Operations Research Letters 51 (2023) 548–554

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Online learning for route planning with on-time arrival reliability

Hongyi Jiang [a,*], Samitha Samaranayake [a], Qing Zhao [b]

[a] School of Civil and Environmental Engineering, Cornell University, Ithaca, NY, United States of America
[b] School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, United States of America

ARTICLE INFO

Article history:
Received 30 August 2022
Received in revised form 9 August 2023
Accepted 8 September 2023
Available online 17 September 2023

Keywords:
On-time arrival
Route planning
Online learning
Stochastic networks
Unknown distributions

ABSTRACT

Consider a network where travel times on edges are i.i.d. over T rounds with unknown distributions. One wishes to choose departure times and routes sequentially between a given origin-destination pair across T rounds to minimize the expectations of: 1) number of rounds when the travel time exceeds an upper bound, and 2) summation over all rounds of the square of the difference between the given target and actual arrival times. We provide an efficient online learning algorithm for this problem.

1. Introduction

For route planning and dispatch scheduling in ride-sharing systems, the goal is to minimize the travel time as well as the wait time of both the driver and the customer. This multi-objective optimization problem can be challenging due to the fact that the travel times are stochastic with unknown probabilistic distributions.

Problem formulation: We formalize the above route planning and dispatch scheduling problem as follows. Let $G = (V, E)$ be a directed graph modeling the transportation network, where $V = (v_1, \ldots, v_{n+1})$ is the node set and $E = (e_1, \ldots, e_m)$ is the edge set. The time to traverse an edge $e \in E$ is a random variable X_e following an *unknown* distribution on $[0, 1]$, independent of the travel times over other edges. Let μ_e and σ_e^2 denote, respectively, the *unknown* expectation and *unknown* variance of X_e. A path $\mathcal{P}$ is given by an ordered sequence of directed edges, and the travel time over $\mathcal{P}$ is given by $\sum_{e\in\mathcal{P}} X_e$.

Suppose that a driver at v_1 is given a target arrival time 0 for pickup at v_{n+1}, and needs to select a *path* $\mathcal{P}$ and a *departure time* $-D$, for some $D > 0$, to go from v_1 to v_{n+1}. The goal is to minimize the second moment of the wait time subject to a minimum requirement on the efficiency of the route in terms of the expected total travel time:

$$\min_{\mathcal{P},D} \mathbb{E}\left[\left(\left(\sum_{e\in\mathcal{P}} X_e\right) - D\right)^2\right] \tag{1a}$$

$$\text{s.t.} \sum_{e\in\mathcal{P}} \mu_e \le \theta L_{\min}, \tag{1b}$$

where $L_{\min}$ is the shortest expected travel time from v_1 to v_{n+1}, i.e. $L_{\min} = \min_{\mathcal{P}} \sum_{e\in\mathcal{P}} \mu_e$, and $\theta \ge 1$ is a given hyperparameter that governs the tradeoff between the two objectives of minimizing the expected travel time and minimizing the uncertainty in wait time and can be chosen based on the underlying application. Note that the objective function captures the wait time for both the driver (when arriving before the target time 0) and the customer (when arriving after the target time 0). Due to the unknown probabilistic models, the above problem is a constrained stochastic optimization with an unknown objective and unknown constraint. Our goal is an online learning algorithm that converges to an optimal solution (defined by the known-model case) with optimal rate.

The known-model case: We now consider the case where the expectations and variances of the travel times of all links are known. This known-model case defines the benchmark performance that an online learning algorithm aims to converge to. In this case, we can minimize the expected penalty by letting $D = \sum_{e\in\mathcal{P}} \mu_e$, which then becomes $\sum_{e\in\mathcal{P}} \sigma_e^2$ for a given path $\mathcal{P}$. Then (1) can be reduced to the following constrained optimization problem known as the restricted shortest path problem (RSPP):

$$\min_{\mathcal{P}} \sum_{e\in\mathcal{P}} \sigma_e^2 \tag{2a}$$

* Corresponding author.
E-mail addresses: hj348@cornell.edu (H. Jiang), samitha@cornell.edu (S. Samaranayake), qz16@cornell.edu (Q. Zhao).

https://doi.org/10.1016/j.orl.2023.09.003
0167-6377/

$$\text{s.t.} \sum_{e\in\mathcal{P}} \mu_e \le \theta L_{\min}. \tag{2b}$$

Let $\text{opt}_{\sigma^2,\mu,\theta}$ be the optimal objective value for (2). The RSPP is known to be $\mathcal{NP}$-complete [5]. Lorenz and Raz [12] developed an approximation algorithm which, for a given approximation constant $\epsilon > 0$, finds a solution with an objective value less than $(1+\epsilon)\text{opt}_{\sigma^2,\mu,\theta}$ with a complexity of $O\left(mn\left(\log\log(n+1)+\frac{1}{\epsilon}\right)\right)$.

The online learning problem: In this work, our aim is to study the cases where μ_e, σ_e^2 and $L_{\min}$ are unknown, while θ and ϵ are given. At each round, a path from v_1 to v_{n+1} and a departure time are selected. Meanwhile, the travel time is observed at each edge of the selected path. The objective is to design a policy of sequentially selecting paths and departure times from v_1 to v_{n+1} such that the cumulative expected penalty and the expected number of rounds that violate the constraint can both be minimized.

To formalize the problem, let $X_{e,t}$ be the random outcome of the travel time on edge e in its t-th trial. Assume that the random variables in the set $\{X_{e,t} : t \ge 1\}$ are independent and identically distributed with unknown expectation μ_e and unknown variance σ_e^2. Define counter variables $T_{e,t}$ as the number of samples collected on edge e after the first t rounds. Let $\mathcal{P}_t$ be the path selected in the t-th round. Let $-D_t^*$ be the departure time in the t-th round (possibly a function of the data observed so far). For each $e \in \mathcal{P}_t$, the realization of $X_{e,T_{e,t}}$ is observed after $\mathcal{P}_t$ is traversed. Then for given $\epsilon > 0$ and $\theta \ge 1$, we define regret after T rounds as

$$\begin{aligned}\text{Reg}(T) =& \mathbb{E}\left[\sum_{t=1}^{T}\left(\left(\sum_{e\in\mathcal{P}_t} X_{e,T_{e,t}}\right) - D_t^*\right)^2\right] \\ &- T\cdot(1+\epsilon)\text{opt}_{\sigma^2,\mu,\theta}\end{aligned} \tag{3}$$

where the second term is from the known-model case as the benchmark. We also define the number of cumulative constraint violations as

$$V(T) = \sum_{t=1}^{T} \mathbb{1}\left\{\sum_{e\in\mathcal{P}_t} \mu_e > \theta L_{\min}\right\}, \tag{4}$$

where $\mathbb{1}\{\cdot\}$ is an indicator function. Note that since μ_e, σ_e^2 and $L_{\min}$ are unknown, it cannot be determined with certainty whether $\sum_{e\in\mathcal{P}_t}\mu_e \le \theta L_{\min}$ is violated for each t.

The goal of our algorithm is to sequentially select paths and departure times based on the previously observed travel times of the traversed edges so that $\text{Reg}(T)$ and $\mathbb{E}[V(T)]$ can both be minimized.

Contributions: We propose a novel formulation of the problem as a constrained online learning problem. To our knowledge, this is the first work with theoretical guarantee addressing route planning and dispatch scheduling under unknown distributions with respect to the expected travel time and the reliability of on-time arrival. (See Sec. 3 for detailed contextualization of this work.)

Our algorithm is designed to address two technical challenges: (i) The tension between the objective of minimizing wait time penalty and the constraint on the route efficiency calls for a different operating point on the exploration-exploitation tradeoff curve. Specifically, to obtain a solution for the RSPP problem, we will need the estimated shortest path length $L_{\min}$ as an input parameter. Significant estimation errors in $L_{\min}$ result in a substantial rise in $\mathbb{E}[V(T)]$. What complicates the estimation of $L_{\min}$ is that the shortest path is not the RSPP solution. The learning algorithm needs to balance the exploration and exploitation of paths under both the wait-time and travel-time metrics. Our approach is an integration of two types of exploration-exploitation control: an open-loop deterministic exploration for learning $L_{\min}$ and an adaptive confidence-bound based exploration for solving the RSPP problem. (ii) In the unknown-model case, the departure time D cannot be set to $\sum_{e\in\mathcal{P}}\mu_e$ since it is unknown. Therefore, the stochastic quadratic objective function of (1) does not reduce to the linear objective function of (2). As a result, we cannot directly apply the technique in [12] to solve the RSPP problem. As far as we are aware, there is no algorithm capable of solving or approximating the problem posed by (1). To address the issue, we determine a suitable estimation so that the regret can be upper-bounded.

We develop a provably efficient algorithm for the sequential route and departure time selection, given an oracle that provides a $(1+\epsilon)$-approximation to the RSPP. In particular, we demonstrate that the regret of our algorithm achieves the optimal logarithmic order over time while the expected cumulative constraint violations resulting from the policy are also of the logarithmic order over time. The challenge in regret analysis is in deriving an upper bound of $\mathbb{E}(V(T))$ separately from $\text{Reg}(T)$ by analyzing the conditions under which the estimate of $L_{\min}$ is of sufficient accuracy.

It is worthwhile to mention that when $\theta = 1$, the optimal solution to (2) is a path from v_1 to v_{n+1} with the shortest expected travel time. The optimal regret for finding a path with the shortest expected travel time is upper bounded by $n\mathbb{E}[V(T)]$, and is lower bounded by $\Omega(\log(T))$ as shown in [6]. Thus $\mathbb{E}[V(T)]$ cannot be better than logarithmic order in terms of T in this case.

At the end of the paper, we also discuss how our framework can be modified to tackle other routing problems.

2. Related work

There is a long line of work on stochastic online route planning problems. Online shortest path routing problems are among those that draw a lot of attention. In this problem, a directed graph is given, and the travel times on edges follow certain distributions. The decision-maker wishes to select in each round a path between the origin and destination on the graph such that the travel time on the selected path, i.e. the sum of the travel times of its component edges, is as small as possible. Numerous studies have been conducted to develop algorithms for online shortest path routing problems [1,4,11,16], to cite a few. In particular, Liu et al. [11] proposed algorithms that are built on the so-called *Deterministic Sequencing of Exploration and Exploitation* (DSEE) approach [10,17] for the problem, which divides the rounds into two interleaving sequences: an exploration sequence and an exploitation sequence. In the former, the decision maker runs the paths formed by some basis, called barycentric spanner [1], of the set of potential routes in a round-robin fashion. In the exploitation sequence, the decision maker runs the paths estimated as the optimal by linearly interpolating the estimated quality of these basis. The approach achieves logarithmic regret order over time for all light-tailed cost distributions on the edges and sublinear regrets over time for heavy-tailed cost distributions on the edges. Our algorithm in Section 3 also incorporates elements of the DSEE approach.

Online shortest path routing problems fall into the class of combinatorial multi-armed problems (CMAB). In CMAB, a decision maker selects a collection of arms per round, referred to as a super arm, which, when combined, gives the decision maker a cost at random. The cost depends on the chosen arms and the cost function that takes the chosen arms as input. Partial or full feedback from the selected arms can be provided per round to the decision maker to assist in her decision making. Regret, which is the difference between the expected total cost of the policy and that when the best arm is always selected, is used to gauge how well the policy is doing. The optimal regret that can be reached has been shown to be logarithmic over time [9]. In the setting of online

shortest path routing problems, each edge and each path between the provided origin-destination pair can be viewed as arms and super arms, respectively. Thus, the algorithms for CMAB can also be used to solve shortest path routing problems, which have been extensively studied in [2,3,18,13,14].

To our knowledge, the joint choice of arms and departure time to optimize the reliability of arrival on time, as well as the challenges mentioned above create obstacles that cannot be overcome via the approaches employed in the aforementioned research.

We found the most related work from Zhou et al. [19]. Using computational experiments, they studied maximizing the on-time arrival problem in the CMAB setting. Here we highlight the significant differences between [19] and our work: (i) In their work, the candidate paths are prespecified, and there is a finite set of candidate departure times. In this setting, each path is an arm for learning, and the number of possible choices for the decision maker is polynomial in the input size. In contrast, in our work, the candidate path set's size is exponential in the number of edges with infinite candidate departure time options, and each edge's travel time is learnt. (ii) In [19], a linear truncated function is used to measure the on-time arrival reliability. The objective function to minimize involves the measurement of on-time arrival reliability and the expected travel time, which are combined with penalty parameters. In contrast, our objective function is quadratic to measure the on-time arrival reliability with a constraint to restrict the expected travel time of the selected path. (iii) In [19], a variant of the UCB algorithm is proposed, and its efficacy is evaluated using real-time travel data in New York City without any theoretical analysis. In our work, we develop an algorithm combining two types of exploration-exploitation control with theoretical guarantees.

Nikolova et al. [15] proposed a decision-theoretic framework to define the optimal route in the presence of uncertainty. In their work, the edge travel times follow known distributions, and the cost is a non-linear function of the total travel time and departure time, including the quadratic cost function utilized in this paper. Adoption of the quadratic cost function is motivated by the need to increase the reliability of a route, as it reflects the variance. The goal is to find the path between a given origin-destination pair such that a chosen cost nonlinear function is minimized. It is shown that in this case the problem with a quadratic cost function can be converted into the classic deterministic shortest-path problem. In [15], other cost functions and complexity results are also discussed. However, their framework assumes known distributions, and there is no discussion of adding constraints to upper bound the expected travel time of the selected paths.

3. Our algorithm and analysis

In this section, we will present and analyze our algorithm. We first make a mild assumption about the graph for the rest of the paper.

Assumption 1. For each edge $e \in E$, there is a simple path from v_1 to v_{n+1} across e, where a simple path is a path that does not involve repeating nodes.

Remark 3.1. Note that for each edge e, we can use the max-flow algorithm with non-zero lower bounds on edge flows [8, Section 7.7] to find whether there is a simple path across the edge in polynomial time, as will be briefly discussed next. For each node v in the original graph G, we create two nodes $a(v)$ and $b(v)$ in the new graph G', and there is an edge from $a(v)$ to $b(v)$. For each edge from v_i to v_j in G, we build an edge from $b(v_i)$ to $a(v_j)$ in G'. Let $a(v_1)$ be the source, and $b(v_{n+1})$ be the sink. For the duplicate of edge e in G', we set its flow lower bounded by 1, while for the other edges, we set their flows upper bounded by 1. When computing a flow, the max-flow algorithm must take into account the duplicate of edge e. Therefore, the flow generated by the algorithm can be used to establish a simple path across the edge e if the max-flow value is 1.

3.1. Algorithm

One of the difficulties as we mentioned is that $L_{\min}$ is not given. Thus, we choose to do part of the explorations in a DSEE manner [17]. We call it type-1 exploration, which will be formalized below.

Definition 3.2. Given $e \in E$, a type-1 exploration for edge e is to run a simple path from v_1 to v_{n+1} across edge e selected as in Remark 3.1.

Let $\mathcal{R}$ be the set of all the simple paths from v_1 to v_{n+1}, and $\mathcal{R}_\theta := \{\mathcal{P} \in \mathcal{R} : \sum_{e\in\mathcal{P}} \mu_e \le \theta L_{\min}\}$. Let Δ_θ be the optimal value for the following problem:

$$\min_{\mathcal{P}\in\mathcal{R}\setminus\mathcal{R}_\theta} \left(\sum_{e\in\mathcal{P}} \mu_e\right) - \theta L_{\min}.$$

Note that if this is an infeasible optimization problem, i.e. $\mathcal{R}\setminus\mathcal{R}_\theta = \emptyset$, its optimal value can be defined as $+\infty$.
We are now prepared to present our algorithm, which is labeled as Algorithm 1.

Algorithm 1

For each edge e, we define the following variables:
- $T_{e,t}$ as the number of samples collected on edge e after the first t rounds.
- $\widehat{\mu}_{e,s} = \left(\sum_{j=1}^{s} X_{e,j}\right)/s$ as the average of the first s realizations of X_e we have observed.
- $\widehat{\sigma}^2_{e,s} := \sum_{j=1}^{s}(X_{e,j} - \widehat{\mu}_{e,s})^2/(s-1)$ as the estimated variance based on the first s realizations of X_e we have observed.

For each edge e, sample two paths from v_1 to v_{n+1} across e as in Remark 3.1. The departure time is set to be n.
Set constant $\epsilon \in (0, +\infty)$ and constant $\Delta \in (0, \Delta_\theta/4)$
for t from $2m+1$ to T **do**
  if there exists $e \in E$ such that $T_{e,t-1} < \frac{3n^2\theta^2 \ln t}{2\Delta^2}$ **then**
    do type-1 exploration for edge e.
  else
    for $e \in E$ **do**
      $\overline{\mu}_{e,t} \leftarrow \max\left\{0, \widehat{\mu}_{e,T_{e,t-1}} - \sqrt{\frac{3\ln t}{2T_{e,t-1}}}\right\}$
      $\overline{\sigma}^2_{e,t} \leftarrow \max\left\{0, \widehat{\sigma}^2_{e,T_{e,t-1}} - \sqrt{\frac{5\ln t}{2T_{e,t-1}}}\right\}$
    end for
    $\widehat{\mathcal{P}}^t_{\min} \leftarrow \operatorname{argmin}_{\mathcal{P}\in\mathcal{R}} \sum_{e\in\mathcal{P}} \widehat{\mu}_{e,T_{e,t-1}}$
    $U \leftarrow \theta\left(\sum_{e\in\widehat{\mathcal{P}}_{\min}} \widehat{\mu}_{e,T_{e,t-1}} + \sqrt{\frac{3\ln t}{2T_{e,t-1}}}\right)$
    $\mathcal{P}_t \leftarrow \text{Oracle}_U(\overline{\mu}_{1,t}, \ldots, \overline{\mu}_{m,t}, \overline{\sigma}^2_{1,t}, \ldots, \overline{\sigma}^2_{m,t}, \epsilon)$
      (See Definition 3.3.)
    Departure time is set to be $D^* = \sum_{e\in\mathcal{P}_t} \widehat{\mu}_{e,T_{e,t-1}}$
    Run path $\mathcal{P}_t$ with departure time D^*.
  end if
end for

Definition 3.3. $\text{Oracle}_U(\overline{\mu}_{1,t}, \ldots, \overline{\mu}_{m,t}, \overline{\sigma}^2_{1,t}, \ldots, \overline{\sigma}^2_{m,t}, \epsilon)$ finds a path by solving the following restricted shortest path problem with the $(1+\epsilon)$ approximation scheme from [12].

$$\min_{\mathcal{P}\in\mathcal{R}} \sum_{e\in\mathcal{P}} \overline{\sigma}^2_{e,t} \tag{5a}$$

$$\text{s.t.} \sum_{e\in\mathcal{P}} \overline{\mu}_{e,t} \le U. \tag{5b}$$

Remark 3.4. Note that our algorithm requires a lower bound on the parameter Δ_θ to determine the value of Δ. In Section 3.5, we

will discuss the cases where no knowledge of the parameter is available. Specifically, we can increase the frequency of the type-1 explorations by an arbitrarily small amount to achieve a regret arbitrarily close to the logarithmic order. The technique is similar to that in [11, Theorem 2].

3.2. Notations for analysis

We will use the notations from Algorithm 1 in our analysis. Furthermore, let

$$\mathcal{B} := \{\mathcal{P} \in \mathcal{R}_\theta : \sum_{e\in\mathcal{P}} \sigma_e^2 > (1+\epsilon)\text{opt}_{\mu,\sigma^2,\theta}\}.$$

For simplicity, let $r^\mu_{e,t} := \sqrt{3\ln t/(2T_{e,t-1})}$ and $r^\sigma_{e,t} := \sqrt{5\ln t/(2T_{e,t-1})}$

For the tth round, define the events

$$H_t^\mu := \{\forall e \in E, \left|\widehat{\mu}_{e,T_{e,t-1}} - \mu_e\right| \le r^\mu_{e,t}\}$$
$$H_t^\sigma := \{\forall e \in E, \left|\widehat{\sigma}^2_{e,T_{e,t-1}} - \sigma_e^2\right| \le r^\sigma_{e,t}\}$$
$$\mathcal{B}_t := \{\mathcal{P}_t \in \mathcal{B}\}$$
$$Q_t := \{\text{Do type-1 exploration at round t}\}$$

We assume that the following optimization problem has at least one solution.

$$\min_{\mathcal{P}} \sum_{e\in\mathcal{P}} \sigma_e^2 - (1+\epsilon)\text{opt}_{\sigma^2,\mu,\theta}$$
$$\text{s.t.} \sum_{e\in\mathcal{P}} \sigma_e^2 > (1+\epsilon)\text{opt}_{\sigma^2,\mu,\theta}$$
$$\sum_{e\in\mathcal{P}} \mu_e \le \theta L_{\min}$$

Note that if it does not have any solutions, it implies that $\sum_{e\in\mathcal{P}} \sigma_e^2 \le (1+\epsilon)\text{opt}_{\sigma^2,\mu,\theta}$ holds true for any path $\mathcal{P}$. In such a case, $\mathcal{B} = \emptyset$ and $P[\mathcal{B}_t] = 0$, then the proof would become straightforward.

Let Δ_ϵ be the optimal value for the optimization problem mentioned above.

3.3. The expected number of constraint violations

In this section, we assume that $\Delta_\theta < +\infty$. Otherwise, $\mathcal{R} = \mathcal{R}_\theta$, and there is no constraint violations.

To bound the expected number of constraint violations, we need the following lemma.

Lemma 3.5. *Let $U_t = \theta\left(\sum_{e\in\widehat{\mathcal{P}}^t_{\min}} \widehat{\mu}_{e,T_{e,t-1}} + r^\mu_{e,t}\right)$, and $\widehat{\mathcal{R}}_\theta = \{\mathcal{P} \in \mathcal{R} : \sum_{e\in\mathcal{P}} \overline{\mu}_{e,T_{e,t-1}} \le U_t\}$. Assume H_t^μ and $\neg Q_t$ hold. Then we have $\mathcal{R}_\theta = \widehat{\mathcal{R}}_\theta$.*

Proof. Since H_t^μ and $\neg Q_t$ hold, we have $\overline{\mu}_{e,t} \le \mu_e$, and

$$\theta\left(\sum_{e\in\widehat{\mathcal{P}}^t_{\min}} \widehat{\mu}_{e,T_{e,t-1}} + r^\mu_{e,t}\right) \ge \theta\left(\sum_{e\in\widehat{\mathcal{P}}^t_{\min}} \mu_{e,T_{e,t-1}}\right) \ge \theta \cdot L_{\min}.$$

Thus $\mathcal{R}_\theta \subseteq \widehat{\mathcal{R}}_\theta$. It remains to show $\mathcal{R}_\theta \supseteq \widehat{\mathcal{R}}_\theta$.

Let $\mathcal{P}_{\min}$ be the true shortest path. Since $\widehat{\mathcal{P}}^t_{\min}$ is the estimated shortest path, then we have the following.

$$\sum_{e\in\widehat{\mathcal{P}}^t_{\min}} \widehat{\mu}_{e,T_{e,t-1}} \le \sum_{e\in\mathcal{P}_{\min}} \widehat{\mu}_{e,T_{e,t-1}} \le \sum_{e\in\mathcal{P}_{\min}} \left(\mu_e + r^\mu_{e,t}\right).$$

Also,

$$\neg Q_t \Rightarrow T_{e,t-1} \ge \frac{3n^2\theta^2\ln t}{2\Delta^2} \Rightarrow r^\mu_{e,t} \le \frac{\Delta}{n\theta} \quad \forall e \in E.$$

Then we have

$$\theta \sum_{e\in\widehat{\mathcal{P}}^t_{\min}} \left(\widehat{\mu}_{e,T_{e,t-1}} + r^\mu_{e,t}\right)$$
$$\le \theta\left(\sum_{e\in\mathcal{P}_{\min}} \left(\mu_e + r^\mu_{e,t}\right) + \left(\sum_{e\in\widehat{\mathcal{P}}^t_{\min}} r^\mu_{e,t}\right)\right)$$
$$\le \theta L_{\min} + \theta\cdot\frac{2\Delta}{\theta} = \theta L_{\min} + 2\Delta < \theta L_{\min} + \frac{\Delta_\theta}{2}.$$

For each path $\mathcal{P} \in \mathcal{R}\backslash\mathcal{R}_\theta$, we have the following.

$$\sum_{e\in\mathcal{P}} \overline{\mu}_{e,T_{e,t-1}} = \sum_{e\in\mathcal{P}} \left(\widehat{\mu}_{e,T_{e,t-1}} - r^\mu_{e,t}\right)$$
$$\ge \sum_{e\in\mathcal{P}} \left(\mu_e - 2r^\mu_{e,t}\right) \ge \left(\sum_{e\in\mathcal{P}} \mu_e\right) - \frac{2\Delta}{\theta}$$
$$> \theta L_{\min} + \Delta_\theta - \frac{\Delta_\theta}{2} = \theta L_{\min} + \frac{\Delta_\theta}{2}$$

Thus for each path $\mathcal{P} \in \mathcal{R}\backslash\mathcal{R}_\theta$, it holds that $\mathcal{P} \notin \widehat{\mathcal{R}}_\theta$. □

The following result is proved in [2] by Hoeffding's inequality.

Lemma 3.6.

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\neg H_t^\mu\right\}\right] \le \frac{\pi^2 m}{3}. \tag{6}$$

Now we are ready to bound the expected number of rounds of violations.

Theorem 3.7. *The expected number of constraint violations is bounded by*

$$m\left\lceil\frac{3n^2\theta^2\ln T}{2\Delta^2}\right\rceil + \frac{\pi^2 m}{3} + 2m.$$

Proof. By definition, at the end of the round T, the number of all type-1 explorations is at most $m\left\lceil\frac{3n^2\theta^2\ln t}{2\Delta^2}\right\rceil$. For those rounds that do not carry out type-1 explorations, Lemma 3.5 implies that there is no violation if H_t^μ holds. We can finish the proof by Lemma 3.6 and including the first $2m$ rounds of explorations. □

3.4. The regret

Next, we will bound the regret. It consists of two parts: (1) the regret produced by choosing wrong paths; (2) the regret generated by setting the wrong departure time. We begin with (1).

The following lemma is proved in [7, Lemma 5].

Lemma 3.8. *For $0 < \delta < 1$, $t \ge 2m$, and $e \in E$, we have*

$$P\left[\left|\widehat{\sigma}^2_{e,T_{e,t-1}} - \sigma_e^2\right| \ge \sqrt{\frac{1}{2T_{e,t-1}}\ln\frac{2(t-1)^3}{\delta}}\right] \le \delta.$$

By letting $\delta = \frac{2(t-1)^3}{t^5} \le 2t^{-2}$, we have

Lemma 3.9. *For $t \geq 2m$, and $e \in E$, we have*

$$P\left[\left|\widehat{\sigma}^2_{e,T_{e,t-1}} - \sigma_e^2\right| \geq \sqrt{\frac{5\ln t}{2T_{e,t-1}}}\right] \leq 2t^{-2}.$$

Next, we will bound the expected number of rounds choosing the paths in $\mathcal{B}$ when H_t^μ and $\neg Q_t$ hold. The proof of the following result was modified from [2].

Lemma 3.10. *Let $\ell_t = \frac{5n^2\ln t}{2(\Delta_\epsilon)^2}$. We have*

$$\mathbb{E}\left[\sum_{t=2m+1}^{T} \mathbb{1}\{H_t^\mu, \neg Q_t, \mathcal{B}_t\}\right] \leq m\cdot\ell_T + \frac{m\pi^2}{3}$$

Proof. We can first derive the following bound.

$$\begin{aligned}
&\sum_{t=2m+1}^{T} \mathbb{1}\{H_t^\mu, \neg Q_t, \mathcal{B}_t\}\\
\leq &\sum_{t=2m+1}^{T} \mathbb{1}\{H_t^\mu, \neg Q_t, \mathcal{B}_t, \forall e\in\mathcal{P}_t, T_{e,t-1} > \ell_T\}\\
&+ \sum_{t=2m+1}^{T} \mathbb{1}\{\exists e\in\mathcal{P}_t, T_{e,t-1} \leq \ell_T\}\\
\leq &\sum_{t=2m+1}^{T} \mathbb{1}\{H_t^\mu, \neg Q_t, \mathcal{B}_t, \forall e\in\mathcal{P}_t, T_{e,t-1} > \ell_t\}\\
&+ \sum_{t=2m+1}^{T}\sum_{e\in E} \mathbb{1}\{T_{e,t-1} < T_{e,t}, T_{e,t-1} \leq \ell_T\}\\
\leq &\sum_{t=2m+1}^{T} \mathbb{1}\{H_t^\mu, \neg Q_t, \mathcal{B}_t, \forall e\in\mathcal{P}_t, T_{e,t-1} > \ell_t\} + m\cdot\ell_T.
\end{aligned}$$

Next we assume $t \geq 2m+1$. By the definition of $\overline{\sigma}^2_{e,t}$, $\left|\widehat{\sigma}^2_{e,T_{e,t-1}} - \sigma_e^2\right| \leq r^\sigma_{e,t}$ implies $\overline{\sigma}^2_{e,t} \leq \sigma_e^2$. Let $r_t = \sqrt{\frac{5\ln t}{2\ell_t}}$. Then

$$T_{e,t-1} > \ell_t, \forall e\in\mathcal{P}_t \Rightarrow r_t > r^\sigma_{e,t}, \forall e\in\mathcal{P}_t \tag{7}$$

$$H_t^\sigma \Rightarrow \forall e\in E, 0 \leq \sigma_e^2 - \overline{\sigma}^2_{e,t} \leq r^\sigma_{e,t} \tag{8}$$

If $\{H_t^\mu, \neg Q_t, H_t^\sigma, \mathcal{B}_t, \forall e\in\mathcal{P}_t, T_{e,t-1} > \ell_t\}$ holds at round t, we have the following derivation:

$$\begin{aligned}
&\sum_{e\in\mathcal{P}_t}(\sigma_e^2 - r_t) < \sum_{e\in\mathcal{P}_t}(\sigma_e^2 - r^\sigma_{e,t}) \leq \sum_{e\in\mathcal{P}_t}\overline{\sigma}^2_{e,t}\\
\leq &(1+\epsilon)\mathrm{opt}_{\overline{\sigma}^2,\mu,\theta} \leq (1+\epsilon)\mathrm{opt}_{\sigma^2,\mu,\theta},
\end{aligned}$$

where the first inequality is due to (7), the second inequality is by (8), the third is by the approximation ratio of the approximation algorithm in [12] and Lemma 3.5, and the last one is due to (8). It leads to

$$\sum_{e\in\mathcal{P}_t}(\sigma_e^2 - r_t) < (1+\epsilon)\mathrm{opt}_{\sigma^2,\mu,\theta}.$$

However, when $\mathcal{P}_t \in \mathcal{B}$, we have

$$\begin{aligned}
&(1+\epsilon)\mathrm{opt}_{\sigma^2,\mu,\theta} \leq \left(\sum_{e\in\mathcal{P}_t}\sigma_e^2\right) - \Delta_\epsilon\\
&= \left(\sum_{e\in\mathcal{P}_t}\sigma_e^2\right) - n\sqrt{\frac{5\ln t}{2\ell_t}} = \left(\sum_{e\in\mathcal{P}_t}\sigma_e^2\right) - nr_t\\
&\leq \sum_{e\in\mathcal{P}_t}(\sigma_e^2 - r_t) < (1+\epsilon)\mathrm{opt}_{\sigma^2,\mu,\theta},
\end{aligned}$$

which leads to a contradiction.

Thus, we have $P[\{H_t^\mu, \neg Q_t, H_t^\sigma, \mathcal{B}_t, \forall e\in\mathcal{P}_t, T_{e,t-1} > \ell_t\}] = 0$, which implies the following due to Lemma 3.9:

$$\begin{aligned}
&P[\{H_t^\mu, \neg Q_t, \mathcal{B}_t, \forall e\in\mathcal{P}_t, T_{e,t-1} > \ell_t\}]\\
=&P\left[\{H_t^\mu, \neg Q_t, H_t^\sigma, \mathcal{B}_t, \forall e\in\mathcal{P}_t, T_{e,t-1} > \ell_t\}\right]\\
&+ P\left[\{H_t^\mu, \neg Q_t, \neg H_t^\sigma, \mathcal{B}_t, \forall e\in\mathcal{P}_t, T_{e,t-1} > \ell_t\}\right]\\
\leq &\sum_{e\in E} P\left[\left|\widehat{\sigma}^2_{e,T_{e,t-1}} - \sigma_e^2\right| > \sqrt{\frac{5\ln t}{2T_{e,t-1}}}\right] \leq 2mt^{-2}.
\end{aligned}$$

Since $\sum_{t=1}^{\infty} 2mt^{-2} \leq \frac{\pi^2 m}{3}$, we can finish the proof. □

Now we can derive a bound on the regret generated after the first $2m$ rounds.

Lemma 3.11. *Let $\mathcal{P}_t$ be the path selected by Algorithm 1 at t-th round. Then we have*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=2m+1}^{T}\sum_{e\in\mathcal{P}_t}\sigma_e^2\right] - (T-2m)\cdot(1+\epsilon)\mathrm{opt}_{\sigma^2,\mu,C}\\
\leq &m\left\lceil\frac{3n^3\theta^2\ln T}{2\Delta^2}\right\rceil + \frac{2\pi^2 mn}{3} + \frac{5mn^3\ln T}{2(\Delta_\epsilon)^2}
\end{aligned}$$

Proof. Since $\sum_{e\in\mathcal{P}}\sigma_e^2 \leq n, \forall\mathcal{P}\in\mathcal{R}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=2m+1}^{T}\sum_{e\in\mathcal{P}_t}\sigma_e^2\right] - (T-2m)\cdot(1+\epsilon)\mathrm{opt}_{\sigma^2,\mu,\theta}\\
\leq &n\mathbb{E}\left[\sum_{t=2m+1}^{T}\left(\mathbb{1}\{\neg H_t^\mu\} + \mathbb{1}\{H_t^\mu, Q_t\} + \mathbb{1}\{H_t^\mu, \neg Q_t, \mathcal{B}_t\}\right)\right]\\
\leq &m\left\lceil\frac{3n^3\theta^2\ln T}{2\Delta^2}\right\rceil + \frac{2\pi^2 mn}{3} + mn\cdot\ell_T\\
= &m\left\lceil\frac{3n^3\theta^2\ln T}{2\Delta^2}\right\rceil + \frac{2\pi^2 mn}{3} + \frac{5mn^3\ln T}{2(\Delta_\epsilon)^2}.
\end{aligned}$$

□

Next, we will bound the regret generated by choosing the wrong departure time. In particular, we prove the following bound.

Lemma 3.12. *Let $\mathcal{P}_t$ be the path selected by Algorithm 1 at t-th round. Then we have*

$$\mathbb{E}\left[\sum_{t=2m+1}^{T}\left(\left(\sum_{e\in\mathcal{P}_t}\left(X_{e,t} - \widehat{\mu}_{e,T_{e,t-1}}\right)\right)^2 - \sum_{e\in\mathcal{P}_t}\sigma_e^2\right)\right] \leq mn\ln T.$$

Proof. Firstly, we have

$$\mathbb{E}\left[\sum_{e\in\mathcal{P}_t}\sigma_e^2\right] \tag{9}$$

$$=\mathbb{E}\left[\sum_{e\in\widehat{\mathcal{P}}_t}(X_{e,t}-\mu_e)^2 \mid \mathcal{P}_t=\widehat{\mathcal{P}}_t\right]P(\mathcal{P}_t=\widehat{\mathcal{P}}_t) \tag{10}$$

$$=\mathbb{E}\left[\sum_{e\in\mathcal{P}_t}(X_{e,t}-\mu_e)^2\right] \tag{11}$$

Here (10) is due to the fact that $X_{e,t}$ is independent of the selection of $\mathcal{P}_t$. Thus, we can rewrite the following term.

$$\mathbb{E}\left[\left(\sum_{e\in\mathcal{P}_t}\left(X_{e,t}-\widehat{\mu}_{e,T_{e,t-1}}\right)\right)^2-\sum_{e\in\mathcal{P}_t}\sigma_e^2\right]$$

$$=\mathbb{E}\left[\sum_{e\in\mathcal{P}_t}\left(2X_{e,t}-\widehat{\mu}_{e,T_{e,t-1}}-\mu_e\right)\cdot\left(\mu_e-\widehat{\mu}_{e,T_{e,t-1}}\right)\right]$$

$$=\mathbb{E}\left[\sum_{e\in\mathcal{P}_t}\left(\mu_e-\widehat{\mu}_{e,T_{e,t-1}}\right)^2\right]. \tag{12}$$

We can obtain (12) by the linearity of expectation and the following.

$$\mathbb{E}\left[\sum_{e\in\mathcal{P}_t}X_{e,t}\cdot\left(\mu_e-\widehat{\mu}_{e,T_{e,t-1}}\right)\mid\mathcal{P}_t=\widehat{\mathcal{P}}_t\right] \tag{13}$$

$$=\sum_{e\in\widehat{\mathcal{P}}_t}\mathbb{E}\left[X_{e,t}\cdot\left(\mu_e-\widehat{\mu}_{e,T_{e,t-1}}\right)\mid\mathcal{P}_t=\widehat{\mathcal{P}}_t\right] \tag{14}$$

$$=\sum_{e\in\widehat{\mathcal{P}}_t}\mathbb{E}\left[X_{e,t}\right]\cdot\mathbb{E}\left[\left(\mu_e-\widehat{\mu}_{e,T_{e,t-1}}\right)\mid\mathcal{P}_t=\widehat{\mathcal{P}}_t\right] \tag{15}$$

$$=\mathbb{E}\left[\sum_{e\in\mathcal{P}_t}\mu_e\cdot\left(\mu_e-\widehat{\mu}_{e,T_{e,t-1}}\right)\mid\mathcal{P}_t=\widehat{\mathcal{P}}_t\right] \tag{16}$$

(15) is due to the fact that $X_{e,t}$ is independent of $\widehat{\mu}_{e,T_{e,t-1}}$ and the selection of $\mathcal{P}_t$. Then we have

$$\mathbb{E}\left[\sum_{t=2m+1}^{T}\left(\left(\sum_{e\in\mathcal{P}_t}\left(X_{e,t}-\widehat{\mu}_{e,T_{e,t-1}}\right)\right)^2-\sum_{e\in\mathcal{P}_t}\sigma_e^2\right)\right]$$

$$=\sum_{t=2m+1}^{T}\mathbb{E}\left[\left(\sum_{e\in\mathcal{P}_t}\left(\mu_e-\widehat{\mu}_{e,T_{e,t-1}}\right)\right)^2\right]$$

$$\le n\sum_{t=2m+1}^{T}\mathbb{E}\left[\sum_{e\in\mathcal{P}_t}\left(\mu_e-\widehat{\mu}_{e,T_{e,t-1}}\right)^2\right]$$

$$=n\sum_{t=2m+1}^{T}\mathbb{E}\left[\sum_{e\in E}\left(\mu_e-\widehat{\mu}_{e,T_{e,t-1}}\right)^2\mathbb{1}\{e\in\mathcal{P}_t\}\right]$$

$$=n\sum_{e\in E}\mathbb{E}\left[\sum_{t=2m+1}^{T}\left(\mu_e-\widehat{\mu}_{e,T_{e,t-1}}\right)^2\mathbb{1}\{T_{e,t}>T_{e,t-1}\}\right]$$

$$\le n\sum_{e\in E}\mathbb{E}\left[\sum_{t=2m}^{T-1}\left(\mu_e-\widehat{\mu}_{e,t}\right)^2\right]$$

$$=n\sum_{e\in E}\sum_{t=2m}^{T-1}\frac{1}{t}\sigma_e^2\le mn\ln T.$$

Here, the last inequality is due to the fact that $\sum_{t=2}^{T}\frac{1}{t}\le\int_1^T\frac{1}{x}dx=\ln T$. □

Thus, we can derive the bound of the regret for Algorithm 1.

Theorem 3.13. *The regret for Algorithm 1 can be bounded by*

$$m\left\lceil\frac{3n^3\theta^2\ln T}{2\Delta^2}\right\rceil+\frac{2\pi^2mn}{3}+\frac{5mn^3\ln T}{2(\Delta_\epsilon)^2}+mn\ln T+2mn^2$$

Proof. The last term is from the regret generated at the first $2m$ rounds, and it is bounded by $2m\cdot n^2$ due to Remark 3.1. Then we can finish the proof by combining the results of Lemma 3.11 and Lemma 3.12. □

3.5. The cases when Δ is not attainable

To determine when to do type-1 exploration, it requires a lower bound of Δ_θ. When it is not available, we can increase the frequency of the type-1 exploration sequence by an arbitrarily small amount to achieve a regret arbitrarily close to the optimal logarithmic order, which is similar to that in [11, Theorem 2].

Specifically, let $f(t)$ be any positive increasing function such that $f(t)\to\infty$ as $t\to\infty$. If there exists $e\in E$ such that $T_{e,t-1}<n^2\theta^2 f(t)\ln t$, then we do type-1 exploration at time t. This is simply because we can find a constant T_0 such that $f(T_0)\ge\frac{3}{2\Delta^2}$. Thus, after a constant number of rounds, we will be able to apply our original analysis. Then the regret will be $O(mn^3\theta^2 f(T)\ln T)$, and the expected number of rounds violating the constraint becomes $O(mn^2\theta^2 f(T)\ln T)$.

4. Discussion

4.1. The cases when $\theta L_{\min}$ is replaced with a known constant

It can be checked that when the right-hand side of (2b), $\theta L_{\min}$, is replaced with a known constant C, and $C\ge L_{\min}$, we can simplify Algorithm 1 by eliminating type-1 explorations to solve this problem. The regret and expected number of constraint violations are still in logarithmic order in terms of the number of rounds T. However, without knowing $L_{\min}$ beforehand, it is unclear how to establish the constant C such that $C\ge L_{\min}$. Moreover, people may be unaware of the scale of C in terms of $L_{\min}$.

4.2. Extensions to other routing problems

Our framework can also be adapted to routing problems in other scenarios. Consider the following illustrative example: given the network G we created in Section 1, there is a random cost c_e for crossing each edge e (on top of the travel time), which follows a distribution on $[0,1]$. Note that the travel time and the cost can be dependent for each edge. The driver leaves for v_{n+1} from v_1, and wishes the path to be the solution of the following optimization problem.

$$\min_{\mathcal{P}}\sum_{e\in\mathcal{P}}\mathbb{E}[c_e]\qquad\text{s.t.}\sum_{e\in\mathcal{P}}\mu_e\le\theta L_{\min}.$$

Assume μ_e and $\mathbb{E}[c_e]$ are unknown. Due to the linearity of expectation, the edges can be dependent in this case. In each round, a path from v_1 to v_{n+1} is selected, and the travel time and the cost are observed at each edge of the selected path. The objective is to design a policy of sequentially selecting paths from v_1 to v_{n+1} such that the cumulative expected cost and the expected number of rounds that violate the constraint can both be minimized. To

measure them, we can define regret and the expected cumulative constraint violations similar to (3) and (4).

In this case, Algorithm 1 can be modified to solve this problem by eliminating the learning for departure time, and replacing the process of learning the variances of travel times with the expected costs of edges. It can be checked that the regret and expected number of constraint violations are still in logarithmic order in terms of the number of rounds T.

Acknowledgements

This work was partially supported by the National Science Foundation [Grants CMMI-1850422 and CMMI-2144127].

References

[1] Baruch Awerbuch, Robert Kleinberg, Online linear optimization and adaptive routing, J. Comput. Syst. Sci. 74 (1) (2008) 97–114.
[2] Wei Chen, Yajun Wang, Yang Yuan, Combinatorial multi-armed bandit: general framework and applications, in: International Conference on Machine Learning, PMLR, 2013, pp. 151–159.
[3] Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, Pinyan Lu, Combinatorial multi-armed bandit with general reward functions, Adv. Neural Inf. Process. Syst. 29 (2016).
[4] András György, Tamás Linder, Gábor Lugosi, György Ottucsák, The on-line shortest path problem under partial monitoring, J. Mach. Learn. Res. 8 (10) (2007).
[5] Refael Hassin, Approximation schemes for the restricted shortest path problem, Math. Oper. Res. 17 (1) (1992) 36–42.
[6] Ting He, Dennis Goeckel, Ramya Raghavendra, Don Towsley, Endhost-based shortest path routing in dynamic networks: an online learning approach, in: 2013 Proceedings IEEE INFOCOM, IEEE, 2013, pp. 2202–2210.
[7] Weiran Huang, Jungseul Ok, Liang Li, Wei Chen, Combinatorial pure exploration with continuous and separable reward functions and its applications (extended version), arXiv preprint, arXiv:1805.01685, 2018.
[8] Jon Kleinberg, Eva Tardos, Algorithm Design, Pearson Education India, 2006.
[9] Tze Leung Lai, Herbert Robbins, et al., Asymptotically efficient adaptive allocation rules, Adv. Appl. Math. 6 (1) (1985) 4–22.
[10] Keqin Liu, Qing Zhao, Multi-armed bandit problems with heavy-tailed reward distributions, in: 2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2011, pp. 485–492.
[11] Keqin Liu, Qing Zhao, Adaptive shortest-path routing under unknown and stochastically varying link states, in: 2012 10th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), IEEE, 2012, pp. 232–237.
[12] Dean H. Lorenz, Danny Raz, A simple efficient approximation scheme for the restricted shortest path problem, Oper. Res. Lett. 28 (5) (2001) 213–219.
[13] Nadav Merlis, Shie Mannor, Batch-size independent regret bounds for the combinatorial multi-armed bandit problem, in: Conference on Learning Theory, PMLR, 2019, pp. 2465–2489.
[14] Nadav Merlis, Shie Mannor, Tight lower bounds for combinatorial multi-armed bandits, in: Conference on Learning Theory, PMLR, 2020, pp. 2830–2857.
[15] Evdokia Nikolova, Matthew Brand, David R. Karger, Optimal route planning under uncertainty, in: Icaps, vol. 6, 2006, pp. 131–141.
[16] Mohammad Sadegh Talebi, Zhenhua Zou, Richard Combes, Alexandre Proutiere, Mikael Johansson, Stochastic online shortest path routing: the value of feedback, IEEE Trans. Autom. Control 63 (4) (2017) 915–930.
[17] Sattar Vakili, Keqin Liu, Qing Zhao, Deterministic sequencing of exploration and exploitation for multi-armed bandit problems, IEEE J. Sel. Top. Signal Process. 7 (5) (2013) 759–767.
[18] Siwei Wang, Wei Chen, Thompson sampling for combinatorial semi-bandits, in: International Conference on Machine Learning, PMLR, 2018, pp. 5114–5122.
[19] Jinkai Zhou, Xuebo Lai, Joseph Y.J. Chow, Multi-armed bandit on-time arrival algorithms for sequential reliable route selection under uncertainty, Transp. Res. Rec. 2673 (10) (2019) 673–682.