Modeling and control of congestion phenomena

Levering, N.A.C.

Publication date
2024

Chapter 6

Input rate control with effective bandwidths

6.1 Introduction

The flow of traffic that road networks can carry is not only determined by physical characteristics of roads (width, curvature, inclination, etc.) and traffic rules (e.g., speed limits, prioritization, overtaking), but to a large extent by traffic itself. As we have seen, in the mathematical analysis of road traffic flow this is captured by the fundamental diagram: large traffic densities may cause traffic speeds to drop rapidly, leading to sudden capacity reductions on individual roads in the network that propagate in space and time, and deteriorate the performance of the entire network. Most road traffic control mechanisms, including the ones studied in the previous chapters, focus on flow management inside the network. Such approaches have proven to be successful in many settings, but are challenged by the worldwide increase in travel demand. In this chapter, we therefore consider the real-time control of the *input* traffic streams, aiming to prevent the above sketched scenario in which the factual capacities on individual network roads are exceeded. By applying such a control procedure, at the cost of some waiting time at the on-ramp boundaries of the network, we guarantee a very small probability of high delays within the network. Reducing traffic collapses in the network also limits additional negative consequences of heavy congestion, such as the aforementioned environmental pollution and economic costs. Similar to the control inside road networks (in terms of routing and departure time advice), the ability to apply control at the boundaries of a road network is facilitated by advances in Intelligent Transportation Systems. Specifically, input rate control may be applied through ramp metering and navigation systems.

When considering traffic flow management, the inherently random nature of vehicle traffic should be taken into account. Specifically, due to, e.g., the heterogeneity in vehicle sizes and individual driving habits, the fraction of the capacity of a road that is unavailable due to the presence of traffic flow on that road suffers from randomness. As the performance of the network, in terms of realized travel times, is significantly affected by catastrophic capacity violations, (deterministic) control policies that solely consider the capacity the *average* traffic flow needs will typically perform poorly. Thus, there is a need for fast control policies that do take the fluctuations in traffic flow, and specifically, random spikes in capacity needs, into account.

In telecommunication networks, similar considerations for the construction of control policies apply: transmission mediums have certain *bandwidths* (capacities in terms of traffic volume per time), and each connection using a medium requests part of this bandwidth.

A control policy may decide whether new connections are allowed, thereby taking into account that the bandwidth requirement of an individual connection may fluctuate during the time it uses a medium. Typical performance control focuses on avoiding that the total demand of simultaneous connections exceeds the available bandwidth. A very influential framework in the control of telecommunication networks is that of *effective bandwidth*, which describes the minimum bandwidth that needs to be reserved for individual connections to guarantee a certain level of service for these connections. The approach leads to linear acceptance regions, making the effective bandwidth framework a fast and easily implemented admission control procedure. In the context of telecommunications, this framework was extended to different arrival and service models; we mention the famous papers by Kelly [125], Gibbens and Hunt [84], and Elwalid and Mitra [62]. For networks with fluctuating bandwidth demands, the notion of effective bandwidths was discussed by Hui [108]. Our goal in this chapter is to explore whether the notion of effective bandwidths can be extended beyond the telecommunications context, specifically, to the context of road traffic networks.

Relevant literature

There is a broad range of work on the external control of traffic streams in networks, in which the control procedures are performed by a traffic planner. Traditional traffic models, such as the Vickrey bottleneck model [4, 211], the kinematic wave model of Lighthill, Whitham, and Richards [148, 185], and its discretized version (recall that this is the cell transmission model introduced by Daganzo [45, 46]), operate in a setting in which both demand and delays are deterministic. However, in the context of road traffic, uncertainty plays a major role. Therefore, there is considerable interest in the stochastic counterparts of these deterministic flow models, which do account for different driver perceptions, moods, car types, etc. Recent contributions include the stochastic traffic flow models of Jabari and Liu [110] and Mandjes and Storm [154] (whose applicability is studied in [200]), the stochastic fundamental diagram of Qu et al. [183] (the model studied in the previous chapter), and the stochastic bottleneck model of Ghanzanfari et al. [82].

For the routing of individual vehicles, taking uncertainty into account amounts to solving a stochastic shortest path problem, in which travel times on arcs are (time-dependent) random variables. Algorithms that minimize the expected travel time or maximize the on-time arrival probability under various conditions have been presented in Chapters 2 and 3. These are stochastic analogues to (a sped-up version of) Dijkstra's algorithm, which yields the optimal route for a vehicle in a deterministic network. For the optimal routing of traffic *streams* in deterministic networks, the seminal work of Wardrop [217] introduces a user equilibrium (i.e., no driver has the incentive to switch routes) and a social equilibrium (i.e., minimizing the total network travel costs). Examples of stochastic counterparts of the Wardrop model, in which the delay is not simply a deterministic function of the traffic flow, are found in, e.g., [3, 44, 168]. Typically, these studies consider the stochastic user equilibrium, the stochastic social optimum, or the best or worst ratio between these as a function of the risk-adverseness of the vehicles.

Whereas the above works do consider uncertainty, their focus lies on the routing of traffic streams. However, in case of high demand, even with an optimal routing scheme, unlimited access to a highway network can lead to capacity violations, which has triggered various

studies about input rate control strategies. Papageorgiou and Kotsialos [171] and Shabaan et al. [193] present overviews of ramp metering strategies, but the referenced works typically consider control in a deterministic setting. In the thesis of Kovács [138, Chapter 7], uncertainty in the arrival stream *is* taken into account, but the framework is limited to a single one-directed road that consists of multiple segments, and the objective of study is a proportionally fair control scheme. A similar objective is studied by Kelly and Williams [126], who do consider uncertainty in the arrival streams for a full network of roads, and who analyze the performance of a Brownian network model – often used for proportionally fair control in telecommunication models – as approximate model for the controlled motorway. Thus, to our knowledge, there are no works that consider ramp metering strategies in road networks that (i) take the stochasticity in the network into account, and (ii) have as objective to limit the number of capacity exceedances in the network.

Contributions

The contributions of this chapter are twofold. In the first place, we present a stochastic model that describes the part of the road capacity that is effectively taken by the traffic input streams on that road. This overcomes the limitations of deterministic models, which do not account for different perceptions, responses, driving habits, car types, etc. Specifically, we model the capacity needs as a compound Poisson process. This model offers great flexibility, as we impose only few assumptions on the jump distribution of this process.

In the second place, we show how the concept of effective bandwidths can be used to construct a fast control policy in the road traffic context. This policy allows waiting at the boundaries of the network, so as to prevent capacity violations within the network. We show that the asymptotic regime suggests a similar notion of effective bandwidths, investigate the properties of these effective bandwidths, and test their optimality through numerical examples. In particular, these experiments show that, by applying an effective bandwidth policy, capacity violations are rare, and that, as a result, the total time it takes the vehicles in the controlled road network to reach their chosen destinations (i.e., their total waiting and driving time) is significantly smaller than for three other control algorithms that serve as benchmarks.

Organization

Section 6.2 presents the utilized capacity model, and introduces the resulting problem of input flow rate control. The application of effective bandwidths for flow control in the context of road traffic is described in Section 6.3. Numerical examples that compare the performance of our method to three benchmark algorithms are given in Section 6.4. We end this chapter with some concluding remarks.

6.2 Preliminaries

In a road traffic network, exceeding the capacity of a road may lead to extreme delays for (a part of) the road users. To keep a handle on congestion, we consider the *control* of the input traffic streams of the network. We are, however, challenged by the fact that, by the heterogeneity within and between traffic input streams, the impact of traffic flows on the available capacity suffers from randomness. We first present a stochastic road traffic model which describes the capacity that is effectively used by the input traffic flows (Section 6.2.1). Then, in this model setting, Section 6.2.2 introduces the problem of controlling the input flow rates of the network, with the aim to avert delays due to exceedingly high traffic loads.

6.2.1 Utilized capacity model

We consider a road network and corresponding graph representation $G = (N, A)$, of which the set of nodes N represents the junctions in the road network and the set of directed arcs $A = \{a_1, \ldots, a_J\}$ represents the roads connecting these junctions. Each directed arc $a \in A$ has a capacity C_a, indicating how much flow can be carried by the arc (i.e., number of vehicles that can enter the arc per time unit). A path from node $n_1 \in N$ to node $n_2 \in N$ is a collection of connected directed arcs that starts at n_1 and ends at n_2, and we denote with $\mathcal{P} = \{P_1, \ldots, P_I\}$ the set of all paths in the network. These paths are described by the route-link incidence matrix B, whose elements indicate which links are part of which paths, i.e., for $i = 1, \ldots, I$ and $j = 1, \ldots, J$,

$$B_{ij} = \begin{cases} 1 & a_j \in P_i, \\ 0 & a_j \notin P_i. \end{cases}$$

Many of the traditional traffic flow models are deterministic. That is, with r_{P_i} the (mean) input flow rate for a path $P_i \in \mathcal{P}$ (i.e., the average number of vehicles per minute traversing P_i), it is commonly assumed that for each $j = 1, \ldots, J$,

$$r_{a_j} = \sum_{i=1}^{I} B_{ij} r_{P_i},$$

and that the capacity on arc a_j is exceeded if $r_{a_j} > C_{a_j}$. However, it is widely recognized that, due to the variation in individual driver behavior and heterogeneity in vehicle sizes, the capacity needed by the flow on path P_i is a stochastic quantity. To capture this uncertainty in capacity needs, we introduce the random variable Y_{a_j}, denoting the so-called *utilized capacity* of arc a_j, i.e., the flow produced by traffic traversing arc a_j. Then, we say that the capacity on arc a_j is exceeded if the total utilized capacity is higher than C_{a_j}.

For a description of the randomness of the utilized capacity, we model Y_{a_j} as a sum of compound Poisson processes. Specifically, we set

$$Y_{a_j} \overset{\mathrm{d}}{=} \sum_{i=1}^{I} B_{ij} \sum_{k=1}^{M_i} D_{ik}, \tag{6.1}$$

with $M_i \sim \text{Poisson}(r_{P_i})$, $r_{P_i} \in \mathbb{R}_{>0}$, and $D_{i1}, \ldots, D_{iM_i}$ i.i.d. non-negative random variables with known distribution. Thus, every path P_i that uses arc a_j generates a Poisson number M_i of vehicles on that arc, with mean r_{P_i}, the average number of vehicles per minute traversing P_i. Describing the random vehicular arrivals with a Poisson process is a natural and widely-used modeling assumption. The applicability of the Poisson distribution stems from the fact that, without congestion or signalized intersections, drivers behave relatively independent. In lightly congested traffic conditions, which form the focus of this chapter, empirical observations have indeed shown that Poisson distributed traffic volumes are realistic.

The amount that one vehicle traversing path P_i contributes to the occupied capacity on arc $a_j \in P_i$ is modeled by $D_{i1} \overset{\text{d}}{=} D_i$. Note that, with limited assumptions on its distribution, this random variable offers great modeling flexibility, and can, e.g., be used to capture the heterogeneity in vehicle sizes. Indeed, modeling the impact of passenger cars and trucks by D_i^{cars} and D_i^{trucks} respectively, the impact of the total traffic mix is well described for D_i a mixture distribution of D_i^{cars} and D_i^{trucks}, the weights set as estimates of their traffic mix proportions. Also note the resemblance with a *passenger car equivalent* factor, which describes the impact that a mode of transport has on a traffic variable (in this case, the flow) compared to a single passenger car [2, 194]. Now, as the different paths may contain different traffic mixes, the occupied capacity is modeled path-dependent. That is, the distribution of D_i may differ from the distribution of D_j for $P_i \neq P_j$.

6.2.2 Avoiding network overflow

We consider a network $G = (N, A)$, with input stream rates $(r_{P_1}, \ldots, r_{P_I})$, whose impact on the utilized flow is given through (6.1). It is assumed that, for these given flow demands, traffic in the network is light, i.e., $\mathbb{P}(Y_a > C_a)$ is sufficiently small for all $a \in A$. Thus, vehicles can travel relatively freely through the network, experiencing only little hindrance from other road users. Now, the goal is to decide on the admissibility of additional traffic flow in a fast and accurate way, such that, with increased loads, there is an acceptable balance between the utilized capacity of an arc and the probability the arc capacity is exceeded, causing congestion. To this end, we manage potential increases in the input rates, so as to limit the number of arc capacity violations.

To avoid reaching the critical capacity, we assume we have control of the input stream rates at the boundaries of the network, i.e., at the starting nodes of the paths. Specifically, for each $P_i \in \mathcal{P}$, we are able to decide if, instead of r_{P_i}, a higher input rate would still be such that capacity violations are rare, i.e., that $\mathbb{P}(Y_a > C_a)$ is sufficiently small for all $a \in A$. This yields, for all $P_i \in \mathcal{P}$, a rule which prescribes whether additional demand can indeed be handled by the network. If not, the input stream rate of this path should not be increased, and additional traffic should queue at the boundary of the network. Our goal is that by controlling traffic in this manner, at the cost of some delay on the boundaries of the network, extreme congestion within the network, leading to high delays for part of the traffic, is prevented.

Now, to construct rules to handle additional demand, we propose the use of effective bandwidths. Effective bandwidths are traditionally applied in the management of communication networks, in which new connections claim part of the available bandwidth, and

bandwidth violations are very undesirable. In these networks, effective bandwidths form a powerful tool for admission control: they efficiently determine a half-space that serves as acceptance region, such that new connections are only accepted if they fall into this region. They are defined arc-wise, as, if congestion is avoided, the arcs in communication networks are relatively independent in terms of throughput. Note that, in vehicle traffic networks, avoiding capacity violations on arcs limits the negative interaction between the different arcs, as there are no traffic jams that affect multiple arcs. Therefore, there is a promise in expanding the use of effective bandwidths to the vehicle network setting, which we explore in this chapter. The application of effective bandwidths in the context of road traffic will be explained in detail in the next section.

Before doing so, it is important to remark that applying access control in vehicular networks is a dynamic procedure. That is, in case there is additional traffic demand on $P_i \in \mathcal{P}$ that is allowed into the network, this yields a new input stream rate $\tilde{r}_{P_i}$. With this new average flow rate, the access rule needs to be updated. Concretely, for $\varepsilon > 0$ small, it should now be decided if an input rate of $\tilde{r}_{P_i}(1+\varepsilon)$ can still be handled by the network. Specifically, for the use of effective bandwidths, this means that, to account for the dynamic updates in traffic streams, the acceptance region should be updated regularly.

6.3 Effective bandwidths in road traffic

A concise overview of the concept of effective bandwidths in their traditional telecommunications context is provided in Section 6.3.1. The framework uses a linear acceptance region, within which the probability of exceeding the network capacity is small, such that the resulting control procedure is simple and fast. Introducing the background of effective bandwidths, the subsection paves the way for Section 6.3.2, in which the notion of effective bandwidths is expanded to the road traffic setting.

6.3.1 Effective bandwidths in telecommunication

In a telecommunication network, different connections are multiplexed over a shared medium. The medium has a total bandwidth, and the individual connections using the medium request part of this bandwidth. However, typically, the connections are bursty, in the sense that the bandwidth requirement may fluctuate during the holding period of the connection. When applying admission control to such a system, i.e., when deciding whether a new connection is accepted, these fluctuations should be taken into account, as exceeding the bandwidth may lead to connection losses or other service level violations. Given that each connection has a mean and a peak rate, an extreme policy would be to accept a new connection if the sum of all peak rates does not exceed the bandwidth, whilst another extreme policy would be to accept if the sum of all mean rates is smaller than the bandwidth. In the first case, there are no service level violations, but a substantial part of the bandwidth may be wasted, given that a connection does not continuously require its peak rate. In the second case, the converse is true: there is little excess in bandwidth use, but there are scenarios in which service levels are violated. The concept of effective bandwidths provides a strategy between these two extremes.

Denote C' as the bandwidth shared by I' types of connections, with $m_i \in \mathbb{N}$ connections of type $i = 1, \ldots, I'$. Let D'_{ij} be the bandwidth requirement for the j-th connection of type i, $D'_{i1}, \ldots, D'_{im_i}$ identically and independently distributed, such that

$$Y' = \sum_{i=1}^{I'} \sum_{j=1}^{m_i} D'_{ij}$$

is the total bandwidth demanded from the medium. Observe the similarity between this bandwidth demand expression and the road traffic capacity demand expression (6.1). Now, parallel to the introduced road traffic setting, the aim in classical effective bandwidth literature is to limit the occurrences in which the capacity is exceeded. For $\gamma > 0$, the admission control policy should guarantee $\mathbb{P}(Y' > C') \leq e^{-\gamma}$. Using a Chernoff bound, it has been proven that

$$\inf_{s \geq 0} \left[\sum_{i=1}^{I'} m_i \mathbb{E}[\exp(sD'_{ij})] - sC' \right] \leq -\gamma \implies \mathbb{P}(Y' > C') \leq e^{-\gamma}, \tag{6.2}$$

such that the probability bound is satisfied when the policy is to only accept a new connection if the new vector of connections falls within the *acceptance region* R':

$$R' = \bigcup_{s \geq 0} R'_s, \qquad R'_s = \left\{ \boldsymbol{m} \in \mathbb{R}^{I'}_{>0} : \sum_{i=1}^{I'} m_i \alpha'_i(s) \leq C' - \gamma/s \right\},$$

with

$$\alpha'_i(s) = \frac{1}{s} \log \mathbb{E}[\exp(sD'_{ij})] \quad , \quad s \geq 0.$$

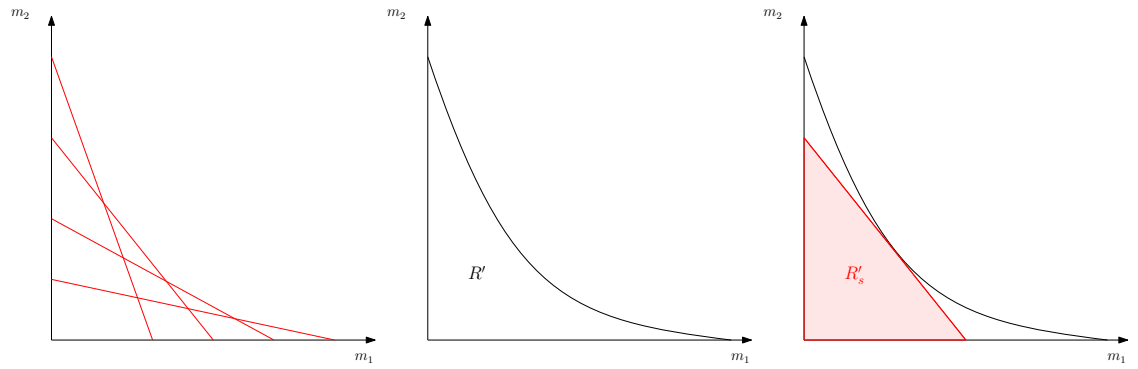
Remark 6.3.1. The condition that the new vector should fall within the region R' is one-way, in that satisfying the probability guarantee does not directly imply that a vector of connections is within the acceptance region. However, an application of Cramér's theorem shows that, for large values of γ, C' and $m_1, \ldots, m_{I'}$, the relation is two-sided. Specifically, as observed in [127], Cramér's theorem implies

$$\lim_{N \to \infty} \frac{1}{M} \log \mathbb{P} \left(\sum_{i=1}^{I'} \sum_{j=1}^{m_i M} D'_{ij} > C'M \right) = \inf_{s \geq 0} \left[s \sum_{i=1}^{I'} m_i \alpha'_i(s) - sC' \right].$$

◇

Example 6.3.2. Observe that the acceptance region R' is a family of half-spaces in $\mathbb{R}^{I'}$, indexed by s. A typical example of the form of R' is presented in Figure 6.1. Figure 6.1a shows half-spaces R'_s constructing R' for some values of s. The resulting acceptance region, being the union of these half-spaces, is displayed in Figure 6.1b. ◇

Unfortunately, the acceptance region R' is often too difficult to work with, as, for a given I'-dimensional vector of connections $\boldsymbol{m}$, it is typically hard to solve the inversion problem, namely, to decide whether there *exists* $s \geq 0$ such that $\boldsymbol{m} \in R'_s$. Therefore, the idea is to approximate the acceptance region R' with a region of simpler size. Specifically, for a chosen $s > 0$, the acceptance region is approximated by R'_s, whose right boundary is linear



(a) (b) (c)

Figure 6.1: Acceptance region and approximation.

(Figure 6.1c). Then, there is a simple and fast admission control procedure: only accept a new connection if the new vector of connections $\boldsymbol{m}$ satisfies

$$\sum_{i=1}^{I'} m_i \alpha'_i(s) \leq C' - \gamma/s.$$

The weight $\alpha'_i(s)$ of connection type i is called the *effective bandwidth* of type i, and has a value between the mean and peak rate of the connection type. Since $R'_s \subseteq R'$ for all $s \geq 0$, any non-negative value of s that is chosen as input for $\alpha'_i(s)$ still guarantees the specified probability bound. As argued in [127], a natural choice is to consider a vector $\boldsymbol{m}$ on the boundary of R' that represents a typical traffic mix, and let s be such that the infimum in (6.2) is attained.

6.3.2 Effective bandwidths in road traffic

We show how to apply the ideas of the previous subsection to the road traffic context introduced in Section 6.2. Similar to the telecommunications setting, the objective is to apply input control to a network, so as to bound the probability of exceeding network capacities. The effective bandwidth framework, working with a linear acceptance region, provides a fast and easily implemented procedure for network admission control. The effective bandwidths capture both the mean, variance and other distributional properties of the capacity requirements of the different traffic streams on an arc, and are computed arc-wise, such that additional traffic on a path is simply accepted once it satisfies the control constraints of the arcs making up the path.

First, with similar techniques as in the traditional effective bandwidth literature, we use a Chernoff bound to bound the probability that the capacity of an arc $a_j \in A$ is exceeded from above:

$$\begin{aligned} \log \mathbb{P}(Y_{a_j} > C_{a_j}) &\leq \inf_{s \geq 0} \log \left(\mathbb{E}\left[e^{sY_{a_j}} \right] e^{-sC_{a_j}} \right) \\ &= \inf_{s \geq 0} \left(\sum_{\substack{i=1,\ldots,I \\ B_{ij}=1}} \log \mathbb{E}\left[\exp\left\{ s \sum_{k=1}^{M_i} D_{ik} \right\} \right] - sC_{a_j} \right). \end{aligned} \tag{6.3}$$

Observing that the expectation in the considered upper bound is the moment-generating function (MGF) of a compound Poisson distribution, we have, for $\gamma > 0$,

$$\inf_{s\geq 0}\left(\sum_{\substack{i=1,\dots,I\\ B_{ij}=1}} r_{P_i}(\mathbb{E}[e^{sD_{i1}}]-1)-sC_{a_j}\right)\leq -\gamma \implies \mathbb{P}(Y_{a_j}>C_{a_j})\leq e^{-\gamma}. \tag{6.4}$$

Now, the probability of exceeding the capacity of arc a_j, and consequently, the probability of delay around this arc, is small if $\boldsymbol{r}$ lies in the acceptance region R_j^α:

$$R_j^\alpha = \bigcup_{s\geq 0} R_{j,s}^\alpha, \qquad R_{j,s}^\alpha = \left\{\boldsymbol{r}\in\mathbb{R}^I_{\geq 0} : \sum_{i=1,\dots,I} B_{ij} r_{P_i}\alpha_i(s)\leq C_{a_j}-\frac{\gamma}{s}\right\},$$

with

$$\alpha_i(s) = \frac{1}{s}(\mathbb{E}[e^{sD_{i1}}]-1) \quad , \quad s\geq 0.$$

Remark 6.3.3. Note that $\alpha_i(s)$ is indeed equivalent to the notion of effective bandwidth as introduced in [127]. We observe that,

$$\alpha_i(s) = \mathbb{E}[e^{sD_{i1}}-1]/s \geq \mathbb{E}[sD_{i1}]/s = \mathbb{E}[D_{i1}],$$

such that the bandwidth is between the mean and peak rate of Y_{a_j}:

$$\mathbb{E}[Y_{a_j}] = \sum_{i=1,\dots,I} B_{ij} r_{P_i}\mathbb{E}[D_{i1}] \leq \sum_{i=1,\dots,I} B_{ij} r_{P_i}\alpha_i(s) < \infty = \sup\{y : \mathbb{P}(Y_{a_j}>y)>0\}.$$

Also, comparing the two different notions of effective bandwidths, it can easily be seen that $\alpha_i(s)\geq \alpha_i'(s)$. This is not surprising, as, instead of only protecting against potential high values of D_{ik}, we need to protect against high values of M_i as well. ◇

Remark 6.3.4. It is easy to see that (6.3) is invariant to the distribution of M_i, in the sense that any distribution over the integers would yield a similar expression for the upper bound. Notably, with $G_X(s) = \mathbb{E}\exp\{sX\}$ the MGF of a random variable X, the implication in (6.4) may be replaced by the more general

$$\inf_{s\geq 0}\Big(\sum_{\substack{i=1,\dots,I\\ B_{ij}=1}} \log G_{M_i}\big(\log G_{D_{i1}}(s)\big)-sC_{a_j}\Big)\leq -\gamma \implies \mathbb{P}(Y_{a_j}>C_{a_j})\leq e^{-\gamma},$$

such that the results are easily extended in case M_i is another distribution from the family of discrete distributions over the integers for which the MGF is well-defined. ◇

Similar to the telecommunications setting, the acceptance region R_j^α is too complex for practical application, i.e., for a given I-dimensional vector of stream rates $\boldsymbol{r}$ it is hard to decide if there exists an $s_j\geq 0$ such that $\boldsymbol{r}\in R_{j,s_j}^\alpha$. Therefore, we approximate the acceptance region R_j^α by one of the regions R_{j,s_j}^α, whose right boundary is again linear. Then, given the regions R_{j,s_j}^α and $\varepsilon>0$, the control procedure is simply to allow an average

input rate of $r_{P_k}(1+\varepsilon)$ instead of r_{P_k} if, for all $j \in \{1,\dots,J\}$, the following inequality is satisfied:

$$\sum_{i=1,\dots,I} B_{ij} r_{P_i} \alpha_i(s_j) + \varepsilon B_{ij} \alpha_k(s_j) \leq C_{a_j} - \gamma/s_j. \tag{6.5}$$

For each $j = 1,\dots,J$, we propose to base the choice of s_j (i.e., R^{α}_{j,s_j}) on the current traffic conditions: given $\boldsymbol{r}$, we let R^{α}_{j,s_j} be such that s_j attains the infimum in (6.4). With s_j given, the effective bandwidths $\alpha_j(s_j)$ are computed, which may then be stored externally, such that it can be decided rapidly if (6.5) is satisfied. Note that, as argued in Section 6.2.2, the application of access control in road networks is a dynamic procedure. On a longer timescale, one should therefore adapt the approximating region to new traffic conditions.

Remark 6.3.5. In the above, γ has the same value for all arcs. We can, however, also work with a bound $\gamma_{a_j} > 0$ per arc $a_j \in A$. This may be preferable if there are arcs in the network that are, e.g., located centrally or have a high degree of neighboring links, such that a congestive setting on these arcs has a more deteriorating effect on the network than congestive settings on other arcs. ◇

Since (6.4) is not an equivalence statement, the described control method is conservative. This is preferred over non-conservative methods, which may exploit resources better, but can no longer guarantee that capacity violations are rare. Moreover, in the classical effective bandwidth notion, the acceptance region R' is conservative as well, but, in the asymptotic regime, $\mathbb{P}(Y' > C') \leq e^{-\gamma}$ is approximately the same as $\boldsymbol{m} \in R'$ (Remark 6.3.1). Notably, such a limit argument carries on to the road traffic setting, as can be deduced from the following theorem. This theorem can be obtained as a consequence of the Bahadur-Rao theorem [7], but also follows straightforwardly from the direct argument presented below.

Theorem 6.3.6. *For $i = 1,\dots,I$, let $\{M_i(t) : t \geq 0\}$ be a Poisson process with rate $\lambda_i > 0$ and $X_{i1}, X_{i2}, \dots$ i.i.d. random variables independent of M_i with $\mathbb{E}[X_{i1}] < c_i$, $\mathbb{P}(X_{i1} > c_i) > 0$, and a log-moment generating function that is finite for real values in an open neighborhood of the origin. Then, with $c = c_1 + \cdots + c_I$,*

$$\lim_{t\to\infty} \frac{1}{t} \log p^c_t \equiv \lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left(\sum_{i=1}^{I} \sum_{k=1}^{M_i(t)} X_{ik} > ct \right) = \inf_{s>0} \sum_{i=1}^{I} [\lambda_i(\mathbb{E}[e^{sX_{i1}}] - 1) - c_i s].$$

Proof. The upper bound is an immediate consequence of the Chernoff bound:

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left(\sum_{i=1}^{I} \sum_{k=1}^{M_i(t)} X_{ik} > ct \right) \leq \lim_{t\to\infty} \inf_{s>0} \left\{ \frac{1}{t} \log \mathbb{E}\Big[\exp\big(s \sum_{i=1}^{I} \sum_{k=1}^{M_i(t)} X_{ik}\big)\Big] - cs \right\}$$

$$= \lim_{t\to\infty} \inf_{s>0} \left\{ \frac{1}{t} \sum_{i=1}^{I} \log \mathbb{E}\Big[\exp\big(s \sum_{k=1}^{M_i(t)} X_{ik}\big)\Big] - cs \right\} = \inf_{s>0} \sum_{i=1}^{I} [\lambda_i\big(\mathbb{E}[e^{sX_{i1}}] - 1\big) - c_i s],$$

where the last step follows from (6.4). For the lower bound, we note that

$$\frac{1}{t} \log p^c_t \geq \frac{1}{t} \log \mathbb{P}\left(\sum_{i=1}^{I} \sum_{k=1}^{M_i(\lfloor t \rfloor)} X_{ik} > c\lfloor t \rfloor \right) + \frac{1}{t} \log \mathbb{P}\left(\sum_{i=1}^{I} \sum_{k=M_i(\lfloor t \rfloor)+1}^{M_i(t)} X_{ik} \geq c(t - \lfloor t \rfloor) \right).$$

For $M_{i,1}, M_{i,2}, M_{i,3}, \ldots$ a sequence of i.i.d. Poisson(λ_i) distributed random variables, and $k_{i,l} \equiv k + \sum_{l'=1}^{l-1} M_{i,l'}$,

$$\begin{aligned}\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left(\sum_{i=1}^{I} \sum_{k=1}^{M_i(\lfloor t \rfloor)} X_{ik} > c\lfloor t \rfloor \right) &= \lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left(\sum_{i=1}^{I} \sum_{k=1}^{\sum_{l=1}^{\lfloor t \rfloor} M_{i,l}} X_{ik} > c\lfloor t \rfloor \right) \\ &\geq \lim_{t\to\infty} \frac{1}{\lfloor t \rfloor} \log \mathbb{P}\left(\sum_{l=1}^{\lfloor t \rfloor} \sum_{i=1}^{I} \sum_{k=1}^{M_{i,l}} X_{ik_{i,l}} > c\lfloor t \rfloor \right) \\ &\equiv \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Z_1 + \cdots + Z_n > cn).\end{aligned}$$

By Cramér's theorem, letting $M_i \sim \text{Poisson}(\lambda_i)$ and using that $Z_1, \ldots, Z_n$ are i.i.d.:

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Z_1 + \cdots + Z_n > cn) &= \inf_{s>0} \left\{ \log \mathbb{E}\big[e^{s \sum_{i=1}^{I} \sum_{k=1}^{M_i} X_{ik}} \big] - cs \right\} \\ &= \inf_{s>0} \left\{ \sum_{i=1}^{I} [\lambda_i \big(\mathbb{E}[e^{sX_{i1}}] - 1 \big) - c_i s] \right\}.\end{aligned}$$

The theorem now follows from noting that for $t \in \mathbb{R}_{\geq 0} \setminus \mathbb{N}$

$$\frac{1}{t} \log \mathbb{P}\left(\sum_{i=1}^{I} \sum_{k=M_i(\lfloor t \rfloor)+1}^{M_i(t)} X_{ik} > c(t - \lfloor t \rfloor) \right) \geq \frac{1}{t} \min_{u\in(0,1]} \log \mathbb{P}\left(\sum_{i=1}^{I} \sum_{k=1}^{M_i(u)} X_{ik} > cu \right) \overset{t\to\infty}{\to} 0.$$

□

6.4 Numerical experiments

Now that we have shown how the notion of effective bandwidths can be used to regulate traffic streams in road networks so as to avoid network congestion, we perform a set of numerical experiments in order to assess the performance of the proposed policies. Specifically, we demonstrate that capacity violations are indeed rare, and that the waiting costs at the boundaries of the network are typically of a smaller scale than the incurred costs of such violations. In the experiments, the policy that follows from the effective bandwidth framework is denoted with EB(γ), where $\gamma > 0$ is such that the probability of capacity violation is upper bounded by $e^{-\gamma}$ (see (6.4)).

We compare the performance of the effective bandwidth framework, in terms of waiting times and number of capacity violations, to three other input rate control algorithms, which serve as natural benchmarks. As a first benchmark, we use the procedure that simply allows all vehicles in the network at all times, which will be called 'No Control' (abbreviated to NC). The second benchmark does incur waiting costs at the boundary, as it will not allow the expected capacity needs on the links to exceed the link capacities, i.e., $\mathbb{E}[Y_{a_j}] < C_{a_j}$ for all $a_j \in A$. This procedure will be called 'Expected Needs', and the corresponding policies are abbreviated with EN. The third benchmark, called 'Random

Needs', takes the random nature of capacity needs into account, and only allows the current input rate when, for some chosen $\alpha > 0$ and all $a_j \in A$,

$$\mathbb{E}[Y_{a_j}] + \alpha\sqrt{\text{Var}\left(Y_{a_j}\right)} < C_{a_j}.$$

The corresponding procedure is abbreviated to RN(α). When aiming for a similar guarantee as in the EB framework, a natural way to calibrate α is to use a normal approximation and set $\alpha \equiv z_{\exp\{-\gamma\}}$, with, for $Z \sim \mathcal{N}(0,1)$ and $\beta > 0$, z_β such that $\mathbb{P}(Z \geq z_\beta) = \beta$. For any $\gamma > 0$, we will denote such a calibration as α_γ.

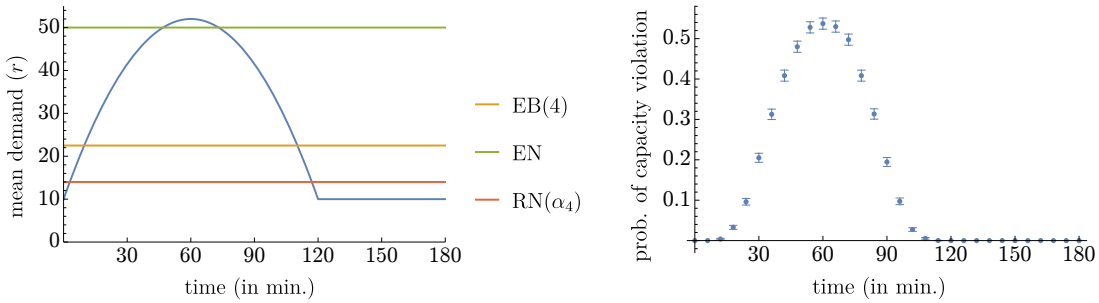
Experiment 1 examines the impact of the different policies on a network consisting of a single link, whereas Experiment 2 considers networks of a larger size. In both experiments, it is shown that with an EB(γ) policy (for γ high enough) capacity violations rarely occur. As a result, a proxy for the total time vehicles spend in the system (i.e., their total waiting and travel time) is typically lower for a such a policy than for policies that do not limit the number of capacity violations, such as the NC and EN policies. Note that, if the capacity needs are approximately Gaussian, the RN($z_{\exp\{-\gamma\}}$) policy will have the same performance as the EB(γ) policy. However, if these are not approximately Gaussian, it may either be that the RN($z_{\exp\{-\gamma\}}$) policy is more conservative (such that the number of capacity violations is still low, but the waiting times may be higher), or that the RN($z_{\exp\{-\gamma\}}$) policy is less conservative (potentially leading to more capacity violations and, consequently, high delays).

Before presenting the experiments, we remark that, as stated above, the experiments contain proxies for the total delay under the different control policies. The interest in these delays stems from the observation that, solely focusing on the number of capacity violations, any policy that limits a (relatively) high amount of traffic performs well, even though some of these policies may result in very high waiting times at the boundaries of the network. By computing proxies for the total of all waiting and driving times, only the policies that find a good balance between these two are considered well-performing. The proxy we put forth for the total time vehicles spend in the system is based on simulations of a queueing system that captures the impact of high volumes on the evolution of traffic. When presenting the experiments, this simulation approach is described in more detail.

Experiment 1

In this experiment, we consider, for illustration purposes, a network of a single link, and study the impact of different input rate control policies in a setting in which traffic demand is rising. Specifically, we compare the performance of an effective bandwidth policy with the presented benchmark policies, in terms of capacity violation instances and total delay. The single network link has a capacity of 50, and the amount that one vehicle contributes to the occupied capacity is modeled as $D \sim \text{Hyperexponential}(\boldsymbol{p}, \boldsymbol{\lambda})$, with $\boldsymbol{p} = (0.7, 0.3)$ and $\boldsymbol{\lambda} = (3/2, 9/16)$. This could, for example, represent a traffic stream for which any vehicle is with 70% probability a car and with 30% probability a truck; cars and trucks occupying a capacity of 2/3 and 16/9, respectively, such that the expected capacity needs of an arbitrary vehicle equal 1.

We consider a rush-hour setting, in which, starting at a low mean demand at time 0 (i.e., the time the rush hour initiates), the mean input rate of the link first monotonically



(a) Mean traffic demand r as a function of time in a rush hour setting, together with the maximum input rates of three control policies.

(b) Simulated probability of capacity violation (95% confidence interval) as a function of time for the policy NC.

Figure 6.2: Example of the mean demand during rush hour on a single link.

increases, and then monotonically decreases, after which the rush hour has passed, and the mean demand stays constant. Specifically, in this experiment, we consider the mean traffic demand to evolve as in Figure 6.2a. For this mean demand curve, Figure 6.2b plots the (simulated) probability of a capacity violation at several points in time, in case there is no input rate control (policy NC). Setting our goal to keep this probability below e^{-4}, we evaluate the network performance for the policies EN, RN(α_4), and EB(4).

Figure 6.2 also shows the maximum input rates for the different control policies. For any of the control policies, the probability of capacity violation can be read off the right hand graph, as long as the mean traffic demand (in the right hand graph) has not yet reached the imposed threshold. For the EN policy this yields a probability larger than 0.45 that a capacity violation occurs during the peak period of the rush hour. For the NC policy, allowing all traffic demand into the network, this probability is larger than 0.5. EB(4) and RN(α_4) are more conservative, and limit access at an earlier stage. Recall that the latter two policies target at a probability of capacity violations of $e^{-4} \approx 0.02$ (for RN this is by approximation using a normal distribution, and for EB it is a bound).

Our aim is to avoid the high delays caused by capacity violations, at the expense of some additional delay on the boundaries of the network. Having shown that capacity violations under the EB(4) and RN(α_4) policies are rare, we assess the gains of such approaches in terms of a proxy for the total delay vehicles experience during their travel. That is, we investigate if, under these policies, the waiting costs at the boundary of the network are of a smaller scale than the travel times under less conservative policies. As stated in the introduction of this section, the waiting-time and travel-time proxies are found using simulations of a queueing system that captures the impact of high volumes on the evolution of traffic. Specifically, the link is associated with a single-server queue whose service rate is a function of the queue length, representing the fact that vehicle speeds on a link are dependent on the capacity needs on that link.

First, we discretize time in steps of size $\delta > 0$, and view the evolution of traffic as a queueing model that consists of a waiting area of infinite size at the starting node of the link, which we will refer to as the *buffer*, and a FCFS *queue* on the link itself. The traffic demand in the buffer and queue at time t are denoted by $B(t)$ and $Q(t)$, respectively. At each time step $t = \delta m$, the mean traffic demand r_t is obtained from the area below the function in Figure 6.2a between $\delta(m-1)$ and δm. The control policy then decides how

	NC	EN	RN(α_4)	EB(4)
Delay estimate (in min.)	49.43	49.41	66.57	41.97

Table 6.1: Estimates for the delays vehicles experience in a single-link network under different input rate control policies, for the mean demand curve of Figure 6.2a.

much of $B(t-\delta)$ and r_t is offered to the queue. Let $\bar{\rho}_t$ be the maximum mean flow rate the control policy allows at time t. Then, a mean flow rate of $\bar{r}_t \equiv \min\{\bar{\rho}_t, B(t-\delta) + r_t\}$ is admitted to the queue, such that $B(t) = B(t-\delta) + r_t - \bar{r}_t$.

Whenever, at time $t = \delta m$, a mean traffic load of $\bar{r}_t$ enters the link, the corresponding number of vehicles is simulated, and they are placed in the queue. For each of these vehicles, a sample of their capacity needs yields their service requirement. The amount of capacity needs that the server is able to process between δm and $\delta(m+1)$ is given through a function $c(\cdot)$, which takes the total capacity needs in the queue at time δm as input. The shape of the function is chosen to represent the impact of high capacity needs on the driveable speeds. That is, whenever the capacity needs are less than the capacity on the link, vehicles are effectively able to drive the free-flow speed, represented by the fact that all capacity can be processed. Whenever the capacity needs exceed the link capacity, the attainable speeds are significantly lower, which is represented by the fact that $c(\cdot)$ outputs a value that is far less than the link capacity. Specifically, in this example, we let

$$c(y) \equiv \begin{cases} y & y \leq 50\delta, \\ \max\{10\delta, 100\delta - y\} & y > 50\delta. \end{cases} \tag{6.6}$$

Note that we multiply the numbers by δ, to account for the fact that we look at time windows of size δ. Furthermore, under congested conditions, we impose a strictly positive service rate, to make sure that, for low arrival rates, the model is able to recover from these congested conditions at a future time point.

With the procedure described above we obtain a proxy for the total delay in the network in the following way. At each time t, an approximation for the total number of customers in the system is given through the sum of $B(t)$ and the number of vehicles in the queue. Computing this sum for each simulation run and each point in time, we obtain an approximation for $\tilde{L}$, the average number of customers in the system. Moreover, averaging over the mean demands r_t of the different time steps yields an estimate $\tilde{r}$ for the average arrival rate. Then, with Little's law, we obtain the proxy $\tilde{L}/\tilde{r}$ for the delay vehicles in the network experience. The proxies for different control procedures, in the setting described above, with the time range of Figure 6.2a, $B(0) = 0$, $Q(0) = 0$, and $\delta = 1$ min, are given in Table 6.1. These proxies reveal that the delay for the NC and EN policies is of the same order, which can be explained by the fact that EN solely limits access in the peak of the rush hour. The delay under EB(4) is significantly lower, whereas the delay under RN(α_4) is of a very high order. Thus, the EB(4) policy balances the waiting and driving costs best.

The differences in experienced delays as presented in Table 6.1 can also be observed in Figure 6.3, which shows the buffer and average queue size for the four policies as a function of time. Note that the mean demand arriving at the buffer is deterministic, such that the buffer size, in terms of mean demand, is a deterministic quantity. The number of vehicles entering the queue being a random variable, the queue size is a random quantity, whose

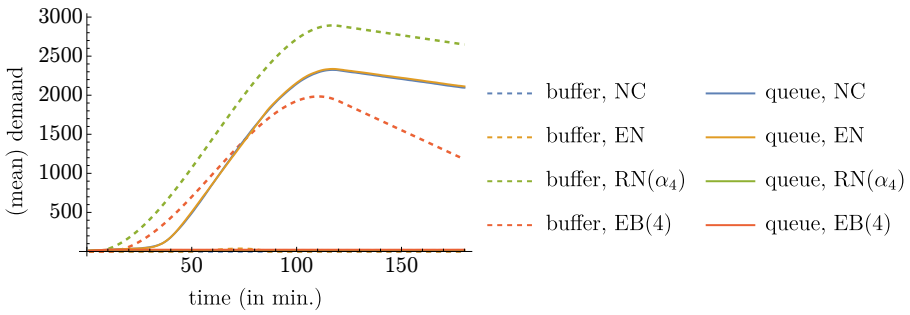


Figure 6.3: Buffer and average queue size as a function of time, for four different control policies.

average is determined with 10.000 simulations. Again, we observe that there is little difference between the NC and EN policies, because both allow a high mean demand onto the link, such that the buffer stays relatively empty. However, on the link itself, due to violations of the link capacity, the traffic becomes, and stays, highly congested. Figure 6.3 shows that both EB(4) and RN(α_4) indeed succeed in avoiding high delays on the link, as the queue sizes under both policies remain of a very small scale. Under EB(4), the demand in the buffer grows during the rush hour period, but as there is no congestion on the link, the total delay is still significantly lower than in the NC and EN regimes. This is not the case for the RN(α_4) policy: being extremely conservative, the buffer size is such that the total delay exceeds the NC and EN regimes.

It is important to remark that, even though the total delay under RN(α_4) is larger than under the NC and EN policies, this conservative policy may still be preferred, as there is a difference between the waiting costs of vehicles on the boundaries of the network and the delay the vehicles experience inside the network. That is, by using Intelligent Transportation Systems, drivers may, for example, request their route from their home, and learn when they will be granted access to the network from there, such that waiting at the boundaries does not automatically yield wasted time for the affected drivers. In contrast, waiting inside the network does yield wasted time, and, moreover, results in additional CO_2 emissions.

Experiment 2

To examine the impact of capacity violations in larger networks, we now consider the linear network of Figure 6.4, and vary m, the number of links of which the network consists. For different values of m, we then consider the performance of the presented control policies, again in terms of the number of capacity violations, and the total time vehicles spend in the system. For the latter, we compute proxies in a similar way as in Experiment 1, in that we construct a queueing system with state-dependent service rates.

Let the network consist of a single traffic stream, with vehicles wanting to travel from the left node to the right node. For this stream, consider the same mean demand curve and the same capacity needs characteristics as in Experiment 1. Moreover, let all link capacities equal 100, except for the last link, whose capacity is chosen to be 50. To compute proxies for the delays vehicles experience under the control procedures, discretize time in steps of size $\delta > 0$, and again view the evolution of traffic as a queueing system, with a single buffer of infinite size at the starting node, and FCFS queues on each of the m links. The

Figure 6.4: A linear network with m links.

	NC	EN	RN(α_4)	EB(4)
$m = 5$	49.36	48.99	68.49	45.01
$m = 10$	49.98	49.73	70.84	48.73
$m = 20$	55.43	56.52	75.30	54.81
$m = 30$	62.23	61.82	79.49	60.43

Table 6.2: Estimates for the delays vehicles experience in the network of Fig. 6.4, under different input rate control policies, for the mean demand curve of Figure 6.2a.

buffer dynamics are similar to those in Experiment 1, with the vehicles that are allowed into the network arriving at the queue that corresponds to link 1. These vehicles visit the queues consecutively: traffic that has been served by the queue of link i ($1 \leq i \leq m-1$) is inserted in the queue of link $i+1$. If a vehicle has been served by the queue of link m, it leaves the system. Now, by summing, for each simulation, the vehicles in all queues at every time step, an application of Little's law again yields the order of travel-time delays.

Let $c_i(\cdot)$ be the amount of capacity needs that the server at the queue on link i is able to process. With $\boldsymbol{y}$ a vector of the capacity needs on each link, y_i denoting its i-th coordinate, and $c(\cdot)$ as in (6.6), we let $c_m(\boldsymbol{y}) = c(y_m)$. Then, to capture the fact that traffic jams propagate through the network, we let, for any link $1 \leq i \leq m-1$, the function $c_i(\cdot)$ explicitly depend on the total capacity needs in its queue, as well as the total capacity needs in the next queue. That is, the shape of the function represents that if the next queue is highly congested, new traffic is not able to enter this queue, and has to stay in its current queue. Specifically, allowing not more than $u_i \equiv 190\delta$ per time step into queue i for any $1 \leq i \leq m-1$, and not more than $u_m \equiv 90\delta$ per time step into queue m,

$$c_i(\boldsymbol{y}) \equiv \begin{cases} \min\{u_{i+1} - y_{i+1} + c_{i+1}(\boldsymbol{y}), y_i\} & y_i \leq 100\delta, \\ \min\{u_{i+1} - y_{i+1} + c_{i+1}(\boldsymbol{y}), \max\{10\delta, 200\delta - y_i\}\} & y_i > 100\delta. \end{cases}$$

With $c_m(\boldsymbol{y})$ known, the amount of vehicles that are served in each queue can be computed recursively.

The proxies under different input rate control schemes are presented in Table 6.2, and are, naturally, increasing functions of m. Just as in Experiment 1, the NC and EN policies behave relatively similar. Their high delay values are caused by the fact that the buffer is relatively empty throughout the complete time window. EB(4) outperforms these policies, but the high buffer value causes quite high delays, especially for large values of m. However, as argued above, waiting at the boundaries of the network is significantly different from waiting within the network, as it is not directly a time waste, and has no negative environmental consequences. A similar argument can be made for RN(α_4), which has the highest delays, but may still be preferred over NC and EN.

6.5 Concluding remarks

In this chapter, we used a compound Poisson process to describe the random part of the road capacity that is effectively taken by the traffic streams using that road. To avoid that, in a given road network, link capacities are exceeded, we constructed an input rate control policy that takes the randomness of these capacity needs into account. This policy guarantees an upper bound on the capacity-violation probability, and is based on the notion of effective bandwidths as originally introduced in the telecommunications context. Numerical experiments demonstrated that, typically, the total delay in the network is of a smaller scale than the delay would be without access control, or with a policy that only takes expected capacity needs into account.

There are a few natural ways in which our modeling procedure may be extended, towards which future work could be specified. For example, in the current procedure, when determining the input rate control, the location of the vehicles on the paths does not play a role. Specifically, if given access to the network, in our model setting, a demand increase on a given path will instantaneously lead to an increase in the capacity needs on all links on this path. In reality, however, there is a time-component, and an increase in traffic demand at the boundary will only lead to an increase in capacity needs on further links at later points in time.

The aim of our control procedure was to limit the negative consequences of capacity violations. Although our procedure does describe the traffic demand that is allowed into the network, it does not consider the fairness in terms of waiting time. That is, if the demand on a single link consists of two traffic streams with the same characteristics, instead of allowing a part of both streams, our procedure may only allow one of the streams. Therefore, in terms of practical operationalization, a potential suggestion for future research would be to adapt our procedure so as to meet certain fairness guarantees.

An interesting application of this chapter would be the identification of network bottlenecks. Since our procedure can identify, for a given demand, the set of links whose capacities have a significant probability of being violated, this will provide an impression of the bottlenecks in the network. Notably, with the characteristics of recurrent traffic demand well known, this may be helpful in deciding on future changes in network infrastructure.