Stochastic Online Shortest Path Routing: The Value of Feedback

Mohammad Sadegh Talebi, Zhenhua Zou, Richard Combes, Alexandre Proutiere, and Mikael Johansson

***Abstract*—This paper studies online shortest path routing over multihop networks. Link costs or delays are time varying and modeled by independent and identically distributed random processes, whose parameters are initially unknown. The parameters, and hence the optimal path, can only be estimated by routing packets through the network and observing the realized delays. Our aim is to find a routing policy that minimizes the regret (the cumulative difference of expected delay) between the path chosen by the policy and the unknown optimal path. We formulate the problem as a combinatorial bandit optimization problem and consider several scenarios that differ in where routing decisions are made and in the information available when making the decisions. For each scenario, we derive a tight asymptotic lower bound on the regret that has to be satisfied by any online routing policy. Three algorithms, with a tradeoff between computational complexity and performance, are proposed. The regret upper bounds of these algorithms improve over those of the existing algorithms. We also assess numerically the performance of the proposed algorithms and compare it to that of existing algorithms.**

***Index Terms*—Online combinatorial optimization, shortest path routing, stochastic multiarmed bandits (MAB).**

I. Introduction

In MOST real-world networks, link delays vary stochastically due to unreliable links and random access protocols (e.g., in wireless networks), mobility (e.g., in mobile ad-hoc networks), randomness of demand (e.g., in overlay networks for peer-to-peer applications), etc. In many cases, the associated parameters with links, e.g., the packet transmission success probabilities in wireless sensor networks, are initially unknown and must be estimated by transmitting packets and observing the outcomes. When designing routing policies, we therefore need to address a challenging tradeoff between exploration and exploitation: On the one hand, it is important to route packets on new or poorly known links to explore the network and ensure that the optimal path is eventually found; on the other hand, it is critical that the accumulated knowledge on link parameters is exploited so that paths with low expected delays are preferred. When designing practical routing schemes, one is mostly concerned about the finite-time behavior of the system and it is crucial to design algorithms that quickly learn link parameters so as to efficiently track the optimal path.

The design of such routing policies is often referred to as an online shortest path routing problem in the literature [2]–[6], and is a particular instance of a combinatorial multiarmed bandit (combinatorial MAB) problem as introduced in [7]. In this paper, we study the *stochastic* version of this problem. More precisely, we consider a network in which the transmission of a packet on a given link is successful with an unknown but fixed probability. A packet is sent on a given link repeatedly until the transmission is successful; the number of time slots to complete the transmission is referred to as the *delay* on this link. We wish to route N packets from a given source to a given destination in a minimum amount of time. A routing policy selects a path to the destination on a packet-by-packet basis. The path selection can be done at the source (source routing), or in the network as the packet progresses toward the destination (hop-by-hop routing). In the case of source routing, some feedback is available when the packet reaches the destination. This feedback can be either the end-to-end delay, or the delays on each link on the path from source to destination. In the MAB literature, the former type of feedback is referred to as *bandit* feedback, whereas the latter is called *semibandit* feedback. The routing policy then selects the path for the next packet based on the feedback gathered from previously transmitted packets. In the case of hop-by-hop routing, routing decisions are taken for each transmission, and the packet is sent over a link selected based on all transmission successes and failures observed so far (for the current packet, and all previously sent packets) on the various links.

The performance of a routing policy is assessed through its expected total delay, i.e., the expected time required to send all N packets to the destination. Equivalently, it can be measured through the notion of *regret*, defined as the difference between the expected total delay under the policy considered and the expected total delay of an oracle policy that would be aware of all link parameters, and would hence always send the packets on the optimal path. Regret conveniently quantifies the loss in performance due to the fact that link parameters are initially unknown and need to be learnt.

It is worth noting that this paper is concerned about a single decision maker or agent learning to route her traffic on the

Manuscript received July 21, 2016; revised January 12, 2017 and June 11, 2017; accepted July 14, 2017. Date of publication August 30, 2017; date of current version March 27, 2018. The work of A. Proutiere was supported by the ERC grant FSA 308267. This paper was presented in part at the 2014 American Control Conference, Portland, OR, USA. Recommended by Associate Editor S. Yuksel. *(Corresponding author: Mohammad Sadegh Talebi.)*

M. S. Talebi, A. Proutiere, and M. Johansson are with the ACCESS Linnaeus Center and the School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm SE-100 44, Sweden (e-mail: mstms@kth.se; alepro@kth.se; mikaelj@kth.se).

Z. Zou is with Ericsson Research, Stockholm SE-164 83, Sweden (e-mail: zhenhua.zou@ericsson.com).

R. Combes is with the Telecommunications Department, CentraleSupelec (L2S), Gif-Sur-Yvette Cedex 91192, France (e-mail: richard.combes@supelec.fr).

Digital Object Identifier 10.1109/TAC.2017.2747409

optimal path. The agent learns to interact with a stochastic environment that is not influenced by the agent's decisions. This setting is relevant in wireless systems as explained above, but also in scenarios where the agent competes with *many* other similar agents strategically routing their traffic (see the literature on mean-field games). Our results do not apply to cases where a *few* selfish agents compete for the network resources. This scenario, often referred to as *adversarial* in the literature, has attracted a lot attention over the past few decades (see, e.g., [8]). In the adversarial setting, there are algorithms approaching Nash equilibria (NE) under some fairly mild assumptions. To the best of our knowledge, when link qualities are stochastically varying, the convergence to NEs has not been investigated.

In this paper, we first address two fundamental questions: 1) What is the benefit of allowing routing decisions at every node, rather than only at the source?; and 2) what is the added value of feeding back the observed delay for every link that a packet has traversed compared to only observing the end-to-end delay?[1] To answer these questions, we derive tight regret lower bounds satisfied by any routing policy in the different scenarios, depending on where routing decisions are made and what information is available to the decision maker when making these decisions. By comparing the different lower bounds, we are able to quantify the value of having semibandit feedback rather than bandit feedback, and the improvements that can possibly be achieved by taking routing decisions hop by hop. We then propose routing policies in the semibandit feedback setting exhibiting better regret upper bounds than existing algorithms. More precisely, our contributions are the following.

1) Regret lower bounds: We derive tight asymptotic (when N grows large) regret lower bounds. The first two bounds concern source-routing policies under bandit and semibandit feedback, respectively, whereas the third bound is satisfied by any hop-by-hop routing policy. As we shall see later, these regret bounds are tight in the sense that one can design actual routing policies, despite being complex and impractical, that achieve these bounds. As it turns out, the regret lower bounds for source-routing policies with semibandit feedback and that for hop-by-hop routing policies are identical, indicating that taking routing decisions hop by hop does not bring any advantage. On the contrary, the regret lower bounds for source-routing policies with bandit and semibandit feedback can be significantly different, illustrating the importance of having information about per-link delays.

2) Routing policies: In the case of semibandit feedback, we propose three online source-routing policies, namely, GeoCombUCB-1, GeoCombUCB-2, and KL-based source routing (KL-SR). Geo refers to the fact that the delay on a given link is geometrically distributed, Comb stands for combinatorial, and UCB (upper confidence bound) indicates that these policies are based on the same "optimism in face of uncertainty" principle as the celebrated UCB algorithm designed for classical MAB problems [9]. KL-SR has already been presented in the conference version of this paper [1]. Here, we improve its regret analysis and show that its regret scales at most as $\mathcal{O}(|E|H\Delta_{\min}^{-1}\theta_{\min}^{-2}\log(N))$,[2] where E is the set of links, H denotes the length (number of links) of the longest path in the network from the source to the destination, $\theta_{\min}$ is the success transmission probability of the link with the worst quality, and $\Delta_{\min}$ is the minimal gap between the average end-to-end delays of a suboptimal and of the optimal path (formal definitions of $\theta_{\min}$ and $\Delta_{\min}$ are provided in Section III-A). We further show that the regret under GeoCombUCB-1 and GeoCombUCB-2 scales at most as $\mathcal{O}(|E|\sqrt{H}\Delta_{\min}^{-1}\theta_{\min}^{-2}\log(N))$. The tradeoff between computational complexity and performance (regret) of online routing policies is certainly hard to characterize, but our policies provide a first insight into such a tradeoff. Furthermore, they exhibit better regret upper bounds than that of the combinatorial UCB (CUCB) algorithm [10], which is, to our knowledge, the state-of-the-art algorithm for stochastic online shortest path routing. Furthermore, we conduct numerical experiments showing that our routing policies perform significantly better than CUCB. The Thompson sampling (TS) algorithm of [11] is applicable to the shortest path problem, but its analysis for general topologies is an open problem. While TS performs slightly better than our algorithms on average, its regret sometimes has a large variance according to our experiments. The regret guarantee of various algorithms and their computational complexity are summarized in Table I.[3]

TABLE I
COMPARISON OF VARIOUS ALGORITHMS FOR SHORTEST PATH ROUTING UNDER SEMIBANDIT FEEDBACK

Algorithm	Regret	Complexity
CUCB [10], [12]	$\mathcal{O}\left(\frac{\lvert E\rvert H}{\Delta_{\min}\theta_{\min}^3}\log(N)\right)$	$\mathcal{O}(\lvert V\rvert\lvert E\rvert)$
GeoCombUCB-1	$\mathcal{O}\left(\frac{\lvert E\rvert\sqrt{H}}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$	$\mathcal{O}(\lvert\mathcal{P}\rvert)$
GeoCombUCB-2	$\mathcal{O}\left(\frac{\lvert E\rvert\sqrt{H}}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$	$\mathcal{O}(\lvert\mathcal{P}\rvert)$
KL-SR	$\mathcal{O}\left(\frac{\lvert E\rvert H}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$	$\mathcal{O}(\lvert V\rvert\lvert E\rvert)$

The remaining of this paper is organized as follows. In Section II, we review the literature related to MAB problems and to online shortest path problems. In Section III, we introduce the network model and formulate our online routing problem. Fundamental performance limits (regret lower bounds) are derived in Section IV. We propose online routing algorithms and evaluate their performance in Section V. Finally, Section VI concludes this paper and provides future research directions. All the proofs are presented in the Appendix.

II. Related Work

Stochastic MAB problems have been introduced by Robbins [13]. In the classical setting, in each round, a decision maker pulls an arm from a set of available arms and observes a realization of the corresponding reward, whose distribution is unknown. The performance of a policy is measured through its regret, defined as the difference between its expected total reward and the optimal reward the decision maker could collect if she knew the reward distributions of all arms. The goal is to find an optimal policy with the smallest regret. This classical stochastic MAB problem was solved by Lai and Robbins in their seminal paper [14], where they derived the asymptotic (when the time horizon grows large) lower bound on regret satisfied by

[1]The effect of different forms of feedback in the adversarial setting was studied in, e.g., [3] and [4].

[2]This improves over the regret upper bound scaling as $\mathcal{O}(\Delta_{\max}|E|H^3\Delta_{\min}^{-1}\theta_{\min}^{-3}\log(N))$ derived in [1], where $\Delta_{\max}$ denotes the maximal gap between the average end-to-end delays of a suboptimal and of the optimal path.

[3]In Table I, V and $\mathcal{P}$, respectively, denote the set of all nodes and the set of all possible paths between the source and the destination.

any algorithm, and proposed an optimal algorithm whose regret asymptotically matches the lower bound.

Online shortest path routing problems fall into the class of combinatorial MAB problems. In these MAB problems, arms are subsets of a set of basic actions (in routing problems, a basic action corresponds to a link), and most existing studies concern the adversarial setting, where the successive rewards of each arm are arbitrary; see, e.g., [7], [15], and [16] for algorithms for generic combinatorial problems, and [2] and [4] for efficient algorithms for routing problems. Stochastic combinatorial MAB problems have received little attention so far. Usually they are investigated in the semibandit feedback setting [10], [12], [17], [18]. Some papers deal with problems, where the set of arms exhibits very specific structures, such as fixed-size sets [19], matroid [20], and permutations [21].

In the case of online shortest path routing problems, as a particular instance of a combinatorial MAB, one could think of modeling each path as an arm, and applying sequential arm selection policies as if arms would yield independent rewards. Such policies would have a regret scaling as $|\mathcal{P}|\log(N)$, where $|\mathcal{P}|$ denotes the number of possible paths from the source to the destination. However, since $|\mathcal{P}|$ grows exponentially with the length H of the paths, treating paths as independent arms would lead to a prohibitive regret. In contrast to classical MAB in [14], where the random rewards from various arms are *independent*, in online routing problems, the end-to-end delays (i.e., the rewards) of the various paths are inherently correlated, since paths may share the same links. It may then be crucial to exploit these correlations, i.e., the structure of the problem, to design efficient routing algorithms, which in turn may have a regret scaling as $C\log(N)$, where C is much smaller than $|\mathcal{P}|$.

Next, we summarize existing results for generic stochastic combinatorial bandits that could be applied to online shortest path routing. Chen *et al.* [10] present CUCB, an algorithm for generic stochastic combinatorial MAB problems under semibandit feedback. When applied to the online routing problem, the best regret upper bound for CUCB presented in [10] scales as $\mathcal{O}(\frac{|E|H}{\Delta_{\min}\theta_{\min}^3}\log(N))$ (see [22, Appendix I] for details). This upper bound constitutes the best existing result for our problem, where the delay on each link is geometrically distributed. It is important to note that most proposed algorithms for combinatorial bandits [12], [17], [18] deal with bounded rewards, i.e., here bounded delays, and are not applicable to geometrically distributed delays. Kveton *et al.* [12] consider the case, where the rewards of basic actions (here links) can be arbitrarily correlated and bounded, and show that the regret under CUCB is $\mathcal{O}(\frac{|E|H}{\Delta_{\min}}\log(N))$. They also prove that this regret scaling has order-optimal regret in terms of $|E|$ and H.[4] In other words, the dependence of their regret upper bound on $|E|$ and H cannot be improved in general. This order-optimality does not contradict our regret upper bound [scaling as $\mathcal{O}(\frac{|E|\sqrt{H}}{\Delta_{\min}}\log(N))$], because Kveton *et al.* [12] consider possibly dependent delays across links. Interestingly, to prove that a regret of $\mathcal{O}(\frac{|E|H}{\Delta_{\min}}\log(N))$ cannot be beaten, they artificially create an instance of the problem, where the rewards of the basic actions of the same arm are identical. In other words, they consider a classical bandit problem, where the rewards of the various arms are either 0 or equal to H. This clearly highlights the fact that the confidence bound derived in [12] cannot be directly applied to our routing problem where delays are unbounded. For bounded rewards, the results of [12] have been recently improved in [18] when the rewards are *independent* across basic actions (links). There, the authors propose an algorithm whose regret scales at most as $\mathcal{O}(\frac{|E|\sqrt{H}}{\Delta_{\min}}\log(N))$. Furthermore, we remark that Wen *et al.* [23] study combinatorial problems under semibandit feedback and provide algorithms with $\mathcal{O}(\sqrt{N})$ regret. Gopalan *et al.* [11] study TS [24] for learning problems with complex arms and provide implicit regret upper bounds with $\mathcal{O}(\log(N))$ regret.

Stochastic online shortest path routing problems have been addressed in [5], [25], and [26]. Liu and Zhao [25] consider routing with bandit (end-to-end) feedback and propose a forced-exploration algorithm with $\mathcal{O}(|E|^3 H\log(N))$ regret in which a random barycentric spanner[5] path is chosen for exploration. He *et al.* [5] consider routing under semibandit feedback, where the source chooses a path for routing and a possibly different path for probing. Our model coincides with the coupled probing/routing case in their paper, for which they derive an asymptotic lower bound on the regret growing logarithmically with time. As we shall see later, their lower bound is not tight.

Finally, it is worth noting that the papers cited above considered source routing only. To the best of our knowledge, this paper is the first to consider online routing problems with hop-by-hop decisions. Such a problem can be formulated as a classical Markov decision process (MDP), in which the states are the packet locations and the actions are the outgoing links of each node. However, most studies consider MDP problems under stricter assumptions than ours and/or targeted different performance measures. Burnetas and Katehakis [27] derive the asymptotic lower bound on the regret and propose an optimal index policy. Their result can be applied only to the so-called ergodic MDPs [28], where the induced Markov chain by any policy is irreducible and consists of a single recurrent class. In hop-by-hop routing, however, the policy that routes packets on a fixed path results in a Markov chain with reducible states that are not in the chosen path. Jaksch *et al.* [29] and Filippi *et al.* [30] study a bigger class of MDPs and present algorithms with finite-time regret upper bounds scaling as $\mathcal{O}(\log(T))$. Nevertheless, these algorithms perform badly when applied to hop-by-hop routing due to loose confidence intervals. Jaksch *et al.* [29] also presents a nonasymptotic, but problem independent (minimax) regret lower bound scaling as $\Omega(\sqrt{T})$. This latter bound does not contradict our problem-dependent lower bounds that grow logarithmically.

III. Online Shortest Path Routing Problems

A. Network Model

The network is modeled as a directed graph $G = (V, E)$, where V is the set of nodes and E is the set of links. Each link $i \in E$ may, for example, represent an unreliable wireless link. Without loss of generality, we assume that time is slotted and that one slot corresponds to the time to send a packet over a single

[4] A policy π is order-optimal in terms of $|E|$ and H, if it satisfies the following: For all problem instances, $R^{\pi}(N) = \mathcal{O}(C_1 g(|E|, H)\log(N))$ with C_1 independent of $|E|$, H, and N, and there exists a problem instance and a constant $C_2 > 0$, independent of $|E|$, H, and N, such that $\liminf_{N\to\infty} R^{\pi'}(N)/\log(N) \geq C_2 g(|E|, H)$ for any uniformly good policy π'.

[5] A barycentric spanner is a set of paths from which the delay of other paths can be computed as its linear combination with coefficients in $[-1, 1]$ [2].

link. Let $X_i(t)$ be a binary random variable indicating whether a transmission on link i at time t is successful. $(X_i(t))_{t\geq 1}$ is a sequence of independent identically distributed Bernoulli variables with initially unknown mean θ_i. Hence, if a packet is sent on link i repeatedly until the transmission is successful, the time to complete the transmission (referred to as the delay on link i) is geometrically distributed with mean $1/\theta_i$. Let $\theta = (\theta_i, i \in E)$ be the vector representing the packet successful transmission probabilities on the various links. We consider a single source–destination pair $(s, d) \in V^2$, and denote by $\mathcal{P} \subseteq \{0, 1\}^{|E|}$ the set of loop-free paths from s to d in G, where each path $p \in \mathcal{P}$ is a $|E|$-dimensional binary vector; for any $i \in E$, $p_i = 1$ if and only if i belongs to p. Let H denote the maximum length of the paths in $\mathcal{P}$, i.e., $H = \max_{p\in\mathcal{P}} \sum_{i\in E} p_i$. For brevity, in what follows, for any binary vector z, we write $i \in z$ to denote $z_i = 1$. Moreover, for any vector z, we use the convention that $z^{-1} = (z_i^{-1})_i$.

For any path p, $D_\theta(p) = \sum_{i\in p} \frac{1}{\theta_i}$ is the average packet delay through path p given link success rates θ. The path with minimal delay is: $p^\star \in \arg\min_{p\in\mathcal{P}} D_\theta(p)$. Moreover, for any path $p \in \mathcal{P}$, we define $\Delta_p = D_\theta(p) - D_\theta(p^\star) = (p - p^\star)^\top \theta^{-1}$. Let $\Delta_{\min} = \min_{\Delta_p \neq 0} \Delta_p$. We let $\theta_{\min} = \min_{i\in E} \theta_i$ and assume that $\theta_{\min} > 0$. Finally, define $D^\star = D_\theta(p^\star)$ and $D^+ = \max_{p\in\mathcal{P}} D_\theta(p)$ the delays of the shortest and longest paths, respectively.

The analysis presented in this paper can be easily extended to more general link models, provided that the (single-link) delay distributions are taken within one-parameter exponential families of distributions.

B. Online Routing Policies and Feedback

We assume that the source is fully backlogged (i.e., it always has packets to send) and that the parameter θ is initially unknown. Packets are sent successively from s to d over various paths, and the outcome of each packet transmission is used to estimate θ, and in turn to learn the path $p^\star$ with the minimum average delay. After a packet is sent, we assume that the source gathers feedback from the network (essentially per-link or end-to-end delays) before sending the next packet.

Our objective is to design and analyze online routing policies, i.e., policies that take routing decisions based on the feedback received for the packets previously sent. We consider and compare three different types of online routing policies depending 1) on where routing decisions are taken (at the source or at each node), and 2) on the received feedback (per-link or end-to-end path delay).

1) Policy set Π_1: The path used by a packet is determined at the source based on the observed end-to-end delays for previous packets. More precisely, for the nth packet, let $p^\pi(n)$ be the path selected under policy π, and let $D^\pi(n)$ denote the corresponding end-to-end delay. Then, $p^\pi(n)$ depends on $p^\pi(1), \ldots, p^\pi(n-1), D^\pi(1), \ldots, D^\pi(n-1)$.
2) Policy set Π_2: The path used by a packet is determined at the source based on the observed per-link delays for previous packets. In other words, under policy π, $p^\pi(n)$ depends on $p^\pi(1), \ldots, p^\pi(n-1), (d_i^\pi(1), i \in p^\pi(1)), \ldots, (d_i^\pi(n-1), i \in p^\pi(n-1))$, where $d_i^\pi(k)$ is the delay experienced on link i for the kth packet (if this packet uses link i at all).
3) Policy set Π_3: Routing decisions are taken at each node in an adaptive manner. At a given time t, the packet is sent over a link selected based on all successes and failures observed on the various links before time t.

In the case of source-routing policies (in $\Pi_1 \cup \Pi_2$), if a transmission on a given link fails, the packet is retransmitted on the same link until it is successfully received (per-link delays are geometric random variables). On the contrary, in the case of hop-by-hop routing policies (in Π_3), the routing decisions at a given node can be adapted to the observed failures on a given link. For example, if transmission attempts on a given link failed, one may well decide to switch link and select a different next-hop node.

C. Performance Metrics and Objectives

1) Regret: Under any reasonably smart routing policy, the parameter θ will eventually be estimated accurately and the minimum delay path will be discovered with high probability after sending a large number of packets. Hence, to quantify the performance of a routing policy, we examine its transient behavior. More precisely, we use the notion of *regret*, a performance metric often used in the MAB literature [14]. The regret $R^\pi(N)$ of policy π up to the Nth packet is the expected difference of delays for the first N packets under π and under the policy that always selects the optimal path $p^\star$ for transmission:

$$R^\pi(N) := \mathbb{E}\left[\sum_{n=1}^{N} D^\pi(n)\right] - N D_\theta(p^\star)$$

where $D^\pi(n)$ denotes the end-to-end delay of the nth packet under policy π and the expectation $\mathbb{E}[\cdot]$ is taken with respect to the random transmission outcomes and possible randomization in the policy π. The regret quantifies the performance loss due to the need to explore suboptimal paths to learn the path with the minimum delay.

2) Objectives: The goal is to design online routing policies in Π_1, Π_2, and Π_3 that minimize regret over the first N packets. As it turns out, there are policies in any Π_j, $j = 1, 2, 3$, whose regrets scale as $\mathcal{O}(\log(N))$ when N grows large, and no policy can have a regret scaling as $o(\log(N))$.

Our objective is to derive, for each $j = 1, 2, 3$, an asymptotic regret lower bound $c_j(\theta)\log(N)$ for policies in Π_j, and then propose simple policies whose regret upper bounds asymptotically approach that of the *optimal* algorithm, i.e., an algorithm whose regret matches the lower bound in Π_j. As we shall discuss later, there exists an algorithm whose regret asymptotically matches these lower bound. Therefore, by comparing $c_1(\theta)$, $c_2(\theta)$, and $c_3(\theta)$, we can quantify the potential performance improvements taking routing decisions at each hop rather than at the source only, and observing per-link delays rather than end-to-end delays.

IV. Fundamental Performance Limits

In this section, we provide fundamental performance limits satisfied by *any* online routing policy in Π_1, Π_2, or Π_3. Specifically, we derive asymptotic (when N grows large) regret lower bounds for our three types of policies. These bounds are obtained exploiting some results and techniques used in the control of Markov chains [31], and they are *tight* in the sense that there exist algorithms achieving these performance limits.

A. Regret Lower Bounds

We restrict our attention to the so-called *uniformly good* policies, under which the number of times suboptimal paths are selected until the transmission of the nth packet is $o(n^\alpha)$ when $n \to \infty$ for any $\alpha > 0$ and for all θ. We know from [31, Th. 2] that such policies exist.

1) Source Routing With Bandit Feedback: Denote by $\psi_\theta^p(k)$ the probability that the delay of a packet sent on path p is k slots, and by $h(p)$ the length (or number of links) of path p. The end-to-end delay is the sum of several independent random geometric variables. If we assume that $\theta_i \neq \theta_j$ for $i \neq j$, according to [32], we have for all $k \geq h(p)$

$$\psi_\theta^p(k) = \sum_{i \in p} \left(\prod_{j \in p, j \neq i} \frac{\theta_j}{\theta_j - \theta_i} \right) \theta_i (1-\theta_i)^{k-1}$$

i.e., the path delay distribution is a weighted average of the individual link delay distributions, where the weights can be negative but always sum to one.

The next theorem provides the fundamental performance limit of online routing policies in Π_1.

Theorem 4.1: For all θ and for any uniformly good policy $\pi \in \Pi_1$, $\liminf_{N\to\infty} \frac{R^\pi(N)}{\log(N)} \geq c_1(\theta)$, where $c_1(\theta)$ is the infimum of the following optimization problem:

$$\inf_{x \geq 0} \quad \sum_{p \in \mathcal{P}} x_p \Delta_p$$

$$\text{subject to:} \quad \inf_{\lambda \in B_1(\theta)} \sum_{p \neq p^\star} x_p \sum_{k=h(p)}^{\infty} \psi_\theta^p(k) \log \frac{\psi_\theta^p(k)}{\psi_\lambda^p(k)} \geq 1 \qquad (1)$$

with

$$B_1(\theta) = \left\{ \lambda : \{\lambda_i, i \in p^\star\} = \{\theta_i, i \in p^\star\}, \ \min_{p \in \mathcal{P}} D_\lambda(p) < D_\lambda(p^\star) \right\}.$$

The variables x_p, $p \in \mathcal{P}$ solving (1) have the following interpretation: For $p \neq p^\star$, $x_p \log(N)$ is the asymptotic number of packets that needs to be sent (up to the Nth packet) on suboptimal path p under optimal routing strategies in Π_1. Hence, x_p determines the optimal rate of *exploration* of suboptimal path p. $B_1(\theta)$ is the set of *bad* network parameters: If $\lambda \in B_1(\theta)$, then the end-to-end delay distribution along the optimal path $p^\star$ is the same under θ or λ (hence by observing the end-to-end delay on path $p^\star$, we cannot distinguish λ or θ), and $p^\star$ is not optimal under λ. It is important to observe that in the definition of $B_1(\theta)$, the equality $\{\lambda_i, i \in p^\star\} = \{\theta_i, i \in p^\star\}$ is a set equality, i.e., order does not matter (e.g., if $p^\star = \{1, 2\}$, the equality means that either $\lambda_1 = \theta_1, \lambda_2 = \theta_2$ or $\lambda_1 = \theta_2, \lambda_2 = \theta_1$).

2) Source Routing With Semibandit (Per-Link) Feedback: We now consider routing policies in Π_2 that make decisions at the source, but have information on the individual link delays. Let $\mathrm{KLG}(u, v)$ denote the Kullback–Leibler (KL) information number between two geometric distributions with parameters u and v as

$$\mathrm{KLG}(u, v) := \sum_{k \geq 1} u(1-u)^{k-1} \log \frac{u(1-u)^{k-1}}{v(1-v)^{k-1}}.$$

Theorem 4.2: For all θ and for any uniformly good policy $\pi \in \Pi_2$, $\liminf_{N\to\infty} \frac{R^\pi(N)}{\log(N)} \geq c_2(\theta)$, where $c_2(\theta)$ is the infimum of the following optimization problem:

$$\inf_{x \geq 0} \quad \sum_{p \in \mathcal{P}} x_p \Delta_p$$

$$\text{subject to:} \quad \inf_{\lambda \in B_2(\theta)} \sum_{p \neq p^\star} x_p \sum_{i \in p} \mathrm{KLG}(\theta_i, \lambda_i) \geq 1 \qquad (2)$$

with

$$B_2(\theta) = \left\{ \lambda : \lambda_i = \theta_i \quad \forall i \in p^\star, \ \min_{p \in \mathcal{P}} D_\lambda(p) < D_\lambda(p^\star) \right\}.$$

The variables x_p, $p \in \mathcal{P}$ solving (2) have the same interpretation as that given previously in the case of bandit feedback. Again, $B_2(\theta)$ is the set of parameters λ such that the distributions of link delays along the optimal path are the same under θ and λ, and $p^\star$ is not the optimal path under λ. The slight difference between the definitions of $B_1(\theta)$ and $B_2(\theta)$ comes from the difference of feedback (bandit versus semibandit). It is also noted that $B_2(\theta) \subset B_1(\theta)$. We stress that by [31, Th. 2], the asymptotic regret lower bounds of Theorems 4.1 and 4.2 are tight, namely, there exists policies that achieve these regret bounds.

Remark 4.1: Of course, we know that $c_1(\theta) \geq c_2(\theta)$, since the lower bounds we derive are tight and getting per-link delay feedback can be exploited to design smarter routing policies than those we can devise using end-to-end delay feedback (i.e., $\Pi_1 \subset \Pi_2$).

Remark 4.2: The asymptotic lower bound proposed in [5] has a similar expression to ours, but the set $B_2(\theta)$ is replaced by $B_2'(\theta) = \bigcup_{i \in E} \{\lambda : \lambda_j = \theta_j \quad \forall j \neq i, \min_{p \in \mathcal{P}} D_\lambda(p) < D_\lambda(p^\star)\}$. Note that $B_2'(\theta) \subset B_2(\theta)$, which implies that the lower bound derived in [5] is smaller than ours. In other words, we propose a regret lower bound that improves that in [5]. Furthermore, our bound is tight (it cannot be improved further).

The proof of Theorems 4.1 and 4.2 leverage techniques from [31] developed for the control of Markov chains, and are presented in Appendix A. Theorem 4.2 can be seen as a direct consequence of [31, Th. 1] (the problem can be easily mapped to a controlled Markov chain). In contrast, the proof of Theorem 4.1 requires a more clever mapping due to the different nature of feedback. To prove Theorem 4.1, we establish Lemma 2, a property for geometric random variables.

3) Hop-by-Hop Routing: Finally, we consider routing policies in Π_3. These policies are more involved to analyze as the routing choices may change at any intermediate node in the network, and they are also more complex to implement. Surprisingly, the next theorem states that the regret lower bound for hop-by-hop routing policies is the same as that derived for strategies in Π_2 (source routing with semibandit feedback). In other words, we cannot improve the performance by taking routing decisions at each hop.

Theorem 4.3: For all θ and for any uniformly good policy $\pi \in \Pi_3$, $\liminf_{N\to\infty} \frac{R^\pi(N)}{\log(N)} \geq c_3(\theta) = c_2(\theta)$.

The proof of Theorem 4.3 is more involved than those of previous theorems, since in the hop-by-hop case, the chosen path could change at intermediate nodes. To overcome this difficulty, we introduce another notion of regret corresponding to the achieved throughput (i.e., the number of packets successfully received by the destination per unit time), which we refer

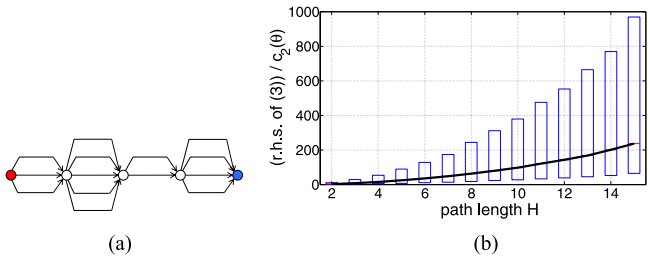


Fig. 1. (a) Line network, (b) semibandit versus bandit feedback: Boxplot for the ratio of the lower bound of $c_1(\theta)$ [the right-hand side of (3)] to $c_2(\theta)$ for a line network. The plot shows the median (black curve), 25% quantile, and 75% quantile.

to as the *throughput regret*. The proof uses the results presented in [31] for throughput regret, but also relies on Lemma 4, which provides an asymptotic relationship between $R^{\pi}(N)$ and the throughput regret.

Remark 4.3: Theorem 4.3, together with the tightness of our regret lower bounds, implies that in spite of exploiting additional feedback, hop-by-hop routing policies cannot yield better regret than source-routing policies at least asymptotically when the number of transmissions grows large. Hop-by-hop routing could provide a significant advantage if the link qualities change on a fast time scale. To achieve a low regret in this situation, an efficient algorithm should not try to retransmit many times over a link whose quality went from good to bad.

As shown in [31, Th. 2], the asymptotic regret lower bounds derived in Theorems 4.1–4.3 are *tight* in the sense that one can design actual routing policies achieving these regret bounds (although these policies might well be extremely complex to compute and impractical to implement). Hence, from the fact that $c_1(\theta) \geq c_2(\theta) = c_3(\theta)$, we conclude the following.

1) The optimal source-routing policy with semibandit feedback asymptotically achieves a lower regret than the optimal source-routing policy with bandit feedback.
2) The optimal hop-by-hop routing policy asymptotically obtains the same regret as the optimal source-routing policy with semibandit feedback.

B. Numerical Example

There are examples of network topologies where the above asymptotic regret lower bounds can be explicitly computed. One such example is the line network[6]; see Fig. 1(a) for an instance of line network. Notice that in line networks, the optimal routing policy consists in selecting the best link in each hop. The following lemma is immediate.

Lemma 1: For any line network with H hops, we have

$$c_1(\theta) \geq \sum_{i \notin p^\star} \frac{\frac{1}{\theta_i} - \frac{1}{\theta_{\zeta(i)}}}{\max_{p: i \in p} \sum_{k=H}^{\infty} \psi_\theta^p(k) \log \frac{\psi_\theta^p(k)}{\psi_{\vartheta^i}^p(k)}} \tag{3}$$

$$c_2(\theta) = c_3(\theta) = \sum_{i \notin p^\star} \frac{\frac{1}{\theta_i} - \frac{1}{\theta_{\zeta(i)}}}{\mathrm{KLG}(\theta_i, \theta_{\zeta(i)})}$$

[6]A line network is a graph of vertices $\{1, \ldots, n\}$, where the neighbors of i are $i-1$ and $i+1$.

where $\zeta(i)$ is the best link on the same hop as link i, and where ϑ^i is a vector of link parameters defined as $\vartheta^i_j = \theta_j$ if $j \neq i$ and $\vartheta^i_i = \theta_{\zeta(i)}$.

The proof of each statement of Lemma 1 involves decomposing the sets of bad parameters as $B(\theta) = \cup_{i \notin p^*} B^i(\theta)$, where $B^i(\theta)$ is the set of parameters $\lambda \in B(\theta)$ such that $\lambda_i > \theta_{\zeta(i)}$. For details, see [22, Appendix B].

Proposition 4.4: There exist problem instances in line networks with arbitrarily small $\theta_{\min}$, for which the regret of any uniformly good policy in $\Pi_2 \cup \Pi_3$ is $\Omega\Big(\frac{|E|-H}{\Delta_{\min}\theta^2_{\min}} \log(N)\Big)$.

For line networks, both $c_1(\theta)$ and $c_2(\theta)$ scale linearly with the number of links in the network. In Fig. 1(b), we present the median, along with 25% and 75% quantiles, for the ratio of the lower bound of $c_1(\theta)$ [i.e., the right-hand side of (3)] to $c_2(\theta)$ for various values of θ (we randomly generated 10^4 link parameters θ) as a function of the network size H in a simple line network, which has two links in the first hop and one link in the rest of hops, and hence, $|E| = H + 1$. These results suggest that collecting semibandit feedback (per-link delays) can significantly improve the performance of routing policies. The gain is significant even for fairly small networks.

V. Routing Policies for Semibandit Feedback

Theorems 4.1–4.3 indicate that within the first N packets, the total amount of packets routed on a suboptimal path p should be of the order of $x_p^\star \log(N)$, where $x_p^\star$ is the optimal solution of the optimization problems in (1) and (2). Graves and Lai [31] present policies that achieve the regret bounds of Theorems 4.1–4.3 (see [31, Th. 2]). These policies suffer from two problems: First, they are computationally infeasible for large problems since their implementation involves solving in each round a semi-infinite linear program [33] similar to those providing the regret lower bounds [defined in (1) and (2)]. Second, these policies have no finite-time performance guarantees, and numerical experiments suggest that their finite-time performance on typical problems is rather poor.

In this section, we present online routing policies for semibandit feedback, which are simple to implement, yet approach the performance limits identified in Section IV. We further analyze their regret and show that they outperform existing algorithms. To present our policies, we introduce additional notations. Under a given policy, we let $t_i(n)$ be the total number of transmission attempts (including retransmissions) on link i before the nth packet is sent. We define $\hat{\theta}_i(n)$ the empirical success rate of link i estimated over the transmissions of the first $(n-1)$ packets. Furthermore, we define the corresponding vectors $t(n) = (t_i(n))_{i \in E}$ and $\hat{\theta}(n) = (\hat{\theta}_i(n))_{i \in E}$.

Note that the proposed policies and regret analysis presented in this section directly apply for generic combinatorial optimization problems with linear objective function and geometrically distributed rewards.

A. Path and Link Indexes

The proposed policies rely on indexes attached either to individual links or paths. Next, we introduce three indexes used in our policies. They depend on the round, i.e., on the number n of the packet to be sent, and on the estimated link parameters $\hat{\theta}(n)$. The three indexes and their properties (i.e., in which policy they are used and how one can compute them) are summarized in

TABLE II
SUMMARY OF INDEXES

Index	Type	Computation	Algorithm
b_p	Path	Line search	GEOCOMBUCB-1
c_p	Path	Explicit	GEOCOMBUCB-2
ω_i	Link	Line search	KL-SR

Table II. Let $n \geq 1$ and assume that the nth packet is to be sent. The indexes are defined as follows.

1) Path Indexes: Let $\lambda \in (0,1]^{|E|}$, $t \in \mathbb{N}^{|E|}$, and $n \in \mathbb{N}$. The first path index, denoted by $b_p(n,\lambda,t)$ for path $p \in \mathcal{P}$, is motivated by the index defined in [18]. $b_p(n,\lambda,t)$ is defined as the infimum of the following optimization problem:

$$\inf_{u \in (0,1]^{|E|}} \quad p^\top u^{-1}$$

$$\text{subject to:} \quad \sum_{i \in p} t_i \mathrm{KL}(\lambda_i, u_i) \leq f_1(n),$$

$$u_i \geq \lambda_i \quad \forall i \in E$$

where $f_1(n) = \log(n) + 4H\log(\log(n))$, and for all $a, b \in [0,1]$, $\mathrm{KL}(a,b)$ is the KL information number between two Bernoulli distributions with respective means a and b, i.e., $\mathrm{KL}(a,b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$.

The second index is denoted by $c_p(n,\lambda,t)$ and defined for path $p \in \mathcal{P}$ as

$$c_p(n,\lambda,t) = p^\top \lambda^{-1} - \sqrt{\sum_{i \in p} \frac{2 f_1(n)}{t_i \lambda_i^3}}.$$

The next theorem provides generic properties of the two indexes b_p and c_p.

Theorem 5.1: (i) For all $n \geq 1$, $p \in \mathcal{P}$, $\lambda \in (0,1]^{|E|}$, and $t \in \mathbb{N}^{|E|}$, we have $b_p(n,\lambda,t) \geq c_p(n,\lambda,t)$.

(ii) There exists a constant $K_H > 0$ depending on H only such that for all $p \in \mathcal{P}$ and $n \geq 2$

$$\mathbb{P}[b_p(n,\hat{\theta}(n),t(n)) > p^\top \theta] \leq K_H n^{-1} (\log(n))^{-2}.$$

Corollary 5.2: We have

$$\sum_{n \geq 1} \mathbb{P}[b_{p^\star}(n,\hat{\theta}(n),t(n)) > p^{\star\top}\theta^{-1}]$$

$$\leq 1 + K_H \sum_{n \geq 2} n^{-1} (\log(n))^{-2} < \infty.$$

2) Link Index: Our third index is a link index. For $n, t \in \mathbb{N}$ and $\lambda \in (0,1]$, the index $\omega_i(n,\lambda,t)$ of link $i \in E$ is defined as

$$\omega_i(n,\lambda,t) = \min\left\{\frac{1}{u} : u \in [\lambda,1], \quad t\mathrm{KL}(\lambda,u) \leq f_2(n)\right\}$$

where $f_2(n) = \log(n) + 4\log(\log(n))$.

B. Routing Policies

We present three routing policies, referred to as GEOCOMBUCB-1, GEOCOMBUCB-2, and KL-SR, respectively. For the transmission of the nth packet, GEOCOMBUCB-1 (resp. GEOCOMBUCB-2) selects the path p with the smallest index $b_p(n) := b_p(n,\hat{\theta}(n),t(n))$ (resp. $c_p(n) := c_p(n,\hat{\theta}(n),t(n))$). KL-SR was initially proposed in [1] and for the transmission of the nth packet, it selects the path $p(n) \in \arg\min_{p \in \mathcal{P}} p^\top \omega(n)$, where $\omega(n) = (\omega_i(n), i \in E)$ and $\omega_i(n) := \omega_i(n,\hat{\theta}_i(n),t_i(n))$. The pseudocodes of GEOCOMBUCB and KL-SR are presented in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1: GEOCOMBUCB

for $n \geq 1$ **do**
Select path $p(n) \in \arg\min_{p \in \mathcal{P}} \xi_p(n)$ (ties are broken arbitrarily), where $\xi_p(n) = b_p(n)$ for GEOCOMBUCB-1, and $\xi_p(n) = c_p(n)$ for GEOCOMBUCB-2.
Collect feedback on links $i \in p(n)$, and update $\hat{\theta}_i(n)$ for $i \in p(n)$.

Algorithm 2: KL-SR

for $n \geq 1$ **do**
Select path $p(n) \in \arg\min_{p \in \mathcal{P}} p^\top \omega(n)$ (ties are broken arbitrarily).
Collect feedback on links $i \in p(n)$, and update $\hat{\theta}_i(n)$ for $i \in p(n)$.

In the following theorems, we provide a finite-time analysis of GEOCOMBUCB and KL-SR and show the optimality of KL-SR in line networks. Define $\varepsilon = (1 - 2^{-\frac{1}{4}})\frac{\Delta_{\min}}{D^+}$.

Theorem 5.3: For all $N \geq 1$, under policies $\pi \in \{$GEOCOMBUCB-1, GEOCOMBUCB-2$\}$, we have

$$R^\pi(N) \leq \frac{32|E|\sqrt{H} f_1(N)}{\Delta_{\min}\theta_{\min}^2} + 2D^+\left(2K_H + \sum_{i \in E} \frac{1}{\varepsilon^2 \theta_i^2}\right).$$

Hence, $R^\pi(N) = \mathcal{O}\left(\frac{|E|\sqrt{H}}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$ when $N \to \infty$.

Theorem 5.4: For all $N \geq 1$, under policy $\pi =$ KL-SR, we have

$$R^\pi(N) \leq \frac{360|E|H f_2(N)}{\Delta_{\min}\theta_{\min}^2} + 2D^+\left(4H + \sum_{i \in E} \frac{1}{\varepsilon^2 \theta_i^2}\right).$$

Hence, $R^\pi(N) = \mathcal{O}(\frac{|E|H}{\Delta_{\min}\theta_{\min}^2}\log(N))$ when $N \to \infty$.

Remark 5.1: The regret bound of KL-SR scales as $\mathcal{O}(H)$ while that of GEOCOMBUCB scales as $\mathcal{O}(\sqrt{H})$. Indeed, the index ω_i used in KL-SR ignores the statistical independence of delays of various links in a given path, whereas the indexes b_p and c_p in GEOCOMBUCB use this independence and yield smaller confidence intervals.

The index b_p is an extension of the KL-based index of [18] to the case of geometrically distributed rewards. However, the proof of Theorem 5.3 is novel and uses the link between b_p and c_p established in Theorem 5.1. The proof of Theorem 5.3 uses some ideas from [18]. The proof of Theorem 5.4 is completely different from the regret analysis of KL-SR in [1]; it relies on Lemma 8, which provides a sharp lower bound for the index ω_i, and borrows some ideas from [12, Th. 5].

Remark 5.2: Theorem 5.4 holds even when the delays on the various links are not independent, as in [12].

The proposed policies have better performance guarantees than existing routing algorithms. Indeed, as shown in [22, Appendix I], the best known regret upper bound for the CUCB

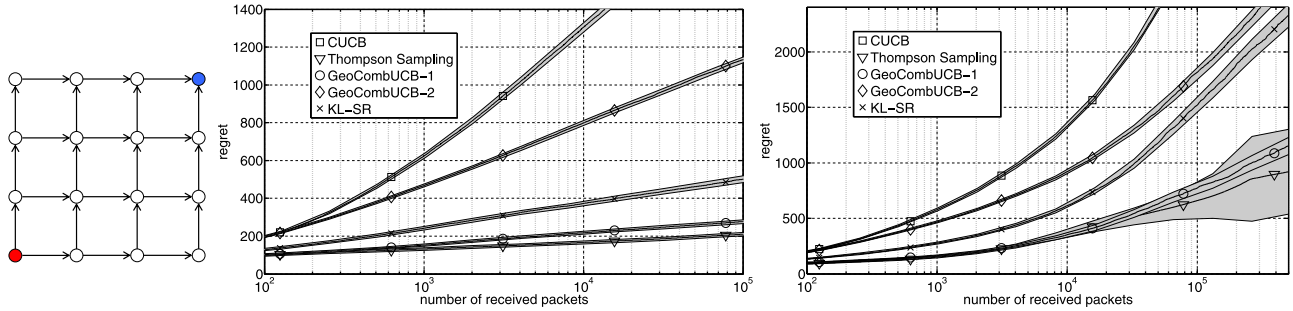


Fig. 2. Network topology, and regret versus number of received packets. (a) Grid network. (b) $\theta_{\min} = 0.18,\ \Delta_{\min} = 0.34$. (c) $\theta_{\min} = 0.1,\ \Delta_{\min} = 0.08$.

algorithm [10] is $\mathcal{O}(\frac{|E|H}{\Delta_{\min}\theta_{\min}^3}\log(N))$, which constitutes a weaker performance guarantee than those of our routing policies. The numerical experiments presented in Section V-D will confirm the superiority of GeoCombUCB and KL-SR over CUCB.

The next proposition states that KL-SR is asymptotically optimal in line networks (see [22, Appendix G] for details).

Proposition 5.5: In line networks, the regret under $\pi =$ KL-SR satisfies $\limsup_{N\to\infty} \dfrac{R^\pi(N)}{\log(N)} \leq c_2(\theta)$. Hence, $R^\pi(N) = \mathcal{O}\left(\frac{|E|-H}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$ when $N \to \infty$.

Remark 5.3: When the link parameters smoothly evolve over time, we can modify the proposed routing policies so that routing decisions are based on past choices and observations over a sliding window consisting of a fixed number of packets, as considered in [34] and [35].

C. Implementation

Next, we discuss the implementation of our routing policies and give simple methods to compute $b_p(n,\lambda,t)$, $c_p(n,\lambda,t)$, and $\omega_i(n,\lambda,t)$ given p, i, n, λ, and t. The path index c_p is explicit and easy to compute. The computation of link index ω_i is also straightforward as it amounts to finding the roots of a strictly convex and increasing function in one variable (note that $v \mapsto \mathrm{KL}(u,v)$ is strictly convex and increasing for $v \geq u$). Hence, the index ω_i can be computed by a simple line search. The path index $b_p(n,\lambda,t)$ can also be computed using a simple line search, as shown below.

Introduce $I_p(\lambda) = \{i \in p : \lambda_i \neq 1\}$, and for $\gamma > 0$, define

$$F(\gamma,\lambda,n,t) = \sum_{i\in I_p(\lambda)} t_i \mathrm{KL}(\lambda_i, g(\gamma,\lambda_i,t_i)) \text{ with}$$

$$g(\gamma,\lambda_i,t_i) = \frac{1}{2\gamma t_i}\Big(\gamma\lambda_i t_i - 1 + \sqrt{(1-\gamma\lambda_i t_i)^2 + 4\gamma t_i}\Big).$$

Proposition 5.6: 1) $\gamma \mapsto F(\gamma,\lambda,n,t)$ is strictly increasing, and $F(\mathbb{R}^+,\lambda,n,t) = \mathbb{R}^+$. 2) If $I_p(\lambda) = \emptyset$, $b_p(n,\lambda,t) = \sum_{i\in E} p_i$. Otherwise, let $\gamma^\star$ be the unique solution to $F(\gamma,\lambda,n,t) = f_1(n)$. Then,

$$b_p(n,\lambda,t) = \sum_{i\in E} p_i - |I_p(\lambda)| + \sum_{i\in I_p(\lambda)} g(\gamma^\star,\lambda_i,t_i).$$

As stated in Proposition 5.6, proven in [22, Appendix H], $\gamma^\star$ can be computed efficiently by a simple line search and b_p is easily deduced. We thus have efficient methods to compute the three indexes. To implement our policies, we then need to find in each round the path minimizing the index (or the sum of link indexes along the path for KL-SR). KL-SR can be implemented (in a distributed fashion) using the Bellman–Ford algorithm, and its complexity is $\mathcal{O}(|V||E|)$ in each round. GeoCombUCB-1 and GeoCombUCB-2 are more computationally involved than KL-SR and have complexity $\mathcal{O}(|\mathcal{P}|)$ in each round.

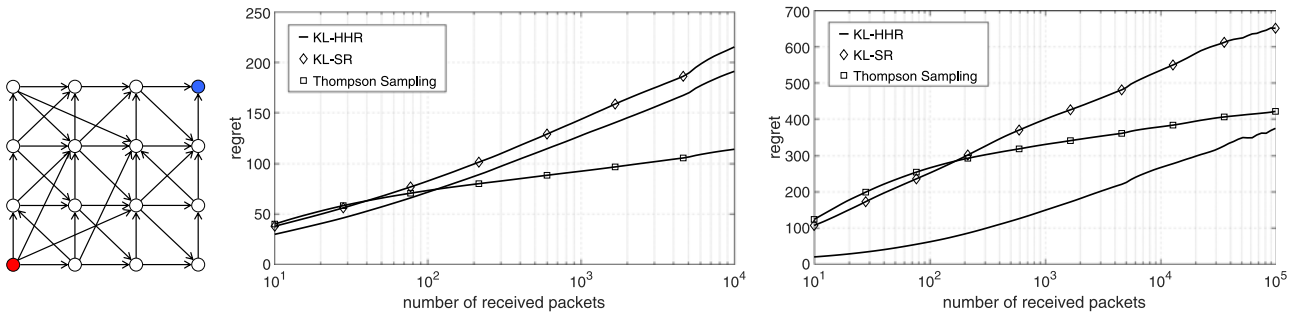
D. Numerical Experiments

In this section, we conduct numerical experiments to compare the performance of the proposed source-routing policies to that of the CUCB algorithm [10] and TS applied to our online routing problem. The CUCB algorithm is an index policy in Π_2 (the set of source-routing policies with semibandit feedback) that selects path $p(n)$ for the transmission of the nth packet:

$$p(n) \in \arg\min_{p\in\mathcal{P}} \sum_{i\in p} \frac{1}{\hat{\theta}_i(n) + \sqrt{1.5\log(n)/t_i(n)}}.$$

We consider a grid network whose topology is depicted in Fig. 2(a), where the node in red (resp. blue) is the source (resp. the destination). In this network, there are $\binom{6}{3} = 20$ possible paths from the source to the destination. Let us compare these algorithms in terms of their per-packet complexity. The complexity of GeoCombUCB-1 and GeoCombUCB-2 is $\mathcal{O}(|\mathcal{P}|)$, whereas that of KL-SR, CUCB, and TS is $\mathcal{O}(|V||E|)$.

In Fig. 2(b) and (c), we plot the regret against the number of the packets N under the various routing policies, and for two sets of link parameters θ. For each set, we choose a value of $\theta_{\min}$ and generate the values of θ_i independently, uniformly at random in $[\theta_{\min}, 1]$. The results are averaged over 100 independent runs, and the 95% confidence intervals are shown using the gray area around curves. The three proposed policies outperform CUCB, and GeoCombUCB-1 attains the smallest regret amongst the proposed policies. The comparison between GeoCombUCB-2 and KL-SR is more subtle and depends on the link parameters: While in Fig. 2(b) KL-SR significantly outperforms GeoCombUCB-2, they attain regrets growing similarly for the link parameter of Fig. 2(c). Yet there are some parameters for which KL-SR is significantly outperformed by GeoCombUCB-2. KL-SR seems to perform better than GeoCombUCB-2 in scenarios where $\Delta_{\min}$ is large. TS performs slightly better than GeoCombUCB-1 on average. Its regret, however, may not be well concentrated around the mean



(a) (b) (c)

Fig. 3. Network topology, and regret versus number of received packets. (a) Topology. (b) $\theta_{\min} = 0.014$. (c) $\theta_{\min} = 0.0056$.

Algorithm 3: KL-HHR for node v

for $\tau \geq 1$ **do**

Select link $(v, v') \in E$, where

$$v' \in \arg \min_{w \in V:(v,w)\in E} \left(\omega_{(v,w)}(\tau, \tilde{\theta}_v(\tau)) + J_w(\tau)\right).$$

Update index of the link (v, v').

for some link parameters, as in Fig. 2(c). Furthermore, the regret analysis of TS for shortest path routing with general topologies is an open problem.

E. Distributed Hop-by-Hop Routing Policy

Next, we propose KL-HHR, a distributed Bellman–Ford implementation of the KL-SR algorithm that belongs to the set of policies Π_3. For any node $v \in V$, we let $\mathcal{P}_v$ denote the set of loop-free paths from node v to the destination. For any time slot τ, we denote by $n(\tau)$ the packet number that is about to be sent or already in the network. For any link i, let $\tilde{\theta}_i(\tau)$ be the empirical success rate of link i *up to time slot* τ, that is $\tilde{\theta}_i(\tau) = s_i(n(\tau))/t'_i(\tau)$, where $t'_i(\tau)$ denotes the total number of transmission attempts on link i up to time slot τ. Moreover, with slight abuse of notation, we denote the index of link i at time τ by $\omega_i(\tau, \tilde{\theta}_i(\tau))$.[7] We define $J_v(\tau)$ as the minimum *cumulative index* from node v to the destination

$$J_v(\tau) = \min_{p\in\mathcal{P}_v} \sum_{i\in p} \omega_i(\tau, \tilde{\theta}_i(\tau)).$$

The index $J_v(\tau)$ can be computed using the Bellman–Ford algorithm. KL-HHR (see Algorithm 3) works based on the following idea: Assuming that the current packet is at node v at time τ, send it to node v' such that $\omega_{(v,v')}(\tau, \tilde{\theta}_v(\tau)) + J_{v'}(\tau)$ is minimal over all outgoing links of node v.

We compare the performance of KL-HHR, KL-SR, and TS through numerical experiments for a network shown in Fig. 3(a), in which there are 40 links and 413 possible paths between the source (in red) and the destination (in blue). Fig. 3(b) and (c) display the regret under KL-HHR, KL-SR, and TS, averaged over 100 independent runs, against the number of the packets N for two sets of link parameters θ. These parameters are generated similarly to the previous experiments. As expected, KL-HHR outperforms KL-SR in both scenarios since it can change routing decisions dynamically at intermediate nodes thus avoiding retransmissions on bad links when they are discovered. Note however that, in both scenarios, the regret of both KL-HHR and KL-SR seems to grow at the same rate as the number of received packets grows large. Moreover, TS would outperform KL-HHR asymptotically, i.e., as the number of received packets grows large. On the other hand, in scenarios where $\theta_{\min}$ is very small, if the number of total packets N is not very large, KL-HHR could outperform TS; see, e.g., Fig. 3(c).

The regret analysis of KL-HHR is beyond the scope of this paper, and is left for future work.

[7]Note that by definition $t'_i(\tau) \geq t_i(n)$ and $\tilde{\theta}_i(\tau)$ is a more accurate estimate of θ_i than $\hat{\theta}_i(n(\tau))$.

VI. Conclusions and Future Work

We have studied online shortest path routing problems in networks with stochastic link delays. Three types of routing policies are analyzed: source routing with semibandit feedback, source-routing with bandit feedback, and hop-by-hop routing. Tight asymptotic lower bounds on the regret for the three types of policies are derived. By comparing these bounds, we observed that semibandit feedback significantly improves performance while hop-by-hop decisions do not. Finally, we proposed several simple routing policies for semibandit feedback that exhibit better regret upper bounds than existing algorithms. As a future work, we plan to propose practical algorithms with provable performance bounds for hop-by-hop routing and source routing with bandit feedback. Furthermore, we would like to study the effect of delayed feedback on the performance as studied in, e.g., [36].

Appendix A
Proofs of Theorems 4.1, 4.2, and 4.3

To derive the asymptotic regret lower bounds, we apply the techniques used by Graves and Lai [31] to investigate efficient adaptive decision rules in controlled Markov chains. We recall here their general framework. Consider a controlled Markov chain $(X_t)_{t\geq 0}$ on a countable state space $\mathcal{S}$ with a control set U. The transition probabilities given control $u \in U$ are parameterized by θ taking values in a compact metric space Θ: The probability to move from state x to state y given the control u and the parameter θ is $P(x, y; u, \theta)$. The parameter θ is not known. The decision maker is provided with a finite set of

stationary control laws $G = \{g_1, \ldots, g_K\}$, where each control law g_j is a mapping from $\mathcal{S}$ to U: When control law g_j is applied in state x, the applied control is $u = g_j(x)$. It is assumed that if the decision maker always selects the same control law g, the Markov chain is irreducible with respect to some maximum irreducibility measure and has stationary distribution π_θ^g. The reward obtained when applying control u in state x is denoted by $r(x, u)$, so that the expected reward achieved under control law g is $\mu_\theta(g) = \sum_x r(x, g(x))\pi_\theta^g(x)$. There is an optimal control law given θ whose expected reward is denoted by $\mu_\theta^\star = \max_{g\in G} \mu_\theta(g)$. Now the objective of the decision maker is to sequentially apply control laws so as to maximize the expected reward up to a given time horizon N. The performance of the decision making scheme can be quantified through the notion of regret, which compares the expected accumulated reward to that obtained by always applying the optimal control law.

A. Source Routing With Bandit Feedback—Theorem 4.1

To prove Theorem 4.1, we construct a controlled Markov chain as follows. The state space is $\mathbb{N}$, the control set is the set of paths $\mathcal{P}$, and the parameter $\theta = (\theta_i, i \in E)$ defines the success rates on the various links. The parameter θ takes value in the compact space $\Theta = [\varepsilon, 1]^{|E|}$ for ε arbitrarily close to zero. Control laws are stationary and each of them corresponds to a given path, i.e., $G = \mathcal{P}$. A transition in the Markov chain occurs at time epochs where a new packet is sent. The state after a transition records the end-to-end delay of the packet. Hence, the transition probabilities are $P(k, l; p, \theta) = \psi_\theta^p(l)$, and do not depend on the starting state. The cost (the opposite of reward) at state l is simply equal to the delay l. For any two sets of parameters θ and λ, we define the KL information number under path (or control law) p as

$$I^p(\theta, \lambda) = \sum_{l=h(p)}^{\infty} \psi_\theta^p(l) \log \frac{\psi_\theta^p(l)}{\psi_\lambda^p(l)}. \tag{4}$$

We have that $I^p(\theta, \lambda) = 0$ if and only if the delays over path p under parameters θ and λ have the same distributions. By Lemma 2, proven at the end of this section, this occurs if and only if the two following sets are identical: $\{\theta_i, i \in p\}$, $\{\lambda_i, i \in p\}$. Let us fix θ and denote by $p^\star$ the corresponding optimal path. We further define $B_1(\theta)$ as the set of bad parameters λ such that under λ, $p^\star$ is not the optimal path, and such that θ and λ are statistically not distinguishable (they lead to the same delay distributions along path $p^\star$). Then

$$B_1(\theta) = \Big\{\lambda : \{\lambda_i, i \in p^\star\} = \{\theta_i, i \in p^\star\}, \min_{p\in\mathcal{P}} D_\lambda(p) < D_\lambda(p^\star)\Big\}.$$

By [31, Th. 1], we conclude that the delay regret scales at least as $c_1(\theta)\log(N)$, where

$$c_1(\theta) = \inf\Big\{\sum_{p\in\mathcal{P}} x_p\Delta_p : x \geq 0, \ \inf_{\lambda\in B_1(\theta)} \sum_{p\neq p^\star} x_p I^p(\theta, \lambda) \geq 1\Big\}$$

where $I^p(\theta, \lambda)$ is given in (4). ■

Lemma 2: Consider $(X_i)_i$ independent with $X_i \sim \text{Geo}(\theta_i)$ and $\theta_i \in (0, 1]$. Consider $(Y_i)_i$ independent with $Y_i \sim \text{Geo}(\lambda_i)$ and $\lambda_i \in (0, 1]$. Define $\overline{X} = \sum_i X_i$ and $\overline{Y} = \sum_i Y_i$. Then, $\overline{X} \overset{d}{=} \overline{Y}$ if and only if $(\theta_i)_i = (\lambda_i)_i$ up to a permutation.[8]

Proof: If $(\theta_i)_i = (\lambda_i)_i$ up to a permutation, then $X \overset{d}{=} Y$ by inspection. Assume that $X \overset{d}{=} Y$. Define $z_m = \min_i(\min(1/(1-\theta_i), 1/(1-\lambda_i))$. For all z such that $|z| < z_m$, we have $\mathbb{E}[z^{\overline{X}}] = \mathbb{E}[z^{\overline{Y}}]$ so that

$$\prod_i \frac{\theta_i}{1-(1-\theta_i)z} = \prod_i \frac{\lambda_i}{1-(1-\lambda_i)z}.$$

Hence

$$\begin{aligned} P_X(z) &:= \prod_i \theta_i(1-(1-\lambda_i)z) \\ &= \prod_i \lambda_i(1-(1-\theta_i)z) := P_Y(z). \end{aligned}$$

Both $P_X(z)$ and $P_X(z)$ are polynomials and are equal on an open set. So they are equal everywhere, and the sets of their roots are equal: $\{1/(1-\theta_i), i\} = \{1/(1-\lambda_i), i\}$. Thus, $(\theta_i)_i = (\lambda_i)_i$ up to a permutation as announced. ■

B. Source Routing With Semibandit Feedback—Theorem 4.2

The proof of Theorem 4.2 is similar to that of Theorem 4.1, except that here we have to account for the fact that the source gets feedback on per-link basis. To this end, we construct a Markov chain that records the delay on each link of a path. The state space is $\mathbb{N}^{|E|}$. Transitions occur when a new packet is sent from the source, and the corresponding state records the observed delays on each link of the chosen path, and the components of the state corresponding to links not involved in the path are set equal to 0. For example, the state (0, 1, 4, 0, 7) indicates that the path consisting of links 2, 3, and 5 has been used, and that the per-links delays are 1, 4, and 7, respectively. The cost of a given state is equal to the sum of its components (total delay). Now assume that path $p = (i_1, \ldots, i_{h(p)})$ is used to send a packet, then the transition probability to a state whose i_kth component is equal to d_k, $k = 1, \ldots, h(p)$ (the other components are 0) is $\prod_{k=1}^{h(p)} q_\theta(i_k, d_k)$, where $q_\theta(i, m) = \theta_i(1-\theta_i)^{m-1}$ for any link i and any delay m. Now the KL information number of (θ, λ) under path p is given by

$$I^p(\theta, \lambda) = \sum_{i\in p} \text{KLG}(\theta_i, \lambda_i) \tag{5}$$

since KL information number is additive for independent random variables. Hence, under semibandit feedback, we have $I^p(\theta, \lambda) = 0$ if and only if $\theta_i = \lambda_i$ for all $i \in p$. The set $B_2(\theta)$ of bad parameters is defined as

$$B_2(\theta) = \Big\{\lambda : \lambda_i = \theta_i \quad \forall i \in p^\star, \ \min_{p\in\mathcal{P}} D_\lambda(p) < D_\lambda(p^\star)\Big\}.$$

[8]The symbol $\overset{d}{=}$ denotes equality in distribution.

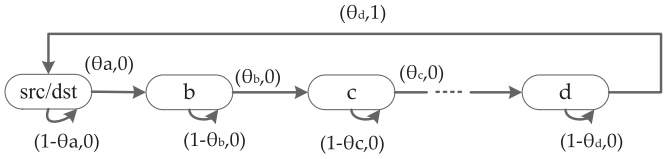


Fig. 4. Markov chain example under a control law p where the values in the parenthesis respectively denote the transition probability and the reward.

Applying [31, Th. 1] gives

$$c_2(\theta) = \inf\left\{\sum_{p\in\mathcal{P}} x_p\Delta_p : x \geq 0,\ \inf_{\lambda\in B_2(\theta)} \sum_{p\neq p^\star} x_p I^p(\theta,\lambda) \geq 1\right\}$$

where $I^p(\theta,\lambda)$ is given in (5). ■

C. Hop-by-Hop Routing—Theorem 4.3

This case is more involved. We first define another notion of regret corresponding to the achieved throughput (i.e., the number of packets successfully received by the destination per unit time). The throughput regret is introduced to ease the analysis since computing the throughput regret is easier in the hop-by-hop case. Define $\mu_\theta(p)$ as the average throughput on path p given link success rates θ: $\mu_\theta(p) = 1/D_\theta(p)$. The *throughput regret* $S^\pi(T)$ of π over time horizon T is: $S^\pi(T) := T\mu_\theta(p^\star) - \mathbb{E}\left[N^\pi(T)\right]$, where $N^\pi(T)$ is the number of packets received up to time T under policy π. Lemma 4, stated at the end of the proof, provides the relation between the asymptotic bound on $R^\pi(N)$ and $S^\pi(T)$.

Now we are ready to prove Theorem 4.3. We let the state of the Markov chain be the packet location. The action is the selected outgoing link. The transition between two states takes one time slot—the time to make a transmission attempt. Hence, the transition probability between state x and y with the action of using link i is denoted by (where $y \neq x$) $P_\theta^i(x,y) = \theta_i$ if link i connects node x and y, and is zero otherwise. On the other hand, the probability of staying at the same state is the transmission failure probability on link i if link i is an outgoing link, that is $P_\theta^i(x,x) = 1-\theta_i$ if link i is an outgoing link, and is zero otherwise.

We assume that the packet is injected at the source immediately after the previous packet is successfully delivered, and we are interested in counting the number of successfully delivered packets. In order not to count the extra time slot we will spend at the destination, we use a single Markov chain state to represent both the source and the destination.

We give a reward of 1 whenever the packet is successfully delivered to the destination. Let $r(x,y,i)$ be the immediate reward after the transition from node x to node y under the action i, i.e., $r(x,y,i) = 1$ if y is the destination node and is zero otherwise (see Fig. 4 for an example). Hence, $r(x,i)$ (i.e., the reward at state x with action i) is

$$r(x,i) = \begin{cases} \theta_i, & \text{if link } i \text{ connects node } x \text{ and the destination} \\ 0, & \text{otherwise.} \end{cases}$$

The stationary control law prescribes the action at each state, i.e., the outgoing link at each node. A stationary control law of this Markov chain is then a path p in the network, and we assign arbitrary actions to the nodes that are not on the path p. The maximal irreducibility measure is then to assign measure zero to the nodes that are not on the path p, and a counting measure to the nodes on the path p. The Markov chain is irreducible with respect to this maximal irreducibility measure, and the stationary distribution of the Markov chain under path p is

$$\pi_\theta^p(x) = \frac{\frac{1}{\theta_{p(x)}}}{\sum_{i\in p}\frac{1}{\theta_i}}\mathbb{1}\{\text{if node } x \text{ is on the path } p\}$$

where $p(x)$ denotes the link we choose at node x. The long-run average reward of the Markov chain under control law p is $\sum_x \pi_\theta^p(x) r(x,p(x)) = 1/\sum_{i\in p}\frac{1}{\theta_i} = \mu_\theta(p)$. The optimal control law is then $p^\star$ with long run average reward $\mu_\theta(p^\star)$.

The throughput regret of a policy $\pi \in \Pi_3$ for this controlled Markov chain at time T is

$$S^\pi(T) = T\mu_\theta(p^\star) - \mathbb{E}_\theta\left[\sum_{t=1}^{T} r(x_t, \pi(t,x_t))\right] \tag{6}$$

where x_t is the state at time t and $\pi(t,x_t)$ is the corresponding action for state x_t at time t. To this end, we construct a controlled Markov chain that corresponds to the hop-by-hop routing in the network. Now define $I^p(\theta,\lambda)$ as the KL information number for a control law p

$$\begin{aligned} I^p(\theta,\lambda) &= \sum_x \pi_\theta^p(x)\sum_y P_\theta^{p(x)}(x,y)\log\frac{P_\theta^{p(x)}(x,y)}{P_\lambda^{p(x)}(x,y)} \\ &= \sum_x \pi_\theta^p(x)\left(\theta_{p(x)}\log\frac{\theta_{p(x)}}{\lambda_{p(x)}} + (1-\theta_{p(x)})\log\frac{1-\theta_{p(x)}}{1-\lambda_{p(x)}}\right) \\ &= \mu_\theta(p)\sum_{i\in p}\frac{\mathrm{KL}(\theta_i,\lambda_i)}{\theta_i} = \mu_\theta(p)\sum_{i\in p}\mathrm{KLG}(\theta_i,\lambda_i) \end{aligned}$$

where we used Lemma 3 in the last equality. Since $I^p(\theta,\lambda) = 0$ if and only if $\theta_i = \lambda_i$ for all $i \in p$, the set $B_2(\theta)$ of bad parameters is

$$\begin{aligned} B_2(\theta) &= \left\{\lambda : \lambda_i = \theta_i \quad \forall i \in p^\star,\ \max_{p\in\mathcal{P}} \mu_\lambda(p) > \mu_\lambda(p^\star)\right\} \\ &= \left\{\lambda : \lambda_i = \theta_i \quad \forall i \in p^\star,\ \min_{p\in\mathcal{P}} D_\lambda(p) < D_\lambda(p^\star)\right\}. \end{aligned}$$

Applying [31, Th. 1], we get

$$\liminf_{T\to\infty} S^\pi(T)/\log(T) \geq c_3'(\theta)$$

with

$$c_3'(\theta) = \inf\left\{\sum_{p\in\mathcal{P}} x_p(\mu_\theta(p^\star) - \mu_\theta(p)) : x \geq 0,\right.$$
$$\left.\inf_{\lambda\in B_2(\theta)}\sum_{p\neq p^\star} x_p\mu_\theta(p)\sum_{i\in p}\mathrm{KLG}(\theta_i,\lambda_i) \geq 1\right\}.$$

By Lemma 4, $c_3(\theta) \geq c_3'(\theta)/\mu_\theta(p^\star)$. Finally, observe that $\mu_\theta(p^\star) - \mu_\theta(p) = \mu_\theta(p^\star)\mu_\theta(p)(D_\theta(p) - D_\theta(p^\star))$. It then

follows that $c_3'(\theta)/\mu_\theta(p^\star) = c_2(\theta)$, and therefore, $c_3(\theta) \geq c_2(\theta)$. On the other hand, $c_3(\theta) \leq c_2(\theta)$ since $\Pi_2 \subset \Pi_3$. As a result, $c_3(\theta) = c_2(\theta)$ and the proof is completed. ■

The following two lemmas prove useful in the proof of Theorem 4.3. Lemma 3, which follows from a straightforward calculation, relates the KL information number between two geometric distributions to that of corresponding Bernoulli distributions. Lemma 4 provides the connection between the throughput regret $S^\pi(T)$ and delay regret $R^\pi(N)$.

Lemma 3: For any $u, v \in (0, 1]$, we have

$$\mathrm{KLG}(u, v) = \frac{\mathrm{KL}(u, v)}{u}.$$

Lemma 4: For any $\pi \in \Pi_i$, $i = 1, 2, 3$, and any $\beta > 0$, we have

$$\liminf_{T\to\infty} \frac{S^\pi(T)}{\log(T)} \geq \beta \implies \mu_\theta(p^\star) \liminf_{N\to\infty} \frac{R^\pi(N)}{\log(N)} \geq \beta.$$

Proof: Define $\mu^\star = \mu_\theta(p^\star)$ and $r_t = \sum_{n=1}^t (D^\pi(n) - D^\star)$. Define $\mathcal{F}_t$ the σ-algebra generated by $(p^\pi(n), (d_i^\pi(n), i \in p^\pi(n)))_{1\leq n\leq t}$, where $d_i^\pi(k)$ is the delay experienced on link i for the kth packet under policy π. Then $p^\pi(t)$ is $\mathcal{F}_{t-1}$ measurable and $\mathbb{E}[r_t - r_{t-1}|\mathcal{F}_{t-1}]$ equals

$$\mathbb{E}[D^\pi(t) - D^\star|\mathcal{F}_{t-1}] = D_\theta(p^\pi(t)) - D^\star \geq 0$$

so $(r_t)_{0\leq t\leq T}$ is a $\mathcal{F}_t$ submartingale.

Since $T \leq \sum_{n=1}^{N^\pi(T)+1} D^\pi(n)$ and $\mu^\star = 1/D^\star$, we have

$$T\mu^\star - N^\pi(T) \leq 1 + \sum_{n=1}^{N^\pi(T)+1} (\mu^\star D^\pi(n) - 1) = 1 + \mu^\star r_{N^\pi(T)+1}.$$

Since $(r_t)_{0\leq t\leq T}$ is a submartingale, $N^\pi(T) \leq T$ is a bounded stopping time, Doob's stopping theorem [37, Th. 5.4.1] gives

$$\mathbb{E}(r_{N^\pi(T)+1}) \leq \mathbb{E}(r_{T+1}) = R^\pi(T + 1).$$

Taking expectations above yields

$$\frac{S^\pi(T)}{\log(T)} \leq \frac{1}{\log(T)} + \mu^\star \frac{R^\pi(T + 1)}{\log(T)}.$$

Letting $T \to \infty$ proves the result since $\frac{\log(T)}{\log(T+1)} \to 1$. ■

Appendix B
Proof of Proposition 4.4

Proof: Consider a problem instance with line topology in which $\theta_i = \alpha$ for all $i \notin p^\star$, and $\theta_i = \alpha + \alpha^2$ for all $i \in p^\star$ for some $\alpha \in (0, 0.36]$. Hence, $\theta_i < 0.5$ for all $i \in p^\star$. For any uniformly good policy $\pi \in \Pi_2 \cup \Pi_3$, by Lemma 1 we have

$$\begin{aligned}
\liminf_{N\to\infty} \frac{R^\pi(N)}{\log(N)} &\geq \sum_{i\notin p^\star} \frac{1}{\mathrm{KLG}(\theta_i, \theta_{\zeta(i)})}\left(\frac{1}{\theta_i} - \frac{1}{\theta_{\zeta(i)}}\right) \\
&\geq \sum_{i\notin p^\star} \frac{1}{2(\theta_{\zeta(i)} - \theta_i)} = \sum_{i\notin p^\star} \frac{1}{2\theta_i\theta_{\zeta(i)}\left(\theta_i^{-1} - \theta_{\zeta(i)}^{-1}\right)} \\
&= \frac{|E| - H}{2\alpha(\alpha + \alpha^2)(\alpha^{-1} - (\alpha + \alpha^2)^{-1})} \\
&= \frac{|E| - H}{2\alpha(\alpha + \alpha^2)\Delta_{\min}} \geq \frac{|E| - H}{4\alpha^2\Delta_{\min}} = \frac{|E| - H}{4\theta_{\min}^2\Delta_{\min}}
\end{aligned}$$

where in the second inequality we used Lemma 3 and $\mathrm{KL}(u, v) \leq \frac{(u-v)^2}{v(1-v)} \leq \frac{2(u-v)^2}{v}$ for $v \leq 0.5$. This implies that the regret of any uniformly good policy $\pi \in \Pi_2 \cup \Pi_3$ for this problem instance is at least $\Omega\left(\frac{|E|-H}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$. ■

Appendix C
Proof of Theorem 5.1

We first recall two results. Lemma 5 is a concentration inequality derived in [38, Th. 2]. Lemma 6, proven in [39, Lemma 2], is a local version of Pinsker's inequality for the KL information number between two Bernoulli distributions.

Lemma 5: There exists a number $K_H > 0$ that only depends on H such that for all p and $n \geq 2$

$$\mathbb{P}\left[\sum_{i\in p} t_i(n)\mathrm{KL}(\hat{\theta}_i(n), \theta_i) > f_1(n)\right] \leq K_H n^{-1}(\log(n))^{-2}.$$

Lemma 6 (see [39, Lemma 2]): For $0 \leq u < v \leq 1$, we have

$$\mathrm{KL}(u, v) \geq \frac{1}{2v}(u - v)^2.$$

Next we prove the theorem.

Statement (i): Let $p \in \mathcal{P}$, $n \in \mathbb{N}$, $t \in \mathbb{N}^{|E|}$, and $u, \lambda \in (0, 1]^{|E|}$ with $u_i \geq \lambda_i$ for all i. By Cauchy–Schwarz inequality, we have

$$\begin{aligned}
p^\top\lambda^{-1} - p^\top u^{-1} &= \sum_{i\in p} \frac{u_i - \lambda_i}{u_i\lambda_i} = \sum_{i\in p} \frac{\sqrt{t_i}(u_i - \lambda_i)}{\sqrt{u_i}} \frac{1}{\lambda_i\sqrt{t_i u_i}} \\
&\leq \sqrt{\sum_{i\in p} \frac{t_i(u_i - \lambda_i)^2}{u_i}} \sqrt{\sum_{i\in p} \frac{1}{t_i u_i \lambda_i^2}} \\
&\leq \sqrt{\sum_{i\in p} \frac{t_i(u_i - \lambda_i)^2}{u_i}} \sqrt{\sum_{i\in p} \frac{1}{t_i \lambda_i^3}}
\end{aligned}$$

where we used $u_i \geq \lambda_i$ for all i in the last step. Using Lemma 6, it then follows that

$$p^\top\lambda^{-1} - p^\top u^{-1} \leq \sqrt{\sum_{i\in p} 2t_i\mathrm{KL}(\lambda_i, u_i)} \sqrt{\sum_{i\in p} \frac{1}{t_i\lambda_i^3}}.$$

Thus, $\sum_{i\in p} t_i\mathrm{KL}(\lambda_i, u_i) \leq f_1(n)$ implies

$$p^\top\lambda^{-1} - p^\top u^{-1} \leq \sqrt{\sum_{i\in p} \frac{2f_1(n)}{t_i\lambda_i^3}}$$

or equivalently, $p^\top u^{-1} \geq c_p(n, \lambda, t)$. Hence, by definition of $b_p(n, \lambda, t)$, we have $b_p(n, \lambda, t) \geq c_p(n, \lambda, t)$.

Statement (ii): If $\sum_{i\in p} t_i(n)\mathrm{KL}(\hat{\theta}_i(n), \theta_i) \leq f_1(n)$, then we have $b_p(n, \hat{\theta}(n), t(n)) \leq p^\top\theta^{-1}$ by definition of b_p. Therefore,

using Lemma 5, there exists K_H such that for all $n \geq 2$

$$\mathbb{P}\left[b_p(n, \hat{\theta}(n), t(n)) > p^\top \theta^{-1}\right] \leq \mathbb{P}\left[\sum_{i\in p} t_i(n)\mathrm{KL}(\hat{\theta}_i(n), \theta_i) > f_1(n)\right] \leq K_H n^{-1}(\log(n))^{-2}.$$

■

Appendix D
Proof of Theorem 5.3

To prove the theorem, we borrow some ideas from the analysis of [12, Th. 3].

A. Preliminary

Define $a = (1 - 2^{-\frac{1}{4}})$ and $\varepsilon = a\frac{\Delta_{\min}}{D^+}$. Note that the definition of D^+, together with the fact that $D_\theta(p^\star) > 0$, implies that $\varepsilon < a$. For $s \in \mathbb{N}^{|E|}$ and $p \in \mathcal{P}$, define $h(s, p) = \sum_{i\in p} \frac{1}{s_i}$. Define $s_i(n) = t_i(n)\hat{\theta}_i(n)$ the number of packets routed through link i before the nth packet is sent and $s(n) = (s_i(n))_{i\in E}$. To ease notation, define $h(n) = h(s(n), p(n))$. We will use the following technical lemma.

Lemma 7: Consider $S \subset \mathbb{N}$, $(s(n))_n$ an integer sequence such that $s(n) \neq s(n')$ for all $(n, n') \in S, n \neq n'$. Consider a constant $C > 0$ and a positive function δ such that $\min_{n\in S} \delta(s(n)) \geq \delta_{\min}$. Then

$$Z := \sum_{n\in S} \delta(s(n))\mathbb{1}\{s(n) \leq C\delta(s(n))^{-2}\} \leq \frac{2C}{\delta_{\min}}.$$

Proof: If $s(n) \leq C\delta(s(n))^{-2}$, we have $\delta(s(n)) \leq \sqrt{C/s(n)}$ and $s(n) \leq C\delta_{\min}^{-2}$. So

$$Z \leq \sum_{n\in S} \sum_{t=1}^{C\delta_{\min}^{-2}} \mathbb{1}\{s(n) = t\}\sqrt{\frac{C}{t}} \leq \sum_{t=1}^{C\delta_{\min}^{-2}} \sqrt{\frac{C}{t}}$$

since $\sum_{n\in S} \mathbb{1}\{s(n) = t\} \leq 1$. Using the inequality $\sum_{t=1}^{T} t^{-\frac{1}{2}} \leq 1 + \int_1^T t^{-\frac{1}{2}} dt \leq 2\sqrt{T}$ yields the result. ■

B. Proof of the Theorem

For any n, introduce the following events:

$$A_n = \left\{\sum_{i\in p^\star} t_i(n)\mathrm{KL}(\hat{\theta}_i(n), \theta_i) > f_1(n)\right\}$$

$$B_{n,i} = \{p_i(n) = 1,\ |\hat{\theta}_i(n) - \theta_i| \geq \varepsilon\theta_i\}, \quad B_n = \bigcup_{i\in E} B_{n,i}$$

$$F_n = \left\{\Delta_{p(n)} \leq (1-a)^{-2}\theta_{\min}^{-1}\sqrt{2f_1(N)h(n)}\right\}.$$

We first prove that $p(n) \neq p^\star$ implies $n \in A_n \cup B_n \cup F_n$. Consider n such that $p(n) \neq p^\star$ and $A_n \cup B_n$ does not occur. By design of the algorithm, $\xi_{p(n)}(n) \leq \xi_{p^\star}(n)$ and $\xi_{p^\star}(n) \leq D^\star$ since A_n does not occur. By Theorem 5.1, we have $c_{p(n)}(n) \leq b_{p(n)}(n)$. Hence $c_{p(n)}(n) \leq D^\star$. This implies

$$p(n)^\top\hat{\theta}(n)^{-1} - \sqrt{\sum_{i\in p}\frac{2f_1(n)}{s_i(n)\hat{\theta}_i(n)^2}} \leq D^\star$$

so that

$$\Delta_{p(n)} \leq p(n)^\top\theta^{-1} - p(n)^\top\hat{\theta}(n)^{-1} + \sqrt{\sum_{i\in p(n)}\frac{2f_1(n)}{s_i(n)\hat{\theta}_i(n)^2}}.$$

Since B_n does not occur, $\hat{\theta}(n)^{-1} \geq \theta^{-1}/(1+\varepsilon)$ and

$$p(n)^\top\theta^{-1} - p(n)^\top\hat{\theta}(n)^{-1} \leq \frac{p(n)^\top\theta^{-1}\varepsilon}{(1+\varepsilon)} \leq D^+\varepsilon = a\Delta_{\min} \leq a\Delta_{p(n)}.$$

Moreover, $\hat{\theta}_i(n) \geq \theta_{\min}(1-a)$ for all $i \in p(n)$, and $f_1(n) \leq f_1(N)$ so

$$\sum_{i\in p(n)}\frac{2f_1(n)}{s_i(n)\hat{\theta}_i(n)^2} \leq \frac{2f_1(N)h(n)}{(1-a)^2\theta_{\min}^2}.$$

Hence

$$\Delta_{p(n)} \leq a\Delta_{p(n)} + \frac{\sqrt{2f_1(N)h(n)}}{(1-a)\theta_{\min}}$$

and $\Delta_{p(n)} \leq (1-a)^{-2}\theta_{\min}^{-1}\sqrt{2f_1(N)h(n)}$. Therefore, $n \in F_n$.

The regret $R^\pi(N)$ is upper bounded by

$$\mathbb{E}\left(\sum_{n=1}^N \Delta_{p(n)}\right) \leq \mathbb{E}\left(\sum_{n=1}^N \Delta_{p(n)}(\mathbb{1}\{A_n\} + \mathbb{1}\{B_n\} + \mathbb{1}\{F_n\})\right).$$

Set A: Using Corollary 5.2 and $K_H > 0$, we have

$$\sum_{n\geq 1}\mathbb{P}(A_n) \leq 1 + K_H\sum_{n\geq 2} n^{-1}(\log(n))^{-2} \leq 4K_H. \tag{7}$$

Set B: Define $\tau_i(n) = \sum_{n'=1}^n \mathbb{1}\{B_{n',i}\}$. Since $B_{n',i}$ implies $p_i(n') = 1$, we have $s_i(n) \geq \tau_i(n)$. Applying [35, Lemma B.1], we have $\sum_{n=1}^N \mathbb{P}(B_{n,i}) \leq 2(\varepsilon\theta_i)^{-2}$. A union bound yields

$$\sum_{n=1}^N \mathbb{P}(B_n) \leq 2\varepsilon^{-2}\sum_{i\in E}\theta_i^{-2}. \tag{8}$$

Set F: Define $U = \frac{4f_1(N)}{(1-a)^4\theta_{\min}^2}$. Define the set

$$S_n = \{i \in p(n) : s_i(n) \leq HU\Delta_{p(n)}^{-2}\}$$

and events

$$G_n = \{|S_n| \geq \sqrt{H}\}$$

$$L_n = \left\{|S_n| < \sqrt{H}, \min_{i\in p(n)} s_i(n) \leq \sqrt{H}U\Delta_{p(n)}^{-2}\right\}.$$

Assume that neither G_n nor L_n occurs. Then

$$h(n) = \sum_{i\in p(n), i\in S_n}\frac{1}{s_i(n)} + \sum_{i\in p(n), i\notin S_n}\frac{1}{s_i(n)} \leq \frac{|S_n|\Delta_{p(n)}^2}{\sqrt{H}U} + \frac{(H - |S_n|)\Delta_{p(n)}^2}{HU} < \frac{2\Delta_{p(n)}^2}{U}$$

since $|S_n| < \sqrt{H}$. Hence, $\Delta^2_{p(n)} > Uh(n)/2$ and F_n does not occur. So $F_n \subset G_n \cup L_n$. Further decompose G_n and L_n as

$$G_{i,n} = G_n \cap \left\{ i \in p(n),\ s_i(n) \leq HU\Delta^{-2}_{p(n)} \right\}$$

$$L_{i,n} = L_n \cap \left\{ i \in p(n),\ s_i(n) \leq \sqrt{H}U\Delta^{-2}_{p(n)} \right\}.$$

Applying Lemma 7 twice, we get:

$$\sum_{n=1}^{N} \Delta_{p(n)} \mathbb{1}\{G_{i,n}\} \leq \frac{2HU}{\Delta_{\min}}\,,\ \sum_{n=1}^{N} \Delta_{p(n)} \mathbb{1}\{L_{i,n}\} \leq \frac{2\sqrt{H}U}{\Delta_{\min}}.$$

We have $\sum_{i\in E} \mathbb{1}\{G_{i,n}\} = |S_n|\mathbb{1}\{G_n\} \geq \sqrt{H}\mathbb{1}\{G_n\}$. So

$$\sum_{n=1}^{N} \Delta_{p(n)} \mathbb{1}\{G_n\} \leq \frac{1}{\sqrt{H}} \sum_{n=1}^{N} \sum_{i\in E} \Delta_{p(n)} \mathbb{1}\{G_{i,n}\} \leq \frac{2|E|\sqrt{H}U}{\Delta_{\min}}.$$

Furthermore

$$\sum_{n=1}^{N} \Delta_{p(n)} \mathbb{1}\{L_n\} \leq \sum_{n=1}^{N} \sum_{i\in E} \Delta_{p(n)} \mathbb{1}\{L_{i,n}\} \leq \frac{2|E|\sqrt{H}U}{\Delta_{\min}}.$$

Since $\mathbb{1}\{F_n\} \leq \mathbb{1}\{G_n\} + \mathbb{1}\{L_n\}$, we get

$$\mathbb{E}\left(\sum_{n=1}^{N} \Delta_{p(n)} \mathbb{1}\{F_n\} \right) \leq \frac{4|E|\sqrt{H}U}{\Delta_{\min}}. \tag{9}$$

Combining (7), (8), and (9) with $\Delta_{p(n)} \leq D^+$ yields the announced result

$$R^\pi(N) \leq \frac{4|E|\sqrt{H}U}{\Delta_{\min}} + 2D^+ \left(2K_H + \varepsilon^{-2} \sum_{i\in E} \theta_i^{-2} \right).$$

■

APPENDIX E
PROOF OF THEOREM 5.4

The proof technique is similar to the analysis of [12, Th. 5].

A. Preliminary

For $s \in \mathbb{N}^{|E|}$ and $p \in \mathcal{P}$, define $h'(s, p) = (\sum_{i\in p} \frac{1}{\sqrt{s_i}})^2$, and as before $s_i(n) = t_i(n)\hat{\theta}_i(n)$ and $s(n) = (s_i(n))_{i\in E}$, and $h'(n) = h'(s(n), p(n))$. We will use the following lemma.

Lemma 8: For all $n, t \in \mathbb{N}$, $\lambda \in (0, 1]$, and $i \in E$

$$\omega_i(n, \lambda, t) \geq \frac{1}{\lambda} - \sqrt{\frac{2f_2(n)}{t\lambda^3}}.$$

Proof: Let $i \in E$, $n, t \in \mathbb{N}$ and $u, \lambda \in (0, 1]$ with $u \geq \lambda$. We have

$$\frac{1}{\lambda} - \frac{1}{u} = \sqrt{\frac{t(u-\lambda)^2}{u}} \cdot \frac{1}{\sqrt{tu\lambda^2}} \leq \sqrt{2t\mathrm{KL}(\lambda, u)} \cdot \frac{1}{\sqrt{t\lambda^3}}$$

where the second inequality follows from Lemma 6 and $u \geq \lambda$. Hence, $t\mathrm{KL}(\lambda, u) \leq f_2(n)$ implies: $\frac{1}{u} \geq \frac{1}{\lambda} - \sqrt{\frac{2f_2(n)}{t\lambda^3}}$. This inequality holds for all $u \in [\lambda, 1]$. Thus, the claim of the lemma follows by definition of $\omega_i(n, \lambda, t)$. ■

B. Proof of the theorem

For any n, we define the following events:

$$A_{n,i} = \left\{ t_i(n)\mathrm{KL}(\hat{\theta}_i(n), \theta_i) > f_2(n) \right\},\quad A_n = \bigcup_{i\in p^\star} A_{n,i}$$

$$B_{n,i} = \{ p_i(n) = 1,\ |\hat{\theta}_i(n) - \theta_i| \geq \varepsilon\theta_i \},\quad B_n = \bigcup_{i\in E} B_{n,i}$$

$$F_n = \{ \Delta_{p(n)} \leq (1-a)^{-2}\theta_{\min}^{-1}\sqrt{2f_2(N)h'(n)} \}.$$

We show that $p(n) \neq p^\star$ implies: $n \in A_n \cup B_n \cup F_n$. Consider n such that $p(n) \neq p^\star$ and $A_n \cup B_n$ does not occur. By design of the algorithm, $p(n)^\top\omega(n) \leq p^{\star\top}\omega(n)$, and $p^{\star\top}\omega(n) \leq D^\star$ since A_n does not occur. Hence, $p(n)^\top\omega(n) \leq D^\star$. By Lemma 8, for all i

$$\omega_i(n) \geq \frac{1}{\hat{\theta}_i(n)} - \sqrt{\frac{2f_2(n)}{s_i(n)\hat{\theta}_i(n)^2}}.$$

Summing over $i \in p(n)$, we get

$$\Delta_{p(n)} \leq p(n)^\top\theta^{-1} - p(n)^\top\hat{\theta}(n)^{-1} + \sum_{i\in p(n)} \sqrt{\frac{2f_2(n)}{s_i(n)\hat{\theta}_i(n)^2}}.$$

As before, when B_n does not occur we have

$$p(n)^\top\theta^{-1} - p(n)^\top\hat{\theta}(n)^{-1} \leq a\Delta_{p(n)}.$$

Furthermore, $\hat{\theta}_i(n) \geq \theta_{\min}(1-a)$ for all $i \in p(n)$, and $f_2(n) \leq f_2(N)$ so that

$$\sum_{i\in p(n)} \sqrt{\frac{2f_2(n)}{s_i(n)\hat{\theta}_i(n)^2}} \leq \sum_{i\in p(n)} \sqrt{\frac{f_2(N)}{s_i(n)\theta_{\min}^2(1-a)^2}}.$$

Hence

$$\Delta_{p(n)} \leq a\Delta_{p(n)} + \frac{\sqrt{2f_2(N)h'(n)}}{(1-a)\theta_{\min}}$$

and $\Delta_{p(n)} \leq (1-a)^{-2}\theta_{\min}^{-1}\sqrt{2f_2(N)h'(n)}$ so that $n \in F_n$. The regret $R^\pi(N)$ is upper bounded by

$$\mathbb{E}\left(\sum_{n=1}^{N} \Delta_{p(n)} \right) \leq \mathbb{E}\left(\sum_{n=1}^{N} \Delta_{p(n)} (\mathbb{1}\{A_n\} + \mathbb{1}\{B_n\} + \mathbb{1}\{F_n\}) \right).$$

Set A: By [40, Th. 10] and a union bound

$$\mathbb{P}(A_n) \leq \sum_{i\in p^\star} \mathbb{P}(A_{n,i}) \leq H\lceil f_2(n)\log(n)\rceil e^{1-f_2(n)}.$$

Hence

$$\sum_{n=1}^{N} \mathbb{P}(A_n) \leq H\left(1 + e\sum_{n\geq 2} \lceil f_2(n)\log(n)\rceil e^{-f_2(n)} \right) \leq 8H. \tag{10}$$

Set B: As in the proof of Theorem 5.3

$$\sum_{n=1}^{N} \mathbb{P}(B_n) \leq 2\varepsilon^{-2} \sum_{i\in E} \theta_i^{-2}. \tag{11}$$

Set F: Define $U' = 2H^2 f_2(N)(1-a)^{-4}\theta_{\min}^{-2}$. Similarly to the proof of [12, Th. 5], consider $\alpha, \beta > 0$, and for $\ell \in \mathbb{N}$ define $\alpha_\ell = \left(\frac{1-\beta}{\sqrt{\alpha}-\beta}\right)^2 \alpha^\ell$ and $\beta_\ell = \beta^\ell$. Introduce set $S_{\ell,n}$ and events $G_{\ell,n}$

$$S_{\ell,n} = \{i \in p(n), s_i(n) \le U'\alpha_\ell \Delta_{p(n)}^{-2}\},$$

$$G_{\ell,n} = \{|S_{\ell,n}| \ge \beta_\ell H\} \cap \{|S_{j,n}| < \beta_j H, j = 1, ..., \ell - 1\}.$$

If $\overline{\cup_{\ell\ge1} G_{\ell,n}} = \{|S_{\ell,n}| < H\beta_\ell, \ell \ge 1\}$ occurs, then

$$\begin{aligned}\sum_{\ell\ge1} \frac{|S_{\ell-1,n}| - |S_{\ell,n}|}{\sqrt{\alpha_\ell}} &= \frac{|S_{0,n}|}{\sqrt{\alpha_1}} + \sum_{\ell\ge1} |S_{\ell,n}|\left(\frac{1}{\sqrt{\alpha_{\ell+1}}} - \frac{1}{\sqrt{\alpha_\ell}}\right) \\ &< \frac{H\beta_0}{\sqrt{\alpha_1}} + \sum_{\ell\ge1} H\beta_\ell\left(\frac{1}{\sqrt{\alpha_{\ell+1}}} - \frac{1}{\sqrt{\alpha_\ell}}\right) \\ &= H\sum_{\ell\ge1} \frac{\beta_\ell - \beta_{\ell-1}}{\sqrt{\alpha_\ell}} \le H\end{aligned}$$

since $\frac{1}{\sqrt{\alpha_{\ell+1}}} - \frac{1}{\sqrt{\alpha_\ell}} \ge 0$. Now

$$|\{i : s_i(n) \in U'\Delta_{p(n)}^{-2}[\alpha_\ell, \alpha_{\ell-1}]\}| = |S_{\ell-1,n}| - |S_{\ell,n}|$$

so that

$$\sqrt{h'(n)} \le \sum_{\ell\ge1} \frac{(|S_{\ell-1,n}| - |S_{\ell,n}|)}{\sqrt{\alpha_\ell}} \frac{\Delta_{p(n)}}{\sqrt{U'}} < H\frac{\Delta_{p(n)}}{\sqrt{U'}}.$$

Hence, $\Delta_{p(n)}^2 > h'(n)U'H^{-2}$ and thus, F_n does not occur. Therefore, $F_n \subset \cup_{\ell\ge1} G_{\ell,n}$ and

$$\sum_{n=1}^{N} \Delta_{p(n)} \mathbb{1}\{F_n\} \le \sum_{n=1}^{N}\sum_{\ell\ge1} \Delta_{p(n)} \mathbb{1}\{G_{\ell,n}\}.$$

Further decompose $G_{\ell,n}$ as

$$G_{i,\ell,n} = G_{\ell,n} \cap \{i \in p(n),\ s_i(n) \le U'\alpha_\ell \Delta_{p(n)}^{-2}\}.$$

Observe that

$$\mathbb{1}\{G_{\ell,n}\} \le \frac{|S_{\ell,n}|}{H\beta_\ell}\mathbb{1}\{G_{\ell,n}\} = \frac{1}{H\beta_\ell}\sum_{i\in E} \mathbb{1}\{G_{i,\ell,n}\}.$$

Applying Lemma 7, we get

$$\begin{aligned}\sum_{n=1}^{N} \Delta_{p(n)} \mathbb{1}\{G_{i,\ell,n}\} &\le \sum_{n=1}^{N} \Delta_{p(n)} \mathbb{1}\left\{s_i(n) \le \frac{U'\alpha_\ell}{\Delta_{p(n)}^2}\right\} \\ &\le \frac{2U'\alpha_\ell}{\Delta_{\min}}.\end{aligned}$$

Putting it together

$$\sum_{n=1}^{N} \Delta_{p(n)} \mathbb{1}\{F_n\} \le \frac{2|E|U'}{H\Delta_{\min}} \sum_{\ell\ge1} \frac{\alpha_\ell}{\beta_\ell} \le \frac{90|E|U'}{H\Delta_{\min}} \qquad (12)$$

by choosing $\alpha = 0.15$ and $\beta = 0.24$ so that $\sum_{\ell\ge1} \frac{\alpha_\ell}{\beta_\ell} \le 45$.

The proof is completed by combining (10), (11), (12), and using the fact that $\Delta_{p(n)} \le D^+$. ■

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the anonymous reviewers for their careful reading and helpful comments.

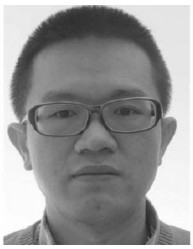
REFERENCES

[1] Z. Zou, A. Proutiere, and M. Johansson, "Online shortest path routing: The value of information," in *Proc. Amer. Control Conf.*, Jun. 2014, pp. 2142–2147.
[2] B. Awerbuch and R. D. Kleinberg, "Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches," in *Proc. 36th Annu. ACM Symp. Theory Comput.*, Jun. 2004, pp. 45–53.
[3] A. György and G. Ottucsák, "Adaptive routing using expert advice," *Comput. J.*, vol. 49, no. 2, pp. 180–189, 2006.
[4] A. György, T. Linder, G. Lugosi, and G. Ottucsák, "The on-line shortest path problem under partial monitoring," *J. Mach. Learn. Res.*, vol. 8, pp. 2369–2403, 2007.
[5] T. He, D. Goeckel, R. Raghavendra, and D. Towsley, "Endhost-based shortest path routing in dynamic networks," in *Proc. 32nd IEEE Int. Conf. Comput. Commun.*, Apr. 2013, pp. 2202–2210.
[6] O. Brun, L. Wang, and E. Gelenbe, "Big data for autonomic intercontinental overlays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 575–583, Mar. 2016.
[7] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," *J. Comput. Syst. Sci.*, vol. 78, no. 5, pp. 1404–1422, 2012.
[8] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic Game Theory*. New York, NY, USA: Cambridge Univ. Press, 2007.
[9] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, pp. 235–256, 2002.
[10] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proc. 30th Int. Conf. Mach. Learn.*, Jun. 2013, pp. 151–159.
[11] A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," in *Proc. 31st Int. Conf. Mach. Learn.*, Jun. 2014, pp. 100–108.
[12] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Tight regret bounds for stochastic combinatorial semi-bandits," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, May 2015, pp. 535–543.
[13] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527–535, 1952.
[14] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.
[15] J.-Y. Audibert, S. Bubeck, and G. Lugosi, "Regret in online combinatorial optimization," *Math. Oper. Res.*, vol. 39, no. 1, pp. 31–45, 2014.
[16] G. Neu and G. Bartók, "An efficient algorithm for learning with semi-bandit feedback," in *Proc. Int. Conf. Algorithmic Learn. Theory*, Oc. 2013, pp. 234–248.
[17] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1466–1478, Oct. 2012.
[18] R. Combes, M. S. Talebi, A. Proutiere, and M. Lelarge, "Combinatorial bandits revisited," in *Proc. Adv. Neural Inf. Process. Syst. 28*, Dec. 2015, pp. 2107–2115.
[19] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: I.I.D. Rewards," *IEEE Trans. Automat. Control*, vol. 32, no. 11, pp. 968–976, Nov. 1987.
[20] B. Kveton, Z. Wen, A. Ashkan, H. Eydgahi, and B. Eriksson, "Matroid bandits: Fast combinatorial optimization with learning," in *Proc. 30th Conf. Uncertainty Artif. Intell.*, Jul. 2014, pp. 420–429.
[21] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *Proc. Symp. New Front. Dyn. Spectr.*, Apr. 2010, pp. 1–9.
[22] M. S. Talebi, Z. Zou, R. Combes, A. Proutiere, and M. Johansson, "Stochastic online shortest path routing: The value of feedback," arXiv: 1309.7367, 2017.
[23] Z. Wen, B. Kveton, and A. Ashkan, "Efficient learning in large-scale combinatorial semi-bandits," in *Proc. 32nd Int. Conf. Mach. Learn.*, Jul. 2015, pp. 1113–1122.
[24] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[25] K. Liu and Q. Zhao, "Adaptive shortest-path routing under unknown and stochastically varying link states," in *Proc. 10th Int. Symp. Model. Optim. Mobile, Ad Hoc Wireless Netw.*, May 2012, pp. 232–237.
[26] P. Tehrani and Q. Zhao, "Distributed online learning of the shortest path under unknown random edge weights," in *Proc. 38th Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 3138–3142.
[27] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for Markov decision processes," *Math. Oper. Res.*, vol. 22, no. 1, pp. 222–255, 1997.
[28] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley-Interscience, 2005.
[29] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *J. Mach. Learn. Res.*, vol. 11, pp. 1563–1600, 2010.
[30] S. Filippi, O. Cappé, and A. Garivier, "Optimism in reinforcement learning and Kullback-Leibler divergence," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput.*, Sep./Oct. 2010, pp. 115–122.
[31] T. L. Graves and T. L. Lai, "Asymptotically efficient adaptive choice of control laws in controlled Markov chains," *SIAM J. Control Optim.*, vol. 35, no. 3, pp. 715–743, 1997.
[32] A. Sen and N. Balakrishnan, "Convolution of geometrics and a reliability problem," *Statist. Probab. Lett.*, vol. 43, no. 4, pp. 421–426, 1999.
[33] A. Shapiro, "Semi-infinite programming, duality, discretization and optimality conditions," *Optimization*, vol. 58, no. 2, pp. 133–161, 2009.
[34] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," *Int. Conf. Algorithmic Learn. Theory*, Springer, 2011, pp. 174–188.
[35] R. Combes and A. Proutiere, "Unimodal bandits: Regret lower bounds and optimal algorithms," *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 521–529. [Online]. Available: http://arxiv.org/abs/1405.5096
[36] P. Joulani, A. György, and C. Szepesvári, "Online learning under delayed feedback," in *Proc. 30th Int. Conf. Mach. Learn.*, Jun. 2013, pp. 1453–1461.
[37] R. Durrett, *Probability: Theory and Examples*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[38] S. Magureanu, R. Combes, and A. Proutiere, "Lipschitz bandits: Regret lower bounds and optimal algorithms," in *Proc. 27th Annu. Conf. Learn. Theory*, Jun. 2014, pp. 975–999.
[39] A. Garivier, P. Ménard, and G. Stoltz, "Explore first, exploit next: The true shape of regret in bandit problems," arXiv:1602.07182, 2016.
[40] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proc. 24th Annu. Conf. Learn. Theory*, Jul. 2011, pp. 359–376.

Mohammad Sadegh Talebi received the B.S. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran, in 2004, and the M.Sc. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2006. He is currently working toward the Ph.D. degree in the Department of Automatic Control, KTH Royal Institute of Technology, Stockholm, Sweden.

His current research interests include resource allocation in networks, reinforcement learning, and learning theory.



Zhenhua Zou received the Ph.D. degree in electrical engineering from the School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden, in 2014.

He is currently at Ericsson Research, Stockholm, Sweden, working on machine-type communications. His research interest includes control, optimization, and learning algorithms applied in wireless communication and communication networks.

Richard Combes received the Engineering degree from Telecom ParisTech, Paris, France, in 2008, the Master's degree in mathematics from the University of Paris VII, Paris, France, in 2009, and the Ph.D. degree in mathematics from the University of Paris VI, Paris, France, in 2013.

He was a Visiting Scientist at INRIA in 2012 and a Post-Doc at KTH in 2013. He is currently an Assistant Professor in Centrale-Supelec (L2S), France. His current research interests include machine learning, networks, and probability.

Dr. Combes received the Best Paper Award at CNSM 2011.

Alexandre Proutiere received the Master's degree in mathematics from École Normale Supérieure, Paris, France, in 1996; the Master's degree in engineering from Télécom ParisTech, Paris, France, in 1998; and the Ph.D. degree in applied mathematics from École Polytechnique, Palaiseau, France, in 2003.

In 2000, he joined France Telecom R&D as a Research Engineer. From 2007 to 2011, he was a Researcher at Microsoft Research, Cambridge, U.K. He is currently a Full Professor in the Department of Automatic Control, KTH Royal Institute of Technology, Stockholm, Sweden. He is an Engineer at Corps of Mines.

Dr. Proutiere received the ACM Sigmetrics Rising Star Award in 2009, and the Best Paper Awards at ACM Sigmetrics Conference in 2004 and 2010 and at the ACM Mobihoc Conference in 2009. He was an Associate Editor of the IEEE/ACM TRANSACTIONS ON NETWORKING and an Editor of the IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS, and is currently an Editor of *Queuing Systems*.

Mikael Johansson received the M.Sc. and Ph.D. degrees in electrical engineering from Lund University, Lund, Sweden, in 1994 and 1999, respectively.

He held postdoctoral positions at Stanford University, Stanford, CA, USA, and the University of California, Berkeley, CA, USA, before joining KTH Royal Institute of Technology, Stockholm, Sweden, in 2002, where he currently serves as a Full Professor. He has published two books and more than a hundred papers, several of which are highly cited and have received recognition in terms of best paper awards. He has served on the editorial boards of *Automatica* and the IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS, as well as on the program committee for several top conferences organized by the IEEE and ACM. He has played a leading role in several national and international research projects in control and communications.