Optimal coordination of variable speed limits to suppress shock waves

A. Hegyi, B. De Schutter, and J. Hellendoorn

If you want to cite this report, please use the following reference instead:

Optimal Coordination of Variable Speed Limits to Suppress Shock Waves

A. Hegyi, B. De Schutter, and J. Hellendoorn
Delft Center for Systems and Control, Delft University of Technology
Mekelweg 2, 2628 CD Delft, The Netherlands
{a.hegyi, b.deschutter, j.hellendoorn}@dsc.tudelft.nl

Abstract—We present a model predictive control (MPC) approach to optimally coordinate variable speed limits for freeway traffic. In particular, we consider discrete-valued variable speed limits. Moreover, we also impose a safety constraint that prevents drivers from encountering speed limit drops larger than, say, 10 km/h. The control objective is to minimize the total time that vehicles spend in the network. This approach results in dynamic speed limits that reduce or even eliminate shock waves.

I. INTRODUCTION

Traffic jams do not only cause considerable costs due to unproductive time losses, but they also augment the possibility of accidents, and they have a negative impact on the environment and on the quality of life. On the short term the most effective measures in the battle against traffic congestion seem to be a selective construction of new roads and a better control of traffic by dynamic traffic management measures. We will concentrate on the latter option.

In practice, dynamic traffic management usually operates based on local data only. However, it has many advantages to consider the effect of the measures on the network level instead. So, a network-wide coordination of control measures based on global data is necessary. Since the effect of a control measure on more distant locations might only be visible after some time, a prediction of the network evolution is also necessary to achieve optimal network control. The approach presented in this paper contains both elements: network-wide coordination and prediction. Furthermore, we consider a special case of traffic control measures: variable speed limits to reduce or eliminate shock waves.

Besides prediction and coordination, the speed limit control problem has other characteristics that impose certain requirements on the control strategy:

1) There is a direct relation between the outflow of a network and the total time spent (TTS) in the network, assuming that the traffic demand is fixed. Papageorgiou [1] showed that in a traffic network, an increase of outflow of 5% may result in a decrease of the total time spent in the network of up to 20%. As the outflow is lower when the traffic is congested, one should try to prevent or postpone a breakdown as much as possible. We can conclude that any control method that resolves (reduces) congestion will at best achieve a flow improvement of approximately 5–10%, but this improvement can decrease the TTS significantly. This also means that the control strategy requires great precision. For this reason and because there are always (unpredictable) disturbances present in a traffic network, feedback control is required.

2) In practice, the speed limit signs usually display speed limits in increments of, e.g., 10 km/h. Therefore, the controller should produce discrete-valued control signals.

3) For safety it is often required (for practical implementation) that the driver should not encounter a decrease in the displayed speed limit larger than a pre-specified amount. The controller should be able to take this kind of constraints into account.

The control strategy presented in this paper includes these requirements.

In the literature, basically two views on the use of speed limits can be found. The first emphasizes the homogenizing effect [2], [3], [4], whereas the second is more focused on the preventing traffic breakdown by reducing the flow by means of speed limits [5], [6]. Several control methodologies are used in literature to find a control law for speed control, such as multi-layer control [7], sliding-mode control [6], and optimal control [2]. In a previous paper [8] we have demonstrated the effectiveness of continuous-valued speed limits against shock waves. In this paper, we focus on discrete-valued speed limits and the constraints following from the safety considerations.

Most of the models used in literature represent the speed limits by a factor that downscales the fundamental diagram (see, e.g., [2], [6]). This can give too optimistic results (see Section III-B.2), and therefore we use the METANET model which we extend with an equation that models the effect of a speed limit. We also introduce an equation to express the difference in the drivers’ anticipation to increasing or decreasing downstream densities.

This paper is organized as follows. In Section II the problem and the basic idea of the solution to moving jams is described. In Section III the basic ingredients of model predictive control are introduced, and the prediction model including the extensions is presented. The proposed control method is applied to a benchmark problem in Section IV.
II. PROBLEM DESCRIPTION

It is well known (see, e.g., [9]) that some types of traffic jams move upstream with approximately 15 km/h. These jams can remain stationary for a long time, so every vehicle that enters the freeway upstream of the jammed area will have to pass through the jammed area, which increases the travel time. Another disadvantage of the moving jams is that they are potentially unsafe. Lighthill and Whitham [10] introduced the term “shock wave” for waves that are formed by several waves running together. At the shock wave, fairly large reductions in velocity occur very quickly. In this paper we will use the term “shock wave” for any wave (the moving jammed areas) and will not distinguish between waves and shock waves, because in practice any wave is undesired.

To suppress shock waves one can use speed limits in the following way. In some sections upstream of a shock wave, speed limits are imposed and consequently the inflow of the jammed area is reduced. When the inflow of the jammed area is smaller than its outflow, the jam will eventually dissolve. In other words, the speed limits create a low-density wave (with a density lower than that in the uncontrolled situation) that propagates downstream. This low-density wave meets the shock wave and compensates its high density, which reduces or eliminates the shock wave.

III. APPROACH

A. Model Predictive Control

We use a model predictive control (MPC) scheme to solve the problem of optimal coordination of speed limits. In MPC, at each time step $k$ the optimal control signal is computed (by numerical optimization) over a prediction horizon $N_p$. A control horizon $N_c$ ($< N_p$) is selected to reduce the number of variables and to improve the stability of the system. (After the control horizon has been passed, the control signal is usually taken to be constant.) From the resulting optimal control signal only the first sample is applied, and so on. This scheme, called rolling horizon, allows for updating the state (from measurements), or even for updating the model in every iteration step. Updating the state results in a controller that has a low sensitivity to prediction errors, and updating the model results in an adaptive control system, which could be useful in situations where the model significantly changes, such as in case of incidents or changing weather conditions.

For more information on MPC we refer the interested reader to [11], [12] and the references therein.

B. Prediction model

The MPC procedure includes a prediction of the network evolution as a function of the current state and a given control input. For this prediction we use a slightly modified version of the (destination-independent) METANET model [13], [14]. The modifications are introduced to model shock waves better and to include the effect of speed limits. Note that the MPC approach is generic and will find the optimal speed limits independently of the model that is used (e.g., independently of the way in which speed limits enter the model), so the modifications are not necessary for the effectiveness of MPC. For the sake of brevity, we describe only those parts of the model that are relevant for the benchmark network of Section IV.

1) Original METANET model: The METANET model represents a network as a directed graph with the links corresponding to freeway stretches. Each freeway link has uniform characteristics, i.e., no on-ramps or off-ramps and no major changes in geometry. Each link $m$ is divided into $N_m$ segments of length $L_m$ (see Figure 1). Each segment $i$ of link $m$ is characterized by the traffic density $\rho_{m,i}(k)$ (veh/ln/km), the mean speed $v_{m,i}(k)$ (km/h), and the traffic volume or flow $q_{m,i}(k)$ (veh/h), where $k$ indicates the time instant $t = kT$, and $T$ is the time step used for the simulation of the traffic flow (typically $T = 10$ s).

The following equations describe the evolution of the network over time [13], [14]:

$$q_{m,i}(k) = \rho_{m,i}(k)v_{m,i}(k)\lambda_m$$

$$\rho_{m,i}(k + 1) = \rho_{m,i}(k) + \frac{T}{L_m}\left(q_{m,i-1}(k) - q_{m,i}(k)\right)$$

$$v_{m,i}(k + 1) = v_{m,i}(k) + \frac{T}{\tau}\left(V\left(\rho_{m,i}(k)\right) - v_{m,i}(k)\right) + \frac{T}{L_m}v_{m,i}(k)\left(v_{m,i-1}(k) - v_{m,i}(k)\right) - \frac{\eta T}{L_m}\frac{\rho_{m,i+1}(k) - \rho_{m,i}(k)}{\rho_{m,i}(k) + \kappa},$$

where $\lambda_m$ is the number of lanes in link $m$; $\tau$, $\eta$ and $\kappa$ are model parameters; and the desired speed $V$ is given by

$$V\left(\rho_{m,i}(k)\right) = \nu_{free,m} \exp\left[-\frac{1}{a_m}\left(\frac{\rho_{m,i}(k)}{\nu_{crit,m}}\right)^a\right],$$

where $\nu_{free,m}$ is the average speed that drivers assume if traffic is flowing freely, and the critical density $\rho_{crit,m}$ is the density at which the traffic flow becomes unstable.

Origins are modeled with a simple queue model:

$$w_o(k + 1) = w_o(k) + T\left(d_o(k) - q_o(k)\right),$$

Fig. 1. In the METANET model, a freeway link is divided into segments.
with $w_o$ the length of the queue at origin $o$, $d_o(k)$ the demand, and $q_o(k)$ the outflow, which is given by

$$q_o(k) = \min \left[ d_o(k) + \frac{w_o(k)}{T} Q_o, \frac{Q_o \rho_{\text{max}} - \rho_{\text{ctrl}, \mu}} {\rho_{\text{max}} - \rho_{\text{crit}, \mu}} \right], \quad (5)$$

where $Q_o$ is the on-ramp capacity (veh/h) under free-flow conditions, $\rho_{\text{max}}$ is the maximum density, and $\mu$ the index of the link to which the on-ramp is connected.

2) Extensions: Since the original METANET model does not describe the effect of speed limits, we will slightly modify the equation for the desired speed (4) to incorporate speed limits. The second extension regards the modeling of a mainstream origin, which has a different nature than an on-ramp origin. The third extension describes the different effects of a positive or negative downstream density gradient on the speed.

In some publications, the effect of the speed limit is expressed by scaling down the desired speed-density diagram [2], [6]. This changes the whole speed-density diagram, also for the states where the speed would otherwise be lower than the value of the speed limit. This means, e.g., that if the free flow speed is 120 km/h and the displayed speed limit is 100 km/h, then the speed and flow of the traffic are reduced even when the vehicles are traveling at 80 km/h. Furthermore, scaling down the desired speed also reduces the capacity, while there is no reason to assume that a speed limit above the critical speed (speeds where the flow has not reached capacity yet) would reduce the capacity of the road. These assumptions are rather unrealistic, and they exaggerate the effect of speed limits. However, to get a more realistic model for the effects of speed limits, we assume that the desired speed is the minimum of the following two quantities: the desired speed based on the experienced density, and the desired speed caused by the speed limit displayed on the variable message sign:

$$V(\rho_{m,i}(k)) = \min \left[ (1 + \alpha) v_{\text{ctrl}, m,i}(k), \right. \left. \frac{v_{\text{free}, m}}{\exp \left[ \frac{\rho_{\text{ctrl}, m}(k) \alpha}{\rho_{\text{ctrl}, m}} \right]} \right], \quad (6)$$

where $v_{\text{ctrl}, m,i}(k)$ is the speed limit imposed on segment $i$, link $m$ at time $k$, and $1 + \alpha$ expresses the non-compliance.

To express the different natures of a mainstream origin link $o$ and a regular on-ramp (the queue at a mainstream origin is in fact an abstraction of the sections upstream of the origin of the part of the freeway network that we are modeling), we use a modified version of (5) with another flow constraint, because the inflow of a segment (and thus the outflow of the mainstream origin) can be limited by an active speed limit or by the actual speed in the first segment (when either of them is lower than the speed at critical density). Hence, we assume that the maximal flow equals the flow that follows from the speed-flow relationship from (1) and (4) with the speed equal to the speed limit or the actual speed in the first segment, whichever is smaller. So if $o$ is the origin of link $\mu$, then we have

$$q_o(k) = \min \left[ d_o(k) + \frac{w_o(k)}{T} q_{\text{lim}, \mu,1}(k), \right. \left. \frac{Q_o \rho_{\text{max}} - \rho_{\text{ctrl}, \mu}} {\rho_{\text{max}} - \rho_{\text{crit}, \mu}} \right],$$

where $q_{\text{lim}, \mu,1}(k)$ is the maximal inflow determined by the limiting speed in the first segment of link $\mu$:

$$q_{\text{lim}, \mu,1}(k) =$$

\begin{align*}
\lambda_{\mu} v_{\text{lim}, \mu,1}(k) \rho_{\text{crit}, \mu} \left[ -a_{\mu} \ln \left( \frac{v_{\text{lim}, \mu,1}(k)}{v_{\text{free}, m}} \right) \right] \right] ^{1/\gamma} \\
&\text{if } v_{\text{lim}, \mu,1}(k) < V(\rho_{\text{crit}, \mu}) \\
&\lambda_{\mu} V(\rho_{\text{ctrl}, \mu}) \rho_{\text{cap}, \mu} \text{ if } v_{\text{lim}, \mu,1}(k) \geq V(\rho_{\text{crit}, \mu}),
\end{align*}

with $v_{\text{lim}, \mu,1}(k) = \min(v_{\text{ctrl}, \mu,1}(k), v_{\mu,1}(k))$ the speed that limits the flow, and $q_{\text{cap}, \mu} = \lambda_{\mu} V(\rho_{\text{ctrl}, \mu}) \rho_{\text{cap}, \mu}$ the capacity flow.

Since the effect of a higher downstream density is usually stronger than the effect of a lower downstream density, we distinguish between these two cases. The sensitivity of the speed to the downstream density is expressed by parameter $\eta$. In (3), $\eta$ is a global parameter and has the same value for all segments. However, here we take different values for $\eta_{m,i}(k)$, depending on whether the downstream density is higher or lower than the density in the actual segment:

$$\eta_{m,i}(k) = \begin{cases} \eta_{\text{high}} & \text{if } \rho_{m,i+1}(k) \geq \rho_{m,i}(k) \\ \eta_{\text{flow}} & \text{if } \rho_{m,i+1}(k) < \rho_{m,i}(k). \end{cases}$$

In addition, when there is no entering link (but a mainstream origin), we assume that the speed of the (virtual) entering link equals the speed of the first segment:

$$v_{m,0}(k) = v_{m,1}(k). \quad (7)$$

This is a good approximation of the speed behavior when there are enough (e.g., three or more) uncontrolled upstream segments.

C. Objective function

We consider the following objective function:

$$J(k) = T \sum_{l=k}^{k+N_o-1} \left\{ \sum_{(m,i) \in \text{Iall}} \rho_{m,i}(l) I_m \lambda_m + \sum_{o \in \text{Oall}} w_o(l) \right\} + \sum_{l=k}^{k+N_o-1} \sum_{(m,i) \in \text{speed}} \left( \frac{v_{\text{ctrl}, m,i}(l) - v_{\text{ctrl}, m,i}(l-1)}{v_{\text{free}, m}} \right)^2,$$

where $I_{\text{all}}$ and $O_{\text{all}}$ are the sets of indices of all pairs of segments and links and of all origins respectively, and $I_{\text{speed}}$ is the set of pairs of indices $(m,i)$ of the links and segments where speed control is applied. This objective function contains a term for the TTS, and a term that penalizes abrupt variations in the speed limit control signal, which is weighted by the nonnegative parameter $\alpha_{\text{speed}}$. 
D. Constraints

In general, for the safe operation of a speed control system, it is required that the maximum decrease in speed limits that a driver can encounter ($v_{\text{maxdiff}}$) is limited. There are three situations where a driver can encounter a different speed limit value: (1) when the speed limit changes in a given segment (and there are more speed limit signs on the same segment), (2) when a driver enters a new segment, (3) when the driver enters a new segment and the speed limit changes. The maximum speed difference constraints for the three situations are formulated as follows:

$$v_{\text{ctrl},m,i}(l-1) - v_{\text{ctrl},m,i}(l) \leq v_{\text{maxdiff}}$$

$$v_{\text{ctrl},m,i}(l) - v_{\text{ctrl},m,i}(l') \leq v_{\text{maxdiff}}$$

$$v_{\text{ctrl},m,i}(l-1) - v_{\text{ctrl},m,i}(l') \leq v_{\text{maxdiff}}$$

for all $(m,i) \in I_{\text{speed}}$ with $(m,i')$ the segment following $(m,i)$ and $(m,i') \in I_{\text{speed}}$ and for $l = k, \ldots, k + N_c - 1$. In addition, the speed limits are often subject to a minimum value $v_{\text{ctrlmin}}$:

$$v_{\text{ctrl},m,i}(l) \geq v_{\text{ctrlmin}}$$

for all $(m,i) \in I_{\text{speed}}$ and $l = k, \ldots, k + N_c - 1$. In practice, the variable speed limit signs display speed limits in increments of, e.g., 10 or 20 km/h. Therefore, the controller should produce discrete-valued control signals:

$$v_{\text{ctrl},m,i}(l) \in V_{m,i}$$

for all $(m,i) \in I_{\text{speed}}$ where $V_{m,i}$ is the set of possible discrete speed limit values in segment $i$ of link $m$.

IV. A BENCHMARK PROBLEM

A. Set-up

The benchmark set-up consists of one origin, one freeway link, and one destination, as in Figure 1 with $N_1 = 12$. The main stream origin $O_1$ has two lanes with a capacity of 2000 veh/h each. The freeway link $L_1$ follows with two lanes, and is 12 km long, consisting of twelve segments of 1 km each. Segments 1 to 5 and segment 12 are uncontrolled, while segments 6 to 11 are equipped with a variable message sign where speed limits can be set. We choose to include the five uncontrolled upstream segments to be sure that boundary condition of equation (7) does not play a dominant role. Link $L_1$ ends in destination $D_1$.

We use the same network parameters as in [13]: $T = 10$ s, $\tau = 18$ s, $\kappa = 40$ veh/lane/km, $\rho_{\text{max}} = 180$ veh/lane/km, $\rho_{\text{crit}} = 33.5$ veh/lane/km, $a_m = 1.867$ and $v_{\text{free}} = 102$ km/h. Furthermore, we take $\eta_{\text{high}} = 65$ km$^2$/h, $\eta_{\text{low}} = 30$ km$^2$/h, $\alpha = 0.05$ and $\delta_{\text{speed}} = 2$. For the variable speed limits we have assumed that they can change only every minute, and that they cannot be less than $v_{\text{ctrlmin}} = 50$ km/h. This is imposed as a hard constraint in the optimization problem. If there is a safety constraint, then $v_{\text{maxdiff}} = 10$ km/h. The input of the system is the traffic demand at the upstream end of the link and the (virtual) downstream density at the downstream end of the link. The traffic demand (inflow) has a constant value of 3900 veh/h, close to capacity (4000 veh/h). The downstream density equals the steady-state value of 28 veh/km, except for the pulse that represents the shock wave. The pulse was chosen large enough to cause a backpropagating wave in the segments, see Figures 2 and 3. It is assumed that the upstream demand and downstream density is known, or predicted by an external algorithm. A combination of traffic measurements outside the controlled area and historical data could be used for prediction.

For the above scenario the performance (TTS) of the continuous and discrete-valued controls with or without safety constraints are examined. In the discrete control case, the control values $v_{\text{ctrl},m,i}$ are in the set $\{50, 60, 70, 80, 90, 100, 110\}$. The solution of the continuous-valued speed control problem is calculated using sequential quadratic programming. The discrete-valued control signal is a rounded version of the continuous optimization result. Three different types of discretization are examined: The first (“round”) rounds the continuous control values to the nearest discrete value, the second (“ceil”) rounds them to the nearest discrete value that is higher than the continuous value, and the third (“floor”) to the nearest discrete value that is lower than the real value. This method of obtaining discrete control signals is heuristic but fast. It is also possible to use discrete optimization techniques such as tabu search, simulated annealing or genetic algorithms, but since for this set-up and input the discretization method results in a performance that is comparable to that of the continuous version, it is not necessary to do so.

The rolling horizon strategy is now implemented as follows. After the discretization, the first sample of the control signal is applied to the traffic system and then the optimization–discretization steps are repeated. Note that this way of rounding is not the same as first rounding the continuous signal and then applying the resulting signal for the whole prediction horizon at once, because in our approach the different traffic behavior caused by the discretization is already taken into account in the next iteration.

Fig. 2. The downstream density scenario considered in the experiments.
B. Results

The results of the simulations of the no-control and the control case with continuous speed limits without constraints is displayed in Figures 3. In the controlled case the shock wave disappears after approximately 2 hours, while in the no-control case, the shock wave travels through the whole link. The speed limits are active in segments 6 up to 10; the speed limit in segment 11 has higher values than the critical speed and is not limiting the flow (see Figure 4). The active speed limits start to limit the flow at \( t = 4 \) min and create a low-density wave traveling downstream (the small dip in Figure 3). This low-density wave meets the shock wave traveling upstream and reduces its density just enough to stop it. So, the tail of the shock wave has a fixed location while the head dissolves into free flow traffic as in the uncontrolled situation, which means that the shock wave eventually dissolves completely.

The speed limits persist until the shock wave (to be precise, the high-density region) is completely dissolved. The speed limits in Figure 4 start to increase after \( t = 17 \) min and return gradually to a high value that is not limiting the flow anymore.

The TTS was 1862.0 veh.hours in the no-control case and 1458.0 veh.hours in the controlled (continuous, unconstrained) case, which is an improvement of 21.7%.

The result of the several types of discretization is shown in Table I. The performance loss caused by the discretized speed limits is small in the “round” and “ceil” cases, but large for “floor”. The performance degradation in case of “floor” can be probably explained by the effect of the safety constraints\(^2\), but the detailed analysis is a topic for future research.

The results of including the safety constraints are comparable to those in Table I, and are not shown here. The performance improvement for \( N_p = 11, N_c = 8 \) in the constrained case is 21.4 %, compared to 21.7 % in the unconstrained case. Figure 4 shows the values of the optimal speed limits discrete (“ceil”) case with safety constraints and \( N_p = 11, N_c = 8 \).

\(^2\)After discretizing the speed limits, the constraints could force the speed limits in the next iteration step to be too low from optimality point of view.
TABLE I
THE RELATIVE IMPROVEMENT OF THE PERFORMANCE (TOTAL TIME SPENT) FOR SEVERAL COMBINATIONS OF $N_p$ AND $N_c$, AND FOR THE DISCRETE-VALEUED SPEED LIMITS: "ROUND", "CEIL", AND "FLOOR"; WITHOUT SAFETY CONSTRAINTS.

<table>
<thead>
<tr>
<th>Horizon</th>
<th>Relative improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N_p$</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS AND FUTURE RESEARCH

We have presented a model predictive control framework to optimally coordinate variable speed limits. The purpose of the control was to find the control signals that minimize the total time that vehicles spend in the network.

We have applied the developed control framework to a benchmark network consisting of a link of 12 km, where 6 segments of 1 km are controlled by speed limits. It was shown that coordinated control with continuous-valued speed limits (base case) is effective against shock waves. The performance of the discrete-valued safety-constrained speed limits was comparable that of the base case if the discrete-valued speed limits were generated by “round” or “ceil”. In all of these cases the coordination of speed limits eliminated the shock wave entering from the downstream end of the link. The coordinated case resulted in a network where the outflow was sooner restored to capacity, and in a decrease of the total time spent of 21 %.

Topics for further research include: explanation of the performance degradation in case of “floor” discretization; comparison of the discrete MPC approach with other existing approaches; further examination of the trade-off between efficiency and optimality for rounding versus full discrete optimization; study of a real freeway stretch, including model calibration with real data.; and inclusion of extra control measures (such as ramp metering, dynamic lane assignment, route information, reversible lanes, etc.).

ACKNOWLEDGMENTS

Research supported by the Traffic Research Centre (AVV) of the Dutch Ministry of Transport, Public Works and Water Management, and by the NWO-CONNEKT project AMICI (014-34-523).

VI. REFERENCES


