Transportation Research Part B 134 (2020) 41–63

Contents lists available at ScienceDirect

Transportation Research Part B

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

Path-based system optimal dynamic traffic assignment: A subgradient approach

Pinchao Zhang[a], Sean Qian[a,b,*]

[a] Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, United States
[b] Heinz College, Carnegie Mellon University, Pittsburgh, PA 15213 United States

ARTICLE INFO

Article history:
Received 1 December 2018
Revised 11 May 2019
Accepted 5 February 2020

Keywords:
System optimum
Dynamic traffic assignment
Subgradient
Social cost minimization
Path Marginal Cost

ABSTRACT

The system-optimal dynamic traffic assignment (SO-DTA) problem aims at solving for the time-dependent link and path flow of a network that yields the minimal total system cost, provided with the Origin-Destination (O-D) demand. The key to solving the path-based formulation of SO-DTA is to efficiently compute the path marginal cost (PMC). Existing studies implicitly assume that the total system cost (TC) is always differentiable with respect to the path flow when computing PMC. We show that the TC could be non-differentiable with respect to the link/path flow in some cases, especially when the flow is close or under the SO conditions. Overlooking this fact can lead to convergence failure or incorrect solutions while numerically solving the SO-DTA problem. In this paper we demonstrate when the TC would be indifferentiable and how to compute the subgradients, namely the lower and upper limit of path marginal costs. We examine the relations between the discontinuity of PMC and the SO conditions, develop PMC-based necessary conditions for SO solutions, and finally design heuristic solution algorithms for solving SO in general networks with multi-origin-multi-destination OD demands. Those algorithms are tested and compared to existing algorithms in four numerical experiments, two toy networks where we compare analytical solutions with numerical solutions, one small network and one sizable real-world network. We show that the proposed heuristic algorithms outperform existing ones, in terms of both the total TC, convergence, and the resultant path/link flow.

1. Introduction

The system-optimal dynamic traffic assignment (SO-DTA) problem aims at solving the time-dependent traffic flows of a network that yield minimal total system cost provided with the Origin-Destination (O-D) demand. The solution provides a benchmark to traffic management applications such as emergency evacuations, construction detour and information provision hence it has attracted significant attentions for decades. Some studies have proposed algorithms for solving the path-based SO-DTA problem that relies on the computation of path marginal cost (PMC). In those studies total system cost (TC) is assumed to be differentiable with respect to the path flow. In this paper we show that the TC is generally not differentiable under system optimum, and propose subgradient based algorithms that achieve better performance in terms of both flow solutions and computational efficiency.

Early work on SO-DTA problems (Vickrey, 1969; Arnott et al., 1990) focused on solving the problem analytically on idealized networks. Merchant and Nemhauser (1978a,b) are among the first studies solving SO-DTA on general networks. They

* Corresponding author at: Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, United States.
E-mail addresses: pinchaoz@andrew.cmu.edu (P. Zhang), seanqian@cmu.edu (S. Qian).

https://doi.org/10.1016/j.trb.2020.02.004
0191-2615/

formulated the SO-DTA problem using a link-based approach in which the base variables are link flows over time. Later studies that using similar approaches under different network settings includes (Carey, 1987; Carey and Subrahmanian, 2000; Friesz et al., 1989; Wie et al., 1994; Ziliaskopoulos, 2000). Intuitively, the SO-DTA problem can be solved directly through mathematical programming with constraints of flow dynamics. However, as pointed out by Shen et al. (2007a) and Qian et al. (2012), the main challenge of the optimization formulation (typically link-based) is the non-convexity and non-smoothness of the constraint set. Relaxations on the constraints were introduced (Carey, 1987; Ziliaskopoulos, 2000) to deal with this issue, Merchant and Nemhauser (1978a,b) formulated the problem as a non-convex programming problem using the exit flow functions for link model. By replacing non-linear equality constraints with linear inequality constraints, Ziliaskopoulos (2000) and Zhu and Ukkusuri (2013) used the kinematic wave model and formulated the SO-DTA problem as a linear programming problem. However this will lead to the "vehicle holding" issue meaning that vehicles will not enter its next link but wait even if the next link has capacity. This is unrealistic in practice. Long and Szeto (2019) addressed this issue by introducing the big M method and cast the original problem into a mixed integer linear programming (MILP) problem. However MILP itself is proved to be NP-hard and it can be mathematically intractable when the dimension of decision variables becomes large in large-scale networks. Another potential issue of the link based formulation is that the first-in-first-out (FIFO) principle can not be guaranteed hence extra constraints are required. This will introduce extra non-convex constraints when multiple Destinations exist (Carey, 1992). These two issues bring challenges on finding the true SO solutions. Recently, Long et al. (2018) proposed a new intersection-movement-based formulation and incorporated non-vehicle-holding and FIFO constraints. In this formulation, the base variables are the time-dependent flow between two links and it was shown using this formulation, an approximate SO solution without vehicle holding and FIFO violation is achieved. However,the formulation itself is a mixed integer non-linear programming. It can be mathematically intractable for large-scale networks. A agent based approach for finding the dynamic traffic assignment such that vehicular emission can be minimized was proposed in Lu et al. (2016) which can also be used to solve the SO-DTA problem. While like other optimization-based models, the solution to this approach will have the "vehicle holding" issue and hence gives a lower bound of the SO-DTA problem.

Another possible approach to formulate the SO-DTA problem is to use the path-based representation of flows initially adopted by Ghali and Smith (1995), Peeta and Mahmassani (1995) and Lo (1999). Using the path-based formulation, all constraints will be linear, thus the constraint set will be a polyhedral. However, the path cost in the objective function (namely TC) in path-based formulation is highly non-linear and thus challenging to compute. Path-based SO-DTA would need to ensure PMC is equalized for network flow across all paths and time. PMC measures the change in TC with respect to a small unit change in the flow on a specific path departing at a specific time, namely the derivative of TC over path flow.

Under the PMC-based method, the difficulties of enforcing realistic traffic flow dynamics and dealing with large-scale decision variables are transformed to the network simulation and heuristic searching of path flows (both path generation and gradient descent directions). The burden of analytical flow dynamics on the optimization models is now on those heuristics. These are two different approaches to address the SO-DTA problem. We fully acknowledge that this PMC-based approach does not utilize sufficient conditions of optimization models nor does it necessarily yield true SO solutions. However, by formulating the SO-DTA problem in terms of path flow and PMC, we are able to solve the problem through a simulation-based heuristics algorithm for large-scale networks, in a tractable manner.

Assuming the TC is differentiable and PMC can be approximated, Shen et al. (2007a) and Qian et al. (2012) showed that the path-based formulation can be cast into a variational inequalities (VI) problem. Hence, algorithms for VI can be used to solve the path-based SO-DTA problem. Therefore, the evaluation of PMC is critical in solving the path-based SO-DTA problem.

We note that beyond solving the SO-DTA problem, evaluating PMC approximately and efficiently is very useful in other network modeling/management applications, such as providing marginal cost based toll, network reliability analysis, and dynamic O-D estimation (Qian and Zhang, 2011), just to name a few. The PMC has strong policy implications to measure the negative social externalities. It also measures how sensitive the network performance or a particular group of travelers could be impacted by a marginal traveler, both temporally and spatially. In cases where a mathematical or policy problem needs to evaluate the (sub)gradient of a function with respect to path/link flow, this paper would also provide insights and knowledge.

Without closed forms, the path cost of network flow is usually evaluated through a Dynamical Network Loading (DNL) process which could be time-consuming. It is oftentimes computational heavy to evaluate PMC by doing DNL repeatedly with small flow perturbations for each time-dependent path. It is computationally infeasible for sizable networks. There are proposed approaches to improve the efficiency of network impact of marginal demand/supply changes without repeating DNL simulations (Corthout et al., 2014). However, the number of DNL needed for the path-based SO-DTA is still substantial, namely a small fraction of the number of potential paths times the number of assignment intervals for each iteration. More importantly, because PMC can be non-smooth with respect to the flow, doing so with a small flow perturbation could sometimes lead to tremendous error in PMC. Ghali and Smith (1995) and Peeta and Mahmassani (1995) evaluated PMC by summing up link marginal costs along the path according to the time-dependent link traversal times. They made the assumption that PMCs are additive which was proven to over-estimate them in dynamic network by Shen et al. (2007b). Shen et al. (2007b) and Qian et al. (2012) showed that the flow perturbation cannot achieve the downstream link until the queue on current link dissipates. A different approximation method with traffic flow modeled by point queue without diverges was proposed in Shen et al. (2007b). Qian et al. (2012) and Qian and Zhang (2011) extended this PMC approximation

method to general networks with spillbacks and general junctions. They showed that using the PMC approximation methods, the TC with respect to the SO solutions can decline effectively and converge in sizable networks.

All existing methods (Ghali and Smith, 1995; Peeta and Mahmassani, 1995; Shen et al., 2007b; Qian et al., 2012; Qian and Zhang, 2011) evaluate the PMC by implicitly applying a positive flow perturbation, with the implicit assumption that the TC is always differentiable. As will be shown in this paper later, this is not always true, especially when the flow reaches system optimum. In this study we propose a method to evaluate the subderivative-based PMC (namely the sub-gradient of TC) and show that the subderivative-based PMC is related to the first order necessary condition of system optimum. Several heuristics that deal with non-differential TC to solve for flow patterns are proposed. Numerical experiments show that the sub-gradient approach can solve the SO-DTA problem resulting solutions close to the analytical SO solutions in the demonstration networks. For a sizable network, we show the subgradient based approach is more efficient and the resultant flow solution does not oscillate over time as exhibited in the case where PMC is assumed to be continuous.

The rest of this paper is organized as follows: Section 2 presents the notations used in this paper and the formulation of the path-based SO-DTA problem. In Section 3, we show when the TC is not differentiable and how to evaluate PMC in these cases. The SO-DTA problem is then formulated in terms of VI with subgradients of TC in Section 4. The SO solutions of two demonstration networks will be solved analytically in Section 5. In Section 6, we propose several SO-DTA algorithms that using heuristics. Two numerical experiments are reported in Section 7 to compare the performance of proposed algorithms with existing algorithms. Section 7 concludes this paper.

2. Notations and formulations

The SO-DTA problem can be formulate in either continuous time or discrete time intervals. Ma et al. (2014, 2017) use the continuous-time setting and formulate the SO-DTA problem as optimal control problems. However, solving continuous-time models in large-scale networks could be challenging and hence numerical solution algorithms are usually developed based on discretized time intervals. Ma et al. (2014, 2017). Another approach is to directly formulate the SO-DTA in discrete time intervals in the first place (e.g. Ziliaskopoulos, 2000; Shen et al., 2007a; Qian et al., 2012; Long et al., 2018). In this paper we will take the latter formulation.

Suppose a general roadway network consists of a set of nodes, N, and a set of links, A. Let a denote the link index, $a \in A$. Let $\mathbb{R}$ and $\mathbb{S}$ denote the set of origin nodes and destination nodes, respectively. $r-s$ represents an O-D pair, where $r \in \mathbb{R}$ and $s \in \mathbb{S}$. Define T_d as the assignment horizon, $T_d = \{1, 2, \ldots, T\}$. Time is drawn from a discrete set $t \in T_d$. K_t^{rs} and q_t^{rs} is the set of paths and O-D demand for an O-D pair $r-s$ departing at time t, respectively. Q^{rs} denotes the total O-D demand for an O-D pair $r-s$. The travel cost of commuters departing at time t on path p of O-D pair rs, $c_p^{rs}(t)$, consists of actual travel time, $w_p^{rs}(t)$, and schedule delay cost,

$$c_p^{rs}(t) = \begin{cases} \alpha w_p^{rs}(t) + \beta[t_{rs}^* - \Delta_{rs} - t - w_p^{rs}(t)] & t_{rs}^* - \Delta_{rs} > t + w_p^{rs}(t) \\ \alpha w_p^{rs}(t) & t_{rs}^* - \Delta_{rs} \le t + w_p^{rs}(t) \le t_{rs}^* + \Delta_{rs} \\ \alpha w_p^{rs}(t) + \gamma[t + w_p^{rs}(t) - t_{rs}^* - \Delta_{rs}] & t + w_p^{rs}(t) > t_{rs}^* + \Delta_{rs} \end{cases} \tag{1}$$

where $[t_{rs}^* - \Delta_{rs}, t_{rs}^* + \Delta_{rs}]$ is the targeted arrival time window for commuters of O-D pair rs. α, β and γ are the weighting scalars of travel time, early arrival and late arrival, respectively. Let f_{pt}^{rs} denote the path flow on path p of O-D pair rs departing at time t and $\mathbf{f} = \{f_{pt}^{rs}\}_{r,s,p,t}$ be the path flow vector (pattern) consisting of path flows across all O-D pairs and departure times. Note that in cases without departure time choice we can simply remove the second term on the right side of Eq. (1). In the reminder of this paper, we use c_{pt}^{rs} instead of $c_p^{rs}(t)$ for clarity. But they are equivalent, both of which are based on discrete time intervals.

A path-based SO-DTA problem optimizing the total travel cost (TC) reads:

Model M1 (with both routes choices and departure time choices):

$$\min \mathrm{TC}(\mathbf{f}) = \sum_{rs \in RS} \sum_{t \in T_d} \sum_{p \in K_t^{rs}} f_{pt}^{rs} c_{pt}^{rs}(\mathbf{f}) \tag{2a}$$

$$s.t. \sum_{t \in T_d} \sum_{p \in K_t^{rs}} f_{pt}^{rs} = Q^{rs}, \ \forall rs \tag{2b}$$

$$f_{pt}^{rs} \ge 0, \forall rs, \forall p, \forall t \tag{2c}$$

Model M2 (with route choices only):

$$\min \mathrm{TC}(\mathbf{f}) = \sum_{rs \in RS} \sum_{t \in T_d} \sum_{p \in K_t^{rs}} f_{pt}^{rs} c_{pt}^{rs}(\mathbf{f}) \tag{3a}$$

$$s.t. \sum_{p \in K_t^{rs}} f_{pt}^{rs} = q_t^{rs}, \ \forall drs, t \tag{3b}$$

$$f_{pt}^{rs} \ge 0, \forall rs, \forall p, \forall t \tag{3c}$$

Path marginal cost (PMC) is precisely the change in the total cost (TC) with respect to one marginal unit change in path flow departing at time t on path p of O-D pair rs. When TC is non-differentiable, the absolute change in TC can be different if one reduces or increases one marginal unit of path flow f_{pt}^{rs}. In general, PMC, in a sub-gradient form, can be represented by a closed interval defined by an upper bound and a lower bound of PMC as follows,

$$\mathrm{PMC}_{pt}^{rs}(\mathbf{f}) \in [\mathrm{PMC}_{pt}^{rs}(\mathbf{f})^{-}, \mathrm{PMC}_{pt}^{rs}(\mathbf{f})^{+}] \tag{4a}$$

$$\mathrm{PMC}_{pt}^{rs}(\mathbf{f})^{+} = \lim_{x \downarrow f_{pt}^{rs}} \frac{TC(\tilde{\mathbf{f}}) - TC(\mathbf{f})}{x - f_{pt}^{rs}} \tag{4b}$$

$$\mathrm{PMC}_{pt}^{rs}(\mathbf{f})^{-} = \lim_{x \uparrow f_{pt}^{rs}} \frac{TC(\tilde{\mathbf{f}}) - TC(\mathbf{f})}{x - f_{pt}^{rs}} \tag{4c}$$

$\tilde{\mathbf{f}}$ denotes the new path flow pattern which reproduces the path flow vector $\mathbf{f}$ except that the element f_{pt}^{rs} of the vector $\mathbf{f}$ is replaced by x.

Note that for a general function, its right derivative is not necessarily less than its left derivative. However, in the case of PMC, it always holds and hence yielding Eq. (4a). This will be shown in Section 3.

Shen et al. (2007b), Nie and Zhang (2010) and Qian et al. (2012) have formulated the path-based SO-DTA problem (2) in a VI problem, i.e. find $\mathbf{f}^* \in \Omega$ defined by Eqs. (2b), (2c) such that,

$$\sum_{rs \in RS} \sum_{t \in T_d} \sum_{p \in K_t^{rs}} \mathrm{PMC}_{pt}^{rs}(\mathbf{f}^*)(f_{pt}^{rs} - f_{pt}^{rs*}) \geq 0, \forall \mathbf{f} \in \Omega \tag{5}$$

Eq. (5) is derived to ensure that PMC is equalized among all time-dependent paths for each O-D pair whenever there is positive flow on a path. Its proof relies on the assumption that TC is differentiable everywhere and any PMC is single-valued. Thus, Eq. (5) does not hold in general. In next section we will revisit this problem, and a general form of the VI formulation will be presented.

This paper will later work with a Mesoscopic dynamic network loading (DNL) process that computes the path cost $\mathfrak{c}$ and TC. However, the method of PMC subgradient approximation can be generally applied to any analytical or simulation process that results in link based cumulative flow curves, where link cost, path cost and TC can be extracted and computed. Link based cumulative flow curves record the number of vehicles arriving and departing each link by each loading interval (typically a few seconds in the DNL), denoting as $\mathbb{A}(t)$ and $\mathbb{D}(t)$. When the First-In-First-Out (FIFO) was met (or assumed), we can retrieve the time-varying traversal time $\tau(t)$ of each link by,

$$\tau(t) = \arg\min_{k}(\mathbb{A}(t) - \mathbb{D}(k)) - t \tag{6}$$

Traffic dynamics in DNL, including models that encapsulate flow propagations across nodes and links, are crucial to DNL models. The implementation of DNL is based on the polymorphic dynamic network loading model in Nie (2006). It is a generic DNL model that can accommodate most commonly-used link models and node models. The FIFO principle on links is automatically maintained in the implementation. In this paper we adopt the classic LWR model with Cell Transmission Model (CTM) representation (Daganzo, 1994; 1995) as the link model in general, but will start presenting the PMC concepts with the point queue model. The DNL model is also generic with any node models such as the ones proposed in Jin and Zhang (2004), Ni and Leonard II (2005) and Tampère et al. (2011). We will use a general node model proposed in Nie (2006) and Nie and Zhang (2010). As shown later, the process of evaluating link marginal cost is completely based on the cumulative arrival/departure curves (also called inflow/outflow curves) of flow for each link. The only requirement here is FIFO on the link, so that the time-dependent travel time can be uniquely determined.

For a general junction connecting with n links, denote the number of vehicles from link i to link j within time interval t as $v_{ij}(t)$. The demand (denoted as $D_i(t)$) and supply (denoted as $S_i(t)$) of a link at time t is defined as the maximum number of vehicles that are able to leave and enter this link. When the proportion of vehicles heading from link i to link j, a_{ij} is known for every link pairs, demand and supply can be computed based on the fundamental diagrams. Then v_{ij} reads (Nie and Zhang, 2010),

$$\textit{virtual demand } \bar{d}_i(t) = \min(D_i(t), \min_j\{\frac{S_j(t)}{a_{ij}(t)}\}) \tag{7}$$

$$\textit{virtual supply } \bar{s}_i(t) = \min(S_i(t), \sum_j a_{ji}(t) D_j(t)) \tag{8}$$

$$v_{ij}(t) = \min(\bar{d}_i(t) a_{ij}(t), \bar{s}_i(t) \frac{\bar{d}_i(t) a_{ij}(t)}{\sum_k \bar{d}_k(t) a_{kj}(t)}) \tag{9}$$

3. Approximating sub-gradient based path marginal cost

This section demonstrates TC can be indifferentiable using two simple examples, followed by a general approach for general networks. To demonstrate the concepts more clearly, we do not discuss the derivatives of the schedule delay in this subsection, since it can be computed in a similar fashion as for the travel time (Qian et al., 2012).

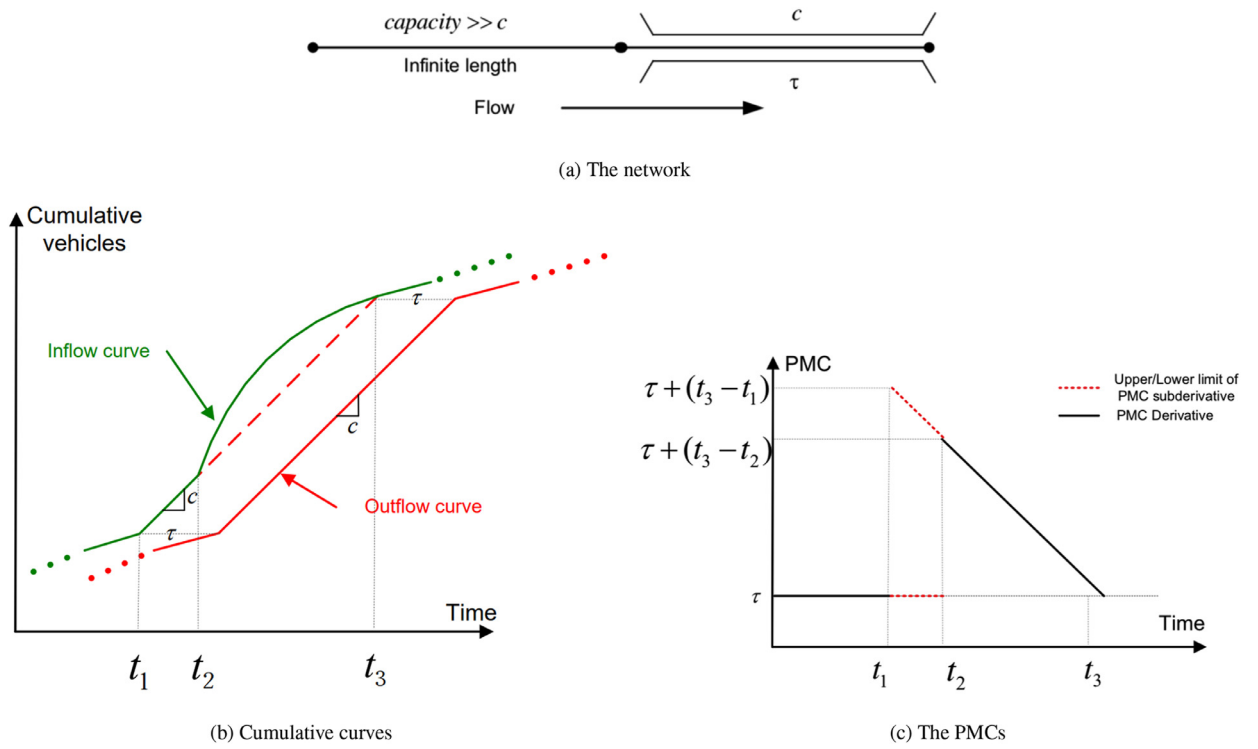


Fig. 1. The case of a single bottleneck.

3.1. A single bottleneck

First consider a small network with two links as in Fig. 1a. The flow comes from the left link and goes through the right link. The capacity of the first/left link is much larger than the right/second link and the first link has infinite length which means no queue spillback will generate from this link to its upstream links. The free flow travel time of the second link is τ and the capacity is c. We want to compute the path marginal cost of the second link.

The free flow travel time is τ and the capacity of the bottleneck is c. When the bottleneck is not congested, the PMC of any flow is always equal to τ. If a flow departing at time t hits the bottleneck when there is a queue Fig. 1b (namely after t_2), neither positive and negative flow perturbation at time t will shift the departure curve until the queue dissipates. In these two cases the TC is differentiable.

TC is not differentiable when the inflow rate equals exactly to the capacity flow rate. Consider a time-dependent flow as shown in Fig. 1b. Before t_1 the inflow rate is smaller than c and the PMC is τ. The inflow rate equals capacity c between t_1 and t_2 and exceeds the capacity after t_2 and a queue forms at the bottleneck. From some time after t_2 the inflow rate starts to decrease. When the inflow rate becomes smaller than the bottleneck rate, the queue starts to dissipate. Finally the queue vanishes at t_3. One can verify that the PMC due to a flow perturbation at $t \in [t_2, t_3]$ reads

$$\mathrm{PMC}_t(\mathbf{f}) = \tau + t_3 - t \tag{10}$$

which is the same as the result from Shen et al. (2007b).

The interesting part is between $t \in [t_1, t_2]$. The PMC_t^+ here is the same as Eq. (10). However if one marginal unit of vehicle was taken out of the link at time $t \in [t_1, t_2]$, then the cumulative inflow curve after t will shift down for one unit until the queue vanishes and so will the cumulative outflow curve. The area change due to the shift of the two curves will cancel out each other and the PMC_t^- will be τ, different from PMC_t^+. Hence in this interval the TC is not differentiable.

PMC can be summarized as shown in Fig. 1c or as follows,

$$\mathrm{PMC}_t(\mathbf{f}) = \begin{cases} \tau & t < t_1 \\ [\tau, t_3 + \tau - t] & t \in [t_1, t_2) \\ t_3 + \tau - t & t \in [t_2, t_3) \\ \tau & t \geq t_3 \end{cases} \tag{11}$$

3.2. Tandem bottlenecks

Now consider a toy network with two tandem bottlenecks as shown in Fig. 2a. The link A has infinite length, but constrained by its outflow rate on Link O, namely the capacity c_1 and free flow travel time τ_1. The downstream link B has also sufficient holding capacity, with an outflow capacity c_2 by Link O'. The free flow travel time on link O' is τ_2. Note that the inflow rate of link B is constrained by the capacity on link O. If $c_1 \le c_2$ then queue will never form on link B. The PMC will be simply the one we computed in single bottleneck case plus τ_2. Here we focus on the cases of $c_1 > c_2$.

First we consider the case without spillback (namely under the point queue model). We assume that link B is long enough or the inflow rate is not so large that the queue on link B will never spillback to link O and link A. When the network is in free flow state or has queues the TC will be differentiable and the computation of PMC is the same as in Shen et al. (2007b). The case where the inflow rates of both link equals to their respective capacities should be examined.

Suppose a demand pattern as shown in Fig. 2b. The inflow rate q_t^A of link A is: Before t_1^A and after t_4^A the inflow rate q_t^A is smaller than c_2. Between t_1^A and t_2^A we have $q_t^A = c_2$. q_t^A equals to c_1 when $t \in [t_2^A, t_3^A)$. The inflow rate exceeds c_1 after t_3^A and a queue will form on A which vanishes at t_4^A. After t_4^A we have $q_t^A < c_2$. Given the inflow pattern, the outflow rate of link A will be c_2 between $t_1^A + \tau_1$ and $t_2^A + \tau_1$, and c_1 between $t_2^A + \tau_1$ and $t_4^A + \tau_1$. The inflow curve of link B is constrained by the outflow curve of link A. Hence the cumulative curves for link B is: Between $t_1^B = t_1^A + \tau_1$ and $t_2^B = t_2^A + \tau_1$ the inflow rate of link B $q_t^B = c_2$. After t_2^B, a queue will form on link B which vanishes at time t_5^B, some time after $t_4^A + \tau_1$. Fig. 2b and d show the cumulative curves of link A and link B respectively.

The link marginal cost of link A, LMC_t^A (defined as the change of the total cost of a link with respect to a unit flow change in the path flow), is the same as the single bottleneck case:

$$\mathrm{LMC}_t^A(\mathbf{f}) = \begin{cases} \tau_1 & t < t_2^A \\ [\tau_1, \tau_1 + t_4^A - t] & t \in [t_2^A, t_3^A) \\ \tau_1 + t_4^A - t & t \in [t_3^A, t_4^A) \\ \tau_1 & t \ge t_4^A \end{cases} \tag{12}$$

Now consider link B. First examine the upper limit of LMC. If an additional vehicle enters link A at $t < t_1^A$ or $t > t_4^A$, it will experience no queue on both links thus $\mathrm{LMC}_t^B(\mathbf{f}) = \tau_2$. If the vehicle enters link A between t_2^A and t_4^A, the perturbation will not arrive link B until $t_4^B = t_4^A + \tau_1$ when the queue on link A vanishes thus $\mathrm{LMC}_t^B(\mathbf{f})^+ = t_5^B + \tau_2 - t_4^B$. If the perturbation enters link A between t_1^A and t_2^A, since q_t^A is smaller than c_1 the perturbation can travel to link B immediately at $t + \tau_1$. Then we have $\mathrm{LMC}_t^B(\mathbf{f})^+ = t_5^B + \tau_2 - t - \tau_1$.

The lower limit of the LMC of link B is the same as the upper LMC when $t < t_1^A$ or $t > t_3^A$. If the flow reduces by one unit between t_2^A and t_3^A, then on link B, the flow reduces by one unit right at $t + \tau_1$ between t_2^B and t_3^B. But the departure curve of link B will not shift due to the queue on link B thus $\mathrm{LMC}_t^B(\mathbf{f})^- = t_5^B + \tau_2 - t - \tau_1$. If the flow reduces by one unit on link A between t_1^A and t_2^A, then at time $t + \tau_1$, both the inflow and outflow curves of link B will shift down for one unit, which leads to $\mathrm{LMC}_t^B(\mathbf{f})^- = \tau_2$. Hence the LMC of link B,

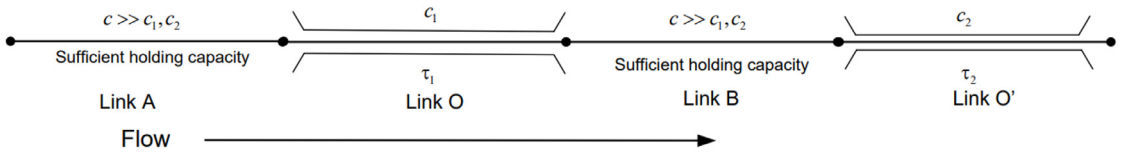
$$\mathrm{LMC}_t^B(\mathbf{f}) = \begin{cases} \tau_2 & t < t_1^A \\ [\tau_2, \tau_2 + t_5^B - t - \tau_1] & t \in [t_1^A, t_2^A) \\ [\tau_2 + t_5^B - t - \tau_1, \tau_2 + t_5^B - t_4^A - \tau_1] & t \in [t_2^A, t_3^A) \\ \tau_2 + t_5^B - t_4^A - \tau_1 & t \in [t_3^A, t_4^A) \\ \tau_2 + t_5^B - t - \tau_1 & t \in [t_4^A, t_5^A) \\ \tau_2 & t \ge t_5^A = t_5^B - \tau_1 \end{cases} \tag{13}$$

Note that in the interval $[t_2^A, t_3^A)$, the lower LMC limit of link B is actually greater than upper LMC limit (see Fig. 2e), with respect to the entering time to link A. Adding up the Eqs. (12), (13) gives the PMC,

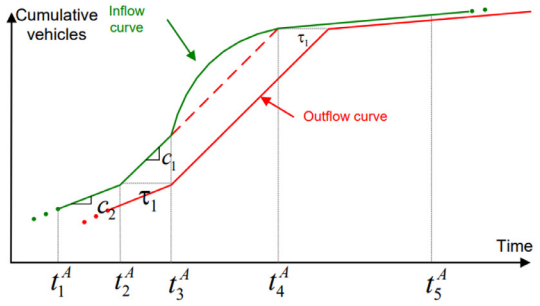
$$\mathrm{PMC}_t(\mathbf{f}) = \begin{cases} \tau_1 + \tau_2 & t < t_1^A \\ [\tau_1 + \tau_2, t_5^B + \tau_2 - t] & t \in [t_1^A, t_2^A) \\ t_5^B + \tau_2 - t & t \in [t_2^A, t_5^A) \\ \tau_1 + \tau_2 & t \ge t_5^A \end{cases} \tag{14}$$

The total PMC is shown in Fig. 2f and the link marginal costs (LMCs) of the two links are shown in Fig. 2c and e. It is interesting to find that though the TC of each link are not differentiable when $t \in [t_2^A, t_3^A)$, the added TC of both links is differentiable. Comparing with the results of a single bottleneck Eq. (14), one can find that the case of two tandem bottlenecks is equivalent to the case of a single bottleneck with the capacity c_2 and free flow travel time $\tau_1 + \tau_2$, regardless of the first link. This implies that the PMC of a network is determined by the most downstream active bottleneck. In Shen et al. (2007b) and Qian et al. (2012) a similar finding is reported when TC is differentiable.

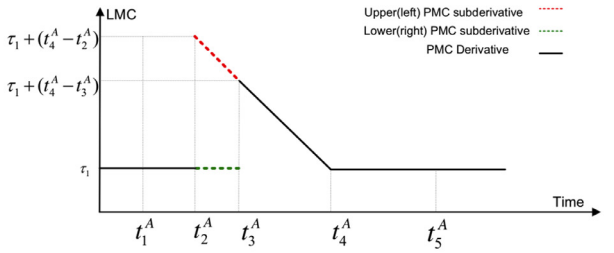
Now consider the case under LWR model with spillover, under which a queue on link B can spillback to link A. Assume the same demand pattern as for the point queue case applies. The cumulative inflow curves remain the same for link A,



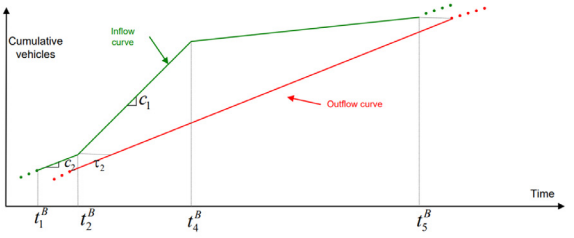
(a) The network



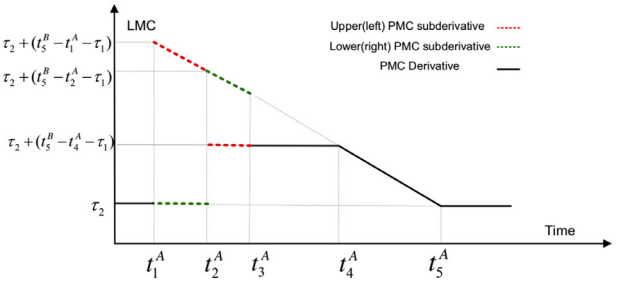
(b) Cumulative curve on link A



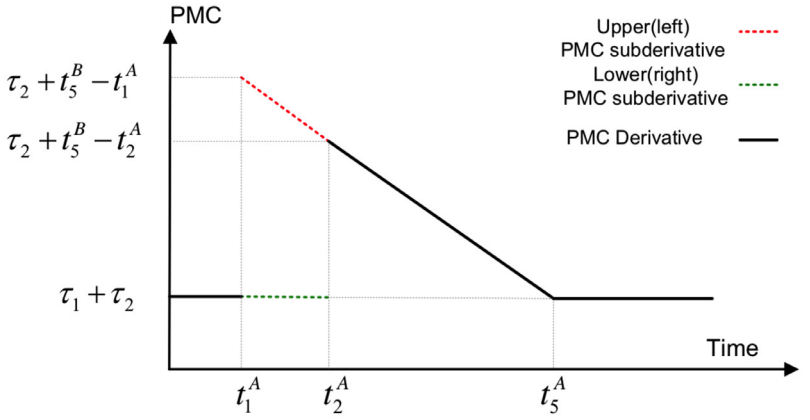
(c) The link marginal cost on link A



(d) Cumulative curve on link B



(e) The link marginal cost on link B



(f) The path marginal cost through the two bottlenecks

Fig. 2. Two tandem bottlenecks.

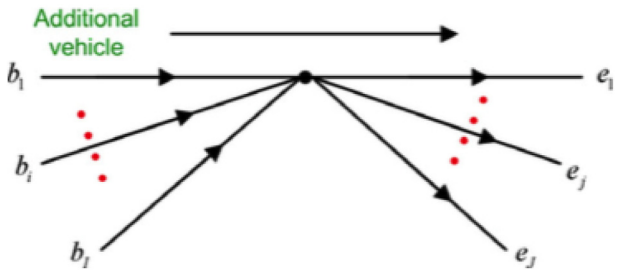


Fig. 3. A general junction (figure from Qian et al., 2012).

except for that a queue forms upstream of link A. Suppose at some time $t+\tau_1 \in [t_2^A+\tau_1, t_5^A+\tau_1]$ that a queue on link B has occupied its physical space, and spill back to link A. Because of the spillback, the outflow rate of link A will drop from c_1 down to c_2 after $t+\tau_1$, so will the inflow rate of link B. One can find that the queue spillover is essentially seen as the interaction between link A and link B, which does not change the flow PMC of the two links combined. The PMC over time resembles Fig. 2f, solely controlled by the most downstream bottleneck.

3.3. Computation of PMC in general case

We show that TC is not differentiable when the following two conditions are met for a path flow,

1. A downstream bottleneck is active on this path for this path flow, namely the discharging flow rate of this bottleneck is precisely its maximum flow rate;
2. The path travel time equals to its free-flow travel time.

If at least one of the two conditions does not hold, the PMC approximation methods in existing studies are still valid. Otherwise, the PMC will be subderivative-based. Let t_{in}^e be the time when the flow perturbation enters link e. $t_{out,+}^e$ is the time when the flow perturbation leaves the current link e if we consider the flow perturbation to be positive. $t_{out,-}^e$ is the time when the flow perturbation leaves the current link e when we consider the flow perturbation to be negative. When link e has an active bottleneck (namely the discharging flow rate of this bottleneck is precisely its maximum flow rate), then $t_{out,+}^e$ is the time when the bottleneck becomes inactive thereafter, since the departure flow from the link e will not see the perturbation till the discharging rate drops below the capacity flow. This is true for $t_{out,-}^e$ with an additional condition: at time t_{in}^e on link e, there exists a queue. If no queue exists (namely the link travel time equals to the free flow travel time), then $t_{out,-}^e$ is simply the departure time of flow arriving at link e at time t_{in}^e, not the time when the perturbation is observed. This can occur when the bottleneck is active and no queue exists, mostly likely under the system optimum.

For a path $p=\{e_1, e_2, \ldots, e_n\}$ without merge or diverge junctions, the PMC reads,

$$PMC_{pt}^{rs}(\mathbf{f})^+ = \sum_i (t_{out,+}^{e_i} - t_{in,+}^{e_i}) + \sum_i fft_{e_i} \tag{15}$$

$$PMC_{pt}^{rs}(\mathbf{f})^- = \sum_i (t_{out,-}^{e_i} - t_{in,-}^{e_i}) + \sum_i fft_{e_i} \tag{16}$$

where fft_e is the free flow travel time on link e. Note that $t_{in,+}^e$, $t_{in,-}^e$, $t_{out,+}^e$ and $t_{out,-}^e$ should be identified from the cumulative curves by tracing each link sequentially along the path. Namely $t_{in,+}^{e_{i+1}} = t_{out,+}^{e_i}$, $t_{in,-}^{e_{i+1}} = t_{out,-}^{e_i}$. If the TC is not differentiable, $t_{in,+}^{e_i} \neq t_{in,-}^{e_i}$.

For a general network with merge/diverge junctions, the results from Qian et al. (2012) with minor modification are sufficient. First consider a general junction under the point queue model with L incoming links $\{b_1, b_2, \ldots, b_l\}$ and J outgoing links $\{e_1, e_2, \ldots, e_j\}$ as shown in Fig. 3. The path that we are tracing goes through link b_1 and e_1. The flow perturbation on link b_1 and e_1 will never affect the flows on links $b_2, b_3, \ldots, b_l$. In other words the link marginal cost of links $b_2, b_3, \ldots, b_l$ w.r.t. the flow perturbation on the path passing b_1 and e_1 are 0.

If at time t link b_1 does not have an active bottleneck, then we can simply set the departure time from b_1 is the entering time of the flow perturbation on link e_1. If b_1 has an active bottleneck, we need to determine if vehicles on link b_1 head for links $e_2, e_3, \ldots, e_j$ are delayed by the flow perturbation. If the flow heading from b_1 for e_1 is the greatest of all flows from b_1 to any e_i, then the LMC on $e_2, e_3, \ldots, e_j$ is assumed to remain the same. If this flow is not the greatest, without loss of generality, assume the flow from b_1 to e_m is the greatest. The LMC on all other downstream links except e_m is assumed to remain the same. Then the link marginal cost for the two links is,

$$LMC_{b_1,t} = (t_{out}^{b_1} - t_{in}^{b_1}) + fft_{b_1} \tag{17}$$

$$LMC_{e_m,t} + LMC_{e_1,t} = (t_{out}^{e_1} - t_{in}^{e_1}) + fft_{e_1} + \frac{1}{c_{b_1}} N'_{b_1,e_m} \tag{18}$$

where N'_{b_1,e_m} is the number of vehicles arriving on link e_m from b_1 between $t_{in}^{e_m}$ and $t_{out}^{e_1}$ that do not encounter any downstream congested links.

Qian et al. (2012) shows that under the LWR model for this general junction, the approximation of PMC is actually easier. Two cases can occur: (1) link b_1 is an active bottleneck and all outgoing links are not; (2) one of the outgoing links is an active bottleneck which lead to the queue spillover to all incoming links including link b_1. In both cases, Eq. (17) remains the same, and LMC of all outgoing links is zero except for link e_1,

$$LMC_{e_1,t} = (t_{out}^{e_1} - t_{in}^{e_1}) + fft_{e_1} \tag{19}$$

Note that when the TC is not differentiable, there could exist the upper and lower limits of the LMC. We would need to substitute t_{out}^l (t_{in}^l) in Eq. (17) to (19) with $t_{out,+}^l$($t_{in,+}^l$) or $t_{out,-}^l$ ($t_{in,-}^l$), where l is a link on the path. $t_{in,+}^e$, $t_{in,-}^e$, $t_{out,+}^l$ and $t_{out,-}^l$ should be identified from the cumulative curves by tracing each link sequentially along the path. When doing this sequentially, $t_{in,+}^l$ and $t_{in,-}^l$ become different in the first place when (under the LWR model),

1. The discharging flow rate of this link l equals to its maximum capacity flow rate; or this link l has a queue spillover from one of its downstream links (but the discharging flow rate can be less than its maximum capacity flow rate);
2. The link travel time equals to its free-flow travel time.

When $t_{in,+}^l \neq t_{in,-}^l$ starting from link l, all flow perturbation times of any downstream links of link l with sequential tracing may be different, and the difference is likely to propagate over links.

4. First order conditions of SO and the VI formulation

At the system optimum (SO), the total travel cost cannot be further reduced by changing any traveler(s)' departure time or routes. We have the following proposition for a necessity condition of SO:

Proposition 1. A necessary condition of SO: *If* $\mathbf{f}$ *is an path flow SO solution, then for any O-D pair* $r-s$*, any path* p*,* p' *and any departure time* t*,* t'*:*

$$f_{pt}^{rs} > 0 \Rightarrow \begin{cases} \min_{p',t'} \mathrm{PMC}_{p't'}^{rs}(\mathbf{f})^+ \geq \mathrm{PMC}_{pt}^{rs}(\mathbf{f})^- \text{ (with departure time choices)} \\ \min_{p'} \mathrm{PMC}_{p't}^{rs}(\mathbf{f})^+ \geq \mathrm{PMC}_{pt}^{rs}(\mathbf{f})^- \text{ (without departure time choices)} \end{cases} \tag{20}$$

Proof. (Only for the M1 model with departure time choices) Suppose there exists an O-D pair $r-s$, two paths $p \neq p'$ and departure time $t \neq t'$ such that $f_{pt}^{rs} > 0$ and $\mathrm{PMC}_{p't'}^{rs}(\mathbf{f})^+ < \mathrm{PMC}_{pt}^{rs}(\mathbf{f})^-$. Then we can always find a small $0 < \delta < f_{pt}^{rs}$ so that we can construct a new path flow pattern $\tilde{\mathbf{f}}$ which resembles $\mathbf{f}$ except that the element f_{pt}^{rs} is replaced by $f_{pt}^{rs} - \delta$ and the element $f_{p't'}^{rs}$ is replaced by $f_{p't'}^{rs} + \delta$. The new flow pattern $\tilde{\mathbf{f}}$ is still feasible since $f_{pt}^{rs} - \delta > 0$ and $\tilde{\mathbf{f}} \in \Omega$. However,

$$TC(\tilde{\mathbf{f}}) - TC(\mathbf{f}) = \delta(\mathrm{PMC}_{p't'}^{rs}(\mathbf{f})^+ - \mathrm{PMC}_{pt}^{rs}(\mathbf{f})^-) < 0$$

implying that moving δ amount of travelers from path p and departure time t to p' and t' can reduce the TC, which contradicts with the SO (namely TC is minimized). □

Proposition 1 can be explained as the first order condition of the SO-DTA problem. The Lagrangian of the M1 model is as follows:

$$L(\mathbf{f}, \mu, \mathbf{v}) = TC(\mathbf{f}) + \sum_{rs \in RS} \mu^{rs}(Q^{rs} - \sum_{t \in T_d} \sum_{p \in K_t^{rs}} f_{pt}^{rs}) - \sum_{rs \in RS} \sum_{t \in T_d} \sum_{p \in K_t^{rs}} v_{pt}^{rs} f_{pt}^{rs} \tag{21}$$

and the Karush-Kuhn-Tucker (KKT) conditions,

$$0 \in \frac{\partial TC(\mathbf{f})}{\partial f_{pt}^{rs}} - \mu^{rs} - v_{pt}^{rs}, \forall rs \in RS, p \in K_t^{rs}, t \in T_d \tag{22}$$

$$v_{pt}^{rs} f_{pt}^{rs} = 0, \forall rs \in RS, p \in K_t^{rs}, t \in T_d \tag{23}$$

$$Q^{rs} - \sum_{t \in T_d} \sum_{p \in K_t^{rs}} f_{pt}^{ts} = 0, f_{pt}^{ts} \geq 0, \forall rs \in RS, p \in K_t^{rs}, t \in T_d \tag{24}$$

$$v_{pt}^{rs} \geq 0, \forall rs \in RS, p \in K_t^{rs}, t \in T_d \tag{25}$$

Eq. (22) means 0 is one of the subgradients of the Lagrangian in terms of time-dependent path flow. Suppose we find $f_{pt}^{rs} > 0$ and $f_{p't'}^{rs} = 0$. Since $f_{p't'}^{rs} = 0$, we have $\mathrm{PMC}_{p't'}^{rs}(\mathbf{f})^+ = \mathrm{PMC}_{p't'}^{rs}(\mathbf{f})^- = \mu^{rs} + v_{p't'}^{rs}$ where $v_{p't'}^{rs} \geq 0$. $v_{pt}^{rs} = 0$ due to $f_{pt}^{rs} > 0$. Hence, $0 \in [\mathrm{PMC}_{pt}^{rs}(\mathbf{f})^- - \mu^{rs}, \mathrm{PMC}_{pt}^{rs}(\mathbf{f})^+ - \mu^{rs}]$, or namely $\mathrm{PMC}_{pt}^{rs}(\mathbf{f})^- \leq \mu^{rs} \leq \mu^{rs} + v_{p't'}^{rs} = \mathrm{PMC}_{p't'}^{rs}(\mathbf{f})^+$, which gives the same result as Proposition 1.

When the TC is differentiable in terms of time-dependent path flow, Eq. (22) becomes $0 = \frac{\partial TC(\mathbf{f})}{\partial f_{pt}^{rs}} - \mu^{rs} - \nu_{pt}^{rs}$ and in this case μ^{rs} is equivalent to the minimum path marginal cost of the O-D pair rs. When TC is not differentiable, let $\mathbb{P}_{rs}^{+}$ denote the set of time-dependent paths connecting r and s that have positive flows at SO. We will have $\mu^{rs} \in [\max_{p\in\mathbb{P}_{rs}^{+}}\{\mathrm{PMC}_{pt}^{rs}(\mathbf{f})^{-}\}, \min_{p\in\mathbb{P}_{rs}^{+}}\{\mathrm{PMC}_{pt}^{rs}(\mathbf{f})^{+}\}]$.

The SO-DTA problem is equivalent to a VI problem, i.e. find $\mathbf{f}^* \in \Omega$ defined by Eqs. (2b) and (2c) such that

$$\sum_{rs\in RS}\sum_{t\in T_d}\sum_{p\in K_t^{rs}} (\mathrm{PMC}_{pt}^{rs}(\mathbf{f}^*)^{+} I_{f_{pt}^{rs} > f_{pt}^{rs*}} + \mathrm{PMC}_{pt}^{rs}(\mathbf{f}^*)^{-} I_{f_{pt}^{rs} < f_{pt}^{rs*}})(f_{pt}^{rs} - f_{pt}^{rs*}) \geq 0, \forall \mathbf{f} \in \Omega \tag{26}$$

where $I_{(\cdot)}$ is the indicator function that equals to 1 where the condition check in the subscript is true otherwise 0.

Proposition 2. *VI problem (26) solves the M1 model defined by Eq. (2).*

Proof. Let $\phi(\zeta) = TC(\mathbf{f}^* + \zeta(\mathbf{f} - \mathbf{f}^*))$, $\mathbf{f} \in \Omega$ and $\zeta \in [0, 1]$. Note that for any value of ζ we have $\mathbf{f}^* + \zeta(\mathbf{f} - \mathbf{f}^*) \in \Omega$ according to the linearity of the constraints defined by Eqs. (2b) and (2c). $\mathbf{f}^*$ minimizes the TC, and thus $\phi(\zeta)$ achieves minimum at $\zeta = 0$. Hence we must have

$$\frac{\partial \phi(0)}{\partial \zeta}^{+} = \sum_{rs\in RS}\sum_{t\in T_d}\sum_{p\in K_t^{rs}} (\mathrm{PMC}_{pt}^{rs}(\mathbf{f}^*)^{+} I_{f_{pt}^{rs} > f_{pt}^{rs*}} + \mathrm{PMC}_{pt}^{rs}(\mathbf{f}^*)^{-} I_{f_{pt}^{rs} < f_{pt}^{rs*}})(f_{pt}^{rs} - f_{pt}^{rs*}) \geq 0$$

□

Based on Proposition 1 two useful corollaries are listed as below:

Corollary 1. *For any two time-dependent paths pt and p′t′ connecting the same O-D pair, moving vehicles from pt to p′t′ can reduce the total cost if and only if*

$$\mathrm{PMC}_{pt}^{rs}(\mathbf{f})^{-} > \mathrm{PMC}_{p't'}^{rs}(\mathbf{f})^{+} \tag{27}$$

Corollary 2. *For any two time-dependent paths pt and p′t′ connecting the same O-D pair, if the flow on both paths are positive at SO, then we have*

$$\min(\mathrm{PMC}_{pt}^{rs}(\mathbf{f})^{+}, \mathrm{PMC}_{p't'}^{rs}(\mathbf{f})^{+}) \geq \max(\mathrm{PMC}_{pt}^{rs}(\mathbf{f})^{-}, \mathrm{PMC}_{p't'}^{rs}(\mathbf{f})^{-}) \tag{28}$$

Similar to the VI formulation of in Shen (2009) which assumed the total cost is always differentiable, the existence and uniqueness of VI problem 26 requires the mapping between the network flow and the resultant path marginal cost (called the dynamic marginal cost mapping in Shen, 2009) satisfies some specific mathematical conditions such as continuity and monotonicity. However, this mapping is related to many factors, including the embedded traffic dynamics models, network topologies, and demand patterns. Therefore the properties of the dynamic marginal cost mapping are difficult to analyze and it may not satisfy the required conditions in general. Since the solution existence and uniqueness is not guaranteed, we turn to solve the problem using heuristics to be discussed in detail in Section 6.

In next section two demonstrative examples will be solved analytically to provide insights on PMC when the SO is achieved.

5. Deriving the SO using subderivative-based PMC

5.1. Analytical SO solution for the M2 model

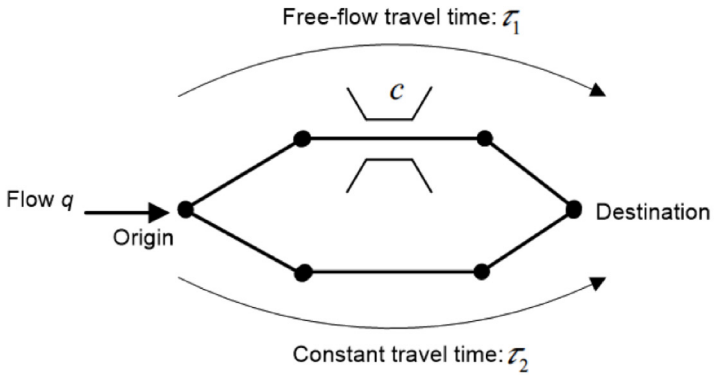
Consider a network with a single O-D pair and two alternative paths as in Fig. 4a. Path 1 has a bottleneck with capacity c and path 2 has no bottleneck. The free flow travel time for path 1 and path 2 are τ_1 and τ_2 respectively. $\tau_1 < \tau_2$. A pre-determined time-dependent flow rate shown in Fig. 4a enters the network.

According to Corollary 2, if the two paths are used at the same time under SO, then we must have path 1 accommodating the flow rate of its exact maximum flow rate c, i.e., $f_{1,t} = c$ and $f_{2,t} = q_t - c$. Since the demand rate is smaller than c after a time point, the queue on path 1 will finally dissipate, say at t_3. Because $\mathrm{PMC}_{2,t} = \tau_2, \forall t$, path 2 would not be used after t_2, when we have $\mathrm{PMC}_{1,t_2}^{+} = \mathrm{PMC}_{2,t_2} = \tau_2$. Given these conditions together with demand constraint, we can solve t_2 and t_3 by,

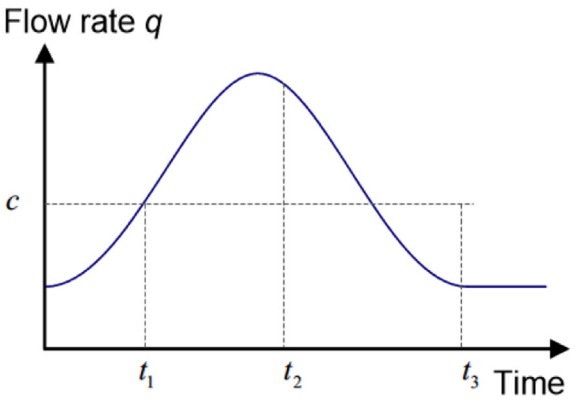
$$t_3 + \tau_1 - t_2 = \tau_2 \tag{29}$$

$$\int_{t_2}^{t_3} q_t dt = (t_3 - t_2)c \tag{30}$$

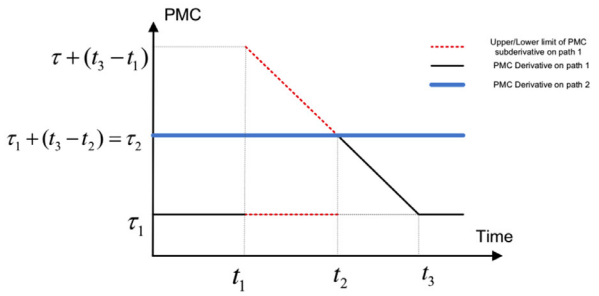
The cumulative curves of the two paths under SO are shown in Fig. 4d and e and the PMCs of the two paths under SO are shown in Fig. 4c. Only during the time period $[t_1, t_2]$ we have $\mathrm{PMC}_{1,t}^{-} < \mathrm{PMC}_{2,t} < \mathrm{PMC}_{1,t}^{+}$, the only time period where both paths are used, consistent with the necessary conditions Eq. (20). In addition, the SO flow pattern is consistent with Munoz and Laval (2006).



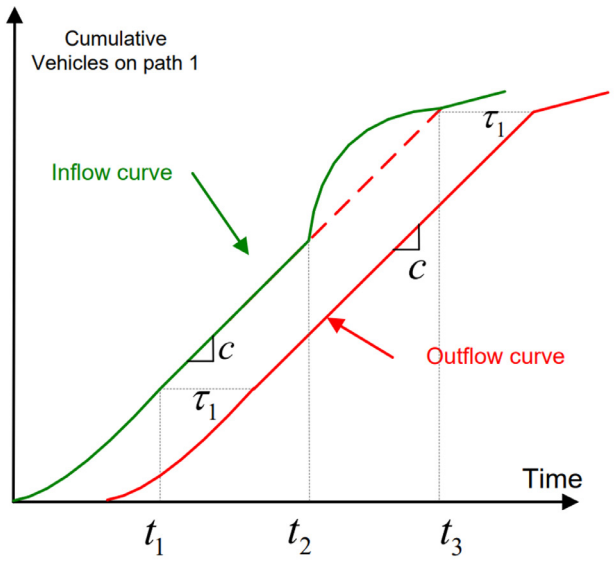
(a)



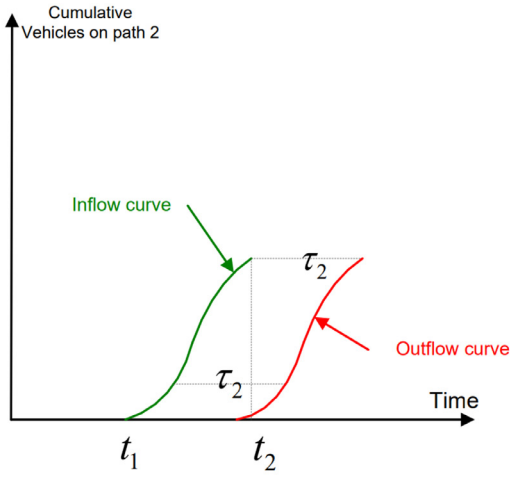
(b)



(c)



(d)



(e)

Fig. 4. An example for M2 model: (a) The network; (b) The time-dependent demand; (c) The PMC of the two paths under SO; (d) The cumulative curve of path 1 under SO; (e) The cumulative curve of path 2 under SO.

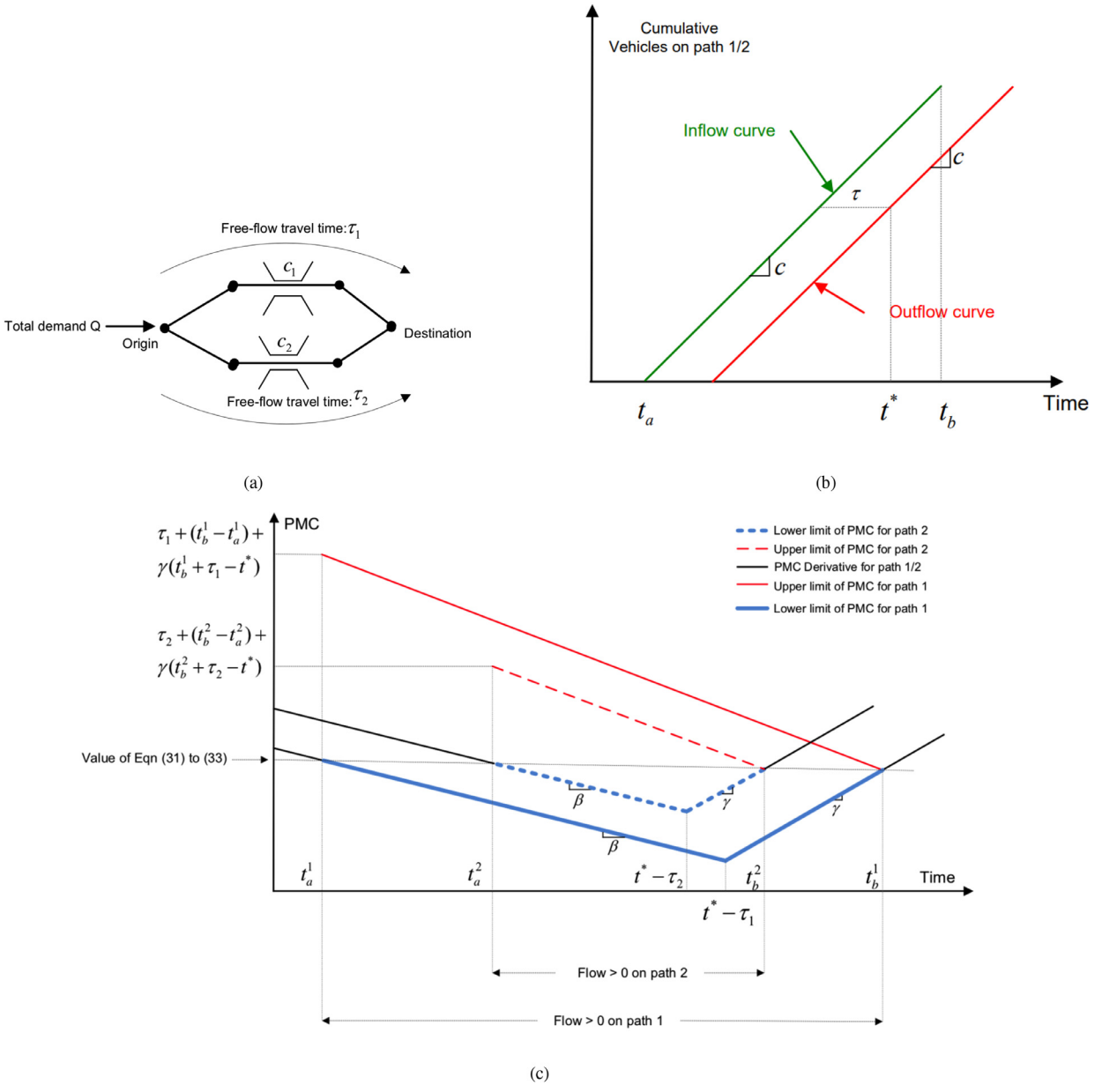


Fig. 5. An example for M1 model: (a) The network; (b) The cumulative curves under SO; (c) The PMC under SO.

5.2. Analytical SO solution for the M1 model

Now consider a network with two alternative paths but both paths have a bottleneck as in Fig. 5a. The capacities of the two bottlenecks are c_1 and c_2 and the free flow travel time are τ_1 and τ_2. In all Q travelers travels from the origin to the destination. They can choose their departure times and routes. All travelers have an identical desired arrival time t^* (namely $\Delta_{rs} = 0$). For simplicity we assume $\alpha = 1$.

From Corollary 2, for any time t_1 that path 1 is used and any time t_2 that path 2 is used, we must have $f_{1,t_1} = c_1$, $f_{2,t_2} = c_2$. Hence, the cumulative curves on both paths 1 and 2 are straight lines with its maximum capacity flow rates, shown in Fig. 5b. The corresponding PMC is shown in Fig. 5c. Let t_a^1 (t_a^2) and t_b^1 (t_b^2) denote the starting and ending times that path 1 (2) is used under SO. For any path, $\text{PMC}_{t_b} = \text{PMC}_{t_a}^-$. Otherwise there must exist a time period $[t_a - \Delta t, t_a)$ where a marginal vehicle at t_b can shift to a time within this time period, which further reduces the TC. Or a marginal vehicle at t_a can shift to t_b to reduce the TC. $\text{PMC}_{1,t_b^1} = \text{PMC}_{2,t_b^2}$ must also hold, because otherwise the flow on the path with a greater PMC can

shift to the other to reduce the TC. Together with the constraint of total demand, SO can be solved by,

$$\tau_1 + \beta(t^* - \tau_1 - t_a^1) = \tau_1 + \gamma(t_b^1 + \tau_1 - t^*) \tag{31}$$

$$\tau_2 + \beta(t^* - \tau_2 - t_a^1) = \tau_2 + \gamma(t_b^1 + \tau_2 - t^*) \tag{32}$$

$$\tau_1 + \beta(t^* - \tau_1 - t_a^1) = \tau_2 + \beta(t^* - \tau_2 - t_a^1) \tag{33}$$

$$(t_b^1 - t_a^1)c_1 + (t_b^2 - t_a^2)c_2 = Q \tag{34}$$

It can be verified that the PMCs depicted in Fig. 5c satisfy Eq. (20).

5.3. Discussions

It is no surprise that the TC becomes not differentiable under SO at some time and along some paths. In a general network, some road segments become bottlenecks when the demand is high. One would naturally think of maximizing the flow throughput for those bottlenecks by, if all possible, controlling the entering flow rate being exactly the bottleneck discharging rate without creating a queue (namely no waste of waiting time nor flow capacity). This is exactly when flow equals to the link capacity flow rate (sometimes the downstream bottleneck's capacity rate) and the TC becomes indifferentiable with respect to those paths and times.

Traditionally, we solve the path-based SO-DTA with the assumption that PMC is differentiable, and in fact that "PMC" refers to the upper limit of PMC in this paper. Since $\text{PMC}^+ \in \text{PMC}$, technically, using only the upper limit of PMC to solve for the VI problem would also work. However, this method suffers from numerical issues, as we will show later in experiments. It is challenging to obtain true SO solutions for flows, even for a toy network. We would use a subgradient based approach to make use of both the upper and lower limits of PMC to solve the SO-DTA problem more effectively and efficiently.

6. A heuristic solution algorithm for path-based SO-DTA

The path-based SO-DTA problem can be cast into the VI problem 26. Any standard solution algorithms can be used to solve the VI problem (Harker and Pang, 1990). Here we propose a heuristic algorithm that specifically considers subgradients of the PMC since TC can be not differentiable.

First, we summarize the method of successive averages (MSA) algorithm in Algorithm 1 that was used by Shen et al. (2007b).

Algorithm 1: A typical MSA (TMSA) algorithm (Parentheses in Step 2 and Step 3 are for SO-DTA without departure time choices).

Initialization; any $\mathbf{f}^0 \in \Omega$; $\nu = 0$; λ^0;**repeat**

1. Load $\mathbf{f}^\nu$ into the network;
2. For all $rs \in RS$ (all $rs \in RS$ given any $t \in T_d$), find the time-dependent path $[p^*, t^*]$ ($[p^*, t]$) with least PMC;
3. Generate an auxiliary path flow pattern $\mathbf{g}(\mathbf{f}^\nu)$ by assigning all demands of Q^{rs} (q_t^{rs}) onto $[p^*, t^*]$ ($[p^*, t]$);
4. Update $\mathbf{f}^{\nu+1} = (1 - \lambda^\nu)\mathbf{f} + \lambda^\nu \mathbf{g}(\mathbf{f}^\nu)$;
5. Update λ^ν to $\lambda^{\nu+1}$, $\nu = \nu + 1$

until *Convergence criteria meets*;

The step size λ in Algorithm 1 can be customized. In this paper, we use the diminishing step size as follows:

$$\lambda^\nu = \frac{1}{\nu + 1} \tag{35}$$

The convergence criteria for Algorithm 1 is defined as either the value of gap function is smaller than a small positive value ϵ or the number of iterations reaches a pre-determined number N, whichever comes first. The gap function is defined as

$$\frac{\sum_t \sum_{rs} \sum_p f_{pt}^{rs}(\text{PMC}_{pt}^{rs}(\mathbf{f}) - \mu^{rs}(t))}{\sum_t \sum_{rs} \sum_p f_{pt}^{rs} \mu^{rs}(t)} \tag{36}$$

This gap function is defined assuming that the path marginal cost is single-valued. $\mu^{rs}(t)$ is the minimum path marginal cost of all paths between O-D pair rs at time t. Note that Eq. (36) is for the case without departure time choices. With departure time choice PMC_{pt}^{rs} and $\mu^{rs}(t)$ should be replaced with PMC_p^{rs} and μ^{rs}.

In the Step 2 of Algorithm 1 when searching for the minimal cost time-dependent path, the algorithm implicitly assumes that the PMC is single-valued. However this is not generally the case under SO. The core to designing a good heuristic is how to find a descent direction with a step size when the TC becomes not differentiable.

We start by introducing a new way of forming a path set.

Definition 1 (Path Set of Minimal PMC (PSMP)). For an O-D pair $rs \in RS$ (and a time $t \in T_d$), the path set of minimal PMC is the set of time-dependent paths $[p,t] \in P^{rs}_{\min}$ ($p \in P^{rs}_{t,\min}$) such that:

(1) for any time-dependent path $[p', t']$ (p') of the same O-D pair (the same O-D pair at the same time t) that are not in the path set of minimal PMC and with a positive flow $f^{rs}_{p't'} > 0$ ($f^{rs}_{p't} > 0$), we always have

$$\mathrm{PMC}^{rs}_{p't'}(\mathbf{f})^- > \mathrm{PMC}^{rs}_{pt}(\mathbf{f})^+ \tag{37}$$

for M1 model and

$$\mathrm{PMC}^{rs}_{p't}(\mathbf{f})^- > \mathrm{PMC}^{rs}_{pt}(\mathbf{f})^+ \tag{38}$$

for M2 model.

(2) for any time-dependent path $[p', t']$ (p') of the same O-D pair (the same O-D pair at the same time t) that are not in the path set of minimal PMC and with a zero flow $f^{rs}_{p't'} = 0$ ($f^{rs}_{p't} = 0$), we always have

$$\mathrm{PMC}^{rs}_{p't'}(\mathbf{f})^+ > \mathrm{PMC}^{rs}_{pt}(\mathbf{f})^- \tag{39}$$

for M1 model and

$$\mathrm{PMC}^{rs}_{p't}(\mathbf{f})^+ > \mathrm{PMC}^{rs}_{pt}(\mathbf{f})^- \tag{40}$$

for M2 model.

(3) for two time-dependent paths $[p, t]$ and $[p'', t'']$ (p and p'') both in the path set of minimal PMC, we always have

$$\min(\mathrm{PMC}^{rs}_{p''t''}(\mathbf{f})^+, \mathrm{PMC}^{rs}_{pt}(\mathbf{f})^+) \geq \max(\mathrm{PMC}^{rs}_{p''t''}(\mathbf{f})^-, \mathrm{PMC}^{rs}_{pt}(\mathbf{f})^-) \tag{41}$$

for M1 model and

$$\min(\mathrm{PMC}^{rs}_{p''t}(\mathbf{f})^+, \mathrm{PMC}^{rs}_{pt}(\mathbf{f})^+) \geq \max(\mathrm{PMC}^{rs}_{p''t}(\mathbf{f})^-, \mathrm{PMC}^{rs}_{pt}(\mathbf{f})^-) \tag{42}$$

for M2 model.

This definition simply states that moving a small tiny fraction of flows within PSMP, or from a path in PSMP to any path outside PSMP, would increase the TC. On the other hand, moving a small fraction of flows from a path outside of PSMP to any path in the PSMP can reduce the TC. If the PSMP is found, we can generate the auxiliary path flow pattern to be assigned to only those paths in the PSMP.

It should be pointed out that finding PSMP is not trivial even when the network is not so large, because the number of time-dependent paths for a single O-D pair can be very large. In practice we can initialize PSMP with a shortest path. Next, in each iteration, we find the time-dependent path with the minimum upper limit PMC for each O-D pair, and examine this (possibly) new path and all existing paths with positive path flow, which altogether are added to the PSMP.

The first proposed algorithm is summarized in Algorithm 2 called PHA1. In the step of generating the auxiliary path flow

Algorithm 2: PSMP-based heuristic algorithm 1 (PHA1) (Parentheses are for SO-DTA without departure time choices). Step 2 is based on path enumeration.

Initialization; any $\mathbf{f}^0 \in \Omega$; $\nu = 0$; λ^0;**repeat**
1. Load $\mathbf{f}^\nu$ into the network;
2. For all $rs \in RS$ ($rs \in RS$ and $t \in T_d$), find the PSMP $P^{rs}_{\min}$ ($P^{rs}_{t,\min}$);
3. Generate an auxiliary path flow pattern $\mathbf{g}(\mathbf{f}^\nu)$ as follows;
For each OD pair rs, go through each path k, t in $P^{rs}_{\min}$ (each path k in $P^{rs}_{t,\min}$):
if *Path k, t (k) has an active bottleneck, i.e., there exists a link along path k where its exit flow rate is less than or equal to its flow capacity* **then**
Assign flow to path k, t (k) that is equal to the flow capacity of the most downstream bottleneck, or f^{rs}_{kt}, whichever is greater.
For all paths in the PSMP that do not have an active bottleneck, assign the remainder flow proportional to the minimum flow capacity of the all links on each of those paths.
4. Update $\mathbf{f}^{\nu+1} = (1-\lambda^\nu)\mathbf{f} + \lambda^\nu \mathbf{g}(\mathbf{f}^\nu)$;
5. Update λ^ν to $\lambda^{\nu+1}$, $\nu = \nu + 1$
until *Convergence criteria meets*;

pattern, the fact that the upper PMC does not equal to lower PMC only if the flow equals to the capacity is considered. PHA1 allows some paths to have constant free-flow travel time at its full capacity, which shows good performance in experiments later.

We also propose a second algorithm PHA2 to improve PHA1 in Algorithm 3. PHA1 could result in assigning excessive demand onto one path that has the minimal upper limite PMC. Because PMC is a subderivative that works only locally, going too farther along the PMC (namely a long step size) at a particular demand pattern could lead to increase in TC. Thus,

Algorithm 3: PSMP-based heuristic algorithm 2 (PHA2) (Parentheses are for SO-DTA without departure time choices). Step 2 is based on path enumeration.

Initialization; any $\mathbf{f}^0 \in \Omega$; $\nu = 0$; λ^0;**repeat**
1. Load $\mathbf{f}^\nu$ into the network;
2. For all $rs \in RS$ ($rs \in RS$ and $t \in T_d$), find the PSMP $P^{rs}_{\min}$ ($P^{rs}_{t,\min}$);
3. Generate an auxiliary path flow pattern $\mathbf{g}(\mathbf{f}^\nu)$ as follows;
For each OD pair rs, go through each path k, t in $P^{rs}_{\min}$ (each path k in $P^{rs}_{t,\min}$):
if *Path k, t (k) has an active bottleneck, i.e., there exists a link along path k where its exit flow rate is less than or equal to its flow capacity* **then**
Assign flow to path k, t (k) that is equal to the flow capacity of the most downstream capacity, if PMC^{rs}_{kt} is not differentiable;
Otherwise, Assign flow f^{rs}_{kt} to path k, t (k).
For all paths in the PSMP that do not have an active bottleneck and all paths outside the PSMP, in the ascent order of upper limit PMC, assign the flow to path k, t (k) that is equal to the minimum flow capacity of all the links on path k. Continue this flow assignment til q^{rs} (q^{rs}_t) is exhausted.
4. Update $\mathbf{f}^{\nu+1} = (1-\lambda^\nu)\mathbf{f} + \lambda^\nu \mathbf{g}(\mathbf{f}^\nu)$;
5. Update λ^ν to $\lambda^{\nu+1}$, $\nu = \nu + 1$
until *Convergence criteria meets*;

in PHA2, the change in flow on a single path is limited from iteration to iteration. The auxiliary flow will be assigned onto the path with the first minimal PMC up to a limit, then onto the path with the second minimal PMC up to a limit, and so on.

Both PHA1 and PHA2 rely on the computation of PSMP. For small networks, we can generate the path set for every O-D pair through path enumeration, following by comparing the PMC of those paths in the set. However this approach is not practical in large networks as the number of paths can be extremely large. On the other hand, although thousands of paths may exist, only a small portion of them may be potentially used under system optimum. Therefore we use some heuristics to identify paths that are likely in PSMP, and then form the PSMP among those selected paths only. The concept of column generation can be useful here. It was proposed for large linear programming problems. We propose a modified version of PHA1 which we call it PHACG (PSMP-based Heuristic Algorithm with Column Generation) in Algorithm 4.

Algorithm 4: PSMP-based Heuristic Algorithm with Column Generation 1 (Parentheses in Step 2 are for SO-DTA without departure time choices).

Initialization: any $\mathbf{f}^0 \in \Omega$; $\nu = 0$; λ^0;For all $rs \in RS$, find the set of n static shortest paths using the free-flow travel time as the weight (n is a small integer). Save this path set as PT^{rs}(PT^{rs}_t).**repeat**
1. Load $\mathbf{f}^\nu$ into the network;
2. For all $rs \in RS$ ($rs \in RS$ and $t \in T_d$), find the PSMP $P^{rs}_{\min}$ ($P^{rs}_{t,\min}$) from PT^{rs} (PT^{rs}_t) ;
3. Generate an auxiliary path flow pattern $\mathbf{g}(\mathbf{f}^\nu)$ as follows;
For each OD pair rs, go through each path k, t in $P^{rs}_{\min}$ (each path k in $P^{rs}_{t,\min}$):
if *Path k, t (k) has an active bottleneck, i.e., there exists a link along path k where its exit flow rate is less than or equal to its flow capacity* **then**
Assign flow to path k, t (k) that is equal to the flow capacity of the most downstream bottleneck, or f^{rs}_{kt}, whichever is greater.
For all paths in the PSMP that do not have an active bottleneck, assign the remainder flow proportional to the minimum flow capacity of the all links on each of those paths.
4. Update $\mathbf{f}^{\nu+1} = (1-\lambda^\nu)\mathbf{f} + \lambda^\nu \mathbf{g}(\mathbf{f}^\nu)$;
5. Update λ^ν to $\lambda^{\nu+1}$, $\nu = \nu + 1$
6. Delete paths in $PT^{rs}(PT^{rs}_t)$ which have no flow.
7. For all $rs \in RS$ ($rs \in RS$ and $t \in T_d$), find the path with the minimal upper PMC using time-dependent shortest path algorithm, and add this path to $PT^{rs}(PT^{rs}_t)$
until *Convergence criteria meets*;

The convergence criteria for Algorithms 2–4 is slightly different from Algorithm 2. Note that the definition of gap function in Eq. (36) assumed that PMC is single-valued. Hence it is not applicable for PHA1 and PHA2. Thus we define a modified version of gap function as below

$$\frac{\sum_t \sum_{rs} \sum_p f^{rs}_{pt}(-\min\{\text{PMC}^{rs}_{pt}(\mathbf{f})^- - \mu^{rs}(t)', 0, \mu^{rs}(t)' - \text{PMC}^{rs}_{pt}(\mathbf{f})^+\})}{\sum_t \sum_{rs} \sum_p f^{rs}_{pt} \mu^{rs}(t)'} \tag{43}$$

where $\mu^{rs}(t)'$ is defined as $\max_{p\in\mathbb{P}_{rs}^{t+}}\{\mathrm{PMC}_{pt}^{rs}(\mathbf{f})^{-}\}$. In other words, $\mu^{rs}(t)'$ is a feasible value of $\mu^{rs}(t)$ in Eq. (22). When first order necessary conditions shown in Section 4 were met, the value of Eq. (43) will be 0. Hence this new version of gap function can be used as a convergence indicator when PMC is not single-valued. Like Algorithm 1, the convergence criteria for Algorithm 2–4 is either the value of Eq. (43) is smaller than a small positive value ϵ or the number of iterations achieves a pre-determined large positive integer N, whichever comes first.

Last but not least, the subgradient method is not a strictly descending algorithm even in the convex case. Since the subgradient is used in our heuristic algorithms, we will save the solution that gives the minimal TC over all iterations as the "optimal" solution. In next section the performance of those algorithms will be examined and presented.

7. Numerical experiments

In this section the two examples that have been solved analytically in Section 5 will be first used to examine the performance of proposed subgradient based algorithms. This is followed by solving SO numerically in a small synthesized network and a sizable real-world network.

7.1. SO-DTA Numerical Example 1: M2 Model

The network is the same as in Fig. 4a. The capacity of the bottleneck on path 1 is set to 5400 veh/h. The free flow travel time for the two paths are $\tau_1 = 15$ min and $\tau_2 = 30$ min respectively. One hour is divided into 600 simulation intervals. Set the maximal number of iteration to be 5000 and $\epsilon = 0.001$. We consider a time-dependent demand pattern that has a single peak in the middle of the assignment assignment horizon. The demand pattern is shown in Fig. 6a as well as the

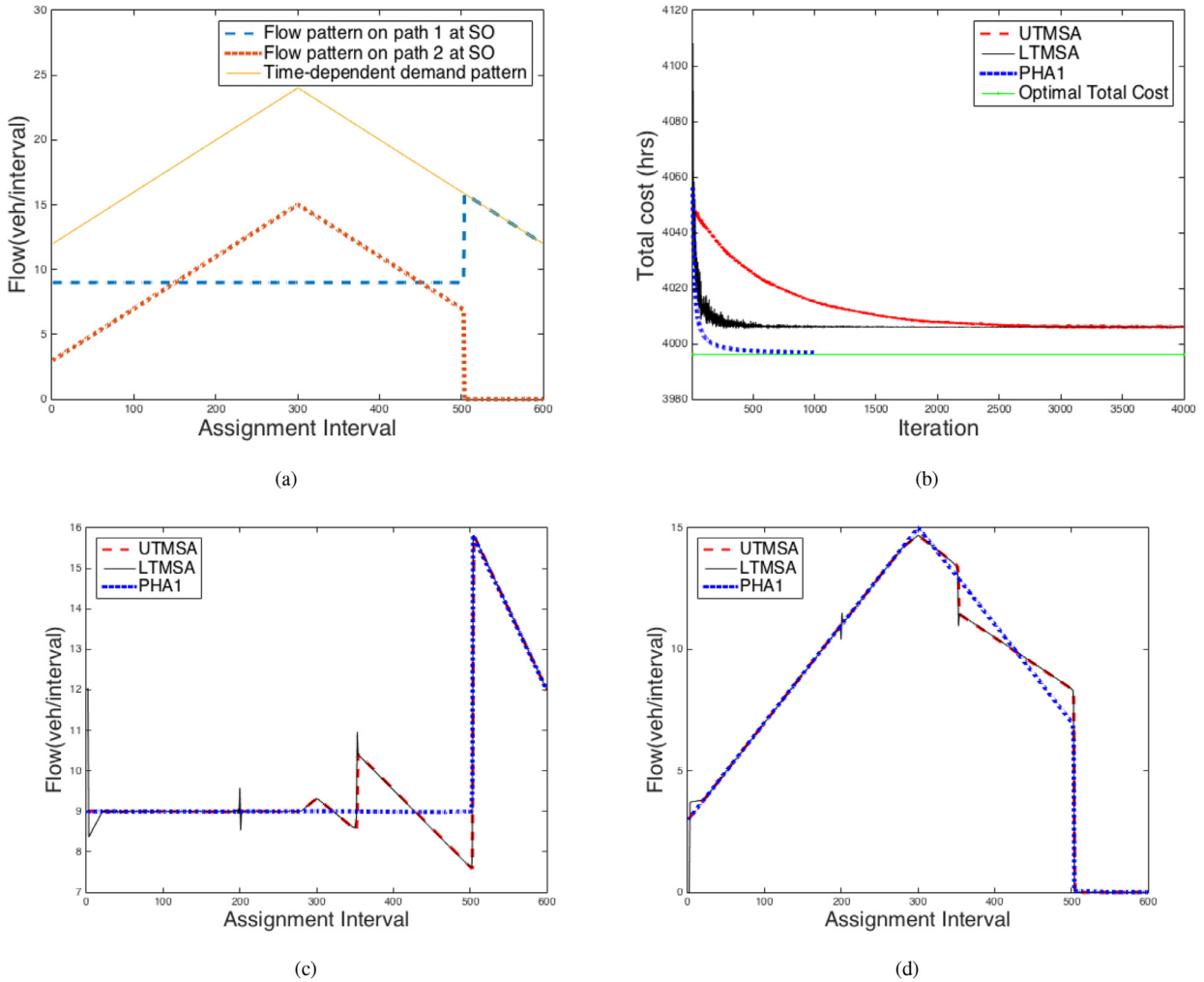


Fig. 6. Numerical example 1: (a) the analytic SO; (b) the TC vs iteration; (c) The resulting flow pattern on path 1; (d) The resulting flow pattern on path 2.

Table 1
Summary of numerical experiments.

	Algorithms	UTMSA	LTMSA	PHA1	PHA2	SO
M2 model	Minimum TC achieved (hr)	4,005.3	4,005.4	3,996.7	–	3,996.1
	TC Error (%)	0.23	0.23	0	–	–
	RMSE of path flow 1 (veh/interval)	0.44	0.46	0.03	–	–
	RMSE of path flow 2 (veh/interval)	0.44	0.47	0.03	–	–
M1 model	Minimum TC achieved (million $)	3.77	3.75	3.75	3.72	3.72
	TC Error (%)	1.42	0.74	0.77	0	–
	RMSE of path flow 1 (veh/interval)	2.78	2.89	0.65	0.33	–
	RMSE of path flow 2 (veh/interval)	1.89	1.79	1.92	0.22	–

system optimal flow pattern solved using the method in Section 5.1. In this case, we assume congestion on path 1 does not spillover to the origin. Therefore, the results would be the same with LWR or with the point queue model. Three algorithms are compared: TMSA that always uses the upper limit of PMC (UTMSA for short), TMSA that always using the lower limit of PMC (LTMSA for short) and PHA1.

The results are reported in Fig. 6 and Table 1. We can see that the TC in UTMSA and LTMSA converge almost to the same value. LTMSA converged slightly faster than UTMSA but the convergence curve is more unstable in the first 1000 iterations. The minimal TC achieved by PHA1 is far closer to the real value than the other two MSA-based algorithms.

PHA1 stopped at about the 1000th iteration while the other two did not stop until reaching the maximal number of iterations. Based on our analysis before, if we only know either the upper limit or lower limit of PMCs at the flow pattern when the total cost is not differentiable, then the solution may be trapped where the shifting flow among several paths would be derived as a result of treating PMC differentiable, but this does not truly reduces the TC. However, if we examine both lower and upper limits of PMCs, the subgradient based algorithm almost ensures TC reduction. This explains why PHA1 could stop earlier. Figs. 6c, and 7 show the resulting flow patterns on the two paths. We see that the MSA based algorithms, though reaching a reasonable minimum TC, do not lead to the correct SO flow solutions, whereas PHA1 does.

7.2. SO-DTA Numerical Example 2: M1 Model

Now we turn to numerically solve the example in Section 5.2. We assume $c_1 = 5400$ veh/h, $c_2 = 3600$ veh/h, $\tau_1 = 15$ min and $\tau_2 = 30$ min. One hour is divided into 600 simulation intervals. Assume a identical desired arrival time $t^* = 1000$th interval and $\beta = \$0.5/\text{interval} < \alpha = \$1/\text{interval} < \gamma = \$2/\text{interval}$. Set the maximal number of iteration to be 5000 and $\epsilon = 0.001$. Using the method presented in Section 5.2 we can solve the SO analytically as shown in Fig. 7a. Four algorithms are compared: PHA1, PHA2, UTMSA, LTMSA.

The convergence curves are shown in Fig. 7b. PHA2 almost converges instantly and PHA1 converges mildly faster than the other two MSA based algorithms. The minimum TC that is ever achieved for each algorithm is reported in Table 1. We can find that the TC still goes up and down even after thousands of iterations under UTMSA and LTMSA. The resulting flow patterns are shown in Figs. 7c and d. Again, both UTMSA and LTMSA cannot achieve the correct SO flow solutions (There results are similar to each other, so we only plot one of them). The PHA2 gave the best result in terms of both minimum TC and SO flow patterns. The variability of the flow pattern on path 1 using PHA1 is less than the other two MSA algorithms, mainly because the PHA1 can split the flow to other paths when the TC is not differentiable (which cannot be done through MSA). The performance of PHA1 in the M1 model is worse than in the M2 model. This is potentially because in the M2 model with 2 alternative paths, clearly the size of PSMP cannot exceed 2, so PHA1's performance is fully explored. Whereas in the M1 model with departure time choice, the size of PSMP can be very large so PHA1 can only explore a very limited set of spatio-temporal paths within PSMP. In this case, PHA2 can help explore other possible paths comparing to PHA1. In general our proposed algorithms outperforms traditional MSA methods.

7.3. SO-DTA Numerical Example 3: many-to-many O-D demands in a small network

We use a small synthetic corridor network which was firstly used by Nie (2006). The network consists of 18 links and 16 nodes as shown in Fig. 8. Nodes 11, 12 and 13 are origin nodes and nodes 14, 15, 16 are destination nodes. The properties, including length, free-flow speed and holding capacity are shown in Table 2. Link 10 to link 15 are O-D connectors. The network attempts to abstract a commuting network. Nodes 11, 12, 13 are residential areas and nodes 14, 15, 16 are workplaces. There is a freeway consisting of links 1, 3, 5, 7 and 9. These links have larger free flow speed. Link 7 has a smaller capacity than downstream links and thus can be treated as the main bottleneck on the freeway. Links 16, 17, 18 are arterial roads with lower free-flow speed than highway roads. Links 2, 4, 6 and 8 are short links representing ramps.

We consider many-to-many OD pairs, and each OD pair has a typical one-peak demand pattern. No departure time choice is considered in this experiment, and thus PHA1 and PHA2 yield similar solutions. The change in TC over the iterations using either the sub-gradient based algorithm (PHA1) or UTMSA are shown in Fig. 9a.

At the beginning of iterations, the two algorithms found identical descent directions as the TC is differential when solutions are far from being SO. The two algorithms start to differ in 3rd iteration in which case the objective function becomes

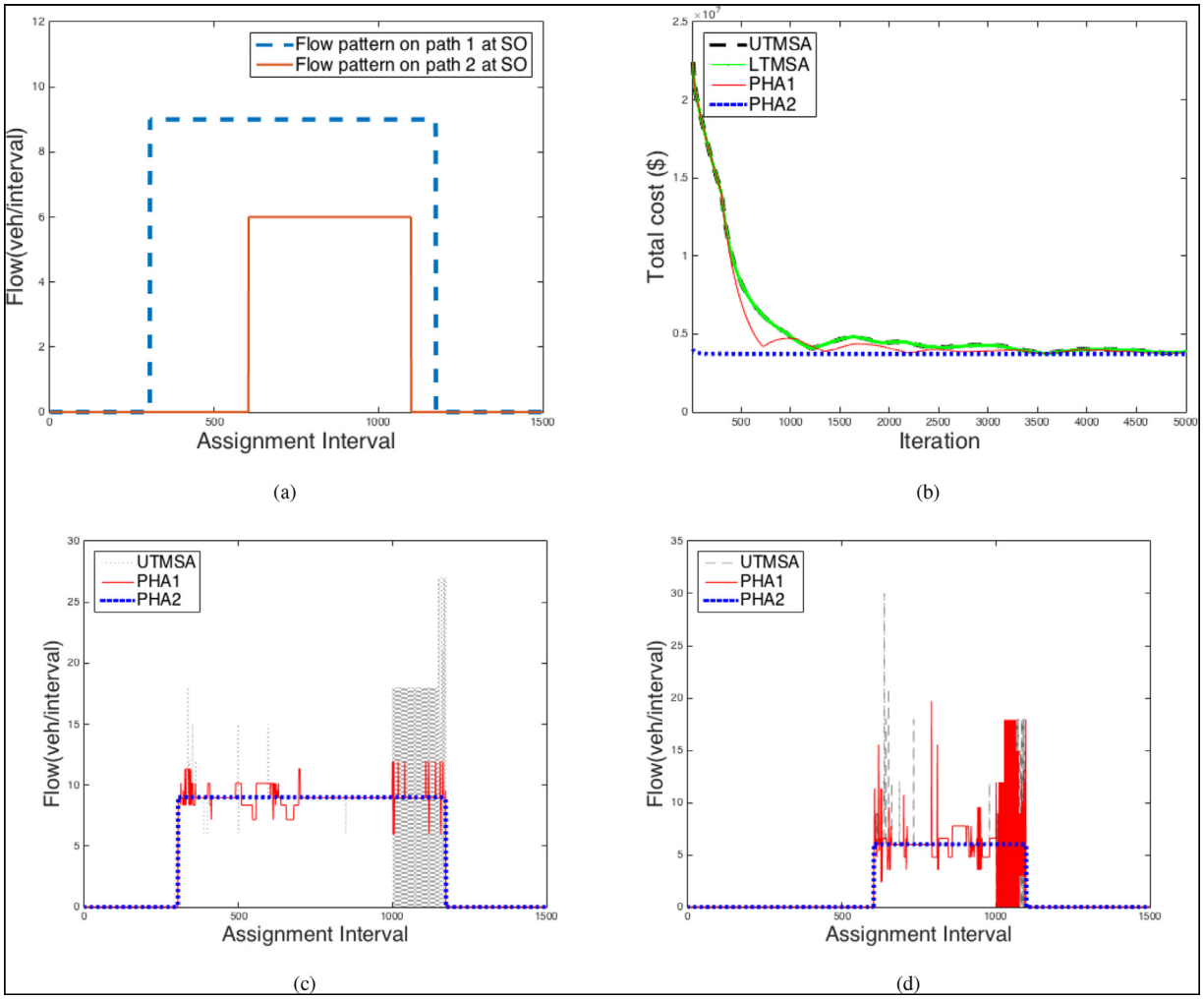


Fig. 7. Numerical example 2: (a) the analytic SO; (b) the TC vs iteration; (c) The resulting flow pattern on path 1 ; (d)The resulting flow pattern on path 2.

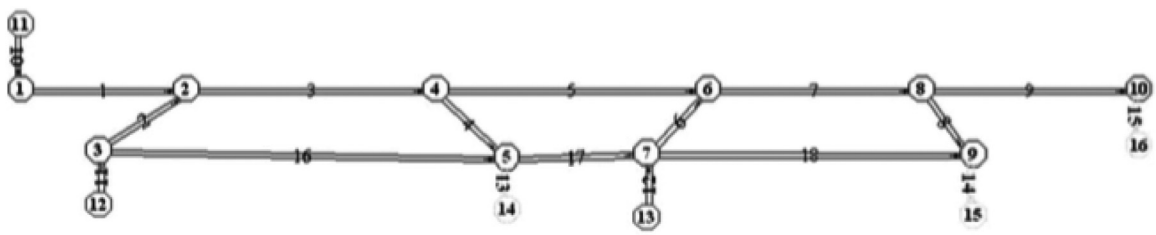


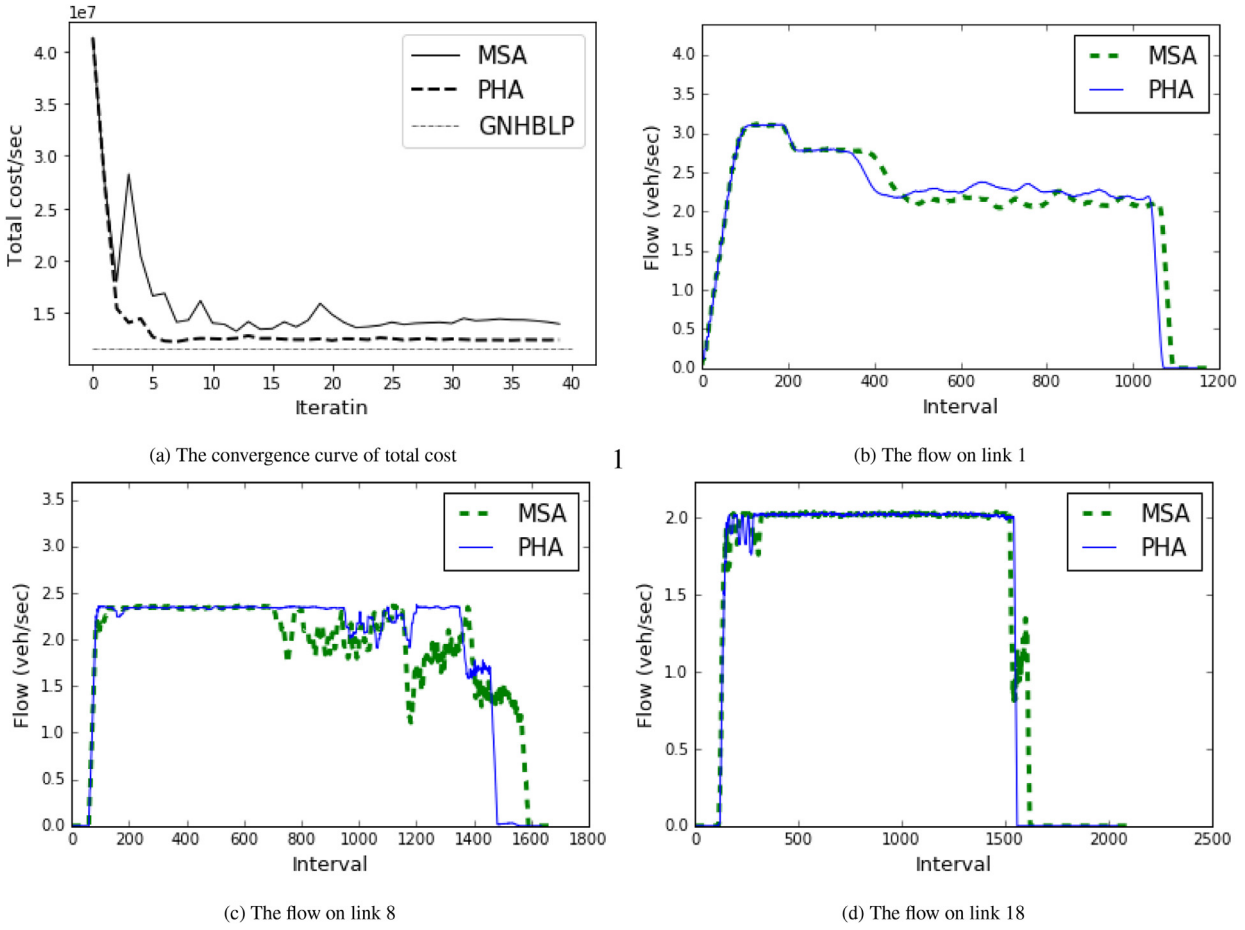
Fig. 8. The synthetic corridor network from Nie (2006).

not differentiable on some time-dependent paths. MSA moves excessive flow to an uncongested path, which could lead to a drastic increase in TC, while the PHA1 ensures a better allocation of flows towards SO. The resultant TC that the PHA converges to is smaller than that of MSA. For PHA, the converged value was almost achieved in the 8th iteration.

Fig. 9b to d shows the resultant flow of representative links 1, 8 and 18, respectively. Since departure time choice was not considered in this experiment, the resultant flows of the two algorithms on link 1 are close (though they can differ slightly due to queue spillover from link 1 to link 10). Links 8 and 18 are the last link of two alternative routes feeding to the destination node 15. The total number of vehicles entering the two links all together should be the same. However, the flow will distribute differently between the two routes. In the first hundreds of time intervals, the resultant flows of MSA and PHA are very close. The flows of two algorithms starts to differ after the 600th interval on link 8 and the 1500th interval on link 18. The flow solution of PHA exhibits less fluctuation over time, which is more reasonable than MSA. For instance, the flow of PHA on link 8 reaches its flow capacity between the 800th and 1300th interval, approximately exhausting the

Table 2
Link properties.

ID	Length(mile)	Free-flow speed (mph)	Holding Capacity (vpm)
1	1	50	540
2	0.5	50	180
3	2	50	540
4	0.5	50	180
5	2	50	540
6	0.5	50	180
7	1	50	360
8	0.5	50	180
9	1	50	540
16	2	30	180
17	2	30	180
18	1	30	180



(a) The convergence curve of total cost

(b) The flow on link 1

(c) The flow on link 8

(d) The flow on link 18

Fig. 9. Numerical example 3.

bottleneck capacity, whereas during the same time, the flow solution of MSA on link 8 is far less than its capacity flow. It has a drastic drop near the 1200th interval, and the flow keeps up and down afterwards, which conflicts with the intuition that this bottleneck link should be fully utilized under SO, provided that our demand has an inverted triangular shape. Similarly flow solutions of PHA on link 18 read more reasonable than MSA solutions.

In general, we conclude that the SO solution of PHA not only provides a lower total cost, but also finds more accurate path/link flow solutions. In practice, the minimum possible total system cost is a benchmark but can be extremely challenging to measure. On the other hand, the optimal link/path flow can have more direct policy and operation implications. We may be able to measure link/path flow and compare them with the optimal (or desired) flow to assess the effectiveness of policies or operation. Having oscillating flow solutions, such as provided by the MSA solution, may not be helpful, nor they are practically realistic. PHA may be able to alleviate this and yield more stabilized flow solutions. This was overlooked in the previous SO-DTA literature.

To further validate our model, we also compared our solution with the solution found through a mathematical programming approach with the link-based formulation (Zhu and Ukkusuri, 2013). Both models are run under the same network and demand settings. Zhu and Ukkusuri (2013) proposed "the most generalized version of the non-holding back LP formulation" (GNHBLP) to handle the SO-DTA problem on a generalized network with multiple OD pairs. It is proved to totally eliminate the vehicle holding-back issues of link based SO-DTA, but FIFO is not considered in this approach. The link model used in GNHBLP is CTM which is the same used in this numerical experiment. The resultant total cost solved by GNHBLP is slightly less than the total travel cost found from the subgradient based SO solutions, shown in Fig. 9a. Note that the GNHBLP does not consider FIFO in the model, hence the solution given by it actually could be a lower bound of the SO-DTA problem with FIFO and no vehicle holding-back.

7.4. SO-DTA Numerical Example 4: many-to-many O-D demands in a sizable network

Now we apply the sub-gradient based algorithm to a real-world sizable network to demonstrate its capability in solving system optimum for general networks. The SR-41 corridor network shown in Fig. 10 is used. It consists of 2065 links, 1441

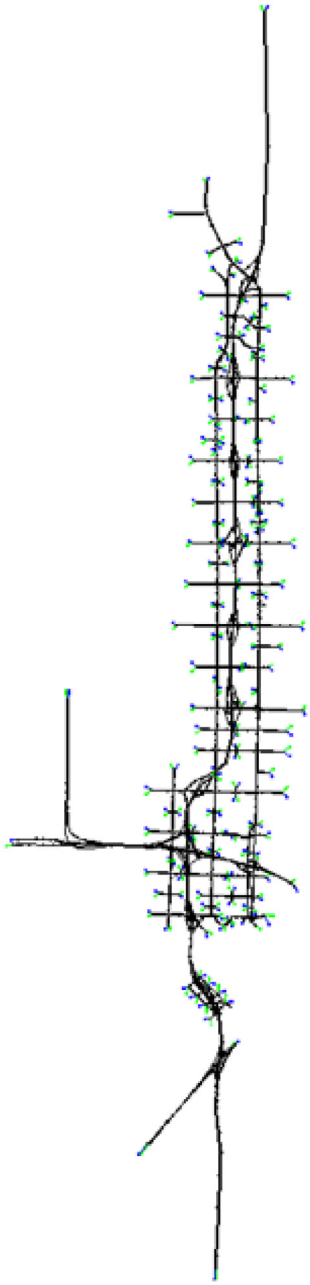
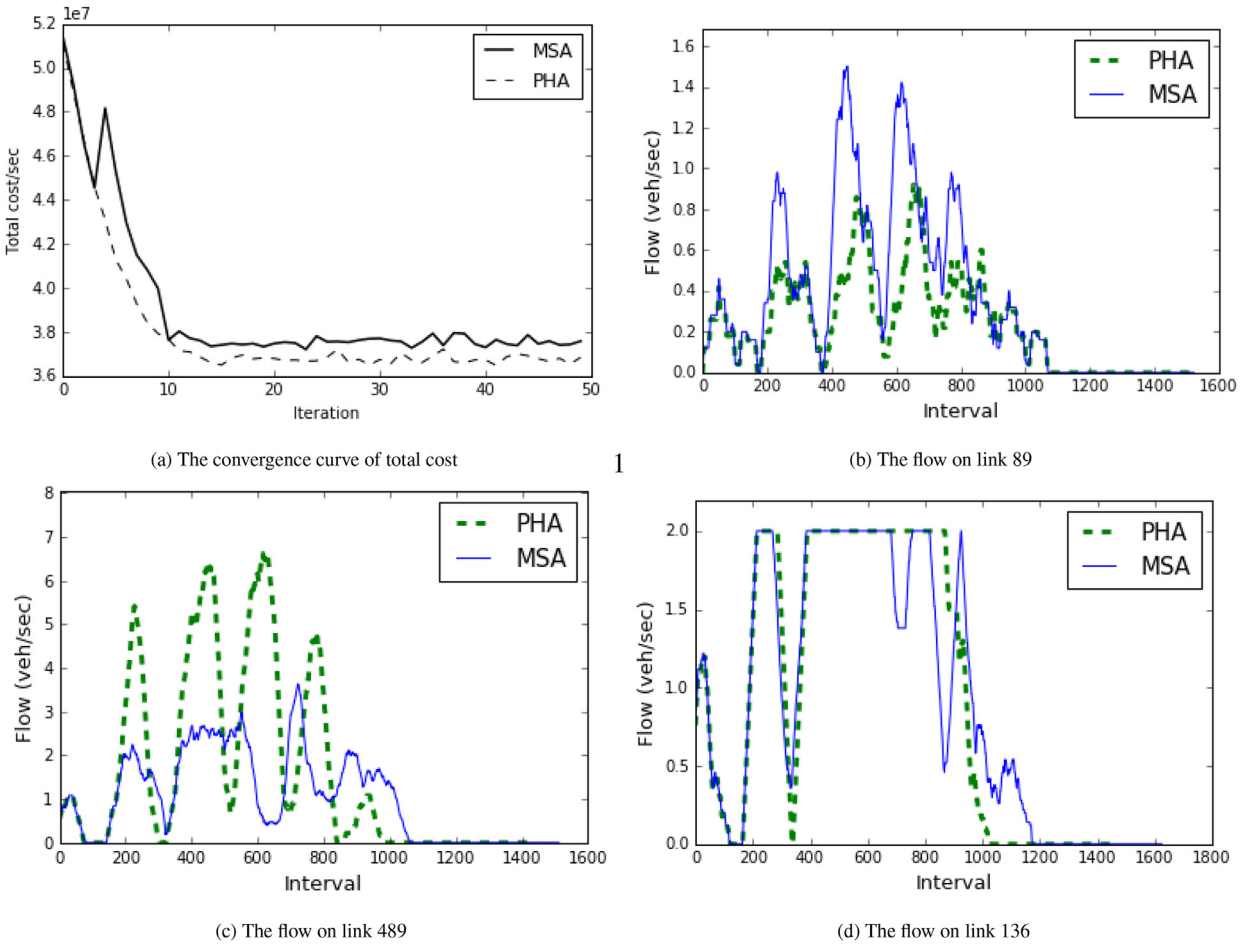


Fig. 10. The SR 41 network Qian et al. (2012).



(a) The convergence curve of total cost

(b) The flow on link 89

(c) The flow on link 489

(d) The flow on link 136

Fig. 11. Numerical experiment 4: an SR-41 network.

nodes and 100 O-D pairs. The SR-41 network is located in the Fresno, California. There is a major freeway in the network with two parallel arterial roads and local streets. This network covers about 4 miles East-West and 16 miles North-South. The time-varying O-D demand was estimated based on field data by Liu et al. (2006) and Zhang et al. (2008). Hence we can solve the system optimum given this calibrated dynamic OD demand. Since the time dependent demand is given, the departure time choice is not considered here. The Algorithm 4 that is designed for large networks is adopted to solve the SO. Since the problem is non-convex and the network consists of a large number of nodes, links and O-D pairs, it is almost sure that a heuristic algorithm will end up with a local minimum. Hence we run the experiments 500 times for both our algorithm and the baseline algorithm, and take the results that gives minimal total cost of each algorithm to compare.

The convergence over iterations for the two algorithms is shown in Fig. 11a. Since we start with a naive assignment for both algorithms - assigning all demand to the one single time-dependent shortest path - the TC declines drastically in the first 5 iterations. Because both algorithms in the first 5 iterations add the same new path to the path set for each OD pair, it can be seen that their performance is very close. The TC resulting from the PHA algorithm keeps declining after 5 iterantions until it finds a minimal value around the 15th iteration. However, MSA stops declining after 10 iterations and stays up and down afterwards. Overall the PHA yields 3% less TC than the MSA after 50 iterations. In other words, the sub-gradient based algorithm can find better descendent directions than the gradient based algorithm in this case.

We found that the link flow solutions can vary substantially on many links for the two algorithms. We choose several of those representative links (on highway or main surface streets) in Fig. 11b, c and –d. Since the resultant flow patterns can be very jumpy over time (possibly due to the algorithms), we apply a moving average smoother just for clear display. The MSA algorithm assigns more flow on link 89 than PHA, whereas less flow on link 489 than PHA. Link 89 is a local street while link 489 is a highway link with a higher capacity. In other words, PHA leads to a better utilization of the freeway link close to its flow capacity. This may contribute to the lower TC resulting from the PHA algorithm. Link 136 is again a surface street with no freeway alternative hence it was heavily used as well. Near the end of the simulation, the solution of PHA gradually decreases from 2 veh/sec (the flow capacity of the link) to 0. However, a drastic drop near the 700th interval and then the 800th interval can be found in the MSA solution, each of which is followed by a jump back to flow capacity. This

would not be reasonable under SO. Similar patterns can be found on some other links under MSA as well. Generally, PHA yields a lower TC as well as more reasonable path/link SO flow.

8. Conclusions

In this paper we study the path-based SO-DTA problem that assigns OD demand over physical paths and time to minimize the total system cost (TC). We show that the TC could be non-differentiable with respect to the link/path flow in some cases, especially when the flow is close or under the SO conditions. This was usually overlooked in previous studies. We demonstrate when the TC would be indifferentiable and how to compute the subgradients, namely the lower and upper limit of path marginal costs. We also examine the relations between the discontinuity of PMC and the SO conditions, develop PMC-based necessary conditions for SO solutions, and finally design heuristic solution algorithms for solving SO in general networks with multi-origin-multi-destination OD demands.

Three heuristic solution algorithms are proposed and tested in four numerical experiments, two toy networks where we compare analytical solutions with numerical solutions, one small network and one sizable real-world network. We show that the proposed heuristic algorithms outperform existing ones by using the upper or lower limit of PMCs, in terms of both the total TC and path/link flow. In toy networks, the results indicate that the proposed PHA1 and PHA2 algorithms can find the theoretical optimum solutions precisely. In two other networks where the real system optimum is hard to identify, the results show that our proposed algorithm can find a lower total system cost, as well as more reasonable link flow patterns.

For large networks the main computational complexity comes from finding the path set of minimal PMC (PSMP) that would be the key to successful solution algorithms. To accelerate the computation, we proposed a heuristic Algorithm 4 by applying the column generation and finding the PSMP among a subset of all paths connecting each O-D pair. The results are promising, indicating a lower TC and more reasonable flow solutions than existing algorithms. However, in large network there is no guarantee that SO can be achieved by far, since (1) the PSMP is not guaranteed to be the true one; (2) the proposed heuristic algorithm do not guarantee a descendent direction from iteration to iteration. Our future research plan is to explore more heuristic algorithms using the sub-gradient information to achieve better SO solutions.

Acknowledgements

This research is funded in part by NSF award CMMI-1751448, and Carnegie Mellon University's Mobility21, a National University Transportation Center for Mobility sponsored by the US Department of Transportation. The contents of this report reflect the views of the authors only. The U.S. Government assumes no liability for the contents or use thereof.

References

Arnott, R., de Palma, A., Lindsey, R., 1990. Departure time and route choice for the morning commute. Transport. Res. Part B 24 (3), 209–228.
Carey, M., 1987. Optimal time-varying flows on congested networks. Oper. Res. 35 (1), 58–69.
Carey, M., 1992. Nonconvexity of the dynamic traffic assignment problem. Transport. Res. Part B 26 (2), 127–133.
Carey, M., Subrahmanian, E., 2000. An approach to modeling time-varying flows on congested networks. Transport. Res. Part B 34 (3), 157–183.
Corthout, R., Himpe, W., Viti, F., Frederix, R., Tampère, C.M., 2014. Improving the efficiency of repeated dynamic network loading through marginal simulation. Transport. Res. Part C 41, 90–109.
Daganzo, C.F., 1994. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transport. Res. Part B 28 (4), 269–287.
Daganzo, C.F., 1995. The cell transmission model, part II: network traffic. Transport. Res. Part B 29 (2), 79–93.
Friesz, T., Bernstein, D., Mehta, N., Tobin, R., Ganjalizadeh, S., 1989. Dynamic network traffic assignment considered as a continuous time optimal control problem. Oper. Res. 37 (6), 893–901.
Ghali, M., Smith, M., 1995. A model for the dynamic system optimum traffic assignment problem. Transport. Res. Part B 29 (3), 155–171.
Harker, P.T., Pang, J.-S., 1990. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. Math. Program. 48 (1–3), 161–220.
Jin, W., Zhang, H.M., 2004. Multicommodity kinematic wave simulation model for network traffic flow. Transp. Res. Rec. 1883 (1), 59–67.
Liu, H.X., Ding, L., Ban, J.X., Chen, A., Chootinan, P., 2006. A streamlined network calibration procedure for california sr41 corridor traffic simulation study. In: Proceedings of the 85th Transportation Research Board Annual Meeting.
Lo, H., 1999. A dynamic traffic assignment formulation that encapsulates the cell-transmission model. 14th International Symposium on Transportation and Traffic Theory.
Long, J., Szeto, W.Y., 2019. Link-based system optimum dynamic traffic assignment problems in general networks. Oper. Res. 67 (1), 167–182.
Long, J., Wang, C., Szeto, W., 2018. Dynamic system optimum simultaneous route and departure time choice problems: intersection-movement-based formulations and comparisons. Transport. Res. Part B 115, 166–206.
Lu, C.-C., Liu, J., Qu, Y., Peeta, S., Rouphail, N.M., Zhou, X., 2016. Eco-system optimal time-dependent flow assignment in a congested network. Transport. Res. Part B 94, 217–239.
Ma, R., Ban, X.J., Pang, J.-S., 2014. Continuous-time dynamic system optimum for single-destination traffic networks with queue spillbacks. Transport. Res. Part B 68, 98–122.
Ma, R., Ban, X.J., Szeto, W., 2017. Emission modeling and pricing on single-destination dynamic traffic networks. Transport. Res. Part B 100, 255–283.
Merchant, D., Nemhauser, G., 1978. A model and an algorithm for the dynamic traffic assignment problem. Transport. Sci. 12 (3), 183–199.
Merchant, D., Nemhauser, G., 1978. Optimality conditions for a dynamic traffic assignment model. Transport. Sci. 12 (3), 200–207.
Munoz, J., Laval, J., 2006. System optimum dynamic traffic assignment graphical solution method for a congested freeway and one destination. Transport. Res. Part B 40 (1), 1–15.
Ni, D., Leonard II, J.D., 2005. A simplified kinematic wave model at a merge bottleneck. Appl Math Model 29 (11), 1054–1072.
Nie, Y., 2006. A Variational Inequality Approach for Inferring Dynamic Origin-Destination Travel Demands. University of California, Davis Ph.D. thesis.
Nie, Y., Zhang, H.M., 2010. Solving the dynamic user optimal assignment problem considering queue spillback. Netw. Spat. Econ. 10 (1), 49–71.
Peeta, S., Mahmassani, H., 1995. System optimal and user equilibrium time-dependent traffic assignment in congested networks. Ann. Oper. Res. 60 (1), 80–113.

Qian, Z., Zhang, H.M., 2011. Computing individual path marginal cost in networks with queue spillbacks. Transport. Res. Rec. (2263) 9–18.
Qian, Z.S., Shen, W., Zhang, H., 2012. System-optimal dynamic traffic assignment with and without queue spillback: its path-based formulation and solution via approximate path marginal cost. Transport. Res. Part B 46 (7), 874–893.
Shen, W., 2009. System Optimal Dynamic Traffic Assignment: A Graph-Theoretic Approach and its Engineering Applications. University of California at Davis Ph.D. thesis.
Shen, W., Nie, Y., Zhang, H.M., 2007. On path marginal cost analysis and its relation to dynamic system-optimal traffic assignment. Transportation and Traffic Theory 2007. Papers Selected for Presentation at ISTTT17.
Shen, W., Nie, Y., Zhang, H.M., 2007. On path marginal cost analysis and its relation to dynamic system-optimal traffic assignment. Procedings of the 17th International Symposium of Transport and Traffic Theory.
Tampère, C.M., Corthout, R., Cattrysse, D., Immers, L.H., 2011. A generic class of first order node models for dynamic macroscopic simulation of traffic flows. Transport. Res. Part B 45 (1), 289–309.
Vickrey, W.S., 1969. Congestion theory and transport investment. Am. Econ. Rev. 59 (2), 251–260.
Wie, B., Tobin, R., Friesz, T., 1994. The augmented Lagrangian method for solving dynamic network traffic assignment models in discrete time. Transport. Scie. 28 (3), 204–220.
Zhang, H.M., Ma, J., Singh, S.P., Chu, L., 2008. Developing Calibration Tools for Microscopic Traffic Simulation Final Report Part III: Global Calibration - O-D Estimation, Traffic Signal Enhancements and a Case Study. Technical Report. California PATH Research Report.
Zhu, F., Ukkusuri, S.V., 2013. A cell based dynamic system optimum model with non-holding back flows. Transport. Res. Part C 36, 367–380.
Ziliaskopoulos, A., 2000. A linear programming model for the single destination system optimum dynamic traffic assignment problem. Transport. Sci. 34 (1), 37–49.