

A VARIATIONAL FORMULATION OF KINEMATIC WAVES: BOTTLENECK PROPERTIES AND EXAMPLES

Carlos F. Daganzo and Monica Menendez, Institute of Transportation Studies, University of California, Berkeley, California, USA

ABSTRACT

It has been recently shown that all kinematic wave traffic problems with concave fundamental diagrams can be reformulated as continuum shortest (least cost) path problems in the time-space domain. This reformulation expands the kinds of kinematic wave problems that can be solved efficiently. Examples are inhomogeneous problems with combinations of gradual, moving and time-dependent bottlenecks.

The paper examines in detail the special case where the fundamental diagram is triangular. It shows that in this case the new procedures compare very favorably with standard methods based on conservation laws. The numerical error is shown to be small and uniformly bounded; zero in important cases. Formulas and examples are given.

The paper also proves that the maximum difference between the vehicle number function (the solution) of any problem with a bottleneck of finite dimension, and the number function of a version of the same problem with a point bottleneck cannot exceed the maximum number of vehicles that fit in the real bottleneck. When this number is small, point-bottleneck idealizations can be used. They require less data and are easier to solve.

INTRODUCTION

This paper proposes two algorithms that, based on a variational version of kinematic wave (KW) theory (Daganzo, 2003, 2003a, 2005), can solve complex inhomogeneous KW problems with triangular fundamental diagrams (FDs) with precision and simplicity. Methods for general concave FDs are discussed in Daganzo (2003, 2003a). The proposed algorithms can be used with any combination of gradual, discontinuous, moving and fixed bottlenecks -- even if the character of the road changes in space-time; e.g., due to snow-plows.

The gradual fixed-bottleneck problem was introduced in Lighthill and Whitham, (1955) and later solved analytically in Newell (1999), but only for some special cases. The discrete moving-bottleneck problem was introduced in Gazis and Herman (1992), and formulated as a KW problem in Newell (1993 and 1998). Muñoz and Daganzo (2002) modeled it in a more general way by treating the moving bottleneck as a special boundary condition. Analytical solutions have only been developed for special cases, however. Existing traffic simulation methods (e.g., Giorgi et al, 2002, and Daganzo and Laval, 2003) only provide first-order approximations for these problems. The new algorithms will be shown to be much better. An upper bound for their numerical error will be presented, along with examples.

The paper will also show that real bottlenecks can be approximated by bottlenecks of zero length, and how to formulate proper boundary conditions to represent discrete bottlenecks. This puts moving bottleneck theory and practice on a solid foundation.

The remainder of this section summarizes the essential facts of variational theory. Following sections present: (i) two algorithms to solve a special family of problems (self-similar problems) without point bottlenecks; (ii) extensions for problems involving moving point bottlenecks; (iii) applications of these results to examine the effect of bottleneck length; and (iv) a simple and very accurate solution method to solve general problems with any number of moving point bottlenecks.

Basic Facts

Consider a one-directional road on which vehicles are conserved, and let x , t , q and k respectively denote the distance along the road (increasing in the direction of travel), time, flow and density. The last two variables are functions of time and space $q(t, x)$ and $k(t, x)$. Given is an FD, $q = Q(k, t, x)$, which is assumed to be concave and piecewise differentiable in k , and to satisfy $q = 0$ for $k = 0$ and $k = \kappa$, where κ is the jam density; see Figure 1(a). As proposed in Newell (1993a), we shall describe the KW solution in terms of a vehicle number function, $N(t, x)$, whose partial derivatives, $q = \partial N / \partial t$ and $k = -\partial N / \partial x$ are the flow and density functions. The KW solution satisfies: $\partial N / \partial t = Q(-\partial N / \partial x, t, x)$.

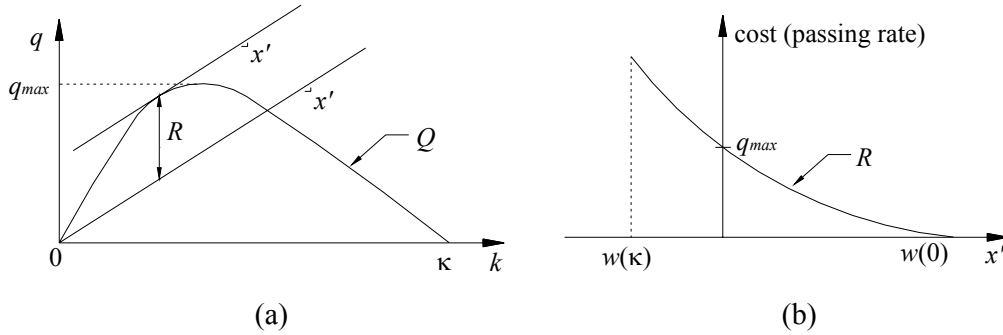


Figure 1. Basic concepts of Variational Theory: (a) fundamental diagram, and (b) cost function.

The shortest path problem is formulated in terms of the following “cost” function, R , which has units of vehicle number per unit time (i.e., flow):

$$R(x', t, x) = \sup_k \{Q(k, t, x) - kx'\}. \quad (1)$$

Figures 1(a) and 1(b) show graphically how Q and R are related; note that $-R$ is the Legendre-Fenchel transform of Q . The cost function is a fundamental property of the highway, just like the fundamental diagram. Physically, it represents the maximum rate at which traffic can pass an observer moving with speed x' . Note that the maximum flow (the highway capacity, q_{max}) is $R(0, t, x)$. For traffic problems R is non-negative, non-increasing, convex (if Q is concave), and only defined for the range of valid wave speeds w . (Recall that the wave speed is $w(k, t, x) = \partial Q / \partial k$.) In the triangular case R turns out to be linear.

The cost function gives the cost per unit time of moving along a space-time path \mathcal{P} with trajectory $x(t)$. We say that a path is “valid” if it is a piecewise differentiable curve with slopes x' in the range of allowable wave-speeds $[w(\kappa, t, x), w(0, t, x)]$. The cost of traversing a valid path is:

$$\Delta(\mathcal{P}) = \int_{t_B}^{t_P} R(x', t, x) dt, \quad (2)$$

where t_B and t_P are the times associated with the path endpoints. Variational theory states that if the vehicle number is given along a boundary curve (or curves), \mathbf{D} , and we think of these as “start up costs”, then the vehicle number at any point in the solution domain is the least cost to reach the point with a valid path from the boundary, including the start up cost; i.e.:

$$N_P = \min_{\mathcal{P} \in V_P} \{ \Delta(\mathcal{P}) + N_{\mathcal{B}_P} \}, \quad (3)$$

where V_P is the set of valid paths \mathcal{P} from B to P , and $\mathcal{B}_P \in \mathbf{D}$ is the start point of \mathcal{P} . Equation (3) is a calculus of variations problem. Its iso-cost contours are the vehicle trajectories.

Networks

In this paper, a network is a digraph with nodes L embedded in time-space, with directed arcs LL' . Arcs are defined only for node pairs that can be connected by a valid path. We call these “valid node pairs,” and say that a network is “validly connected” if all its valid node pairs are connected by a walk. Each arc is assigned a cost, $c_{LL'}$, equal to that of an optimum continuum path between its end nodes, and a duration and distance, $t_{LL'} > 0$ and $x_{LL'}$, consistent with the node coordinates.

It is shown in Daganzo (2003, 2003a) that if the FD is piecewise linear there are networks whose shortest “walks” (network paths) between all valid node pairs are shortest continuum paths. These networks are said to be “sufficient” because by solving the shortest path problem on the network one solves the continuum problem exactly for all its valid node pairs. This is advantageous but does not imply exactness. Prediction errors still arise if a network does not have enough nodes on the boundary (origins) to sample the data with enough frequency. We call these discrepancies “sampling errors.” The rest of this section presents some key results.

Error Bounds

To quantify sampling errors, e_s , we shall assume (reasonably) that the cost function is bounded, with $R(k, t, x) \leq \tilde{R}$, and that the boundary data satisfy a Lipschitz continuity condition; i.e. that there is a $\beta > 0$ such that

$$|N_B - N_{B'}| \leq \beta |B - B'|, \quad \forall B, B' \in \mathbf{D} \quad (4)$$

where $|B - B'| = |x_B - x_{B'}| + |t_B - t_{B'}|$. We quantify sampling frequency with two constants, h and τ , as follows. For every valid path to a node L with beginning point $\mathcal{B}_P \in \mathbf{D}$ there should be: (i) a network origin B such that $|B - \mathcal{B}_P| \leq h$, and (ii) a short valid path that intercepts the original path from B in time less than τ . Figure 2 illustrates the concept. For any h and τ with these properties, sampling errors satisfy: $e_s \leq \tau \tilde{R} + \beta h$.

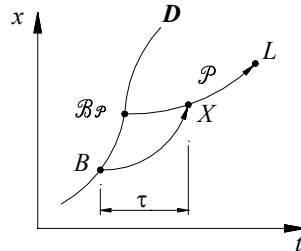


Figure 2. Interception of an optimum path.

If a network is not sufficient the solution will also include network errors, e_n . A network where the best walk between any valid node pair is within z cost units of the continuum

shortest path (i.e., where $e_n \leq z$) is said to be z -sufficient. It is shown in Daganzo (2003) that the total prediction error e for a z -sufficient network satisfies $e \leq e_s + e_n$; i.e., that:

$$e \leq \tau \tilde{R} + \beta h + z \tag{5}$$

Triangular FDs

The two basic wave speeds of a triangular FD, whether homogeneous or not, will be denoted $w_1 \equiv w(0, t, x)$ and $w_2 \equiv w(\kappa, t, x)$; w_1 is the free-flow speed and w_2 the “backward” wave speed. We consider first problems with homogeneous FDs, $q = Q(k)$. The following two results, proven in Daganzo (2003a), will turn out to be useful.

Result 1 (sufficiency): If the FD is triangular and homogeneous, all valid paths are optimum and all validly connected networks sufficient. \square

Since all valid paths are optimum, a straight path is optimum. Therefore, in this case it is easy to set up networks and calculate their arc costs; the arc cost formula is:

$$c_{LL'} = t_{LL'} R(x'), \quad \text{where } x' = x_{LL'} / t_{LL'}. \tag{6}$$

This leads to the following:

Result 2 (exactness): If (i) the FD is triangular and homogeneous, (ii) the network is validly connected, (iii) the data and the boundary are linear between network origins, and (iv) the network contains straight walks with speeds w_1 and w_2 from the boundary to every node, then $e = 0$. \square

Result 2 is true because problem (3) is a linear program, so that its optimum paths must either emerge from vertices or have extreme speeds.

Inhomogeneous roads where the FD changes with space by a similarity transformation also have useful properties. In these cases, the FD is $q \equiv Q(k, x) = \alpha(t, x) Q_o(k / \alpha(t, x))$, where $\alpha(t, x) > 0$ is a proxy for the number of available lanes, which can vary with t and x . The similarity relation implies that the wave speeds are space-independent and that the cost function is self-similar: $R(x', t, x) = \alpha(t, x) R_o(x')$. If a highway is homogeneous then $\alpha(t, x)$ is constant, $\alpha(t, x) = \alpha$. If it includes a moving bottleneck then $\alpha(t, x) < \alpha$ in a space-time swath along the bottleneck trajectory of a width comparable with the length of the bottleneck. The following result, proven in Daganzo (2003a) with calculus of variations, will be useful.

Result 3 (shortest paths): If (i) the FD is triangular and self-similar, and (ii) $\alpha(x)$ is non-increasing (non-decreasing), then the valid path that connects P and P' with the highest (lowest) possible trajectory is shortest. \square

Because the wave speeds are constant, these extreme paths are bilinear (piecewise linear with two components) as shown in Figure 3. Note that if $\alpha(x)$ decreases (the highway narrows), the optimum path from P to P' is concave - leaving P with slope w_1 and reaching P' with slope w_2 when the slope of $\overline{PP'}$ is neither w_1 nor w_2 . If the highway widens the path is convex. We now turn our attention to more practical matters.

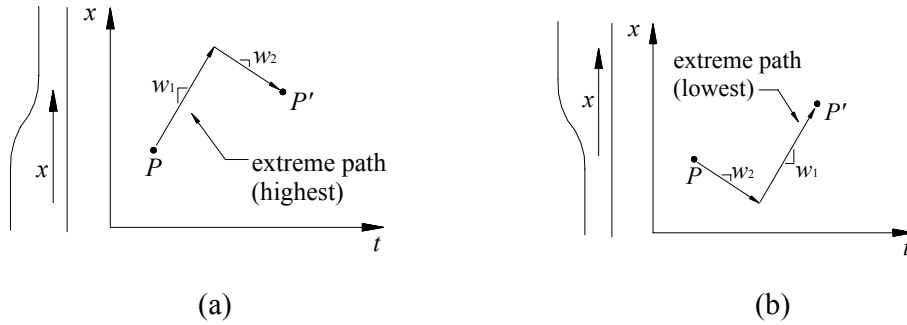


Figure 3. Shortest paths for similar-monotonic cost functions: (a) decreasing $\alpha(x)$ (road narrows), and (b) increasing $\alpha(x)$ (road widens).

SELF-SIMILAR PROBLEMS

We examine here procedures for solving general self-similar problems, and then analyze their accuracy.

Solution Methods

The procedures in question have two basic features: an efficient network geometry and a fast shortest path algorithm. They are now described.

Geometric networks: A geometric digraph formed by two families of parallel equidistant lines with slopes w_i and time separations ε_i ($i = 1, 2$) is called a *geometric network* if all its arcs point in the direction of increasing time. Figure 4(a) displays an example. Note that nodes exist at every intersection and that they form an oblique lattice. Note too that any connected subgraph of a geometric network is a validly connected network. In the special case where $\varepsilon_1 = \varepsilon_2 \equiv \varepsilon$, the lattice includes space-rows, as shown in Figure 4(b). The row spacing is denoted δ .

Lopsided networks: Assume now that we delete any number of space rows from a geometric network with space rows, but maintain all the node connections across the deleted rows as shown in Figure 4(c). Assume too that we introduce horizontal links along all the remaining rows, also as shown. It then turns out that the resulting network is validly connected, as the reader can easily verify.

We call these networks “lopsided” because the row spacing can be increased as much as desired without changing ε by selective deletion of rows. This is of considerable computational advantage when results are only sought at widely spaced locations. Note that every node L of a lopsided network has exactly three “from” nodes. In this case too, the row spacing is denoted δ .

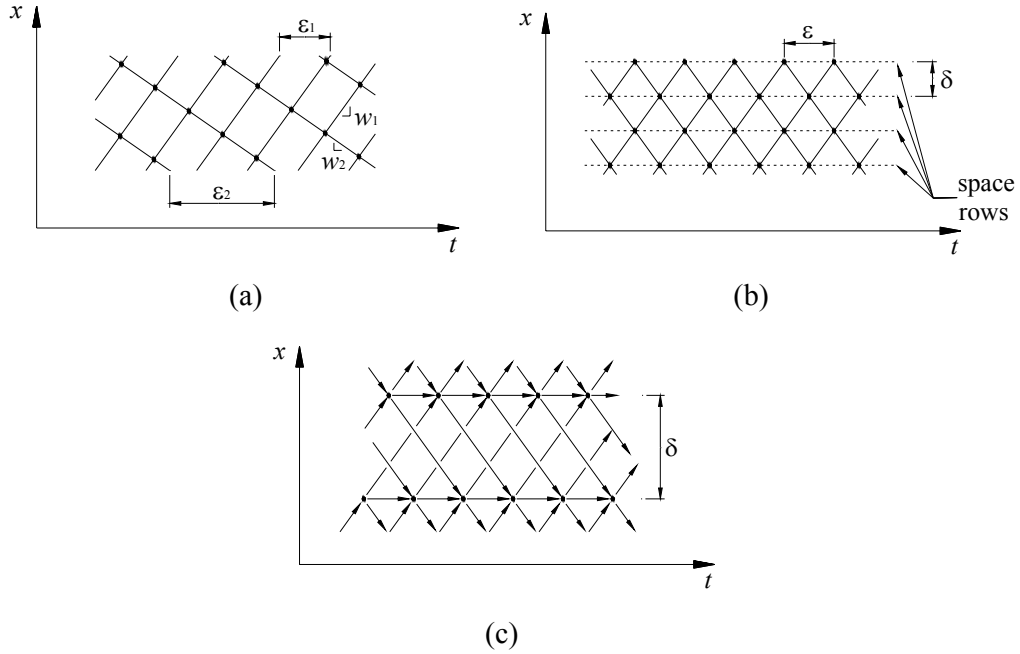


Figure 4. Network types: (a) geometric, (b) geometric with space-rows, and (c) lopsided.

Shortest path method: To solve a problem we first assign costs $c_{LL'}$ to every link, and then use a shortest path algorithm to compute the optimum cost for every node in the solution domain, c_L . Costs for origin nodes are known since they are part of the boundary data: $c_L = N_L$. Link costs should be computed by minimizing Eq. (2). In the homogeneous case the answer is Eq. (6). In the self-similar case we can simply evaluate the cost of an extremal path for each arc; see Result 3. If we use a geometric network with short arcs we can also use Eq. (6) as an approximation with $R(x')$ replaced by $R(x', t, x)$ and evaluated at the midpoint of the arc. This only introduces errors of order $O(\varepsilon^2)$ as $\varepsilon \rightarrow 0$, and can be done even in the general non-self-similar case.

Since our networks have translational symmetry, dynamic programming (DP) can be used for the optimization. As a result, the calculation effort to obtain all the c_L only increases linearly with the number of arcs. The recursion is:

$$c_{L'} = \min_{L \in F(L')} \{c_L + c_{LL'}\}, \quad (7)$$

where $F(L')$ is the set of “from” nodes for L' . When the DP algorithm is used with a geometric (or lopsided) network it will be respectively called the GDP (or LDP) algorithm. Since both network types are sparse (with only 2 or 3 links per node) the calculations of Eq. (7) involve a very small number of comparisons and additions per node. The effort per node is comparable to that required by solution methods based on conservation laws. The LDP method requires considerably fewer nodes, though.

Accuracy

Since a geometric network includes walks with the extremal property of Result 3 for every valid node pair, it follows that a geometric network is sufficient for the self-similar problem if α is non-decreasing (or non-increasing). Consideration shows that this continues to be true for problems with general but time-independent $\alpha(x)$ if we add nodes and horizontal links with appropriate costs at all the locations where $d\alpha(x)/dx$ is a minimum; i.e., at bottlenecks. Therefore, the GDP algorithm avoids network errors in the self-similar case. Note too from Result 1 that both DP algorithms avoid network errors in the homogeneous case.

In the above instances only sampling errors remain. According to Eq. (5) these sampling errors are bounded by: $\tau\tilde{R} + \beta h$. For a geometric network with space rows the parameters τ and h are respectively bounded by ε and $(\varepsilon+2\delta)$. Thus, the sampling error is bounded by $\varepsilon[\tilde{R} + \beta + 2\beta(\delta/\varepsilon)]$. Since $\delta/\varepsilon = (w_1^{-1} - w_2^{-1})^{-1}$, we see that the quantity in brackets is a constant. Hence, the error bound is $O(\varepsilon)$ and uniform; i.e., it is independent of the size of the solution domain. The bound can be tightened in two important cases: (i) homogeneous problems with linear data between origin nodes (when $e = 0$); and(ii) problems with smooth data (when $e = O(\varepsilon^2)$ for $\varepsilon \rightarrow 0$).

The exactness of case (i) is a direct consequence of Result 2. The situation arises frequently, since data for typical problems is usually available in piecewise linear (PWL) form.

Case (ii) applies when both the data and the boundary can be expressed as smooth functions of a parameter, π , such that the maximum jump in π across consecutive origins is $O(\varepsilon)$. The result is true because under its conditions the least cost to reach L from a point on the boundary, $N_L(\pi)$, is a smooth function, with a smooth global minimum, N_L . Since sampling errors for smooth minima become quadratic in the sampling interval if the sampling intervals are reduced toward zero, and since the sampling intervals are $O(\varepsilon)$ for $\varepsilon \rightarrow 0$, it follows that the error is $O(\varepsilon^2)$ for $\varepsilon \rightarrow 0$.

As an illustration of this result, Figure 5 shows how the GDP algorithm performs with one of the examples in Newell (1999). The example involves a self-similar highway with a re-scaled triangular FD that varies with x ($w_1 = \infty$, $w_2 = -1$ and $q_{max} = \frac{1}{2} x^2$), and gradually increasing upstream flow ($q = t$ for $t \geq 0$). Note that Newell's figure (and ours as a result) display distance in the reverse direction, using $x^* = -x$ instead of x . The example is of interest for its difficulty. In particular note that conventional methods for conservation laws would either

introduce infinite errors or use an infinite computation time, since the maximum wave speed is infinite. The GDP algorithm, however, works without a glitch. Part (a) of the figure is a reproduction of the exact wave map in Newell (1999) including the shock-path (q^* on the abscissa stands for either flow or time in this reference.) Figure 5(b) is the wave-map produced by the GDP algorithm. It matches Figure 5(a) remarkably well in all respects, including the shock-path and the position where the shock first develops at $(t, x^*) = (0.5, 1)$. Figure 5(c) gives the vehicle trajectories produced by the new algorithm. Note how the shock develops. Figures 5(b) and 5(c) were obtained with a very dense network that allowed us to estimate the solution at many relevant points appearing in Figure 5(a). Nevertheless the GDP algorithm is very accurate with large time steps. Figure 5(d) compares the estimated and exact vehicle number across time at three different locations when the time step of the GDP procedure is 0.1 time units. The error bound in this case turns out to be 0.00125 vehicles everywhere in the solution domain - imperceptible to the eye.

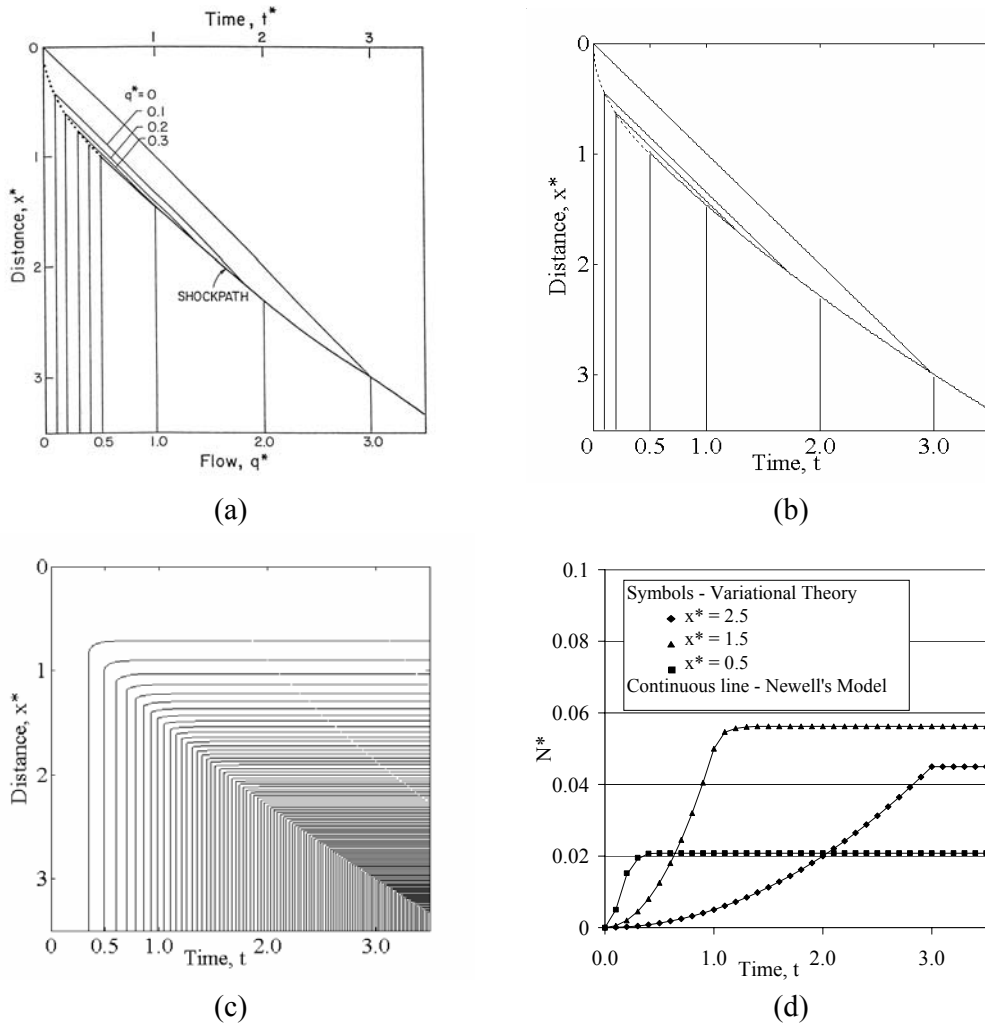


Figure 5. Solution of a gradual bottleneck problem: (a) exact wave map, (b) numerical wave map, (c) vehicle trajectories, and (d) normalized vehicle numbers, N^* , with $N/N^*=1, 10, 100$ for $x^*=0.5, 1.5, 2.5$, respectively.

MORE GENERAL PROBLEMS WITH POINT BOTTLENECKS

This section examines the structure of optimum continuum paths in a very general class of problems involving non-similar highway sections and any number of point bottlenecks. The results are then used in the following section to examine the effect of bottleneck length, and later to show how this class of general problems can be solved.

We consider first a homogeneous highway that includes a single bottleneck. We assume that the bottleneck describes a valid path \mathcal{S} with trajectory $x_B(t)$, and that the unit cost along it $r_B(t)$ is less than or equal to $R(x'_B(t))$ for all t . This allows us to state the following; see Figure 6(a):

Shortcut Theorem: If a problem is homogeneous and two points, P and P' , can be connected by a path that touches the bottleneck, then a path from P to P' is optimal if it has the following structure: (i) an access path from P to the earliest possible $S \in \mathcal{S}$; (ii) a continuous portion of the shortcut from S to the latest possible $S' \in \mathcal{S}$; and (iii) an egress path from S' to P' .

Proof: We show first that an optimum path does not have to leave and return to the shortcut. This is true because the cost of any bypass is always matched by the cost of a competing path that is infinitely close to the shortcut (see Result 1), and because the cost of the competing path cannot be less than the cost of a similar path that would use the shortcut - since the shortcut has an equal or smaller unit cost than parallel, neighboring paths. Hence, the shortcut only needs to be left once. It follows that there is an optimum path with only three components for: (i) access, (ii) use and (iii) egress from the shortcut.

To complete the proof we now show that the access and egress components with the shortest possible durations do the job. This is true for the access part because if the path were to join the shortcut at a later point than S , $S_b \in \mathcal{S}$, as shown in Figure 6(a), its cost could not decrease. The reason is that before the change, the cost of reaching S_b was the sum of the cost of the path from P to S plus the cost of the shortcut from S to S_b . But after the change, the cost is given by any valid path from P to S_b that does not use the shortcut (as per Result 1); e.g., by a path that shadows the “before” path arbitrarily close to the shortcut without using it. The cost of the new path is obviously higher or equal. The same lines of reasoning also reveal that the egress part should be of the shortest possible duration. \square

Note that the access and egress portions of the optimum paths identified by the Shortcut Theorem are straight and have slopes w_1 or w_2 , unless they connect P to the beginning of the shortcut or the end of the shortcut to P' .

The result also holds for self-similar highways where $\alpha(t, x)$ is non-increasing in x (narrowing) upstream of \mathcal{S} , and non-decreasing (widening) downstream. The reason is that under these conditions, paths on the upstream side should be as high as possible and on the

downstream side as low as possible (recall Figure 3). When this happens the steps of the proof continue to hold. The steps hold even if the upstream portion of the highway is not similar to the downstream portion, as suggested by Figure 6(b). In this figure, arrows indicate the direction in which $\alpha(t,x)$ decreases with x . Cost functions where α is monotonic in x are said to be “similar-monotonic” (SM). The following corollary summarizes these ideas:

Corollary 1: The result of the shortcut theorem also holds if the cost function is: (a) homogeneous outside the time range of the bottleneck; and (b) SM upstream and downstream within the range, with the following properties:

- (i) Upstream: $\underline{R} = \underline{\alpha}(t,x)\underline{R}_o(x')$, $\underline{\alpha}(t,x)$ non-increasing in x for $x \leq x_B(t)$;
- (ii) Downstream: $\bar{R} = \bar{\alpha}(t,x)\bar{R}_o(x')$, $\bar{\alpha}(t,x)$ non-decreasing in x for $x \geq x_B(t)$;
- (iii) $r_B(t) \leq \min\{\underline{\alpha}(t,x_B(t))\underline{R}_o(x'_B(t)), \bar{\alpha}(t,x_B(t))\bar{R}_o(x'_B(t))\}$. \square

Note that the corollary applies to shortcuts of both, finite and infinite range. Note too that the access and egress portions of the optimum paths identified by the Shortcut Theorem and this corollary have maximum slopes, unless, as before, they connect P to the beginning of the shortcut or the end of the shortcut to P' . The same reasoning of the shortcut theorem also reveals that paths where $\alpha(t,x)$ is a maximum with respect to x , “reverse-bottlenecks”, never have to be used.

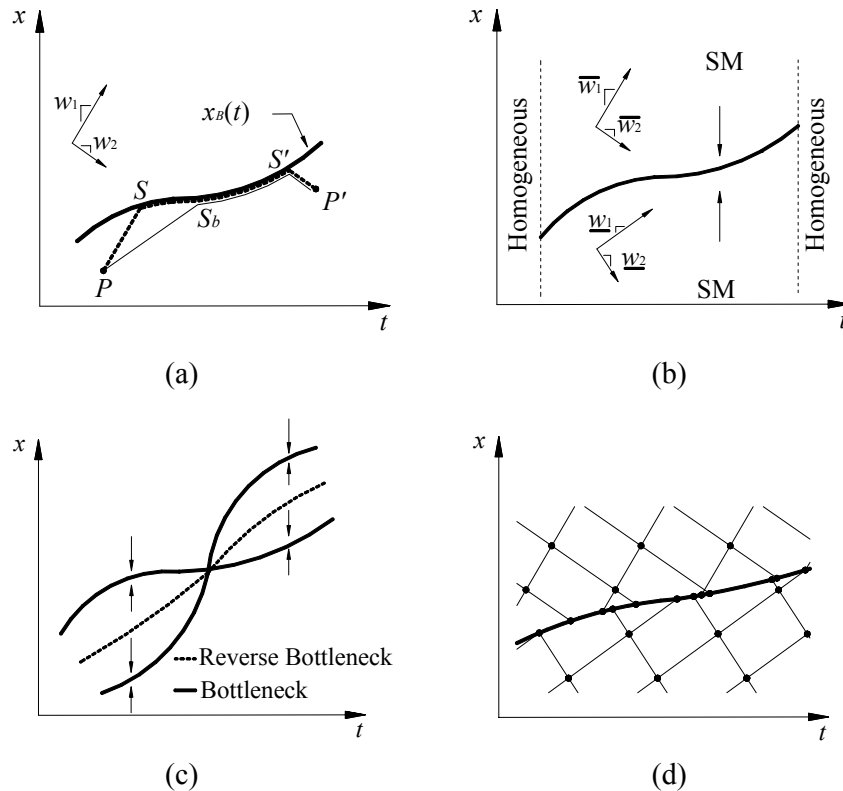


Figure 6. General problems with point bottlenecks: (a) Shortcut Theorem, (b) Corollary 1, (c) Optimum Path Theorem, and (d) composite geometric network.

All of the above suggests that problems with multiple bottlenecks can perhaps be analyzed by decomposing their solution domain into regions describable by Corollary 1, as in Figure 6(c)--in this case too, arrows indicate the direction in which $\alpha(t, x)$ decreases with x . To formalize this idea, let us define a *quasi-monotonic* problem as one whose solution domain can be partitioned by continuous curves into subregions where the cost function is similar-monotonic. Then, we have:

Optimum Path Theorem: If a problem is quasi-monotonic, then an optimum path between two points exists that is only composed of one or more pieces of the following two types: (i) continuous sections of bottlenecks, and (ii) straight segments with extremal slopes, w_1 or w_2 , appropriate for the subregion containing the segment.

Proof: The non-bottleneck portion of an optimum path can be divided into continuous parts, each embedded in a similar-monotonic subregion. Every one of these parts must minimize the cost between its endpoints. Result 3 guarantees that this is achieved for each part by a bilinear sub-path with extremal slopes w_1 and w_2 , appropriate for its subregion. \square

The class of quasi-monotone problems is very broad; it seems to encompass all KW problems of practical interest, or at least all those formulated to date. For example, the highway can be piecewise similar or piecewise homogeneous with time-dependent features; and any of these, not just the capacity, can be changed by moving bottlenecks. An example of the latter is a snowplow that changes with its passage both, the free-flow speed and the number of available lanes. The theorem underpins the justification given later in this paper for using geometric networks to solve quasi-monotone problems. The basic idea consists in using a geometric network with appropriate slopes and costs within each monotone subregion of the solution domain, and then stitching together these networks by adding along the sub-regional interfaces nodes and links with relevant costs; see Figure 6(d). Nodes and links with reduced costs should also be added along the trajectories of point bottlenecks.

If the same composition idea is used with the Shortcut Theorem instead of Corollary 1 we find the following stronger result for problems where bottlenecks and reverse bottlenecks partition the solution domain into homogeneous parts (quasi-homogeneous problems).

Corollary 2: If a problem is quasi-homogeneous, then an optimum path to a point exists that is only composed by one or more pieces of the following three types: (i) continuous sections of bottlenecks; (ii) least-duration access and egress sub-paths as in the Shortcut Theorem; and (iii) inter-bottleneck sub-paths, each with either maximal or minimal slopes.

Proof: Only (iii) needs proof. Geometrical considerations show that if a connector between bottlenecks were not to have either maximal or minimal slope, we could then find a cheaper connector, of greater or smaller slope, that would shadow one of the bottlenecks for a short time. \square

Corollary 2 applies to piecewise homogeneous highways. It will be the justification for modeling them with lopsided networks. Before this is done, we examine a question that seems to have generated some confusion in the literature. What effect can the length of a bottleneck have on traffic flow?

THE EFFECT OF BOTTLENECK LENGTH

The results of the previous section are now used to answer the question. We will assume in this section that the bottleneck moves with non-negative speed and that its trajectory \mathcal{S} is contained in a space-time swath of width $\ell(t) \leq \ell$ distance units.

Our problem is defined as follows. Inside the swath, R is defined as in Corollary 1. Outside and upstream of the swath, $\underline{R} = \underline{\alpha}_o \underline{R}_o(x')$, with $\underline{\alpha}_o \geq \underline{\alpha}(t, x)$. Outside and downstream, $\overline{R} = \overline{\alpha}_o \overline{R}_o(x')$, with $\overline{\alpha}_o \geq \overline{\alpha}(t, x)$. We shall denote the upstream and downstream jam densities outside the swath $\underline{\kappa}$ and $\overline{\kappa}$, respectively. Obviously, Corollary 1 applies to the complete problem.

We are interested in evaluating the accuracy of a point-bottleneck approximation with the same \mathcal{S} and $r_B(t)$, but a new cost function $\hat{R} = \underline{\alpha}_o \underline{R}_o(x')$ for $x < x_B(t)$ and $\hat{R} = \overline{\alpha}_o \overline{R}_o(x')$ for $x > x_B(t)$. The new cost function is homogeneous both, upstream and downstream of \mathcal{S} . Corollary 1 also applies to this approximation. We are now in position to establish the main result of this section.

Result 4 (bottleneck length): The difference in vehicle number between the solutions of the original problem and the point-bottleneck approximation is bounded from above by $\ell \max(\underline{\kappa}, \overline{\kappa})$ at every point of the solution domain.

Proof: Since Corollary 1 applies to both problems, the paths of the Shortcut Theorem are optimum for both problems. Their unit costs can only differ for the access/egress portions internal to the swath. The unit costs vanish for paths with slope \underline{w}_1 or \overline{w}_1 . Therefore, only access/egress portions with slope \underline{w}_2 or \overline{w}_2 contribute to the difference. If we let \underline{t}_2 and \overline{t}_2 be the time duration of the portions with slopes \underline{w}_2 and \overline{w}_2 , we see that the cost difference cannot exceed $\hat{R}(\underline{w}_2)\underline{t}_2 + \hat{R}(\overline{w}_2)\overline{t}_2$. Note, however, that for homogeneous FDs $R(w_2) = w_2 \kappa$. Thus, the bound is $\underline{w}_2 \underline{\kappa} \underline{t}_2 + \overline{w}_2 \overline{\kappa} \overline{t}_2$. Since the speed of the swath is non-negative we can write $\underline{w}_2 \underline{t}_2 + \overline{w}_2 \overline{t}_2 \leq \ell$. Hence, the bound satisfies $\underline{w}_2 \underline{\kappa} \underline{t}_2 + \overline{w}_2 \overline{\kappa} \overline{t}_2 \leq \ell \max(\underline{\kappa}, \overline{\kappa})$. Since the bound defined by the right side of this inequality applies to every pair of points, it follows that the difference in the predicted vehicle numbers for both problems, including the effect of the boundary data, cannot exceed $\ell \max(\underline{\kappa}, \overline{\kappa})$. \square

The quantity $\ell \max(\underline{\kappa}, \bar{\kappa})$ is a uniform error bound when one uses a point-bottleneck approximation. Thus, a simple rule is that the error in vehicle number cannot exceed the maximum number of vehicles that fit alongside the bottleneck. Obviously, the error is negligible for bottlenecks such as traffic signals, trucks, buses, snowplows, convoys of a few vehicles, and other localized restrictions. Figure 7 shows the exact and approximated solution for an example involving a gradual lane drop spanning $\frac{1}{4}$ mi.

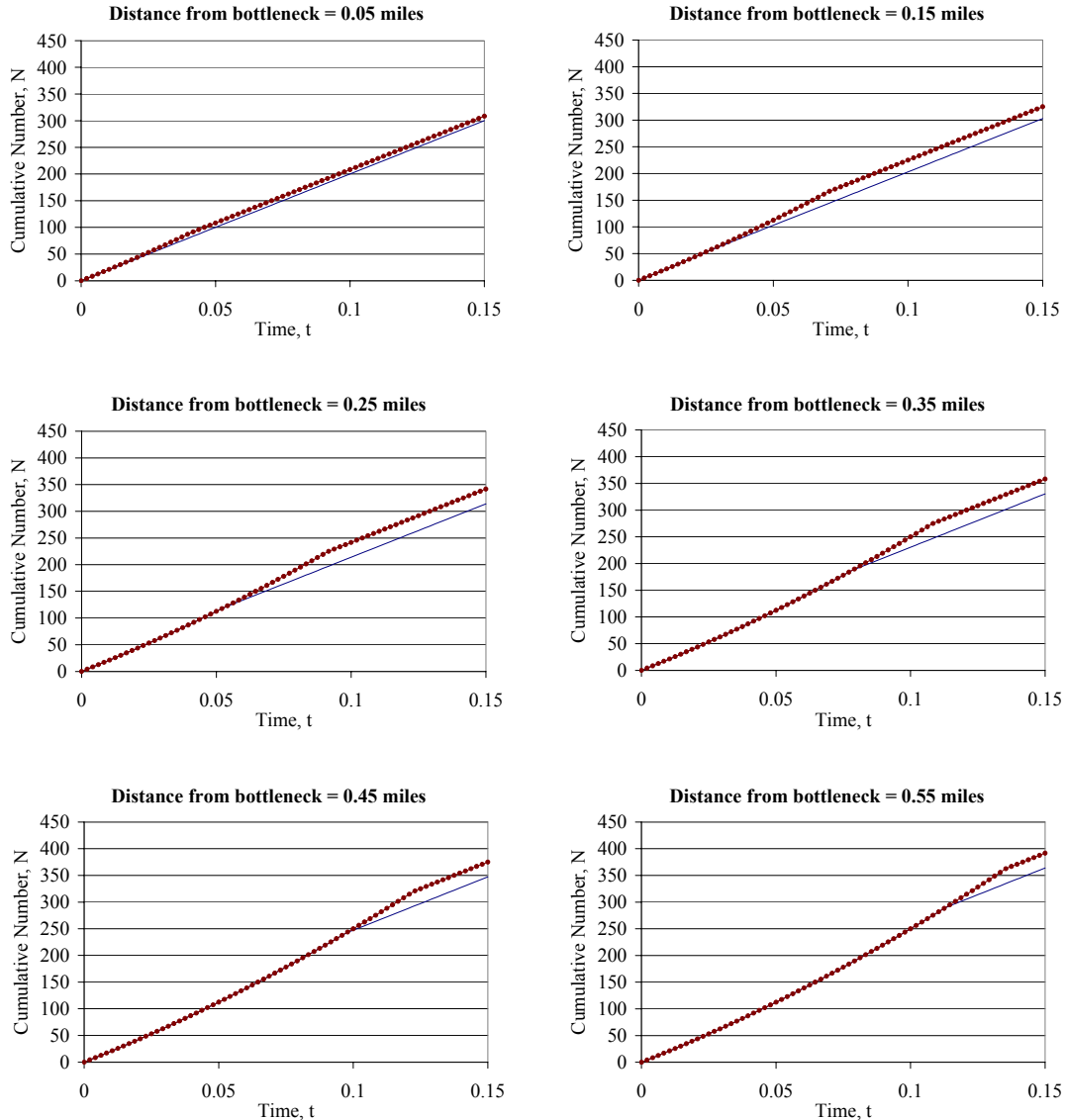


Figure 7. Error of point-bottleneck approximation. The smooth lines correspond to the exact solution (0.25 mile gradual bottleneck) and the dotted lines to the approximation (point-bottleneck).

The highway is self-similar with a triangular FD that varies with the upstream distance from the bottleneck:

$$w_1 = 60 \text{ mph}$$

$$w_2 = -15 \text{ mph}$$

$$q_{max} = \gamma_0 + \gamma_1(\ell - x)^2 \text{ for } 0 \leq x \leq \ell, \quad q_{max} = 2\gamma_0 \text{ for } x \leq 0, \text{ and } q_{max} = \gamma_0 \text{ for } x \geq \ell,$$

where $\gamma_0 = 2000$ vph, $\gamma_1 = 32000$ vph/mi², and $\ell = 1/4$ mi.

The upstream flow is assumed to increase gradually with time:

$$q = \gamma_2 + \gamma_3 t,$$

where $\gamma_2 = 2000$ vph and $\gamma_3 = 10000$ vph/hr.

The figure shows the cumulative vehicle number at six different locations as a function of time. According to Result 4, the error from an approximation using a point-bottleneck cannot exceed $\ell \max(\underline{\kappa}, \bar{\kappa}) = 83$ vehicles. This is consistent with the numerical results of the figure, which show a maximum error of around 40 vehicles. The error grows initially as we move away from the bottleneck, but after a certain point it stabilizes, as expected.

GENERAL SOLUTION METHOD

We now show and demonstrate with examples how to solve general quasi-monotonic problems with point bottlenecks. General quasi-monotonic problems with multiple point bottlenecks can be handled by superimposing all the point bottleneck shortcuts on a stitched network with geometric components, such as the one in Figure 6(d), adding nodes at the points of intersection, and connecting these with links of proper cost. Consideration shows that any such network contains walks parallel and very close to the relevant paths specified by the Optimum Path Theorem. This is also true of networks composed of lopsided components if the problem is quasi-homogeneous, by virtue of Corollary 2.

Accuracy

The error introduced by these procedures at a point P is bounded by a quantity that tends to zero as $\varepsilon \rightarrow 0$ if both the duration and number of sub-paths of the relevant optimum path specified by the Optimum Path Theorem are bounded. Consider any such relevant path. For sufficiently small ε there always is a “nearest” walk that differs from the shortest path only in the duration and time-displacement of corresponding sub-paths, and not in their slopes. These discrepancies are of order $O(\varepsilon)$. Since the number of sub-paths is $O(t_p)$, the cost difference caused by all the duration discrepancies should be $O(\varepsilon t_p)$ as $\varepsilon \rightarrow 0$. The cost difference caused by each displacement can grow at most with the product of displacement and sub-path duration. Therefore, the total displacement error is also of order $O(\varepsilon t_p)$ as $\varepsilon \rightarrow 0$. Thus, both, the combined network error and the prediction error in N_P must be $O(\varepsilon t_p)$ as $\varepsilon \rightarrow 0$.

For piecewise homogeneous problem the displacement error vanishes. Then, if the number of sub-paths in the relevant optimal path to P is uniformly bounded across all points -- a usual case in practical applications -- the network error is $O(\varepsilon)$ as $\varepsilon \rightarrow 0$. This bound is problem-dependent, but uniform.

Examples

We illustrate the accuracy and flexibility of the method with two examples. The first one, adapted from Daganzo and Laval (2003), includes a moving bottleneck on a homogeneous highway. The results of the proposed method are favorably compared to those in Daganzo and Laval (2003). The second example is more complicated; it assumes that the bottleneck changes the character of the road (as a snowplow or a police car would) and that the highway is inhomogeneous.

Example 1:

Consider a one-mile homogeneous freeway whose FD is an isosceles triangle with free flow speed, $w_1 = 60$ mph, backward wave speed, $w_2 = -60$ mph, and capacity, $q_{max} = 9000$ vph. These parameters are not realistic, but were chosen in Daganzo and Laval (2003) to separate the error of their procedure from those of the numerical KW solver—which could only handle isosceles FDs exactly. The freeway is initially in a steady state flowing at capacity when, at $t_0 = .3$ min and $x = .3$ mi, a truck traveling at $v_B = 20$ mph enters the road. The truck maintains this speed until $t = 2.1$ min, when it leaves. It is assumed that the maximum rate at which traffic can pass the truck is $r_B = 3000$ vph.

Figure 8(a) shows a time-space diagram with the exact solution of this problem, including all its shocks and interfaces, and the moving bottleneck. The figure also shows by means of solid horizontal lines the trajectories of six detectors. Figure 8(b) shows the (geometric) network we used to solve this problem with a discretization of $\Delta t = 6$ secs, including the shortcut for the moving bottleneck. Figure 9(a) shows the N -curves for both, the solution in Daganzo and Laval (2003), which used $\Delta t = 6$ secs (wiggly lines), and the exact solution (smooth lines). Figure 9(b) shows the results from variational theory. They match the exact solution without any error.

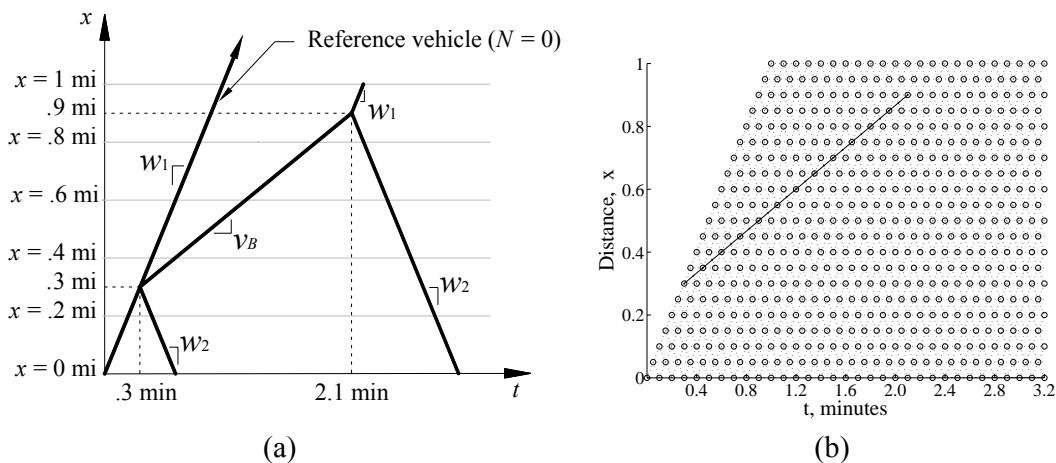


Figure 8. Example 1: (a) time-space diagram; (b) network with shortcut.

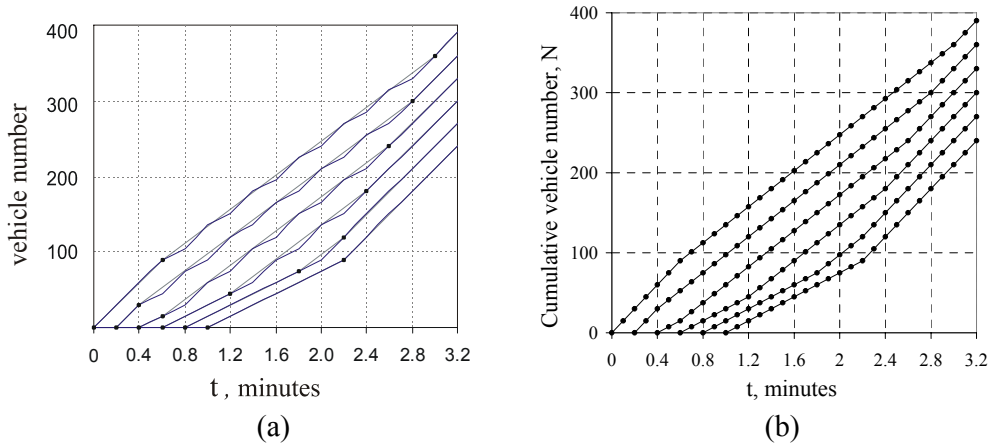


Figure 9. Example 1: (a) N -curves at six locations from Daganzo and Laval (2003); (b) N -curves at six locations from variational theory. The smooth lines are the exact solution.

Example 2:

This example includes a moving bottleneck that changes the character of the road as it travels through it. We still consider a one-mile homogeneous freeway flowing at capacity, described by an isosceles FD diagram with free flow speed, $\underline{w}_1 = 60$ mph, backward wave speed, $\underline{w}_2 = -60$ mph, and capacity, $q_{max} = 9000$ vph. In the present case, however, a slow moving police car with $v_B = 20$ mph enters the highway at $t_0 = .3$ min and $x = .3$ mi, “reminding” every vehicle that passes it to adopt the speed limit, $\bar{w}_1 = 30$ mph. (These unrealistic numbers have been chosen to dramatize the effects in our figures, but as we have proven in previous sections the procedure would perform equally well with realistic data.) In essence, the police car changes the FD downstream of its position. The downstream FD is not an isosceles triangle any more: its jam density and the backward wave speed remain unchanged (e.g., $\bar{w}_2 = \underline{w}_2 = -60$ mph but its free-flow speed becomes $\bar{w}_1 = 30$ mph $\neq \underline{w}_1$). To complete the formulation we assume that “rubbernecking” restricts the maximum rate at which traffic can pass the police car to $r_B = 1800$ vph.

As in the previous example, Figure 10(a) is a time-space diagram with the exact solution, including relevant interfaces and the trajectories for the police car and six detectors. Figure 10(b) shows the composite network for this problem, including the shortcut. Note the different slopes of the “free-flow” arcs, upstream and downstream of the shortcut. Different time increments were also used on each side of the shortcut for convenience. Figure 10(c) shows by means of continuous lines the exact N -curves for this problem, plotted on an oblique coordinate system. Dots display the results of the procedure. There is no error.

The procedure is just as quick for problems that cannot be solved easily by other means. Consider the same police car as before, but assume now that the highway is self-similar and inhomogeneous, including a gradual bottleneck:

$q_{max} = \gamma_0 + \gamma_1(\ell - x)^2$ for $0 \leq x \leq \ell$, $q_{max} = 1.5\gamma_0$ for $x \leq 0$, and $q_{max} = \gamma_0$ for $x \geq \ell$, where $\gamma_0 = 6000$ vph, $\gamma_1 = 3000$ vph/mi², and $\ell = 1$ mi.

The backward wave speed and the free-flow speed are assumed to be the same as before -- the latter dropping to 30 mph downstream of the police car. We assume that the maximum bottleneck passing rate depends on location and obeys: $r_B = q_{max}/5$. It is still 1800 vph upstream of the bottleneck, i.e., $x \leq 0$. To solve this problem we can still use the network of Figure 10(b), but with updated link costs, of course. Figure 10(d) shows the solution. A comparison with the exact solution is not made, but we know from the results of this paper that the numerical solution has a negligible error. In comparing with part (c), note the curvature of the new N -curves, and their uneven separations.

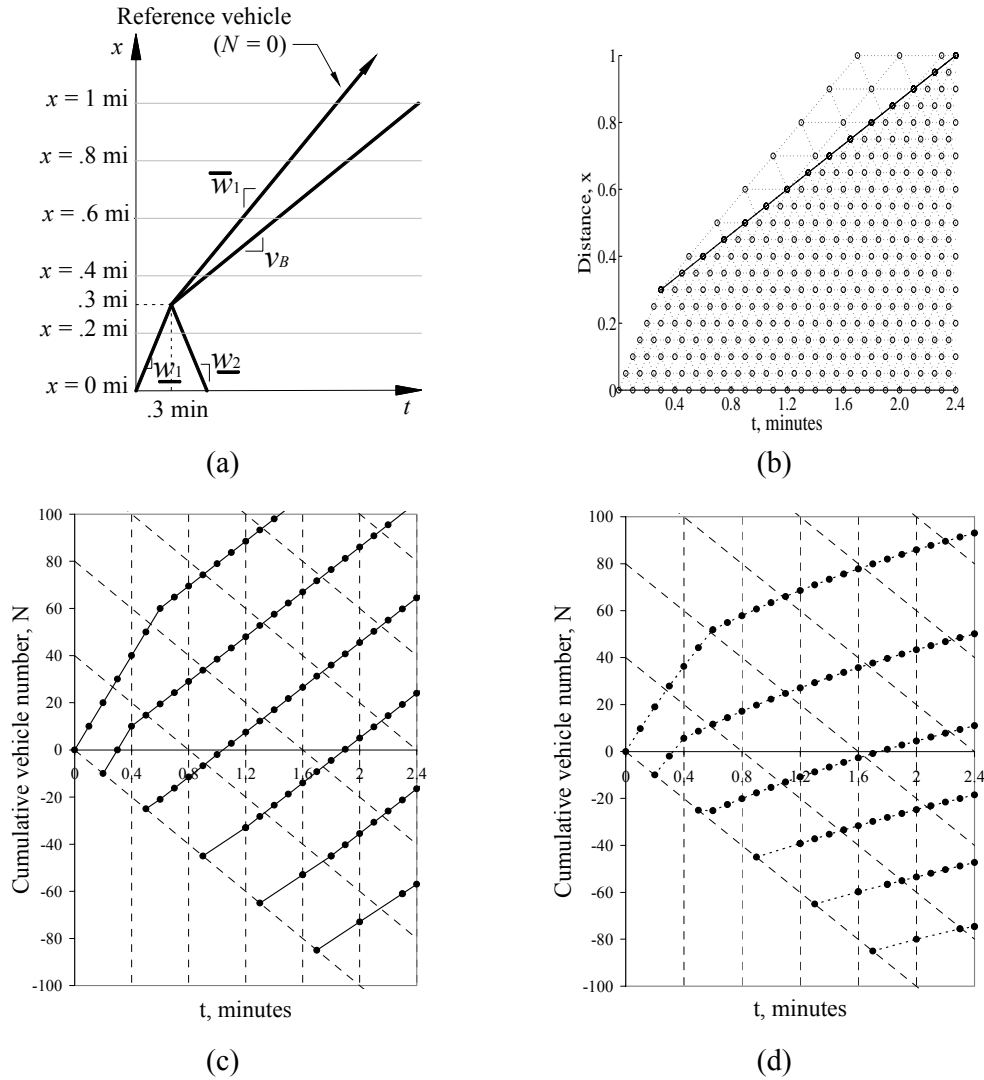


Figure 10. Example 2: (a) time-space diagram, (b) composite network grid and shortcut, (c) N -curves at six locations in oblique system: homogeneous highway, and (d) N -curves at six locations in oblique system: inhomogeneous highway. Continuous lines in part (c) correspond to the exact solution.

FINAL REMARKS

This paper has introduced recipes to solve complex KW problems with precision and simplicity. The techniques are especially useful for inhomogeneous KW problems with multiple moving bottlenecks. They can be used even if the bottlenecks change the character of the road as they move through it.

The new methods are now being applied to complex real-life problems with combinations of gradual, moving and time-dependent bottlenecks. They can help improve hybrid (discrete/continuous) models of traffic flow, where trucks and other slow vehicles are modeled as underpowered discrete particles that can both generate queues and be slowed by them; see Laval and Daganzo (2004). Hybrid models that include lane changing are currently being implemented, and initial tests with real data are very encouraging (Laval and Daganzo, 2004a). These models, for example, appear to explain the strange observations of moving bottlenecks reported in Muñoz and Daganzo (2002), and the detailed merge bottleneck phenomena in Cassidy and Rudjanakanoknad (2004). The new methods can also be used to simulate intermittent bus lanes and HOV lanes.

ACKNOWLEDGMENT

Research supported by NSF Grant CMS-0313317 to the University of California, Berkeley, and by an NSF Graduate Research Fellowship.

REFERENCES

- Cassidy, M. and J. Rudjanakanoknad. (2004). Increasing Capacity of an Isolated Merge by Metering its On-ramp. Institute of Transportation Studies Report UCB-ITS-RR-2004-03, U. of California, Berkeley, CA. *Transportation Research B* (in press).
- Daganzo, C. F. (2003). A Variational Formulation for a Class of First Order PDEs. Institute of Transportation Studies Report UCB-ITS-RR-2003-03, U. of California, Berkeley, CA.
- Daganzo, C. F. (2003a). A Variational Formulation of Kinematic Waves: Solution Methods. Institute of Transportation Studies Report UCB-ITS-RR-2003-07, U. of California, Berkeley, CA. *Transportation Research B* (in press).
- Daganzo, C. F. (2005). A Variational Formulation of Kinematic Waves: Basic Theory and Complex Boundary Conditions. *Transportation Research B* **39**(2), 187-196.
- Daganzo, C. F. and J. A. Laval. (2003). On the Numerical Treatment of Moving Bottlenecks. Institute of Transportation Studies Report UCB-ITS-PWP-2003-10, U. of California, Berkeley, CA. *Transportation Research B* **39**(1), 31-46.
- Gazis, D. C. and R. Herman. (1992). The Moving and “Phantom” Bottlenecks. *Transportation Science* **26**, 223-229.

- Giorgi, F., L. Leclercq, and J. B. Lesort. (2002). A Traffic Flow Model for Urban Traffic Analysis: Extensions of the LWR Model for Urban and Environmental Applications. *Proceedings of the 15th ISTTT*, Oxford, U.K., 393-415.
- Laval, J. A. and C. F. Daganzo. (2004). A Hybrid Model of Traffic Flow: Impacts of Roadway Geometry on Capacity. Presented at the 83rd Annual Meeting of the Transportation Research Board, Washington D.C. (submitted for publication).
- Laval, J. A. and C. F. Daganzo. (2004a). Multi-Lane Traffic Flow Hybrid Model: Quantifying the Impacts of Lane Changing Maneuvers on Traffic Flow” Institute of Transportation Studies Working Paper UCB-ITS-WP-2004-1, U. of California. (Presented at the 84th Annual Meeting of the Transportation Research Board, Washington D.C.)
- Lighthill, M. J. and G. B. Whitham. (1955). On Kinematic Waves. i Flow Movement in Long Rivers. ii A Theory of Traffic Flow on Long Crowded Roads. *Proceedings of the Royal Society A* **229**, 281-345.
- Muñoz, J. C. and C. F. Daganzo. (2002). Moving Bottlenecks: a Theory Grounded on Experimental Observation. *Proceedings of the 15th ISTTT*, Oxford, U.K., 441-462.
- Newell, G. F. (1993). A Moving Bottleneck. Institute of Transportation Studies Report UCB-ITS-RR-93-3, U. of California, Berkeley, CA.
- Newell, G. F. (1993a). A Simplified Theory of Kinematic Waves in Highway Traffic: (i) General Theory; (ii) Queuing at Freeway Bottlenecks; (iii) Multi-Dimensional Flows. *Transportation Research B* **27**, 281-313.
- Newell, G. F. (1998). A Moving Bottleneck. *Transportation Research B* **32**, 531-537.
- Newell, G. F. (1999). Flows Upstream of a Highway Bottleneck. *Proceedings of the 14th ISTTT*, Jerusalem, Israel, 125-146.