Optimal coordination of ramp metering and variable speed control – An MPC approach

A. Hegyi, B. De Schutter, H. Hellendoorn, and T. van den Boom

If you want to cite this report, please use the following reference instead:
Abstract
We present a model predictive control (MPC) approach to optimally coordinate variable speed limits and ramp metering for highway traffic. The basic idea is that speed limits can increase the range in which ramp metering is useful. The control objective is to minimize the total time that vehicles spend in the network. For the prediction of the evolution of the traffic flows in the network we use an adapted version of the METANET model that takes the variable speed limits into account. The coordinated control results in a network with less congestion, a higher outflow, and a lower total time spent. In addition, the receding horizon approach of MPC results in an adaptive, on-line control strategy that automatically takes changes in the system parameters into account.

1 Introduction
As the number of vehicles and the need for transportation grow, cities around the world face considerable traffic congestion problems: almost every weekday morning and evening during rush hours the saturation point of the highways and the main roads in and around the city is attained. 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, considering the effect of the measures on the network level has many advantages compared to local control. E.g., solving a local congestion may have as a consequence that the vehicles run faster into another downstream congestion because of the improved flow, whereas still the same number of vehicles have to pass the bottleneck (which has a fixed capacity). So, the average travel time in the network will still be the same. Another reason is that in a dense network the effect of a local control measure could also influence the traffic flows in more distant parts of the network: an improved (or delayed) flow could cause (or prevent) congestion somewhere else in the network. Furthermore, if dynamic origin-destination (OD) data is available, control on the network level can take the predicted flows in the network into account. Local controllers are not able to use OD information because the traffic flow arriving at the local controller depends on the actions of other controllers in the network, which are unknown. E.g., during peak hours the density on the highway can be so high that the queue on an on-ramp spills back to the surface streets, whereas pro-active, coordinated metering of upstream on-ramps could reduce the density of the mainstream flow and prevent the on-ramp queue from spilling back. Another source of degradation of network performance is that congestion might block traffic directions that have nothing to do with the congestion (e.g., in case of congested off-ramps).

To address the problems sketched above a control strategy that operates on the network level is needed, i.e., there should be a network-wide coordination of control measures, based on global data. 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 needed to achieve optimal network control. To predict the effects of a control measure several techniques can be used, such as case-based reasoning, rule-based systems, or model-based prediction. In this paper we opt for the latter approach. More specifically, we use the METANET traffic simulation model [2, 8] for the predictions, and we apply a model predictive control framework [1, 4] to find the optimal control inputs. In practice, the prediction model used by the controller is always different from the real system, and the disturbances are only partially known. Model predictive control is known to perform well when this occurs.

The main control objective is to minimize the total time spent (TTS) by the vehicles in the network, but we will add an extra term to the objective function to penalize abrupt changes in the control signal. Papageorgiou [6] showed that, under the condition that the network inflow is known or can be predicted accurately, minimizing TTS is equivalent to maximizing the time-weighted outflow of the network. That means that a controller with this objective function (minimize TTS) will tend to maximize the outflow as soon as possible.

Ramp metering is used for two different purposes. First,
The fundamental diagram with an illustration that speed limits increase the range where ramp metering is useful.

In this paper we use a combined approach in which ramp metering and dynamic speed limits are coordinated to increase the range in which ramp metering is useful.

2 Problem description

The fundamental diagram is one of the basic tools for understanding the behavior of traffic systems: it relates the traffic flow and the traffic density on a highway (a typical behavior is shown in Figure 1). Ramp metering is only useful when traffic is not too light (otherwise ramp metering is not needed) and not too dense (otherwise breakdown will happen anyway). The corresponding region is on the stable (no breakdown) side of the fundamental diagram, and close to the top (see Figure 1), because that is where a breakdown can happen.

The main idea of this paper is that the combination of ramp metering with variable speed limits increases the range where ramp metering is useful. This prevents or postpones a traffic breakdown when traffic is getting dense. It is important to note that the congestion after a breakdown has usually a flow that is 5-10% lower than the available capacity. Papageorgiou et al. [7] have shown that a decrease of outflow of 5% can result in a TTS increase of 20%. This effect can be explained by the fact that the number of vehicles in the network is equal to the accumulated net inflow of the network (where the net inflow is the difference between the inflow and the outflow). But the outflow is lower when there is congestion (capacity-drop phenomenon), so the queue grows faster, and consequently congestion will last longer, and the outflow will be low for a longer time. This is why one should try to prevent or postpone a breakdown as much as possible.

In addition to ramp metering for the on-ramps, we also consider variable speed limits that are imposed on the mainstream traffic. The speed limits change the shape of the fundamental diagram (see Figure 1), which increases the range where ramp metering is useful. Indeed, suppose that the flow on the mainstream road is close to capacity and that the density is close to the critical density (state 1 in Figure 1). In this state even a small flow from the on-ramp can cause a breakdown. Now assume that we apply speed limits of 70 km/h to the sections of the main road upstream of the on-ramp (see Figure 2). The density in these sections will remain approximately the same, and the drivers will experience relatively large headway distances (approximately the inverse of the density), because their headway distance was chosen to match a higher speed. So, the state in Figure 1 changes from 1 to 2, and the shape of the fundamental diagram changes from the solid gray line to the dashed black line. This increases the stable region and creates some space for the on-ramp traffic. The additional vehicles from the on-ramp change the state from 2 into the direction of 3.

A drawback is that the flow will also decrease by the same factor as the speed. But if the control is optimized properly, this flow drop will always be less than or equal to the flow drop of a breakdown, since otherwise breakdown would be the optimal situation. Another point of criticism could be that the approach would keep the controlled network free of congestion, but at the cost of creating congestion at the entrances of the network. This is only partially true, because the controller will indeed sometimes delay the traffic to prevent a breakdown in the network, but afterward the flow will be higher than if the breakdown would have occurred. So the inflow of the network will be decreased by the speed limits for a short period of time only. Unfortunately, this could still cause congestion on upstream sections, but if this would happen we can use a larger region to evaluate the performance or to control, so that effects outside the original region are also taken into account and/or so that the effects of the control reach beyond the bounds of the original controlled region.
3 Approach

3.1 Model Predictive Control

We use a model predictive control (MPC) scheme to solve the problem of optimal coordination of speed limits and ramp metering. We assume that the reader is familiar with the basic ingredients of the MPC approach: at each sample step \( k \) the optimal control signal is computed over a prediction horizon \( N_p \), a control horizon \( N_c \) is selected to reduce the number of variables\(^1\) and to improve the stability of the system, and a rolling horizon\(^2\) strategy is used. For more information see [1, 4] and the references therein.

3.2 Prediction model

The MPC procedure includes a prediction of the network evolution as a function of the current state and a given control signal. We describe only those parts of the model that are relevant for interpreting and understanding the simulation results of our benchmark network (see Section 4). The METANET model represents a network as a directed graph with the links corresponding to highway stretches. Each highway stretch has uniform characteristics, i.e. no on-ramps or offramps and no major changes in geometry. Where major changes occur in the characteristics of the link or in the road geometry (on/off-ramp), a node is placed. Each link \( m \) is divided into \( N_m \) segments of length \( L_m \) (see Figure 3). Each segment \( i \) of link \( m \) is characterized by the traffic density \( \rho_{m,i}(k) \) (veh/lane/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. The outflow of each segment is equal to the density multiplied by the mean speed and the number of lanes on that segment (denoted by \( \lambda_m \)):

\[
q_{m,i}(k) = \rho_{m,i}(k) v_{m,i}(k) \lambda_m . \tag{1}
\]

The density of a segment equals the previous density plus the inflow from the upstream segment, minus the outflow of the segment itself (conservation of vehicles):

\[
\rho_{m,i}(k + 1) = \rho_{m,i}(k) + \frac{T}{L_m \lambda_m} (q_{m,i-1}(k) - q_{m,i}(k)) .
\]

\(^1\)After the control horizon has been passed the control signal is usually taken to be constant.

\(^2\)At each time step only the first sample of the optimal control signal is applied to the system; afterward the time axis is shifted one sample step, the model is updated, and the procedure is restarted.

The mean speed equals the previous mean speed plus a relaxation term that expresses that the drivers try to achieve a desired speed \( V(\rho) \), a convection term that expresses the speed increase (or decrease) caused by the inflow of vehicles, and an anticipation term that expresses the speed decrease (increase) as drivers experience a density increase (decrease) downstream:

\[
v_{m,i}(k + 1) = v_{m,i}(k) + \frac{T}{\tau_m} \left( V(\rho_{m,i}(k)) - v_{m,i}(k) \right) + \frac{T}{L_m} \left( v_{m,i}(k)(v_{m,i-1}(k) - v_{m,i}(k)) - \nu T \left( \rho_{m,i+1}(k) - \rho_{m,i}(k) \right) \right) . \tag{2}
\]

where \( \tau, \nu \) and \( \kappa \) are model parameters, and with

\[
V(\rho_{m,i}(k)) = v_{\text{free},m} \exp \left[ -\frac{1}{a_m} \left( \frac{\rho_{m,i}(k)}{\rho_{\text{crit},m}} \right)^{\nu_m} \right] . \tag{3}
\]

with \( a_m \) a model parameter, and where the free-flow speed \( v_{\text{free},m} \) is the average speed that drivers assume if traffic is flowing freely, and the critical density \( \rho_{\text{crit},m} \) is the density at which the traffic flow becomes unstable (cf. Figure 1).

On-ramps are modeled as origin links and a simple queue model is used to describe their dynamics. The length of the queue at the on-ramp equals the previous queue length plus the demand\(^3\) \( d_o(k) \), minus the outflow \( q_o(k) \):

\[
w_o(k + 1) = w_o(k) + T(d_o(k) - q_o(k)) .
\]

The outflow depends on the traffic conditions on the highway and the ramp metering rate\(^4\) \( r_o(k) \in [r_{\text{min}}, 1] \).

The flow \( q_o(k) \) is the minimum of the demand plus the queue at the on-ramp, and the maximal flow that can enter the highway because of the mainstream conditions and the ramp metering:

\[
q_o(k) = \min \left[ d_o(k) + w_o(k) \frac{T}{\nu_T}, \quad Q_o \min \left( r_o(k), \frac{\rho_{\text{max}} - \rho_{\text{crit},1}(k)}{\rho_{\text{max}} - \rho_{\text{crit},\mu}} \right) \right] . \tag{4}
\]

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.

Since in order to evaluate the evolution equations for a segment we need an upstream speed and flow, and a downstream density, the nodes (that connect the links) in the network should provide the entering and leaving links (the last and first segments) with the appropriate values. In our case the speed of the last segment of the entering link is simply passed to first segment of the leaving link:

\[
v_{m,0}(k) = v_{\mu,N_p}(k) ,
\]

\(^3\)Just as in [2, 3, 9] we assume that the demand is independent of any control actions taken in the network. Otherwise, a larger network should be considered.

\(^4\)The metering rate \( r_o(k) \) determines the fraction of the flow that is allowed to enter the highway. So no ramp metering corresponds to \( r_o(k) = 1 \); \( r_{\text{min}} \geq 0 \) is the minimal allowed ramp metering rate.
where \( m \) is the leaving link, \( \mu \) the entering link, and \( N_\mu \) the index of the last segment of link \( \mu \).

Furthermore, the sum of the flows of the entering links equals the inflow of the leaving link:

\[
q_{m,0}(k) = \begin{cases} 
 q_{\mu,N_\mu}(k) & \text{if there is no on-ramp} \\
 q_{\mu,N_\mu}(k) + q_o(k) & \text{if there is an on-ramp},
\end{cases}
\]

where \( q_o(k) \) is the flow from the on-ramp, if any, connected to the node, and \( N_\mu \) is the index of the last segment of the link \( \mu \) entering the node. The downstream density of the last segment \( N_\mu + 1 \) of link \( \mu \) is the density of the first segment of the leaving link \( m \):

\[
\rho_{\mu,N_\mu+1}(k) = \rho_{m,1}(k) .
\]

Since the original METANET model does not describe the effect of speed limits, we have slightly modified the desired speed equation (3) to incorporate speed limits: we assume that the desired speed is always less than or equal to the speed limit \( v_{\text{ctrl}}(k) \) displayed on the variable message sign (VMS):

\[
V(\rho_{m,i}(k)) = \min \left( v_{\text{ctrl},m,i}(k), v_{\text{free},m} \exp \left( -\frac{1}{a_m} \left( \frac{\rho_{m,i}(k)}{\rho_{\text{crit},m}} \right)^{a_m} \right) \right).
\] (5)

In addition, to express the different nature of a mainstream origin link \( o \) compared to a regular on-ramp (the queue at a mainstream origin is in fact an abstraction of the sections upstream of the origin of part of highway network that we are modeling), we use a modified version of (4), without the ramp metering term but 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 on the first segment (when either of them is lower than the speed at critical density). We assume that the maximal flow equals the flow that follows from the speed-flow relationship from 3 and 1 with the speed equal to the speed limit or the actual speed on the first segment whichever is the smaller. So if \( o \) is the origin of link \( \mu \), then we have

\[
q_o(k) = \min \left[ d_o(k) + \frac{w_o(k)}{T}, q_{\text{lim},\mu,1}(k) \right],
\]

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

\[
q_{\text{lim},\mu,1}(k) = \begin{cases} 
 \lambda_{\mu} v_{\text{ctrl},\mu,1}(k) \rho_{\text{crit},\mu} & \text{if } v_{\text{ctrl},\mu,1}(k) < V(\rho_{\text{crit},\mu}) \\
 q_{\text{cap},\mu} & \text{if } v_{\text{ctrl},\mu,1}(k) \geq V(\rho_{\text{crit},\mu}),
\end{cases}
\]

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

### 3.3 Objective function

We consider the following objective function:

\[
J(k) = \sum_{l=0}^{k+N_p-1} \left\{ T \sum_{m,i,l} \rho_{m,i,l}(l) L_m \lambda_m + \sum_o w_o(l) + a_{\text{ramp}} \sum_{o \in O_{\text{ramp}}} \left( r_o(l) - r_o(l-1) \right)^2 + a_{\text{speed}} \sum_{(m,i) \in I_{\text{speed}}} \left( v_{\text{ctrl},m,i}(l) - v_{\text{free},m}(l) \right)^2 \right\},
\]

where \( O_{\text{ramp}} \) is the set of indices \( o \) of those on-ramps where ramp metering is present, 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 two terms for the TTS (one term for the mainstream flow and one term for the on-ramp queue), and two terms that penalize abrupt variations in the ramp metering and speed limit control signals respectively. These terms are weighted by the nonnegative weight parameters \( a_{\text{ramp}} \) and \( a_{\text{speed}} \).

### 3.4 Tuning of \( N_p \) and \( N_c \)

In conventional MPC heuristic tuning rules have been developed to select appropriate values for \( N_p \) and \( N_c \) [1, 4]. However, these rules cannot be straightforwardly applied to the traffic flow control framework presented above. If we take the prediction horizon \( N_p \) shorter than the typical travel time in the network, then the effect of the vehicles that are influenced by the current control measure and — as a consequence — have an effect on the network performance before they exit the network, will not be taken into account. Furthermore, a control action may affect the traffic state (by improved flows, etc.) even when the actually affected vehicles have already exited the network. On the other hand, \( N_p \) should not be too large because of the computational complexity of the MPC optimization problem. So based on this heuristic reasoning we select \( N_p \) to be about the typical travel time in the network. For the control horizon \( N_c \) we will select a value that represents a trade-off between the computational effort and the performance.

### 4 A benchmark problem

In order to illustrate the control framework presented above we will now apply it to a simple traffic network.

#### 4.1 Set-up

The network for the experiment (Figure 4) was chosen as simple as possible. It basically consists of a mainstream with speed limits, and a metered on-ramp. The choice for the second speed limit was made to have more control over the state (speed and density) in the segment that is just before the on-ramp. The benchmark network consists of two
origins, two highway links, and one destination. $O_1$ is the mainstream origin and has two lanes with a capacity of 2000 veh/h each. The highway link $L_1$ follows with two lanes, and is 2 km long consisting of two segments of 1 km each. Both segments are equipped with a VMS where speed limits can be set. At the end of $L_1$ a single-lane on-ramp ($O_2$) with a capacity of 2000 veh/h is attached. Link $L_2$ follows with two lanes and a length of 1 km, and ends in destination $D_1$ with unrestricted outflow.

We have used the same network parameters as in [2]: $T = 10$ s, $\tau = 18$ s, $\nu = 60$ km/h, $\kappa = 40$ veh/h, $\rho_{\text{max}} = 180$ veh/lane/km, $\rho_{\text{crit}} = 33.5$ veh/lane/km, $\alpha_1 = 1.867$ and $v_{\text{free}} = 102$ km/h. Furthermore, we have set $\alpha_{\text{ramp}} = \alpha_{\text{speed}} = 0.4$. To examine the effect of the combination of variable speed limits and ramp metering a typical demand scenario was considered (see Figure 5): The mainstream demand has a constant, relatively high level, while the demand on the on-ramp increases to near capacity, remains constant for half an hour, and decreases finally to a constant low value. Important quantities in assessing the coordinated control are the TTS (related to the outflow) and the duration of the congestion (a queue on either origins). For the above scenario these quantities will be compared for the ‘coordinated speed limits and ramp metering’ (CSLRM), and the ‘ramp metering only’ (RMO) case.

### 4.2 Results

The optimal prediction horizon was found to be approximately $N_p = 42$ (7 min.), which is in the order of the typical travel time through the network (4 km / 40 km/h). Shorter prediction horizons did not take the whole response of the system into account and resulted in insufficient control actions. Longer prediction horizons tended to take the future demand too much into account, which degraded the performance. When the difference $N_p - N_c$ was kept constant (2 min.), a further increase of $N_p$ caused only a small decrease of the TTS. For the control signals we have assumed that they can change only every minute, which is more realistic than every 10 seconds. A control horizon $N_c = 3$ (3 min.) was sufficient for the RMO case, for the CSLRM case $N_c = 5$ was necessary. Longer control horizons tended to give the control signal optimization too much freedom, which resulted in more variance in the signals.

The results of the two cases are displayed in Figures 6 and 7. In the plots we can see that RMO performs well until the maximum queue length is reached. After that point ramp metering is unable to prevent congestion and the performance breaks down in the RMO case. In the CSLRM case the speed limits become active to prevent breakdown, by which a higher outflow is reached. The queues at both origins never disappear in the RMO case; but in the CSLRM case they are resolved after approximately 2.5 hours. The TTS was 860 veh hours in the RMO case and 734 veh hours in the CSLRM case, which is an improvement of about 15%.

### 5 Conclusions and future research

We have applied model predictive control (MPC) to optimally coordinate variable speed limits and ramp metering. The combination of speed limits with ramp metering increased the range in which ramp metering is useful. This idea is illustrated by a simple benchmark network, where the cases ‘ramp metering only’ and ‘coordinated ramp metering and speed limits’ are compared for a given demand scenario. For both cases the control signal was optimized such that the total time spent (TTS) in the network is minimal. The coordinated case resulted in a network with a higher outflow and a lower TTS.

Topics for further research include: selecting other methods to model the effect of a speed limit; further investigation of the effectiveness of MPC for optimal coordination of speed limits and ramp metering for a wider range of scenarios, networks, traffic flow models and/or model parameters; explicit inclusion of modeling errors and disturbances; and — for networks in which not all on-ramps are metered — taking the modified routing behavior of the drivers into account. Furthermore, including extra control measures in addition to speed limits and ramp metering (such as peak-lanes, route information, reversible lanes, etc.) is also a topic for future research.

### Acknowledgments

Research supported by the Traffic Research Centre (AVV) of the Dutch Ministry of Transport, Public Works and Water Management.

---

Footnote: The effect of a speed limit was modeled in (5) as a change in the desired speed of the drivers, whereas using a stochastic distribution may be a better way of modeling the (partial) compliance to the speed limit.
Figure 6: The simulation results for the RMO case. When the density approaches the critical density (33.5 veh/km), the ramp metering gradually switches on, and keeps the flow high (around 4200 veh/h). After the queue length on the on-ramp has reached the maximum queue length (100 veh), the ramp metering is forced to let the complete demand enter the highway. This causes a congestion, and hence results in a reduced outflow. The queue at $O_1$ is never resolved.

Figure 7: The results for the CSLRM case. In the beginning the ramp metering signal is similar to that of the RMO case, but when the density becomes too high (which endangers the high outflow) the speed limits become active and reduce the inflow from the mainstream. This initially causes a longer queue on the mainstream, but the outflow is kept higher and both queues ($O_1$ and $O_2$) are resolved within approximately 2.5 hours.

References