View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics

Department of Economics

Network development under a strict self-financing constraint

by

André DE PALMA
Stef PROOST
Saskia VAN DER LOO

ETE

Center for Economic Studies
Discussions Paper Series (DPS) 08.29
http://www.econ.kuleuven.be/ces/discussionpapers/default.htm

October 2008

DISCUSSION PAPER

Network development under a strict self-financing constraint

André de Palma[1], Stef Proost[2] and Saskia van der Loo[3]

Abstract

This paper offers a stylized model in which an agency is in charge of investing in road capacity and maintain it but cannot use the capital market so that the only sources of funds are the toll revenues. We call this the strict self-financing constraint in opposition to the traditional self financing constraint where implicitly 100% of the investment needs can be financed by loans. Two stylised problems are analysed: the one link problem and the problem of two parallel links with one link untolled. The numerical illustrations show the cost of the strict self-financing constraint as a function of the importance of the initial infrastructure stock, the rate of growth of demand, the price elasticity of demand and the flexibility in the pricing instruments.

Keywords: Cost-benefit analysis, road tolling, self-financing, infrastructure investments, congestion, bottleneck model

JEL Classification: R42, L91, R40

Acknowledgements and author information

[1]André de Palma: Ecole Normale Supérieure de Cachan. andré.depalma@ens-cachan.fr

[2] Stef Proost: Center for Economic Studies, KULeuven. Naamse straat 69 3000 Leuven, Belgium. stef.proost@econ.kuleuven.be

[3] Saskia van der Loo: Center for Economic Studies, KULeuven. Naamse straat 69 3000 Leuven, Belgium. saskia.vanderloo@econ.kuleuven.be

We thank David Levinson and anonymous referees for helpful comments. The authors would like to acknowledge the French Ministry of Ecology and sustainable development (Land transportation research and development program, PREDIT) for supporting the Project "Gestion du transport et de la mobilité dans le cadre du changement climatique" (Ref. 07MTS063), the FUNDING consortium and the Policy research Center Budgetary & Tax Policy, Government of Flanders (B) for financial support.

1 Introduction

The objective of the paper is to better understand the growth of a transportation network, when it operates under "strict" self-financing constraints. By "strict" self-financing constraint we understand a combination of two requirements. First that an agency, in charge of the maintenance and construction of new infrastructure, cannot borrow on the capital market and has to finance maintenance and extensions out of current toll revenues. Second the agency is also subject to an earmarking or hypothecation constraint as it is forced to spend all the toll revenues on maintenance and construction of roads.

This is a polar case and in strong contrast to the traditional version of the self-financing constraint. The traditional version is however also a polar case as it implicitly assumes that the infrastructure agency can borrow the full amount of the infrastructure cost and can also return any excess revenues to the rest of the economy. In many countries (France, Japan, Germany, Norway, US ..), an agency receives, at its creation, an initial infrastructure stock and corresponding outstanding debt on which it has to pay interest. It also receives the right to toll roads but it has to break even while its call upon the capital market is limited and sometimes reduced to zero. Two examples: in France, the road agency had to respect a debt to earnings ratio of 7, in the US the Federal Highway Fund cannot borrow at all on the capital market. There are two reasons for this limited borrowing capacity. First, as in the case of a private firm, only part of the total investment can be borrowed on the capital market as there remains a risk for the lender. Second, in the case of a public agency, voters are reluctant to give an independent agency the power to build up a parallel public debt without their consent.

We consider the extreme case where the toll revenues can only be used to pay the maintenance, to pay existing debts and to pay the investments in the extension of the network. The main reasons for this earmarking practice and the limited capacity to take loans are institutional (see [7]). Many road agencies are created as a reaction to a period where the government messed up infrastructure policy resulting in insufficient and badly maintained road infrastructure. By creating a separate, independent agency with its own sources of income that cannot be diverted to the government, there is a guarantee that investments are made in road infrastructure. The institutional details differ strongly among countries. Raux et al (2007) [8] discuss the case of France where a Highway fund was created in 1955 and where the use of the toll revenues was restricted to maintenance and extensions. In France, like in many countries, tolling a road is only possible if an untolled alternative is available. Doll et al (2007) [4] discuss the recent German toll on highways and its best use in their Highway Fund. In the US, the Highway Trust fund is fed by excise taxes on gasoline and the resources of the Fund are in principle hypothecated to highway infrastructure.

The polar case we develop in this paper serves to illustrate what are the effects and costs of the two institutional restrictions on debt finance and hypothecation of funds we discussed. In order to gain insight we focus on the explicit dynamic modelling of one link or at most two links (one tolled and one

parallel untolled link), homogeneous users and a bottleneck representation of congestion. One agency is responsible for maintenance, tolling and investment, has to break even in every period, and cannot call on the capital market to smoothen infrastructure decisions by taking on loans. The initial conditions of infrastructure and debt are important for the outcome. Most agencies are created when there is dissatisfaction about the level of infrastructure supply. For this reason we will use as initial condition a level of infrastructure supply that is smaller than the optimal level.

More specifically we address the following research questions. First what type of capacity development can we expect on a single link when there is no possibility to borrow on the capital market? What is the impact of the initial conditions (i.e. the configuration of the initial network) on the future network growth? How many years does it take to reach the optimal capacity size? Does the fine toll, that is optimal in the static bottleneck model, lead to provision of too much capacity or to too low capacity levels? How does this depend on the type of tolling and on the growth rate and price elasticity of demand that are in place? The same questions can be raised for the two link case where only one link can be tolled and extended. The revenue capacity of the optimal second best toll is now smaller and the potential for network extension is more limited.

We show that the smaller the initial infrastructure stock (relative to the first best optimal stock), the larger will be the welfare losses in the period where the infrastructure stock is catching up but also the larger will be the welfare losses once the optimal infrastructure stock has been reached because one ends up in a regime with larger overinvestments in capacity. The numerical illustrations show that the inefficiency of the strict financing constraint is equivalent to an increase of 5% to 120% of the capacity costs in the first best where the strict self-financing constraint does not apply. This implies that only severe institutional failures would make a strict self-financing constraint acceptable.

The link between tolling revenues and investment needs is a theme that has been explored thoroughly in the literature. The "traditional" self-financing theorem (see [6] , for a recent and comprehensive literature overview of self-financing constraints see Verhoef and Mohring (2007) [9])) holds for our case with one bottleneck link where there are constant returns to scale in congestion technology and constant average costs of capacity extension. Toll revenues from fine tolling are sufficient to pay the rental cost of capacity. But the use of a rental cost concept for capacity presupposes that a financial sector is ready to pay for the investment in exchange for a share of the future toll revenues. Under our strict self-financing constraint the rental cost of capacity loses its meaning and the build up of new capacity is more difficult. Verhoef and Mohring (2007) [9] discuss the problem where one starts with an optimal capacity but where all toll revenues have to be reinvested into new capacity. They call this the "naive interpretation" of the self-financing theorem. In fact this can also be seen as a re-investment rule. Our case is different and more realistic: we start from a too low capacity, interest is paid on the initial capacity, there is ear-marking of toll revenues for investments and we consider two-part toll as a way out for excessive investments.

Section 2 of this paper defines the one link model, the optimal fine toll and the optimal capacity in the absence of a strict self-financing constraint. Section 3 discusses the properties of the stationary state that results from fine tolling under a strict self-financing constraint and when demand is time independent and price inelastic. In section 4 we consider the more general case of growing and price elastic demand and analyze the case of a two-part toll consisting of a fine toll (toll varying within the day) and a fixed component that is independent of the timing of his trip within the day. We show that this two-part toll can bring large welfare gains. Section 5 presents the two link case where one link is untolled. Section 6 presents numerical simulations to illustrate the orders of magnitude associated to the strict self-financing constraint. Section 7 concludes.

2 The one link model

Consider a route, joining an origin to one destination. The capacity of this route at the calendar time t is denoted by $s(t)$. Capacity can be varied continuously.

We use the reduced form of a dynamic (bottleneck) model with homogenous users, endogenous trip times and one desired arrival time (see [1]). In this model identical individuals all want to arrive at the same desired arrival time but this is impossible because the capacity of the bottleneck is too small. In the no-toll solution, the Nash equilibrium implies that all users have the same total cost that consists of the sum of travel cost, queueing cost and schedule delay cost. The schedule delay cost consists of either the cost of being too late or the cost of being too early. The optimal toll solution for given capacity consists of using a fine toll that varies over the day in function of the departure time. This type of toll is capable of eliminating all queueing costs but the schedule delay costs remain. In addition, because the average queueing cost is transformed into average toll revenue, total demand is identical to the demand level without the fine toll. For given capacity and given demand function, the fine toll is the most efficient way to deal with the intraday congestion problem. In this paper we are mainly interested in the optimization of the capacity level and for that reason we can leave the intraday dimension of the fine toll implicit and concentrate on the intertemporal evolution of the average revenue of the fine toll, the total demand per day and the capacity.

The elementary period t we consider is a day or a year in which the capacity and demand conditions (with its intraday variations) can be considered constant. In this equilibrium, the average (over the users within one day) of the cost function during the time interval $[t, t + dt)$ dt is (up to an additive constant which equals the travel time at maximum speed, omitted):

$$C(t)dt = \delta \frac{N(t)}{s(t)} dt \tag{1}$$

where δ is a summary measure of the schedule delay costs $\delta = \beta\gamma/(\beta + \gamma)$, β being the unit cost for early arrivals and γ being the unit cost for late arrivals, with typically $\beta < \gamma$. Eq(1) is the average cost in equilibrium where none of

the users can reduce their travel cost (schedule delay and queueing) by choosing another departure time within the day. Usage at time t is denoted by $N(t)$, that is growing over time: $dN(t)/dt \geq 0$; we assume first that demand is time dependent but price inelastic. In the bottleneck model, half of this average travel cost is due to schedule delay costs, while the other half are the queueing costs. It is well known that the optimal toll is an intraday varying toll, called fine toll, that totally eliminates queueing. Alternatively, a coarse toll, with steps, eliminates a smaller fraction of the total cost. As the number of steps goes to infinity, the solution converges towards the fine toll. In the rest of the paper we concentrate on the fine toll.

Maintenance cost $M(t)dt$ during the interval $[t, t+dt)$ is the sum of a term which depends on usage, and a term which is independent of usage. The second term depends on natural degradation, due for example to bad weather conditions (induced potholes, cracks and the like) (see [5]). We have:

$$M(t)dt = a(Q)N(t)dt + b(Q)s(t)dt, \text{with } a(Q) > 0, b(Q) > 0 \tag{2}$$

where Q is the quality of the road. In the sequel, we will omit the quality dependency. The optimal level of the toll (in fact an average over the day), computed on a marginal cost pricing principle, is the sum of the optimal congestion charge and the maintenance cost due to usage:

$$\tau(t) = \delta \frac{N(t)}{2s(t)} + a. \tag{3}$$

This is the toll that optimizes the welfare for a given capacity. Total toll revenue collected during $[t, t+dt)$ is:

$$TR(t)dt = \left\{ \delta \frac{N(t)}{2s(t)} + a \right\} N(t)dt. \tag{4}$$

Without a strict financial constraint, the first best capacity $s(t)$ is the capacity that minimizes total costs which are (assuming optimal tolling), the sum of the total scheduled delay costs, the rental cost of capacity and the maintenance costs:

$$\text{Total Costs} = \delta \frac{N(t)}{2s(t)} N(t) + i\kappa s(t) + aN(t) + bs(t) \tag{5}$$

where i is the interest rate and κ is the unit construction cost of new capacity. We assume constant returns to scale for road construction. With constant input prices, this implies that the construction cost of a unit of capacity ds is κds. Capacity, once constructed, has an infinite lifetime. This implies that only interest is due. The absence of a strict financial constraint means that in every period, capacity can be adapted optimally in function of demand at a rental price equal to the interest rate.
Solving the first order condition yields the first best capacity:

$$s^{opt}(t) = N(t) \sqrt{\frac{\delta}{2(i\kappa + b)}}. \tag{6}$$

Optimal capacity is an increasing function of the number of users and the scheduled delay cost but a decreasing function of the cost of installing and maintaining capacity.

3 Fine tolling and the strict self-financing constraint

Although the toll given in eq(3) optimizes the welfare for a given capacity, it is not necessarily the toll that maximizes welfare under a strict self-financing constraint. The strict self-financing constraint considered in this paper requires that the total toll revenue collected during the infinitesimal time interval $[t, t+dt)$ is used for maintenance, for payment of the annual interest on the initial capacity (s_0) and for construction of new capacity. We have therefore the following accounting balance:

$$\left\{\delta\frac{N(t)}{2s(t)}+a\right\}N(t)dt = aN(t)dt + bs(t)dt + i\kappa s_0 dt + \kappa ds, \text{or}$$

$$\delta\frac{[N(t)]^2}{2s(t)}dt = bs(t)dt + i\kappa s_0 dt + \kappa(ds). \tag{7}$$

We assume that the sum raised by the toll is large enough to cover the maintenance cost, so that the residual is used to build extra capacity. In the limiting case, the amount raised is exactly equal to the maintenance cost. It may be the case that when the government wishes to start with an autonomous agency and a strict self-financing constraint, the fixed maintenance cost is larger than the toll revenue. Such situations are not impossible, but disregarded here.
Expression (7) is a differential equation, which can be written as:

$$\frac{ds}{dt} = \frac{\delta}{\kappa}\frac{[N(t)]^2}{2s(t)} - \frac{b}{\kappa}s(t) - is_0. \tag{8}$$

We have assumed that the initial capacity (at $t=0$) is small enough (or maintenance cost b is small enough), so that $ds/dt > 0$, i.e.

$$s_0 < N(0)\sqrt{\frac{\delta}{2(i\kappa+b)}}. \tag{9}$$

Note that the left hand side of the inequality is just the first best capacity at $t=0$ (see eq(6)).

For time-dependent demand functions we have to rely on numerical simulations, but when demand is constant over time and inelastic $\left(N(t)=\bar{N}\right)$ we have a stationary state for eq(8):

Proposition 1 *For time independent and inelastic demand $\bar{N}$ and if the usage independent maintenance costs $b = 0$, the stationary state $\widehat{s}$ is given by*

$$\widehat{s} = \frac{\left(s^{opt}\right)^2}{s_0}, \tag{10}$$

if $b \neq 0$ then

$$\widehat{s} = \frac{-i\kappa s_0 + \sqrt{(i\kappa)^2\left(s_0^2 - \left(s^{opt}\right)^2\right) + (i\kappa + 2b)^2\left(s^{opt}\right)^2}}{2b} \tag{11}$$

where s^{opt} is given by eq(6) where $N(t) = \bar{N}$.
Proof. *If $b = 0$: the stationary state is the solution of following equation*

$$\frac{ds}{dt} = 0 \Rightarrow \frac{\delta}{\kappa}\frac{\bar{N}^2}{2} - is_0\widehat{s} = 0$$

solving for $\widehat{s}$ we get

$$\widehat{s} = \bar{N}^2\frac{\delta}{2i\kappa s_0} = \frac{\left(s^{opt}\right)^2}{s_0}$$

It is obvious that $\widehat{s} > s^{opt}$ if $s_0 < s^{opt}$.
To find the stationary state for $b \neq 0$ we need to solve

$$b\widehat{s}^2 + i\kappa s_0\widehat{s} - \frac{\delta\bar{N}^2}{2} = 0$$

and

$$\widehat{s} = \frac{-i\kappa s_0 + \sqrt{\left(i\kappa s_0\right)^2 + 2b\delta\bar{N}^2}}{2b}$$

From eq(6) we get that $\delta\bar{N}^2 = 2(i\kappa + b)\left(s^{opt}\right)^2$, substituting this into the previous equation and after some rearrangement of the terms we get eq(11).
To prove $\widehat{s} > s^{opt}$ if $s_0 < s^{opt}$ rewrite the inequality as

$$\widehat{s} > s^{opt} \Leftrightarrow \sqrt{\left(i\kappa s_0\right)^2 + 2b\delta\bar{N}^2} > 2bs^{opt} + i\kappa s_0$$

squaring both sides and substituting $\delta\bar{N}^2$ this is equivalent with

$$\left(i\kappa s_0\right)^2 + 4b(i\kappa + b)\left(s^{opt}\right)^2 > 4b^2\left(s^{opt}\right)^2 + \left(i\kappa s_0\right)^2 + 4bi\kappa s_0 s^{opt}$$

which reduces to

$$4bi\kappa\left(s^{opt}\right)^2 > 4bi\kappa s_0 s^{opt}$$
$$s^{opt} > s_0$$

which is true by assumption. ■

To understand the intuition of this result, assume first $s_0 = s^{opt}$ and remember that the level of congestion determines the level of the toll revenues, then no capacity additions are needed and all toll revenue is used to pay for the initial capacity stock at a cost $i\kappa s_0$. In this case one stays at the optimum capacity. Take now a somewhat smaller initial capacity. The budget surplus that remains after paying for the initial capacity is reinvested every year even when the optimal capacity is reached. The capacity keeps growing until the capacity is so large that the fine toll revenues equal the cost of paying for the initial infrastructure. Take now a very low initial capacity. This implies that the yearly cost of this initial capacity is also very low. The fine toll now leaves a larger budget surplus for investments and capacity extension can go on for much longer. The resulting steady state infrastructure stock implies a much more excessive capacity level.

So the initial capacity plays a double role. First a higher initial capacity allows to reach the optimal capacity more quickly as the need for investments in the short run is more limited. Second, and this is less obvious, it limits the surplus available for capacity extension and this helps to limit the construction of excessive capacity in the long run. As we will later demonstrate numerically, using price elastic demands, this insight has important policy implications. Tolling is often introduced when capacity is much too small and then an agency is often created with a strict self-financing constraint and a low initial debt in order to guarantee a quick build up of capacity. However, it is precisely under these conditions that the costs of overinvestment will be largest.

One could think of other initial conditions for the infrastructure capacity. Given that the agency receives the power to toll it is logical to require it to pay the cost of the initially received infrastructure under the form of interest. One alternative assumption is that the agency starts with no infrastructure at all but then the model is no longer defined. Another alternative is that the agency receives for free a small initial infrastructure. Then the strict self-financing constraint would generate even more extreme results.

4 One Mode with two-part tolling and financing constraint

The fine toll is not necessarily the toll that maximizes welfare under a strict self-financing constraint. It is obvious that the introduction of an additional fixed term, not varying within the day, could bring in extra financial resources or allows to decrease toll revenues when they are not longer needed. The optimal fixed term is difficult to determine but one can state that it will always be optimal to charge the fine toll in order to eliminate the queuing and convert the queuing into toll revenues. We therefore consider a two-part toll whose first part consists of the fine toll $\tau^{\text{fine}}(t)$ and a second part that is a fixed part or

"base toll" $F(t)$ (see de Palma, Lindsey (2000) [3]). :

$$toll(t) = \tau^{\text{fine}}(t) + F(t). \tag{12}$$

The fixed part $F(t)$ can be negative or positive and allows to either raise more revenue when large investments are needed or to limit revenue raising once the optimal capacity is reached. We need to distinguish the case of price inelastic and price elastic demand.

When demand is inelastic and varying over time, the optimal tolling and investment regimes can be achieved as follows: if we start with a capacity s_0 with $s_0 < s^{opt}$ then invest in first period $s^{opt}(t) - s_0$. This investment is financed by a fine toll plus a flat toll:

$$\left[F(t) + \tau^{\text{fine}}(t)\right] N(t) \geq i\kappa s_0(t) + \left[s^{opt}(t) - s_0\right]\kappa + bs^{opt}(t), \tag{13}$$

in the second period, when the first best capacity is reached the flat toll becomes a subsidy

$$F(t) N(t) = -\tau^{\text{fine}}(t) N(t) + bs^{opt}(t) + i\kappa s_0. \tag{14}$$

Because demand is inelastic the flat part of the toll allows to overcome the strict self-financing constraint by simply charging all users the sums necessary to reach optimal capacity and by redistributing the surplus once the optimal capacity is reached. We assumed implicitly that the fixed part of the toll does not exceed the income of the representative consumer, if this would be the case, one has to proceed more gradually.

When demand is elastic and time-dependent, finding the optimal two-part tariff is a difficult optimal control problem. We limit ourselves here to a heuristic approach for case where demand is elastic but time-independent . The search for the optimal structure of the fixed part of the toll ($F(t)$) relies on two principles. First, the fixed part of the toll can be used to raise more revenues and increase the rate of the investments; we will assume the fixed part to be proportional to the difference between the actual capacity and some capacity ($s^*(t)$) which we will call third-best capacity for reasons that will become clear later. Second, once the third best level of capacity ($s^*(t)$) is reached, we would like to stay at that level and not invest anymore; the fixed part of the toll can then be used to set the level of the total toll revenues equal to the cost of the initial capacity and the maintenance cost so that no residual toll revenues are left . To be more precise, $F(t)$ is of the form:

$$F(t) N(t) = \left\{ \begin{array}{ll} \alpha\kappa\left(s^*(t) - s(t)\right) & \text{if } s(t) < s^*(t) \\ -\tau^{\text{fine}}(t) N(t) + i\kappa s_0 + bs^*(t) & \text{if } s(t) = s^*(t) \end{array} \right\}. \tag{15}$$

The first best capacity derived in eq(6) is the capacity that maximizes welfare given that the toll equals the fine toll and that there is no strict financing constraint. In our case, however, the toll is not equal to the fine toll so the "optimal" capacity level can be different from the first best capacity derived in

eq(6)). The capacity $s^*(t)$ is the capacity that maximizes welfare given that the toll is equal to $i\kappa s_0$:

$$W = \int_p^\infty N(p')\,dp' - i\kappa s - bs. \tag{16}$$

Proposition 2 *If demand is elastic $N(p) = p^{-\eta}$, then the optimal capacity $s^*(t)$ given a toll equal to the sum of the fine toll and $F(t)$ given in (15) is larger (or smaller) than the first best capacity, if $\eta < 1$ ($\eta > 1$). More precisely:*

$$s^*(t) = N(t)\sqrt{\frac{\delta}{2(1+\eta)(i\kappa+b)}}. \tag{17}$$

This capacity level is third-best because three constraints are present: first there is the strict financial constraint, second the initial capacity is assumed to be inefficiently low and third it maximizes welfare in a myopic way.

5 Two Modes: the untolled alternative

Consider the case of two parallel modes connecting an origin O to a destination D. Suppose one mode is untolled and has a fixed capacity (s_U). The second mode, on the other hand, can be tolled and has a capacity s_T. Consider any period t so that we can save on notation by dropping the time index. Assume both alternatives to be perfect substitutes so that in equilibrium both modes will have equal generalized prices:

$$p_U(N_U) = p_T(N_T), \tag{18}$$

where the total demand for trips from O to D is $N = N_U + N_T$. From de Palma and Lindsey [3] we know that the second-best toll is the fine toll to eliminate queueing on mode T corrected by a term (the flat toll τ_T) to control the number of users on route T :

$$\tau_T = -\frac{|p_N|}{|p_N| + p_N^U} p_N^U N_U, \tag{19}$$

where $p_N^U = \frac{\partial p_U(N_U)}{\partial N_U}$ and $p_N = \frac{\partial p}{\partial N}$.The intuition for the use of a fine toll structure on mode T is easy: a fine toll makes sure the queueing costs of the existing users is transformed into toll revenue without any change in total use of mode T. The average fine toll has to be smaller than in the one link case because this way one can attract users of the other (congested and inefficiently managed) mode U to the better managed mode T.

If $\tau_T = 0$ so that there is only a fine toll on mode T, then

$$N_T = \frac{s_T}{s_T + s_U} N. \tag{20}$$

Since the fraction $\frac{s_T}{s_T+s_U} < 1$, this is always smaller then the demand on a single mode with the same total capacity.

Toll revenues will also be smaller by a fraction $\left(\frac{s_T}{s_T+s_U}\right)^{\frac{2}{1+\eta}}$. For a given capacity s_T, it is thus clear that toll revenues in the single mode case will always exceed toll revenues in the parallel case. We therefore expect the capacity build up over time under the strict financial constraint to give a similar pattern in both cases but capacity in the parallel case will increase more slowly. If the second best toll is charged, we see that demand and thus toll revenues and investments will be even smaller for mode T .

In the two modes case it is difficult to get analytical results and we therefore refer to the simulation results.

6 Numerical simulations

For the numerical simulations in the single mode case we assume the following functional form for the demand function

$$N\left(p\left(t\right)\right) = \left(1+t\right)^{\theta} p\left(t\right)^{-\eta}, \tag{21}$$

where θ and η are positive. The generalized price is equal to the time cost plus the toll

$$p\left(t\right) = \frac{\delta N\left(p\left(t\right)\right)}{2s\left(t\right)} + \tau^{\text{fine}}\left(t\right) + F\left(t\right), \tag{22}$$

with

$$\tau^{\text{fine}}\left(t\right) = \frac{\delta N\left(p\left(t\right)\right)}{2s\left(t\right)}. \tag{23}$$

In the numerical simulations we assume that maintenance costs are zero $\left(a=b=0\right)$.

Welfare $TW\left(t\right)$ is the discounted integral over time of the sum of the consumer surplus $\left(CS\left(t\right)\right)$ and the toll revenues $\left(TR\left(t\right)\right)$ minus the cost of capacity $\left(CC\left(t\right)\right)$ at a given time t. Since the last two are constrained to be equal we end up with

$$TW = \int_0^{\infty} W\left(s\left(t\right)\right) e^{-it} dt \tag{24}$$

where

$$W\left(s\left(t\right)\right) = CS\left(s\left(t\right)\right) = \int_{p(t)}^{\infty} N\left(p'\left(t\right)\right) dp'. \tag{25}$$

We call this the "real" welfare, i.e. the welfare under the strict financial constraint. As a benchmark or reference we use the first best welfare (i.e. the welfare if there is no financial constraint, the toll equals the fine toll and capacity is always at its first best level given in eq(6)):

$$TW^{opt} = \int_0^{\infty} W^{opt}\left(s^{opt}\left(t\right)\right) e^{-it} dt \tag{26}$$

$$W^{opt}\left(s^{opt}\left(t\right)\right) = CS^{opt}\left(t\right) + TR^{opt}\left(t\right) - CC^{opt}\left(t\right). \tag{27}$$

In order to compare the welfare in both cases, we normalize the welfare loss by the cost of capacity in the first best. For some factor Γ_{cs} the first best welfare will become equal to the "real" welfare:

$$CS^{opt}\left(s^{opt}\left(t\right)\right)+TR^{opt}\left(s^{opt}\left(t\right)\right)-\Gamma_{cs}CC^{opt}\left(s^{opt}\left(t\right)\right)=CS\left(s\left(t\right)\right). \quad (28)$$

Using $TR^{opt}\left(s^{opt}\left(t\right)\right)=CC^{opt}\left(s^{opt}\left(t\right)\right)$, the equation can be rewritten as:

$$CS^{opt}\left(t\right)-CS\left(t\right)=CC^{opt}\left(t\right)\left(\Gamma_{cs}-1\right) \quad (29)$$

and

$$\Gamma_{cs}-1=\frac{CS^{opt}\left(t\right)-CS\left(t\right)}{CC^{opt}\left(t\right)}. \quad (30)$$

Define

$$\Gamma_{CS}\equiv i\int_{0}^{\infty}\frac{CS^{opt}\left(t\right)-CS\left(t\right)}{CC^{opt}\left(t\right)}e^{-it}dt+1 \quad (31)$$

We use this index Γ_{CS} as a measure of the welfare loss when the strict financial constraint is imposed and one starts with a suboptimal level of capacity. To understand better the meaning of Γ_{CS} consider first $\Gamma_{CS}=1$, then there is no welfare loss, if $\Gamma_{CS}=2$, this means that if in the first best, capacity costs were doubled, one ends up with the same level of welfare as in the case with strict financial constraint. Or put differently; the fact that one starts with a suboptimal level of capacity and there is a strict self-financing rule, is equivalent to a doubling of the cost of capacity if $\Gamma_{CS}=2$.

For the case with two perfectly substitutable modes, we assume similar expressions for the demand function. The generalized prices for the tolled and untolled mode are respectively:

$$p^{T}\left(N_{T}\right)=\frac{\delta N_{T}}{2s_{T}}+\tau_{T}^{\text{fine}}+\tau_{T}, \quad (32)$$

$$p^{U}\left(N_{U}\right)=\frac{\delta N_{U}}{s_{U}} \quad (33)$$

where

$$\tau_{T}^{\text{fine}}=\frac{\delta N_{T}}{2s_{T}}. \quad (34)$$

The welfare is now the sum of the consumer surplus on both modes, the toll revenues $\left(TR^{T}\right)$ on the tolled mode and the investment cost of the tolled mode (we again neglect maintenance costs):

$$SS=\int_{0}^{N}p\left(n\right)dn-p^{T}\left(N_{T}\right)N_{T}-p^{U}\left(N_{U}\right)N_{U}+TR^{T}-i\kappa s_{T} \quad (35)$$

where the toll revenues are

$$TR^{T}=\tau_{T}N_{T}+\frac{\delta N_{T}^{2}}{2s_{T}}. \quad (36)$$

Substituting this expression in SS :

$$SS = \int_0^N p(n)\, dn - \frac{\delta N_T^2}{2s_T} - \frac{\delta N_U^2}{s_U} - i\kappa s_T \tag{37}$$

The index which gives us a measure of the welfare loss is now defined as

$$\Gamma_{SS} \equiv i \int_0^\infty \frac{SS^{opt}(t) - SS(t)}{CC^{opt}(t)} e^{-it} dt + 1 \tag{38}$$

where $SS(t)$ is given by eq(35) with $SS^{opt}(t) = SS\left(s = s_T^{opt}\right)$ and $CC^{opt}(t) = i\kappa s_T^{opt}$.

We discuss first the case of a single link with a fine toll and a two-part toll. We continue with the case of two parallel modes where one mode is untolled.

6.1 Numerical simulations for the single link and fine toll

In order to make the simulation one needs some assumptions on the values of the parameters. We continue to assume maintenance costs equal to zero: $a = b = 0$. The value of the parameter δ is based on the value found in [1] and is: $\delta = 2.427$. For the costs of capacity we assume $i\kappa = 15.157$, where i is the interest rate and is equal to 5%. The initial capacity s_0 is taken to be a percentage of the first best capacity at $t = 0$. We will take 20%, 50% and 80%. Finally, the time horizon is 50 years.

First we consider the case where demand is time-independent ($\theta = 0$) and inelastic ($\eta = 0$), with other words; demand is fixed and normalized to one.

The evolution of the capacity level (vertical axis) over time (horizontal axis) for different initial capacities is given in the Figure 6.1. The dotted horizontal line is the optimal capacity, the full lines represent capacity extensions when initial capacity is 20, 50 and 80% of the optimal capacity.

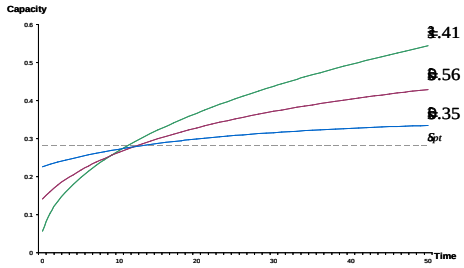

Figure 6.1: The one mode capacity evolution with fine toll for different initial capacities when $\theta = \eta = 0$.

The first-best capacity ($s^{opt} = 0.283$) will be reached more or less after the same period ($t \sim 10$) in the three cases. The steady state levels ($\hat{s}$) are, however very different. For initial capacities of respectively $20\%, 50\%$ and 80% of the optimal

capacity level, the steady states are $\hat{s} = 1.41$, $\hat{s} = 0.56$ and $\hat{s} = 0.35$. These values are consistent with eq(10). For example, for an initial capacity equal to 20% of the first best capacity the steady state will according to eq(10) be equal to; $\hat{s} = \frac{\left(s^{opt}\right)^2}{s_0} = \frac{\left(s^{opt}\right)^2}{0.2s^{opt}} = 5s^{opt} = 1.41$.
We see that starting with a very low capacity (20% of the optimal capacity) generates more revenues (higher congestion and lower interest payments) and the first best level of capacity will therefore be reached more quickly. The low cost of the initial capacity will, however, lead to a larger toll revenue surplus when optimal capacity is reached which generates a larger overinvestment in the long run.
Values of Γ_{CS} for different combinations of growth (θ) and price elasticity (η) are provided in Table 1:

$(\boldsymbol{\theta}, \boldsymbol{\eta})$	$(\mathbf{0}, \mathbf{0})$	$(\mathbf{0.1}, \mathbf{0})$	$(\mathbf{0.2}, \mathbf{0})$	$(\mathbf{0.5}, \mathbf{0})$
$\mathbf{0.2}\boldsymbol{s}^{opt}$	1.46	1.61	1.75	2.17
$\mathbf{0.5}\boldsymbol{s}^{opt}$	1.18	1.3	1.43	1.81
$\mathbf{0.8}\boldsymbol{s}^{opt}$	1.05	1.14	1.25	1.61
$(\boldsymbol{\theta}, \boldsymbol{\eta})$	$(\mathbf{0}, \mathbf{0.5})$	$(\mathbf{0}, \mathbf{1})$	$(\mathbf{0}, \mathbf{2})$	
$\mathbf{0.2}\boldsymbol{s}^{opt}$	1.42	1.39	1.33	
$\mathbf{0.5}\boldsymbol{s}^{opt}$	1.18	1.17	1.16	
$\mathbf{0.8}\boldsymbol{s}^{opt}$	1.05	1.05	1.05	
$(\boldsymbol{\theta}, \boldsymbol{\eta})$	$(\mathbf{0.2}, \mathbf{0.5})$	$(\mathbf{0.2}, \mathbf{1})$	$(\mathbf{0.5}, \mathbf{0.5})$	$(\mathbf{0.5}, \mathbf{1})$
$\mathbf{0.2}\boldsymbol{s}^{opt}$	1.64	1.56	1.93	1.79
$\mathbf{0.5}\boldsymbol{s}^{opt}$	1.39	1.35	1.68	1.59
$\mathbf{0.8}\boldsymbol{s}^{opt}$	1.23	1.22	1.53	1.46

Table1: Relative efficiency losses of a strict financing constraint for different initial capacities, different growth rates of demand (θ) and different price elasticities (η)

From the table we see that starting with a capacity level which is only 20% of the optimal capacity level and imposing a strict self-financing constraint is equivalent with increasing the capacity costs by $40\% - 120\%$. Starting with half of the optimal level reduces this to $20\% - 80\%$. Larger initial capacity (80%) will create only small welfare losses (around $5 - 60\%$). Larger growth in demand will increase the welfare loss. A higher price elasticity, however, will correspond to slightly lower values for Γ_{cs}. This can be expected as the higher price elasticity means better substitutes in case of sub-optimal capacity levels.
The lower the initial capacity, the larger will be the excessive capacity in the long run. There are, however, no general statements to be made about the moment that capacity reaches s^{opt} : defining t_p as the time where the capacity reaches the first best level starting from an initial capacity $s_0 = p * s^{opt}$, we have; if $\theta = 0.2, \eta = 0.2$: then $t_{0.2} = 17$, while $t_{0.8} = 21$; the lower the initial capacity the quicker the first best level is reached. For the same growth parameter but an elasticity of $\eta = 4$, the opposite is true $t_{0.2} = 42$, while $t_{0.8} = 41$.

6.2 Simulation results for the two-part toll

In the following simulations the fixed part of the toll, $F(t)$ is different from zero and is given by eq(15). First we consider the case without growth ($\theta = 0$), as price elasticity we take $\eta = 0.5$ and the initial capacity is 20% of the first best capacity: $s_0 = 0.2 * s^{opt}$, where $s^{opt} = 0.096$. The third best capacity as defined in eq(17) is for these parameter values equal to $s^* = 0.1$. We see that this is higher than the first best capacity.

We first assume that we have fine tolling until the third best capacity is reached followed by a fixed toll that returns all surplus toll revenues under the form of a flat subsidy per trip ($\alpha = 0$ in eq(15)). In this case the capacity will reach the third best level at $t = 18$. The value of Γ_{CS} is now 1.32 which is only a very slight improvement compared to the case where we had fine tolling and we allowed the capacity to continue to increase above the optimal level ($\Gamma_{CS} = 1.42$). Things can be improved by allowing α to be different from zero. The best performing α turns out to be $\alpha = 0.07$. For this value of α, Γ_{CS} is equal to 1.3. As can be seen in Figure 6.2 (where the dotted line corresponds to $\alpha = 0$, and the dashed line to $\alpha = 0.07$) the extra fixed toll can be used to reach the third best capacity at an earlier point in time ($t = 13$).

As a second illustration we take the same parameter values as above but now $\eta = 0.2$ instead of 0.5. Again we compare the case when $\alpha = 0$ (no fixed part) with the best performing α. When $\alpha = 0$; $\Gamma_{CS} = 1.32$ (compared to 1.44 in the case with only fine tolling). After 15 periods the third best capacity level is reached (dotted line in Figure 6.3). The best performing $\alpha = 0.19$ yields a $\Gamma_{CS} = 1.26$ which is a 10% improvement compared to $\alpha = 0$, the evolution of the capacity in this case is given by the dashed line in Figure 6.3 (where again the dotted line corresponds to $\alpha = 0$, and the full line to $\alpha = 0.19$).

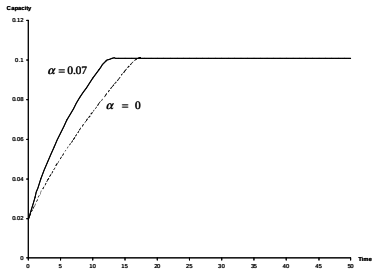

Figure 6.2: The one mode capacity evolution with two part toll for different values of α when $\theta = 0$ and $\eta = 0.5$.

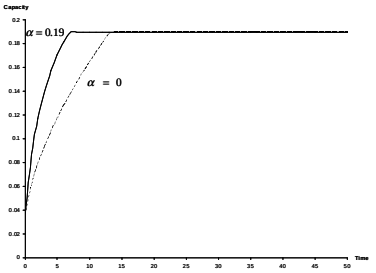

Figure 6.3: The one mode capacity evolution with two part toll for different values of α when $\theta = 0$ and $\eta = 0.2$.

Concluding the numerical simulations of the two-part toll we see that when demand is price elastic, a two-part toll is a limited instrument to reduce the

welfare losses associated to the strict self-financing constraint. The main reason is that the fixed toll (or subsidy) can only transform the efficiency loss associated to an excessive investment into an inefficient pricing regime as user prices will be different from marginal social costs.

6.3 Simulation results for two modes

We take the same parameter values for δ, κ and i as previously. We restrict ourselves to the case without growth ($\theta = 0$) and $\eta = 0.5$ and where the initial capacity of the tolled mode is half of the optimal level.

The results will depend on the characteristics of the untolled alternative, more precisely on the relative importance of the untolled capacity compared to the tolled initial capacity. In Table 2 we summarize the results for different capacities of the untolled mode and for the case where only the fine toll can be used on the tolled link (fixed part=0):

s_U	s_T^{opt}	s_0	Γ_{SS}
0.0001	0.0966	0.0483	1.18
0.01	0.0956	0.0481	1.67
0.04	0.086	0.043	1.58
0.07	0.0713	0.0365	1.31

Table 2: Relative efficiency losses of a strict self-financing constraint in the two mode case for different initial capacities of the tolled mode and different capacities of the untolled mode when the toll is equal to the fine toll.

In the second column we report the optimal capacity of the tolled mode given the capacity of the untolled mode and given that the toll equals the fine toll. Column three gives the different initial capacities of the tolled mode and the last column gives the values of the efficiency parameter Γ_{SS}. We see that when capacity of the untolled mode increases, the welfare loss increases until a certain point where it decreases again.

If instead of the fine toll one charges the second best toll that takes into account the existence of an untolled alternative (see eq(19)) we get the results reported in Table 3:

s_U	s_T^{opt}	τ_T	Γ_{SS}
0.01	0.094	-1.084	1.15
0.04	0.082	-2.207	1.12
0.07	0.068	-2.67	1.28

Table 3: Relative efficiency losses of a strict self-financing constraint in the two mode case for different initial capacities of the tolled mode and different capacities of the untolled mode for the second best toll.

Here the second column corresponds to the optimal capacity for the tolled mode given that this mode applies the second best tolling. From Table 2 and 3 we see that the optimal capacities are slightly smaller and the welfare losses are smaller with second best tolling than with fine tolling only. Simple fine tolling with an untolled alternative has the following drawbacks in comparison to the second best toll. First, the second best toll can be set at a lower value in order to attract more users to the tolled alternative that has better congestion management. Moreover, the second best toll produces less toll revenues so that there are less excessive investments.

7 Conclusions

This paper has analyzed analytically and numerically the effect of a strict self-financing constraint on the development of a network of the bottleneck type consisting of one link or one link plus an untolled alternative. Tolls are either fine tolls or a two-part toll consisting of a fine toll plus a fixed part. The strict self-financing constraint forces the operator to spend the surplus revenues on new capacities. We find that the ratio between the initial infrastructure capacity and the optimal capacity is the dominant factor to explain the inefficiency associated to the strict self-financing constraint. There are two reasons for this. First starting with a very low capacity makes it more difficult to reach the optimal capacity. Second, a very low initial capacity means that the toll surplus will be larger and that investments continue until capacity is really excessive. The use of a two-part toll where a fixed part can help to build capacity more quickly and to return excessive toll revenues once optimal capacity is reached can in theory reduce these inefficiencies. Numerical simulations show that the advantage of this more complex tolling regime is limited and that the additional cost of the strict self-financing constraint can be reduced of the order of 20 to 100% of the cost of capacity. When there is an untolled alternative in place, the strict self-financing constraint is a handicap when one starts with a too low capacity as the optimal fine toll generates less revenues and it therefore takes more time to reach the optimal capacity. On the other hand it limits excessive capacities as revenues are more limited.

Our paper is limited to the analysis of simple one or two link cases. In the real world, agencies have often the responsibility of a complex network. If the complex network is an aggregate of one link problems with identical structure our results continue to apply. Things would be different if part of the network is close to optimal capacity while another part requires large investments. In this case, the pooling of revenues over links relaxes the strict self-financing constraint. A second limitation of our analysis is that the strict self-financing constraint is exogenous. The strict self-financing constraint is the result of a principal agent problem between voters, politicians and agencies (see [2]). One of the results of our paper, the efficiency loss of this constraint, is then an important input for the principal agency game.

References

[1] Arnott R., de Palma A. and Lindsey R. (1993),"A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand", The American Review, Vol 83 (1), pg 161-179.

[2] Besley T. (2006), "Principled agents? the Political Economy of Good Government", The Lindahl Lectures, Oxford: Oxford University Press.

[3] de Palma A. and Lindsey R. (2000), "Private toll roads: Competition under various ownership regimes", The Annals of Regional Science, Vol 34, pg 13-35.

[4] Doll C. and Link H., (2007) "The German HGV Motorway Toll", Chapter 10 in de Palma, Proost and Lindsey (Eds): "Investment and the use of tax and toll revenues in the transport sector.", Elsevier Science 5.

[5] Newbery D.C. (1988), Road damage externalities and road user charges, *Econometrica*, 56 (2).

[6] Mohring H. and Harwitz,M., (1962) "Highway Benefits, an analytical framework", Evanston, IL: Northwestern University Press.

[7] Proost S., De Borger B. and Koskenoia P., (2007) "Public Finance Aspects of transport charging and investments", Chapter 3 in de Palma, Proost and Lindsey (Eds): "Investment and the use of tax and toll revenues in the transport sector.", Elsevier Science 5.

[8] Raux C., Mercier A. and Souche S., (2007) "French Multi-modal Transport Funds: Issues of Cross-financing and Pricing.", Chapter 11 in de Palma, Proost and Lindsey (Eds): "Investment and the use of tax and toll revenues in the transport sector.", Elsevier Science 5.

[9] Verhoef, E. and Mohring, H., (September 2007) "Self-Financing Roads" . Tinbergen Institute Discussion Paper No. 2007-068/3