Title
A Variable Formulation of Kinematic Waves: Solution Methods

Author
Daganzo, Carlos F.

Publication Date
2003-11-01

Institute of Transportation Studies
University of California at Berkeley

A Variable Formulation of Kinematic Waves: Solution Methods

Carlos F. Daganzo

RESEARCH REPORT
UCB-ITS-RR-2003-7

November 2003
ISSN 0192 4095

A Variational Formulation of Kinematic Waves: Solution Methods

Carlos F. Daganzo
Institute of Transportation Studies and
Department of Civil and Environmental Engineering
University of California, Berkeley CA 94720

November 13, 2003

Abstract

This paper presents improved solution methods for kinematic wave traffic problems with concave flow-density relations. As explained in part I of this work, the solution of a kinematic wave problem is a set of continuum least-cost paths in space-time. The least cost to reach a point is the vehicle number. The idea here consists in overlaying a dense but discrete network with appropriate costs in the solution region and then using a shortest-path algorithm to estimate vehicle numbers. With properly designed networks, this procedure is more accurate than existing methods and can be applied to more complicated problems. In many important cases its results are exact.

Contents

1 Introduction

It was shown in part I of this work [1] that kinematic wave (KW) traffic problems with a concave flow-density relation are shortest (least cost) path problems. This paper shows how the theory can be put into practice. As in that reference, t and x denote time and space; $N(t, x)$, $q = \partial N/\partial t$ and $k = -\partial N/\partial x$ the vehicle number, flow and density functions; and

$$q = Q(k, t, x). \tag{1}$$

a "fundamental diagram" that can vary with t and x.

The function Q was assumed to be differentiable and concave in k in [1], but the results extend trivially to piecewise differentiable functions, possibly discontinuous in t and x. This will be the working assumption now. Concavity implies that the wave velocity $w = \partial Q/\partial k$ is non-increasing in k. The density can take any real value, although in well posed traffic problems it automatically remains in a range $[0, \kappa]$, where $q \geq 0$; see Fig. 1a. It will also be assumed that $Q(0, t, x) = 0$, as shown in the figure.[1]

The main result of part I is that if N has been defined for all points B on a boundary $\boldsymbol{D}$ (i.e., N_B is known for all $B \in \boldsymbol{D}$) and a function Q is given, then the value of N at any point P in the solution domain, N_P, is given by

$$N_P = \min_{\mathcal{P} \in V_P} \{\mathcal{B}_{\mathcal{P}} + \Delta(\mathcal{P})\}, \tag{2}$$

where $\mathcal{P}$ is a "valid" space-time path $x(t)$ from the boundary to P, $\mathcal{B}_{\mathcal{P}}$ is the vehicle number at the beginning of $\mathcal{P}$, V_P is the set of all valid paths from the boundary to P, and $\Delta(\mathcal{P})$ is the "cost" of $\mathcal{P}$, defined below. Valid paths are piecewise differentiable curves with slopes x' in the range of allowable wave-speeds $[w(\infty, t, x),\ w(-\infty, t, x)]$. If we use Δ_{BP} for the minimum of $\Delta(\mathcal{P})$ across all valid paths from a generic point B to P, with $\Delta_{BP} = \infty$ if there is no valid path from B top P, then (2) can also be expressed as an ordinary minimization:

$$N_P = \min_{B \in \boldsymbol{D}} \{N_B + \Delta_{BP}\}. \tag{3}$$

The expression for the cost of a valid path is:

$$\Delta(\mathcal{P}) = \int_{t_B}^{t_P} R(x', t, x)dt. \tag{4}$$

[1] No generality is lost because one can always find the solution of a problem with $Q(0, t, x) = F(t, x) \neq 0$ by superposing the solutions of two problems with fundamental diagrams $Q - F$ and F, respectively. The first problem is of the type discussed in this paper, and the second one is trivial since its wave speeds are identically 0.

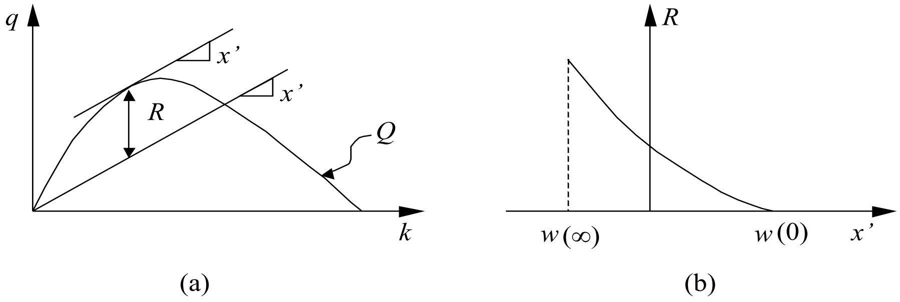

Figure 1: Features of a homogeneous problem: (a) fundamental diagram; (b) cost function.

where t_B and t_P are the times associated with the path endpoints and R is given by:

$$R(x', t, x) = \sup_k \{Q(k, t, x) - kx'\}; \tag{5}$$

see Fig. 1. Given our assumptions, it is easy to show that in the range of valid wave velocities, R is non-increasing and convex in x', and satisfies $R \geq 0$ everywhere, as shown by the figure. Definition (4) applies whether or not $B \in \boldsymbol{D}$.

It was explained in part I that R is the maximum rate at which traffic can pass an observer moving with speed x'. Like Q, R has units of vehicles/time and can be defined a priori independently of the boundary data; i.e., it is a property of the highway. Thus, (2) - (5) define a continuum shortest path problem from $\boldsymbol{D}$ to the solution domain. It was also explained that if Q is strictly concave shortest paths are waves, and that while this is also true if Q is merely concave, in this case there can also be non-differentiable shortest paths.

These results apply to physically "well-posed" problems, which satisfy the following conditions: (i) the solution domain only includes points P that can be reached by a valid path from the boundary; and (ii) if a valid path $\mathcal{P}$ issued from $B \in \boldsymbol{D}$ ends at $B' \in \boldsymbol{D}$ then $N_{B'} \leq N_B + \Delta(\mathcal{P})$. Condition (i) ensures that N_P can be calculated; and (ii) that there are no ambiguities. We also assume , that the continuum cost function is uniformly bounded, $0 \leq R(x', t, x) \leq \widetilde{R}$, as occurs in traffic flow; and that the boundary data satisfy a Lipschitz continuity condition for some $\alpha > 0$; i.e.:

$$|N_B - N_{B'}| \leq \alpha |B - B'| \ , \quad \forall B, B' \in \boldsymbol{D} \tag{6}$$

where $|B - B'|$ is the L_1 distance between B and B': $|B - B'| = |x_B - x_{B'}| + |t_B - t_{B'}|$.

The rest of this paper shows how to solve KW problems with discrete networks and evaluates the method's accuracy. Errors are shown to be uniformly bounded – zero in many important cases. The new method improves on conservation-law procedures because it exploits more fully the properties of Q. Section 2, below, characterizes useful networks. Sections 3 and 4 describe the procedure for homogeneous and inhomogeneous problems.

2 Networks

This section introduces some definitions and six related facts that indicate when/how a continuum problem can be replaced by a network problem.

2.1 Definitions and terminology

The set of (t, x) points that can be reached by valid paths from a point P – the "to-points" of P – is denoted $\boldsymbol{T}(P)$ and called the *range of influence* of P. The set of points that can be reached by valid paths from P – the "from-points" of P – is denoted $\boldsymbol{F}(P)$ and called the *domain of dependence* of P. An ordered pair of points $P'P$ is said to be a *valid pair* if there is a valid path from P' to P; i.e., $P' \in \boldsymbol{F}(P)$. Associated with every valid pair $P'P$ there are three quantities: a time duration $t_{P'P} = t_P - t_{P'} > 0$, a distance $x_{P'P} = x_P - x_{P'}$ and a slope $v_{P'P} = x_{P'P}/t_{P'P}$.

A *network* is a digraph of nodes $L \in \boldsymbol{L}$ and arcs embedded in the (t, x) plane, where nodes are points and arcs are valid paths. The set of nodes on the boundary will be denoted $\boldsymbol{B} \equiv \boldsymbol{L} \cap \boldsymbol{D}$. These nodes will be called *origins*. The conventional "from-to" notation is used to label arcs, so that $L'L$ denotes the arc pointing from L' to L. For each $L \in \boldsymbol{L}$, $F(L) \subset \boldsymbol{F}(L)$ is the set of predecessor ("from") nodes, and $T(L) \subset \boldsymbol{T}(L)$ the set of successor ("to") nodes. Associated with each arc, $L'L$, there are four quantities: its time duration $t_{L'L} > 0$, distance $x_{L'L}$, slope $v_{L'L}$, and a cost $c_{L'L}$ given by the path-cost of the arc. We shall always choose least-cost arcs. Therefore, $c_{L'L} \equiv \Delta_{L'L}$. A walk is a series of arcs defining a path between two nodes. The cost of the cheapest walk from $B \in \boldsymbol{L}$ to $E \in \boldsymbol{L}$ will be denoted c_{BE}.

A network is said to be *geometric* if the underlying graph is geometric; i.e., if its arcs do not intersect – except at end-points. A network is said to be *validly connected* if every valid pair of nodes is connected by a walk. A network is said to be *sufficient* if the optimum walk between every valid pair of nodes is an optimum path. A network is said to be *z-sufficient* with sufficiency *level* z if its best walks are within z units of optimality; i.e., if $E \in \boldsymbol{L}$ and $B \in \boldsymbol{F}(E) \cap \boldsymbol{L} \Rightarrow c_{BE} \leq \Delta_{BE} + z$. Sufficient networks will play an important role.

2.2 General results

In this subsection R is an arbitrary (bounded) function that can be discontinuous. In subsequent sections, R has the form of the model under consideration.

Let c_L denote the least cost of reaching a node L with a walk from $\boldsymbol{B}$; i.e.,

$$c_L = \min_{B \in \boldsymbol{B}} \{N_B + c_{BL}\}. \tag{7}$$

This expression can be evaluated for all $L \in \boldsymbol{L}$ with efficient shortest path algorithms in a time that increases with the number of arcs. We look for networks with few arcs for which $c_L = N_L$ or, failing this, $c_L \approx N_L$. The first two basic facts give sufficient conditions for this to happen.

Let $\mathcal{B}_{\mathcal{P}^*} \in \boldsymbol{D}$ be the source of a path $\mathcal{P}^*$ that solves (3) for a node L. The following fact should be obvious from our definitions, and a comparison of (7) and (3):

> **Fact 1**: If $\mathcal{P}^*$ emanates from an origin of a $z-$sufficient network ($\mathcal{B}_{\mathcal{P}^*} \in \boldsymbol{B}$) then $c_L - z \leq N_L \leq c_L$. If the network is sufficient, $c_L = N_L$. □

The second fact specifies a condition for *all* paths emitted by the boundary. This is more restrictive but easier to check. The following lemma is proven first. Refer to Fig. 2.

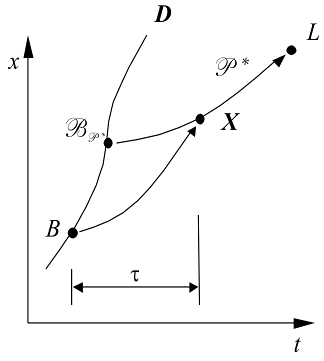

Figure 2: Interception of an optimum path.

> **Lemma 1**: If $\mathcal{P}^*$ can be intercepted in time less than or equal to τ by a valid path from an origin B of a z-sufficient network, then:
>
> $$c_L - (\tau \widetilde{R} + \alpha |B - \mathcal{B}_{\mathcal{P}^*}| + z) \ \leq \ N_L \ \leq \ c_L. \tag{8}$$

Proof: The second inequality is obviously true since optimization problem (2) is a relaxation of (7). To prove the first inequality, let X be the point of interception and note: $\Delta_{\mathcal{B}_{\mathcal{P}^*}L} \geq \Delta_{XL} \geq \Delta_{BL} - \Delta_{BX} \geq \Delta_{BL} - \tau\widetilde{R} \geq c_{BL} - \tau\widetilde{R} - z$. [The first inequality is based on the non-negativity of path costs; the second on the triangle inequality; the third on the definition of $\widetilde{R}$ as an upper bound; and the fourth on z-sufficiency.] If to inequality $\Delta_{\mathcal{B}_{\mathcal{P}^*}L} \geq c_{BL} - \tau\widetilde{R} - z$ we add the relation $N_{\mathcal{B}_{\mathcal{P}^*}} \geq N_B - \alpha|B - \mathcal{B}_{\mathcal{P}^*}|$, derived from (6), the result is $N_{\mathcal{B}_{\mathcal{P}^*}} + \Delta_{\mathcal{B}_{\mathcal{P}^*}L} \geq N_B + c_{BL} - \tau\widetilde{R} - \alpha|B - \mathcal{B}_{\mathcal{P}^*}| - z$. Since the LHS is N_L and the RHS terms $N_B + c_{BL}$ satisfy $N_B + c_{BL} \geq c_L$, (8) follows. □

Clearly, if the condition of the Lemma applies uniformly to all paths, it will apply to $\mathcal{P}^*$ and (8) will hold. Thus, the following is true. Here, $\mathcal{P}$ is a valid path rooted at the boundary, and $\mathcal{B}_{\mathcal{P}} \in \boldsymbol{D}$ its beginning-point.

Fact 2: If all $\mathcal{P}$ can be intercepted in time less than or equal to τ by a valid path from an origin B of a z-sufficient network such that $|B - \mathcal{B}_{\mathcal{P}}| \leq h$, then

$$c_L - (\tau\widetilde{R} + \alpha h + z) \ \leq \ N_L \ \leq \ c_L, \quad \forall L \in \boldsymbol{L}. \ \square \tag{9}$$

The quantity $\tau\widetilde{R} + \alpha h + z$ in (9) is a uniform error bound. Its first two terms should be small for dense networks, and the third vanishes for sufficient networks. This suggests that to solve practical problems, we should look for dense sufficient networks with nodes placed where the solution is desired. This turns out to be easy for homogeneous problems with piecewise linear (PWL) fundamental diagrams.

2.3 Additional results for homogeneous problems

Since Q is homogeneous, it is now written as a function of a single argument, $q = Q(k)$. The same convention is used for the cost function $R(x')$ and the wave velocities $w(k)$. Four additional facts are now presented. The first one (Fact 3) does not assume that Q is PWL.

Fact 3 (straight paths): If Q is homogeneous, the cost between two arbitrary points P and P' is minimized by a straight path.

Proof: This is true because for homogeneous problems all valid straight paths are waves and all waves are least cost paths; see [1].[2] □

[2] Alternatively, note $R(x')$ is convex. Thus, Jensen's inequality implies $(t_{P'} - t_P)\langle R(x')\rangle \geq (t_B - t_{P'})R(\langle x'\rangle)$,

The remaining results pertain to cases where Q is PWL. As we shall see, the PWL assumption allows optimum paths in the continuum to have corners and form a network.

Let $m \geq 1$ be the number of linear segments in Q, w_i the slope (wave velocity) of segment i, and q_i and k_i the flow and density at the right end of the segment. It is understood that $k_i < k_{i+1}$, $k_0 = -\infty$, and $k_m = +\infty$. See Fig. 3. Note that $w_i > w_{i+1}$ since Q is concave and that, as shown by the figure, R is itself PWL with m segments and corners where $x' = w_i$.

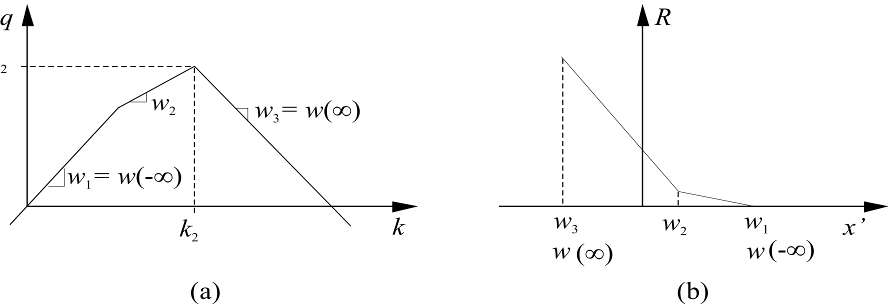

Figure 3: Piecewise linear problem: (a) fundamental diagram; (b) cost function.

Refer now to Fig. 4a. A key result is:

Fact 4 (path equivalence): If $q = Q(k)$ is piecewise linear and a path $x(t)$ with endpoints P and P' is such that $x' \in [w_{i+1}, w_i]$ then $x(t)$ is optimal with cost:

$$\Delta_{PP'} = (t_{P'} - t_P)R(v_{P'P}) = (t_{P'} - t_P)q_i - (x_{P'} - x_P)k_i. \tag{10}$$

Proof: Note first that the linear path from P to P', which is optimal by virtue of Fact 3, satisfies $x' \in [w_{i+1}, w_i]$. Thus, we just have to show that if $x(t)$ satisfies $x' \in [w_{i+1}, w_i]$ its cost is the RHS of (10). However, if $x' \in [w_{i+1}, w_i]$ the derivative of the objective function in (5) with respect to k, $w(k) - x'$, changes from a value $w_{i+1} - x' < 0$ to a value $w_i - x' \geq 0$ at $k = k_i$. Thus, the maximum of (5) is k_i, and $R(x') = Q(k_i) - k_i x' = q_i - k_i x'$ everywhere along $x(t)$. Inserting this expression in (4), assuming a straight path where $R(x') \equiv R(v_{P'P})$, we obtain the middle member of (10). If the same is done for a generic path one obtains the last member. □

where $\langle \cdot \rangle$ is the time-averaging operation along the path. This proves the result too since the LHS of this inequality is another way of writing (4) for a path with our two endpoints, and the RHS is the cost of the straight path.

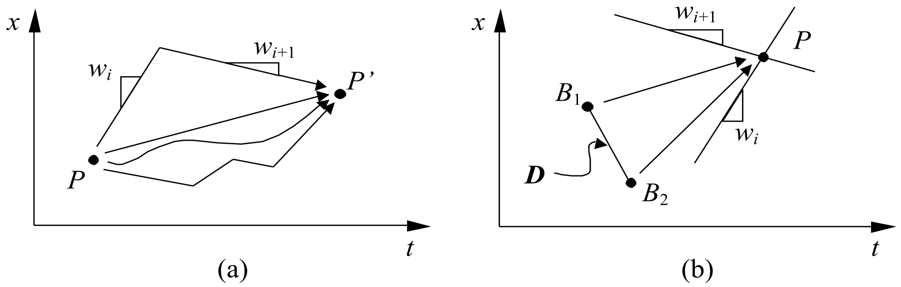

Figure 4: Optimality conditions for PWL problems: (a) paths; (b) sources.

The four paths of Fig. 4a are optimum between P and P'. Note in particular that *any* PWL path with slopes w_i and w_{i+1} going from P to P' is optimal. In the important case where $m = 2$ this means the following:

Fact 5 (path equivalence; $m = 2$): If $m = 2$ all valid paths are optimal, all validly connected networks are sufficient, and the least cost between two points P and $P' \in \boldsymbol{T}(P)$ is a linear function of their coordinates:

$$\Delta_{PP'} = (t_{P'} - t_P)(w_1 - v_{P'P})k_1 = (t_{P'} - t_P)q_1 - (x_{P'} - x_P)k_1. \ \square \tag{11}$$

Now, refer to Fig. 4b and note:

Lemma 2: If a boundary $\boldsymbol{D}$ is a linear segment between points B_1 and B_2, and P is such that $v_{B_1P}, v_{B_2P} \in [w_{i+1}, w_i]$ for some i, then the minimum of (3) is reached at either B_1, B_2, or both. $\square$

This is true because (10) is now linear in the coordinates of B and, therefore, (3) is a linear program. As a result, all intermediate data for $B \in (B_1, B_2)$ can be ignored. This allows us to establish:

Fact 6 (exact solution, $m = 2$): The prediction error $|c_L - N_L|$ for a node L of a validly connected network vanishes if: (i) the boundary and data are linear in (t, x) between consecutive origins; and (ii) the endpoints of $\boldsymbol{F}(L) \cap \boldsymbol{D}$ (i.e., the relevant part of the boundary for node L) are origins.

Proof: These two conditions guarantee that the relevant portion of the boundary can be partitioned into linear segments with origins at their extremes and linear data. For $m = 2$ all the segments of the boundary satisfy the conditions of Lemma

2. Thus, an optimum path to L must emanate from one of the origins. Since validly connected networks are sufficient for $m = 2$ (see Fact 5) Fact 1 can be applied to point L. Hence, $c_L = N_L$. □

The rest of this paper shows how well problems of increasing complexity can be solved with networks.

3 Homogeneous problems

Section 3.1 defines relevant network structures, and Sec. 3.2 how they should be adapted to specific problems. In Sec. 3.1 and the first two subsections of Sec. 3.1, Q is piecewise linear; this is relaxed in Sec. 3.1.3.

3.1 Sufficient network structures

We examine here possible network geometries, starting with the case of $m = 2$.

3.1.1 Geometric networks ($m = 2$)

Any geometric network formed by two families of parallel equidistant lines with slopes w_i and time separations ϵ_i ($i = 1, 2$) is validly connected; see Fig. 5a. Networks of this type are sufficient, as per Fact 5. With only two arcs per node they are also efficient for computation.

The nodes of any such network form an oblique lattice and its arcs partition the plane into similar cells. If one sets $\epsilon_1 = \epsilon_2 \equiv \epsilon$, as in Fig. 5b, then nodes also line-up along "space-rows." This is useful when the boundary has sections where t is constant. The distance between space-rows is $q_1 \, \epsilon/\kappa$. If $\frac{q_1}{\kappa w_i}$ is rational for $i = 1, 2$, then a subset of the nodes also define a rectangular lattice, as shown by the circled dots of Fig. 5c.

3.1.2 Lopsided networks ($m = 2$)

If the solution is only sought for the (circled) nodes of the rectangular lattice, all other nodes and many arcs can be eliminated while retaining sufficiency. The result of this trimming is shown in Fig. 5d. Note: (i) horizontal arcs have been introduced to preserve sufficiency; (ii) the new network is not geometric; (iii) its nodes form a rectangular lattice with 3 arcs per node; and (iv)

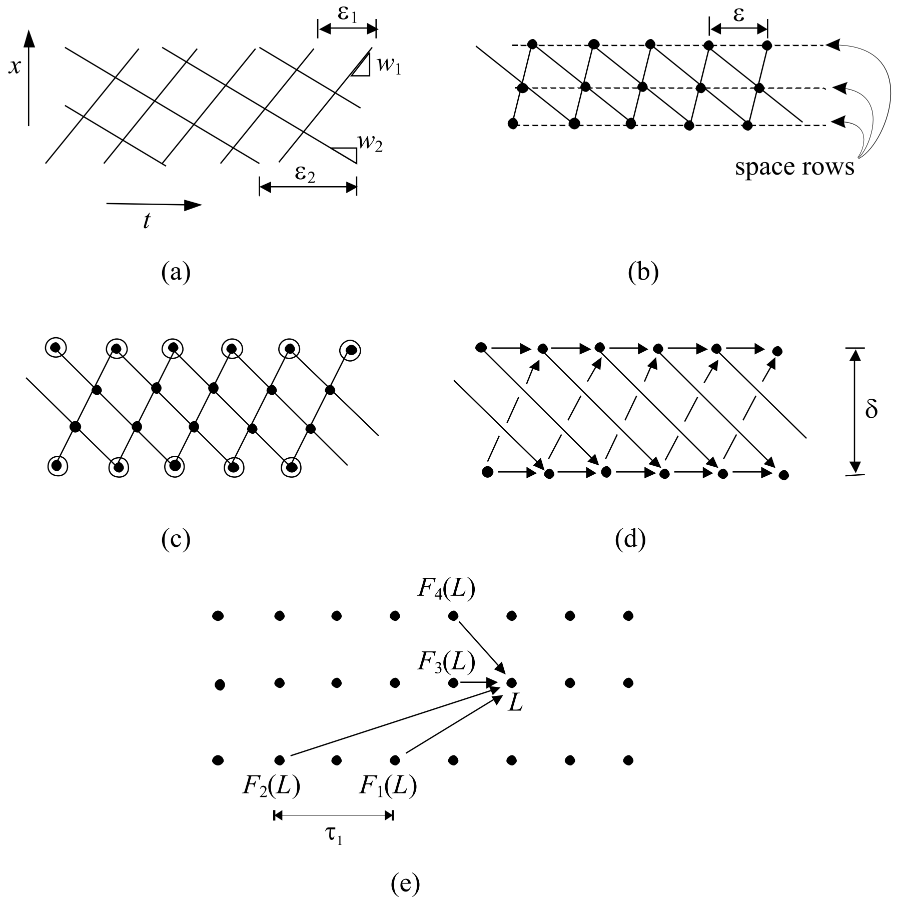

Figure 5: Network types: (a) geometric; (b) geometric with space-rows; (c) geometric with rectangular lattice; (d) lopsided ($m = 2$); (e) lopsided ($m = 3$).

nodes continue to line up along lines with slopes w_1 and w_2. The trimming procedure does not depend on specific values of w_1 and w_2. Thus, it applies generally to any problem with $m = 2$.

Let δ be the lattice interval between space-rows, and note that additional rows of nodes could have been deleted from the rectangular-lattice without destroying sufficiency or changing the ratio of arcs to nodes. Thus, δ can be increased. By deleting rows selectively, one could vary it in space. This flexibility and the lopsided meshes that non-geometric networks allow are of much value in traffic applications, where one is often only interested in the solution at widely spaced points such as freeway ramps. Lopsided meshes allow us to place solution nodes only where desired without the computational burden of intermediate nodes and links. The advantage is considerable when one wishes to use small time increments because by holding δ constant, the number of arcs is only proportional to ϵ^{-1} and not to ϵ^{-2}, as would occur with a geometric mesh.

3.1.3 Lopsided networks ($m > 2$)

Useful networks with translational symmetry can also be defined for the case with $m > 2$. Here we only analyze lopsided networks with one arc per wave speed. Figure 5e is an example. It depicts the set of arcs pointing to a generic node. Arcs in networks of this type never span more than one space row. For the pattern to be possible δ/ϵ must be an integer multiple of all the w_i. (Appropriate δ and ϵ always exist if the w_i are rational numbers.) As in the triangular case, we always include an arc with zero slope, $w_i = 0$, whether or not the fundamental diagram requires it, and index it appropriately. The cost rate for these arcs, $R(0)$, is the maximum flow.

We now show that the resulting network is z-sufficient. To this end, let τ_i be zero if the i and $i+1$ predecessor nodes of L are on different space rows, and equal their time separation otherwise; i.e., $\tau_i = 0$ if $w_i \cdot w_{i+1} \leq 0$, and $\tau_i = |\frac{\delta}{w_i} - \frac{\delta}{w_{i+1}}|$, otherwise; see Fig. 5e.

> **Fact 7:** The lopsided network is z-sufficient with $z = \max_i\{\tau_i(R(0) - q_i)\} \leq \delta\kappa$.
>
> **Proof:** We establish z-sufficiency first, and then prove the last inequality. Every node pair $L'L$ with $v_{L'L} \in [w_{i+1}, w_i]$ can be joined by a walk consisting of a horizontal sub-walk from L' to X along the space-row of L' and an optimum sub-walk from X to L with slopes w_{i+1} and w_i. Since $c_{L'L}$ cannot exceed the sum of the costs of the two optimum sub-walks, we can write $c_{L'L} \leq \Delta_{L'X} + \Delta_{XL}$. However, (10) yields $\Delta_{L'X} = t_{L'X}R(0)$ and $\Delta_{XL} = \Delta_{L'L} - t_{L'X}q_i$. Thus, $c_{L'L} \leq \Delta_{L'L} + t_{L'X}(R(0) - q_i)$. Consideration of the network geometry shows that the duration of the horizontal sub-walk satisfies $t_{L'X} \leq \tau_i$. Thus, $c_{L'L} \leq \Delta_{L'L} + \tau_i(R(0) - q_i) \leq \Delta_{L'L} + z$. Thus, the network is z-sufficient. To establish the last inequality note that $\tau_i \leq |\frac{\delta}{w_i}|$; thus, $\max_i\{\tau_i(R(0) - q_i)\} \leq \max_i\{|\frac{\delta}{w_i}|(R(0) - q_i)\}$. Furthermore, $\frac{R(0)-q_i}{|w_i|} \leq \kappa$. Obviously then, $\max_i\{\tau_i(R(0) - q_i)\} \leq \max_i\{|\frac{\delta}{w_i}|(R(0) - q_i)\} \leq \delta\kappa$. □

Note that $z = 0$ for $m = 2$ since all the τ_i vanish in this case. This agrees with the findings of Sec. 3.1.2.

Since the geometric and lopsided networks are translationally symmetric, dynamic programming yields all the c_L in a time proportional to the number of arcs. The basic recursion is:

$$N_L = \min_i\{N_{F_i(L)} + R_i\}, \tag{12}$$

where $F_i(L)$ is the predecessor of node L corresponding to slope w_i, and R_i the cost the corresponding arc, $R_i = (t_L - t_{F_i(L)})R(w_i)$.

3.2 Solution properties

We consider here the precision of the method for increasing values of m.

3.2.1 Exact solutions ($m = 2$)

Problems with $m = 2$ and simple boundaries, such as the initial value problem or the finite highway problem can be solved by aligning the nodes of a dense sufficient network with the boundary and evaluating its shortest paths. If the nodes of these networks are aligned with lines of slopes w_1 and w_2 (as occurs with the geometric or lopsided networks) condition (ii) of Fact 6 will hold for most nodes. If in addition the boundary data are PWL, condition (i) should also hold. The procedure is then exact for these nodes.

Figure 6 shows three networks that solve PWL initial-value problems exactly for most of their nodes. Only the lopsided mesh introduces errors – in the shaded area, where condition (ii) is violated.

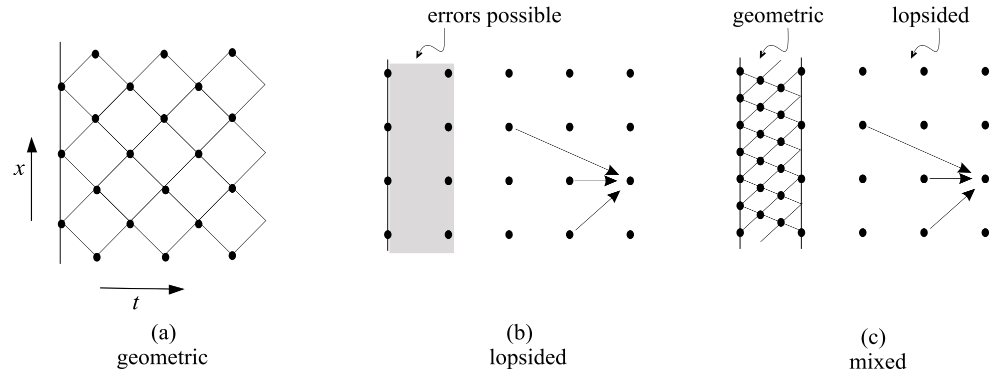

Figure 6: Three network geometries that produce exact solutions for an IVP ($m = 2$).

Note that the number of nodes (arcs) with both geometric and lopsided networks is $O(\delta^{-1}\epsilon^{-1})$. Thus, the computational effort needed to cover a solution domain is $O(\epsilon^{-2})$ for a geometric network and only $O(\epsilon^{-1})$ for a lopsided network with fixed δ. This improves on methods based on conservation laws.

3.2.2 Solutions with uniformly bounded error ($m \geq 2$)

For general problems with complex boundaries it may not be possible to align network nodes with the boundary. But if a network is expanded by adding extra origins on the boundary with a maximum L_1 separation $h = O(\delta)$, and enough connector arcs to preserve the original z-sufficiency level $z \leq \delta\kappa$, then the error in the network solution will be $O(\delta)$, independent of problem size. This is guaranteed by Fact 2. Proper connection methods are described below.

Consider first the case with $m = 2$. An expanded network will continue to be sufficient if it continues to be validly connected. To ensure this, connect each new origin B to all the "orphan" nodes in its range of influence (i.e., all $L \in \boldsymbol{T}(B)$ that cannot be reached from another node in $\boldsymbol{T}(B)$: $\boldsymbol{T}(B) \cap F(L) = \emptyset$); and from all the "childless" nodes in its domain of dependence (i.e., all $L \in \boldsymbol{F}(B)$ such that $\boldsymbol{F}(B) \cap T(L) = \emptyset$). If any boundary nodes are in $\boldsymbol{T}(B)$ or $\boldsymbol{F}(B)$, they must also be considered for connection, unless already connected by a prior step.

Since geometric and lopsided networks have translational symmetry in the directions with slopes w_1 and w_2 the number of connectors is small. For the lopsided network there can be at most two orphan nodes and two childless nodes for every B; see Fig. 7a. For the geometric network there can be at most one node of each type. Thus, an expanded network can be constructed simply by overlaying the network on the solution domain, placing a node at every point where the boundary $\boldsymbol{D}$ intersects an arc and replacing the intersected arc by two (connector) arcs. If the network is geometric we are done. If the network is lopsided two more connector arcs per intersection should be added.

Lopsided networks with $m > 2$ can be similarly connected to preserve their z-sufficiency level. Consider all the nodes in the wedge formed by two rays with slopes w_i and w_{i+1} emanating from a new origin B. Optimum walks from B to all the nodes in this wedge, satisfying the conditions of Fact 4, will exist if a connector is added from B to every node L that is an orphan with respect to directions i and $i+1$; i.e., where $F_i(L)$ and $F_{i+1}(L)$ are outside $\boldsymbol{T}(B)$. The connecting patterns are as in Fig. 7b and c – depending on whether one of the slopes is zero or not. Circled dots have been used to denote the nodes to which B should be connected.[3] Using this convention, Fig. 7d displays the complete set of nodes from/to which B should be connected for a problem with $m = 3$. Note that connectors are only needed for next-neighbor space-rows and that an upper bound for the duration of a connector is $\delta/\underline{w} + \epsilon$, where $\underline{w}$ is the smallest non-zero $|w_i|$.

[3] Of course, if any boundary nodes are in $\boldsymbol{T}(B)$ or $\boldsymbol{F}(B)$, they too should be considered for connection.

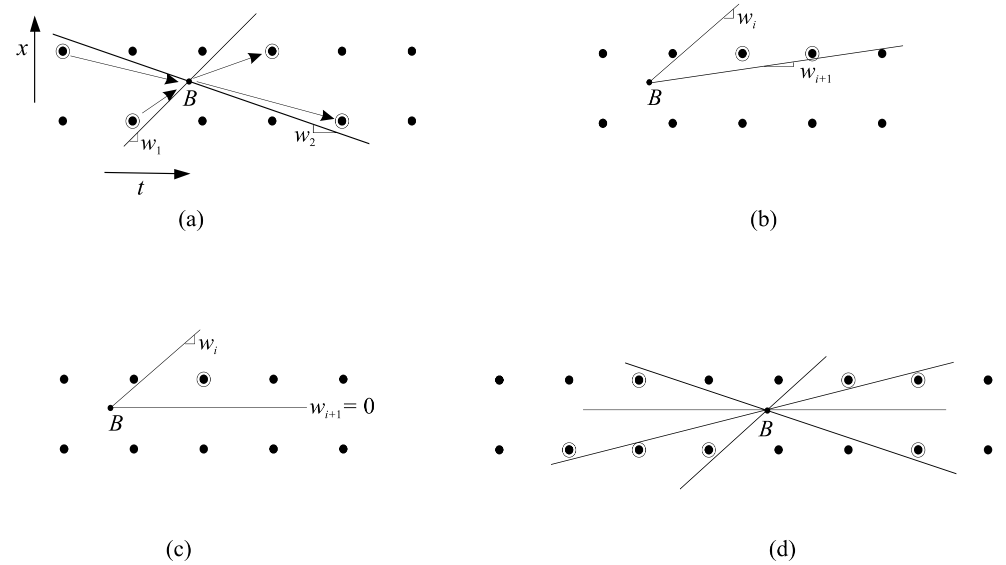

Figure 7: Connected nodes for different wave-speed ranges.

3.2.3 Solutions with bounded error for smooth FD's ($m = \infty$)

Here, we evaluate the error committed when one approximates a smooth R by a PWL $R^{(m)}$ with m segments and uses a lopsided network to make a prediction. The following is true:

> **Fact 8:** Assume m is even and the slopes w_i are evenly spaced as follows: $w_i - w_{i+1} = w(-\infty)(2/m)$ if $w_i, w_{i+1} \geq 0$; and $w_i - w_{i+1} = -w(\infty)(2/m)$ if $w_i, w_{i+1} \leq 0$. Then, if the oldest point on the boundary is at $t = 0$: $|c_L - N_L| = O(m\delta) + O(t_L m^{-2})$.
>
> **Proof:** For this sequence of slopes the sufficiency level of the lopsided network given by Fact 7 can be shown to satisfy $z \leq \delta m R(0)/(w(-\infty) - w(\infty))$; i.e., $z = O(m\delta)$. Thus, according to Fact 2, the network error – i.e., the absolute difference between the exact cost for the approximate problem $N_L^{(m)}$ and our prediction $c_L^{(m)}$ – should be $|N_L^{(m)} - c_L^{(m)}| = O(m\delta)$. Since R is smooth the maximum deviation between R and $R^{(m)}$ should be quadratic in $1/m$; i.e., $\sup |R - R^{(m)}| = O(m^{-2})$. Thus, the error due to the PWL approximation is $|N_L^{(m)} - N_L| = O(t_L m^{-2})$.[4] Obviously, the total prediction error is $|c_L^{(m)} - N_L| = O(m\delta) + O(t_L m^{-2})$, as claimed.□

[4] This is true because both problems are homogeneous. Hence, a straight segment joining any valid node-pair $L'L$ is an optimum path in both cases; see Fact 3. Since the exact and approximate costs for this path can differ at most by $t_{L'L} \sup |R - R^{(m)}|$ it follows that $|N_L^{(m)} - N_L| = O(t_L m^{-2})$.

Note that the error is globally bounded but increases with t_L.

Consider now the calculation effort. We recognize that to accommodate large m with a lopsided mesh with fixed δ, ϵ must be reduced; e.g., by halving it for every doubling of m. Thus, the number of arcs, and therefore the computation effort, is of order $O(m\delta^{-1}\epsilon^{-1}) = O(m^2\delta^{-1})$.

This shows that the asymptotic error for $t_L \to \infty$, $O(t_L m^{-2})$, varies inversely with the computation error. This is the same error-effort ratio as in second-order methods based on conservation laws. The advantage of the network method is that it is proven to converge, has a quantifiable global error and does not yield negative flows.

4 Inhomogeneous Problems

This section examines four generalizations: (i) discrete shortcuts in a homogeneous continuum; (ii) piecewise homogeneous problems; (iii) self-similar highways; and (iv) general problems.

4.1 Shortcuts ($m = 2$)

We consider here homogeneous problems with shortcuts (moving bottlenecks in traffic flow lingo) and only address the issue when $m = 2$ because this is of most relevance to traffic flow, but a generalization is possible. Shortcuts are valid paths, $x = C(t)$, with a reduced cost per unit time $0 \leq R_c(t) \leq R(\frac{dC}{dt})$. The cost function for a homogeneous problem with shortcuts is: $R^*(x', t, x) = R_c(t)$, if $x = C(t)$ and $x' = R(\frac{dC}{dt})$; and $R^*(x', t, x) = R(x')$, otherwise.

Assume first that there is only one continuous shortcut. The triangle inequality ensures that an optimum path form P' to P must use the shortcut continuously, or not at all, joining and leaving the shortcut just once. Furthermore, if access and egress is only allowed at specific points along the shortcut, and the longest duration between consecutive points is ϵ' or less, then the cost of the shortest path is increased by at most $2\epsilon'\widetilde{R}$. This bound is now exploited.

Consider a pre-existing validly connected network, which is sufficient for the original problem, and form an expanded network by adding the following items: (i) a set of closely spaced nodes along the shortcut; (ii) arcs with cost given by R^* between consecutive new node pairs along the shortcut; and (iii) enough connectors with cost R between these nodes and the nodes of the original network to ensure that the expanded network is validly connected. Since the expanded network contains shortest paths for all valid pairs $L'L$ where either L or L' is on the shortcut, our bound guarantees that it is z-sufficient for the shortcut problem with $z = 2\epsilon'\widetilde{R}$.

If one overlays the shortcut on either a geometric or lopsided network and adds nodes at the points of intersection, few connectors are needed; see Sec. 3.2.2. The construction is particularly easy for the geometric case because in this case only the first and last points of the shortcut need a connector; see Fig. 8.

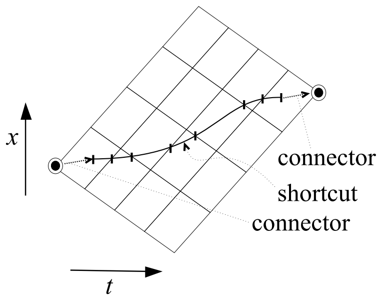

Figure 8: Network representation of a shortcut ($m = 2$).

Define now $\max\{\epsilon_1,\ \epsilon_2\} \equiv \epsilon$ in the geometric case. Consideration shows that the longest duration of a shortcut arc is $\epsilon' \leq \epsilon$ in, both, the geometric and lopsided cases. Thus, the expanded network is z-sufficient,[5] with $z = 2\epsilon\widetilde{R}$. As a result, the least cost walks of the expanded network solve any well-posed problem with a uniformly bounded error $O(2\epsilon\widetilde{R})$; see Fact 2.

Multiple shortcuts can be treated individually as suggested above – if shortcuts intersect, new nodes can optionally be included at the points of intersection. Obviously, an error of order $z = \epsilon\widetilde{R}$ arises every time an optimum path would join or leave a shortcut. Therefore, if J_L is the number of times that an optimum path to node L joins or leaves a shortcut, the error in N_L is $O(J_L\epsilon\widetilde{R})$. Of course, if $J_L \leq J$, the error is uniformly bounded. A uniform bound should exist for problems with a finite number of slow-varying shortcuts of finite duration, as often occurs in traffic flow applications.

We assumed in the above that new nodes are only placed at points where the shortcut(s) intersect the network, and this led to errors of order ϵ. But this is not the only option. Accuracy can be improved without much extra computational burden by decreasing the maximum spacing between shortcut nodes, ϵ', further – since errors are $O(\epsilon'\widetilde{R})$ but the total computational effort is less sensitive to ϵ'.

[5] If C_1 is linear and R_1 time-independent the expanded geometric network can be shown to be sufficient.

4.2 Piecewise homogeneous problems ($m \geq 2$)

The results of the previous sections can be applied to piecewise homogeneous (PWH) problems consisting of different but homogeneous sections, l, with PWL fundamental diagrams, $Q^l(k)$. A superscript l is used to index the sections, so that for example the length of the l^{th} section is L^l.

Since the results of Sec.2.2 apply to inhomogeneous problems, they guarantee that to solve a PWH problem with uniformly bounded error it suffices to construct a dense (z)-sufficient network for the PWH problem. This, of course can be done by constructing homogeneous (z)-sufficient networks for the pieces, putting extra nodes and connectors at the boundaries as needed, and joining the pieces together. For problems where L^l is an integer multiple of the separation(s) between-space-rows, δ (or δ^l), no extra nodes and connectors are necessary. If two parallel links would run between consecutive nodes along the boundary, only one link – with the least cost of the two – would have to be retained. These ideas are illustrated in Fig. 9a, which is a geometric network representation ($m = 2$) of a PWH problem including a traffic signal followed by a lane-drop. Arcs upstream and downstream of the lane drop (the dotted line) would have to be associated with costs $R^{(1)}$ and $R^{(2)}$ consistent with the number of lanes.

The procedure can be used with moving bottlenecks; including cases where the bottleneck changes the character of the road ahead and behind it. Examples of this application are snowplows and intermittent bus lanes which preclude usage of the bus lane by other traffic, e.g., 2 minutes prior to a bus arrival, on a rolling basis. The network of Fig. 9b shows how an intermittent bus lane could be modelled in the neighborhood of a traffic signal, explicitly recognizing that traffic is allowed to follow the bus in the express bus lane. All arcs upstream of the bus trajectory would receive costs based on the full complement of lanes, and all arcs downstream (in the shaded area) based on one less lane.

The next two subsections examine time-independent problems where $Q(k, x)$ is not piecewise homogeneous in space. This creates difficulties because it prevents the discretization of space on a fixed scale, e.g., with $\delta \approx 1$ km.

4.3 Self-similar highways ($m \geq 2$)

We consider here (self-similar) highways where flow and density are related by the same FD everywhere, except for a space-dependent scaling transformation; i.e., Q is a function of two arguments with the following form: $Q(k, x) = \alpha(x)Q_o(k/\alpha(x))$. The function $\alpha(x)$ is a proxy

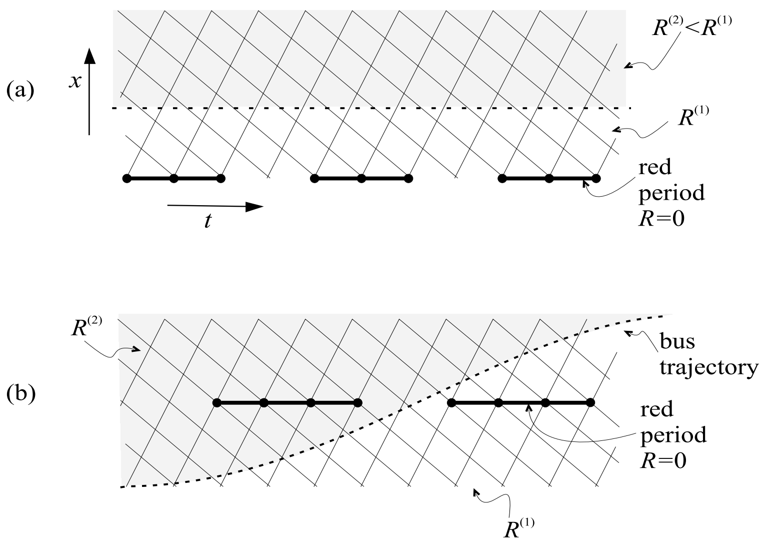

Figure 9: Inhomogeneous highways: (a) lane drop and traffic signal; (b) snow-plow and traffic signal.

for the number of lanes and Q_o for the behavior of a single lane. Consideration of (5) reveals that the cost function is $R(x', x) = \alpha(x)R_o(x')$, in agreement with the physics of the problem.

Assume first that Q_o is triangular and use $\alpha'(x)$ for $d\alpha(x)/dx$. Then we have:

Fact 9 ($m = 2$): If either $\alpha'(x) > 0$ or $\alpha'(x) < 0$, the geometric network is sufficient.

Proof: Consideration of the Euler equation of the calculus of variations for this problem shows that all optimum paths must be piecewise linear with $x'(t) \in \{w_i\}$ for all segments. Furthermore, the optimum path from L' to L can only have one corner. It must follow the lower boundary of $\boldsymbol{F}(L) \cap \boldsymbol{T}(L')$ if $\alpha'(x) > 0$, or the upper boundary if $\alpha'(x) < 0$. Clearly, if L' and L are nodes of a geometric network, the corner is also on the network and the optimum path is a walk. □

It follows, in view of Fact 2 that these types of problems are solved with a uniformly bounded error $O(\delta)$ if one assigns the exact cost to the arcs of the network. This cost is: $c_{L'L} = \Delta_{L'L} = [A(x_{L'}) - A(x_L)]R_o(v_{L'L})t_{L'L}$, where $A(x)$ is the indefinite integral of $\alpha(x)$.[6]

[6] If one assigns to the arcs an approximate cost based on the value of R at their mid-point instead of the above, an error $O(\delta^2)$ is introduced for every arc and the total error in the final solution is $O(t_L\delta^2)$. The error is of second order, but not uniformly bounded.

If $\alpha'(x)$ changes sign the highway can be divided into sections where either $\alpha'(x) < 0$, $\alpha'(x) > 0$, or $\alpha'(x) = 0$, and treated as in the previous subsection. The accuracy of the method is not affected by the decomposition.

Calculus of variations considerations also show that if $m > 2$ optimum paths can have more corners but they are still convex (or concave), with at most $m - 1$ corners, if $\alpha > 0$ (or $\alpha < 0$). Therefore, lopsided networks with small δ can be used to address this type of problem because they always contain piecewise linear walks with proper slopes and corners close to those of an optimum path. Lopsided networks are not sufficient, however, and the accuracy of this method not as good as in the triangular case.

By increasing m one can approximate a smooth FD as closely as desired, and in this way solve the general self-similar highway problem. Unfortunately, the accuracy to efficiency ratio deteriorates with increasing m, and the method does not compare as favorably with conventional first-order methods for these types of problems. First order methods are discussed below.

4.4 First-order procedures for general problems and discussion

First order accuracy for general problems can also be obtained using lopsided networks and rectangular lattices (t_r, x_s) with $\delta \to 0$ by successive linearizations, as in Godunov's method [2]. The discrete solution for the nodes L at time t_{r+1} is obtained from the discrete solution at t_r by solving a *continuum* shortest path problem from the line $t = t_r$ to all the nodes on the line $t = t_{r+1}$. To do this, it is assumed that $N(t_r, x)$ is linear in x between nodes and that the cost function is homogeneous within each cell of the mesh; i.e., that $R(x', t, x) = R(x', x_s + \delta/2)$ for $t \in [t_r, t_r + \epsilon]$, $x \in [x_s, x_s + \delta]$. Shortcuts can be superimposed on the mesh. The continuum shortest path problem to each L is simple if just a few shortcuts cross $\boldsymbol{F}(L)$ in the relevant time interval. This condition can be enforced by decreasing ϵ. To improve efficiency, ϵ can be varied from step to step.

If for a problem with shortcuts and $m = 2$ one uses variable time steps to avoid three-way shortcut interactions in $\boldsymbol{F}(L)$, the method reduces to that in [3]. If for a problem without shortcuts one uses short steps consistent with "Courant's stability condition", the method reduces to Godunov's. The proposed method, however, is numerically stable with larger steps – since the linearization and shortest path processes are non-expansion mappings. Larger steps increase accuracy at the expense of additional calculations, but do not remove the first-order limitation.

The methods in this paper can be used to enhance hybrid models of traffic flow [4] that consider the interactions between (discrete) slow vehicles such as trucks and buses and a continuum KW stream of automobiles.

Acknowledgment

PhD candidate Jorge Laval offered many constructive comments. Research supported by NSF Grant CMS-0313317 to the University of California, Berkeley.

References

[1] C F Daganzo. A variational formulation for a class of first order PDEs. Technical Report UCB-ITS-RR-2003-03, Inst. Trans. Studies, Univ. of California, Berkeley, CA, 2003.

[2] S Godunov. A difference scheme for numerical computation of discontinuous solutions of equations of fluid dynamics. *Mat. Sb.*, 89(47):271–306, 1959.

[3] C F Daganzo and J A Laval. Moving bottlenecks: A numerical method that converges in flows. Technical Report UCB-ITS-RR-2003-2, Inst. Trans. Studies, Univ. of California, Berkeley, CA, 2003.

[4] J A Laval and C F Daganzo. A hybrid model of traffic flow: Impacts of roadway geometry on capacity. *Trans. Res. Rec. (submitted)*, 2003.