Integrating Congestion Pricing, Transit Subsidies and Mode Choice Models: Beyond the Downs-Tompson Hypothesis[*]

Leonardo J. Basso[†] Sergio R. Jara-Díaz[‡]

First Draft: December 2010
This Revision: March 2011

Abstract

We propose a model to study analytically the problem of the optimal pricing and design of transport services in a bimodal context. All the usual properties of the car congestion process and the transit design problem are simultaneously introduced, consumers can choose between the two modes based on the full price they perceive, and the planner maximizes social welfare optimizing three variables: the congestion toll, the transit fare (and hence the level of subsidies), and transit frequency. We obtain five main results: introducing behavior makes the optimal car-transit split generally different from the total cost minimizing one; there is no optimal congestion price independent of transit price, as it is the difference what matters; the optimal money price difference has an optimal transit frequency attached, and together they yield the optimal modal split; once the optimal modal split is found, congestion tolls and transit subsidies and fares from stand-alone formulations –where each mode is price independently– are consistent with the optimal money price difference; it is possible to establish prices such that the total subsidy for transit equals the total (congestion) tolls collected, i.e. self financing of the transport sector is feasible; and, finally, investment in car infrastructure induces an increase in generalized cost for all public transport users.

[*] *Acknowledgments*: This research was partially funded by FONDECYT-Chile, Grants 1090187 and 1080140 and by the Millennium Institute 'Complex Engineering Systems', Grants ICM: P-05-004-F and CONICYT: FB016. We are indebted to Hugo Silva, the Editor and two anonymous referees for helpful comments and suggestions.

[†] *Civil Engineering Department, Universidad de Chile; lbasso@ing.uchile.cl*

[‡] *Civil Engineering Department, Universidad de Chile; jaradiaz@ing.uchile.cl*

1. Introduction

Congestion is, undoubtedly, the externality caused by urban transportation that has attracted the largest share of attention from economists and engineers. In addition to the obvious idea of increasing capacity, the two most popular ways to deal with congestion that have been suggested in the literature are congestion pricing and giving priority to public transportation. Congestion pricing has been analyzed in a very large number of settings and a large number of articles, books and reviews have been written on the topic (for reviews see Small and Verhoef, 2007, and Tsekeris and Voβ, 2008). Two particular results can be stressed: first, that if congestion pricing is implemented, travelers surplus will decrease since the full price consumers pay (time costs plus the tax) is larger than the time costs they pay without congestion pricing; however, total social welfare would be increased because tax collection dominates the travelers' surplus reduction, making revenue recycling an important issue if political support is to be raised. And, second, that in most cases the change in surplus is worse for poorer people, making congestion pricing a regressive measure prior to recycling (Arnott et al. 1994; Hau, 1988). It is because of this that there is a quite sizeable literature on the best way to recycle congestion pricing revenues in order to make everyone better off, that is, to establish a Pareto improving policy (see e.g. Kockelman & Kalmanje, 2005, for a recent article).

On the other hand, many authors have studied the optimal design of scheduled public transport services (Mohring, 1972; Jansson, 1984; Jara-Díaz and Gschwender, 2003b, 2009), seeking frequencies, vehicle sizes, spacing of bus stops and spatial structure of lines that minimize total costs. Although depending on the specific setting, there are two main results here. First, that just as in the case of cars, one needs to take into consideration the resources supplied by all parts: operators (energy, crew, maintenance, administration, infrastructure, rolling stock and so on) and users (waiting, access and in-vehicle times), something that in the case of automobiles is done naturally since both roles are played by the same agent. And, second, that when one does consider users' resources, the efficient cost minimizing service requires subsidies, since the sum of operators' and users' costs yields a total cost that grows less than proportionally with the demand, implying scale economies. This is sometimes known as the Mohring effect and the direct implication is that transit subsidies would be optimal (for review and basic theory see Jara-Díaz and Gschwender, 2003a; Jara-Díaz, 2007).

Now, most of the studies mentioned above consider systems where each mode is isolated. Yet, as it is evident, in most cities people have a choice between using a car or public transportation and, in making a choice, people would consider not only direct monetary costs but also travel times, comfort and so on; therefore, demands for car and buses have cross-elasticities not only with respect to price, but also with respect to quality of service. Therefore, congestion tolls, transit subsidies, prices and service variables –such as frequency– are closely interrelated. Moreover, their optimal values and the modal split should strongly depend on the way mode choice occurs. Yet, in the literature these issues are rarely studied together, which raises the obvious question of whether results coming from a single mode analysis hold in a bi-modal system. There are indeed some papers that have looked at two-mode systems from this perspective before, but many of them focus on numerical results rather than on theoretical ones (Mohring, 1972; Jackson, 1975; Small, 1983; Viton, 1983; Huang, 2000; Proost and Van Dender, 2008; Basso et al., 2011; Basso and Silva, 2011); or focus on the Pareto improving aspect and therefore only allow for changes in transit fares but not service levels (Nie and Liu, 2010). They do not, therefore, help to understand in a simple, tractable and analytical way how optimal congestion tolls and transit design interact, and what is the role of the way people choose between modes, i.e. the mode choice model.

The closest paper to what we intend to do here is the diagrammatic analysis of modal split by Mogridge (1997), who explains the Downs-Thomson hypothesis, namely, that at equilibrium modal generalized costs have to be equal. That analysis captures car congestion (car costs increases with car use) and the fact that transit frequencies are adjusted according to ridership, causing transit costs reductions as the number of users increase. It implicitly considers a rather limited mode-split model, which does not consider elements beyond time and cost nor heterogeneous commuters, while ignoring the effects of ridership on transit design variables and transit service levels. Tolls and subsidies are also not explicitly treated. The Downs- Thompson hypothesis is used by Mogridge to rescue and present what is known as the Dawns-Thomson paradox, which shows that an investment in urban roads capacity would provoke a vicious chain of consequences: a reduction of public transport users, an increase in public transport costs, a further increase in car users leading to an increase in perceived travel costs, such that the final result finds every user worse off.

What we do in this paper is to propose a simple tractable way to study analytically the problem of the optimal pricing and design of transport services in a context where consumers can choose between the two modes based on the full price they perceive, and the planner maximizes social welfare optimizing three variables: the congestion toll, the transit fare (and hence the level of subsidies), and transit frequency. We solve the model analytically, while providing intuition through the use of graphs. We do not commit to a specific mode-split model but show the impacts of using different models. Particularly important is the analysis of the optimal modal split as compared to that which minimizes costs. We use the model to discuss usual transport policies such as self-financing rules and touch on the effect of capacity expansions.

2. Model Formulation

2.1 Stand-alone models and the Downs-Thomson equilibrium concept

Let us consider users in an urban area that choose between two modes: transit (T) and auto (A). We analyze a case where total demand for trips is perfectly inelastic, that is, the total number of commuters is not affected by prices or quality of service. This assumption –which amounts to assuming away trip generation– is quite reasonable for commuting to work, and enables a better comparison with the Mogridge-Downs-Thomson analysis. Thus, let Y be the total demand for travel, Y_i the number of users of mode i, and g_i the generalized cost (perceived by users) of mode i, which depend on price and time as follows:

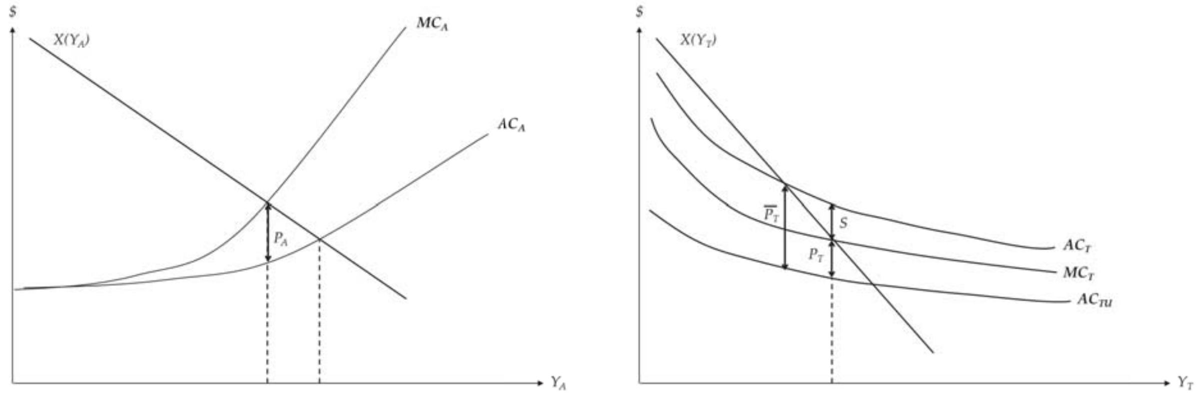
$$g_A \equiv P_A + AC_A = P_A + OC_A + \alpha\, t_A(Y_A) \tag{1}$$

$$g_T \equiv P_T + AC_{TU} = P_T + \alpha \left(T_M + Y_T \frac{\mu}{f} \right) + \alpha \frac{\beta}{2f} \tag{2}$$

$$Y = Y_A + Y_T \tag{3}$$

where P_i is the money price charged for mode i, AC_A stand for average cost for automobile drivers and AC_{TU} is average cost for transit users; these concepts will play a key role in the model. Auto average cost includes the operating cost of auto users OC_A , and the monetary cost of travel time captured by the product of the value of saving in-vehicle travel time α and travel time $t_A(Y_A)$. Congestion among cars is captured by the assumptions that $t'_A > 0$ and $t''_A > 0$. On the other hand, the average cost for transit users includes the time they spend in the vehicle and walking and waiting times. The time in the vehicle is the addition of time the vehicle is in motion T_M, and the time the bus is stopped for boarding and alighting, given by $Y_T \frac{\mu}{f}$, where μ is boarding/alighting time per passenger, and f is transit frequency. Assuming homogeneity of arrivals to the bus stop, average waiting time is given by $1/2f$, which when multiplied by the value of saving waiting time $\alpha\,\beta$ gives a monetary cost. We assume that walking times are constant and normalize them to zero without further loss of generality. As in Mogridge (1997) no congestion between buses or across modes is considered.

These concepts and equations (1) to (3) are enough to present the idea behind congestion pricing and transit subsidies in stand alone models, as well as to explain the Downs-Thomson concept of equilibrium, named by Mogridge (1997) the *Downs-Thomson hypothesis*. Let us start with congestion pricing and suppose first that $P_a = 0$. Drivers total cost is given by $C_A = Y_A\, AC_A = Y_A\,(OC_A + \alpha\, t_A(Y_A))$ while an individual driver's cost is $AC_A = OC_A + \alpha\, t_A(Y_A)$. If a new driver enters the flow, she will experience



(a) (b)

Figure 1: Optimal congestion pricing and transit subsidy and fare

average cost but will impose $Y_A \alpha\, t'_A(Y_A)$ extra costs on the other drivers. This externality problem can be solved following the Pigouvian tradition of charging this *external marginal cost*, making $P_A = Y_A \alpha\, t'_A(Y_A)$. Note that marginal total cost is given by $MC_A = OC_A + \alpha\, t_A(Y_A) + Y_A \alpha\, t'_A(Y_A)$ and, therefore, $P_A(Y_A) = AC_A(Y_A) - MC_A(Y_A)$. The only remaining issue to have a well defined optimal congestion toll is the efficient level of traffic which is given by the point where demand $x(Y_A)$ equals marginal cost, thus leading to an optimal level of traffic that is smaller than the original one. A graphical exposition is shown in Figure 1a.

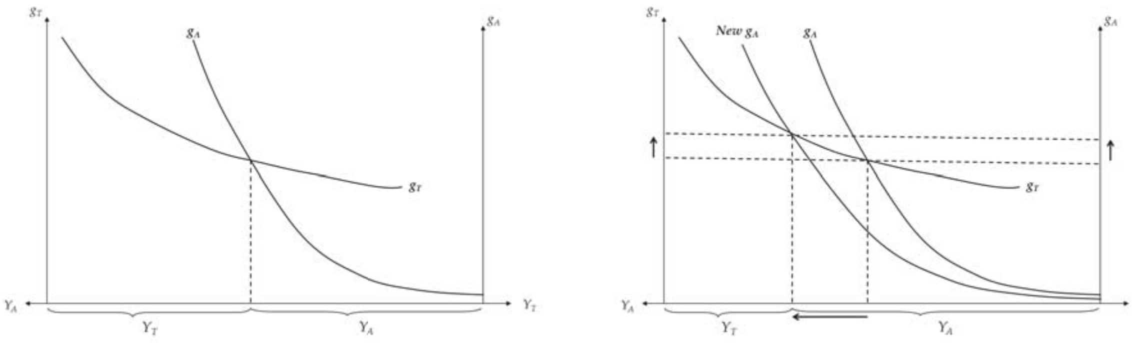
Let us now consider a simple analysis of efficient pricing in transit systems (Jansson, 1984). Just as in the case of private transport, one needs to consider both operational costs and users costs. The difference here is that these are paid by different agents: transit operators on one hand and commuters on the other. From (2) it is easy to see that individual commuter cost, AC_{TU}, decreases when frequency is increased; but since one expects that $\partial f / \partial Y_T > 0$ –i.e. frequency increases with ridership– it happens that as the number of transit users increase, the cost for each transit user decreases: average user cost is a decreasing function of Y_T, thus displaying economies of scale. Operators costs, on the other hand, include energy, labor, maintenance, administration, vehicles and in some cases infrastructure. It is commonly accepted, as a result of empirical studies, that operators costs present either increasing or constant returns to scale, that is, that operators average costs are either decreasing with Y_T or flat. If we then define the total cost of the transit system as the sum of users' and operators' costs, it follows that the average cost of the transit system AC_T is decreasing in Y_T and, furthermore, the marginal total cost of transit, MC_T goes below average cost. Both curves are shown in Figure 1b, together with the average cost of transit users AC_{TU}. Note that, by tdefinition, the vertical differen between AC_{TU} and AC_T is operators' average cost.

Then, if transit inverse demand is given by $x(Y_T)$, the efficient generalized price of transit should equal total marginal cost, that is $x(Y_T^*) = MC(Y_T^*)$, but since part of the marginal total cost are user costs, the optimal toll to be charged to users is $P_T^* = MC_T - AC_{TU}$. That optimal toll, however, is not enough to cover the operators average cost, from where it flows the need of an optimal subsidy given by $S^* = AC_T - MC_T$. Figure 1b shows both the optimal toll and the optimal subsidy. If tolls are not feasible, then self-financing requires setting a price $\overline{P}_T$ equal to operators average cost (see Jara-Díaz and Gschwender, 2009).

From both panels of Figure 1, one can see that if optimal congestion tolls or optimum subsides are implemented, the number of users change. Yet these stand-alone models do not address the question of where do these users go or come from. And this is indeed relevant. Suppose for example, that congestion tolls are implemented. That would decrease the number of car users as explained but, in an urban context, it is very likely that those commuters moved to a differetn transport mode, in particular public transportation. And according to the transit model, this would imply an increase in frequency, which would diminish its generalized cost. The change of price of a substitute mode will affect car demand but now in a different way: it will shift the demand curve inwards, affecting again the number of car users. So, it is clear that stand alone models are not a good tool if one wants to deal with modal substitution, because the way the equilibrum modal split is achieved is not clear. Mogridge (1997) proposes a way to diagramatically represent the problem of modal split arguing that "at equilibrium, the generalised costs of car and collective transport will be equal", which he calls the *Downs-Thomson hypothesis.* The intuition is that if generalized costs differ then users would move towards the mode with smaller generalized cost, indicating that the original modal split was not an equilibrium. Graphically, generalized costs and the equilibrium modal split (Y_A, Y_T) would look like in Figure 2a.

Although in this paper we focus mainly on the Downs-Thomson equilibrium concept as a useful reference, for completeness we present the Downs-Thompson paradox that follows this equilibrium concept. The paradox occurs because while a road capacity expansion implies that for a given number of car users, generalized costs diminish, this reduction will attract new car users from the public transport system, which would increase transit generalized cost through depressed frequencies. The new equilibrium modal split –the one that equalizes the new generalized costs– would occur at larger generalized costs than before. The process is shown in Figure 2b and is, according to Mogridge (1997), the self-defeating nature of road capacity policies.[1]

[1] The empirical literature on induced traffic seems to ratify the intuition behind the paradox in that capacity



(a) (b)

Figure 2: The Downs-Thomson equilibrium and paradox.

The intuition behind the Downs-Thomson equilibrium concept seems certainly compelling, and one could think to use it in order to analyze the interaction between congestion tolls and transit subsidies, but there are two important limitations to note. First, the transit design (i.e. frequency and vehicle capacity) is not analytically incorporated, so the shape of the transit generalized costs assumes that frequency increases with ridership but ignores the actual relation between optimal transit design and how it affects consumers and operators. Second, the mode choice implicit behind the analysis is quite limited, being of the all-or-nothing type: if the generalized cost of one mode is smaller than the other, then everyone will choose that mode and vice-versa. Yet, one can easily picture the following case: when generalized costs are close, some users will choose car while other will use transit because they have different valuations of mode attributes other than price and time, such as environmental impact or social status. These attributes, together, may represent some form of intrinsic car attractiveness, which differ across the population. This heterogeneity will make the number of users of one mode diminish smoothly as its generalized cost grows farther away from that of the other mode (something that popular mode choice models do capture, as the binomial Logit). What we show in what follows, is that moving away from the all-or-nothing choice towards more realistic mode choice models has sizeable impacts in the characteristics of the optimal modal split, that is, the one that maximizes welfare as opposed to minimize costs.

2.2 Optimal tolls and subsidies in the two-modes system

We start by allowing a more general mode choice model at the outset, as follows:

expansions seem to have worsen the congestion problem instead of solving it. One of the latest empirical efforts is Duranton and Turner (2011) who analzyed US cities, finding that the elasticity of vehicle-kilometers driven to kilometers of lane is on average larger than one in urban settings. For other articles that study induced traffic phenomena see references therein.

$$Y_T(g_A, g_T) = Y \cdot H\,(g_A - g_T) \tag{4}$$

$$Y_A(g_A, g_T) = Y - Y_T$$

Function H is the fraction of users choosing transit as a function of the difference in generalized costs. The Downs Thomson equilibrium would have a function H which is equal to 1 if $g_A > g_T$, equal to 0 if $g_A < g_T$ and equal to $(g_A - g_T + Y_T)\,/Y$ if $g_A = g_T$. We instead assume that H is increasing, concave and twice differentiable. Note that this demand formulation is quite general and it encompasses, for example, the binomial Logit model.

What is of interest now that we have a demand model that captures substitution, is to optimize tolls –for both transit and cars– and public transport design, i.e. frequency and vehicle size. In order to do this, we first need to establish an objective function; the welfare function we consider will be the sum of of consumers' surplus, money collected by government and transit agency profit, i.e.

$$W(P_A, P_T, f) = -\int_{(g_A^0, g_T^0)}^{(g_A, g_T)} \sum_i Y_i(g_A, g_T)dg_i + P_A Y_A(P_A, P_T, f) + P_T Y_T(P_A, P_T, f) - c\,f \tag{5}$$

where we have recognized that modal demands depend on (P_A, P_T, f) through generalized prices. The first term on the-right hand side represents consumer surplus. Note that, despite the fact that total demand is perfectly inelastic, the fact that there is substition and interdependency makes this term relevant; moreover, since when one generalized price changes that affect the demand of the other mode, what needs to be calculated is the generalized Marshallian Consumer Surplus given by the line integral above.[2] The second and third terms on the right-hand side are revenues from car tolls and transit fares. Finally, we have simplified transit costs to a constant marginal cost per movement. The optimization problem, then consists on maximizing (5) subject to total transit capacity large enough to carry all transit passengers, that is:

$$\underset{P_A, P_T, f, K}{Max} \; -\int_{(g_A^0, g_T^0)}^{(g_A, g_T)} \sum_i Y_i(g_A, g_T)dg_i + P_A Y_A(P_A, P_T, f) + P_T Y_T(P_A, P_T, f) - c\,f \tag{6}$$

$$s.t \quad K\,f \geq Y_T(P_A, P_T, f)$$

[2] This line integral simplifies to the sum of two one-dimensional integrals when demands are independent. The expression is a generalization of the well-known consumer surplus measure that captures the change of the area under the demand curve. See Mas-Collel et al. (1995, p.80) for a general discussion and Jara-Díaz (2007, p. 91) for an application to mode choice changes.

where K is vehicle capacity. In the absence of crowding costs, it will never be optimal to choose a vehicle size larger than what is strictley needed, and therefore it holds that at the optimum $K = Y_T/f$. We therefore need only to worry about f and then obtain K as result. The first-order conditions of problem (6) are calculated in the Appendix, and lead to:

$$\frac{\partial W}{\partial P_A} = \frac{\partial W}{\partial P_T} = \left(Y_T^* \frac{\alpha\,\mu}{f^*} - Y_A^* \alpha t'_A(Y_A^*) + P_A^* - P_T^* \right) \frac{\partial Y_A}{\partial P_A} = 0 \tag{7}$$

$$\frac{\partial W}{\partial f} = \frac{\partial Y_A}{\partial f} \left(Y_T^* \frac{\alpha\,\mu}{f^*} - Y_A^* \alpha t'_A(Y_A^*) + P_A^* - P_T^* \right) + Y_T^* \frac{\alpha\,\beta}{2f^2} + Y_T^{*2} \frac{\alpha\,\mu}{f^2} - c = 0 \tag{8}$$

It is not suprising that the first-order conditions of P_A and P_T lead to the same equation: as the model features two modes and perfectly inelastic total number of trips, demand depends on the difference of generalized costs, which makes demand dependent on the difference of prices. Replacing (7) in (8) one obtains:

$$f^* = \sqrt{\frac{\alpha Y_T^*}{c} \left(\frac{\beta}{2} + Y_T^* \mu \right)} \tag{9}$$

while simplifying (7) we get:

$$P_A^* - P_T^* = \triangle P^* = \frac{\alpha\,\mu Y_T^*}{f^*} - \alpha Y_A^* t'_A(Y_A^*) \tag{10}$$

Note that these relations do not provide explicit expressions for f^* and $\triangle P^*$ since the demand function H has not been given a specific form yet and the modal split depends on prices and frequency, i.e. $Y_T^* = Y \cdot H\left(g_A\left(\triangle P^*, f^*\right) - g_T\left(\triangle P^*, f^*\right)\right)$. Nevertheless, these intermediate results are revealing: equation (9) replicates the *square root rule* obtained in the public transport literature after Mohring (1972) and Jansson (1984), which leads to the presence of economies of scale when both transit users and transit operators costs are taken into account. To see this consider for a second a given transit demand in order to solve parametrically on Y_T for the frequency that minimizes the sum of operators and users costs, that is:

$$\underbrace{\left[\alpha \left(T_C + Y_T \frac{\mu}{f} \right) + \alpha \frac{\beta}{2f} \right] Y_T}_{users\ costs} + \underbrace{c\,f}_{operators\ costs} \tag{11}$$

The result is indeed Equation (9) which fulfils $\partial f / \partial Y_T > 0$. Furthermore, replacing f^* back in (11) it is a matter of algebra to calculate the average cost of the transit system AC_T, the marginal total cost of transit, MC_T and the average cost of transit users AC_{TU}, all as functions Y_T. And they indeed behave as shown in Figure 1b: the square root formula leads to the economies of scale to which Downs-Thomson and Modgridge refer to, something

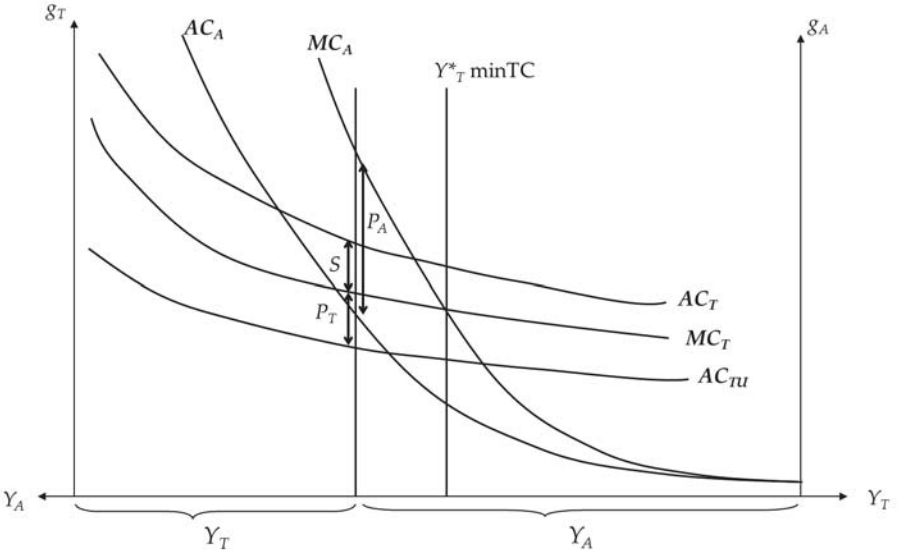


Figure 3: Average and marginal costs, prices and subsidies for auto and transit

also known as the Mohring effect. What it is more important, is that straighforward calculations allow us to obtain that $MC_T - AC_{TU} = \frac{\alpha\,\mu Y_T^*}{f^*}$, which is the first term on the right hand side of (10), while the second term on the right-hand side of (10) happens to be the difference between marginal and average costs for auto for a given demand or, in other words, the Pigouvian toll.

Hence, overall, equation (10) indicates that the optimal price difference $\triangle P^*$ replicates what one would obtain if car congestion pricing and public transport were priced independently by minimizing social cost including users' time as a resource, i.e.:

$$\triangle P^* = \frac{\alpha\,\mu Y_T^*}{f^*} - \alpha Y_A^* t_A'(Y_A^*) \equiv (MC_T - AC_{TU}) + (MC_A - AC_A) \tag{12}$$

Note however, that equation (12) depends on Y_i^*, whose calculation, up to this point has not occurred. In the cases of stand-alone models, the optimal number of users is determined by the interaction of optimal tolls and demand functions. But in this case it is not obvious how such an approach would work and, in fact, what is needed is the choice model. We devote the next section to show how the (optimal) modal split (Y_T^*, Y_A^*)depends on the mode choice model but before, we present a figure wich we hope will help clarify matters.

In Figure 3 we have chosen to represent optimal prices for a level of transit users that is to the left of the total cost minimizing modal split $Y_T^* minTC$, where marginal costs for transit and auto intersect. As discussed earlier, all pairs P_A and P_T calculated as the difference between marginal and average costs for a given transit level fulfill optimal condition (10), but maximum welfare will take place at a modal split that will depend on the properties of the H function in equation (2). In Figure 3 one can also find the point that Modgridge (1997) would describe as an equilibrium. In the absence of transit subsidies, transit would be priced at average (operators) cost; thus, the cost that would be perceived by transit users would be their time costs (AC_{TU}) plus the transit fare, i.e. total average cost. If there is no congestion pricing either, then automobile users would perceive AC_A, making the *equilibrium point á-la-Mogridge* the intersection between AC_A and AC_T. If on the other hand, optimal pricing policies (congestion pricing and transit subsidies) are in place, then users would perceive marginal costs, and the *equilibrium point á-la-Mogridge* would be the the intersection between MC_A and MC_T. This modal split, although minimizing costs, is not necessarily optimal from a welfare maximizing point of view: the latter will depend on the properties of the mode choice model H, while Mogridge's equilibrium points are those for a particular H function, identified earlier. Thus, in principle, there is no clear intuition on whether the optimal modal split (OMS) Y_T^* would be to the left or to the right of the total cost minimizing modal split $Y_T^* minTC$, something that is most important for good policy design, as welfare maximization is what planners usually seek.

3. The optimal modal split

Let us first find some general properties of the welfare maximizing level of auto and transit users. From equations (2), (3) and (4) we have that $Y_T = Y \cdot H\left(g_A - g_T\right) = Y \cdot H\left(P_A - P_T + AvC_A - AvC_U\right)$. Replacing equation (12) for $\triangle P^*$, we obtain:

$$Y_T^* = Y \cdot H\left(g_A\left(\triangle P^*, f^*\right) - g_T\left(\triangle P^*, f^*\right)\right) = Y \cdot H\left(MC_A\left(Y - Y_T^*\right) - MC_T\left(Y_T^*\right)\right) \qquad (13)$$

As $H' > 0$ and $\triangle MC$ decreases with Y_T, equation (13) has a single solution for Y_T^* if $H(0) > 0$. Note that, in general, the optimal modal split (OMS) given by Y_T^* will be different from the cost minimizing one, which is given by the point where $\triangle MC = 0$ and that we have called $Y_T^* minTC$. Obviously, it will be also different from the modal split that equates average costs. And since, it is obvious from (13) that what drives the OMS is H, we examine now some particular mode choice functions.

3.1 Deterministic linear utility mode choice

Let us first assume that choice is commanded by a deterministic linear utility where intrinsic mode attractiveness is considered. Utilities from using auto and transit are given

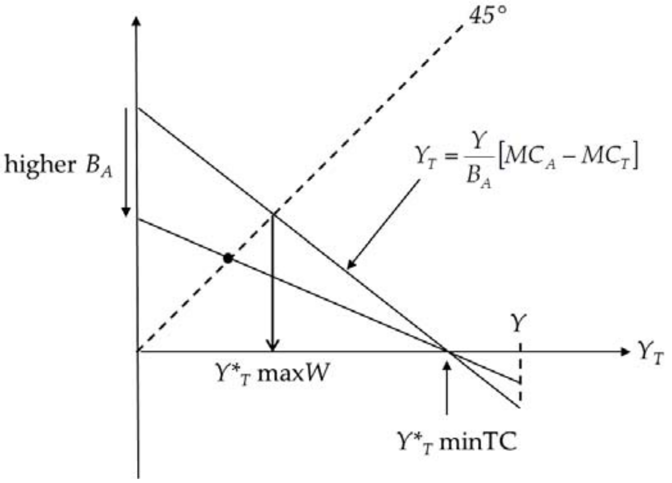


Figure 4: Optimal transit demand for deterministic linear utility.

respectively by:

$$U_A = \theta\, B_A - g_A \qquad U_T = -g_T$$

where, $B_A > 0$ represents positive atributes that the car has over transit and that are not captured in g (for example status, reliablity or flexibility) and θ is the valuation that a consumer places on those attributes. The population has differences in taste, which are captured by the fact that θ takes different values for different people. Modal split then is defined simply by comparisons of utilities; if for simplicity we assume that θ is uniformly distributed between [0,1], then the resulting modal split model is given by $Y_T = Y \cdot H\left(g_A - g_T\right) = \frac{Y}{B_A}\left[g_A - g_T\right]$ and thus, from equation (13), the fixed-point equation that determines the optimal modal split is:

$$Y_T^* = \frac{Y}{B_A}\left[MC_A\left(Y - Y_T^*\right) - MC_T\left(Y_T^*\right)\right] \tag{14}$$

The OMS for this mode choice model is graphically represented in Figure 4 as the intersection between the 45° line and the right hand side of equation (14).

In Figure 4 the minimum cost solution corresponds to the level of Y_T where marginal costs are equal ($Y_T^* minTC$), which will always be to the right of the welfare maximizing solution, as can be easily seen graphically. Therefore if demand is well described through linear deterministic utilities, the optimal modal split will always have more people on cars than the one that minimizes total costs showing that –when optimal prices are in place– the optimal modal split is less prone to transit usage than what the Downs-Thomson equilibrium would have predicted graphically. We can also use Figure 4 to perform one basic comparative static: if the intrinsic attractiveness of auto, B_A, increases the absolute value of the slope of the line representing the right hand side of equation (14) diminishes, and therefore Y_T^* decreases and Y_A^* increases as expected.

3.2 Binomial Logit mode choice

Consider now that mode choice is assumed to be probabilistic, where the probability depends on the difference between the generalized costs. In this case the H transformation will still be decreasing in Y_T but, as probabilities are always positive, the function will never reach zero. Thus, let us examine the optimal transit level of users when choice is represented by the (binomial) logit model, probably the most used model in transport demand applications. The probability of choosing transit is given by

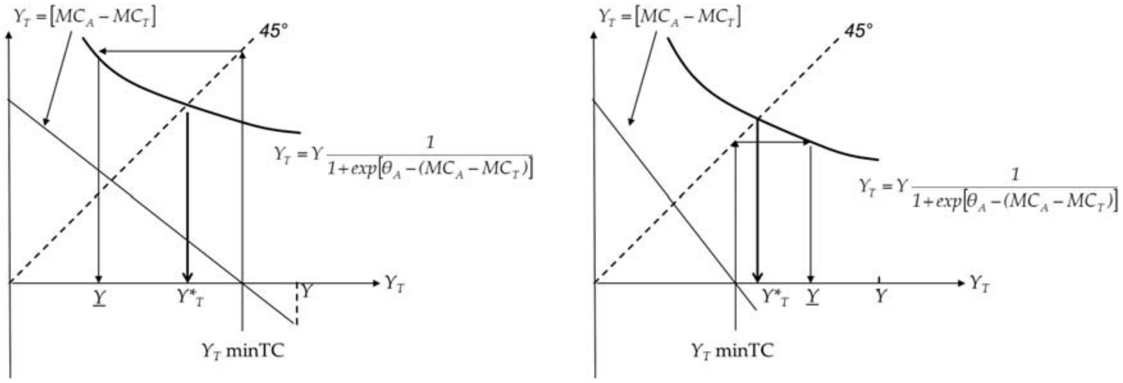
$$Y_T = Y \cdot H\left(g_A - g_T\right) = Y \frac{exp(-g_T)}{exp(-g_T) + exp(\theta_A - g_A)} = Y \frac{1}{1 + exp\left(\theta_A - \left(g_A - g_T\right)\right)}$$

θ_A is the automobile modal constant indicating its basic relative attractiveness with respect to transit. Then, as we did before, we use equation (13) to obtain the fixed-point equation that determines the optimal modal split:

$$Y_T^* = \frac{Y}{1 + exp\left[\theta_A - MC_A\left(Y - Y_T^*\right) - MC_T\left(Y_T^*\right)\right]} \tag{15}$$

The optimal modal split for this mode choice model can be, again, represented graphically as a fixed-point problem whose solution is given by the intersection between the 45° line and the right hand side of equation (15), represented as the downward sloping curve in Figure 5. The difficulty here, as compared to the deterministic case, is that it is not clear at the outset how the right hand side of (15) compares to the function $MC_A\left(Y - Y_T\right) - MC_T\left(Y_T\right)$. Hence, in Figure 5 we have also included the line representing the difference between marginal costs as a function of Y_T, and the particular point $Y_T^* minTC$.

Panels (a) and (b) in Figure 5 show that, with a probabilistic mode choice model, the number of transit users in the optimal modal split might be larger or smaller than in the case of cost minimization. One way to test which case prevails is shown in the Figure panels. As explained before, $Y_T^* minTC$ is the number of transit users that make marginal costs equal. Replacing this value on the right hand side of (15) instead of Y_T^* leads to a number $\underline{Y}$ –not dependent on a optimization variables– given by $\underline{Y} = Y\left[1 + exp(\theta_A)\right]^{-1}$. In Figure 3 this number is obtained by taking $Y_T^* minTC$ to the 45° line and then to the choice curve. Hence there are three possible cases:



(a) (b)

Figure 5: Probabilistic choice

1. If the modal split that minimizes total costs (where marginal costs are equal) is larger than $\underline{Y} = Y\left[1+exp(\theta_A)\right]^{-1}$, then $\underline{Y} < Y_T^* < Y_T^*minTC$ and therefore the optimal modal split has less people using transit than when total cost is minimized (as in Figure 3a).

2. If, on the other hand, the modal split that minimizes total costs is smaller than $\underline{Y}$, then $Y_T^*minTC < Y_T^* < \underline{Y}$ and therefore the optimal modal split has more people using transit than when total cost is minimized (as in Figure 3b).

3. If the modal split that minimizes total costs is equal to $\underline{Y}$, then $Y_T^*minTC = Y_T^* = \underline{Y}$.

4. Financial and other properties

The general conditions for optimality for the model formulated in Section 2 are synthesized by equations (9), (10) and either (14) or (15). Qualitatively we have found that the key element is that the usual rules –meaning as in separate models– to calculate the optimal fares and subsidies for transit and optimal congestion tolls for cars hold for a modal split whose optimal values have to be found depending on the specific mode choice model. However, equation (7) shows that in this model what matters is actually the difference between these prices and not the price levels themselves, such that it is more accurate to refer to an optimal price difference than to optimal prices. As discussed previously, this feature follows from the assumption of only two substitutes modes and a perfectly inelastic total demand, and has been found previosuly in the literature (e.g. Danielis and Marcucci, 2002).

The implication of this for the financial result of the transport sector is direct: if total demand is quite inelastic, once the optimal modal split is calculated one can plug in these values into the specific (isolated) models for the transit and private transport systems, and calculate the optimal toll, fare and subsidy per user but, if doing this does not lead to a desired financial result –for example that congestion pricing revenues cover transit subsidies– the planner can raise both the transit fare and the congestion toll by the same amount until reaching the desired outcome. This will keep the optimal price difference unchanged and, therefore, the modal split will also remain unchanged, at its optimal level. In particular, this implies that there would not be a need to raise income from other sectors of the economy in order to achieve the first-best in the urban transport sector, a result that, we stress for clarity, hinges on the fact that total travel is perfectly inelastic. If there are margins of adjustment other than mode choice, such as elastic total travel or a third free mode –such as biking– increasing fares would affect welfare.

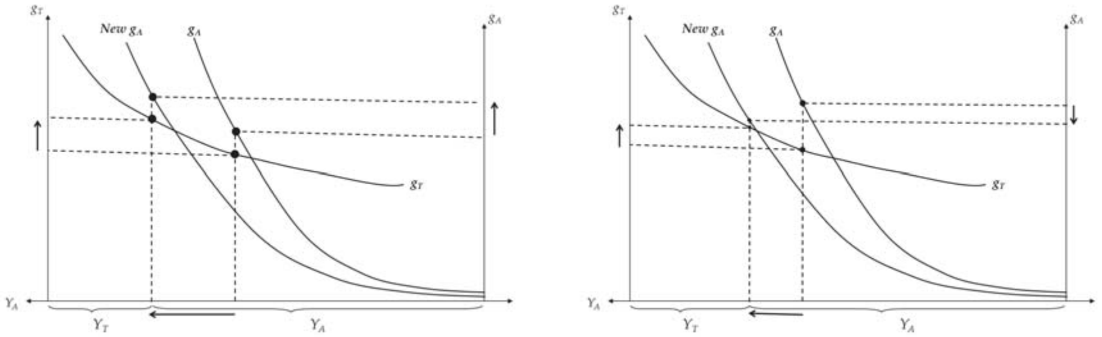
Note also that if there were no congestion tolls in place, that is $P_A = 0$, then achieving the optimal price difference would require a quite low transit fare, even perhaps a negative one, and therefore large subsidies. If that low –yet optimal– transit fare were not feasible, either because is politically difficult or because it is actually negative, then the modal split would not be optimal and social welfare might be increased by enlarging transit subsidies in order to decrease the fare. This is in line with what Parry and Small (2009) found empirically.

A second most important analysis we can use this model for relates to the Downs-Thomson paradox, synthesized by Mogridge (1997) and explained in Section 2.1. Recall that, in essence, the paradox makes reference to the negative effect of an investment in roads that will cause a reduction in auto users' average cost (time) for every demand level, i.e. a downward displacement of the average cost for auto users as in Figure 2b. Since in Mogridge's formulation the modal split takes place where this curve intersects the total average costs for transit (if it is priced to self-finance), then at this *equilibrium* the generalized cost for all users increase.

So, what happens when both an explicit choice model and optimal transit design are introduced as we have done here? The first thing to note is that an investment in auto infrastructure changes –conditional on usage– car travel time such that $t_A(Y_A) > t_A^I(Y_A)$, where I stands for investment. But it is also true that $t'_A(Y_A) > t_A^I{}'(Y_A)$, that is, for a given level of traffic, the marginal contribution to congestion of an extra car is smaller after the investment. This implies that $\frac{\partial MgC_A}{\partial I} < 0$ and, since $H' > 0$, one can see from the equation that describes the optimal modal split in general (equation 13), that:

$$\frac{\partial Y_T^*}{\partial I} = Y \underbrace{\frac{\partial H}{\partial MgC_A}}_{+} \underbrace{\frac{\partial MgC_A}{\partial I}}_{-} < 0$$

That is, the number of transit users do diminish and the number of car users increase, a process similar to what happens in the Downs-Thomson paradox. Given this, the aver-



(a) (b)

Figure 6: Average costs for transit and car users after road investment

age cost for transit users will certainly increase along g_t, as shown in Figures 6a and 6b. Yet what happens to the average cost for car users is undetermined. The reason is that it is not clear where the initial and new optimal modal split are. A situation where both type of users are worse off after investment is depicted in Figure 6a, and another where car users improve their situation is depicted in Figure 6b. The actual specific conditions under which one or the other situation prevails are, we believe, a quite important avenue for future research with important policy implications.

Finally, we want to stress that a more complex model that has transit costs depending on the vehicle size K would lead us to same results. Consider for example, the operators costs proposed by Jara-Díaz and Gschwender (2009), $(c_0 + c_1 K) f$. Since it would still be true that having excess capacity is not optimal in the absence of crwoding costs, then it still holds that $K = Y_T / f$. Thus, one could replace K and obtain the following transit cost: $c_0 f + c_1 Y_T$, such that the objective function still depends explicitely on P_A, P_T, f. The only difference would be that the conditional optimal frequency would be a bit more complex than (9), but it would still feature the square root shape.

5. Conclusions

We have formulated a model where all the usual properties of the car congestion process and the transit design problem have been simultaneously introduced. These technical modal descriptions have been linked trough users' behavior by means of a mode choice model between auto and transit. Finding the maximum welfare problem for this setting showed that:

1. Introducing behavior makes the optimal car-transit split generally different from the total cost minimizing one. Whether the difference will favor car or transit will de-

pend on the specific mode choice model.

2. When total travel demand is inelastic (as it its usually the case on peak-hours of working days) there is no optimal congestion price independent of transit price; it is the difference what matters.

3. The optimal money price difference has an optimal transit frequency attached, and together they yield the optimal modal split.

4. Once the optimal modal split is found, the corresponding congestion tolls and transit subsidies and fares from stand-alone formulations –where each mode is priced independently– are consistent with the optimal money price difference.

5. It is possible to establish prices that, while keeping the optimal price difference, make the total subsidy for transit equal to the total (congestion) tolls collected, i.e. self financing of the transport sector is feasible.

6. Investment in car infrastructure will induce an increase in generalized cost for transit users; the effect on car users however is unclear. Taking into account users behaviour might imply that the Downs-Thomson paradox survive only if more specific conditions are fulfilled.

Finally, in our opinion this analysis opens a number of avenues for future research. First, there have been a number of papers that have looked at ways of recycling congestion tolls to ensure a Pareto superior result, or the support of a majority for the policy (the so-called political economy of congestion pricing). This recycling might include using revenues for transit subsidies, yet these papers have not considered that financial and transit design aspects are indeed interrelated, as shown here; the literature on congestion tolls recycling would benefit from a framework encompassing these interrelations. Second, it seems necessary to understand the exact circumstances that make the Downs-Thomson paradox survive, that is, when does an investment in road infrastructure hurts not only transit users (through equilibrium adjustments) but also car users. Our theoretical findings suggest that individual parameters and their distribution (car attractiveness, values of time) could play a role in the welfare properties of such investment but the exact way they play that role is far from clear. It might be also important to take into account that in many cases transit and cars share capacity and therefore there are sizeable cross-congestion (see Basso et al. 2011 for a paper that considers these effect in a fixed-capacity setting). Finally, issues of distributional impacts arise indeed when the population is heterogeneous and this should be explored following a framework where congestion pricing, transit subsidies, transit design and mode choice models interact.

References

Arnott, R., de Palma, A., Lindsey, R., (1994) The welfare effects of congestion tolls with heterogeneous commuters. Journal of Transport Economics and Policy 28 (2), 139–161.

Basso, LJ., Guevara, AC., Gschwender, A., Fuster, M. (2011) Congestion pricing, transit subsidies and dedicated bus lanes: Efficient and practical solutions to congestion. Transport Policy, in press, doi:10.1016/j.tranpol.2011.01.002

Danielis, R. and Marcucci, E. (2002), Bottleneck road congestion pricing with a competing railroad service. Transportation Research Part E: Logistics and Transportation Review 38(5), 379-388.

Duranton, G. and Turner, MA. (2011) The Fundamental Law of Road Congestion: Evidence from US Cities. American Economic Review, forthcoming.

Hau, T.D., (1998) Congestion pricing and road investment. In: Road Pricing, Traffic Congestion and the Environment. Edward Elgar, Cheltenham, UK, pp. 39–78.

Huang. H-J. (2000) Fares and Tolls in a Competitive System with Transit and Highway: the case with two groups of commuters. Transportation Research Part E 36. 267-284

Jackson, R. (1975) Optimal subsidies for public transport. Journal of Transport Economics and Policy 9 (1), 3-15.

Jansson. J.O. (1984) Transport System Optimization and Pricing. Wiley

Jara-Díaz. S.R. and Gschwender. A. (2003a) Towards a general microeconomic model for the operation of public transport. Transport Reviews 23, 453–469

Jara-Díaz. S.R. and Gschwender. A. (2003b) From the single line model to the spatial structure of transit services: corridors or direct? Journal of Transport Economics and Policy 37-2, 261-277.

Jara-Díaz. S.R. (2007) Transport Economic Theory, Elsevier.

Jara-Díaz. S.R. and Gschwender. A. (2009) The Effect of Financial Constraints on the Optimal Design of Public Transport Services. Transportation 36 (1), 65-75.

Kockelman, K.M., Kalmanje, S., 2005. Credit-based congestion pricing: a policy proposal and the public's response. Transportation Research Part A 39, 671–690.

Mohring. H. (1972) Optimization and Scale Economies in Urban Bus Transportation. American Economic Review 62, 591–604.

Mohring. H. (1979) The Benefits of Reserved Bus Lanes. Mass Transit Subsidies and Marginal Cost Pricing in Alleviating Traffic Congestion. In Current Issues in Urban Economics. Mieskowosky. P. y Straszheim M. editors.

Mogridge, M (1997) The self-defeating nature of urban road capacity policy - A review of theories, disputes and available evidence. Transport Policy, 4(1) 5-23.

Parry, I., Small, KA. (2009) Should urban transit be reduced? The American Economic Review 99, 700–724.

Proost, S. and Van Dender, K. (2008) Optimal Urban Transport Pricing in the Presence of Congestion, Economies of Density and Costly Public Funds. Transportation Research Part A 42, 1220–1230.

Small. K. (1983) Bus Priority and Congestion Pricing on Urban Highways. In Research in Transportation Economics. Keeler. T. Ed. JAI.

Small. KA. and Verhoef. ET. (2007) The Economics of Urban Transportation. London. Routledge.

Tsekeris. T. and Voβ. S. (2009) Design and Evaluation of Road Pricing: state-of-the-art and methodological advances. Netnomics 10(1), 5-52.

Viton. P.. (1983) Pareto Optimal Urban Transportation Equilibria. In Research in Transportation Economics. Keeler. T. Ed. JAI.

Nie, Y. and Liu. Y, J. (2010) Existence of self-financing and Pareto-improving congestion pricing: impact of value of time distribution. Transportation Research Part A 44, 39-51.

Appendix

Int this appendix we detail how first-order conditions (7) and (8) are obtained from the optimization problem (6). Consider first that in the absence of crowding costs, it will never be optimal to choose a vehicle size larger than what is strictley needed, and therefore it holds that at the optimum $K = Y_T/f$. We therefore need only to worry about f and then obtain K as result. First-order conditions are then obtained by differentiating:

$$W(P_A, P_T, f) = -\int_{(g_A^0, g_T^0)}^{(g_A, g_T)} \sum_i Y_i(g_A, g_T) dg_i + P_A Y_A(P_A, P_T, f) + P_T Y_T(P_A, P_T, f) - c\,f$$

with respect to P_A, P_A and f. Taking derivative with respect to P_A we get:

$$\frac{\partial W}{\partial P_A} = \frac{\partial W}{\partial g_A}\frac{\partial g_A}{\partial P_A} + \frac{\partial W}{\partial g_T}\frac{\partial g_T}{\partial P_A} + \frac{\partial W}{\partial P_A}$$

$$\frac{\partial W}{\partial P_A} = \frac{\partial CS}{\partial g_A}\frac{\partial g_A}{\partial P_A} + \frac{\partial CS}{\partial g_T}\frac{\partial g_T}{\partial P_A} + P_A\frac{\partial Y_A}{\partial P_A} + P_T\frac{\partial Y_T}{\partial P_A}$$

Since Consumer Surplus is represented by a line-integral that is path independent, the vector version of the fundamental theorem of calculus hold, leading to:

$$\frac{\partial W}{\partial P_A} = -Y_A\frac{\partial g_A}{\partial P_A} - Y_T\frac{\partial g_T}{\partial P_A} + Y_A + P_A\frac{\partial Y_A}{\partial P_A} + P_T\frac{\partial Y_T}{\partial P_A}$$

from the definitions of g_A and g_T in equations (1) and (2), and noting that $\frac{\partial Y_T}{\partial P_A} = \frac{\partial (Y - Y_A)}{\partial P_A} = -\frac{\partial Y_A}{\partial P_A}$ we then get

$$\frac{\partial W}{\partial P_A} = -Y_A\left(1 + \alpha\, t_A'(Y_A)\frac{\partial Y_A}{\partial P_A}\right) + Y_T\frac{\alpha\,\mu}{f}\frac{\partial Y_A}{\partial P_A} + Y_A + P_A\frac{\partial Y_A}{\partial P_A} - P_T\frac{\partial Y_A}{\partial P_A}$$

$$\frac{\partial W}{\partial P_A} = \left(Y_T\frac{\alpha\,\mu}{f} - Y_A\alpha\, t_A'(Y_A) + P_A - P_T\right)\frac{\partial Y_A}{\partial P_A}$$

And then equating $\frac{\partial W}{\partial P_A}$ to zero and following the same procedure for $\frac{\partial W}{\partial P_T}$ allows to get the first order conditions (7).

Next, differentiating W with respect to f we get:

$$\frac{\partial W}{\partial f} = \frac{\partial CS}{\partial g_A}\frac{\partial g_A}{\partial f} + \frac{\partial CS}{\partial g_T}\frac{\partial g_T}{\partial f} + P_A\frac{\partial Y_A}{\partial f} - P_T\frac{\partial Y_A}{\partial f} - c$$

Again, we use the vector version of the fundamental theorem of calculus and get:

$$\frac{\partial W}{\partial f} = -Y_A\frac{\partial g_A}{\partial f} - Y_T\frac{\partial g_T}{\partial f} + (P_A - P_T)\frac{\partial Y_A}{\partial f} - c$$

From the definitions of g_A and g_T in equations (1) and (2), we obtain

$$\frac{\partial W}{\partial f} = -Y_A \alpha\, t'_A(Y_A) \frac{\partial Y_A}{\partial f} - Y_T \left(-\frac{\alpha\, \beta}{2 f^2} + \alpha\, \mu \frac{-\frac{\partial Y_A}{\partial f} f - Y_T}{f^2} \right) + (P_A - P_T) \frac{\partial Y_A}{\partial f} - c$$

which simplifies to

$$\frac{\partial W}{\partial f} = \frac{\partial Y_A}{\partial f} \left(Y_T \frac{\alpha\, \mu}{f} - Y_A \alpha\, t'_A(Y_A) + P_A - P_T \right) + Y_T \frac{\alpha\, \beta}{2 f^2} + Y_T^2 \frac{\alpha\, \mu}{f^2} - c$$

And then equating $\frac{\partial W}{\partial f}$ to zero leads to the first order condition (8).