How to Estimate, Take Into Account, and Improve Travel Time Reliability in Transportation Networks

Ruey L. Cheu, Vladik Kreinovich, François Modave, Gang Xiang, Tao Li, and Tanja Magoc
Center for Transportation Infrastructure Systems, University of Texas, El Paso, TX 79968, USA, contact vladik@utep.edu

Abstract. Many urban areas suffer from traffic congestion. Intuitively, it may seem that a road expansion (e.g., the opening of a new road) should always improve the traffic conditions. However, in reality, a new road can actually worsen traffic congestion. It is therefore extremely important that before we start a road expansion project, we first predict the effect of this project on traffic congestion.

Traditional approach to this prediction is based on the assumption that for any time of the day, we know the exact amount of traffic that needs to go from each *o*rigin city zone A to every other *d*estination city zone B (these values form an *OD-matrix*), and that we know the exact capacity of each road segment. Under this assumption, known efficient algorithms produce the equilibrium traffic flows.

In reality, the road capacity may unpredictably change due to weather conditions, accidents, etc. Drivers take this uncertainty into account when planning their trips: e.g., if a driver does not want to be late, he or she may follow a slower route but with a guaranteed arrival time instead of a (on average) faster but unpredictable one. We must therefore take this uncertainty into account in traffic simulations. In this paper, we describe algorithms that take this uncertainty into account.

Keywords: transportation networks, traffic assignment, reliability, risk-taking behavior

1. Decreasing Traffic Congestion: Formulation of the Problem

Decreasing traffic congestion: a practical problem. Many urban areas suffer from traffic congestion. It is therefore desirable to decrease this congestion: e.g., by building new roads, or by adding new lanes to the existing roads.

Important difficulty: a new road can worsen traffic congestion. Intuitively, it may seem that a road expansion (e.g., the opening of a new road) should always improve the traffic conditions. However, in reality, a new road can actually worsen traffic congestion. Specifically, if too many cars move to a new road, this road may become even more congested than the old roads initially were, and so the traffic situation will actually decrease – prompting people to abandon this new road. This possible negative effect of a new road on congestion is a very well known "paradox" of transportation science, a paradox which explains the need for a detailed analysis in the planning of the new road; see, e.g, (Ahuja et al., 1993; Sheffi, 1985). This paradox was first discovered by A.

Doig (see (Appa, 1973)) and first published in (Braess, 1968; Charnes and Klingman, 1971; Szwarc, 1971).

Importance of the preliminary analysis of the results of road expansion. Our objective is to decrease traffic congestion. We have just mentioned that an addition of a new road can actually worsen the traffic congestion. It is therefore extremely important that before we start a road expansion project, we first predict the effect of this project on traffic congestion.

Traditional approach to predicting the results of road expansion. Traditional approach to predicting the results of road expansion is based on the assumption that for any time of the day, we know the exact amount of traffic that needs to go from each *o*rigin city zone A to every other *d*estination city zone B (these values form an *OD-matrix*), and that we know the exact capacity of each road segment. Under this assumption, known efficient algorithms produce the equilibrium traffic flows; see, e.g., (Sheffi, 1985).

Limitations of the traditional approach to predicting the results of road expansion. In reality, the road capacity may unpredictably change due to weather conditions, accidents, etc. Drivers take this uncertainty into account when planning their trips: e.g., if a driver does not want to be late, he or she may follow a slower route but with a guaranteed arrival time instead of a (on average) faster but unpredictable one.

We must therefore take this uncertainty into account in traffic simulations.

What we do in this paper. In this paper, we describe algorithms that take the above uncertainty into account.

Comment. Some of the results presented in this paper first appeared in our research report (Cheu et al., 2007). This report also describes a software package that implements our algorithms.

2. Traffic Assignment: Brief Reminder

Road assignment problem: informal description. In order to select the best road expansion project, we must be able to predict how different projects will affect road congestion. For that, we need to be able, based on the traffic demand and on the road capacities, to predict the traffic on different places of different roads at different times of the day. This prediction problem is called the *traffic assignment* problem.

To describe this problem in precise terms, we need to describe how exactly the traffic demand is described, how the road capacities are described, and what exactly assumptions do we make about the drivers' behavior.

Granulation. To describe traffic demand, we divide the urban area into *zones* and describe how many drivers need to get from one zone to another.

Similarly, to describe road capacity, we divide all the roads into road *segments* (*links*), and describe the capacity of each link.

The time of the day is similarly divided into *time intervals*.

Comment. How to select an appropriate size of a zone, of a road link, and of a time interval?

– On the one hand, the finer the division, we more accurate is the resulting traffic picture.

– On the other hand, the finer the division, the more zones and links we need to consider and hence, the more computations we need to perform.

Thus, the granularity of the traffic problem should be determined by the trade-off between accuracy and computational complexity.

For example, for the city of El Paso with a population of 700,000, a standard road network model consists of 681 zones and 4836 road links.

How to describe traffic demand? Once we divided the urban area into n zones, we must describe, for every two zones i and j, the number of drivers d_{ij} who need to go from zone i to zone j. The corresponding $n \times n$ matrix is called an *origin-to-destination* matrix, or an O-D matrix, for short.

So, to the traffic demand is described by the O-D matrices corresponding to different times of the day.

How to describe road capacity? For each road link, the road capacity is usually described by the number c of cars per hour which can pass through this road link.

How to describe travel time along a road link? Every road link has a posted speed limit. When there are few cars of this road, then these few cars can safely travel at the speed limit s. The resulting travel time t^f along this road link can be estimated as L/s, where L is the length of this road link. This travel time t^f is called a *free flow* travel time.

When the traffic volume v increases, congestions starts, the cars start slowing each other down. As a result, the travel time t along the road link increases. The dependence of the travel time on the volume is usually described by the Bureau of Public Roads (BPR) formula

$$t = t^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^{\beta}\right].$$

The parameters a and β are determined experimentally; usually, $a \approx 0.15$ and $\beta \approx 4$.

Equilibrium. When a new road is built, some traffic moves to this road to avoid congestion on the other roads; this causes congestion on the new road, which, in its turn, leads drivers to go back to their previous routes, etc. These changes continue until there are alternative routes in which the overall travel time is larger.

Eventually, this process converges to an *equilibrium*, i.e., to a situation in which the travel time along all used alternative routes is exactly the same – and the travel times along other un-used routes is higher; see, e.g., (Sheffi, 1985).

There exist efficient algorithms which, given the traffic demand (i.e., the O-D matrices) and the road capacity, computes the corresponding equilibrium (Sheffi, 1985). This algorithm computes the traffic volume along each road link, the travel time between every two zones, etc.

3. How We Can Use the Existing Traffic Assignment Algorithms to Solve Our Problem: Analysis

Our main objective: reminder. Our main objective is to predict how different road projects will affect future traffic congestion – so that we will be able to select a project which provides the best congestion relief.

To be able to do that, we must predict the traffic congestion resulting from the implementation of each of the road projects.

How we can predict the traffic congestion resulting from different road projects. As we have mentioned, to apply the existing traffic assignment algorithms, we need to know the traffic capacities and traffic demands.

The traffic capacities of the improved road network come directly from the road project – we know which new road links we build, what is their capacity, and which existing links are expanded. So, to solve our problem, we need to find the traffic demands.

Future traffic demands: what is known. There exist tools and techniques for predicting population growth in different zones, and for describing how this population growth will affect the overall traffic demand. Texas Department of Transportation (TxDOT) have been using the resulting predictions of daily O-D matrices corresponding to different future times (such as the year 2030).

Future traffic demands: what is lacking. To get a better understanding of the future traffic patterns, we must be able to describe how this daily traffic is distributed over different time intervals, in particular, how much of this traffic occurs during the critical time intervals corresponding to the morning rush hour. In other words, we need to "decompose" the daily O-D matrix into O-D matrices corresponding to different time intervals, e.g., 1 hour or 15 minute intervals.

How to find traffic demands corresponding to different times of the day: first approximation. In the first approximation, we can determine these O-D matrices by simply assuming that the proportion of drivers starts their trip at different times (such as 7 to 7:15 am, 7:15 to 7:20 am, etc.) as now. This first approximation is described in the next chapter.

Limitation of the first approximation predictions– and the need for better predictions. the problem with this first approximation is that the existing traffic pattern is based on the current traffic congestion. For example, if traveling from zone A to zone B takes a long time (say, 1 hour), drivers who need to drive from A to B and reach B by 9 am leave early, at 8 am, so as to be at their destination on time. As a result, in the existing traffic pattern, we have a lot of drivers leaving from A to B at 8 am.

If we simply use the existing travel pattern, we will therefore predict that in the future, a similarly big portion of drivers going from A to B also leaves at 8 am.

If we build a new road segment that eases this congestion, then there is no longer a need for these drivers to leave earlier. As a result, the actual O-D value corresponding to leaving at 8 am will be much smaller than according to our first approximation prediction.

To provide a more accurate prediction of the future traffic demand, we must therefore take into account the road improvements. In the following sections,w e describe how this can be taken into account.

Taking uncertainty into account. Finally, as we mentioned earlier, we need to take into account the uncertainty which which we can predict travel times. This is taken into account in the final sections of this paper.

4. How to Predict Future Traffic Demand: First Approximation

Main idea behind the first approximation: reminder. To predict the effect of different road projects on the future traffic congestion, we need to know future traffic demand, i.e., we need to know how many drivers will go from every zone to every other zones at different moments of time.

We usually have *daily* predictions, i.e., predictions describing the overall daily traffic for every origin-destination (O-D) pair. Based on these daily O-D matrices, we must predict O-D matrices corresponding to different time intervals.

It is reasonable to assume that in the planned future, the distribution of departure times will be approximately the same as at present. Under this assumption, we can estimate the O-D matrix corresponding to a certain time interval by simply multiplying the (future) daily O-D matrix by the corresponding *K-factor* – portion of traffic which occurs during this time interval. These K-factors can be determined by an empirical analysis of the current traffic: a K-factor corresponding to a certain time interval can be estimated as a ratio between

- the number of trips which start at this time interval, and
- the overall number of trips.

Use of empirical K-factors and linear interpolation. At present, the empirical values of the K-factor are only available for hourly intervals. If we want to find the K-factors corresponding to half-hours or 15 minute intervals, it is reasonable to use linear interpolation. Let us illustrate linear interpolation on a simple example. Let us assume that we know K-factors corresponding to the hourly traffic, in particular, we know that:

- at 7:00 am, the K-factor is 6.0%, meaning that at this moment of time, the traffic volume (in terms of vehicles per hour) is equal to 6.0% of the daily traffic volume (in terms of vehicles per day); and
- at 8:00 am, the K-factor is 8.0%, meaning that at this moment of time, the traffic volume (in terms of vehicles per hour) is equal to 8.0% of the daily traffic volume (in terms of vehicles per day).

For example, if for some O-D pair, the daily traffic volume is 1,000 vehicles per day, then:

- at 7:00 am, the traffic volume will be $6.0\% \cdot 1000 = 60$ vehicles per hour, and

– at 8:00 am, the traffic volume will be $8.0\% \cdot 1000 = 80$ vehicles per hour.

If we are interested in half-hour intervals, then we need to also estimate the traffic volume at the intermediate moment of time 7:30 am. Linear interpolation means that as such an estimate, we use the value $(6.0 + 8.0)/2 = 7\%$. So, we get the following K-factors for the half-hour time intervals:

– at 7:00 am, the K-factor is 6.0%;

– at 7:30 am, the K-factor is 7.0%;

– at 8:00 am, the K-factor is 8.0%.

Similarly, to extrapolate into 15 minute intervals, we use $(6.0 + 7.0)/2 = 6.5\%$ for 7:15 am and $(7.0 + 8.0)/2 = 7.5\%$ for 7:45 am. So, we get the following K-factors for the 15 minute time intervals:

– at 7:00 am, the K-factor is 6.0%;

– at 7:15 am, the K-factor is 6.5%;

– at 7:30 am, the K-factor is 7.0%;

– at 7:45 am, the K-factor is 7.5%;

– at 8:00 am, the K-factor is 8.0%.

In the above example, in which for some O-D pair the daily traffic volume is 1,000 vehicles per day:

– at 7:00 am the traffic volume is $6.0\% \cdot 1000 = 60$ vehicles per hour,

– at 7:15 am the traffic volume is $6.5\% \cdot 1000 = 65$ vehicles per hour,

– etc.

5. How to Take Departure Time Choice into Account

Need to take departure time choice into consideration. To understand how different road projects will affect the future traffic, we need to estimate the O-D matrices for different time intervals. At present, we usually only have estimates for the daily O-D matrices. In the previous section, we described how to use the current K-factors to divide the daily O-D matrices into O-D matrices for different time intervals.

The resulting O-D matrices are, however, only a first approximation to the actual O-D matrices. Indeed, the existing O-D matrices and the existing values of the K-factor are based on the experience of the drivers under current driving conditions. A driver selects his or her departure time based on the time that the driver needs to reach the destination (e.g., the work-start time), and the expected

travel time. For example, if the driver needs to be at work at 8:00am, and the travel time to his or her destination is 30 minutes, then the driver leaves at 7:30 am.

Population changes and new roads will change expected travel time. For example, if due to the increased population and the resulting increase road congestion the expected travel time increases to 45 minutes, then the same driver leaves at 7:15 am instead of the previous 7:30 am. So, the corresponding entry in O-D matrix corresponding to 7:30 am will decrease while a similar entry in the O-D matrix corresponding to 7:15 am will increase.

Similarly, if a new freeway decreases the expected travel time to 15 minutes, then the driver will leave at 7:45 am instead of the original 7:30 am. In this case, the corresponding entry in O-D matrix corresponding to 7:30 am will decrease while a similar entry in the O-D matrix corresponding to 7:45 am will increase.

In general, the change in a transport network and/or the change in travel time will change the departure time choice and thus, change the resulting O-D matrix. Let us describe how we can take this departure time choice into consideration.

The use of logit model: general idea. In transportation engineering, the most widely used model for describing the general choice (especially the choice in transportation-related situations) is the logit model. In the logit model, the probability of departure in different time intervals is determined by the utility of different departure times to the driver. According to this model, the probability P_i that a driver will choose the i-th time interval is proportional to $\exp(u_i)$, where u_i is the expected utility of selecting this time interval. The coefficient at $\exp(u_i)$ must be chosen from the requirement that the sum of these probabilities be equal to 1. So, the desired probability has the form $P_i = \exp(u_i)/s$, where $s \stackrel{\text{def}}{=} \exp(u_1) + \ldots + \exp(u_n)$. (Motivation for this model is presented in Appendix A.)

To apply the logit model, we must be able to estimate the utilities of different departure time choices. According to (Noland and Small, 1995; Noland et al., 1998), the utility u_i of choosing the i-th time interval is determined by the following formula:

$$u_i = -0.1051 \cdot E(T) - 0.0931 \cdot E(SDE) - 0.1299 \cdot E(SDL) - 1.3466 \cdot P_L - 0.3463 \cdot \frac{S}{E(T)},$$

where $E(T)$ is the expected value of travel time T, $E(SDE)$ is the expected value of the wait time SDE when arriving early, $E(SDL)$ is the expected value of the delay SDL when arriving late, P_L is the probability of arriving late, and S is the variance of the travel time. If we denote departure time by t_d, and the desired arrival time by t_a, then we can express SDE as $SDE = \max(t_a - (t_d + T), 0)$, and SDL as $SDL = \max((t_d + T) - t_a, 0)$. So, to estimate the values of the utilities, we must be able to estimate the values of all these auxiliary characteristics.

How to estimate the expected travel time, expected wait and delay times, and the probability of arriving late. The first of these auxiliary values – the expected value $E(T)$ of the traffic time T – is the most straightforward to compute: we can find it by simply applying a standard traffic assignment procedure (e.g., the one implemented in the standard package TransCAD) to the original O-D matrices.

To estimate the expected value $E(SDE)$ of the wait time SDE and the expected value $E(SDL)$ of the time delay SDL, in addition to the travel time, we must also know the departure time t_d and the desired arrival time t_a.

Let us start our analysis with the departure time t_d. For simplicity, for all the traffic originating during a certain time interval, as a departure time, we take the midpoint of the corresponding time interval. For example, for all the traffic originating between 7:00 am and 7:15 am, we take 7:07.5 am as the departure time.

The analysis of the desired arrival time t_a is slightly more complicated. The desired arrival time depends on the time of the day. In the morning, the desired arrival time is the time when the drivers need to be at work or in school. During the evening rush hour, the desired arrival time is the time by which the drivers want to get back home, etc.

In terms of traffic congestion, the most crucial time interval is the morning rush hour, when for most drivers, the desired arrival time is the work-start time. In view of this, in the following text, we will refer to all desired arrival times as work-start times.

The work-start time usually depends on the destination zone. For example, in El Paso, most zones have the same work-start time with the exception of a few zones such as:

- the Fort Bliss zones where the military workday starts earlier, and
- the University zone(s) where the school day usually starts somewhat later.

For every zone, we therefore usually know the (average) work-start time, i.e., the (average) desired arrival time for all the trips with the destination in this zone.

Of course, the actual work-start time for different drivers arriving in the zone may somewhat differ from the average work-start time for this zone. To take this difference into consideration, we assume that the distribution of the actual works-start time follows a bell-shaped distribution around the average. We only consider discrete time moments, e.g., time moments separated by 15 minute time intervals. It makes sense to assume that:

- for the 40% of the drivers, the actual work-start time is the average for this zone,
- for 20%, the work-start time is 15 minute later,
- for another 20%, the work-start time is 15 minutes earlier,
- for 10%, it is 30 minutes later, and
- for the remaining 10%, it is 30 minutes earlier.

For example, if the average work-start time for a zone is 8:00 am, then the assumed work-start times are as follows

- for 10% of the drivers, the work-start time is 7:30 am;
- for 20% of the drivers, the work-start time is 7:45 am;
- for 40% of the drivers, the work-start time is 8:00 am;

– for 20% of the drivers, the work-start time is 8:15 am; and, finally,

– for 10% of the drivers, the work-start time is 8:30 am.

For each of these 5 groups, we can estimate the corresponding value of SDE as $SDE(t_a) = \max(t_a - (t_d + T), 0)$. To get the desired value of the expected wait time $E(SDE)$, we need to combine these values $SDE(t_a)$ with the corresponding probabilities. For example, when the average work-start time is 8:00 am, the expected value of SDE is equal to

$$E(SDE) = 0.1 \cdot SDE(7{:}30) + 0.2 \cdot SDE(7{:}45) + 0.4 \cdot SDE(8{:}00) + 0.2 \cdot SDE(8{:}15) + 0.1 \cdot SDE(8{:}15).$$

Similarly, we can estimate the expected value $E(SDL)$ of the delay SDL. By adding the probabilities corresponding to different work-start times, we can also estimate the probability P_L of being late.

How to estimate the variance of the travel time. In the previous paragraphs, we described how to estimate the expected values $E(T)$, $E(SDE)$, $E(SDL)$, and the probability P_L. To compute the desired utility value, we only need one more characteristic: the variance S of the travel time. Let us analyze how we can estimate the variance S.

In the deterministic traffic assignment model, once we know the capacities of all the road links and the traffic flows (i.e., the values of the O-D matrix), we can uniquely determine the traffic times for all O-D pairs. In practice, the travel time can change from day to day. Some changes in travel time are caused by a change in weather, by special events, etc.; the resulting deviations from travel time are usually minor. The only case when travel times change drastically is when there is a serious road incident somewhere in the network. Since incidents are the major source of travel time delays, it is reasonable to analyze incidents to estimate the variance S of the travel time.

For this analysis, we need to have a record of incidents which occurred during a certain period of time (e.g., 90 days). The record of each incident typically includes the location and time of this incident, and the number of lanes of the corresponding road which were closed because of this incident. To estimate the variance S corresponding to a certain time interval (e.g., from 8:00 to 8:15 am), we should only consider the incidents which occurred during that time interval. Based on the incident location, we can find the link on which this incident occurred. The incident decreases the capacity of this link. This decrease can be estimated based on the original number of lanes and on the number of lanes closed by this incident.

Comment. If all the lanes were closed by the incident, then the capacity of the link goes down to 0. A reader should be cautioned that the TransCAD software tool does not allow us to enter 0 value of a link capacity. To overcome this problem, we set the capacity to the smallest possible value (such 1 vehicle per hour). For all practical purposes, this is equivalent to setting this capacity to 0.

Let us now provide heuristic arguments for estimating the decrease in capacity in situations in which some lanes remain open. Let us start with the simplest case of a 1-lane road. In reality, depending on the severity of an incident, the factor from 0 to 1 describing the decreased capacity can take all possible values from the interval $[0, 1]$. In the incident record, we only mark whether the incident actually led to the lane closure or not. In other words, instead of the actual value of the capacity-decrease factor, we only keep, in effect, 0 or 1, with

– 0 corresponding to the closed lane, and

– 1 corresponding to the open lane.

In yet another terms, we approximate the actual value of the factor by 0 or 1. It is reasonable to assume that:

– factors 0.5 or higher get approximated by 1 (lane open), while

– factors below 0.5 are approximated by 0 (lane closed).

So, the incident records in which the lane remained open correspond to all possible values of the capacity-decrease factor from the interval $[0.5, 1]$. As a reasonable average value of this factor for the case when the lane remained open, we can therefore take the midpoint of this interval, i.e., the value 0.75.

In multi-lane roads, an incident usually disrupts the traffic on all the lanes. It is therefore reasonable to assume that if no lanes were closed, then the capacity of each lane was decreased to 75% of its original value. Thus, for minor incidents in which no lanes were closed, we set the resulting capacity to 3/4 of the original capacity of the link.

For a 2-lane road, if one lane is closed and another lane remain open, then we have one lane with 0 capacity and one lane with 3/4 of the original capacity; the resulting capacity is 3/4 of the capacity of a single lane, i.e., 3/8 of the original capacity of the 2-lane road.

For a 3-lane road, if one lane is closed this means that we retain only 2/3 of the incident-reduced 75% capacity, i.e., 1/2 of the original capacity. If two lanes are closed, this means that we retain only 1/3 of the reduced capacity, i.e., 1/4 of the original capacity.

Similar values can be estimated for 4-lane roads and, if necessary, for roads with a larger number of lanes.

For each recorded incident occurring at a given time interval, we replace the original capacity in the incident-affected link by the correspondingly reduced value, and solve the traffic assignment problem for thus reduced capacity. As a result, for each O-D pair, we get a new value of the travel time.

– when the incident is far away from the route, this travel time may be the same as in the original (no-incidents) traffic assignment;

– however, if the incident is close to the route (or on this route), this travel time is larger than in the no-incidents case.

Thus, for each O-D pair and for each time interval, for each day d during the selected time period P (e.g., 90 days), we have a value of the travel time $t(d)$:

– if there was no incident on this day, the value of the travel time comes from the original traffic assignment;

– for the days on which there was an incident during the given time interval, the travel time comes from the analysis of the network with the correspondingly reduced capacity.

Based on these values $t(d)$, we compute the mean value E of the travel time as $E = \frac{1}{P} \cdot \sum_{d=1}^{P} t(d)$, and then the desired variance S as $S = \frac{1}{P} \cdot \sum_{d=1}^{P} (t(d) - E)^2$.

How to take into account departure time choice when making traffic assignments: a seemingly natural idea and its limitations. In the two previous text, we described how we can compute the characteristics which are needed to estimate the utility related to each departure time. Let us now assume that we know the original O-D matrices for each time interval i. For each time interval i, we can use the corresponding O-D matrix and solve the traffic assignment problem corresponding to this time interval. From the resulting traffic assignment, we can compute the values of the desired auxiliary characteristics, and thus, estimate the expected utility u_i of departing at this time interval i. The logit formula $P_i = \exp(u_i)/s$, where $s = \exp(u_1) + \ldots + \exp(u_n)$, enables us to compute the probability P_i that the driver will actually select departure time interval i.

The probability P_i means that out of N drivers who travel from the given origin zone to the given destination zone, $N \cdot P_i$ leave during the i-th time interval. The overall number of drivers who leave from the given origin zone to the given destination zone can be computed by adding the corresponding values in the original O-D matrices for all time intervals. Multiplying this sum by P_i, we get the new value. These new values form the new O-D matrices for different time intervals i.

These new O-D matrices take into account the departure time choice. However, they are not the ultimate O-D matrices. Indeed, since we have changed the O-D matrices, we thus changed the traffic assignments at different moments of time; this will lead to different values of utilities u_i and probabilities P_i.

As an example, let us assume that there is an O-D pair for which the free-flow travel time is 30 minutes. Let us also assume that for the corresponding destination, everyone needs to be at work at 8 am. Let us also assume that at present, there is not much traffic congestion between the origin and destination zones, so everyone leaves around 7:30 am and gets to work on time. Since we are estimating the distribution of traffic flow over time intervals based on the existing traffic, we will thus conclude that

- in the O-D matrix corresponding to 7:30 am, we will have all the drivers, while
- in the O-D matrices corresponding to earlier time intervals, we will have no drivers at all.

Let us now apply these O-D matrices to the future traffic, when due to the population increase, the traffic volume becomes much higher. Due to this higher traffic volume, the traffic time will drastically exceed 30 minutes, so all the drivers leaving at 7:30 am will be, e.g., 15 minutes late.

On the other hand, drivers who happen to leave at 7:15 am encounter practically no traffic – because there was no one needing to drive at this time in the original O-D matrix, so their travel time is exactly 30 minutes, and they get to work by 7:45 am, 15 minutes earlier. As we have seen in the above empirical formula (and in full accordance with common sense), the penalty for being 15 minutes late is much higher than the penalty of being 15 minutes early. As a result, the utility corresponding to leaving at 7:15 am is higher than the probability of leaving at 7:30 am. Hence, in

accordance with the logit formula, the probability that a driver will select to leave at 7:15 am is much higher than the probability that this driver will leave at 7:30 am.

So, in the new O-D matrices, most drivers will leave at 7:15 am, and the values corresponding to leaving at 7:30 am will be much lower. If the drivers really follow the pattern corresponding to the new O-D matrix, then the traffic congestion corresponding to 7:30 am will be much lighter than before, so the utility of leaving at 7:30 am will become higher and thus, the probability of leaving at 7:30 am will increase again. It is reasonable to expect that if we repeat this procedure several times, we will eventually reach the desired stable values of the O-D matrix.

Let us describe these ideas in precise term. In essence, we have described a procedure which transforms the original set M of O-D matrices into a new set $F(M)$ of O-D matrices, a set which takes into account departure time choice based on the traffic assignments generated by the original O-D matrices. To completely take into account the departure time choice means to find the O-D matrices which already incorporate the departure time choice, i.e., the matrices M which do not change after this transformation: $F(M) = M$.

At first glance, it seems reasonable to find these "stable" O-D matrices M by using a reasonable iterative procedure:

- we start with the set of first-approximation O-D matrices M_1 which are obtained by multiplying the new O-D daily matrix by the original K-factors;
- then, we apply the transformation F again and again: $M_2 = F(M_1)$, $M_3 = F(M_2)$, ..., until the procedure converges, i.e., until the new set of matrices M_{i+1} becomes close to the previous set M_i.

This procedure seems even more reasonable if we recall that a similar iterative procedure is successfully used in TransCAD to find the traffic assignment. However, we found out that this seemingly reasonable procedure often does not converge.

This lack of convergence can be illustrated on a "toy" example in which we have a single origin, single destination, and two possible departure times. Similarly to the above example, let us assume that the work starts at 8 am, that the free-flow traffic time is 30 minutes, and that we consider two possible departure times 7:30 am and 7:15 am. Again, just like in the above example, we assume that the original O-D matrices are based on the existing low-congestion networks in which everyone leaves at 7:30 am and nobody leaves at 7:15 am. In other words, we assume that the K-factor for 7:30 am is 1, and the K-factor for 7:15 am is 0. We also assume that there are high penalties for being late and for spending too much time in traffic.

In accordance with the above iterative procedure, we start with the O-D matrices M1 in which everyone leaves for work at 7:30 am, and nobody leaves for work at 7:15 am. The only difference with the current situation is that we are applying the same K-factors to the future, more heavy traffic.

- For those departing at 7:15 am, there is no traffic, so the travel time is equal to the free-flow time of 30 minutes.
- The drivers departing at 7:30 am face a much heavier traffic, so we get a traffic congestion. As a result of this congestion, the travel time increases to 45 minutes.

So:

– drivers who leave at 7:15 am spend only 30 minutes in traffic and arrive 15 minutes early, while

– drivers who leave at 7:30 am spend 45 minutes on the road and are 15 minutes late.

Since we assumed that the penalties for being late are heavy, the expected utility of leaving at 7:15 am is much higher than the expected utility of leaving at 7:30 am. Thus, the probability of leaving at 7:15 am is overwhelmingly higher than the probability of leaving at 7:30 am. As a result, we arrive at the new O-D matrices $M_2 = F(M_1)$ in which almost everyone leaves at 7:15 am and practically no one leaves at 7:30 am.

For these new O-D matrices M_2:

– for those departing at 7:30 am, there is no traffic, so the travel time is equal to the free-flow time of 30 minutes;

– the drivers departing at 7:15 am face a much heavier traffic, so we get a traffic congestion; as a result of this congestion, the travel time increases to 45 minutes.

So:

– drivers who leave at 7:30 am spend only 30 minutes in traffic and arrive on time, while

– drivers who leave at 7:15 am spend 45 minutes on the road.

Since we assumed that the penalties for spending extra time on the road are heavy, the expected utility of leaving at 7:30 am is much higher than the expected utility of leaving at 7:15 am. Thus, the probability of leaving at 7:30 am is overwhelmingly higher than the probability of leaving at 7:15 am. As a result, we arrive at the new O-D matrices $M_3 = F(M_2)$ in which almost everyone leaves at 7:30 am and practically no one leaves at 7:15 am.

In other words, we are back to the original O-D matrices $M_3 \approx M_1$. These "flip-flop" changes continue without any convergence. How can we modify the above idea so as to enhance convergence?

How to take into account departure time choice when making traffic assignments: a more realistic approach. We started with the O-D matrices M_1 which describe the existing traffic behavior. We want to predict how a change in traffic volume and in road network will affect the driver's behavior. To do that, let us analyze

– how the actual drivers change their behavior if the road congestion and road conditions change, and

– how we can simulate this behavior in a computer model so as to predict these changes.

At first, the drivers simply try to follow the same traffic patterns as before, i.e., depart at the same times as before. In terms of the computer representation of the drivers' behavior, this means that the proportion of the drivers departing at different time intervals remains the same as in the original traffic. In other words, this behavior corresponds to what we described as the first approximation

M_1 – when we take the new daily O-D matrix and multiply it by the K-factors corresponding to the original traffic.

As we have mentioned, due to the change in traffic volume and in road capacity, this first-approximation behavior may lead to congestions and delays. When drivers realize this, they will change their departure time so as to avoid these new delays. The drivers will use the traffic patterns and delays caused by M_1 to decide on the new departure times. The resulting change in the O-D matrix is what we described in the previous section as a transformation F. In other words, the resulting O-D matrix is $M_2 = F(M_1)$.

The change of departure times, as reflected by the move from the original O-D matrices M_1 to the new O-D matrices M_2, will again change the traffic patterns and delay times, so again, there will be a need to change the departure times based on the new traffic delays.

In these terms, the above iterative process $M_{i+1} = F(M_i)$ corresponds to the situation when the drivers only use the experience of their most recent traffic behavior and ignore the rest of the traffic history. Let us illustrate this idea on the above "toy" example.

In this example, the drivers used to go to work at 7:30 am. For the original traffic volume, this was a reasonable departure time because it allowed them to be at work exactly at the desired time 8:00 am, and to spend as little time on the road as possible – exactly 30 minutes, the free-flow traffic time.

When the traffic volume increases, in Day 1 of this new arrangement, the drivers follow the same departure time as before, i.e., they all leave for work at 7:30 am. Since the traffic volume has increased, this departure time no longer lead to the desired results – most of the drivers are 15 minutes late for work.

Since in the first day, most drivers were 15 minutes late, on the second day they leave 15 minutes earlier, at 7:15 am, so as to be at work on time. They do reach work on time, but at the expense of driving 15 minutes longer than they used to. A few drivers, however, still leave at 7:30 am. To their pleasant surprise, they experience a smooth and fast ride and arrive at work exactly on time.

The other drivers learn about the negative experience of those who left at 7:15 am and of the positive experience of those who left at 7:30 am. In our iterative model, we assume that when the drivers decide on departure time at Day 3, they only take into account delays on the previous Day 2. Under this assumption, to select the departure time on Day 3, the drivers only use the Day 2 experience. On Day 2, departing at 7:30 am certainly led to much better results that leaving for work at 7:15 am. So, under this assumption, on Day 3, most drivers will switch to 7:30 am departure time. As a result, most of them will be again 15 minutes late for work, with the exception of those who left home earlier, at 7:15 am. Since on Day 3, leaving at 7:15 am was clearly much preferable than leaving for work at 7:30 am, on the next Day 4, most drivers will again leave at 7:15 am, etc.

In this analysis, we get the same non-converging fluctuations as we had in the previous section, but this time, we understand the reason for these fluctuations: the fluctuations are caused by the simplifying assumption that the drivers' behavior is determined only by the previous moment of time.

In reality, when the drivers choose departure times, they take into account not only the traffic congestions on the day before, but also traffic congestions on several previous days. When a driver adjusts to the new environment (e.g., to the new city), he or she takes into account not just a single previous day, but rather all the previous days of driving in this new environment.

It is reasonable to assume that all these previous days are weighted equally. Let us describe this assumption in precise terms. We start with the set M_1 of O-D matrices which describe the number of drivers leaving at different time intervals on Day 1, when the drivers follow their original departure times. Similarly to the above text, let us denote the set of O-D matrices describing the drivers on Day i by M_i.

Suppose that we already know the O-D matrices M_1, M_2, ..., M_i which describe the number of drivers leaving at different time intervals at days $1, \ldots, i$. Since the drivers weigh all these previous days equally, they estimate the expected traffic E_i as the average of the previous traffics: $E_i = \frac{1}{i} \cdot (M_1 + \ldots + M_i)$.

The drivers use this expected traffic E_i to make their departure time choices. We have already described the corresponding procedure, and we have denoted the resulting transformation of O-D matrices by F. So, we can conclude that the O-D matrices M_{i+1} corresponding to the new departure times have the form $M_{i+1} = F(E_i)$.

Thus, we arrive at a new iterative procedure that takes into account departure time choice when making traffic assignments. In this procedure,

- we start with the O-D matrices M_1 which describe the original departure times; these O-D matrices can be obtained if we multiply the daily O-D matrix by the original values of the K-factors;
- then, for $i = 2, 3, \ldots$, we repeat the following procedure: first, we compute the average $E_i = \frac{1}{i} \cdot (M_1 + \ldots + M_i)$, and then we compute $M_{i+1} = F(E_i)$;
- after the iterations stop, we use the resulting set of O-D matrices to describe the resulting traffic assignments.

Our experiments on the "toy" road network and on the actual El Paso road network confirmed that this procedure converges. An important question is when to stop iterations:

- The more iterations we perform, the closer we are to the desired "equilibrium" traffic assignment.
- However, each iteration requires a reasonably large computation time on TransCAD, so it is desirable to limit the number of iterations.

To find a reasonable stopping criterion, let us recall that the main objective of our task is to help with traffic planning decisions. To help with these decisions, we must be able to predict future consequences of different road improvement plans. Thus, the objective is to deal with the O-D matrices which describe future drivers' behavior. The only way to get such future matrices is by prediction. Prediction cannot be very accurate. At best, we can predict the accuracy of the future traffic with the accuracy of 10–15%. Thus, it makes sense to stop iterations when we have already achieved this accuracy, i.e., when the difference between the O-D matrices E_i (based on which we make the plans at moment $i + 1$) and the resulting matrices M_{i+1} is smaller than (or equal to) 10–15% of the size of the matrices themselves.

As a measure of the difference between the matrices E_i and M_{i+1}, it is reasonable to take the root mean square difference, i.e., the value $d(E_i, M_{i+1})$ determined by the formula $d^2(E_i, M_{i+1}) = \frac{1}{N} \cdot \sum_{j=1}^{N} (e_j - m_j)^2$, where N is the total number of components in the corresponding matrices (i.e., of all tuples consisting of a time interval and an O-D pair), and e_j and m_j are these components. Similarly, as a measure of the size of a set E of matrices, it is reasonable to take its root mean square value, i.e., the value $v(E)$ determined by the formula $v^2(E) = \frac{1}{N} \cdot \sum_{j=1}^{N} e_j^2$. To speed up computations, we only compute the sizes $v(M_1)$ and $v(M_2)$ for the first two iterations, and use the largest of the two resulting sizes as an estimate for the size in general. In other words, we stop when $d(E_i, M_{i+1}) \le 0.1 \cdot \max(v(M_1), v(M_2))$.

How to take into account departure time choice when making traffic assignments: final idea and the resulting algorithm. In the previous text, we described the algorithm for taking into account departure time choice when making traffic assignments. The advantage of this algorithm is that it converges. However, from the computational viewpoint, this algorithm has a serious limitation. To implement the above algorithm, we must store the sets of O-D matrices M_1, M_2, ..., M_i corresponding to different iterations. For a large city-wide road network, we need to store information about many O-D pairs at several different time intervals. For example, the standard El Paso network has 681 zones, so we need to store the information about each of the 681×681 O-D pairs at each of, say, 12 time intervals, and we most store as many different pieces of this information as there are iterations – which may be in dozens. Storing, accessing, and processing all this information requires a large amount of computation time.

It is therefore desirable to reformulate the above algorithm in such a way as to avoid this excessive storage. We will show that such a simplification is indeed possible. The idea for this simplification comes from the fact that once we know the previous average value $E_i = \frac{1}{i} \cdot (M_1 + \ldots + M_i)$, and we have computed the new matrices $M_{i+1} = F(E_i)$, we do not need to repeat all the additions to compute the new average $E_{i+1} = \frac{1}{i+1} \cdot (M_1 + \ldots + M_i + M_{i+1})$.

Indeed, the expression for E_{i+1} can be reformulated as follows:

$$E_{i+1} = \frac{1}{i+1} \cdot ((M_1 + \ldots + M_i) + M_{i+1}),$$

and, by definition of E_i, we have $M_1 + \ldots + M_i = i \cdot E_i$. Thus, to compute the new average E_{i+1}, we can use the simplified formula

$$E_{i+1} = \frac{1}{i+1} \cdot (i \cdot E_i + M_{i+1}) = E_i \cdot \left(1 - \frac{1}{i+1}\right) + M_{i+1} \cdot \frac{1}{i+1}.$$

Since $M_{i+1} = F(E_i)$, we can reformulate the iterative procedure in terms of the average matrices E_i as follows: $E_{i+1} = E_i \cdot \left(1 - \frac{1}{i+1}\right) + F(E_i) \cdot \frac{1}{i+1}$. Taking into account that $E_1 = M_1$, we arrive at the following algorithm:

- we start with the O-D matrices E_1 which describe the original departure times; these O-D matrices can be obtained if we multiply the daily O-D matrix by the original values of the K-factors;
- then, for $i = 2, 3, \ldots$, we repeat the following procedure: first, we compute $F(E_i)$, and then $E_{i+1} = E_i \cdot \left(1 - \dfrac{1}{i+1}\right) + F(E_i) \cdot \dfrac{1}{i+1}$;
- we stop when $d(E_i, F(E_i)) \le 0.1 \cdot \max(v(E_1), v(E_2))$.
- after the iteration stop, we use the resulting set of O-D matrices E_i to describe the resulting traffic assignments.

Comment. As we show in Appendix B, this iterative procedure is, in some reasonable sense, an optimal algorithm for computing the fixed point of the mapping F.

6. Taking Uncertainty into Account

Need to consider uncertainty. In the previous text, we consider deterministic traffic models, in which the link travel time is uniquely determined by the traffic volume. Real-life traffic, however, is non-deterministic. To have more accurate predictions of travel times, we must take this non-determinism into account and consider stochastic traffic models.

In a stochastic traffic model, the BPR formula only describes the *average* travel time $\bar{t}$:

$$\bar{t} = t^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^{\beta}\right].$$

The stochastic nature of traffic means the actual travel time t may differ from this average value $\bar{t}$. We must therefore describe not only how the *average* travel time $\bar{t}$ depends on t_f, v, and c, but also how the *deviations* $t - \bar{t}$ from this average depend on these parameters. For example, we may want to describe how the standard deviation of the travel time t – or some other statistical characteristic – depends on these parameters.

It turns out that several seemingly reasonable models of this dependence are faulty because the predicted travel times drastically change when we simply subdivide the road links without making any changes in the actual traffic.

In this text, we describe this phenomenon, and we describe how to set up this dependence in such a way that a simple subdivision of a road link will no longer affect the resulting travel times.

We can have different subdivision into road links. Traffic networks in a big city are usually very complicated, with lots of small roads. As a result, the fully detailed simulation of a traffic network would require a large amount of computation time.

It is well known, however, that in practice, there is no need for such a detailed simulation: it is well known that it is sufficient to divide the city into zones and consider only traffic between the zones. The size of the zone depends on the amount and direction of traffic in this zone.

Once we decided how to divide the city into zones, each major road is then naturally subdivided into *road links*, i.e., pieces of this road within each zone.

In busy downtown areas, we may have a popular restaurant in one block and a big office in a neighboring block, with completely different traffic patterns. So, in order to accurately predict downtown-related traffic, we may need to have zones of the size of a few city blocks.

On the other hand, e.g., in a large residential area, we usually get the same pattern of traffic in all its parts: traffic leaving to work in the morning and traffic coming back in the afternoon. As a result, for such areas, it is sufficient to consider larger residential communities as single zones.

For deterministic traffic models, the resulting travel times do not change much if we switch to a finer subdivision into zones: a known fact. Once we come up with zones which provide a reasonable description of the traffic patterns, we can get reasonably good predictions of the traffic volumes and travel times.

If we still have additional computational power, we can consider smaller-size zones. In this case, the original road links are further subdivided into smaller-size links. If we use such a refined model, we get an even more accurate prediction of the travel times.

However, we know that the estimates coming from the original model still provide a reasonably accurate description of the travel times.

For deterministic traffic models, the resulting travel times do not change much if we switch to a finer subdivision into zones: a mathematical explanation. For a deterministic model, one of the reasons for this accuracy is that, because of the above formula for t, the travel time t predicted by the model does not depend on how exactly we subdivide the road into road links – as long as this subdivision remains reasonable in the sense that the traffic volume and the traffic capacity does not change much within this link.

Indeed, let us assume that we start with a single link of length L in the original model and then decided to subdivide it into several sublinks of length $L_1, \ldots, L_n$ – for which $L = L_1 + \ldots + L_n$. In the original model, the travel time t along this link is predicted directly – by using the above formula

$$t = t^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^{\beta}\right].$$

In the new model, we predict individual travel times $t_1, \ldots, t_n$ along different sublinks and then predict the resulting overall travel time as $t_1 + \ldots + t_n$.

Let us show that in this case, the originally predicted travel time t is equal to the total travel time $t_1 + \ldots + t_n$ predicted by the new model.

We assume that the traffic volume v and traffic capacity c are the same for all these sublinks, the only think which is different is the free flow travel time. In other words, the predicted travel times along sublinks take the form

$$t_i = t_i^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^{\beta}\right].$$

In general, the free flow travel time t^f is determined by the length L of the road link and the speed limit s along this link: $t^f = \dfrac{L}{s}$. Similarly, for each sublink, we have $t_i^f = \dfrac{L_i}{s}$.

Since $L = L_1 + \ldots + L_n$, we conclude that $t^f = t_1^f + \ldots + t_n^f$. Thus, we conclude that

$$t_1 + \ldots + t_n = t_1^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^\beta\right] + \ldots + t_n^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^\beta\right] =$$

$$(t_1^f + \ldots + t_n^f) \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^\beta\right] = t^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^\beta\right].$$

So, the originally predicted travel time t is indeed equal to the total travel time $t_1 + \ldots + t_n$ predicted by the new model.

Stochastic case: brief introduction. In the deterministic case, the driver selects a route for which the expected travel time is the shortest.

According to decision theory, in the general situation with stochastic uncertainty case, preferences of a person can be described by a special *utility function* which assigns, to each possible result x, a number $U(x)$ describing the "utility" of this result for this person; a person then selects an action for which the expect value of utility is the largest.

In transportation situations, the main parameter of interest to the drive is the overall travel time, so the utility depends on the travel time t: $U = U(t)$. To make the stochastic formulation of the transportation problems similar to the deterministic ones (in which the objective is to *minimize* travel time), researchers usually replace the problem of maximizing utility with an equivalent problem of minimizing disutility $u(t)$ which is defined as $u(t) = -U(t)$. Usually, an exponential disutility function is used $u(t) = A \cdot \exp(\alpha \cdot t)$; see, e.g., (Mirchandani and Soroush, 1987; Tatineni, 1996; Tatineni et al., 1997). The justification for using such functions is given in Appendix C.

Random deviations $t_i - \overline{t}_i$ for different links are usually caused by different reasons; so traditionally, the travel times t_i on different links $t_1, \ldots, t_n$ along the path are assumed to be independent random variables. Thus, the expected disutility of a path

$$\overline{u} = E[\exp(\alpha \cdot t)] = E[\exp(\alpha \cdot (t_1 + \ldots + t_n)] = E[\exp(\alpha \cdot t_1) \cdot \ldots \cdot \exp(\alpha \cdot t_n)]$$

can be represented as a product

$$\overline{u} = E[\exp(\alpha \cdot t_1)] \cdot \ldots \cdot E[\exp(\alpha \cdot t_n)].$$

Minimizing the product is equivalent to minimizing its logarithm, i.e., the sum

$$s = \ln(E[\exp(\alpha \cdot t_1)]) + \ldots + \ln(E[\exp(\alpha \cdot t_n)]).$$

In the deterministic case, $E[\exp(\alpha \cdot t)] = \exp(\alpha \cdot t)$ hence $\ln(E[\exp(\alpha \cdot t)]) = \alpha \cdot t$. So, to make the problem more similar to the deterministic one, we can divide each logarithm by α – dividing the minimizing function by a positive function does not change where the minimum is attained.

Thus, selecting of a route can be described in a form which is very similar to selecting a deterministic route, but with $\widetilde{t}_i \stackrel{\text{def}}{=} \frac{1}{\alpha} \cdot \ln(E[\exp(\alpha \cdot t_1)])$ instead of the original travel times.

We know that the deviations $t - \overline{t}$ are usually relatively small. Thus, to simplify the above expression, we can substitute $t = \overline{t} + (t - \overline{t})$ into the formula, expand the functions $\exp(z)$ and $\ln(z)$ into Taylor series and keep only the few first (major) terms in the expansion. Specifically, we have

$$\exp(\alpha \cdot t) = \exp(\alpha \cdot \overline{t}) \cdot \exp(\alpha \cdot (t - \overline{t})).$$

Here, the first factor does not depend on the random variable at all, so from the viewpoint of taking an expected value, it is simply a constant:

$$E[\exp(\alpha \cdot t)] = \exp(\alpha \cdot \overline{t}) \cdot E[\exp(\alpha \cdot (t - \overline{t}))].$$

We use the Taylor expansion of the exponential function:

$$\exp(z) = 1 + z + \frac{z^2}{2!} + \ldots = 1 + z + \frac{z^2}{2} + \ldots$$

Thus,

$$\exp(\alpha \cdot (t - \overline{t})) \approx 1 + \alpha \cdot (t - \overline{t}) + \frac{\alpha^2 \cdot (t - \overline{t})^2}{2},$$

and

$$E[\exp(\alpha \cdot (t - \overline{t}))] \approx 1 + \alpha \cdot E[t - \overline{t}] + \frac{\alpha^2 \cdot E[t - \overline{t})^2]}{2}.$$

By definition, $E[t - \overline{t}] = \overline{t} - \overline{t} = 0$, and $E[(t - \overline{t})^2]$ is the variance V. Thus, in our approximation,

$$E[\exp(\alpha \cdot t)] = \exp(\alpha \cdot \overline{t}) \cdot \left(1 + \frac{\alpha^2}{2} \cdot V\right).$$

So,

$$\frac{1}{\alpha} \cdot \ln(E[\exp(\alpha \cdot t)]) = \widetilde{t} + \frac{1}{\alpha} \cdot \ln\left(1 + \frac{\alpha^2}{2} \cdot V\right).$$

Using the Taylor expansion of the logarithm function $\ln(1 + z) = z + \ldots$, we conclude that

$$\frac{1}{\alpha} \cdot \ln(E[\exp(\alpha \cdot t)]) = \widetilde{t} + \frac{\alpha}{2} \cdot V.$$

Thus, minimizing the sum of these logarithmic expressions is equivalent to minimizing the sum of the expressions

$$\widetilde{t} = \overline{t} + \frac{\alpha}{2} \cdot V.$$

In other words, to make stochasticity into account, to each link's travel time, we add its variance (with an appropriate weight $\alpha/2$).

A seemingly natural description. In the case of the free flow traffic, there is no uncertainty; uncertainty occurs only if we have some volume on the road link – i.e., when the travel time t exceeds the free flow travel time t^f. Intuitively, the larger this excess $t - t^f$, the larger this uncertainty.

At first glance, it may seem natural to pick a proportion r_0 (e.g., 20%) and assume that for every link, the actual value $t - t_f$ can deviate by about $\pm 20\%$ (or whatever r is) from the average. In more precise terms, the standard deviation $\sigma \stackrel{\text{def}}{=} \sqrt{V}$ of the travel time is equal to $r_0 \cdot (\overline{t} - t^f)$.

Since $\sigma = \sqrt{V} = r_0 \cdot (\overline{t} - t^f)$, we conclude that $V = r_0^2 \cdot (\overline{t} - t_f)^2$.

Problem with seemingly natural assumption. Let us show that this seemingly natural assumption leads to counter-intuitive conclusions. Indeed, let us assume that we have two one-link

routes of equal quality leading from point A to point B, with the same free flow time t^f, same capacity c, and the same traffic volume v. In this case, for both links, we have the same expected travel time $\overline{t}$ and hence, the same variance – so, the values of the resulting minimized function are the same for both routes:

$$\widetilde{t}^{(1)} = \widetilde{t}^{(2)} = \overline{t} + \frac{\alpha}{2} \cdot r_0^2 \cdot (\overline{t} - t_f)^2.$$

Intuitively, if we subdivide one of the links into two equal sublinks of equal length (without changing anything of substance) we should end up with exactly the same selection. In reality, if we subdivide the first link, then for this link, we will have both t and t_f divided by 2: $\overline{t}_1 = \overline{t}_2 = \dfrac{\overline{t}}{2}$ and $t_1^f = t_2^f = \dfrac{t^f}{2}$. Hence, the variance V (proportional to $(t - t_f)^2$) will divide by 4. As a result, for each of these links, we get

$$\widetilde{t}_1 = \widetilde{t}_2 = \frac{\overline{t}}{2} + \frac{\alpha}{2} \cdot r_0^2 \cdot \frac{(\overline{t} - t_f)^2}{4}.$$

By adding these two values, we get the minimized value $\widetilde{t} = \widetilde{t}_1 + \widetilde{t}_2$ for the whole two-link route:

$$\widetilde{t} = \overline{t} + \frac{\alpha}{2} \cdot r_0^2 \cdot \frac{(\overline{t} - t_f)^2}{2}.$$

In this expression, the term proportional to the variance is twice smaller than for the second route, so this route will be selected.

Alternatively, if we keep the first route whole but subdivide the second route, we get a clear preference for the second route. Thus, the route selection depends on the exact subdivision into links – hence our seemingly natural assumption is really counter-intuitive.

Proposed solution. Our objective is to find a reasonable expression for the term

$$\widetilde{t} = \frac{1}{\alpha} \cdot \ln(E[\exp(\alpha \cdot t)]).$$

In general, this expression can depend on the free flow time t^f and on the average time $\overline{t}$.

As we have mentioned, in the absence of the traffic flow, when the travel time consists 100% of the free flow time t^f, there is no stochasticity. The larger the proportion of the excess time, i.e., the larger the ratio $r \stackrel{\text{def}}{=} \dfrac{\overline{t} - t^f}{t^f}$, the more stochasticity there is. Thus, it is reasonable to describe the desired expression for $\widetilde{t}$ in terms of t^f and r.

By definition of r, we have $\overline{t} - t^f = r \cdot t^f$ hence $\overline{f} = (1 + r) \cdot t^f$; so, once we know the dependence of $\widetilde{t}$ on t^f and $\overline{t}$, we can find its dependence on t^f and r as well. Thus, it is reasonable to claim that $\widetilde{t} = F(t^f, r)$ for some yet-to-be-determined function F.

The first desired property of the function F is that if the average time coincides with the free flow time, then there is no stochasticity, and $\widetilde{t} = t$. In other words, we must have $F(t, 0) = t$ for all t.

The second desired property is that when we subdivide a link into two sublinks, without changing the flow or capacity (and hence, without changing the ratio r), then the sum of the resulting values

$\widetilde{t}_1 + \widetilde{t}_2$ should be equal to the original value $\widetilde{t}$: $F(t_1^f, r) + F(t_2^f, r) = F(t_1^f + t_2^f, r)$. For each r, we get an equation $F'(a+b) = F'(a) + F'(b)$ for a monotonic function $F'(a) \stackrel{\text{def}}{=} F(a, r)$ hence (Aczel, 2006) $F'(a) = k \cdot a$ for some constant $k(r)$ which may depend on r. The fact that $F(t, 0) = t$ means that $k(0) = 1$.

In other words, we conclude that $\widetilde{t} = F(t^f, r) = t^f \cdot k(r)$. We know that $r = a \cdot \left(\frac{v}{c}\right)^\beta$, thus,

$$\widetilde{t} = t^f \cdot k\left(\left(\frac{v}{c}\right)^\beta\right).$$

Similarly to the above case, we can expand the dependence $k(r)$ into Taylor series and keep the first few terms in this expansion. Since $k(0) = r$, we conclude that $k(r) = 1 + a_1 \cdot r + a_2 \cdot r^2 + \ldots$, hence

$$\widetilde{t} = 1 + a_1 \cdot a \cdot \left(\frac{v}{c}\right)^\beta + a_2 \cdot a^2 \cdot \left(\frac{v}{c}\right)^{2\beta}.$$

Conclusion. The effect of stochasticity on the transportation problem can be described as follows:

- in the deterministic case, drivers select a route for which the overall travel time $\overline{t} = \overline{t}_1 + \ldots + \overline{t}_n$ is the smallest, where $\overline{t}_i = t_i^f \cdot \left[1 + a \cdot \left(\frac{v_i}{c_i}\right)^\beta\right]$;
- in the stochastic case, drivers select a route for which the expression $\widetilde{t} = \widetilde{t}_1 + \ldots + \widetilde{t}_n$ is the smallest, where $\widetilde{t}_i = t_i^f \cdot \left[1 + a_1 \cdot a \cdot \left(\frac{v_i}{c_i}\right)^\beta + b \cdot \left(\frac{v_i}{c_i}\right)^{2\beta}\right]$.

Thus, we can use the standard traffic assignment algorithms with a modified travel time function to find the corresponding traffic assignment.

Comment. Our experiments show that $a_1 \approx 1.4$ and $b \cdot 0$. So, to take the uncertainty into account, it is sufficient to replace the original value $a \approx 0.15$ in the BPR formula with the new value $a_1 \cdot a \approx 0.21$.

Acknowledgements

This work was supported in part by NSF grants HRD-0734825, EAR-0225670, and EIA-0080940, by Texas Department of Transportation grant No. 0-5453, by the Japan Advanced Institute of Science and Technology (JAIST) International Joint Research Grant 2006-08, and by the Max Planck Institut für Mathematik.

References

Aczel, J. *Lectures on Functional Equations and Their Applications*, Dover, New York, 2006.

Ahuja, R. K., T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

Appa, G. M. The transportation problem and its variants. *Oper. Res. Quarterly*, 24:79–99, 636–639, 1973.

Berinde, V. *Iterative approximation of fixed points*, Editura Efemeride, Baia Mare, 2002.

Braess, D. Über ein Paradox der Verkerhsplannung. *Unternehmenstorchung*, 12:258–268, 1968.

Charnes, A., and D. Klingman. The more-for-less paradox in the distribution model. *Cahiers du Centre d'Etudes de Recherche Operationelle*, 13:11–22, 1971.

Cheu, R., V. Kreinovich, Y.-C. Chiu, R. Pan, G. Xiang, S. Bhupathiraju, and S. R. Manduva. *Strategies for Improving Travel Time Reliability*, Texas Department of Transportation, Research Report 0-5453-R2, August 2007.

Chipman, J. The foundations of utility. *Econometrica*, 28:193–224, 1960.

Debreu, G. Review of R. D. Luce, "Individual Choice Behavior". *American Economic Review*, 50:186–188, 1960.

Jaynes, E. T., and G. L. Bretthorst (ed.), *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

Keeney, R. L., and H. Raiffa, *Decisions with Multiple Objectives*, John Wiley and Sons, New York, 1976.

Kohlenbach, U. Uniform asymptotic regularity for Mann iterates. *J. Math. Anal. Appl.*, 279(8):531–544, 2003.

Kohlenbach, U. Some computational aspects of metric fixed point theory. *Nonlinear Analysis*, 61(5):823–837, 2005.

Kohlenbach, U. Effective uniform bounds from proofs in abstract functional analysis", In: B. Cooper, B. Loewe, and A. Sorbi (eds.), *New Computational Paradigms: Changing Conceptions of What is Computable*, Springer Verlag, Berlin-Heidelberg-New York, 2007.

Kohlenbach, U.: *Applied Proof Theory: Proof Interpretations and their Use in Mathematics*. Springer Verlag, Berlin-Heidelberg, 2008.

Kohlenbach. U., and B. Lambov. Bounds on iterations of asymptotically quasi-nonexpansive mappings. In: J. G. Falset, E. L. Fuster, and B. Sims (eds.), *Proc. International Conference on Fixed Point Theory and Applications, Valencia 2003*, pages 143–172, Yokohama Publishers, 2004.

Luce, D. *Individual Choice Behavior*, John Wiley and Sons, New York, 1959.

Luce, D., and P. Suppes, Preference, utility, and subjective probability, In: D. Luce, R. Bush, and E. Galanter (eds.), *Handbook on Mathematical Psychology*, pages 249–410, John Wiley and Sons, New York, 1965.

McFadden, D. Conditional logit analysis of qualitative choice behavior, In: P. Zarembka (ed.), *Frontiers in Econometrics*, pages 105–142, Academic Press, New York, 1974.

McFadden, D. Economic choices, *American Economic Review*, 91:351–378, 2001.

Mirchandani, P., and H. Soroush, Generalized traffic equilibrium with probabilistic travle times and perceptions, *Transportation Science*, 21(3):133–152, 1987.

Noland, R. B. and K. A. Small. Travel time uncertainty, departure time choice, and the cost of morning commutes, *Transportation Researh Record*, 1493:150–158, 1995.

Noland, R. B., K. A. Small, P. M. Koseknoja, and X. Chu. Simulating travel reliability, *Regional Science and Urban Economics*, 28:535–564, 1998.

Pratt, J. W. Risk Aversion in the Small and in the Large. Econometrica, 32:122–136, 1964.

Raiffa, H. *Decision Analysis*, Addison-Wesley, Reading, Massachusetts, 1970.

Sheffi, Y. *Urban Transportation Networks*, Prentice Hall, Englewood Cliffs, NJ, 1985.

Su, Y., and X. Qin, Strong convergence theorems for asymptotically nonexpansive mappings and asymptotically nonexpansive semigroups. *Fixed Point Theory and Applications*, 2006, Article ID 96215, pp. 1–11.

Szwarc, W. The transportation paradox. *Naval Res. Logist. Quarterly*, 18:185–202, 1971.

Tatineni, M. *Solution properties of stochastic route choice models*, Ph.D. Dissertation, Department of Civil Engineering, University of Illinois at Chicago, 1996.

Tatineni, M., D. E. Boyce, and P. Mirchandani, Comparison of deterministic and stochastic traffic loading models, *Transportation Research Record*, 1607:16–23, 1997

Train, K. *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, Massachusetts, 2003.

Wadsworth, H. M. (ed.), *Handbook of statistical methods for engineers and scientists*, McGraw-Hill Publishing Co., New York, 1990.

Appendix

A. Logit Discrete Choice Model: Towards a New Justification

Traditional approach to decision making. In decision making theory, it is proven that under certain reasonable assumption, a person's preferences are defined by his or her *utility function* $U(x)$ which assigns to each possible outcome x a real number $U(x)$ called *utility*; see, e.g., (Keeney and Raiffa, 1976; Raiffa, 1970). In many real-life situations, a person's choice s does not determine the outcome uniquely, we may have different outcomes $x_1, \ldots, x_n$ with probabilities, correspondingly, $p_1, \ldots, p_n$.

For example, drivers usually select the path with the shortest travel time. However, when a driver selects a path s, the travel time is often not uniquely determined: we may have different travel times $x_1, \ldots, x_n$ with corresponding probabilities $p_1, \ldots, p_n$.

For such a choice, we can describe the utility $U(s)$ associated with this choice as the expected value of the utility of outcomes: $U(s) = E[U(x)] = p_1 \cdot U(x_1) + \ldots + p_n \cdot U(x_n)$. Among several possible choices, a user selects the one for which the utility is the largest: a possible choice s is preferred to a possible choice s' (denoted $s > s'$) if and only if $U(s) > U(s')$.

It is important to mention that the utility function is not uniquely determined by the preference relation. Namely, for every two real numbers $a > 0$ and b, if we replace the original utility function $U(x)$ with the new one $U'(x) \stackrel{\text{def}}{=} a \cdot U(x) + b$, then for each choice s, we will have

$$U'(s) = E[a \cdot U(x) + b] = a \cdot E[U(x)] + b = a \cdot U(s) + b$$

and thus, $U'(s) > U'(s')$ if and only if $U(s) > U(s')$.

Situations in which we can only predict probabilities of different decision. One important application of decision making theory is predicting the user decisions. If we know the exact values $U(s)$ of the utilities, then we can predict the exact choice. For example, if the user has to choose between alternatives s and s', then the user chooses s if $U(s) \geq U(s')$ and s' if $U(s) \leq U(s')$.

In practice, we do not know the exact values $U(s)$ of the user's utility, we only know the approximate values $V(s) \approx U(s)$. Due to the difference between the observed (approximate) values $V(s)$ and the actual (unknown) values $U(s)$, we are no longer able to uniquely predict the user's behavior: e.g., even when $V(s) > V(s')$, we may still have $U(s) < U(s')$, and thus, it is possible that the user will prefer s.

If the differences $V(s) - U(s)$ and $V(s') - U(s')$ are small, then for $V(s) \gg V(s')$, we can be reasonably sure that $U(s) > U(s')$ and thus, that the user will select s. Similarly, if $V(s) \ll V(s')$, we can be reasonably sure that $U(s) < U(s')$ and thus, that the user will select s'. However, when the values $V(s)$ and $V(s')$ are close, then there is a certain probability that $U(s) > U(s')$ and thus, that the user will select s, and there is also a certain probability that $U(s) < U(s')$ and thus, that the user will select s'.

In this situation, based on the (approximate) utility values $V(s)$ and $V(s')$, we cannot exactly predict whether the user will prefer s or s' – because for the same values of $V(s)$ and $V(s')$, the user can prefer s and the user can also prefer s'. The best we can do in this situation is to predict the *probability* $P(s > s')$ of selecting s over s'.

Discrete choice: a formal description of the problem. Let us formulate the problem in precise terms. We have n different alternatives $s_1, \ldots, s_n$. For each of these alternative s_i, we know the (approximate) utility value $V_i \stackrel{\text{def}}{=} V(s_i)$. Based on these utility values $V(s_1), \ldots, V(s_n)$, we would like to predict the probability p_i that a user will select the alternative s_i.

Models used for such prediction are usually called *discrete choice models* (Train, 2003).

Invariance requirements in discrete choice models. As we have mentioned, the utility function is not uniquely determined by the preference relation. Namely, whenever the original utility function $U(s)$ describes the user's preference, then, for every $a > 0$ and b, the new function $U'(s) = a \cdot U(s) + b$ also describes the same preference. In other words, we can shift all the values of the utility function $u(s) \to U(s) + b$, and we can re-scale all the values $U(s) \to a \cdot u(s)$, and the resulting utility function will still describe the same preferences.

It is therefore reasonable to assume that if we shift the values of the approximate utility function, i.e., if we replace the original values $V(s_i)$ with the new values $V'(s_i) = V(s_i) + b$, then we should get the same preference probabilities:

$$p_i(V(s_1), V(s_2) \ldots, V(s_n)) = p_i(V(s_1) + b, V(s_2) + b, \ldots, V(s_n) + b).$$

In particular, if we take $b = -V(s_1)$, then we conclude that

$$p_i(V(s_1), V(s_2) \ldots, V(s_n)) = p_i(0, V(s_2) - V(s_1), \ldots, V(s_n) - V(s_1)),$$

i.e., that the probabilities depend only on the *differences* between the utility values – but not on the values themselves.

At first glance, it may seem reasonable to similarly require that the probability not change under re-scaling. However, in this case, re-scaling does not make intuitive sense, because we have a natural scale. For example, as a unit for such a scale, we can choose a standard deviation of the difference $U(s) - V(s)$ between the (unknown) actual utility $U(s)$ and the (known) approximate value of this utility $V(s)$.

In line with this analysis, in discrete choice models, it is usually assumed that the probabilities do not change with shift but it is *not* assumed that these probabilities are scale-invariant.

Logit: the most widely used discrete choice model. The most widely used discrete choice model is a *logit* model in which

$$p_i(V_1, \ldots, V_n) = \frac{e^{\beta \cdot V_i}}{\sum\limits_{j=1}^{n} e^{\beta \cdot V_j}} \tag{1}$$

for some parameter β. This model was first proposes in (Luce, 1959).

Logit: original justification. In (Luce, 1959), this model was justified based on the assumption of *independence of irrelevant alternatives*, according to which the relative probability of selecting s_1 or s_2 should not change if we add a third alternative s_3. In formal terms, this means that the probability of selecting s_1 out of two alternatives s_1 and s_2 should be equal to the conditional probability of selecting s_1 from three alternatives s_1, s_2, and s_3 under the condition that either s_1 or s_2 are selected.

It can be proven that under this assumption, the ratio p_i/p_j of the probabilities p_i and p_j should only depend on V_i and V_j; moreover, that we must have $p_i/p_j = f(V_i)/f(V_j)$ for some function $f(z)$. The requirement that this ratio be shift-invariant then leads to the conclusion that $f(z) = e^{\beta \cdot z}$ for some β – and thus, to the logit model.

Limitations of the original justification. At first glance, the above independence assumption sounds reasonable (and it is often reasonable). However, there are reasonable situations where this assumption is counter-intuitive; see, e.g., (Chipman, 1960; Debreu, 1960; Train, 2003).

For example, assume that in some cities, all the buses were originally blue. To get from point A to point B, a user can choose between taking a taxi (s_1) and taking a blue bus (s_2). A taxi is somewhat better to this user, so he selects a taxi with probability $p_1 = 0.6$ and a blue bus with the remaining probability $p_2 = 1 - 0.6 = 0.4$. In this case, the ratio p_1/p_2 is equal to 1.5.

Suppose now that the city decided to buy some new buses, and to paint them red. Let us also suppose that the comfort of the travel did not change, the buses are exactly the same. From the common sense viewpoint, it does not matter to the user whether buses are blue or red, so he should still select a taxi with probability $p_1 = 0.6$ and buses with probability 0.4. However, from the purely mathematical viewpoint, we now have *three* options: taking a taxi (s_1), taking a blue bus (s_2), and taking a red bus (s_3). Here, the probability of taking a bus is now $p_2 + p_3 = 0.4$. Hence, $p_2 < 0.4$ and so, the ratio p_1/p_2 is different from what we had before – contrary to the above independence assumption.

Current justification. An alternative justification for logit started with the unpublished result of Marley first cited in (Luce and Suppes, 1965). Marley has shown that if we assume that the approximation errors $\varepsilon(s) \stackrel{\text{def}}{=} U(s) - V(s)$ are independent and identically distributed, and if this distribution is the Gumbel distrubution, then the probability of selecting s_i indeed follows the logit formula.

Gumbel distribution can be characterized by the cumulative distribution function $F(\varepsilon) = e^{-e^{-\varepsilon}}$; it is a known distribution of extreme values.

In 1974, McFadden (McFadden, 2001) showed that, vice versa, if we assumed that the approximation errors $\varepsilon(s)$ are independent and identically distributed, and the choice probabilities are described by the logit formula, then the errors $\varepsilon(s)$ must follow the extreme value (Gumbel) distribution.

This justification was one of the main achievements for which D. McFadden received a Nobel prize in 2001 (McFadden, 2001).

Limitations of the current justification. The problem with this justification is that the logit model is known to work well even in the cases when different approximation errors are differently distributed; see, e.g., (Train, 2003).

For such situations, the only known alternative explanation is Luce's original one. The main limitation of this explanation was that it is based on the independence assumption. This is not so critical if we have three or more alternatives. Indeed, in this case, the empirical logit formula (that we are trying to explain) satisfies this assumption, so making this assumption in the situations when the logit formula holds makes sense.

This limitation, however, becomes crucial if we only consider the case of two alternatives. In this case, the independence assumption cannot even be formulated and therefore, Luce's justification does not apply. So, we arrive at the following problem.

Formulation of the problem. We need to come up with a new distribution-free justification for the logit formula, i.e., with a justification that does not depend on the assumption that approximation errors are independent and identically distributed. Such a justification is provided in this paper.

Preliminary analysis. In accordance with the above formulation of the problem, we are interested in the case of $n = 2$ alternatives s_1 and s_2. We know the approximate utility values V_1 and V_2, and we know that the probability p_1 of selecting the first alternative p_1 should only depend on the difference $V_1 - V_2$: $p_1 = F(V_1 - V_2)$ for some function $F(z)$. Our objective is to find this function $F(z)$. Let us first describe reasonable properties of this function $F(z)$.

When s_2 is fixed (hence V_2 is fixed) but the alternative s_1 is improving (i.e., V_1 is increasing), then the probability of selecting s_1 can only increase (or at least remain the same – e.g., if that probability was already equal to 1, it cannot further increase). In other words, as the difference $V_1 - V_2$ increases, the probability $p_1 = F(V_1 - V_2)$ should also increase (or at least remain the same). Thus, it is reasonable to require that the function $F(z)$ should be (non-strictly) increasing.

When s_2 and V_2 are fixed and s_1 becomes better and better, i.e., $V_1 \to +\infty$, then we should select s_1 with probability tending to 1. So, we must have $F(z) \to 1$ as $z \to +\infty$.

Similarly, s_2 and V_2 are fixed, and s_1 becomes worse and worse, i.e., $V_1 - V_2 \to -\infty$, then we should prefer s_2. So, we must have $F(z) \to 0$ as $z \to -\infty$.

Since we only have two alternatives, the probability $p_1 = F(V_1 - V_2)$ and the probability $p_2 = F(V_2 - V_1)$ must always add up to 1. Thus, we must have $F(z) + F(-z) = 1$ for all z.

So, we arrive at the following definition.

Definition 1. *By a* choice function*, we mean a function $F : R \to [0, 1]$ which is (non-strictly) increasing, and for which $F(z) \to 1$ as $z \to +\infty$, $F(z) \to 0$ as $z \to -\infty$, and $F(z) + F(-z) = 1$ for all z.*

Main idea. Our main idea is as follows. Up to now, we have discussed how to *describe* the user's behavior, but often, the ultimate objective is how to *modify* this behavior. For example, in transportation problems, the goal is often to use public transportation to relieve traffic congestion and related pollution. In this case, the problem is not just to estimate the probability of people using public transportation, but to find out how to increase this probability.

One way to increase this probability is to provide incentives. If we want to encourage people to prefer alternative s_1, then we can provide those who select this alternative with an additional benefit of value v_0. In this case, for alternatives $s_i \neq s_1$, the corresponding utility V_i remains the same, but for the alternative s_1, we have a new value of utility $V_1' = V_1 + v_0$.

After this addition, the original probability

$$p_1 = F(V_1 - V_2) \tag{2}$$

of selecting the alternative s_1 changes to a new value

$$p_1' = F(V_1' - V_2) = F(V_1 + v_0 - V_2). \tag{3}$$

These formulas can be simplified if we denote the difference $V_1 - V_2$ between the approximate utility values by ΔV. In these new notations, the original probability

$$p_1 = F(\Delta V) \tag{4}$$

is replaced by the new probability

$$p_1' = F(\Delta V + v_0). \tag{5}$$

This change of probability can be described in general terms: we receive new information – that there are now incentives. Based on this new information, we update our original probabilities p_i of selecting different alternatives s_i.

From the statistical viewpoint (see, e.g., (Jaynes and Bretthorst, 2003; Wadsworth, 1990)), when we receive new information, the correct way of updating probabilities is by using the Bayes formula. Specifically, if we have n incompatible hypotheses $H_1, \ldots, H_n$ with initial probabilities

$$P_0(H_1), \ldots, P_0(H_n), \tag{6}$$

then, after observations E, we update the initial probabilities to the new values:

$$P(H_i \,|\, E) = \frac{P(E \,|\, H_i) \cdot P_0(H_i)}{P(E \,|\, H_1) \cdot P_0(H_1) + \ldots + P(E \,|\, H_n) \cdot P_0(H_n)}. \tag{7}$$

Thus, we should require that the function $F(z)$ be such for which the transition from the old probability (4) to the new probability (5) can be described by the (fractionally linear) Bayes formula (7).

From the main idea to the exact formulas. Let us formalize the above requirement. In the case of two alternatives s_1 and s_2, we have two hypotheses: the hypothesis H_1 that the user will prefer s_1 and the opposite hypothesis H_2 that the user will prefer s_2. Initially, we did now know about any incentives, we only knew the approximate utility V_1 of the first alternative and the approximate utility V_2 of the second alternative. Based on the information that we initially had, we concluded that the probability of the hypothesis H_1 is equal to $p_1 = p(H_1) = F(\Delta V)$ (where $\Delta V = V_1 - V_2$), and the probability of the opposite hypothesis H_2 is equal to $p_2 = p(H_2) = 1 - p_1$.

Now, suppose that learn that there was no incentive to select alternative s_2 and an incentive of size v_0 to select alternative s_1. This new information E changes the probabilities of our hypotheses H_1 and H_2. Namely, according to Bayes formula, after the new information E, the probability p_1 should be updated to the following new value $p_1' = P(H_1 \,|\, E)$:

$$p_1' = \frac{P(E \,|\, H_1) \cdot P(H_1)}{P(E \,|\, H_1) \cdot p_1 + P(E \,|\, H_2) \cdot P(H_2)}. \tag{8}$$

The probability $P(E \,|\, H_1)$ is the conditional probability with which we can conclude that there was an incentive of size v_0 based on the fact that the user actually selected the alternative s_1. This conditional probability is, in general, different for different values v_0. To take this dependence into account, we will denote this conditional probability $P(E \,|\, H_1)$ by $A(v_0)$.

Similarly, the probability $P(E \,|\, H_2)$ is the conditional probability with which we can conclude that there was an incentive of size v_0 for alternative s_1 based on the fact that the user actually

selected the alternative s_2. This conditional probability is also, in general, different for different values v_0. To take this dependence into account, we will denote this conditional probability $P(E \,|\, H_2)$ by $B(v_0)$.

If we substitute the expressions $P(E \,|\, H_1) = A(v_0)$, $P(E \,|\, H_2) = B(v_0)$, $P(H_1) = F(\Delta V)$, and $P(H_2) = 1 - P(H_1) = 1 - F(\Delta V)$ into the above formula (8), then we conclude that

$$p_1' = \frac{A(v_0) \cdot F(\Delta V)}{A(v_0) \cdot F(\Delta V) + B(v_0) \cdot (1 - F(\Delta V))}. \tag{9}$$

On the other hand, once we know that there was an incentive v_0 to select the alternative s_1 and no incentive for the alternative s_2, then we have a better idea of the resulting utilities of the user: namely, the new value of the approximate utility is $V_1 + v_0$ for alternative s_1 and V_2 for the alternative s_2. In accordance with our expression for the choice probability based on the approximate utility values, the new probability of selecting s_1 should be equal to $F((V_1 + v_0) - V_2)$, i.e., to $F(\Delta V + v_0)$ (expression (4)).

If the probability update was done correctly, in full accordance with the Bayes formula, then this new value (4) must be equal to the value (9) that comes from using the Bayes formula. So, we arrive at the following definition:

Definition 2. *A choice function $F(z)$ is called* Bayes correct *if, for every v_0, there exist values $A(v_0)$ and $B(v_0)$ for which*

$$F(\Delta V + v_0) = \frac{A(v_0) \cdot F(\Delta V)}{A(v_0) \cdot F(\Delta V) + B(v_0) \cdot (1 - F(\Delta V))} \tag{10}$$

for all ΔV.

Comment. In other words, we require that the 2-parametric family of functions $F = \left\{ \dfrac{A \cdot F(\Delta V)}{A \cdot F(\Delta V) + B} \right\}$ corresponding to Bayesian updates be *shift-invariant* under a shift $\Delta V \to \Delta V + v_0$.

Theorem 1. *Every Bayes correct choice function $F(z)$ has the form*

$$F(\Delta V) = \frac{1}{1 + e^{-\beta \cdot \Delta V}} \tag{11}$$

for some real number β.

If we substitute $\Delta V = V_1 - V_2$ into this formula, and multiply the numerator and the denominator of the resulting formula by $e^{\beta \cdot V_1}$, then we conclude that for every Bayes correct choice function $F(z)$, we have

$$p_1 = F(V_1 - V_2) = \frac{e^{\beta \cdot V_1}}{e^{\beta \cdot V_1} + e^{\beta \cdot V_2}}. \tag{12}$$

Thus, for the desired case of two alternatives, we indeed provide a new distribution-free justification of the logit formula.

Proof. It is known that many formulas in probability theory can be simplified if instead of the probability p, we consider the corresponding odds

$$O = \frac{p}{1 - p}. \tag{13}$$

(If we know the odds O, then we can reconstruct the probability p as $p = O/(1+O)$.) The right-hand side of the formula (10) can be represented in terms of odds $O(\Delta V)$, if we divide both the numerator and the denominators by $1 - F(\Delta V)$. As a result, we get the following formula:

$$F(\Delta V + v_0) = \frac{A(v_0) \cdot O(\Delta V)}{A(v_0) \cdot O(\Delta V) + B(v_0)}. \tag{14}$$

Based on this formula, we can compute the corresponding odds $O(\Delta V + v_0)$: first, we compute the value

$$1 - F(\Delta V + v_0) = \frac{B(v_0)}{A(v_0) \cdot O(\Delta V) + B(v_0)}, \tag{15}$$

and then divide (14) by (15), resulting in:

$$O(\Delta V + v_0) = c(v_0) \cdot O(\Delta V), \tag{16}$$

where we denoted $c(v_0) \stackrel{\text{def}}{=} A(v_0)/B(v_0)$. It is known (see, e.g., (Aczel, 2006)) that all monotonic solutions of the functional equation (16) are of the form $O(\Delta V) = C \cdot e^{\beta \cdot \Delta V}$. Therefore, we can reconstruct the probability $F(\Delta V)$ as

$$F(\Delta V) = \frac{O(\Delta V)}{O(\Delta V) + 1} = \frac{C \cdot e^{\beta \cdot \Delta V}}{C \cdot e^{\beta \cdot \Delta V} + 1}. \tag{17}$$

The condition $F(z) + F(-z) = 1$ leads to $C = 1$. Dividing both the numerator and the denominator of the right-hand side by $e^{\beta \cdot \Delta V}$, we get the desired formula (11). Q.E.D.

B. Towards an Optimal Algorithm for Computing Fixed Points

Many practical situations eventually reach equilibrium. In many real-life situations, we have dynamical situations which eventually reach an equilibrium.

For example, in economics, when a situation changes, prices start changing (often fluctuating) until they reach an equilibrium between supply and demand.

In transportation, as we have mentioned, when a new road is built, some traffic moves to this road to avoid congestion on the other roads; this causes congestion on the new road, which, in its turn, leads drivers to go back to their previous routes, etc. (Sheffi, 1985).

It is often desirable to predict the corresponding equilibrium. For the purposes of the long-term planning, it is desirable to find the corresponding equilibrium. For example, for the purposes of economic planning, it is desirable to know how, in the long run, oil prices will change if we start exploring new oil fields in Alaska. For transportation planning, it is desirable to find out to what extent the introduction of a new road will relieve the traffic congestion, etc.

In order to describe how we can solve this practically important problem, let us describe this equilibrium prediction problem in precise terms.

Finding an equilibrium as a mathematical problem. To describe the problem of finding the *equilibrium* state(s), we must first be able to describe *all possible* states. In this paper, we assume that we already have such a description, i.e., that we know the set X of all possible states.

We must also be able to describe the fact that many states $x \in X$ are not equilibrium states. For example, if the price of some commodity (like oil) is set up too high, it will become profitable to explore difficult-to-extract oil fields; as a new result, the supply of oil will increase, and the prices will drop.

Similarly, as we have mentioned in the main text, if too many cars move to a new road, this road may become even more congested than the old roads initially were, and so the traffic situation will actually decrease – prompting people to abandon this new road.

To describe this instability, we must be able to describe how, due to this instability, the original state x gets transformed in the next moment of time. In other words, we assume that for every state $x \in X$, we know the corresponding state $f(x)$ at the next moment of time.

For non-equilibrium states x, the change is inevitable, so we have $f(x) \neq x$. The equilibrium state x is the state which does not change, i.e., for which $f(x) = x$. Thus, we arrive at the following problem: We are given a set X and a function $f : X \to X$; we need to find an element x for which $f(x) = x$.

In mathematical terms, an element x for which $f(x) = x$ is called a *fixed point* of the mapping f. So, there is a practical need to find fixed points.

The problem of computing fixed points. Since there is a practical need to compute the fixed points, let us give a brief description of the existing algorithms for computing these fixed points. Readers interested in more detailed description can look, e.g., in (Berinde, 2002).

Straightforward algorithm: Picard iterations. At first glance, the situation seems very simple and straightforward. We know that if we start with a state x at some moment of time, then in the next moment of time, we will get a state $f(x)$. We also know that eventually, we will get an equilibrium. So, a natural thing to do is to simulate how the actual equilibrium will be reached.

In other words, we start with an arbitrary (reasonable) state x_0. After we know the state x_k at the moment k, we predict the state x_{k+1} at the next moment of time as $x_{k+1} = f(x_k)$. This algorithm is called *Picard iterations* after a mathematician who started efficiently using it in the 19 century.

If the equilibrium is eventually achieved, i.e., if in real life the process converges to an equilibrium point x, then Picard's iterations are guaranteed to converge. Their convergence may be somewhat slow – since they simulate all the fluctuations of the actual convergence – but eventually, we get convergence.

Situations when Picard's iterations do not converge. In some important practical situations, Picard iterations do not converge.

The main reason is that in practice, we can have panicky fluctuations which prevent convergence. Of course, one expects fluctuations. For example, if the price of oil is high, then it will become profitable for companies to explore and exploit new oil fields. As a result, the supply of oil will drastically increase, and the price of oil will go down. Since this is all done in a unplanned way, with different companies making very rough predictions, it is highly probable that the resulting oil supply will exceed the demand. As a result, prices will go down, oil production in difficult-to-produce oil areas will become unprofitable, supply will go down, etc.

Such fluctuations have happened in economics in the past, and sometimes, not only they did not lead to an equilibrium, they actually led to deep economic crises.

As we have seen, similar situations happen in transportation as well.

How can we handle these situation: a natural practical solution. If the natural Picard iterations do not converge, this means that in practice, there is too much of a fluctuation. When at some moment k, the state x_k is not an equilibrium, then at the next moment of time, we have a state $x_{k+1} = f(x_k) \neq x_k$. However, this new state x_{k+1} is an not necessarily closer to the equilibrium: it "over-compensates" by going too far to the other side of the desired equilibrium.

For example, we started with a price x_k which was too high. At the next moment of time, instead of having a price which is closer to the equilibrium, we may get a new price x_{k+1} which is too low – and may even be further away from the equilibrium than the previous price.

In practical situations, such things do happen. In this case, to avoid such weird fluctuations and to guarantee that we eventually converge to the equilibrium point, a natural thing is to "dampen" these fluctuations: we know that a transition from x_k to x_{k+1} has gone too far, so we should only go "halfway" (or even smaller piece of the way) towards x_{k+1}.

How can we describe it in natural terms? In many practical situations, there is a reasonable linear structure on the set X on all the states, i.e., X is a linear space. In this case, going from x_k to $f(x_k)$ means adding, to the original state x_k, a displacement $f(x_k) - x_k$. Going halfway would then mean that we are only adding a half of this displacement, i.e., that we go from x_k to $x_{k+1} = x_k + \frac{1}{2} \cdot (f(x_k) - x_k)$, i.e., to

$$x_{k+1} = \frac{1}{2} \cdot x_k + \frac{1}{2} \cdot f(x_k).$$

The corresponding iteration process is called *Krasnoselskii iterations*. In general, we can use a different portions $\alpha \neq 1/2$, and we can also use different portions α_k on different moments of time. In general, we thus go from x_k to $x_{k+1} = x_k + \alpha_k \cdot (f(x_k) - x_k)$, i.e., to

$$x_{k+1} = (1 - \alpha_k) \cdot x_k + \alpha_k \cdot f(x_k).$$

These iterations are called *Krasnoselski-Mann iterations*.

Practical problem: the rate of convergence drastically depends on α_i. The above convergence results show that under certain conditions on the parameters α_i, there is a convergence. From the viewpoint of guaranteeing this convergence, we can select any sequence α_i which satisfies these conditions. However, in practice, different choice of α_i often result in drastically different rate of convergence.

To illustrate this difference, let us consider the simplest situation when already Picard iterations $x_{n+1} = f(x_n)$ converge, and converge monotonically. Then, in principle, we can have the same convergence if instead we use Krasnoselski-Mann iterations with $\alpha_n = 0.01$. Crudely speaking, this means that we replace each original step $x_n \to x_{n+1} = f(x_n)$, which bring x_n directly into x_{n+1}, by a hundred new smaller steps. Thus, while we still have convergence, we will need 100 times more iterations than before – and thus, we require a hundred times more computation time.

Since different values α_i lead to different rates of convergence, ranging from reasonably efficient to very inefficient, it is important to make sure that we select *optimal* values of the parameters α_i, values which guarantee the fastest convergence.

First idea: from the discrete iterations to the continuous dynamical system. In this section, we will describe the values α_i which are optimal in some reasonable sense. To describe this sense, let us go back to our description of the dynamical situation. In the above text, we considered observations made at discrete moments of time; this is why we talked about current moment of time, next moment of time, etc. In precise terms, we considered moments t_0, $t_1 = t_0 + \Delta t$, $t_2 = t_0 + 2\Delta t$, etc.

In principle, the selection of Δt is rather arbitrary. For example, in terms of prices, we can consider weekly prices (for which Δt is one week), monthly prices, yearly prices, etc. Similarly, for transportation, we can consider daily, hourly, etc. descriptions. The above discrete-time description is, in effect, a discrete approximation to an actual continuous-time system.

Similarly, Krasnoselski-Mann iterations $x_{k+1} - x_k = \alpha_k \cdot (f(x_k) - x_k)$ can be viewed as a discrete-time approximations to a continuous dynamical system which leads to the desired equilibrium. Specifically, the difference $x_{k+1} - x_k$ is a natural discrete analogue of the derivative $\dfrac{dx}{dt}$, the values α_k can be viewed as discretized values of an unknown function $\alpha(t)$, and so the corresponding continuous system takes the form

$$\frac{dx}{dt} = \alpha(t) \cdot (f(x) - x). \tag{18}$$

A discrete-time system is usually a good approximation to the corresponding continuous-time system. Thus, we can assume that, vice versa, the above continuous system is a good approximation for Krasnoselski-Mann iterations.

In view of this fact, in the following text, we will look for an appropriate (optimal) continuous-time system (18).

Scale invariance: natural requirement on a continuous-time system. In deriving the continuous system (18) from the formula for Krasnoselski-Mann iterations, we assumed that the original time interval Δt between the two consecutive iterations is 1. This means, in effect, that to measure time, we use a scale in which this interval Δt is a unit interval.

As we have mentioned earlier, the choice of the time interval Δt is rather arbitrary. If we make a different choice of this discretization time interval $\Delta t' \neq \Delta t$, then we would get a similar dynamical system, but described in a different time scale, with a different time interval $\Delta t'$ taken as a measuring unit. As a result of "de-discretizing" this new system, we would get a different continuous system of type (18) – a system which differs from the original one by a change in scale.

In the original scale, we identified the time interval Δt with 1. Thus, the time t in the original scale means physical time $T = t \cdot \Delta t$. In the new scale, this same physical time corresponds to the time $t' = \dfrac{T}{\Delta t'} = t \cdot \dfrac{\Delta t}{\Delta t'}$.

If we denote by $\lambda = \dfrac{\Delta t'}{\Delta t}$ the ratio of the corresponding units, then we conclude that the time t in the original scale corresponds to the time $t' = t/\lambda$ in the new scale. Let us describe the system (18) in terms of this new time coordinate t'. From the above formula, we conclude that $t = \lambda \cdot t'$; substituting $t = \lambda \cdot t'$ and $dt = \lambda \cdot dt'$ into the formula (18), we conclude that

$$\frac{1}{\lambda} \cdot \frac{dx}{dt'} = \alpha(\lambda \cdot t') \cdot (f(x) - x),$$

i.e., that

$$\frac{dx}{dt'} = (\lambda \cdot \alpha(\lambda \cdot t')) \cdot (f(x) - x). \tag{19}$$

It is reasonable to require that the optimal system of type (18) should not depend on what exactly time interval Δt we used for discretization.

Conclusion: optimal Krasnoselski-Mann iterations correspond to $\alpha_k = c/k$. Since a change of the time interval corresponds to re-scaling, this means the system (18) must be scale-invariant, i.e., to be more precise, the system (19) must have exactly the same form as the system (18) but with t' instead of t, i.e., the form

$$\frac{dx}{dt'} = \alpha(t') \cdot (f(x) - x). \tag{20}$$

By comparing the systems (19) and (20), we conclude that we must have

$$\lambda \cdot \alpha(\lambda \cdot t') = a(t')$$

for all t' and λ. In particular, if we take $\lambda = 1/t'$, then we get $\alpha(t') = \dfrac{\alpha(1)}{t'}$, i.e., $\alpha(t') = c/t'$ for some constant c $(= \alpha(1))$.

With respect to the corresponding discretized system, this means that we take $\alpha_k = \alpha(k) = c/k$.

Comment. The formula $\alpha_k = c/k$ is not exact: it comes form approximating the actual continuous dependence by a discrete one. This approximation makes asymptotic sense, but this formula cannot be applied for $k = 0$. To make this formula applicable, we must start with $k = 1$ – or, equivalently, start with $k = 0$ (since this is how most descriptions of iterations work), but use the expression $\alpha_k = c/(k + 1)$ instead.

Reasonable choice of the constant c and its interpretation. As we have mentioned, a reasonable idea is to use Picard iterations. This is not always a good idea, because we may get wild fluctuations. However, it makes some sense to start with the Picard iteration first, to get away from the initial state.

Picard iterations correspond to $\alpha_k = 1$; so, if we want $\alpha_0 = 1$, i.e., $c/(0 + 1) = 1$, we must take $c = 1$. The resulting iterations take the form

$$x_{k+1} = \left(1 - \frac{1}{k+1}\right) \cdot x_k + \frac{1}{k+1} \cdot f(x_k).$$

This formula (corresponding to $c = 1$) has a natural commonsense interpretation.

Namely, in Picard iterations, as a next iteration x_{k+1}, we take $f(x_k)$. When there are wild oscillations, these iterations do not converge. We expect, however, that these oscillations are going on around the equilibrium point. So, while the values x_i are oscillating and not converging at all, their averages

$$\frac{x_0 + \ldots + x_k}{k+1}$$

and the corresponding values

$$\frac{f(x_0) + \ldots + f(x_k)}{k+1}$$

will be getting closer and closer to the desired equilibrium. Thus, if we want to enhance convergence, then, instead of taking $f(x_k)$ as the next iteration, it makes sense to take an *average* of the previous values of $f(x_k)$:

$$x_{k+1} = \frac{f(x_0) + \ldots + f(x_{k-1}) + f(x_k)}{k+1}.$$

Let us show that this idea leads exactly to our choice $\alpha_k = 1/(k+1)$. Indeed, from $x_k = \frac{f(x_0) + \ldots + f(x_{k-1})}{k}$, we conclude that $f(x_0) + \ldots + f(x_{k-1}) = k \cdot x_k$, hence $f(x_0) + \ldots + f(x_{k-1}) + f(x_k) = k \cdot x_k + f(x_k)$ and thus,

$$x_{k+1} = \frac{f(x_0) + \ldots + f(x_{k-1}) + f(x_k)}{k+1} = \frac{k \cdot x_k + f(x_k)}{k+1} = \left(1 - \frac{1}{k+1}\right) \cdot x_k + \frac{1}{k+1} \cdot f(x_k).$$

This selection seems to work well. The choice $a_k = 1/k$ have been successfully used and shown to be efficient. We have shown this on the example of our transportation problem. For other examples, see, e.g., (Su and Qin, 2006) and references therein.

C. Exponential Disutility Functions in Transportation Modeling: Justification

Stochastic approach, and the need to use utility or disutility functions. In real life, travel times are non-deterministic (*stochastic*): on each link, for the same capacity and flow, we may have somewhat different travel times (Sheffi, 1985).

In other words, for each link, the travel time t_i is no longer a uniquely determined real number, it is a *random variable* whose characteristics may depend on the capacity and flow along this link. As a result, the overall travel time t is also a random variable.

If we take this uncertainty into account, then it is no longer easy to predict which path will be selected: if we have two alternative paths, then it often happens that with some probability, the time along the first path is smaller, but with some other probability, the first path may turn out to be longer. How can we describe decision making under such uncertainty?

In decision making theory, it is proven that under certain reasonable assumption, a person's preferences are defined by his or her *utility function* $U(x)$ which assigns to each possible outcome x a real number $U(x)$ called *utility*; see, e.g., (Keeney and Raiffa, 1976; Raiffa, 1970). In many real-life situations, a person's choice s does not determine the outcome uniquely, we may have different outcomes $x_1, \ldots, x_n$ with probabilities, correspondingly, $p_1, \ldots, p_n$. For example, when a driver selects a path s, the travel time is often not uniquely determined: we may have different travel times $x_1, \ldots, x_n$ with corresponding probabilities $p_1, \ldots, p_n$. For such a choice, we can describe the utility $U(s)$ associated with this choice as the expected value of the utility of outcomes: $U(s) = E[U(x)] = p_1 \cdot U(x_1) + \ldots + p_n \cdot U(x_n)$. Among several possible choices, a user selects the one

for which the utility is the largest: a possible choice s is preferred to a possible choice s' (denoted $s > s'$) if and only if $U(s) > U(s')$.

For the applications presented in this paper, it is important to emphasize that the utility function is not uniquely determined by the preference relation. Namely, for every two real numbers $a > 0$ and b, if we replace the original utility function $U(x)$ with the new one $V(x) \stackrel{\text{def}}{=} a \cdot U(x) + b$, then for each choice s, we will have

$$V(s) = E[a \cdot U(x) + b] = a \cdot E[U(x)] + b = a \cdot U(s) + b$$

and thus, $V(s) > V(s')$ if and only if $U(s) > U(s')$.

In transportation, the main concern is travel time t, so the utility depends on time: $U = U(t)$. Of course, all else being equal, the longer it takes to travel, the less preferable the choice of a path; so, the utility function $U(t)$ must be strictly increasing: if $t < t'$, then $U(t) > U(t')$.

In general, decision making is formulated in terms of *maximizing* a utility function $U(x)$. In traditional (deterministic) transportation problems, however, decision making is formulated in terms of *minimization*: we select a route with the smallest possible travel time. Thus, when people apply decision making theory in transportation problems, they reformulate these problems in terms of a *disutility* function $u(x) \stackrel{\text{def}}{=} -U(x)$. Clearly, for every choice s, we have

$$u(s) \stackrel{\text{def}}{=} E[u(x)] = E[-U(x)] = -E[U(x)] = -U(s).$$

So, selecting the route with the *largest* value of expected utility $U(s)$ is equivalent to selecting the route with the *smallest* value of expected disutility $u(s)$. In line with this usage, in this paper, we will talk about disutility functions.

Since a disutility function $U(t)$ is strictly decreasing, the corresponding utility function $u(t) = -U(t)$ must be strictly increasing: if $t < t'$ then $u(t) < u(t')$.

Disutility functions traditionallly used in transportation: description and reasons. In transportation, traditionally, three types of disutility functions are used; see, e.g., (Mirchandani and Soroush, 1987; Tatineni, 1996; Tatineni et al., 1997).

First, we can use *linear* disutility functions $u(t) = a \cdot t + b$, with $a > 0$. As we have mentioned, multiplication by a constant $a > 0$ and addition of a constant b does not change the preferences, so we can safely assume that the utility function simply conincides with the travel time $u(t) = t$.

Second, we can use *risk-prone exponential* disutility functions

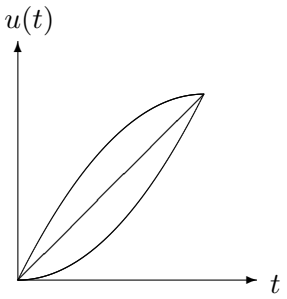
$$u(t) = -a \cdot \exp(-c \cdot t) + b$$

for some $a > 0$ and $c > 0$. This is equivalent to using $u(t) = -\exp(-c \cdot t)$.

Third, we can use *risk-averse exponential* disutility functions

$$u(t) = a \cdot \exp(c \cdot t) + b$$

for some $a > 0$ and $c > 0$. This is equivalent to using $u(t) = \exp(c \cdot t)$.



Several other possible disutility functions have been proposed, e.g., quadratic functions $u(t) = t + c \cdot t^2$; see, e.g., (Mirchandani and Soroush, 1987).

In practice, mostly linear and exponential functions are used. Actually, a linear function can be viewed as a limit of exponential functions:

$$t = \lim_{\alpha \to 0} \frac{1}{\alpha} \cdot (\exp(\alpha \cdot t) - 1),$$

so we can say that mostly exponential functions are used.

The main reason for using exponential disutility functions is that these functions are in accordance with common sense (Mirchandani and Soroush, 1987; Tatineni et al., 1997). Indeed:

- functions $-\exp(-c \cdot t)$ indeed lead to risk-prone behavior, i.e., crudely speaking, a behavior in which a person, when choosing between two paths, one with a deterministic time t_1 and another with a stochastic time t_2, prefers the second path if there is a large enough probability that $t_2 < t_1$ – even when the average time of the second path may be larger $\bar{t}_2 > t_1$;

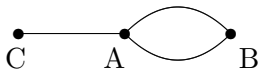
- functions $\exp(c \cdot t)$ indeed lead to risk-averse behavior, i.e., crudely speaking, a behavior in which a person, when choosing between two paths, one with a deterministic time t_1 and another with a stochastic time t_2, prefers the first path if there is a reasonable probability that $t_2 > t_1$ – even when the average time of the second path may be smaller: $\bar{t}_2 < t_1$.

This accordance, however, does not limit us to only exponential functions: e.g., quadratic functions are also in reasonably good accordance with common sense.

However, there is another common sense requirements that leads to linear or exponential functions.

A common sense assumption about the driver's preferences. Let us assume that we have several routes going from point A to point B, and a driver selected one of these routes as the best for him/her. For example, A may be a place at the entrance to the driver's department, and B is a similar department at another university located in a nearby town.

Let us now imagine a similar situation, in which the driver is also interested in reaching the point B, but this time, the driver starts at some prior point C. At this point C, there is only one possible way, and it leads to the point A; after A, we still have several possible routes. We can also assume that the time t_0 that it takes to get from C to A is deterministic. For example, C may be a place in the parking garage from where there is only one exit.

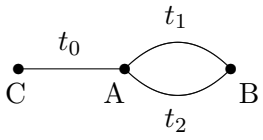


It is reasonable to assume that if the road conditions did not change, then, after getting to the point A, the driver will select the exact same route as last time, when this driver started at A.

Comment. Similarly, if two routes from A to B were equally preferable to the driver, then both routes should be equally preferable after we add a deterministic link from C to A to both routes.

In the deterministic case, this assumption is automatically satisfied. In the deterministic case, the travel time along each route is deterministic, and the driver selects a route with the shortest travel time.

Let us assume when going from A to B, the drive prefers the first route because its travel time t_1 is smaller than the travel time t_2 of the second route: $t_1 < t_2$. In this case, next time, when the travel starts from the point C, we have time $t_1 + t_0$ along the first route and $t_2 + t_0$ along the second route. Since we had $t_1 < t_2$, we thus have $t_1 + t_0 < t_2 + t_0$ – and therefore, the driver will still select the first route.



In the stochastic case, this assumption is not necessarily automatically satisfied. In the stochastic case, when going from A to B, the driver selects the first route if $E[u(t_1)] < E[u(t_2)]$, where $u(t)$ is the corresponding disutility function.

Next time, when the driver goes from C to B, the choice between the two routes depends on comparing different expected values: $E[u(t_1 + t_0)]$ and $E[u(t_2 + t_0)]$, where t_0 is the (deterministic) time of traveling from C to A. In principle, it may be possible that $E[u(t_1)] < E[u(t_2)]$ but

$$E[u(t_1 + t_0)] > E[u(t_2 + t_0)].$$

Let us describe a simple numerical example when this counter-intuitive phenomenon happens. In this example, we will use a simple non-linear disutility function: namely, the quadratic function $u(t) = t^2$. Let us assume that the first route from A to B is deterministic, with $t_1 = 7$, and the second route from A to B is highly stochastic: with equal probability 0.5, we may have $t_2 = 1$ and $t_2 = 10$. In this case, $E[u(t_1)] = t_1^2 = 49$ and

$$E[u(t_2)] = E[t_2^2] = \frac{1}{2} \cdot 1^2 + \frac{1}{2} \cdot 10^2 = 0.5 + 50 = 50.5.$$

Here, $E[u(t_1)] < E[u(t_2)]$, so the driver will prefer the first route.

However, if we add the same constant time $t_0 = 1$ for going from C to A to both routes, then in the first route, we will have $t_1 + t_0 = 7 + 1 = 8$, while in the second route, we will have $t_2 + t_0 = 1 + 1 = 2$ and $t_2 + t_0 = 10 + 1 = 11$ with equal probability 0.5. In this case,

$$E[u(t_1 + t_0)] = (t_1 + t_0)^2 = 8^2 = 64,$$

while

$$E[u(t_2+t_0)] = \frac{1}{2}\cdot 2^2 + \frac{1}{2}\cdot 11^2 = 2 + 60.5 = 62.5.$$

We see that here, $E[u(t_2+t_0)] < E[u(t_1+t_0)]$, i.e., the drive will select the second route instead of the first one.

This counter-intuitive phenomenon does not happen for linear or exponential disutility functions. Indeed, for a linear disutility function $u(t) = t$, we have $u(t_1+t_0) = t_1+t_0 = u(t_1)+t_0$; therefore, $E[u(t_1+t_0)] = E[u(t_1)]+t_0$ and similarly, $E[u(t_2+t_0)] = E[u(t_2)]+t_0$. Thus, if the driver selected the first route, i.e., if $E[u(t_1)] < E[u(t_2)]$, then by adding t_0 to both sides of this inequality, we can conclude that $E[u(t_1+t_0)] < E[u(t_2+t_0)]$ – i.e., that, in accordance with common sense, the same route will be selected if we start at the point C.

For the exponential disutility function $u(t) = \exp(\alpha\cdot t)$, we have $u(t_1+t_0) = \exp(\alpha\cdot(t_1+t_0)) = \exp(\alpha\cdot t_1)\cdot\exp(\alpha\cdot t_0)$ and therefore, $u(t_1+t_0) = u(t_1)\cdot\exp(\alpha\cdot t_0)$. Similarly, for the exponential disutility function $u(t) = -\exp(\alpha\cdot t)$, we have $u(t_1+t_0) = -\exp(\alpha\cdot(t_1+t_0)) = -\exp(\alpha\cdot t_1)\cdot\exp(\alpha\cdot t_0)$ and thus, $u(t_1+t_0) = u(t_1)\cdot\exp(\alpha\cdot t_0)$;

For both types of exponential disutility function, we have $E[u(t_1+t_0)] = \exp(\alpha\cdot t_0)\cdot E[u(t_1)]$ and similarly, $E[u(t_2+t_0)] = \exp(\alpha\cdot t_0)\cdot E[u(t_2)]$. Thus, if the driver selected the first route, i.e., if $E[u(t_1)] < E[u(t_2)]$, then by multiplying both sides of this inequality by the same constant $\exp(\alpha\cdot t_0)$, we can conclude that $E[u(t_1+t_0)] < E[u(t_2+t_0)]$ – i.e., that, in accordance with common sense, the same route will be selected if we start at the point C.

Resulting justification of exponential utility fuinctions. It turns out linear and exponential disutility functions are the only ones which are consistent with the above common sense requirement – for every other disutility function, a paradoxical counter-intuitive situation like the one described above is quite possible.

Let us describe this result in precise terms.

Definition 3. *By a* disutility function*, we mean a strictly increasing function $u(t)$ from non-negative real numbers to real numbers.*

Definition 4. *We say that two disutility functions $u(t)$ and $v(t)$ are* equivalent *if there exist real numbers $a > 0$ and b such that $v(t) = a\cdot u(t) + b$ for all t.*

Definition 5. *We say that a disutility function is* consistent with common sense *if it has the following property: let t_1 and t_2 be random variables with non-negative values, and let t_0 be an arbitrary (deterministic) non-negative real number; then,*

- *if $E[u(t_1)] < E[u(t_2)]$, then $E[u(t_1+t_0)] < E[u(t_2+t_0)]$;*
- *if $E[u(t_1)] = E[u(t_2)]$, then $E[u(t_1+t_0)] = E[u(t_2+t_0)]$.*

Theorem 2. *A disutility function is consistent with common sense if and only if it is equivalent to either the linear function $u(t) = t$, or to an exponential function $u(t) = \exp(c\cdot t)$ or $-\exp(-c\cdot t)$.*

Proof. Under an additional conditions of differentiability of the function $u(t)$, this result has been proven in (Pratt, 1964). For reader's convenience, we provide a new proof which does not require differentiability.

1°. We already know that linear and exponential disutility functions are consistent with common sense in the sense of Definition 5. It is therefore sufficient to prove that every disutility function $u(t)$ which is consistent with common sense is equivalent either to a linear one or to an exponential one.

2°. Let $u(t)$ be a disutility function which is consistent with common sense. By definition of computational simplicity, for every random variables t_1, once we know the values $u_1 = E[u(t_1)]$ and t_0, we can uniquely determine the value $E[u(t_1 + t_0)]$. Let us denote the value $E[u(t_1 + t_0)]$ corresponding to u_1 and t_0 by $F(u_1, t_0)$.

3°. Let t_1' be a non-negative number. For the case when $t_1 = t_1'$ with probability 1, we have $u_1' = E[u(t_1)] = u(t_1')$. In this case, $t_1 + t_0 = t_1' + t_0$ with probability 1, so $E[u(t_1 + t_0)] = u(t_1' + t_0)$. Thus, in this case, $u(t_1' + t_0) = F(u_1', t_0)$, where $u_1' = u(t_1')$.

4°. Let us now consider the case when t_1 is equal to t_1' with some probability $p_1' \in [0, 1]$, and to some smaller value $t_1'' < t_1'$ with the remaining probability $p_1'' = 1 - p_1'$. In this case,

$$u_1 = E[u(t_1)] = p_1' \cdot u(t_1') + (1 - p_1') \cdot u(t_1'').$$

We have already denoted $u(t_1')$ by u_1'; so, if we denote $u_1'' \stackrel{\text{def}}{=} u(t_1'')$, we can rewrite the above expression as

$$u_1 = p_1' \cdot u_1' + (1 - p_1') \cdot u_1''.$$

In this situation, $t_1 + t_0$ is equal to $t_1' + t_0$ with probability p_1' and to $t_1'' + t_0$ with probability $1 - p_1'$. Thus,

$$E[u(t_1 + t_0)] = p_1' \cdot u(t_1' + t_0) + (1 - p_1') \cdot u(t_1'' + t_0).$$

We already know that $u(t_1' + t_0) = F(u_1', t_0)$ and $u(t_1'' + t_0) = F(u_1'', t_0)$. So, we can conclude that

$$E[u(t_1 + t_0)] = p_1' \cdot F(u_1', t_0) + (1 - p_1') \cdot F(u_1'', t_0). \tag{21}$$

On the other hand, by the definition of the function F as $F(u_1, t_0) = E[u(t_1 + t_0)]$, we conclude that

$$E[u(t_1 + t_0)] = F(u_1, t_0),$$

i.e.,

$$E[u(t_1 + t_0)] = F(p_1' \cdot u_1' + (1 - p_1') \cdot u_1'', t_0). \tag{22}$$

Comparing the expressions (21) and (22) for $E[u(t_1 + t_0)]$, we conclude that

$$F(p_1' \cdot u_1' + (1 - p_1') \cdot u_1'', t_0) = p_1' \cdot F(u_1', t_0) + (1 - p_1') \cdot F(u_1'', t_0).$$

Let us analyze this formula. For every value $u_1 \in [u_1'', u_1']$, we can find the probability p_1' for which $u_1 = p_1' \cdot u_1' + (1 - p_1') \cdot u_1''$: namely, the desired equation means that $u_1 = p_1' \cdot u_1' + u_1'' - p_1' \cdot u_1''$; rearranging the terms, we get $u_1 - u_1'' = p_1' \cdot (u_1' - u_1'')$ and hence, the value $p_1' = \dfrac{u_1 - u_1''}{u_1' - u_1''}$. Substituting this expression into the above formula, we conclude that for a fixed t_0, the function $F(u_1, t_0)$ is a linear function of u_1:

$$F(u_1, t_0) = A(t_0) \cdot u_1 + B(t_0)$$

for some constants $A(t_0)$ and $B(t_0)$ which, in general, depend on t_0.

5°. We have already shown, in Part 3 of this proof, that $u(t'_1 + t_0) = F(u'_1, t_0)$. Thus, we conclude that for every $t'_1 \geq 0$ and $t_0 \geq 0$, we have

$$u(t'_1 + t_0) = A(t_0) \cdot u(t'_1) + B(t_0).$$

6°. For an arbitrary function $u(t)$, by introducing an appropriate constant $b = -u(0)$, we can always find an equivalent function $v(t)$ for which $v(0) = 0$. So, without losing generality, we can assume that $u(0) = 0$ for our original disutility function $u(t)$.

Since the disutility function is strictly increasing, we have $u(t) > 0$ for all $t > 0$.

For $t'_1 = 0$, the above formula takes the form $u(t_0) = B(t_0)$. Substituting this expression for $B(t_0)$ into the above formula, we conclude that

$$u(t'_1 + t_0) = A(t_0) \cdot u(t'_1) + u(t_0).$$

7°. The above property has to be true to arbitrary values of $t'_1 \geq 0$ and $t_0 \geq 0$. Swapping these values, we conclude that

$$u(t_0 + t'_1) = A(t'_1) \cdot u(t_0) + u(t'_1).$$

Since $t'_1 + t_0 = t_0 + t'_1$, we have $u(t'_1 + t_0) = u(t_0 + t'_1)$, hence

$$A(t_0) \cdot u(t'_1) + u(t_0) = A(t'_1) \cdot u(t_0) + u(t'_1).$$

Moving terms proportional to $u(t'_1)$ to the left hand side and terms proportional to $u(t_0)$ to the right hand side, we conclude that

$$(A(t_0) - 1) \cdot u(t'_1) = (A(t'_1) - 1) \cdot u(t_0). \tag{23}$$

In the following text, we will consider two possible situations:

- the first situation is when $A(t_0) = 1$ for some $t_0 > 0$;
- the second situation is when $A(t_0) \neq 1$ for all $t_0 > 0$.

In the first situation, $A(t_0) = 1$ for some $t_0 > 0$. For this t_0, the equation (23) takes the form $(A(t'_1) - 1) \cdot u(t_0) = 0$ for all t'_1. Since $u(t_0) > 0$ for $t_0 > 0$, we conclude that $A(t'_1) - 1 = 0$ for every real number $t'_1 \geq 0$, i.e., that the function $A(t)$ is identical to a constant function 1.

So, we have two possible situations:

- the first situation is when $A(t_0) = 1$ for some $t_0 > 0$; we have just shown that in this case, $A(t) = 1$ for all t; in the following text, we will show that in this situation, the disutility function $u(t)$ is linear;
- the second situation is when $A(t_0) \neq 1$ for all $t_0 > 0$; we will show that in this situation, the disutility function $u(t)$ is exponential.

8°. Let us first consider the situation in which $A(t)$ is always equal to 1. In this case, the above equation takes the form

$$u(t_0 + t_1') = u(t_0) + u(t_1').$$

In other words, in this case,

$$u(t_1 + t_2) = u(t_1) + u(t_2)$$

for all possible values $t_1 > 0$ and $t_2 > 0$.

In particular, for every $t_0 > 0$, we get:

- first, $u(2t_0) = u(t_0) + u(t_0) = 2u(t_0)$,
- then $u(3t_0) = u(2t_0) + u(t_0) = 2u(t_0) + u(t_0) = 3u(t_0)$, and,
- in general, $u(k \cdot t_0) = k \cdot u(t_0)$ for all integers k.

For every integer n and for $t_0 = 1/n$, we have $u(n \cdot t_0) = u(1) = n \cdot u(1/n)$, hence $u(1/n) = u(1)/n$. Then, for an arbitrary non-negative rational number k/n, we get

$$u(k/n) = u(k \cdot (1/n)) = k \cdot u(1/n) = k \cdot (1/n) \cdot u(1) = k/n \cdot u(1).$$

In other words, for every rational number $r = k/n$, we have $u(r) = r \cdot u(1)$.

Every real value t can be bounded, with arbitrary accuracy, by rational numbers k_n/n and $(k_n + 1)/n$: $k_n/n \le t \le (k_n + 1)/n$, where $k_n/n \to t$ and $(k_n + 1)/n \to t$ as $n \to \infty$. Since the disutility function $u(t)$ is strictly increasing, we conclude that $u(k_n/n) \le u(t) \le u((k_n + 1)/n)$. We already know that for rational values r, we have $u(r) = r \cdot u(1)$, so we have

$$k_n/n \cdot u(1) \le u(t) \le (k_n + 1)/n \cdot u(1).$$

In the limit $n \to \infty$, both sides of this inequality converge to $t \cdot u(1)$, hence $u(t) = t \cdot u(1)$.

So, in this case, we get a linear disutility function.

9°. Let us now analyze the case when $A(t) \ne 1$ for all $t > 0$. Since the values $u(t)$ are positive for all $t > 0$, we can divide both sides of the equality

$$(A(t_0) - 1) \cdot u(t_1') = (A(t_1') - 1) \cdot u(t_0)$$

by $u(t_0)$ and $u(t_1')$, and conclude that

$$\frac{A(t_0) - 1}{u(t_0)} = \frac{A(t_1') - 1}{u(t_1')}.$$

The ratio $\dfrac{A(t) - 1}{u(t)}$ has the same value for arbitrary two numbers $t = t_0$ and $t = t_1'$; thus, this ratio is a constant. Let us denote this constant by k; then, $A(t) - 1 = k \cdot u(t)$ for all $t > 0$. Since $A(t) \ne 1$, this constant k is different from 0.

Substituting the resulting expression $A(t) = 1 + k \cdot u(t)$ into the formula $u(t_1' + t_0) = A(t_0) \cdot u(t_1') + u(t_0)$, we conclude that

$$u(t_1' + t_0) = u(t_0) + u(t_1') + k \cdot u(t_0) \cdot u(t_1'),$$

i.e., that

$$u(t_1 + t_2) = u(t_1) + u(t_2) + k \cdot u(t_1) \cdot u(t_2)$$

for arbitrary numbers $t_1 > 0$ and $t_2 > 0$.

10°. Let us now consider a re-scaled function $v(t) \stackrel{\text{def}}{=} 1 + k \cdot u(t)$.

For this function $v(t)$, from the above formula, we conclude that

$$v(t_1 + t_2) = 1 + k \cdot u(t_1 + t_2) = 1 + k \cdot (u(t_1) + u(t_2)) + k^2 \cdot u(t_1) \cdot u(t_2).$$

On the other hand, we have

$$v(t_1) \cdot v(t_2) = (1 + k \cdot u(t_1)) \cdot (1 + k \cdot u(t_2)) =$$

$$1 + k \cdot (u(t_1) + u(t_2)) + k^2 \cdot u(t_1) \cdot u(t_2).$$

The expression for $v(t_1 + t_2)$ and for $v(t_1) \cdot v(t_2)$ coincide, so we conclude that

$$v(t_1 + t_2) = v(t_1) \cdot v(t_2)$$

for all possible values $t_1 > 0$ and $t_2 > 0$.

11°. When $k > 0$, then the new function $v(t)$ is an equivalent disutility function. We know that $u(0) = 0$ hence $v(0) = 1 + k \cdot 0 = 1$. Since $v(t)$ is a strictly increasing function, we thus conclude that $v(t) \geq v(0) > 0$ for all $t \geq 0$.

Thus, we can take a logarithm of all the values, and for the new function $w(t) \stackrel{\text{def}}{=} \ln(v(t))$, get an equation

$$w(t_1 + t_2) = \ln(v(t_1 + t_2)) = \ln(v(t_1) \cdot v(t_2)) = \ln(v(t_1)) + \ln(v(t_2)) = w(t_1) + w(t_2),$$

i.e., $w(t_1 + t_2) = w(t_1) + w(t_2)$ for all t_1 and t_2. The function $w(t)$ is increasing – as the logarithm of an increasing function. Thus, as we have already shown, $w(t) = c \cdot t$ for some $c > 0$.

From the logarithm $w(t) = \ln(v(t))$, we can reconstruct the original disutility function $v(t)$ as $v(t) = \exp(w(t))$. Since $w(t) = c \cdot t$, we conclude that the disuility function $v(t)$ has the desired risk-averse exponential form

$$v(t) = \exp(c \cdot t).$$

12°. When $k < 0$, the new function is strictly decreasing (and is thus not a disutility function; its opposite $-v(t)$ is a disutility function).

For the function $v(t)$, we cannot have $v(t_0) = 0$ for any t_0 – because otherwise we would have

$$v(t) = v(t_0 + (t - t_0)) = v(t_0) \cdot v(t - t_0) = 0$$

for all $t \geq t_0$ which contradicts to our conclusion that the function $v(t)$ should be strictly decreasing.

13°. For the function $v(t)$, we cannot have $v(t_0) < 0$ for any $t_0 > 0$ – because otherwise, we would have $v(2t_0) = v(t_0)^2 > 0$ hence $v(2t_0) > v(t_0)$ – which, since $2t_0 > t_0$, also contradicts to our conclusion that the function $v(t)$ should be strictly decreasing.

We thus conclude that $v(t) > 0$ for all t.

14°. Thus, we can take a logarithm of all the values, and for the new function $w(t) \stackrel{\text{def}}{=} \ln(v(t))$, get the equation $w(t_1 + t_2) = w(t_1) + w(t_2)$ for all t_1 and t_2. The function $w(t)$ is decreasing – as the logarithm of a decreasing function. Thus, $w(t) = -c \cdot t$ for some $c > 0$.

From the logarithm $w(t) = \ln(v(t))$, we can reconstruct the original function $v(t)$ as $v(t) = \exp(w(t)) = \exp(-c \cdot t)$, and the disutility function $u(t)$ as $-v(t) = -\exp(-c \cdot t)$.

So, we conclude that the disuility function $v(t)$ has the desired risk-prone exponential form $v(t) = -\exp(-c \cdot t)$.

The theorem is proven.