+(v^*)} + \widetilde{C}_{N_t^+(v^*),t} \right\}.$$

We then follow the argument by Auer et al. (2002a), and bound (39) by allowing $N_t^+(v)$ to take any values in $[w_{v,t}, t]$ (due to Eq. 38), and allowing $N_t^+(v^*)$ to take any values in $[1, t]$.

This gives,

$$\begin{aligned}
&\mathbb{P}\left(\widetilde{Z}_{v,N_t^+(v)} + \widetilde{C}_{N_t^+(v),t} \geq \widetilde{Z}_{v^*,N_t^+(v^*)} + \widetilde{C}_{N_t^+(v^*),t},\ t \geq T^{(2)}_{\min,v} \middle| \mathcal{E}_{v,t} \right) \\
\leq &\sum_{s=w_{v,t}}^{t} \sum_{s^*=1}^{t} \mathbb{P}\left(\widetilde{Z}_{v,s} + \widetilde{C}_{s,t} \geq \widetilde{Z}_{v^*,s^*} + \widetilde{C}_{s^*,t},\ t \geq T^{(2)}_{\min,v} \middle| \mathcal{E}_{v,t} \right). \qquad (40)
\end{aligned}$$

As is used by Auer et al. (2002a), when the event

$$\left\{ \widetilde{Z}_{v,s} + \widetilde{C}_{s,t} \geq \widetilde{Z}_{v^*,s^*} + \widetilde{C}_{s^*,t} \right\}$$

is true, at least one of the following three must be true:

$$\widetilde{Z}_{v,s} - \widetilde{C}_{s,t} \leq \mathbb{E}\left[Z_v\right], \qquad (41)$$

$$\widetilde{Z}_{v^*,s^*} + \widetilde{C}_{s^*,t} \geq \mathbb{E}\left[Z_{v^*}\right], \qquad (42)$$

$$\mathbb{E}\left[Z_{v^*}\right] < \mathbb{E}\left[Z_v\right] + 2\widetilde{C}_{s,t}. \qquad (43)$$

By Step 1 (Eq. 35), we know (43) is false for $s \geq w_{v,t}$ and $t \geq T^{(2)}_{\min,v}$. Thus one of (41) and (42) must be true. Therefore we can continue from (40) to get

$$\begin{aligned}
&\sum_{s=w_{v,t}}^{t} \sum_{s^*=1}^{t} \mathbb{P}\left(\widetilde{Z}_{v,s} + \widetilde{C}_{s,t} \geq \widetilde{Z}_{v^*,s^*} + \widetilde{C}_{s^*,t}, t \geq T^{(2)}_{\min,v} \middle| \mathcal{E}_{v,t} \right) \\
\leq &\sum_{s=w_{v,t}}^{t} \sum_{s^*=1}^{t} \left[\mathbb{P}\left(\widetilde{Z}_{v,s} + \widetilde{C}_{s,t} \leq \mathbb{E}\left[Z_v\right] \right) + \mathbb{P}\left(\widetilde{Z}_{v,s} + \widetilde{C}_{s,t} \leq \mathbb{E}\left[Z_v\right] \right) \right] \\
\leq &\sum_{s=w_{v,t}}^{t} \sum_{s^*=1}^{t} \left(\frac{3}{t^4} + \frac{3}{t^4} \right) \qquad (44) \\
\leq &\frac{6}{t^2} \qquad (45)
\end{aligned}$$

where (44) uses Lemma 3.

Combining (39) and (45) gives

$$\mathbb{E}\left[\sum_{t=T^{(2)}_{\min,v}}^{T} \mathbb{I}_{[J_t=v]} \right] \leq \sum_{t=T^{(2)}_{\min,v}}^{T} \frac{6}{t^2} + \frac{\pi^2}{6} \leq \frac{7\pi^2}{6}, \qquad (46)$$

which concludes Step 2.

Step 3: Up to time $T^{(2)}_{\min,v}$, a sub-optimal node v is played $\mathcal{O}\left(polylog\left(T^{(2)}_{\min,v}, \frac{1}{\Delta_v}\right)\right)$ number of times

Recall $N_t(v)$ is the number of times we play node v up to time t.

For any integer w, we have

$$\begin{aligned}
\mathbb{E}\left[N_T(v)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}_{[J_t=v]}\right] \\
&\leq w + \mathbb{E}\left[\sum_{t=1}^{t=T^{(2)}_{\min,v}} \mathbb{I}_{[J_t=v,\, N_t(v)\geq w]} + \sum_{t=t=T^{(2)}_{\min,v}}^{T} \mathbb{I}_{[J_t=v]}\right] \\
&\leq w + \mathbb{E}\left[\sum_{t=1}^{t=T^{(2)}_{\min,v}} \mathbb{I}_{[J_t=v,\, N_t(v)\geq w]}\right] + \frac{7\pi^2}{6} \qquad (47) \\
&\leq w + \sum_{t=1}^{t=T^{(2)}_{\min,v}} \sum_{s^*=1}^{t} \sum_{s=w}^{t} \mathbb{P}\left(\widetilde{Z}_{v,s} + \widetilde{C}_{s,t} \geq \widetilde{Z}_{v^*,s^*} + \widetilde{C}_{s^*,t}\right) + \frac{7\pi^2}{6}, \qquad (48)
\end{aligned}$$

where (47) uses the result of Step 2, and (48) uses similar argument in Step 2 (Eq. 39 - 40).

We set $w := \left\lceil \frac{32\xi_{T^{(2)}_{\min,v}} \log T^{(2)}_{\min,v}}{\Delta_v^2} \right\rceil$, so that at time $t \leq T^{(2)}_{\min,v}$, for $s \geq w$,

$$\widetilde{C}_{s,t} = \sqrt{\frac{8\xi_t \log t}{s}} \leq \frac{1}{2}\Delta_v,$$

which mean (43) is false.

Also, by Lemma 3, we have

$$\mathbb{P}\left[\widetilde{Z}_{v,s_v} - \widetilde{C}_{s_v,t} \leq \mathbb{E}\left[Z_v\right]\right] \leq \frac{3}{t^4} \qquad \mathbb{P}\left[\widetilde{Z}_{v^*,s^*} + \widetilde{C}_{s^*,t} \geq \mathbb{E}\left[Z_{v^*}\right]\right] \leq \frac{3}{t^4}. \qquad (49)$$

Again we continue from (48) and use (41), (42) and (43) to get

$$\begin{aligned}
&\mathbb{E}\left[N_T(v)\right] \\
\leq &\left\lceil \frac{32\xi_{T^{(2)}_{\min,v}} \log T^{(2)}_{\min,v}}{\Delta_v^2} \right\rceil \\
&+ \sum_{t=1}^{T} \sum_{s^*=1}^{t} \sum_{s_v=w}^{t} \left(\mathbb{P}\left[\widetilde{Z}_{v,s_v} - \widetilde{C}_{s_v,t} \leq \mathbb{E}\left[Z_v\right]\right] + \mathbb{P}\left[\widetilde{Z}_{v^*,s^*} + \widetilde{C}_{s^*,t} \geq \mathbb{E}\left[Z_{v^*}\right]\right]\right) + \frac{7\pi^2}{6} \qquad (50) \\
\leq &\frac{32\xi_{T^{(2)}_{\min,v}} \log T^{(2)}_{\min,v}}{\Delta_v^2} + 1 + \frac{13\pi^2}{6}, \qquad (51)
\end{aligned}$$

where on the last line we use (49).

Finally, since $\mathrm{Reg}(T) = \sum_{v\in[K]} \Delta_v \mathbb{E}\left[N_T(v)\right]$, we have

$$Reg(T) \leq \sum_{v:\Delta_v>0} \left[\frac{32\xi_{T^{(2)}_{\min,v}} \log T^{(2)}_{\min,v}}{\Delta_v} + \left(1 + \frac{13\pi^2}{6}\right)\Delta_v\right],$$

where $T^{(2)}_{\min,v} = \widetilde{\mathcal{O}}\left(\frac{1}{\alpha_v(1-\rho)^2\Delta_v^2}\right)$ and $\xi_{T^{(2)}_{\min,v}} = \widetilde{\mathcal{O}}\left(\frac{1}{(1-\rho)^2}\right)$.

□

B.5 Proof of Theorem 3

Theorem 3. *Fix any $\beta \in \left(0, \frac{1}{2}\right)$. Define $X_{\min,v} := \min_{t\in\mathbb{N}}\left\{\frac{\sqrt{8\xi_t \log t}}{\sqrt{\alpha_v t - t^{\frac{1}{2}+\beta}}} \leq \frac{1}{2}\Delta_v\right\}$. For any $x, T \geq X_{\min,v}$ and any sub-optimal node v, we have*

$$\mathbb{P}\left(N_T(v) \geq x\right) \lesssim \frac{\sqrt{x}}{\Delta_v\sqrt{\alpha_v \log 1/\rho}} \exp\left(-\Delta_v\sqrt{\alpha_v x \log 1/\rho}\right) + \frac{1}{(1+\alpha_v)x^3} + mx\exp\left(-x^{2\beta}\right), \qquad (52)$$

where $\lesssim$ omits constants and terms that are polynomial in β. In particular, setting $\beta = \frac{1}{4}$ recovers the version in the main paper.

Proof. For any positive integer u, and $\beta \in \left(0, \frac{1}{2}\right)$, consider event

$$\mathcal{E}_{u,\beta} := \left\{ N_t^+(v) - N_t(v) - (t - N_t(v))\alpha_v \geq -t^{\frac{1}{2}+\beta}, \text{ for all } t \geq u \text{ and all } v \in V \right\}.$$

By Lemma 4 and a union bound, we have $\mathbb{P}\left(\mathcal{E}_{u,\beta}\right) \geq 1 - m\sum_{t=u}^{\infty} \exp\left(-u^{2\beta}\right) \gtrsim 1 - mx\exp\left(-x^{2\beta}\right)$. Under event $\mathcal{E}_{u,\beta}$, at any time $t \in [u, T]$, we have $N_t^+(v) \geq \alpha_v t - t^{\frac{1}{2}+\beta}$ for all $v \in V$.

We define an "alternative" index at time t by:

$$I'_{v,t} := \widetilde{Z}_{v,N_t^+(v)} + \sqrt{\frac{8\xi_t \log t}{\alpha_v t - t^{\frac{1}{2}+\beta}}}. \tag{53}$$

By Lemma 4, we know $I'_{v,t} \geq I_{v,N_t^+(v),t}$ with high probability.

Fix $\beta \in \left(0, \frac{1}{2}\right)$. For any $\tau \in \mathbb{R}$, any positive integer w, and any sub-optimal node v, we have

$$\begin{aligned}\left\{ N_T(v) \leq w \right\} \Leftarrow & \left\{ I'_{v,t} \leq \tau \text{ for all } t \in [w, T] \right\} \\ & \cap \mathcal{E}_{w,\beta} \\ & \cap \left\{ I_{v^*,s,s+w} > \tau \text{ for all } s \in \left[\alpha_v w - w^{\frac{1}{2}+\beta}, T\right] \right\}. \end{aligned} \tag{54}$$

This is because (54) ensures that after time w, the index of node v is always lower than the index of node v^*. Thus (54) implies that v is never played after time w.

By taking the contrapositive of (54), we have, for arbitrary $\tau \in \mathbb{R}$,

$$\begin{aligned}\{N_T(v) > w\} \Rightarrow & \left\{ \exists t \in [w, T] \ \ s.t.\ I'_{v,t} > \tau \right\} \\ & \cup \left\{ \exists s \in \left[\alpha_v w - w^{\frac{1}{2}+\beta}, T\right] \ \ s.t.\ I_{v^*,s,s+w} \leq \tau \right\} \\ & \cup \overline{\mathcal{E}_{w,\beta}}. \end{aligned} \tag{55}$$

By taking probability on both sides of (55), we get, for any $\tau \in \mathbb{R}$ and $w \in \mathbb{N}$, we have

$$\mathbb{P}\left(N_T(v) > w\right) \leq \sum_{t=w}^{T} \mathbb{P}\left(I'_{v,t} > \tau\right) + \sum_{s=\alpha_v w - w^{\frac{1}{2}+\beta}}^{T} \mathbb{P}\left(I_{v^*,s,w+s} \leq \tau\right) + \mathbb{P}\left(\overline{\mathcal{E}_{w,\beta}}\right).$$

By taking $\tau = \mathbb{E}\left[Z_{v^*}\right]$, and $w = x$, we get

$$\mathbb{P}\left(N_T(v) > x\right) \lesssim \sum_{t=x}^{T} \mathbb{P}\left(I'_{v,t} > \mathbb{E}\left[Z_{v^*}\right]\right) + \sum_{s=\alpha_v x - x^{\frac{1}{2}+\beta}}^{T} \mathbb{P}\left(I_{v^*,s,w+s} \leq \mathbb{E}\left[Z_{v^*}\right]\right) + mx\exp\left(-x^{2\beta}\right). \tag{56}$$

For $t \geq x \geq X_{\min,v}$, we have

$$\sqrt{\frac{8\xi_t \log t}{\alpha_v t - t^{\frac{1}{2}+\beta}}} \leq \frac{1}{2}\Delta_v. \tag{57}$$

Thus by Lemma 3, under event $\mathcal{E}_{w,t}$, we have $N_t^+(v) \geq \alpha_v t - t^{\frac{1}{2}+\beta}$. Thus

$$\begin{aligned}\mathbb{P}\left(I'_{v,t} > \mathbb{E}\left[Z_{v^*}\right]\right) &= \mathbb{P}\left(\widetilde{Z}_{v,N_t^+(v)} + \sqrt{\frac{8\xi_t \log t}{\alpha_v t - t^{\frac{1}{2}+\beta}}} > \mathbb{E}\left[Z_v\right] + \Delta_v\right) \\ &\leq \mathbb{P}\left(\widetilde{Z}_{v,N_t^+(v)} > \mathbb{E}\left[Z_v\right] + \frac{1}{2}\Delta_v\right) \\ &\leq \max\left\{ 3\exp\left(-\frac{\Delta_v^2(1-\rho)^3}{8\rho} N_t^+(v)\right), 3\exp\left(-\Delta_v\sqrt{\frac{N_t^+(v)\log 1/\rho}{10}}\right)\right\} \\ &\lesssim \max\left\{ 3\exp\left(-\frac{\Delta_v^2(1-\rho)^3}{8\rho} \alpha_v t\right), 3\exp\left(-\Delta_v\sqrt{\frac{\alpha_v t\log 1/\rho}{10}}\right)\right\}. \end{aligned}$$

The above gives

$$\sum_{t=x}^{\infty} \mathbb{P}\left(I'_{v,t} > \mathbb{E}\left[Z_{v^*}\right]\right) \lesssim \sum_{t=x}^{\infty} \max\left\{3\exp\left(-\frac{\Delta_v^2(1-\rho)^3}{8\rho}\alpha_v t\right), 3\exp\left(-\Delta_v\sqrt{\frac{\alpha_v t\log 1/\rho}{10}}\right)\right\}$$
$$\lesssim \frac{\sqrt{x}}{\Delta_v\sqrt{\alpha_v \log 1/\rho}}\exp\left(-\Delta_v\sqrt{\alpha_v x\log 1/\rho}\right). \tag{58}$$

Also by Lemma 3, we have

$$\mathbb{P}\left(I_{v^*,s,s+u} \geq \mathbb{E}\left[Z_{v^*}\right]\right) = \mathbb{P}\left(\widetilde{Z}_{v^*,s} + \widetilde{C}_{s,s+u} \geq \mathbb{E}\left[Z_{v^*}\right]\right) \leq (u+s)^{-4}.$$

This gives

$$\sum_{s=\alpha_v x - x^{2\beta}}^{\infty} \mathbb{P}\left(I_{v^*,s,s+u} \geq \mathbb{E}\left[Z_{v^*}\right]\right) = \sum_{s=\alpha_v x - x^{2\beta}}^{\infty} \mathbb{P}\left(\widetilde{Z}_{v^*,s} + \widetilde{C}_{s,s+u} \geq \mathbb{E}\left[Z_{v^*}\right]\right)$$
$$\leq \sum_{s=\alpha_v x - x^{2\beta}}^{\infty} (x+s)^{-4}$$
$$\lesssim \frac{1}{(1+\alpha_v)x^3}. \tag{59}$$

Collecting terms from (56), (58), (59), and rearranging concludes the proof.

□

B.6 Proof of Theorem 4

Firstly, we bound the step regret at time t by the confidence radius at time t, as is common in bandit regret analysis (e.g., Srinivas et al. (2010b)).

Lemma 5. *With probability at least $1 - \frac{2}{t^4}$, Algorithm 2 satisfies, for any $t \in \mathbb{N}$,*

$$\mathbb{E}\left[Z_{v^*}\right] - \mathbb{E}\left[Z_{J_t}\right] \leq \mathcal{O}\left(\frac{\textit{poly-log}(t)}{(1-\rho)^2\sqrt{N_t^+(J_t)}}\right), \tag{60}$$

where $\mathcal{O}$ omits absolute constants and logarithmic factors in problem intrinsics.

Proof. By Lemma 3, with probability at least $1 - \frac{2}{t^4}$, we have

$$\mathbb{E}\left[Z_{J_t}\right] \leq \widetilde{Z}_{J_t,N_t^+(J_t)} + \widetilde{C}_{N_t^+(J_t),t} \quad \text{and} \quad \mathbb{E}\left[Z_{v^*}\right] \geq \widetilde{Z}_{v^*,N_t^+(v^*)} - \widetilde{C}_{N_t^+(v^*),t} \tag{61}$$

Thus, with probability at least $1 - \frac{2}{t^4}$,

$$\mathbb{E}\left[Z_{J_t}\right] - \mathbb{E}\left[Z_{v^*}\right] = \widetilde{Z}_{J_t,N_t^+(J_t)} + \widetilde{C}_{N_t^+(J_t),t} - \left(\widetilde{Z}_{v^*,N_t^+(v^*)} + \widetilde{C}_{N_t^+(v^*),t}\right) \tag{62}$$
$$\leq \widetilde{Z}_{J_t,N_t^+(J_t)} + \widetilde{C}_{N_t^+(J_t),t} - \left(\widetilde{Z}_{J_t,N_t^+(J_t)} + \widetilde{C}_{N_t^+(J_t),t}\right) \tag{63}$$
$$\leq 2\widetilde{C}_{N_t^+(J_t),t} \tag{64}$$
$$\leq \mathcal{O}\left(\frac{\textit{poly-log}(t)}{(1-\rho)^2\sqrt{N_t^+(J_t)}}\right), \tag{65}$$

where (63) uses that $J_t \in \arg\max_{v\in V}\left[\widetilde{Z}_{J_t,N_t^+(v)} + \widetilde{C}_{N_t^+(v),t}\right]$.

□

Theorem 4. *Let T be any positive integer. Under Assumption 1, Algorithm 2 admits regret of order*

$$\textit{Reg}(T) = \widetilde{\mathcal{O}}\left(\min\left\{\frac{1}{(1-\rho)^2}\sqrt{\frac{T}{\alpha}}, \frac{1}{(1-\rho)^2}\sqrt{mT}\right\}\right),$$

where $\alpha = \min_{v\in V}\alpha_v$, and α_v is defined in Definition 1, and $\widetilde{\mathcal{O}}$ omits poly-logarithmical factors in T.

In this bound the dependence on optimality gaps Δ_v are removed and all problem intrinsics are global. In other words, Algorithm 2 achieves this regret rate no matter how identical the nodes are.

Proof. **Part I:** $\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\frac{1}{(1-\rho)^2}\sqrt{\frac{T}{\alpha}}\right)$**.** For any $t, T \in \mathbb{N}$, consider events

$$\widetilde{\mathcal{E}}_t := \left\{\mathbb{E}\left[Z_{J_t}\right] \le \widetilde{Z}_{J_t, N_t^+(J_t)} + \widetilde{C}_{N_t^+(J_t),t} \quad \text{and} \quad \mathbb{E}\left[Z_{v^*}\right] \ge \widetilde{Z}_{v^*, N_t^+(v^*)} - \widetilde{C}_{N_t^+(v^*),t}\right\},$$
$$\mathcal{E} := \left\{N_t^+(v) - N_t(v) - (t - N_t(v))\alpha_v \ge -\sqrt{t\log(mT)} \quad \forall t \in [T], v \in V\right\}.$$

By Lemmas 3 and 4,

$$\mathbb{P}\left(\widetilde{\mathcal{E}}_t \cap \mathcal{E}\right) \ge 1 - \frac{4}{t^4} - \frac{1}{T}. \tag{66}$$

By Lemma 1,

$$\mathbb{E}\left[Z_{v^*}\right] = \sum_{x=0}^{\infty} \mathbb{P}\left(Z_{v^*} > x\right) \le 1 + \frac{\rho}{(1-\rho)^2}. \tag{67}$$

Thus, by (66), (67), we have

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T}\left(\mathbb{E}\left[Z_{v^*}\right] - \mathbb{E}\left[Z_{J_t}\right]\right) \\
&\le \left\lceil \frac{\log T}{\min_v \alpha_v^2} \right\rceil \cdot \mathbb{E}\left[Z_{v^*}\right] + \sum_{t=\left\lceil \frac{\log T}{\min_v \alpha_v^2}\right\rceil}^{T} \mathbb{E}\left[Z_{v^*} - Z_{J_t} \middle| \widetilde{\mathcal{E}}_t \cap \mathcal{E}\right] \mathbb{P}\left(\widetilde{\mathcal{E}}_t \cap \mathcal{E}\right) \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\left(Z_{v^*} - Z_{J_t}\right) \middle| \overline{\widetilde{\mathcal{E}}_t \cap \mathcal{E}}\right]\left(1 - \mathbb{P}\left(\widetilde{\mathcal{E}}_t \cap \mathcal{E}\right)\right) \\
&\le \left\lceil \frac{\log T}{\min_v \alpha_v^2} \right\rceil \cdot \frac{\rho}{(1-\rho)^2} + \sum_{t=\left\lceil \frac{\log T}{\min_v \alpha_v^2}\right\rceil}^{T} \mathcal{O}\left(\frac{\textit{poly-log}(t)}{(1-\rho)^2\sqrt{N_t^+(J_t)}}\right) \\
&\quad + \frac{\rho^2}{(1-\rho)^2}\sum_{t=1}^{T}\left(\frac{4}{t^4} + \frac{1}{T}\right) && \text{(by Lemma 5 and (66))} \\
&\le \widetilde{\mathcal{O}}\left(\sum_{t=1}^{T} \frac{\textit{poly-log}(t)}{(1-\rho)^2\sqrt{N_t^+(J_t)}}\right), \\
&\le \widetilde{\mathcal{O}}\left(\sum_{t=1}^{T} \frac{\textit{poly-log}(t)}{(1-\rho)^2\sqrt{N_t(J_t) + \alpha_{J_t}(t - N_t(J_t)) - \sqrt{t\log(mT)}}}\right), && \text{(under event } \mathcal{E}\text{)} \\
&\le \widetilde{\mathcal{O}}\left(\sum_{t=1}^{T} \frac{\textit{poly-log}(t)}{(1-\rho)^2\sqrt{\alpha_{J_t} t - \sqrt{t\log(mT)}}}\right) \\
&\le \widetilde{\mathcal{O}}\left(\frac{1}{(1-\rho)^2}\sqrt{\frac{T}{\alpha}}\right),
\end{aligned}$$

where $\alpha := \min_{v\in V} \alpha_v$.

Part II: $\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\frac{1}{(1-\rho)^2}\sqrt{mT}\right)$**.**

By definitions in (4), we know $N_t^+(v) \ge N_t(v)$ for all $t \in \mathbb{N}$ and $v \in V$. From Lemma 5, we have

$$\mathit{Reg}(T) \le \sum_{t=1}^{T} \mathcal{O}\left(\frac{\textit{poly-log}(t)}{(1-\rho)^2\sqrt{N_t^+(J_t)}}\right) \le \sum_{t=1}^{T} \mathcal{O}\left(\frac{\textit{poly-log}(t)}{(1-\rho)^2\sqrt{N_t(J_t)}}\right). \tag{68}$$

Let $t_{i,v}$ be the i-th time the node v is played. Let B_v be the total number of time node v is played. Then we can regroup the sum in (68) by

$$\sum_{t=1}^{T} \frac{1}{N_t(J_t)} = \sum_{v \in V} \sum_{i=1}^{B_v} \frac{1}{N_{t_{i,v}}(v)}.$$

Since $t_{i,v}$ is the i-th time v is played, we have $n_{t_{i,v}}(v) = i$. Thus we have

$$\sum_{t=1}^{T} \frac{1}{N_t(J_t)} = \sum_{v \in V} \sum_{i=1}^{B_v} \frac{1}{N_{t_{i,v}}(v)} = \sum_{v \in V} \sum_{i=1}^{B_v} \frac{1}{i} \leq \sum_{v \in V} (1 + \log B_v) \leq m + m \log \frac{T}{m}, \tag{69}$$

where the last inequality uses $\sum B_v = T$ and the AM-GM inequality.

We then insert the above results into (68) to get

$$\begin{aligned}
\mathrm{Reg}(T) \leq & \sum_{t=1}^{T} \mathcal{O}\left(\frac{\textit{poly-log}(t)}{(1-\rho)^2 \sqrt{N_t^+(J_t)}} \right) \\
\leq & \widetilde{\mathcal{O}}\left(\frac{1}{(1-\rho)^2} \sqrt{T \sum_{t=1}^{T} \frac{1}{N_t(J_t)}} \right) && \text{(use the Cauchy-Schwarz inequality)} \\
\leq & \widetilde{\mathcal{O}}\left(\frac{1}{(1-\rho)^2} \sqrt{Tm\left(1 + \log \frac{T}{m}\right)} \right) && \text{(use (69))} \\
\leq & \widetilde{\mathcal{O}}\left(\frac{1}{(1-\rho)^2} \sqrt{mT} \right),
\end{aligned}$$

which concludes this part.

□

B.7 Proof of Theorem 5

Theorem 5. *For any given T and any policy π, there exists a problem instance $\mathfrak{I}$ satisfying Assumption 1 such that the T step regret of π on instance $\mathfrak{I}$ is lower bounded by $\Omega\left(T^{1/2}\right)$.*

Proof. We construct two "symmetric" problem instances $\mathfrak{J}$ and $\mathfrak{J}'$ both on two transient nodes $\{0, 1\}$ and one absorbing node $*$. All edges in both instances are of length 1. We use $M == [m_{ij}]$ (resp. $M' = [m'_{ij}]$) to denote the transition probabilities among transient nodes in instance $\mathfrak{J}$ (resp. $\mathfrak{J}'$). We construct instances $\mathfrak{J}$ and $\mathfrak{J}'$ so that $M = \begin{bmatrix} \frac{1}{2} & \epsilon \\ \epsilon & \frac{1}{2} + \epsilon \end{bmatrix}$ and $M' = \begin{bmatrix} \frac{1}{2} + \epsilon & \epsilon \\ \epsilon & \frac{1}{2} \end{bmatrix}$. In other words, M' is the anti-transpose (transpose with respect to the anti-diagonal) of M.

We use Z_v (resp. Z'_v) to denote the random variable $\mathcal{L}\left(\mathcal{P}_{t,v}\right)$ in problem instance $\mathfrak{J}$ (resp. $\mathfrak{J}'$). Then we have

$$\begin{bmatrix} \mathbb{E}[Z_0] \\ \mathbb{E}[Z_1] \end{bmatrix} = M \begin{bmatrix} \mathbb{E}[Z_0] \\ \mathbb{E}[Z_1] \end{bmatrix} + \mathbf{1}, \qquad \begin{bmatrix} \mathbb{E}[Z'_0] \\ \mathbb{E}[Z'_1] \end{bmatrix} = M' \begin{bmatrix} \mathbb{E}[Z'_0] \\ \mathbb{E}[Z'_1] \end{bmatrix} + \mathbf{1} \tag{70}$$

where $\mathbf{1}$ is the all-one vector. Solving the above equations gives, for both instances $\mathfrak{J}$ and $\mathfrak{J}'$, the optimality gap Δ is

$$\Delta := |\mathbb{E}[Z_0] - \mathbb{E}[Z_1]| = 8\epsilon + O\left(\epsilon^2\right). \tag{71}$$

Let π be any fixed algorithm and let T be any fixed time horizon, we use $\mathbb{P}_{\mathfrak{J},\pi}$ (resp. $\mathbb{P}_{\mathfrak{J}',\pi}$) to denote the probability measure of running π on instance $\mathfrak{J}$ (resp. $\mathfrak{J}'$) for T epochs.

Since the event $\{J_t = 1\}$ $(t \leq T)$ is measurable by both $\mathbb{P}_{\mathfrak{J},\pi}$ and $\mathbb{P}_{\mathfrak{J}',\pi}$, by Pinsker's inequality we have

$$|\mathbb{P}_{\mathfrak{J},\pi}(J_t = 1) - \mathbb{P}_{\mathfrak{J}',\pi}(J_t = 1)| \leq d_{TV}\left(\mathbb{P}_{\mathfrak{J},\pi}, \mathbb{P}_{\mathfrak{J},\pi'}\right) \leq \sqrt{2 D_{KL}(\mathbb{P}_{\mathfrak{J},\pi} \| \mathbb{P}_{\mathfrak{J}',\pi})}. \tag{72}$$

where we use Pinsker's inequality for the last inequality.

Let $\mathbb{Q}_i$ (resp. $\mathbb{Q}'_i$) be the probability measure generated by playing node i in instance $\mathfrak{J}$ (resp. $\mathfrak{J}'$). We can then decompose $\mathbb{P}_{\mathfrak{J},\pi}$ by

$$\mathbb{P}_{\mathfrak{J},\pi} = \mathbb{Q}_{J_1}\mathbb{P}\left(J_1|\pi\right)\mathbb{Q}_{J_2}\mathbb{P}(J_2|\pi,J_1)\cdots\mathbb{Q}_{J_T}\mathbb{P}(J_T|\pi,J_1,J_2,\ldots,J_{t-1}),$$
$$\mathbb{P}_{\mathfrak{J}',\pi} = \mathbb{Q}'_{J_1}\mathbb{P}\left(J_1|\pi\right)\mathbb{Q}'_{J_2}\mathbb{P}(J_2|\pi,J_1)\cdots\mathbb{Q}'_{J_T}\mathbb{P}(J_T|\pi,J_1,J_2,\ldots,J_{t-1}).$$

By chain rule for KL-divergence, we have

$$\begin{aligned}&D_{KL}(\mathbb{P}_{\mathfrak{J},\pi}\|\mathbb{P}_{\mathfrak{J}',\pi})\\ =&\sum_{J_1\in\{0,1\}}\mathbb{P}(J_1|\pi)D_{KL}(\mathbb{Q}_{J_1}\|\mathbb{Q}'_{J_1})+\sum_{t=2}^{T}\sum_{J_t\in\{0,1\}}\mathbb{P}(J_t|\pi,J_1,\cdots,J_{t-1})D_{KL}(\mathbb{Q}_{J_t}\|\mathbb{Q}'_{J_t}).\end{aligned}\tag{73}$$

Since the policy must pick one of node 1 and node 2, from nonnegativity of KL-divergence and (73) we have

$$D_{KL}(\mathbb{P}_{\mathfrak{J},\pi}\|\mathbb{P}_{\mathfrak{J}',\pi})\le\sum_{t=1}^{T}\sum_{i=0}^{1}D_{KL}(\mathbb{Q}_i\|\mathbb{Q}'_i),\tag{74}$$

which allows us to remove dependence on policy π.

Next we study the distributions $\mathbb{Q}_i$. With edge lengths fixed, the sample space of this distribution is $\cup_{h=1}^{\infty}\{0,1\}^h$, since length of the trajectory can be arbitrarily long, and each node on the trajectory can be either of $\{0,1\}$. To describe the distribution $\mathbb{Q}_i$ and $\mathbb{Q}'_i$, we use random variables $X_0, X_1, X_2, X_3, \cdots \in \{0,1,*\}$ (with $X_0 = i$), where X_k is the k-th node in the trajectory generated by playing i.

By Markov property we have, for $i, j \in \{0,1\}$,

$$\mathbb{Q}_i\left(X_{k+1}, X_{k+2}, \ldots|X_k = j\right) = \mathbb{Q}_j \quad \text{and} \quad \mathbb{Q}'_i\left(X_{k+1}, X_{k+2}, \ldots|X_k = j\right) = \mathbb{Q}'_j, \quad \forall k \in \mathbb{N}_+.$$

In words, conditioning on k-th node being j, the distribution generated by subsequent nodes are the same as $\mathbb{Q}_j$.

Note we can decompose $\mathbb{Q}_i$ by $\mathbb{Q}_i = \mathbb{Q}_i(X_1)\mathbb{Q}_i(X_2, X_3, \ldots, |X_1)$. Thus by chain rule of KL-divergence, for $i \in \{0,1\}$,

$$\begin{aligned}&D_{KL}(\mathbb{Q}_i\|\mathbb{Q}'_i)\\ =&D_{KL}(\mathbb{Q}_i(X_1)\|\mathbb{Q}'_i(X_1))\\ &+\sum_{x_1\in\{0,1\}}\mathbb{Q}_i(X_1=x_1)D_{KL}(\mathbb{Q}_i(X_2,X_3,\cdots|X_1=x_1)\|\mathbb{Q}'_i(X_2,X_3,\cdots|X_1=x_1))\\ =&D_{KL}(\mathbb{Q}_i(X_1)\|\mathbb{Q}'_i(X_1))+\mathbb{Q}_i(X_1=i)D_{KL}(\mathbb{Q}_i\|\mathbb{Q}'_i)+\mathbb{Q}_i(X_1=1-i)D_{KL}(\mathbb{Q}_{1-i}\|\mathbb{Q}'_{1-i}),\end{aligned}$$

where the last step uses the Markov property.

Since $D_{KL}(\mathbb{Q}_i(X_1)\|\mathbb{Q}'_i(X_1)) = 2\epsilon^2 + o\left(\epsilon^3\right)$ for $i \in \{0,1\}$, the above gives

$$D_{KL}(\mathbb{Q}_i\|\mathbb{Q}'_i) = 2\epsilon^2 + O\left(\epsilon^3\right).\tag{75}$$

Combining the above results with (72) and (75), we get

$$\left|\mathbb{P}_{\mathfrak{J},\pi}(J_t=1)-\mathbb{P}_{\mathfrak{J}',\pi}(J_t=1)\right|\le\sqrt{2D_{KL}\left(\mathbb{P}_{\mathfrak{J},\pi}\|\mathbb{P}_{\mathfrak{J}',\pi}\right)}\le 2\sqrt{T\epsilon^2}+O\left(\sqrt{T\epsilon^3}\right).\tag{76}$$

Let $\mathrm{Reg}(T)$ (resp. $\mathrm{Reg}'(T)$) be the T epoch regret in instance $\mathfrak{J}$ (resp. $\mathfrak{J}'$).

Recall, by our construction, node 1 is suboptimal in instance $\mathfrak{J}$ and node 1 is optimal in instance $\mathfrak{J}'$. Since the optimality gaps in $\mathfrak{J}$ and $\mathfrak{J}'$ are the same (Eq. 71), we have,

$$\begin{aligned}\mathrm{Reg}(T)+\mathrm{Reg}'(T)&\ge\Delta\sum_{t=1}^{T}\mathbb{P}_{\mathfrak{J},\pi}\left(J_t=1\right)+\Delta\sum_{t=1}^{T}\mathbb{P}_{\mathfrak{J}',\pi}\left(J_t=2\right)\\ &=\Delta\sum_{t=1}^{T}\left(1-\mathbb{P}_{\mathfrak{J},\pi}\left(J_t=1\right)+\mathbb{P}_{\mathfrak{J}',\pi}\left(J_t=1\right)\right)\\ &\ge\left(8\epsilon+O\left(\epsilon^2\right)\right)\left(T-2\sqrt{T\epsilon^2}\cdot T+O\left(\sqrt{T\epsilon^3}\cdot T\right)\right).\end{aligned}$$

Now, we set $\epsilon = \frac{1}{4}T^{-1/2}$, which is the largest (in terms of order in T) value allowed by a constant probability gap in (76), and get

$$\mathrm{Reg}(T) + \mathrm{Reg}'(T) \geq \left(2T^{-1/2} + O\left(T^{-1}\right)\right)\left(T - \frac{1}{2}T + O\left(\sqrt{T}\right)\right) \geq \sqrt{T} + O(1),$$

which concludes the proof.

□

C Proofs for the Adversarial Setting

Note. Unless otherwise noted, we use $\mathcal{F}_t$ to denote the σ-algebra generated by all randomness up to the end of the t-th trajectory. Also, we use $\mathbb{E}_t$ to denote the expectation conditioning on $\mathcal{F}_t$, i.e., $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$. Also, we write $q_{ij} := \mathbb{P}(j \in \mathcal{P}_{t,i})$, which is the probability of j being visited by a trajectory from i. Since the transition probabilities do not change over time, we do not need to use t in the subscript of q_{ij}.

We start with the following lemma.

Lemma 6. *Let $f_a(x) = \frac{1}{x^a}$ for some real-number a, and $\mu \in (0,1)$. Let $X \in [0,1]$ be a random variable such that, $\mathbb{E}[X] = \mu$. Then it holds that*

$$\mathbb{E}[f_a(X)] = \sum_{p=0}^{\infty} \frac{f_a^{(p)}(\mu)}{p!}\mathbb{E}[(X-\mu)^p] = f_a(\mu) + \mathcal{O}(\mathbb{V}[X]), \tag{77}$$

where $\mathbb{V}$ is the variance operator.

Proof. The series $\left\{\sum_{p=1}^{\infty} \frac{|f_a^{(p)}(\mu)|}{p!}\mathbb{E}[|X-\mu|^p]\right\}$ convergences, since it is bounded and increasing. Thus, by dominated convergence test, the Taylor expansion of f_a converges in expectation, and

$$\mathbb{E}[f_a(X)] = \sum_{p=0}^{\infty} \frac{f_a^{(p)}(\mu)}{p!}\mathbb{E}[(X-\mu)^p] = f_a(\mu) + \mathcal{O}(\mathbb{V}[X]), \tag{78}$$

since $\mathbb{E}[(X-\mu)^p] \leq \mathbb{E}\left[(X-\mu)^2\right]$. □

Next, we study some properties of the estimator $\widehat{q}_{t,ij}$. Firstly, it is an unbiased estimator of q_{ij}.

Lemma 7. *For any t, i, j, it holds that*

$$\mathbb{E}[\widehat{q}_{t,ij}] = q_{ij}.$$

Proof. For any $i \in [K]$, let $\mathcal{F}_{t,i}$ be the σ-algebra generated by all randomness up to the first occurrence of i in the epoch t or the end of t-th epoch, whichever occurs earlier. Recall $\mathbb{I}_{t,ij} := \mathbb{I}_{\left[i\in\mathcal{P}_{t,J_t}\right]}\mathbb{I}_{\left[j\in\mathcal{P}_{t,J_t}\right]}\mathbb{I}_{\left[Y_i(\mathcal{P}_{t,J_t})>Y_j(\mathcal{P}_{t,J_t})\right]}$. By Markov property, we know that given $\mathcal{F}_{t,i}$, the probability of $\mathbb{I}_{\left[j\in\mathcal{P}_{t,J_t}\right]}\mathbb{I}_{\left[Y_i(\mathcal{P}_{t,J_t})>Y_j(\mathcal{P}_{t,J_t})\right]}$ being 1 is q_{ij}. Writing this down, we have

$$\mathbb{E}[\mathbb{I}_{t,ij}|\mathcal{F}_{t,i}] = \mathbb{E}\left[\mathbb{I}_{\left[i\in\mathcal{P}_{t,J_t}\right]}\mathbb{I}_{\left[j\in\mathcal{P}_{t,J_t}\right]}\mathbb{I}_{\left[Y_i(\mathcal{P}_{t,J_t})>Y_j(\mathcal{P}_{t,J_t})\right]}|\mathcal{F}_{t,i}\right] = \mathbb{I}_{\left[i\in\mathcal{P}_{t,J_t}\right]}q_{ij}$$

Thus, by tower law

$$\begin{aligned}
\mathbb{E}[\widehat{q}_{t,ij}] &= \mathbb{E}\left[\frac{\sum_{s=1}^{t-1}\mathbb{I}_{s,ij}}{\sum_{s=1}^{t-1}\mathbb{I}_{[i\in\mathcal{P}_{s,J_s}]}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{s=1}^{t-1}\mathbb{I}_{s,ij}}{\sum_{s=1}^{t-1}\mathbb{I}_{[i\in\mathcal{P}_{s,J_s}]}}\Big|\mathcal{F}_{t-1,i}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{s=1}^{t-2}\mathbb{I}_{s,ij} + \mathbb{I}_{\left[i\in\mathcal{P}_{t-1,J_{t-1}}\right]}q_{ij}}{\sum_{s=1}^{t-1}\mathbb{I}_{[i\in\mathcal{P}_{s,J_s}]}}\Big|\mathcal{F}_{t-1,i}\right]\right].
\end{aligned}$$

We can then repeatedly apply tower rule (conditioning on $\mathcal{F}_{t,i}$, $\mathcal{F}_{t-1,i}$, ..., $\mathcal{F}_{1,i}$) to get

$$\mathbb{E}[\widehat{q}_{t,ij}] = \mathbb{E}\left[\frac{\sum_{s=1}^{t-2}\mathbb{I}_{s,ij} + \mathbb{I}_{\left[i\in\mathcal{P}_{t-1,J_{t-1}}\right]}q_{ij}}{\sum_{s=1}^{t-1}\mathbb{I}_{[i\in\mathcal{P}_{s,J_s}]}}\right] = \cdots = \mathbb{E}\left[\frac{\sum_{s=1}^{t-1}\mathbb{I}_{[i\in\mathcal{P}_{s,J_s}]}q_{ij}}{\sum_{s=1}^{t-1}\mathbb{I}_{[i\in\mathcal{P}_{s,J_s}]}}\right] = q_{ij}.$$

□

Also, the variance of our estimators $\widehat{q}_{t,ij}$ is small.

Lemma 8. *For any T, it holds that*

$$\sum_{t=1}^{T}\sum_{i,j\in[K]}\mathbb{V}\left[\widehat{q}_{t,ij}\right]\leq\mathcal{O}\left(\textit{poly-log}(T)\right).$$

Proof. For the variance, we have

$$\mathbb{V}\left[\widehat{q}_{t,ij}\right]=\sum_{m=1}^{t}\mathbb{V}\left[\widehat{q}_{t,ij}\middle|N_t^+(i)=m\right]\mathbb{P}\left(N_t^+(i)=m\right)\leq\sum_{m=1}^{t}\frac{1}{m}\mathbb{P}\left(N_t^+(i)=m\right)=\mathbb{E}\left[\frac{1}{N_t^+(i)}\right].$$

By Lemma 4 and a union bound, we know, for any $\delta\in(0,1)$,

$$\mathbb{P}\Big(N_t^+(i)\geq\alpha t-\sqrt{2t\log(2TK/\delta)},\quad\forall i\in[K],\ t\in[T]\Big)\leq\delta.$$

Thus it holds that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{N_t^+(i)}\right]=&\mathbb{E}\left[\frac{1}{N_t^+(i)}\middle|N_t^+(i)\geq\alpha t-\sqrt{2t\log(TK/\delta)}\right]\mathbb{P}\left(N_t^+(i)\geq\alpha t-\sqrt{2t\log(TK/\delta)}\right)\\&+\mathbb{E}\left[\frac{1}{N_t^+(i)}\middle|N_t^+(i)<\alpha t-\sqrt{2t\log(TK/\delta)}\right]\mathbb{P}\left(N_t^+(i)<\alpha t-\sqrt{2t\log(TK/\delta)}\right)\\\leq&\frac{1}{\max\left\{1,\alpha t-\sqrt{2t\log(TK/\delta)}\right\}}+\delta.\end{aligned}$$

Taking summation for the above equation gives

$$\sum_{t=1}^{T}\sum_{i,j\in[K]}\mathbb{V}\left[\widehat{q}_{t,ij}\right]\leq\sum_{t=1}^{T}\frac{K^2}{\max\left\{1,\alpha t-\sqrt{2t\log(TK/\delta)}\right\}}+\delta K^2T.$$

Setting $\delta=\frac{1}{KT^2}$ concludes the proof. □

Lemma 9. *Let $l_{t,j}:=\mathbb{E}\left[\mathcal{L}\left(\mathcal{P}_{t,j}\right)\right]$ for notational ease. Let B be the constant used in defining $\widehat{Z}_{t,j}$. For all t, i, it holds that*

$$\begin{aligned}&\sum_t l_{t,i}-\mathbb{E}\left[\sum_t\sum_j p_{tj}l_{t,i}\right]\\=&\mathbb{E}\left[\sum_t\widehat{Z}_{t,i}\right]-\mathbb{E}\left[\sum_t\sum_j p_{tj}\widehat{Z}_{t,j}\right]+B\,\mathcal{O}\left(\textit{poly-log}(T)\right)\end{aligned}$$

If the estimation for $\widehat{q}_{t,ij}$ is exact, we have

$$l_{t,j}-B=\mathbb{E}\left[\widehat{Z}_{t,j}\right],\quad\forall t,j.$$

Proof. By the observation in Proposition 1, we know $Z_{t,j}$ is a sample of $\mathcal{L}\left(\mathcal{P}_{t,j}\right)$ if $j\in\mathcal{P}_{t,J_t}$. Thus, $\mathbb{E}_{t-1}\left[Z_{t,j}|j\in\mathcal{P}_{t,J_t}\right]=l_{t,j}$ and

$$\begin{aligned}\mathbb{E}_{t-1}\left[\widehat{Z}_{t,j}\right]=&\mathbb{E}_{t-1}\left[\widehat{Z}_{t,j}|j\in\mathcal{P}_{t,J_t}\right]\mathbb{P}\left(j\in\mathcal{P}_{t,J_t}\right)+\mathbb{E}_{t-1}\left[\widehat{Z}_{t,j}|j\notin\mathcal{P}_{t,J_t}\right]\mathbb{P}\left(j\notin\mathcal{P}_{t,J_t}\right)\\=&\mathbb{E}_{t-1}\left[\frac{Z_{t,j}-B}{p_{tj}+\sum_{i\neq j}\widehat{q}_{t,ij}p_{ti}}\middle|j\in\mathcal{P}_{t,J_t}\right]\mathbb{P}\left(j\in\mathcal{P}_{t,J_t}\right)\\=&\left(l_{t,j}-B\right)\frac{p_{tj}+\sum_{i\neq j}p_{ti}q_{ij}}{p_{tj}+\sum_{i\neq j}\widehat{q}_{t,ij}p_{ti}}.\end{aligned}$$

Thus, by tower law of total expectation

$$\mathbb{E}_{t-2}\left[\widehat{Z}_{t,j}\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{Z}_{t,j}|\mathcal{F}_{t-1}\right]|\mathcal{F}_{t-2}\right] \tag{79}$$

$$= (l_{t,j} - B)\,\mathbb{E}\left[\frac{p_{tj} + \sum_{i\neq j} p_{ti}q_{ij}}{p_{tj} + \sum_{i\neq j}\widehat{q}_{t,ij}p_{ti}}\middle|\mathcal{F}_{t-2}\right]. \tag{80}$$

When the estimation for $\widehat{q}_{t,ij}$ is exact, the above equation translates to $\mathbb{E}\left[\widehat{Z}_{t,j}\right] = l_{t,j} - B$.

We apply Lemma 6, with X being $p_{tj} + \sum_{i\neq j}\widehat{q}_{t,ij}p_{ti}$ and $f(x) = \frac{1}{x}$, and get

$$\mathbb{E}_{t-2}\left[\widehat{Z}_{t,j}\right] = (l_{t,j} - B)\left(1 + \mathcal{O}\left(\mathbb{V}\left[p_{tj} + \sum_{i\neq j}\widehat{q}_{t,ij}p_{ti}|\mathcal{F}_{t-2}\right]\right)\right)$$

$$= (l_{t,j} - B)\left(1 + \mathcal{O}\left(\mathbb{V}\left[\sum_{i\neq j}\widehat{q}_{t,ij}p_{ti}\middle|\mathcal{F}_{t-2}\right]\right)\right). \tag{81}$$

Taking total expectation on both sides gives

$$\mathbb{E}\left[\widehat{Z}_{t,j}\right] = (l_{t,j} - B)\left(1 + \mathcal{O}\left(\mathbb{V}\left[\sum_{i\neq j}\widehat{q}_{t,ij}p_{ti}\right]\right)\right).$$

For the variance term $\mathbb{V}\left[\sum_{i\neq j}\widehat{q}_{t,ij}p_{ti}\right]$, we have

$$\mathbb{V}\left[\sum_{i\neq j}p_{ti}\widehat{q}_{t,ij}\right] = \sum_{i\neq j}\mathbb{V}\left[p_{ti}\widehat{q}_{t,ij}\right] + \sum_{i\neq j,k\neq j,i\neq k} Cov\left(p_{tk}\widehat{q}_{t,kj}, p_{ti}\widehat{q}_{t,ij}\right)$$

$$\leq K\sum_{i\neq j}\mathbb{V}\left[p_{ti}\widehat{q}_{t,ij}\right] \leq K\sum_{i\neq j}\mathbb{V}\left[\widehat{q}_{t,ij}\right]. \tag{82}$$

We can insert (82) into (81) to get

$$\mathbb{E}\left[\widehat{Z}_{t,j}\right] - (l_{t,j} - B) = \mathcal{O}\left((l_{t,j} - B)\sum_{i\neq j}\mathbb{V}\left[\widehat{q}_{t,ij}\right]\right).$$

Summing over t, it holds that

$$\sum_t \mathbb{E}\left[\widehat{Z}_{t,j}\right] = \sum_t l_{t,i} - BT + B\,\mathcal{O}\left(\textit{poly-log}(T)\right), \tag{83}$$

where we use Lemma 8 on the last line.

Similarly, we get

$$\mathbb{E}\left[\sum_j p_{tj}\widehat{Z}_{t,j}\right] = \mathbb{E}\left[\sum_t\sum_j p_{tj}l_{t,j}\right] - BT + \sum_t\sum_{i,j}\mathcal{O}\left((l_{t,j} - B)K\sum_{i\neq j}\mathbb{V}\left[\widehat{q}_{t,ij}\right]\right).$$

Again we use Lemma 8 to get, for any $\delta \in (0,1)$,

$$\mathbb{E}\left[\sum_j p_{tj}\widehat{Z}_{t,j}\right] = \mathbb{E}\left[\sum_t\sum_j p_{tj}l_{t,j}\right] - BT + B\,\mathcal{O}\left(\textit{poly-log}(T)\right). \tag{84}$$

Combining (83) and (84) concludes the proof.

□

As a final preparation step, we state a high probability event.

Lemma 10. *For any B, let $\mathcal{E}_T(B) := \{Z_{t,j} \le B \text{ for all } t = 1, 2, \cdots, T, \text{ and } j \in [K]\}$. For any $\epsilon \in (0,1)$ and $B = \frac{\log \frac{(1-\rho)\epsilon}{KT}}{\log \rho}$, it holds that*

$$\mathbb{P}\left(\mathcal{E}_T\left(B\right)\right) \ge 1 - \epsilon,$$

and

$$\mathbb{E}\left[\widehat{Z}_{t,i} | \textit{not } \mathcal{E}_T(B)\right] \le \mathcal{O}\left(\textit{poly-log}(T/\epsilon)\right).$$

Proof. Since all edge lengths are smaller than 1, we have, for any integer B,

$$\begin{aligned}
\mathbb{P}\left(Z_{t,i} > B\right) &\le \mathbb{P}\left(\{\text{random walk starting from } i \text{ does not terminate in } B \text{ steps}\}\right) \\
&= \sum_{l=B}^{\infty} \mathbb{P}\left(\{\text{random walk starting from } i \text{ terminates at step } l\}\right) \\
&= \sum_{l=B}^{\infty} \sum_{j=1}^{K} \left[M^l\right]_{ij} \le \sum_{l=B}^{\infty} \|M^l\|_{\infty} \le \sum_{l=B}^{\infty} \rho^l \le \frac{\rho^B}{1-\rho}.
\end{aligned}$$

Thus with probability at least $1 - \frac{\rho^B}{1-\rho}$, we have $Z_{t,i} \le B$. We define

$$\mathcal{E}_T(B) := \{Z_{t,j} \le B \text{ for all } t = 1, 2, \cdots, T, \text{ and } j \in [K]\}.$$

By a union bound, $\mathbb{P}\left(\mathcal{E}_T(B)\right) \ge 1 - \frac{KT\rho^B}{1-\rho}$. Now we can set $B = \frac{\log \frac{(1-\rho)\epsilon}{KT}}{\log \rho}$ so that $\mathbb{P}\left(\mathcal{E}_T(B)\right) \ge 1 - \epsilon$.

The random variables $Z_{t,i}$ also has the memorylessness-type property:

$$\begin{aligned}
&\mathbb{E}\left[Z_{t,i} | \text{not } \mathcal{E}_T(B)\right] \\
\le &\mathbb{E}\left[Z_{t,i} | Z_{t,i} > B\right] \le \sum_{l=B+1}^{\infty} l \frac{\mathbb{P}\left(\{\mathcal{P}_{t,i} \text{ terminates at step } l\} \cap \{Z_{t,i} > B\}\right)}{\mathbb{P}\left(Z_{t,i} > B\right)} \\
= &\sum_{l=B+1}^{\infty} l \frac{\sum_j \mathbb{P}\left(\{\mathcal{P}_{t,i} \text{ terminates at step } l\} \cap \{\text{the } (B+1)\text{-th step is at } j\}\right)}{\sum_j \mathbb{P}\left(\{\text{the } (B+1)\text{-th step is at } j\}\right)} \\
\le &\sum_{l=B+1}^{\infty} l \sum_j \mathbb{P}\left(\{\mathcal{P}_{t,i} \text{ terminates at step } l\} \middle| \{\text{the } (B+1)\text{-th step is at } j\}\right) \\
= &\sum_j \sum_{l=1}^{\infty} (l+B) \mathbb{P}\left(\{\mathcal{P}_{t,j} \text{ terminates at step } l\}\right) \\
= &\sum_j \mathbb{E}\left[Z_{t,j} + B\right] \le KB + \sum_j \mathbb{E}\left[Z_{t,j}\right]
\end{aligned}$$

where we use Markov property on the second last line.

Since $\mathbb{E}\left[Z_{t,j}\right] \le \mathcal{O}\left(\frac{1}{(1-\rho)^2}\right)$, we insert this into the above equation to get

$$\mathbb{E}\left[Z_{t,i} | \text{not } \mathcal{E}_T(B)\right] \le \mathcal{O}\left(KB + \frac{K}{(1-\rho)^2}\right). \tag{85}$$

Similarly, we have

$$\mathbb{E}\left[Z_{t,i}^2 | \text{not } \mathcal{E}_T(B)\right] \le \mathcal{O}\left(KB^2 + \frac{K}{(1-\rho)^3}\right). \tag{86}$$

Next we turn to $\mathbb{E}\left[\widehat{Z}_{t,i} | \text{not } \mathcal{E}_T(B)\right]$. By Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\widehat{Z}_{t,i} \middle| \text{not } \mathcal{E}_T(B)\right] \le \sqrt{\mathbb{E}\left[\left(Z_{t,i} - B\right)^2 \middle| \text{not } \mathcal{E}_T(B)\right] \mathbb{E}\left[\left(\frac{1}{p_{ti} + \sum_{j \ne i} \widehat{q}_{t,ji} p_{tj}}\right)^2 \middle| \text{not } \mathcal{E}_T(B)\right]}. \tag{87}$$

By again applying Lemma 6, we get

$$\mathbb{E}\left[\left(\frac{1}{p_{ti}+\sum_{j\neq i}\widehat{q}_{t,ji}p_{tj}}\right)^2\Big|\text{not }\mathcal{E}_T(B)\right]\leq\mathcal{O}\left(1\right).$$

Combine this with (87), (85) and (86), we get, when $B=\frac{\log\frac{(1-\rho)\epsilon}{KT}}{\log\rho}$,

$$\mathbb{E}\left[\widehat{Z}_{t,i}|\text{not }\mathcal{E}_T(B)\right]\leq\mathcal{O}\left(\textit{poly-log}(T/\epsilon)\right). \tag{88}$$

□

C.1 Proof of Theorem 6

Theorem 6. *Fix any T. Algorithm 3 satisfies*

$$\mathbb{E}\left[\mathrm{Reg}_i^{\mathrm{adv}}(T)\right]\leq\widetilde{\mathcal{O}}\left(\sqrt{\left(1+\sum_{j\in[K]}\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right)T}\right),\quad\forall i\in[K].$$

In addition, if the estimators $\widehat{q}_{t,ij}$ are exact, then with probability at least $1-7\epsilon$, Algorithm 3 achieves, for $\forall i\in[K]$,

$$\mathrm{Reg}_i^{\mathrm{adv}}(T)\leq\widetilde{\mathcal{O}}\left(\sqrt{\left(1+2\sum_{j\in[K]}\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right)T\log(KT/\delta)}\right),\quad\forall i\in[K].$$

Proof. For completeness, we repeat some classic arguments for exponential weights algorithms until Eq. 90 (Littlestone and Warmuth, 1994; Auer et al., 2002b). Recall, for $t\in\mathbb{N}$, $\widehat{S}_{t,j}:=\sum_{s:s\text{ is even, }0\leq s\leq 2t}\widehat{Z}_{s,j}$, $\widehat{S}_t:=\sum_{s:s\text{ is odd, }1\leq s\leq t}\sum_{i\in[K]}p_{si}\widehat{Z}_{s,i}$. Define $W_{2m}:=\sum_i\exp\left(\eta\widehat{S}_{2m,i}\right)$ (resp. $W_{2m+1}:=\sum_i\exp\left(\eta\widehat{S}_{2m+1,i}\right)$), and $W_0=W_1=K$. Then we have

$$\frac{W_t}{W_{t-2}}:=\sum_i\frac{\exp\left(\eta\widehat{S}_{t-2,i}\right)}{W_{t-2}}\exp\left(\eta\widehat{Z}_{t,i}\right)=\sum_i p_{ti}\exp\left(\eta\widehat{Z}_{t,i}\right).$$

Since $\exp(x)\leq 1+x+\frac{x^2}{2}$ for $x\leq 0$, and $\widehat{Z}_{t,j}\leq 0$ (under event $\mathcal{E}_T(B)$), we have

$$\exp\left(\eta\widehat{Z}_{t,j}\right)\leq\left\{1+\eta(\widehat{Z}_{t,j})+\frac{\eta^2}{2}\left(\widehat{Z}_{t,j}\right)^2\right\}.$$

Continuing from above, we have, for T being even,

$$\begin{aligned}
\exp\left(\eta\widehat{S}_{T,i}\right)&\leq W_0\frac{W_2}{W_0}\cdots\frac{W_T}{W_{T-2}}\\
&\leq W_0\prod_{t:t\text{ is even, }2\leq t\leq T}\left(\sum_{j=1}^K p_{tj}\exp\left(\eta\widehat{Z}_{t,j}\right)\right)\\
&\leq W_0\prod_{t:t\text{ is even, }2\leq t\leq T}\left(\sum_{j=1}^K p_{tj}\left(1+\eta(\widehat{Z}_{t,j})+\frac{\eta^2}{2}\left(\widehat{Z}_{t,j}\right)^2\right)\right)\\
&\leq W_0\prod_{t:t\text{ is even, }2\leq t\leq T}\left(1+\sum_{j=1}^K p_{tj}\left(\eta(\widehat{Z}_{t,j})+\frac{\eta^2}{2}\left(\widehat{Z}_{t,j}\right)^2\right)\right)\\
&\leq W_0\prod_{t:t\text{ is even, }2\leq t\leq T}\exp\left(\sum_{j=1}^K p_{tj}\left(\eta(\widehat{Z}_{t,j})+\frac{\eta^2}{2}\left(\widehat{Z}_{t,j}\right)^2\right)\right)\qquad\text{(by }1+x\leq\exp(x).\text{)}\\
&=W_0\exp\left(\eta\widehat{S}_T+\frac{\eta^2}{2}\sum_{t:t\text{ is even, }2\leq t\leq T}\sum_j p_{tj}\widehat{Z}_{t,j}^2\right).
\end{aligned}\tag{89}$$

We repeat the above for the odd indices, take logarithm on both sides, and combine all the indices. By doing this, we get

$$\sum_{t=1}^{T} \widehat{Z}_{t,i} - \sum_{t=1}^{T} \sum_{j} p_{tj} \widehat{Z}_{t,j} \leq \frac{2 \log K}{\eta} + \frac{\eta B^2}{2} \sum_{t=1}^{T} \sum_{j} \frac{p_{tj}}{\left(p_{tj} + \sum_{k \neq j} \widehat{q}_{t,kj} p_{tk}\right)^2} \mathbb{I}_{\left[j \in \mathcal{P}_{t,J_t}\right]}. \tag{90}$$

Next we split the proof into two cases.

Case I: We do not have oracle access to q_{ij}.
For simplicity, let $\widetilde{p}_{tj} = p_{tj} + \sum_{i \neq j} p_{ti} q_{ij}$ and let $\widehat{p}_{tj} = p_{tj} + \sum_{i \neq j} p_{ti} \widehat{q}_{t,ij}$.
For the last term in (90), its conditional expectation is

$$\mathbb{E}_{t-2}\left[\frac{p_{tj}}{\widehat{p}_{tj}^2} \mathbb{I}_{\left[j \in \mathcal{P}_{t,J_t}\right]}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{p_{tj}}{\widehat{p}_{tj}^2} \mathbb{I}_{\left[j \in \mathcal{P}_{t,J_t}\right]} \middle| \mathcal{F}_{t-1}\right] \middle| \mathcal{F}_{t-2}\right] = \mathbb{E}\left[\frac{p_{tj}}{\widehat{p}_{tj}^2} \widetilde{p}_{tj} \middle| \mathcal{F}_{t-2}\right] \tag{91}$$

By Lemma 6, we can write the above as

$$\begin{aligned}
\mathbb{E}\left[\frac{p_{tj}}{\widehat{p}_{tj}^2} \widetilde{p}_{tj} \middle| \mathcal{F}_{t-2}\right] &= \frac{p_{tj}}{p_{tj} + \sum_{i \neq j} q_{ij} p_{ti}} + \mathcal{O}\left(\mathbb{V}\left[p_{tj} + \sum_{i \neq j} \widehat{q}_{t,ij} p_{ti} \middle| \mathcal{F}_{t-2}\right]\right) \\
&\leq \frac{p_{tj}}{p_{tj} + (1 - p_{tj}) \alpha_j} + \mathcal{O}\left(\mathbb{V}\left[p_{tj} + \sum_{i \neq j} \widehat{q}_{t,ij} p_{ti} \middle| \mathcal{F}_{t-2}\right]\right) \qquad \text{(since } q_{ij} \leq \alpha_j \text{ by definition)} \\
&\leq p_{tj} + \frac{1 - \sqrt{\alpha_j}}{1 + \sqrt{\alpha_j}} + \mathcal{O}\left(\sum_{i \neq j} \mathbb{V}\left[\widehat{q}_{t,ij} \middle| \mathcal{F}_{t-2}\right]\right),
\end{aligned} \tag{92}$$

where the last equation uses $\frac{x}{x + a(1-x)} \leq x + \frac{1 - \sqrt{a}}{1 + \sqrt{a}}$ for $x \in [0,1]$ and $a \in [0,1]$ (Proposition 2 in Appendix C.3). We take summation over j on both sides of (92) and use the identity in (91), to get

$$\mathbb{E}\left[\sum_{j} \frac{p_{tj}}{\widehat{p}_{tj}^2} \mathbb{I}_{\left[j \in \mathcal{P}_{t,J_t}\right]} \middle| \mathcal{F}_{t-2}\right] \leq 1 + \sum_{j} \frac{1 - \sqrt{\alpha_j}}{1 + \sqrt{\alpha_j}} + \sum_{j} \mathcal{O}\left(\sum_{i \neq j} \mathbb{V}\left[\widehat{q}_{t,ij} \middle| \mathcal{F}_{t-2}\right]\right).$$

Now all the randomness in the above equation are in the third term only. By another use of total law of expectation, we have

$$\mathbb{E}\left[\sum_{j} \frac{p_{tj}}{\widehat{p}_{tj}^2} \mathbb{I}_{\left[j \in \mathcal{P}_{t,J_t}\right]}\right] \leq 1 + \sum_{j} \frac{1 - \sqrt{\alpha_j}}{1 + \sqrt{\alpha_j}} + \sum_{j} \mathcal{O}\left(\sum_{i \neq j} \mathbb{V}\left[\widehat{q}_{t,ij}\right]\right).$$

We can take summation over t (in the above equation) and apply Lemma 8 to get, for any $\delta \in (0,1)$,

$$\sum_{t=1}^{T} \mathbb{E}\left[\sum_{j} \frac{p_{tj}}{\widehat{p}_{tj}^2} \mathbb{I}_{\left[j \in \mathcal{P}_{t,J_t}\right]}\right] \leq T\left(1 + \sum_{j} \frac{1 - \sqrt{\alpha_j}}{1 + \sqrt{\alpha_j}}\right) + \mathcal{O}\left(\textit{poly-log}(T)\right). \tag{93}$$

Since $\mathbb{P}\left(\mathcal{E}_T(B)\right) \geq 1 - \epsilon$, it holds that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \sum_{j} \frac{p_{tj}}{\widehat{p}_{tj}^2} \mathbb{I}_{\left[j \in \mathcal{P}_{t,J_t}\right]}\right] =& \mathbb{E}\left[\sum_{t=1}^{T} \sum_{j} \frac{p_{tj}}{\widehat{p}_{tj}^2} \mathbb{I}_{\left[j \in \mathcal{P}_{t,J_t}\right]} \middle| \mathcal{E}_T(B)\right] \mathbb{P}\left(\mathcal{E}_T(B)\right) \\
&+ \mathbb{E}\left[\sum_{t=1}^{T} \sum_{j} \frac{p_{tj}}{\widehat{p}_{tj}^2} \mathbb{I}_{\left[j \in \mathcal{P}_{t,J_t}\right]} \middle| \text{not } \mathcal{E}_T(B)\right] \left(1 - \mathbb{P}\left(\mathcal{E}_T(B)\right)\right) \\
\geq& \mathbb{E}\left[\sum_{t=1}^{T} \sum_{j} \frac{p_{tj}}{\widehat{p}_{tj}^2} \mathbb{I}_{\left[j \in \mathcal{P}_{t,J_t}\right]} \middle| \mathcal{E}_T(B)\right] (1 - \epsilon),
\end{aligned}$$

which means, for any $\epsilon \in \left(0, \frac{1}{2}\right)$

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j}\frac{p_{tj}}{\widehat{p}_{tj}^2}\mathbb{I}_{\left[j\in\mathcal{P}_{t,J_t}\right]}\middle|\mathcal{E}_T(B)\right] \leq 2\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j}\frac{p_{tj}}{\widehat{p}_{tj}^2}\mathbb{I}_{\left[j\in\mathcal{P}_{t,J_t}\right]}\right]. \tag{94}$$

Thus, we take conditional expectation on both sides of (90), and use (94) and (93), to get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\widehat{Z}_{t,i} - \sum_{t=1}^{T}\sum_{j}p_{tj}\widehat{Z}_{t,j}\middle|\mathcal{E}_T(B)\right] &(95)\\
\leq&\frac{2\log K}{\eta}2\mathbb{E}\left[\frac{\eta B^2}{2}\sum_{t=1}^{T}\sum_{j}\frac{p_{tj}}{\widehat{p}_{tj}^2}\mathbb{I}_{\left[j\in\mathcal{P}_{t,J_t}\right]}\middle|\mathcal{E}_T(B)\right]\\
\leq&\frac{2\log K}{\eta} + \eta B^2\left(1+\sum_{j}\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right)T + \mathcal{O}\left(\textit{poly-log}(T)\right). &(96)
\end{aligned}$$

Now, by total law of expectation, we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\widehat{Z}_{t,i} - \sum_{t=1}^{T}\sum_{j}p_{tj}\widehat{Z}_{t,j}\right]\\
=&\mathbb{E}\left[\sum_{t=1}^{T}\widehat{Z}_{t,i} - \sum_{t=1}^{T}\sum_{j}p_{tj}\widehat{Z}_{t,j}\middle|\mathcal{E}_T(B)\right]\mathbb{P}\left(\mathcal{E}_T(B)\right)\\
&+\mathbb{E}\left[\sum_{t=1}^{T}\widehat{Z}_{t,i} - \sum_{t=1}^{T}\sum_{j}p_{tj}\widehat{Z}_{t,j}\middle|\text{not }\mathcal{E}_T(B)\right]\left(1-\mathbb{P}\left(\mathcal{E}_T(B)\right)\right)\\
\leq&\mathbb{E}\left[\sum_{t=1}^{T}\widehat{Z}_{t,i} - \sum_{t=1}^{T}\sum_{j}p_{tj}\widehat{Z}_{t,j}\middle|\mathcal{E}_T(B)\right]\\
&+\epsilon\left(\mathbb{E}\left[\sum_{t=1}^{T}\widehat{Z}_{t,i}\middle|\text{not }\mathcal{E}_T(B)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{j}\widehat{Z}_{t,j}\middle|\text{not }\mathcal{E}_T(B)\right]\right)\\
\leq&\frac{2\log K}{\eta} + \eta B^2\left(1+\sum_{j}\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right)T\\
&+\mathcal{O}\left(\textit{poly-log}(T)\right) + \epsilon T\mathcal{O}\left(\textit{poly-log}(T/\epsilon)\right),
\end{aligned}$$

where the last line uses (96) and Lemma 10.

We set $\epsilon = \frac{1}{T^2}$ to get

$$\mathbb{E}\left[\sum_{t=1}^{T}\widehat{Z}_{t,i} - \sum_{t=1}^{T}\sum_{j}p_{tj}\widehat{Z}_{t,j}\right] \lesssim \frac{2\log K}{\eta} + \eta B^2\left(1+\sum_{j}\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right)T, \tag{97}$$

where $\lesssim$ drops terms of constant order. A use of Lemma 9 and setting $\eta = \frac{1}{B\sqrt{T}}\cdot\frac{\sqrt{\log K}}{\sqrt{1+2\sum_{j\in[K]}\frac{1-\alpha_j}{1+\alpha_j}}}$ conclude this case.

Case II: We have oracle access to q_{ij}. We pick up the argument from (90).

When the estimates for q_{ij} is exact, it holds that $\mathbb{E}\left[\frac{\mathbb{I}_{\left[j\in\mathcal{P}_{t,J_t}\right]}}{\widetilde{p}_{tj}}\right] = 1$ and $\left|\frac{\mathbb{I}_{\left[j\in\mathcal{P}_{t,J_t}\right]}}{\widetilde{p}_{tj}} - 1\right| \leq \frac{1}{\alpha}$. Thus by Azuma's inequality, we have, with probability at least $1-\epsilon$,

$$\sum_{t=1}^{T}\sum_{j}\frac{\mathbb{I}_{\left[j\in\mathcal{P}_{t,J_t}\right]}}{\widetilde{p}_{tj}}\cdot\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}} \leq \sum_{t=1}^{T}\sum_{j}\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}} + \sqrt{\frac{T}{\alpha}\left(\sum_{j}\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right)\log\frac{1}{\epsilon}}.$$

Plug this back to (90) and we get, with probability at least $1-\epsilon$,

$$\begin{aligned}
&\exp\left(\eta\sum_t \widehat{Z}_{t,i}\right) \qquad (98)\\
\leq& W_0\exp\left(\eta\widehat{S}_T+\frac{\eta^2B^2}{2}\sum_{t=1}^T\left(1+\sum_j\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}+\sum_j\frac{\mathbb{I}_{[j\in\mathcal{P}_{t,J_t}]}}{\widetilde{p}_{tj}}\cdot\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right)\right)\\
\leq& W_0\exp\left(\eta\widehat{S}_T+\frac{\eta^2B^2}{2}\left(T+2\sum_{t=1}^T\sum_j\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}+\sqrt{\frac{T}{\alpha}\left(\sum_j\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right)\log\frac{1}{\epsilon}}\right)\right).
\end{aligned}$$

Taking logarithm and rearranging terms give, under event $\mathcal{E}_T(B)$, with probability at least $1-\epsilon$,

$$\widehat{S}_{T,j}-\widehat{S}_T\leq\frac{1}{\eta}\log W_0+\frac{\eta B^2}{2}\left(T+2\sum_{t=1}^T\sum_j\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}+\sqrt{T\left(\sum_j\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right)\log\frac{1}{\epsilon}}\right).$$

Consider the martingale difference sequence $\left\{\left(\mathcal{L}\left(\mathcal{P}_{t,j}\right)-B-\widehat{Z}_{t,j}\right)\right\}_t$. We can apply Lemma 2 to this sequence and get, for any $\epsilon\in(0,1)$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{t=1}^T\left(\mathcal{L}\left(\mathcal{P}_{t,j}\right)-\widehat{Z}_{t,j}\right)\right|>\sqrt{2BT\log\left(1/\epsilon\right)}\right) \qquad (99)\\
\leq&\mathbb{P}\left(\left|\sum_{t=1}^T\left(\mathcal{L}\left(\mathcal{P}_{t,j}\right)-\widehat{Z}_{t,j}\right)\mathbb{I}_{[Z_{t,j}\leq B,\mathcal{L}(\mathcal{P}_{t,j})\leq B]}\right|>2\sqrt{BT\log\left(1/\epsilon\right)}\right)\\
&\quad+\sum_{t=1}^T\mathbb{P}\left(Z_{t,j}>B\right)+\sum_{t=1}^T\mathbb{P}\left(\mathcal{L}\left(\mathcal{P}_{t,j}\right)>B\right)\\
\leq&\epsilon+2\frac{\epsilon}{K}\leq 3\epsilon.
\end{aligned}$$

A similar argument gives

$$\mathbb{P}\left(\left|\sum_{t=1}^T\sum_j p_{tj}\widehat{Z}_{t,j}-\sum_{t=1}^T\sum_j p_{tj}\widehat{Z}_{t,j}\right|>\sqrt{2BT\log\left(1/\epsilon\right)}\right)\leq 3\epsilon.$$

Collecting above terms gives, with probability at least $1-7\delta$,

$$\begin{aligned}
S_{T,j}-S_T\leq&2\sqrt{2BT\log\left(1/\epsilon\right)}+\frac{1}{\eta}\log W_0\\
&+\frac{\eta B^2}{2}\left(\left(1+2\sum_j\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right)T+\sqrt{T\left(\sum_j\frac{1-\sqrt{\alpha_j}}{1+\sqrt{\alpha_j}}\right)\log\frac{1}{\epsilon}}\right).
\end{aligned}$$

Setting $\eta=\frac{1}{B\sqrt{T}}\cdot\frac{\sqrt{\log K}}{\sqrt{1+2\sum_{j\in[K]}\frac{1-\alpha_j}{1+\alpha_j}}}$ concludes this case.

□

C.2 Proof of Theorem 7

Theorem 7. *Fix any $T>\sqrt{128\log 8}$ and $\sigma<\frac{1}{7}$. On a graph of K nodes and any pair of nodes are connected with probability p, there exists a sequence of edge lengths, such that regret incurred by any policy satisfies*

$$\mathbb{P}\left(\mathbb{E}\left[\mathrm{Reg}_j^{\mathrm{adv}}(T)\right]\geq\frac{1}{4}\sqrt{\frac{(1-Kp)\,\sigma^2T}{\left(1+\frac{2p}{1-Kp}\right)}}\right)\geq\frac{3}{16},\quad\forall j\in[K]. \qquad (100)$$

Proof. A deterministic problem instance $\mathfrak{J}$ is represented by T graphs $\mathfrak{J} = (G_1, G_2, \cdots, G_T)$. For our purpose, a graph G_t consists of edge lengths: $G_t := \left(\{l_{i*}^{(t)}\}_{i\in[K]}, \{l_{ij}^{(t)}\}_{i,j\in[K]}\right)$, where $l_{i*}^{(t)}$ is the length from i to $*$ in G_t, and $l_{ij}^{(t)}$ is the length from i to j in G_t. A stochastic problem instance is represented by a distribution over deterministic problem instances.

By Proposition 4 (in Appendix C.3), it suffices to consider stochastic instances.

Next we construct stochastic problem instances to prove Theorem 7.

We first sample T *i.i.d.* Gaussian random variables $\eta_t \sim \mathcal{N}\left(0, \sigma^2\right)$ (σ to be specified later). Consider the stochastic problem instance $\mathfrak{J}$: $\mathfrak{J} = (G_1, G_2, \cdots, G_T)$. In G_t, $l_{0*}^{(t)} = \text{clip}\left(\frac{1}{2} + \frac{\epsilon}{1-Kp} + \eta_t\right)$, $l_{i*}^{(t)} = \text{clip}\left(\frac{1}{2} + \eta_t\right)$ $(i = 1, 2, \cdots, K-1)$, $l_{ij}^{(t)} = \text{clip}\left(\frac{1}{2} + \eta_t\right)$ $(i, j \in [K])$, where "clip" takes a number and clip its value to $[0, 1]$.

By this construction, the hitting times have the following properties. If no clipping happens, the hitting times $\boldsymbol{Z}_t = [Z_{t,0}, Z_{t,1}, \cdots, Z_{t,K-1}]$ at time t satisfies

$$\mathbb{E}\left[\boldsymbol{Z}_t\right] = (1-Kp)\left(\frac{1}{2} + \frac{\epsilon}{1-Kp}\boldsymbol{e}_1\eta_t\mathbf{1}\right) \tag{101}$$

$$+ M\mathbb{E}\left[\boldsymbol{Z}_t\right] + Kp\left(\frac{1}{2} + \eta_t\mathbf{1}\right), \tag{102}$$

where $\boldsymbol{e}_1 = [1, 0, 0, \cdots, 0]^\top$ and $\mathbf{1} = [1, 1, \cdots, 1]^\top$. The right-hand-side in (101) accounts for the edge lengths (hitting time) from hitting absorbing node. The terms in (101) accounts for the edge lengths (hitting time) from remaining in the transient nodes.

This gives,

$$(I - M)\,\mathbb{E}\left[\boldsymbol{Z}_t\right] = \frac{1}{2} + \epsilon\boldsymbol{e}_1 + \eta_t\mathbf{1}, \quad \mathbb{E}\left[\boldsymbol{Z}_t\right] = \left(I + \frac{M}{1-Kp}\right)\left(\frac{1}{2} + \epsilon\boldsymbol{e}_1 + \eta_t\mathbf{1}\right).$$

If $\eta_t \in \left[-\frac{1}{2}, \frac{1}{2} - \frac{\epsilon}{1-Kp}\right]$, no clipping happens and $\mathbb{E}\left[Z_{t,0}\right] \geq \mathbb{E}\left[Z_{t,j}\right] + \epsilon$ for all $j = 1, 2, \cdots, K-1$. If $\eta_t \notin \left[-\frac{1}{2}, \frac{1}{2} - \frac{\epsilon}{1-Kp}\right]$, some edges are clipped and $\mathbb{E}\left[Z_{t,0}\right] \geq \mathbb{E}\left[Z_{t,j}\right]$. Thus we have, for all $j \geq 1$,

$$\mathbb{E}\left[Z_{t,0}\right] \geq \mathbb{E}\left[Z_{t,j}\right] + \epsilon\mathbb{I}_{\left[\eta_t \in \left[-\frac{1}{2}, \frac{1}{2} - \frac{\epsilon}{1-Kp}\right]\right]}. \tag{103}$$

Consider another instance $\mathfrak{J}'$ where everything is identical with the previously constructed instance $\mathfrak{J}$ except that node 0 is optimal in $\mathfrak{J}$ and node 1 is optimal in $\mathfrak{J}'$. More specifically, $\mathfrak{J}' = (G_1', G_2', \cdots, G_T')$ and $G_t' = \left(\{l_{i*}^{(t)'}\}_{i\in[K]}, \{l_{ij}^{(t)'}\}_{i,j\in[K]}\right)$, where $l_{i*}^{(t)'}$ is the length from i to $*$ in G_t', and $l_{ij}^{(t)'}$ is the length from i to j in G_t'. We again sample $\eta_t' \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ (σ to be specified later). The instance $\mathfrak{J}'$ is constructed by $l_{1*}^{(t)'} = \text{clip}\left(\frac{1}{2} + \frac{\epsilon}{1-Kp} + \eta_t'\right)$, $l_{i*}^{(t)'} = \text{clip}\left(\frac{1}{2} + \eta_t'\right)$ $(i \neq 1)$, and $l_{ij}^{(t)'} = \text{clip}\left(\frac{1}{2} + \eta_t'\right)$ for all $i, j \in [K]$.

Let $\boldsymbol{Z}_t' = \left[Z_{t,0}', Z_{t,1}', \cdots, Z_{t,K-1}'\right]$ be the hitting times at time t in instance $\mathfrak{J}'$. Thus for $j \neq 1$, in instance $\mathfrak{J}'$,

$$\mathbb{E}\left[Z_{t,1}'\right] \geq \mathbb{E}\left[Z_{t,j}'\right] + \epsilon\mathbb{I}_{\left[\eta_t' \in \left[-\frac{1}{2}, \frac{1}{2} - \frac{\epsilon}{1-Kp}\right]\right]} \tag{104}$$

From (103), we know, in instance $\mathfrak{J}$, when there are at least $\frac{3}{4}T$ unclipped rounds and arm 1 is played *no more than* $\frac{1}{2}T$ times, then in at least $\frac{1}{4}T$ rounds, a regret of ϵ is incurred. Similarly, in instance $\mathfrak{J}'$, when there are at least $\frac{3}{4}T$ unclipped rounds and arm 1 is played *more than* $\frac{1}{2}T$ times, then in at least $\frac{1}{4}T$ rounds, a regret of ϵ is incurred.

Now we define some notations to write the above observations symbolically. Let $\mathbb{P}_{\mathfrak{J},\pi}$ (resp. $\mathbb{P}_{\mathfrak{J}',\pi}$) be the probability measure on running π on $\mathfrak{J}$ (resp. $\mathfrak{J}'$). Let N_i (resp. N_i') be the number of times i is played in $\mathfrak{J}$ (resp. $\mathfrak{J}'$). Let $\text{Reg}_0^{\text{adv}}(T)$ (resp. $\text{Reg}_1^{\text{adv}'}(T)$) be the regret in $\mathfrak{J}$ against node 0 (resp. in $\mathfrak{J}'$ against node 1). Let $u = \frac{1}{4}\epsilon T$. Let $W = \sum_t \mathbb{I}_{\left[\eta_t \in \left[-\frac{1}{2}, \frac{1}{2} + \frac{\epsilon}{1-Kp}\right]\right]}$, and $W' = \sum_t \mathbb{I}_{\left[\eta_t' \in \left[-\frac{1}{2}, \frac{1}{2} + \frac{\epsilon}{1-Kp}\right]\right]}$. Using the above notations, we have

$$\begin{aligned}
&\mathbb{P}_{\mathfrak{J},\pi}\left(\mathbb{E}\left[\text{Reg}_0^{\text{adv}}(T)\right] \geq u\right) \\
\geq &\mathbb{P}_{\mathfrak{J},\pi}\left(N_1 < T/2 \quad \text{and} \quad W \geq 3T/4\right) \geq \mathbb{P}_{\mathfrak{J},\pi}\left(N_1 < T/2\right) - \mathbb{P}_{\mathfrak{J},\pi}\left(W < 3T/4\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}_{\mathfrak{J}',\pi}\left(\mathbb{E}\left[\text{Reg}_1^{\text{adv}'}(T)\right] \geq u\right) \\
\geq &\mathbb{P}_{\mathfrak{J}',\pi}\left(N_1' \geq T/2 \quad \text{and} \quad W' \geq 3T/4\right) \geq \mathbb{P}_{\mathfrak{J}',\pi}\left(N_1' \geq T/2\right) - \mathbb{P}_{\mathfrak{J}',\pi}\left(W' < 3T/4\right),
\end{aligned}$$

which give

$$\begin{aligned}&\mathbb{P}_{\mathfrak{J},\pi}\left(\mathbb{E}\left[\mathrm{Reg}_0^{\mathrm{adv}}(T)\right]\geq u\right)+\mathbb{P}_{\mathfrak{J}',\pi}\left(\mathbb{E}\left[\mathrm{Reg}_1^{\mathrm{adv}'}(T)\right]\geq u\right)\\ \geq&\mathbb{P}_{\mathfrak{J},\pi}\left(N_1<T/2\right)-\mathbb{P}_{\mathfrak{J},\pi}\left(W<3T/4\right)+\mathbb{P}_{\mathfrak{J}',\pi}\left(N_1'\geq T/2\right)-\mathbb{P}_{\mathfrak{J}',\pi}\left(W'<3T/4\right).\end{aligned}\tag{105}$$

The quantities $\mathbb{P}_{\mathfrak{J},\pi}\left(W<3T/4\right)$ and $\mathbb{P}_{\mathfrak{J}',\pi}\left(W'<3T/4\right)$ can be easily handled since η_t are Gaussian (Proposition 3). Now we turn to lower bound $\mathbb{P}_{\mathfrak{J},\pi}\left(N_1<T/2\right)+\mathbb{P}_{\mathfrak{J}',\pi}\left(N_1'\geq T/2\right)$, and then select a proper ϵ to maximize this lower bound.

By the definition of total variation and the Pinsker's inequality,

$$\begin{aligned}\mathbb{P}_{\mathfrak{J},\pi}\left(N_1\geq T/2\right)+\mathbb{P}_{\mathfrak{J}',\pi}\left(N_1'<T/2\right)&=1+\mathbb{P}_{\mathfrak{J},\pi}\left(N_1\geq T/2\right)-\mathbb{P}_{\mathfrak{J}',\pi}\left(N_1'\geq T/2\right)\\ &\geq 1-d_{TV}\left(\mathbb{P}_{\mathfrak{J},\pi},\mathbb{P}_{\mathfrak{J}',\pi}\right)\\ &\geq 1-\sqrt{2D_{KL}\left(\mathbb{P}_{\mathfrak{J},\pi},\mathbb{P}_{\mathfrak{J},\pi}\right)}.\end{aligned}$$

Let $\mathbb{Q}_{t,j}$ (resp. $\mathbb{Q}'_{t,j}$) be the probability space generated by playing j at t in instance $\mathfrak{J}$ (resp. $\mathfrak{J}'$). By chain rule, we have

$$D_{KL}\left(\mathbb{P}_{\mathfrak{J},\pi}\|\mathbb{P}_{\mathfrak{J},\pi}\right)=\sum_{t=1}^{T}\sum_{j\in[K]}\mathbb{P}\left(J_t=j\right)D_{KL}\left(\mathbb{Q}_{t,j}\|\mathbb{Q}'_{t,j}\right).\tag{106}$$

Let $X_0,L_1,X_1,L_2,\cdots$ be the nodes and edge length of each step in the trajectory after playing a node. The sample space of $\mathbb{Q}_{t,j}$ and $\mathbb{Q}'_{t,j}$ is spanned by $X_0,L_1,X_1,L_2,\cdots$.

By Markov property, we have, for all $i,j\in[K]$ and $k\in\mathbb{N}_+$,

$$\begin{aligned}&\mathbb{Q}_{t,i}\left(L_{k+1},X_{k+1},L_{k+2},X_{k+2},\ldots|X_k=j\right)=\mathbb{Q}_{t,j}.\\ &\mathbb{Q}'_{t,j}\left(L_{k+1},X_{k+1},L_{k+2},X_{k+2},\ldots|X_k=j\right)=\mathbb{Q}'_{t,j}.\end{aligned}\tag{107}$$

Thus by chain rule,

$$\begin{aligned}&D_{KL}\left(\mathbb{Q}_{t,i}\|\mathbb{Q}'_{t,i}\right)\\ =&D_{KL}\left(\mathbb{Q}_{t,i}(X_1,L_1)\|\mathbb{Q}'_{t,i}(X_1,L_1)\right)\\ +&\sum_{x\in[K]}\int_0^1\mathbb{P}(X_1=x,L_1=l)D_{KL}\big(\mathbb{Q}_{t,i}(X_2,L_2,\ldots|X_1=x,L_1=l)\|\mathbb{Q}'_{t,i}(X_2,L_2,\ldots|X_1=x,L_1=l)\big)dl\\ =&D_{KL}\left(\mathbb{Q}_{t,i}(X_1,L_1)\|\mathbb{Q}'_{t,i}(X_1,L_1)\right)+\sum_{j\in[K]}m_{ji}D_{KL}\left(\mathbb{Q}_{t,j}\|\mathbb{Q}'_{t,j}\right).\end{aligned}\tag{108}$$

Let f be the *p.d.f.* of $\mathcal{N}\left(\frac{1}{2},\sigma^2\right)$ truncated to $[0,1]$. Let f^* be the *p.d.f.* of $\mathcal{N}\left(\frac{1}{2}+\frac{\epsilon}{1-Kp},\sigma^2\right)$ clipped to $[0,1]$. Let ϕ (resp. Φ) be the *p.d.f.* (resp. *c.d.f.*) of the standard normal distribution. Thus we have

$$\begin{aligned}&D_{KL}\left(\mathbb{Q}_{t,1}(X_1,L_1)\|\mathbb{Q}'_{t,1}(X_1,L_1)\right)\\ =&\int_0^1(1-Kp)f(z)\log\frac{(1-Kp)f(z)}{(1-Kp)f^*(z)}\,dz+K\int_0^1 pf(z)\log\frac{pf(z)}{pf(z)}\,dz\\ =&(1-Kp)\int_0^1 f(z)\log\frac{f(z)}{f^*(z)}\,dz\\ =&(1-Kp)D_{KL}\left(\mathcal{N}\left(\frac{1}{2},\sigma^2\right)\left|\mathrm{clip},\mathcal{N}\left(\frac{1}{2}+\frac{\epsilon}{1-Kp},\sigma^2\right)\right|\mathrm{clip}\right)\\ \leq&(1-Kp)D_{KL}\left(\mathcal{N}\left(\frac{1}{2},\sigma^2\right),\mathcal{N}\left(\frac{1}{2}+\frac{\epsilon}{1-Kp},\sigma^2\right)\right)\\ =&\frac{\epsilon^2}{2(1-Kp)\sigma^2},\end{aligned}\tag{109}$$

where (109) uses monotonicity of f-divergence (e.g., Csiszár and Shields (2004)).

Similarly,

$$D_{KL}\left(\mathbb{Q}'_{t,1}(X_1,L_1)\|\mathbb{Q}_{t,1}(X_1,L_1)\right)\leq\frac{\epsilon^2}{2(1-Kp)\sigma^2}$$

and

$$D_{KL}\left(\mathbb{Q}_{t,i}(X_1,L_1)\|\mathbb{Q}'_{t,i}(X_1,L_1)\right)=0$$

for $i\geq 2$.

Next, define

$$D=\left[D_{KL}\left(\mathbb{Q}_{t,0}\|\mathbb{Q}'_{t,1}\right),D_{KL}\left(\mathbb{Q}_{t,2}\|\mathbb{Q}'_{t,2}\right),\ldots,D_{KL}\left(\mathbb{Q}_{t,K-1}\|\mathbb{Q}'_{t,K-1}\right)\right]^\top,$$

and

$$c=\left[D_{KL}\left(\mathbb{Q}_{t,0}(X_1,L_1)\|\mathbb{Q}'_{t,0}(X_1,L_1)\right),\ldots,D_{KL}\left(\mathbb{Q}_{t,K-1}(X_1,L_1)\|\mathbb{Q}'_{t,K-1}(X_1,L_1)\right)\right]^\top.$$

Then we can rewrite (108) as

$$\begin{aligned}D&=MD+c\\ D&=\left(I+\frac{M}{1-Kp}\right)c.\end{aligned}$$

Solving this gives, for all $i\in[K]$,

$$\begin{aligned}&D_{KL}\left(\mathbb{Q}_{t,i}\|\mathbb{Q}'_{t,i}\right)\\ =&D_{KL}\left(\mathbb{Q}_{t,i}(X_1,L_1)\|\mathbb{Q}'_{t,i}(X_1,L_1)\right)+\frac{p}{1-Kp}\sum_{j\in[K]}D_{KL}\left(\mathbb{Q}_{t,j}(X_1,L_1)\|\mathbb{Q}'_{t,j}(X_1,L_1)\right)\\ \leq&\begin{cases}\left(1+\frac{2p}{1-Kp}\right)\frac{\epsilon^2}{2(1-Kp)\sigma^2}, & \text{if } i=0,1,\\ \frac{p\epsilon^2}{(1-Kp)^2\sigma^2}, & \text{otherwise.}\end{cases}\end{aligned}$$

Plugging above computation back to (106), we have

$$\begin{aligned}D_{KL}\left(\mathbb{P}_{\mathfrak{J},\pi}\|\mathbb{P}_{\mathfrak{J}',\pi}\right)&=\sum_{t=1}^{T}\sum_{i\in[K]}\mathbb{P}\left(J_t=j\right)D_{KL}\left(\mathbb{Q}_{t,i}\|\mathbb{Q}'_{t,i}\right)\\ &\leq\sum_{i\in[K]}\sum_{t=1}^{T}\mathbb{P}\left(J_t=j\right)\left(1+\frac{2p}{1-Kp}\right)\frac{\epsilon^2}{2(1-Kp)\sigma^2}\\ &=T\left(1+\frac{2p}{1-Kp}\right)\frac{\epsilon^2}{2(1-Kp)\sigma^2}.\end{aligned}$$

Thus by Pinsker's inequality,

$$d_{TV}\left(\mathbb{P}_{\mathfrak{J},\pi},\mathbb{P}_{\mathfrak{J}',\pi}\right)\leq\sqrt{2T\left(1+\frac{2p}{1-Kp}\right)\frac{\epsilon^2}{2(1-Kp)\sigma^2}}.\tag{110}$$

Thus, from the definition of total variation,

$$\begin{aligned}1+\mathbb{P}_{\mathfrak{J},\pi}\left(N_1\geq T/2\right)-\mathbb{P}_{\mathfrak{J}',\pi}\left(N_1<T/2\right)&\geq 1-d_{TV}(\mathbb{P}_{\mathfrak{J},\pi},Pr_{\mathfrak{J}',\pi})\\ &\geq 1-\sqrt{T\left(1+\frac{2p}{1-Kp}\right)\frac{\epsilon^2}{(1-Kp)\sigma^2}}\end{aligned}$$

By picking $\epsilon=\sqrt{\frac{4(1-Kp)\sigma^2}{T\left(1+\frac{2p}{1-Kp}\right)}}$, $1+\mathbb{P}_{\mathfrak{J},\pi}\left(N_1\geq T/2\right)-\mathbb{P}_{\mathfrak{J}',\pi}\left(N_1<T/2\right)\geq\frac{1}{2}$. Applying the above results to (105) gives,

$$\begin{aligned}&\mathbb{P}_{\mathfrak{J},\pi}\left(\mathbb{E}\left[\mathrm{Reg}_0^{\mathrm{adv}}(T)\right]\geq\frac{1}{2}\sqrt{\frac{(1-Kp)\sigma^2T}{\left(1+\frac{2p}{1-Kp}\right)}}\right)\\ &+\mathbb{P}_{\mathfrak{J}',\pi}\left(\mathbb{E}\left[\mathrm{Reg}_1^{\mathrm{adv}'}(T)\right]\geq\frac{1}{2}\sqrt{\frac{(1-Kp)\sigma^2T}{\left(1+\frac{2p}{1-Kp}\right)}}\right)\\ \geq&1+\mathbb{P}_{\mathfrak{J},\pi}\left(N_1\geq T/2\right)-\mathbb{P}_{\mathfrak{J}',\pi}\left(N_1'\geq T/2\right)-\mathbb{P}\left(W<3T/4\right)-\mathbb{P}\left(W'<3T/4\right)\\ \geq&\frac{3}{8},\end{aligned}$$

where we use Proposition 2 to get terms involving W and W'.

This means either

$$\mathbb{P}_{\mathfrak{J},\pi}\left(\mathbb{E}\left[\mathrm{Reg}_0^{\mathrm{adv}}(T)\right] \geq \frac{1}{2}\sqrt{\frac{(1-Kp)\,\sigma^2 T}{\left(1+\frac{2p}{1-Kp}\right)}}\right) \geq \frac{3}{16}$$

or

$$\mathbb{P}_{\mathfrak{J}',\pi}\left(\mathbb{E}\left[\mathrm{Reg}_1^{\mathrm{adv}'}(T)\right] \geq \frac{1}{2}\sqrt{\frac{(1-Kp)\,\sigma^2 T}{\left(1+\frac{2p}{1-Kp}\right)}}\right) \geq \frac{3}{16},$$

which concludes the proof.

□

C.3 Additional Propositions

Proposition 2. *Fix any $a \in (0,1]$. We have*

$$\frac{x}{x+(1-a)x} \leq x + \frac{1-\sqrt{a}}{1+\sqrt{a}}, \qquad \forall x \in (0,1). \tag{111}$$

Proof. If suffices to show, for any $a \in (0,1]$, the function $f_a(x) := \frac{x}{x+(1-x)a} - x$ is upper bounded by $\frac{1-\sqrt{a}}{1+\sqrt{a}}$. This can be shown via a quick first-order test. At $x_{\max} = \frac{\sqrt{a}}{1+\sqrt{a}}$, the maximum of f_a is achieved, and $f_a(x_{\max}) = \frac{1-\sqrt{a}}{1+\sqrt{a}}$. □

Proposition 3. *Fix any $\sigma < \frac{1}{7}$. Pick T such that $T > 128\log 8$, and ϵ such that $\frac{\epsilon}{1-Kp} \leq \frac{1}{4}$. Then $\mathbb{P}\left(W < \frac{3}{4}T\right) \leq \frac{1}{8}$ and $\mathbb{P}\left(W' < \frac{3}{4}T\right) \leq \frac{1}{8}$.*

Proof. Recall $W = \sum_{t=1}^{T} \mathbb{I}_{\left[\eta_t \in \left[-\frac{1}{2}, \frac{1}{2} - \frac{\epsilon}{1-Kp}\right]\right]}$. Since $\eta_t \in \mathcal{N}\left(0, \sigma^2\right)$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{I}_{\left[\eta_t \in \left[-\frac{1}{2}, \frac{1}{2} - \frac{\epsilon}{1-Kp}\right]\right]}\right] &= \mathbb{P}\left(\eta_t \in \left[-\frac{1}{2}, \frac{1}{2} - \frac{\epsilon}{1-Kp}\right]\right) \\
&\geq \mathbb{P}\left(\eta_t \in \left[-\frac{1}{4}, \frac{1}{4}\right]\right) && \left(\text{since } \tfrac{\epsilon}{1-Kp} \leq \tfrac{1}{4}\right) \\
&= 1 - \mathbb{P}\left(|\eta_t| > \frac{1}{4}\right) \\
&= 1 - 2\exp\left(-\frac{1}{16\sigma^2}\right) && \left(\text{since } \eta_t \text{ is } \sigma^2\text{-sub-Gaussian}\right) \\
&\geq \frac{7}{8}. && \left(\text{since } \sigma \leq \tfrac{1}{7}\right)
\end{aligned}$$

By Hoeffding's inequality,

$$\mathbb{P}\left(W < \frac{3}{4}T\right) \leq \mathbb{P}\left(W < \frac{7}{8}T - \sqrt{2T\log 8}\right) \leq \mathbb{P}\left(W < \mathbb{E}\left[W\right] - \sqrt{2T\log 8}\right) \leq \frac{1}{8}. \tag{112}$$

□

Proposition 4. *For any distribution Q over problem instances and policy π, let $\mathbb{P}_{Q,\pi}$ be the probability of running π for T steps on a problem instance sampled from Q. For any problem instance $\mathfrak{J}$, let $\mathbb{P}_{\mathfrak{J},\pi}$ be the probability of running π for T steps on problem instance $\mathfrak{J}$. Then for any Q, π and event A and $u \in (0,1)$, if $\mathbb{P}_{Q,\pi}(Q) \geq u$, then there exists $\mathfrak{J} \in support(Q)$, such that $\mathbb{P}_{\mathfrak{J},\pi}(A) \geq u$.*

Proof. For any event A,

$$\mathbb{P}_{Q,\pi}(A) = \int_{\mathfrak{J} \in support(Q)} \mathbb{P}_{\mathfrak{J},\pi}(A) dQ(\mathfrak{J}). \tag{113}$$

From above, it is clear that if $\mathbb{P}_{\mathfrak{J},\pi}(A) \leq u$ for all $\mathfrak{J} \in support(Q)$, then it is impossible to have $\mathbb{P}_{Q,\pi}(A) > u$. □