Adaptive Shortest-Path Routing under Unknown and Stochastically Varying Link States

Keqin Liu, Qing Zhao

Adaptive Shortest-Path Routing under Unknown and Stochastically Varying Link States

Keqin Liu, Qing Zhao
Dept. of Elec. and Comp. Eng., Univ. of California, Davis
Email: {kqliu,qzhao}@ucdavis.edu

***Abstract*—We consider adaptive shortest-path routing in wireless networks. In this problem, we aim to optimize the quality of communication between a source and a destination through adaptive path selection. Due to the randomness and uncertainties in the network dynamics, the state of each communication link varies over time according to a stochastic process with unknown distributions. The link states are not directly observable. The aggregated end-to-end cost of a path from the source to the destination is revealed after the path is chosen for communication. The objective is an adaptive path selection algorithm that minimizes regret defined as the additional cost over the ideal scenario where the best path is known *a priori*. This problem can be cast as a variation of the classic multi-armed bandit (MAB) problem with each path as an arm and arms dependent through common links. We show that by exploiting arm dependencies, a regret polynomial with the network size can be achieved while maintaining the optimal logarithmic order with time. This is in sharp contrast with the exponential regret order with the network size offered by a direct application of the classic MAB policies that ignores arm dependencies. Furthermore, our results are obtained under a general model of link state distributions (including heavy-tailed distributions). These results find applications in cognitive radio and ad hoc networks with unknown and dynamic communication environments.**

I. Introduction

We consider the communication between a source and a destination in a wireless network. The state of each communication link is modeled as a random cost (or reward) that evolves according to an i.i.d. random process with an unknown distribution. At each time, a path from the source to the destination is selected as the communication route, and the *total end-to-end cost*, given by the sum of the costs of all links on the path, is subsequently observed. The cost of each individual link is not observable. The objective is to design an optimal sequential path selection policy to minimize the long-term total cost.

[0]This work was supported by the National Science Foundation under Grant CCF-0830685 and by the Army Research Office under Grant W911NF-08-1-0467.

A. Stochastic Online Learning based on Multi-Armed Bandit

The above problem can be modeled as a variation of the classic Multi-Armed Bandit (MAB) problem. In the classic MAB [1]–[5], there are N independent arms and a player needs to decide which arm to play at each time. An arm, when played, incurs a random cost drawn from an unknown distribution. The performance of a sequential arm selection policy is measured by regret, defined as the difference in expected total cost with respect to the optimal policy in the ideal scenario with known cost models where the player always plays the best arm. The optimal regret was shown by Lai and Robbins to be logarithmic with time [1]. Since arms are assumed independent, observations from one arm do not provide information about other arms. It is then readily seen that the optimal regret grows linearly with the number of arms.

The adaptive shortest-path routing problem can be cast as an MAB by treating each path as an arm. The difference is that paths are dependent through common links. While the dependency across paths can be ignored in learning and the policies for the classic MAB proposed in [1]–[5] directly apply,

978-3-901882-45-6/2012 - Copyright is with IFIP

such a naive approach yields poor performance with a regret growing linearly with the number of paths, thus exponentially with the network size (*i.e.,* the number of links which determines the number of unknowns) in the worst case.

In this paper, we show that by exploiting the structure of the path dependencies, a regret polynomial with the network size can be achieved while preserving the optimal logarithmic order with time. Specifically, we propose an algorithm that achieves $O(md^3 \log T)$ regret for all light-tailed cost distributions, where m is the number of edges, d the dimension of the path set that is upper bounded by m, and T the time horizon length. We further show that by sacrificing an arbitrarily small regret order with time, the proposed algorithm achieves regret linear with d. Specifically, the algorithm offers $O(d\, f(T) \log T)$ regret where $f(T)$ is an arbitrarily slowly diverging function with $f(T) \to \infty$ as $T \to \infty$. The proposed algorithm thus offers a performance tradeoff in terms of the network size and the time horizon length. For heavy-tailed cost distributions, a regret linear with the size of the network and sublinear with time can be achieved. Specifically, the algorithm offers $O(dT^{1/q})$ regret when the moments of the cost distributions exist up to the qth order. We point out that any regret sublinear with time implies the convergence of the time-average cost to the minimum one of the best path.

B. Applications

One application example is adaptive routing in cognitive radio networks where secondary users communicate by exploiting channels temporarily unoccupied by primary users. In this case, the availability of each link dynamically varies according to the communication activities of nearby primary users. The delay on each link can thus be modeled as a stochastic process unknown to the secondary users. The objective is to route through the path with the smallest latency (*i.e.,* the lightest primary traffic) through stochastic online learning.

Other applications include ad hoc networks where link states vary stochastically due to channel fading or random contentions with other users.

C. Related Work

Several policies exist for the classic MAB with independent arms [1]–[4]. These policies achieve the optimal logarithmic regret order for certain specific light-tailed cost/reward distributions. In [6], we extended the UCB policy proposed by Auer *et al.* in [4] for distributions with finite support to all light-tailed distributions. In [5], we proposed a learning algorithm based on a deterministic separation of exploration and exploitation (DSEE) that achieves the optimal logarithmic regret order for all light-tailed distributions and sublinear regret order for heavy-tailed distributions.

The adaptive shortest path problem considered in this paper is a special class of the so-called stochastic online linear optimization problem that considers a general compact action space (rather than a finite number of discrete choices such as paths). In [7], an algorithm was proposed to achieve $O(d^{3/2} \log^{3/2} T\sqrt{T})$ regret for distributions with finite support, where d is the dimension of the action space. In [8], we proposed an algorithm that achieves $O(d\, f(T) T^{2/3} \log^{1/3} T)$ regret for all light-tailed distributions, where $f(T)$ is an arbitrary function with $f(T) \to \infty$ as $T \to \infty$. The algorithm proposed in [8] thus achieves a regret better in d but worse in T.

In the context of adaptive shortest path routing, the work most relevant to this paper is [9]. Under the assumption of fully observable link states, a learning algorithm was proposed in [9] based on the UCB policy in [4] to achieve $O(m^4 \log T)$ regret, where m is the number of links in the network. The learning algorithm proposed in this paper improves the regret order under a less informative observation model. Furthermore, our algorithm directly applies to the model with link-level observability as adopted in [9] and achieves $O(md \log T)$ regret. The stochastic online linear optimization algorithms proposed in [7] apply to adaptive shortest path routing and would offer $O(d^2 \log^3 T)$ regret for cost distributions with finite support. The learning algorithm proposed in this paper, with $O(d\, f(T) \log T)$ regret, improves the regret order in both d and T and applies to all light-tailed distributions. In [10], the problem was addressed under an adversarial bandit model

in which the link costs are chosen by an adversary and are treated as arbitrary bounded deterministic quantities. An algorithm was proposed to achieve regret sublinear with time and polynomial with the network size [10].

II. Problem Formulation

A. Adaptive Shortest-Path Routing

Consider the communication between a source s and a destination r in a network. Let $G = (V, E)$ denote the directed graph consisting of all simple paths from s to r (see Fig. 1). Let m and n denote, respectively, the number of edges and vertices in graph G.

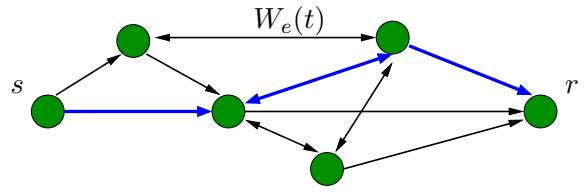

Fig. 1. The network model for adaptive shortest-path routing.

Let $W_e(t)$ denote the weight of edge e at time t. We assume that for each edge e, $W_e(t)$ is an i.i.d. process in time with an unknown distribution. The weights are independent across edges with potentially different distributions.

At the beginning of each time slot t, a path $i \in \mathcal{P}$ ($i \in \{1, 2, \ldots, |\mathcal{P}|\}$) from s to r is chosen, where $\mathcal{P}$ is the set of all paths from s to r in G. Subsequently, the *total end-to-end cost* $C_t(i)$ of path i, given by the sum of the weights of all edges on the path, is revealed at the end of the slot. The edge cost $W_e(t)$ is not directly observable.

The regret of a path selection policy π is defined as the expected additional cost incurred over time T compared to the optimal policy that always selects the best path (*i.e.,* the path with the minimum expected total end-to-end cost). The objective is to design a path selection policy π to minimize the growth rate of the regret with respect to both the horizon length T and the network size m (the number of unknowns). Let σ be a permutation on all paths such that

$$\mathbb{E}[C_t(\sigma(1))] \leq \mathbb{E}[C_t(\sigma(2))] \leq \cdots \leq \mathbb{E}[C_t(\sigma(|\mathcal{P}|))].$$

We define regret

$$\mathcal{R}^{\pi}(T) \triangleq \mathbb{E}[\sum_{t=1}^{T}(C_t(\pi) - C_t(\sigma(1)))],$$

where $C_t(\pi)$ denotes the total end-to-end cost of the selected path under policy π at time t.

B. Vector Representation

We represent each path i as a vector $\mathbf{p}_i$ with m entries consisting of 0's and 1's representing whether or not an edge is on the path. The vector space of all paths is embedded in a d-dimensional ($d \leq m$) subspace[1] of $\mathcal{R}^m$. The cost of path i at time t is thus given by the linear function

$$C_t(i) =< [W_1(t), W_2(t), \ldots, W_m(t)],\ \mathbf{p}_i > .$$

The vector space all paths is a compact subset of $\mathcal{R}^d$ with dimension $d \leq m$. For any compact subset of $\mathcal{R}^d$, there exists a barycentric spanner (see Definition 1), which can be efficiently constructed [10].

Definition 1: A set $\mathbf{B} = \{\mathbf{p}_1, \ldots, \mathbf{p}_d\}$ is a *barycentric spanner* for a d-dimensional set $\mathcal{P}$ ($\mathbf{B} \subset \mathcal{P}$) if every $\mathbf{p} \in \mathcal{P}$ can be expressed as a linear combination of elements of $\mathbf{B}$ using coefficients in $[-1, 1]$.

III. Adaptive Shortest-Path Routing Algorithm

The general structure of the algorithm follows the DSEE framework established in our prior work [5] for the classic MAB. More specifically, we partition time into interleaving exploration and exploitation sequences. In the exploration sequence, we sample the d basis vectors in the barycentric spanner in a round-robin fashion to obtain an empirical mean of the cost of each basis path. In the exploitation sequence, we obtain an estimate of the cost of each path as a linear combination of the empirical mean of the basis paths and select the path with the minimum estimated cost. Specifically, in a slot t that belongs to the exploitation sequence, let $\overline{\theta}_{\mathbf{p}_k}(t)$ denote the

[1] If graph G is acyclic, then $d = m - n + 2$.

current empirical mean of the cost of the basis path $\mathbf{p}_k$ $(k = 1, 2, \ldots, d)$, where we have labeled the paths in the barycentric spanner as the first d paths in $\mathcal{P}$. For any path $\mathbf{p}_i \in \mathcal{P}$, let $\{a_k : |a_k| \leq 1\}_{k=1}^d$ be the coefficients such that $\mathbf{p}_i = \sum_{k=1}^d a_k \mathbf{p}_k$. The estimated cost of this path is then given by

$$\overline{\theta}_{\mathbf{p}_i}(t) = \sum_{k=1}^{d} a_k \overline{\theta}_{\mathbf{p}_k}(t). \tag{1}$$

The path $\mathbf{p}^*(t)$ with

$$\mathbf{p}^*(t) = \arg\min\{\overline{\theta}_{\mathbf{p}_i}(t),\ i = 1, 2, \ldots, |\mathcal{P}|\}. \tag{2}$$

is then chosen at this exploitation time t. A detailed algorithm is given in Fig. 2.

With the above structure, the only remaining issue is the design of the cardinality of the exploration sequence which balances the tradeoff between learning and exploitation. As shown in the next section, the cardinality of the exploration sequence can be chosen according to the heaviness of the tail distribution to achieve the optimal logarithmic regret order with T for all light-trailed distributions and sublinear regret order with T for heavy-tailed distributions. For both light-tailed and heavy-tailed distributions, the regret order is polynomial with the network size.

Adaptive Shortest-Path Routing Algorithm

- Notations and Inputs: Construct a barycentric spanner

  $$\mathbf{B} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_d\}$$

  of the vector space of all paths. Let $\mathcal{A}(t)$ denote the set of time indices that belong to the exploration sequence up to (and including) time t and $\mathcal{A}(1) = \{1\}$. Let $|\mathcal{A}(t)|$ denote the cardinality of $\mathcal{A}(t)$. Let $\overline{\theta}_{\mathbf{p}_k}(t)$ denote the sample mean of path $\mathbf{p}_k$ $(k \in \{1, \ldots, d\})$ computed from the past cost observations on the path. For two positive integers k and l, define $k \oslash l \triangleq ((k-1) \bmod l) + 1$, which is an integer taking values from $1, 2, \cdots, l$.

- At time t,
  1. if $t \in \mathcal{A}(t)$, choose path $\mathbf{p}_k$ with $k = |\mathcal{A}(t)| \oslash d$;
  2. if $t \notin \mathcal{A}(t)$, estimate the cost of each path $\mathbf{p}_i \in \mathcal{P}$ according to (1). Choose path $\mathbf{p}^*(t)$ as given in (2).

Fig. 2. The general structure of the algorithm.

IV. Regret Analysis

A. Light-Tailed Cost Distributions

We consider the light-tailed cost distributions as defined below.

Definition 2: A random variable X is light-tailed if its moment-generating function exists, *i.e.,* there exists a $u_0 > 0$ such that for all $u \leq |u_0|$,

$$M(u) \triangleq \mathbb{E}[\exp(uX)] < \infty.$$

Otherwise X is heavy-tailed.

For a zero-mean light-tailed random variable X, we have [11], for all $u \leq |u_0|$,

$$M(u) \leq \exp(\zeta u^2/2), \tag{3}$$

where $\zeta \geq \sup\{M^{(2)}(u), \ -u_0 \leq u \leq u_0\}$ with $M^{(2)}(\cdot)$ denoting the second derivative of $M(\cdot)$ and u_0 the parameter specified in Definition 2. From (3), we have the following extended Chernoff-Hoeffding bound on the deviation of the sample mean from the true mean for light-tailed random variables [5].

Lemma 1 [5]: Let $\{X(t)\}_{t=1}^{\infty}$ be i.i.d. light-tailed random variables. Let $\overline{X_s} = (\Sigma_{t=1}^{s} X(t))/s$ and $\theta = \mathbb{E}[X(1)]$. We have, for all $\delta \in [0, \zeta u_0], a \in (0, 1/(2\zeta)]$,

$$\Pr(|\overline{X_s} - \theta| \geq \delta) \leq 2\exp(-a\delta^2 s). \tag{4}$$

Note that the path cost is the sum of the costs of all edges on the path. Since the number of edges in a path is upper bounded by m, the bound on the moment generating function in (3) holds on the path cost by replacing ζ by $m\zeta$, and so does the Chernoff-Hoeffding bound in (4).

Theorem 1: Construct an exploration sequence as follows. Let a, ζ, u_0 be the constants such that (4) holds for each edge cost. Let $d \leq m$ be the dimension of the path set. Choose a constant

$b > 2m/a$, a constant

$$c \in (0, \min_{j:\mathbb{E}[C_t(\sigma(j))-C_t(\sigma(1))]>0}\{\mathbb{E}[C_t(\sigma(j))-C_t(\sigma(1))]\}),$$

and a constant $w \geq \max\{b/(md\zeta u_0)^2, 4b/c^2\}$. For each $t > 1$, if $|\mathcal{A}(t-1)| < d\lceil d^2 w \log t\rceil$, then include t in $\mathcal{A}(t)$. Under this exploration sequence, the DSEE-based adaptive routing algorithm has regret

$$R(T) \leq Amd^3 \log T$$

for some constant A independent of d and m.

Proof: Since $\mathcal{A}(T) \leq d\lceil d^2 w \log T\rceil$, the regret caused in the exploration sequence is at the order of $md^3 \log T$. Now we consider the regret caused in the exploitation sequence. Let E_k denote the kth exploitation period which is the kth contiguous segment in the exploitation sequence. Let E_k denote the kth exploitation period. Similar to the proof of Theorem 3.1 in [5], we have

$$|E_k| \leq ht_k \tag{5}$$

for some constant h independent of d and m. Let $t_k > 1$ denote the starting time of the kth exploitation period. Next, we show that by applying the Chernoff-Hoeffding bound in (4) on the path cost, for any t in the kth exploitation period and $i = 1, \ldots, d$, we have

$$\Pr(|\mathbb{E}[C_t(\mathbf{p}_i)] - \overline{\theta}_{\mathbf{p}_i}(t)| \geq c/(2d)) \leq 2t_k^{-ab/m}.$$

To show this, we define the parameter $\epsilon_i(t) \triangleq \sqrt{b \log t/\tau_i(t)}$, where $\tau_i(t)$ is the number of times that path i has been sampled up to time t. From the definition of parameter b, we have

$$\epsilon_i(t) \leq \min\{m\zeta u_0, c/(2d)\}. \tag{6}$$

Applying the Chernoff-Hoeffding bound, we arrive at

$$\begin{aligned} & \Pr(|\mathbb{E}[C_t(\mathbf{p}_i)] - \overline{\theta}_{\mathbf{p}_i}(t)| \geq c/(2d)) \\ \leq\ & \Pr(|\mathbb{E}[C_t(\mathbf{p}_i)] - \overline{\theta}_{\mathbf{p}_i}(t)| \geq \epsilon_i(t)) \\ \leq\ & 2t_k^{-ab/m}. \end{aligned}$$

In the exploitation sequence, the expected times that at least one path in B has a sample mean deviating from its true mean cost by $c/(2d)$ is thus bounded by

$$\Sigma_{k=1}^{\infty} 2dt_k^{-ab/m} ht_k \leq \Sigma_{t=1}^{\infty} 2hdt^{1-ab/m} = gd. \tag{7}$$

for some constant g independent of d and m. Based on the property of the barycentric spanner, the best path would not be selected in the exploitation sequence only if one of the basis vector in B has a sample mean deviating from its true mean cost by at least $c/(2d)$. We thus proved the theorem. ■

We point out that under the model with link-level observability as adopted in [9], a slight modification to the proof of Theorem 1 leads to $O(md \log T)$ regret.

In Theorem 1, we need a lower bound (parameter c) on the difference in the cost means of the best and the second best paths. We also need to know the bounds on parameters ζ and u_0 such that the Chernoff-Hoeffding bound (4) holds. These bounds are required in defining w that specifies the minimum leading constant of the logarithmic cardinality of the exploration sequence necessary for identifying the best path. Similar to [5], we can show that without any knowledge of the cost models, increasing the cardinality of the exploration sequence of π^* by an arbitrarily small amount leads to a regret linear with d and arbitrarily close to the logarithmic order with time.

Theorem 2: Let $f(t)$ be any positive increasing sequence with $f(t) \to \infty$ as $t \to \infty$. Define the exploration sequence in the DSEE-based adaptive routing algorithm as follows: include t $(t > 1)$ in $\mathcal{A}(t)$ if $|\mathcal{A}(t-1)| < d\lceil f(t) \log t\rceil$. The resulting regret is given by

$$R(T) = O(d\, f(T) \log T).$$

Proof: It is sufficient to show that the regret caused in the exploitation sequence is bounded by $O(d)$, independent of T. Since the exploration sequence is denser than the logarithmic order as in Theorem 1, it is not difficult to show that the bound on $|E_k|$ given in (5) still holds with a different value of h.

We consider any positive increasing sequence $b(t)$ such that $b(t) = o(f(t))$ and $b(t) \to \infty$ as $t \to \infty$. By replacing b in the proof of Theorem 1 with $b(t)$, we notice that after some finite time T_0, the parameter $\epsilon_i(t)$ will be small enough to ensure (6) holds and $b(t)$ will be large enough to ensure (7) holds. The proof thus follows. ■

B. Heavy-Tailed Cost Distributions

We now consider the heavy-tailed cost distributions whose moment-generating functions do not exist and the moments exist up to the q-th ($q > 1$) order. From [5], we have the following bound on the deviation of the sample mean from the true mean for heavy-tailed cost distributions.

Lemma 2 [5]: Let $\{X(t)\}_{t=1}^{\infty}$ be i.i.d. random variables drawn from a distribution with the q-th moment ($q > 1$). Let $\overline{X}_t = (\Sigma_{k=1}^{t} X(k))/t$ and $\theta = \mathbb{E}[X(1)]$. We have, for all $\epsilon > 0$,

$$\Pr(|\overline{X}_t - \theta| > \epsilon) = o(t^{1-q}). \tag{8}$$

Based on Lemma 2, we arrive at the following theorem that states the sublinear regret order of the DSEE-based adaptive routing algorithm with a properly chosen exploration sequence.

Theorem 3: Construct an exploration sequence as follows. Choose a constant $v > 0$. For each $t > 1$, if $|\mathcal{A}(t-1)| < vdt^{1/q}$, then include t in $\mathcal{A}(t)$. the DSEE-based adaptive routing algorithm has regret

$$R(T) \leq DdT^{1/q}$$

for some constant D independent of d and T.

Proof: Based on the construction of the exploration sequence, it is sufficient to show that the regret in the exploitation sequence is $o(T^{1/q}) \cdot d$. From (8), we have, for any $i = 1, \ldots, d$,

$$\Pr(|\mathbb{E}[C_t(\mathbf{p}_i)] - \overline{\theta}_{\mathbf{p}_i}(t)| \geq c/(2d)) = o(|\mathcal{A}(t)/d|^{1-q}).$$

For any exploitation slot $t \in \mathcal{A}(t)$, we have $|\mathcal{A}(t)| \geq vdt^{1/q}$. We arrive at

$$\Pr(|\mathbb{E}[C_t(\mathbf{p}_i)] - \overline{\theta}_{\mathbf{p}_i}(t)| \geq c/(2d)) = o(t^{(1-q)/q}).$$

Since the best path will not be chosen only if at least one of the basis vector has the sample deviating from the true mean by $c/(2d)$, the regret in the exploitation sequence is thus bounded by

$$\sum_{t=1}^{T} o(t^{(1-q)/q}) \cdot d = o(T^{1/q}) \cdot d.$$

■

V. Conclusion

In this paper, we considered the adaptive routing problem in networks with unknown and stochastically varying link states, where only the total end-to-end cost of a path is observable after the path is selected for routing. For both light-tailed and heavy-tailed link-state distributions, we proposed a stochastic online learning algorithm that minimizes the regret in terms of both time and the network size.

References

[1] T. Lai and H. Robbins, "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4-22, 1985.

[2] T. Lai, "Adaptive Treatment Allocation and The Multi-Armed Bandit Problem," *Ann. Statist.*, vol 15, pp. 1091-1114, 1987.

[3] R. Agrawal, "Sample Mean Based Index Policies with $O(logn)$ Regret for the Multi-armed Bandit Problem," *Advances in Applied Probability*, vol. 27, pp. 1054-1078, 1995.

[4] P. Auer, N. Cesa-Bianchi, P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, vol. 47, pp. 235-256, 2002.

[5] K. Liu and Q. Zhao, "Multi-Armed Bandit Problems with Heavy Tail Reward Distributions," in *Proc. of Allerton Conference on Communications, Control, and Computing*, September, 2011.

[6] K. Liu and Q. Zhao, "Extended UCB1 for Light-Tailed Reward Distributions," available at http://arxiv.org/abs/1112.1768.

[7] V. Dani, T. Hayes, S. Kakade, "Stochastic Linear Optimization under Bandit Feedback," in *Proc. of the 21st Annual Conference on Learning Theory*, 2008.

[8] K. Liu and Q. Zhao, "Online Learning for Stochastic Linear Optimization Problems," in *Proc. of Information Theory and Applications Workshop (ITA)*, February, 2012.

[9] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial Network Optimization with Unknown Variables: Multi-Armed Bandits with Linear Rewards and Individual Observations," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, 2012.

[10] B. Awerbuch, R. Kleinberg, "Online Linear Optimization and Adaptive Routing," *Journal of Computer and System Sciences*, pp. 97-114, 2008.

[11] P. Chareka, O. Chareka, S. Kennendy, "Locally Sub-Gaussian Random Variable and the Stong Law of Large Numbers," *Atlantic Electronic Journal of Mathematics*, vol. 1, no. 1, pp. 75-81, 2006.