Journal of Machine Learning Research 8 (2007) 2369-2403 Submitted 4/07; Revised 8/07; Published 10/07

The On-Line Shortest Path Problem Under Partial Monitoring

András György GYA@SZIT.BME.HU
Machine Learning Research Group
Computer and Automation Research Institute
Hungarian Academy of Sciences
Kende u. 13-17, Budapest, Hungary, H-1111

Tamás Linder LINDER@MAST.QUEENSU.CA
Department of Mathematics and Statistics
Queen's University, Kingston, Ontario
Canada K7L 3N6

Gábor Lugosi GABOR.LUGOSI@GMAIL.COM
ICREA and Department of Economics
Universitat Pompeu Fabra
Ramon Trias Fargas 25-27
08005 Barcelona, Spain

György Ottucsák OTI@SZIT.BME.HU
Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Magyar Tudósok Körútja 2.
Budapest, Hungary, H-1117

Editor: Leslie Pack Kaelbling

Abstract

The on-line shortest path problem is considered under various models of partial monitoring. Given a weighted directed acyclic graph whose edge weights can change in an arbitrary (adversarial) way, a decision maker has to choose in each round of a game a path between two distinguished vertices such that the loss of the chosen path (defined as the sum of the weights of its composing edges) be as small as possible. In a setting generalizing the multi-armed bandit problem, after choosing a path, the decision maker learns only the weights of those edges that belong to the chosen path. For this problem, an algorithm is given whose average cumulative loss in n rounds exceeds that of the best path, matched off-line to the entire sequence of the edge weights, by a quantity that is proportional to $1/\sqrt{n}$ and depends only polynomially on the number of edges of the graph. The algorithm can be implemented with complexity that is linear in the number of rounds n (i.e., the average complexity per round is constant) and in the number of edges. An extension to the so-called label efficient setting is also given, in which the decision maker is informed about the weights of the edges corresponding to the chosen path at a total of $m \ll n$ time instances. Another extension is shown where the decision maker competes against a time-varying path, a generalization of the problem of tracking the best expert. A version of the multi-armed bandit setting for shortest path is also discussed where the decision maker learns only the total weight of the chosen path but not the weights of the individual edges on the path. Applications to routing in packet switched networks along with simulation results are also presented.

Keywords: on-line learning, shortest path problem, multi-armed bandit problem

1. Introduction

In a sequential decision problem, a decision maker (or forecaster) performs a sequence of actions. After each action the decision maker suffers some loss, depending on the response (or state) of the environment, and its goal is to minimize its cumulative loss over a certain period of time. In the setting considered here, no probabilistic assumption is made on how the losses corresponding to different actions are generated. In particular, the losses may depend on the previous actions of the decision maker, whose goal is to perform well relative to a set of reference forecasters (the so-called "experts") for any possible behavior of the environment. More precisely, the aim of the decision maker is to achieve asymptotically the same average (per round) loss as the best expert.

Research into this problem started in the 1950s (see, for example, Blackwell, 1956 and Hannan, 1957 for some of the basic results) and gained new life in the 1990s following the work of Vovk (1990), Littlestone and Warmuth (1994), and Cesa-Bianchi et al. (1997). These results show that for any bounded loss function, if the decision maker has access to the past losses of all experts, then it is possible to construct on-line algorithms that perform, for any possible behavior of the environment, almost as well as the best of N experts. More precisely, the per round cumulative loss of these algorithms is at most as large as that of the best expert plus a quantity proportional to $\sqrt{\ln N/n}$ for any bounded loss function, where n is the number of rounds in the decision game. The logarithmic dependence on the number of experts makes it possible to obtain meaningful bounds even if the pool of experts is very large.

In certain situations the decision maker has only limited knowledge about the losses of all possible actions. For example, it is often natural to assume that the decision maker gets to know only the loss corresponding to the action it has made, and has no information about the loss it would have suffered had it made a different decision. This setup is referred to as the *multi-armed bandit problem*, and was considered, in the adversarial setting, by Auer et al. (2002) who gave an algorithm whose normalized regret (the difference of the algorithm's average loss and that of the best expert) is upper bounded by a quantity which is proportional to $\sqrt{N \ln N/n}$. Note that, compared to the *full information* case described above where the losses of all possible actions are revealed to the decision maker, there is an extra $\sqrt{N}$ factor in the performance bound, which seriously limits the usefulness of the bound if the number of experts is large.

Another interesting example for the limited information case is the so-called *label efficient decision problem* (see Helmbold and Panizza, 1997) in which it is too costly to observe the state of the environment, and so the decision maker can query the losses of all possible actions for only a limited number of times. A recent result of Cesa-Bianchi, Lugosi, and Stoltz (2005) shows that in this case, if the decision maker can query the losses m times during a period of length n, then it can achieve $O(\sqrt{\ln N/m})$ normalized regret relative to the best expert.

In many applications the set of experts has a certain structure that may be exploited to construct efficient on-line decision algorithms. The construction of such algorithms has been of great interest in computational learning theory. A partial list of works dealing with this problem includes Herbster and Warmuth (1998), Vovk (1999), Bousquet and Warmuth (2002), Schapire and Helmbold (1997), Takimoto and Warmuth (2003), Kalai and Vempala (2003) and György et al. (2004a,b, 2005a). For a more complete survey, we refer to Cesa-Bianchi and Lugosi (2006, Chapter 5).

In this paper we study the on-line shortest path problem, a representative example of structured expert classes that has received attention in the literature for its many applications, including, among others, routing in communication networks; see, for example, Takimoto and Warmuth (2003), Awer-

buch et al. (2005), or György and Ottucsák (2006), and adaptive quantizer design in zero-delay lossy source coding; see, György et al. (2004a,b, 2005b). In this problem, a weighted directed (acyclic) graph is given whose edge weights can change in an arbitrary manner, and the decision maker has to pick in each round a path between two given vertices, such that the weight of this path (the sum of the weights of its composing edges) be as small as possible.

Efficient solutions, with time and space complexity proportional to the number of edges rather than to the number of paths (the latter typically being exponential in the number of edges), have been given in the full information case, where in each round the weights of all the edges are revealed after a path has been chosen; see, for example, Mohri (1998), Takimoto and Warmuth (2003), Kalai and Vempala (2003), and György et al. (2005a).

In the bandit setting only the weights of the edges or just the sum of the weights of the edges composing the chosen path are revealed to the decision maker. If one applies the general bandit algorithm of Auer et al. (2002), the resulting bound will be too large to be of practical use because of its square-root-type dependence on the number of paths N. On the other hand, using the special graph structure in the problem, Awerbuch and Kleinberg (2004) and McMahan and Blum (2004) managed to get rid of the exponential dependence on the number of edges in the performance bound. They achieved this by extending the exponentially weighted average predictor and the follow-the-perturbed-leader algorithm of Hannan (1957) to the generalization of the multi-armed bandit setting for shortest paths, when only the sum of the weights of the edges is available for the algorithm. However, the dependence of the bounds obtained in Awerbuch and Kleinberg (2004) and McMahan and Blum (2004) on the number of rounds n is significantly worse than the $O(1/\sqrt{n})$ bound of Auer et al. (2002). Awerbuch and Kleinberg (2004) consider the model of "non-oblivious" adversaries for shortest path (i.e., the losses assigned to the edges can depend on the previous actions of the forecaster) and prove an $O(n^{-1/3})$ bound for the expected normalized regret. McMahan and Blum (2004) give a simpler algorithm than in Awerbuch and Kleinberg (2004) however obtain a bound of the order of $O(n^{-1/4})$ for the expected regret.

In this paper we provide an extension of the bandit algorithm of Auer et al. (2002) unifying the advantages of the above approaches, with a performance bound that is polynomial in the number of edges, and converges to zero at the right $O(1/\sqrt{n})$ rate as the number of rounds increases. We achieve this bound in a model which assumes that the losses of all edges on the path chosen by the forecaster are available separately after making the decision. We also discuss the case (considered by Awerbuch and Kleinberg, 2004 and McMahan and Blum, 2004) in which only the total loss (i.e., the sum of the losses on the chosen path) is known to the decision maker. We exhibit a simple algorithm which achieves an $O(n^{-1/3})$ normalized regret *with high probability* against "non-oblivious" adversary. In this case it remains an open problem to find an algorithm whose cumulative loss is polynomial in the number of edges of the graph and decreases as $O(n^{-1/2})$ with the number of rounds. Throughout the paper we assume that the number of rounds n in the prediction game is known in advance to the decision maker.

In Section 2 we formally define the on-line shortest path problem, which is extended to the multi-armed bandit setting in Section 3. Our new algorithm for the shortest path problem in the bandit setting is given in Section 4 together with its performance analysis. The algorithm is extended to solve the shortest path problem in a combined label efficient multi-armed bandit setting in Section 5. Another extension, when the algorithm competes against a time-varying path is studied in Section 6. An algorithm for the "restricted" multi-armed bandit setting (when only the sums

of the losses of the edges are available) is given in Section 7. Simulation results are presented in Section 8.

2. The Shortest Path Problem

Consider a network represented by a set of vertices connected by edges, and assume that we have to send a stream of packets from a distinguished vertex, called *source*, to another distinguished vertex, called *destination*. At each time slot a packet is sent along a chosen route connecting source and destination. Depending on the traffic, each edge in the network may have a different delay, and the total delay the packet suffers on the chosen route is the sum of delays of the edges composing the route. The delays may change from one time slot to the next one in an arbitrary way, and our goal is to find a way of choosing the route in each time slot such that the sum of the total delays over time is not significantly more than that of the best fixed route in the network. This adversarial version of the routing problem is most useful when the delays on the edges can change dynamically, even depending on our previous routing decisions. This is the situation in the case of ad-hoc networks, where the network topology can change rapidly, or in certain secure networks, where the algorithm has to be prepared to handle denial of service attacks, that is, situations where willingly malfunctioning vertices and links increase the delay; see, for example, Awerbuch et al. (2005).

This problem can be cast naturally as a sequential decision problem in which each possible route is represented by an action. However, the number of routes is typically exponentially large in the number of edges, and therefore computationally efficient algorithms are called for. Two solutions of different flavor have been proposed. One of them is based on a follow-the-perturbed-leader forecaster, see Kalai and Vempala (2003), while the other is based on an efficient computation of the exponentially weighted average forecaster, see, for example, Takimoto and Warmuth (2003). Both solutions have different advantages and may be generalized in different directions.

To formalize the problem, consider a (finite) directed acyclic graph with a set of edges $E = \{e_1, \ldots, e_{|E|}\}$ and a set of vertices V. Thus, each edge $e \in E$ is an ordered pair of vertices (v_1, v_2). Let u and v be two distinguished vertices in V. A *path* from u to v is a sequence of edges $e^{(1)}, \ldots, e^{(k)}$ such that $e^{(1)} = (u, v_1)$, $e^{(j)} = (v_{j-1}, v_j)$ for all $j = 2, \ldots, k-1$, and $e^{(k)} = (v_{k-1}, v)$. Let $\mathcal{P} = \{i_1, \ldots, i_N\}$ denote the set of all such paths. For simplicity, we assume that every edge in E is on some path from u to v and every vertex in V is an endpoint of an edge (see Figure 1 for examples).

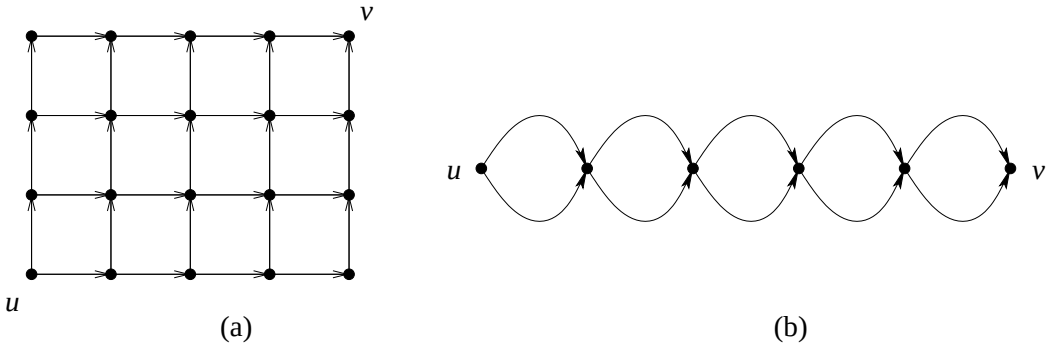

Figure 1: Two examples of directed acyclic graphs for the shortest path problem.

In each round $t = 1, \ldots, n$ of the decision game, the decision maker chooses a path I_t among all paths from u to v. Then a loss $\ell_{e,t} \in [0,1]$ is assigned to each edge $e \in E$. We write $e \in i$ if the edge $e \in E$ belongs to the path $i \in \mathcal{P}$, and with a slight abuse of notation the loss of a path i at time slot t is also represented by $\ell_{i,t}$. Then $\ell_{i,t}$ is given as

$$\ell_{i,t} = \sum_{e \in i} \ell_{e,t}$$

and therefore the cumulative loss up to time t of each path i takes the additive form

$$L_{i,t} = \sum_{s=1}^{t} \ell_{i,s} = \sum_{e \in i} \sum_{s=1}^{t} \ell_{e,s}$$

where the inner sum on the right-hand side is the loss accumulated by edge e during the first t rounds of the game. The cumulative loss of the algorithm is

$$\widehat{L}_t = \sum_{s=1}^{t} \ell_{I_s,s} = \sum_{s=1}^{t} \sum_{e \in I_s} \ell_{e,s} \ .$$

It is well known that for a general loss sequence, the decision maker must be allowed to use randomization to be able to approximate the performance of the best expert (see, e.g., Cesa-Bianchi and Lugosi, 2006). Therefore, the path I_t is chosen randomly according to some distribution p_t over all paths from u to v. We study the normalized regret over n rounds of the game

$$\frac{1}{n}\left(\widehat{L}_n - \min_{i \in \mathcal{P}} L_{i,n}\right)$$

where the minimum is taken over all paths i from u to v.

For example, the exponentially weighted average forecaster (Vovk, 1990; Littlestone and Warmuth, 1994; Cesa-Bianchi et al., 1997), calculated over all possible paths, has regret

$$\frac{1}{n}\left(\widehat{L}_n - \min_{i \in \mathcal{P}} L_{i,n}\right) \leq K\left(\sqrt{\frac{\ln N}{2n}} + \sqrt{\frac{\ln(1/\delta)}{2n}}\right)$$

with probability at least $1-\delta$, where N is the total number of paths from u to v in the graph and K is the length of the longest path.

3. The Multi-Armed Bandit Setting

In this section we discuss the "bandit" version of the shortest path problem. In this setup, which is more realistic in many applications, the decision maker has only access to the losses corresponding to the paths it has chosen. For example, in the routing problem this means that information is available on the delay of the route the packet is sent on, and not on other routes in the network.

We distinguish between two types of bandit problems, both of which are natural generalizations of the simple bandit problem to the shortest path problem. In the first variant, the decision maker has access to the losses of those edges that are on the path it has chosen. That is, after choosing a path I_t at time t, the value of the loss $\ell_{e,t}$ is revealed to the decision maker if and only if $e \in I_t$. We study this case and its extensions in Sections 4, 5, and 6.

The second variant is a more restricted version in which the loss of the chosen path is observed, but no information is available on the individual losses of the edges belonging to the path. That is, after choosing a path I_t at time t, only the value of the loss of the path $\ell_{I_t,t}$ is revealed to the decision maker. Further on we call this setting as the *restricted* bandit problem for shortest path. We consider this restricted problem in Section 7.

Formally, the on-line shortest path problem in the multi-armed bandit setting is described as follows: at each time instance $t = 1,\ldots,n$, the decision maker picks a path $I_t \in \mathcal{P}$ from u to v. Then the environment assigns loss $\ell_{e,t} \in [0,1]$ to each edge $e \in E$, and the decision maker suffers loss $\ell_{I_t,t} = \sum_{e\in I_t} \ell_{e,t}$. In the unrestricted case the losses $\ell_{e,t}$ are revealed for all $e \in I_t$, while in the restricted case only $\ell_{I_t,t}$ is revealed. Note that in both cases $\ell_{e,t}$ may depend on $I_1,\ldots,I_{t-1}$, the earlier choices of the decision maker.

For the basic multi-armed bandit problem, Auer et al. (2002) gave an algorithm, based on exponential weighting with a biased estimate of the gains combined with uniform exploration. Applying their algorithm to the on-line shortest path problem in the bandit setting results in a performance that can be bounded, for any $0 < \delta < 1$ and fixed time horizon n, with probability at least $1-\delta$, by

$$\frac{1}{n}\left(\widehat{L}_n - \min_{i\in\mathcal{P}} L_{i,n}\right) \le \frac{11K}{2}\sqrt{\frac{N\ln(N/\delta)}{n}} + \frac{K\ln N}{2n} \ .$$

(The constants follow from a slightly improved version; see Cesa-Bianchi and Lugosi (2006).)

However, for the shortest path problem this bound is unacceptably large because, unlike in the full information case, here the dependence on the number of all paths N is not merely logarithmic, while N is typically exponentially large in the size of the graph (as in the two simple examples of Figure 1). Note that this bound also holds for the restricted setting as only the total losses on the paths are used. In order to achieve a bound that does not grow exponentially with the number of edges of the graph, it is imperative to make use of the dependence structure of the losses of the different actions (i.e., paths). Awerbuch and Kleinberg (2004) and McMahan and Blum (2004) do this by extending low complexity predictors, such as the follow-the-perturbed-leader forecaster (Hannan, 1957; Kalai and Vempala, 2003) to the restricted bandit setting. However, in both cases the price to pay for the polynomial dependence on the number of edges is a worse dependence on the length n of the game.

4. A Bandit Algorithm for Shortest Paths

In this section we describe a variant of the bandit algorithm of Auer et al. (2002) which achieves the desired performance for the shortest path problem. The new algorithm uses the fact that when the losses of the edges of the chosen path are revealed, then this also provides some information about the losses of each path sharing common edges with the chosen path.

For each edge $e \in E$, and $t = 1,2,\ldots$, introduce the *gain* $g_{e,t} = 1 - \ell_{e,t}$, and for each path $i \in \mathcal{P}$, let the gain be the sum of the gains of the edges on the path, that is,

$$g_{i,t} = \sum_{e\in i} g_{e,t} \ .$$

The conversion from losses to gains is done in order to facilitate the subsequent performance analysis. This has technical reasons. For the ordinary bandit problem the regret bounds of the order of $O(\sqrt{n^{-1}N\log N})$ were proved based on gains by Auer et al. (2002) and it was only recently shown

by Allenberg et al. (2006) and Auer and Ottucsák (2006) that it is possible to achieve the same type of bound for an algorithm based on losses. However, we do not know how to convert the latter algorithm into one that is efficiently computable for the shortest path problem.

To simplify the conversion, we assume that each path $i \in \mathcal{P}$ is of the same length K for some $K > 0$. Note that although this assumption may seem to be restrictive at the first glance, from each acyclic directed graph (V,E) one can construct a new graph by adding at most $(K-2)(|V|-2)+1$ vertices and edges (with constant weight zero) to the graph without modifying the weights of the paths such that each path from u to v will be of length K, where K denotes the length of the longest path of the original graph. If the number of edges is quadratic in the number of vertices, the size of the graph is not increased substantially. We describe a simple algorithm to do this in the Appendix.

A main feature of the algorithm, shown in Figure 2, is that the gains are estimated for each edge and not for each path. This modification results in an improved upper bound on the performance with the number of edges in place of the number of paths. Moreover, using dynamic programming as in Takimoto and Warmuth (2003), the algorithm can be computed efficiently. Another important ingredient of the algorithm is that one needs to make sure that every edge is sampled sufficiently often. To this end, we introduce a set $\mathcal{C}$ of *covering paths* with the property that for each edge $e \in E$ there is a path $i \in \mathcal{C}$ such that $e \in i$. Observe that one can always find such a covering set of cardinality $|\mathcal{C}| \leq |E|$.

We note that the algorithm of Auer et al. (2002) is a special case of the algorithm below: For any multi-armed bandit problem with N experts, one can define a graph with two vertices u and v, and N directed edges from u to v with weights corresponding to the losses of the experts. The solution of the shortest path problem in this case is equivalent to that of the original bandit problem with choosing expert i if the corresponding edge is chosen. For this graph, our algorithm reduces to the original algorithm of Auer et al. (2002).

Note that the algorithm can be efficiently implemented using dynamic programming, similarly to Takimoto and Warmuth [28]. See the upcoming Theorem 2 for the formal statement.

The main result of the paper is the following performance bound for the shortest-path bandit algorithm. It states that the normalized regret of the algorithm, after n rounds of play, is, roughly, of the order of $K\sqrt{|E|\ln N/n}$ where $|E|$ is the number of edges of the graph, K is the length of the paths, and N is the total number of paths.

Theorem 1 *For any $\delta \in (0,1)$ and parameters $0 \leq \gamma < 1/2$, $0 < \beta \leq 1$, and $\eta > 0$ satisfying $2\eta K|\mathcal{C}| \leq \gamma$, the performance of the algorithm defined above can be bounded, with probability at least $1-\delta$, as*

$$\frac{1}{n}\left(\widehat{L}_n - \min_{i\in\mathcal{P}} L_{i,n}\right) \leq K\gamma + 2\eta K^2|\mathcal{C}| + \frac{K}{n\beta}\ln\frac{|E|}{\delta} + \frac{\ln N}{n\eta} + |E|\beta.$$

In particular, choosing $\beta = \sqrt{\frac{K}{n|E|}\ln\frac{|E|}{\delta}}$, $\gamma = 2\eta K|\mathcal{C}|$, and $\eta = \sqrt{\frac{\ln N}{4nK^2|\mathcal{C}|}}$ yields for all $n \geq \max\left\{\frac{K}{|E|}\ln\frac{|E|}{\delta}, 4|\mathcal{C}|\ln N\right\}$,

$$\frac{1}{n}\left(\widehat{L}_n - \min_{i\in\mathcal{P}} L_{i,n}\right) \leq 2\sqrt{\frac{K}{n}}\left(\sqrt{4K|\mathcal{C}|\ln N} + \sqrt{|E|\ln\frac{|E|}{\delta}}\right).$$

Parameters: real numbers $\beta > 0$, $0 < \eta, \gamma < 1$.
Initialization: Set $w_{e,0} = 1$ for each $e \in E$, $\overline{w}_{i,0} = 1$ for each $i \in \mathcal{P}$, and $\overline{W}_0 = N$. For each round $t = 1, 2, \ldots$

(a) Choose a path I_t at random according to the distribution p_t on $\mathcal{P}$, defined by

$$p_{i,t} = \begin{cases} (1-\gamma)\frac{\overline{w}_{i,t-1}}{\overline{W}_{t-1}} + \frac{\gamma}{|\mathcal{C}|} & \text{if } i \in \mathcal{C} \\ (1-\gamma)\frac{\overline{w}_{i,t-1}}{\overline{W}_{t-1}} & \text{if } i \notin \mathcal{C}. \end{cases}$$

(b) Compute the probability of choosing each edge e as

$$q_{e,t} = \sum_{i:e\in i} p_{i,t} = (1-\gamma)\frac{\sum_{i:e\in i}\overline{w}_{i,t-1}}{\overline{W}_{t-1}} + \gamma\frac{|\{i \in \mathcal{C} : e \in i\}|}{|\mathcal{C}|}.$$

(c) Calculate the estimated gains

$$g'_{e,t} = \begin{cases} \frac{g_{e,t}+\beta}{q_{e,t}} & \text{if } e \in I_t \\ \frac{\beta}{q_{e,t}} & \text{otherwise.} \end{cases}$$

(d) Compute the updated weights

$$\begin{aligned} w_{e,t} &= w_{e,t-1}e^{\eta g'_{e,t}} \\ \overline{w}_{i,t} &= \prod_{e\in i} w_{e,t} = \overline{w}_{i,t-1}e^{\eta g'_{i,t}} \end{aligned}$$

where $g'_{i,t} = \sum_{e\in i} g'_{e,t}$, and the sum of the total weights of the paths

$$\overline{W}_t = \sum_{i\in\mathcal{P}} \overline{w}_{i,t}.$$

Figure 2: A bandit algorithm for shortest path problems

The proof of the theorem is based on the analysis of the original algorithm of Auer et al. (2002) with necessary modifications required to transform parts of the argument from paths to edges, and to use the connection between the gains of paths sharing common edges.

For the analysis we introduce some notation:

$$G_{i,n} = \sum_{t=1}^{n} g_{i,t} \qquad \text{and} \qquad G'_{i,n} = \sum_{t=1}^{n} g'_{i,t}$$

for each $i \in \mathcal{P}$ and

$$G_{e,n} = \sum_{t=1}^{n} g_{e,t} \qquad \text{and} \qquad G'_{e,n} = \sum_{t=1}^{n} g'_{e,t}$$

for each $e \in E$, and

$$\widehat{G}_n = \sum_{t=1}^{n} g_{I_t,t}.$$

Note that $g'_{e,t}$, $g'_{i,t}$, $G'_{e,n}$, and $G'_{i,n}$ are random variables that depend on I_t.

The following lemma, shows that the deviation of the true cumulative gain from the estimated cumulative gain is of the order of $\sqrt{n}$. The proof is a modification of Cesa-Bianchi and Lugosi (2006, Lemma 6.7).

Lemma 2 *For any $\delta \in (0,1)$, $0 \leq \beta < 1$ and $e \in E$ we have*

$$\mathbb{P}\left[G_{e,n} > G'_{e,n} + \frac{1}{\beta}\ln\frac{|E|}{\delta}\right] \leq \frac{\delta}{|E|} .$$

Proof Fix $e \in E$. For any $u > 0$ and $c > 0$, by the Chernoff bound we have

$$\mathbb{P}[G_{e,n} > G'_{e,n} + u] \leq e^{-cu}\mathbb{E}e^{c(G_{e,n}-G'_{e,n})} . \tag{1}$$

Letting $u = \ln(|E|/\delta)/\beta$ and $c = \beta$, we get

$$e^{-cu}\mathbb{E}e^{c(G_{e,n}-G'_{e,n})} = e^{-\ln(|E|/\delta)}\mathbb{E}e^{\beta(G_{e,n}-G'_{e,n})} = \frac{\delta}{|E|}\mathbb{E}e^{\beta(G_{e,n}-G'_{e,n})} ,$$

so it suffices to prove that $\mathbb{E}e^{\beta(G_{e,n}-G'_{e,n})} \leq 1$ for all n. To this end, introduce

$$Z_t = e^{\beta(G_{e,t}-G'_{e,t})} .$$

Below we show that $\mathbb{E}_t[Z_t] \leq Z_{t-1}$ for $t \geq 2$ where $\mathbb{E}_t$ denotes the conditional expectation $\mathbb{E}[\cdot|I_1,\ldots,I_{t-1}]$. Clearly,

$$Z_t = Z_{t-1}\exp\left(\beta\left(g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}}g_{e,t}+\beta}{q_{e,t}}\right)\right) .$$

Taking conditional expectations, we obtain

$$\begin{aligned}
&\mathbb{E}_t[Z_t] \\
&= Z_{t-1}\mathbb{E}_t\left[\exp\left(\beta\left(g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}}g_{e,t}+\beta}{q_{e,t}}\right)\right)\right] \\
&= Z_{t-1}e^{-\frac{\beta^2}{q_{e,t}}}\mathbb{E}_t\left[\exp\left(\beta\left(g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}}g_{e,t}}{q_{e,t}}\right)\right)\right] \\
&\leq Z_{t-1}e^{-\frac{\beta^2}{q_{e,t}}}\mathbb{E}_t\left[1+\beta\left(g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}}g_{e,t}}{q_{e,t}}\right)+\beta^2\left(g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}}g_{e,t}}{q_{e,t}}\right)^2\right] &(2)\\
&= Z_{t-1}e^{-\frac{\beta^2}{q_{e,t}}}\mathbb{E}_t\left[1+\beta^2\left(g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}}g_{e,t}}{q_{e,t}}\right)^2\right] &(3)\\
&\leq Z_{t-1}e^{-\frac{\beta^2}{q_{e,t}}}\mathbb{E}_t\left[1+\beta^2\left(\frac{\mathbb{1}_{\{e\in I_t\}}g_{e,t}}{q_{e,t}}\right)^2\right] \\
&\leq Z_{t-1}e^{-\frac{\beta^2}{q_{e,t}}}\left(1+\frac{\beta^2}{q_{e,t}}\right) \\
&\leq Z_{t-1}. &(4)
\end{aligned}$$

Here (2) holds since $\beta \leq 1$, $g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}}g_{e,t}}{q_{e,t}} \leq 1$ and $e^x \leq 1+x+x^2$ for $x \leq 1$. (3) follows from $\mathbb{E}_t\left[\frac{\mathbb{1}_{\{e\in I_t\}}g_{e,t}}{q_{e,t}}\right] = g_{e,t}$. Finally, (4) holds by the inequality $1+x \leq e^x$. Taking expectations on both

sides proves $\mathbb{E}[Z_t] \le \mathbb{E}[Z_{t-1}]$. A similar argument shows that $\mathbb{E}[Z_1] \le 1$, implying $\mathbb{E}[Z_n] \le 1$ as desired. □

Proof of Theorem 1. As usual in the analysis of exponentially weighted average forecasters, we start with bounding the quantity $\ln \frac{\overline{W}_n}{\overline{W}_0}$. On the one hand, we have the lower bound

$$\ln \frac{\overline{W}_n}{\overline{W}_0} = \ln \sum_{i \in \mathcal{P}} e^{\eta G'_{i,n}} - \ln N \ge \eta \max_{i \in \mathcal{P}} G'_{i,n} - \ln N \,. \tag{5}$$

To derive a suitable upper bound, first notice that the condition $\eta \le \frac{\gamma}{2K|C|}$ implies $\eta g'_{i,t} \le 1$ for all i and t, since

$$\eta g'_{i,t} = \eta \sum_{e \in i} g'_{e,t} \le \eta \sum_{e \in i} \frac{1+\beta}{q_{e,t}} \le \frac{\eta K(1+\beta)|C|}{\gamma} \le 1$$

where the second inequality follows because $q_{e,t} \ge \gamma/|C|$ for each $e \in E$.
Therefore, using the fact that $e^x \le 1 + x + x^2$ for all $x \le 1$, for all $t = 1, 2, \ldots$ we have

$$\begin{aligned}
\ln \frac{\overline{W}_t}{\overline{W}_{t-1}} &= \ln \sum_{i \in \mathcal{P}} \frac{\overline{w}_{i,t-1}}{\overline{W}_{t-1}} e^{\eta g'_{i,t}} \\
&= \ln \left(\sum_{i \in \mathcal{P}} \frac{p_{i,t} - \frac{\gamma}{|C|} \mathbb{1}_{\{i \in C\}}}{1-\gamma} e^{\eta g'_{i,t}} \right) \qquad (6) \\
&\le \ln \left(\sum_{i \in \mathcal{P}} \frac{p_{i,t} - \frac{\gamma}{|C|} \mathbb{1}_{\{i \in C\}}}{1-\gamma} \left(1 + \eta g'_{i,t} + \eta^2 g'^2_{i,t} \right) \right) \\
&\le \ln \left(1 + \sum_{i \in \mathcal{P}} \frac{p_{i,t}}{1-\gamma} \left(\eta g'_{i,t} + \eta^2 g'^2_{i,t} \right) \right) \\
&\le \frac{\eta}{1-\gamma} \sum_{i \in \mathcal{P}} p_{i,t} g'_{i,t} + \frac{\eta^2}{1-\gamma} \sum_{i \in \mathcal{P}} p_{i,t} g'^2_{i,t} \qquad (7)
\end{aligned}$$

where (6) follows form the definition of $p_{i,t}$, and (7) holds by the inequality $\ln(1+x) \le x$ for all $x > -1$.

Next we bound the sums in (7). On the one hand,

$$\begin{aligned}
\sum_{i \in \mathcal{P}} p_{i,t} g'_{i,t} &= \sum_{i \in \mathcal{P}} p_{i,t} \sum_{e \in i} g'_{e,t} = \sum_{e \in E} g'_{e,t} \sum_{i \in \mathcal{P}: e \in i} p_{i,t} \\
&= \sum_{e \in E} g'_{e,t} q_{e,t} = g_{I_t,t} + |E|\beta.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\sum_{i\in\mathcal{P}} p_{i,t} g'^2_{i,t} &= \sum_{i\in\mathcal{P}} p_{i,t} \left(\sum_{e\in i} g'_{e,t}\right)^2 \\
&\leq \sum_{i\in\mathcal{P}} p_{i,t} K \sum_{e\in i} g'^2_{e,t} \\
&= K \sum_{e\in E} g'^2_{e,t} \sum_{i\in\mathcal{P}: e\in i} p_{i,t} \\
&= K \sum_{e\in E} g'^2_{e,t} q_{e,t} \\
&= K \sum_{e\in E} q_{e,t} g'_{e,t} \frac{\beta + \mathbb{1}_{\{e\in I_t\}} g_{e,t}}{q_{e,t}} \\
&\leq K(1+\beta) \sum_{e\in E} g'_{e,t}
\end{aligned}$$

where the first inequality is due to the inequality between the arithmetic and quadratic mean, and the second one holds because $g_{e,t} \leq 1$. Therefore,

$$\ln \frac{\overline{W}_t}{\overline{W}_{t-1}} \leq \frac{\eta}{1-\gamma} (g_{I_t,t} + |E|\beta) + \frac{\eta^2 K(1+\beta)}{1-\gamma} \sum_{e\in E} g'_{e,t} \,.$$

Summing for $t = 1, \ldots, n$, we obtain

$$\begin{aligned}
\ln \frac{\overline{W}_n}{\overline{W}_0} &\leq \frac{\eta}{1-\gamma} \left(\widehat{G}_n + n|E|\beta\right) + \frac{\eta^2 K(1+\beta)}{1-\gamma} \sum_{e\in E} G'_{e,n} \\
&\leq \frac{\eta}{1-\gamma} \left(\widehat{G}_n + n|E|\beta\right) + \frac{\eta^2 K(1+\beta)}{1-\gamma} |\mathcal{C}| \max_{i\in\mathcal{P}} G'_{i,n}
\end{aligned}$$

where the second inequality follows since $\sum_{e\in E} G'_{e,n} \leq \sum_{i\in\mathcal{C}} G'_{i,n}$. Combining the upper bound with the lower bound (5), we obtain

$$\widehat{G}_n \geq (1-\gamma-\eta K(1+\beta)|\mathcal{C}|) \max_{i\in\mathcal{P}} G'_{i,n} - \frac{1-\gamma}{\eta} \ln N - n|E|\beta.$$

Now using Lemma 2 and applying the union bound, for any $\delta \in (0,1)$ we have that, with probability at least $1-\delta$,

$$\widehat{G}_n \geq (1-\gamma-\eta K(1+\beta)|\mathcal{C}|) \left(\max_{i\in\mathcal{P}} G_{i,n} - \frac{K}{\beta} \ln \frac{|E|}{\delta}\right) - \frac{1-\gamma}{\eta} \ln N - n|E|\beta \,,$$

where we used $1-\gamma-\eta K(1+\beta)|\mathcal{C}| \geq 0$ which follows from the assumptions of the theorem. Since $\widehat{G}_n = Kn - \widehat{L}_n$ and $G_{i,n} = Kn - L_{i,n}$ for all $i \in \mathcal{P}$, we have

$$\begin{aligned}
\widehat{L}_n &\leq Kn(\gamma + \eta(1+\beta)K|\mathcal{C}|) + (1-\gamma-\eta(1+\beta)K|\mathcal{C}|) \min_{i\in\mathcal{P}} L_{i,n} \\
&\quad + (1-\gamma-\eta(1+\beta)K|\mathcal{C}|) \frac{K}{\beta} \ln \frac{|E|}{\delta} + \frac{1-\gamma}{\eta} \ln N + n|E|\beta
\end{aligned}$$

with probability at least $1-\delta$. This implies

$$\begin{aligned}\widehat{L}_n - \min_{i\in\mathcal{P}} L_{i,n} &\le K n\gamma + \eta(1+\beta)nK^2|\mathcal{C}| + \frac{K}{\beta}\ln\frac{|E|}{\delta} + \frac{1-\gamma}{\eta}\ln N + n|E|\beta \\ &\le K n\gamma + 2\eta nK^2|\mathcal{C}| + \frac{K}{\beta}\ln\frac{|E|}{\delta} + \frac{\ln N}{\eta} + n|E|\beta\end{aligned}$$

with probability at least $1-\delta$, which is the first statement of the theorem. Setting

$$\beta = \sqrt{\frac{K}{n|E|}\ln\frac{|E|}{\delta}} \qquad \text{and} \qquad \gamma = 2\eta K|\mathcal{C}|$$

results in the inequality

$$\widehat{L}_n - \min_{i\in\mathcal{P}} L_{i,n} \le 4\eta nK^2|\mathcal{C}| + \frac{\ln N}{\eta} + 2\sqrt{nK|E|\ln\frac{|E|}{\delta}}$$

which holds with probability at least $1-\delta$ if $n \ge (K/|E|)\ln(|E|/\delta)$ (to ensure $\beta \le 1$). Finally, setting

$$\eta = \sqrt{\frac{\ln N}{4nK^2|\mathcal{C}|}}$$

yields the last statement of the theorem ($n \ge 4\ln N|\mathcal{C}|$ is required to ensure $\gamma \le 1/2$). □

Next we analyze the computational complexity of the algorithm. The next result shows that the algorithm is feasible as its complexity is linear in the size (number of edges) of the graph.

Theorem 3 *The proposed algorithm can be implemented efficiently with time complexity $O(n|E|)$ and space complexity $O(|E|)$.*

Proof The two complex steps of the algorithm are steps (a) and (b), both of which can be computed, similarly to Takimoto and Warmuth (2003), using dynamic programming. To perform these steps efficiently, first we order the vertices of the graph. Since we have an acyclic directed graph, its vertices can be labeled (in $O(|E|)$ time) from 1 to $|V|$ such that $u = 1$, $v = |V|$, and if $(v_1, v_2) \in E$, then $v_1 < v_2$. For any pair of vertices $u_1 < v_1$ let $\mathcal{P}_{u_1,v_1}$ denote the set of paths from u_1 to v_1, and for any vertex $s \in V$, let

$$H_t(s) = \sum_{i\in\mathcal{P}_{s,v}} \prod_{e\in i} w_{e,t}$$

and

$$\widehat{H}_t(s) = \sum_{i\in\mathcal{P}_{u,s}} \prod_{e\in i} w_{e,t}\ .$$

Given the edge weights $\{w_{e,t}\}$, $H_t(s)$ can be computed recursively for $s = |V|-1,\dots,1$, and $\widehat{H}_t(s)$ can be computed recursively for $s = 2,\dots,|V|$ in $O(|E|)$ time (letting $H_t(v) = \widehat{H}_t(u) = 1$ by definition). In step (a), first one has to decide with probability γ whether I_t is generated according to the graph weights, or it is chosen uniformly from $\mathcal{C}$. If I_t is to be drawn according to the graph weights, it can be shown that its vertices can be chosen one by one such that if the first k vertices

of I_t are $v_0 = u, v_1, \ldots, v_{k-1}$, then the next vertex of I_t can be chosen to be any $v_k > v_{k-1}$, satisfying $(v_{k-1}, v_k) \in E$, with probability $w_{(v_{k-1},v_k),t-1} H_{t-1}(v_k)/H_{t-1}(v_{k-1})$. The other computationally demanding step, namely step (b), can be performed easily by noting that for any edge (v_1, v_2),

$$\begin{aligned} q_{(v_1,v_2),t} &= (1-\gamma)\frac{\widehat{H}_{t-1}(v_1) w_{(v_1,v_2),t-1} H_{t-1}(v_2)}{H_{t-1}(u)} \\ &\quad + \gamma \frac{|\{i \in \mathcal{C} : (v_1,v_2) \in i\}|}{|\mathcal{C}|} \end{aligned}$$

as desired. □

5. A Combination of the Label Efficient and Bandit Settings

In this section we investigate a combination of the multi-armed bandit and the label efficient problems. This means that the decision maker only has access to the losses of all the edges on the chosen path upon request and the total number of requests must be bounded by a constant m. This combination is motivated by some applications, in which feedback information is costly to obtain.

In the general label efficient decision problem, after taking an action, the decision maker has the option to query the losses of all possible actions. For this problem, Cesa-Bianchi et al. (2005) proved an upper bound on the normalized regret of order $O(K\sqrt{\ln(4N/\delta)/(m)})$ which holds with probability at least $1-\delta$, where K is the length of the longest path in the graph.

Our model of the label-efficient bandit problem for shortest paths is motivated by an application to a particular packet switched network model. This model, called the cognitive packet network, was introduced by Gelenbe et al. (2004, 2001). In these networks a particular type of packets, called smart packets, are used to explore the network (e.g., the delay of the chosen path). These packets do not carry any useful data; they are merely used for exploring the network. The other type of packets are the data packets, which do not collect any information about their paths. The task of the decision maker is to send packets from the source to the destination over routes with minimum average transmission delay (or packet loss). In this scenario, smart packets are used to query the delay (or loss) of the chosen path. However, as these packets do not transport information, there is a tradeoff between the number of queries and the usage of the network. If data packets are on the average α times larger than smart packets (note that typically $\alpha \gg 1$) and ε is the proportion of time instances when smart packets are used to explore the network, then $\varepsilon/(\varepsilon + \alpha(1-\varepsilon))$ is the proportion of the bandwidth sacrificed for well informed routing decisions.

We study a combined algorithm which, at each time slot t, queries the loss of the chosen path with probability ε (as in the solution of the label efficient problem proposed in Cesa-Bianchi et al., 2005), and, similarly to the multi-armed bandit case, computes biased estimates $g'_{i,t}$ of the true gains $g_{i,t}$. Just as in the previous section, it is assumed that each path of the graph is of the same length K.

The algorithm differs from our bandit algorithm of the previous section only in step (c), which is modified in the spirit of Cesa-Bianchi et al. (2005). The modified step is given in Figure 3.

The performance of the algorithm is analyzed in the next theorem, which can be viewed as a combination of Theorem 1 in the preceding section and Theorem 2 of Cesa-Bianchi et al. (2005).

(c') Draw a Bernoulli random variable S_t with $\mathbb{P}(S_t = 1) = \varepsilon$, and compute the estimated gains

$$g'_{e,t} = \begin{cases} \frac{g_{e,t}+\beta}{\varepsilon q_{e,t}} S_t & \text{if } e \in I_t \\ \frac{\beta}{\varepsilon q_{e,t}} S_t & \text{if } e \notin I_t \end{cases} .$$

Figure 3: Modified step for the label efficient bandit algorithm for shortest paths

Theorem 4 *For any* $\delta \in (0,1)$, $\varepsilon \in (0,1]$, *parameters* $\eta = \sqrt{\frac{\varepsilon \ln N}{4nK^2|C|}}$, $\gamma = \frac{2\eta K|C|}{\varepsilon} \le 1/2$, *and* $\beta = \sqrt{\frac{K}{n|E|\varepsilon} \ln \frac{2|E|}{\delta}} \le 1$, *and for all*

$$n \ge \frac{1}{\varepsilon} \max \left\{ \frac{K^2 \ln^2(2|E|/\delta)}{|E| \ln N}, \frac{|E| \ln(2|E|/\delta)}{K}, 4|C| \ln N \right\}$$

the performance of the algorithm defined above can be bounded, with probability at least $1-\delta$, *as*

$$\begin{aligned} &\frac{1}{n}\left(\widehat{L}_n - \min_{i \in \mathcal{P}} L_{i,t}\right) \\ &\quad \le \sqrt{\frac{K}{n\varepsilon}} \left(4\sqrt{K|C| \ln N} + 5\sqrt{|E| \ln \frac{2|E|}{\delta}} + \sqrt{8K \ln \frac{2}{\delta}} \right) + \frac{4K}{3n\varepsilon} \ln \frac{2N}{\delta} \\ &\quad \le \frac{27K}{2} \sqrt{\frac{|E| \ln \frac{2N}{\delta}}{n\varepsilon}} . \end{aligned}$$

If ε is chosen as $(m - \sqrt{2m \ln(1/\delta)})/n$ then, with probability at least $1-\delta$, the total number of queries is bounded by m (Cesa-Bianchi and Lugosi, 2006, Lemma 6.1) and the performance bound above is of the order of $K\sqrt{|E| \ln(N/\delta)/m}$.

Similarly to Theorem 1, we need a lemma which reveals the connection between the true and the estimated cumulative losses. However, here we need a more careful analysis because the "shifting term" $\frac{\beta}{\varepsilon q_{e,t}} S_t$, is a random variable.

Lemma 5 *For any* $0 < \delta < 1$, $0 < \varepsilon \le 1$, *for any*

$$n \ge \frac{1}{\varepsilon} \max \left\{ \frac{K^2 \ln^2(2|E|/\delta)}{|E| \ln N}, \frac{K \ln(2|E|/\delta)}{|E|} \right\} ,$$

parameters

$$\frac{2\eta K|C|}{\varepsilon} \le \gamma, \qquad \eta = \sqrt{\frac{\varepsilon \ln N}{4nK^2|C|}} \quad \textit{and} \quad \beta = \sqrt{\frac{K}{n|E|\varepsilon} \ln \frac{2|E|}{\delta}} \le 1 ,$$

and $e \in E$, *we have*

$$\mathbb{P}\left[G_{e,n} > G'_{e,n} + \frac{4}{\beta\varepsilon} \ln \frac{2|E|}{\delta} \right] \le \frac{\delta}{2|E|} .$$

Proof Fix $e \in E$. Using (1) with $u = \frac{4}{\beta\varepsilon} \ln \frac{2|E|}{\delta}$ and $c = \frac{\beta\varepsilon}{4}$, it suffices to prove for all n that

$$\mathbb{E}\left[e^{c(G_{e,n} - G'_{e,n})}\right] \leq 1 \,.$$

Similarly to Lemma 2 we introduce $Z_t = e^{c(G_{e,t} - G'_{e,t})}$ and we show that $Z_1, \ldots, Z_n$ is a supermartingale, that is $\mathbb{E}_t[Z_t] \leq Z_{t-1}$ for $t \geq 2$ where $\mathbb{E}_t$ denotes $\mathbb{E}[\cdot|(I_1, S_1), \ldots, (I_{t-1}, S_{t-1})]$. Taking conditional expectations, we obtain

$$\begin{aligned}
\mathbb{E}_t[Z_t] &= Z_{t-1}\mathbb{E}_t\left[e^{c\left(g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}} S_t g_{e,t} + S_t\beta}{q_{e,t}\varepsilon}\right)}\right] \\
&\leq Z_{t-1}\mathbb{E}_t\left[1 + c\left(g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}} S_t g_{e,t} + S_t\beta}{q_{e,t}\varepsilon}\right)\right. \\
&\qquad \left. + c^2\left(g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}} S_t g_{e,t} + S_t\beta}{q_{e,t}\varepsilon}\right)^2\right]. \qquad (8)
\end{aligned}$$

Since

$$\mathbb{E}_t\left[g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}} S_t g_{e,t} + S_t\beta}{q_{e,t}\varepsilon}\right] = -\frac{\beta}{q_{e,t}}$$

and

$$\mathbb{E}_t\left[\left(g_{e,t} - \frac{\mathbb{1}_{\{e\in I_t\}} S_t g_{e,t}}{q_{e,t}\varepsilon}\right)^2\right] \leq \mathbb{E}_t\left[\left(\frac{\mathbb{1}_{\{e\in I_t\}} S_t g_{e,t}}{q_{e,t}\varepsilon}\right)^2\right] \leq \frac{1}{q_{e,t}\varepsilon}$$

we get from (8) that

$$\begin{aligned}
&\mathbb{E}_t[Z_t] \\
&\leq Z_{t-1}\mathbb{E}_t\left[1 - \frac{c\beta}{q_{e,t}} + \frac{c^2}{q_{e,t}\varepsilon} + c^2\left(\frac{2\mathbb{1}_{\{e\in I_t\}} S_t g_{e,t}\beta}{q_{e,t}^2\varepsilon^2} - \frac{2g_{e,t}S_t\beta}{q_{e,t}\varepsilon} + \frac{S_t\beta^2}{q_{e,t}^2\varepsilon^2}\right)\right] \\
&\leq Z_{t-1}\left(1 + \frac{c}{q_{e,t}}\left(-\beta + \frac{c}{\varepsilon} + c\beta\left(\frac{2}{\varepsilon} + \frac{\beta}{q_{e,t}\varepsilon}\right)\right)\right). \qquad (9)
\end{aligned}$$

Since $c = \beta\varepsilon/4$ we have

$$\begin{aligned}
-\beta + \frac{c}{\varepsilon} + c\beta\left(\frac{2}{\varepsilon} + \frac{\beta}{q_{e,t}\varepsilon}\right) &= -\frac{3\beta}{4} + \frac{\beta^2\varepsilon}{4}\left(\frac{2}{\varepsilon} + \frac{\beta}{q_{e,t}\varepsilon}\right) \\
&= -\frac{3\beta}{4} + \frac{\beta^2}{2} + \frac{\beta^3}{4q_{e,t}} \\
&\leq -\frac{\beta}{4} + \frac{\beta^3}{4q_{e,t}} \\
&\leq -\frac{\beta}{4} + \frac{\beta^3|C|}{4\gamma} \qquad (10) \\
&\leq 0, \qquad (11)
\end{aligned}$$

where (10) follows from $q_{e,t} \geq \frac{\gamma}{|\mathcal{C}|}$ and (11) holds since $\beta \leq 1$ and by

$$\frac{\beta^2|\mathcal{C}|}{\gamma} \leq \frac{\beta^2\varepsilon}{2\eta K} \leq 1 \,,$$

and the last inequality is ensured by $n \geq \frac{K^2 \ln^2(2|E|/\delta)}{\varepsilon|E|\ln N}$, the assumption of the lemma.

Combining (9) and (11) we get that $\mathbb{E}_t[Z_t] \leq Z_{t-1}$. Taking expectations on both sides of the inequality, we get $\mathbb{E}[Z_t] \leq \mathbb{E}[Z_{t-1}]$ and since $\mathbb{E}[Z_1] \leq 1$, we obtain $\mathbb{E}[Z_n] \leq 1$ as desired. □

Proof of Theorem 4. The proof of the theorem is a generalization of that of Theorem 1, and follows the same lines with some extra technicalities to handle the effects of the modified step (c'). Therefore, in the following we emphasize only the differences. First note that (5) and (7) also hold in this case. Bounding the sums in (7), one obtains

$$\sum_{i\in\mathcal{P}} p_{i,t} g'_{i,t} = \frac{S_t}{\varepsilon}\left(g_{I_t,t} + |E|\beta\right)$$

and

$$\sum_{i\in\mathcal{P}} p_{i,t} g'^2_{i,t} \leq \frac{1}{\varepsilon} K(1+\beta) \sum_{e\in E} g'_{e,t} \,.$$

Plugging these bounds into (7) and summing for $t = 1,\ldots,n$, we obtain

$$\ln\frac{\overline{W}_n}{\overline{W}_0} \quad \leq \quad \frac{\eta}{1-\gamma}\sum_{t=1}^{n}\frac{S_t}{\varepsilon}\left(g_{I_t,t} + |E|\beta\right) + \frac{\eta^2 K(1+\beta)}{(1-\gamma)\varepsilon}|\mathcal{C}|\max_{i\in\mathcal{P}} G'_{i,n} \,.$$

Combining the upper bound with the lower bound (5), we obtain

$$\sum_{t=1}^{n}\frac{S_t}{\varepsilon}\left(g_{I_t,t} + |E|\beta\right) \qu\geq \quad \left(1-\gamma-\frac{\eta K(1+\beta)|\mathcal{C}|}{\varepsilon}\right)\max_{i\in\mathcal{P}} G'_{i,n} - \frac{\ln N}{\eta} \,. \tag{12}$$

To relate the left-hand side of the above inequality to the real gain $\sum_{t=1}^{n} g_{I_t,t}$, notice that

$$X_t = \frac{S_t}{\varepsilon}\left(g_{I_t,t} + |E|\beta\right) - \left(g_{I_t,t} + |E|\beta\right)$$

is a martingale difference sequence with respect to $(I_1,S_1),(I_2,S_2),\ldots$. Now for all $t = 1,\ldots,n$, we have the bound

$$\begin{aligned}
\mathbb{E}\left[X_t^2|(I_1,S_1),\ldots,(I_{t-1},S_{t-1})\right] \quad &\leq \quad \mathbb{E}\left[\frac{S_t}{\varepsilon^2}(g_{I_t,t}+|E|\beta)^2\middle|(I_1,S_1),\ldots,(I_{t-1},S_{t-1})\right] \\
&\leq \quad \frac{(K+|E|\beta)^2}{\varepsilon} \\
&\leq \quad \frac{4K^2}{\varepsilon} \stackrel{\text{def}}{=} \sigma^2,
\end{aligned} \tag{13}$$

where (13) holds by $n \geq \frac{|E|\ln(2|E|/\delta)}{K\varepsilon}$ (to ensure $\beta|E| \leq K$). We know that

$$X_t \in \left[-2K, \left(\frac{1}{\varepsilon}-1\right)2K\right]$$

for all t. Now apply Bernstein's inequality for martingale differences (see Lemma 14 in the Appendix) to obtain

$$\mathbb{P}\left[\sum_{t=1}^{n} X_t > u\right] \leq \frac{\delta}{2}\,, \tag{14}$$

where

$$u = \sqrt{2n\frac{4K^2}{\varepsilon}\ln\left(\frac{2}{\delta}\right)} + \frac{4K}{3\varepsilon}\ln\left(\frac{2}{\delta}\right).$$

From (14) we get

$$\mathbb{P}\left[\sum_{t=1}^{n} \frac{S_t}{\varepsilon}\left(g_{I_t,t} + |E|\beta\right) \geq \widehat{G}_n + \beta n|E| + u\right] \leq \frac{\delta}{2}\,. \tag{15}$$

Now Lemma 5, the union bound, and (15) combined with (12) yield, with probability at least $1-\delta$,

$$\begin{aligned}\widehat{G}_n \;\geq\; & \left(1-\gamma-\frac{\eta K(1+\beta)|\mathcal{C}|}{\varepsilon}\right)\left(\max_{i\in\mathcal{P}} G_{i,n} - \frac{4K}{\beta\varepsilon}\ln\frac{2|E|}{\delta}\right)\\ & -\frac{\ln N}{\eta} - \beta n|E| - u\end{aligned}$$

since the coefficient of $G'_{i,n}$ is greater than zero by the assumptions of the theorem.

Since $\widehat{G}_n = Kn - \widehat{L}_n$ and $G_{i,n} = Kn - L_{i,n}$, we have

$$\begin{aligned}\widehat{L}_n \;\leq\; & \left(1-\gamma-\frac{K(1+\beta)\eta|\mathcal{C}|}{\varepsilon}\right)\min_{i\in\mathcal{P}} L_{i,n} + Kn\left(\gamma+\frac{K(1+\beta)\eta|\mathcal{C}|}{\varepsilon}\right)\\ & +\left(1-\gamma-\frac{K(1+\beta)\eta|\mathcal{C}|}{\varepsilon}\right)\frac{4K}{\beta\varepsilon}\ln\frac{2|E|}{\delta} + \beta n|E| + \frac{\ln N}{\eta} + u\\ \leq\; & \min_{i\in\mathcal{P}} L_{i,n} + Kn\left(\gamma+\frac{K(1+\beta)\eta|\mathcal{C}|}{\varepsilon}\right) + 5\beta n|E| + \frac{\ln N}{\eta} + u\,,\end{aligned}$$

where we used the fact that $\frac{K}{\beta\varepsilon}\ln\frac{2|E|}{\delta} = \beta n|E|$.

Substituting the value of β, η and γ, we have

$$\begin{aligned}\widehat{L}_n - \min_{i\in\mathcal{P}} L_{i,n} \leq & Kn\frac{2K\eta|\mathcal{C}|}{\varepsilon} + Kn\frac{2K\eta|\mathcal{C}|}{\varepsilon} + \frac{\ln N}{\eta} + 5\beta n|E| + u\\ \leq & 4K\sqrt{\frac{n|\mathcal{C}|\ln N}{\varepsilon}} + 5\sqrt{\frac{n|E|K\ln(2|E|/\delta)}{\varepsilon}} + u\\ \leq & \sqrt{\frac{nK}{\varepsilon}}\left(4\sqrt{K|\mathcal{C}|\ln N} + 5\sqrt{|E|\ln(2|E|/\delta)} + \sqrt{8K\ln(2/\delta)}\right)\\ & + \frac{4K}{3\varepsilon}\ln(2/\delta)\end{aligned}$$

as desired. □

6. A Bandit Algorithm for Tracking the Shortest Path

Our goal in this section is to extend the bandit algorithm so that it is able to compete with time-varying paths under small computational complexity. This is a variant of the problem known as *tracking the best expert*; see, for example, Herbster and Warmuth (1998), Vovk (1999), Auer and Warmuth (1998), Bousquet and Warmuth (2002) and Herbster and Warmuth (2001).

To describe the loss the decision maker is compared to, consider the following "m-partition" prediction scheme: the sequence of paths is partitioned into $m+1$ contiguous segments, and on each segment the scheme assigns exactly one of the N paths. Formally, an m-partition $\texttt{Part}(n,m,\mathbf{t},\underline{i})$ of the n paths is given by an m-tuple $\mathbf{t}=(t_1,\ldots,t_m)$ such that $t_0=1<t_1<\cdots<t_m<n+1=t_{m+1}$, and an $(m+1)$-vector $\underline{i}=(i_0,\ldots,i_m)$ where $i_j\in\mathcal{P}$. At each time instant t, $t_j\le t<t_{j+1}$, path i_j is used to predict the best path. The cumulative loss of a partition $\texttt{Part}(n,m,\mathbf{t},\underline{i})$ is

$$L(\texttt{Part}(n,m,\mathbf{t},\underline{i}))=\sum_{j=0}^{m}\sum_{t=t_j}^{t_{j+1}-1}\ell_{i_j,t}=\sum_{j=0}^{m}\sum_{t=t_j}^{t_{j+1}-1}\sum_{e\in i_j}\ell_{e,t}.$$

The goal of the decision maker is to perform as well as the best time-varying path (partition), that is, to keep the normalized regret

$$\frac{1}{n}\left(\widehat{L}_n-\min_{\mathbf{t},\underline{i}}L(\texttt{Part}(n,m,\mathbf{t},\underline{i}))\right)$$

as small as possible (with high probability) for all possible outcome sequences.

In the "classical" tracking problem there is a relatively small number of "base" experts and the goal of the decision maker is to predict as well as the best "compound" expert (i.e., time-varying expert). However in our case, base experts correspond to all paths of the graph between source and destination whose number is typically exponentially large in the number of edges, and therefore we cannot directly apply the computationally efficient methods for tracking the best expert. György et al. (2005a) develop efficient algorithms for tracking the best expert for certain large and structured classes of base experts, including the shortest path problem. The purpose of the following algorithm, shown in Figure 4, is to extend the methods of György et al. (2005a) to the bandit setting when the forecaster only observes the losses of the edges on the chosen path.

The following performance bounds shows that the normalized regret with respect to the best time-varying path which is allowed to switch paths m times is roughly of the order of $K\sqrt{(m/n)|\mathcal{C}|\ln N}$.

Theorem 6 *For any $\delta\in(0,1)$ and parameters $0\le\gamma<1/2$, $\alpha,\beta\in[0,1]$, and $\eta>0$ satisfying $2\eta K|\mathcal{C}|\le\gamma$, the performance of the algorithm defined above can be bounded, with probability at least $1-\delta$, as*

$$\begin{aligned}\frac{1}{n}&\left(\widehat{L}_n-\min_{\mathbf{t},\underline{i}}L(\texttt{Part}(n,m,\mathbf{t},\underline{i}))\right)\\ &\le\ K(\gamma+\eta(1+\beta)K|\mathcal{C}|)+\frac{K(m+1)}{n\beta}\ln\frac{|E|(m+1)}{\delta}\\ &\quad+\beta|E|+\frac{1}{n\eta}\ln\left(\frac{N^{m+1}}{\alpha^m(1-\alpha)^{n-m-1}}\right).\end{aligned}$$

Parameters: real numbers $\beta > 0$, $0 < \eta, \gamma < 1$, $0 \le \alpha \le 1$.
Initialization: Set $w_{e,0} = 1$ for each $e \in E$, $\overline{w}_{i,0} = 1$ for each $i \in \mathcal{P}$, and $\overline{W}_0 = N$. For each round $t = 1, 2, \ldots$

(a) Choose a path I_t according to the distribution p_t defined by

$$p_{i,t} = \begin{cases} (1-\gamma)\frac{\overline{w}_{i,t-1}}{\overline{W}_{t-1}} + \frac{\gamma}{|\mathcal{C}|} & \text{if } i \in \mathcal{C}; \\ (1-\gamma)\frac{\overline{w}_{i,t-1}}{\overline{W}_{t-1}} & \text{if } i \notin \mathcal{C}. \end{cases}$$

(b) Compute the probability of choosing each edge e as

$$q_{e,t} = \sum_{i:e\in i} p_{i,t} = (1-\gamma)\frac{\sum_{i:e\in i}\overline{w}_{i,t-1}}{\overline{W}_{t-1}} + \gamma\frac{|\{i \in \mathcal{C} : e \in i\}|}{|\mathcal{C}|}.$$

(c) Calculate the estimated gains

$$g'_{e,t} = \begin{cases} \frac{g_{e,t}+\beta}{q_{e,t}} & \text{if } e \in I_t; \\ \frac{\beta}{q_{e,t}} & \text{otherwise.} \end{cases}$$

(d) Compute the updated weights

$$\begin{aligned} \overline{v}_{i,t} &= \overline{w}_{i,t-1} e^{\eta g'_{i,t}} \\ \overline{w}_{i,t} &= (1-\alpha)\overline{v}_{i,t} + \frac{\alpha}{N}\overline{W}_t \end{aligned}$$

where $g'_{i,t} = \sum_{e\in i} g'_{e,t}$ and $\overline{W}_t$ is the sum of the total weights of the paths, that is,

$$\overline{W}_t = \sum_{i\in\mathcal{P}} \overline{v}_{i,t} = \sum_{i\in\mathcal{P}} \overline{w}_{i,t}.$$

Figure 4: A bandit algorithm for tracking shortest paths

In particular, choosing

$$\beta = \sqrt{\frac{K(m+1)}{n|E|}\ln\frac{|E|(m+1)}{\delta}}, \qquad \gamma = 2\eta K|\mathcal{C}|, \qquad \alpha = \frac{m}{n-1},$$

and

$$\eta = \sqrt{\frac{(m+1)\ln N + m\ln\frac{e(n-1)}{m}}{4nK^2|\mathcal{C}|}}$$

we have, for all $n \ge \max\left\{\frac{K(m+1)}{|E|}\ln\frac{|E|(m+1)}{\delta}, 4|\mathcal{C}|D\right\}$,

$$\frac{1}{n}\left(\widehat{L}_n - \min_{\mathbf{t},\underline{i}} L(\mathtt{Part}(n,m,\mathbf{t},\underline{i}))\right) \le 2\sqrt{\frac{K}{n}}\left(\sqrt{4K|\mathcal{C}|D} + \sqrt{|E|(m+1)\ln\frac{|E|(m+1)}{\delta}}\right),$$

where

$$D = (m+1)\ln N + m\left(1 + \ln\frac{n-1}{m}\right).$$

The proof of the theorem is a combination of that of our Theorem 1 and Theorem 1 of György et al. (2005a). We will need the following three lemmas.

Lemma 7 *For any $1 \le t \le t' \le n$ and any $i \in \mathcal{P}$,*

$$\frac{\overline{v}_{i,t'}}{\overline{w}_{i,t-1}} \ge e^{\eta G'_{i,[t,t']}}(1-\alpha)^{t'-t}$$

where $G'_{i,[t,t']} = \sum_{\tau=t}^{t'} g'_{i,\tau}$.

Proof The proof is a straightforward modification of the one in Herbster and Warmuth (1998). From the definitions of $v_{i,t}$ and $w_{i,t}$ (see step (d) of the algorithm) it is clear that for any $\tau \ge 1$,

$$\overline{w}_{i,\tau} = (1-\alpha)\overline{v}_{i,\tau} + \frac{\alpha}{N}\overline{W}_\tau \ge (1-\alpha)e^{\eta g'_{i,\tau}}\overline{w}_{i,\tau-1}\ .$$

Applying this equation iteratively for $\tau = t, t+1, \ldots, t'-1$, and the definition of $\overline{v}_{i,t}$ (step (d)) for $\tau = t'$, we obtain

$$\begin{aligned}\overline{v}_{i,t'} = \overline{w}_{i,t'-1}e^{\eta g'_{i,t'}} &\ge e^{\eta g'_{i,t'}}\prod_{\tau=t}^{t'-1}\left((1-\alpha)e^{\eta g'_{i,\tau}}\right)\overline{w}_{i,t-1} \\ &= e^{\eta G'_{i,[t,t']}}(1-\alpha)^{t'-t}\overline{w}_{i,t-1}\end{aligned}$$

which implies the statement of the lemma. □

Lemma 8 *For any $t \ge 1$ and $i, j \in \mathcal{P}$, we have*

$$\frac{\overline{w}_{i,t}}{\overline{v}_{j,t}} \ge \frac{\alpha}{N}$$

Proof By the definition of $\overline{w}_{i,t}$ we have

$$\overline{w}_{i,t} = (1-\alpha)\overline{v}_{i,t} + \frac{\alpha}{N}\overline{W}_t \ge \frac{\alpha}{N}\overline{W}_t \ge \frac{\alpha}{N}\overline{v}_{j,t}\ .$$

This completes the proof of the lemma. □

The next lemma is a simple corollary of Lemma 2.

Lemma 9 *For any $\delta \in (0,1)$, $0 \le \beta \le 1$, $t \ge 1$ and $e \in E$ we have*

$$\mathbb{P}\left[G_{e,t} > G'_{e,t} + \frac{1}{\beta}\ln\frac{|E|(m+1)}{\delta}\right] \le \frac{\delta}{|E|(m+1)}\ .$$

Proof of Theorem 6. Again, we upper bound $\ln \overline{W}_n/\overline{W}_0$ the same way as in Theorem 1. Then we get

$$\ln \frac{\overline{W}_n}{\overline{W}_0} \quad \leq \quad \frac{\eta}{1-\gamma}\left(\widehat{G}_n + n|E|\beta\right) + \frac{\eta^2 K(1+\beta)}{1-\gamma}|\mathcal{C}| \max_{i\in\mathcal{P}} G'_{i,n} \,. \tag{16}$$

Let $\texttt{Part}(n,m,\mathbf{t},\underline{i})$ be an arbitrary partition. Then the lower bound is obtained as

$$\ln \frac{\overline{W}_n}{\overline{W}_0} = \ln \sum_{j\in\mathcal{P}} \frac{\overline{w}_{j,n}}{N} = \ln \sum_{j\in\mathcal{P}} \frac{\overline{v}_{j,n}}{N} \geq \ln \frac{\overline{v}_{i_m,n}}{N}$$

(recall that i_m denotes the path used in the last segment of the partition). Now $v_{i_m,n}$ can be rewritten in the form of the following telescoping product

$$\overline{v}_{i_m,n} = \overline{w}_{i_0,t_0-1} \frac{\overline{v}_{i_0,t_1-1}}{\overline{w}_{i_0,t_0-1}} \prod_{j=1}^{m} \left(\frac{\overline{w}_{i_j,t_j-1}}{\overline{v}_{i_{j-1},t_j-1}} \frac{\overline{v}_{i_j,t_{j+1}-1}}{\overline{w}_{i_j,t_j-1}} \right).$$

Therefore, applying Lemmas 7 and 8, we have

$$\begin{aligned} \overline{v}_{i_m,n} &\geq \overline{w}_{i_0,t_0-1} \left(\frac{\alpha}{N}\right)^m \prod_{j=0}^{m} \left((1-\alpha)^{t_{j+1}-1-t_j} e^{\eta G'_{i_j,[t_j,t_{j+1}-1]}} \right) \\ &= \left(\frac{\alpha}{N}\right)^m e^{\eta G'(\texttt{Part}(n,m,\mathbf{t},\underline{i}))} (1-\alpha)^{n-m-1}. \end{aligned}$$

Combining the lower bound with the upper bound (16), we have

$$\begin{aligned} &\ln \left(\frac{\alpha^m (1-\alpha)^{n-m-1}}{N^{m+1}} \right) + \max_{\mathbf{t},\underline{i}} \eta G'(\texttt{Part}(n,m,\mathbf{t},\underline{i})) \\ &\leq \tfrac{\eta}{1-\gamma}\left(\widehat{G}_n + n|E|\beta\right) + \tfrac{\eta^2 K(1+\beta)}{1-\gamma}|\mathcal{C}| \max_{i\in\mathcal{P}} G'_{i,n} \,, \end{aligned}$$

where we used the fact that $\texttt{Part}(n,m,\mathbf{t},\underline{i})$ is an arbitrary partition. After rearranging and using $\max_{i\in\mathcal{P}} G'_{i,n} \leq \max_{\mathbf{t},\underline{i}} G'(\texttt{Part}(n,m,\mathbf{t},\underline{i}))$ we get

$$\begin{aligned} \widehat{G}_n \quad \geq \quad & (1-\gamma-\eta K(1+\beta)|\mathcal{C}|) \max_{\mathbf{t},\underline{i}} G'(\texttt{Part}(n,m,\mathbf{t},\underline{i})) \\ & -n|E|\beta - \frac{1-\gamma}{\eta} \ln \left(\frac{N^{m+1}}{\alpha^m (1-\alpha)^{n-m-1}} \right). \end{aligned}$$

Now since $1-\gamma-\eta K(1+\beta)|\mathcal{C}| \geq 0$, by the assumptions of the theorem and from Lemma 9 with an application of the union bound we obtain that, with probability at least $1-\delta$,

$$\begin{aligned} \widehat{G}_n \geq\, & (1-\gamma-\eta K(1+\beta)|\mathcal{C}|) \left(\max_{\mathbf{t},\underline{i}} G(\texttt{Part}(n,m,\mathbf{t},\underline{i})) - \frac{K(m+1)}{\beta} \ln \frac{|E|(m+1)}{\delta} \right) \\ & - n|E|\beta - \frac{1-\gamma}{\eta} \ln \left(\frac{N^{m+1}}{\alpha^m (1-\alpha)^{n-m-1}} \right). \end{aligned}$$

Since $\widehat{G}_n = Kn - \widehat{L}_n$ and $G(\mathtt{Part}(n,m,\mathbf{t},\underline{i})) = Kn - L(\mathtt{Part}(n,m,\mathbf{t},\underline{i}))$, we have

$$\begin{aligned}\widehat{L}_n &\le (1-\gamma-\eta K(1+\beta)|\mathcal{C}|)\min_{\mathbf{t},\underline{i}} L(\mathtt{Part}(n,m,\mathbf{t},\underline{i})) + Kn(\gamma+\eta(1+\beta)K|\mathcal{C}|) \\ &\quad + (1-\gamma-\eta(1+\beta)K|\mathcal{C}|)\,\frac{K(m+1)}{\beta}\ln\frac{|E|(m+1)}{\delta} + n|E|\beta \\ &\quad + \frac{1}{\eta}\ln\left(\frac{N^{m+1}}{\alpha^m(1-\alpha)^{n-m-1}}\right).\end{aligned}$$

This implies that, with probability at least $1-\delta$,

$$\begin{aligned}&\widehat{L}_n - \min_{\mathbf{t},\underline{i}} L(\mathtt{Part}(n,m,\mathbf{t},\underline{i})) \\ &\quad\le Kn(\gamma+\eta(1+\beta)K|\mathcal{C}|) + \frac{K(m+1)}{\beta}\ln\frac{|E|(m+1)}{\delta} \\ &\qquad + n|E|\beta + \frac{1}{\eta}\ln\left(\frac{N^{m+1}}{\alpha^m(1-\alpha)^{n-m-1}}\right). \qquad (17)\end{aligned}$$

To prove the second statement, let $H(p) = -p\ln p - (1-p)\ln(1-p)$ and $D(p\,\|\,q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$. Optimizing the value of α in the last term of (17) gives

$$\begin{aligned}&\frac{1}{\eta}\ln\left(\frac{N^{m+1}}{\alpha^m(1-\alpha)^{n-m-1}}\right) \\ &\quad= \frac{1}{\eta}\left((m+1)\ln(N) + m\ln\frac{1}{\alpha} + (n-m-1)\ln\frac{1}{1-\alpha}\right) \\ &\quad= \frac{1}{\eta}\big((m+1)\ln(N) + (n-1)(D_b(\alpha^*\,\|\,\alpha) + H_b(\alpha^*))\big)\end{aligned}$$

where $\alpha^* = \frac{m}{n-1}$. For $\alpha = \alpha^*$ we obtain

$$\begin{aligned}&\frac{1}{\eta}\ln\left(\frac{N^{m+1}}{\alpha^m(1-\alpha)^{n-m-1}}\right) \\ &\quad= \frac{1}{\eta}\left((m+1)\ln(N) + (n-1)(H_b(\alpha^*))\right) \\ &\quad= \frac{1}{\eta}\left((m+1)\ln(N) + m\ln((n-1)/m)\right. \\ &\qquad \left.+(n-m-1)\ln(1+m/(n-m-1))\right) \\ &\quad\le \frac{1}{\eta}\left((m+1)\ln(N) + m\ln((n-1)/m) + m\right) \\ &\quad= \frac{1}{\eta}\left((m+1)\ln(N) + m\ln\frac{e(n-1)}{m}\right) \stackrel{\text{def}}{=} \frac{1}{\eta}D\end{aligned}$$

where the inequality follows since $\ln(1+x) \le x$ for $x > 0$. Therefore

$$\begin{aligned}&\widehat{L}_n - \min_{\mathbf{t},\underline{i}} L(\mathtt{Part}(n,m,\mathbf{t},\underline{i})) \\ &\quad\le Kn(\gamma+\eta(1+\beta)K|\mathcal{C}|) + \frac{K(m+1)}{\beta}\ln\frac{|E|(m+1)}{\delta} + n|E|\beta + \frac{1}{\eta}D\,.\end{aligned}$$

which is the first statement of the theorem. Setting

$$\beta = \sqrt{\frac{K(m+1)}{n|E|}\ln\frac{|E|(m+1)}{\delta}},\ \gamma = 2\eta K|\mathcal{C}|, \text{ and } \eta = \sqrt{\frac{D}{4nK^2|\mathcal{C}|}}$$

results in the second statement of the theorem, that is,

$$\begin{aligned}\widehat{L}_n - \min_{\mathbf{t},\underline{i}} L(\mathtt{Part}(n,m,\mathbf{t},\underline{i})) \\ \leq\ 2\sqrt{nK}\left(\sqrt{4K|\mathcal{C}|D} + \sqrt{|E|(m+1)\ln\frac{|E|(m+1)}{\delta}}\right). \quad \square\end{aligned}$$

For $t = 1$, choose I_1 uniformly from the set $\mathcal{P}$. For $t \geq 2$,

(a) Draw a Bernoulli random variable Γ_t with $\mathbb{P}(\Gamma_t = 1) = \gamma$.

(b) If $\Gamma_t = 1$, then choose I_t uniformly from $\mathcal{C}$.

(c) If $\Gamma_t = 0$,

(c1) choose τ_t randomly according to the distribution

$$\mathbb{P}\{\tau_t = t'\} = \begin{cases} \frac{(1-\alpha)^{t-1}Z_{1,t-1}}{\overline{W}_{t-1}} & \text{for } t' = 1 \\ \frac{\alpha(1-\alpha)^{t-t'}\overline{W}_{t'}Z_{t',t-1}}{N\overline{W}_t} & \text{for } t' = 2,\dots,t \end{cases}$$

where $Z_{t',t-1} = \sum_{i\in\mathcal{P}} e^{\eta G'_{i,[t',t-1]}}$ for $t' = 1,\dots,t-1$, and $Z_{t,t-1} = N$;

(c2) given $\tau_t = t'$, choose I_t randomly according to the probabilities

$$\mathbb{P}\{I_t = i|\tau_t = t'\} = \begin{cases} \frac{e^{\eta G'_{i,[t',t-1]}}}{Z_{t',t-1}} & \text{for } t' = 1,\dots,t-1 \\ \frac{1}{N} & \text{for } t' = t. \end{cases}$$

Figure 5: An alternative bandit algorithm for tracking shortest paths

Similarly to György et al. (2005a), the proposed algorithm has an alternative version, shown in Figure 5, which is efficiently computable. With a slight modification of the proof of Theorem 2 in György et al. (2005a), it can be shown that the alternative and the original algorithms are equivalent. Moreover, in a way completely analogous to György et al. (2005a), in this alternative formulation of the algorithm one can compute the probabilities the normalization factors $Z_{t',t-1}$ efficiently, as the baseline bandit algorithm for shortest paths has an $O(n|E|)$ time complexity by Theorem 3. Therefore the factors $\overline{W}_t$ and hence the probabilities $\mathbb{P}\{I_t = i|\tau_t = t'\}$ can also be computed efficiently as in György et al. (2005a). In particular, it follows from Theorem 3 of György et al. (2005a) that the time complexity of the alternative bandit algorithm for tracking the shortest path is $O(n^2|E|)$.

7. An Algorithm for the Restricted Multi-Armed Bandit Problem

In this section we consider the situation where the decision maker receives information only about the performance of the whole chosen path, but the individual edge losses are not available. That is, the forecaster has access to the sum $\ell_{I_t,t}$ of losses over the chosen path I_t but not to the losses $\{\ell_{e,t}\}_{e\in I_t}$ of the edges belonging to I_t.

This is the problem formulation considered by McMahan and Blum (2004) and Awerbuch and Kleinberg (2004). McMahan and Blum provided a relatively simple algorithm whose regret is at most of the order of $n^{-1/4}$, while Awerbuch and Kleinberg gave a more complex algorithm to achieve $O(n^{-1/3})$ regret. In this section we combine the strengths of these papers, and propose a simple algorithm with regret at most of the order of $n^{-1/3}$. Moreover, our bound holds with high probability, while the above-mentioned papers prove bounds for the expected regret only. The proposed algorithm uses ideas very similar to those of McMahan and Blum (2004). The algorithm alternates between choosing a path from a "basis" B to obtain unbiased estimates of the loss (exploration step), and choosing a path according to exponential weighting based on these estimates.

A simple way to describe a path $i \in \mathcal{P}$ is a binary row vector with $|E|$ components which are indexed by the edges of the graph such that, for each $e \in E$, the eth entry of the vector is 1 if $e \in i$ and 0 otherwise. With a slight abuse of notation we will also denote by i the binary row vector representing path i. In the previous sections, where the loss of each edge along the chosen path is available to the decision maker, the complexity stemming from the large number of paths was reduced by representing all information in terms of the edges, as the set of edges spans the set of paths. That is, the vector corresponding to a given path can be expressed as the linear combination of the unit vectors associated with the edges (the eth component of the unit vector representing edge e is 1, while the other components are 0). While the losses corresponding to such a spanning set are not observable in the restricted setting of this section, one can choose a subset of $\mathcal{P}$ that forms a *basis*, that is, a collection of b paths which are linearly independent and each path in $\mathcal{P}$ can be expressed as a linear combination of the paths in the basis. We denote by B the $b \times |E|$ matrix whose rows $b^1,\dots,b^b$ represent the paths in the basis. Note that b is equal to the maximum number of linearly independent vectors in $\{i : i \in \mathcal{P}\}$, so $b \le |E|$.

Let $\ell_t^{(E)}$ denote the (column) vector of the edge losses $\{\ell_{e,t}\}_{e\in E}$ at time t, and let $\ell_t^{(B)} = (\ell_{b^1,t},\dots,\ell_{b^b,t})^T$ be a b-dimensional column vector whose components are the losses of the paths in the basis B at time t. If $\alpha_{b^1}^{(i,B)},\dots,\alpha_{b^b}^{(i,B)}$ are the coefficients in the linear combination of the basis paths expressing path $i \in \mathcal{P}$, that is, $i = \sum_{j=1}^b \alpha_{b^j}^{(i,B)} b^j$, then the loss of path $i \in \mathcal{P}$ at time t is given by

$$\ell_{i,t} = \langle i, \ell_t^{(E)} \rangle = \sum_{j=1}^b \alpha_{b^j}^{(i,B)} \langle b^j, \ell_t^{(E)} \rangle = \sum_{j=1}^b \alpha_{b^j}^{(i,B)} \ell_{b^j,t} \tag{18}$$

where $\langle \cdot,\cdot \rangle$ denotes the standard inner product in $\mathbb{R}^{|E|}$. In the algorithm we obtain estimates $\tilde{\ell}_{b^j,t}$ of the losses of the basis paths and use (18) to estimate the loss of any $i \in \mathcal{P}$ as

$$\tilde{\ell}_{i,t} = \sum_{j=1}^b \alpha_{b^j}^{(i,B)} \tilde{\ell}_{b^j,t} \,. \tag{19}$$

It is algorithmically advantageous to calculate the estimated path losses $\tilde{\ell}_{i,t}$ from an intermediate estimate of the individual edge losses. Let B^+ denote the Moore-Penrose inverse of B defined by

$B^+ = B^T(BB^T)^{-1}$, where B^T denotes the transpose of B and BB^T is invertible since the rows of B are linearly independent. (Note that $BB^+ = I_b$, the $b \times b$ identity matrix, and $B^+ = B^{-1}$ if $b = |E|$.) Then letting $\tilde{\ell}_t^{(B)} = (\tilde{\ell}_{b^1,t}, \ldots, \tilde{\ell}_{b^b,t})^T$ and

$$\tilde{\ell}_t^{(E)} = B^+ \tilde{\ell}_t^{(B)}$$

it is easy to see that $\tilde{\ell}_{i,t}$ in (19) can be obtained as $\tilde{\ell}_{i,t} = \langle i, \tilde{\ell}_t^{(E)} \rangle$, or equivalently

$$\tilde{\ell}_{i,t} = \sum_{e \in i} \tilde{\ell}_{e,t}.$$

This form of the path losses allows for an efficient implementation of exponential weighting via dynamic programming Takimoto and Warmuth (2003).

To analyze the algorithm we need an upper bound on the magnitude of the coefficients $\alpha_{b^j}^{(i,B)}$. For this, we invoke the definition of a barycentric spanner from Awerbuch and Kleinberg (2004): the basis B is called a *C-barycentric spanner* if $|\alpha_{b^j}^{(i,B)}| \le C$ for all $i \in \mathcal{P}$ and $j = 1, \ldots, b$. Awerbuch and Kleinberg (2004) show that a 1-barycentric spanner exists if B is a square matrix (i.e., $b = |E|$) and give a low-complexity algorithm which finds a C-barycentric spanner for $C > 1$. We use their technique to show that a 1-barycentric spanner also exists in case of a non-square B, when the basis is chosen to maximize the absolute value of the determinant of BB^T. As before, b denotes the maximum number of linearly independent vectors (paths) in $\mathcal{P}$.

Lemma 10 *For a directed acyclic graph, the set of paths $\mathcal{P}$ between two dedicated nodes has a 1-barycentric spanner. Moreover, let B be a $b \times |E|$ matrix with rows from $\mathcal{P}$ such that* $\det[BB^T] \neq 0$. *If $B_{-j,i}$ is the matrix obtained from B by replacing its jth row by $i \in \mathcal{P}$ and*

$$\left|\det\left[B_{-j,i}B_{-j,i}^T\right]\right| \le C^2 \left|\det\left[BB^T\right]\right| \tag{20}$$

for all $j = 1, \ldots, b$ and $i \in \mathcal{P}$, then B is a C-barycentric spanner.

Proof Let B be a basis of $\mathcal{P}$ with rows $b^1, \ldots, b^b \in \mathcal{P}$ that maximizes $|\det[BB^T]|$. Then, for any path $i \in \mathcal{P}$, we have $i = \sum_{j=1}^b \alpha_{b^j}^{(i,B)} b^j$ for some coefficients $\{\alpha_{b^j}^{(i,B)}\}$. Now for the matrix $B_{-1,i} = [i^T, (b^2)^T, \ldots, (b^b)^T]^T$ we have

$$\begin{aligned}
&\left|\det\left[B_{-1,i}B_{-1,i}^T\right]\right| \\
&= \left|\det\left[B_{-1,i}i^T, B_{-1,i}(b^2)^T, B_{-1,i}(b^3)^T, \ldots, B_{-1,i}(b^b)^T\right]\right| \\
&= \left|\det\left[\left(\sum_{j=1}^b \alpha_{b^j}^{(i,B)} B_{-1,i} b^j\right)^T, B_{-1,i}(b^2)^T, B_{-1,i}(b^3)^T, \ldots, B_{-1,i}(b^b)^T\right]\right| \\
&= \left|\sum_{j=1}^b \alpha_{b^j}^{(i,B)} \det\left[B_{-1,i}(b^j)^T, B_{-1,i}(b^2)^T, B_{-1,i}(b^3)^T, \ldots, B_{-1,i}(b^b)^T\right]\right| \\
&= |\alpha_{b^1}^{(i,B)}| \left|\det\left[B_{-1,i}B^T\right]\right| \\
&= \left(\alpha_{b^1}^{(i,B)}\right)^2 \left|\det\left[BB^T\right]\right|
\end{aligned}$$

where last equality follows by the same argument the penultimate equality was obtained. Repeating the same argument for $B_{-j,i}$, $j = 2,\dots,b$ we obtain

$$\left|\det\left[B_{-j,i}B_{-j,i}^T\right]\right| = \left(\alpha_{b^j}^{(i,B)}\right)^2 \left|\det\left[BB^T\right]\right|. \tag{21}$$

Thus the maximal property of $|\det[BB^T]|$ implies $|\alpha_{b^j}^{(i,B)}| \le 1$ for all $j = 1,\dots,b$. The second statement follows trivially from (20) and (21). □

Awerbuch and Kleinberg (2004, Proposition 2.4) also present an iterative algorithm to find a C-barycentric spanner if B is a square matrix. Their algorithm has two parts. First, starting from the identity matrix, the algorithm replaces sequentially the rows of the matrix in each step by maximizing the determinant with respect to the given row. This is done by calling b times an optimization oracle, to compute $\arg\max_{i\in\mathcal{P}} |\det[B_{-j,i}]|$ for $j = 1,2,\dots,b$. In the second part the algorithm replaces an arbitrarily row j of the matrix in each iteration with some $i \in \mathcal{P}$ if $|\det[B_{-j,i}]| > C|\det[B]|$. It is shown that the oracle is called in the second part $O(b\log_C b)$ times for $C > 1$. In case B is not a square matrix, the algorithm carries over if we have access to an alternative optimization oracle that can compute $\arg\max_{i\in\mathcal{P}} |\det[B_{-j,i}B_{-j,i}^T]|$: In the first b steps, all the rows of the matrix are replaced (first part), then we can iteratively replace one row in each step, using the oracle, to maximize the determinant $|\det[B_{-j,i}B_{-j,i}^T]|$ in i until (20) is satisfied for all j and i. By Lemma 10, this results is a C-barycentric spanner. Similarly to Awerbuch and Kleinberg (2004, Lemma 2.5), it can be shown that the alternative optimization oracle is called $O(b\log_C b)$ times for $C > 1$.

For simplicity (to avoid carrying the constant C), assume that we have a 2-barycentric spanner B. Based on the ideas of label efficient prediction, the next algorithm, shown in Figure 6, gives a simple solution to the restricted shortest path problem. The algorithm is very similar to that of the algorithm in the label efficient case, but here we cannot estimate the edge losses directly. Therefore, we query the loss of a (random) basis vector from time to time, and create unbiased estimates $\tilde{\ell}_{b^j,t}$ of the losses of basis paths $\ell_{b^j,t}$, which are then transformed into edge-loss estimates.

The performance of the algorithm is analyzed in the next theorem. The proof follows the argument of Cesa-Bianchi et al. (2005), but we also have to deal with some additional technical difficulties. Note that in the theorem we do not assume that all paths between u and v have equal length.

Theorem 11 *Let K denote the length of the longest path in the graph. For any $\delta \in (0,1)$, parameters $0 < \varepsilon \le \frac{1}{K}$ and $\eta > 0$ satisfying $\eta \le \varepsilon^2$, and $n \ge \frac{8b}{\varepsilon^2}\ln\frac{4bN}{\delta}$, the performance of the algorithm defined above can be bounded, with probability at least $1-\delta$, as*

$$\widehat{L}_n - \min_{i\in\mathcal{P}} L_{i,n} \le K\left(\frac{\eta b}{\varepsilon}Kn + \sqrt{\frac{n}{2}\ln\frac{4}{\delta}} + n\varepsilon + \frac{\sqrt{2n\varepsilon\ln\frac{4}{\delta}}}{K} + \frac{16}{3}b\sqrt{2n\frac{b}{\varepsilon}\ln\frac{4bN}{\delta}}\right) + \frac{\ln N}{\eta}$$

In particular, choosing

$$\varepsilon = \left(\frac{Kb}{n}\ln\frac{4bN}{\delta}\right)^{1/3} \quad \text{and} \quad \eta = \varepsilon^2$$

we obtain

$$\widehat{L}_n - \min_{i\in\mathcal{P}} L_{i,n} \le 9.1K^2 b\left(Kb\ln(4bN/\delta)\right)^{1/3} n^{2/3}.$$

Parameters: $0 < \varepsilon, \eta \leq 1$.
Initialization: Set $w_{e,0} = 1$ for each $e \in E$, $\overline{w}_{i,0} = 1$ for each $i \in \mathcal{P}$, $\overline{W}_0 = N$. Fix a basis B, which is a 2-barycentric spanner. For each round $t = 1, 2, \ldots$

(a) Draw a Bernoulli random variable S_t such that $\mathbb{P}(S_t = 1) = \varepsilon$;

(b) If $S_t = 1$, then choose the path I_t uniformly from the basis B. If $S_t = 0$, then choose I_t according to the distribution $\{p_{i,t}\}$, defined by

$$p_{i,t} = \frac{\overline{w}_{i,t-1}}{\overline{W}_{t-1}} .$$

(c) Calculate the estimated loss of all edges according to

$$\tilde{\ell}_t^{(E)} = B^+ \tilde{\ell}_t^{(B)} ,$$

where $\tilde{\ell}_t^{(E)} = \{\tilde{\ell}_{e,t}^{(E)}\}_{e \in E}$, and $\tilde{\ell}_t^{(B)} = (\tilde{\ell}_{b^1,t}^{(B)}, \ldots, \tilde{\ell}_{b^b,t}^{(B)})$ is the vector of the estimated losses

$$\tilde{\ell}_{b^j,t} = \frac{S_t}{\varepsilon} \ell_{b^j,t} \mathbb{1}_{\{I_t = b^j\}} b$$

for $j = 1, \ldots, b$.

(d) Compute the updated weights

$$\begin{aligned} w_{e,t} &= w_{e,t-1} e^{-\eta \tilde{\ell}_{e,t}} , \\ \overline{w}_{i,t} &= \prod_{e \in i} w_{e,t} = \overline{w}_{i,t-1} e^{-\eta \sum_{e \in i} \tilde{\ell}_{e,t}} , \end{aligned}$$

and the sum of the total weights of the paths

$$\overline{W}_t = \sum_{i \in \mathcal{P}} \overline{w}_{i,t} .$$

Figure 6: A bandit algorithm for the restricted shortest path problem

The theorem is proved using the following two lemmas. The first one is an easy consequence of Bernstein's inequality:

Lemma 12 *Under the assumptions of Theorem 11, the probability that the algorithm queries the basis more than* $n\varepsilon + \sqrt{2n\varepsilon \ln \frac{4}{\delta}}$ *times is at most* $\delta/4$.

Using the estimated loss of a path $i \in \mathcal{P}$ given in (19), we can estimate the cumulative loss of i up to time n as

$$\tilde{L}_{i,n} = \sum_{t=1}^{n} \tilde{\ell}_{i,t} .$$

The next lemma demonstrates the quality of these estimates.

Lemma 13 *Let $0 < \delta < 1$ and assume $n \geq \frac{8b}{\varepsilon} \ln \frac{4bN}{\delta}$. For any $i \in \mathcal{P}$, with probability at least $1 - \delta/4$,*

$$\sum_{t=1}^{n} \sum_{i \in \mathcal{P}} p_{i,t} \ell_{i,t} - \sum_{t=1}^{n} \sum_{i \in \mathcal{P}} p_{i,t} \tilde{\ell}_{i,t} \leq \frac{8}{3} b \sqrt{2n \frac{bK^2}{\varepsilon} \ln \frac{4b}{\delta}} \, .$$

Furthermore, with probability at least $1 - \delta/(4N)$,

$$\tilde{L}_{i,n} - L_{i,n} \leq \frac{8}{3} b \sqrt{2n \frac{bK^2}{\varepsilon} \ln \frac{4bN}{\delta}} \, .$$

Proof We may write

$$\begin{aligned}
\sum_{t=1}^{n} \sum_{i \in \mathcal{P}} p_{i,t} \ell_{i,t} - \sum_{t=1}^{n} \sum_{i \in \mathcal{P}} p_{i,t} \tilde{\ell}_{i,t} &= \sum_{t=1}^{n} \sum_{i \in \mathcal{P}} p_{i,t} \sum_{j=1}^{b} \alpha_{b^j}^{(i,B)} \left(\ell_{b^j,t} - \tilde{\ell}_{b^j,t} \right) \\
&= \sum_{j=1}^{b} \sum_{t=1}^{n} \left[\sum_{i \in \mathcal{P}} p_{i,t} \alpha_{b^j}^{(i,B)} \left(\ell_{b^j,t} - \tilde{\ell}_{b^j,t} \right) \right] \\
&\stackrel{\text{def}}{=} \sum_{j=1}^{b} \sum_{t=1}^{n} X_{b^j,t} \, . \qquad (22)
\end{aligned}$$

Note that for any b^j, $X_{b^j,t}$, $t = 1, 2, \ldots$ is a martingale difference sequence with respect to (I_t, S_t), $t = 1, 2, \ldots$ as $\mathbb{E}_t \tilde{\ell}_{b^j,t} = \ell_{b^j,t}$. Also,

$$\mathbb{E}_t[X_{b^j,t}^2] \leq \left(\sum_{i \in \mathcal{P}} p_{i,t} \alpha_{b^j}^{(i,B)} \right)^2 \mathbb{E}_t \left[\left(\tilde{\ell}_{b^j,t} \right)^2 \right] \leq \sum_{i \in \mathcal{P}} p_{i,t} \left(\alpha_{b^j}^{(i,B)} \right)^2 \frac{K^2 b}{\varepsilon} \leq 4 \frac{K^2 b}{\varepsilon} \qquad (23)$$

and

$$|X_{b^j,t}| \leq \left| \sum_{i \in \mathcal{P}} p_{i,t} \alpha_{b^j}^{(i,B)} \right| \left| \ell_{b^j,t} - \tilde{\ell}_{b^j,t} \right| \leq \sum_{i \in \mathcal{P}} p_{i,t} \left| \alpha_{b^j}^{(i,B)} \right| \frac{Kb}{\varepsilon} \leq 2 \frac{Kb}{\varepsilon} \qquad (24)$$

where the last inequalities in both cases follow from the fact that B is a 2-barycentric spanner. Then, using Bernstein's inequality for martingale differences (Lemma 14), we have, for any fixed b^j,

$$\mathbb{P} \left[\sum_{t=1}^{n} X_{b^j,t} \geq \frac{8}{3} \sqrt{2n \frac{bK^2}{\varepsilon} \ln \frac{4b}{\delta}} \right] \leq \frac{\delta}{4b}$$

where we used (23), (24) and the assumption of the lemma on n. The proof of the first statement is finished with an application of the union bound and its combination with (22).

For the second statement we use a similar argument, that is,

$$\begin{aligned}
\sum_{t=1}^{n} (\tilde{\ell}_{i,t} - \ell_{i,t}) = \sum_{j=1}^{b} \alpha_{b^j}^{(i,B)} \sum_{t=1}^{n} (\tilde{\ell}_{b^j,t} - \ell_{b^j,t}) &\leq \sum_{j=1}^{b} \left| \alpha_{b^j}^{(i,B)} \right| \left| \sum_{t=1}^{n} (\tilde{\ell}_{b^j,t} - \ell_{b^j,t}) \right| \\
&\leq 2 \sum_{j=1}^{b} \left| \sum_{t=1}^{n} (\tilde{\ell}_{b^j,t} - \ell_{b^j,t}) \right| \, . \qquad (25)
\end{aligned}$$

Now applying Lemma 14 for a fixed b^j we get

$$\mathbb{P}\left[\sum_{t=1}^{n}(\tilde{\ell}_{b^j,t}-\ell_{b^j,t})\geq\frac{4}{3}\sqrt{2n\frac{K^2b}{\varepsilon}\ln\frac{4bN}{\delta}}\right]\leq\frac{\delta}{4bN} \tag{26}$$

because of $\mathbb{E}_t[(\tilde{\ell}_{b^j,t}-\ell_{b^j,t})^2]\leq\frac{K^2b}{\varepsilon}$ and $-K\leq\tilde{\ell}_{b^j,t}-\ell_{b^j,t}\leq K\left(\frac{b}{\varepsilon}-1\right)$. The proof is completed by applying the union bound to (26) and combining the result with (25). □

Proof of Theorem 11. Similarly to earlier proofs, we follow the evolution of the term $\ln\frac{\overline{W}_n}{\overline{W}_0}$. In the same way as we obtained (5) and (7), we have

$$\ln\frac{\overline{W}_n}{\overline{W}_0}\geq-\eta\min_{i\in\mathcal{P}}\widetilde{L}_{i,n}-\ln N$$

and

$$\ln\frac{\overline{W}_n}{\overline{W}_0}\quad\leq\quad\sum_{t=1}^{n}\left(-\eta\sum_{i\in\mathcal{P}}p_{i,t}\tilde{\ell}_{i,t}+\frac{\eta^2}{2}\sum_{i\in\mathcal{P}}p_{i,t}\tilde{\ell}_{i,t}^2\right)\,.$$

Combining these bounds, we obtain

$$\begin{aligned}-\min_{i\in\mathcal{P}}\widetilde{L}_{i,n}-\frac{\ln N}{\eta}\quad&\leq\quad\sum_{t=1}^{n}\left(-\sum_{i\in\mathcal{P}}p_{i,t}\tilde{\ell}_{i,t}+\frac{\eta}{2}\sum_{i\in\mathcal{P}}p_{i,t}\tilde{\ell}_{i,t}^2\right)\\&\leq\quad\left(-1+\frac{\eta Kb}{\varepsilon}\right)\sum_{t=1}^{n}\sum_{i\in\mathcal{P}}p_{i,t}\tilde{\ell}_{i,t}\,,\end{aligned}$$

because $0\leq\tilde{\ell}_{i,t}\leq\frac{2Kb}{\varepsilon}$. Applying the results of Lemma 13 and the union bound, we have, with probability $1-\delta/2$,

$$\begin{aligned}&-\min_{i\in\mathcal{P}}L_{i,n}-\frac{8}{3}b\sqrt{2n\frac{K^2b}{\varepsilon}\ln\frac{4bN}{\delta}}\\&\leq\quad\left(-1+\frac{\eta Kb}{\varepsilon}\right)\left(\sum_{t=1}^{n}\sum_{i\in\mathcal{P}}p_{i,t}\ell_{i,t}-\frac{8}{3}b\sqrt{2n\frac{K^2b}{\varepsilon}\ln\frac{4b}{\delta}}\right)+\frac{\ln N}{\eta}\\&\leq\quad\left(-1+\frac{\eta Kb}{\varepsilon}\right)\sum_{t=1}^{n}\sum_{i\in\mathcal{P}}p_{i,t}\ell_{i,t}+\frac{8}{3}b\sqrt{2n\frac{K^2b}{\varepsilon}\ln\frac{4b}{\delta}}+\frac{\ln N}{\eta}\,.\end{aligned} \tag{27}$$

Introduce the sets

$$\mathcal{T}_n\stackrel{\text{def}}{=}\{t:1\leq t\leq n\text{ and }S_t=0\}\qquad\text{and}\qquad\overline{\mathcal{T}}_n\stackrel{\text{def}}{=}\{t:1\leq t\leq n\text{ and }S_t=1\}$$

of "exploitation" and "exploration" steps, respectively. Then, by the Hoeffding-Azuma inequality (Hoeffding, 1963) we obtain that, with probability at least $1-\delta/4$,

$$\sum_{t\in\mathcal{T}_n}\sum_{i\in\mathcal{P}}p_{i,t}\ell_{i,t}\geq\sum_{t\in\mathcal{T}_n}\ell_{I_t,t}-\sqrt{\frac{|\mathcal{T}_n|K^2}{2}\ln\frac{4}{\delta}}\,.$$

Note that for the exploration steps $t \in \overline{\mathcal{T}}_n$, as the algorithm plays according to a uniform distribution instead of $p_{i,t}$, we can only use the trivial lower bound zero on the losses, that is,

$$\sum_{t\in\overline{\mathcal{T}}_n}\sum_{i\in\mathcal{P}} p_{i,t}\ell_{i,t} \geq \sum_{t\in\overline{\mathcal{T}}_n} \ell_{I_t,t} - K|\overline{\mathcal{T}}_n| \ .$$

The last two inequalities imply

$$\sum_{t=1}^{n}\sum_{i\in\mathcal{P}} p_{i,t}\ell_{i,t} \geq \widehat{L}_n - \sqrt{\frac{|\mathcal{T}_n|K^2}{2}\ln\frac{4}{\delta}} - K|\overline{\mathcal{T}}_n| \ . \tag{28}$$

Then, by (27), (28) and Lemma 12 we obtain, with probability at least $1-\delta$,

$$\begin{aligned}
&\widehat{L}_n - \min_{i\in\mathcal{P}} L_{i,n} \\
&\quad\leq\ K\left(\frac{\eta b}{\varepsilon}Kn + \sqrt{\frac{n}{2}\ln\frac{4}{\delta}} + n\varepsilon + \frac{\sqrt{2n\varepsilon\ln\frac{4}{\delta}}}{K} + \frac{16}{3}b\sqrt{2n\frac{b}{\varepsilon}\ln\frac{4bN}{\delta}}\right) + \frac{\ln N}{\eta}
\end{aligned}$$

where we used $\widehat{L}_n \leq Kn$ and $|\mathcal{T}_n| \leq n$. Substituting the values of ε and η gives

$$\begin{aligned}
\widehat{L}_n - \min_{i\in\mathcal{P}} L_{i,n} &\leq K^2 bn\varepsilon + \frac{1}{4}Kn\varepsilon + Kn\varepsilon + \frac{1}{2}n\varepsilon + \frac{16}{3}b\sqrt{K}n\varepsilon + n\varepsilon \\
&\leq 9.1K^2 bn\varepsilon
\end{aligned}$$

where we used $\sqrt{\frac{n}{2}\ln\frac{4}{\delta}} \leq \frac{1}{4}n\varepsilon$, $\sqrt{2n\varepsilon\ln\frac{4}{\delta}} \leq \frac{1}{2}n\varepsilon$, $\sqrt{n\frac{bK}{\varepsilon}\ln\frac{4N}{\delta}} = n\varepsilon$, and $\frac{\ln N}{\eta} \leq n\varepsilon$ (from the assumptions of the theorem). □

8. Simulation Results

To further investigate our new algorithms, we have conducted some simple simulations. As the main motivation of this work is to improve earlier algorithms in case the number of paths is exponentially large in the number of edges, we tested the algorithms on the small graph shown in Figure 1 (b), which has one of the simplest structures with exponentially many (namely $2^{|E|/2}$) paths.

The losses on the edges were generated by a sequence of independent and uniform random variables, with values from $[0,1]$ on the upper edges, and from $[0.32,1]$ on the lower edges, resulting in a (long-term) optimal path consisting of the upper edges. We ran the tests for $n = 10000$ steps, with confidence value $\delta = 0.001$. To establish baseline performance, we also tested the EXP3 algorithm of Auer et al. (2002) (note that this algorithm does not need edge losses, only the loss of the chosen path). For the version of our bandit algorithm that is informed of the individual edge losses (edge-bandit), we used the simple 2-element cover set of the paths consisting of the upper and lower edges, respectively (other 2-element cover sets give similar performance). For our restricted shortest path algorithm (path-bandit) the basis $\{uuuuu, uuuul, uuull, uulll, ullll, lllll\}$ was used, where u (resp. l) in the kth position denotes the upper (resp. lower) edge connecting v_{k-1} and v_k. In this example the performance of the algorithm appeared to be independent of the actual choice of the basis; however, in general we do not expect this behavior. Two versions of the algorithm of

Awerbuch and Kleinberg (2004) were also simulated. With its original parameter setting (AwKl), the algorithm did not perform well. However, after optimizing its parameters off-line (AwKl tuned), substantially better performance was achieved. The normalized regret of the above algorithms, averaged over 30 runs, as well as the regret of the fixed paths in the graph are shown in Figure 7.

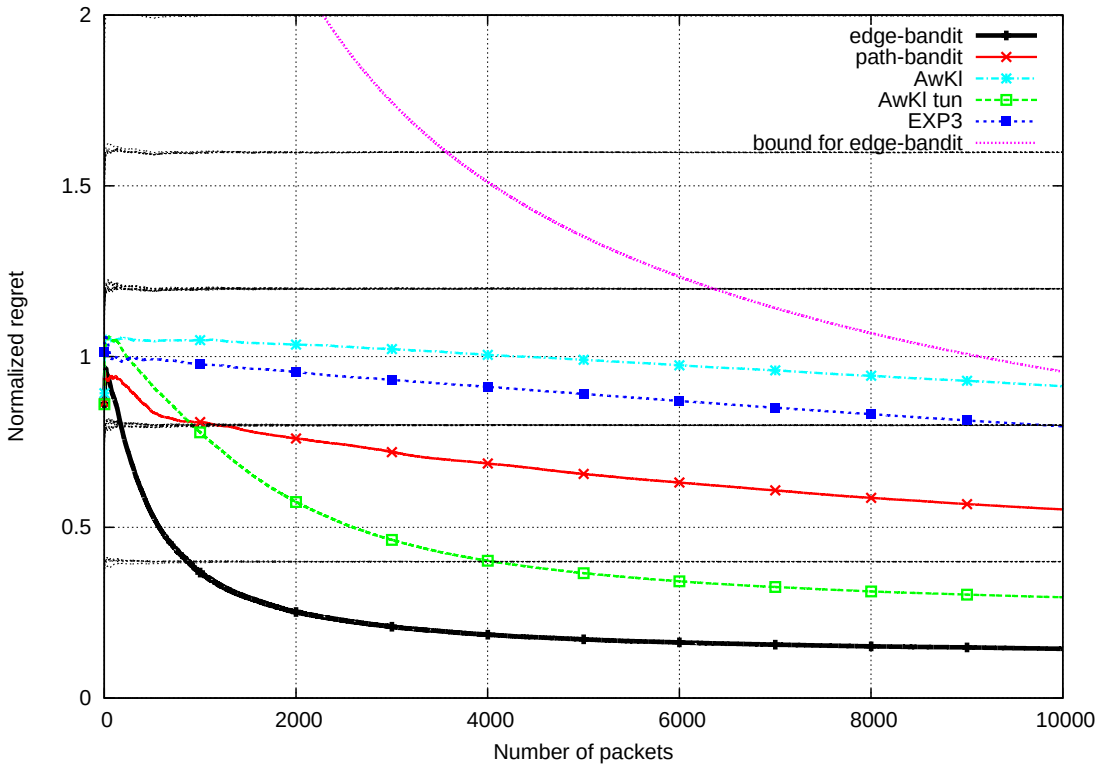

Figure 7: Normalized regret of several algorithms for the shortest path problem. The gray dotted lines show the normalized regret of fixed paths in the graph.

Although all algorithms showed better performance than the bound for the edge-bandit algorithm, the latter showed the expected superior performance in the simulations. Furthermore, our algorithm for the restricted shortest path problem outperformed Awerbuch and Kleinberg's (AwKl) algorithm, while being inferior to its off-line tuned version (AwKl tuned). It must be noted that similar parameter optimization did not improve the performance of our path-bandit algorithm, which showed robust behavior with respect to parameter tuning.

9. Conclusions

We considered different versions of the on-line shortest path problem with limited feedback. These problems are motivated by realistic scenarios, such as routing in communication networks, where the vertices do not have all the information about the state of the network. We have addressed the problem in the adversarial setting where the edge losses may vary in an arbitrary way; in particular, they may depend on previous routing decisions of the algorithm. Although this assumption may

neglect natural correlation in the loss sequence, it suits applications in mobile ad-hoc networks, where the network topology changes dynamically in time, and also in certain secure networks that has to be able to handle denial of service attacks.

Efficient algorithms have been provided for the multi-armed bandit setting and in a combined label efficient multi-armed bandit setting, provided the individual edge losses along the chosen path are revealed to the algorithms. The normalized regrets of the algorithms, compared to the performance of the best fixed path, converge to zero at an $O(1/\sqrt{n})$ rate as the time horizon n grows to infinity, and increases only polynomially in the number of edges (and vertices) of the graph. Earlier methods for the multi-armed bandit problem either do not have the right $O(1/\sqrt{n})$ convergence rate, or their regret increase exponentially in the number of edges for typical graphs. The algorithm has also been extended so that it can compete with time varying paths, that is, to handle situations when the best path can change from time to time (for consistency, the number of changes must be sublinear in n).

In the restricted version of the shortest path problem, where only the losses of the whole paths are revealed, an algorithm with a worse $O(n^{-1/3})$ normalized regret was provided. This algorithm has comparable performance to that of the best earlier algorithm for this problem Awerbuch and Kleinberg (2004), however, our algorithm is significantly simpler. Simulation results are also given to assess the practical performance and compare it to the theoretical bounds as well as other competing algorithms.

It should be noted that the results are not entirely satisfactory in the restricted version of the problem, as it remains an open question whether the $O(1/\sqrt{n})$ regret can be achieved without the exponential dependence on the size of the graph. Although we expect that this is the case, we have not been able to construct an algorithm with such a proven performance bound.

Acknowledgments

This research was supported in part by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences, the Mobile Innovation Center of Hungary, by the Hungarian Scientific Research Fund (OTKA F60787), by the Natural Sciences and Engineering Research Council (NSERC) of Canada, by the Spanish Ministry of Science and Technology grant MTM2006-05650, by Fundación BBVA, by the PASCAL Network of Excellence under EC grant no. 506778 and by the High Speed Networks Laboratory, Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics. Parts of this paper have been presented at COLT'06.

Appendix A.

First we describe a simple algorithm that, given a directed acyclic graph (V,E) with a source vertex u and destination vertex v, constructs a graph by adding at most $(K-2)(|V|-2)+1$ vertices and edges (with constant weight zero) to the graph without modifying the weights of the paths between u and v such that each path from u to v is of length K, where K denotes the length of the longest path of the original graph.

Order the vertices v_i, $i=1,\ldots,|V|$ of the graph such that if $(v_i,v_j)\in E$ then $i<j$. Replace the destination vertex $v=v_{|V|}$ with a chain of K vertices $v_{|V|,0},\ldots,v_{|V|,K-1}$ and vertices v_i, $i=3,\ldots,|V|-1$ with a chain of $K-1$ vertices $v_{i,0},\ldots,v_{i,K-2}$ such that in the chains the only edges

are of the form $(v_{i,k+1}, v_{i,k})$ (for each possible value of k), and these edges are of constant weight zero. Furthermore, if $(v_i, v_j) \in E$ is such that the length of the longest path from v_i (resp. v_j) to the destination is K_i (resp. K_j), then this edge is replaced in the new graph by $(v_{i,0}, v_{j,K_j-K_i-1})$ whose weight equals that of the original at each time instant. (Note that here $v_{1,0} = v_1 = u$ and $v_{2,0} = v_2$ and $K_i < K_j$.) It is easy to see that each path from source to destination is of length K in the new graph and the weights of the new paths are the same as those of the corresponding originals.

Next we recall a martingale inequality used in the proofs:

Lemma 14 *(Bernstein's inequality for martingale differences (Freedman, 1975).) Let $X_1, \ldots, X_n$ be a martingale difference sequence such that $X_t \in [a,b]$ with probability one ($t = 1, \ldots, n$). Assume that, for all t,*

$$\mathbb{E}\left[X_t^2 | X_{t-1}, \ldots, X_1\right] \leq \sigma^2 \ a.s.$$

Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left\{\sum_{t=1}^{n} X_t > \varepsilon\right\} \leq e^{\frac{-\varepsilon^2}{2n\sigma^2 + 2\varepsilon(b-a)/3}}$$

and therefore

$$\mathbb{P}\left\{\sum_{t=1}^{n} X_t > \sqrt{2n\sigma^2 \ln \delta^{-1}} + 2\ln \delta^{-1}(b-a)/3\right\} \leq \delta.$$

References

C. Allenberg, P. Auer, L. Györfi, and Gy. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Proceedings of 17th International Conference on Algorithmic Learning Theory, ALT 2006, Lecture Notes in Computer Science 4264*, pages 229–243, Barcelona, Spain, Oct. 2006.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

P. Auer and M. Warmuth. Tracking the best disjunction. *Machine Learning*, 32(2):127–150, 1998.

P. Auer and Gy. Ottucsák. Bound on high-probability regret in loss-bandit game. Preprint, 2006. http://www.szit.bme.hu/õti/green.pdf.

B. Awerbuch, D. Holmer, H. Rubens, and R. Kleinberg. Provably competitive adaptive routing. In *Proceedings of IEEE INFOCOM 2005*, volume 1, pages 631–641, March 2005.

B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the 36th Annual ACM Symposium on the Theory of Computing, STOC 2004*, pages 45–53, Chicago, IL, USA, Jun. 2004. ACM Press.

D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

O. Bousquet and M. K. Warmuth. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3:363–396, Nov. 2002.

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

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE Trans. Inform. Theory*, IT-51:2152–2162, June 2005.

D.A. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3:100–118, 1975.

E. Gelenbe, M. Gellman, R. Lent, P. Liu, and P. Su. Autonomous smart routing for network QoS. In *Proceedings of First International Conference on Autonomic Computing*, pages 232–239, New York, May 2004. IEEE Computer Society.

E. Gelenbe, R. Lent, and Z. Xhu. Measurement and performance of a cognitive packet network. *Journal of Computer Networks*, 37:691–701, 2001.

A. György, T. Linder, and G. Lugosi. Efficient algorithms and minimax bounds for zero-delay lossy source coding. *IEEE Transactions on Signal Processing*, 52:2337–2347, Aug. 2004a.

A. György, T. Linder, and G. Lugosi. A "follow the perturbed leader"-type algorithm for zero-delay quantization of individual sequences. In *Proc. Data Compression Conference*, pages 342–351, Snowbird, UT, USA, Mar. 2004b.

A. György, T. Linder, and G. Lugosi. Tracking the best of many experts. In *Proceedings of the 18th Annual Conference on Learning Theory, COLT 2005, Lecture Notes in Computer Science 3559*, pages 204–216, Bertinoro, Italy, Jun. 2005a. Springer.

A. György, T. Linder, and G. Lugosi. Tracking the best quantizer. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 1163–1167, Adelaide, Australia, June-July 2005b.

A. György and Gy. Ottucsák. Adaptive routing using expert advice. *The Computer Journal*, 49(2): 180–189, 2006.

J. Hannan. Approximation to Bayes risk in repeated plays. In M. Dresher, A. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games*, volume 3, pages 97–139. Princeton University Press, 1957.

D.P. Helmbold and S. Panizza. Some label efficient learning results. In *Proceedings of the 10th Annual Conference on Computational Learning Theory*, pages 218–230. ACM Press, 1997.

M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.

M. Herbster and M. K. Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1:281–309, 2001.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

A. Kalai and S Vempala. Efficient algorithms for the online decision problem. In B. Schölkopf and M. Warmuth, editors, *Proceedings of the 16th Annual Conference on Learning Theory and the 7th Kernel Workshop, COLT-Kernel 2003, Lecture Notes in Computer Science 2777*, pages 26–40, New York, USA, Aug. 2003. Springer.

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

H. B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004, Lecture Notes in Computer Science 3120*, pages 109–123, Banff, Canada, Jul. 2004. Springer.

M. Mohri. General algebraic frameworks and algorithms for shortest distance problems. Technical Report 981219-10TM, AT&T Labs Research, 1998.

R. E. Schapire and D. P. Helmbold. Predicting nearly as well as the best pruning of a decision tree. *Machine Learning*, 27:51–68, 1997.

E. Takimoto and M. K. Warmuth. Path kernels and multiplicative updates. *Journal of Machine Learning Research*, 4:773–818, 2003.

V. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 372–383, Rochester, NY, Aug. 1990. Morgan Kaufmann.

V. Vovk. Derandomizing stochastic prediction strategies. *Machine Learning*, 35(3):247–282, Jun. 1999.