UNIVERSITÀ DI TRIESTE
Sede amministrativa del Dottorato di Ricerca

XIX Ciclo del Dottorato di Ricerca in
Ingegneria delle Infrastrutture, delle Strutture e dei Trasporti

ESTIMATION OF RAILWAY CAPACITY CONSUMPTION USING STOCHASTIC DIFFERENTIAL EQUATIONS

(Settore scientifico-disciplinare: ICAR/05 TRASPORTI)

Dottorando:
Roberto Stok

Coordinatore:
Chiar.mo Prof.
Roberto Camus
(Università di Trieste)

Relatore:
Chiar.mo Prof.
Giovanni Longo
(Università di Trieste)

Contents

Introduction

The efficient utilization of railway infrastructures is a primary objective in an open-market context like the European one. The capacity consumption, that is the infrastructure occupation augmented with buffers to avoid delays (referred to a time window, e.g. peak hour or day), is a measure of the utilization level of a given timetable.

The standard UIC leaflet "Capacity" recommends a procedure to evaluate the infrastructure occupation, without buffers, by compressing the timetable until the blocking time stairways touch each other in the critical section. There is no recommendation about buffer times, except a well known rule of thumb about the running time supplement, which is often set to 5% of the journey time.

Buffer times choice is a trade-off between efficient utilization and stability, to avoid secondary delays caused by primary delays. Given the probability distribution of the primary delays, it is possible to estimate the distribution of secondary delays and hence the buffers.

In this work the primary delays are modelled following an innovative approach, that is using a family of stochastic processes called Lévy processes. These stochastic processes are defined through a very simple and general assumption: stationary independent increments. A disturbance on which there is few knowledge may reasonably be assumed to satisfy independence and stationarity properties, because independence means the future doesn't depend on the past and stationarity means the process doesn't change its structure over the time (often it is true only on large time windows). Another reasonable assumption is the path continuity, so the Lévy process reduces to a Brownian motion.

The train movement is therefore modelled with a differential equation to which a Brownian motion is added, leading to a stochastic differential equation (SDE). Given the numerical approximation of the Brownian motion sample path, which is pseudo-random generated, it is possible to solve the SDE and obtain the sample path of the train.

The analysis of the stochastic phenomenon requires the replication (called

Monte Carlo) of the pseudo-random generation of the approximated Brownian motion sample path and the solution of a SDE to be computed many times. The result is a collection of simulated sample paths for each train scheduled.

This set of collections may be used in two different ways:

- estimation of the probabilistic distributions of the blocking times and consideration of the stochastic version of the blocking time stairways;
- estimation of the risk, that is the probability of hindrance which corresponds to a given timetable, with the trains running in free mode (no external control, signals ignored) but counting the risky events highlighted by the signalling system.

The estimation of the risk is repeated varying the number of trains, so that a relationship is built between the risk and the number of the trains or headways, from which a measure of capacity consumption is obtained given the risk level.

The thesis consists of two parts, the first is made of preparation chapters while the second one is devoted to models and applications.

The work begins with a chapter on capacity issues, where concepts, definitions and standards are illustrated to establish the research framework; the chapter ends with a literature review where the recent approaches are discussed and the lack of an SDE approach is highlighted. There is one stochastic approach which uses differential equations, but they are deterministic and combined with stochastic boarding time.

Then a brief theoretical chapter on the Lévy processes is presented to justify the modelling choice and to introduce the Brownian motion in a reasonably acceptable way; unavoidable definitions of stochastic entities and theorems bring the reader to the presentation of the central result of the Lévy theory, that is the decomposition theorem. The Lévy-Itô decomposition theorem states that a Lévy process always decomposes as the sum of a Brownian motion with drift, a compound Poisson process and a martingale: if the sample paths are continuous, then there is only the Brownian component because the other ones have jumps in their paths.

The Brownian motion introduction is followed by a theoretical chapter on the SDEs, where the symbolic meaning of dW_t is explained, together with two popular formulations of the stochastic integral, Itô and Stratonovich and the choice of Itô's one is justified on modelling basis. The Itô formula, that is the stochastic chain rule of calculus, and the stochastic Itô-Taylor expansion are presented as essential tools for understanding stochastic numerical methods. The chapter ends with the theorem that states the existence and

uniqueness of a strong solution, that is the solution of the SDE given the driving Brownian motion W_t.

The following chapter is about the numerical methods that are available to solve an SDE; at the beginning a brief review of the concepts and methods of deterministic ordinary differential equations is given. The most known and used schemes - i.e. Euler and Milstein - are presented together with their convergence and stability properties. Other schemes are briefly cited for sake of completeness.

The preparation part closes with a chapter devoted to Monte Carlo simulations and the type of possible applications of the resulting collections of train paths to capacity assessment, that is estimation of the probabilistic distributions of the blocking times and estimation of the risk as probability of hindrance. Replications of Monte Carlo simulations with different timetables allow the building of the capacity-risk relationship.

In the models and applications part two SDE-based models are presented, together with case studies: the first model is simple but allows some theoretical considerations to validate (in the form of bounds) the simulation results; the second model is a stochastic optimal control model. In both cases the model parameters are estimated using real life data and then the capacity-risk relationship is build through simulations. Another result of the simulation is the set of blocking/clearing time distributions for each section, which is graphically represented by plotting their key points (the mean value and extremes of the almost-sure range estimated by taking three times the standard deviation) at each section for a group of train paths.

This second model describes in a more realistic way the train journey, because the mechanical equation is more suitable and the driving machine produces a force following an optimal control rule which considers both the distance from the timetable and the energy consumption.

The optimal control law of the exact stochastic optimal control problem may be found by solving the Hamilton Jacobi Bellman equation, which is numerically heavy as well as difficult to solve because of instability and nonlinearities. An approximated stochastic optimal control problem is solved for the more relevant part of the the train travel, that is the steady state maintenance stage (initial acceleration stage and final stop stage are excluded), where the driver tries to reach the steady state determined from the planned timetable: the mechanical equation is linearized near the steady state speed and the optimal control law expression in terms of state variables is found and therefore substituted in the SDE. A parameter, the *driving style*, defined as the ratio of the schedule cost and the energy cost, is introduced to describe the different weights the two objectives may be given. Sensitivity analysis has been performed to determine the parameters' ranges for model

applicability.

The final chapter summarizes findings and conclusions of this research work.

Chapter 1

Railway capacity issues

Railway capacity is a topic of interest in railway infrastructure planning and management as well, especially in Europe where the transport policy aims to revitalize the rail transport sector through opening-up the markets (ref: White Paper of the European Commission on the Transport Policy [EC, 2001]): the European infrastructure managers have to carefully evaluate the existing capacity and efficiently determine the infrastructure access to operators. The rail sector needs the standardization of incompatible national systems to become more dynamic and competitive, so recently (2004) the European Railway Agency was established to create an integrated railway area by reinforcing safety and interoperability and it also acts as the system authority for the European Rail Traffic Management System (ERTMS) project, which has been set up to create unique signalling standards throughout Europe.

The natural and intuitive definition of *capacity* of a railway line is the maximum number of trains that can be operated over it in a given period of time. The concept seems simple and clear but it becomes elusive when considering a lot of factors that influence the measure, like physical infrastructure limits, signalling system design, trains belonging to different speed classes and service reliability.

The upper limit of this measure is called *theoretical capacity* and it is the maximum number of trains (per hour) operated over the line in ideal conditions, that is identical trains, evenly spaced, permanently running at the minimum possible temporal distance between them. This number is easily evaluated but it is only a bound: the *practical capacity*, that is the capacity evaluated under more realistic assumptions, is usually around 60-75% of the theoretical one. The practical capacity is so hard to define that the UIC (International Union of Railways or *Union Internationale des Chemins de fer*) states that "Capacity as such does not exist; railway infrastructure

capacity depends on the way it is utilised" in its last standard document about capacity, UIC Leaflet 406 R [UIC, 2004b].

The UIC is the world-wide organisation for international cooperation among railways and its purpose is the standardisation and improvement of conditions for railway construction and operations. Its best-practice documents are recognized as world-wide standards and has been adopted by many countries.

A good coverage of the railway capacity issues may be given by illustrating the following topics:

- definition of the concepts related to capacity and timetable;
- the analytical method UIC Leaflet 405 (dropped standard);
- the optimization method UIC Leaflet 406 (current standard);
- recent literature review.

Simulation methods, that is computer methods based on a model of the infrastructure very close to the reality, are not taken in consideration because the analysis of the commercial simulation environments is outside the scope of this work, although they are the only way to validate a timetable.

1.1 Concepts related to capacity and timetable

First of all, the different flavours of railway capacity must be defined:

- *Transport Capacity* - general definition used in transportation science: it is the maximum transport volume per route, that is the maximum number of passengers or tons of freight that could be moved over the line in a given time period. In the railway environment it is less used because the focus is on the number of trains. Obviously the transport capacity may be easily calculated as the product of the number of trains and the maximum number of passengers or tons of freight per train.
- *Theoretical capacity*: it is the maximum number of trains that could be operated over the route in ideal conditions, during a given time interval. The ideal (mathematically generated) environment has identical trains, evenly spaced, permanently running at the minimum possible temporal distance between them. It is only an upper bound for capacity because it is not possible to run the trains in reality.

- *Practical capacity*: it is the practical limit of train traffic volume that can be moved on a line at a reasonable level of reliability. It is "the" measure of capacity, because it is calculated under realistic assumptions and it depends on the way the infrastructure is utilised (UIC 406R). It is usually around 60-75% of the theoretical capacity.

- *Used capacity*: it is the actual traffic volume moved on the line. It reflects the actual infrastructure occupation determined by the actual timetable. It is usually lower than the practical capacity.

- *Available capacity*: it is the difference between the Practical capacity and the Used capacity and it represents the additional traffic volume that could be handled by the infrastructure.

The attributes of the railway as a transportation system must be defined:

- *Reliability*: it is the ability of a system (or component) to perform its required functions under states conditions for a specified period of time (IEEE 1990). Measures of reliability are MTBF and percentage of process completed in time. A railway system is reliable when the trains run according to the timetable most of the time. There a lot of measures for the reliability of a railway system; the most used and useful are the mean and the standard deviation of the difference between the expected and the scheduled arrival time.

- *Robustness*: it indicates how much the railway system is influenced by disturbances. If the system is not robust then small disturbances cause large delays which propagate quickly.

- *Stability*: a stable system absorbs quickly the disturbances; it returns to normal operations quickly after disturbances.

- *Delay of an event*: it is the positive difference between the actual realization time and the planned time of the event (e.g. arrival/departure delay).

- *Primary delay*: it is a delay not caused by other delayed trains, but only by disturbances. It is also called initial delay or source delay.

- *Secondary delay or knock-on delay*: it is a delay caused by other delayed trains.

- *Punctuality*: it is the percentage of trains arriving within a certain margin from the scheduled arrival time and it is one of the most used performance measures used in railway systems.

The modern approach to capacity estimation refers to the Blocking Time and Headway Theory [Pachl, 2002], which requires the introduction of the following concepts:

- *Fixed Block System*: it is the typical signalling system, where the track is divided in fixed block sections protected by signals. The signal gives information about the occupation of the following block sections. Typically the signals have three aspects:
  - *red* - must stop (the section protected by the signal is occupied by a train)
  - *yellow* - must prepare to stop at next signal (the section after that protected by the signal is occupied by a train)
  - *green* - clear (the next two sections are free)
- *Block section*: it is the length of a track between two block signals.
- *Block occupation* or *blocking time*: it is the time a block section is occupied by a train, more precisely it is the time interval in which the section is exclusively allocated to a train and therefore blocked for other trains. The occupation begins when the train sees the signal (at the beginning of the preceding section) that allows approaching the section and it ends when the train passes the clearing point (as shown in Figure 1.1). Block occupations of two consecutive trains limit the minimum headway because the blocking times must not overlap to avoid train hindrance.
- *Headway distance*: it is the distance between the front ends of two consecutive trains moving along the same track in the same direction. The minimum headway is the shortest possible distance at a certain travel speed allowed by the signalling/safety system.
- *Headway time*: it is the time interval between two trains, more precisely it is the time interval between the passing of the front ends of two consecutive trains moving along the same track in the same direction.
- *Blocking time stairway*: with reference to the time-over-distance diagram, it is the sequence of the blocking times of all the sections the train passes. It allows to determine the minimum headway of the line (see Figure 1.2).
- *Signal headway*: it is the minimum time interval between two following trains and it is measured in a critical section where the two blocking

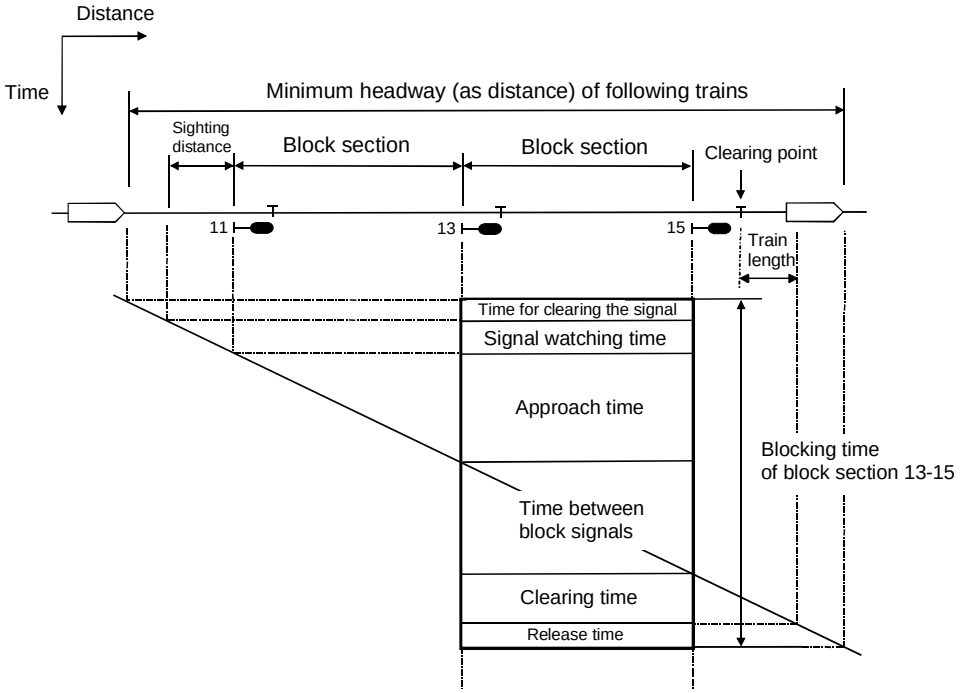


Figure 1.1: Blocking time [Pachl, 2002]

times touch each other (see Figure 1.2). For trains with different speeds, it is measured at the beginning of the line (the critical section is at the end of the line, where the stairways touch themselves).

- *Minimum Line Headway*: it is the minimum headway between two trains considering the whole blocking time stairways when they touch in a critical blocking section (see Figure 1.2).

- *Buffer time*: the time difference between actual headway (established by the timetable) and the minimum allowable headway.

- *Running time supplement*: the difference between the planned running time (established by the timetable) and the minimum running time. Infrastructure Managers typically apply a supplement of 3%-7% of the minimal running time to cope with minor delays.

- *Dwell time*: the total elapsed time the train stays in a station, from the time it stops until it resumes moving.

- *Timetable capacity*: it is the maximum number of train paths that could be scheduled considering the block occupations without considering

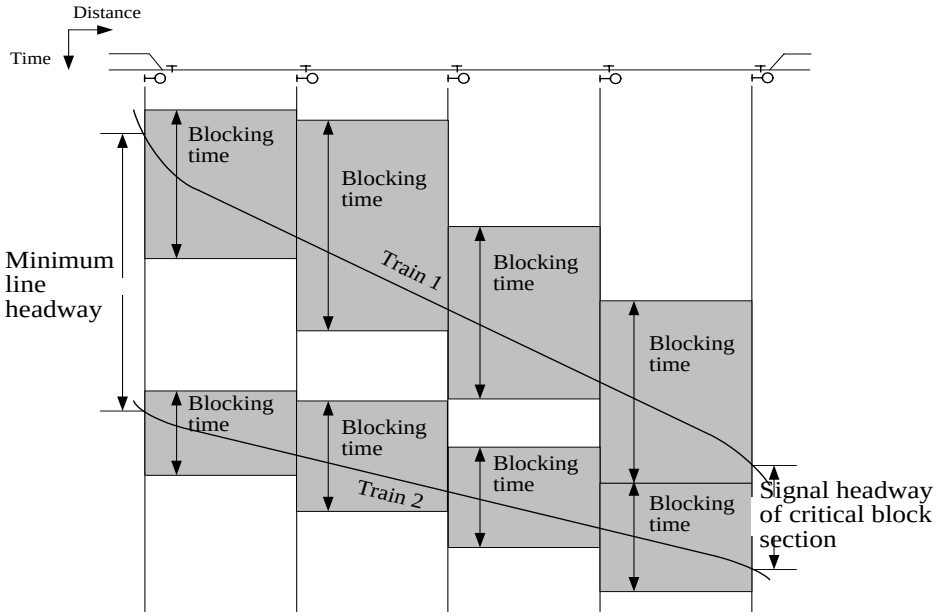


Figure 1.2: Blocking time stairways, signal headways and minimum line headway [Pachl, 2002]

buffer times.

- *Infrastructure occupation*: given a timetable, it is the result of the timetable compression described in UIC 406 R, that is the occupation time to be measured at the beginning of the first block section of the line after the timetable compression, with reference to a time window (peak hours or whole day). Detailed explanations of UIC standards are given in the next sections.

- *Total consumption time*: given a timetable, it is the sum of the Infrastructure occupation and buffer times (and other supplements) as defined in UIC 406 R ($t_{cons} = t_{occ} + t_{buff} + t_{supp}$).

- *Capacity consumption*: given a timetable, it is the ratio of the total consumption time and the time window chosen as reference, expressed as a percentage of consumption ($K = \frac{t_{cons}}{t_{win}} \cdot 100$) - def. in UIC 406 R.

- *Unused capacity*: given a timetable, it is the difference between the capacity consumption and the chosen time window. It may be divided into:

  - *usable capacity* - it is the fraction that can be used for additional train paths, providing they meet customer requirements;
  - *lost capacity* - it is the remaining time, when no further train paths can be added.

- *Moving Block System*: it is a signalling system where there are no fixed sections and the system calculates a blocking area around each moving train that no other train is allowed to enter. The lengths of the block sections are reduced to zero and the blocking time stairway becomes a continuous time channel.

1.2 Analytical method - UIC Leaflet 405

Twenty years ago the UIC proposed an analytical method [UIC, 1983], known as the UIC formula:

$$C = \frac{T}{t_{fm} + t_r + t_{zu}}$$

where:

- C = capacity (Number of trains operated in the time T)
- T = 1440 minutes (for one day)
- t_{fm} = average of minimum train headways
- t_r = running time margin
- t_{zu} = additional time

The exact calculation of t_{fm} takes into account the order of dispatch of trains belonging to different speed classes. The timetable is required to know the sequence cases where one train of type j follows one train of type i. The first version of the formula (UIC Leaflet 405-1, 1983) required the exact number of sequence cases n_{ij}, while the revised version (UIC Leaflet 405-OR, 1996) suggests a calculation of t_{fm} based only on the number of trains of each type, with a random approach for scheduling, using:

$$t_{fm} = \frac{\sum n_i \times n_j \times t_{fij}}{\sum n_i \times n_j}$$

where t_{fij} is the minimum headway between a train of type j following a train of type i.

The running time margin t_r is a breathing space added to train headways to reduce knock-on delays and to achieve an acceptable quality of service. The UIC proposed two expressions for this extra time margin:

- $t_r = 0.67 \times t_{fm}$, when the desired utilization is 0.6 ($\frac{C_{tr}}{C_{max}} = \frac{1}{1+2/3}$)
- $t_r = 0.33 \times t_{fm}$, when the desired utilization is 0.75 ($\frac{C_{tr}}{C_{max}} = \frac{1}{1+1/3}$)

The additional time t_{zu} is added to take in account the fact that the capacity decreases when the number of sections increases

- $t_{zu} = 0.25 \times a$, where a is the number of sections.

This method was officially dropped as a standard on capacity some years ago, but it lets an efficient estimation of the capacity of a line and it can be used as a reference measure and for identifying bottlenecks too.

1.3 Optimization method - UIC Leaflet 406

The UIC established the project "Capacity Management" to produce a common methodology to assess the capacity of railway infrastructures [UIC, 2004a]. The project was carried out in three phases:

- Phase 1 (2001) - Methodology: clarification of definitions, understanding of capacity and work out of a methodology to assess it.
- Phase 2 (2002) - UIC leaflet: test of the methodology (over 5000 km in Europe) and writing of the "Capacity" leaflet UIC 406 R.
- Phase 3 (2003) - Quality: get deeper knowledge of the connection between the level of the infrastructure occupation and punctuality. It is based on real time simulations to test different timetables (and simulate different kind of delays).

In the UIC leaflet 406 R (2004) the capacity knowledge is summarized as follows: "Capacity as such does not exist. Railway infrastructure capacity depends on the way it is utilised. The basic parameters underpinning capacity are the infrastructure characteristics themselves and these include the signalling system, the transport schedule and the imposed punctuality level". A qualitative model, the *capacity balance*, is introduced to represent the relationship between the capacity parameters: number of train, average speed, stability, heterogeneity. The capacity balance is illustrated in Figure

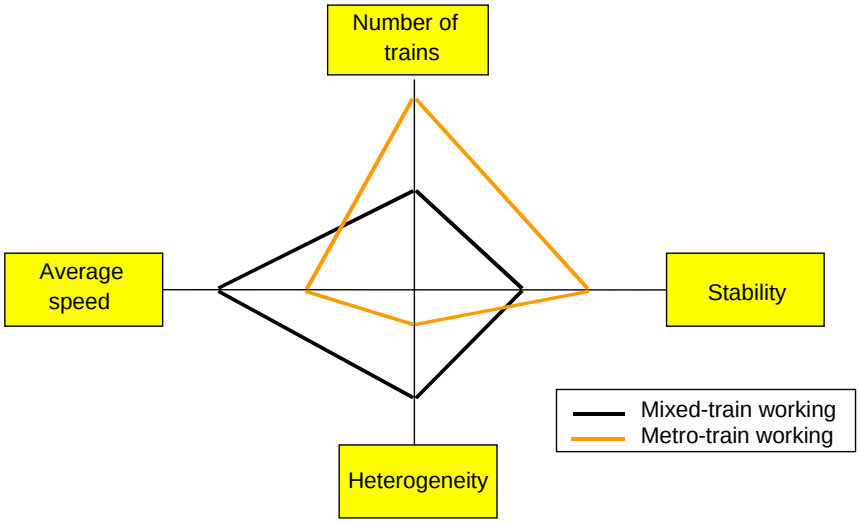


Figure 1.3: Capacity Balance - UIC 406 R

1.3. The length of the chord linking the four axes represents the capacity, while the chord positions on the four axis represent capacity utilization.

The concept of capacity depends on the point of view adopted: the different views of capacity are summarized in Figure 1.4. The market is interested in peak values, the objective of the infrastructure planner is its profitable utilization, timetabling must satisfy train paths demand and the operating view is the real time one (focused on the actual traffic and delays). Each viewpoint leads to different capacity requirements.

The Leaflet identifies some capacity-relevant constraints: priority regulations, timetable structure (integration of local and long-distance timetables), the methodology used for capacity allocation, design rules (determination of physical journey times), environmental protection (noise emissions), safety and technical constrains (specific traffic routes, braking system, power supply) and the theoretical capacity (upper bound).

After the deep illustration of the concept, the *capacity of a railway infrastructure* is defined as the total number of possible path in a defined time window, considering the actual path mix with market-oriented quality (and taking account of the Infrastructure Manager requirements). It coincides with the *Timetable Capacity*, as defined in the concepts section.

The market needs are represented by customer (Railway Undertaking) requirements, expressed in form of path requests characterized by two parameters: typical running time (depending on rolling stock) and departure/arrival

Market (customer needs)	Infrastructure planning	Timetable planning	Operations
expected number of train paths (peak)	expected number of train paths (average)	requested number of train paths	actual number of trains
expected mix of traffic and speed (peak)	expected mix of traffic and speed (average)	requested mix of traffic and speed	actual mix of traffic and speed
infrastructure quality need	expected conditions of infrastructure	existing conditions of infrastructure	actual conditions of infrastructure
journey times as short as possible	time supplements for expected disruptions	time supplements for expected disruptions	delays caused by operational disruptions
translation of all short and long-term market-induced demands to reach optimised load	maintenance strategies	time supplements for maintenance	delays caused by track works
		connecting services in stations	delays caused by missed connections
		requests out of regular interval timetables (system times, train stops, ...)	additional capacity by time supplements not needed

Figure 1.4: Different views of capacity - UIC 406 R

time. Capacity is allocated through an iteration process that occurs between RU requirements and IM offers and which always takes in account punctuality requirements (time supplements and buffer time).

Additional definitions given in the Leaflet refer to the railway network elements (corridor, route, line, nodes, stations, junctions, line sections, relevant block section).

The central topic of the Leaflet is the **Calculation of capacity consumption**. The capacity examination requires an existing pre-constructed timetable for the examined infrastructure. The analysis is performed by calcultating the capacity consumption within a line of the infrastructure and it is based on the **compression** of the **timetable** train paths in a pre-defined time window.

The analysis of a route requires the calculation of the capacity consumption of the lines of the route: the capacity consumption of the route is the highest value of the line-consumptions. The time window to be chosen shall be a peak period (one or two hours long) in one representative day (e.g. Thursday) or the whole representative day.

The *compression process* must follow some rules:

- all train paths are pushed together up to the minimum headway according to their timetable order, without buffer times;

- the running times, overtakings, crossings, stopping times (requested by RUs) are not changeable;

- any occupation times must be incorporated, also indirect occupation times (times occupied and not available for further train paths).

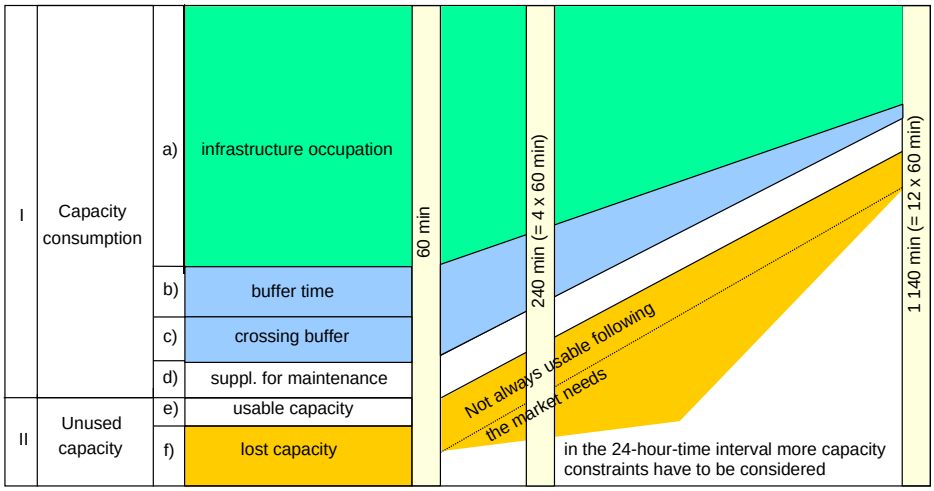


Figure 1.5: Determination of capacity consumption - UIC 406 R

The *capacity consumption* is the sum of the infrastructure occupation and the buffer times (plus other supplements) and it may be determined on the basis of the compressed timetable, using the following formula:

$$K = \frac{k}{U} \cdot 100$$

where

- K = capacity consumption [%]
- U = chosen time window [min] (I+II in Fig. 1.5)

and

$$k = A + B + C + D$$

where

- k = total consumption time [min]
- A = infrastructure occupation [min]
- B = buffer time [min]

- C = supplement for single-track lines [min]

- D = supplement for maintenance [min]

Running times are not changed by the compression. **Running time supplements** are assumed to be 5% of journey times and must be considered as part of the buffers. *Buffer times* serve to reduce transfer of delays from one train to the next (secondary delays) and shall be introduced either between each path or globally.

The **Infrastructure Occupation** is the result of the compression process and it is measured at the beginning of the first block of the line; it may contain the supplement for maintenance (otherwise it must be specified separately). The *Unused Capacity* can be used for additional train paths - satisfying customer requirements - until a maximum fraction called *Usable Capacity*; the remaining useless fraction is called *Lost Capacity*.

The Leaflet defines the condition under which an *infrastructure* is declared *congested*, that is the infrastructure is congested or there is a bottleneck if:

- the infrastructure occupation is greater than a "typical value" (no additional route can be "easily" inserted), or

- shifting of routes to add a path is so extensive that market requirements can no longer be met.

The **iterative application** of capacity consumption is a methodology of capacity assessment: after the compression of the given timetable, the unused capacity must be checked by trying to incorporate further additional train paths of appropriate type (it depends on the possible RU demands) and recompressing the new timetable. This procedure shall be repeated until either the infrastructure becomes congested or no more paths can be incorporated.

The compression methodology was applied on over 5000 km, to test it and to obtain *recommended values* of the infrastructure occupation time that must not be exceeded. The results are figures of 75%-85% in the peak hour and 60%-70% in a daily period.

The Leaflet ends with some considerations about the length of the line on which the compression is applied. The recommendations are to use the compression on short line setions (e.g. line between two stations) and to apply the enrichment process (new path addition trial) on longer line sections taking into account the type of traffic.

1.4 Recent literature review

The most recent and relevant literature reference is the paper "An assessment of railway capacity" [Abril et al., 2007]. It reviews the main concepts and methods to perform capacity analyses and it also briefly describes computer-based systems. After defining the concepts, it illustrates three classes of methods used to evaluate railway capacity:

- *Analytical methods*: they model the railway infrastructure by means of mathematical expressions, in a simple manner so that a preliminary solution is easily determined. The UIC 405 method and the "absolute capacity" [Burdett and Kozan, 2006] on theoretical capacity are worth mentioning. They are a good start point for identifying bottlenecks and constraints, although analytical model outputs are very sensitive to parameter inputs and train mix scheduling.

- *Optimization methods*: they are based on obtaining optimal saturated timetables. There are a lot of methods based on Operational Research methods, but "The method" is the UIC timetable compression method where the saturation is reached by iterative addition of train paths to the compressed scheduling (UIC 406 R). Some papers by Landex describe the application of this method in Denmark.

- *Simulation methods*: they provide a model, very close to reality, to validate a given timetable. Various commercial simulation environments (Multirail, OpenTrack, SIMONE) have been produced and they normally generate timetables by simulation using train motion differential equations.

The current trend is to develop tools that integrate the three approaches, each covering a phase of capacity management: analytical for the preliminary solution, optimization for timetable generation and simulation for timetable validation.

Finally the paper illustrates capacity analyses on the ERTMS (European Rail Traffic Management System), which features are: interoperability, highest speed (up to 500 km/h), automatic train protection (ATP prevents collision by automatic braking), smaller headways and moving block operation.

The focus of the present research work is the relationship between capacity (consumption) and actual running times, as will be illustrated in the next section. Recent works about running times, primary and secondary delays that are worth mentioning:

- [Mattson, 2004] - it is a review of the literature on delays and the relationships between them and the actions that may be taken. The analytical section is mainly based on the following work:

- [Huisman and Boucherie, 2001] - this paper investigates the distributions of the running times when the train traffic is heterogeneous, that is the trains have different speeds. In the paper the free running times are deterministic, so two scenarios are considered: in the first all the delays are secondary delays, in the second the distribution of primary delays is assumed to be exponential (delay caused by longer boarding at the stations); in both cases a relationship of the mean delay versus a parameter (train flow or mean primary delay) is obtained.

- [Yuan and Hansen, 2007] - the paper proposes an analytical stochastic model of train delay propagations by estimating the secondary delays caused by route conflicts and late transfer connections. The conditional distribution of the arrival time of an approaching train at the platform is obtained as the convolution of several individual distributions, each referred to a condition.

- [Yuan et al., 2006] - the paper evaluates the fitting of train process time distributions to commonly applied distribution models. The fitting statistics used is the Kolmogorov-Smirnov at a certain significance level. The log-normal distribution is the best fitted one for arrival times of trains.

- [Meester and Muns, 2007] - the paper discuss a fairly general model for delay propagation. In essence the model is that of Carey and Kwiecinski, 1995 and it could be called a *stochastic event graph*. It shows that it is possible to derive secondary delay distributions from primary delay distribution, assuming the so-called phase-type distributions environment. The paper distinguishes the *free running time*, that is the running time that results if the effects of other trains are not taken into account, from the *hindered running time* that incorporates influences of other trains. The process times of the model are the free process times, only their distributions are to be specified and the mixture of these distributions gives the probability of hindrance.

1.5 Actual running times and capacity consumption

While many efforts have been made to find relationships between primary and secondary delays, the actual running time has not been fully investigated from the stochastic modelling point of view:

- [Huisman and Boucherie, 2001] solve deterministic differential equations and add an exponential distributed primary delay which only models the stochastic boarding time;
- [Yuan et al., 2006] consider the arrival time distributions, that is the final act of a travel during which many stochastic events may alter the deterministic running time;
- other works consider the primary delay distribution without trying to model its generation process.

In the present work the actual free running time will be modelled using stochastic differential equations, that is differential equations describing the movement of the train with a stochastic perturbation driven by a Brownian motion, which is a continuous paths Levý process (a stochastic process with stationary independent increments).

It is possible to obtain a measure of the infrastructure occupation (and of the buffer times needed) using the stochastic free running times, by solving the stochastic differential equations and evaluating the probability of hindrance between trains.

Instead of using the deterministic blocking time stairways, the capacity will be put in relationship with the probability of hindrance, thought as a risk probability. It is also possible to obtain the stochastic distributions of the blocking times at the set and clearing points, so that the timetable may be compressed "in a stochastic way" considering the level of hindrance when the blocking sections get closer.

Buffer times are implicitly considered when working in such a stochastic framework (if the traffic flow is low there is no hindrance), more precisely they are considered together with the infrastructure occupation at a given level of hindrance risk.

In conclusion: given a risk level of hindrance, the capacity consumption (infrastructure occupation plus buffer times) may be estimated using free running times modelling based on stochastic differential equations.

Chapter 2

Modelling the stochastic primary delay with Lévy processes

When modelling a physical phenomenon with strong stochastic characterization like the primary delay, there is a small amount of knowledge about the data generation process, so it is reasonable to make as few assumptions as possible, which corresponds to the choice of a very wide class of stochastic processes. The family of Lévy processes may be used in primary delay stochastic modelling because a Lévy process is defined only through the following simple conditions:

- the initial value of the process is zero almost surely, i.e. the start value is zero with probability one;
- independent increments, i.e. an increment (difference of the values taken in two different times) is independent from the past values of the process;
- stationary increments, i.e. increments' distributions depend only on time distances;
- stochastic continuity, i.e. the probability of a non-zero increment tends to zero with the time distance.

The first property makes the environment more comfortable because it frees the theory of Lévy processes from the need of highlighting the initial value. The second property is often referred to as the "Markov property" because it can be also thought as "the future outcomes only depend on the present value" and it needs a more formal definition to be well understood. The third

property states that the structure of the evolution process is the same at all the time and it seems to be a quite natural assumption. The fourth property speaks about continuity in a probabilistic sense, which is quite natural but allows the process to have non-continuous sample paths. A more formal definition of the Lévy processes is necessary for a better understanding of the properties and of a fundamental theorem (Lévy-Itô) about the decomposition of the generic Lévy process into components belonging to the sub-families of Brownian motions and Poisson processes.

2.1 Basic notions about Lévy processes

A quick analysis of the properties of the Lévy processes class and its subclasses (Brownian motions and Poisson processes) needs a mathematical framework which uses notions from Measure Theory.

The following definitions are taken from the literature on Lévy processes [Protter, 2004] [Karatzas and Shreve, 1998] [Kyprianou, 2005] [Applebaum, 2004].

Definition 2.1 *A complete probability space $(\Omega, \mathcal{F}, P)$ is assumed as given. It is a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, P)$ if a filtration $\mathbb{F}$ is defined on it. A **filtration** $\mathbb{F}$ is a family of sub-σ-algebras of $\mathcal{F}$ that is increasing on the positive time line $T=R_+$*

$\mathbb{F} = \{\mathcal{F}_t : \mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}\ \forall s < t,\ s, t \in T\}$

Definition 2.2 *A filtered complete probability space is said to satisfy the **usual hypothesis** if*

(i) $\mathcal{F}_0$ contains all the P-null sets

(ii) $\mathcal{F}_t = \cap_{u>t}\mathcal{F}_u\ \forall t,\ 0 \leq t < \infty$; that is, the filtration $\mathbb{F}$ is right continuous

The assumption that the usual hypothesis hold and all stochastic processes are defined on a complete stochastic basis $(\Omega, \mathcal{F}, \mathbb{F}, P, T)$ is made throughout the thesis.

Definition 2.3 *A stochastic process X_t on $(\Omega, \mathcal{F}, P)$ is said to be **adapted to the filtration** $\mathbb{F}$ if it is $\mathcal{F}_t$ measurable for each t.*

The typical interpretation of the σ-algebra $\mathcal{F}_t$ is the accumulated information up to time t, so it represents the history (up to time t) of the process in the following definitions.

Definition 2.4 *A real-valued, adapted process $X=(X_t)_{0\leq t<\infty}$ is called a **martingale** (resp. **supermartingale**, **submartingale**) with respect to the filtration $\mathbb{F}$ if:*

(i) $X_t \in L^1(dP)$*; that is,* $E\{|X_t|\} < \infty$*;*

(ii) if $s \leq t$*, then* $E\{X_t|\mathcal{F}_s\}$=X_s *a.s ; (resp.* $E\{X_t|\mathcal{F}_s\} \leq X_s$*, resp.* $\geq X_s$*)*

Definition 2.5 *A real-valued stochastic process* X_t*, adapted to the filtration* $\mathbb{F}$ *is said to be* ***Markov with respect to*** $\mathbb{F}$ *if*

$$P[X_t \in \Gamma|\mathcal{F}_s] = P[X_t \in \Gamma|X_s] \quad \forall \Gamma \subset \mathbb{R}$$

The Markov property is true when the information given by the value X_s is equivalent to the complete information up to time s, i.e. the σ-algebra $\mathcal{F}_s$.

Definition 2.6 *Two stochastic processes are* ***modifications*** *if* $X_t = Y_t$ *almost surely, each t.*

If X and Y are modifications there exists a null set which depends on t, N_t, such that if $\omega \notin N_t$ then $X_t(\omega) = Y_t(\omega)$. The union $\cup_{0 \leq t < \infty} N_t$ could even be non-measurable.

Definition 2.7 *Two stochastic processes are* ***indistinguishable*** *if almost surely, for all t,* $X_t = Y_t$*.*

If X and Y are indistinguishable there exists a null set which doesn't depend on t, N, such that if $\omega \notin N$ then $X_t(\omega) = Y_t(\omega)$. The functions $t \mapsto X_t$ and $t \mapsto Y_t$ are the same for all $\omega \notin N$, where $P(N) = 0$.

Definition 2.8 *The functions* $t \mapsto X_t$ *are called the* ***sample paths*** *of the stochastic process X.*

The sample paths play a fundamental role in the analysis of the risky events on a railway, where the distance of two trains is computed using their sample positions.

Definition 2.9 *A stochastic process X is said to be* ***càdlàg*** *(continu à droite, limites à gauche) if it almost surely has sample paths which are right continuos, with left limits (****rcll****).*

Theorem 2.10 *If X is a martingale, then there exists a unique modification Y of X which is càdlàg.*

The notion of càdlàg is necessary to "capture" possible jumps in sample paths and it turns out that it covers all the processes of the Lévy family through modifications:

Definition 2.11 *An adapted process* $X = (X_t)_{t \geq 0}$ *is a* ***Lévy process*** *if*

(i) $X_0 = 0$ *a.s.; that is* $P(\ X_0 = 0\ ) = 1$ *;*

(ii) *X has independent increments; that is,* $X_t - X_s$ *is independent of* $\mathcal{F}_s$*,* $0 \leq s < t < \infty$*;*

(iii) *X has stationary increments; that is,* $X_t - X_s$ *has the same distribution as* X_{t-s}*,* $0 \leq s < t < \infty$*;*

(iv) *X is stochastically continuous (or continuous in probability); that is, for every* $t \geq 0$ *and* $\epsilon > 0$ $lim_{s \to t} P(|X_s - X_t| > \epsilon) = 0$

Theorem 2.12 *Let X be a Lévy process. There exists a unique modification Y of X which is càdlàg and which is also a Lévy process.*

It will be always assumed to work with the càdlàg modification of a Lévy process, so $X(t) - X(t_-)$ will be non-zero in the event of a jump. It's worth noticing that some authors use the *càdlàg* property directly as part of the Lévy family definition, but the *stochastically continuous* property is wider and therefore more suitable for modelling purposes. Two important sub-families are identified by adding properties like continuity and distribution law:

- Brownian motions: a.s. continuous paths and normal distribution;
- Poisson processes: (càdlàg and) Poisson distribution.

2.2 Brownian motion

A formal definition for the sub-family of the Brownian motions is:

Definition 2.13 *An adapted process* $B = (B_t)_{0 \leq t \leq \infty}$ *taking values in* $\mathbb{R}^n$ *is called an n-****dimensional Brownian motion*** *if*

(i) *for* $0 \leq s < t < \infty$*,* $B_t - B_s$ *is independent of* $\mathcal{F}_s$ *(increments are independent of the past);*

(ii) *for* $0 \leq s < t < \infty$*,* $B_t - B_s$ *is a Gaussian random variable with mean zero and variance matrix (t-s)C, for a given, non-random matrix C.*

Theorem 2.14 *Let B be a Brownian motion. There exists a modification of B which has continuous paths a.s.*

It will be always assumed to work with the version of a Brownian motion with continuous paths. If C is the identity matrix, i.e. C=I, the process is called a **standard Brownian motion** or **standard Wiener process**.

Some authors follow another way, assuming the *continuous sample paths* property and deriving normality, e.g. Jackel in [Jackel, 2003] presents a theorem that states that "If Y is a continuous process with stationary independent increments, then Y is a Brownian motion" and the following citation from Harrison: "This beautiful theorem shows that Brownian motion can actually be defined by stationary independent increments and path continuity alone, with normality following as a consequence of these assumptions. This may do more than any other characterization to explain the significance of Brownian motion for probabilistic modeling".

In the present work the normality is part of the Brownian motion definition and the importance of its continuous paths will be highlighted by the decomposition theorem.

2.3 Poisson process

The other interesting sub-family is that of Poisson processes:

Definition 2.15 *A process valued on the non-negative integers $N = \{N_t : t \geq 0\}$, is said to be a **Poisson process** with intensity $\lambda > 0$ if the following hold:*

(i) $P(N_0 = 0) = 1$;

(ii) *N has independent increments; that is, $N_t - N_s$ is independent of $\mathcal{F}_s$, $0 \leq s < t < \infty$;*

(iii) *N has stationary increments; that is, $N_t - N_s$ has the same distribution as N_{t-s}, $0 \leq s < t < \infty$;*

(iv) *the paths of N are almost surely right continuous with left limits;*

(v) *for each $t > 0$, N_t is equal in distribution to a Poisson random variable with parameter λt, i.e. $P(N_t = n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$.*

A counting process like the Poisson process leads to the definition of the compound Poisson process, which is a Lévy process too.

Definition 2.16 *A **compound Poisson process** is defined as:*

$$C_t = \sum_{k=1}^{N_t} Y_k \quad t \geq 0$$

where N_t is a Poisson process and $Y_k, k \geq 1$ are independent identically distributed random variables.

While the Poisson process may count the jumps in the sample path of a Lévy process, the compound Poisson process may be used to measure the total size of these jumps up to time t.

The rigorous treatment of the jumps requires the notion of **Poisson point process**. The full precise definition is not presented here because it is far beyond the purpose of this work, but the underlying idea is very simple: it is written as $\Delta X_s = X_s - X_{s-}$ and it represents a jump at time s, so the compound Poisson process $\sum_{s \leq t} \Delta X_s$ is the stochastic process performing all these jumps.

2.4 Lévy-Itô decomposition

The notion of infinitely divisible distribution must be introduced to understand process decomposition:

Definition 2.17 *A real-valued random variable Θ has an **infinitely divisible distribution** if for each n=1,2... there exist a sequence of i.i.d random variables $\Theta_{1,n}$,..., $\Theta_{n,n}$ such that $\Theta \overset{d}{=} \sum_i \Theta_{i,n}$ where $\overset{d}{=}$ is equality in distribution.*

For a Lévy process, X_t, $t > 0$, must be divisible into $n \geq 2$ i.i.d. (independent identically distributed) random variables

$$X_t = \sum_{k=1}^{n} (X_{tk/n} - X_{t(k-1)/n})$$

since these are successive increments (hence independent) of equal length (hence identically distributed, by stationarity).

Theorem 2.18 (Lévy-Khintchine) *A random variable Z is infinitely divisible if and only if its characteristic function $\mathbb{E}[e^{i\theta Z}]$ is of the form*

$$exp\left\{ i\beta\theta - \frac{1}{2}\sigma^2\theta^2 + \int_{\mathbb{R}\setminus\{0\}} (e^{i\theta x} - 1 - i\theta x 1_{\{|x|\leq 1\}})\nu(dx) \right\}$$

where $\beta \in \mathbb{R}$, $\sigma^2 \geq 0$ and ν is a measure on $\mathbb{R}\setminus\{0\}$ such that

$$\int_{\mathbb{R}\setminus\{0\}} (1 \wedge x^2)\nu(dx) < \infty$$

For a Lévy process the random variable X_1 (i.e. X_t when t=1) has an infinitely divisible distribution, so its characteristic function has a compound expression which corresponds to a decomposition of the process. The exponent of the characteristic function can be split in four parts: the first two addends, which correspond to a scaled brownian motion with drift, and the integral, which represents the collection of jumps of the process and can be divided in two parts distinguishing the small jumps from the big ones.

Theorem 2.19 (Lévy-Itô) *Let X be a Lévy process, the distribution of X_1 parametrized by (β, σ^2, ν). Then X decomposes*

$$X_t = \beta t + \sigma B_t + J_t + M_t$$

where B is a Brownian motion, and $\Delta X_t = X_t - X_{t-}, t \geq 0$, an independent Poisson point process with intensity measure ν,

$$J_t = \sum_{s \leq t} \Delta X_s 1_{\{|\Delta X_s| > 1\}}$$

and M is a martingale with jumps $\Delta M_t = \Delta X_t 1_{\{|\Delta X_t| \leq 1\}}$

The theorem states that a generic Lévy process decomposes in three independent Lévy processes:

- a scaled Brownian motion with drift: $\beta t + \sigma B_t$;
- a compound Poisson process J_t with jumps which are of magnitude greater or equal than unity;
- a square integrable martingale M_t with an almost surely countable number of jumps in each finite interval which are of magnitude less than unity and a characteristic function given by the Lévy-Khintchine theorem considering only the integral restricted to the sub-domain of small jumps $\int_{0<|x|<1} \nu(dx)$.

The Brownian motion has continuous sample paths, while the other two components have jumps in their paths. The Poisson jumps have magnitude greater than one and the martingale component is necessary to make fractional jump size adjustments. The set of jumps has been split in two subsets for countability purposes: in any finite time the process can have only a finite number of jumps of size greater than one, so it is possible to write J_t as a sum; the other subset may contain a large number of very small jumps and it is possible to write the limit of their sum as a square integrable martingale M_t. Paul Lévy argued that the accumulation of a large number of very small jumps may be difficult to distinguish from bursts of deterministic motion and K. Itô found the expression of the limit.

2.5 Lévy processes and stochastic modelling

Lévy processes are a family of stochastic processes characterized by very few assumptions - independent stationary increments and stochastic continuity - which make them very suitable for modelling real stochastic processes. The decomposition theorem gives some invaluable indications for stochastic modelling:

- if the real process to be modelled is supposed to have continuous sample paths then the Lévy process will have only the Brownian motion component, hence the Brownian motion is the cornerstone of stochastic modelling in a continuous environment
- if the real process to be modelled is supposed to have jumps in the sample paths then there are two possible situations to manage:
  - if the process exhibits finitely many jumps per unit time, then the compound Poisson process is easy to manage by equations and simulations;
  - if the process exhibits infinitely many jumps per unit time, then everything needs a special treatment.

Chapter 3

Stochastic differential equations

The position of a train along a line in a deterministic environment can be described through the general differential equation taken from the classical mechanics theory

$$\ddot{x}(t) = f(x(t), \dot{x}(t), \vec{u}(t), t)$$

where $\vec{u}(t) = (u_1(t), \ldots, u_m(t))$ represents the inputs to the mechanical system (external agents, i.e. forces). For the sake of compactness it can be written as

$$\dot{\vec{x}}(t) = \vec{f}(\vec{x}(t), \vec{u}(t), t)$$

where $\dot{\vec{x}} = (x_1, x_2)' \triangleq (\dot{x}, x)'$ and $\vec{f} = (f_1, f_2)' \triangleq (f, x_1)'$.

The model may be enriched with the introduction of a stochastic component to describe the travel in a more realistic way. Lévy processes are characterized by independent stationary increments so it seems natural to introduce randomness in the model by adding a Lévy process component to the deterministic increment of $\vec{x}$, which is given by $\vec{f}(\vec{x}(t), \vec{u}(t), t)dt$:

$$d\vec{x}(t) = \vec{f}(\vec{x}(t), \vec{u}(t), t)dt + d\vec{L}_t$$

If L_t is assumed to have continuous sample paths then it is a Brownian motion, so its increment can be written as $L_t = \sigma W_t$ where W_t is the standard Brownian motion or Wiener process. The following properties hold:

- continuous sample paths $\Rightarrow dL_t = \sigma dW_t$
- $dW_t \sim N(0, dt)$ (normal distribution, zero mean, variance dt)
- $E[dW_t] = 0$
- $E[dW_t^2] = dt$

- $E[(\frac{dW_t}{dt})^2] = \frac{1}{dt} \to \infty$ if $dt \to 0$, i.e. nowhere differentiability

Brownian motion cannot be differentiated, therefore the differential relationships cannot be written using derivatives but only differentials and they have only a symbolic meaning which must be related to the corresponding stochastic integral equations. In the general case, the stochastic differential equation

$$dX_t(\omega) = a(t, X_t(\omega))dt + b(t, X_t(\omega))dW_t(\omega)$$

is symbolic for the stochastic integral equation

$$X_t(\omega) = X_{t_0}(\omega) + \int_{t_0}^{t} a(s, X_s(\omega))ds + \int_{t_0}^{t} b(s, X_s(\omega))dW_s(\omega)$$

where $\omega \in \Omega$ has been highlighted to point out that there is an integral for each sample path. For each sample path the first integral is an ordinary deterministic integral, while the second one is not ordinary and may not exist pathwise: it is stochastic and there is more than one way to define it.

3.1 Itô and Stratonovich integrals

The literature about stochastic integration [Kloeden and Platen, 1999] [Oksendal, 2000] offers two popular formulations corresponding to two different ways of building the sequence of random variables leading to a stochastic limit: Itô integral and Stratonovich one.

3.2 Itô integral

The Itô integral is defined in a manner similar to the Riemann-Stieltjes integral, that is as a limit in probability of Riemann sums; such a limit does not necessarily exist pathwise. The definition steps are:

- Suppose that $W : [t_0, t] \times \Omega \to R$ is a Wiener process
- Suppose that $X : [t_0, t] \times \Omega \to R$ is a stochastic process adapted to the filtration $(\mathcal{F}_s)_{t_0 \le s \le t}$ generated by the Wiener process, that is X_s is $\mathcal{F}_s$ measurable for all $s \in [t_0, t]$
- Partition $[t_0, t]$ as $t_0 = t_0^{(n)} < t_1^{(n)} < \cdots < t_{n-1}^{(n)} < t_n^{(n)} = t$ with $\delta_n = max(t_k^{(n)} - t_{k-1}^{(n)})$ such that $\lim_{n\to\infty} \delta_n = 0$
- Form the sum $I_n(t) = \sum_{k=1}^{n} X_{t_{k-1}^{(n)}} \left[W_{t_k^{(n)}} - W_{t_{k-1}^{(n)}} \right]$

- the Itô integral of X with respect to W is the limit random variable to which the sequence $I_n(t)$ converges:

$$I_I(t) = \int_{t_0}^{t} X_s dW_s \triangleq \underset{n\to\infty}{plim} I_n(t)$$

- Itô showed that sequence $I_n(t)$ converges in probability and in mean square. Kunita and Watanabe showed that the sequence converges in the general case that W is a martingale.

Itô integral properties:

- $I_I(t)$ is $\mathcal{F}_t$ measurable (and hence non-anticipating)
- $I_I(t)$ is a martingale, i.e. $E[I_I(t)|\mathcal{F}_s] = I_I(s)$
- $I_I(t)$ has continuous sample paths with probability 1

3.3 Stratonovich and other integrals

The definition steps are the same of the Itô integral, but the sums are built using $X_{\frac{t_k^{(n)}+t_{k-1}^{(n)}}{2}}$ instead of $X_{t_{k-1}^{(n)}}$

- Suppose that $W : [t_0, t] \times \Omega \to R$ is a Wiener process
- Suppose that $X : [t_0, t] \times \Omega \to R$ is a stochastic process adapted to the filtration $(\mathcal{F}_s)_{t_0 \le s \le t}$ generated by the Wiener process, that is X_s is $\mathcal{F}_s$ measurable for all $s \in [t_0, t]$
- Partition $[t_0, t]$ as $t_0 = t_0^{(n)} < t_1^{(n)} < \cdots < t_{n-1}^{(n)} < t_n^{(n)} = t$ with $\delta_n = max(t_k^{(n)} - t_{k-1}^{(n)})$ such that $\lim_{n\to\infty} \delta_n = 0$
- Form the sum $S_n(t) = \sum_{k=1}^{n} X_{\frac{t_k^{(n)}+t_{k-1}^{(n)}}{2}} \left[W_{t_k^{(n)}} - W_{t_{k-1}^{(n)}} \right]$
- the Stratonovich integral of X with respect to W is the limit random variable to which the sequence $I_n(t)$ converges:

$$I_S(t) = \int_{t_0}^{t} X_s \circ dW_s \triangleq \underset{n\to\infty}{plim} S_n(t)$$

- It is possible to show that the sequence $I_n(t)$ converges in probability and in mean square.

Stratonovich integral has continuous sample paths but it is not a martingale and it is anticipating (it "looks into the future").

It is possible to define other stochastic integrals changing the choice of the evaluation points $\xi_k^{(n)} \in [t_{k-1}^{(n)}, t_k^{(n)}]$ and generally every choice leads to a different limit. A family of integrals is defined by the evaluation points systematically chosen as follows:

$$\xi_k^{(n)} = (1-\lambda)t_{k-1}^{(n)} + \lambda t_k^{(n)}$$

for the same fixed $0 \leq \lambda \leq 1$.

Itô and Stratonovich integrals correspond to $\lambda = 0$ and $\lambda = \frac{1}{2}$ respectively. The sums use $X_{(1-\lambda)t_{k-1}^{(n)}+\lambda t_k^{(n)}}$ but frequently the λ-integrals are defined using $(1-\lambda)X_{t_{k-1}^{(n)}} + \lambda X_{t_k^{(n)}}$, which is equivalent (Taylor expansion), it is often simpler to evaluate and leads to a relationship between the integrals of the family: $I_\lambda = (1-\lambda)I_0 + \lambda I_1$.

For a Wiener process the integrals $\int_{t_0}^t w(s)dw_s$ of the family are:

$$I_\lambda = \frac{1}{2}(w^2(t) - w^2(t_0)) + (\lambda - \frac{1}{2})(t - t_0)$$

therefore the normal rules of calculus only work for the Stratonovich integral.

3.4 Itô and Stratonovich formulations comparison

The Stratonovich framework - stochastich differential equations and related integrals - has some advantages:

- the formal rules of ordinary calculus such as integration by parts, changes of variables, and the chain rule hold for the Stratonovich approach;

- the numerical schemes of deterministic differential equations may be used to find numerical solutions of Stratonovich stochastich differential equations;

- it is possible to approximate a Wiener process with a smooth process, solve the approximating differential equation using Lebesgue integration, and then consider the limit of the solution processes as the smooth process converges to the Wiener process.

The disadvantage of the Stratonovich approach is that the integral is anticipating and it does not yield a martingale. This is an important issue when modelling random phenomena in physical systems, where it is natural to make the assumption of future conditional expectations equal to the last known value of the process.

The martingale property of the Itô integral is, by far, the most important from the modelling point of view and it justifies the choice of the Itô framework from now on.

The Itô framework - stochastich differential equations and related integrals - has the following fundamental advantages:

- the integral is non-anticipating and it is a martingale;
- there is no need for approximations: it is possible to work with the original Wiener process.

The disadvantages of the Itô formulation are:

- the formal rules of ordinary calculus do not hold: stochastic differentials, which are interpreted in terms of stochastic integrals, do not transform according to the chain rule of classical calculus but follow a modified one. The stochastic chain rule, called the Itô formula, has an additional term due to the first order magnitude (Wiener process: $E[dW_t^2] = dt$) of a second order term dX_t^2 in the Taylor expansion.
- the numerical schemes of deterministic differential equations must be modified according to the Itô formula.

It is possible to switch from one formulation to the other using a drift correction. It easy to show that if the process X_t satifies the Itô SDE

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t$$

then it also satisifies the Stratonovich SDE

$$dX_t = \underline{a}(t, X_t)dt + b(t, X_t) \circ dW_t$$

with modified drift $\underline{a}$ defined by

$$\underline{a}(t, X_t) = a(t, X_t) - \frac{1}{2}b(t, x)\frac{\partial b}{\partial x}(t, x)$$

The Itô and Stratonovich SDEs have the same coefficients if the diffusion coefficient b is independent of x.

The natural-for-modelling Itô interpretation will be assumed for every stochastic integral and differential equation that will be introduced from now on.

3.5 Itô formula (stochastic chain rule)

The Itô formulation introduces a modification in the chain rule of calculus, which needs an additional term due to the first order magnitude of a second order term $E[dW_t^2] = dt$ (Wiener process) in the Taylor expansion of the stochastic differential.

- In a deterministic environment the differential of X_t is

$$dX_t = a(t, X_t)dt$$

and the chain rule of ordinary calculus states that if Y_t is a continuously differentiable function

$$Y_t = f(t, X_t)$$

then the differential is

$$\begin{aligned} dY_t &= \tfrac{\partial f}{\partial t}dt + \tfrac{\partial f}{\partial x}dX_t \\ &= \left[\tfrac{\partial f}{\partial t} + \tfrac{\partial f}{\partial x}\,\tfrac{dX_t}{dt}\right]dt \quad (chain\ rule) \\ &= \left[\tfrac{\partial f}{\partial t} + a\,\tfrac{\partial f}{\partial x}\right]dt \end{aligned}$$

- In a stochastic environment the differential of X_t is

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t$$

and the stochastic differential of a twice continuously differentiable function

$$Y_t = f(t, X_t)$$

is obtained by taking into account the first and the second order terms of the Taylor expansion, and considering only the second order term dX_t^2, which contains dW_t^2 and it is of first order magnitude dt because $E[dW_t^2] = dt$ for the Wiener process:

$$\begin{aligned} dY_t &= \tfrac{\partial f}{\partial t}dt + \tfrac{\partial f}{\partial x}dX_t + \tfrac{1}{2}\left\{\tfrac{\partial^2 f}{\partial t^2}dt^2 + 2\tfrac{\partial^2 f}{\partial t \partial x}dt\,dX_t + \tfrac{\partial^2 f}{\partial x^2}dX_t^2\right\} \\ &= \tfrac{\partial f}{\partial t}dt + \tfrac{\partial f}{\partial x}dX_t + \tfrac{1}{2}\tfrac{\partial^2 f}{\partial x^2}\,dX_t^2 \\ &= \tfrac{\partial f}{\partial t}dt + \tfrac{\partial f}{\partial x}dX_t + \tfrac{1}{2}\,\tfrac{\partial^2 f}{\partial x^2}\,b^2\,dW_t^2 \\ &= \left[\tfrac{\partial f}{\partial t} + \tfrac{1}{2}\,\tfrac{\partial^2 f}{\partial x^2}\,b^2\right]dt + \tfrac{\partial f}{\partial x}dX_t \\ &= \left[\tfrac{\partial f}{\partial t} + a\tfrac{\partial f}{\partial x} + \tfrac{1}{2}\,b^2\,\tfrac{\partial^2 f}{\partial x^2}\right]dt + b\tfrac{\partial f}{\partial x}\,dW_t \quad (stochastic\ chain\ rule) \end{aligned}$$

The term $\frac{1}{2}\,b^2\,\frac{\partial^2 f}{\partial x^2}$ is the only true modification due to the stochastic environment driven by a Wiener process, because the other terms would be always present in a deterministic environment when developing the differential dY_t in the general case of $X_t = X(t, W)$ with t and W independent variables, i.e. $dX_t = a\,dt + b\,dW$ with $a = \frac{\partial X}{\partial t}$ and $b = \frac{\partial X}{\partial W}$.

The Itô formula or stochastic chain rule may be written more compactly using L^0 and L^1 operators:

$$dY_t = L^0 f(t, X_t)dt + L^1 f(t, X_t)dW_t$$

where

$$L^0 = \frac{\partial}{\partial t} + a\frac{\partial}{\partial x} + \frac{1}{2}\,b^2\,\frac{\partial^2}{\partial x^2}$$

$$L^1 = b\,\frac{\partial}{\partial x}$$

3.6 Stochastic Taylor expansions

An important application of the Itô formula is the derivation of the stochastic Taylor expansion. In a deterministic environment the Taylor expansion formula is a fundamental tool for developing numerical methods to solve ordinary differential equations. In a stochastic environment the knowledge on deterministic numerical methods may be applied on condition that a stochastic expansion formula is available. Consider an Itô SDE in integral form and apply the Itô formula to the integrand functions $a(s, X_s)$ and $b(s, X_s)$:

$$\begin{aligned}
X_t &= X_{t_0} + \textstyle\int_{t_0}^{t} a(s, X_s)ds + \int_{t_0}^{t} b(s, X_s)dW_s \\
\textstyle\int_{t_0}^{s} dY_s &= \textstyle\int_{t_0}^{s} L^0 f(\tau, X_\tau)d\tau + \int_{t_0}^{s} L^1 f(\tau, X_\tau)dW_\tau \qquad (Ito) \\
f = a \xrightarrow{Ito} a(s, X_s) &= a(t_0, X_{t_0}) + \textstyle\int_{t_0}^{s} L^0 a(\tau, X_\tau)d\tau + \int_{t_0}^{s} L^1 a(\tau, X_\tau)dW_\tau \\
f = b \xrightarrow{Ito} b(s, X_s) &= b(t_0, X_{t_0}) + \textstyle\int_{t_0}^{s} L^0 b(\tau, X_\tau)d\tau + \int_{t_0}^{s} L^1 b(\tau, X_\tau)dW_\tau \\
X_t &= X_{t_0} + a(t_0, X_{t_0}) \textstyle\int_{t_0}^{t} ds + b(t_0, X_{t_0}) \int_{t_0}^{t} dW_s + R
\end{aligned}$$

where the remainder terms are:

$$\begin{aligned} R &= \int_{t_0}^{t}\int_{t_0}^{s} L^0 a \, d\tau ds + \int_{t_0}^{t}\int_{t_0}^{s} L^1 a \, dW_\tau ds + \int_{t_0}^{t}\int_{t_0}^{s} L^0 b \, d\tau dW_s + \\ &+ \int_{t_0}^{t}\int_{t_0}^{s} L^1 b \, dW_\tau dW_s \end{aligned}$$

Applying the Itô formula again to expand the last term of the double integrals group

$$f = L^1 b \xrightarrow{Ito} L^1 b(\tau, X_\tau) = L^1 b(t_0, X_{t_0}) + \int_{t_0}^{\tau} L^0 L^1 b(u, X_u) du + \int_{t_0}^{\tau} L^1 L^1 b(u, X_u) dW_u$$

leads to the (first order) **stochastic Itô Taylor expansion**:

$$X_t = X_{t_0} + a(t_0, X_{t_0}) \int_{t_0}^{t} ds + b(t_0, X_{t_0}) \int_{t_0}^{t} dW_s + L^1 b(t_0, X_{t_0}) \int_{t_0}^{t}\int_{t_0}^{s} dW_\tau dW_s + \tilde{R}$$

The remainder terms may be expanded again (infinitely) using the Itô formula. The expansion without the remainder terms is called the **Itô Taylor expansion truncated** and it converges to the Itô process X_t both in the mean-square sense, i.e. $E[|X_t - X_n(t)|^2] \underset{t \to t_0}{\longrightarrow} 0$, and uniformly in $t \in [t_0, T]$ with probability one. These "good approximation" properties of the truncated expansion are the foundations of time discrete approximations schemes for the numerical solutions of SDEs.

3.7 Itô Taylor expansion - general form

In the general case where X_t is d-dimensional, W_t is m-dimensional and the expansion of the integrals is iterated, the **general Itô Taylor expansion** of $f(t, X_t)$ with respect to (t_0, X_{t_0}) is:

$$f(t, X_t) = \sum_{\alpha \in \mathcal{A}} f_\alpha(t_0, X_{t_0}) I_{\alpha, t_0, t} + \sum_{\alpha \in \mathcal{B}(\mathcal{A})} I_\alpha[f_\alpha(\cdot, X_\cdot)]_{t_0, t}$$

where

- α is a multi-index of length l, that is $\alpha = (j_1, j_2, ..., j_l) \in \mathcal{M}$
- $\mathcal{M}$ is the set of all multi-indices

  $\mathcal{M} = \{(j_1, j_2, \ldots, j_l) : j_i \in \{0, 1, \ldots, m\}, i \in \{0, 1, \ldots, l\} \text{ for } l = 1, 2, 3, \ldots\}$

- $\mathcal{A}$ is a hierarchical set, that is a subset of $\mathcal{M}$ with the following properties

  (i) $\mathcal{A}$ is non empty: $\mathcal{A} \neq \emptyset$

  (ii) the multi-indices in $\mathcal{A}$ are uniformly bounded in length:

  $$\sup_{\alpha \in \mathcal{A}} l(\alpha) < \infty$$

  (iii) if the multi-index α belongs to $\mathcal{A}$ then $-\alpha$ belongs to $\mathcal{A}$:

  $$-\alpha \in \mathcal{A} \quad \forall \alpha \in \mathcal{A} \backslash \{v\}$$

  where v is the multi-index of length zero and $-\alpha$ is the multi-index obtained deleting the first component of α

- $\mathcal{B}(\mathcal{A})$ is the remainder set

  $$\mathcal{B}(\mathcal{A}) = \{\alpha \in \mathcal{M} \backslash \mathcal{A} : -\alpha \in \mathcal{A}\}$$

  that is, the remainder set is built with the multi-indices not belonging to $\mathcal{A}$ ($\mathcal{M}\backslash\mathcal{A}$ is the complement of $\mathcal{A}$ with respect to $\mathcal{M}$) but with the "tail" in $\mathcal{A}$ (think at $-\alpha$ as the tail of α, i.e. the multi-index α without its first-index head)

- $f_\alpha : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ is recursively defined:

  $$f_\emptyset = f$$
  $$f_\alpha = L^{j_1} f_{-\alpha}$$

  where j_1 is the first index of α and the operator L^j is the generalized version of L^1 defined as follows (b_j = j-th column of b):

  $$L^0 = \frac{\partial}{\partial t} + a' \frac{\partial}{\partial x} + \frac{1}{2} b b' \frac{\partial^2}{\partial x^2}$$
  $$L^j = b'_j \frac{\partial}{\partial x} \qquad j \in \{1, \ldots, m\}$$

- $I_\alpha[g(\cdot)]$ is the Itô multiple stochastic integral of a function $g : [0,T] \times \Omega \to \mathbb{R}$ and it is recursively defined as follows:

  $$I_\emptyset[g]_{t_0,t} = g(t)$$
  $$I_\alpha[g]_{t_0,t} = \int_{t_0}^{t} I_{-\alpha}[g]_{t_0,s} \; dW_s^{j_l}$$

  where j_l is the last index of α, with $dW^0 = dt$ (if $j_l = 0$ the integral is deterministic otherwise it is stochastic).

The **truncated Ito-Taylor expansions** are approximated expressions of X_t derived from the general form, which is an exact relationship, by choosing $f(t,x) = x$ and discarding the remainder terms. The approximations are classified as strong or weak depending on which is the satisfied convergence criterion:

- **strong approximations** - the hierarchical set is

  $$\mathcal{A} = \Lambda_k \triangleq \{\alpha \in \mathcal{M} : l(\alpha) + n(\alpha) \leq 2k\}$$

  where $l(\alpha)$ = length of α and $n(\alpha)$ = number of zeros of α, the criterion is the mean square error convergence

  $$E(|X_t - X_t^{(k)}|^2)^{\frac{1}{2}} \leq C(t - t_0)^k \qquad \forall t \in [t_0, T]$$

  and the approximation

  $$X_t^{(k)} = \sum_{\alpha \in \Lambda_k} f_\alpha(X_{t_0}) I_{\alpha,t_0,t}$$

  tends to X_t with probability one uniformly in $[t_0, T]$; the condition $l(\alpha) + n(\alpha) \leq 2k$ puts a bound $2k$ on a modified length of α where the zeroes count twice and it highlights the different weights of dt terms (zero indexes) and dW $(=dt^{\frac{1}{2}})$ ones,

- **weak approximations** - the hierarchical set is

  $$\mathcal{A} = \Gamma_k \triangleq \{\alpha \in \mathcal{M} : l(\alpha) \leq k\}$$

  the criterion is the moment convergence

  $$|E[g(X_t)] - E[g(X_t^{(k)})]| \leq C(t - t_0)^k \qquad \forall t \in [t_0, T]$$

  where g is continuous, differentiable with polynomial growth and the approximation

  $$X_t^{(k)} = \sum_{\alpha \in \Gamma_k} f_\alpha(X_{t_0}) I_{\alpha,t_0,t}$$

  is said to be weakly convergent to X_t in $[t_0, T]$

3.8 SDE strong and weak solutions

An Itô stochastic differential equation has a **solution** if the integrals X_t exist, at least w.p.1 (with probability 1), for t belonging to an interval $[t_0, T]$. The literature on SDEs distinguishes strong solutions from weak ones:

Definition 3.1 (Strong solution) *Given the Itô stochastic differential equation*

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t \tag{3.1}$$

given the initial value X_{t_0} and the Brownian motion W_t, a ***strong solution*** *on an interval $[t_0, T]$ is a process $\{X(t), t_0 \leq t \leq T\}$ for which*

$$\begin{cases} P(X(t_0) = X_{t_0}) = 1 \\ \\ X_t = X_{t_0} + \int_{t_0}^{t} a(s, X_s)ds + \int_{t_0}^{t} b(s, X_s)dW_s \end{cases}$$

Definition 3.2 (Weak solution) *A solution of the SDE* (3.1) *is called a* ***weak solution*** *if the coefficients a and b are specified, but not the Brownian motion.*

The strong solution can be roughly thought of as a functional of the initial value and of the specified values of the Wiener process over the interval $[t_0, T]$. In the case of the weak solution, the Wiener process is not specified but is free-choice and then the solution corresponding to the chosen Wiener process may be found. Some stochastic differential equations may only have weak solutions and no strong solutions. The numerical methods for SDE solving will use Monte Carlo iterations where a set of Brownian motion values will be pseudo-randomly generated each iteration. The Wiener process sample path is always thought as specified in a simulation environment, therefore strong solutions and conditions for their existence and uniqueness are what really matter.

In the stochastic environment the uniqueness of a solution must be defined with reference to the P-almost-surely equivalence:

Definition 3.3 *If two solutions X_t and $\tilde{X}_t$ have, almost surely, the same sample paths on $[t_0, T]$, that is if*

$$P(\sup_{t_0 \leq t \leq T} |X_t - \tilde{X}_t|^2 > 0) = 0$$

the solutions are called ***pathwise unique****.*

The existence and the uniqueness of the solution of a SDE with an initial value is guaranteed by the Lipschitz condition, as it happens in ordinary differential equations, with the addition of a linear growth condition and some measurability conditions:

A1 (Measurability) a=a(t,x) and b=b(t,x) are jointly ($\mathcal{L}^2-$) measurable functions, i.e. they are enough regular (smooth) to be integrated

A2 (Lipschitz condition) there exists a constant $K > 0$ such that

$$\begin{aligned} |a(t,x) - a(t,y)| &\leq K|x-y| \\ |b(t,x) - b(t,y)| &\leq K|x-y| \end{aligned}$$

A3 (Linear growth bound) there exists a constant $K > 0$ such that

$$\begin{aligned} |a(t,x)|^2 &\leq K^2(1+|x|^2) \\ |b(t,x)|^2 &\leq K^2(1+|x|^2) \end{aligned}$$

A4 (Initial value) for the initial value X_{t_0} the following assumptions hold:

- X_{t_0} is $\mathcal{F}_{t_0}$ measurable (it is non-anticipative)
- X_{t_0} has finite second moment, i.e. $E(|X_{t_0}|^2) < \infty$

Theorem 3.4 (Existence and Uniqueness of Solutions) *Under assumptions A1-A4 the stochastic differential equation* (3.1) *has a pathwise unique strong solution* X_t *on* $[t_0, T]$ *with*

$$\sup_{t_0 \leq t \leq T} E(|X_t|^2) < \infty$$

The theorem states the existence of a process X_t with the following properties:

- X_t is the unique solution w.p. 1
- X_t is non-anticipative
- X_t has finite second moment, i.e. $E(|X_t|^2) < \infty$
- its sample paths are continuous (almost surely)

If the coefficients a and b are continuous, then a fifth property holds:

- X_t is a **diffusion process** with **drift** $a(t,x)$ and **diffusion coefficient** $b(t,x)$, i.e. it is a Markov process "without instantaneous jumps" (a more precise definition requires the explanation of some conditions on the transition densities of the Markov process)

Chapter 4

Numerical solution of SDEs

There is a short list of explicitly solvable stochastic differential equations, so finding the numeric approximation of the solution of a SDE is often the only way of analyze it. The numeric approximation of continuous variables implies their discretization and it can be done by making different choices leading to different approaches. Two approaches are worth considering:

- discretization of both time and space variables, which leads to follow the evolution of finite Markov chains by means of their transition matrices;

- discretization of the time variable and pseudo-generation of the Brownian motion, which leads to find approximate values of the sample paths at the discretization times.

The first approach is applicable only for low dimensional problems on bounded domains, because the computations require the repeated processing of the transition matrices which is very heavy in terms of computer resources consumption. This approach is not very efficient because of the processing of superfluous information contained in the transition matrices, too. The second one is the most efficient and widely applicable approach to solving Stochastic Differential Equations, because it focuses on finding good approximations of the sample paths, without useless processing [Kloeden and Platen, 1999] [Milstein and Tretyakov, 2003]. The typical steps of this approach are:

- discretization of the time variable;

- pseudo-generation of the Wiener process at the discretization times: W_t is given when looking for strong solutions and it is called the "driving" Wiener process;

- choice of a stochastic time discrete approximation scheme (e.g.: Euler, Milstein) on the basis of the desired goodness of approximation and the available computer resources;
- simulation (exploration of the sample space Ω) of approximating time discrete trajectories of the solution X_t, called sample paths of X_t.

Concepts and results from the theory of numerical solution of deterministic ordinary differential equations are a useful framework of reference for developing similar concepts and results for SDEs.

4.1 Numerical solution of deterministic ODE

The deterministic initial value problem (IVP)

$$\begin{cases} \dot{x}(t) &= a(t,x) \\ x(t_0) &= x_0 \end{cases}$$

has a unique solution $x(t)$ provided $a(t,x)$ satisfies a simple smoothness condition called Lipschitz continuity. In general the solution does not have a closed-form expression, therefore the problem requires a numerical solution, that is a sequence of values (y_n) close to the solution $x(t)$ at the times' sequence (t_n), discretization of the time interval $[t_0, T]$. The discretization usually uses N equal sized steps, hence $t_{n+1} = t_n + \delta$ with $\delta = \frac{T}{N}$.

The sequence (y_n) is built using a **numerical scheme** or **method**. Frequently used schemes are quickly presented, together with the definition of their properties: consistency, convergence and stability.

The **time discrete approximation or difference methods** use the exact relationship

$$x(t_{n+1}) = x(t_n) + \int_{t_n}^{t_{n+1}} a(t, x(\tau))d\tau$$

to derive a **one-step method**, that is a recursive rule for y_{n+1} computation which uses the last computed value y_n and has the following typical form:

$$\begin{aligned} y_{n+1} &= y_n + \phi(a(\cdot), y_n, \delta) \\ &= y_n + \Psi(t_n, y_n, \delta)\, \delta \end{aligned}$$

In the general case the recursive rule which defines y_{n+1} uses more computed values of the solution and it is called **multi-step method**:

$$y_{n+1} = y_n + \Phi(a(\cdot), y_n, y_{n-1}, \ldots, y_{n-k}, \delta)$$

Different choices of the functional Ψ lead to different **schemes** or methods:

- Euler method: $\Psi = a(t_n, y_n)$;
- *predictor - corrector methods* like the Heun method:

$$\Psi = \frac{1}{2}[a(t_n, y_n) + a(t_{n+1}, \hat{y}_{n+1})]$$

  where the estimate $\hat{y}_{n+1}$ is computed via Euler;

- Runge-Kutta methods:

$$\Psi = \sum_i \beta_i a(t_n + \alpha_i \delta, \hat{y}_{t_n+\alpha_i\delta})$$

  where the weights β_i, α_i are "carefully" chosen, the estimates $\hat{y}_{t_n+\alpha_i\delta}$ are computed using:

$$\hat{y}_{t_n+\alpha_i\delta} = y_n + c_{i,j} \sum_{j=1}^{s} a(t_n + \alpha_j\delta,\ \hat{y}_{t_n+\alpha_,\delta}) \quad i = 1 \ldots s$$

  and each RK method will be *explicit* or *implicit* depending on the coefficient matrix $c_{i,j}$ being triangular or not.

The goodness of the approximation can be measured each step or globally:

- the *local discretization error* is the approximation error for the local IVP with initial value y_n

$$l_{n+1} = y_{n+1} - x(t_{n+1};\ IVP\, x(t_n) = y_n)$$

- the *global discretization error* is the approximation error for the global original IVP

$$e_{n+1} = y_{n+1} - x(t_{n+1};\ IVP\, x(t_0) = x_0)$$

A discretization methods is **consistent** if the derivative of the solution is well approximated, that is if

$$\lim_{\delta \to 0} \frac{y_{n+1} - y_n}{\delta} = a(t_n, x_n)$$

The consistency requires the local discretization error to converge to zero with order greater than 1. For a one-step method the consistency translates into a property of the Ψ function:

$$\lim_{\delta \to 0} \frac{y_{n+1} - y_n}{\delta} = \lim_{\delta \to 0} \Psi(t_n, y_n, \delta) = a(t_n, y_n)$$

A discretization methods is **convergent** if the global discretization error converges to zero

$$\lim_{\delta \to 0} |e_{n+1}| = \lim_{\delta \to 0} |y_{n+1} - x(t_{n+1};\ IVP\, x(t_0) = x_0)| = 0$$

Consistency and Convergence are, in fact, equivalent: for a one-step method they are equivalent if the increment function Ψ satisfies a global Lipschitz condition

$$|\Psi(t', x', \delta') - \Psi(t, x, \delta)| \leq K(|t' - t| + |x' - x| + |\delta' - \delta|).$$

Besides, a one-step method with increment function Ψ satisfying a global Lipschitz condition and with local discretization error of order p+1, i.e. $|l_n| \leq \delta^{p+1}$, has global discretization error of order p, i.e. $|e_n| \leq \delta^p$. Convergence is a fundamental requirement for a numerical method to be applicable, because it assures the solution may be reached by step size shrinking.

Another important property to be investigated is the numerical stability of the method: a one-step method is **numerically stable** if for each interval $[t_0, T]$ and differential equation (see [Kloeden and Platen, 1999]) with $a(t, x)$ satisfying a Lipschitz condition there exist positive constants δ_0 and M such that

$$|y_n - \tilde{y}_n| \leq M|y_0 - \tilde{y}_0|$$

for n=1,...,N and any two solutions y_n, $\tilde{y}_n$ corresponding to any time discretizations with $\delta = \frac{T}{N} < \delta_0$.

The idea of stability is that the error (difference of two solutions) doesn't grow too much, but remains bounded with respect to its initial value if the step size is below an appropriate threshold δ_0. A theorem states that a one-step method is numerically stable if the increment function Ψ satisfies a global Lipschitz condition.

If the time horizon is infinite then the **asymptotically numerical stability** must be considered, which is defined using the same inequality and taking the limit $n \to \infty$ of the left hand side.

The threshold δ_0 depends on the particular differential equation under consideration and it defines the stability region of the numerical method for that equation. Typically the threshold is searched for a class of test equations $\dot{x} = \lambda x$ and usually the requirement on the growth is strong, i.e. $M \leq 1$, which means no growth at all. In this case the set of values of δ below the no-growth threshold δ_0 (which depends on λ) is called **region of absolute stability**. For the Euler method the region of absolute stability is characterized by $|1 + \lambda\delta| \leq 1$, which corresponds to the unit disc in the complex plane $z = \lambda\delta$ centered in -1.

4.2 Strong and Weak convergence criteria of a time discrete stochastic approximation

Consider an Itô process $\{X(t), t_0 \leq t \leq T\}$ satisfying the scalar SDE:

$$dX_t(\omega) = a(t, X_t(\omega))dt + b(t, X_t(\omega))dW_t(\omega)$$

with the initial value X_{t_0}
Consider a **time discretization** $t_0 < t_1 < \cdots < t_n < \cdots < t_{N-1} < t_N = T$ of the time interval $[0, T]$, which in the simplest equidistant case has step size $\delta = \frac{T}{N}$.
Consider an **approximation** Y_n of the solution X_t at the discretization times t_n, that is a sequence of random variables $(Y_0, \ldots, Y_n, \ldots, Y_N)$ with values "close to" $(X_t)_{(t_0,\cdots,t_n,\cdots,t_N)}$. A definition of approximation is meaningful if some characterization of the error $Y_n - X_{t_n}$ is given, because its intuitive meaning is "close to X" but it needs to be coupled to a convergence criterion with respect to the discretization refinement, i.e. the step size shrinking (consistency requires approximation quality improving under discretization refinement). The convergence criteria presented refer to the approximation error at the end of the time interval, because there is an implicit assumption that limiting the final error implies limiting the error at the generic discretization time.

Definition 4.1 *An approximation process Y* ***converges in the strong sense with order*** *$\gamma > 0$ if there exists a finite constant K and a positive constant δ_0 such that*

$$\epsilon_s(\delta) \triangleq E(|X_T - Y_N|) \leq K\delta^\gamma$$

for any time discretization with maximum step size $\delta \in (0, \delta_0)$

The order γ of the strong convergence criterion gives a measure of how much close pathwise is the approximation process Y_n to the Itô process X_t. The increments ΔW_n of the Wiener process are of root mean square order $\delta^{\frac{1}{2}}$ and not δ, therefore the order of a scheme is sometimes less in the stochastic case than in the corresponding deterministic one.

Sometimes the approximation pathwise is not required and the approximation of the mean (of a functional) of X is enough, that is only a weak approximation is required:

Definition 4.2 *An time discrete approximation Y* ***converges in the weak sense with order*** *$\beta > 0$ if for any polynomial g there exists a finite constant K and a positive constant δ_0 such that*

$$\epsilon_w(\delta) \triangleq |E(g(X_T)) - E(g(Y_N))|) \leq K\delta^\beta$$

for any time discretization with maximum step size $\delta \in (0, \delta_0)$

The order β of the weak convergence criterion gives a measure of how much close is the probability distribution of the random variable Y_N to that of X_T.

4.3 Consistency and numerical stability

The **consistency** of a numerical method in the stochastic framework is defined as in the deterministic one with some stochastic adjustments, that is evaluating the conditional expectation of the numerical derivative $\frac{1}{\delta}[Y_{n+1}^{(\delta)} - Y_n^{(\delta)}]$ of the solution and its "consistency" or fitting goodness when substituted into the differential equation. Consistency and convergence are equivalent under the assumptions of the existence and uniqueness theorem for the solution of a SDE, so no formal definition of weak and strong consistency will be given.

Both in the deterministic and in the stochastic framework, the concept of **numerical stability** of a numerical method is that the error, i.e. the difference of two numerical solutions, doesn't grow too much, but remains bounded with respect to its initial value if the step size is below an appropriate threshold δ_0. The formal definition of numerical stability in the stochastic environment refers to a specific SDE

Definition 4.3 *A time discrete approximation* $Y^{(\delta)}$ *is* ***stochastically numerically stable*** *for a given stochastic differential equation if for any finite interval* $[T_0, T]$ *there exists a positive constant* δ_0 *such that for each* $\epsilon > 0$ *and each* $\delta \in (0, \delta_0)$*.*

$$\lim_{|Y_0^{(\delta)} - \tilde{Y}_0^{(\delta)}| \to 0} \sup_{t_0 \le t \le T} P(|Y_{N_T}^{(\delta)} - \tilde{Y}_{N_T}^{(\delta)}| \ge \epsilon) = 0$$

As in the deterministic case an approximation method is simply said *numerically stable* if the property holds for the class of SDEs for which the approximation converges to the solution. The concept of asymptotically numerically stable must be defined when the time horizon T is not fixed or known:

Definition 4.4 *A time discrete approximation* $Y^{(\delta)}$ *is* ***asymptotically stochastically numerically stable*** *for a given stochastic differential equation if it is numerically stable and there exists a positive constant* δ_0 *such that for each* $\epsilon > 0$ *and each* $\delta \in (0, \delta_0)$*.*

$$\lim_{\left|Y_0^{(\delta)} - \tilde{Y}_0^{(\delta)}\right| \to 0} \lim_{T \to \infty} P(|Y_{N_T}^{(\delta)} - \tilde{Y}_{N_T}^{(\delta)}| \ge \epsilon) = 0$$

As in the deterministic case, it is possible to determine the stability threshold δ_0 for a class of test equations

$$dX_t = \lambda X_t dt + dW_t$$

and hence the **absolute stability region**, that is the set of δ for each λ or, more compactly, the set $\lambda\delta \in \mathbb{C}$ for which the approximation method is asymptotically numerically stable. The method is said to be **A-stable** (absolutely stable) if its region of absolute stability contains the left half complex plain $\Re(\lambda\delta) < 0$.

4.4 Numerical schemes for SDEs

The more common numerical schemes will be presented together with their weak and strong convergence orders. For the sake of simplicity the discretization of the time interval is always assumed as "equidistant", that is the time increment is constant ($\delta_n = \delta \ \forall n$ where $\delta_n \triangleq t_{n+1} - t_n$). Some schemes are the same used in the deterministic environment, while others are specifically made for the stochastic one.

4.4.1 The stochastic Euler scheme

The stochastic *Euler scheme* has the form:

$$Y_{n+1} = Y_n + a(t_n, Y_n)\,\delta + b(t_n, Y_n)\,\Delta W_n$$

where

$$\delta \triangleq t_{n+1} - t_n = \frac{T}{N}$$

is the length of the discretization sub-interval and

$$\Delta W_n = W_{n+1} - W_n$$

is the increment of the Wiener process on the sub-interval. In the one-dimensional case everything is scalar, while in the multi-dimensional case Y,a,W are vectors and b is a matrix.

Theorem 4.5 *The Euler scheme converges* ***strongly*** *with* ***order*** $\gamma = 0.5$*, that is*

$$E(|X_T - Y_N^{(\delta)}|) \leq K_T\,\delta^{\frac{1}{2}}\ ,$$

provided a and b satisfy the conditions:

- *existence and uniqueness conditions (Lipschitz and linear growth) for the solution*
- *spatial linear growth mixed with temporal square root growth condition*

$$|a(s,x) - a(t,x)| + |b(s,x) - b(t,x)| \leq K(1 + |x|)|s - t|^{\frac{1}{2}}$$

and under the following assumptions

- *finiteness of inial second order moment:* $E[X_0^2] < \infty$
- *the initial mean square error is of order 0.5*

$$E(|X_0 - Y_0^{(\delta)}|^2)^{\frac{1}{2}} \leq K_0\, \delta^{\frac{1}{2}}$$

It is worth noting that in the proof of the theorem a stronger result is proved, because the error is bounded not only at the final time but the bound is uniform over the whole time interval [0,T] too.

Another theorem states that the Euler scheme converges **weakly** with order $\beta = 1$ under appropriate conditions.

The stochastic Euler scheme has the same region of absolute stability $|1 + \lambda\delta| \leq 1$ as the deterministic case, because the additive noise ΔW_n vanishes when computing the difference $Y_n - \tilde{Y}_n$.

4.4.2 The Milstein scheme

The *Milstein scheme* is derived from the Itô-Taylor expansion and it is equal to the Euler scheme enriched with the term $\frac{1}{2}bb'\left\{(\Delta W)^2 - \delta\right\}$, which arises in the stochastic environment where the Itô formula holds. The Milstein scheme has the form:

$$Y_{n+1} = Y_n + a(t_n, Y_n)\,\delta + b(t_n, Y_n)\,\Delta W_n + \frac{1}{2}b\frac{\partial b}{\partial x}\left\{(\Delta W_n)^2 - \delta\right\}$$

where

$$\delta \triangleq t_{n+1} - t_n = \frac{T}{N} \qquad \Delta W_n = W_{n+1} - W_n$$

have the same meaning as the Euler ones. In the one-dimensional case everything is scalar, while in the multi-dimensional case the scheme has the same form only if the driving Wiener process is one-dimensional, hence Y,a,b are vectors and W is scalar. In the general multi-dimensional case the Taylor expansion contains multiple Itô integrals and the scheme may become unpracticable.

Theorem 4.6 *The Milstein scheme converges* ***strongly*** *with* ***order*** $\gamma = 1$, *that is*

$$E(|X_T - Y_N^{(\delta)}|) \leq K_T\ \delta\ ,$$

provided the ***coefficient functions*** $a - \frac{1}{2}bb'$, b, $\frac{1}{2}bb'$ *satisfy the Lipschitz and linear growth conditions and under the following assumptions*

- *finiteness of initial second order moment:* $E[X_0^2] < \infty$
- *the initial mean square error is of order* $\gamma = 1$

$$E(|X_0 - Y_0^{(\delta)}|^2)^{\frac{1}{2}} \leq K_0\ \delta$$

It is worth noting that there is strong convergence of $\gamma = 1$ order uniformly within the whole time interval [0,T].

Another theorem states that the Milstein scheme converges **weakly** with order $\beta = 1$ under appropriate conditions.

It is possible to prove that the Milstein scheme has the same region of absolute stability $|1 + \lambda\delta| \leq 1$ as the Euler scheme.

4.4.3 General Strong Itô-Taylor schemes

The general form of the Itô-Taylor expansion suggests approximations with respect to the strong convergence criterion. The strong Itô-Taylor schemes are derived from the strong approximations by evaluating them at the discretization times.
For $\gamma = 0.5, 1.0, ...$ the **order** γ **strong Itô-Taylor scheme** is:

$$\begin{aligned} Y_{n+1} &= \sum_{\alpha \in \mathcal{A}_\gamma^s} f_\alpha(t_n, Y_n) I_{\alpha, t_n, t_{n+1}} \\ &= Y_n + \sum_{\alpha \in \mathcal{A}_\gamma^s \setminus \{v\}} f_\alpha(t_n, Y_n) I_{\alpha, t_n, t_{n+1}} \end{aligned}$$

where the hierarchical set $\mathcal{A}_\gamma^s$ is "strong-approximation" defined

$$\mathcal{A}_\gamma^s \triangleq \{\alpha \in \mathcal{M} : l(\alpha) + n(\alpha) \leq 2\gamma\}$$

and f_α and I_α have the same definition of the Itô-Taylor expansion.

[See SDE chapter]

If $\gamma = 0.5$ then $\mathcal{A}_{0.5}^s = \{\emptyset, (0), (1)\}$: it is the Euler scheme.
If $\gamma = 1.0$ then $\mathcal{A}_{1.0}^s = \{\emptyset, (0), (1), (1,1)\}$: it is the Milstein scheme.

A theorem states that there is strong convergence of γ order uniformly within the whole time interval [0,T] under the assumptions of "regularity" of the coefficients (the requirement is Lipschitz and linear growth of f_α) plus the γ-order of the initial mean square error:

$$E(|X_0 - Y_0^{(\delta)}|^2)^{\frac{1}{2}} \leq K_0 \, \delta^\gamma$$

4.4.4 General Weak Itô-Taylor schemes

The general form of the Itô-Taylor expansion suggests approximations with respect to the weak convergence criterion. The weak Itô-Taylor schemes have the same form of the strong ones with a different multi-index set. For the weak schemes the hierarchical set $\mathcal{A}_\beta^w$ is "weak-approximation" defined

$$\mathcal{A}_\beta^w \triangleq \{\alpha \in \mathcal{M} : l(\alpha) \leq \beta\}$$

where β is the order of the scheme, β=1.0, 2.0, ... (integers only). In this work the weak convergence is not so meaningful and all the "weak" stuff will be given a light treatment.

4.4.5 Strong explicit and implicit Runge-Kutta schemes

The schemes based on Taylor approximations have one major drawback: the evaluation of the derivatives of a and b at each step. In the deterministic case the Runge-Kutta methods avoid the use of derivatives and estimate as good as possible a set of values of the solution at some points within the $[t_n, t_{n+1}]$ interval. A scheme is classified as explicit if all the quantities of its generic iteration are explicitly defined, otherwise it is classified as implicit. In the implicit case there is at least one quantity which is implicitly defined and requires the solution of an equation which may involve random variables. In the stochastic framework the adaptation of a deterministic Runge-Kutta scheme is not always possible, hence only few particular schemes will be presented. The following stochastic version of the deterministic Heun method

$$\begin{aligned} Y_{n+1} &= Y_n + \frac{1}{2}\left\{a(\bar{\Upsilon}_n) + a\right\}\delta + \frac{1}{2}\left\{b(\bar{\Upsilon}_n) + b\right\}\Delta W_n \\ \bar{\Upsilon}_n &= (\hat{Y}_{n+1}) = Y_n + a\delta + b\Delta W_n \end{aligned}$$

is generally not strongly consistent and proves how much misleading could be a heuristic stochastic adaptation: it is worth noting that it becomes strongly consistent when $b(\bar{\Upsilon}_n)$ is simply replaced with $b(Y_n)$ so that the coefficient of ΔW_n simply goes back the "ante-Heun" value b.

A **strong order 1.0 Runge-Kutta scheme** is given by

$$Y_{n+1} = Y_n + a\delta + b\Delta W_n + \frac{1}{2\sqrt{\delta}} \left\{ b(\bar{\Upsilon}_n) - b \right\} \left\{ (\Delta W_n)^2 - \delta \right\}$$

with supporting value

$$\bar{\Upsilon}_n = (\hat{Y}_{n+1}) = Y_n + a\delta + b\sqrt{\delta}$$

Explicit strong Runge-Kutta schemes of order $\gamma > 1.0$ may be found in literature, but the convergence benefits are small compared to the computational efforts required. When numerical stability is a primary issue, as in stiff stochastic differential equations, an implicit scheme must be used.

A typical **implicit strong order 1.5 Runge-Kutta scheme** for additive noise (b constant) is

$$\begin{aligned} Y_{n+1} = Y_n &+ \frac{1}{2} \left\{ a(Y_{n+1}) + a \right\} \delta + b\Delta W_n \\ &+ \frac{1}{2\sqrt{\delta}} \left\{ a(\bar{\Upsilon}_n^+) - a(\bar{\Upsilon}_n^-) \right\} \left\{ \Delta Z_n - \frac{1}{2} \Delta W_n \delta \right\} \end{aligned}$$

with supporting values

$$\bar{\Upsilon}_n^{\pm} = Y_n + a\delta \pm b\sqrt{\delta}$$

and random increments where

$$\begin{aligned} \Delta W_n &= \zeta_n^{(1)} \sqrt{\delta} \\ \Delta Z_n &= \frac{1}{2} \left(\zeta_n^{(1)} + \frac{1}{\sqrt{3}} \zeta_n^{(2)} \right) \sqrt{\delta^3} \end{aligned}$$

where $\zeta_n^{(1)}$ and $\zeta_n^{(2)}$ are independent standard normally distributed. If b is not constant the scheme is more complicated. This scheme is A-stable and it is the best compromise when absolute stability is desired: it converges well with less additional computational effort compared to a simpler explicit strong Taylor scheme.

4.4.6 Implicit non Runge-Kutta schemes

It is not usual to group all the non-RK implicit schemes together, but it happens! Some families of implicit strong Taylor schemes can be quickly presented.

The explicit Euler scheme suggests a **family of implicit Euler schemes**

$$Y_{n+1} = Y_n + \left\{ \alpha\, a(t_{n+1}, Y_{n+1}) + (1 - \alpha)\, a \right\} \delta + b\Delta W_n$$

where the parameter $\alpha \in [0, 1]$ characterizes the degree of implicitness.

The explicit Milstein scheme suggests a **family of implicit Milstein schemes** the same way, that is by the substitution of the coefficient $a(t_n, Y_n)$ with the convex combination of it and $a(t_{n+1}, Y_{n+1})$.

The implicit versions of Euler and Milstein have the same strong convergence order of the explicit schemes. The stability improves depending on α. For $\frac{1}{2} \leq \alpha \leq 1$ both are A-stable schemes, otherwise the absolute region is a disc in the complex plane.

Multi-step schemes may be used to reduce the number of evaluations of a,b and their derivatives, such as in the following **implicit two-step order 1.0 strong scheme**

$$Y_{n+1} = Y_n + \{a(t_{n+1}, Y_{n+1}) + a\}\,\delta + V_n + V_{n-1}$$

with

$$V_n = b\Delta W_n + \frac{1}{2} b \frac{\partial b}{\partial x} \left\{(\Delta W_n)^2 - \delta\right\}$$

A final remark is mandatory: an implicit scheme must be chosen only if a good explicit one exhibits numerical instability. Instability in the behaviour of the numerical solution requires a check of the stability region bounds and of the values of $\lambda\delta$ involved. The value of δ is obvious (and usually fixed), while the value of λ is not unique but there is a value for each step and they form a set. The set is made of the values λ_n such that each test equation $y_{i+1} = \lambda_n y_i$ best approximates the SDE locally, at the point (t_n, Y_n).

Chapter 5

Monte Carlo simulations

The generic numerical procedure for the strong solution of a stochastic differential equation may be seen as a black box system which transforms an input sequence, i.e. the numerical approximation W_n of a Brownian trajectory $W(t, \bar{\omega})$, into an output sequence X_n, i.e. the numerical approximation of the trajectory of the strong solution $X(t, \bar{\omega})$ of the SDE given W_t.

Considering one realization of the Brownian process, which is equivalent to one only value $\bar{\omega} \in \Omega$, is meaningless because the rich information bind to the stochastic objects may be appreciated only by exploring the sample space Ω.

A *Monte Carlo simulation* is a wide exploration (i.e. its coverage is very good) of the sample space Ω, performed through *Monte Carlo iterations*: in each iteration a realization of the stochastic objects is considered, i.e. all the stochastic objects are evaluated at a point $w_i \in \Omega$. The iterations are considered as realizations and are also called *replications*, thinking at the replication of an experiment (see [Asmussen and Glynn, 2007], [Platen and Heath, 2006], [Glasserman, 2004], [Fishman, 1996]).

The result of the Monte Carlo simulation is a collection of the outcomes (or realized values or draws) of the stochastic objects under analysis; e.g.: $\{(X_t)(\omega_i), i = 1 \dots M\}$ is a collection of SDE solution trajectories.

This collection of values are typically used to compute an estimate for the probability of an event of interest or for the expectation of a stochastic quantitative object, because these are the true objectives of the stochastic simulation. The usual measures of interest are:

- estimation of $z = \mathbb{P}(W_n > x)$ using

$$\hat{z} = \frac{1}{M} \sum 1\{W_n(\omega_i) > x\}$$

- estimation of $z = \mathbb{E}(Z)$ using

$$\hat{z} = \frac{1}{M} \sum Z(\omega_i)$$

The convergence of the estimator is proved by the central limit theorem, or law of large numbers, a well known result in probability theory. Monte Carlo estimators may be computed in more than one way. The aspects to be considered when making a comparison are:

- *computing time*, i.e. the number of heavy operations (e.g. multiplications) required to perform the computation
- *bias*, i.e. the difference between the expectation of the estimator and the true value, $E[\hat{\alpha}] - \alpha$; typically bias can be eliminated by increasing computational effort (estimators considered are asymptotically unbiased) and the order of convergence is $n^{-\frac{1}{2}}$
- *variance* of the estimator, i.e. $E[(\hat{\alpha} - E[\hat{\alpha}])^2]$
- *mean square error* of the estimator, i.e. the expectation of the square error, $E[(\hat{\alpha} - \alpha)^2]$; this is a measure that balances bias and variance, because the following relationship holds:

$$MSE(\hat{\alpha}) = Bias^2(\hat{\alpha}) + Variance(\hat{\alpha})$$

5.1 Hindrance probability as a measure of risk

One of the main objectives of this research work is to find a way to build a relationship between the number of the trains moved over a line and the hindrance risk, defined as the probability of risky events.

A risky event is assumed to be a *dangerous interaction* between two subsequent trains *running* in *free* mode, that is to say the signalling system have no control over them, because the aim is to consider only primary delays. A primary delays is a delay which do not depend on other trains, but it is unavoidable because the train journey is always different from the planned one because of perturbations - always present in a real life environment - modelled (in this work) using Lévy processes.

The interaction is defined in terms of the distance between the two trains: the interaction is classified as dangerous when the distance falls under a threshold called critical distance.

$$Risk \triangleq Prob(Hindrance) = Prob(X_i - X_{i+1} < D_{cr})$$

The threshold D_{cr} is set up on the basis of the blocking scheme but the context (here called "'approach"') in which it is used also counts:

- sections with three aspects signalling system, theoretical approach: the threshold is assumed fixed for analytical purposes

$$D_{cr} \cong 2L_{section} + L_{train}$$

  where $L_{section}$ is the length of a blocking section and L_{train} is the length of the train (the signal is green if the preceding train has cleared two sections behind);

- sections with three aspects signalling system, simulation approach: the threshold D_{cr} is dynamic because it depends on the distances of the moving trains from the fixed signals - the aspect of the signal seen by the second train is considered (that is the blocking status of the section in front of it) together with the blocking status of the section currently occupied by it;

- mobile blocking: the threshold D_{cr} is dynamic and it is a function of the train speed.

In each Monte Carlo iteration the SDEs describing train positions are solved and the risky events are counted. The probability of hindrance is estimated at the end of the iterations, out of the Monte Carlo loop.

5.2 Risk - capacity (consumption) relationship

Performing another loop, extern to the simulation, it is possible to vary the number of scheduled trains and redo the Monte Carlo simulation with the modified "capacity" hypothesis: at the end of the loop a relationship **risk-capacity** is built and also the correspondent **risk-headways**, easily obtained using $T_{window} = N_{trains} \cdot T_{headway}$ where T_{window} is the chosen reference time window (peak hour or day).

When the infrastructure is not saturated, the term *capacity consumption* is more appropriate than capacity, but if the maximum level of risk is fixed then the corresponding capacity level obtained from the relationship represents the maximum number of trains that can be moved with the given maximum level of risk, that is "the capacity"; the "capacity consumption" is the number of movable trains corresponding to a level of risk lower then the fixed bound.

Standard definitions are given with reference to a (chosen) time window, in a blocking-sections controlled deterministic (referred to train paths in the timetable) environment, and must be adapted to the simulated no-blocking stochastic environment:

- (Timetable) Capacity = maximum number of train paths that could be scheduled considering block occupation without buffers. The critical blocking sections touch each other, so a deterministic moving train is practically always in a risky situation; a proxy for this configuration in the no-blocking stochastic environment may be: very low headways corresponding to a very high level of risk. Fixing the maximum acceptable level of risk means fixing the minimum headway $T_{h_{MIN}}$ corresponding to the timetable capacity and that saturates the chosen time window.

- Capacity consumption = percentage of the time window filled by the infrastructure occupation with buffers; in the stochastic environment the time measure of an unsaturated situation is the headway $T_{headway} > T_{h_{MIN}}$; the difference $T_{headway} - T_{h_{MIN}}$ is the time the infrastructure is not occupied in a deterministic environment with respect to a time interval of size $T_{headway}$, that is

  - the infrastructure occupation is $\frac{T_{h_{MIN}}}{T_{headway}}$
  - the capacity consumption is $\frac{T_{h^*_{MIN}}}{T_{headway}}$

  where a modified minimum headway is considered

  $$T_{h^*_{MIN}} \triangleq T_{h_{cons}} = T_{h_{MIN}} + T_{buffers}$$

  Obviously, increasing the headway implies decreasing the capacity consumption $\frac{T_{h^*_{MIN}}}{T_{headway}}$ and leaving more space to eventually enlarge the buffers (up to the headway) and lower the risk.

The "stochastic" definitions may be given in terms of number of trains:

- Capacity = Maximum number of trains N_{max} that can be moved with the given maximum level of risk R_{max} in the time window T_{window}, it corresponds to the headway $T_{h_{MIN}}$, which is computed with blocking time without buffers; choosing the buffers implies lowering the maximum number of movable trains $N^*_{max} = \frac{T_{window}}{T_{h_{MIN}}}$

- Capacity consumption = fraction of the capacity $\frac{N}{N^*_{max}}$ which represents the infrastructure occupation with buffers; the corresponding level of risk $R = R(N)$ represents the residual level of risk of the headway-enlarged buffers

5.3 Distributions of Primary and Secondary delays

The main output of the Monte Carlo loop is the collection of the free running times: for each train there is a set of free running times, one for each iteration, so it is possible to estimate the probability distribution of the primary delay respect the planned running time for each train scheduled. The primary delay distributions may be employed as follows:

- test model behaviour against real life system behaviour, by delay distribution comparison;
- estimation of the secondary delay distributions as described in a recent work [Meester and Muns, 2007].

The first option is used also for parameters' model estimation, as illustrated in the chapters devoted to models. The second option will be not investigated, though it is very interesting: secondary delay distributions are important, especially when deciding buffer time sizes.

5.4 Distributions of blocking times

The collection of free running trajectories may be used to estimate the distributions of the blocking times of the timetable stairways.

In the deterministic environment the minimum headway of a line is evaluated considering the whole blocking stairways of the line, determined by the planned - deterministic and free running - train paths.

In the stochastic (simulated) environment the collection of sample free-running train paths is available and therefore it is possible to estimate the distributions of both the blocking time at the beginning of the sections and of the clearing (unblocking) time at the end of the sections.

Considering the blocking time stairways as stochastic objects forces to consider the minimum headway in a different way, it is no more an absolute quantity but a stochastic one, seeing that given a partially compressed timetable there could be hindrance because of the overlapping of the distributions of the clearing and blocking times of consequent sections.

Again, the remark is that in the stochastic environment the minimum headway and the correspondent capacity must be considered at a given level of hindrance.

It is worth noting that considering the whole time-distance graph as a stochastic object implies that not only the critical but every section has a

blocking time distribution which hinders the clearing time distribution of the preceding train, hence there is a different hindrance probability for each section. This set of different section risks suggests a reasonable rule for buffer times allocation, that is the allocation process should equalize the risks (highly hindered sections should receive bigger buffers).

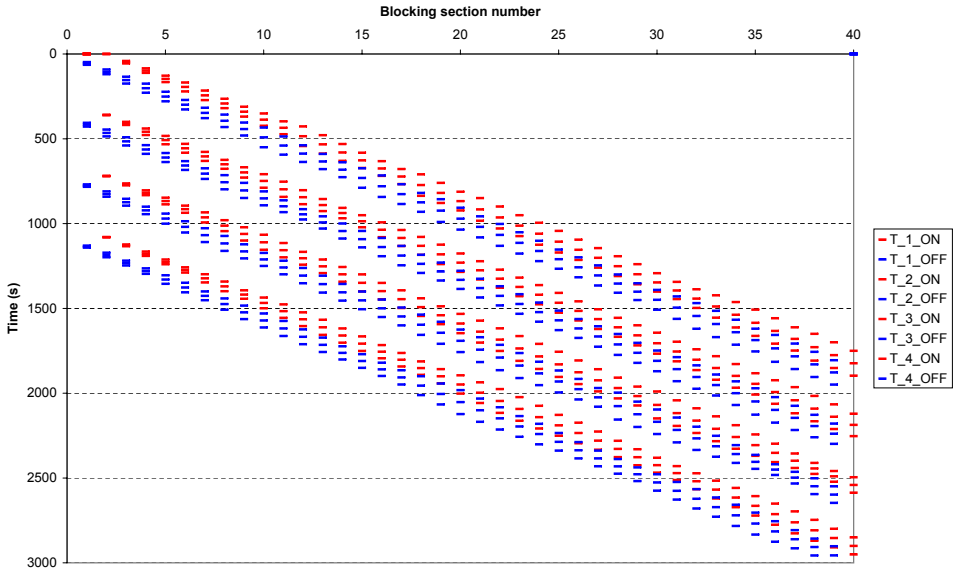


Figure 5.1: Distributions key points (mean, $\pm 3\sigma$) of Blocking/Clearing times distributions - stochastic optimal control model

5.5 Monte Carlo simulations procedure

The steps of the simulation are:

(i) Cycle on the event space Ω: in each iteration, choose $\omega_k \in \Omega, k = 1 \dots M_c$ (where usually $M_c = 1000$ iterations), that practically corresponds to draw a sample path of the Brownian motion using a pseudo-random number generator.

(ii) Compute the numerical solution of the SDE equation that describes the train journey using one of the available numerical schemes, e.g. Euler or Milstein: the result is a strong approximation of the sample path of the train; this step must be repeated for each scheduled train, to obtain the set of train paths $\{X_i(t)\}$ used in the following step.

(iii) Count the number of risky events (yellow traffic-light seen or distance under threshold, $X_i - X_{i+1} < D_{cr}$) seen by each train in its travel $N_y(i, \omega_k)$; the train sample path $X(t)$ is approximated by the numerical solution of the SDE - i.e. the sequence $X_{t_n} \cong X(t_n)$ at discretization points t_n where the steps have the same size $\delta = t_{n+1} - t_n$ - therefore every risk event seen represents a real time-continuous risky situation which is assumed to have a time length equal to δ. The ratio of the value of the counter $N_y(i, \omega_k)$ and the total number of steps $\frac{T_{sched}}{\delta}$ represents the risky fraction of the scheduled journey time (practically, this ratio may be used to estimate the probability of risk).

(iv) Compute the frequency array of the events of the type "a train sees exactly n risky events"

$$f_y(n, \omega_k) = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbf{1}_{\{j:N_y(j,\omega_k)=n\}}(i)$$

where

- $\mathbf{1}_A(x)$ is the indicator function: $\mathbf{1}_A(x) = 1$ if $x \in A$, $\mathbf{1}_A(x) = 0$ if $x \notin A$; in the formula the set A is made of indices, more precisely it contains an index if the corresponding counter is equal to n
- N_T is the total number of trains scheduled;
- practically, the sum in the formula counts the number of counters that are exactly equal to n for the group of trains scheduled, that is the number of trains that have seen exactly n risky events;

the number n of events that may be seen is very high because a "potentially hindered" train can see a risky event at each discretization point, that is $n_{max} = \frac{T_{actual}}{\delta}$.

(v) Outside the Monte Carlo loop, estimate the probability distribution that a train sees exactly n events using the "natural" estimator that computes the mean of the iteration frequencies

$$\hat{p}_y(n) = \frac{1}{M_c} \sum_{k=1}^{M_c} f_y(n, \omega_k) \qquad n = 0 \dots N_{max_E}$$

where N_{max_E} is the maximum number of risky events that can be seen by a train and M_c is the number of Monte Carlo iterations.

The probability of a risky a event may be estimated as $1 - \hat{p_y}(0)$ where $\hat{p_y}(0)$ is the estimated probability of no risky events seen or by summing up all the probability distribution except $\hat{p_y}(0)$:

$$Prob(risky\,event) = \sum_{n=1}^{N_{max_E}} \hat{p_y}(n)$$

The relationships **risk-capacity** and **risk-headways** are built by performing sensitivity analysis by repeating the simulation with N_T taking values in a range of investigation.

5.6 Pseudo-random number generators

Each Monte Carlo iteration needs the numerical approximation $W(t_n)$ of a sample path of a Brownian motion $W(t)$ as the input sequence for the SDE solver. The natural path construction of a standard Brownian motion is incremental, based on the known incremental property $\Delta W_t \sim N(0, \Delta t)$:

$$W(t_{n+1}) = W(t_n) + \sqrt{\delta}\ Z_n$$

with

- $\delta = t_{n+1} - t_n$, it is the step used in the SDE numerical solver;
- $Z_n \sim N(0,1)$, i.e. Z_n should be distributed as a standard normal, which can be easily (but carefully) pseudo-generated.

The pseudo-generation of numbers drawn from a **standard normal distribution** can be performed using one of the following methods:

- inverse transform sampling (poor efficiency);
- Box-Muller transform (good efficiency);
- Marsaglia polar or Box-Muller polar method (good efficiency, faster than Box-Muller);
- the Ziggurat algorithm (fastest, but requires precomputed tables: good for large number of draws).

Each method is based on a source of numbers drawn from a *uniform distribution*, which pseudo-generation is the real problem. Details about the four methods may be found in the cited literature on random generator numbers

[Press et al., 1992] [Gentle, 2003] [Jackel, 2003]; what really matters is the knowledge of their existence for simulation choice purposes: the MATLAB software implements both Polar and Ziggurat algorithms and the choice is left to the user.

The most commonly methods used for the pseudo-generation of numbers drawn from a **uniform distribution** are *linear congruential generators*. The integer variates are calculated starting from a seed m_0 by iteration:

$$m_{n+1} = (a\ m_c + c)\ \ mod\ M$$

and then the uniform variates $u_n \in [0, 1]$ are obtained by rescaling $u_n = \frac{m_n}{M}$. The minimal standard generator ran0 and other generators that derive from it use carefully chosen parameters a and M:

- ran0 - Park and Miller choice ($a = 7^5, c = 0, M = 2^{31} - 1$), the period is equal to M-1, i.e. $T_{seq} \approx M \approx 2 \cdot 10^9$;
- ran1 - enhancement of ran0 with shuffling: same period but slower (1.3 times) then ran0, passes more statistical tests because of the lower serial correlation;
- ran2 - coupling of two linear congruential generators with different periods (provides "perfect" random numbers), much longer period and twice slower then ran0, by l'Ecuyer;
- AWC/SWB - generators that combine congruential ones using Add-with-Carry/Subtract-with-Borrow operations, by Marsaglia and Zaman (high period $\approx 10^{43}$, available in MATLAB);

The recommendation is to use ran1 for "few" draws and ran2 when the number of draws is greater than 10^8, i.e. 5% of ran1's period.

Another generator, ran3, is worth citing, because it is twice faster than ran0 (it is good for "few" draws); it is not congruential based but uses a subtractive method suggested by Knuth.

The generator that has become increasingly popular in the last years is the *Mersenne twister*, which is available in the MATLAB software from the version 7.4 (R2007a). It produces pseudo-random numbers using the Mersenne Twister algorithm by Nishimura and Matsumoto, and is an alternative to the SWB algorithm available in the built-in function RAND in MATLAB. The period of the sequence is a Mersenne number, that is a prime number that can be written as $2^n - 1$ for some n. It creates double precision values in the closed interval $[0, 1 - 2^{-53}]$, and can generate $2^{19937} - 1$ values before repeating itself (the period is huge $\approx 10^{6000}$, i.e. infinite in practice). It has

many theoretical good properties, especially for what concerns Monte Carlo simulations and it is no slower than any other generator illustrated, so no penalty derives from using it: it is the finally recommended choice.

The C source code for implementing the pseudo-random generators described is freely available, so MATLAB computation bottlenecks may be removed by combining the flexibility of MATLAB code and the compiled code speed, that is by rewriting critical sections in C language, compiling them in DLLs (Dynamic Load Libraries) and calling them from the MATLAB code.

In this work the SDE solver and the interacting blocking system manager have been rewritten in C and compiled, with a speed gain of about 10 times with respect to pure MATLAB code.

5.7 Stochastic Models for train movement

In the stochastic environment the minimum headway and the correspondent capacity must be considered at a given level of hindrance. Obviously the stochastic model underlying the computation of train paths is a key factor: different models lead to different risks.

In the following chapters two SDE-based models are presented, together with case studies: the first model is simple but allows some theoretical considerations to validate (in the form of bounds) the simulation results; the second model is a stochastic optimal control model.

This second model describes in a more realistic way the train journey, because the driving machine produces a force following an optimal control rule which considers both the distance from the timetable and the energy consumption. A parameter, the *driving sytle*, defined as the ratio of the schedule cost and the energy cost, will be introduced to describe the different weights the two objectives may be given. Sensitivity analysis will been performed to determine the parameters' ranges for model applicability.

Chapter 6

A simple SDE model

6.1 Introduction

A simple model based on stochastic differential equations will be presented, together with the procedure based on Monte Carlo simulations used to build the risk-capacity relationship. The simulated results will be compared with analytical ones given by capacity estimation formulae. The equation (6.1) shows an example of capacity estimation formula based on a static approach, where the maximum number of available trains on a line depends of the ratio between the considered time interval (for example a day) and the time interval occupied by a single train.

$$P = \frac{V \cdot T}{\lambda + d + D + V \cdot t_m + l} \tag{6.1}$$

where P is the maximum static capacity of the railway line, V the track speed, T the reference time period (usually one day), λ is the visual distance of the signal, d is the distance between secondary signal and main one, D is the block distance, l is the length of the train and t_m is the time required for technical operations.

Another approach is that proposed by UIC in [UIC, 1983]:

$$L = \frac{T}{t_{fm} + t_r + t_{zu}} \tag{6.2}$$

where L is the dynamic capacity, T is again the reference time period, t_{fm} is the average of minimum headways, t_r is an additional time buffer to prevent delay propagation and t_{zu} is another additional time interval to ensure the desired quality over the global considered railway line. This approach tried to consider also some stochastic aspects of the train circulation within the methods suggested for the calculation of the two additional time intervals:

the t_r buffer is calculated as a percentage of t_{fm} (e.g. 33%) corresponding to a desired level of infrastructure occupation (e.g. 75%). The buffer typically considered is very large, so the equation (6.2) defines a lower bound for capacity and together with the upper bound defined by the equation (6.1) determines the range of acceptance for capacity values.

The objective of Monte Carlo simulations are the estimates of the values of risk, defined as probability of hindrance, associated with different levels of train circulation. The estimation of the probability of hindrance in a simulation environment is relatively simple, being based on a counting of risky events identified by entering a blocked zone. The estimation is more difficult if the approach is theoretical, more precisely analytical: an analytical proxy must be defined to work out something and this leads to the concept of critical distance, that is the minimum distance between two trains under which the configuration is considered risky. For analytical purposes it is better to choose it as fixed and give it a reasonable value, for example based on blocking time considerations.

Both in capacity and in timetabling research, the blocking time (and therefore the minimum headway) is a central point. The blocking time is the time interval in which a section is exclusively allocated to a train and it determines a minimum headway, as shown in 6.1 from , which suggests the choice $D_{cr} = 2L_{section} + L_{train}$.

If the train speed is known, it is quite simple to determine a critical distance D_{cr} between trains corresponding to the blocking time. Moreover other values of D_{cr} may be defined for different blocking systems [Pachl, 2002]. The minimum headway is the time interval between two consecutive trains which enable the second train to run at unrestricted speed. But as the travel time is a stochastic variable, usually in real life running time supplement and buffer times are considered so that the real headway is quite higher than the minimum one. Of course the higher is the headway, the higher is the probability of unrestricted speed for the second train and the lower is the capacity and vice versa. So a trade off exists between capacity and timetable reliability and then buffer times (and capacity) should be determined according to acceptable hindrance. Given a probability level of timetable failure (e.g. the speed of the following train could be restricted by the preceding one) the right additional running and buffer times should be determined to avoid primary delays.

In this first SDE model, the train is assumed to move following a very simple model that is the planned movement at a constant speed is perturbed by a Brownian motion.

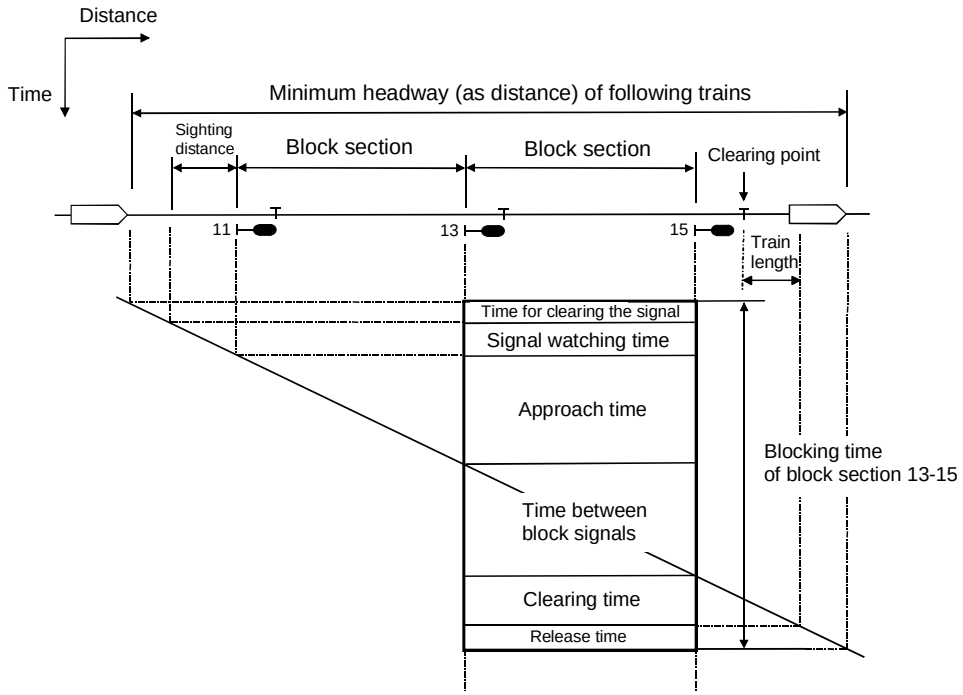


Figure 6.1: Blocking time - Pahl, 2002

6.2 The model

The simplest way to model a stochastic difference that may be very large but may shrink to zero in a finite time is the Wiener process or standard Brownian motion, which is a continuous-time Gaussian stochastic process with independent increments used in modelling real Brownian motion in physics and some random phenomena observed in finance. Let W_t be the value of the process at time t, for each positive number t, then the process is characterized by the following two conditions:

(i) if $0 < s < t$, then $W_t - W_s \sim N(0, t-s)$ where $N(\mu, \sigma^2)$ denotes the normal distribution with expected value μ and variance σ^2;

(ii) if $0 \leq s < t \leq u < v$, (i.e. the two intervals [s, t) and [u, v) do not overlap) then $W_t - W_s$ and $W_v - W_u$ are independent random variables.

The following properties hold for sample paths: continuity, nowhere differentiability and unbounded variation.

The first model considered is driven by simplicity: a very simple stochastic differential equation is used for each travel on the track, in which all stochastic processes are defined on a complete stochastic basis $(\Omega, F, \mathcal{F}, P, T)$ with

(Ω, F, P) a complete probability space and $\mathcal{F}$ a filtration, that is an increasing sequence of $sub-\sigma-algebras$ of F, $\mathcal{F} = \{F_t : F_s \subset F_t \subset F \; \forall s < t, \; s, t \in T\}$ on the positive time line $T = R_+$.
Consider an N-train evolution model that is represented by stochastic differential equations

$$dX_i(t) = V_{M\,i}(t)\, dt + \sigma_i(t)\, dW_i(t) \qquad i \in I \tag{6.3}$$

where $W_i(t)$ is a one-dimensional standard Brownian motion and $I = \{1, \ldots, N_T\}$. The trains leave the start point $X_i = 0$ at time $t_{departure}(i) = i\, T_{dep}$

$$X_i(i\, T_{dep}) = 0 \quad , \quad T_{dep} = \frac{T_{day}}{N_T}$$

because the departures of the N_T trains are scheduled at T_{dep} intervals and cover T_{day} seconds in a day.

6.2.1 Hypotheses: one-speed, one-class, independence

The following assumption are made to perform a first investigation of the model without useless complications:

$$V_{M\,i}(t) = V_M \qquad \forall (i,t) \in I \times T \qquad (one - speed) \tag{6.4}$$

$$\sigma_i(t) = \sigma \qquad \forall (i,t) \in I \times T \qquad (one - class) \tag{6.5}$$

$$W_i(t), W_j(t) \;\; independent \quad if \;\; i \neq j \qquad \forall t \in T \qquad (independence) \tag{6.6}$$

that is the parameters $V_{M\,i}(t)$ and σ_i are supposed to be time- and train-independent and the Brownian motions are supposed to be independent one from each other. The symbol μ will be used instead of V_M, to follow standards used in Brownian motion literature [Oksendal, 2000] [Shreve, 2004].

6.2.2 Brownian motion equation

The generic stochastic differential equation of the model (6.3) under the hypotheses (6.4),(6.5),(6.6) may be written in a standard form known as *Brownian motion with drift*:

$$dX_t = \mu\, dt + \sigma\, dW_t \tag{6.7}$$

where the random variable W(t) is a Brownian motion.
W(t) is called a Brownian motion if it satisfies the following properties:

(i) W(0) = 0 ;

(ii) W(t) is a continuous function of t ;

(iii) W(t) has independent, normally distributed increments $W(t)-W(s) \sim N(0, t-s)$.

The literature [Oksendal, 2000], [Shreve, 2004] shows some interesting properties. Let $0 \leq s \leq t$ be given, then

(i) W(s) and W(t)=(W(t) - W(s)) + W(s) are jointly normal and E[W(s)W(t)]=s;

(ii) the Brownian motion is a martingale, i.e. $E[W(t)|\mathcal{F}(s)] = W(s)$;

(iii) nowhere differentiability: $E[\frac{(W(t)-W(s))^2}{(t-s)^2}] = \frac{1}{t-s} \to \infty \;\; if \; t \to s$;

(iv) unbounded variation: fix $x > 0 \Rightarrow W(t)$ reaches level x with probability 1 and $E[\tau] = \infty$ where τ is the time needed to reach level x.

Some of these properties may be used to derive interesting results about risky situations.

6.3 Risk - theoretical approach

The aim is to find a relationship between track capacity (measured by N_T) and risk, defined as the probability of risky events. A risky event is assumed to be a dangerous interaction between two subsequent trains, that is their distance falls under a threshold called critical distance. The stochastic process described by equation (6.7) is known in literature as *Brownian motion with drift* and it has some properties described by closed formulas. The formulas may be derived through a rigorous path, but there is a "substitution map" between *standard Brownian* and *Brownian with drift* formulas, which is $x \mapsto x - \mu t$ and $t \mapsto \sigma^2 t$. Let $p_W(x;t)$ be the transition density of a standard Brownian motion, that is the probability that the standard Brownian motion changes value from 0 to x in time t:

$$p_W(x;t) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} \tag{6.8}$$

Let $t = i \cdot T_{dep}$ be the departure time of the i-th train, whose motion is Brownian with drift, then the probability that the i-th train arrives in x_i at time t is

$$p_{X_i}(x_i;t) = \frac{1}{\sqrt{2\pi\sigma^2(t - i\,T_{dep})}}\; e^{-\frac{(x-\mu(t-i\,T_{dep}))^2}{2\sigma^2(t-i\,T_{dep})}} \tag{6.9}$$

then X_i is normally distributed

$$X_i \sim N(\ \mu(t - i\ T_{dep})\ ,\ \sigma^2(t - i\ T_{dep})\ )$$

The distance between two trains $X_i - X_{i+1}$ is normally distributed too:

$$X_i - X_{i+1} \sim N(\ \mu\ T_{dep}\ ,\ \sigma^2(2t - (2i+1)\ T_{dep})\ )$$

A configuration $\{X_i\}$ becomes risky when the trains are too much close one to each other, that is $X_i - X_{i+1} < D_{cr}$ where D_{cr} is a critical distance (critical criteria are listed below):

$$Prob(X_i - X_{i+1} < D_{cr}\ ;\ t\ ) = \Phi(\frac{D_{cr} - \mu T_{dep}}{\sigma\sqrt{2t - (2i+1)\ T_{dep}}}) \tag{6.10}$$

where $\Phi(x)$ is the standard normal cumulative distribution function:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

There are about T_{sched}/T_{dep} trains to be considered at time t:

$$\begin{aligned} Prob(risky\ event; t) &= Prob(\cup\{X_i - X_{i+1} < D_{cr}\}\ ;\ t\ ) && (6.11)\\ &\leq \textstyle\sum_{i\ run} \Phi(\frac{D_{cr} - \mu T_{dep}}{\sigma\sqrt{2t - (2i+1)\ T_{dep}}}) && (6.12)\\ &\leq \tfrac{T_{sched}}{T_{dep}} \Phi(\tfrac{D_{cr} - \mu T_{dep}}{\sigma\sqrt{2T_{sched}}}) && (6.13) \end{aligned}$$

(the last inequality holds because $D_{cr} - \mu T_{dep} < 0$ and $\Phi(\cdot)$ is a strictly growing function)

The number of trains running in a day is $N_T = \frac{T_{day}}{T_{dep}}$ and it is a measure of capacity:

$$Prob(risky\ event) \leq \frac{T_{sched} N_T}{T_{day}} \Phi(\frac{D_{cr} - \mu \frac{T_{day}}{N_T}}{\sigma\sqrt{2T_{sched}}}) \tag{6.14}$$

The threshold D_{cr} must be set on the blocking scheme basis, as stated in the introduction:

- sections with three aspects signalling system: fixed $D_{cr} = 2L_{section} + L_{train}$
- mobile blocking (theoretical approach): $D_{cr} = f_{stop}(\mu)$
- mobile blocking (simulation approach): $D_{cr} = f_{stop}(\dot{x}_{i+1})$ or $D_{cr} = f_{stop}(\dot{x}_i, \dot{x}_{i+1})$

6.3.1 Multi-speed environment

The formula (6.14) may be used for trains with different speeds, because the distance is

$$X_i - X_{i+1} \sim N((\mu_i - \mu_{i+1})(t - iT_{dep}) + \mu_{i+1}T_{dep}\ ,\ \sigma^2(2t - (2i+1)\ T_{dep})) \quad (6.15)$$

Its mean takes values in a range with bounds which depends on μ_i dispersion

$$E[X_i - X_{i+1}] \in [\mu_{min}T_{dep} - T_{sched}\ min(\Delta\mu_i), \mu_{max}T_{dep} + T_{sched}\ max(\Delta\mu_i)] \quad (6.16)$$

where

$$min(\Delta\mu_i) = min\ \{\mu_i - \mu_{i+1}\} \quad , \quad max(\Delta\mu_i) = max\ \{\mu_i - \mu_{i+1}\}$$

The lower bound of the range (6.16) to which the mean of the distribution (6.15) belongs is a good substitute for the one-speed mean μT_{dep} of (6.14), hence in a multi-speed environment the formula (6.14) may be used with $\mu T_{dep} \mapsto \mu_{min}T_{dep} - T_{sched}\ min(\Delta\mu_i)$. The substitution is possible because the aim of (6.14) is to give an upper bound of the probability of a risky event.

6.3.2 Multi-class environment

The formula (6.14) may be used for trains of different classes, that is with different σ_i, because the distance is

$$X_i - X_{i+1} \sim N(\mu T_{dep},\ \sigma_i^2(t - iT_{dep}) + \sigma_{i+1}^2(t - (i+1)T_{dep})\ ) \quad (6.17)$$

The variance of the distribution (6.17) is upper bounded by $(\sigma_i^2 + \sigma_{i+1}^2)T_{sched}$, hence in a multi-class environment the formula (6.14) may be used with the following substitution:

$$\sigma\sqrt{2T_{sched}} \mapsto \sqrt{(\sigma_i^2 + \sigma_{i+1}^2)\ T_{sched}}$$

because the aim is to give an upper bound, $D_{cr} - \mu T_{dep} < 0$ and $\Phi(\cdot)$ is a stricly growing function.

6.3.3 Correlation between Brownian motions

Let be $\vec{W}(t)$ the N-dimensional Brownian motion with correlated components $W_i(t)$, $i \in I$ and let be the stochastic dependence between W_i enough simple to be modelled as follows

$$\exists\ \vec{W}^*,\ \exists\ \Gamma \in \mathbf{R}^{N_T \times N_T} : \vec{W}(t) = \Gamma\vec{W}^*(t)$$

where $\vec{W}^*(t)$ is a N_T-dimensional Brownian motion with uncorrelated components. The Γ matrix may be used to transform the original system of stochastic differential equations (6.3) into an uncorrelated one.

6.4 Risk - Monte Carlo simulation approach

A Montecarlo simulation has been performed to compute the probability of a risky event, given the capacity N_T and the σ of the Brownian motion. The simulation needs a set of equations derived from the stochastic differential equations (6.3), therefore a numerical scheme - like the Euler one - must be applied to solve the SDE numerically:

$$\Delta X_i(t + \Delta t) = \mu \, \Delta t + \sigma \, \Delta W_i(t) \qquad i \in I \tag{6.18}$$

where $\Delta W_i(t)$ is a pseudo-random standard Brownian generated sequence. The steps of the simulation are:

(i) Cycle on the event space Ω : choose $\omega_k \in \Omega, k = 1 \dots M_c$ (typically 1000 iterations)

(ii) Compute the numerical solution of the (6.3) using (6.18)

(iii) Count the number of risky events (yellow traffic-light or $X_i - X_{i+1} < D_{cr}$) seen by each train in its travel $N_y(i, \omega_k)$

(iv) Compute the frequency array of events "a train sees exactly n risky events"

$$f_y(n, \omega_k) = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbf{1}_{\{j:N_y(j,\omega_k)=n\}}(i)$$

where $\mathbf{1}_A(x)$ is the indicator function: $\mathbf{1}_A(x) = 1 \; if \; x \in A$, $\mathbf{1}_A(x) = 0 \; if \; x \notin A$

(v) Estimate the probability that a train sees exactly n events

$$\hat{p_y}(n) = \frac{1}{M_c} \sum_{k=1}^{M_c} f_y(n, \omega_k) \qquad n = 0 \dots N_{max_E}$$

where N_{max_E} is the maximum number of risky events that can be seen by a train.
The probability of a risky a event may be estimated as $1 - \hat{p_y}(0)$ that is

$$Prob(risky \, event) = \sum_{n=1}^{N_{max_E}} \hat{p_y}(n) \tag{6.19}$$

The relationship capacity-risk is built by repeating the simulation with N_T taking values in the range 300-500. A family of curves may be obtained

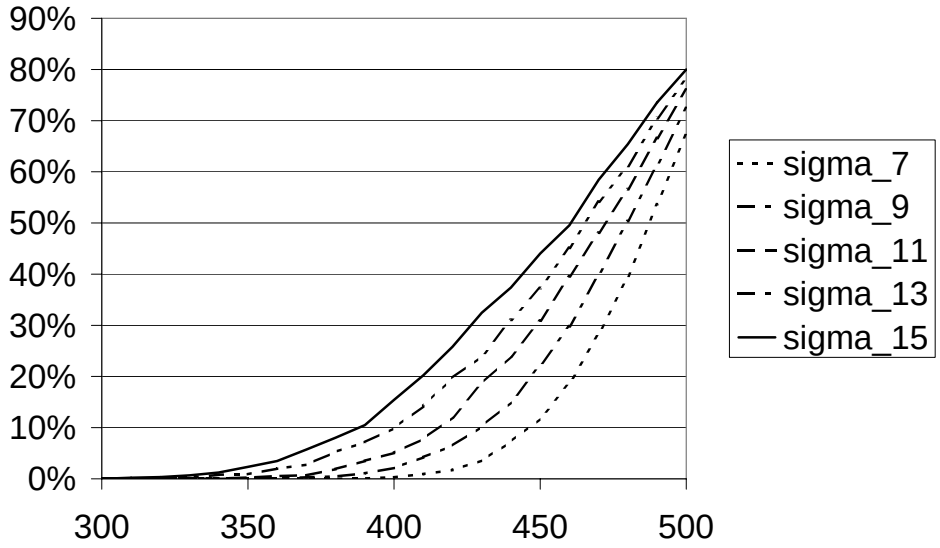


Figure 6.2: Capacity - Risk relationships from simulations

varying the value of the parameter σ of the Brownian motion. The results are shown in Fig (6.2). The comparison between theoretical probability, its bounds - from (6.11) to (6.14) - and the output (7.63) of the simulation may be seen in figure (6.3). The curves of figure (6.3) have the value of parameter $\sigma = 30.02$ which is the output of an estimation procedure applied to real observed data (see section Application). In the figure (6.3) four curves are shown:

(i) Bound P(risky) which corresponds to the equation (6.14) and $D_{cr} = 2L_{section} + L_{train}$. This is an upper bound of risk probability;

(ii) Sum phi P(Risky) which corresponds to the equation (6.11);

(iii) One train P(Risky) which corresponds to the main term of the sum in the equation (6.11);

(iv) Simul P(Risky) which shows the simulated results.

It could be noticed that the upper bound is quite close to the simulated values at least for low N_T. Moreover it is a good substitute of the sum of Φs which is a more precise way to estimate the risk probability. Finally the simulated curve could be obtained starting from One train P(Risky) curve through a suitable factor. This factor may represent a sort of running trains equivalent factor and it may be empirically found.

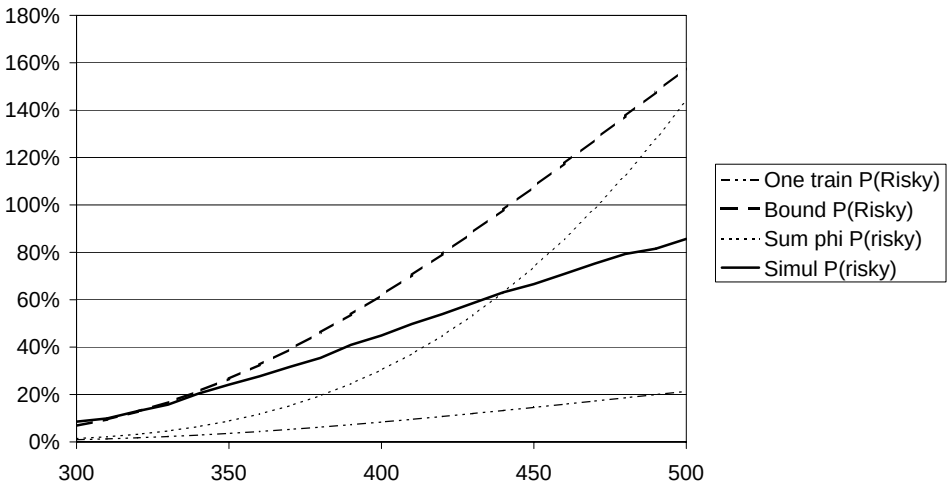


Figure 6.3: Capacity - Risk simulated vs theoretical

6.5 Model estimation procedures

The simple model (6.3) has two parameters μ and σ that must be estimated from observed data. In the real world the set of hypotheses (6.4), (6.5) and (6.6) does not hold but it is possible to find a space-limited and time-limited environment where the hypo-set is almost true and real world data like travel times may be gathered and used in an estimation procedure.

The classical estimation procedure for a brownian motion requires a set of sampled positions, but it is better to employ a procedure which requires a set of travel times:

- Parameter Estimation using the sets $\{x_i(t_k), k = 1...N_{smp}, i = 1...N_T\}$ of sampled positions ($\Rightarrow$ the observed data sets are difficult to capture in real life)

$$\widehat{\mu_i} = \frac{1}{t_{N_{smp}}} \sum_{k=1}^{N_{smp}} x_i(t_k)$$

$$\widehat{\sigma_i^2} = \frac{1}{N_{smp}} \sum_{k=1}^{N_{smp}} \Delta t_k \left[\frac{x_i(t_k) - x_i(t_{k-1})}{\Delta t_k} - \widehat{\mu_i} \right]^2$$

- Parameter Estimation using arrival time distribution: the moments of the theoretical distribution are related to μ and σ, so the observed distribution may be used to estimate them ($\Rightarrow$ the arrival time distribution is easy to capture in real life)

The arrival time of a train is a random variable τ that may be thought as a Brownian motion hitting time, that is the first time that X_t reaches a position Θ :

$$\tau(\Theta) = inf(t : X_t = \Theta) \tag{6.20}$$

The density function of the hitting time is

$$f_\tau(t, \Theta) = \frac{\Theta}{\sigma\sqrt{2\pi t^3}} e^{-\frac{(\Theta-\mu t)^2}{2\sigma^2 t}} \tag{6.21}$$

The expectation of the hitting time is

$$E\left[\tau\left(\Theta\right)\right] = \frac{\Theta}{\mu} = T_{schedule} \tag{6.22}$$

The variance is

$$Var\left[\tau\left(\Theta\right)\right] = \frac{\Theta}{\mu}\frac{\sigma^2}{\mu^2} \tag{6.23}$$

The length of the path Θ is fixed and known $\Rightarrow$ the estimates of the parameters are

$$\widehat{\mu} = \frac{\Theta}{\hat{E}\left[\tau\left(\Theta\right)\right]} = \frac{\Theta}{\hat{T}_{schedule}}$$

$$\widehat{\sigma^2} = \frac{\hat{\mu}^3}{\Theta}\widehat{Var}\left[\tau\left(\Theta\right)\right]$$

The estimates from observed data are presented in the Application section.

The arrival time may be thought as the sum of the scheduled time and the delay

$$t = T_{schedule} + t_{delay} \tag{6.24}$$

The density function of the delay time τ^* of a train is

$$f_{\tau^*}(t_{delay}, \Theta) = \frac{\Theta}{\sigma\sqrt{2\pi(T_{schedule} + t_{delay})^3}} e^{-\frac{t_{delay}^2}{2\sigma^2(T_{schedule}+t_{delay})}} \tag{6.25}$$

6.5.1 Estimation of $E[\tau]$ and $Var[\tau]$

The observed data of the random variable τ in real life (or in one Monte Carlo simulation) are

$$T_1, T_2, ..., T_{N_T} \tag{6.26}$$

The expectation may be "well" estimated with the mean of the observed data

$$\widehat{\mu}_\tau = \widehat{E[\tau]} = \frac{1}{N_T}\sum_{i\in I} T_i = \bar{T} \tag{6.27}$$

The variance of the mean estimator is a measure of its dispersion needed to construct its confidence interval and it is N times smaller than the variance of the observed data :

$$Var[\hat{\mu}] = \frac{1}{N_T} Var[T]$$

The variance of the observed data may be "well" estimated with the jackknife procedure [Efron and Tibshirani, 1993] to obtain the estimator and an estimate of its variance (the value of the estimator is meaningless without its variance):

(i) Estimate variance (as usual) using N values:

$$\widehat{V}_T = \widehat{Var[\tau]} = \frac{1}{N-1}\sum_{i\in I}(T_i - \bar{T})^2 = \frac{N}{N-1}(\ \overline{T^2} - \bar{T}^2\ ) \tag{6.28}$$

(ii) Estimate N variances $\widehat{V}_T^i$ using N-1 values:

$$M_{1,i} = \frac{1}{N-1}\sum_{j\neq i} T_j \tag{6.29}$$

$$M_{2,i} = \frac{1}{N-1}\sum_{j\neq i} T_j^2 \tag{6.30}$$

$$\widehat{V}_T^i = \frac{1}{N-2}\sum_{j\neq i}(T_j - \widehat{\mu}_T^i)^2 = \frac{N-1}{N-2}(\ M_{2,i} - M_{1,i}^2\ ) \tag{6.31}$$

(iii) Compute N pseudo-values $\theta_{V_T}^i$ (their expectation is V_T)

$$\theta_{V_T}^i = N\ \widehat{V}_T - (N-1)\ \widehat{V}_T^i \tag{6.32}$$

$$E[\theta_{V_T}^i] = N\ E[\widehat{V}_T] - (N-1)\ E[\widehat{V}_T^i] = V_T$$

(iv) The jackknife estimator of the variance is the mean of the pseudo-values:

$$\widehat{V}_T^{jack} = \frac{1}{N}\sum_{i\in I}\theta_{V_T}^i \tag{6.33}$$

$$E[\widehat{V}_T^{jack}] = \frac{1}{N}\sum_{i\in I} E[\theta_{V_T}^i] = V_T$$

(v) The variance of the pseudo-values may be estimated as usual :

$$\widehat{V_{\theta_V}} = \widehat{Var[\theta^i_{V_T}]} = \frac{1}{N-1}\sum_{i\in I}(\theta^i_{V_T} - \widehat{V}^{jack}_T)^2 \tag{6.34}$$

(vi) The variance of the jackknife estimator may be estimated as $\frac{1}{N}$ the variance estimator of the pseudo-values:

$$Var[\widehat{V}^{jack}_T] = \frac{1}{N^2}\sum_{i\in I} Var[\theta^i_{V_T}] = \frac{1}{N^2}\, N\, V_{\theta_V} = \frac{1}{N}\, V_{\theta_V} \tag{6.35}$$

6.6 Application

The proposed model has been tested in a simple case study from the Italian railway network. The line is 24,3 km long and the average length of the existing sections is 1350 m. A traditional three aspects signalling system is present. In this first simple application, homotachical services have been considered, so that only one train category is present and all trains travel at the same average speed. Starting from observed travel time data, referred to a specific train over 4 months, a travel time distribution has been pointed out.

In this simple case, first the railway capacity has been estimated according to (6.1) and to (6.2). In both cases, the headway has been easily determined too, as the time interval between departures is constant.

$$P = 527\ trains/day \qquad Headway = 2.54\ minutes$$

$$L = 184\ trains/day \qquad Headway = 12.22\ minutes$$

Of course the results are quite different because the basic hypotheses of the considered approaches are very different. In fact, in the first case, which is a sort of extreme case, almost all the trains suffer a hindrance due to the preceding ones, while in the second case, that is the other extreme, the formula considers two kinds of time buffers so that any train hindrance is avoided.

In the same scenario, the proposed model has been applied. The model parameter has been determined using the real life travel time distribution. The estimates from observed data are:

$$\hat{T}_{schedule} = 858.58\ s$$

$$\widehat{\mu} = \frac{18 \cdot 1350}{858.58} = 28.30\ m/s$$

$$\widehat{Var}\left[\tau\left(\Theta\right)\right] = 966.47\ s^2$$
$$\widehat{\sigma^2} = \frac{28.30^3}{18 \cdot 1350} 966.47 = 901.70\ m^2/s$$
$$\widehat{\sigma} = \sqrt{901.70} = 30.02\ m/s^{\frac{1}{2}}$$

Table 6.1: Risk probability vs. Capacity and Headway

Risk probability %	**Capacity** (trains/day)	**Headway** (minutes)
9	300	4,4
10	310	4,3
13	320	4,1
16	330	4,0
20	340	3,9
24	350	3,8
28	360	3,7
32	370	3,6
35	380	3,5
41	390	3,4
45	400	3,3
50	410	3,2
54	420	3,1
59	430	3,1
63	440	3,0
67	450	2,9
71	460	2,9
75	470	2,8
79	480	2,8
82	490	2,7
86	500	2,6

The table 6.1 shows the simulation results in the case studied. The table represents a relationship between the risk probability and capacity (and headway). So for a given risk level it allows to find out the corresponding capacity or headway which could be used in timetable design. Of course both capacity and headway values are included respectively into their ranges found with classical formulas (6.1) and (6.2). Finally these values depend heavily on the stochastic parameter σ, which depends mainly on the characteristics of the travel time distribution. Lower headways could be acceptable only if travel time dispersion is narrow enough. In other words, capacity and headways depend heavily on the precision of train circulation.

6.7 Concluding remarks

A new stochastic approach is proposed, which allows to link together railway capacity (train headway) and the probability that a train would suffer a speed limitation due to preceding trains. The introduction of the Brownian motion component is the simplest way to consider the lot of stochastic elements which may heavily influence the train circulation. This approach could be a first attempt to consider the existing trade-off between railway capacity and timetable stability. The relationship between capacity and probability of a risky event has been investigated from a point of view both theoretical and empirical through Monte Carlo simulations. The model has been tested in a simple case study and its results have been compared to the results of other existing approaches. For a given risk level they allow to find out the corresponding capacity or headway which could be used in timetable design.

Chapter 7

A stochastic optimal control model

The purpose is to establish a link between line capacity, that is the number of trains that can run a simple line between two stations, and the risk of a "crash", that is the probability of a risky event (e.g.: train that sees a yellow). A model is presented in which stochastic differential equations describe the movement of the train where the driving machine produces a force following an optimal control rule. The optimal control rule takes in account the gap between planned and real timetable and the amount of energy spent to control the train.

A stochastic component has been introduced to model every unknown force that can influence the deterministic motion and it describes the difference between the real impulse and the deterministic one.

Lévy processes are a family of stochastically continuous processes which is very suitable to model a stochastic difference because they are defined only through the property of independent stationary increments.

A theorem states that a Lévy process with continuous sample paths is a Brownian motion (also called Wiener process), which is a continuous-time Gaussian stochastic process with independent stationary increments used in modelling real Brownian motion in physics and some random phenomena observed in finance. It is the simplest way to model a difference that may be very large but may shrink to zero in a finite time [Shreve, 2004].

The model considered is a stochastic differential equation with a control term U(t) and it is used for each travel on the track, all stochastic processes are defined on a complete stochastic basis $(\Omega, F, \mathcal{F}, P, T)$ with (Ω, F, P) a complete probability space and $\mathcal{F}$ a filtration, that is an increasing sequence of sub-σ-algebras of F, $\mathcal{F} = \{F_t : F_s \subseteq F_t \subseteq F \ \forall s < t, \ s, t \in T\}$ on the positive time line T=R_+ An N-train evolution model represented by "controlled"

stochastic differential equations:

$$\begin{cases} dV_i(t) &= f(X_i(t), U_i(t), t) \cdot dt + \sigma \cdot dW_i(t) \qquad i = 1 \dots N \\ dX_i(t) &= V_i(t) \cdot dt \end{cases} \tag{7.1}$$

where $W_i(t)$ is a standard Brownian motion and $U_i(t)$ an optimal (by a timetable criterion) control law, is considered to perform Monte Carlo simulations to analyze the interactions between trains. The outputs are the probability of a "crash" using the number of events in which a train is in a risky situation (e.g.: it sees a yellow or a red signal) and the delay distribution. A capacity-risk (=hindrance probability) curve has been built, with capacity expressed in terms of headway. Sensitivity analysis has been performed by varying the diffusion coefficient σ of the Brownian term and/or other parameters. In real life cases the diffusion coefficient must be estimated and the model must be tested for goodness of fit. Italian railway data will be used to estimate the parameters of the model and validate its fitting within a statistical framework.

7.1 The stochastic model

A Monte Carlo simulation consists of iterations with the purpose of exploring the behavior of a stochastic model while moving in the Ω space. The stochastic model analyzed is quite simple from the mechanical point of view but it contains a control input that models the driver behavior. The driver is assumed to be optimal in the sense of the timetable observance and energy consumption, so a stochastic optimal control problem has to be solved to know which is the best control law to use.

7.1.1 The stochastic optimal control problem

The key idea is to model the system using simple mechanical equations for the train and then introducing stochastic perturbations, because in real life the external forces acting on the train have a stochastic component.
The deterministic mechanical equations are:

$$\begin{cases} F &= m \cdot \frac{dv}{dt} \\ F &= F_{machine} - F_{reaction} \\ F_{machine} &= u(t) = \text{optimal control law} \\ F_{reaction} &= \alpha + \beta \cdot v(t) + \gamma \cdot v(t)^2 \end{cases} \tag{7.2}$$

The train is driven following an input law $u(t)$, which can be thought being optimal by some criteria like minimize energy consumption or differences from timetable. The deterministic optimal control problem is formulated as a minimization problem over the control domain where the objective function J is a measure of the criteria meeting degree, typically written as a sum of cost functions. The optimal J is marked as J^* (the optimal control law is marked as u^*):

$$\begin{cases} dv &= \frac{1}{m}[u(t) - \alpha - \beta \cdot v(t) - \gamma \cdot v(t)^2] \cdot dt \\ dx &= v \cdot dt \\ J &= \int_{t_0}^{t_F} C(\vec{x}(t), u(t), t)\ dt + S(\vec{x}(t_F), t_F)) \\ J^* &= \min_{u(\cdot)} J \end{cases} \tag{7.3}$$

where $\vec{x} = (v \quad x)'$.

In real life a stochastic perturbation of the forces acting on the train must be taken in account. A stochastic component must be introduced to model every unknown force that can influence the deterministic motion and it describes the difference between the real impulse and the deterministic one. The stochastic optimal control problem has the Brownian perturbation term $\frac{\sigma}{m} dW$ and the objective function is the expectation of the sum of the cost functions:

$$\begin{cases} dv &= \frac{1}{m}[u(t) - \alpha - \beta \cdot v(t) - \gamma \cdot v(t)^2] \cdot dt + \frac{1}{m}\sigma\ dW \\ dx &= v \cdot dt \\ & \min_{u(\cdot)} E\left[\int_{t_0}^{t_F} C(\vec{x}(t), u(t), t)\ dt + S(\vec{x}(t_F), t_F))\right] \end{cases} \tag{7.4}$$

where $W(\cdot)$ is a standard Brownian motion.

The cost functions $C(\cdot)$ and $S(\cdot)$ implement the criteria followed by the "optimal driver":

- cost function $C(\cdot)$ - criteria and components:
  - minimize position differences: the timetable constraint may be tought in terms of a continuous scheduled position determined by a desired average speed: $x_{sched} = V_{mean} \cdot t_{sched}$, so it is desiderable to keep the difference $x - V_M \cdot t$ as small as possible : term $(x - V_M \cdot t)^2$
  - It is possible to use the difference between real and desired speed $v - v_M$ but is not so effective.
  - minimize energy: term u^2

$$
\begin{aligned}
C(\vec{x}, u, t) &= C_0(\vec{x}, t) + \frac{1}{2} C_2 u^2 \\
&= \frac{1}{2} C_{sched} (\frac{x}{V_M} - t)^2 + \frac{1}{2} C_2 u^2 \\
&= \frac{1}{2} C_{sched} (\frac{x_2}{V_M} - t)^2 + \frac{1}{2} C_2 u^2
\end{aligned} \tag{7.5}
$$

- final state cost function $S(\cdot)$ - criteria and components:
  In the final state there are three conditions to be met

  (i) v = 0 - the train must stop : term v^2

  (ii) $x = L_{tratta}$ - the desired final position : term $(x - L_{tratta})^2$

  (iii) $t = t_{sched}$ - scheduled timetable : term $(t - t_{sched})^2$

$$
\begin{aligned}
S(\vec{x}, t_F) &= \frac{1}{2} C_{xf} (x - L_{tratta})^2 + \frac{1}{2} C_{vf} v^2 + \frac{1}{2} C_{tf} (\frac{x}{V_M} - t_F)^2 \\
&= \frac{1}{2} C_{xf} (x_2 - L_{tratta})^2 + \frac{1}{2} C_{vf} x_1^2 + \frac{1}{2} C_{tf} (\frac{x_2}{V_M} - t_F)^2
\end{aligned} \tag{7.6}
$$

7.1.2 Hamilton Jacobi Bellman equation

The optimal control law u^* of a stochastic optimal control problem may be found solving the HJB equation (see Appendix and [Yong and Zhou, 1999]; [Oksendal, 2000]):

$$
\begin{cases}
\frac{\partial J^*}{\partial t} + H^* = 0 \\
J^*(\vec{x}, t_F) = S(\vec{x})
\end{cases} \tag{7.7}
$$

where J measures the expected cost to go from $\vec{x}_0$ at t_0 to $\vec{x}(t_f)$ at fixed time t_f :

$$
\begin{cases}
J^*(\vec{x_0}, t_0) &= \min\limits_{u(\cdot)} E \left[\int_{t_0}^{t_F} C(\vec{x}(t), u(t), t)\, dt + S(\vec{x}(t_F), t_F)) \right] \\
\vec{x_0} &= \vec{x}(t_0)
\end{cases} \tag{7.8}
$$

The functional H^* contains the terms of "state variations" that compensate the "temporal variation" J_t^* when the control law is optimal. The stochastic system to be controlled has a general form but is assumed to be linear in the control variable $u(\cdot)$ and the cost function is assumed (see (7.5)) to be

quadratic in the control variable $u(\cdot)$, so H^* has the following expression:

$$\begin{cases} H^*(x,t) &= H(x,u^*,t) \\ H(x,u,t) &= H_0 + H_1 u + \frac{1}{2} u' H_2 u \\ u^* &= \underset{u(\cdot)}{argmin}\,[H] = -H_2^{-1} H_1 \\ H_0 &= C_0 + f_0' \frac{\partial J^*}{\partial x} + \frac{1}{2} tr(G' J_{xx} G) \\ H_1 &= f_1' \frac{\partial J^*}{\partial x} \\ H_2 &= C_2 \\ G &= \begin{bmatrix} g_v & 0 \\ 0 & g_x \end{bmatrix} \end{cases} \tag{7.9}$$

The stochastic system to be controlled is linear in the control variable $u(\cdot)$ and the stochastic term doesn't contain the state variable $\vec{x}$ (nor the control, which could be a structural heavy complication):

$$\begin{cases} d\vec{x} = [f_0(\vec{x},t) + f_1(\vec{x},t) \cdot u]dt + g \cdot dW \\ \\ f_0 = \begin{bmatrix} f_{0v} \\ f_{0x} \end{bmatrix} \quad f_1 = \begin{bmatrix} f_{1v} \\ f_{1x} \end{bmatrix} \quad g = \begin{bmatrix} g_v \\ g_x \end{bmatrix} \\ \\ f_{0v} = \frac{1}{m}[-\alpha - \beta \cdot v(t) - \gamma \cdot v(t)^2] \\ f_{0x} = v \\ f_{1v} = \frac{1}{m} u(t) \\ f_{1x} = 0 \\ g_v = \frac{1}{m}\sigma \\ g_x = 0 \end{cases} \tag{7.10}$$

The solution of the HJB equation is the optimal cost-to-go function $J^*(\vec{x},t)$. The optimal control law $u^*(\vec{x},t)$ to use as input when the train is in state $\vec{x}$ at time t is:

$$u^*(\vec{x},t) = \underset{u(\cdot)}{argmin}\,[H] \tag{7.11}$$

$$= -H_2^{-1} H_1 \tag{7.12}$$

$$= -C_2^{-1} \cdot f_1(\vec{x},t)' \cdot \frac{\partial J^*}{\partial x} \tag{7.13}$$

Physical constraints like $u \in [u_{min}, u_{max}]$ are to be considered when solving the equation. The numerical solution of the general problem may be found but it needs a lot of techniques to avoid instability, non-linear behavior complications, resources greedness (see [Hanson, 2007]; [Osher and Fedkiw, 2003]). It is possible to transform the original problem in order to apply a linear approximation and solve a simpler problem.

7.1.3 Physical Model

Running Resistance

The running resistance of the train is assumed to follow the law (see professional reference for Italian trains: [Piro, 2001]), where r is in N/t and V in km/h.:

$$r = 20 + 28(\frac{V}{100})^2 \tag{7.14}$$

Using SI units:

$$\begin{aligned} r &= 2.0 \cdot 10^{-2} + 2.8 \cdot 10^{-2}(3.6 \cdot 10^{-2}V)^2 \\ &= 2.0 \cdot 10^{-2} + 3.6288 \cdot 10^{-5} \cdot V^2 \end{aligned} \tag{7.15}$$

where r is in N/kg (or m/s^2 , because r is really a decelaration) and V in m/s.

Upper bound for the control

The upper bound for the control is given by the maximum traction effort expressed in terms of mass unit. Typical values (see prof. ref. for italian trains) of traction effort and mass have been considered and for simplicity's sake the upper bound is assumed constant and equal to:

$$u_{max} = \frac{200\ kN}{800 \cdot 10^3 kg} \ = \ 0.25\ m/s^2 \tag{7.16}$$

Lower bound for the control

The lower bound for the control is given by the maximum braking effort (expressed in terms of mass unit) and it must be lower than the adhesion limit, which is a fraction of the weight g · M of the train :

$$\begin{aligned} F_{braking} &\leq f_{adhesion} \cdot g \cdot M \\ |u_{min}| &\leq f_{adhesion} \cdot g \end{aligned} \tag{7.17}$$

where g = 9,81 m/s^2 and $f_{adhesion}$ may vary a lot with train speed and track condition (its lower limit is 0.1). For simplicity's sake the lower bound is assumed constant and equal to:

$$u_{min} = -0.7 \cdot m/s^2 \tag{7.18}$$

7.1.4 Transformation to timetable coordinates

The observance of the timetable is the main objective. The typical train travel would be made of three stages: acceleration, steady state maintenance to observe the timetable, deceleration. The steady state stage is by far the most important and its analysis requires the transformation of the problem to timetable coordinates by the introduction of a new variable:

$$z \triangleq x - V_M \cdot t \tag{7.19}$$

where V_M is the desired average speed and it is the only timetable parameter of the problem. The velocity and acceleration transformations are:

$$\dot{z} = \dot{x} - V_M \qquad \ddot{z} = \ddot{x} \tag{7.20}$$

The transformation of the problem may be done by expliciting the old coordinates :

$$v_x = v_z + V_M \qquad \dot{v}_x = \dot{v}_z \qquad dv_x = dv_z \tag{7.21}$$

The transformed state keeps its structure (velocity,position), that is $\hat{x} = (v_z \;\; z)$.

Model equations in the timetable frame

Transformation steps of the main equation of (7.4) in the timetable frame:

$$\begin{aligned}
dv &= \frac{1}{m}[u(t) - \alpha - \beta \cdot v(t) - \gamma \cdot v(t)^2] \cdot dt + \frac{1}{m}\sigma \; dW \qquad (7.22)\\
dv_z &= \frac{1}{m}[u(t) - \alpha - \beta \cdot (v_z + V_M) - \gamma \cdot (v_z + V_M)^2] \cdot dt + \frac{1}{m}\sigma \; dW \\
&= \frac{1}{m}[u(t) - (\alpha + \beta V_M + \gamma V_M^2) - (\beta + 2\gamma V_M) \cdot v_z - \gamma \cdot v_z^2] \cdot dt + \frac{1}{m}\sigma \; dW \\
&= \frac{1}{m}[u(t) - \alpha_z - \beta_z \cdot v_z - \gamma_z v_z^2] \cdot dt + \frac{1}{m}\sigma \; dW
\end{aligned}$$

where

$$\begin{aligned}
\alpha_z &\triangleq \alpha + \beta V_M + \gamma V_M^2 \\
\beta_z &\triangleq \beta + 2\gamma V_M \\
\gamma_z &\triangleq \gamma
\end{aligned}$$

Assuming the value $V_M = 100$ km/h $= 10^2 \cdot (3.6)^{-1}$ m/s for the scheduled travel mean speed, the parameters' values in the timetable frame are:

$$\begin{array}{rcllcl}
\alpha_z &=& \alpha + \beta V_M + \gamma V_M^2 &=& 2.0 \cdot 10^{-2} + 0 + 2.8 \cdot 10^{-2} &=& 4.8 \cdot 10^{-2} \\
\beta_z &=& \beta + 2\gamma V_M &=& 0 + 2 \cdot 2.8 \cdot 10^{-2} \cdot 3.6 \cdot 10^{-2} &=& 2.016 \cdot 10^{-3} \\
\gamma_z &=& \gamma &=& 3.6288 \cdot 10^{-5} &&
\end{array} \tag{7.23}$$

The stochastic optimal control problem doesn't change its structure but only the parameters' values, as usual with transformations between inertial systems.

$$\begin{cases} dv_z &= \frac{1}{m}[u(t) - \alpha_z - \beta_z \cdot v_z(t) - \gamma_z \cdot v_z(t)^2] \cdot dt + \frac{1}{m}\sigma \; dW \\ dz &= v_z \cdot dt \\ & \min_{u(\cdot)} E\left[\int_{t_0}^{t_F} C(\hat{x}(t), u(t), t) \; dt + S(\hat{x}(t_F), t_F))\right] \end{cases} \tag{7.24}$$

with $C(\cdot)$ and $S(\cdot)$ defined in the following sections. The new frame has two key benefits: simple cost functions and steady state analysis through linearization.

Cost function in the timetable frame

Transformation of the cost function defined by (7.5):

$$\begin{aligned} C(\vec{x}, u, t) &= \frac{1}{2}\frac{C_{sched}}{V_M^2} \cdot z^2 + \frac{1}{2} \cdot C_2 u^2 \\ &= \frac{1}{2} C_z \cdot z^2 + \frac{1}{2} C_u \cdot u^2 \end{aligned} \tag{7.25}$$

The parameters C_z and C_u may be joined together (it is assumed they are not null) by taking their ratio, which describes the desired *driving sytle* of the train:

$$\delta \triangleq \frac{schedule \;\; cost}{energy \;\; cost} \tag{7.26}$$

The cases where the cost has one only component can be studied exploring boundary behavior:

$\delta \to 0 \quad \Longrightarrow \quad$ schedule $\ll$ energy, that is only energy consumption counts

$\delta \to \infty \quad \Longrightarrow \quad$ schedule $\gg$ energy, that is only timetable counts

The cost function will be expressed using only one parameter without lack of generality:

$$C(\hat{x}, u, t) = \frac{1}{2} \cdot z^2 + \frac{1}{2} \cdot \frac{1}{\delta} \cdot u^2 \tag{7.27}$$

Final state Cost function in the timetable frame

If the train travel is divided into three stages then the desired final state must be considered with reference to each stage. The first and the second stages have the same desired final state set (the speed v_z doesn't matter) in the timetable frame $\hat{x}(t_F) = (\cdot, 0)$. The train must stop at the end of the third stage, so its desired final transformed state is $\hat{x}(t_F) = (-V_M, 0)$.

- First and second stage (acceleration and steady state maintenance) final cost:

$$S(\hat{x}, t_F) = \frac{1}{2} C_{zf} \cdot z^2 \tag{7.28}$$

- Third stage (decelaration and stop) final cost:

$$S(\hat{x}, t_F) = \frac{1}{2} C_{zf} \cdot z^2 + \frac{1}{2} C_{vf} \cdot (v_z + V_M)^2 \tag{7.29}$$

7.2 Steady state maintenance analysis

The steady state maintenance stage of the travel is by far the most important to our stochastic analysis of capacity, because the other two stages are very short and often "boundary driven" (the input control is equal to one of the bounds). The bounds (7.16) and (7.18) may be used to estimate the time needed to accelerate from 0 km/h to 100 km/h and then to decelerate to stop: 110 s and 40 s respectively.

7.2.1 Steady state and linearization

Steady state

The steady state is characterized by the condition $d\hat{x} = 0$:

$$\begin{array}{rcccrcl} dz & = & 0 & \Rightarrow & 0 & = & v_z \\ dv_z & = & 0 & \Rightarrow & 0 & = & \frac{1}{m}[u(t) - \alpha_z] \cdot dt + \frac{1}{m}\sigma \; dW \end{array} \tag{7.30}$$

The relationships bring to the only characterizing condition of the steady state $v_z = 0$, which also means $z = constant$, but nothing can be said on the value of the z coordinate. The second equation brings to a constant input control $u(t) = \alpha_z$ if there is no stochastic perturbation ($\sigma = 0$).

Linearization

The linearization of the model (7.22) is possible and meaningful if the quadratic term $\gamma_z \cdot v_z^2$ is negligible compared to the linear one, that is when the condition $\gamma_z \cdot v_z^2 \ll \beta_z \cdot v_z$ holds. If the running resistance is assumed to have the functional form (7.15) then $\beta = 0$, $\beta_z = 2\gamma V_M$ and the condition simplifies to:

$$v_z \ll 2 \cdot V_M \tag{7.31}$$

The linearized version of the stochastic control problem (7.24) is:

$$\begin{aligned} dx &= [Ax + Bu + b]\ dt + \sigma\ dW \\ J^*(t_0, x_0) &= \min_{u(\cdot)} E\left\{\frac{1}{2}\int_{t_0}^{t_F} [x'Qx + u'Ru]dt + \frac{1}{2}x'(t_F)Gx(t_F)\right\} \end{aligned} \tag{7.32}$$

where

$$\begin{aligned} A &= \begin{bmatrix} -\beta_z & 0 \\ 1 & 0 \end{bmatrix} \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad b = \begin{bmatrix} -\alpha_z \\ 0 \end{bmatrix} \\ Q &= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad R = \begin{bmatrix} \frac{1}{\delta} \end{bmatrix} \quad G = \begin{bmatrix} 0 & 0 \\ 0 & c_{zf} \end{bmatrix} \end{aligned} \tag{7.33}$$

and for sake of simplicity the hat of $\hat{x}$ has been thrown away, that is $x = (v_z\ z)'$ and u, α_z, β_z, σ are measured with reference to the unit mass.

7.2.2 Optimal control for the linear problem

Literature results

The literature ([Yong and Zhou, 1999]; [Oksendal, 2000]) on stochastic optimal control gives the following results:

$$\begin{aligned} J^*(t, x) &= \frac{1}{2}x'P(t)x + \varphi(t)'x + \psi(t) \\ u^* &= -R^{-1}B'\frac{\partial J}{\partial x} \\ &= -R^{-1}B'[P(t)x + \varphi(t)] \end{aligned} \tag{7.34}$$

where $P(t)$ and $\varphi(t)$ are the solutions of a couple of matrix differential equations called *stochastic Riccati equations*:

$$\begin{cases} \dot{P} + PA + A'P + Q - PBR^{-1}B'P = 0 & P(t_F) = G \\ \dot{\varphi} + [A' - PBR^{-1}B']\ \varphi + Pb = 0 & \varphi(t_F) = 0 \end{cases} \tag{7.35}$$

Differential Riccati equation transformation

The solution of (7.35) is reachable in two steps: in the first one the P(t) matrix is found solving the Riccati differential equation, in the second one P(t) is known and φ is the solution of an ordinary linear differential equation.

The condition $\dot{P} = 0$ identifies the steady state solution and it leads to the *algebraic Riccati equation*:

$$PA + A'P + Q - PBR^{-1}B'P = 0 \tag{7.36}$$

The solution P_0 of the algebric equation (7.36) may be used to split the solution of the Riccati differential equation in two pieces, dividing the dynamic component from the static one:

$$P(t) = X(t) + P_0 \tag{7.37}$$

The first equation of (7.35) may be transformed

$$\begin{aligned} \dot{X} &= -Q - (X + P_0)A - A'(X + P_0) + (X + P_0)M(X + P_0) \\ &= -X(A - MP_0) - (A - MP_0)'X + XMX \\ &\quad -Q - P_0A - A'P_0 + P_0MP_0 \end{aligned} \tag{7.38}$$

into a (homogeneous) Riccati differential equation in X(t) :

$$\begin{cases} \dot{X}(t) &= -X(t)A_{cl} - A'_{cl}X(t) + X(t)MX(t) \\ X(t_f) &= G - P_0 \end{cases} \tag{7.39}$$

where :

$$M = BR^{-1}B' \qquad A_{cl} = A - MP_0 \tag{7.40}$$

The matrix $A_{cl} = A - MP_0$ is called the "closed loop" matrix because it refers to a "closed" system $\dot{x} = (A - MP_0)x$ which is the result of closing the "open loop" system $\dot{x} = Ax + Bu$ by using the current state as input through the feedback law $u = -R^{-1}B'P_0x$. The dynamic behaviour of X(t) is determined by the closed loop matrix A_{cl}.

Algebraic Riccati equation

The matrix equation (7.36) is equivalent to 2×2 scalar equations, one for each element of the matrix (2 are equal, because of the simmetry of P) :

$$\begin{aligned} elem_{1,1} &\rightarrow 2a_{11}p_{11} + 2p - \delta p_{11}^2 = 0 \\ elem_{1,2} &\rightarrow a_{11}p + p_{22} - \delta p_{11}p = 0 \\ elem_{2,2} &\rightarrow 1 - \delta p^2 = 0 \end{aligned} \tag{7.41}$$

where assumptions (7.33) were used, joined with :

$$a_{11} = -\beta_z \qquad P = \begin{bmatrix} p_{11} & p \\ p & p_{22} \end{bmatrix} \qquad M \triangleq BR^{-1}B' = \begin{bmatrix} \delta & 0 \\ 0 & 0 \end{bmatrix} \tag{7.42}$$

The solution is determined by the triple (p, p_{11}, p_{22}):

- $p^2 = \frac{1}{\delta}$ and the sign of p must be positive, see below - condition (7.45)
- p_{11} is the solution of the second order algebraic equation (7.41)
- $p_{22} = p(\delta p_{11} - a_{11})$

The "closed loop" matrix (which is what really matters to the analysis) must be considered to decide about the signs of p and p_{11}:

$$A_{cl} = A - MP_0 = \begin{bmatrix} a_{11} - \delta p_{11} & -\delta p \\ 1 & 0 \end{bmatrix} \tag{7.43}$$

Its eigenvalues are the solutions s_1 and s_2 of the characteristic equation:

$$0 = det(sI - A_{cl}) = s^2 - (a_{11} - \delta p_{11})s + \delta p \tag{7.44}$$

They must be both negative because P_0 must be the *stabilizing* solution of the algebraic Riccati equation:

$$\begin{array}{rcll} \delta p & = & s_1 \cdot s_2 > 0 & \rightarrow \quad p = \frac{1}{\sqrt{\delta}} > 0 \\ a_{11} - \delta p_{11} & = & s_1 + s_2 < 0 & \rightarrow \quad p_{11} > \frac{a_{11}}{\delta} \end{array} \tag{7.45}$$

The elements of P_0 have the following closed-form expressions, derived from equations (7.41) with conditions (7.45):

$$\left\{ \begin{array}{rcl} p_{11} & = & \frac{1}{\delta}\left\{ a_{11} + \sqrt{a_{11}^2 + 2\sqrt{\delta}} \right\} \\ p & = & \frac{1}{\sqrt{\delta}} \\ p_{22} & = & p(\delta p_{11} - a_{11}) = \frac{1}{\sqrt{\delta}}\sqrt{a_{11}^2 + 2\sqrt{\delta}} \end{array} \right. \tag{7.46}$$

Solution of the transformed Riccati differential equation

The existence of P_0 makes the transformation (7.39) possible. The transformed differential equation has a closed-form solution $X = \hat{Z} \cdot \hat{Y}^{-1}$ whose components come from the linear system which has the Hamiltonian matrix made with A_{cl}:

$$\begin{bmatrix} \dot{\hat{Y}} \\ \dot{\hat{Z}} \end{bmatrix} = \begin{bmatrix} A_{cl} & -M \\ 0 & -A_{cl}' \end{bmatrix} \begin{bmatrix} \hat{Y} \\ \hat{Z} \end{bmatrix} \qquad \begin{bmatrix} \hat{Y}(t_f) \\ \hat{Z}(t_f) \end{bmatrix} = \begin{bmatrix} I \\ X(t_f) \end{bmatrix} \tag{7.47}$$

The expression of the solution $\hat{Z}(t)$ is easily derived because its differential equation is autonomous, that is it contains only $\hat{Z}$. The expression of $\hat{Y}(t)$

is easily derived too, assuming $\hat{Z}(t)$ as known input:

$$\begin{aligned}\hat{Z}(t) &= e^{-A'_{cl}(t-t_f)} \cdot X(t_f) \qquad (7.48)\\ \hat{Y}(t) &= e^{A_{cl}(t-t_f)} \cdot I + \int_{t_f}^{t} e^{A_{cl}(t-\sigma)}(-M)\hat{Z}(\sigma)d\sigma \\ &= e^{A_{cl}(t-t_f)} - e^{A_{cl}t}\left[\int_{t_f}^{t} e^{-A_{cl}\sigma} M e^{-A'_{cl}\sigma} d\sigma\right] e^{A'_{cl}t_f} X(t_f)\end{aligned}$$

The Woodbury formula $(A+BC)^{-1} = A^{-1} - A^{-1}B(I+CA^{-1}B)^{-1}CA^{-1}$ is useful to evaluate Y^{-1} and then X(t):

$$\begin{aligned}X(t) &= \hat{Z} \cdot \hat{Y}^{-1} \qquad (7.49)\\ &= e^{-A'_{cl}(t-t_f)} X(t_f) \cdot \left[e^{-A_{cl}(t-t_f)} + e^{A_{cl}t_f}(I - \Psi e^{A_{cl}t_f})^{-1}\Psi e^{-A_{cl}(t-t_f)}\right]\end{aligned}$$

where

$$\Psi = \left[\int_{t_f}^{t} e^{-A_{cl}\sigma} M e^{-A'_{cl}\sigma} d\sigma\right] e^{A'_{cl}t_f} X(t_f)$$

If $X_f \to 0$ then $\Psi \to 0$ and $X(t) \approx e^{-A'_{cl}(t-t_f)} X(t_f) e^{-A_{cl}(t-t_f)}$, that is X(t) moves backward in time from value $X(t_f) = G - P_0$ at time t_f towards a zero matrix two times faster than the closed loop A_{cl} system [Anderson and Moore, 1989]).

Steady state stochastic Riccati solutions

The solution $P(t)$ of (7.35) reaches the steady state condition $P(t) = P_0$ very quickly, moving backwards in time and starting from the final value G, that is P(t) has always the steady state value P_0 except in a neighbourhood of the final time t_F. This behaviour suggests choosing the final cost matrix G very close to P_0 so that $P(t)$ may be assumed constant and equal to P_0.

The second equation of (7.35) is an ordinary linear differential matrix equation if $P(t) = P_0$:

$$\begin{aligned}\dot{\varphi}(t) &= -A'_{cl}\ \varphi(t) - P_0\ b \qquad (7.50)\\ \varphi(t_F) &= 0\end{aligned}$$

The steady state solution is the value φ_0 such that $\dot{\varphi}(t) = 0$ and it can be used to split $\varphi(t)$ into two components, the static and the dynamic one:

$$\begin{aligned}\varphi_0 &= -(A'_{cl})^{-1} P_0\ b \qquad (7.51)\\ \varphi(t) &= \varphi_0 + e^{-A'_{cl}(t-t_F)}(\varphi(t_F) - \varphi_0)\end{aligned}$$

The dynamic component moves backward in time from value $\varphi(t_F) - \varphi_0$, towards zero at the speed allowed by the closed loop A_{cl} system, that is $\varphi(t)$

has always the steady state value φ_0 except in a neighbourhood of the final time t_F. The expression (7.50) of φ_0 suggests explicit evaluation of A'_{cl} and its inverse:

$$A'_{cl} = \begin{bmatrix} a_{11} - \delta p_{11} & 1 \\ -\delta p & 0 \end{bmatrix} \qquad (A'_{cl})^{-1} = \frac{1}{\delta p} \begin{bmatrix} 0 & -1 \\ \delta p & a_{11} - \delta p_{11} \end{bmatrix} \tag{7.52}$$

The substitution in (7.51) leads to a very simple expression of the steady state value φ_0:

$$\begin{aligned} \varphi_0 &= -(A'_{cl})^{-1} P_0\, b = -\frac{1}{\delta p} \begin{bmatrix} 0 & -1 \\ \delta p & a_{11} - \delta p_{11} \end{bmatrix} \begin{bmatrix} p_{11} & p \\ p & p_{22} \end{bmatrix} \begin{bmatrix} b_1 \\ 0 \end{bmatrix} \\ \varphi_0 &= \frac{b_1}{\delta} \begin{bmatrix} 1 \\ -a_{11} \end{bmatrix} \end{aligned} \tag{7.53}$$

where $b_1 = -\alpha_z$, according to the definition of b in (7.33) .

The Stochatic Optimal control Law

The optimal control law (7.34) may be expressed in closed-form using the explicit expressions of $P(t)$ and $\varphi(t)$ when $G = P_0$:

$$\begin{aligned} u^* &= -R^{-1}B'P_0\hat{x} - R^{-1}B'\varphi(t) \\ &= -\delta \begin{bmatrix} 1 & 0 \end{bmatrix} P_0 \begin{bmatrix} v_z \\ z \end{bmatrix} - \delta \left\{ [1\ 0]\, \varphi_0 - [1\ 0] e^{-A'_{cl}(t-t_F)} \varphi_0 \right\} \\ &= -\delta \left\{ p_{11}\ v_z + p\ z + [1\ 0]\, \varphi_0 - [1\ 0] e^{-A'_{cl}(t-t_F)} \varphi_0 \right\} \\ &= -\left\{ (a_{11} + \sqrt{a_{11}^2 + 2\sqrt{\delta}})\ v_z + \sqrt{\delta}\ z + b_1 - b_1\ [1\ 0]\ e^{-A'_{cl}(t-t_F)} \begin{bmatrix} 1 \\ -a_{11} \end{bmatrix} \right\} \end{aligned} \tag{7.54}$$

where :

$$A'_{cl} = \begin{bmatrix} a_{11} - \delta p_{11} & 1 \\ -\delta p & 0 \end{bmatrix} = \begin{bmatrix} -\sqrt{a_{11}^2 + 2\sqrt{\delta}} & 1 \\ -\sqrt{\delta} & 0 \end{bmatrix} \tag{7.55}$$

The dynamic component of $\varphi(t)$ vanishes moving backwards in time from the final time, at a speed determined by the eigenvalues s_1,s_2 of the A'_{cl} matrix:

$$s_{1,2} = \frac{1}{2} \left\{ -\sqrt{a_{11}^2 + 2\sqrt{\delta}} \pm \sqrt{a_{11}^2 - 2\sqrt{\delta}} \right\} \tag{7.56}$$

The eigenvalues are a couple of complex conjugate values, because the condition $a_{11}^2 - 2\sqrt{\delta} < 0$ is met not only for high values of δ but for low values too,

given the very small boundary value for the positiveness of the square root argument, $\delta_{bound} = 0.25 \cdot a_{11}^4 = 4.13 \cdot 10^{-12}$. The following approximation of s_1, s_2 is very good for a wide range of δ, when the condition $\sqrt{\delta} \gg \sqrt{\delta_{bound}}$ holds (e.g.: $\delta > 10^4 \delta_{bound}$, holds whenever timetable is considered):

$$s_{1,2} \approx \frac{\sqrt{2}}{2} \left\{ -\sqrt[4]{\delta} \pm j \cdot \sqrt[4]{\delta} \right\} \tag{7.57}$$

The time constant $\sqrt[4]{\delta}$ is a measure of the distance from the final time at which the dynamic component vanishes: before the time $t = t_F - \sqrt[4]{\delta}$ the *optimal control* is in *steady state* too:

$$u_0^* = -\left\{ (a_{11} + \sqrt{a_{11}^2 + 2\sqrt{\delta}})\ v_z + \sqrt{\delta}\ z + b_1 \right\} \tag{7.58}$$

The term b_1 has a straightforward physical meaning because it is the force needed to keep the system still in the timetable coordinates. This is the law to be used in maintenance-stage simulations, so it is better to restore the original parameters using the definitions (7.33):

$$u_0^* = (\beta_z - \sqrt{\beta_z^2 + 2\sqrt{\delta}})\ v_z - \sqrt{\delta}\ z + \alpha_z \tag{7.59}$$

The form of the control law u_0^* is linear, time-invariant (obvious, the model has been linearized, the objective is linear and this is the steady state) and it is reasonable because everything has a physical meaning:

- the coefficient of v_z is negative, that is u_0^* increases if v_z decreases (accelerated driving to fight the deceleration of the system)
- the coefficient of z is negative, that is u_0^* increases if z decreases (accelerated driving to fight the delay of the system)
- the constant term is positive, that is if $v_z = 0$ and $z = 0$ there is a positive acceleration needed to keep the system still in the timetable coordinates (accelerated driving to fight the resistance of the train)

Physical bounds and linear control domain

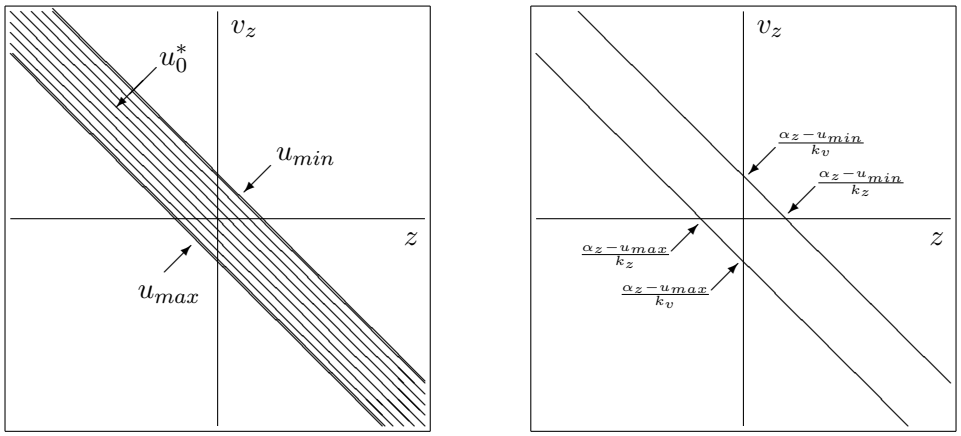
The physical bounds u_{min} and u_{max} (see (7.18) and (7.16)) must be applied to the linear optimal control law (7.59):

$$u_{linear}^* = -k_v(\delta)\ v_z - k_z(\delta)\ z + \alpha_z \tag{7.60}$$

$$u_0^* = \begin{cases} u_{linear}^* & \text{if} \quad u_{min} \le u_{linear}^* \le u_{max} \\ u_{min} & \text{if} \quad u_{linear}^* < u_{min} \\ u_{max} & \text{if} \quad u_{linear}^* > u_{max} \end{cases} \tag{7.61}$$

where the linear coefficients of v_z and z are written as $k_v(\delta)$ and $k_z(\delta)$ for sake of simplicity.
The control law u_0^* is linear within the bounds applied to u_{linear}, which implies a bounded domain of "linear working" for the state $(v_z\ z)$. If u_0^* is thought as a fixed value within the bounds, then the reverse image of u_0^* is a line in the plane z, v_z . The intersections of this line with the axes z and v_z are:

$$z = 0 \ \rightarrow \ v_z = \frac{\alpha_z - u_0^*}{k_v(\delta)} \qquad v_z = 0 \ \rightarrow \ z = \frac{\alpha_z - u_0^*}{k_z(\delta)} \qquad u_0^* \text{ fixed} \tag{7.62}$$



The projection of the linear domain region on the z-axis has a size equal to $\frac{u_{max} - u_{min}}{k_z(\delta)}$, that is proportional to $\frac{1}{\sqrt{\delta}}$: high values of δ shrink the region, while small values enlarge it. When the region shrinks enough, the state $(v_z\ z)$ is always out of it, so the value of u_0^* alternates between u_{min} and u_{max} and the control is called *Bang-Bang control*. The physical meaning is that a high value of the driving style δ is equivalent to taking in account only the timetable, so the bang-bang control is the "best" and "comfortable" way of driving when the energy consumption doesn't care.

A measure of the usefulness of the linear control approach may be given using an estimate of the time spent in the linear region compared to the total travel time. The estimation can be done by performing Monte Carlo simulations (see later), measuring the percentage of linear travel time spent in each iteration and taking the mean of the percentages.

A rule of thumb may be: "do not use the model when the couple (δ, σ) corresponds to a linear working time less than 20%". Low levels of δ like 10^{-5} or 10^{-6} are consistent with physical constraints, because δ=1 means that the objective function weights equally 1 m of delay and 1 ms^{-2} of driver acceleration (it couldn't be a real driving style). Monte Carlo iterations produce an arrival time distribution for every couple (δ, σ), which can be

Table 7.1: Control Coefficients and Domain boundary

δ	k_v	k_z	z_{min}	z_{max}	$v_{z_{min}}$	$v_{z_{max}}$
0.00001	0.08	0.00	-63.88	236.54	-2.61	9.65
0.0001	0.14	0.01	-20.20	74.80	-1.45	5.37
0.001	0.25	0.03	-6.39	23.65	-0.81	3.00
0.01	0.45	0.10	-2.02	7.48	-0.45	1.68
0.1	0.79	0.32	-0.64	2.37	-0.25	0.94
1	1.41	1.00	-0.20	0.75	-0.14	0.53
5	2.11	2.24	-0.09	0.33	-0.10	0.35
10	2.51	3.16	-0.06	0.24	-0.08	0.30

Table 7.2: Simulations - Linear Region occupation time%

$\delta \diagdown \sigma$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
10^{-5}	100	100	100	98	95	89	83	77	72	68
10^{-4}	100	100	97	92	86	77	70	63	57	52
10^{-3}	100	99	91	82	73	63	54	46	40	34
10^{-2}	100	94	81	68	57	46	37	27	21	16
10^{-1}	99	85	67	52	39	27	18	11	7	5
1	96	72	51	33	19	10	5	3	2	1
2	94	68	45	26	14	6	3	2	1	1

used to check the goodness of the model: the shape is similar to a real one, because the mean is positive (expected delay) and it has a fat tail on the right as depicted in literature [Yuan et al., 2006]. The mean and standard deviations of the arrival time distributions obtained from simulations make possible to build the functions Mean(δ, σ) and Std(δ, σ), which can be use to estimate the model parameters from real data.

7.3 Capacity - Risk relationship

The model proposed may be used to obtain a relationship between capacity and risk (see [Longo and Stok, 2007]) or between buffer time consumption and risk. The focus is on the capacity consumption due to primary delay only - the theoretical and practical capacity are widely discussed in literature ([UIC, 2004b]; [Abril et al., 2007]; [Yuan and Hansen, 2007]). The link is established via Monte Carlo simulations, under the hypothesis that the trains are free to move, subject to internal control only. The external input control (blocking scheme) is not applied but it is considered to compute the frequency of risky situations, where "risky" means hindrance between trains.

7.3.1 Monte Carlo simulation

A Monte Carlo simulation has been performed to compute the probability of a risky event, given the capacity N_T, the driving style δ and the σ of the Brownian motion. The simulation needs a set of equations derived from the stochastic differential equations (7.1) by means of "discrete substitutions" suggested by the Euler scheme for the numerical solution of an SDE, that is $dx \mapsto \Delta x$, $dW_i \mapsto \Delta W_i(t)$ where $\Delta W_i(t)$ is a pseudo-random standard Brownian generated sequence. The steps of the simulation are:

(i) Cycle on the event space Ω : choose $\omega_k \in \Omega, k = 1 \dots M_c$ (typically 1000 iterations)

(ii) Compute the numerical solution of the model using the optimal control law u_0^* of the steady state (7.60)

(iii) Count the number of risky events (blocking scheme would act) seen by each train in his travel $N_y(i, \omega_k)$

(iv) Compute the frequency array of events "a train sees exactly n risky events"

$$f_y(n, \omega_k) = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbf{1}_{\{j:N_y(j,\omega_k)=n\}}(i)$$

where $\mathbf{1}_A(x)$ is the indicator function: $\mathbf{1}_A(x) = 1 \; if \; x \in A$, $\mathbf{1}_A(x) = 0 \; if \; x \notin A$

(v) Estimate the probability that a train sees exactly n events

$$\hat{p_y}(n) = \frac{1}{M_c} \sum_{k=1}^{M_c} f_y(n, \omega_k) \qquad n = 0 \dots N_{max_E}$$

where N_{max_E} is the maximum number of risky events that can be seen by a train.
The probability of a risky a event may be estimated as $1 - \hat{p_y}(0)$ that is

$$Prob(risky\, event) = \sum_{n=1}^{N_{max_E}} \hat{p_y}(n) \tag{7.63}$$

The relationship capacity-risk is built by repeating the simulation with N_T taking values in the range 10-25 trains/hour. A family of curves may be obtained varying the values of δ and σ. The mean time delay and its standard deviation are outputs of a simulation: the curves Mean(δ, σ) and

Std (δ, σ) are built by repeating the simulation with different values of δ and σ and are used in the estimation procedure performed with real data.

The collection of sample free-running train paths obtained in a simulation may be used to estimate the distributions of both the blocking time at the beginning of the sections and of the clearing (unblocking) time at the end of the sections, as shown in figures 7.2 and 7.3.

7.3.2 Case Study

Capacity estimation procedure has been performed from the following real life Italian data:

(i) mean delay and its standard deviation : 25,3 s and 81,3 s

(ii) estimation of δ and σ from Mean(δ, σ) and Std (δ, σ) : $\delta = 3 \cdot 10^{-6}$, $\sigma = 3.19$.

More precisely, the family of curves Mean(δ, σ) and Std (δ, σ) have to be built by simulation, the real values $mean_{real}$ and std_{real} are estimated using the real life observed distribution; they identify two level curves $mean_{real}$=Mean(δ, σ) and Std $(\delta, \sigma) = std_{real}$, whose intersections are the estimates $\hat{\delta}$ and $\hat{\sigma}$.

Finally, the proposed model has been applied through Monte Carlo simulation with fixed $(\hat{\delta}, \hat{\sigma})$. The figure (7.1) shows the relationship between capacity and risk probability. Capacity is given by train headway and risk probability means the probability of train hindrance. These results may be very useful in timetable planning.

7.4 Concluding remarks

A model is presented which allows to establish a link between railway capacity and the probability of hindrance between trains. The proposed model uses stochastic differential equations to describe train movement and the driving machine produces a force following an optimal control rule, which considers the dynamic distance from the planned timetable and the energy consumption. A stochastic (Brownian motion) component has been also introduced to model every unknown force that can influence the deterministic train movement.

The optimal control law of the exact stochastic optimal control problem may be found solving the Hamilton Jacobi Bellman equation, which is numerically heavy as well as difficult to resolve because of instability and nonlinearities. An approximated stochastic optimal control problem is solved for

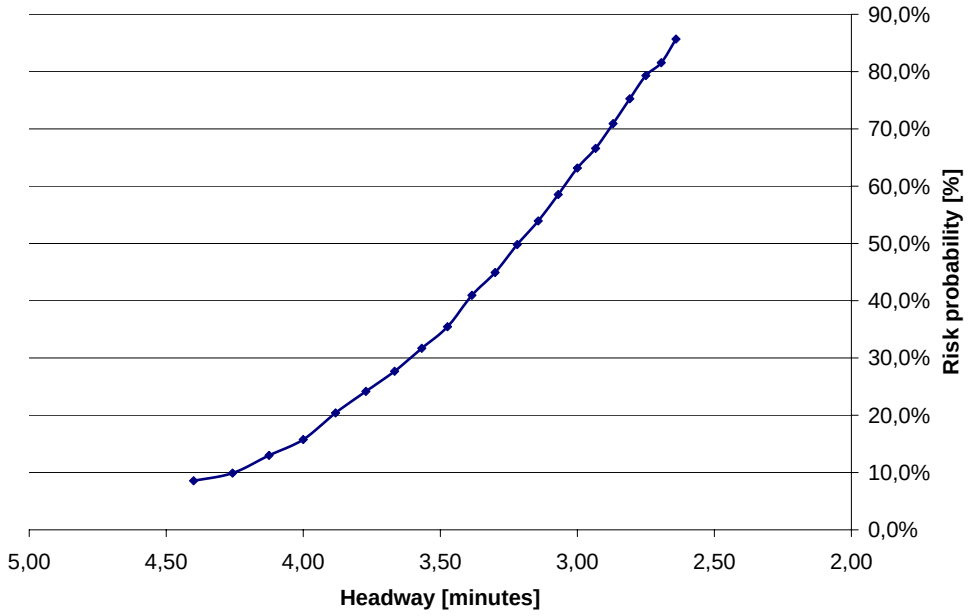


Figure 7.1: capacity vs. risk

the more relevant part of the the train travel, that is the steady state maintenance stage (initial acceleration stage and final stop stage are excluded), where the driver tries to reach the steady state determined from the planned timetable: the mechanical equation is linearized near the steady state speed and the optimal control law expression in terms of state variables is found and therefore substituted in the SDE.

A parameter, the *driving sytle*, defined as the ratio of the schedule cost and the energy cost, is introduced to describe the different weights the two objectives may be given.

Capacity-risk curves have been built and sensitivity analysis has been performed by varying the diffusion coefficient of the Brownian term and other parameters. The sensitivity analysis allows to determine the parameters' ranges for model applicability. As an example, Italian railway data have been used to estimate the model parameters and to estimate a capacity-risk probability curve, which may be very useful in timetable planning.

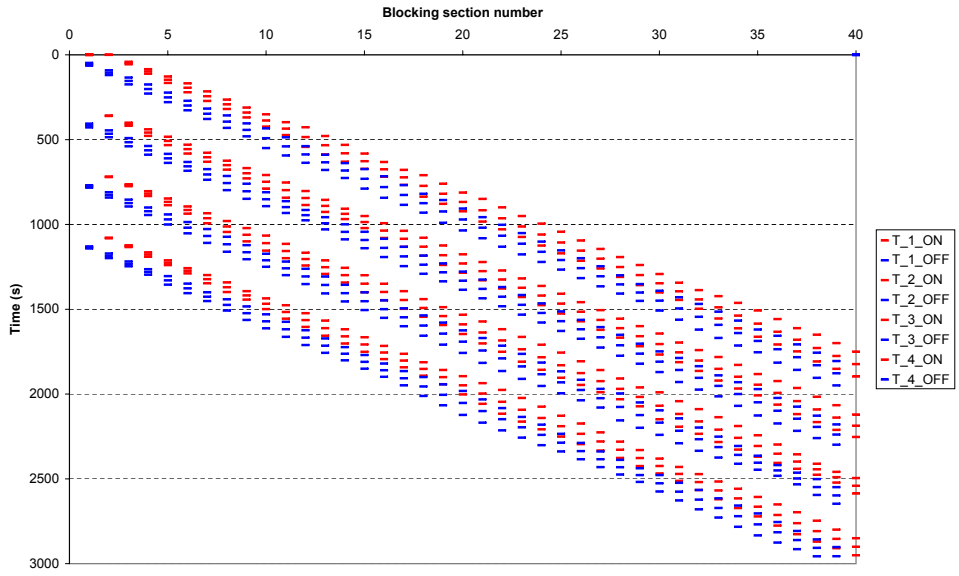


Figure 7.2: Distributions key points (mean, ± 3 stddev) of Blocking/Clearing times distributions, $\delta = 0.1$, $\sigma = 1$

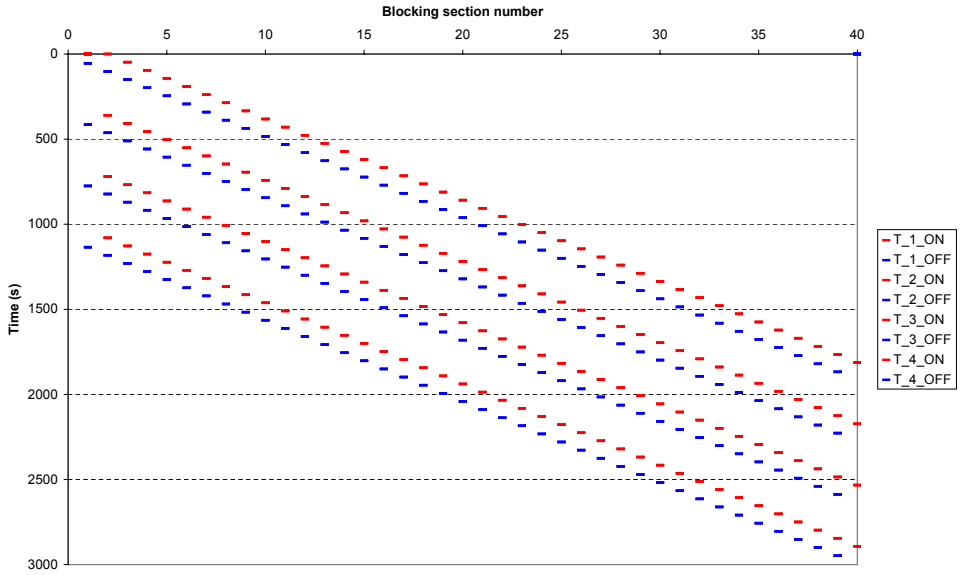


Figure 7.3: Distributions key points (mean, ± 3 stddev) of Blocking/Clearing times distributions, $\delta = 0.00001$, $\sigma = 0.1$

Chapter 8

Conclusions

The efficient utilization of railway infrastructures is a primary objective in an open-market context like the European one. The capacity consumption, that is the infrastructure occupation augmented with buffers to avoid delays, is a measure of the utilization level of a given timetable.

The standard UIC leaflet "Capacity" recommends a procedure to evaluate the infrastructure occupation, without buffers, by compressing the timetable until the blocking time stairways touch each other in the critical section. There is no recommendation about buffer times, except a well known rule of thumb about the running time supplement, which is often set to 5% of the journey time.

Buffer times choice is a trade-off between efficient utilization and stability, to avoid secondary delays caused by primary delays. Given the probability distribution of the primary delays, it is possible to estimate the distribution of secondary delays and hence the buffers.

In the recent literature review on capacity the recent approaches are discussed and the lack of an SDE approach for the primary delay is highlighted: there is only one stochastic approach which uses differential equations, but they are deterministic and combined with stochastic boarding time.

In this work the primary delays are modelled following an innovative approach, that is using a family of stochastic processes called Lévy processes. These stochastic processes are defined through a very simple and general assumption: stationary independent increments. A disturbance on which there is few knowledge may reasonably be assumed to satisfy independence and stationarity properties, because independence means the future doesn't depend on the past and stationarity means the process doesn't change its structure over the time. Another reasonable assumption is the path continuity, so the Lévy process reduces to a Brownian motion.

The train movement is therefore modelled with a differential equation

to which a Brownian motion is added, leading to a stochastic differential equation (SDE). The SDEs are formulated in the Itô sense and given the Brownian motion it is possible to solve the SDE using either the Euler or the Milstein scheme, whose convergence and stability properties are pretty good. The solution of the SDE is the strong approximation of the sample path of the train, given the numerical approximation of a Brownian motion sample path, which is pseudo-random generated.

The analysis of the stochastic phenomenon requires the Monte Carlo replications of the pseudo-random generation of the approximated Brownian motion sample path and the solution of a SDE to be computed many times. The result is a collection of paths for each train scheduled.

Monte Carlo simulation topics are illustrated together with the possible applications of the resulting collections of train paths to capacity assessment, that is to say the estimation of the probabilistic distributions of the blocking times and the estimation of the risk as probability of hindrance which corresponds to a given timetable, with the trains running in free mode (no external control, signals ignored) but counting the risky events highlighted by the signalling system.

The estimation of the risk is repeated varying the number of trains, so that a relationship is built between the risk and the number of the trains or between headways, from which a measure of capacity consumption is obtained given the risk level.

Two SDE-based models are presented, together with estimation procedures and case studies: the first model is simple but allows some theoretical considerations to validate (in the form of bounds) the simulation results; the second model is a stochastic optimal control model.

This second model describes in a more realistic way the train journey, because the driving machine produces a force following an optimal control rule which considers both the distance from the timetable and the energy consumption. A parameter, the *driving sytle*, defined as the ratio of the schedule cost and the energy cost, is introduced to describe the different weights the two objectives may be given. Sensitivity analysis has been performed to determine the parameters' ranges for model applicability.

In both cases the model parameters are estimated using real life data and then the capacity-risk relationship is build through simulations.

The more relevant aspects of this work from a transportation research point of view are:

- a new approach of computing free running times by means of stochastic differential equations has been introduced, after deep considerations about the characteristics of the stochastic perturbation;

- the collection of free running times, result of a Monte Carlo simulation, can be used to estimate the distribution of primary delays, from which it is possible to derive the distribution of the secondary delays and hence choose the buffers;
- the collection of free running trajectories can be used to estimate the distributions of the blocking times of the timetable stairways;
- a new approach of capacity assessment based on the estimation of a relationship with the probability of hindrance by performing Monte Carlo simulations in different conditions has been introduced, together with the concept of risk-coupled capacity: capacity (and capacity consumption) depends on the maximum acceptable level of risk;
- two SDE models have been introduced, together with their parameters' estimation procedures and applicability rules;
- the second model is obtained solving a stochastic optimal control problem which models real life needs such as timetable observation and low energy consumption; once its closed-form expression is found, the optimal control law can also be applied in real life, provided the continuous measures of train state, i.e. its speed and position.

Appendix A

Appendix - HJB equation

The HJB equation is a partial differential equation where the unknown is the optimal cost-to-go function $J^*(\vec{x}_0, t_0)$ - also called the value function - that is the minimun expectation of the cost function achievable when starting from $(\vec{x}_0, t_0)$. The differential relationship may be obtained thinking at what happens at the time $t_0 + \Delta t$: the residual path has $J^*(\vec{x}(t_0 + \Delta t), t_0 + \Delta t)$ as cost-to-go function, which can be written in differential terms using Taylor expansion, with second order terms in $\vec{x}$ because they aren't small, since $E[dW^2] = dt$.

$$
\begin{aligned}
J^*(\vec{x}_0, t_0) &= \min_{u(\cdot)} E\left[\int_{t_0}^{t_F} C(\vec{x}(t), u(t), t)\ dt + S(\vec{x}(t_F), t_F))\right] \\
&= \min_{u(\cdot)} E\left[\int_{t_0}^{t_0+\Delta t} C(\vec{x}(t), u(t), t)dt + \int_{t_0+\Delta t}^{t_F} C(\vec{x}(t), u(t), t)dt + S(\vec{x}(t_F), t_F))\right] \\
&= \min_{u(\cdot)} E\left[\int_{t_0}^{t_0+\Delta t} C(\vec{x}(t), u(t), t)dt\right] + \\
&\quad + \min_{u(\cdot)} E\left[\int_{t_0+\Delta t}^{t_F} C(\vec{x}(t), u(t), t)dt + S(\vec{x}(t_F), t_F))\right] \\
&= \min_{u(\cdot)} \{\ E\left[C(\vec{x}(t_\xi), u(t_\xi), t_\xi)\Delta t\right] + E\left[J^*(\vec{x}(t_0 + \Delta t), t_0 + \Delta t)\right]\ \} \qquad \text{(A.1)}
\end{aligned}
$$

The right-hand side terms may be Taylor-expanded before taking expectations. The $o(\Delta t^2)$ vanishes when taking the $\Delta t \to 0$ limit of the expansion:

$$\begin{aligned} J^*(\vec{x}_0, t_0) &= J^*(\vec{x}_0, t_0) + \min_{u(\cdot)} \{ C(\vec{x}_0, u(t_0), t_0)\, \Delta t + \\ &\quad + \left[J_x^{*\top} f_0(\vec{x}_0, t_0) + J_x^{*\top} f_1(\vec{x}_0, t_0) \cdot u(t_0) + \frac{1}{2} tr[G' J_{xx}^* G] + J_t^* \right] \Delta t \; + \\ &\quad + o(\Delta t^2) \} \\ 0 &= \min_{u(\cdot)} \{ C(\vec{x}_0, u(t_0), t_0) + \\ &\quad + J_x^{*\top} f_0(\vec{x}_0, t_0) + J_x^{*\top} f_1(\vec{x}_0, t_0) \cdot u(t_0) + \frac{1}{2} tr[G' J_{xx}^* G] + J_t^* \; \} \end{aligned} \tag{A.2}$$

The equation (A.2) is the Hamilton-Jacobj-Bellman equation for the stochastic control problem. The HJB is a partial differential equation for the unknown J^* with a $\min(\cdot)$ operator, which may be thrown away if it is possible to get the minimizing optimal control u^* in closed form. The HJB solution procedure may clarify the statement:

(i) the minimun $u^*(t_0)$ may be found solving the one-variable minimum problem, with all the variables - $\vec{x}_0, t_0, J^*(\vec{x}_0, t_0)$ except $u(t_0)$ thought fixed;

(ii) the optimal control $u^*(t_0)$, which is function of $\vec{x}_0, t_0, J^*(\vec{x}_0, t_0)$, is put back in the HJB equation so that the $\min(.)$ operator can be thrown away and the pde solved;

(iii) the solution $J^*(\vec{x}, t)$ of the pde is used to get the explicit expression of the optimal control u^* for every couple $(\vec{x}, t)$.

The procedure is applicable when the minimizing u^* can be expressed explicitly as a function of $\vec{x}_0, t_0, J^*$ and it happens when the cost function $C(\vec{x}, u, t)$ is assumed to be quadratic in $u(\cdot)$, so that the HJB becomes:

$$\begin{aligned} 0 &= \min_{u(\cdot)} \{ C_0(\vec{x}_0, t_0) + \frac{1}{2} C_2(\vec{x}_0, t_0) \cdot u^2(t_0) + J_x^{*\top} f_0(\vec{x}_0, t_0) \\ &\quad + J_x^{*\top} f_1(\vec{x}_0, t_0) \cdot u(t_0) + \frac{1}{2} tr[G' J_{xx}^* G] \; \} \; + \; J_t^* \end{aligned} \tag{A.3}$$

The functional to be minimized may be written as a quadratic expression of the variable $u(t_0)$:

$$\begin{aligned} 0 &= \min_{u(\cdot)} \{ H_0 + H_1 \cdot u(t_0) + \frac{1}{2} H_2 \cdot u^2(t_0) \} + J_t^* \qquad \text{(A.4)} \\ H_0 &= C_0(\vec{x}_0, t_0) + J_x^*(\vec{x}_0, t_0)' f_0(\vec{x}_0, t_0) + \frac{1}{2} tr[G' J_{xx}^* G] \\ H_1 &= J_x^*(\vec{x}_0, t_0)' \cdot f_1(\vec{x}_0, t_0) \\ H_2 &= C_2(\vec{x}_0, t_0) \end{aligned}$$

The optimal control may be found by minimizing this very simple quadratic expression in u:

$$\begin{aligned} u^*(t_0) &= \underset{u(\cdot)}{argmin} \{ H_0 + H_1 \cdot u(t_0) + \frac{1}{2} H_2 \cdot u^2(t_0) \} \\ &= -H_2^{-1} H_1 \\ &= -C_2(\vec{x}_0, t_0)^{-1} \cdot J_x^*(\vec{x}_0, t_0)' \cdot f_1(\vec{x}_0, t_0) \end{aligned}$$

The min-pde equation (A.4) may be transformed in a pde equation:

$$\begin{aligned} 0 &= H_0 + H_1 \cdot (-H_2^{-1} H_1) + \frac{1}{2} H_2 \cdot H_2^{-2} H_1^2 + J_t^* \\ &= H_0 - \frac{1}{2} H_2^{-1} H_1^2 + J_t^* \\ &= C_0 + J_x^{*\top} f_0 + \frac{1}{2} tr[G' J_{xx}^* G] - \frac{1}{2} C_2^{-1} \cdot (J_x^{*\top} \cdot f_1)^2 + J_t^* \qquad \text{(A.5)} \end{aligned}$$

Taylor stuff The terms of the Taylor expansion and their expectations are:

$$\begin{aligned} J^*(\vec{x}_0 + \Delta\vec{x}, t_0 + \Delta t) &\approxeq J^*(\vec{x}_0, t_0) + J_x^{*\top}(\vec{x}_0, t_0)\Delta\vec{x} + \frac{1}{2}\Delta\vec{x}' J_{xx}^*(\vec{x}_0, t_0)\Delta\vec{x} + J_t^*(\vec{x}_0, t_0)\Delta t \\ d\vec{x} &= [f_0(\vec{x}, t) + f_1(\vec{x}, t) \cdot u]dt + G \cdot d\vec{W} \\ d\vec{x}' J_{xx}^*(\vec{x}_0, t_0) d\vec{x} &= d\vec{W}' \cdot G' J_{xx}^* G \cdot d\vec{W} + o(dt^2) \\ E\left[d\vec{x}' J_{xx}^*(\vec{x}_0, t_0) d\vec{x}\right] &= tr[G' J_{xx}^* G] \cdot E[dW^2] + o(dt^2) \\ &= tr[G' J_{xx}^* G] \cdot dt + o(dt^2) \\ J_x^{*\top} d\vec{x} &= J_x^{*\top}[f_0 + f_1 \cdot u]dt + J_x^{*\top} G \cdot d\vec{W} \\ E\left[J_x^{*\top} d\vec{x}\right] &= J_x^{*\top}[f_0 + f_1 \cdot u]dt + J_x^{*\top} G \cdot E[d\vec{W}] \\ E\left[J^*(\vec{x}_0 + \Delta\vec{x}, t_0 + \Delta t)\right] &\approxeq J^*(\vec{x}_0, t_0) + \\ &\quad + \left[J_x^{*\top} f_0(\vec{x}_0, t_0) + J_x^{*\top} f_1(\vec{x}_0, t_0) \cdot u(t_0) + \frac{1}{2} tr[G' J_{xx}^* G] + J_t^* \right] \Delta t \end{aligned}$$

Bibliography

[Abril et al., 2007] Abril, M., Barbers, F., Lingolotti, L., Salido, M. A., Tormos, P., and Lova, A. (2007). An assessment of railway capacity. *Transport. Res. Part E*, doi:10.1016/j.tre.2007.04.001.

[Anderson and Moore, 1989] Anderson, B. and Moore, J. (1989). *Optimal Controls, Linear Quadratic Methods*. Prentice-Hall International, Englewood Cliffs, NJ.

[Applebaum, 2004] Applebaum, D. (2004). Levy processes - from probability fo finance and quantum groups. *Notices of AMS*, 51:1336–1347.

[Asmussen and Glynn, 2007] Asmussen, S. and Glynn, P. (2007). *Stochastic Simulation: Algorithms and Analysis*. Springer-Science, New York.

[Burdett and Kozan, 2006] Burdett, R. L. and Kozan, E. (2006). Techniques for absolute capacity determination in railways. *Transport. Res. Part B*, 40:616–632.

[EC, 2001] EC (2001). White paper: European transport policy for 2010: time to decide. European Commission. http://ec.europa.eu/transport/white-paper/documents/doc/lb-texte-complet-en.pdf.

[Efron and Tibshirani, 1993] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. CRC Press, Boca Raton, Florida.

[Fishman, 1996] Fishman, G. S. (1996). *Monte Carlo, Concepts, Algorithms, and Applications*. Springer-Verlag, New York.

[Gentle, 2003] Gentle, J. E. (2003). *Random Number Generation and Monte Carlo Methods, 2nd ed.* Springer-Science, New York.

[Glasserman, 2004] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York.

[Hanson, 2007] Hanson, F. (2007). *Applied Stochastic Processes and Control for Jump Diffusions: Modeling, Analyis, and Computation*. SIAM, Philadelphia, PA.

[Huisman and Boucherie, 2001] Huisman, T. and Boucherie, R. J. (2001). Running times on railway sections with heterogeneous train traffic. *Transport. Res. Part B*, 35:271–292.

[Jackel, 2003] Jackel, P. (2003). *Monte Carlo Methods in Finance*. Springer-Science, New York.

[Karatzas and Shreve, 1998] Karatzas, I. and Shreve, S. E. (1998). *Brownian motion and Stochastic Calculus*. Springer-Verlag, New York.

[Kloeden and Platen, 1999] Kloeden, P. E. and Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations, 3rd printing*. Springer-Verlag, Heidelberg, Germany.

[Kyprianou, 2005] Kyprianou, A. E. (2005). *Introductory Lectures on Fluctuation of Levy Processes with Applications*. Springer-Verlag, Heidelberg, Germany.

[Longo and Stok, 2007] Longo, G. and Stok, R. (2007). Estimation of railway capacity using stochastic differential equations. In *Proceedings of the 2nd International Seminar on Railway Operations Modelling and Analysis*, Hannover, Germany. eds. I.A. Hansen, J.P. Pachl, A. Radtke, E. Wendler.

[Mattson, 2004] Mattson, L. G. (2004). Train service reliability: A survey of methods for deriving relationship for train delays. Technical report, Department of Transport and Economics, Royal Institute of Technology, Stockholm. http://users.du.se/∼jen/Seminarieuppsatser/Forsening-tag-Mattsson.pdf.

[Meester and Muns, 2007] Meester, L. E. and Muns, S. (2007). Stochastic delay propagation in railway networks and phase-type distributions. *Transport. Res. Part B*, 41:218–230.

[Milstein and Tretyakov, 2003] Milstein, G. N. and Tretyakov, M. V. (2003). *Stochastic Numerics for Mathematical Physics*. Springer-Verlag, Heidelberg, Germany.

[Oksendal, 2000] Oksendal, B. (2000). *Stochastic Differential Equations, An Introduction with Applications, 5th Edition*. Springer-Verlag, Heidelberg, Germany.

[Osher and Fedkiw, 2003] Osher, S. and Fedkiw, R. (2003). *Level Set Methods and Dynamic Implicit Surfaces*. Springer-Verlag, New York.

[Pachl, 2002] Pachl, J. (2002). *Railway Operation and Control*. VTD Rail Publishing, Mountlake Terrace (USA).

[Piro, 2001] Piro, G. (2001). *Materiale Rotabile e Norme di esercizio FS*. CIFI Collegio Ingegneri Ferroviari Italiani, Italy.

[Platen and Heath, 2006] Platen, E. and Heath, D. (2006). *A Benchmark Approach to Quantitative Finance*. Springer-Verlag, Heidelberg, Germany.

[Press et al., 1992] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

[Protter, 2004] Protter, P. E. (2004). *Stochastic Integration and Differential Equations*. Springer-Verlag, Heidelberg, Germany.

[Shreve, 2004] Shreve, S. (2004). *Stochastic Calculus for Finance II, Continuous-Time models*. Springer-Science, New York.

[UIC, 1983] UIC (1983). *UIC leaflet 405-1, Method to be used for the Determination of the Capacity of Lines*. UIC International Union of Railways, France.

[UIC, 2004a] UIC (2004a). Capacity management (capman phase 3), summary report. Technical report, UIC International Union of Railways, France. http://www.uic.asso.fr/download.php/infra/2004-Capman-3-Summary.pdf.

[UIC, 2004b] UIC (2004b). *UIC leaflet 406 R, Capacity*. UIC International Union of Railways, France.

[Yong and Zhou, 1999] Yong, J. and Zhou, X. Y. (1999). *Stochastic Controls, Hamiltonian Systems and HJB Equations*. Springer-Verlag, New York.

[Yuan et al., 2006] Yuan, J., Goverde, M. P., and Hansen, I. A. (2006). Evaluating stochastic train process time distribution models on the basis of empirical detection data. In *Proceedings of the 10th International conference on Computers in Railways*. eds. J. Allan, C.A. Brebbia, A. F., Rumsey, G. Sciutto, S. Sone, C.J. Goodman.

[Yuan and Hansen, 2007] Yuan, J. and Hansen, I. A. (2007). Optimizing capacity utilization of stations by estimating knock-on train delays. *Transport. Res. Part B*, 41:202–217.

Acknowledgements

I would like to thank all the people working at the Department of Civil and Environmental Engineering in Trieste. Special thanks to my dear friend Christian for his suggestions about theoretical stochastic stuff. My patient wife Dorina and my lovely son Umberto deserve to be thanked and to receive the dedication of this work as a present (real gifts will follow!).