LOOKING THROUGH THE CUTOFF WINDOW
LOOKING THROUGH THE CUTOFF WINDOW

CARLO LANCIA
The main topic of the present dissertation is the cutoff phenomenon for discrete-time Markov chains. We speak about cutoff when a random process experiences a sudden, abrupt convergence at a deterministic time after having been for a long time arbitrarily far from equilibrium.

The first part of this essay builds a general framework for studying cutoff behaviour. In the recent literature many works have appeared on the topic of cutoff, but very few of them exploit any intuition from statistical mechanics in establishing the phenomenon. Cutoff is completely understood nowadays for the class of birth-and-death chains. Unfortunately, in many cases of interest the process under examination is not within that class. Nevertheless, it is often the case that a projection of the original chain returns a birth-and-death chain or a more opportune process. If this action is performed capitalising on the entropic properties of the original process, the resulting projected chain is likely to exhibit a drift towards a relatively small region of the projected state space. Such a region will correspond to those states that are most likely to be visited under equilibrium conditions. The aforementioned drift, in turn, can provide a quasi-deterministic trajectory to the portion of the state space where the stationary distribution is most concentrated, implying cutoff-like behaviour.

The role played by entropy in highlighting the drift is extensively studied through a number of examples like many-particles systems, card shuffling models, birth-and-death chains and random walks on higher-dimensional structures. The main results provided therein use the language and framework of hitting times to characterise the cutoff phenomenon and the existence of two contributions to the cutoff window, in which the sharp convergence takes place. The cutoff window arises as an interplay between the strength of the drift and the thermalisation time that is needed to relax inside the region of the state space where the stationary distribution is mostly concentrated.

The second part of the research focuses on a queueing system where the arrivals are given by a point process called Pre-Scheduled Random Arrivals. The arrival pattern is obtained by superimposing random fluctuations to a constant stream of customers. This problem was posed and intensively studied in the late 50’s by the founders of queueing theory, but so far it remains unsolved. This essay proposes a method to approximate the stationary distribution up to the desired precision when the fluctuations are exponentially distributed. The results obtained suggest that this queueing system exhibits cutoff as in the first part of the dissertation.

The study of this model is eventually motivated by comparing the theoretical queue length with actual data from air traffic applications.
L’argomento principale di questa ricerca è il cutoff per catene di Markov a tempo discreto. Si parla di cutoff quando un processo stocastico converge improvvisamente allo stato di equilibrio dopo essere stato arbitrariamente lontano da tale stato per un tempo molto lungo.

La tesi è divisa in due parti. La prima fornisce una serie di strumenti generali, utili per lo studio del cutoff. Su tale tema sono apparsi recentemente molti articoli in letteratura, ma la maggior parte di essi studia il fenomeno senza sfruttare l’intuizione che può fornire la meccanica statistica. Il cutoff è ormai completamente caratterizzato per la classe di processi che va sotto il nome di catene di nascita e morte. In molti casi di interesse, purtroppo, il processo in esame non appartiene a tale classe ma è tuttavia possibile ottenere una catena di nascita e morte o un processo più conveniente attraverso una proiezione. Se questa operazione viene fatta utilizzando al meglio le proprietà entropiche del processo originale, il processo proiettato presenterà una tendenza a muoversi verso una regione ridotta dello spazio degli stati che corrisponde all’insieme degli stati che hanno maggiore probabilità di essere visitati in condizioni di equilibrio. Questo effetto di deriva è in grado di garantire una traiettoria quasi deterministica verso la porzione dello spazio degli stati dove la misura di equilibrio è maggiormente concentrata, producendo la convergenza tipica del cutoff.

Il ruolo giocato dall’entropia nell’evidenziare la deriva viene studiato in maniera approfondita attraverso una serie di esempi, quali sistemi con un grande numero di particelle, modelli di mescolamento delle carte, catene di nascita e morte, e cammini aleatori su strutture multidimensionali. Il risultato principale usa il linguaggio e gli strumenti degli hitting times per caratterizzare il fenomeno del cutoff e per mostrare l’esistenza di due contributi alla finestra di cutoff. La finestra è l’intervallo temporale in cui avviene la convergenza allo stato di equilibrio, la quale si configura come mutua interazione tra l’intensità della deriva e il processo di termalizzazione, durante il quale il processo perde memoria della traiettoria passata e diffonde nella regione dove la misura stazionaria è concentrata.

La seconda parte di questa ricerca è focalizzata su un modello di code in cui gli arrivi sono la realizzazione di un processo stocastico chiamato Pre-Scheduled Random Arrivals. Il processo degli arrivi è ottenuto imponendo fluttuazioni aleatorie ad un flusso costante di clienti. Tale modello è stato introdotto negli anni ’50 dai fondatori della teoria delle code e da allora è stato studiato intensamente, ma una caratterizzazione completa dello stato di equilibrio non è stata ancora trovata. In questa tesi viene proposto un metodo per approssimare la misura stazionaria alla precisione desiderata nel caso speciale in cui i ritardi imposti siano distribuiti esponenzialmente. I risultati ottenuti sembrano suggerire che tale modello di code esibisca cutoff nel senso illustrato nella prima parte della tesi.

Lo studio di questo modello di code è infine motivato da applicazioni in ambiti di traffico aereo. In particolare viene confrontata la lunghezza della coda teorica con quella proveniente da un database di dati reali.
PUBLICATIONS

Some ideas and figures have previously appeared in the following publications:


7.3 London Heathrow Airport 111
7.4 Insensitivity to the delays’ distribution 116
7.5 The SESAR Programme 119
7.6 Mixed-traffic scenarios 120
7.7 Python code for PSRA simulation 123

CONCLUSIONS 131

APPENDICES 133

A HITTING TIME OF THE BULK FOR THE MEAN-FIELD ISING MODEL 135
B HITTING TIME OF THE BULK FOR THE PARTIALLY DIFFUSIVE RANDOM WALK 141

BIBLIOGRAPHY 143
LIST OF FIGURES

Figure 1  Graphical interpretation of TV-distance.  8
Figure 2  Cutoff for the biased random walk.  12
Figure 3  Top-in-at-random shuffle.  13
Figure 4  Rising sequences in a deck of 13 cards while it is riffle-shuffled.  21
Figure 5  Approach to equilibrium vs. behaviour of rising sequences.  22
Figure 6  Distribution of the number of rising sequences in a deck of 52 cards.  23
Figure 11  Approach to equilibrium of a biased random walk of size 10,000.  30
Figure 12  Approach to equilibrium of a biased random walk of size 100,000.  31
Figure 13  The Ehrenfest Urn: approach to equilibrium depending on the initial state.  46
Figure 14  Evolute measure and approach to equilibrium of an Ehrenfest Urn.  48
Figure 15  Coupling scheme for the random walk on the cylinder.  59
Figure 16  Transitions of EDA/D/1 in the quarter plane.  69
Figure 17  Stationary queue of EDA/D/1 : simulation vs. truncated expansion.  87
Figure 18  Paths of the typical trajectories from (0, 0) to T_l for an EDA/D/1 queue.  100
Figure 19  Example of PSRA arrival pattern.  108
Figure 20  STARS of London Heathrow airport.  111
Figure 21  Qualitative layout of the inbound air traffic over London Heathrow.  113
Figure 22  Fit of the queue distribution at London Heathrow airport.  114
Figure 24  Output of a PSRA/D/1 queueing system for different delays’ PDF.  117
Figure 26  PSRA vs. 4D Trajectory, FIFO policy.  121
Figure 27  PSRA vs. 4D Trajectory, BEBS policy.  122
LIST OF TABLES

Table 1  Goodness of fit for the queue distribution at London Heathrow airport  116

LISTINGS

Listing 1  Sage code to compute $a_k^j(z)$.  82
Listing 2  Sage code to compute $A_k^j(z)$.  83
Listing 3  Sage code to compute $Q_k^j(z,y)$.  84
Listing 4  Sage code to compute $Q(z,y)$ up to any prescribed order $n$.  84
Listing 5  Python code for simulations of different single-server queue models  123

ACRONYMS

ASMA  Arrival Sequencing and Metering Area
ATA  Actual Time of Arrival
ATC  Air Traffic Control
ATM  Air Traffic Management
BEBS  Best Equipped Best Served
BDC  Birth-and-Death Chain
BT  Business Trajectory
BVP  Boundary Value Problem
EDA  Exponentially Delayed Arrivals
EU  European Union
FIFO  First In First Out
GDP  Ground Delay Program
HDR  High Density Rule
IATA  International Air Transport Association
ICAO  International Civil Aviation Organization
IID  Independent and Identically Distributed
IFR  Instrument Flight Rules
LACC  London Area Control Centre
LTCC  London Terminal Control Centre
MC  Markov Chain
MCMC  Markov Chain Monte Carlo
MMQS  Markov Modulated Queueing System
PDF  Probability Density Function
PSA  Power Series Approximation
PSRA  Pre-Scheduled Random Arrivals
RHS  Right Hand Side
SESAR  Single European Sky ATM Research
STAR  Standard Terminal Arrival Route
SWIM  System Wide Information Management
TMA  Terminal Manoeuvring Area
VFR  Visual Flight Rules

SYMBOLS

The following is a list of frequently used symbols, their meaning is invariant throughout the essay.

\( \delta_{i,j} \)  Kronecker’s delta;
\( P(A) \)  Probability of event \( A \);
\( E[X] \)  Expectation of random variable \( X \);
\( Var[X] \)  Variance of random variable \( X \);
\( \sigma[X] \)  Standard deviation of random variable \( X \);
\( X^t_n \)  Generic Markov Chain (MC);
\( \Omega_n \)  State space of a generic MC;
\( P_n \)  Transition matrix of a generic MC;
\( \mu_0^n \)  Initial measure of a generic MC;
\( \mu^t_n \)  Evolved measure after \( t \) steps of a generic MC;
\( \pi_n \)  Stationary measure of a generic MC;
\( \|\mu_1 - \mu_2\|_{TV} \)  Total-variation distance between \( \mu_1 \) and \( \mu_2 \);
\( \tau \)  
Hitting time of a generic set, or  
Generic stopping time;

\( \gamma \)  
Coalescence time of a generic coupling;

\( \{ A_{n,0} \}_0 \)  
Family of nested subsets;

\( \zeta_n^{\theta} \)  
Hitting time of the set \( A_{n,0} \);

\( \mathcal{C} \)  
Superscript to denote complementary set;

\( \equiv \)  
Asymptotic equivalence;

\( o(\cdot) \)  
Asymptotically dominated by \( \cdot \);

\( O(\cdot) \)  
Asymptotically bounded above by \( \cdot \);

\( \Theta(\cdot) \)  
Asymptotically bounded above and below by \( \cdot \);

\( \sim \)  
Equivalence relation, or  
Distribution of a random variable;

\( \check{\cdot} \)  
Superscript in projected quantities;

\( \rho \)  
Traffic index, load;

\( Q_{n,1} \)  
Stationary distribution of \( \text{EDA/D/1} \);

\( Q(z, y) \)  
Generating function of \( Q_{n,1} \).
When I was a bachelor student, the odds of me getting a PhD in Mathematics were two to the power of two hundred and seventy-six thousand seven hundred and nine to one against. My favourite activity was playing tressette, an Italian game close to bridge. The breaks between lectures were perfect occasions to play small tournaments. I was indeed so addicted as a player that at the final exam of Discrete Mathematics I was asked to compute the probability of a ten-card hand* with at least one void suit. It was my baptism.

In the Fall 2004 I met Benedetto Scoppola, who taught me a first course in Probability and Markov Chains. In the following years I have taken his courses in Graph Theory, Advanced Markov Chains, and Queueing Theory. He has also supervised both my bachelor’s and master’s theses, respectively, an application of a perfect sampling algorithm, the Randomness Recycler, to the Clique Problem on Erdős-Rényi random graphs, and the development of an original MCMC algorithm for the so-called Terminal Steiner Tree Problem. These works have the idea of minimising an object function by sampling from a Gibbs measure in common, the so-called Statistical Mechanics approach to Operation Research problems.

Regarding the subject of this essay there is a nice story to tell. In the final year of my master I had to complete a self-study activity. I thought I would ask Gianluca Guadagni, my former teacher of Stochastic Differential Equations, for a small research topic. My request was moved by a desire for payback: I felt he had not fairly graded my exam with respect to some other colleagues. Expecting something about Stochastic Differential Equations, I went to him and asked for a research subject. I had clearly counted my chickens before having hatched them because he suggested me to study cutoff instead. Quite reluctantly, I accepted – I had just shaped my future years, but who could have ever foretold that?

I started reading [DLP10], which was the most recent paper on cutoff at that time. Quite soon, though, I encountered the wonderful papers on cutoff by Persi Diaconis, [AD86] and [BD92] in particular, and I decided that I liked them the most. In those papers cutoff was proved for two deck-shuffling models, the perfect match for my cards passion.

After graduating with a master’s degree in Mathematical Engineering, I went for a few job interviews without the necessary determination: I had more than half a mind to try an academic career, and in the end I entered the open competitions for a PhD position. Before applying for it, I went to my former teacher in Quantum Mechanics and Statistical Physics, Andrey Varlamov, and took counsel with him. He is a great scientist and wonderful speaker, one of the best I have ever encountered in my life, and a very pragmatic person into the bargain. “For being a researcher you need two vectors,” he explained to me, his thumb and forefinger forming an L. “The first is the vector of exploration,” he continued, twisting his wrist as if his forefinger were a drill, “and you need it for getting inside things. However, this vector is rather useless without the second one, the vector of money, which is orthogonal to the former. Only by working in this direction you can live on research. I think you already possess the first vector but regarding the second, I can give no guarantees.”

* In Italy regular decks are composed of 40 cards.
I was quite lucky to win a position in the same university I had just graduated from, so that I could continue working with Benedetto. Together we decided that the best option for my PhD project was to continue studying cutoff phenomena. At that time, a paper by Javiera Barrera, Olivier Bertoncini, and Roberto Fernández had appeared at arXiv.org. In [BBF09] cutoff was investigated in the class of Birth-and-Death Chains (BDCs) with the explicit use of hitting times and the concept of drift. For BDCs there already existed a complete characterisation of cutoff, but this was actually more focused on the spectral properties of the transition matrix [DSC06, DLP10], and hitting times could be found only by looking under the hood. We thought of bridging the two approaches and started to develop a general methodology for proving cutoff. The idea was to tackle the problem in a way more familiar to statistical mechanics, keeping at the same time the probabilistic view of a sudden convergence of the evolved measure to the equilibrium one.

After one year, we had written the draft of a paper which contained an interesting result, the forebear of Theorem 3.2. It is quite a pity that its proof was formally incorrect. Even worse, I discovered this rather important fact during a seminar.

It was 2011, and Francesca Nardi had invited me to present my research at EURANDOM. I was actually a bit nervous before the presentation, but I never expected I was going to argue with Sergey Foss, who was among the listeners. The incident happened because of a wrong formula: I had bounded a random variable instead of its expectation. As a matter of fact, it was not such a big error, because the random variable was supposed to be quasi-deterministic; it was an informal seminar into the bargain, and I was just supposed to present my ideas in a more-colloquial-than-detailed way. At that point, the optimal strategy would have been admitting some sloppiness in the formulas and be safe.

Pride goes before a fall, they say, and I had the very brilliant idea of claiming I was right. Sure enough, a quarrel started, during which I managed to score no points for myself but a few for Sergey. After a while I was deeply buried in the troubles I had just looked for. In a desperate try for solving the situation I scribbled something at the blackboard, but things did only worsen as he left his seat and came to the board – “Surely not to negotiate my surrender, but to finish me off”, I thought. I was about to give up when Francesca intervened in the dispute and snoozed it. I completed the talk nearly whispering, running ashamedly through the remaining slides.

Later that year, Francesca became co-supervisor of my PhD project, together with Benedetto we fixed the draft and published our first paper. It was dedicated to the loving memory of Roberta Dal Passo, a former professor of the University of Rome Tor Vergata. I will be forever grateful to her, who bred my confidence and mindset.

I visited EURANDOM once more during the spring of 2012 to continue my collaboration with Francesca. The travel also gave me the opportunity to discuss several times with Roberto Fernández about cutoff, diffusion and thermalisation. The title of this dissertation was actually coined by him during an enlightening conversation on the meaning of cutoff. Almost at the end of that visit we asked for a joint degree agreement. Thanks to Francesca’s dogged determination we quickly found a joint agreement between TU/e

* Apparently, this is the standard way for Russian mathematicians to express they have an interest in the topic. At the moment, though, I was mainly concerned over the possibility to get just kicked out.
and Tor Vergata. In November 2012 I moved to Eindhoven and started writing this dissertation.

The essay spans aspects of Theoretical and Applied Probability with the topic of cutoff phenomena; a little Statistics appears as well in the last chapter, which covers an actual Air Traffic Management problem. The dissertation also bridges the way cutoff is approached by different communities of probabilists, namely, those researchers who study cutoff as a singular kind of convergence to the equilibrium measure, and those researchers who are interested in cutoff as the potential counterpart (and trigger) of metastable behaviours. Looking through the cutoff window, those who describe the phenomenon as the former see it with the eyes of the latter.

Extending across many different subjects, the trademark of this dissertation just lies in the width of the topics it traverses. Personally, I really like how smooth the transition is from the initial chapters, purely theoretical, to the last one, entirely devoted to an important applied problem. As Andrey said, research means to delve into problems by the vector of exploration, and I had to find a trade-off between width and depth, or I would have never completed this work. While writing it, I have tried my best to keep the formalism at bay, so as to enhance its bridging qualities. The essay should be mostly accessible even to an undergraduate student, at any rate I have written it having this target in mind.

Eindhoven, 2 October 2013
Um, um, um. Stop that thunder!
Plenty too much thunder up here.

What’s the use of thunder? Um, um, um.
We don’t want thunder; we want rum;
give us a glass of rum. Um, um, um!

— Herman Melville [Mel51]

ACKNOWLEDGEMENTS

There are many people I owe much to. I guess this is the right place where to write that I am really fond of them.

First of all, I wish to thank both my supervisors. Francesca, this joint project was made possible only by your calm and perseverance. It is a pity that you have taken only a little part in it, I sincerely wish you all the best and a fast recovery. Benedetto, you truly represents to me what being a scientist means. I have been a student of yours for ten years or so, it has been a constant growth for me ever since. Should I ever be parted from you, be sure that I will invoke un po’ di puma first.

Next, there are a few people I wish to credit for their precious help, which has been important, if not fundamental, in the realisation of this essay. I would like to start expressing my gratefulness to Remco van der Hofstad and Roberto Fernández, discussing with them has been for me a source of positive inspiration. I would like to address earnest thanks to Elisabetta Scoppola, for the words of encouragement she had for me at the early stage of this work. I wish to thank also Maria Vlasiou, Stella Kapodistria, and Serban Badila, for the useful conversations we had on the topics of Chapter 5. Many thanks to Sokol Ndreca, Gianluca Guadagni, Antonio Iovanella and Guglielmo Lulli, for their precious friendship and the collaboration on the subjects of Chapter 5 and 7. A special mention is due to Emilio Cirillo: our walks along the Dommel have proved really helpful in clearing up the heuristics of Chapter 6. I would also like to thank Maria Virginia Caccavale, who processed the Heathrow dataset, and Damiano Taurino, for his helpful comments on the SESAR programme and the 4D Trajectories.

I wish to express my deepest gratitude to the espresso brigade, that is to say, Alessandro, Enrico, Julien, Maria Luisa, Rui, and my officemate, Martin: hard times await me without your sweet company. I also owe an apology to my dearest friend, Ale, for the unfair words I said regarding the style of his master’s thesis, they were only sour grapes. In Italian we say “Chi disprezza compra”, and it looks like I am no exception to this rule after all. Also, I wish to address heartfelt thanks to Alex, who read the non-technical parts and gave me important feedback on the cover design.

Finally, I want to mention my family, Piero, Giuliana, Michele, and Giacomo. I could write double the amount of these pages and give just an introductory view of how complex and difficult my feelings for them are; but my equilibrium state, what really lies at the bottom of my heart, is just love and serenity. Vi voglio bene sempre.

Last but not least, I wish to thank Silvia, my sweet love. However rough might it be the path that lies in front of us, if you are at my side I will never falter. This small work of mine is dedicated to you.
CALL ME CUTOFF. Some instants ago – never mind how long precisely – having little or no states in my trajectory, and nothing particular to interest me on the tails, I thought I would converge about a little and see the relevant part of the state space.
The present essay is mainly focused on the cutoff phenomenon for finite Markov Chains (MCs). A MC is a model for random dynamics, i.e., a probabilistic description of the evolution of an observable. The signature feature of a MC is the dependence on the past observations only through the current state. In other words, the observable will take the next value by running a probabilistic update rule that depends on the current value only and comes typically bundled in a square matrix called transition kernel, or transition matrix. Such a characteristic is known as markovianess, or memoryless property, and it can be summarised in the following way: whatever the past sequence of values already taken, only the very last matters.

Under mild assumptions on the transition matrix, a MC exhibits a unique stationary state (or equilibrium state), that is, a probability distribution that gives the likelihood of measuring a certain value of the observable in equilibrium conditions. The stationary state is an asymptotic characterisation of the MC, in the sense that it describes the long-run behaviour of the observable. The existence of a unique equilibrium distribution is often quite easy to prove, this partly explains why MCs are important sources of elegant mathematics in addition to being widely deployed tools for modelling stochastic evolutions.

It is also possible to use MCs as efficient computational devices. Indeed, the stationary distribution of a MC can be easily sampled by simulating the chain evolution until equilibrium is reached. This idea has given rise to the so-called Markov Chain Monte Carlo (MCMC) Paradigm, with plenty of applications ranging from approximate counting and integration to combinatorial optimisation, statistical physics, and statistical inference. The interested reader is referred to [Jer93, JS96, Jer03] for further details. However, the validity of MCMC algorithms crucially depends on how fast the MC being run reaches its equilibrium state. Slow convergence to equilibrium will, in practice, result in a useless algorithm that requires too large a computational time to be run. On the other hand, it is quite dangerous to stop the algorithm without precise knowledge of whether equilibrium has been reached or not. Indeed, it may happen that the output is unreliable, due to the sampling from a distribution which is not the targeted one.

There exists an entire industry devoted to the characterisation of the convergence to equilibrium, an exhausting exposition of these techniques can be found in [MT06, LPW06]. Estimating $\lambda_2$, the second eigenvalue of the transition matrix, is the most frequently addressed problem since in many instances $\lambda_2$ is a satisfactory proxy for the actual distance from equilibrium after the chain has evolved for $t$ steps. The picture is radically different when the MC exhibits cutoff because the cutoff phenomenon gives an extremely accurate description of the convergence to the stationary state. More precisely, a MC exhibits cutoff behaviour if the distance from the equilibrium state suddenly drops from almost the maximum to almost the minimum value. Such an abrupt convergence takes place over a short time-window $b$, negligible with respect to the time $a$ for the phenomenon to arise. In this respect, cutoff is much more informative than the classical bound in terms of the second
eigenvalue cited above. In particular, for MCMC algorithms the presence of cutoff immediately advises

- not to run the algorithm execution for times much larger than \( a \) as it would be a waste of computational resources to simulate the \( MC \) much longer than necessary;

- not to stop it before \( a \) steps have passed, for in this case the output would be sampled from a probability distribution which is as far as possible from the targeted one.

The cutoff phenomenon naturally arises in many models of interest, see for instance [Dia96]. Hence, the characterisation of cutoff phenomena is a favourable addition to the study of non-asymptotic behaviour and convergence to equilibrium, a central topic in the modern theory of MCs.

The name cutoff phenomenon first appeared in the literature in [AD86], although the first results in this subject were obtained a few years earlier [DS81]. In 1994 Yuval Peres conjectured that cutoff occurs if and only if the time to reach equilibrium is much larger than the ratio \( 1/\lambda \), see [Per04]. It took more than ten years to have that conjecture proved within a special class of MCs, known as Birth-and-Death Chains (BDCs) [DSC06, DLP10]. This was possible because BDCs manifest a peculiar link between the transition matrix and the stationary distribution, and between the spectral properties of the transition kernel and the typical evolution of the chain. Whether Yuval Peres’ characterisation holds for a wider class of chains still remains an open question.

The remaining literature on the cutoff phenomenon is mainly composed of model-dependent results. In his 1996 survey Persi Diaconis wrote: “At present writing, proof of a cutoff is a difficult, delicate affair, requiring detailed knowledge of the chain, such as all eigenvalues and eigenvectors. Most of the examples where this can be pushed through arise from random walk on groups, with the walk having a fair amount of symmetry”. Since then the picture has essentially remained unchanged, see for instance the discussion in [LS13]. However, in the last few years cutoff has been investigated in its generality for its relation to the exponential escape, a feature of metastable behaviour. Metastable behaviour can be roughly described as a (exponentially) long sojourn in a state of apparent equilibrium followed by a quick transition to the stable equilibrium. In [BBF09] the authors showed that under suitable hypotheses, cutoff and exponential behaviour are two sides of the same coin. To study both phenomena in a common framework, they had to renounce the description of cutoff in terms of an abrupt fall-off in the distance from equilibrium. They chose, instead, the common, unifying language of hitting times. A hitting time is a random variable representing the first time a state is visited by the \( MC \). For the exponential behaviour, hitting times are used to mark the first moment the chain visits a state sufficiently far from the metastable state for the escape to be successful. When studying cutoff, hitting times are used to flag the first time the chain visits target quantiles of the stationary distribution.

This research is conveniently situated at the interface between the two approaches to cutoff described above. On one hand it tackles the phenomenon classically, i.e., with the formalism of the distance from the equilibrium state, on the other it explicitly speaks the language of hitting times. It also uses the idiom of statistical physics. The quote above by Persi Diaconis mentions a fair amount of symmetry; it turns out that these symmetries, if any, may often be used to define an equivalence relation on the set of all possible values
of a MC, the state space. Projecting the chain onto the quotient state space by means of this equivalence relation often leads to a simpler process for which cutoff is easier to prove. The projection highlights the entropy of each equivalence class and helps to establish the paths of the typical trajectories in the state space. In this way it is therefore possible to characterise how the chain approaches the relevant portion of the state space, i.e., the one that corresponds to the typical values taken by the MC at equilibrium. In terms of the stationary state the relevant part of the state space translates to the appropriate quantiles of the equilibrium distribution. The cutoff time, \( a \), can then be interpreted as the expectation of the hitting time \( \zeta \) of this relevant part, whilst the cutoff window, \( b \), is discovered to arise as the intertwining of two separate contributions: the standard deviation of \( \zeta \), and the thermalisation.

Among the original contributions of this work, the thermalisation is one of the most important, it is the new thing we see looking through the cutoff window. Roughly speaking, it is the time to reach equilibrium starting from within the relevant part of the state space. This means that the typical trajectory of a MC exhibiting cutoff can be decomposed in two parts: the approach to the relevant quantities of the state space and the relaxation to equilibrium once they have been reached. Each of these parts is naturally studied over the corresponding time scale, easing up the task of proving cutoff. In fact, a quite common approach to the proof of cutoff is the design of a coupling, sufficiently clever to allow estimates on the overall time scale \( a \pm b \). Such a detailed inspection is not needed here, being intrinsic to the modus operandi developed. The proposed new methodology is shown at work in a variety of examples, mainly classical and non-classical models of random walks.

The essay continues with the study of the cutoff phenomenon for a family of queueing systems, extremely important for both historical and applied arguments. These are single-server queues with fixed-length, deterministic service time, and arrival process obtained as the superposition of Independent and Identically Distributed (IID) random shifts to a pre-scheduled stream of customers. An arrival stream of this kind was introduced for the first time by C.B. Winsten in the 50s, who named it the problem of the late customer [Win59], it was then studied by the pioneers of queueing theory, in particular by D.G. Kendall [Ken64]. Currently it is better known by the name of Pre-Scheduled Random Arrivals (PSRA) [GNS11, Gwi11, NH12]. It is very fitting in describing the actual stream of arrivals in many situations where planned inflow of customers is inherently subject to random fluctuations, e.g. transportation systems.

The queueing system obtained in the special case of exponentially distributed delay can be easily described by mean of a bivariate MC, i.e. a two-component chain. Finding the stationary distribution for this MC is a very difficult and still open problem. The equilibrium state can be investigated using the bivariate generating function to produce an iterative functional scheme, able to approximate the generating function to the desired order. The analysis of the generating function also leads to a fair location of the relevant quantiles of the stationary distribution in the quarter plane. Then, using the already developed methodology, a cutoff-like abrupt convergence can be shown.

The research is completed with a study of the inbound air traffic at the London Heathrow Airport. The analysis of a data set of actual arrivals gives a description of the airport congestion. A comparison of the latter with the output of the family of queues described above shows a surprising goodness
of fit. The existence of the cutoff behaviour for such a system is a synonym of resilience, a key performance of models for Air Traffic Management (ATM). According to [Glu12], resilience means that starting from a stress situation, like peaks of traffic load, time deficit, operational procedures, limitation and reliability of equipment, or abnormal/emergency situations, the system steadily reaches equilibrium, that is to say, normal operation condition. Due to the presence of cutoff the system can cope very well with congestion, and the time needed to recover from a stress situation can be estimated with high precision.
Cutoff belongs to the mixing properties of a Markov Chain (MC), a sequence of random objects describing the evolution of a possibly complex system. The mixing properties of a MC specify the existence of a stationary distribution and the speed of convergence to it. The cutoff phenomenon is a strong realisation of the former and a sharp characterisation of the latter at the same time.

For these statements to make sense from a mathematical point of view we start with some definitions.

### 2.1 Representation of a Markov Chain

Let \( \Omega_n \) be a sequence of finite sets. A MC is a collection of \( \Omega_n \)-valued random variables \( X^0_n, X^1_n, X^{n-1}_n, \ldots \) satisfying the so-called Markov (or memoryless) property: for each \( i, j \in \Omega_n \) and for each sequence of states \( \{k_s\}_{0 \leq s \leq t} \subseteq \Omega_n \) with \( k_{t-1} = i \) and \( k_t = j \),

\[
\mathbb{P} \left( X^t_n = k_t \mid X^0_n = k_0, X^1_n = k_1, \ldots, X^{t-1}_n = k_{t-1} \right) = \mathbb{P} \left( X^t_n = j \mid X^{t-1}_n = i \right), \tag{2.1}
\]

\[
= P_n(i, j). \tag{2.2}
\]

The meaning of (2.1) is that the evolution of a MC, i.e., its next value \( X^t_n \), depends only on the current position \( X^{t-1}_n \) whatever may it be its past trajectory \( X^0_n, X^1_n, \ldots, X^{t-2}_n \) in the state space. Equation (2.2) points out that the probability of a transition from state \( i \) to state \( j \) does not depend on time, in this case the MC is said to be homogeneous. Although inhomogeneous MCs can be very useful to model relevant systems, we will not consider them and henceforth restrict ourselves to the homogeneous case.

**Remark 2.1.** All the quantities introduced so far display a subscript \( n \). This is indeed the usual notation used to represent families of MCs. Families of MCs are a key ingredient in the definition of cutoff, to be introduced later on in Section 2.6.

By (2.2), a MC can be represented as a \( |\Omega_n| \times |\Omega_n| \) square matrix \( P_n = \{P_n(i, j)\} \), usually called the transition matrix or transition kernel. The elements of a transition matrix are non-negative, that is, \( P_n(i, j) \geq 0 \), and if summed over any row they equal to unity, i.e., \( \sum_j P_n(i, j) = 1 \). Another interesting object to consider is the probability law the initial state \( X^0_n \) is chosen according to. This is called the initial distribution of the MC and is indicated by \( \mu^0_n \). The initial distribution can be thought of as a vector \( \mu^0_n \in [0, 1]^{|\Omega_n|} \) such that its \( i \)-th component is \( \mu^0_n(i) = P (X^0_n = i) \).

Once \( P_n \) and \( \mu^0_n \) are prescribed, the evolved measure after \( t \) steps can be computed. The evolved measure, \( \mu^t_n \), is a vector whose \( t \)-th entry is \( \mu^t_n(i) = P (X^n_i = i) \). It can be found by multiplying the \( t \)-th power of \( P_n \) by \( \mu^0_n \) to the left [Höf2], i.e.,

\[
\mu^t_n = \mu^0_n P^n_t.
\]
Remark 2.2. When $\mu_n$ is set, the evolved measures $\mu_n^t$ is given by a deterministic recursive relation and, in principle, could be computed at any specified time.

2.2 Topology of the State Space

Any MC can be naturally represented as a graph $\mathcal{G}(V, E)$, where the vertex set is $V = \Omega_n$ and the edge set $E \subseteq V \times V$ contains all the couples of states $(i, j)$ such that $P_n(i, j) > 0$. According to the geometrical intuition, two states $i$ and $j$ are said to be adjacent if the edge $(i, j)$ is in $E$. The state $i$ is said to communicate with state $j$ if there exists a path $L \subseteq \mathcal{G}$ that joins $i$ with $j$. If all vertices in $V$ communicate each other then the graph $\mathcal{G}$ is connected and the transition matrix $P_n$ is said to be irreducible. The interested reader is referred to [Bol08] for a comprehensive introduction on graph theory.

Each chain $X_n^i$ induces a neighbourhood structure on $\Omega_n$ through its unique graph representation $\mathcal{G}(V, E)$. Given a couple of states $(i, j)$, let $\mathcal{P}_{i,j}$ be the set of all paths joining $i$ with $j$, if any, and define the following function on $\Omega_n \times \Omega_n$:

$$d(i, j) = \begin{cases} \min_{L \in \mathcal{P}_{i,j}} \text{len}(L), & \text{if } \mathcal{P}_{i,j} \neq \emptyset, \\ \infty, & \text{otherwise}, \end{cases}$$

(2.3)

where $\text{len}(L)$ is the length of $L$, i.e., the number of edges that compose the path.

Remark 2.3. We note that ‘state $i$ is adjacent to state $j$’ and ‘state $i$ is communicating with state $j$’ may not be symmetric relations. In graph theory terms, $\mathcal{G}(V, E)$ is a directed graph. As such, the equality $d(i, j) = d(j, i)$ fails in general to hold, and the function defined by (2.3) is not a metric on $\Omega_n$.

Given a subset of the space state $A \subset \Omega_n$, we define its boundary as the set of states which are not in $A$ but adjacent to it, i.e.,

$$\partial A = \{ i \in A : d(i, j) = 1, j \in \Omega_n \setminus A \}. \quad \text{(2.4)}$$

In the following we will encounter very often a special type of MCs, named Birth-and-Death Chains (BDCs). The graph representation $\mathcal{G}(V, E)$ of a BDC is isomorphic to a segment, that is, the state space of a BDC can be put in a one-to-one correspondence with the set $\{1, 2, \ldots, |\Omega_n|\}$; from state $i$ only transitions to states $i, i - 1,$ and $i + 1$ are allowed. BDCs are of great interest in applications and are frequently used to model queueing systems. For a BDC, the set $\partial A$ contains two elements for each connected component of the restricted graph $\mathcal{G}(A, E)$, and if the starting point $i \in \Omega_n \setminus A$ is known then it is possible to establish in which state the chain will be found the first time it visits $A$. We come back to this very important remark later on in Section 3.3.

2.3 The Equilibrium Distribution

In Section 2.1 we have pointed out that the transition kernel of a MC is a non-negative matrix. If the chain happens also to be irreducible, i.e., if

$$\forall i, j \in \Omega_n \ \exists t \geq 1 \ \text{such that } P_n^t(i, j) > 0,$$

(2.5)

* The irreducibility condition (2.3) is equivalent to all couple of states being communicating in the sense of Section 2.2.
then the Perron-Frobenius Theorem states that the spectrum of $P_n$ is included in $(-1, 1]$ and that 1 a simple eigenvalue [Sen81]. Therefore,

$$
\exists ! \pi_n \quad \text{such that} \quad \pi_n = \pi_n P_n .
$$

(2.6)

Formula (2.6) establishes that there exists an invariant probability distribution for the MC. If a chain is started with initial distribution $\mu_n^0 = \pi_n$, then $\mu_n^t = \pi_n$ for all $t \geq 0$. Because of this property, $\pi_n$ is called the stationary (or equilibrium) distribution. A classical result about $\pi_n$ is the so-called Markov Chain Convergence Theorem [Gne66, Brégo, Hőz2]. It states that under the further assumption of the chain being aperiodic (see below for the definition),

$$
\pi_n = \lim_{t \to \infty} \mu_n^t .
$$

(2.7)

A MC $X_n^t$ is said to be aperiodic if

$$\gcd\{t \geq 0 : P_n^t(i, i) > 0\} = 1 ,$$

where $\gcd$ stands for the greatest common divisor.

By (2.7), a MC that is both irreducible and aperiodic progressively loses memory of its past, in the sense that the probability of finding the chain in state $i$ after a sufficiently large number of steps becomes independent of the starting state after a sufficiently large number of steps. Therefore, an irreducible and aperiodic chain is often called ergodic. The convergence expressed by (2.7) is usually meant to be component-wise, as in [Gne66]. In this case the theorem states the convergence in the sense of the $\ell_\infty$-norm, i.e.,

$$
\lim_{t \to \infty} \|\mu_n^t - \pi_n\|_\infty = \lim_{t \to \infty} \sup_{t \in \Omega_n} |\mu_n^t(i) - \pi_n(i)| = 0 .
$$

(2.7a)

However, any distance between probability distributions is suitable since the set of all probability measure on $\Omega_n$ is finite-dimensional. The most often used distance between probability measures is probably the so-called total-variation distance, frequently abbreviated to TV-distance,

$$
\|\mu_n^t - \pi_n\|_{TV} = \max_{A \subseteq \Omega_n} |\mu_n^t(A) - \pi_n(A)| ,
$$

(2.8)

$$
\frac{1}{2} \sum_{i \in \Omega_n} |\mu_n^t(i) - \pi_n(i)| .
$$

(2.9)

Remark 2.4. Figure 1 gives a graphical interpretation of total-variation distance between two probability distributions $\lambda$ and $\mu$. In particular, the overlap region has area $1 - |\lambda - \mu|_{TV}$. This means that if $\lambda$ and $\mu$ are supported on disjoint regions then their TV-distance is equal to 1.

Another distance of common use is the $\ell_p$-distance

$$
\|\mu_n^t - \pi_n\|_p = \left( \sum_{i \in \Omega_n} \left| \frac{\mu_n^t(i)}{\pi_n(i)} - 1 \right|^p \pi_n(i) \right)^{\frac{1}{p}} .
$$

(2.10)

Remark 2.5. The TV-distance is a number between 0 and 1, which satisfies $\|\mu_n^t - \pi_n\|_{TV} = 2 \|\mu_n^t - \pi_n\|_{TV}$. In addition, regarded as a function of time, the TV-distance is non-increasing, i.e.,

$$
\|\mu_n^t - \pi_n\|_{TV} \leq \|\mu_n^{t+1} - \pi_n\|_{TV} .
$$

Surprisingly this easy and useful property of TV-distance is not mentioned at all in many textbooks. A nice exception is [Jero3].
Separation is a further possibility for quantifying the likeness of two distributions. It is defined as

\[
\text{sep} (\mu, \pi) = \max_{i \in \Omega_n} \left\{ 1 - \frac{\mu_n(i)}{\pi_n(i)} \right\}.
\]

Although separation is not formally a distance (it is not symmetric), it is important for historical reasons. In the early studies on the cutoff phenomenon many important results were initially obtained for separation, see e.g. [AD86, AD87]. In particular, separation was used in the fundamental paper [DSC06] that fully characterises the cutoff phenomenon for the whole class of BDCs, see Section 2.8 below.

### 2.4 Reversibility

A MC is reversible with respect to a given probability measure \(\nu\) on \(\Omega\) if

\[
\nu(i) P_n(i,j) = \nu(j) P_n(j,i),
\]

in which case \(\nu\) is said to be a reversible measure with respect to \(P\). A reversible chain owns its name to the time reversal property it exhibits: if the initial distribution is \(\nu_0 = \nu\) then, for each sequence of states \(\{k_m\}_{m=1}^{\infty}\),

\[
\mathbb{P} \left( X_n^0 = k_0, X_{n+1}^1 = k_1, \ldots, X_{n+t}^t = k_t \right) = \mathbb{P} \left( X_n^0 = k_t, X_{n+1}^1 = k_{t-1}, \ldots, X_{n+t}^t = k_0 \right). \tag{2.11}
\]

Roughly speaking, equation (2.11) states that a reversible MC behaves the same regardless of whether time runs backwards or forwards. Another extremely important feature of reversible measures is that they are also stationary [Hö2].

For a reversible chain the following formula for the \(t\)-th power of the transition matrix can be proved:

\[
P_n^t(i,j) = \sqrt{\frac{\pi_n(j)}{\pi_n(i)}} \sum_{m=1}^{n} \lambda_{n,m}^t u_{n,m}(i) u_{n,m}(j), \tag{2.12}
\]

where \(\lambda_{n,m}\) and \(u_{n,m}\) are the \(m\)-th eigenvalue and eigenvector of the symmetric matrix \(P'_n = D P_n D^{-1}\), respectively, and \(D\) is a diagonal matrix such
that \( D(i, i) = \sqrt{\pi_n(i)} \), see [Fil91, LPWo6]. An almost immediate consequence of (2.12) is a bound on the total-variation distance from stationarity after \( t \) steps. For a reversible and ergodic MC started at time \( t = 0 \) from state \( i \),
\[
\| \mu_n^t - \pi_n \|_{TV} \leq \frac{1}{2} \left[ \lambda_{n, 2}^t - \frac{1 - \pi_n(i)}{-\pi_n(i)} \right]
\]
(2.13)
where \( \lambda_{n, 2} \) is the second largest* eigenvalue of \( P_n \) in absolute value, i.e., \( |\lambda_2| = \max_{m > 1} |\lambda_m| \).

2.5 MARKOV CHAIN MIXING

Given an ergodic MC, the Convergence Theorem entails the existence of a unique stationary measure \( \pi_n \) that represents the asymptotic distribution of \( X_n^t \) when \( t \) is large, ideally infinite. To have a more precise control on the convergence, the mixing time is typically introduced. The mixing time is the first time such that the distance from equilibrium drops below a fixed threshold \( \varepsilon \), i.e.,
\[
t_n^{\text{mix}}(\varepsilon) = \min \left\{ t \geq 0 : \| \mu_n^t - \pi_n \|_{TV} \leq \varepsilon \right\}
\]
(2.14)

Remark 2.6. The mixing time of a MC is a purely deterministic quantity which is not affected by the realisation of the MC whatsoever.

Alternative definitions of the mixing time can be obtained by using a different metric on the set of probability measures on \( \Omega_n \), but (2.14) is fairly the most common way to define the mixing time. There are, however, dissenting opinions on whether the TV-distance is the right distance to measure how far from equilibrium a MC is. The reason for that is the tendency of the TV-distance to be very ‘unforgiving’ even of small deviations from stationarity. An example may be useful.

Suppose that, after perfectly shuffling a deck of 52 cards, we happen to see the bottom card, say \( Q♠ \). The distribution of our deck is now no longer uniform over the set of 52! cards permutations, but over the set of 51! permutations having \( Q♠ \) at the bottom-most position. From (2.8) the distance between the uniform distribution and the biased one is
\[
\left| \left( \frac{1}{51!} - \frac{1}{52!} \right) 51! \right| = 1 - \frac{1}{52} \approx 0.98,
\]
which is quite close to 1, the distance from uniformity of a brand-new deck.

The mixing time is completely determined by the transition matrix \( P_n \), and by the initial distribution \( \mu_0^\delta \). In order to drop the latter dependence, the worst-case scenario is often considered†, i.e.,
\[
t_n^{\text{mix}}(\varepsilon) = \max_{\mu_0^\delta} \min \left\{ t \geq 0 : \| \mu_n^t - \pi_n \|_{TV} \leq \varepsilon \right\}
\]
(2.14a)

Accordingly, the worst-case mixing time can be related to the spectral properties of the transition matrix. For example, a direct consequence of (2.13) is the following bound on the mixing time:
\[
t_n^{\text{mix}}(\varepsilon) \leq \log \frac{1}{\pi \varepsilon} + \log \frac{1 - \min \pi_n(i)}{\min \pi_n(i)}
\]
\[
\log \frac{1}{\lambda_{n, 2}}.
\]
(2.15)

* The largest being \( 1 \) from Perron-Frobenius theorem.
† We note that maximising \( \| \mu_n^t - \pi_n \|_{TV} \) over the initial distribution is equivalent to setting \( \mu_0^\delta(i) = \delta_{i,j} \) (Kronecker’s delta) and maximising over \( j \), see [MT06].
Both (2.13) and (2.15) suggest that the closer is $\lambda_{n,2}$ to 1, the larger is the time needed to ensure convergence within a tolerance $\epsilon$. Let us define the \textit{spectral gap} as $\text{gap}_n = 1 - \lambda_{n,2}$ and the \textit{relaxation time} as $t_{\text{rel}}^n = \text{gap}_n^{-1}$. Then, the following lower bound holds in fact for ergodic reversible MCs:

$$t_{\text{mix}}^n(\epsilon) \geq (t_{\text{rel}}^n - 1) \log \frac{1}{\epsilon}.$$ 

Spectral methods for bounding the mixing time are highly relevant in the treatment of the cutoff phenomenon. In particular, for the class of BDCs there exists a complete characterisation of cutoff in terms of the asymptotic behaviour of the product $t_{\text{rel}} \cdot t_{\text{mix}}$, see [DLP10] and Section 2.8 below. However, the approach followed in this thesis does not focus on spectral methods. The interested reader is referred to [MT06, LPW06] for a comprehensive discussion of these techniques.

Another widely used method for bounding the total-variation distance from equilibrium is the \textit{Coupling Lemma}. A coupling of two MCs with the same transition probabilities but distinct initial distributions is a bivariate chain $(X^t_n, Y^t_n)$ such that the marginals of the joint transition probability $P_n(i, i', (j, j'))$ are $P_n(i, j)$ and $P_n(i', j')$, respectively. In other words, both components $X^t_n$ and $Y^t_n$ are MCs with transition matrix $P_n$, but the initial distribution of $X^0_n$ may differ in general from that of $Y^0_n$. Couplings are generally designed in such a way that the two components stay glued together after they have met for the first time, i.e.,

$$\text{if } X^t_n = Y^t_n \text{ then } X^s_n = Y^s_n \quad \forall \; t \geq s.$$  \hspace{1cm} (2.16)

In this way, if $Y^0_n$ is started according to the stationary distribution $\pi_n$ and the chains evolve together after they meet for the first time, the mixing time of $X^t_n$ is dominated by the coalescence time $\gamma = \min\{t : X^t_n = Y^t_n\}$. The Coupling Lemma states in fact that for a coupling satisfying (2.16), with starting positions $X^0_n = i$ and $Y^0_n \sim \pi_n$,

$$\left\| \mu^t_n - \pi_n \right\|_{\text{TV}} \leq \mathbb{P}(\gamma \geq t).$$ \hspace{1cm} (2.17)

We will make frequent use of couplings, for they allow to transform the problem of estimating a deterministic object like the TV-distance into the problem of estimating the coalescence time, $\gamma$. The interested reader is referred to [Lin02] or [Tho00] for a more detailed treatment of coupling methods for finite MCs.

2.6 THE CUTOFF PHENOMENON

Hereafter we consider families of finite ergodic discrete-time MCs, that is sextets of the form

$$\{\Omega_n, X^t_n, P_n, \pi_n, \mu^t_n, \mu^0_n\},$$

where $\Omega_n$ is the finite state space of the $n$-th chain $X^t_n$, which has transition matrix $P_n$ and unique stationary measure $\pi_n$. The symbols $\mu^0_n$ and $\mu^t_n$ stand for the initial distribution of the $n$-th chain and its probability distribution after $t$ steps; the time $t$ is a discrete quantity. For the sake of simplicity we will drop the ‘finite ergodic discrete-time’ specification and simply refer to them as families of MCs only:
Definition 2.1. A family of Markov chains is said to exhibit a total-variation cutoff if there exist two sequences of integers, \( \{a_n\} \) and \( \{b_n\} \) such that
\[
\frac{b_n}{a_n} \to 0 \quad (2.18)
\]
and
\[
\lim_{\theta \to \infty} \liminf_{n \to \infty} \|\mu_n^{a_n-\theta b_n} - \tau_n\|_{TV} = 1, \quad (2.19)
\]
\[
\lim_{\theta \to \infty} \limsup_{n \to \infty} \|\mu_n^{a_n+\theta b_n} - \tau_n\|_{TV} = 0. \quad (2.20)
\]
In this case \( a_n \) and \( b_n \) are called cutoff time and cutoff window, respectively.

Remark 2.7. If a family of MC exhibits an \( (a_n, b_n) \)-cutoff then it will also exhibit an \( (a_n, b'_n) \)-cutoff for every sequence of windows \( b'_n = O(b_n) \).

Remark 2.8. Definition 2.1 is specific for the TV-distance and was first introduced in [AD86]. It can easily be adapted to any other notion of distance \( \|\mu_n - \tau_n\| \), in this case the family is said to exhibit a \( \|\cdot\| \)-cutoff. However, an arbitrary distance could in principle be not bounded above by 1 or even be unbounded. For this reason a bit of care may be required in adapting the definition. For example, in the case of the \( \ell_2 \) distance equations (2.19) and (2.20) become
\[
\lim_{\theta \to \infty} \liminf_{n \to \infty} \|\mu_n^{a_n-\theta b_n} - \tau_n\|_2 = +\infty, \quad (2.19a)
\]
\[
\lim_{\theta \to \infty} \limsup_{n \to \infty} \|\mu_n^{a_n+\theta b_n} - \tau_n\|_2 = 0, \quad (2.20a)
\]
see [SC97] and [DSC06] for more details.

Equations (2.19) and (2.20) represent the sharp convergence to the equilibrium distribution in a narrow window of order \( b_n \) centred about the cutoff time \( a_n \). Figure 2 displays \( \|\mu_n - \tau_n\|_{TV} \) as a function of time for a biased random walk. A biased random walk is a BDC on the segment \( \Omega_n = \{0, 1, \ldots, n\} \) whose transition probabilities present a constant unbalance (bias) towards one of the extreme points of \( \Omega_n \). The transition probabilities are displayed by (3.1)–(3.2) in Section 3.1; with respect to Figure 2, the bias is \( \beta = 1/6 \). We clearly see that the system abruptly converges in a small window centred at \( a_n = n/2\beta \). The actual number of steps needed to achieve equilibrium is \( b_n = \Theta(\sqrt{n}) \), and the size of the window is negligible with respect to the length of the plateau.

The cutoff phenomenon is a crisp asymptotic picture of the mixing time of a family of MCs. The following bound is valid for every ergodic MC (see for example [LPW06]):
\[
\|\mu_n^{k\cdot t_{\text{mix}}^{\text{min}}(\epsilon)} - \tau_n\|_{TV} \leq (2\epsilon)^k. \quad (2.21)
\]

Take \( \epsilon = 1/4 \), then equation (2.21) states that in a time equal to \( t_{\text{mix}}^{\text{min}}(1/4) \) the chain is at a distance from equilibrium lower than 1/2. After that, it is sufficient to wait another \( t_{\text{mix}}^{\text{min}}(1/4) \) steps to see that distance reduced by a factor 2. Conversely, if cutoff is present then the time to go from distance 1/2 to distance 1/4 is proportional to the window size \( b_n \), an infinitesimal time lapse with respect to the number of steps already waited to reach distance 1/2. Thus, establishing cutoff for a given chain is much stronger a characterisation than providing any estimates of the mixing time. In [AD86], the first
The cutoff phenomenon appears in many natural examples ranging from cards shuffling models and random walks on the symmetric group \([BD92, DMP95, DS81, Hil92, Ros94, Por95]\) to statistical mechanics models \([Ald83, DS87, DGM90, LS10, LS13]\), list management problems \([DFP92]\) and many other. An excellent and exhaustive review is given in \([Dia96]\) by Persi Diaconis. We next discuss some examples.

2.7 Collecting and Shuffling

The literature about the cutoff phenomenon starts in 1981 with the paper by P. Diaconis and M. Shahshahani. In \([DS81]\) they investigated the convergence to uniformity of a shuffling method called random transpositions, a MC on the symmetric group \(S_n\). Given a set of \(n\) objects – cards in this case – the symmetric group is the set of all \(n!\) possible permutations of those objects. As for each random walk on \(S_n\), the equilibrium distribution is uniform \([LPW06]\). Performing the random transposition shuffle, the cards of a deck are initially displaced in a row. Two cards are then chosen uniformly at random and transposed – the cards may possibly coincide, in this case no transposition is made. These operations are repeated until the deck is shuffled.
For $c = c(t, n) = \frac{t-1/2 \log n}{n}$, P. Diaconis and M. Shahshahani proved that

$$\exists b \in \mathbb{R} \text{ s.t. for } n \geq 10, c > 0, \quad \|\mu_n^t - \pi_n\|_\text{TV} \leq b e^{-2c},$$

(2.22)

$$\forall t, \|\mu_n^t - \pi_n\|_\text{TV} \geq 2 \left( \frac{1}{e} - e^{-e^{-2c}} \right) + o(1) \quad \text{as } n \to \infty.$$  

(2.23)

If the upper bound (2.22) is evaluated in $t = \frac{1}{2} n \log n + \theta n$ then (2.20) is obtained.

The original proof of (2.22)–(2.20) is rather technical and uses the tools of group representations. An alternative and much easier approach is via strong stationary times, see [LPW06]. Stationary times are a special kind of stopping times and play a key role in many proofs of cutoff. The first use of stopping times in establishing cutoff was made by D. Aldous and P. Diaconis, in [AD86] they proved cutoff for the top-in-at-random shuffle. Incidentally, that was also the first paper in which the name ‘cutoff’ was used.

**Definition 2.2.** A non-negative discrete random variable $\tau_n$ is a stopping time for the MC $X^t_n$ if the event $\{\tau_n = s\}$ depends on the trajectory of $X^t_n$ only up to time $s$. In other words, the indicator function $\mathbb{I}_{\{\tau_n = s\}}$ is a function of $X^0_n, X^1_n, \ldots, X^s_n$ only.

**Definition 2.3.** A stopping time $\tau_n$ is called a stationary time if $X^t_n$ evaluated at time $\tau_n$ is at equilibrium. In formulas, $\tau_n$ is a stationary time if

$$\mathbb{P}(X^t_n = i, \tau_n = t) = \pi_n(i).$$

**Definition 2.4.** A strong stationary time is a stationary time with the additional property of $X^t_n$ being independent of $\tau_n$, i.e.,

$$\mathbb{P}(X^t_n = i | \tau_n = t) = \pi_n(i).$$

The top-in-at-random shuffle is performed by repeatedly inserting the topmost card back in a position of the deck picked out uniformly at random, see Figure 3. Also the MC that models the top-in-at-random is a random walk on the symmetric group, so its equilibrium distribution is uniform. For the top-in-at-random shuffle, D. Aldous and P. Diaconis proved that

$$\forall \theta \geq 0, n \geq 2, \quad \|\mu_n \log n + \theta_n - \pi_n\|_\text{TV} \leq e^{-\theta},$$

$$\forall \theta_n \to \infty, \quad \|\mu_n \log n - \theta_n n - \pi_n\|_\text{TV} \to 1.$$  

(2.24)
The key to (2.24) is the following Lemma:

**Lemma 2.1** Let $X_n^t$ be a MC with state space $\Omega_n$. Let $\tau_n$ be a strong stationary time for $X_n^t$. Then, $\forall t \geq 0$,

$$\|\mu_n^t - \pi_n\|_\infty \leq \mathbb{P}(\tau_n > t).$$

**Proof.** For any $A \subseteq \Omega_n$,

$$\mu_n^t(A) = \mathbb{P}(X_n^t \in A),$$

$$= \sum_{s \leq t} \mathbb{P}(X_n^t \in A, \tau_n = s) + \mathbb{P}(X_n^t \in A, \tau_n > t),$$

$$= \sum_{s \leq t} \pi_n(A) \mathbb{P}(\tau_n = s) + \mathbb{P}(X_n^t \in A | \tau_n > t) \mathbb{P}(\tau_n > t),$$

$$= \pi_n(A) + \mathbb{P}(X_n^t \in A | \tau_n > t) - \pi_n(A) \mathbb{P}(\tau_n > t),$$

which yields $|\mu_n^t(A) - \pi_n(A)| \leq \mathbb{P}(\tau_n > t).$ \hfill \Box

A strong stationary time for the top-in-at-random model is the integer following the first time the original bottom card reaches the topmost position. Let us imagine to perform the top-in-at-random shuffle until $T_n^1$, the first time when $j$ cards have been re-inserted into the deck below the card that at time $t = 0$ was the bottommost one. Clearly, the $j$ cards are equally distributed due to the randomness of the inserting position. The number of steps between the first moment $j - 1$ cards are below the original bottom card and the first moment $j$ cards are below it is $T_n^1 - T_n^{j-1}$, a geometric random variable. In formulas,

$$\mathbb{P}(T_n^1 - T_n^{j-1} = t) = \frac{j}{n} \left(1 - \frac{j}{n}\right)^{t-1},$$

$$T_n^0 = 0.$$  \hfill (2.25)

At time $T_n^{n-1}$, when $n - 1$ cards have been placed below the original bottom card, the latter has reached the topmost position of the deck, and the $n - 1$ cards below it are uniformly permuted. Performing another shuffle, the system eventually loses memory of the starting position and reaches uniformity. Thus, a strong stationary time for the top-in-at-random is $\tau_n^{\text{top}} = T_n^{n-1} + 1$. To obtain (2.24) is sufficient to show that for $t = n \log n + \theta n$,

$$\mathbb{P}(\tau_n^{\text{top}} > t) \leq e^{-\theta}.$$  \hfill (2.26)

Inequality (2.26) is easily obtained via the coupon collector’s problem, as we next explain.

The coupon collector draws with equal probability from a set of $n$ different coupons, and the drawn coupons are immediately replaced. The collector wins as soon as he draws all the $n$ different coupons. The question typically asked by the collector, especially if he pays a fee for each draw he makes, is ‘how many draws do I need to win?’ The answer is rather assertive due to the cutoff phenomenon.

The coupon collector’s model is a BDC on the segment $\{0, 1, \ldots, n\}$, $X_n^t = i$ meaning that at time $t$ the collector is still missing $i$ coupons. As soon as $X_n^t = 0$ the collector wins, so we define

$$\tau_n^{cc} = \min\{t \geq 0 : X_n^t = 0\}.$$
The transition rates for the coupon collector’s chain are

$$P \left( X^n_t = j \mid X^n_{t-1} = i \right) = \begin{cases} \frac{i}{n}, & \text{if } j = i - 1, \\ 1 - \frac{i}{n}, & \text{if } j = i, \\ 0, & \text{otherwise}. \end{cases}$$

The coupon collector’s chain is a biased random walk, in the sense that the probability to go left (from $i$ to $j < i$) is larger than the probability to go right (from $i$ to $j > i$). More precisely, it is the extreme case of the biased random walk we have already discussed*, the probability to go right is in fact null.

For the coupon collector’s model, the number of steps $S^n_i$ before a move takes place from state $i$ to $i - 1$ represents the number of draws between two successful extractions of missing coupons. The random variable $S^n_i$ is geometrically distributed, i.e.,

$$P \left( S^n_i = s \right) = \frac{i^n}{(1 - \frac{i}{n})^s - 1}. \quad (2.27)$$

According to (2.25) and (2.27), the random variables $T^n_i - T^n_{i-1}$ and $S^n_i$ are identically distributed and so are the random variables $\tau^{\text{top}}_n = 1 + \sum_{i=1}^{n-1} \left( T^n_i - T^n_{i-1} \right)$ and $\tau^{\text{cc}}_n = 1 + \sum_{i=1}^{n-1} S^n_i$. Therefore,

$$P \left( \tau^{\text{top}}_n > 1 \right) = P \left( \tau^{\text{cc}}_n > t \right). \quad (2.28)$$

Now, if we call $A_j$ the event ‘the $j$-th coupon is not drawn in the first $t$ trials’ then

$$P \left( \tau^{\text{cc}}_n > t \right) = P \left( \bigcup_{j=1}^{n} A_j \right) \leq \sum_{j=1}^{n} P \left( A_j \right),$$

$$= n \left( 1 - \frac{1}{n} \right)^t \leq n e^{-t/n}, \quad (2.29)$$

which gives (2.26) for $t = n \log n + \theta n$.

Provided the total number of coupons is large, the coupon collector knows hitherto that $n \log n + cn$ draws will be enough to win. Can he expect to win having drawn much less than $n \log n$ coupons? No, he definitely can not. According to Cantelli’s inequality† the tail probabilities of a real random variable $X$ with finite mean $\mu$ and variance $\sigma^2$ can be estimated as follows:

$$P \left( X - \mu \geq \theta \sigma \right) \leq \frac{1}{1 + \theta^2},$$

$$P \left( \mu - X \geq \theta \sigma \right) \leq \frac{1}{1 + \theta^2}. \quad (2.30)$$

The time to win, $\tau^{\text{cc}}_n$, has mean $n \log n$ and variance $O(n^2)$, so that (2.30) infers

$$P \left( \tau^{\text{cc}}_n \geq n \log n - \theta n \right) \leq \frac{1}{1 + \theta^2}.$$

The stationary distribution for the coupon collector’s chain is a mass concentrated in state 0, $\pi_n(0) = \delta_{0,0}$. This means that $\tau^{\text{cc}}_n$ is a stationary time

* Cf. Figure 2 on page 12.
† Named after Italian mathematician Francesco Paolo Cantelli, inequality (2.30) is the one-sided version of Chebyshev’s inequality. A proof of this inequality can be found in [Fel68b].
for the coupon collector because when the chain reaches state 0, only self-transitions are available (absorbing state). Further, by means of (2.8),
\[
\|\mu_n^t - \pi_n\|_{TV} = 1 - \mathbb{P}(X_n^t = 0),
\]
\[
= 1 - \mathbb{P}(\tau_n^c < t).
\]

If \( t = n \log n - \theta n \), then (2.31) yields
\[
\|\mu_n^{n \log n - \theta n} - \pi_n\|_{TV} \geq 1 - \frac{1}{1 + \theta^2},
\]
which is equivalent to (2.20) for \( a_n = n \log n \) and \( b_n = \Theta(n) \). Lemma 2.1 applied to the stationary time \( \tau_n^c \) and (2.28)–(2.29) (or Cantelli again) leads to (2.19) and a \((a_n, b_n)\)-cutoff.

## 2.8 Cutoff for Birth-and-Death Chains

In Section 2.5 we reviewed some methods for bounding the distance of the evolved measure \( \mu_n^t \) of an ergodic MC from the unique stationary distribution \( \pi_n \). For our purposes we can sort those method in two distinct classes, namely, spectral methods and random times methods. Random times comprehend coalescence and stopping/stationary times. Section 2.7 explained how the existence of a concentrated strong stationary time can be exploited to prove cutoff in some classical examples. On the other hand, the evolution of a MC and its mixing properties do depend only on \( P_n \), as we already noted at the end of Section 2.1 and in Remark 2.6. A question naturally arising is whether, with respect to the cutoff phenomenon, the spectral properties of \( P_n \) and the presence of random times particularly meaningful for the evolution of the chain are in fact two sides of the same coin. Strongly expected and sought for, so far a positive answer could be found only in the class of Birth-and-Death Chains (BDCs).

The state space \( \Omega_n \) of a BDC can be put in a one-to-one correspondence with the segment \( \{0, 1, \ldots, |\Omega_n| - 1\} \). Only transitions to nearest-neighbour are allowed, that is \( P_n(i, j) = 0 \) if \( |i - j| > 1 \). It is a common habit to indicate the non-zero transition probabilities with the following symbols:

\[
p_i = P_n(i, i + 1),
\]
\[
r_i = P_n(i, i),
\]
\[
q_i = P_n(i, i - 1).
\]

If \( p_i, q_i > 0 \) \( \forall i = 0, 1, \ldots, |\Omega_n| - 1 \) and \( r_i \neq 0 \) for at least a state \( i \in \Omega_n \) then a BDC is ergodic, and the stationary distribution can be written as
\[
\pi_n(i) = \pi_n(0) \prod_{k=1}^{i} \frac{p_{k-1}}{q_k}.
\]

In [Pero4] Yuval Peres conjectured that in many natural chains ‘cutoff occurs if and only if \( \tau_n^{\text{rel}} = o(\min(\{1/4\})) \)’ asymptotically, for the parameter \( n \to \infty \). This conjecture was proved true within the class of BDCs both for cutoff in separation and TV-distance in [DSC06] and [DLP10], respectively.

In order to better understand the proof of Yuval Peres’s claim and how it links the spectral properties of the transition matrix to the behaviour of the distance from stationarity, let us start surveying the proof of cutoff for the coupon collector’s chain. Carried out at the end of Section 2.7, it is based on
two facts, namely, the stationary distribution is a mass in state 0 and there exists a stationary time, $\tau_n^c$, which is concentrated in the sense that

$$\frac{\sigma^2[\tau_n^c]}{\mathbb{E}[\tau_n^c]} \xrightarrow{n \to \infty} 0. \quad (2.33)$$

For the examples presented in Section 2.7 the stationary time $\tau_n^c$ is also the hitting time of state 0. Hitting times, defined below, play a fundamental role in the methodology to be developed in Chapter 3.

**Definition 2.5.** The hitting time of a set $A \subset \Omega_n$ is defined as

$$\tau_n(A) = \min \{ t \geq 0 : X_n^t \in A \} .$$

**Remark 2.9.** Hitting times are stopping times.

With respect to Definition 2.5, the proof of cutoff for the coupon collector can then be sketched as follows:

1. The stationary measure $\pi_n$ is concentrated in the subset $A^* = \{ 0 \}$. As a consequence, equation (2.8) yields to

$$|\mu_n^1 - \pi_n|_\nu = \max_{A \subseteq \Omega_n} \{ \pi_n(A) - \mu_n^1(A) \},$$

$$\geq \pi_n(A^*) - \mathbb{P}(X_n^t \in A^*),$$

$$= 1 - \mathbb{P}(X_n^t \in A^*),$$

$$\geq 1 - \mathbb{P}(\tau_n(A^*) \leq t) ;$$

2. The hitting time of 0 is a stationary time for the chain. Lemma 2.1 then provides the bound

$$|\mu_n^1 - \pi_n|_\nu \leq \mathbb{P}(\tau(A^*) > t) ;$$

3. The hitting time $\tau_n(A^*)$ is concentrated as in (2.33), so if we take $a_n = \mathbb{E}[\tau_n(A^*)]$ and $b_n = \sigma[\tau_n(A^*)]$ then the probability of both the events $\{ \tau_n(A^*) \geq a_n - \theta b_n \}$ and $\{ \tau_n(A^*) \leq a_n + \theta b_n \}$ is asymptotically small in $n$ and $\theta$ by (2.30).

**Remark 2.10.** Having in mind a possible generalisation of this structure to other-than-BDCs, the most critical point is 2. Indeed, the stationary measure of a BDC, as well as the first two moments of any hitting time, depend only on the transition rates $p_i$ and $q_i$. The last statement is justified by (2.32) and equations (2.43)–(2.46) below.

In [DSCo6] P. Diaconis and L. Saloff-Coste remarked that if $D(\cdot, \cdot)$ is either separation or TV-distance then there exists a sequence of random times $T_n^D$ such that

$$D(\mu_n^1, \pi_n) = \mathbb{P}(T_n^D > t) . \quad (2.34)$$

If $D$ is total-variation then the elements of the family $T_n^D$ are understood to be ‘optimal coupling times’ (optimal in the sense that (2.17) is satisfied as an equality), while if $D$ is separation then the $T_n^D$’s are ‘optimal strong stationary times’. They also stressed that once mean and variance of $T_n^D$ are computed, the distance from stationarity can be easily bounded by means of Cantelli’s inequality (2.30). In the case of separation $T_n^D$ is the first time an auxiliary chain, named strong stationary dual, hits the state which is furthest

* In the sense of (2.3).
with respect to the starting state, conventionally taken to be $X^0_n = 0$. The dual chain is still a BDC and keeps the eigenvalues of the original chain, see [DF90]. Then, using the first passage time\(^*\) distribution in terms of the spectral properties of $P_n$ provided by [Kei79, BS87], P. Diaconis and L. Saloff-Coste obtained

$$
\mathbb{E} \left[ T^D_n \right] = \sum_{i=2}^{\Omega_n} \frac{1}{\lambda_i},
$$

$$
\sigma^2 \left[ T^D_n \right] = \sum_{i=2}^{\Omega_n} \left( 1 - \lambda_i \right) \frac{1}{\lambda_i^2},
$$

where $\{\lambda_i\}_{2 \leq i \leq \Omega_n}$ are the eigenvalues of $P_n$ different from 1. The variance of $T^D_n$ can be easily bounded by

$$
\sigma^2 \left[ T^D_n \right] \leq \tau^2 \left[ T^D_n \right],
$$

(2.35)

or alternatively as

$$
\sigma^2 \left[ T^D_n \right] \leq \mathbb{E}^2 \left[ T^D_n \right].
$$

(2.36)

Via (2.30) and (2.34) it is immediate to show that

$$
\text{sep} \left( \mu_n^{1 - \theta \mathbb{E} [ T^D_n ]}, \pi_n \right) \geq 1 - \frac{1}{1 + \theta^2},
$$

(2.37)

$$
\text{sep} \left( \mu_n^{1 + \theta \mathbb{E} [ T^D_n ]}, \pi_n \right) \leq \frac{1}{1 + \theta^2},
$$

(2.38)

that is to say, a cutoff with $a_n = \mathbb{E} \left[ T^D_n \right]$ and $b_n = \Theta \left( \mathbb{E} \left[ T^D_n \right] \right)$, provided $\mathbb{E}^{\tau^2} = o \left( \mathbb{E} \left[ T^D_n \right] \right)$. Conversely, suppose the chain exhibits an $(a'_n, b'_n)$-cutoff with cutoff time $a'_n \rightarrow \infty$. The separation distance can be alternatively bounded by (2.30), (2.34), and (2.35). For all $\theta > 0$,

$$
\text{sep} \left( \mu_n^{1 - \theta \mathbb{E} [ T^D_n ]}, \pi_n \right) \geq 1 - \frac{1}{1 + \theta^2} \text{gap} \mathbb{E} \left[ T^D_n \right],
$$

(2.39)

$$
\text{sep} \left( \mu_n^{1 + \theta \mathbb{E} [ T^D_n ]}, \pi_n \right) \leq \frac{1}{1 + \theta^2} \text{gap} \mathbb{E} \left[ T^D_n \right].
$$

(2.40)

Inequalities (2.39) and (2.40) could be interpreted as a cutoff with cutoff window proportional to the cutoff time. This means that the length of the time intervals $[a_n, a_n + \theta b_n]$ and $[(1 - \theta) \mathbb{E} [ T^D_n ], (1 + \theta) \mathbb{E} [ T^D_n ]]$ must be of the same order for $n$ large. In other words, $a_n = \mathbb{E} \left[ T^D_n \right)$. Inequalities (2.36)-(2.40) now infer that $\text{gap} \cdot \mathbb{E} \left[ T^D_n \right] \rightarrow \infty$. The proof of Yuval Peres's conjecture for the separation case is concluded by showing that $\mathbb{E} \left[ T^D_n \right]$ is proportional to the mixing time and $\text{gap} \cdot \mathbb{E} \left[ T^D_n \right] \rightarrow \infty$ if and only if $\text{gap} \cdot t_n^{\text{mix}} (\varepsilon) \rightarrow \infty$.

In [DLP10] the conjecture is again proved true in the class of BDCs but for TV-distance. Here the authors used the following equivalent definition of cutoff:

**Definition 2.6.** A family of ergodic MCs is said to exhibit cutoff if

$$
\lim_{n \rightarrow \infty} \frac{t_n^{\text{mix}} (\varepsilon)}{\tau_n (1 - \varepsilon)} = 1 \quad \forall \varepsilon \in (0, 1/2).
$$

The structure of the proof is similar to the one we have outlined on page 17, the main difference being that point 2 is replaced by

$$
\| \mu_n - \pi_n \|_\text{TV} \leq \mathbb{P} \left( \max \{ \tau_n (Q(\varepsilon)), \tau_n (Q(1 - \varepsilon)) \} > t \right) + 2 \varepsilon,
$$

\(^*\) Another name for hitting time.
where \( \tau_n(Q(\varepsilon)) \) is the hitting time (starting from state 0) of the \( \varepsilon \)-quantile of the stationary distribution, i.e.,

\[
Q(\varepsilon) = \min \left\{ k \geq 0 : \sum_{i=0}^{k} \pi_n(i) \geq \varepsilon \right\}.
\]

Then, the first passage time distribution provided by [Kei79, BS87] is used as in [DSC06] to obtain the following bound of \( \tau_n(Q(\varepsilon)) \):

\[
\sigma^2[\tau_n(Q(\varepsilon))] \leq \frac{\mathbb{E}[\tau_n(Q(\varepsilon))] - \varepsilon \text{ gap}_n}{\varepsilon \text{ gap}_n}.
\]

Similarly to what we have seen for \( T_n^D \), the hitting time \( \tau_n(Q(\varepsilon)) \) is concentrated if the product \( \text{gap}_n \cdot \mathbb{E}[\tau_n(Q(\varepsilon))] \) diverges. It turns out that again this condition is also necessary for cutoff, and \( \mathbb{E}[\tau_n(Q(\varepsilon))] \) is proportional to the mixing time. Thus, Yuval Peres’s conjecture is proved.

A radically different approach to the cutoff phenomenon for BDCs is the one proposed in [MY04, BB08]:

**Definition 2.7.** A family of chains is said to exhibit cutoff at mean times if there exists a sequence of random times \( T_n \) such that

\[
\frac{T_n}{\mathbb{E}[T_n]} \xrightarrow{\text{Prob}} 1.
\]

This alternative definition is adopted for several reasons. It allows to release the phenomenon from the notion of distance in use, the TV-distance, in particular, can greatly penalise the vision of the approach to stationarity*. Definition 2.7 does not require the knowledge of the stationary distribution and let us look at the cutoff phenomenon as a physical phenomenon rather than as a purely probabilistic one. Moreover, it gives the possibility to set up a common framework to study and characterise cutoff and metastability at the same time. In [BB05] the authors indeed derive some sufficient conditions for both cutoff phenomenon (in the sense of (2.42)) and exponential escape (a fundamental feature of metastable behaviour) to arise. Last but not least, if we compare the proof of cutoff for the coupon collector’s chain and the characterisation of cutoff for BDCs both in separation and TV-distance, then we can fairly say that (2.42) captures the soul of the phenomenon. Condition (2.42) is in fact inferred by a concentration property like \( \sigma[T_n] = o(\mathbb{E}[T_n]) \), a cornerstone of all the discussions so far.

In the analysis carried out in [BB08] a key role is played by formulas for the first and second moment of the hitting times of any state. We introduce these formulas here for future reference. Let us suppose that a BDC has state space \( \Omega_n = \{0, 1, \ldots, n\} \) and let \( T_{i \to j} \) be the hitting time of \( j \) starting from \( i \), that is,

\[
T_{i \to j} = \min\{t \geq 0 : X_n^t = j, X_n^0 = i\}.
\]

For \( 0 \leq j < i \leq n \),

\[
\mathbb{E}[T_{i \to j}] = \sum_{k=j+1}^{i} \frac{1}{q_k} \sum_{m=k}^{n} \pi_n(m) \pi_n(k),
\]

\[
\mathbb{E}[T_{i \to j}^2] = \sum_{k=j+1}^{i} \frac{2}{q_k} \sum_{m=k}^{n} \mathbb{E}[T_{m \to j}] \frac{\pi_n(m)}{\pi_n(k)} \mathbb{E}[T_{i \to j}] - \mathbb{E}[T_{i \to j}],
\]

* See the discussion on page 9 on the 'unforgiving' nature of TV-distance.
We end this chapter with the riffle shuffle, also known as dovetail shuffle, which gives a mathematical foundation to the well-known statement 'seven shuffles are enough'.

A proof of formulas (2.43)–(2.46) is found in [Fel68a, BBF09]. From (2.43) and (2.44) the following formula can be easily obtained for the variance of \( T_{i \rightarrow j} \) when \( 0 \leq i < j \leq n \):

\[
\sigma^2 [T_{i \rightarrow j}] = \sum_{k=1}^{i} \sigma^2 [T_{k \rightarrow k-1}]
\]

\[
= \sum_{k=j+1}^{i} \frac{1}{4k} \sum_{m=k}^{n} \left( 2E[T_{m \rightarrow k-1}] - E[T_{k \rightarrow k-1}] \right) \frac{\pi_n(m)}{\pi_n(k)} - E[T_{i \rightarrow j}] .
\]

An analogous formula holds if \( 0 \leq i < j \leq n \).

2.9 Seven Shuffles are Enough

We end this chapter with the riffle shuffle, also known as dovetail shuffle. Experiments in [Dia88] show that a good mathematical model for this shuffling technique is the following. A deck of \( n \) cards is split in two according to a binomial distribution \( \mathcal{B}(n, \frac{1}{2}) \), then the two parts are riffled together in such a way that the next card drops from one of the two heaps with probability proportional to the number of cards still present in each heap, see [Gil55] for more details. The mathematical analysis of riffle shuffle is presented in [BD92], the seminal paper by D. Bayer and P. Diaconis. Apparently, the riffle shuffle is the first example that comes to the mind when speaking of cutoff. A possible explanation for that is equation (2.48) below, which gives a mathematical foundation to the well-known statement 'seven shuffles are enough'.

The analysis of the riffle shuffle is based on the concept of rising sequences. A rising sequence is a maximal subset of an arrangement of cards, consisting of successive face values displayed in order. Given the permutation \( A \ 4 \ 5 \ 2 \ 6 \ 3 \ 7 \) we recognise two interleaved rising sequences, namely, \( A \ 2 \ 3 \) and \( 4 \ 5 \ 6 \ 7 \). As pointed out in [BD92], rising sequences do not intersect, thus any arrangement of cards is the union of its rising sequences. Figure 4 shows a single iteration of riffle shuffle. The deck, initially arranged in ascending order, is cut in two parts following a binomial rule, that is, the probability of one of the two packets having \( k \) cards is \( \binom{n}{k} \frac{1}{2^k} \). This first stage selects two rising sequences, namely, \( 1, 2, \ldots, k \) and \( k + 1, k + 2, \ldots, n \). Then, the two packets are riffled together in such a way that, if the two heaps have respectively \( A \) and \( B \) cards, the next card will fall from the first heap with probability \( \frac{A}{A+B} \). While the riffling takes place, the cards coming from the two heaps keep their relative order so that the final arrangement is just the interleaving of the mentioned sequences. If successive shuffles are performed then the number of rising sequences in the deck will initially tend to double. Conversely, the length of any sequence will tend to decrease exponentially fast.

In their delightful paper, D. Bayer and P. Diaconis proved that if a deck of \( n \) cards (originally in ascending ordered) is given a sequence of \( t \) rif-
seven shuffles are enough

A 2 3 4 5 6 7 8 9 10 J Q K

A 2 3 4 5 6 7 8 9 10 J Q K

A 8 2 3 4 5 6 10 J Q K 7

A 8 2 3 9 4 10 J 5 Q K 6 K 7

Figure 4: Riffle shuffling and rising sequences. (a) Initial deck arrangement, only one rising sequence is present; (b) The deck is divided in two parts according to a binomial distribution; (c) The packets are riffled together; (d) New deck arrangement, two rising sequences are now present.

The riffle shuffle model exhibits cutoff

2.9 Seven Shuffles Are Enough

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff

The riffle shuffle model exhibits cutoff
The stationary distribution is hence no longer uniform, instead it can be asymptotically treated as a normal. On the other hand, the evolved measure $\mu_t (\rho)$ shows a sort of energy-entropy conflict. $Q^t (r)$ gives in fact the largest weight to the initial permutation $1, 2, \ldots, n$ even for those values of $t$ for which stationarity is achieved, see Figures 6h–6j. Nonetheless, the configurations with a number of rising sequences more than $O(\sqrt{n})$ away from $n/2$ are greatly penalised. As a result, after $\log_2 n^{3/2}$ steps the evolved measure is nearly uniform in the bulk of (2.49). The bulk is the family of subsets

$$A_{n,0} = \left\{ r : \left| r - \frac{n}{2} \right| \leq \theta \sqrt{ \frac{n}{12} } \right\}.$$ 

Since the stationary measure is concentrated on $A_{n,0}$, the difference between the evolved and the stationary measure in $A_{n,0}$ can be neglected.

It is very difficult to exploit an approach to cutoff like the one illustrated in Section 2.8 in the case of riffle shuffle. Although the stationary measure present a concentration property in the sense discussed above, it is rather hard to compute the hitting time of the region $A_{n,0}$. Among the technical reasons, the complete expression of the transition matrix for the riffle shuffle model is not known in literature [BD92]. However, Figure 5 gives strong evidence that the picture we have started building since Section 2.8 is quite general. Indeed, the convergence is triggered in the first time the number of rising sequences hits the set $A_{n,0}$. 

Figure 5: Riffle shuffling of a deck of 52 cards. The chart displays the TV-distance from stationarity computed from (2.48) and a simulation of the number of rising sequences.
2.9 Seven Shuffles Are Enough

Figure 6: $Q^t(r)$ for a regular deck of 52 cards, $t = 1, \ldots, 10$. 
Figure 7: (cont.) $Q^1(t)$ for a regular deck of 52 cards, $t = 1, \ldots, 10$. 
2.9 Seven Shuffles Are Enough

Figure 8: (cont.) $Q^t(r)$ for a regular deck of 52 cards, $t = 1, \ldots, 10$. 
Figure 9: (cont.) $Q^t(r)$ for a regular deck of 52 cards, $t = 1, \ldots, 10$. 

(g) $Q^7(r)$ after 7 shuffles.

(h) $Q^8(r)$ after 8 shuffles.
2.9 Seven Shuffles Are Enough

Figure 10: (cont.) \(Q^t(r)\) for a regular deck of 52 cards, \(t = 1, \ldots, 10\).
In Chapter 2 we have introduced the cutoff phenomenon and presented the fundamental results that characterise it for the class of Birth-and-Death Chains (BDCs). We have also outlined a common thread to all chains exhibiting cutoff behaviour, i.e., the existence of a stopping time that encapsulates all the features of the approach to stationarity. When such a stopping time is a strong stationary time, a concentration property of the form (2.33) or (2.42) easily gives cutoff, and the proof, mutatis mutandis, is the same as for the coupon collector’s chain, Section 2.7. In Remark 2.10 we have stressed that the upper bound to the TV-distance is the most critical point to deal with for a generalisation of the method. We begin this chapter, looking for a possible way to overcome that point, from the biased random walk. This is a model that has much in common with the coupon collector’s chain from the point of view of the description of cutoff that we are building, yet it does not present any evident strong stationary time.

### 3.1 The Drift Triggers Cutoff: Biased Random Walk on a Segment

The biased random walk on a segment is a BDC on the state space \( \Omega_n = \{0, 1, \ldots, n\} \). For \( i \in \{1, 2, \ldots, n-1\} \), the transition probabilities are

\[
\begin{align*}
P_n(i, i+1) &= p_i = \frac{1}{3} - \beta, \\
P_n(i, i) &= r_i = \frac{1}{3}, \\
P_n(i, i-1) &= q_i = \frac{1}{3} + \beta,
\end{align*}
\]

(3.1)

whereas at the extreme points 0 and n,

\[
\begin{align*}
P_n(0, 1) &= p_0 = \frac{1}{3} - \beta, \\
P_n(0, 0) &= r_0 = \frac{2}{3} + \beta, \\
P_n(n, n-1) &= q_n = \frac{1}{3} + \beta, \\
P_n(n, n) &= r_n = \frac{2}{3} - \beta.
\end{align*}
\]

(3.2)

The parameter \( \beta \in (0, 1/3) \) is called bias, or drift. Since \( \beta > 0 \), the chain defined by (3.1) and (3.2) is more likely to take transitions to the left rather than to the right. The imbalance in the transition probabilities does not present any spatial dependency, so we say that the chain has a constant drift to the left. According to (2.32),

\[
\pi_n(i) = \pi_n(0) \left( \frac{1/3 - \beta}{1/3 + \beta} \right)^i,
\]

where, by normalisation,

\[
\pi_n(0) = \frac{2\beta}{1/3 + \beta}.
\]

The presence of a drift positively affects the stochastic stability of each Markov Chain (MC) \([MT93]\). This is especially true in BDCs, due to formulas (2.32) and (2.43)–(2.47). Figure 11 shows how the evolved measure \( \mu_n^t \)
approaches $\pi_n$ for increasing values of the bias, including the case $\beta = 0$. When $\beta = 0$ we have a uniform random walk because the probability to go left equals the probability to go right, and the chain has a diffusive behaviour. In diffusive processes the distance from stationarity decreases exponentially*, but not abruptly. In other words, diffusion means no cutoff. Conversely, for every drift $\beta$ larger than 0, the diffusion behaviour is destroyed and a cutoff-like curve arises. In Figure 11, though, the smallest value of $\beta$ seems to be problematic, for the convergence does not look abrupt at all. However, we must not forget that cutoff is an asymptotic feature. Increasing the the size of the state space, Figure 12 shows in fact the correct curve for the formerly problematic value of the bias. Both Figures 11 and 12 are drawn starting from the initial distribution $\mu_0^n(i) = \delta_{i,0}$.

The loss of diffusive behaviour for any finite positive value of $\beta$ can be intuitively explained as follows. As soon as $\beta > 0$, the stationary measure $\pi_n$ is concentrated around the state 0. This means that the bulk of $\pi_n$ is a finite region containing the state 0. Up to the time of $X_t^n$ having a reasonable chance to visit the bulk, the overlap between $\mu^n_t$ and $\pi_n$ is negligible. Therefore, the TV-distance from equilibrium is nearly 1 from Remark 2.4. As we show later, the hitting time of the bulk is quasi-deterministic, in the sense that its standard deviation is of smaller order with respect to the mean. Thus, the statement ‘up to the time of the chain $X_t^n$ having a reasonable chance to visit the bulk’ can be given a sharp probabilistic interpretation via Cantelli’s inequality (2.30). This remark is clarified by the following proof of the cutoff phenomenon for the biased random walk.

Let us take $X_0^n = n$ and consider $A_\theta = \{0, 1, \ldots, \theta - 1\}$, a finite segment of length $\theta$. For $\beta > 0$, the measure of this subset of the state space is

$$\pi_n(A_\theta) = \pi_n(0) \sum_{i=0}^{\theta-1} \left( \frac{1/3 - \beta}{1/3 + \beta} \right)^i = 1 - \left( \frac{1/3 - \beta}{1/3 + \beta} \right)^\theta.$$ 

* See (2.21) on page 11 and the subsequent comment.
Then,
\[
\|\mu^t - \pi_n\|_{tv} = \max_{A \subset \Omega_n} \{\pi_n(A) - \mu^t_n(A)\},
\]
\[
\geq \pi_n(A_{\theta}) - \mu^t_n(A_{\theta}),
\]
\[
= 1 - \left(\frac{1/3 - \beta}{1/3 + \beta}\right)^t - \mathbb{P}(X_n^t \in A_{\theta}),
\]
\[
\geq 1 - \left(\frac{1/3 - \beta}{1/3 + \beta}\right)^t - \mathbb{P}(\tau_n(A_{\theta}) < t),
\]  
where \(\tau_n(A_{\theta})\) is the hitting time of \(A_{\theta}\), that is,
\[
\tau_n(A_{\theta}) = \min\{t \geq 0 : X_n^t \in A_{\theta}\}.
\]

Inequality (3.6) is obtained from (3.5) by noticing that the event \(\{X_n^t \in A_{\theta}\}\) infers the event \(\{\tau_n(A_{\theta}) < t\}\). Suppose now the chain is started with initial measure \(\mu_n^0(i) = \delta_{i_0},\) that is \(X_0^t = 0\). Then, by (2.43) with \(i = n\) and \(j = \theta,\)
\[
\mathbb{E}[\tau_n(A_{\theta})] = \sum_{k=0}^{n} \frac{1}{q^k} \sum_{m=k}^{n} \pi_n(m) \pi_n[k],
\]
\[
= \sum_{k=0}^{n} \frac{1}{1/3 + \beta} \sum_{m=0}^{n-k} \left(\frac{1/3 - \beta}{1/3 + \beta}\right)^m,
\]
\[
= \sum_{k=0}^{n} \frac{1 - \left(\frac{1/3 - \beta}{1/3 + \beta}\right)^{n-k+1}}{2\beta},
\]
which leads to the estimates
\[
\mathbb{E}[\tau_n(A_{\theta})] \leq \frac{n - \theta}{2\beta},
\]
\[
\mathbb{E}[\tau_n(A_{\theta})] \geq \frac{n - \theta}{2\beta} - \frac{1/3 + \beta}{4\beta^2}.
\]

Figure 12: Biased random walk on a segment of size 100,000.
Thus, for \( n \) and \( \theta \) sufficiently large, the leading term of the expected value of \( \tau_n(A_0) \) is \( \frac{n^2 - \theta}{2 \beta} \). Formulas (2.47), (3.7), and (3.8) yield the following estimate for the variance

\[
\sigma^2 [\tau_n(A_0)] \leq \sum_{k=0}^{n} \frac{1}{\frac{1}{2} + \beta} \sum_{m=0}^{k} \left[ \frac{m}{\beta} + \frac{1}{2} + \beta \right] \left( \frac{1}{\frac{1}{2} + \beta} \right)^{m},
\]

so inequalities (2.30) and (3.6) give

\[
\lim_{\theta \to \infty} \lim_{n \to \infty} \| \mu_n - \pi_n \|_{TV} = 1
\]

for \( t = \frac{n^2 - \theta}{2 \beta} - \theta \sqrt{n} = \frac{n^2}{2 \beta} - \theta \left( \sqrt{n} + \frac{1}{2 \beta} \right) \).

**Remark 3.1.** We could have chosen a different definition for the bulk. However, setting \( A_\theta = \{0, 1, \ldots, \theta - 1\} \) is particularly convenient. From (3.7)–(3.8) we see in fact that the leading order of \( \mathbb{E} [\tau_n(A_\theta)] \) is \( O(\theta) \). This term can hence be safely embodied into the window since it is a negligible contribution as a function of \( n \). That definition of the bulk has another great advantage: it allows to link together space and time. When the ‘temporal’ limit for \( \theta \to \infty \) is taken, the ‘spatial’ error we make approximating (3.3) by (3.4) vanishes. This idea is the core of Theorem 3.3, Section 3.3.

To complete the proof of cutoff we still need (2.20), which we obtain by means of the Coupling Lemma (2.17). According to the random mapping representation, a MC with transition kernel \( P_n \) can be thought of as a function \( f : \Omega_n \times [0, 1] \to \Omega_n \) such that

\[
P(f(i, u) = j) = P_n(i, j),
\]

where \( u \) is a uniform continuous random variable taking value in \([0, 1] \), see [Ho2, LPW06]. The random variable \( u \) is called random update. Given a sequence of random updates \( \{u_n\} \) and the initial state \( X_0^n \), the evolution of any MC is completely determined by the recursive relation

\[
X_{n+1}^t = f(X_n^t, u_t), \quad t \geq 0.
\]

Let us consider the following coupling \( (X_n^t, Y_n^t) \). Both \( X_n^t \) and \( Y_n^t \) are copies of the biased random walk defined by (3.1)–(3.2), and their joint evolution happens according to (3.10) using the same random updates. Clearly, the evolution of \( Y_n^t \) is not independent of that of \( X_n^t \). Provided both chains are not in state \( 0 \) or in state \( n \), if \( X_n^t \) moves to the right then so does \( Y_n^t \) and vice versa. In particular, the distance \( |X_n^t - Y_n^t| \) is non-increasing in time. It can decrease only when either \( X_n^t \) or \( Y_n^t \) are in states \( 0 \) and \( n \).

Let us now suppose that \( X_0^n = n \) while \( Y_0^n \) is chosen according to \( \tau_n \). The coalescence time \( \gamma = \min\{t \geq 0 : X_n^t = Y_n^t\} \) is stochastically dominated by \( \tau_n(A_0) = \min\{t \geq 0 : X_n^t = 0\} \), the hitting time of state \( 0 \) by \( X_n^t \), in the sense that

\[
P(\gamma \geq t) \leq P(\tau_n(A_0) \geq t) \quad \forall t \geq 0.
\]
In fact, if \( X_t^n \) has reached state 0 then the two copies have surely met. The Coupling Lemma (2.17) then yields
\[
\|\mu_t^n - \pi_n\|_{IV} \leq P(\tau_n(A_0) \geq t).
\]
For \( t = \frac{n}{2\beta} + \theta \sqrt{n} \),
\[
\lim_{\theta \to \infty} \limsup_{n \to \infty} \|\mu_t^n - \pi_n\|_{IV} = 0
\]
by (2.30) and (3.7)–(3.9). Therefore, provided \( \beta > 0 \), the biased random walk (3.1) exhibits cutoff, with cutoff time \( a_n = n/2\beta \) and cutoff window \( b_n = \Theta(\sqrt{n}) \).

Remark 3.2. In this work cutoff is considered as a phenomenon that may depend also on the initial measure \( \mu_0^n \). It is to this purpose that we have defined families of MCs as sextets including the initial state. For example, if the biased random walk is started in \( X_0^n = n/2 \) then it still exhibits cutoff, but the cutoff time is halved. There are situations, like those presented in Chapter 4, where the order of the cutoff window can be controlled by varying a parameter which do not affect the drift of the chain towards the bulk. Then, by suitable choosing the initial measure and the parameter, it is possible to infringe (2.18). On the other hand, the system will continue exhibiting a quasi-deterministic hitting of the bulk, a strong symptom of cutoff-like phenomena†, due to the presence of the drift. Roughly speaking, what happens is that the time scale of the quasi-deterministic hitting is literally swallowed-up by the increasing window. In all the models we have encountered so far, the cutoff window has been identified with the standard deviation of the hitting time of the bulk, so it may seem contradictory to claim that (2.18) fails while a limit like (3.9) continues to hold. As next section explains, there is more going on than meets the eye.

3.2 A WINDOW WITH TWO SHUTTERS

For all chains we have studied so far, it has always been possible to identify a small region of the state space \( \Omega_n \) where the stationary measure is mostly concentrated. We have called such a region the bulk. The existence of the bulk has always been associated to a quasi-deterministic hitting of this region in the sense of (2.33). Those two ingredients are sufficient to assert condition (2.19) of Definition 2.1, we have experienced it in Section 3.1 and we formalise it later in Theorem 3.3. Therefore, the standard deviation of the hitting time of the bulk must be accounted for as a component of the cutoff window whenever the phenomenon is present. In this respect, it is quite remarkable that in neither [DSC06] or [DLP10] any explicit mention is made of the standard deviation contribution to the cutoff window, although from the technical details of the proofs‡ it is quite obvious that this contribution is implicitly being considered. To our knowledge the first paper in which the contribution due to the standard deviation is explicitly noticed and developed is [LNS12].

The standard deviation of the hitting time of the bulk is not the only contribution to be considered, though. As a matter of fact, the cutoff window has two hinged doors: opening the first we have found the standard deviation contribution, looking through the second we see the thermalisation term. Thermalisation is a term borrowed from the language of thermal physics,*

\* See Remark 2.7 on page 11.
† See the discussion at the end of Section 2.8.
‡ See, for example, (2.37) and (2.38) on page 18 as well as (2.41) on page 19.
where it is used for the process of particles reaching thermal equilibrium through mutual interaction. The exact meaning we give to ‘thermalisation’ is made clear in the present and next chapters through a series of examples. Roughly speaking, it is the time needed for the chain to lose memory of its past once the bulk is reached.

To improve our understanding of this contribution, it may be useful to imagine an energy landscape associated to the stationary distribution. The mode of $\pi_n$ is the state having the highest probability to be visited at equilibrium, it hence ideally corresponds to the ground state of a physical system. For the sake of simplicity, let us assume that the modal value of $\pi_n$ is unique. Then, the energy landscape is actually an energy well, and what we have so far called the bulk can be identified with the bottom of the well. The steepness of the walls of this well depends on the gradient of the stationary measure, i.e. on the drift. The higher the drift, the steeper the walls. For example, in the coupon collector’s chain the drift is the highest possible, since the probability to go right is $p_1 = P_n(i, i + 1) = 0$. Accordingly, the energy landscape associated to the coupon collector is an infinite potential well.

The biased random walk from Section 3.4 has walls with fixed steepness, but there are also cases in which the chain has a drift that vanishes approaching the ground state like in the lazy Ehrenfest Urn, Section 3.5. The Ehrenfest chain has state space $\Omega_n = \{0, 1, \ldots, n\}$, and its stationary measure is a binomial $\mathcal{B}(n, 1/2)$. Close to the ground state $\frac{n}{2}$ and in a region of size $O(\sqrt{n})$ centred about it, the transition probabilities $p_1$ and $q_1$ are $1/4 \pm O(1/\sqrt{n})$. This makes the energy well look like a bowl with a rather flat bottom. In this model the thermalisation term is then the amount of time that the chain needs to diffuse inside the bottom, for flatness means no drift and lack of drift means diffusion.

Let us consider again the biased random walk \((3.1)\). What is the meaning of thermalisation in this model? To prove \((2.20)\), we have used a coupling $\left(X^0_n, Y^1_n\right)$ such that the second component $Y^0_n \sim \pi_n$, this means that $P \left(Y^0_n = i\right) = \pi_n(i)$ for all $i \geq 0$. In particular, with respect to the notation introduced in Section 3.1, $P \left(Y^1_n \notin A_0\right)$ is exponentially small in $\theta$. Therefore, when the first component $X^0_n$ hits the bottom of the well $A_0$, the probability of the two copies having not yet coalesced is nearly 1 if $\theta$ is sufficiently large. The time can now be reset by means of the Strong Markov Property, and with $X^0_n$ starting from the boundary* of the bulk $A_0$, the thermalisation term can be related to the coalescence time. All these ideas are formalised in Section 3.3 by Theorem 3.4.

We end the present section briefly recalling the Strong Markov Property. The meaning of the Markov property \((2.1)\) is that the evolution of a MC is the same whether we condition on the whole past trajectory or on the final state only. The Strong Markov Property extends this lack of memory from deterministic to random times. More precisely,

**Theorem 3.1** (Strong Markov Property \cite{Brégg}) Let $X^t_n$ be a MC with countable state space $\Omega_n$ and transition matrix $P_n$. Let $\tau_n$ be a stopping time with respect $X^t_n$. Then for any state $i \in \Omega_n$, given that $X^\tau_n = i$, the following hold:

1. The process after $\tau_n$ and the process before $\tau_n$ are independent;
2. The process after $\tau_n$ is a MC with transition matrix $P_n$.

* In the sense of \((2.4)\), page 6.
All the elements that have emerged in the previous discussions are formalised in this section to an original methodology for proving cutoff. Let us take a moment to survey them. Cutoff is a purely deterministic phenomenon, in the sense that it depends only on the spectral properties of the transition kernel $P_n$, yet many tools are available to relate cutoff to the dynamical properties of the chain, e.g. Lemma 2.1, Coupling Lemma (2.17), equation (2.34), and arguments like (3.3)–(3.4). All those tools feature a stopping time, and their use in a cutoff context relies on the crucial hypothesis of that stopping time being concentrated in the sense of (3.42) below. The reason for such an hypothesis is that a deterministic phenomenon can be studied using random quantities only if it can be interpreted as a typical behaviour. By means of an unlikely realisation, it may happen that the coupon collector’s chain hits the state 0 (and reach equilibrium) after $n^2$ or just n steps. For all finite values of $n$, both events are quite rare but still possible. On the other hand, for $n$ large enough, the trajectories that do not hit 0 in the time interval $n \log n \pm \Theta(n)$ do not represent a typical behaviour and can therefore be neglected. In this sense, what Theorem 3.2 below seems to suggest is a methodology similar to the pathwise approach, introduced in [CGOV84] to study metastability in statistical mechanics.

This section presents sufficient conditions for a family of finite ergodic MCs to exhibit cutoff. Theorems 3.3 and 3.4 below clearly identifies the cutoff time as the expectation of the hitting time of the relevant part of the state space, i.e. the bulk. The bulk may be related to entropic considerations that ease its identification and definition. Theorems 3.3 and 3.4 also give evidence of the nature of the cutoff window, which is in turn kindred to the standard deviation of the hitting time mentioned above or to the mixing features of the thermalisation. The level of generality of the results presented herein gives the possibility to use statistical-mechanics-based ideas to prove cutoff for a variety of models known in literature such as Top-in-at-random, Ehrenfest Urn, Random walk on the hypercube, and mean-field Ising model.

In Chapter 4 we apply the key results of the present chapter to a couple of previously unpublished one-parameter families of random walks, partially biased and partially diffusive, with the peculiar feature to have a parameter-dependent cutoff window.

As mentioned in Sections 2.7 and 2.8, there exists a precise connection between the cutoff time and the expectation of certain hitting times. The following is a possible way to exploit that connection keeping the focus on the dynamics rather than on the spectral properties of the transition kernel. Let us think of the TV-distance from equilibrium, a deterministic object that could in principle be exactly computed at any given time, as a random variable, or better as a deterministic object computed at a stochastic time. This idea motivates the subsequent definition:

**Definition 3.1.** Given a random variable $\xi$, the total-variation distance from equilibrium at time $\xi$ is the random variable

$$\left\| \mu_n^\xi - \pi_n \right\|_{TV} = \sum_{t \in \mathbb{Z}} \left\| \mu_n^{t+} - \pi_n \right\|_{TV} \mathbbm{1}_{\{\xi = t\}},$$

where $t_+ = \max\{0, t\}$. 

---

**3.3 ENTROPY-DRIVEN CUTOFF PHENOMENA**

---
When $\xi$ takes values in $[-a, +\infty)$, $a \in \mathbb{R}^+$, (3.11) is equivalent to
\[
\|\mu_n^t - \pi_n\|_{TV} = \sum_{t \geq -a} \left[ \|\mu_n^t - \pi_n\|_{TV} \mathbb{P}(\xi = t, \xi \geq 0) + \|\mu_n^0 - \pi_n\|_{TV} \mathbb{P}(\xi = t, \xi < 0) \right].
\] (3.11a)

We need this definition because later we want to consider the expectation of (3.11a) at the stochastic time $\tau - a$, where $\tau$ is a hitting time. The expectation of (3.11a) can be computed as
\[
\mathbb{E} \left[ \|\mu_n^t - \pi_n\|_{TV} \right] = \sum_{t \geq 0} \|\mu_n^t - \pi_n\|_{TV} \mathbb{P}(\xi = t, \xi \geq 0) + \|\mu_n^0 - \pi_n\|_{TV} \mathbb{P}(\xi < 0).
\] (3.12)

Although the condition $\xi \geq 0$ could be dropped in the first sum of (3.12), we keep it for notational purposes that will become clear in the proof of Theorem 3.3.

The idea of relating cutoff to the hitting time of the appropriate quantiles of the stationary distribution is already present in the literature for the special case of BDCs. The most relevant examples are those already outlined in Section 2.8. A comparison between the results of this chapter and those from [DLP10] is presented after the former have been stated and proved.

The first theorem, which is the main ingredient of the proof of Theorem 3.3, relates what may be called cutoff in the mean to the standard cutoff phenomenon, as defined by (2.18)–(2.20):

**Theorem 3.2** Let \( \{\Omega_n, X^t_n, P_n, \pi_n, \mu_n^t, \mu_n^0\} \) be a family of MCs and \( \{\tau_n\}_n \) a family of non-negative random variables with finite mean \( E_n = E[\tau_n] \) and standard deviation \( \sigma_n = \sigma[\tau_n] \) such that
\[
\lim_{n \to \infty} \frac{\sigma_n}{E_n} = 0.
\] (3.13)

Suppose there exists a sequence of positive numbers \( \{\delta_n\}_n \) such that
\[
\lim_{n \to \infty} \frac{\delta_n}{E_n} = 0,
\] (3.14)

asympt., for \( n \to \infty \), \( \mathbb{E} \left[ \|\mu_n^t - \pi_n\|_{TV} \right] \geq 1 - f(\theta), \) (3.15)

asympt., for \( n \to \infty \), \( \mathbb{E} \left[ \|\mu_n^t + \theta \delta_n - \pi_n\|_{TV} \right] \leq g(\theta), \) (3.16)

where \( f \) and \( g \) are two functions tending to 0 as \( \theta \to \infty \). Then, the family exhibits cutoff with
\[
a_n = E_n, \\
b_n = \Theta(\sigma_n + \delta_n).
\]

**Proof.** To have a more compact notation, let \( D(t) = \|\mu_n^t - \pi_n\|_{TV} \) and \( \xi = \tau_n - \theta \delta_n \). Then, \( \mathbb{E} [\xi] = \theta \sigma_n = a_n - \theta b_n \), and using (3.12),
\[
\mathbb{E} [D(\xi)] = \sum_{t \geq 0} D(t) \mathbb{P}(\xi = t, \xi \geq 0) + D(0) \mathbb{P}(\xi < 0),
\]
\[
\leq \sum_{t \geq 0} D(t) \mathbb{P}(\xi = t, \xi \geq 0) + \mathbb{P}(\xi < 0).
\] (3.17)
We can estimate the sum in (3.17) as follows

\[
\sum_{t \geq 0} D(t) \mathbb{P}(\xi = t, \xi \geq 0) \leq \sum_{t=0}^{\xi \leq \theta \sigma_n} D(t) \mathbb{P}(\xi = t, \xi \geq 0) + \sum_{t \geq \xi \leq \theta \sigma_n} D(t) \mathbb{P}(\xi = t, \xi \geq 0),
\]

(3.18)

where we have used (2.10) to obtain (3.19) from (3.18). Substituting equation (3.19) in (3.17),

\[
\mathbb{E}[D(\xi)] \leq \mathbb{P} \{ \tau_n \leq \mathbb{E}[\tau_n] - \theta \sigma_n \} + D(\mathbb{E}[\xi] - \theta \sigma_n),
\]

that is,

\[
\mathbb{E} \left[ \left\| \mu_n^{\tau_n - \theta \delta_n} - \pi_n \right\|_{TV} \right] \leq \left\| \mu_n^{\tau_n - \theta b_n} - \pi_n \right\|_{TV} + \mathbb{P} (\tau_n \leq \mathbb{E}[\tau_n] - \theta \sigma_n).
\]

Reverting the last inequality, by (3.15) and (2.30),

\[
1 \geq \liminf_{n \to \infty} \left[ \left\| \mu_n^{\tau_n - \theta b_n} - \pi_n \right\|_{TV} \right] \geq 1 - f(\theta) - \frac{1}{1 + \theta^2}.
\]

(3.20)

Now let \( \eta = \tau_n + \theta \delta_n \) and notice that \( \mathbb{E}[\eta] + \theta \sigma_n = \eta_n + \theta b_n \). Since \( \eta \geq \theta \delta_n \), by (3.12),

\[
\mathbb{E}[D(\eta)] = \sum_{t \geq \theta \delta_n} D(t) \mathbb{P}(t = \eta) \geq \sum_{t = \theta \delta_n}^{\eta \geq \mathbb{E}[\eta] + \theta \sigma_n} D(t) \mathbb{P}(t = \eta),
\]

(3.21)

\[
\geq D(\mathbb{E}[\eta] + \theta \sigma_n) \sum_{t = \theta \delta_n}^{\eta \geq \mathbb{E}[\eta] + \theta \sigma_n} \mathbb{P}(t = \eta),
\]

(3.22)

\[
\geq D(\mathbb{E}[\eta] + \theta \sigma_n) \left( 1 - \frac{1}{1 + \theta^2} \right),
\]

where we have used (2.10) to obtain (3.22) from (3.21). Reverting the last inequality,

\[
\left\| \mu_n^{\eta_n + \theta b_n} - \pi_n \right\|_{TV} \leq \mathbb{E} \left[ \left\| \mu_n^{\tau_n + \theta \delta_n} - \pi_n \right\|_{TV} \right] + \frac{1}{1 + \theta^2}.
\]

Therefore, by (3.16) and (2.30),

\[
0 \leq \limsup_{n \to \infty} \left\| \mu_n^{\eta_n + \theta b_n} - \pi_n \right\|_{TV} \leq g(\theta) + \frac{1}{1 + \theta^2}.
\]

(3.23)

Finally, note that (3.13) and (3.14) infer (2.18). Passing to the limits for \( \theta \to \infty \) in (3.20) and (3.23) concludes the proof.

Before we move to the statement of Theorem 3.3 we need a couple of definitions to formally define what we have called so far bulk, relevant part of the state space, or bottom of the energy well.

**Definition 3.2.** A family of nested subsets is a sequence \( \{ A_n, \theta \} \_{\theta \geq 1} \) having the following properties: for an arbitrarily fixed value of \( \theta \), there exists \( N \geq 0 \) such that \( \forall n \geq N \),

\[
A_n, \theta' \subset \Omega_n \quad \forall 1 \leq \theta' \leq \theta,
\]

(3.24)

\[
A_n, \theta'' \subset A_n, \theta' \quad \forall 1 \leq \theta' \leq \theta'' \leq \theta.
\]
Definition 3.3. Given a family of nested subsets, the stationary measure $\pi_n$ is said to be $h$-concentrated on $A_{n,\theta}$ if there exists a function $h(\theta)$ tending to zero as $\theta \to \infty$ such that

\begin{equation}
\text{asymptotically, for } n \to \infty, \quad \pi_n(A_{n,\theta}^c) < h(\theta),
\end{equation}

where $A_{n,\theta}^c = \Omega_n \setminus A_{n,\theta}$.

Definitions 3.2 and 3.3 capture and formalise the properties of what we have called the relevant part of the state space, or bulk. According to the ideas we have developed so far, the bulk is a set which, at least asymptotically, is given weight close to 1 by the stationary distribution. The reason why we build a family of nested subsets indexed by $\theta$ with this concentration property is summarised by Remark 3.1 above, where it is explained that such a formulation allows to create a useful link between space and time. The idea is that, when the bulk is defined as a family of nested subsets where $\pi_n$ is $h$-concentrated on, one can regard $A_{n,\theta}$, rather than $\Omega_n$, as the support of the stationary measure $\pi_n$. As long as the probability of $X_t^n$ being in the bulk is negligible, the parts of the space state which are relevant with respect to $\pi_n$ and $\mu_n^\theta$ are disjoint, and the TV-distance is asymptotically 1 (Remark 2.4) save for an error that vanishes in the limit $\theta \to \infty$.

We need some notation before we can state Theorem 3.3. Let

$$\zeta_n^\theta = \min\{t \geq 0 : X_t^n \in A_{n,\theta}\}$$

be the hitting time of $A_{n,\theta}$.

Remark 3.3. We note that $\zeta_n^\theta \geq \zeta_n^{\theta'}$ whenever $\theta \leq \theta'$.

**Theorem 3.3** Let \{\Omega_n, X_n^t, P_n, \pi_n, \mu_n^t, \mu_n^\theta\} be a family of MCS. Suppose that $\mu_n^\theta$ is such that there exists a family of nested subsets \{\A_{n,\theta}\}_{\theta \geq 1} \subset \Omega_n$ with the following properties:

\begin{align}
\pi_n \text{ is } h\text{-concentrated in } A_{n,\theta}, \\
\sigma \left[ \zeta_n^1 \right] \to 0, \\
\sigma \left[ \zeta_n^\theta \right] \leq \sigma \left[ \zeta_n^1 \right],
\end{align}

and there exists a sequence of positive integers \{\Delta_n\} such that

\begin{align}
\lim_{n \to \infty} \frac{\Delta_n}{\sigma \left[ \zeta_n^1 \right]} &\in \mathbb{R}^+ \cup \{+\infty\}, \\
\lim_{n \to \infty} \frac{\Delta_n}{\mathbb{E} \left[ \zeta_n^1 \right]} &\to 0, \\
\lim_{n \to \infty} \limsup_{\theta \to \infty} \frac{\mathbb{E} \left[ \zeta_n^1 - \zeta_n^\theta \right]}{\theta \Delta_n} &= 0.
\end{align}

Then, there exists a function $f(\theta)$, tending to 0 as $\theta \to \infty$, such that

\begin{equation}
\text{asympt., for } n \to \infty, \quad \mathbb{E} \left[ \left\| \mu_n^{1-\theta} - \pi_n \right\|_{TV} \right] \geq 1 - f(\theta),
\end{equation}

where

\begin{equation}
\delta_n = 2 \left( \Delta_n + \sigma \left[ \zeta_n^1 \right] \right).
\end{equation}
Proof. Let us fix an arbitrary \( \theta > 1 \) and consider \( n \) sufficiently large to ensure (3.24). As in the proof of Thm. 3.2, let \( D(t) = \|\mu_n^t - \pi_n\|_\nu \) and \( \xi = \zeta_n^1 - \theta \delta_n \). By (3.12),

\[
\mathbb{E} [D(\xi)] = \sum_{t\geq 0} D(t) \mathbb{P} (\xi = t, \xi \geq 0) + D(0) \mathbb{P} (\xi < 0) ,
\]

\[
\geq \sum_{t\geq 0} D(t) \mathbb{P} (\xi = t, \xi \geq 0) ,
\]

\[
= \sum_{t\geq 0} \mathbb{P} (\xi = t, \xi \geq 0) \frac{1}{2} \sum_{i \in \Omega_n} |\mu_n^t(i) - \pi_n(i)| ,
\]

\[
= \mathbb{P} (\xi \geq 0) \frac{1}{2} \sum_{i \in \Omega_n} \left( \pi_n(i) - \sum_{t\geq 0} \mu_n^t(i) \mathbb{P} (\xi = t | \xi \geq 0) \right) ,
\]

\[
\geq \mathbb{P} (\xi \geq 0) \frac{1}{2} \sum_{i \in \Omega_n} \left( \frac{1}{\sqrt{n}} \right) .
\]

At this point we note that \( \rho_n(i) = \sum_{t\geq 0} \mu_n^t(i) \mathbb{P} (\xi = t | \xi \geq 0) \) is a probability distribution on \( \Omega_n \) because

\[
\sum_{i \in \Omega_n} \rho_n(i) = \sum_{t\geq 0} \mathbb{P} (\xi = t | \xi \geq 0) \sum_{i \in \Omega_n} \mu_n^t(i) = 1 .
\]

Hence, using (2.8) and (2.9),

\[
\mathbb{E} [D(\xi)] \geq \mathbb{P} (\xi \geq 0) \max_{A \subseteq \Omega_n} \left[ \pi_n(A) - \sum_{t\geq 0} \mu_n^t(A) \mathbb{P} (\xi = t | \xi \geq 0) \right] ,
\]

\[
\geq \mathbb{P} (\xi \geq 0) \left[ \pi_n(A_{n,0}) - \sum_{t\geq 0} \mu_n^t(A_{n,0}) \mathbb{P} (\xi = t | \xi \geq 0) \right] ,
\]

\[
\geq \mathbb{P} (\xi \geq 0) (1 - h(\theta)) - \sum_{t\geq 0} \mu_n^t(A_{n,0}) \mathbb{P} (\xi = t, \xi \geq 0) .
\]  

(3.34)

(3.35)

Using (2.30), we can estimate the first term of the Right Hand Side (RHS) of (3.35) as

\[
(1 - h(\theta)) \mathbb{P} (\xi \geq 0) = (1 - h(\theta)) \mathbb{P} \left( \mathbb{E} \left[ \zeta_n^1 \right] - \zeta_n^1 \leq \mathbb{E} \left[ \zeta_n^1 \right] - \theta \delta_n \right) ,
\]

\[
\geq (1 - h(\theta)) \left( 1 - \frac{\text{Var}[\zeta_n^1]}{\text{Var}[\zeta_n^1] + (\mathbb{E} [\zeta_n^1] - \theta \delta_n)^2} \right) .
\]

By (3.27), (3.30), and (3.33), \( \mathbb{P} (\xi \geq 0) \) is asymptotically greater than any function of \( \theta \) tending to one, say \( 1 - \frac{1}{\theta} \). Thus, for \( n \) sufficiently large,

\[
(1 - h(\theta)) \mathbb{P} (\xi \geq 0) \geq 1 - h(\theta) - \frac{1}{\theta} .
\]

Next, consider the remaining term of (3.35).

\[
\sum_{t\geq 0} \mu_n^t(A_{n,0}) \mathbb{P} \left( \zeta_n^1 - \theta \delta_n = t, \zeta_n^1 - \theta \delta_n \geq 0 \right)
\]

\[
\leq \sum_{t\geq 0} \mathbb{P} (t \geq \zeta_n^0) \mathbb{P} \left( \zeta_n^1 - \theta \delta_n = t \right) ,
\]
Comparison with \[ \text{DLP} \]

sizing up the cutoff window exhibiting cutoff. Although quite general, it is most useful when we face a

\[ \text{Cf. Remark} \]

Theorem \[ 3 \]

Thus, for \( n \)

the limit in (A)

\[ \text{E} \]

easier task than estimating

Moreover, the computation of the spectral gap of a chain is not always an

is found. On the other hand, provided a suitable definition

* \[ \text{Remark} \]

hitting time of

\[ \text{BDCs} \]

As we have already noted, \( \pi_n \) being concentrated in \( A_{n,0} \) is equivalent to a drift of the chain towards \( A_{n,0} \) itself, and such a drift is likely to ensure

the limit in (3.27). The latter means in turn that, for \( n \) sufficiently large, the

\[ \text{Comparison with \[ \text{DLP} \]} \]

\[ \text{3.10} \]

\[ \text{40} \]

\[ \text{SIZING UP THE CUTOFF WINDOW} \]
chain hits $A_{n,0}$ in a quasi-deterministic way. In other words, the probability of $X_t^n$ being inside $A_{n,0}$ suddenly rises from 0 to 1 in a time-window of size $\sigma[\zeta_n^0]$ centred around $E[\zeta_n^0]$. This means that, if the system was started outside $A_{n,0}$, it is undergoing the first part of the cutoff curve, i.e., it satisfies (2.19). In addition, if the system relaxes inside $A_{n,0}$ in a time interval that is comparable with $\sigma[\zeta_n^0]$ then it experiences cutoff with a window of order $\sigma[\zeta_n^0]$. It is also possible that the time $t_{\text{therm}}$ needed for the system to thermalise inside $A_{n,0}$ is larger than $\sigma[\zeta_n^0]$ but smaller than $E[\zeta_n^0]$, implying then cutoff with a cutoff window of order $t_{\text{therm}}$.

The technical problem that was encountered while conceiving Theorem 3.3 is that $E[\zeta_n^0]$ is not a good candidate to the cutoff time $a_n$, being $\theta$-dependent. For this reason, the trajectory has been split in two parts, namely, the approach to $A_{n,0}$, the brim of the bottom part of the energy well, and the subsequent fall deep inside a region $A_{n,1}$, included in the bulk and independent of $\theta$. Conditions (3.30) and (3.31) model the requirement of the fall-off part of the trajectory being much faster with respect to the time scale of $\zeta_n^1$.

The contribution to the cutoff window due to the thermalisation inside $A_{n,1}$ is dealt with by the next theorem:

**Theorem 3.4** Assume that all the hypothesis of Theorem 3.3 hold for a given family of MCs. In addition, suppose that given two identical copies of the $n$-th chain $Z_n^0$ and $W_n^0$, there exists a coupling $(Z_n^1, W_n^1)$ such that
\[
Z_n^0 = z_0 \in A_{n,1}, \quad W_n^0 \sim \pi_n,
\]
if $Z_n^s = W_n^s$ then $Z_n^t = W_n^t$, $\forall s \geq s^*$,
\[
\gamma_n = \min\{t \geq 0 : Z_n^t = W_n^t\} \quad \text{is such that asympt., for } n \to \infty,
\]
\[
\max_{z_0 \in A_{n,1}} \mathbb{P}\left(\gamma_n > \theta\delta_n \mid Z_n^0 = z_0\right) \leq g(\theta), \quad (3.37)
\]
where $\delta_n$ is defined in (3.33) and $g(\theta) \to 0$. Then, the family exhibits cutoff with
\[
a_n = E\left[\zeta_n^1\right],
b_n = O(\delta_n).
\]

As it will become clear in the proof, the idea of Theorem 3.4 is to characterise the contribution to the cutoff window due to the thermalisation as the time needed for the chain to achieve equilibrium starting from the bottom of the well. The main ingredient is $(Z_n^0, W_n^0)$, a coupling such that $W_n^0$ has initial position drawn according to the stationary measure $\pi_n$ and $Z_n^0$ starts on the boundary of $A_{n,1}$, the inner part of the bottom of the energy well. Being started according to $\pi_n$, $W_n^0$ belongs to the bottom of the well with very high probability for all $t \geq 0$. This coupling is not sufficient to bound the total-variation distance between $\mu_n^t$ and $\pi_n$, though. The reason is that $Z_n^t$ is not distributed according to $\mu_n^t = P_n^t \mu_n^0$, simply because the initial measure of $Z_n^0$ is not $\mu_n^0$ in general. However, a coupling between $\mu_n^t$ and $\pi_n$ can be constructed in such a way to exploit the coalescence properties of $(Z_n^0, W_n^0)$. The latter is the actual target of the Coupling Lemma, and we denote it by $(X_t^n, Y_t^n)$. Before moving to the proof of Theorem 3.4, we need the equality
\[
\lim_{M \to +\infty} \mathbb{P}\left(\zeta_n^1 \geq M\right) = 0, \quad (3.38)
\]
an easy consequence of (2.30) and (3.27).
Proof. The idea of the proof is to use the Coupling Lemma (2.17). To this purpose, we construct as follows a coupling \((X^t_n, Y^t_n)\) of \(\mu^t_n\) and \(\pi_n\), based on the coupling \((Z^t_n, W^t_n)\):

1. let \(X^0_n \sim \mu^0_n\) and \(Y^0_n \sim \pi_n\), and define the first coalescence time as
   \[
   \hat{\gamma}_n = \min \{ t \geq 0 : X^t_n = Y^t_n \} ;
   \]
2. for \(0 \leq t \leq \hat{\zeta}_n^1\),
   a) \(X^t_n\) and \(Y^t_n\) evolve independently until \(\hat{\gamma}_n\), if \(\hat{\gamma}_n < \hat{\zeta}_n^1\);
   b) \(X^t_n = Y^t_n\) \forall \hat{\gamma}_n \leq t \leq \hat{\zeta}_n^1\), if any;
3. let \(Z^0_n = X^1_n\) and \(W^0_n = Y^1_n\), then for all \(t > \hat{\zeta}_n^1\) run the coupling of \(Z^t_n\) and \(Y^t_n\) and set \((X^t_n, Y^t_n) = (Z^t_n, W^t_n)\).

We have built the coupling \((X^t_n, Y^t_n)\) in this fashion because it has the following property as a consequence of Theorem 3.1. Given \(\hat{\zeta}_n^1 = T < \infty\), for all \(z_0 \in A_n,\)

\[
\mathbb{P} \left( \hat{\gamma}_n > T + \theta \delta_n \mid X_n^T = z_0 \right) = \mathbb{P} \left( \gamma_n > \theta \delta_n \mid Z^0_n = z_0 \right),
\]

(3.39)

where, according to the notation introduced in Theorem 3.4, \(\gamma_n\) is the first coalescence time of \(Z_n\) and \(W_n\). Let us take an arbitrary \(M\). Then, by (2.17),

\[
\left\| \mu^{T + \theta \delta_n}_n - \pi_n \right\|_{TV} = \sum_{T \geq 0} \left\| \mu^{T + \theta \delta_n}_n - \pi_n \right\|_{TV} I_{\{\hat{\zeta}_n = T\}},
\]

(4.40)

\[
\leq \sum_{T = 0}^M \left\| \mu^{T + \theta \delta_n}_n - \pi_n \right\|_{TV} I_{\{\hat{\zeta}_n = T\}} + I_{\{\hat{\zeta}_n \geq M\}},
\]

\[
\leq \sum_{T = 0}^M \mathbb{P} \left( \hat{\gamma}_n > T + \theta \delta_n \mid X^0_n = x_0 \right) I_{\{\hat{\zeta}_n = T\}} + I_{\{\hat{\zeta}_n \geq M\}},
\]

\[
= \sum_{T = 0}^M \sum_{z_0 \in A_n, i} \left[ \mathbb{P} \left( \hat{\gamma}_n > T + \theta \delta_n \mid X^0_n = x_0, X_n^T = z_0 \right) \times \right]
\]

\[
\mathbb{P} \left( X_n^0 = x_0, X_n^T = z_0 \right) I_{\{\hat{\zeta}_n = T\}} + I_{\{\hat{\zeta}_n \geq M\}},
\]

\[
\leq \sum_{T = 0}^M \max_{z_0 \in A_n, i} \mathbb{P} \left( \hat{\gamma}_n > T + \theta \delta_n \mid X_n^T = z_0 \right) I_{\{\hat{\zeta}_n = T\}} + I_{\{\hat{\zeta}_n \geq M\}}.
\]

By (3.37) and (3.39), for \(n\) sufficiently large,

\[
\left\| \mu^{T + \theta \delta_n}_n - \pi_n \right\|_{TV} \leq g(\theta) I_{\{\hat{\zeta}_n \leq M\}} + I_{\{\hat{\zeta}_n \geq M\}}.
\]

(3.41)

Finally, passing to the expectation in (3.41), by (3.38),

\[
\mathbb{E} \left[ \left\| \mu^{T + \theta \delta_n}_n - \pi_n \right\|_{TV} \right] \leq g(\theta) \text{ asymptotically, for } n \to \infty.
\]

Identifying \(\hat{\zeta}_n^1\) with \(\tau_n\) from Theorem 3.2, we obtain (3.16). Theorem 3.3 gives (3.13)-(3.15) and the definition of \(\delta_n\) via (3.33). Hence, Theorem 3.2 infers that the family of MCS exhibits cutoff with cutoff time \(a_n = \mathbb{E} [\hat{\zeta}_n]\) and cutoff window of order \(b_n = \Theta(2\Delta_n + 3\sigma [\hat{\zeta}_n]) = \Theta(\delta_n)\).
Theorems 3.3 and 3.4 clarify the link between the cutoff phenomenon and the hitting of the relevant part of the state space, or bulk. Relevant part stands for that subset of the state space where the stationary distribution $\pi_n$ is mostly concentrated, a concept which has been formalised by Definitions 3.2 and 3.3 above. This seems quite a natural approach when we realise that nearly every chain known to exhibit cutoff hits the relevant part of the state space in a quasi-deterministic way, i.e., the hitting time $\tau_n$ of the bulk satisfies

$$\frac{\sigma[\tau_n]}{E[\tau_n]} \to 0, \quad n \to \infty$$

(3.42)

For BDCs, the presence of a drift and a limit as in (3.42) are two sides of the same coin. In a more general framework we can expect (3.42) to be relatively easy to prove whenever the chain presents a drift towards the relevant part of the state space.

The picture of a quasi-deterministic hitting time that we have described so far holds as well for systems with uniform stationary measure, for which the relevant part of the state space would be $\Omega_n$ itself. As a matter of fact, if we focus on a suitable projection of the chain then we may find that the original stationary distribution is no longer uniform. For example, the uniform distribution on the symmetric group $S_n$ becomes (2.49) when each permutation is identified with the number of rising sequences it is composed of. In general, the projected stationary distribution $\pi_n^\sigma(x)$ is proportional to the number of states $i \in \Omega_n$ which correspond to $x$ according to the equivalence relation used to perform the projection. Consequently, since $-\pi_n^\sigma(x) \log \pi_n^\sigma(x)$ is the contribution given by the $x$-th equivalence class to the entropy of $\pi_n^\sigma$, the relevant part of the state space is composed of those classes providing the leading contribution to the entropy. In these cases the drift mentioned above is therefore supplied by entropic considerations’. Having this idea in mind, Theorems 3.3 and 3.4 may then represent a possible trait d’union between two different classes of MCs exhibiting cutoff: those chains having stationary measure concentrated in a small subset of the state space, like BDCs with drift, and the class of chains with stationary measure uniform or spread over $\Omega_n$, like the random walk on the hypercube, many card-shuffling models and some high-temperature statistical mechanics models.

We end the present section with some very important remarks.

Remark 3.4. There is no universal choice for the family $\Lambda_{n,0}$. Multiple definitions are possible and each of them indirectly affects the size of the cutoff window. Remark 4.7 in Section 4.2 shows a choice of $\Lambda_{n,0}$ which leads to a non-optimal cutoff window. The applications presented in the remainder of this chapter and at the beginning of the following suggest that the key to obtain an optimal cutoff window is designing the family $\Lambda_{n,0}$ in such a way that the expectation of the travel time $E[\zeta_n - \zeta_n^0]$ is of the same order in $n$ as the thermalisation time $\theta_0 \delta_n$, sufficient to achieve equilibrium in the sense of Theorem 3.4.

Remark 3.5. In the case of BDCs, hypotheses (3.28) is trivial.

Remark 3.6. It is important to stress here that the task of showing the cutoff behaviour is usually accomplished in the literature by means of a coupling argument. In most situations the coupling argument needs to be sufficiently fine since the desired estimates are to be performed at times $a_n \pm \theta_0 b_n$, i.e.,

- We have already encountered this idea in Section 2.9, and we return to it in Sections 3.6 and 3.7 where the name entropy-driven cutoff phenomena is fully explained.
with two very different time scales involved. In our approach this time scale issue does not appear as we split the study of the cutoff in two phases, namely, the hitting of $A_{n,1}$ and the subsequent thermalisation phase. The approaching phase takes place on a time scale of order $a_n$ while the thermalisation phase is on a time scale of order $b_n$, much smaller than $a_n$. In Sections 3.4–4.2 we see that only very basic and intuitive couplings are demanded by Theorem 3.4.

**Remark 3.7.** If the family of MCs is a nearest-neighbour dynamics then $X_n^t$ can not *jump* inside $A_{n,1}$, but only hit it on the boundary*. In other words, $X_n^t \in \partial A_{n,1}$. Therefore, we can safely replace (3.37) by

$$\max_{z_0 \in \partial A_{n,1}} \mathbb{P} \left( \gamma_n > \theta \delta_n \mid Z_n^0 = z_0 \right) < g(\theta)$$

(3.37a)

in the statement of Theorem 3.4. Further, if the family under consideration is a family of BDCs then the set $\partial A_{n,1}$ is composed of just two points for each connected components of $A_{n,1}$. Depending on $\mu_n^0$, in those situations we could be able to determine which point of $\partial A_{n,1}$ will be hit by the chain. If this is the case, the max in (3.37) can be profitably dropped.

### 3.4 COUPON COLLECTOR’S CHAIN REVISITED

The coupon collector’s model is a *pure-death* chain on the state space $\Omega_n = \{0, 1, 2, \ldots, n\}$. Its transition rates are

$$q_i = \mathbb{P}_n(i, i-1) = \frac{i}{n},$$
$$r_i = \mathbb{P}_n(i, i) = \frac{n-i}{n},$$
$$p_i = \mathbb{P}_n(i, i+1) = 0.$$

The coupon collector’s problem was first introduced in [ER61] and is discussed in many probability books, e.g. [Stio3, LPWo6]. We have already met this model in Section 2.7, here we give an alternative description of the cutoff by means of Theorem 3.2.

The chain clearly has a drift towards the state 0, for it just can not move to the right. The equilibrium distribution is $\pi_n = \delta_{i,0}$, while the initial distribution is taken to be $\mu_n^0 = \delta_{i,n}$. The hitting time of the state 0 is $\tau_n(0)$, which happens to be a strong stationary time. Thus, for any finite time $t$,

$$\mathbb{P} \left( X_n^t = i, \ t \geq \tau_n(0) \right) = \pi_n(i).$$

(3.43)

The hitting time of 0 satisfies

$$\mathbb{E}[\tau_n(0)] = n \log n,$$
$$\sigma[\tau_n(0)] = n,$$

where the symbol $\sim$ stands for asymptotic equivalence. By (3.40), (3.41), and (3.43),

$$\mathbb{E} \left[ \left\| \mu_n^{\tau_n(0)} + c - \pi_n \right\|_{TV} \right] = 0 \quad \forall c \geq 0.$$

Let us next recall that $D(t) = \| \mu_n^t - \pi_n \|_{TV}$ and take $\xi = \tau_n(0) - 2\theta n$ and $A_{n,0} = \{0\}$. Then, from (3.34)

$$\mathbb{E}[D(\xi)] \geq \mathbb{P}(\xi \geq 0) - \sum_{t \geq \theta} \mathbb{P}(X_n^t = 0) \mathbb{P}(\xi = t, \xi \geq 0).$$

---

* As defined by (2.4) on page 6.
Now,
\[ P(\xi \geq 0) = P(n \log n - \tau_n(0) \leq n(\log n - 2\theta)) \geq 1 - \frac{1}{1 + (\log n - 2\theta)^2}, \]
and
\[
\sum_{t \geq 0} P(X_n^t = 0) P(\xi = t, \xi \geq 0) \\
\leq \sum_{t = n \log n - 3\theta n}^{n \log n - \theta n} P(t \geq \tau_n(0)) P(t = \tau_n(0) - 2\theta n) + \frac{1}{\theta^2}, \\
\leq P(n \log n - \tau_n(0) \geq \theta n) + \frac{1}{\theta^2}, \\
\leq \frac{1}{1 + \theta^2} + \frac{1}{\theta^2}.
\]
Thus, for \( n \) sufficiently large, there exists a function \( f(\theta) \), tending to 0 as \( \theta \to \infty \), such that
\[ E \left[ \| \mu_n^{\tau_n(0) - \theta n} - \mu_n \|_{TV} \right] \geq 1 - f(\theta). \]

Then, using Theorem 3.2, we find that the coupon collector exhibits cutoff with cutoff time \( a_n = E[\tau_n(0)] = n \log n \) and cutoff window \( b_n = \Theta(\sigma[\tau_n(0)]) = \Theta(n) \).

### 3.5 The Ehrenfest Urn Model

The Ehrenfest Urn is one of the most famous models in statistical mechanics. It was proposed in 1907 by Tatiana and Paul Ehrenfest [EE07] to explain the second law of thermodynamics and solve its apparent clash with Poincaré’s Recurrence Theorem. The cutoff phenomenon for this chain is shown in many papers, e.g. [Ald83, DS87, DGM90]. The review paper [Dia96] gives a detailed explanation of the phenomenon and offers a number of interesting references.

The Ehrenfest Urn model is composed of two boxes, Urns 1 and 2, containing a total amount of \( n \) particles. Each particle may change container at each step independently with probability \( \frac{1}{2n} \). Let \( X_n^i \) count the number of particles in Urn 1, and suppose it contains \( i \) particles. Then, the transition rates of the Ehrenfest chain are
\[
q_i = P_n(i, i - 1) = \frac{i}{2n}, \\
r_i = P_n(i, i) = \frac{1}{2}, \\
p_i = P_n(i, i + 1) = \frac{n-i}{2n}.
\]

According to the description above, the Ehrenfest chain is a lazy BDC on the segment \( \Omega_n = \{0, 1, \ldots, n\} \) and, by reversibility, its stationary distribution is a binomial \( B(n, \frac{1}{2}) \).

We now prove cutoff for the Lazy Ehrenfest Urn using Theorems 3.3 and 3.4. A good choice for the family of nested subsets is
\[
A_{n, \theta} = \left\{ i \in \Omega_n : \left| i - \frac{n}{2} \right| \leq \frac{\theta}{2} \sqrt{n} \right\}
\]

* A MC is called lazy if \( P_n(i, i) \geq 1/2 \) for each \( i \in \Omega_n \).
46 SIZING UP THE CUTOFF WINDOW

Figure 13: TV-distance from equilibrium for an Ehrenfest Urn depending on the initial state, \( \mu_0 \). The Urn contains 10,000 particles. The red line refers to the case where Urn 1 is started empty, whereas the blue line refer to the case where at time zero Urn 1 and Urn 2 contain the same number of particles.

since \( \pi_n(A_{n,0}) < \frac{1}{\sigma^2} \) by means of Chebyshev’s inequality. Let us also suppose that \( \mu_0 = \delta_{\text{Urn}} \), which is to say, at time \( t = 0 \) Urn 1 is empty. Plain but lengthy calculations, presented for the sake of completeness in Appendix A, show that

\[
\begin{align*}
E\left[\zeta_n^1\right] & = \frac{1}{2}n \log n, \\
E\left[\zeta_n^1 - \zeta_n^0\right] & = n \log \theta, \\
\sigma\left[\zeta_n^1\right] & = O(n).
\end{align*}
\]

By Remark 3.5, the hypotheses of Theorem 3.3 are fulfilled if we take \( \Delta_n = \Theta(n) \). Then, \( \delta_n = \Theta(n) \) as a consequence of this choice, and we are left with showing that there exists a coupling verifying Theorem 3.4. The Lazy Ehrenfest Urn shares the latter feature with the Mean-field Ising model so we defer the matter to Section 3.7, see in particular Remark 3.12. As a result, we have proved that the Lazy Ehrenfest Urn exhibit cutoff with \( a_n = \frac{1}{2}n \log n \) and \( b_n = \Theta(n) \).

In Remark 3.2 we have mentioned that we are building a framework where to study cutoff as a phenomenon depending on the initial measure \( \mu_0 \). Figure 13 shows the behaviour of the TV-distance for two different initial configurations, namely, Urn 1 completely empty and Urn 1 with half the total amount of particles. In the former case there is cutoff, whereas in latter there is not. With respect to the energy-well picture we have discussed in Section 3.2, starting the Ehrenfest chain with the same number of particles in Urn 1 and Urn 2, i.e. \( \mu_0 = \delta_{\text{Urn}} \), means that the system is initially at the bottom of the well. In a region of order \( \sqrt{n} \) around \( n/2 \) the energy landscape is flat because \( p_i = q_i \) except for an infinitesimal correction of order \( \frac{1}{\sqrt{n}} \), see (3.44).

On the contrary, the chain exhibits cutoff in the case of an initially empty Urn 1. A clear explanation of this is given by Figure 14. As long as \( X_n^1 \) has no reasonable chance to be inside the relevant part of the state space \( A_{n,\theta} \), the
evolute and the stationary measure have negligible overlap (see Figure 14(a)).
In other words, the part of the state space relevant to \( \mu_n^1 \) is disjoint from that part which is relevant according to \( \pi_n \). From Remark 2.4 the TV-distance \( \|\mu_n^1 - \pi_n\|_{TV} \) is then asymptotically 1.

Remark 3.8. Figures 13 and 14 refer not to the Ehrenfest dynamics defined in (3.44). As a matter of fact, they refer to the following transition kernel

\[
\begin{align*}
q_i &= P_n(i, i-1) = \frac{i}{n+1}, \\
r_i &= P_n(i, i) = \frac{1}{n+1}, \\
p_i &= P_n(i, i+1) = \frac{n-i}{n+1}.
\end{align*}
\]

The transition matrix (3.44a) is the one used in [Ald83, DS87, Dia96], where it is proved that the associated family of MCs exhibits cutoff with cutoff time \( a_n = \frac{1}{4} n \log n \) and cutoff window \( b_n = \Theta(n) \). The lazy formulation (3.44) is typically preferred because of the spectral properties of lazy chains, see [LPW06]. Roughly speaking, the main difference between the two dynamics defined by (3.44) and (3.44a) is that the former takes an average amount of self-transitions which is double with respect to the latter. This explains, at least at an intuitive level, why the cutoff time of the two models have ratio 2.

### 3.6 The Lazy Random Walk on the Hypercube

In this model the state space is \( \Omega_n = \{0,1\}^n \), the \( n \)-dimensional hypercube. Each state can be represented as a binary \( n \)-tuple \( x = (x_1, \ldots, x_n) \). Without loss of generality, let the chain be at time zero in the vertex \( (0, \ldots, 0) \). Then, at each step a component of the tuple chosen uniformly at random is switched with probability \( \frac{1}{2} \). This corresponds to the following update procedure: at each step a fair coin is flipped, if heads comes up then the chain stands still on current vertex (self-transition); if tails comes up instead, then one of the possible \( n \) directions is chosen uniformly at random, and the chain moves along it to the new vertex. The equilibrium distribution is \( \pi_n(i) = \frac{1}{2^n} \) for all \( i \).

Generally, the proof of cutoff for this model proceeds via the projection onto a BDC. Let us consider the following equivalence relation:

\[ x \sim y \iff \|x\|_{\ell_1} = \|y\|_{\ell_1}, \]

where \( \|x\|_{\ell_1} = \sum_i x_i \) is the Hamming weight of the vertex \( x \). The quotient state space \( \Omega_n/\sim \) can be put into a one-to-one correspondence with the state space \( \Omega_n^a = \{0,1, \ldots, n\} \) of a new chain \( X_n^a \) having transition rates given by (3.44) and equilibrium distribution equal to a binomial \( \mathcal{B}(n, \frac{1}{2}) \).

Let us denote by \( \mu_n^a \) the evolved measure of the projected chain \( X_n^a \) after \( t \) steps, and by \( \pi_n \) its equilibrium distribution. Then, it is a standard task\(^*\) to show that

\[
\|\mu_n^a - \pi_n\|_{TV} = \|\mu_n^a - \pi_n^a\|_{TV}.
\]

Hence, the Lazy Random Walk on the Hypercube exhibits cutoff with the same cutoff time and cutoff window as the Lazy Ehrenfest Urn.

Remark 3.9. Since \( \pi_n \) is uniform, the projected stationary distribution \( \pi_n^a(i) \)

\[^*\text{See, e.g., [LPW06, Thm. 18.3].} \]
(a) Until the chain has no reasonable chance to be in the bulk $\mathcal{A}_{n,0}$ of the stationary measure, $\mu_n^t$ and $\pi_n$ have negligible overlap and we are on the plateau part of the cutoff curve.

(b) After the bulk $\mathcal{A}_{n,0}$ is hit, the chain thermalises at the bottom of the energy well.

Figure 14: Evolved measure of the Ehrenfest Urn and approach to stationarity. The total number of particles is 10,000.
is clearly proportional to the number of vertices having Hamming weight equal to \( i \). Therefore, \( \pi_n^i \) is binomial and supported on \( \Lambda_{n,0} \) in the sense of (3.25). Since the configurations in \( \Lambda_{n,0} \) give the leading contribution to the entropy of the distribution \( \pi_n^i \), we say that the cutoff is entropy-driven. The role of the projection is outlining the drift, which enables to easily check the conditions of Theorem 3.3. We note that the drift can not be immediately detected from the original stationary distribution because this is uniform over the vertices of the hypercube.

### 3.7 MEAN-FIELD ISING MODEL WITH GLAUBER DYNAMICS

In the mean-field Ising model \( n \) binary spins are displaced on the vertices of a complete graph \( K_n \). The complete graph gives the neighbourhood structure of the physical model, i.e., each spin ‘feels’ and interacts with the remaining \( n-1 \). The set of all possible \textit{configurations} is \( \mathcal{X}_n = \{ +1, -1 \}^n \). The energy of a configuration \( \mathcal{X}_n \ni \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \) is

\[
H(\sigma) = -\frac{1}{n} \sum_{i<j} \sigma_i \sigma_j . \tag{3.46}
\]

The Glauber dynamics for this model is defined as follows. At each step a site \( i \in \{1,2,\ldots,n\} \) is picked up uniformly at random, and then \( \sigma_i \) is updated to the values \( +1 \) or \( -1 \) according to the following probabilities:

\[
p_+ = \frac{e^{\beta S(i)}}{e^{\beta S(i)} + e^{-\beta S(i)}} , \tag{3.47}
\]

\[
p_- = \frac{e^{-\beta S(i)}}{e^{\beta S(i)} + e^{-\beta S(i)}} , \tag{3.48}
\]

where \( S(i) = \frac{1}{n} \sum_{j \neq i} \sigma_j \) is the so-called \textit{local field}. The parameter \( \beta \) has the physical meaning of the system inverse temperature; the higher its value, the stronger the role of the energy over the entropy in establishing the equilibrium state. The limiting case \( \beta = 0 \) coincides with the lazy random walk on the hypercube discussed in Section 3.6, where all the spins are independently updated and equivalent from an energy-landscape point of view. By reversibility, the unique stationary measure of the Glauber chain is

\[
\pi_n(\sigma) = \frac{e^{-\beta H(\sigma)}}{Z_{n,\beta}} , \tag{3.49}
\]

where \( Z_{n,\beta} = \sum_{\sigma' \in \Omega_n} e^{-\beta H(\sigma')} \) is the so-called \textit{partition function}.

In [LLP10] the authors proved cutoff for the mean-field Ising model evolving in accordance with the Glauber dynamics. Here we give an alternative proof of the existence of the cutoff by means of Theorems 3.3 and 3.4. The computations needed to achieve the goal within this framework are a bit shorter. A generalisation of this result to the non-symmetrical case, i.e., when a constant magnetic field is added, is likely to be obtained with little effort.

Let us define the \textit{magnetisation} of a configuration \( \sigma \) as

\[
m(\sigma) = \frac{1}{2} \sum_i \sigma_i . \tag{3.50}
\]
Remark 3.10. The magnetisation takes values on the set \( \Omega_n = \{-\frac{n}{2}, -\frac{n}{2} + 1, \ldots, \frac{n}{2} - 1, \frac{n}{2}\} \). We note that the magnetisation is usually defined as

\[
m(\sigma) = \frac{1}{2n} \sum_{i=1}^{n} \sigma_i, \tag{3.50a}
\]

while (3.50) is typically referred to as the total spin. For computational convenience we do not follow the usual nomenclature, but we consider (3.50) rather than (3.50a) as the magnetisation of the configuration \( \sigma \).

We can rewrite the Hamiltonian (3.46) in terms of the magnetisation by

\[
m^2(\sigma) = \frac{1}{4} \left( \sum_i \sigma_i \right) \left( \sum_j \sigma_j \right) = \frac{n}{4} - \frac{n}{2} H(\sigma),
\]

so that

\[
H(m(\sigma)) = -\frac{2m^2(\sigma)}{n} + \frac{1}{2}.
\]

The update probabilities and the stationary distribution now take the form

\[
p_+ = \frac{\frac{2n}{n} (m(\sigma) - \sigma_i)}{e^{\frac{2n}{n} (m(\sigma) - \sigma_i)} + e^{-\frac{2n}{n} (m(\sigma) - \sigma_i)}},
\]

\[
= \frac{1}{1 + e^{-\frac{2n}{n} (m(\sigma) - \sigma_i)}}, \tag{3.47a}
\]

\[
p_- = \frac{\frac{2n}{n} (m(\sigma) - \sigma_i)}{e^{\frac{2n}{n} (m(\sigma) - \sigma_i)} + e^{-\frac{2n}{n} (m(\sigma) - \sigma_i)}},
\]

\[
= \frac{1}{1 + e^{\frac{2n}{n} (m(\sigma) - \sigma_i)}}, \tag{3.48a}
\]

\[
\tau_n(m(\sigma)) = \frac{e^{\frac{2n}{n} m^2(\sigma)}}{Z_{n,\beta}}. \tag{3.49a}
\]

We can consider the magnetisation chain alongside the Glauber chain. This is a new BDC \( X_n^\sigma \) with state space \( \Omega_n^\sigma = \{-\frac{n}{2}, -\frac{n}{2} + 1, \ldots, \frac{n}{2} - 1, \frac{n}{2}\} \) and transition rates

\[
p_n^\sigma(k, k + 1) = p_k = \frac{n-k}{n} \frac{1}{1 + e^{-\frac{4\beta}{n} (k+1)}},
\]

\[
p_n^\sigma(k, k) = r_k = \frac{1}{2} + \frac{k}{n} \tanh \left( \frac{4\beta}{n} (k+1) \right), \tag{3.51}
\]

\[
p_n^\sigma(k, k - 1) = q_k = \frac{n-k}{n} \frac{1}{1 + e^{-\frac{4\beta}{n} (k-1)}}.
\]

Using standard techniques, it is possible to show that the magnetisation chain is actually the projection of the Glauber chain according to the equivalence relation

\[
\sigma \sim \sigma' \iff m(\sigma) = m(\sigma'),
\]

see for example [Pre08, Thm. 5.1.4.1]. Let us then consider the Glauber chain started with an initial distribution \( \mu_0^\sigma \) on \( X_n^\sigma \) such that \( \mu_0^\sigma(\sigma) = \mu_0^\sigma(\sigma') \) whenever \( \sigma \sim \sigma' \). At first glance this condition may look quite strong, but there are a few reasons why to require it. First of all, it makes sense from a physical point of view because we can measure the magnetisation of the system but can not discover by any means the exact spin configuration. Also, according to the statement of Theorems 3.3 and 3.4, we are free to choose the initial
Therefore, for such a \( \mu_n^0 \) the hypotheses hold. Eventually, if we decided to fix \( \mu_n^0 \) and adopt the worst-case scenario introduced in (2.148), the initial configuration to be considered would be either \( \sigma_+ = (1, 1, \ldots, 1) \) or \( \sigma_- = (-1, -1, \ldots, -1) \). Both configurations satisfy the requirement above as the equivalence classes they belong to are composed of a single element, themselves.

Let us consider together the Glauber chain and its projection, i.e., the magnetisation chain that has initial distribution \( \mu_n^{t,0} \) and stationary measure \( \pi_n^\circ \) equal to

\[
\mu_n^{0,\circ}(k) = \sum_{\sigma : m(\sigma) = k} \mu_n^0(\sigma),
\]

\[
\pi_n^\circ(k) = \sum_{\sigma : m(\sigma) = k} \pi_n(\sigma) = \frac{e^{2n \beta k}}{Z_n,\beta} \left( \frac{n}{2} + k \right).
\]

It is not difficult to prove that the condition \( \mu_n^{t,0}(\sigma) = \mu_n^0(\sigma') \) for \( \sigma \sim \sigma' \) leads to \( \mu_n^t(\sigma) = \mu_n^t(\sigma') \) for each \( t \geq 0 \) and \( \sigma \sim \sigma' \), which in turn infers that

\[
\| \mu_n^t - \pi_n \|_{TV} = \| \mu_n^{t,0} - \pi_n^\circ \|_{TV} \quad \forall t \geq 0.
\] (3.52)

In other words, the Glauber chain exhibit cutoff if and only if the magnetisation chain does. We next prove cutoff for the magnetisation chain.

Let us fix \( \theta \geq 1 \) and define

\[
A_{n,\theta} = \left\{ k \in \Omega_n : -\theta \sqrt{\frac{n}{1 - \beta}} \leq k \leq \theta \sqrt{\frac{n}{1 - \beta}} \right\}.
\]

For \( k \in A_{n,\theta} \), we can estimate \( \pi_n^\circ(k) \) as follows by Stirling’s formula:

\[
\left( \frac{n}{2} + k \right) = \frac{n^n \sqrt{2\pi n} (1 + O(n^{-1}))}{(\frac{n}{2} + k)^{\frac{n}{2} + k} (\frac{n}{2} - k)^{\frac{n}{2} - k} 2\pi \sqrt{n^\frac{n^2}{4} - k^2}}
\]

\[
= \frac{2^{n+\frac{1}{2}}}{\sqrt{\pi n} \left(1 - \frac{4k^2}{n^2}\right)} \left(1 + O(n^{-1})\right)
\]

Then, we pass to the logarithm and use its analytic expansion

\[
\log \left( \frac{1}{1 + \frac{2k}{n} \left(1 + \frac{4k^2}{n^2}\right)} \left(1 - \frac{2k}{n} \left(1 - \frac{4k^2}{n^2}\right)\right) \right)
\]

\[
= - \frac{n}{2} \left[ \log \left(1 - \frac{4k^2}{n^2}\right) + \frac{2k}{n} \log \left(1 + \frac{2k}{n}\right) - \frac{2k}{n} \log \left(1 - \frac{2k}{n}\right) \right],
\]

\[
= - \frac{n}{2} \left[ \sum_{i \geq 1} \left( \frac{2k}{n} \right)^{2i} \frac{1}{2i^2 - 1} \right].
\]

Therefore, for \( k \in A_{n,\theta} \),

\[
\pi_n^\circ(k) = \frac{2^{n+\frac{1}{2}}}{Z_{n,\beta} \sqrt{\pi n}} e^{-\frac{2k(1-\beta)}{n} \sum_{i \geq 1} \left( \frac{2k}{n} \right)^{2i} 2i^{-1} \log \left(1 + O(n^{-1})\right)},
\]

\[
= \frac{2^{n+\frac{1}{2}}}{Z_{n,\beta} \sqrt{\pi n}} e^{-\frac{2k(1-\beta)}{n} k^2 \left(1 + O(n^{-1})\right)},
\] (3.53)
that is to say, \( \tau_n^\beta(k) \) is very close to a normal law \( N\left(0, \frac{1}{2} \sqrt{\frac{n}{\beta}}\right) \) for \( k \in \mathcal{A}_{n,0} \).

This means that (3.26) holds because there exists a positive constant \( c_\beta \) such that, for \( n \) sufficiently large,

\[
\tau_n \left(A^c_{n,0}\right) < \frac{c_\beta}{n^2}.
\]  

(3.54)

**Remark 3.11.** In this model the Gaussian structure of \( \pi_n^\beta \) is given both by energetic and entropic contributions. Energy and entropy appear together at the exponent of \( e^{-\frac{2(1-\beta)}{n}k^2} \) in the guise of the Helmholtz free-energy times \( \beta \). Similarly to the Ehrenfest Urn model, we say in this case that the cutoff is free-energy-driven.

Let us now suppose that the Glauber chain is started at time zero with magnetisation \( \frac{n}{2} \), that is, \( \mu_n^0 = \delta_{\sigma,\sigma^*} \) and \( \mu_n^{1,0} = \delta_{\sigma,2} \). This assumption guarantees that (3.52) holds. As usual, let us define \( \zeta_n^\beta \) as the hitting time of \( \mathcal{A}_{n,0} \) and \( \zeta_n^1 \) as the hitting time of \( \mathcal{A}_{n,1} \). Lengthy but straightforward calculations (deferred to Appendix A) show that

\[
\mathbb{E} \left[ \zeta_n^1 \right] = \frac{1}{2(1-\beta)} n \log n,
\]

\[
\mathbb{E} \left[ \zeta_n^1 - \zeta_n^0 \right] = (1 + \log \theta) O(n),
\]

and that \( \text{Var}[\zeta_n^1] = O(n^2) \). Therefore, hypotheses (3.27)-(3.38) are satisfied. Moreover, if we take \( \Delta_n = \Theta(n) \) then also \( \delta_n \) is of order \( n \), and both (3.30) and (3.31) are fulfilled. Theorem 3.3 then gives us (3.32), and we are left to verify the hypotheses of Theorem 3.4 with \( \delta_n = \Theta(n) \).

**Remark 3.12.** Since for \( \beta = 0 \) the magnetisation chain reduces to the Ehrenfest chain, the following argument holds as well for the Ehrenfest Urn model presented in Section 3.5.

Let us take \( h(\theta) = \theta^\frac{1}{4} \) and consider the next coupling, a tetra-variate process \( (Z_n^0, W_n^0, Z_n^+, Z_n^-) \) where each component is a copy of the magnetisation chain and

\[
\begin{align*}
Z_n^0 &= z_0 = \frac{1}{2} \sqrt{-\frac{n}{\beta}}, & Z_n^{+,-} &= z_0 = h(\theta), \\
W_n^0 &= \pi_n^\beta, & Z_n^{0,-} &= -z_0 = h(\theta),
\end{align*}
\]

for a given fixed \( \theta > 1 \). Let any of the four chains evolve using the same random mapping representation (3.10) and the same sequence of Independent and Identically Distributed (IID) random updates \( u_t \sim \mathcal{U}(0,1) \). To illustrate the transition probabilities, let us consider for instance the chain \( Z_n^+ \) and suppose that \( Z_n^+ = k \). Then,

\[
\begin{align*}
\text{for } k \geq 0 & \quad \begin{cases} 
Z_n^{t+1} = Z_n^t + 1, & \text{if } 0 \leq u < p_k, \\
Z_n^{t+1} = Z_n^t, & \text{if } p_k \leq u \leq 1 - q_k, \\
Z_n^{t+1} = Z_n^t - 1, & \text{if } 1 - q_k < u \leq 1;
\end{cases} \\
\text{for } k < 0 & \quad \begin{cases} 
Z_n^{t+1} = Z_n^t - 1, & \text{if } 0 \leq u < q_k, \\
Z_n^{t+1} = Z_n^t, & \text{if } q_k \leq u \leq 1 - p_k, \\
Z_n^{t+1} = Z_n^t + 1, & \text{if } 1 - p_k < u \leq 1.
\end{cases}
\end{align*}
\]

\* Cf. Remark 3.9 on page 47.
The coupling that we use to apply Theorem 3.4 is the restriction of the coupling \((Z_n^0, W_n^0, Z_n^{+1}, Z_n^{-1})\) to its first two components, \(Z_n^0\) and \(W_n^0\). Thus, we define \(\gamma_n = \min\{t \geq 0 : Z_n^t = W_n^t\}\) and recall Remark 3.7. By a careful analysis of (3.51) – noticing that \(r_k \geq \frac{1}{Z}\) and \(p_k = q_{-k}\), in particular – such a scheme ensures that any two components of the coupling undergoing a single-step transition maintain their relative partial order. Indeed, according to the random mapping representation above, it is impossible that two chains at distance 1 will undergo a one step transition that would change their relative order.

Hence, the evolution scheme described above has the following **sandwiching properties**

1. \(Z_n^{+t} = -Z_n^{-t}\);
2. \(Z_n^{-t} \leq Z_n^t \leq Z_n^{+t}\);
3. \(Z_n^{-t} \leq W_n^t \leq Z_n^{+t}\) provided that \(W_n^0 \in A_{n,h(\theta)}\).

Using (3.54),

\[
\mathbb{P} \left( \gamma_n > t \mid Z_n^0 = z_0 \right) = \mathbb{P} \left( \gamma_n > t \mid Z_n^0 = z_0, W_n^0 \in A_{n,h(\theta)} \right) \mathbb{P} \left( W_n^0 \in A_{n,h(\theta)} \right) + \mathbb{P} \left( \gamma_n > t \mid Z_n^0 = z_0, W_n^0 \notin A_{n,h(\theta)} \right) \mathbb{P} \left( W_n^0 \notin A_{n,h(\theta)} \right),
\]

\[
\leq \mathbb{P} \left( \gamma_n > t \mid Z_n^0 = z_0, W_n^0 \in A_{n,h(\theta)} \right) + \frac{c_\beta}{h(\theta)^2}.
\]

Therefore, by the sandwiching properties,

\[
\mathbb{P} \left( \gamma_n > t \mid Z_n^0 = z_0, W_n^0 \in A_{n,h(\theta)} \right) \leq \mathbb{P} \left( \tau_n(0) > t \mid Z_n^{+0} = z_0^+ \right),
\]

where \(\tau_n(0) = \min\{t \geq 0 : Z_n^{+t} = Z_n^{-t} = 0\}\). We note that \(Z_n^{+t}\) has a drift towards 0. Accordingly, it can be coupled with a lazy uniform random walk \(R_n^0\) such that

\[
R_n^0 = z_0^+,
\]

\[
\mathbb{P} \left( \tau_n(0) > t \mid Z_n^{+0} = z_0^+ \right) \leq \mathbb{P} \left( \tau_n(0) > t \mid R_n^0 = z_0^+ \right),
\]

where \(\tau_n(0) = \min\{t \geq 0 : R_n^t = 0\}\). It is now possible to use the classical result for random walks:

\[
\mathbb{P} \left( \tau_n(0) > t \mid R_n^0 = z_0^+ \right) \leq \frac{c z_0^+}{\sqrt{t}}.
\]

Thus, we have found that Theorem 3.4 holds with \(\delta_n = n\).
This chapter focuses on the thermalisation contribution we have introduced in Chapter 3. In Section 3.2 we have described it as the time necessary for the chain to forget the past trajectory after it has reached the bulk of the stationary measure. We have also interpreted the thermalisation in terms of energy landscapes, within this picture the bulk is the bottom of an energy well.

In many statistical mechanics models there exists a natural link between the stationary measure and the energy well. The mean-field Ising model, Section 3.7, gives an example of such a link. After the projection onto the magnetisation chain, the Gibbs measure (\(3.49\)) can be written as (\(3.53\)), from which we immediately see that the bulk is a region of size \(O(\sqrt{n})\) centred around the state \(m(\sigma) = 0\). In such a region the energy well is close to a parabola, and its bottom is flat. In particular, according to (\(3.51\)), for \(k = O(\sqrt{n})\),

\[
\begin{align*}
p_k & = P_n(k,k+1) = \frac{1}{4} + O\left(n^{-1/2}\right), \\
r_k & = P_n(k,k) = \frac{1}{2} + O\left(n^{-1}\right), \\
q_k & = P_n(k,k-1) = \frac{1}{4} + O\left(n^{-1/2}\right).
\end{align*}
\]

In other words, the trajectory of the chain within the bulk will asymptotically look like the trajectory of a lazy random walk, i.e., a diffusion process. The thermalisation can then be interpreted as the diffusive behaviour at the bottom of the well.

In the following we study two non-classical models designed to highlight the meaning of the thermalisation contribution. The first is a non-reversible random walk on a cylindrical lattice. Here the chain has a constant drift along the axis that is responsible for the stationary measure to concentrate on the bulk, a small volume at one end of the cylinder for this model. The study of the BDC obtained by projecting the model onto the cylinder axis ensures that the chain hits the bulk in a quasi-deterministic way; further, the projected chain exhibits cutoff. However, the same result is not automatically guaranteed for the original chain as it needs more time to thermalise at the bottom of the cylinder. Depending on the growth rate of the bulk, both in the radial and axis directions, the thermalisation contribution may grow faster than the expected hitting time of the bulk itself. In this case the quasi-deterministic hitting of the bulk – key feature of cutoff – is still present, but the phenomenon is swallowed up by the window because (2.18) fails to hold.

The second model we analyse is a one-dimensional random walk, partially biased and partially diffusive. In this model the MC has a strong drift towards a region where the transition probabilities are perfectly symmetric. Due to the presence of such a strong drift, the cutoff phenomenon can not be destroyed by a mechanism like the one in the previous example. In addition, the window order may be controlled by suitably choosing the size of the diffusive region.

* Cf. the Gibbs measure (3.49) and the energy (3.46).
4.1 Non-reversible Random Walk on a Cylinder

Consider a family of Markov chains \( \{ \Omega_n, X^t_n, P_n, \pi_n, \mu_n^0, \mu_n^1 \} \) having space state

\[
\Omega_n = \{ (h, \phi) : h \in \{0, 1, \ldots, l - 1\}, \phi \in \{0, 1, \ldots, m - 1\} \},
\]

with \( |\Omega_n| = n = l \cdot m \). As stated more precisely below, we think of \( \Omega_n \) as a cylindrical lattice of volume \( n \) having height \( l \) and base circumference of length \( m \). The entries of the transition kernel \( P_n \) are the following: for \( i, j \in \Omega_n \) with \( i = (h, \phi) \) and \( j = (h', \phi') \),

\[
P_n(i, j) = \begin{cases} \frac{4}{2r}, & \text{if } \phi' = \phi, h' = h - 1 \text{ and } h \neq 0, \\ & \text{or } \phi' = \phi, h' = h \text{ and } h = 0, \\ \frac{1-q}{2r}, & \text{if } \phi' = \phi, h' = h + 1 \text{ and } h \neq l - 1, \\ & \text{or } \phi' = \phi, h' = h \text{ and } h = l - 1, \\ \frac{r}{2}, & \text{if } h' = h, \phi' = \phi + 1 \text{ mod } m, \\ \frac{1-r}{2}, & \text{if } h' = h, \phi' = \phi - 1 \text{ mod } m, \\ 0, & \text{otherwise}, \end{cases}
\]

where \( r \in (0, 1) \) and \( q \in (\frac{1}{2}, 1) \). Let us define the net vertical drift felt by the chain as \( \beta = \frac{2q - 1}{2} \).

Remark 4.1. The transition matrix (4.1) naturally induces on \( \Omega_n \) an undirected graph \( G(V, E) \), with vertex set is \( V = \Omega_n \) and edge set \( \Omega_n \times \Omega_n E = \{(u, v) \text{ s.t. } P(X^t_n = u | X^t_n = v) > 0\} \). Such a graph can be thought of as a cylindrical lattice of volume \( n \), with \( l \) layers composed of \( m \) points each. The neighbourhood structure induced by the undirected graph \( G(V, E) \) naturally introduces on \( \Omega_n \) the metric of the shortest path (2.3).

Each chain of the family is an irreducible and aperiodic chain, so there exists a unique invariant measure \( \pi_n = \pi_n P_n \). Since the model has an evident radial symmetry, it is reasonable to expect that

\[
\pi_n(h, \phi) = \pi_n(h, \phi') \ \forall \ \phi, \phi' \in \{0, 1, \ldots, m - 1\}.
\]

Hence, let us look for \( \pi_n \) in the form

\[
\pi_n(h, \phi) = f_n(h) \quad \text{where} \quad f_n(h + 1) = \alpha f_n(h).
\]

By definition of \( \pi_n \) and (4.1), for \( h \neq 0, l - 1 \),

\[
\pi_n(h, \phi) = \frac{1-r}{2} \pi_n(h, \phi + 1 \text{ mod } m) + \frac{r}{2} \pi_n(h, \phi - 1 \text{ mod } m) + \frac{q}{2} \pi_n(h + 1, \phi) + \frac{1-q}{2} \pi_n(h - 1, \phi).
\]

A stationary distribution of the form (4.2) is a solution of the last equation only if

\[
\alpha = 1 \quad \text{or} \quad \alpha = \frac{1-q}{q}.
\]

The value \( \alpha = \frac{1-q}{q} \) satisfies \( \pi_n = \pi_n P_n \) also for \( h = 0 \) and \( h = l - 1 \). Thus,

\[
\pi_n(h, \phi) = \alpha^h f_n(0) = \left( \frac{1-q}{q} \right)^h f_n(0).
\]
By normalisation,
\[ f_n(0) = \pi_n(0, \phi) = \frac{1 - \alpha}{m(1 - \alpha^l)} \approx \frac{2q - 1}{m q}, \]
where the last approximation holds for sufficiently large \( l \).

Remark 4. Therefore we can study the hitting time of any layer \( X \) and consider a new MC \( X^\pi_n \), which takes values on \( \Omega^\pi_n = \{0, 1, \ldots, l - 1\} \) and has transition matrix
\[
P^\pi_n(i, j) = \begin{cases} 
\frac{1}{2}, & \text{if } i = j \text{ and } i \neq 0, 1 - 1, \\
\frac{1}{2} + q, & \text{if } i = j = 0, \\
\frac{2}{2} - q, & \text{if } i = j = 1 - 1, \\
\frac{q}{2}, & \text{if } j = i - 1 \text{ and } i \neq 0, \\
\frac{1}{2} - q, & \text{if } j = i + 1 \text{ and } i \neq 1 - 1, \\
0, & \text{otherwise}.
\end{cases}
\]

The chain \( X^\pi_n \) is the projection of \( X^t_n \) according to \( \sim \). The stationary measure \( \pi^\pi_n(i) \) of the lumped chain is found summing \( \pi_n(v) \) over the elements \( v \) that belong to the equivalence class \([i]\). Since every equivalence class, i.e., every layer, contains exactly \( m \) points,
\[
\pi^\pi_n(i) \approx \frac{2q - 1}{q} \left( \frac{1 - q}{q} \right)^i, \quad i \in \{0, 1, \ldots, l - 1\}.
\]

Remark 4.2. The stationary measure \( \pi^\pi_n \) is obviously reversible with respect to \( P^\pi_n \). The latter property does not hold in general for the original chain \( X^t_n \), i.e., the equilibrium measure \( \pi_n \) is not reversible with respect to \( P_n \). To see this, it suffices to take any two states \( v, w \in \Omega_n \) such that \( h(v) = h(w) \) and \( |\phi(v) - \phi(w)| = 1 \). By (4.2), \( \pi_n(v) = \pi_n(w) \) but according to (4.1), \( P(v, w) \neq P(w, v) \) unless \( r = \frac{1}{2} \).

Remark 4.3. We have introduced the lumped chain \( X^\pi_n \) since it can be coupled to \( X^t_n \) in such a way that
\[
h(X^t_n) = X^\pi_n \quad \forall \ t \geq 0.
\]

Therefore we can study the hitting time of any layer by simply considering a one-dimensional chain. Nevertheless, it is opportune to stress that the study of the cutoff phenomenon for \( X^t_n \) can not be reduced to the study of the cutoff for \( X^\pi_n \) since in general the identity (3.45) does not hold. Let us consider, indeed, the initial distribution \( \mu^t_0 = \delta_{v_0} \) with \( h(v_0) = l - 1 \). It represents the worst case scenario for the behaviour of the total-variation distance. Then, (3.45) is false for any finite \( t \), but it is still possible to prove cutoff with relative ease using Theorems 3.3 and 3.4.
We are now ready to prove cutoff for the MC family defined by \((4.1)\). The outline of the proof is the following. We use Theorem 3.3 and Theorem 3.4, checking all the hypotheses on the hitting time of the family \(A_{n,0}\) by means of the projected chain and Remark 4.3. Since the projected chain is a BDC, formulas \((2.43)\)–\((2.47)\) can be exploited. This means that we implicitly prove cutoff for the projected chain too. We then focus on the thermalisation contribution to the cutoff window. What we discover is that the thermalisation term is always negligible for the projection \(X^\Delta_{n,t}\) while its behaviour is critical in establishing cutoff for the original chain \(X^\gamma_n\).

Let us define the following family of sets:

\[ A_{n,0} = \{ v \in \Omega_n : h(v) < \sqrt{n} \}. \]

With this definition, \(A_{n,0}\) is the union of the \(\sqrt{n}\) bottom layers and \(A_{n,1}\) is just the bottommost layer. From Remark 4.3 and formulas \((2.43)\)–\((2.47)\), the hitting time \(\zeta_n^0\) of the set \(A_{n,0}\) has the following expectation and variance:

\[
\mathbb{E}\left[ \zeta_n^0 \right] = \beta^{-1}(1 - \sqrt{\delta}) + O(\alpha^1),
\]

\[
\text{Var}\left[ \zeta_n^0 \right] = O(1).
\]

To use Theorem 3.3 we want to study the behaviour of these quantities in the limit for \(n \to \infty\). Since \(n = 1 \cdot m\), we can let the volume of the cylinder grow by extending its height or enlarging its diameter, or yet increasing both simultaneously. To this extent, let us fix \(\varepsilon > 0\) and consider

\[ m = m(n) = n^{1-\varepsilon} \quad \text{and} \quad l = l(n) = n^{1-\varepsilon}.
\]

The main result of the present section is the cutoff for \(X^t_n\) with \(a_n = \beta^{-1} n^{1-\varepsilon}\) and \(b_n = \Theta(n^{2\varepsilon} + n^{\frac{1\varepsilon}{2}})\).

With the usual notation take \(\Delta_n = m^2 = n^{2\varepsilon}\), this choice verifies the remaining hypothesis of Theorem 3.3, i.e., \((3.30)\) and \((3.31)\), and sets the candidate cutoff-window order to

\[ \delta_n = \Theta(m^2 + \sqrt{l}) = \Theta(n^{2\varepsilon} + n^{\frac{1\varepsilon}{2}}). \]

Having in mind the proof of cutoff for \(X^t_n\), all we are left to deal with is the existence of a coupling \((Z^t_n, W^t_n)\) that satisfies the hypotheses of Theorem 3.4. This means we need to exhibit a coupling of \(Z^t_n\) and \(W^t_n\), both identical copies of \(X^t_n\), such that

\[ \lim_{\theta \to \infty} \lim_{n \to \infty} \mathbb{P}(\gamma_n > \theta \delta_n) = 0, \]

where \(\gamma_n = \min\{t \geq 0 : Z^t_n = W^t_n\}\) is the coalescence time. According to the statement of Theorem 3.4, \(W^0_n \sim \pi_n\) and \(Z^0_n\) is located on a point of the bottommost layer, i.e., \(h(Z^0_n) = 0\). The former condition is equivalent to \(h(W^0_n) \geq 0\) and distributed as a truncated geometric.

Consider the following distance\(^*\) between \(Z^t_n\) and \(W^t_n\):

\[ D^t_n = |h(Z^t_n) - h(W^t_n)| \]

\[ + \min\{|\phi(Z^t_n) - \phi(W^t_n)|, m - |\phi(Z^t_n) - \phi(W^t_n)|\}. \]

There exists a coupling \((Z^t_n, W^t_n)\), sketched for reference in Figure 15, such that

\[ \text{Cf. Remark 4.1} \]

\[ ^* \]
1. \( H^n_t = |h(Z^{1}_n) - h(W^{1}_n)| \) is a death-only chain on \( \{0, 1, \ldots, l - 1\} \), that is to say, \( H^{t+1} \leq H^n_t \).

2. \( H^n_s = 0 \) for each \( s \geq \gamma^H_n = \min\{t \geq 0 : H^n_t = 0\} \);

3. \( \gamma^H_n = \min\{t \geq 0 : h(W^n_t) = 0\} \);

4. \( \Phi^n_t = \min(|\phi(Z^n_t) - \phi(W^n_t)|, m - |\phi(Z^n_t) - \phi(W^n_t)|) \) is a symmetric \( r\)-lazy* random walk on \( \{0, 1, \ldots, \lceil \frac{m}{2} \rceil\} \);

5. \( \Phi^n_s = 0 \) for each \( s \geq \gamma^\Phi_n = \min\{t \geq 0 : \Phi^n_t = 0\} \).

The coalescence time satisfies

\[ \gamma_n = \max\{\gamma^H_n, \gamma^\Phi_n\} \leq \gamma^H_n + \gamma^\Phi_n, \]

and, by Markov’s inequality,

\[ \mathbb{P}(\gamma_n \geq \theta \delta_n) \leq \frac{\mathbb{E}[\gamma_n]}{\theta \delta_n} \leq \frac{\mathbb{E}[\gamma^H_n] + \mathbb{E}[\gamma^\Phi_n]}{\theta \delta_n}. \]

According to point 3 listed above and the transition probabilities of \( W^{n,t}_n = h(W^n_t) \),

\[ \mathbb{E}\left[ \gamma^H_n \left| h(W^n_t) = h' \right. \right] = \beta^{-1} h', \]

which yields,

\[ \mathbb{E}\left[ \gamma^H_n \right] = \beta^{-1} \mathbb{E}\left[ h(W^n_t) \right] \]

\[ = \beta^{-1} \sum_x \pi^n_0(x) \leq \beta^{-1} \frac{1 - q}{2q - 1}. \]  

* A MC is called \( r\)-lazy if \( P_n(i, i) \geq r \) for each \( i \in \Omega_n \).

Figure 15: Coupling scheme, the same random update is used for both \( Z^n_t \) and \( W^n_t \).
The two copies have the same probability to move to the upper or lower layer, except when one of the chains is on the topmost or bottommost layer. In the latter case the distance \( H^n_t \) has probability \( \frac{q}{2} \) to reduce by 1 while in the former it has probability \( \frac{1 - q}{2} \).
According to point 4, we get
\[ \mathbb{E} \left[ \gamma_n^\Phi \right] = \Theta(m^2). \] (4.5)

Equations (4.4) and (4.5) clearly infer (4.3), and the proof of cutoff for \( X^t_n \) is complete: the model exhibits cutoff at time
\[ a_n = \beta^{-1} \mathcal{t} = \beta^{-1} n^{1-\epsilon} \]
with window order
\[ b_n = \Theta(m^2 + \sqrt{\mathcal{t}}) = \Theta \left( n^{2\epsilon} + n^{1-\epsilon} \right). \]

Remark 4.4. The condition \( \frac{b_n}{a_n} = o(1) \) for \( n \to \infty \) is fulfilled if and only if \( \epsilon < \frac{1}{3} \). Within this constraint the chain exhibits cutoff with window order
\[
\begin{cases}
  b_n = \Theta \left( n^{1-\epsilon} \right), & \text{if } 0 < \epsilon \leq \frac{1}{3}, \\
  b_n = \Theta \left( n^{2\epsilon} \right), & \text{if } \frac{1}{3} \leq \epsilon < \frac{1}{2}.
\end{cases}
\]

Therefore, the value \( \epsilon = \frac{1}{3} \) gives the smallest cutoff window order achievable.

Remark 4.5. The case \( \epsilon = 0 \) is equivalent to increasing the cylinder volume by extending its height while keeping fixed its base diameter. It then corresponds to a higher-dimensional generalisation of the biased random walk on a segment, Section 3.1. Due to the obvious symmetry of the state space \( \Omega_n \) and Remark 4.3, the proof of cutoff for \( \epsilon = 0 \) can be either carried out as in [LPW06, §18.2.1] or by means of Theorems 3.3 and 3.4. The general case \( \epsilon > 0 \) represents a possible non-reversible, higher-dimensional extension of the biased random walk, where we clearly see the interplay discussed in Section 3.2 between the two contributions to the cutoff window.

The model we have just studied shows that the mixing mechanism operates on two different time scales, namely, the time needed by the chain to reach \( \Lambda^t_{n,1} \) and the time required for the thermalisation inside it. If the thermalisation contribution \( \Delta_n \) is of a smaller order with respect to the cutoff time \( \mathbb{E} \left[ \zeta^\Phi_n \right] \) then the process exhibits cutoff. Conversely, when \( \epsilon \) is such that the major contribution to the mixing time is the thermalisation at the bottom of the cylinder then the cutoff window grows too wide and blows out the cutoff effect. It may be pointed out that the projection is unaffected by the thermalisation issue, so one could believe that the thermalisation matters only in models with dimensionality higher than one. In the next section we disprove this potential claim.

4.2 Partially-diffusive random walk

Fix \( \epsilon \in (0, \frac{1}{3}) \), and consider the BDC chain \( X^t_n \) defined on the state space \( \Omega_n = \{0, 1, \ldots, n\} \) with initial position \( X^0_n = n \) and transition rates
\[
\begin{align*}
p_i &= P_n(i, i + 1) = \begin{cases} \frac{1}{n}, & \text{if } n^\epsilon \leq i \leq n, \\ \frac{1}{2}, & \text{if } 0 \leq i < n^\epsilon, \end{cases} \\
r_i &= P_n(i, i) = 1 - p_i - q_i, \\
q_i &= P_n(i, i - 1) = \begin{cases} \frac{1}{n}, & \text{if } n^\epsilon < i \leq n, \\ \frac{1}{2}, & \text{if } 1 \leq i \leq n^\epsilon. \end{cases}
\end{align*}
\] (4.6)
Outside the interval $[0, n^\varepsilon]$ the chain behaves like a biased random walk while it behaves like an unbiased one for $i \in [n^\varepsilon, n]$. At the end of the section we show that this model does not satisfies the **Strong Drift Condition**, sufficient to prove cutoff in the sense of (2.42) as shown in [BBF09]. Using Theorems 3.3 and 3.4, it is easy to demonstrate that this model actually does exhibit cutoff. In particular, we prove that the chain $X_n^t$, defined by (4.6), exhibits cutoff with $a_n = 2(1 - \varepsilon) n \log_2 n$ and $b_n = \Theta(n^{2\varepsilon} + n^{1-\varepsilon/2})$.

The stationary distribution $\pi_n$ is found to satisfy

$$\pi_n(i) = \begin{cases} c, & \text{for } 0 \leq i \leq n^\varepsilon, \\ c \left(\frac{1}{2}\right)^{i-n^\varepsilon}, & \text{for } n^\varepsilon < i \leq n, \end{cases}$$

where the constant $c$ is $\frac{1}{n^{1-\varepsilon}} + O \left(\frac{1}{n}\right)$. In order to use Theorem 3.3, it is enough to take the following family of nested subsets:

$$A_{n,0} = \{ i : 0 \leq i \leq n^{\varepsilon} \theta^{n^{2\varepsilon-1}} \}.$$

With this choice (3.26) holds and

$$\begin{align*}
E \left[ \zeta_n \right] &= 2(1 - \varepsilon) n \log_2 n, \\
E \left[ \zeta_n^0 - \zeta_0 \right] &= 2 n^{2\varepsilon} \log_2 \theta, \\
\text{Var} \left[ \zeta_n \right] &= O \left( n^{1-2\varepsilon} \right).
\end{align*}$$

The details of the calculations are presented for the sake of completeness in Appendix B. Choosing $\Delta_n = n^{2\varepsilon}$, we verify (3.30) and (3.31). Then, by Remark 3.5, we know that all the hypotheses of Theorem 3.3 hold.

Let us now consider a coupling $(Z_n^t, W_n^t)$ such that $Z_n^t$ and $W_n^t$ are two copies of $X_n^t$ with initial positions $Z_n^0 = n^\varepsilon$ and $W_n^0 \sim \pi_n$, respectively. Then, provided that the two chains have not yet collided, at each time we let the two copies evolve using the same sequence of IID random updates. Let $\gamma_n = \min\{ t \geq 0 : Z_n^t = W_n^t \}$ be the coalescence time and $Z_n^t = W_n^t$ for any $t \geq \gamma_n$. Then,

$$\begin{align*}
\mathbb{P} \left( \gamma_n > t \mid Z_n^0 = n^\varepsilon \right) &= \mathbb{P} \left( \gamma_n > t \mid Z_n^0 = n^\varepsilon, W_n^0 \leq n^\varepsilon \right) \mathbb{P} \left( W_n^0 \leq n^\varepsilon \right) \\
&\quad + \mathbb{P} \left( \gamma_n > t \mid Z_n^0 = n^\varepsilon, W_n^0 > n^\varepsilon \right) \mathbb{P} \left( W_n^0 > n^\varepsilon \right), \\
&\leq \mathbb{P} \left( \gamma_n > t \mid Z_n^0 = n^\varepsilon, W_n^0 \leq n^\varepsilon \right) + \frac{1}{n^\varepsilon},
\end{align*}$$

where (4.7) we have used $\mathbb{P} \left( W_n^0 > n^\varepsilon \right) = \frac{1}{n^\varepsilon}$ and bounded the other terms by 1.

Let us define

$$\tau_n^0 = \min\{ t \geq 0 : Z_n^t = 0 \}.$$

From Markov’s inequality,

$$\begin{align*}
\mathbb{P} \left( \gamma > t \mid Z_n^0 = n^\varepsilon, W_n^0 \leq n^\varepsilon \right) &= \mathbb{P} \left( \tau_n^0 > t \mid Z_n^t = n^\varepsilon \right), \\
&\leq \frac{\mathbb{E} \left[ \tau_n^0 \mid Z_n^0 = n^\varepsilon \right]}{t}.
\end{align*}$$

(4.8)
Formula \((2.43)\) yields
\[
E \left[ \tau_n^0 \bigg| Z_n^0 = n^\epsilon \right] = n^{2\epsilon} + O(n^{\epsilon}),
\]
and for \(t = \theta n^{2\epsilon}\), by Remark 3.7,
\[
\max_{z_0 \in \mathcal{A}_{n,1}} P \left( Y_n > \theta n^{2\epsilon} \bigg| Z_n^0 = z_0 \right) \leq \frac{2}{\theta}
\]
asymptotically, for \(n \to \infty\).

Therefore \((3.27)\) is verified and, with respect to the coupling above, \((3.37)\) follows from \((4.7)-(4.8)\) by choosing
\[
t = \theta \delta_n = 2\theta \left( n^{2\epsilon} + n^{1-\epsilon/2} \right).
\]
Thus, by Theorem 3.3 and Theorem 3.4, \(X_n^1\) exhibits cutoff with cutoff time
\[
a_n = E \left[ \ell_n^1 \right] = 2(1-\epsilon) n \log_2 n
\]
and cutoff window
\[
b_n = \begin{cases} 
\Theta(n^{1-\epsilon}), & \text{if } 0 < \epsilon \leq \frac{2}{5}, \\
\Theta(n^{2\epsilon}), & \text{if } \frac{2}{5} < \epsilon \leq \frac{1}{2}.
\end{cases}
\]

**Remark 4.6.** From \((4.9)\) we see that the choice \(\epsilon = \frac{2}{5}\) gives the smallest cutoff window order possible.

**Remark 4.7.** This example shows how decisive the definition of \(\{A_{n,0}\}\) is. One could be tempted in fact to try
\[
A_{n,0} = \{ i : 0 \leq i \leq \theta n^\epsilon \}
\]
because that scaling, linear in \(\theta\), has worked well for the Lazy Ehrenfest Urn model. This alternative definition would lead to an expected travelling time
\[
E \left[ \ell_n^1 - \ell_n^0 \right] = n \log \theta
\]
and force \(\Lambda_n\) (and consequently \(\delta_n\)) to be of order \(n\). Since \(\theta n\) steps are clearly sufficient for the chain started in \(n^\epsilon\) to achieve equilibrium, a non-optimal \(\Theta(n)\) cutoff window would be obtained.

At the beginning of the section we have mentioned that this model does not satisfy the Strong Drift Condition, i.e.,
\[
K_q = \inf_{n \in \mathbb{N}} \inf_{0 \leq i \leq n} q_i > 0,
\]
\[
\frac{K_n^2}{E \left[ T_{n-0}^{(n)} \right]} \to 0, \quad \text{as } n \to \infty,
\]
where \(q_i\) is the transition probability defined by \((4.6)\), \(T_{n-0}^{(n)}\) is the hitting time of zero starting from \(n\), and
\[
K_n = \sup_{1 \leq i \leq n} q_i E \left[ T_{i-1}^{(n)} \right] = \sup_{0 \leq i \leq n} \frac{\pi_n([i+1])}{\pi_n([i])}.
\]

Let us explain why the Strong Drift Condition does not hold here. First of all, \(X_n^1\) fails \((4.10)\) as \(K_q = 0\) for the model presented in this section. Nevertheless, it is clear from the discussion carried out in [BBF9] that the condition \(K_q > 0\) can be actually dropped if one replaces \((4.11)\) with
\[
\frac{K_n^2}{K_q E \left[ T_{n-0}^{(n)} \right]} \to 0, \quad \text{as } n \to \infty,
\]

\[\text{Thermalisation and Strong Drift Condition}\]
4.2 PARTIALLY-DIFFUSIVE RANDOM WALK

where $K^*_{q} = \inf_{0 \leq i \leq n} q_i$. The expected value of $T_{n \rightarrow 0}^{(n)}$ can be easily estimated as

$$E \left[ T_{n \rightarrow 0}^{(n)} \right] = O(n \log n + n^{2\epsilon}),$$

while $K_n$ can be bounded from below by $n^{\epsilon}$. Since $\epsilon > 0$,

$$\frac{K^2_n}{K_q n E \left[ T_{n \rightarrow 0}^{(n)} \right]} \geq \frac{n^{2\epsilon}}{2n} O \left( n \log n + n^{2\epsilon} \right) \xrightarrow{n \rightarrow \infty} \infty,$$

On the other hand, if we consider the chain (4.6) with $\epsilon = 0$ then the strong drift condition holds. Therefore, the causes of the Strong Drift Condition fail are to be looked for in the thermalisation phase. The energy landscape associated to the stationary distribution $\pi_n$ is an energy well whose bottom is the diffusive region $0 \leq i < n^{\epsilon}$. Hence, we see that the Strong Drift Condition is designed for energy wells with a non-flat bottom. In this framework it works perfectly well and gives cutoff according to Definition 2.7.

This definition has some very strong points, which we have already discussed at the end of Section 2.8. On the other hand, it completely overlook the cutoff window and it cannot take into account the potential thermalisation effects.
EXponentially Delayed Arrivals

In this chapter we temporarily leave aside the study of cutoff phenomena to introduce a discrete time queueing system with correlated arrivals. The arrival pattern originates as the random shifting of a deterministic schedule, the shifts being modelled as IID exponential delays $\xi_i$. The standard deviation of the delays is $\sigma$, its value is finite but much larger than the deterministic unit service time. For such a system, we find the (bivariate) generating function and solve the resulting functional equation in terms of a power series expansion in a parameter related to $\sigma^{-1}$. Later, in Chapter 6, we discuss some aspects of the cutoff phenomenon for this queueing system. For an application like a queue model, cutoff has a very interesting meaning; starting from a very congested situation, the system reaches equilibrium operational levels in a quasi-deterministic way. A generalisation of this arrival process to other-than-exponential delays is presented in Chapter 7.

5.1 THE EDA/D/1 QUEUEING SYSTEM

Let us consider a single-server queue with deterministic service time. The $i$-th customer arrives to the system at time $t_i = i + \xi_i$, \hspace{1cm} (5.1)

where $\xi_i$ are IID exponential random variables with parameter $\beta$. The delays’ Probability Density Function (PDF) is

\[
f_{\xi}(t) = \begin{cases} 
\beta e^{-\beta t}, & \text{if } t > 0, \\
0, & \text{otherwise}.
\end{cases}
\]

In the limit $\beta \to 0$ the arrival process (5.1) weakly converges to a Poisson process, whereas for small but fixed $\beta$ the arrivals are negatively autocorrelated, see [GNS11, Col76] and references therein. It may be noted that all the results in [GNS11] are given under the hypothesis of $f_\xi$ having compact support. As a matter of fact, that hypothesis does not play any role in establishing the convergence to a Poisson process so the very same analysis applies here too.

In this chapter we are interested in studying the system for fixed $\beta$. We also assume throughout the chapter an independent thinning to the arrival process, i.e., each customer can be deleted before joining the queue independently with probability $1 - \rho$. The thinning is used to model empty intervals in a constant stream of customers, see Section 7.1 for a discussion about this point in an applied context. The thinning is also a mathematical expedient to ensure the ergodicity of the resulting MC, see Remark 5.1 below. Also the thinned arrival process weakly converges to a Poisson process, but with parameter $\beta \rho$, see again [GNS11]. After D.G. Kendall (see below for his exact words) we name this arrival process Exponentially Delayed Arrivals (EDA).

Service can be delivered by the unique server only at discrete times. The length of the queue at time $t$, denoted by $n_t$, is the number of customers waiting to be served including the customer that is served precisely at time
Due to the thinning procedure, it is immediate to see that the traffic intensity of the system is given by $\rho$, see [GNS11] for details.

Using Kendall’s notation*, we refer to the system described above as EDA/D/1. This model is motivated by the description of many transportation systems, including vessel queueing in marine ports [GL63, JA03] and Air Traffic Management (ATM) [BVH01, GN10, GNS11, ISPT11], crane handling in dock operations [Dag90, Edm75], outpatient services [Bai52, CV03, Mer60, Mer73], and in general any system where scheduled arrivals are intrinsically subject to random variations. Further, the results presented in Chapter 7 strongly suggest that the model described above is very suitable for the description of the inbound air traffic over a large hub.

The same problem was investigated also by D.G. Kendall. In [Ken64, page 11] he remarked the great importance of systems with arrival pattern like (5.1): "[...] perhaps too much attention has been paid to rather uninteresting variations on the fundamental Poisson stream. As soon as one considers variations dictated by the exigencies of the real world, rather than by the pursuit of mathematical elegance, severe difficulties are encountered; this is particularly well illustrated by the notoriously difficult problem of late arrivals." Kendall also provided the following elegant interpretation: if the random variables $\xi_i$ are non-negative then the process defined by (5.1) is the output of the stationary D/G/$\infty$ queueing system. In particular, if the random variables $\xi_i$ are exponentially distributed then EDA/D/1 can be viewed as a 2-stage tandem queueing network

$$D/M/\infty \rightarrow D/1.$$  

However, this is not the approach followed in this research.

Some years later, under the hypothesis $\xi_i > 0$, Nelsen and Williams exactly characterised in [NW70] the distribution of the inter-arrival time intervals and the correlation coefficient between successive inter-arrival time intervals. They also gave an explicit expression of these quantities in the particular case of $\xi_i$’s exponentially distributed.

After the ’70s only approximations of the arrival process (5.1) [BC72, SD86] or numerical studies of its output [AA12, BVH01, NH12] seem to have appeared in the literature. In particular, in [GNS11] the authors presented a self-contained study of an arrival process like (5.1), assuming for $\xi_i$ a compact-support distribution. They also proposed an approximation scheme that keeps the correlation of the arrivals and is able to compute in a quite accurate way the quantitative features of the queue. To the best of our knowledge, a queueing system with arrivals described by (5.1) still remains an open problem and the best results obtained so far are due to C.B. Winsten in 1959.

EDA/D/1 is an example of a queueing system with correlated arrivals, a subject broadly studied in past years. There are many ways to impose

* According to the Kendall’s notation, any queueing system with an infinite queue capacity may be described by the code $a/b/c$. The symbol $a$ represents the inter-arrival distribution, $b$ the service-time distribution, and $c$ the number of servers. Some notable values of $a$ and $b$ are: M, Poissonian distribution; D, deterministic; G, general distribution.
a correlation to the arrival process. For instance, the parameters of the process may depend on their past realisation, as in [Dre99], or on some on/off sources, as in [WB99]. Another relevant example of a queue model with correlated arrivals is the so-called Markov Modulated Queueing System (MMQS). In MMQSs the parameters are driven by an independent external Markovian process, see [AK03, AK00, CB98, Luc91, NBH08, PPT09] and references therein. Our model shares with MMQSs the property that one can define an external and independent Markovian process that drives the arrival rates. However, we see in Section 5.2 that the output of this external drive also determines the evolution of the queue length. More precisely, our system can be seen as a single-server queue with deterministic service time while the arrivals are given by the reneging* from an auxiliary queue representing the customers that are late at time, see (5.3) below. Due to the memoryless property of the exponential delays, each customer late at time \( t \) may be late in the unit time slot \( (t, t+1] \) independently and with probability \( q = e^{-\beta} \). This means that the aforesaid reneging only happens at integer times, and clients perform synchronous independent abandonments leading to binomial transitions in the number of late customers.

In Section 5.2 we show that the EDA/D/1 model can be described as a two-dimensional Markov chain representing the queue length and the number of late customers. There exists an extensive literature about two-dimensional Markov models. Many methods for attacking the problem are available under two assumptions, namely, spatial homogeneity and finiteness of at least one marginal chain, see [BLM05, GHT00, Gra02, LR99, MC95, Neu89, Neu95]. Unfortunately, the Markov chain defined in Section 5.2 does not satisfy any of the mentioned requirements.

When both components of the Markov chain are infinite but space homogeneity is still ensured, the problem is typically attacked by reduction to a Riemann-Hilbert Boundary Value Problem (BVP). BVPs represent a broadly studied subject and several techniques have been developed in the last decades to solve them; among these we mention the uniformisation technique [Kin61], conformal mappings [Coh88, CB83, Fl79], the compensation method [AWZ93], and the Power Series Approximation (PSA) [Bla87a, Bla87b, Koo97, HKVDR88]. Usually, PSA is used to obtain the generating function in terms of a power series in the load \( \rho \) although different parameters may be used, see e.g. [WvLB10]. A power series expansion in a parameter different from \( \rho \) is also the strategy we adopt in what follows.

The second constraint that EDA/D/1 fails to meet is spatial homogeneity. The lack of homogeneity is due to the reneging with binomial transitions mentioned above. This kind of transitions are often encountered in Mathematical Biology [AELH07, BGR82, Eco04].

To the best of our knowledge, the functional equation (5.9) below has never previously appeared in the literature. Yet it is possible to mark some analogies with the functional equation in [EK09, EKR10, Kap11], the most important being the following. In both equations the right hand side exhibits the generating function computed in a convex combination in the parameter \( q = e^{-\beta} \), that is, the probability of each independent abandonment; this feature plays an important role in the proof of Theorem 5.2. Other examples of chains with binomial transitions may be found in [AEK09, AY06, Neu94, Yec07].

The remainder of the chapter is organised as follows. In Section 5.2 we derive an equation for the bivariate generating function step by step. Then

---

* A customer performs a reneging when it abandons the queue before being served.
in Sections 5.3-5.5 we develop an iterative method to compute the explicit expression for the generating function in terms of a power series and discuss some aspects of the solution, explicitly computing the first terms of its expansion; Section 5.6 contains some additional results on the coefficients of the power series, and a proof of the analyticity of the generating function with respect to the parameter \( q = e^{-\beta} \); finally, Sections 5.7-5.9 are devoted to the discussion of supplementary details of the method and to some closing remarks.

### 5.2 The Generating Function of an EDA/D/1 Queue

The process \( n_t \) describes the length of the queue at time \( t \) and is governed by the recursion

\[
n_{t+1} = n_t + m_{t,t+1} - (1 - \delta_{n_t,0}),
\]

where \( m_{t,t+1} \) is the number of arrivals in the interval \( (t, t+1] \), and the term \( 1 - \delta_{n_t,0} \) represents the action of the service: if at time \( t \) the queue is non-empty then the first waiting customer is served.

The quantity \( m_{t,t+1} \) depends in general on the whole previous history of the system. Indeed, if for some large value of \( T \) \( m_{t,s+1} = 0 \) for any \( s \in \{ t - T, t - T + 1, \ldots, t - 1 \} \) then \( m_{t,t+1} \) is large with great probability. Conversely, if in the recent past the values of \( m_{t,s+1} \) have been large then \( m_{t,t+1} \) is expected to be small. This suggests that the arrival process is negatively autocorrelated, proof of this property can be found in [GNS11]. Hence, the recursion (5.2) does not depend only on the present value of \( n_t \), and the memory of the process is infinite since \( T \) can be arbitrarily large.

Let us now denote by \( l_t \) the number of customers that have not yet arrived at time \( t \), that is to say,

\[
l_t = \left| \left\{ 0 \leq i \leq t \quad \text{s.t.} \quad \xi_i > t - i \right\} \right|.
\]

Let us next define \( p = \int_0^t f_\xi(t) \, dt = 1 - e^{-\beta} \) and \( q = e^{-\beta} \). Given the value of \( l_t \), the random variable \( m_{t,t+1} \) is binomially distributed with parameters \( l_t \) and \( 1 - q \). According to the memoryless property of the exponential delays \( \xi_i \), each customer which is late at time \( t \) has probability \( q \) to continue being late in each of the following time slots. For the sake of simplicity, let us use the notation \( m_{t,t+1} = m_t \). If the customer expected to arrive in the slot \( (t, t+1] \) has been deleted by the thinning procedure then

\[
P(m_t = j | l_t = 1) = \binom{l_t}{j} p^j q^{l_t-j} = b_{j,1},
\]

otherwise

\[
P(m_t = j | l_t = 1) = \binom{l_t+1}{j} p^j q^{l_t+1-j} = b_{j,l+1}.
\]

All in all,

\[
P(m_t = j | l_t = 1) = b_{j,1} (1 - \rho) + b_{j,l+1} \rho.
\]

If we assume that the state of the system is determined by \( l_t \) \( n_t \) then the evolution of the system is Markovian due to (5.5). Thus, the bivariate process \( (n_t, l_t), t \in \mathbb{N} \) is a discrete-time MC. Since we look at embedded points on the time axis, i.e., at departure instants, the process \( (n_t, l_t) \) is usually called embedded chain.
The embedded MC has the following transition probabilities:

For \( i > 0 \),
\[
P((i, j), (i + a - 1, j - a + 1)) = \rho b_{a,j+1}, \quad 0 \leq a \leq j+1,
\]
\[
P((i, j), (i + a - 1, j - a)) = \left(1 - \rho\right) b_{a,j}, \quad 0 \leq a \leq j;
\]

For \( i = 0 \),
\[
P((i, j), (a, j - a + 1)) = \rho b_{a,j+1}, \quad 0 \leq a \leq j+1,
\]
\[
P((i, j), (a, j - a)) = \left(1 - \rho\right) b_{a,j}, \quad 0 \leq a \leq j.
\]

Figure 16 displays those transitions having non-zero probability according to (5.6)–(5.7).

Remark 5.1. Provided \( \rho < 1 \), the bivariate MC \((n_t, l_t)\) is irreducible and positive recurrent. The latter property can be easily checked through the so-called Foster-Lyapunov criteria. Let us consider the Lyapunov function \( V((n_t, l_t)) = l_t + 1 \). It is immediate to prove that
\[
\mathbb{E}[V((n_1, l_1)) - V((n_0, l_0)) \mid l_0 = l] = \rho q - l(1 - q).
\]
Then, \( \mathbb{E}[V((n_1, l_1)) - V((n_0, l_0)) \mid l_0 = l] \) is easily bounded from above
- by a positive constant uniformly in \( l \);
- by a negative constant asymptotically for \( l \to \infty \).

This is sufficient to infer positive recurrence in the whole quarter plane, see [Twe76]. The aperiodicity is trivial. Thus, \((n_t, l_t)\) is an ergodic MC, and there exists a unique stationary measure \( Q_{n,l} \). We note that the irreducibility no longer holds for \( \rho = 1 \) because in this case the system is unstable and will eventually leave any finite region of the quarter plane.

* An infinite MC \( X^t \) is said to be positive recurrent if for all \( l \), the expectation of the first return time \( \tau_i^l = \min\{t > 0 : X^t = l\} \) is finite.
Let us now consider the following bivariate generating function:

\[ Q(z, y) = \sum_{n,l \geq 0} z^n y^l Q_{n,l}, \quad |z|, |y| \leq 1. \]  
(5.8)

**Theorem 5.1** The bivariate generating function (5.8) satisfies

\[ Q(z, y) = \frac{1 + \rho (v - 1) - v}{z} \left[ (z - 1) Q(0, v) + Q(z, v) \right], \]  
(5.9)

where

\[ v = v(z, y) = z + q (y - z). \]

**Proof.** The balance equations for EDA/D/1 are easily seen to be the following:

\[
Q_{n,l} = (1 - \rho) \left( \sum_{j=0}^{n} Q_{j+1,l+1-n-j} b_{n-j,l+1-n-j} + Q_{0,1+n} b_{n,1+n} \right) \\
+ \rho \left( \sum_{j=0}^{n} Q_{j+1,l+1-n-j-1} b_{n-j,l+1-n-j} + Q_{0,1+n-1} b_{n,1+n} \right),
\]
(5.10)

where \( b_{j,l} \) are given by (5.4) and we agree that \( Q_{n,1} = 0 \) whenever \( n, l < 0 \). The special cases \( n = 0 \) and \( n = l = 0 \) respectively lead to

\[
Q_{0,1} = (1 - \rho) \left( Q_{1,1} + Q_{0,1} \right) b_{0,1} + \rho \left( Q_{1,1} + Q_{0,1} \right) b_{0,1},
\]
(5.11)

\[
Q_{0,0} = (1 - \rho) \left( Q_{1,0} + Q_{0,0} \right).
\]
(5.12)

To show that (5.10)-(5.12) hold, it suffices to write \( Q_{n+1,l+1} \) in terms of \( Q_{n,l} \) and then neglect the time dependency. Take for example (5.12). The system is found at time \( t + 1 \) in state \( (0,0) \), i.e., with empty queue and no late customers, only if at time \( t \) it was either in state \( (0,0) \) or in state \( (1,0) \), and the \( (t+1) \)-th customer is deleted by the thinning. Indeed, if at time \( t \) the system was in state \( (0,0) \) then nothing happens and the state remains unchanged, whereas if it was in state \( (1,0) \) then the customer in queue is served and at time \( t+1 \) the system is in state \( (0,0) \). Similarly, there are four cases such that the system is found at time \( t+1 \) in state \( (0,1) \), i.e., with an empty queue and \( l \) customers late. In the first two cases the system is in state \( (1,1) \) or in state \( (0,1) \) at time \( t \), the \( (t+1) \)-th customer is deleted, and no one of the \( l \) late customers arrives in the interval \( [t, t+1) \). In the remaining cases the system is in state \( (1,1) \) or in state \( (0,1) \) at time \( t \), the \( (t+1) \)-th customer is not deleted (in which case he is added to the set of the \( l+1 \) already late customers), and no one of the \( (l+1) \) late customers arrives in the interval \( [t, t+1) \). The latter argument gives (5.11) while an easy generalisation to the case \( n \geq 1 \) leads to (5.10).

Let us take (5.10), multiply both sides by \( z^n y^l \), and then sum over \( n \) and \( l \). The term in \( (1 - \rho) \), appearing at the RHS after the summation, is

\[
(1 - \rho) \left\{ \sum_{n,l \geq 0} \left[ \sum_{j=0}^{n} Q_{j+1,l+1-n-j} \binom{n}{n-j} (z p)^{n-j} (y q)^l \right] \right. \\
+ Q_{0,1+n} \left( \frac{1+n}{n} \right) (z p)^n (y q)^l \}.
\]
or equivalently,
\[
(1 - \rho) \left\{ \sum_{j \geq 0} \sum_{n \geq j} \left[ \sum_{l \geq 0} Q_{l+1, l+n-j} \binom{n+l-j}{n-j} z^l (zp)^{n-j} (yq)^l + Q_{0, l+n} \binom{l+n}{n} (zp)^n (yq)^l \right] \right\}.
\]

The change of variable \( k = n - j \) and \( m = l + n - j = l + k \) yield
\[
(1 - \rho) \left\{ \sum_{j \geq 0} z^j \sum_{m \geq 0} \left[ \sum_{k=0}^n Q_{j+1, m} \binom{m}{k} (zp)^k (yq)^m \right] \right\}
= \frac{1 - \rho}{z} \sum_{j \geq 0} \sum_{m \geq 0} Q_{j, m} z^j (zp + yq)^m,
= \frac{1 - \rho}{z} \left[ Q(z, zp + yq) - Q(0, zp + yq) \right],
\]
to which we still have to sum the contribution
\[
(1 - \rho) \sum_{n, l \geq 0} Q_{0, l+n} \binom{l+n}{n} (zp)^n (yq)^l = (1 - \rho)Q(0, zp + yq).
\]

All in all, the term in \((1 - \rho)\) is
\[
(1 - \rho) \left[ Q(0, zp + yq) + \frac{1}{z} \left( Q(z, zp + yq) - Q(0, zp + yq) \right) \right].
\]

In a completely analogous way we can compute the term in \( \rho \), which turns out to be
\[
\rho(zp + yq) \left[ Q(0, zp + yq) + \frac{1}{z} \left( Q(z, zp + yq) - Q(0, zp + yq) \right) \right].
\]

Summing up the two contributions we get (5.9).

### 5.3 Intermediate Results on \( Q(z, y) \)

In the following sections we study the functional equation (5.9) given by Theorem 5.1 for the bivariate generating function (5.8). Some results may be directly obtained by iterating (5.9). Let us start by defining
\[
u^{(0)}(y) = y, \quad \nu^{(k)}(y) = q^k (y - z) + z, \quad k \geq 1,
\]
and let us compute (5.9) in \( y = \nu^{(k)} \). The obvious relation
\[
\nu^{(k+1)} = z + q(\nu^{(k)} - z)
\]
then yields
\[
Q(z, \nu^{(k)}) - \frac{1 - \rho + \rho \nu^{(k+1)}}{z} Q(z, \nu^{(k+1)}) = \frac{(1 - \rho + \rho \nu^{(k+1)})(z - 1)}{z} Q(0, \nu^{(k+1)}), \quad (5.13)
\]
Keeping now \( z \) fixed as a parameter,
\[
\lim_{k \to \infty} u^{(k)}(y) = z \quad \text{uniformly in } y.
\]
For \( k \to \infty \), equation (5.13) gives
\[
Q(z, z) = Q(0, z) \frac{1 - \rho + \rho z}{1 - \rho}.
\] (5.14)

An immediate consequence of equation (5.14) is
\[
Q(0, z) + \frac{Q(z, z) - Q(0, z)}{z} = \frac{Q(0, z)}{1 - \rho},
\] (5.15)
which we note here for future use.

By iteratively substituting (5.13) into (5.9),
\[
Q(z, y) = (z - 1) \left[ \sum_{j \geq 1} Q(0, u^{(j)}) \prod_{k=1}^{j} \frac{1 - \rho + \rho u^{(k)}}{z} \right] + \prod_{k \geq 1} \frac{1 - \rho + \rho u^{(k)}}{z} Q(z, z).
\] (5.16)

The behaviour of the infinite product at the RHS of (5.16) depends on the position of \( z \) with respect to the curve
\[
\frac{1 - \rho(z - 1)}{z} = 1.
\]
For \( a + ib = z \in \mathbb{C} \), that curve is the circle
\[
\Gamma: \quad \left( a - \frac{\rho}{1 + \rho} \right)^2 + b^2 = \frac{1}{(1 + \rho)^2}.
\]
\( \Gamma \) lies inside the unit disc \(|z| \leq 1\) for each value of \( \rho \) within the interval \([0, 1]\).
For \( z \) outside \( \Gamma \),
\[
\prod_{k \geq 1} \frac{1 - \rho + \rho u^{(k)}}{z} = 0,
\]
whereas for \( z \) inside \( \Gamma \) the product diverges. Since \( Q(z, y) \) is analytic in its arguments in the unit bi-disc, such a divergence is cancelled by the divergence of the series at the RHS of (5.16). Looking for cancellations and imposing the regularity of \( Q(z, y) \) could well lead to a reduction of (5.9) into an explicit BVP. However, this is not the approach we choose here.

The point \( \Gamma \ni z = 1 \) is very interesting for evaluating (5.16) because we find
\[
Q(1, y) = \prod_{k=0}^{\infty} [1 + \rho q^{k+1}(y - 1)].
\] (5.17)

Formula (5.17) is a simple and compact expression for \( Q(1, y) \), the generating function of the marginal distribution
\[
Q_{.,1} = \sum_{n \geq 0} Q_{n,.}.
\]
We come back to (5.17) in Chapter 6 where we reformulate \( Q(1, y) \) as a series and discuss the asymptotic behaviour of \( Q_{.,1} \).
Remark 5.2. The infinite product (5.17) is known in combinatorics with the name of \( q \)-Pochhammer symbol of the pair \( (\rho(y - 1), q) \). It is the \( q \)-analog of the descending factorial, also known as Pochhammer symbol. When \( \rho(y - 1) \) is substituted by 1 it gives back the well-known Euler function. In the disc \( |q| < 1 \) the \( q \)-Pochhammer symbol is analytic for all the values of \( y \).

We end this section with a discussion of the special case \( q = 0 \). In this regime, the RHS of equation (5.9) does not depend on \( y \) anymore and it is \( Q(z, y) = Q(z) \). In other words, \( l_i \) is constantly null as the \( i \)-th customer cannot have a delay \( \xi_i \geq 1 \). The system reduces to a simple D/D/1 queue with balking, and the stationary probability to have a void queue is \( 1 - \rho \). Then, equation (5.9) yields directly

\[
Q(z) = \frac{1 + \rho(z - 1)}{z}([z - 1]Q(0) + Q(z)),
\]

(5.18)

where \( Q(0) = 1 - \rho \) is the stationary probability of a void queue. Therefore, equation (5.18) is equivalent to

\[
Q(z) = 1 + \rho(z - 1),
\]

the classical result of a D/D/1 queue with balking.

5.4 Power series expansion of \( Q(z, y) \)

Let us look for the solution of the functional equation (5.9) in the class of functions which are analytic in the region

\[
\{(z, y, q) \in \mathbb{C}^2 \times [0, \phi) \text{ s.t. } |z| \leq 1, |y| \leq 1\}.
\]

Any solution found within this class is automatically the unique solution of (5.9) by the uniqueness of the stationary distribution \( Q_{n, l} \).

Let us consider the power expansion

\[
Q(z, y) = \sum_{k \geq 0} q^k Q^{(k)}(z, y),
\]

(5.19)

convergent for \( 0 \leq q < \phi, \phi > 0 \) (an estimate of \( \phi \) is provided in Section 5.6).

The \( k \)-th coefficient of the power expansion is represented by

\[
Q^{(k)}(z, y) = [q^k]Q(z, y) = \left. \frac{1}{k!} \frac{\partial^k}{\partial q^k} Q(z, y) \right|_{q = 0}.
\]

Equation (5.19) can be combined with a Taylor expansion centered in the point \( y = z \) to obtain

\[
Q(z, y) = \sum_{j \geq 0} \frac{(y - z)^j}{j!} \left. \frac{\partial^j}{\partial y^j} Q(z, y) \right|_{y = z}.
\]

(5.20)

To find the coefficients \( Q^{(k)}(z, y) \), we take the \( j \)-th order derivative of the RHS in (5.9). This produces a factor \( q^j \), and a subsequent comparison of (5.19) with (5.20) yields to a set of equations for the coefficients \( Q^{(k)}(z, y) \).

Let us introduce the symbol \( \partial_y^j Q(z, y) \) to indicate the \( j \)-th order partial derivative of \( Q(z, y) \) taken with respect to \( y \) and evaluated for \( z = z_0 \) and \( y = y_0 \). Equation (5.9) gives

\[
\partial_y^j Q(z, z) = \left. \frac{\partial^j}{\partial y^j} Q(z, y) \right|_{y = z},
\]

(5.21)

* Cf. Remark 5.1 on page 69
74 exponentially delayed arrivals

whereas the second gives $j$ where we agree that $\rho = 0$. This infers that $[q^0]Q(z,y) = Q^{(0)}(z,y)$, the coefficient of the zero-th order, is independent of $y$ and

$$Q^{(0)}(z,y) = 1 + \rho(z - 1).$$

A key role is played throughout the chapter by the functions

$$a_k(z) = \frac{\partial_y^j Q^{(k)}(0,z)}{z} + \frac{\partial_y^j Q^{(k)}(z,z) - \partial_y^j Q^{(k)}(0,z)}{z}, \quad (5.22)$$

where we agree that $j = 0$ means no differentiation and that $a_k(z) = 0$ whenever $j < 0$. The functions $a_k(z)$ are important because they form a contracting array, recursively computable; we discuss this point in Section 5.6. The substitution of (5.19) into (5.15) and a term-by-term comparison give

$$a_k(z) = \frac{Q^{(k)}(0,z)}{1 - \rho} \quad \forall k \geq 0. \quad (5.23)$$

For each $q \geq 0$, the following boundary conditions must be imposed to the bivariate generating function $Q(z,y)$:

$$
\begin{cases}
Q(0,1) = \sum_n Q_{0,n} = 1 - \rho, & \text{Little’s Law}, \\
Q(1,1) = \sum_n Q_{n,1} = 1, & \text{Normalisation}.
\end{cases} \quad (5.24)
$$

From (5.19) and Remark 5.3,

$$Q(z,y) = 1 - \rho(z - 1) + \sum_{k \geq 1} q^k Q^{(k)}(z,y) \quad \forall 0 \leq q < 1.$$

Then, the first of (5.24) infers

$$Q^{(k)}(0,1) = 0 \quad \forall k \geq 1, \quad (5.25)$$

whereas the second give

$$Q^{(k)}(1,1) = 0 \quad \forall k \geq 1. \quad (5.26)$$

The next theorem gives an explicit formula for the coefficients $Q^{(k)}(z,y)$ of the power expansion (5.19):

**Theorem 5.2** The coefficients $Q^{(k)}(z,y)$ satisfy

$$Q^{(0)}(z,y) = 1 + \rho(z - 1), \quad (5.27)$$

$$Q^{(k)}(z,y) = \sum_{j=1}^{k} \frac{(y - z)^{j}}{j!} \left\{ \rho a_{k-j}^{j-1}(z) + (1 + \rho(z - 1)a_{k-j}^{j-1}(z) \right\}$$

$$+ (1 + \rho(z - 1)) \frac{Q^{(k)}(0,z)}{1 - \rho} \quad \forall k \geq 1. \quad (5.28)$$
Proof. Equation (5.27) has been already proved by Remark 5.3, so the proof only covers (5.28). Equating (5.19) and (5.20), and using (5.21),

\[
\sum_{k \geq 0} q^k Q^{(k)}(z, y) = \frac{q^l (y - z)^j}{j!} \sum_{l \geq 0} \left[ 1 + \rho(z - 1) \right] \times \\
\left( c^l_{y} Q(0, z) + c^l_{y} Q(z, z) - c^l_{y} Q(0, z) \right) \\
+ \rho \left( c^l_{y} Q(0, z) + c^l_{y} Q(z, z) - c^l_{y} Q(0, z) \right) \right]. \tag{5.29}
\]

By the analyticity we can now substitute (5.19) into the RHS of (5.29) and rearrange the expression to find

\[
\sum_{k \geq 0} q^k Q^{(k)}(z, y) = \sum_{j \geq 0} \frac{(y - z)^j}{j!} \sum_{l \geq 0} q^{j+l} \left[ \rho a^j_{l-1} \right] + [1 + \rho(z - 1)] a^j_0. \tag{5.30}
\]

For \( k \geq 1 \), let us operate in (5.30) the change of indices \( i + j = k \)

\[
\sum_{k \geq 0} q^k Q^{(k)}(z, y) = \sum_{k \geq 0} q^k \sum_{j = 0}^k \frac{(y - z)^j}{j!} \left\{ \rho a^{k-j}_{j-1} + [1 + \rho(z - 1)] a^{k-j}_j \right\}. \tag{5.31}
\]

By (5.23), the RHS of (5.31) becomes

\[
\sum_{k \geq 0} q^k \left\{ \sum_{j = 1}^k \frac{(y - z)^j}{j!} \left\{ \rho a^{k-j}_{j-1} + [1 + \rho(z - 1)] a^{k-j}_j \right\} + (1 + \rho(z - 1)) \frac{Q^{(k)}(0, z)}{1 - \rho} \right\} \tag{5.32}
\]

because \( a^k_j(z) = 0 \) whenever \( j < 0 \). Equations (5.31) and (5.32) immediately give (5.28).

Remark 5.4. Equation (5.28) represents a recursive relation that gives the coefficients \( Q^{(k)}(z, y) \) in terms of \( Q^{(k)}(0, z) \) and the derivatives of \( Q^{(k)}(z, y) \), \( l < k \). Lemma 5.3 below allows to write also \( Q^{(k)}(0, z) \) in terms of the derivatives of lower-order coefficients. This gives a recursion to compute the expansion (5.19) to any desired order.

Remark 5.5. From (5.22) and (5.23),

\[
a^0_j(z) = \delta_{j, 0}. \tag{5.33}
\]

We next introduce a symbol that keeps the notation sufficiently compact. Let \( k \geq 1 \), and for \( 1 \leq j \leq k \), define

\[
A^k_j(z) = \rho a^{k-j}_{j-1}(z) + [1 + \rho(z - 1)] a^{k-j}_j(z). \tag{5.34}
\]

According to (5.34), formula (5.28) can be rewritten as

\[
Q^{(k)}(z, y) = \sum_{j = 1}^k \frac{(y - z)^j}{j!} A^k_j(z) + (1 + \rho(z - 1)) \frac{Q^{(k)}(0, z)}{1 - \rho}. \tag{5.35}
\]
The following result gives \( Q^{(k)}(z, y) \) in terms of \( A^k_j(z) \):

**Lemma 5.3** For any \( k \geq 1 \),

\[
Q^{(k)}(0, z) = \sum_{j=1}^{k} \frac{(z - 1)}{j!} A^k_j(0). \tag{5.36}
\]

**Proof.** Take \( k \geq 1 \), from (5.25) and (5.35)

\[
Q^{(k)}(0, 1) = \sum_{j=1}^{k} \frac{A^k_j(0)}{j!} + Q^{(k)}(0, 0) = 0,
\]

that yields

\[
Q^{(k)}(0, 0) = -\sum_{j=1}^{k} \frac{A^k_j(0)}{j!}.
\]

The value of \( Q^{(k)}(0, 0) \) can now be used to compute \( Q^{(k)}(0, z) \) via (5.35).

The result is

\[
Q^{(k)}(0, z) = \sum_{j=1}^{k} \frac{z^j}{j!} A^k_j(0) + Q^{(k)}(0, 0),
\]

\[
= \sum_{j=1}^{k} \frac{z^j}{j!} A^k_j(0) - \sum_{j=1}^{k} \frac{A^k_j(0)}{j!}.
\]

\[ \Box \]

### 5.5 Recursive Computation of \( Q^{(k)}(z, y) \)

In Section 5.4 we have derived a formula for \( Q^{(k)}(z, y) \) in terms of the derivatives of the coefficients \( Q^{(1)}(z, y) \), \( 1 \leq k \). In this section we show how to actually exploit this recursive relation. The actual computations required to obtain \( Q^{(k)}(z, y) \) are presented in full detail for \( k = 1, 2, 3 \).

Combining equations (5.35) and (5.36) we get the following general form of \( Q^{(k)}(z, y) \):

\[
Q^{(k)}(z, y) = \sum_{j=1}^{k} \left[ (y - z)^j \frac{1}{j!} A^k_j(z) + \frac{1 + \rho(z - 1)}{1 - \rho} \frac{z^j - 1}{j!} A^0_j(0) \right]. \tag{5.37}
\]

This formula will be used in the computations to come. The recursive computation starts from (5.33); it gives \( Q^{(1)}(z, y) \), from which is possible to compute \( a^1_1(z) \) and \( A^1_1(z) \); from these quantities we can compute \( Q^{(2)}(z, y) \), and so on so forth. The iteration just described can be used to obtain the coefficients of the power series (5.10) up to any desired order \( k \). Before moving to the actual computations we note the following:

**Remark 5.6.** By (5.33),

\[
A^k_j(z) = \rho a^0_{j-1}(z) + [1 + \rho(z - 1)] a^0_j(z) = 0 \quad \forall k \geq 2.
\]

Therefore, for any \( k \geq 2 \), in formulas (5.28), (5.35), and (5.36) the index \( j \) actually runs from 1 to \( k - 1 \).
For $k = 1$, equation (5.37) gives

$$Q^{(1)}(z, y) = (y - z) A_1^1(z) + \frac{1 + \rho(z - 1)}{1 - \rho}(z - 1) A_1^1(0), \quad (5.38)$$

where $A_1^1(z) = \rho$ by (5.33). Thus,

$$Q^{(1)}(z, y) = (y - z)\rho + \frac{1 + \rho(z - 1)}{1 - \rho}\rho(z - 1).$$

We note that $Q^{(1)}(1, 1) = 0$ in accordance with (5.26). Differentiating with respect to $y$,

$$a_0^1(z) = \frac{\rho}{1 - \rho}(z - 1),$$

$$a_1^1(z) = \rho,$$

$$a_2^1(z) = 0 \quad \forall j \geqslant 2.$$  

Using (5.34),

$$A_1^2(z) = \rho \left( 1 + \rho(z - 1) \frac{2 - \rho}{1 - \rho} \right).$$

The last formula is needed in the computations of the second-order coefficient $Q^{(2)}(z, y)$. From (5.35), (5.36), and Remark 5.6,

$$Q^{(2)}(z, y) = (y - z) A_1^2(z) + \frac{1 + \rho(z - 1)}{1 - \rho}(z - 1) A_1^2(0),$$

that gives

$$Q^{(2)}(z, y) = (y - z)\rho \left[ 1 + \rho(z - 1) \frac{2 - \rho}{1 - \rho} \right] + \frac{1 + \rho(z - 1)}{1 - \rho}\rho(z - 1) \left[ 1 - \rho \frac{2 - \rho}{1 - \rho} \right].$$

We note again that $Q^{(2)}(1, 1) = 0$ in accordance with (5.26). Differentiating,

$$a_0^2(z) = \frac{\rho}{1 - \rho} \left[ 1 - \rho \left( 1 + \frac{1}{1 - \rho} \right) \right] (z - 1),$$

$$a_1^2(z) = \rho,$$

$$a_2^2(z) = 0 \quad \forall j \geqslant 2.$$  

By (5.34), the coefficients needed for the third-order computation are found:

$$A_1^3(z) = \rho \left[ 1 + \rho(z - 1) \left( 2 - \frac{\rho}{(1 - \rho)^2} \right) \right],$$

$$A_2^3(z) = 2 \rho^2.$$  

Remark 5.7. Since $Q^{(1)}(z, y)$ and $Q^{(2)}(z, y)$ are linear in $y$,

$$A_3^4(z) = A_4^5(z) = \cdots = A_j^j(z) = 0 \quad \forall j \geqslant 3,$$

$$A_3^5(z) = A_4^6(z) = \cdots = A_j^{j+2}(z) = 0 \quad \forall j \geqslant 3.$$
In particular, \( k = 6 \) is the lowest \( k \) such that \( Q^{(k)}(z, y) \) is more than quadratic in \( y \). This could be directly deduced from the memoryless property of the delays. Indeed, the most likely situation leading to \( t_1 = 3 \) (see (5.3) above) is that none of those customers supposed to arrive at times \( t - 1, t - 2 \) and \( t - 3 \) has yet arrived at time \( t \). This event has probability of order \( q^6 \). Similarly, the lowest \( k \) such that \( Q^{(k)}(z, y) \) is of order \( m \) in \( y \) is the \( m \)-th triangular number, \( k = \frac{m(m+1)}{2} \).

**Third order in \( q \)**

From (5.35), (5.36) and Remark 5.6,

\[
Q^{(3)}(z, y) = \frac{(y-z)^2}{2} A_2^3(z) + (y-z) A_1^3(z) + \frac{1 + \rho(z-1)}{1 - \rho} \left[ (z-1) A_1^3(0) + \frac{z^2 - 1}{2} A_2^3(0) \right].
\]

Thus,

\[
Q^{(3)}(z, y) = (y-z)^2 \rho^2 + (y-z) \rho \left[ 1 + \rho(z-1) \left( 2 - \frac{\rho}{(1-\rho)^2} \right) \right] + \frac{1 + \rho(z-1)}{1 - \rho} \rho(z-1) \left[ 1 - \rho \left( 2 - \frac{\rho}{(1-\rho)^2} \right) + \rho(z+1) \right].
\]

We note again that \( Q^{(3)}(1,1) = 0 \), in accordance with (5.26). Differentiating,

\[
a_3^3(z) = \rho(z-1) \left\{ \rho(z+1) + \left[ 1 - \rho \left( 2 - \frac{\rho}{(1-\rho)^2} \right) \right] \right\},
\]

\[
a_2^3(z) = 2 \rho^2 (z-1) + \rho,
\]

\[
a_1^3(z) = 2 \rho^2,
\]

\[
a_0^3(z) = 0 \quad \forall j \geq 3.
\]

The recursive method we have just outlined can be applied to compute a truncated expansion of the generating function \( Q(z, y) \). Section 5.7 illustrates how to implement the computation of \( Q^{(k)}(z, y) \) in a fully automated way. Section 5.8 successively demonstrates that such an automated computation is equivalent to the solution of a linear system. Before moving to these topics, we study in more detail the functions \( a_j^k(z) \).

### 5.6 Analytic Results on \( a_j^k(z) \)

In Remark 5.7, we have claimed that, fixed an integer \( m \), all the coefficients \( Q^{(k)}(z, y) \) with \( k < \frac{m(m+1)}{2} \) are polynomials of degree strictly less than \( m \) in \( y \). The next argument proves this claim. Let us define the following symbol:

\[
j_k = \max \left\{ j \geq 0 : \left( \frac{j + 1}{2} \right) \leq k \right\}.
\]

It is straightforward to show that

\[
j_k = \left\lfloor \frac{-1 + \sqrt{1 + 8k}}{2} \right\rfloor,
\]

where \( \lfloor \cdot \rfloor \) stands for the lower integer part.

**Lemma 5.4** With respect to the notation introduced above,

\[
a_j^k(z) = 0 \quad \forall k \geq 2, \ j > j_k.
\]
Proof. We prove (5.40) by induction. For \( k = 2 \), it is \( j_2 = 1 \) and \( a_j^2(z) = 0 \).

Let us now make the following induction hypothesis:

\[
a_j^m(z) = 0 \quad \forall 2 \leq m \leq k, \quad j > j_m.
\]  

(5.41)

Arbitrarily fixed \( l > j_{k+1} > 1 \),

\[
\partial_y^l Q^{(k+1)}(z, y) = \sum_{j=1}^{k+1} \frac{(y-z)^{j-1}}{(j-1)!} \left[ j \rho a_j^{k+1-j}(z) + (1 + \rho(z-1)) a_j^{k+1-j}(z) \right].
\]

It is now sufficient to check that \( j - 1 > j_{k+1-j} \) for \( j = 1 \). If this is proved then \( j - 1 > j_{k+1-j} \) for \( j \geq 1 \), and \( \partial_y^l Q^{(k+1)}(z, y) \) is identically zero by the induction hypothesis, and the thesis follows from (5.22). There are two possibilities for \( k + 1 \):

a. \( j_{k+1} = j_k + 1 \);

b. \( j_{k+1} = j_k \).

Case a. yields

\[
k + 1 - 1 < k + 1 - j_{k+1} = k - j_k, \\
1 - 1 < j_{k+1} - 1 = j_k.
\]

Since \( j_k - j_k \leq j_k \), equation (5.41) infers \( a_{k-1}^{k+1-j}(z) = 0 \).

Case b. gives

\[
k + 1 - 1 < k + 1 - j_{k+1} \leq k + 1 - j_k - 1, \\
1 - 1 > j_{k+1} - 1 \geq j_k.
\]

(5.42) (5.43)

Recall that \( j_{k+1} = j_k \), then

\[
k + 1 = \left( \frac{j_k + 1}{2} \right) + r + 1, \quad r + 1 < j_k + 1.
\]

From (5.42),

\[
k + 1 - 1 \leq \left( \frac{j_k + 1}{2} \right) + r - j_k = \left( \frac{j_k - 1}{2} + 1 \right) + r.
\]

and \( r < j_k \). Thus, \( j_{k+1} - 1 \leq j_k - 1 \), and \( a_{k-1}^{k+1-j}(z) = 0 \) follows by (5.41) and (5.43).

Remark 5.8. An immediate consequence of Lemma 5.4 is that the index \( j \) actually runs from 1 to \( j_k \) in the expressions (5.36) and (5.37).

The proof of Lemma 5.4 has clearly shown that the values of \( k \) of the form \( \binom{m+1}{2} \) are in some sense special. The next result gives the exact characterization of \( a_{j_k}^k \) when \( k \) has that special form, which is to say, \( k \) is a triangular number.

**Lemma 5.5** For \( k = \binom{m+1}{2} \) and \( m \geq 1 \),

\[
a_{j_k}^k(z) = j_k! \rho^{j_k} = m! \rho^m.
\]  

(5.44)
The proof of Lemma 5.5 makes use of the recursive structure of the functions \(a^k_l(z)\). Combining (5.22) with formulas (5.34) and (5.37), another recursive relation can be obtained. For \(l \geq 1\),

\[
a^k_l(z) = \sum_{j=1}^{k} \frac{z^{j-1}(z-1)}{(j-1)!}\left[jp a^k_{j-1}(0) + (1 - \rho) a^k_{j-1}(0)\right]
+ l \rho \left[ a^k_{l-1}(0) + \frac{a^k_{l-1}(z) - a^k_{l-1}(0)}{z} \right] + \rho a^k_{l-1}(z)
+ (1 - \rho) \left[ a^k_{l-1}(0) + \frac{a^k_{l-1}(z) - a^k_{l-1}(0)}{z} \right],
\]

(5.45)

while for \(l = 0\), equations (5.23), (5.34), and (5.36) give

\[
a^0_l(z) = \sum_{j=1}^{k} \frac{z^{j-1}}{j!} \left[jp a^k_{j-1}(0) + (1 - \rho) a^k_{j-1}(0)\right].
\]

(5.46)

Equations (5.45) and (5.46) demonstrate that the functions \(a^k_l(z)\) form a contracting array which is recursively computable.

**Proof of Lemma 5.5.** We prove (5.44) by induction. For \(m = 1\), it is \(k = 1, j_k = 1,\) and \(a^1_1(z) = \rho\). Let now \(m \geq 3\) and suppose that (5.44) holds for \(k = \binom{m}{2}\), i.e.,

\[
a^k_{j_k}(z) = (m-1)! \rho^{m-1}.
\]

Taking the next value of \(m\),

\[
k' = \left(\frac{m+1}{2}\right),
\]

\[
j_{k'} = m,
\]

\[
k' - j_{k'} = k,
\]

\[
j_{k'} - 1 = j_k.
\]

Equation (5.45) and Lemma 5.4 now give

\[
a^k_{j_{k'}}(z) = j_{k'} \rho a^{k'-j_{k'}}_{j_{k'}-1}(0) + (1 - \rho) a^{k'-j_{k'}}_{j_{k'}-1}(0)
+ j \rho \frac{a^{k'-j_{k'}}_{j_{k'}-1}(z) - a^{k'-j_{k'}}_{j_{k'}-1}(0)}{z}
+ (1 - \rho) \frac{a^{k'-j_{k'}}_{j_{k'}}(z) - a^{k'-j_{k'}}_{j_{k'}}(0)}{z}
+ \rho a^{k'-j_{k'}}_{j_{k'}}(z),
\]

\[
= j_{k'} \rho a^{k'-j_{k'}}_{j_{k'}-1}(0).
\]

This concludes the proof. \(\square\)

We end the section with an estimate of the radius of analyticity of (5.19).

They key to such an estimate is the following:

**Lemma 5.6** For all \(k, l, m \in \mathbb{N}\) and \(|z| \leq 1\),

\[
\left| \frac{d^m}{dz^m} a^k_l(z) \right| \leq C^k,
\]

(5.47)

where \(C = \max \left\{12, \frac{2}{1 - \rho}\right\} \).
Proof. The proof is carried out by induction. For $k = 0, 1$, inequality (5.47) is clearly satisfied. In addition, from Remark 5.6 it suffices to prove (5.47) for $1 \leq k$. Thus let us assume that (5.47) is fulfilled in the polyhedral lattice

$$L_k = \left\{(j, l, m) \in \mathbb{N}^3, j \leq k, l \leq k\right\}.$$ 

We show that (5.47) holds as well in

$$L_{k+1} = L_k \cup \left\{(j, l, m) \in \mathbb{N}^3, j = k + 1, l \leq k + 1\right\}.$$ 

From (5.22), (5.34) and (5.37) it easily follows that $a_k^l(z)$ is a polynomial in $z$ with degree not larger than $k$. Using the induction hypothesis and a Taylor expansion, $\forall m < k$,

$$\frac{d^m}{dz^m} a_k^l(z) - a_k^l(0) = \frac{d^m}{dz^m} \left[ \sum_{j=1}^{\infty} \frac{z^{j-1}}{j!} \frac{d^j}{dz^j} a_k^l(0) \right] = \left[ \sum_{j=m-1}^{k} \frac{z^{j-m+1}}{j!(j-m+1)!} \frac{d^j}{dz^j} a_k^l(0) \right],$$

$$\leq C^k \epsilon,$$

where we agree that $m = 0$ means no differentiation. Then, from (5.46),

$$\left| a_k^l(z) \right| \leq \sum_{j=1}^{k} \frac{2^j}{j!(1-\rho)} \left( j\rho C^{k-j} + (1-\rho)C^{k-j} \right) \leq C^k \frac{2\epsilon\rho}{C(1-\rho)}.$$  

If $C \geq \frac{2}{1-\rho}$ then (5.48) infers $|a_k^l(z)| \leq C^k$.

From (5.45),

$$\left| a_k^l(z) \right| \leq \sum_{j=1}^{l+1} \frac{2}{(j-1)!} \left\{ [C^{k-j} + l\rho(1+\epsilon)C^{k-1} + \rho C^{k-1} + (1-\rho)(1+\epsilon)C^{k-1}] \right\} \leq 2C^{k-1} \sum_{j=1}^{l+1} \frac{1+j}{j!(C^l\epsilon)} \leq C^k \frac{5 + 4l + 2\epsilon\rho}{C^l} (1 + \frac{1}{C^l}).$$

Provided $C$ is sufficiently large, e.g. $C \geq 12$, inequality (5.48) infers $|a_k^l(z)| \leq C^k, 1 \geq 1$. The proof of (5.47) for $m \geq 1$ is completely analogous.

If we plug the estimate provided by Lemma 5.6 into equation (5.37) then we obtain that $A_k^l(z) = O(C^{k-1})$. The last estimate and equation (5.34), now give $Q^{(k)}(z, y) = O(C^{k-1})$. As a consequence, Lemma 5.6 implies that the analyticity radius in $q$ is at least $\phi = C^{-1}$. Indeed, formulas (5.50)-(5.78) below show that the functions $a_k^l(z)$ can be either positive or negative depending on the value of $\rho$. The very same computations also suggest that for
Listing 1: Sage code to compute $a_k^j(z)$.

```python
def ajk(j,k,P):
    """Return $a_j^k(z)$"""
    if j == 0 and k == 0:
        return 1
    elif j == 0 and k >= 1:
        return P[k].subs(z==0).subs(y==z)/(1-rho)
    if j > 0 and k < j:
        return 0
    else:
        return (((P[k].subs(z==0) + (P[k] - P[k].subs(z==0))/z).diff(y,j)).subs(y==z)).simplify_full()
```

a fixed $k$, the fastest-growing coefficient $a_k^0(z)$ is the one with $j = 0$, which seems to diverge as $(1 - \rho)^{-k}$ for $\rho \to 1$. On the other hand, Figure 17 shows that a truncated expansion well approximates the equilibrium distribution $Q_{n,0}$ even for values of $q$ greater than $C^{-1}$. Therefore, it seems likely that the coefficients $Q^{(k)}(z,y)$ present some cancellations and the analyticity of the bivariate generating function $Q(z,y)$ may be extended to $q \in [0,1)$. The problem of proving this claim still remains open, though.

5.7 AUTOMATED COMPUTATION OF THE SOLUTION

The recursion relations (5.27) and (5.28) can give a series expansion truncated to a given order $L$, i.e.,

$$Q(z,y) = \sum_{k=0}^{L} Q^{(k)}(z,y) q^k. \quad (5.49)$$

The computation of an expression like (5.49) can be automated in a symbolic calculus environment, e.g. MATLAB, Mathematica, or Sage. In the following, some snippets of Sage† code are listed. These can behave used to compute formulas (5.50)–(5.78) below.

Listing 1 displays a function that can compute $a_k^j(z)$. The input arguments are the indices $j$ and $k$, and a list $P$ containing the coefficients $Q^{(k)}(z,y)$. Listing 2 presents a function that computes $A_k^j(z)$ and $A_k^0(0)$, and return them wrapped in a tuple. The arguments are the same as the function $ajk(j,k,P)$. Listing 3 displays a function that computes the $k$-th coefficient $Q^{(k)}(z,y)$. The input arguments are the order $k$ and a list $P$ of already computed coefficients $Q^{(l)}(z,y), l < k$. Finally, Listing 4 shows an example of how to use the functions defined above. This code iteratively builds a list $Q$ of length $n+1$ that contains all the coefficients $Q^{(k)}$ up to the order $n$. It also com-

---

* In Figure 17 the histogram drawn for $\rho = 0.25$, $q = 0.50$ implies a very good accordance between simulations and a truncated expansion. Nevertheless, $q$ is larger than the estimated analyticity radius $\phi$.

† There are several reasons to prefer Sage over the many other viable alternatives. For example, Sage is free, open source software, wrapped in a Python-based interface, and it gives the possibility to build high-end scientific applications. More information on Sage at the project home page, \url{http://www.sagemath.org/}. 
computes and displays the functions $a^k_j(z)$ for $k = 1, \ldots, n$ and $j = 1, \ldots, k$. The coefficients computed this way are listed below.

\begin{align*}
  a^1_0(z) &= \frac{\rho}{1 - \rho} (z - 1), \\
  a^1_1(z) &= \rho; \\
  a^2_0(z) &= \frac{\rho^2 - 3\rho + 1}{(1 - \rho)^2} (z - 1), \\
  a^2_1(z) &= \rho; \\
  a^3_0(z) &= \frac{\rho^2}{1 - \rho} (z^2 - 1) - \rho \frac{2\rho^3 - 6\rho^2 + 4\rho - 1}{(1 - \rho)^3} (z - 1), \\
  a^3_1(z) &= 2\rho^2 (z - 1) + \rho, \\
  a^3_2(z) &= 2\rho^2; \\
  a^4_0(z) &= -\rho^2 \frac{2\rho^2 - \rho - 2}{1 - \rho} (z - 1) + \rho, \\
  a^4_1(z) &= 2\rho^2; \\
  a^4_2(z) &= 2\rho^2; \\
  a^5_0(z) &= -\rho^2 \frac{3\rho - 2}{1 - \rho} (z^2 - 1) - \rho \frac{2\rho^7 - 7\rho^6 + 4\rho^5 + 16\rho^4}{(1 - \rho)^5} (z - 1), \\
  a^5_1(z) &= -\rho^2 \frac{2\rho^3 - 7\rho^2 + 7\rho - 4}{1 - \rho} (z - 1) + \rho, \\
  a^5_2(z) &= 4\rho^2; \\
  a^5_3(z) &= 4\rho^2; \\
  a^5_4(z) &= 4\rho^2; \\
  a^5_5(z) &= 4\rho^2;
\end{align*}

Listing 2: Sage code to compute $A^k_j(z)$.

```python
def bigAjk(j, k, P):
    """
    Return a tuple $(A^j_k(z), A^j_k(0))$
    """
    if j == 0 and k == 0:
        return (1 + rho * (z-1), 1 - rho)
    elif j == 0:
        return (0,0)
    elif j <= k:
        m = j * rho * ajk(j-1,k-j,P) + (1 + rho * (z - 1)) * ajk(j,k-j,P)
        return (m, m.subs(z==0))
    else:
        return (0,0)
```
Listing 3: Sage code to compute $Q^k(z,y)$.

```python
def orderofQ(k, P):
    """
    Return $Q^{(k)}(z,y)$
    """
    sum = 0
    for j in range(1,k+1):
        b = bigAjk(j,k,P)
        sum += (y - z)^j/factorial(j) * bigAjk(j,k,P)[0] + ( 1 + rho * (z - 1))/(1 - rho) * (z^j - 1)/factorial(j) * bigAjk(j,k,P)[1]
    return sum
```

Listing 4: Sage code to compute $Q(z,y)$ up to any prescribed order $n$.

```python
# order of the expansion
n = 10
# variables
rho = var('rho')
z = var('z')
y = var('y')
q = var('q')
Q = [-1 for i in range(n+1)]
Q[0] = 1 + rho * (z - 1)
for k in range(1,n+1):
    Q[k] = orderofQ(k,Q)
# Compute and show $a_j^k(z)$, for $k=0,...,n$
for k in range(1,n+1):
    L = []
    for j in range(1,k+1):
        L.append(ajk(j,k,Q).simplify_full())
    show(L)
```
\[ a_0(z) = \frac{\rho^3}{1 - \rho} (z^3 - 1) + \rho^2 \frac{2\rho^3 - 5\rho + 2}{(1 - \rho)^2} (z^2 - 1) - \rho \frac{2\rho^3 - 15\rho^2 + 52\rho + 112\rho^6 + 166\rho^6}{(1 - \rho)^6} (z - 1), \]

\[ a_1(z) = 3\rho^3(z^2 - 1) - \rho^3 \frac{2\rho^4 - 11\rho^3 + 7\rho^2 + 7\rho - 4}{1 - \rho} (z - 1) + \rho, \]

\[ a_2(z) = 6\rho^3 (z - 1) + 4 \rho^2, \]

\[ a_3(z) = 6\rho^3; \]

\[ a_4(z) = \frac{\rho^3}{1 - \rho} (z^3 - 1) - \rho^2 \frac{2\rho^4 - 7\rho^3 + 9\rho^2 - 9\rho + 3}{(1 - \rho)^2} (z^2 - 1) + \rho \frac{2\rho^{11} - 21\rho^{10} + 89\rho^9 - 200\rho^8 + 247\rho^7 - 124\rho^6}{(1 - \rho)^7} (z - 1), \]

\[ a_5(z) = -\rho^3 \frac{3\rho^2 - \rho - 3}{1 - \rho} (z^2 - 1) + \rho^2 \frac{2\rho^6 - 17\rho^5 + 35\rho^4 - 26\rho^3 + 17\rho^2 - 18\rho + 6}{(1 - \rho)^2} (z - 1) + \rho, \]

\[ a_6(z) = 6\rho^3 (z - 1) + 6 \rho^2, \]

\[ a_7(z) = 6\rho^3; \]

\[ a_8(z) = 2\rho^3 \frac{1}{1 - \rho} (z^3 - 1) + \rho^2 \frac{+12\rho^2 - 12\rho + 3}{(1 - \rho)^2} (z^2 - 1) - \rho \frac{2\rho^3 - 27\rho^2 + 147\rho^1 + 47\rho^10 + 872\rho^9 - 1189\rho^6 + 1726\rho^7}{(1 - \rho)^8} (z - 1), \]

\[ a_9(z) = -\rho^3 \frac{3\rho^3 - \rho^2 + 2\rho - 6}{1 - \rho} (z^2 - 1) + \rho^2 \frac{2\rho^7 - 21\rho^6 + 60\rho^5 - 62\rho^4 + 8\rho^3 + 28\rho^2 - 24\rho + 6}{(1 - \rho)^2} (z - 1) + \rho, \]

\[ a_{10}(z) = 12\rho^3 (\rho + 1) (z - 1) + 6 \rho^2, \]

\[ a_{11}(z) = 12\rho^3; \]
\[ a_0(z) = \rho^3 \frac{3 - 4\rho}{1 - \rho} (z^3 - 1) - \rho^2 \frac{2\rho^2 - 17\rho^6 + 38\rho^5 - 33\rho^4 + 27\rho^3 - 34\rho^2 + 20\rho - 4}{(1 - \rho)^3} (z^2 - 1) \]
\[ + \rho \frac{2\rho^{15} - 35\rho^{14} + 233\rho^{13} - 848\rho^{12} + 2051\rho^{11} - 3316\rho^{10} + 3559\rho^9 - 2226\rho^8 + 115\rho^7 + 1369\rho^6 - 1582\rho^5 + 1014\rho^4 - 416\rho^3 + 108\rho^2 - 16\rho + 1}{(1 - \rho)^9} (z - 1), \]
\[ (5.75) \]
\[ a_1(z) = -\rho^3 \frac{3\rho^4 - \rho^3 - 10\rho^2 + 13\rho - 9}{1 - \rho} (z^2 - 1) - \rho^2 \frac{2\rho^7 - 23\rho^6 + 70\rho^5 - 27\rho^4 + 14\rho^3 - 22\rho^2 + 26\rho - 8}{1 - \rho} (z - 1) + \rho, \]
\[ (5.76) \]
\[ a_2(z) = -2 \rho^3 \frac{3\rho^3 - 10\rho^2 + 15\rho - 9}{1 - \rho} (z - 1) + 8 \rho^2, \]
\[ (5.77) \]
\[ a_3(z) = 18 \rho^3. \]
\[ (5.78) \]
Figure 17: Comparison of $Q_n = \sum_t Q_{n,t}$ for interesting scenarios. For each different couple of values $(\rho, q)$ the 'Empirical' series represents the empirical distribution while 'Theoretical' is obtained by means of (5.89).
5.8 General expression of $Q(z, y)$

An inspection of formulas (5.50)–(5.78) reveals that the coefficients $a^k_j(z)$ have regular structure. It seems reasonable to make the following:

**Ansatz**

$$ a^k_j(z) = c^k_j + \sum_{r=1}^{j_k-j} c^k_r (z^r - 1). \quad (5.79) $$

**Remark** 5.9. The coefficients $c^k_r$ appearing in (5.79) are only those satisfying $r + j \leq j_k$.

Therefore, we can safely define $c^k_r = 0$ for $r + j > j_k$.

Combining (5.45) and (5.46) with (5.79), we obtain a set of equations. For $j = 0$ and all $k \geq 1$,

$$ c^k_0 + \sum_{r=1}^{j_k} c^k_r (z^r - 1) = \sum_{j=1}^{j_k} \frac{z^j - 1}{j!} \left[ \frac{j \rho}{1 - \rho} (c^k_{j-1,j-1} - \sum_{r=1}^{j_k-j+1} c^k_r (z^r - 1)) \right], \quad (5.80) $$

while for all $k, j \geq 1$,

$$ c^k_0 + \sum_{r=1}^{j_k} c^k_r (z^r - 1) = \sum_{l=1}^{j_k} \frac{(z^{1-j} - 1) - (z^{1-j-1} - 1)}{(1 - j)!} \times $$

$$ \times \left[ \frac{1}{1 - \rho} \left( c^k_{l-1,j-1} - \sum_{r=1}^{j_k-j+1} c^k_r (z^r - 1)) \right) \right] $$

$$ + (1 - \rho) \left( c^k_{j-1,l} - \sum_{r=1}^{j_k-l-1} c^k_r (z^r - 1)) \right] $$

$$ + \rho j \left( c^k_{j-1,j} - \sum_{r=2}^{j_k-j+1} c^k_r (z^r - 1)) \right) $$

$$ + (1 - \rho) \left( c^k_{j-1,j} + \sum_{r=2}^{j_k-j} c^k_r (z^r - 1)) \right) $$

$$ + \rho \left( c^k_{j-1,j} + \sum_{r=1}^{j_k-j} c^k_r (z^r - 1)) \right). \quad (5.81) $$

The set of equations above is then completed by the normalisation condition

$$ c^k_0 = \delta_{k,0} = \begin{cases} 1, & \text{if } k = 0, \\ 0, & \text{otherwise}. \end{cases} \quad (5.82) $$

Condition (5.82) is a direct consequence of (5.25).
Let us now fix a value \( m \geq 0 \) and look for an expansion of \( Q(z, y) \) of order at most \( m \) in \( y \). This means that we are neglecting \( Q_{n,l} \) whenever \( l > m \), i.e.,

\[
Q(z, y) = \sum_{k \geq 0} q^k Q^{(k)}(z, y) + o(q^k),
\]

(5.83)

where

\[
\kappa = \left( \frac{m + 2}{2} \right) - 1.
\]

We note that \( j_k = m \) and \( \kappa \) is the largest integer satisfying this relation. According to (5.79), the number of coefficients \( c^{k,j}_r \) appearing in the expansion (5.83) can be computed as follows. The index \( k \) varies in the range \( \{0, 1, \ldots, \kappa\} \), while \( j \) varies in \( \{0, 1, \ldots, j_k\} \) and \( r \in \{0, 1, \ldots, j_k - j\} \). Thus,

\[
|\{c^{k,j}_r\}| = \sum_{k=0}^{\kappa} \sum_{j=0}^{j_k} \sum_{r=0}^{j_k - j} 1,
\]

\[
= \sum_{k=0}^{\kappa} \sum_{j=0}^{j_k} j_k - j + 1,
\]

\[
= \sum_{k=0}^{\kappa} \sum_{j=1}^{j_k + 1} j,
\]

\[
= \sum_{k=0}^{\kappa} \frac{(j_k + 2)}{2},
\]

\[
= \frac{(2)}{2} + 2 \frac{(3)}{2} + \cdots + (m + 1) \frac{(m + 2)}{2},
\]

\[
= \sum_{j=1}^{m+1} j \frac{j(j+1)}{2},
\]

\[
= \frac{1}{24} (m+1)(m+2)(m+3)(3m+4).
\]

We equate the coefficients of \((z^r - 1)\) in (5.80) and drop \( c^{k,0}_0 \) by (5.82). Then, for \( 1 \leq r \leq j_k \),

\[
c^{k,0}_r = \frac{1}{r!} \left[ \frac{r \rho}{1 - \rho} \left( c^{k-r,r-1}_0 - \sum_{m=1}^{j_k-r-1} c^{k-r,r-1}_m \right) + \right.
\]

\[
\left. + \left( c^{k-r,r}_0 - \sum_{m=1}^{j_k-r} c^{k-r,r}_m \right) \right],
\]

(5.84)

where we agree that the sum \( \sum_{m=a}^{b} \) is null whenever \( b < a \) and that \( c^{k,j}_r = 0 \) whenever \( j + r > j_k \).
Equating the coefficients of \((z^r - 1)\) in (5.81),
\[
c^{k,j}_0 = \rho j c^{k+1,j-1}_0 + c^{k,j}_0, \tag{5.85}
\]
\[
c^{k,j}_r = \frac{(r+1)\rho}{r!} \left( c^{k-r,j+r-1}_0 - \sum_{m=1}^{j_k-r} c^{k-r,j+r-1}_m \right) + (1 - \rho) c^{k-r,j+r}_0 - \frac{(r+j+1)\rho}{(r+1)!} c^{k-r+1,j+r-2}_0 - \sum_{m=1}^{j_k-r+1} c^{k-r+1,j+r-1}_m + (1 - \rho) c^{k-r+1,j+r}_0 + \left( \rho j c^{k-1,j-1}_r + (1 - \rho) c^{k,j}_r \right) \chi_{r \geq 2} + \rho c^{k,j}_r, \tag{5.86}
\]
Again we agree that the sum \(\sum_{m=a}^{b} c^{k,j}_m\) is null whenever \(b < a\) and that \(c^{k,j}_r = 0\) whenever \(j + r = j_k\); further, \(\chi_{r \geq 2}\) is 1 if \(r \geq 2\), and 0 otherwise.

Remark 5.10. By Remark 5.9 and (5.85),
\[
c_{0,1}^k = c_{0,1}^{k-1,1} = c_{0,1}^{1,1}.
\]
We note that last equation is verified by all coefficients \(a^k(z)\) computed above.

Remark 5.11. For orders \(k = (m+1)\) and \(r = j_k - j = m - j > 0\), equation (5.86) yields
\[
c^{k,j}_r = \frac{1}{[j_k - j]!} \rho j_k c^{k-1,j,k-1}_0 = \frac{j_k!}{[j_k - j]!} \rho^k, \tag{5.87}
\]
where we have used Lemma 5.4 from (5.87) to (5.88). Again we note that last equation is verified by all coefficients \(a^k(z)\) we have already computed.

The recursive structure of (5.84) and (5.86) indicates that there exist a one-to-one mapping
\[
(k, j, r) \mapsto \left\{ 1, 2, \ldots, \frac{(m+1)(m+2)(m+3)(3m+4)}{24} \right\}
\]
such that the associated linear system for \(c^{k,j}_r\) is lower triangular. One possibility to achieve this mapping is to define the following strict total order relation: \((k, j, r) < (k', j', r')\) if and only if
\[
k < k' \lor k = k', j < j' \lor k = k', j = j', r < r'.
\]
The linear system has a non-trivial solution due to (5.82). Therefore, the ansatz (5.79) is justified ex post, and we have proved the following result:

**Lemma 5.7** Let \(0 \leq \rho < \phi\). The bivariate generating function \(Q(z, y)\) defined in Section 5.2 can be written as
\[
Q(z, y) = 1 + \rho (z - 1) + \sum_{j \geq 1} (y - z)^j \sum_{k \in \left(\frac{1}{2} \right)^j} q^k p_{j,k}(z)
+ \frac{1 + \rho (z - 1)}{1 - \rho} \sum_{j \geq 1} (z - 1)^j \sum_{k \in \left(\frac{1}{2} \right)^j} q^k p_{j,k}(0),
\]
where $P_{j,k}(z)$ is a polynomial of degree $j_{k-j} - j$ in $z$.

**Proof.** Substituting (5.79) into (5.34) we easily see that $A^k_j(z)$ is a polynomial in $z$ of degree $j_{k-j} - j$. Let $P_{j,k} = A^k_j(z)/j!$. Inverting the order of the sums in (5.35), we arrive to the thesis.

## 5.9 Closing Remarks

Although equation (5.28) represents the complete solution of the problem, its use is quite involved due the recursive nature of the relation. The discussion in Section 5.8 has shown that a truncated expansion in powers of $q$ can be computed by solving a linear system. Finding a closed expression for $Q(z,y)$ still remains an open problem, though.

Numerical simulations can give an idea of the quality of the approximation given by the truncated expansion (5.49). The distribution of the queue length can be found as the marginal of the empirical distribution $Q_{n,l}$, or it can be found by

$$Q_{n,l} = \frac{1}{n!} \frac{\partial^n}{\partial z^n} Q(z,y) \bigg|_{(z,y) = (0,1)},$$

$$= \frac{1}{n!} \sum_{k=0}^l q^k \frac{\partial^n}{\partial z^n} Q^{(k)}(z,y) \bigg|_{(z,y) = (0,1)}. \quad (5.89)$$

Figure 17 shows a comparison between the distribution of the queue length empirically obtained via numerical simulations of an EDA/D/1 system and the distribution obtained by means of (5.89). Here we see that if the average load of the system is kept at a moderate level then the truncated model is capable to deliver a very good accordance with the simulated data. This holds true even for scenarios where each customer has a great probability to arrive out of its pre-scheduled slot. Conversely, if the traffic index $\rho$ is close to 1, then even small values of $q$ are sufficient for discrepancies to arise between the simulated and the theoretical distributions. In this cases a higher-order truncation should be enough to achieve a better approximation.

The EDA/D/1 queue demonstrate a deep combinatorial structure, which emerges quite soon in the study of the model. At the present stage it is not possible to exclude that a bright analysis of this combinatorial structure can lead to a better solution. Here we just want to give a glimpse of the combinatorial-structure richness through the following remark. The generating function of the marginal distribution of the number of late customers is $Q(1,y)$. According to (5.25) and (5.28),

$$Q^{(k)}(1,y) = \sum_{j \geq 0} \frac{(y-1)^j}{j!} \left( j \rho \ a_{j-1}^{k-1}(1) + a_j^{k-1}(1) \right). \quad (5.90)$$

From (5.17) $a_j^1(1) = \frac{\rho^j}{j!} Q^{(1)}(1,y) \bigg|_{y=1}$, which is to say, $a_j^1(1)$ is the coefficient of $q^j(y-1)^j$ appearing in (5.17). Since a contribution $q^j(y-1)^j$ can be obtained from the infinite product (5.17) by using only $j$ terms different from 1,

$$a_j^1(1) = j! \rho^j d(l,j),$$

**Empirical queue vs. truncated expansion**
where \( d(l, j) \) is the number of partition of the integer \( l \) in terms of \( j \) distinct parts, i.e.,

\[
d(l, j) = \sum_{l_1 < \ldots < l_j} \sum_{m} t_m = l
\]

Hence, (5.90) can be rewritten as

\[
Q^{(k)}(1, y) = \sum_{j \geq 0} \frac{(\rho y - 1)^j}{j!} [d(k - j, j - 1) + d(k - j, j)].
\]

Using Ferrers diagrams, see e.g. [Tat05], it is possible to prove

\[
d(k, j) = d(k - j, j - 1) + d(k - j, j),
\]

which immediately yields the equivalent expression

\[
Q^{(k)}(1, y) = \sum_{j \geq 0} \frac{(\rho y - 1)^j}{j!} d(k, j).
\]
In this chapter we study the EDA/D/1 queueing system approaches equilibrium. In particular, we are interested in investigating the possibility to use the ideas from Chapter 3 to show the presence of cutoff behaviour. The main challenge here is that we do not have complete knowledge of the stationary distribution of the process. Nevertheless, cutoff is related more to the transient behaviour than to the stationary state. According to the picture we have developed in Chapters 2 and 3, cutoff is often triggered by a ballistic approach of the chain to the relevant quantiles of the stationary distribution.

In Section 6.1 we introduce \( \alpha_t \), an auxiliary process which can be used to retrieve some properties of the transience of the EDA/D/1 queue. Section 6.2 gives strong evidence that the stationary measure of EDA/D/1 is concentrated in a certain region, and that \( \alpha_t \) has a drift towards it. This approach closely recalls the use of the projection we have already exploited in Chapters 3 and 4. Using the auxiliary MC \( \alpha_t \) we prove cutoff in Section 6.4 for a simplified model. Then, in Section 6.5, we show that many of the ideas presented in Section 6.4 can be extended to the general case as well.

### 6.1 THE \( \alpha_t \) AUXILIARY CHAIN

In Chapter 5 we have conveniently described EDA/D/1 as the bivariate chain \((n_t, l_t)\), defined by (5.6) and (5.7). The feasible transitions of this chain are displayed by Figure 16 above. In particular, we note that the quarter plane can be decomposed into lines inclined at an angle of \(3\pi/4\) such that only transitions between states that lie on adjacent lines are allowed.

It is then natural to define the auxiliary process \( \alpha_t = n_t + l_t \). This is not a MC but rather a Markov modulated chain that takes non-negative integer values. Given the realisation of the current state \((n_t, l_t)\), the auxiliary process \( \alpha_t \) is a MC whose transition probabilities can be explicitly computed, see (6.1) below. The chain \( \alpha_t \) represents which diagonal the bivariate chain \((n_t, l_t)\) lies on at time \( t \). The chain \((n_t, l_t)\) can take transitions only to states that lie on the same or nearest-neighbour diagonal. Thus, it makes sense to take the following equivalence relation between states of \( \mathbb{N} \times \mathbb{N} \):

\[
(i, j) \sim (h, k) \quad \iff \quad i + j = h + k. \tag{6.1}
\]

Then \( \alpha_t \) is a nearest-neighbour chain on \( \mathbb{N} \times \mathbb{N} / \sim \).

The bivariate chain \((n_t, l_t)\) governs the evolution of \( \alpha_t \) because the transitions to an outer* diagonal are prohibited unless \( n_t = 0 \). Conversely, if at time \( t \) the EDA/D/1 system is in a state with an empty queue, all transitions to an inner diagonal are precluded. Figure 16 clearly shows these features. The transition probabilities of \( \alpha_t \) are the following:

\[
\begin{align*}
P(\alpha_{t+1} = i - 1 \mid \alpha_t = i) &= (1 - \rho)(1 - \bar{P}(i, t)), \\
P(\alpha_{t+1} = i \mid \alpha_t = i) &= \rho + \bar{P}(i, t) - 2\rho \bar{P}(i, t), \\
P(\alpha_{t+1} = i + 1 \mid \alpha_t = i) &= \rho \bar{P}(i, t),
\end{align*}
\]

* A given diagonal line divides the quarter plane into the inner part, which contains the origin, and the outer part, complementary to the former.
where
\[
P(i, t) = \mathbb{P}(n_t = 0, l_t = i | \alpha_t = i),
\]
\[
= \frac{\mathbb{P}(n_t = 0, l_t = i)}{\sum_{k=0}^{l_t} \mathbb{P}(n_t = k, l_t = i - k)}.
\] (6.3)

As we have discussed in Chapter 5, the bivariate chain \((n_t, l_t)\) has a unique joint stationary distribution \(Q_{n,l}\) under the hypothesis that \(\rho < 1\). Therefore, there exists a unique stationary distribution for the chain \(\alpha_t\), given by the formula
\[
\pi_\alpha = \sum_{n,l \geq 0} Q_{n,l}, \quad \alpha \geq 0.
\]

Then, we have the following:

**Lemma 6.1** With respect to the notation introduced in Chapter 5, the generating function of \(\pi_\alpha\) is \(Q(z, z)\).

**Proof.** Let us call \(A(z)\) the generating function of \(\pi_\alpha\). Then,
\[
A(z) = \sum_{\alpha \geq 0} \pi_\alpha z^\alpha,
\]
\[
= \sum_{\alpha \geq 0} \sum_{n,l \geq 0} Q_{n,l} z^n z^l,
\]
\[
= Q(z, z).
\]

Another interesting result about the chain \(\alpha_t\) is the characterisation of the transition probabilities when the system is at equilibrium. For \((n_t, l_t)\) distributed according to the stationary distribution, equation (6.3) yields
\[
\hat{p}(i) = \frac{Q_{0,i}}{\sum_{k=0}^{l_t} Q_{k,i-k}} = \frac{Q_{0,i}}{\frac{1}{i!} \left( \frac{1 + \rho (z - 1)}{1 - \rho} \frac{\partial^i}{\partial z^i} Q(0, z) \right)_{z=0}}.
\] (6.3a)

Differentiating (5.14),
\[
\frac{1}{i!} \left( \frac{1 + \rho (z - 1)}{1 - \rho} \frac{\partial^i}{\partial z^i} Q(0, z) \right)_{z=0} = \frac{1}{i!} \left[ \frac{1 + \rho (z - 1)}{1 - \rho} \frac{\partial^i}{\partial z^i} Q(0, z) + \rho \frac{i - 1}{1 - \rho} \frac{\partial^{i-1}}{\partial z^{i-1}} Q(0, z) \right]_{z=0},
\]
\[
= Q_{0,i} + \frac{\rho}{1 - \rho} Q_{0,i-1}.
\]

For \((n_t, l_t) \sim Q_{n,l}\), the transition probabilities of \(\alpha_t\) are then the following. For \(i \geq 1\),
\[
P(i, i - 1) = q_i = \frac{\rho}{1 + \frac{\rho}{1 - \rho} Q_{0,i-1}} Q_{0,i-1},
\] (6.4)
\[
P(i, i) = r_i = \frac{1 - \rho}{1 + \frac{\rho}{1 - \rho} Q_{0,i-1}} Q_{0,i-1},
\] (6.5)
\[
P(i, i + 1) = p_i = \frac{\rho}{1 + \frac{\rho}{1 - \rho} Q_{0,i-1}} Q_{0,i-1},
\] (6.6)
whereas for $i = 0,$
\[ P(0, 0) = r_0 = 1 - \rho, \]  
\[ P(0, 1) = p_0 = \rho. \]  

**Remark 6.1.** Equations (6.4)-(6.6) demonstrate that for $(n_t, 1_t) \sim Q_{n,1}$, the process $\alpha_t$ is a homogeneous BDC.

### 6.2 The Bulk of the Stationary Distribution

Suppose that the bivariate chain $(n_t, 1_t)$ is distributed according the stationary joint measure $Q_{n,1}$. Then, $\alpha_t = n_t + 1_t$ is a homogeneous BDC and its transition rates are given by (6.4)-(6.6). From the standard formula for the stationary distribution of a BDC,
\[ \pi_\alpha = \sum_{l=0}^{\alpha} Q_{\alpha-1,l} = Q_{0,0} \prod_{l=1}^{\alpha} \frac{p_{l-1}}{q_l}, \]
\[ = \rho Q_{0,0} \prod_{l=1}^{\alpha-1} \frac{Q_{0,1}}{Q_{0,l-1}} q_{\alpha-1}^{-1}. \]  

Aiming at the proof of cutoff, equation (6.7) is extremely important because it is a tool to bound the measure $Q_{n,1}(T_L)$ of a triangle
\[ T_L = \{(n, 1) \in \mathbb{N} : n + 1 \leq L\}. \]  

In particular, we can estimate the asymptotic behaviour of $\frac{Q_{0,1}}{Q_{0,0}} - L$ then easily obtain a concentration property like (3.26) for the family of nested triangles $(T_L)_L$.

In Chapter 5 we have seen that $Q(1, y)$, the generating function of the number of late customers, can be written in the form of an infinite product. If we expand the product and rearrange in powers of $\rho(y - 1)$ then
\[ Q(1, y) = \prod_{k \geq 0} \left(1 + \rho q^{k+1}(y - 1)\right), \]
\[ = 1 + \sum_{k \geq 1} \rho^k(y - 1)^k \sum_{m \geq \binom{k+1}{2}} d(m, k) q^m, \]  

where $d(m, k)$ is the number of partitions of $m$ in $k$ distinct parts. We have already encountered (6.6) in Section 5.9, although disguised as (5.90). We can justify (6.9) by expanding the infinite product and look, for example, at the coefficient of $q^2(y - 1)^2$. This is given by
\[ \sum_{k_1 \geq 1} \sum_{k_2 \geq k_1} q^{k_1+k_2} = \sum_{m \geq 3} q^m d(m, 2). \]

Our goal is to obtain an expression for the $l$-th derivative of $Q(1, y)$ because it is proportional to the marginal probability $Q_{n,l} = \sum_{n \geq 0} Q_{n,1}$. In addition, $Q_{n,1}$ is an upper bound on $Q_{0,1}$, the probability to have null queue and $1$ late customers at stationarity. Therefore let us look at the term in $y^l$ of $Q(1, y)$, i.e. $[y^l]Q(1, y)$. We have the following:

**Theorem 6.2** With respect to the notation introduced in Chapter 5.1, let $Q_{n,l}$ be the marginal distribution of the number of late customers, i.e. $Q_{n,l} = \sum_{n = 0} Q_{n,1}$. Then,
\[ Q_{n,l} = \sum_{k \geq 1} \rho^k q^{k+1} \binom{k}{1} (-1)^{k-l} \left[ 1 + \prod_{l=1}^{k} \frac{1}{1-q^l} \right]. \]
Theorem 6.2 is a direct consequence of two results from number theory. The first result can be found in [YY64] and links the number of partitions in k distinct parts with the number of partitions into at most k parts.

**Lemma 6.3** If \( m > \binom{k+1}{2} \) then the number of partitions of \( m \) in \( k \) distinct parts equals the number of partitions of \( m - \binom{k+1}{2} \) into at most \( k \) parts (not necessarily distinct).

The second lemma states that the number of partitions into at most \( k \) parts equals the number of partitions in parts less or equal than \( k \).

**Lemma 6.4** Let \( p_{\leq k}(m) \) be the number of partitions of \( m \) in parts that do not exceed \( k \). Then \( p_{\leq k}(m) \) equals the number of partitions of \( m \) into at most \( k \) parts.

The generating function of the number of partitions of \( m \) into at most \( k \) parts is

\[
P_{\leq k}(x) = \sum_{m \geq 1} p_{\leq k}(m) x^m = \prod_{i=1}^{k} \frac{1}{1 - x^i}.
\]

Lemma 6.4 is relatively easy to prove. Using Ferrers diagrams, it is readily seen that a partition in parts that do not exceed \( k \) and a partition into at most \( k \) parts are conjugate to one another. For more details, see [And98, HW79].

**Proof of Theorem 6.2.** The marginal distribution of the number of late customers \( Q_{0,1} \) is the coefficient of \( y^1 \) in the expression of \( Q(1, y) \). By the Binomial Theorem, from (6.9)

\[
[y^1]Q(1, y) = \sum_{k \geq 1} \rho^k \binom{k}{1} (-1)^{k-1} q^{\binom{k+1}{2}} \times

1 + \sum_{m > \frac{k+1}{2}} d(m,k) q^{m-\binom{k+1}{2}}, \quad (6.11)
\]

where we have used the obvious fact that \( d(m,k) = 1 \) when \( m \) is the \( k \)-th triangular number. We can now use Lemma 6.3 and Lemma 6.4 to rewrite (6.11) as

\[
[y^1]Q(1, y) = \sum_{k \geq 1} \rho^k \binom{k}{1} (-1)^{k-1} q^{\binom{k+1}{2}} \left[ 1 + \sum_{m \geq 1} p_{\leq k}(m) q^m \right],
\]

\[
= \sum_{k \geq 1} \rho^k q^{\binom{k+1}{2}} \binom{k}{1} (-1)^{k-1} \left[ 1 + \prod_{i=1}^{k} \frac{1}{1 - q^i} \right].
\]

Using (6.10), we can derive an upper bound on \( Q_{0,1} \), namely,

\[
Q_{0,1} \leq \psi_1 \rho^1 q^{\binom{1}{2}}, \quad (6.12)
\]

where

\[
\psi_1 = 1 + \sum_{k \geq 1} q^{\frac{k^2 + 2kt + k}{2}} (-\rho)^k \binom{k+1}{1} \left[ 1 + \prod_{i=1}^{k} \frac{1}{1 - q^i} \right].
\]

Last series is clearly dominated by the first term, so \( \psi_1 \) can be bounded from above by a constant.
Let us now suppose that $q$ is smaller than the analyticity radius we have estimated in Section 5.6. By (5.36) and Remark 5.8, for $l \geq 1$,

\[
Q_{0,1} = \frac{1}{l!} \frac{\partial^l}{\partial z^l} Q(0,z) \bigg|_{z=0},
\]

\[
= \frac{1}{l!} \frac{\partial^l}{\partial z^l} \left[ \sum_{k \geq 0} q^k \sum_{j=1}^{l_k} \frac{z^j - 1}{j!} A^j_k(0) \right] \bigg|_{z=0},
\]

\[
= \frac{q^{(1 + l)/2}}{l!} \sum_{k \geq l} q^{-k(1 + l)/2} A_k^l(0). \tag{6.14}
\]

To get (6.14) from (6.13), we just note that the non-zero terms come from those values of $k$ such that $j_k \geq 1$, i.e. $k \geq \frac{l(l+1)}{2}$. From (5.34), (6.14), Lemma 5.4, and Lemma 5.5 the first term of the series appearing at the RHS of (6.14) is $\mathcal{l}!p^l$. Then,

\[
Q_{0,1} = q^{(1 + l)/2} p^l \left[ 1 + \frac{p^{-1}}{l!} \sum_{k \geq 1} q^k A_k^{(l+1)/2}(0) \right]. \tag{6.15}
\]

By comparison of (6.12) with (6.15),

\[
\frac{p^{-1}}{l!} \sum_{k \geq 1} q^k A_k^{(l+1)/2}(0) < \infty.
\]

On the other hand, the chain $(n_t, l_t)$ is irreducible, and as such it must be $Q_{0,1} > 0$. Therefore, there exists a positive $\psi'_l$ such that

\[
1 + \frac{p^{-1}}{l!} \sum_{k \geq 1} q^k A_k^{(l+1)/2}(0) > \psi'_l.
\]

Opposite to the case of $\psi_l$, we do not know the asymptotic behaviour of $\psi'_l$ for large $l$. Nevertheless, there are a few heuristics that let us believe $\psi'_l \geq c$, where $c$ is a positive constant. They are discussed in the next section.

### 6.3 HEURISTIC LOWER BOUND ON $Q_{0,1}$

A first intuitive argument in favour of the asymptotic lower bound

\[
Q_{0,1} \geq cp^l q^{(l+1)/2}
\]

uses the well-known relation between stationary probabilities and expected return times. Due to the positive recurrence proved in Chapter 5,

\[
Q_{0,1} = \frac{C}{\mathbb{E}[\tau_{(0,1)}^C]},
\]

where $C$ is a numerical constant and

\[
\tau_{(0,1)}^C = \min\{t > 0 : n_t = n_0 = 0, l_t = l_0 = l\}
\]

is the first return time to the state $(0,1)$. In the next discussion we try to heuristically characterise the tube of typical trajectories, that is, the set of trajectories that the system is most likely to follow while it returns to the initial state $(0,1)$. The whole analysis is carried out for $l$ large enough.
The key ingredient of the analysis is the motion of the vertical component $l_t$ of the bivariate chain $[n_t, l_t]$. This is a MC with transition probabilities

\begin{align}
P(k, k + 1) &= \rho q^{k+1}, \\
P(k, k) &= \rho k(1 - q)q^k + (1 - \rho)q^k, \\
P(k, k - j) &= \rho \binom{k+1}{j}(1 - q)^j q^{k+1-j} + (1 - \rho) \binom{k}{j}(1 - q)^j q^{k-j}, \quad j = 1, \ldots, k.
\end{align}

In particular, the downward transition rates (6.18) satisfy

\begin{equation}
\sum_{j=1}^{k} P(k, k - j) = 1 - q^k [1 + \rho(1 - q)(k - 1)].
\end{equation}

Hence, $l_t$ has a very sound downward drift.

An indication of the strength of the drift comes from $l_1$. Due to the binomial nature of the transition probabilities (6.16)-(6.18), the expected position after one step is $(1 - \rho)q$ with an uncertainty $O(1)$. Moreover, let us imagine that the customers of EDA/D/1 have not too large average delays $\bar{\xi}_t$. Depending on how small $q$ is, it may well be that $1/2$ is on the tail of the binomial $\mathcal{B}(l, q)$. In this case the probability of all the transitions $l_0 \rightarrow l_1$, $l_1 \geq 1/2$, is exponentially small. The presence of such a tough drift, stronger than in any other model we have already encountered, is responsible for the vertical component $l_t$ to quickly drop from the initial height $l$ to a much lower level. The very same drift is likely to ensure vertical cutoff in the sense of (2.42), but the latter is just a mere conjecture.

The paths of typical trajectories can be ideally divided in two, namely, a first part in which the system returns to the triangle $T_l$ defined by (6.8) and the subsequent reaching of the upper vertex $(0, l)$ from within $T_l$. The dynamics of $(n_t, l_t)$ is indeed such that, starting from inside $T_l$, the only way to visit a state which does not belong to this set is through the state $(0, l)$.

The arguments that support this vision of the typical trajectories are due to the chain drift, both in the vertical and horizontal directions. Some words have been already spent on the existence of the vertical drift. We now study the motion of $l_t$ by coupling it with $\tilde{l}_t$, a BDC defined by the transition probabilities (6.16), (6.17), and $P(k, k - 1)$ given by the RHS of (6.19). It is clear from the structure of the transition probabilities that there exists coupling a such that

- $l_0 = \tilde{l}_0$;
- $\tilde{l}_t \geq l_t$ for all $t \geq 0$.

For large $k$, the chain $\tilde{l}_t$ clearly has a very strong tendency to take transitions $k \rightarrow k - 1$. The stationary distribution of the auxiliary BDC $\tilde{l}_t$ is

$$\tilde{\pi}_k = \tilde{\pi}_0 \rho^k q^{k+1} \prod_{j=1}^{k} \frac{1}{1 - q^k [1 + \rho(1 - q)(k - 1)]}.$$ 

Last formula gives the energetic landscape where the motion of the auxiliary BDC $\tilde{l}_t$ takes place. This landscape is an energy well and the position of

\* The scenario with small delays’ variability is notably fitting in ATM context, see Chapter 7 and in particular Sections 7.5 and 7.6.
the bottom depends on the interplay between $\rho$ and $q$. The walls of such an energy well rapidly get steeper and steeper with the distance from the bottom and become asymptotically vertical. The picture remembers with certain respects the Freidlin-Wentzell regime [OVos, Ch. 6], where the walls become vertical in the limit for the inverse temperature $\beta \to \infty$. The chain $l_0$, started at time $t = 0$ from $l$, is then likely to hit the bottom of the well before reaching the state $l + 1$. It is also reasonable to expect that the probability of such an event tends exponentially fast to 1 for $l \to \infty$. Further, it is more than well-thought to guess that, starting from the well bottom, the chain visits for the first time the state $l$ in a time interval comparable to $\rho^1 q^{(1+\epsilon)}$. This is for example the leading order of $E[\tau_{0-l}]$ given by (2.45).

Looking at $l_0$, this picture can be only sturdier. In fact, every time $l_0$ takes a downward transition, i.e. $k \to k - 1$, the chain $l_0$ jumps an average number levels downwards equal to $k(1 - q) - \rho q$. If $q$ is small, the mean of the difference $l_0 - l$ is expected to be quite large.

Let us now focus on the horizontal drift. As we have seen in Section 6.2, conditioned on the event $n_0 \neq 0$, the chain $\alpha_t$ can only take a self-transition or move to the left, i.e. $\alpha \to \alpha - 1$; the two transitions have probability $\rho$ and $1 - \rho$, respectively. Conversely, conditioned on the event $n_0 = 0$, the chain $\alpha_t$ takes a self-transition or move to the right, i.e. $\alpha \to \alpha + 1$; the two transitions have probability $1 - \rho$ and $\rho$, respectively.

All in all, our guess for the paths of the typical trajectories of the bivariate chain starting from $(0,1)$ is the following:

1. the second component of the bivariate chain, $l_t$, quickly drops to a level much smaller than $l$;
2. the queue length increases by an amount which is about the loss in height of $l_t$;
3. $l_t$ does not escape far from this new position until a time is passed which is super-exponential in the distance to cover;
4. as a consequence, the probability $\hat{P}(i,t)$, defined by (6.3), is negligible for $i \geq 1$ and $t < O(q^{-\frac{1}{2}})$.

Then, the mean number of steps to see $\alpha_t$ decrease by 1 when $\alpha_t \neq T_t$ is approximately $\frac{1}{1-\rho}$ from (6.2). The mismatch between the times scales $1/(1-\rho)$ and $q^{-\frac{1}{2}}$ is very strong. It is then natural to guess that after an initial phase where $l_t$ drops to a level lower than $l$, the motion of $(n_t, l_t)$ takes place in a horizontal conduit. This idea is sketched by Figure 18.

Having in mind the description of the typical trajectories that return to $(0,1)$, we can completely neglect the description of the $n_t$ component after the chain $(n_t, l_t)$ has reached the triangle $T_t$. Indeed, the chain has no possibility to leave such a triangle before having hit the state $(0,1)$. A rough upper bound to the hitting time of $(0,1)$ starting from inside $T_t$ is easy to obtain. It suffices to mark that the chain will surely hit the state $(0,1)$ if a sequence of 1 consecutive upward transitions happens. A transition $k \to k + 1$ has probability $\rho q^{k+1}$, then a sequence of 1 consecutive transitions happens with probability

$$\Psi = \prod_{k=0}^{l-1} \rho q^{k+1} = \rho^1 q^{\frac{l-1}{2}}.$$
Figure 18: Paths of the typical trajectories from \((0, 1)\) back to \(T_l\). The initial fall from \((0, 1)\) into the conduit is idealised. The blue line illustrates the case where the chain \(l_t\) goes against the drift for a short while. The component \(l_t\) initially increases for a certain number of steps, it eventually reaches the conduit, and from this moment on the motion is confined therein.

The average time to see such an event is proportional to \(\Psi^{-1}\), which is a contribution to the mean return time \(\mathbb{E} \left[ \tau_{(0,1)} \right] \) much larger than the time spent in the conduit. As a result,

\[
\mathbb{E} \left[ \tau_{(0,1)} \right] = O \left( \Psi^{-1} \right),
\]

so that

\[
Q_{0,1} = c \rho l q^{l+1}. 
\]

Another argument that can be used to support the conjecture \(Q_{0,1} \geq c \rho l q^{l+1}\) comes from the study of the marginal distribution \(Q_{1,1}\). For a fixed, large value of \(l\), simulations show that the fraction of time spent by the system in states \((n, l)\) with \(n \neq 0\) is negligible with respect to the fraction of time spent in \((0, 1)\). In other words, the probability mass \(Q_{1,1}\) is asymptotically concentrated in \((0, 1)\). This remark is not surprising whatsoever. Let us briefly consider \(\mathbb{E} \left[ \tau_{(1,1)} \right] \), i.e., the return time to the state \((1, 1)\). The mechanism by which the chain returns there is pretty much the same as the one we have just discussed for \((0, 1)\). Under the influence of the vertical drift, the system quickly reaches the conduit and from there it enters \(T_{l+1}\). The key point in the analysis of the return trajectory is that the chain does not reach \((1, 1)\) by \textit{climbing} the \(\alpha\)-line \(\{(i, j) : i + j = l + 1\}\). In fact, this trajectory would require a super-exponential time in \(l\); meanwhile, by the effect of the horizontal drift, the system moves to an inner triangle. Using the very same argument as above we expect the chain to visit \((0, 1)\) before reaching \((1, 1)\).

The most likely way to reach \((1, 1)\) from state \((0, 1)\) is by means of an eastward transition in the quarter plane. This transition has probability \(\rho(1+1)(1-q)q^l\) to happen, and any other path is much less likely. For example, the trajectory \((0, l) \rightarrow (0, l+1) \rightarrow (1, 1)\) has probability

\[
\rho q^{l+1} \cdot (1-\rho)(l+1)(1-q)q^l = O(l q^{2l+1}).
\]
Therefore we intuitively expect that
\[ \mathbb{E} \left[ Y_{(1,1)}^c \right] = O \left( \frac{q^{-1}}{T} \mathbb{E} \left[ Y_{(0,1)}^c \right] \right), \]
and so
\[ \frac{Q_{n,1}}{Q_{0,1}} \xrightarrow{n \to \infty} 0 \quad n \neq 0. \]

### 6.4 Cutoff for the Truncated Model

In the present section we prove cutoff for a simplified version of EDA/D/1, a model studied in Chapter 5. In the EDA arrival stream (5.1) customers may be late for an arbitrary number of unitary time intervals. Such an assumption might be not acceptable in models of actual queues. Chapter 7 deals with a generalisation of EDA/D/1 to delays other-than-exponential. It shows as well that the resulting class of models is really fitting for actual ATM systems. The latter are examples of a real situation where, due to safety issues and global regulations, a customer can not mature too large a delay without being deleted.

Following this remark, let us fix a large, arbitrary integer number \( L \) and let us consider an EDA/D/1 system. From the arrival stream (5.1) we remove all the customers with a delay \( \xi \) larger than \( L \). According to Remark 5.7, this means we are truncating the model by neglecting all the events of probability \( o \left( q^{l+1}/2 \right) \).

**Remark 6.2.** According to the usual fashion in the cutoff literature, a subscript \( n \) has been added in Chapters 2–4 to all the quantities considered therein. In this section we consider a chain which is defined on an infinite state space, and the subscript is not needed anymore as we do not take any limit for \( n \to \infty \). This new setting requires also a modification of Definitions 2.1 and 2.7, that we discuss after formally defining the truncated model.

The state space of the truncated model is the infinite lattice strip
\[ \Omega = \{ (i,j) \text{ s.t. } 0 \leq i, 0 \leq j \leq L \}\]

Let us define a new bivariate MC \((\hat{n}_t, \hat{t}_t)\) that takes transitions according to the transition kernel listed below:

- **For** \( i > 0 \text{ AND } j \neq L \),
  \[ \begin{align*}
P((i,j), (i+a-1, j+1-a)) &= \rho b_{a,j+1}, & 0 \leq a \leq j+1, \\
P((i,j), (i+a-1, j-a)) &= (1-\rho) b_{a,j}, & 0 \leq a \leq j;
\end{align*} \]

- **For** \( i > 0 \text{ AND } j = L \),
  \[ \begin{align*}
P((i,j), (i+a-1, j-a)) &= \rho b_{a,j}, & 0 \leq a \leq L, \\
P((i,j), (i+a-1, j-1-a)) &= (1-\rho) b_{a,j-1}, & 0 \leq a \leq L-1;
\end{align*} \]

- **For** \( i = 0 \text{ AND } j \neq L \),
  \[ \begin{align*}
P((i,j), (a,j+1-a)) &= \rho b_{a,j+1}, & 0 \leq a \leq j+1, \\
P((i,j), (a,j-a)) &= (1-\rho) b_{a,j}, & 0 \leq a \leq j;
\end{align*} \]

- **For** \( i = 0 \text{ AND } j = L \),
  \[ \begin{align*}
P((i,j), (a,j-a)) &= \rho b_{a,j}, & 0 \leq a \leq L, \\
P((i,j), (a,j-1-a)) &= (1-\rho) b_{a,j-1}, & 0 \leq a \leq L-1;
\end{align*} \]
where \( b_{k,1} = \left( \frac{1}{q} \right) (1 - q)^k q^{1-k} \).

Definitions 2.1 and 2.7 are tailored for finite MCs but not fitting in the present framework. The idea proposed in [MY01] to overcome this problem is defining the initial measure as a mass concentrated in a given point \( a \), and then replace the limits for \( n \to \infty \) appearing in (2.18)–(2.20) and (2.42) with a limit for \( a \to \infty \). Following this idea, we consider hereafter the initial measure

\[
\mu^0((i,j)) = \delta_{(L_j),(a,0)} \quad \forall (i,j) \in \Omega. \tag{6.22}
\]

Remark 6.3. In the language of queues (6.22) corresponds to starting the system with a large queue length \( a \) and no late customer. The proof of cutoff in the sense described above sharply characterises the transience of the queueing system under conditions of high congestion. In particular, cutoff means that the system steadily reaches normal operation conditions in a time which is very well approximated by a deterministic quantity. Moreover, it is clear from the proof of cutoff below that the latter can be carried out in exactly the same way even if the system is started uniformly at random from the states \{ \( (i,j) \) s.t. \( \ i + j = a \) \}.

Similarly to the approach followed in Section 6.1, we can define the auxiliary chain \( \bar{\alpha}_t = \bar{n}_t + \bar{l}_t \) having transition probabilities (6.2)–(6.3). Conditioned on the event \( \bar{\alpha}_0 \geq L + 1 \), the probability of \( \bar{n}_t \) being \( 0 \) is null by the truncation. For \( i \geq L + 1 \), \( \bar{\alpha}_t \) is actually a BDC with transition probabilities

\[
\hat{P}(k,k+1) = \hat{p}_k = 0, \\
\hat{P}(k,k) = \hat{\rho}_k = \rho, \\
\hat{P}(k,k-1) = \hat{q}_k = 1 - \rho. \tag{6.23}
\]

Let us define the subset

\[ \hat{T}_{L+1} = \{(i,j) \in \Omega \text{ s.t. } j \leq L, i + j \leq L + 1 \}. \]

From (6.23) we see that the states in \( \Omega \setminus \hat{T}_{L+1} \) are clearly null recurrent. On the contrary, by a similar argument as the one we have used in Chapter 5, we see that the states in \( \hat{T}_{L+1} \) are positive recurrent. This is quite natural as \( \hat{T}_{L+1} \) is positive invariant for the truncated chain \( (\bar{n}_t, \bar{l}_t) \). The chain is also clearly aperiodic so that there exists a unique stationary measure \( \hat{Q}_{n_0} \) and \( \hat{Q}_{n_1} \hat{T}_{L+1} = 1 \).

Now, let \( \hat{\tau}_{L+1} \) be the hitting time of \( \hat{T}_{L+1} \) and take \( a \) sufficiently large. Using the auxiliary chain \( \bar{\alpha}_t \) it is trivial to show that the truncated model exhibits cutoff at mean times. From equations (6.22) and (6.23),

\[
\mathbb{E}[\hat{\tau}_{L+1}] = \frac{a}{1 - \rho}, \tag{6.24}
\]

\[
\sigma[\hat{\tau}_{L+1}] = O(\sqrt{a}). \tag{6.25}
\]

Equations (6.24) and (6.25) immediately give the convergence in probability of Definition 2.7.

Cutoff in total variation can be proved using Theorems 3.3 and 3.4 with \( A_{\bar{n},0} = \hat{T}_{L+1} \) for all \( \bar{n} \geq 1 \). We must bear in mind that the limits for \( n \to \infty \) are now limits for \( a \to \infty \). All the hypotheses save those regarding \( \Delta_n \) can be trivially checked by means of the auxiliary chain \( \bar{\alpha}_t \), thus we skip

\[ \begin{align*}
\text{Remark 6.3.} & \quad \text{In the language of queues (6.22) corresponds to starting the system with a large queue length } a \text{ and no late customer. The proof of cutoff in the sense described above sharply characterises the transience of the queueing system under conditions of high congestion. In particular, cutoff means that the system steadily reaches normal operation conditions in a time which is very well approximated by a deterministic quantity. Moreover, it is clear from the proof of cutoff below that the latter can be carried out in exactly the same way even if the system is started uniformly at random from the states } \{(i,j) \text{ s.t. } i + j = a\}. \\
\text{Similarly to the approach followed in Section 6.1, we can define the auxiliary chain } \bar{\alpha}_t = n_t + l_t \text{ having transition probabilities (6.2)–(6.3). Conditioned on the event } \bar{\alpha}_0 \geq L + 1, \text{ the probability of } n_t \text{ being } 0 \text{ is null by the truncation. For } i \geq L + 1, \text{ } \bar{\alpha}_t \text{ is actually a BDC with transition probabilities } \hat{P}(k,k+1) = \hat{p}_k = 0, \hat{P}(k,k) = \hat{\rho}_k = \rho, \hat{P}(k,k-1) = \hat{q}_k = 1 - \rho. \text{ Let us define the subset } \hat{T}_{L+1} = \{(i,j) \in \Omega \text{ s.t. } j \leq L, i + j \leq L + 1\}. \text{ From (6.23) we see that the states in } \Omega \setminus \hat{T}_{L+1} \text{ are clearly null recurrent. On the contrary, by a similar argument as the one we have used in Chapter 5, we see that the states in } \hat{T}_{L+1} \text{ are positive recurrent. This is quite natural as } \hat{T}_{L+1} \text{ is positive invariant for the truncated chain } (n_t, l_t). \text{ The chain is also clearly aperiodic so that there exists a unique stationary measure } \hat{Q}_{n_0}, \text{ and } \hat{Q}_{n_1} \hat{T}_{L+1} = 1. \text{ Now, let } \hat{\tau}_{L+1} \text{ be the hitting time of } \hat{T}_{L+1} \text{ and take } a \text{ sufficiently large. Using the auxiliary chain } \bar{\alpha}_t \text{ it is trivial to show that the truncated model exhibits cutoff at mean times. From equations (6.22) and (6.23), } \mathbb{E}[\hat{\tau}_{L+1}] = \frac{a}{1 - \rho}, \sigma[\hat{\tau}_{L+1}] = O(\sqrt{a}). \text{ Equations (6.24) and (6.25) immediately give the convergence in probability of Definition 2.7. Cutoff in total variation can be proved using Theorems 3.3 and 3.4 with } A_{\bar{n},0} = \hat{T}_{L+1} \text{ for all } \bar{n} \geq 1. \text{ We must bear in mind that the limits for } n \to \infty \text{ are now limits for } a \to \infty. \text{ All the hypotheses save those regarding } \Delta_n \text{ can be trivially checked by means of the auxiliary chain } \bar{\alpha}_t, \text{ thus we skip.} \end{align*} \]
this point. The following coupling argument show that the thermalisation happens on a time scale which is independent of $a$, so one can choose $\Delta_n = \Theta(L)$.

Let $(h_t, k_t), (i_t, j_t)$ be two identical copies of the truncated chain $(h_t, i_t)$ and let $\hat{\gamma}$ be their coalescence time. The first copy is started in state $(h_0, k_0)$ such that $h_0 + k_0 = L + 1$, whereas $\hat{t}_{L+1} \equiv (i_0, j_0) \sim Q_{n,1}$. The chains are coupled by using the same realisation of the arrival stream $(5.1)$. Since a customer is deleted if more than $L$ time units late, such a coupling guarantees that $k_s = j_s$ for all $s \geq L + 1$. The coupling we have chosen is such that the number of arrivals $m(t+1; t+1)$ is the same for both copies, so formula $(5.2)$ ensures that $|h_t - i_t|$ can only decrease.

Let us suppose that $h_{L+1} \geq i_{L+1}$ for the time being. After the partial coalescence has happened on the second component, in the most unlikely situation the two copies reach full coalescence the first time $s \geq L + 1$ such that $h_s = 0$. If, conversely, $h_{L+1} < i_{L+1}$, then the coalescence time is stochastically dominated by the first time $s \geq L + 1$ such that $i_s = 0$. Let

$$
\tau^h_0 = \min\{t \geq L + 1, h_s = 0\},
\tau^i_0 = \min\{t \geq L + 1, i_s = 0\},
$$

then

$$
P(\hat{\gamma} \geq t) \leq P\left(\max\{\tau^h_0, \tau^i_0\} \geq t\right),
\leq P\left(\tau^h_0 \geq t\right) + P\left(\tau^i_0 \geq t\right).
$$

Now we can use the Strong Markov Property and the auxiliary chain $\delta_t$ to see that the expectation of both $\tau^h_0$ and $\tau^i_0$ are less than or equal to $\frac{L+1}{\rho}$. Take for example $\tau^h_0$ and suppose that at time $L + 1$ it is in state $\eta \in \hat{T}_{L+1}$. The Strong Markov Property yields

$$
P\left(\tau^h_0 \geq t\right) = P(\tau_0 \geq t + L + 1),
$$

where $\tau_0 = \min\{t \geq 0, h_t = 0, h_0 = \eta\}$. Since $E[\tau_0] = \frac{\eta}{1 - \rho}$, Markov’s inequality gives

$$
P\left(\tau^h_0 \geq t\right) \leq \frac{\frac{L+1}{\rho}}{\frac{L+1}{\rho} + \frac{L}{1 - \rho}}.
$$

Choosing $\Delta_n = \Theta(L)$, we set $\delta_n = \Theta(\sqrt{a})$ and then all the hypotheses of Theorem 3.4 are satisfied. So the truncated model exhibits cutoff with cutoff time $\frac{a}{1 - \rho}$ and window order $\Theta(\sqrt{a})$.

Remark 6.4. In the computations above we have made no use at all of the stationary measure save the information about its support. A detailed knowledge of the stationary measure $Q_{n,1}$ of the truncated chain could be derived from the truncated generating function $(5.49)$; for a given $L$ this requires only a finite number of computations. However, this machinery is not actually needed, and the possibility to skip all those computations reveals a very strong point of the method illustrated in Chapters 3 and 4.

### 6.5 Extending the Proof to the General Case

In this final section we try and sketch a possible argument for the cutoff of the MC $(n_t, l_t)$, which represents the state of the EDA/D/1 system. The
idea is to recycle as much as possible the arguments we have presented in Section 6.4, possibly adjusting them with the intuition we have used in Section 6.3. We have explained in Section 6.4 that both definitions and results given in Chapters 2 and 3 for finite MCs work in the infinite case if we formally replace the limits for $n \to \infty$ with limits for $a \to \infty$, where $a$ is related to the initial measure of the system through (6.22). More in general, for a fixed $b$, one can consider the initial measure

$$\mu^0(i, j) = \delta_{(i, j), (a-b, b)}.$$  

(6.26)

Then, equation (6.22) corresponds to the particular case $b = 0$ of (6.26), which we start examining.

First of all, we need to guess the structure of the sets $A_{n, b}$. By the analysis carried out in Sections 6.2 and 6.3 we have strong evidence that

$$Q_{0,1} = \Theta \left( \rho^1 q^{(l_z^3)} \right).$$

Then, equation (6.7) infers that the equilibrium probability mass of the $\alpha$-line $\{(i, j) : \ i+j = L\}$ decreases at least exponentially fast with $L$. Let us consider the triangle $T_0 = \{(i, j) : \ i+j \leq L\}$. We expect that the measure of the complementary set $T_0^c$ satisfies

$$Q_{n,1} \left( T_0^c \right) = O(q^{L+1}).$$

Taking last equation for granted, let $L = L(a) = \log_{1/b} a$. Then, $Q_{n,1} \left( T_0^c \right) = O \left( a^{-1} \right) \to 0$, that is to say, the stationary measure $Q_{n,1}$ of the EDA/D/1 queue is concentrated on the family of triangles $\{T_l\}_L$ in the sense of Definition 3.3.

Next, we discuss the possibility to apply Theorems 3.3 and 3.4, using $\{T_l\}_L$ as the family of nested subsets. As usual, the key questions to answer are two, namely, 'does the system hit the triangle $T_{l, 1}$ in a quasideterministic way for $a$ large?', and 'how long does it take to thermalise therein?'. We have seen in Section 6.4 that $\alpha_t$ may be used to study the motion of $[n_t, l_t]$ from the initial state to the bulk of $Q_{n,1}$. In Section 6.4 we were guaranteed that $\alpha_t$ can only decrease for all $t$ smaller than the hitting time of the bulk*. Now $\alpha_t$ can increase but only if $n_t = 0$, and according to the intuition we have developed so far,

$$\hat{P}(i, t) \approx 0 \quad \forall \log_{1/q} a \leq i \leq a, \ 0 \leq t \leq \frac{a - \log_{1/q} a}{1 - \rho},$$

where $\hat{P}(i, t)$ is defined by (6.3).

Let us explain better the meaning of last claim. We have highlighted the vertical drift of $l_t$ in Section 6.3. This drift is really strong, it increases with the height and is responsible for the motion of $l_t$ to be mostly confined in a small interval, see Figure 18. A comparison with $l_t$, the BDC-ised version of $l_t$, has given a strong indication that the time needed to reach the height $l$ is comparable to $\rho^{-1} q^{-l/2}$ if $l$ is large. This means that climbing the $k$-th $\alpha$-line, $\{(i, j) : \ i+j = k\}$, i.e., reaching the empty queue, is an event that happens on a time scale proportional to $q^{-l/2}$. More precisely, our guess is that the chain $l_t$ does not reach a level higher than $\log_{1/q} a$ before a time

$$t_{\alpha} \geq \frac{\log_{1/q} a}{2}.$$

* In this respect, the truncated model is like a coupon collector’s chain, Section 3.4, with a thermalisation phase internal to the state 0.
The latter is much larger with respect to \((1 - \rho)^{-1}\), the typical time required for \(\alpha_t\) to take a transition \(k \rightarrow k - 1\) conditioned to \(n_t \neq 0\). It is also much larger than the time to make \(a - \log_{1/q} a\) consecutive transitions. As a consequence the probability to see a null queue for \(0 \leq t \leq \frac{a - \log_{1/q} a}{1 - \rho}\) is negligible. Then, \((6.2)\) implies that before the system hits \(T_{\log_{1/q} a}\), the motion of \(\alpha_t\) is actually ballistic. This analysis takes care of the quasi-deterministic hitting of the triangle \(T_{\log_{1/q} a}\). More precisely, we expect \(\tau_L\), the hitting time of \(T_{\log_{1/q} a}\), to satisfy

\[
\mathbb{E} [\tau_L] = \frac{a - \log_{1/q} a}{1 - \rho},
\]

\[
\sigma [\tau_L] = O(\sqrt{a}).
\]

The other fundamental question regards the thermalisation time. The idea is that the coupling argument used in Section 6.4 works here too. For the truncated model, the key to thermalisation contribution was the impossibility for a customer to be more than \(L\) slots late, \(L\) being the dimension of the stationary measure bulk. In the present framework \(L = \log_{1/q} a\) so if we consider two copies of the bivariate chain, \((h_t, k_t)\) and \((i_t, j_t)\), such that \(h_0 + k_0 = L\) and \((i_0, j_0) \sim Q_{n,t}\), then the probability of the vertical components having not yet coalesced after a time \(L\) can be computed as follows. Let \(\gamma\) be the coalescence time. Except for contributions vanishing for a sufficiently large,

\[
P (\gamma \geq L) \leq P (\gamma \geq L | |k_0 - j_0| \leq L),
\]

\[
\leq P (\text{at least a customer late more than } L \text{ slots}),
\]

\[
\leq \sum_{k=1}^{L} P (k\text{-th customer late more than } L \text{ slots}),
\]

\[= L q^{-1}.
\]

Hence, the probability not to coalesce in a time \(\Delta_n = \Theta(L)\) goes to zero as fast as \(\frac{\log_{1/q} a}{a}\). It is then reasonable to expect that the very same argument used for the truncated model works out-of-the-box, and give a thermalisation contribution \(\Theta(\log_{1/q} a)\). All in all, we expect that the bivariate chain \((n_t, l_t)\) exhibits cutoff with cutoff time \(\frac{a}{1 - \rho}\) and window order \(\Theta(\sqrt{a})\).

Eventually let us consider the family of initial measures \((5.49)\), \(b \neq 0\). With respect to the case \(b = 0\), the new starting point clearly affects only the approach to \(T_{\log_{1/q} a}\). However, according to the discussion in Section 6.3, and in particular the great strength of the vertical drift, the contribution to be added to the mean hitting time of the bulk is negligible with respect to \(a(1 - \rho)^{-1}\). Therefore we expect that the picture just given for \(b = 0\) also applies to the general case.

**Remark 6.5.** In both the truncated and non-truncated models we have shown how to apply Theorems 3.3 and 3.4 using a family \(A_{n,\theta}\) which actually does not scale with \(\theta\). In the truncated model the family does not even scale with the initial position\(^*\). This choice is justified by the finiteness of the support of stationary measure \(Q_{n,t}\). In the non-truncated model this is no longer true and the bulk of the stationary measure has to scale with \(a\) and/or \(\theta\). The approach we have followed in the present section is to let

\(^*\) For infinite MCs the asymptotics for \(n \rightarrow \infty\) are replaced with \(a \rightarrow \infty\), where \(a\) represents the initial position, see \((6.22)\).
it scale with $a$ but not with $\theta$. This leads to a possible overestimate of the thermalisation contribution which, however, do not affect the final size of the cutoff window. Since the stationary measure $Q_{n,1}$ exponentially decays with the side of the triangles $T_L$, we could try for example to take $L$ as a function of $\theta$ only. This would be similar to the approach used in Section 3.1, and would likely give a more precise estimation of the thermalisation time. On the other hand, the very same choice of $L$ would require a much more detailed knowledge of the motion of $(n_t, l_t)$ in the quarter plane, which at the present time we have not. This remark highlights the flexibility of the methodology developed in Chapter 3.
This chapter considers a class of arrival processes that goes under the name of Pre-Scheduled Random Arrivals (PSRA). PSRA are a generalisation of Exponentially Delayed Arrivals (EDA) to other-than-exponential delays. We have introduced and studied EDA in Chapters 5 and 6. Let $\frac{1}{\lambda}$ be the expected inter-arrival time between two consecutive customers. In PSRA the actual arrival time of the $i$-th customer is defined by

$$t_i = \frac{i}{\lambda} + \xi_i, \quad i \in \mathbb{Z}, t_i \in \mathbb{R},$$

where $\xi_i$ are IID continuous random variables whose probability density and (finite) variance are $f_{\xi}(\sigma)(t)$ and $\sigma^2$, respectively. Without loss of generality it can be assumed $E(\xi_i) = 0$ since $E(\xi_i) \neq 0$ affects only the initial configuration of the system.

The idea of the process is easy. An ideal, deterministic schedule is organised in such a way that customers arrive as a constant, homogeneous stream. Then an independent random variable is added to each of the scheduled arrivals. As a result, the arrival times lose their appointed ordering and get mixed up by the random delays. The process obtained in this way has a long history, part of which we have discussed in Chapter 5. It is easy to study numerically but quite difficult to treat from a mathematical point of view although some significant progress has been recently made in [GNS11, GLNS13].

This final chapter focuses on the applications of PSRA in an applied contexts. More precisely, we are interested in PSRA as the actual arrival pattern of aircraft at a congested airport. We consider the case of the Heathrow airport and present a comparison with actual data. Then we move to the SESAR project, the European Air Traffic Control (ATC) infrastructure modernisation programme, and remark some issues that may arise in the introduction of a new ATM concept.

## 7.1 THE PSRA/D/1 QUEUEING SYSTEM

This chapter focuses on the statistical properties of the queueing system PSRA/D/1, that is, arrivals according to (7.1) and single server with deterministic service time. In particular, we are interested in the possibility to use a PSRA/D/1 in the description of congestion problems arising in Air Traffic Management (ATM). As a matter of fact, PSRA represent a very natural point process to model aircraft arrivals. The arrival stream is very different from a Poisson process, and yet it presents a distribution of the inter-arrival times very close to an exponential law*. The main contribution of this chapter is to demonstrate that a PSRA/D/1 model can give a very fitting description of the queueing process at a very congested airport. To this extent, we examine the inbound air traffic at the international airport of London Heathrow. A statistic from actual data is compared in Section 7.3 with simulations of PSRA/D/1. This comparison clearly shows that PSRA/D/1 is a good model for ATM congestion problems.

* See Section 7.2 below.
The arrival pattern (7.1) has the same properties exhibited by (5.1). In particular, when $\sigma$ is large the arrival process weakly converges to the Poisson process, see the discussion at the beginning of Section 5.1. This property also holds for a variant of the PSRA process that takes into account the possibility of customers cancellation as in [BVH01]. The variant is obtained by an independent thinning of (7.1), i.e., a process in which each arrival has an independent probability $1 - r$ to be cancelled (and the complementary probability $r$ to be a true arrival).

**Remark 7.1.** The thinning procedure described above does not reflect actual flight cancellations. Rather, it can be considered as a fictitious procedure to model empty slots in a constant stream of customers.

Figure 19 sketches the output of a thinned PSRA. The actual stream of aircraft (ATA) is the consequence of the random delays $\xi_i$ mixing-up the pre-scheduled, expected stream (ETA).

![Figure 19: The actual stream of arrival times (ATA) arise from the action of random delays and thinning (white dots) of a deterministic schedule (ETA).](image)

In ATM contexts it is natural to couple a PSRA arrival process with a deterministic service process with expected service rate $\lambda$. In this case the traffic intensity of the thinned process is clearly $\rho = r$. According to the usual ATM terminology, we refer to a time interval of length $1/\lambda$ as a slot. ATM regulations forbid two consecutive landings in a time lapse shorter than a fixed quantity that depends on the airport capacity and the weather conditions. Having in mind the problem of modelling these features, a solution is to take a regular stream of customers at rate $\lambda$ and to allow for gaps in such a stream via the thinning procedure described above. Then, it is natural to consider a deterministic service policy with constant service rate $\lambda$ because, neglecting the thinning, a larger service rate would result in a waste of resources, whereas a smaller service rate in an unstable system. In summary, we study throughout the chapter a queueing model with traffic intensity $\rho$. The arrivals to this queue are given by the thinned version of (7.1) with rate $\rho$, while the service, delivered by a unique server, has a constant, fixed duration $\lambda^{-1}$.

In a normal operations day, with fair wind and visibility conditions, the typical duration of a slot is about 90 seconds. Whenever any of the following situations arise, critical weather conditions, large volumes of traffic in the same line of flight, aircraft incidents, closed runways, or whenever the approach to the airport is switched between Instrument Flight Rules (IFR) and Visual Flight Rules (VFR), the slot duration is varied to accomplish safety regulations. In those cases a Ground Delay Program (GDP) may also be applied by the ATC. All the situations described above may be well modelled by simply varying the value of $\lambda$.

Although both the PSRA process and its thinned version are very similar to the Poisson process for large $\sigma$, they present a crucial difference with the
latter. As long as \( \sigma \) remains finite, the PSRA process is negatively autocorrelated. Fix \( T \) and let \( n_1 \) be the number of arrivals in \([t, t + T]\) and \( n_2 \) the number of arrivals in \([t + T, t + 2T]\). The covariance \( \text{Cov}(n_1, n_2) \) between the number of arrivals within the two consecutive time intervals \( n_1 \) and \( n_2 \) is given by

\[
\text{Cov}(n_1, n_2) = \mathbb{E}(n_1 n_2) - \mathbb{E}(n_1) \mathbb{E}(n_2) = - \sum_i p_i^\sigma(t, t + T) p_i^\sigma(t + T, t + 2T)
\]

where \( p_i^\sigma(t_1, t_2) \) is the probability that the \( i \)-th aircraft arrives in the interval \([t_1, t_2]\), see [GNS11] for details. A negative covariance between \( n_1 \) and \( n_2 \) means that a congested time slot is more likely to be followed or preceded by a slot with less-than-expected arrivals. Moreover, this proves that the hypothesis of independence for \( n_1 \) and \( n_2 \) is not correct unless we are in the limit \( \sigma \to \infty \). Simulations show that if we neglect the correlation (7.2) and try to describe the inbound air traffic at an airport with independent inter-arrival times, then we get a gross overestimate for the average length of the queue. When \( \rho \) is near to 1, the system is congested and that error is particularly large, see again [GNS11] and Figure 23 below.

7.2 THE PROBLEM OF AIRPORT CONGESTION

Airport congestion is a persistent phenomenon in air traffic. Air traffic congestion is significant even if the principal airports in Western and Central Europe are treated as *fully coordinated*, meaning that the number of flights scheduled there per hour (or other unit of time) is not allowed to exceed the *declared capacity* of the airports [dNO03]. In 2011, the average Arrival Sequencing and Metering Area (ASMA) additional time† at the top 30 European airports amounted to 2.9 minutes per arrival increasing by +5% with respect to the previous year. On this statistic, London Heathrow is a clear outlier having by far the highest level of additional time within the last 40 nautical miles with 8.2 minutes per arrival, followed by Frankfurt and Madrid [EUR12]. Similar situations occur in the US [BVH01]. Quoting the 2011 Performance Review Report [EUR12], “Airports are key nodes of the aviation network and airport capacity is considered to be one of the main challenges to future air traffic growth. This requires an increased focus on the integration of airports in the ATM network and the optimisation of operations at and around airports”.

Several approaches have been proposed to mitigate congestion and resolve demand-capacity imbalances. At operational level (short-term), these approaches consider the adjustment of air traffic flows to match the available capacity. So far, the most popular approach to resolve these short-term periods of congestion has proved to be the allocation of ground delays [Odo87]. The *Ground Holding Problem* considers the development of strategies for allocating ground delays to aircraft and has received considerable attention [RO94, DL03, BHM10, ADL11]. However, these ATM strategies might be sub-optimal because they do not capture the inherent unpredictability of

---

* In the U.S., scheduling limits are applied only to the New York region airports, and the Washington-Reagan and Chicago-O’Hare airports, under the High Density Rule (HDR).

† The ASMA is the airspace within a radius of 40NM around an airport. The ASMA additional time is a proxy for the average arrival runway queuing time on the inbound traffic flow, during times when the airport is congested.
arrivals at airports. In [BVH01] the authors showed that changes in the current practice for setting airport arrival rates can lead to significant benefits in terms of additional ASMA times.

In view of the current situation, it is extremely important to have a reliable tool to measure and forecast congestion in the Air Traffic System. However, in developing such a tool there are some issues to address. First of all, the stochastic models developed so far to describe air traffic congestion are not reliable. In [WFM04] it is showed that the estimated inter-arrival times at a distance of 100NM from the final destination are nearly exponential. In other words, the arrival stream can be considered Poissonian when entering the control zone. The aircraft stream is successively rearranged to meet the ATC rules and needs. It is then natural to expect that Poissonian arrivals to give a bad fit with the actual arrivals at a congested airport†. Nevertheless, Poissonian arrivals are very often considered as the actual arrival stream when studying actual scenarios [BC06, BEFK07, Dun76, MS03].

A second issue regards the validation of stochastic models. It is not easy to draw a comparison between observed and forecast congestion, simply because it is hard to retrieve the number of aircraft in queue from a database of flight traces. Indeed, it is not clear a priori which procedure should be used to extract information about the airport congestion from a data set of waypoints passage times. In particular, the problem we address is how to determine the correct fraction of time each aircraft actually spends in queue.

The present chapter proposes a possible answer to both the aforementioned problems. Regarding the former, Section 7.1 has already presented a description of the arrival pattern and it has built a mathematical model for the queue at the airport. The latter issue is attacked by computing as follows the time a single aircraft spent in queue. Given a terminal route and a database of flight records, the minimum time lapse between the entrance in the control zone and the touch-down is subtracted from the time flown by any other aircraft in that terminal route. In this way the time spent by the aircraft in one or more stacks is found. A detailed discussion of this approach is given in Section 7.3. The analysis is completed by comparing the stationary output of the mathematical queueing model defined in Section 7.1 with the distribution of the queue obtained from a data set of arrivals at London Heathrow airport. In particular, we show that the fit of the actual Heathrow data with the output of the proposed model is really excellent, incomparably better than the fit with the assumption of Poissonian arrivals. Finally, in Section 7.4 we see by numerical simulations that the output of the model depends very weakly on the kind of delays added to each arrival time. In other words, the PDF of the random variables $\xi_i$ has a rather small impact on the distribution of the observed queue. The only relevant parameter appears to be the variance of the random delays. Even uniform random variables may be then used to implement a PSRA/D/1 and obtain a reliable forecast of the traffic over a congested hub.

---

* For flights landing in any of the London airports, the control zone is a large area covering England and Wales, operated by the London Area Control Centre (LACC). See Section 7.3 for more details.
† Cf. Figure 23a below.
The London Heathrow airport has IATA* airport code LHR, and ICAO† airport code EGLL. It is served by two parallel independent runways which are used in a separate mode, i.e., one runway is used only for departures while the other only for arrivals. Aircraft are occasionally cleared to land on the departure runway to minimise delay subject to certain criteria. The approach to Heathrow is managed by the London Area Control Centre (LACC) and by the London Terminal Control Centre (LTCC), both operated by NATS‡ and located in Swanwick, Hampshire. LACC operates the control zone of England and Wales, whereas LTCC provides an area service in and around the London Terminal Manoeuvring Area (TMA). The TMA is a block of airspace covering most of the South East of England that encloses all London airports. Arrivals to the London airports are handed over from Area Control to Terminal Control, usually following Standard Terminal Arrival Routes (STARs). According to the ICAO Standard Arrival Chart, Heathrow has 28 STARs, starting from 11 entry points and ending in 4 stacks, namely, LAM, BIG, BNN, and OCK. The position of each entry point and the geometry of the STARs are sketched by Figure 20. In 2007 the airport of London Heathrow operated at an actual flow rate between 97 to 98% of its runway capacity [UK 08]. This sets \(0.97 \leq \rho \leq 0.98\) for the time being, while an operative definition of \(\rho\) based on actual data is given by (7.3) below.

**Figure 20:** The 28 STAR of the London Heathrow Airport.

In the intent of this analysis, we consider actual data ranging from July 20 to July 30, 2010. The initial database is composed of 7,140 flights. For each flight in the database, let us define the *approaching time* as the difference between the entrance time and the landing time. By entrance time we mean the first passage time to one of the entry points depicted in Figure 20, while the landing time is just the touch-down time. Starting from the database mentioned above, we classify the records by the STAR flown. For each of these classes we then compute the approaching time of each aircraft in the class.

---

* International Air Transport Association
† International Civil Aviation Organization
‡ National Air Traffic Services
Successively, for all aircraft flying a given STAR, we subtract from each approaching time the minimum value of the approaching time for the selected STAR and consider the exceeding time as spent in queue.

**Remark 7.2.** The procedure described above returns an approximation of the time spent in queue by each aircraft. Indeed, the approaching time varies between different types of aircraft, with particular respect to the phase between leaving the stack and landing (service time). This is the same as considering slots with fixed duration and modelling the runway as a server that can start the service only at integer multiples of the slot duration.

Figure 21 shows a qualitative layout of the incoming air traffic in the period July 20 to 30, 2010. For each STAR, the thickness of the arrows is an indication of the traffic volume flowing therein. Figure 21 shows also that for each of the stacks LAM, BIG, and BNN, we can identify one STAR which absorbs nearly all the inbound traffic to the stack. In particular, the majority of the flights (about 80%) approach those stacks following one of the following three STARs:

- LOGAN, TIRPO, SABER, BRASO, LAMBORNE (LAM);
- ALESO, TIGER, BIGGIN (BIG);
- NUGRA, TOBID, SOPIT, WCO, BOVINGDON (BNN).

The picture is radically different for the stack OCK because the traffic to the stack is split in 3 streams, and none of them is negligible with respect to the others. As a consequence, for each of the routes ending at OCK, the number of available records form too small a database to obtain an adequate statistic of the time spent in queue. Thus, we limit the analysis to the three STARs mentioned above. Starting from such a restricted database, we observe that the traffic intensity of the system can be considered constant throughout the periods 6:00 a.m. to 10:30 a.m. and 4:00 p.m. to 8:00 p.m. Therefore, we restrict the database to these time intervals. All in all, the final number of flights considered is 4139.

Once we have obtained an approximation of the time spent in queue by each aircraft, we can easily compute the queue length. In fact, we are assuming that the service time is nearly deterministic, this means that the length of the queue a given aircraft had to wait in is proportional to the time spent in queue by the very same aircraft. Hence, the empirical distribution of the time spent in queue is equivalent to the empirical distribution of the queue length. As a final step of our analysis, we gather the congestion data from each of the three STARs and bin the time spent in queue. The bin size is equal to the slot duration, which we can approximate by

\[
1/\lambda = \frac{3600 \text{ sec}}{\text{average numb. of landings per hour}} = 90 \text{ sec}.
\]

The final histogram is normalised to the probability distribution presented in Figure 23.

We need \( \rho \), the traffic load, to run the simulations and make a comparison with the output of PSRA/D/1. The traffic index can be approximated by

\[
\rho = \frac{\text{average numb. of landings per hour}}{\text{max numb. of landings per hour}} = \frac{40}{41} \approx 0.976. \tag{7.3}
\]

Before moving to the comparison with the simulated scenarios, a note is needed on the choice of the data set. It is rather difficult to obtain detailed
Figure 21: Qualitative layout of the incoming air traffic over the London Heathrow Airport during the time lapse under examination. The thickness of the arrows is an indication of the traffic volume in the period July 20 to 30, 2010.

data sets such as the one used in the previous analysis because they are typically not publicly available. However, the data set used here can be considered representative of the airport in stable or quasi-stationary conditions in the following sense. The period considered is reasonably long and guarantees fine weather conditions – in particular, wind not exceeding 2 Bft and no relevant precipitation phenomena; moreover, season and worldwide importance of the destination make sure the system is highly congested. These remarks guarantee that the parameters $\lambda$ and $\rho$ do not vary over the period under investigation. It is then reasonable to expect that any other data set providing equivalent stability of traffic and weather conditions will give a qualitatively similar distribution for the queue length.

In Figure 23 the empirical distribution derived from actual data is compared with the empirical distribution obtained by simulation of models of interest. Figure 22a refers to the M/D/1 model. Poissonian arrivals are de facto a standard assumption in many works and studies in the ATM field \cite{BC06, BEFK07, Dun76, MS03, WFM04} but they do not give a satisfactory prediction of the queue distribution tail whatsoever. Figure 22b and 22c refer to the PSRA with $\sigma = \frac{2\lambda}{\Lambda}$ and $\sigma = \frac{3\lambda}{\Lambda}$, respectively. Although these values are just reasonable guesses of the actual value of the standard deviation, we clearly see that the PSRA model gives a very good accordance with empirical data. A precise estimation of $\sigma$ would require to cross-check a data set containing actual landing times with a data set containing the scheduled landing times. In this way it could be possible to compute for each flight the discrepancy between actual and scheduled landing time, i.e., the realisation of the delays $\xi_i$. If the initial data set is large – larger than the one used here – an accurate statistic of $\sigma$ can be obtained.

In real-life operations the duration of a service is not fixed, and the service providers adjust the landings time-sequence on demand within the limits of the separation rules. The PSRA/D/1 model can be easily adapted to this new scenario by considering a random service time. Figure 23d shows the stationary output of a queueing model with PSRA-like arrivals and a service
114 pre-scheduled random arrivals (a) M/D/1 vs. London Heathrow. Uniform delays with $\sigma = 30/\lambda$.

(b) PSRA/D/1 vs. London Heathrow. Uniform delays with $\sigma = 30/\lambda$.

Figure 22: Empirical distribution from actual Heathrow data (red) and from simulations (yellow) of M/D/1 and PSRA/D/1 models.
(c) PSRA/D/1 vs. London Heathrow. Uniform delays with $\sigma = 20/\lambda$.

(d) PSRA/G/1 vs. London Heathrow. Uniform delays with $\sigma = 20/\lambda$, service time has a triangular PDF with mean $1/\lambda$ and mode $0.8/\lambda$.

Figure 23: Empirical distribution from actual Heathrow data (red) and from simulations (yellow) of M/D/1 and PSRA/D/1 models.
time modelled as a triangular random variable with mean $\frac{1}{\lambda}$ and mode $\frac{0.8}{\lambda}$. This is only one of the possible variants that may be considered.

Remark 7.3. All the models described above have been simulated using a single, relatively easy-to-code Python module that is presented for completeness in Listing 5, Section 7.7. The possibility of being easily coded and simulated is an important feature of PSRA.

Table 1 measures the accordance between the observed and simulated distributions using both total-variation and Hellinger distance. We note that with memoryless arrivals the distance is an order of magnitude larger than with PSRA arrivals.

<table>
<thead>
<tr>
<th>Model</th>
<th>TV-distance</th>
<th>Hellinger distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/D/1</td>
<td>0.41067</td>
<td>0.43903</td>
</tr>
<tr>
<td>PSRA/D/1*a</td>
<td>0.03938</td>
<td>0.04565</td>
</tr>
<tr>
<td>PSRA/D/1*b</td>
<td>0.07516</td>
<td>0.08133</td>
</tr>
<tr>
<td>PSRA/G/1*c</td>
<td>0.05723</td>
<td>0.05814</td>
</tr>
</tbody>
</table>

*a $\sigma = \frac{30}{\lambda}$.
*b $\sigma = \frac{20}{\lambda}$.
*c $\sigma = \frac{20}{\lambda}$ and service times modelled as triangular random variable with mean $\frac{1}{\lambda}$ and mode $\frac{0.8}{\lambda}$.

7.4 In sensitivity to the delays’ distribution

Simulations can also be used to study the robustness of the PSRA/D/1 model with respect to the choice of $f_\xi$, the probability law of the delays. Figure 25 compares delays of different type, namely,

- uniform (Fig. 24a),
- normal (Fig. 24b),
- triangular (Fig. 24c),
- and exponential (Fig. 24d),

all with zero mean and fixed standard deviation $\sigma$.

Figure 25 suggests that the standard deviation of the delay is the actual parameter of the model and that the resulting queue distribution is insensitive of the nature of the delays. For symmetrical probability laws such as uniform, triangular and normal the histograms in Figure 25 present an astonishing resemblance and differ only for small fluctuations.

Even more surprisingly, we obtain the same layout with a skew PDF like the exponential distribution. Figure 25 clearly shows that the queue distribution obtained using exponentially delayed arrivals is qualitatively the same as the uniformly delayed case. This circumstance is of extreme interest for applications, for the equilibrium distribution of the model can be well approximated in the case of exponential delays with the methods presented in Chapter 5.
7.4 Inensitivity to the delays’ distribution

Figure 24: Output of PSRA/D/1 for different delays’ PDF, $f^{(\sigma)}_\lambda$. The delays have zero mean and $\sigma = \frac{20}{\lambda}$. 
Figure 25: (cont.) Output of PSRA/D/1 for different delays’ PDF, \( t_{\xi}^{(\sigma)} \). The delays have zero mean and \( \sigma = \frac{20}{\lambda} \).
Since 2004 the European Union (EU) gained competences in ATM and decided to launch an ambitious initiative to reform the structure of European ATM. The EU Single European Sky aims at providing a uniform, high level of safety and efficiency through a globally, rather than locally, managed ATC. In contrast to the United States, the European approach to ATM is rather fragmented in the sense that there is not a full synchronisation between all stakeholders and federate resources. The European airspace is one of the busiest in the world too, with traffic peaks of over 33,000 flights per day. Together with a very high airport density this makes ATC even more complex.

In order to meet future capacity and air safety needs, a project called Single European Sky ATM Research (SESAR) was started, founded by the EU, EUROCONTROL* and the Industry. The SESAR programme – technological and operational core of the Single European Sky initiative – works to give Europe a high-performance ATC infrastructure which will bring important benefits from the point of view of safety, cost efficiency, and environmental impact. The objectives of the SESAR programme for 2020 are to save

- 8 to 14 minutes
- 300 to 500 kg of fuel
- 948 to 1575 kg of CO₂

on average per flight†.

The ‘4D’ or Business Trajectory (BT) is one of the cornerstones of the future ATM system being developed by the SESAR programme. The BT is a trajectory which the airspace user agrees to fly and the service provider agrees to facilitate. The name Business Trajectory applies to civil aviation whereas for military flights it becomes ‘Mission Trajectory’. The BT represents the key of concept to realise a ‘paradigm shift’ from aircraft-based to trajectory-based operations. This innovative approach consists of a four-dimensional trajectory (space and time) including ascent and descent paths, that will be negotiated in advance with departure and arrival airports and constantly monitored afterwards. The BT is expected to allow for much more flexibility without any loss in operation safety. Within the SESAR programme it will be possible to take care of the need for some flights and activities to be managed within defined volumes in certain situations, e.g. military operations, search and rescue, internal security, and special events [SES10]. This goal will be achieved by continuously revising and keeping up-to-date the BT during the flight, to apply all the possible clearances and constraints.

Detailed information about the aircraft position will be shared throughout the flight among all service providers on the route by means of System Wide Information Management (SWIM), another key element of SESAR. With airborne and ground systems exchanging information in a straight way through SWIM, the information-sharing process will become more direct.

According to SESAR, the adoption of 4D Trajectories and SWIM will lead to a greatly increased certainty about the positions of every airspace user in the sky at any given moment. The global, centralised knowledge of position and velocity of every aircraft, together with the possibility to realise on-route tactical interventions on single flights, will bring large improvements

* The European Organisation for the Safety of Air Navigation. For additional information, see http://www.eurocontrol.int/content/about-us.
† Source: http://www.sesarju.eu/about
in safety and flight predictability. The next section considers BTs in mixed-
traffic scenarios, where only a portion of the airspace users fly following a
4D Trajectory.

7.6 MIXED-TRAFFIC SCENARIOS

The Business Trajectory (BT), a key concept in future ATM, was presented
and discussed in Section 7.5. In particular, we have described how the ac-
complishment of SESAR’s concepts of operations will influence flight pre-
dictability. According to [SES06], BTs will attain a minimisation of punctu-
ality fluctuations in both departure and arrival, and also in flight phases
duration. A more controlled variability is expected in turn to avoid delays
and prevent knock-on effects on other flights. At a practical level, this means
that the probability for the aircraft to be late and miss the assigned slot is
negligibly small.

In what follows we consider the PSRA/D/1 system we have introduced in
Section 7.1 under a condition of mixed traffic, meaning that a fraction of the
inbound aircraft follow a 4D Trajectory and arrive in the pre-scheduled slot.
The percentage of 4D aircraft is varied from zero (current scenario) to 100%
(SESAR target). The arrival process for the remaining aircraft is taken to be a
PSRA. As we have discussed in Section 7.5, at the core of SESAR there is an
information management system called SWIM. The implementation of SWIM
requires the airspace users to get equipped with a compatible technology,
and the cost of such an equipment is likely to slow down the achievement
of SESAR’s targets. Thus, the study of mixed-traffic scenarios is of crucial
importance to analyse the resulting transience from the actual to the future
ATM system.

Regarding which service policy should apply during the transition phase,
the SESAR community has long discussed about the possibility to choose
a standard First In First Out (FIFO) policy over a more controversial Best
Equipped Best Served (BEBS) policy. So far, consensus has not yet been
reached.

Figure 26 shows a simulation of a mixed-traffic scenario under the FIFO
policy. The average queue length and time spent in queue decrease as the
number of 4D aircraft grows. As it is natural to expect, the presence of an in-
creasing fraction of on-time customers results in a reduced variability of the
system. Under a BEBS principle, early adopters of SESAR avionics receive a
“preferential service” over non-equipped aircraft. In the language of queues,
the 4D aircraft have priority over the normal PSRA customers. The simula-
tion of such a scenario is presented in Figure 27. The time spent in queue
by non-4D aircraft increases with the traffic index ρ and/or the fraction
of 4D aircraft. The reason why this happens is that the non-4D customers
are served only when a 4D aircraft is deleted. Conditioned on the aircraft
being a 4D customer, that event has probability 1 − ρ. Therefore, at an ad-
vanced stage of the introduction of the SESAR technology, especially if ρ is
large, much care is required in the choice of the service principle to avoid
potential safety issues.
Figure 26: Performance of a mixed-traffic scenario for increasing percentage of 4D-aircraft. A FIFO policy service is considered. The PSRA arrivals have delays $\xi_i$ uniformly distributed between $-10/\lambda$ and $10/\lambda$. 

(a) Average queue length.

(b) Average time spent in queue, measured in unit slots.
Figure 27: Performance of a mixed-traffic scenario for increasing percentage of 4D-aircraft. A BEBS policy service is considered. The PSRA arrivals have delays $\xi_i$ uniformly distributed between $-\frac{10}{\lambda}$ and $\frac{10}{\lambda}$. 

(a) Average queue length of non-4D aircraft.

(b) Average time spent in queue by non-4D aircraft, measured in unit slots.
This final section is devoted to listing a code that can be used to simulate the PSRA/D/1 and other single-server queueing systems. The module presented in Listing 5 defines a base class ooController which serves as the base class. It provides a set of tools to run the model and generate a statistic from the simulated data.

A set of arrival and service models are created by subclassing the base class. The subclass XoController provides a method genATA to generate the next arrival time with respect to the arrival process X, while the subclass oYController implements the genNextServTime method, specific to service policy Y.

Remark 7.4. We note that the base class needs both the genATA and the genNextServTime methods which are not implemented by the class itself. Their implementation is delegated to the subclasses.

Eventually, the desired model X/Y/1 is coded using multiple inheritance and a further subclassing. In other words, a class named XYController is declared as a subclass of both classes XoController and oYController. The newly defined class inherits the method genATA from XoController and the method genNextServTime from oYController.

Listing 5: Python code for simulations of different single-server queue models

```python
#!/usr/bin/env python

# This module provides tools to simulate various single-server queueing systems. The name of each class is composed of two letters and the name 'Controller'. The first letter indicates the type of arrivals while the second refers to the service policy. Thus, class XYController models a X/Y/1 queueing system.

from random import seed, random
from matplotlib.pyplot import show, plot, legend, xlabel, ylabel, title
from scientifice import exp, mean, deviation, normal, normpdf
from numpy import *
import math

class ooController:
    """Base class for the single-server queue. Provides methods for generating actual times of arrival and simulating the queue, as well as computing and plotting the empirical stationary distribution. All other classes are subclasses of this one."""
    def __init__(self, r=0.9, s=10.0, t=1000):
        self.rho = r
        self.sigma = s
        self.slots = t
        np.random.seed()
        self.queuehistory = np.zeros((2,2*self.slots))
        self.sawc = 0 # sawc = set_atas was called
        self.ghwc = 0 # ghwc = genHistory was called

# \* ATA stands for Actual Time of Arrival.
```
self.hswc = 0 # hswc = genHistogram was called
print self

def set_atas(self):
    """
    Build the list self.atas, which contains the actual times of arrivals. This function uses the genATA() method provided by the calling subclass.
    """
    self.atas = [self.genATA(i) if np.random.random_sample() <= self.rho else None for i in xrange(self.slots)]
tpc = int(self.slots * 0.1)
    self.atas = self.atas[tpc:-tpc]
    self.atas.sort()
    while True:
        if self.atas[0] is None or self.atas[0] < 0:
            self.atas.pop(0)
        else:
            break
    self.sawc = 1

def genHistory(self):
    """
    Build the list self.queuehistory, that contains all the times when a customer leaves or enter the system and the corresponding queue length at that moment. This function uses the genNextServTime() method provided by the calling subclass.
    """
    if self.sawc != 1:
        self.set_atas()
    service = self.atas[0]
    j = 0
    index = 0
    buflen = 0
    while j<len(self.atas):
        if service >= self.atas[j]:
            buflen += 1
            self.queuehistory[0, index] = self.atas[j]
            self.queuehistory[1, index] = buflen
            j += 1
            index += 1
        else:
            buflen -= 1
            self.queuehistory[0, index] = service
            self.queuehistory[1, index] = buflen
            index += 1
        if buflen == 0:
            buflen += 1
            service = self.atas[j] + self.genNextServTime()
            self.queuehistory[0, index] = self.atas[j]
            self.queuehistory[1, index] = buflen
            j += 1
            index += 1
        else:
            service += self.genNextServTime()
    self.ghwc = 1
def genHistogram(self):
    """
    Build the array self.hist, whose i-th component
    contains the fraction of time the system has spent
    in state i, that is with queue length equal to i.
    """
    if self.sawc != 1:
        self.set_atas()
    if self.ghwc != 1:
        self.genHistory()
    self.maxindx = 0
    for i in self.queuehistory[0,:]:
        if i != 0.0:
            self.maxindx += 1
        else:
            break
    self.hist = np.zeros(int(max(self.queuehistory[1,:])+1))
    for j in xrange(self.maxindx-2):
        indx = int(self.queuehistory[1,j+1])
        timeelaps = self.queuehistory[0,j+1] - self.queuehistory[0,j+2]
        self.hist[indx] += timeelaps
    self.hist /= sum(self.hist)
    self.hswc = 1

def main(self):
    """
    Run the simulation of the system.
    Compute the queue and build its histogram.
    """
    self.set_atas()
    self.genHistory()
    self.genHistogram()

def plotting(self):
    """
    Compute and plot the histogram of the equilibrium distribution.
    """
    if self.sawc != 1:
        self.set_atas()
    if self.ghwc != 1:
        self.genHistory()
    if self.hswc != 1:
        self.genHistogram()
    import matplotlib.pyplot as plt
    plt.figure(1)
    if self.slots <= 1000:
        plt.subplot(211, title='History')
        line, = plt.plot(self.queuehistory[0,1:self.maxindx],self.
                         queuehistory[1,1:self.maxindx],'r-')
        line.set_linestyle('steps-post-')
        line.set_linewidth(1.5)
        line.set_marker('s')
        line.set_markerfacecolor('#FFBF00')
        line.set_markersize(4.0)
        plt.axis([math.floor(min(self.queuehistory[0,1:self.maxindx]))
                   ]-0.5,math.floor(max(self.queuehistory[0,1:self.maxindx]))
                   )+0.5,math.floor(max(self.queuehistory[0,1:self.maxindx])
                   )+0.5,
### Customized Service Time Models

#### class oDController(ooController):

```
Subclass of the base controller providing
deterministic service time
```

```
def __init__(self, r=0.9, s=10.0, t=1000):
    ooController.__init__(self, r, s, t)

def genNextServTime(self):
    return 1.0
```

#### class oUController(ooController):

```
Subclass of the base controller providing uniform service
time. The mean service time is 1.
```

```
def __init__(self, r=0.9, s=10.0, t=1000, s2 = 1.0):
    ooController.__init__(self, r, s, t)
    self.sl = math.sqrt(3)*s2
    # support of a uniform distribution with standard deviation s2

def genNextServTime(self):
    return np.random.uniform(1.0-self.sl, 1.0+self.sl)
```

#### class oTController(ooController):

```
Subclass of the base controller providing symmetric-
triangular service time. The mean service time is 1.
```

```
def __init__(self, r=0.9, s=10.0, t=1000, s2=1.0):
    ooController.__init__(self, r, s, t)
    self.sl = math.sqrt(6)*s2
    # half-support of a triangular distribution with standard
deviation s2

def genNextServTime(self):
    return np.random.triangular(1.0-self.sl, 1.0, 1.0+self.sl)
```

#### class oSController(ooController):

```
Subclass of the base controller providing skew-triangular
service time. The mean service time is 1.
```

```
```
def __init__(self, r=0.9, s=10.0, t=1000, a=0.666, m=0.9):
    ooController.__init__(self, r, s, t)
    self.leftb = a  
    # left border of the support of the distribution
    self.mode = m  
    # mode of the distribution
    self.rightb = 3.0 - self.leftb - self.mode  
    # right border of the support of the distribution
    self.s2 = math.sqrt((self.leftb**2 + self.mode**2 + self.rightb**2 - self.leftb*self.mode - self.leftb*self.rightb - self.mode*self.rightb )/18.0)  
    #standard deviation of the distribution

def genNextServTime(self):
    return np.random.triangular(self.leftb, self.mode, self.rightb)

# PSRAs models

class UoController(ooController):
    """Subclass of the base controller providing PSRA arrivals. Delays are uniform with mean 0 and standard deviation s."""
    def __init__(self, r=0.9, s=10.0, t=1000, **kwargs):
        ooController.__init__(self, r, s, t)
        self.bl = math.sqrt(3)*self.sigma  
        # half-support of a uniform distribution with stand.dev. sigma
        self.tlaw = 'Uniform'

    def genATA(self, i):
        return i + np.random.uniform(-self.bl, self.bl)

class ToController(ooController):
    """Subclass of the base controller providing PSRA arrivals. Delays are triangular with mean 0 and standard deviation s."""
    def __init__(self, r=0.9, s=10.0, t=1000, **kwargs):
        ooController.__init__(self, r, s, t)
        self.bl = math.sqrt(6)*self.sigma  
        self.tlaw = 'Triangular'

    def genATA(self, i):
        return i + np.random.triangular(-self.bl,0.0,self.bl)

class EoController(ooController):
    """Subclass of the base controller providing PSRA arrivals. Delays are uniform with mean 0 and standard deviation s."""
    def __init__(self, r=0.9, s=10.0, t=1000, **kwargs):
        ooController.__init__(self, r, s, t)
        self.scale = self.sigma  
        self.tlaw = 'Exponential'

    def genATA(self, i):
        return i + np.random.exponential(self.scale) - self.scale
class NoController(ooController):
    """
    Subclass of the base controller providing PSRA arrivals. Delays are normal with mean 0 and standard
    deviation s.
    """
    def __init__(self, r=0.9, s=10.0, t=1000, **kwargs):
        ooController.__init__(self, r, s, t)
        self.scale = self.sigma
        self.tlaw = 'Normal'

    def genATA(self, i):
        return i + np.random.normal(scale=self.scale)

#############################################################
# Memoryless Arrivals
#############################################################

class MoController(ooController):
    """
    Subclass of the base controller providing Poissonian arrivals.
    """
    def __init__(self, r=0.9, s=10.0, t=1000, **kwargs):
        ooController.__init__(self, r, s, t)
        self.scale = 1.0/self.rho

    def genATA(self, i):
        return i + np.random.exponential(self.scale)

    def set_atas(self):
        self.atas = [self.genATA(0.0)]
        for i in xrange(self.slots - 1):
            self.atas.append(self.genATA(self.atas[-1]))
        tpc = int(self.slots * 0.1)
        self.atas = self.atas[tpc:-tpc]
        self.sawc = 1

#############################################################
# Usable Models
#############################################################

class UDController(UoController, oDController):
    """
    PSRA arrivals with uniform law
    Deterministic service time
    """
    def __repr__(self):
        return '<Unf-Det Controller : : rho = %.3f, sigma = %.3f>' % (self.rho, self.sigma)

class UUController(UoController, oUController):
    """
    PSRA arrivals with uniform law
    Uniform service time
    """
    def __init__(self, r=0.9, s=10.0, t=1000, s2=1.0):
        ooController.__init__(self, r, s, t)
        self.sl = math.sqrt(3)*s2
        self.bl = math.sqrt(3)*self.sigma
```python
def __repr__(self):
    return '<Unf-Unf Controller :: rho = %3f, sigma = %3f>' % (self.rho, self.sigma)

class USController(UoController, oSController):
    """
    PSRA arrivals with uniform law
    Skew-triangular service time
    """
    def __init__(self, r=0.9, s=10.0, t=1000, a=0.6666, m=0.9, **kwargs):
        oSController.__init__(self, r, s, t, a, m)
        self.bl = math.sqrt(3)*self.sigma
    def __repr__(self):
        return '<Unf-Trg Controller :: rho = %3f, sigma = %3f>' % (self.rho, self.sigma)

class TDController(ToController, oDController):
    """
    PSRA Arrivals with Triangular law
    Deterministic service
    """
    def __repr__(self):
        return '<Trg-Det Controller :: rho = %3f, sigma = %3f>' % (self.rho, self.sigma)

class EDController(EoController, oDController):
    """
    PSRA Arrivals with Exponential law
    Deterministic service
    """
    def __repr__(self):
        return '<Exp-Det Controller :: rho = %3f, sigma = %3f>' % (self.rho, self.sigma)

class NDController(NoController, oDController):
    """
    PSRA Arrivals with Normal law
    Deterministic service time
    """
    def __repr__(self):
        return '<Nrm-Det Controller :: rho = %3f, sigma = %3f>' % (self.rho, self.sigma)

# M/D/1 model

class MDController(MoController, oDController):
    """
    M/D/1 model
    """
    def __repr__(self):
        return '<M/D/1 Controller :: rho = %3f, sigma = %3f>' % (self.rho, self.sigma)
```
Looking through the cutoff window we have discovered how to develop an original methodology for proving cutoff. This new approach speaks the language of statistical physics and exploits entropy to highlight the drift of the chain, actual trigger of cutoff phenomena. Using the results presented in the previous chapters, we have proved cutoff for a series of interesting models and we have given a very strong evidence of cutoff behaviour for the EDA/D/1 queue. The latter is particularly significant for the quality of the proposed method because the stationary distribution of the chain is not completely known. The analysis was possible because the method makes little use of the equilibrium properties, focusing rather on the dynamics evolution.

Looking through the cutoff window we have found the thermalisation contribution, i.e., the time scale on which the chain manifests a diffusive behaviour. This contribution may be used to crack the window; either in the sense of opening it a bit, obtaining the desired window order, or in the sense of destroying it completely. The wreckage of the cutoff phenomenon happens when the thermalisation contribution grows so large to become the leading time scale. It swallows up the time scale of the quasi-deterministic hitting of the stationary distribution bulk, but the feature is kept nonetheless. The following dichotomy seems then to arise between diffusive and cutoff behaviours: if the chain has a sufficiently strong drift then cutoff prevails over diffusion; conversely, if the drift is weak in contrast to the thermalisation, the latter is the dominant behaviour. A formalisation of this picture, giving precise meaning to what ‘weak’ is, needs further research and possibly a revision of the phenomenological definition of cutoff. We have discussed two different definitions of the cutoff phenomenon, namely, the Diaconis’ definition, based on total-variation convergence to equilibrium, and the cutoff at mean times, based on stopping times. The former has the great disadvantage to be tied to the distance from stationarity – in the case of TV-distance this can potentially lead to a distorted view of the phenomenon. On the other hand, the latter does not care about the cutoff window, and the thermalisation effects in particular. Hence, a new concept that hybridises the two approaches above is probably advisable.

Looking through the cutoff window we also look towards the future. The study of the PSRA/D/1 family of queues is far from being mature; in particular, the correct solution of the EDA/D/1 system is not yet known. Apart from being a result of great importance, a complete characterisation of the stationary distribution of the EDA/D/1 queue would incidentally provide solid arguments to replace the heuristics we have relied upon in the discussion of cutoff.

Looking through the cutoff window we have seen the horizon, but many details still remain cut-off from our view.
APPENDICES
In this appendix we present in full details the estimates for $\mathbb{E} \left[ \zeta_1^n \right]$ and $\text{Var} \left[ \zeta_1^n \right]$ we have used to apply Theorem 3.4 to the magnetization chain, Section 3.7. Since for $\beta = 0$ the magnetization chain reduces to the Ehrenfest chain, the following estimates hold as well for the Ehrenfest Urn model presented in Section 3.5. From formulas (2.43)–(2.47),

\[
\mathbb{E} \left[ \zeta_1^n \right] = \frac{1}{2} \sum_{k=1}^{n} \mathbb{E} \left[ \zeta_{k \to k-1} \right],
\]

(A.1)

\[
\text{Var} \left[ \zeta_1^n \right] = \frac{1}{2} \sum_{k=1}^{n} \text{Var} \left[ \zeta_{k \to k-1} \right],
\]

(A.2)

where $\zeta_{k \to k-1}$ is the first time the chain visits $k-1$ after visiting $k$ and

\[
q_k = \frac{n}{2} + k \left( 1 + e^{\frac{4\beta}{n} (k-1)} \right),
\]

\[
\frac{\pi_n(j)}{\pi_n(k)} = \left( \frac{n}{2} + j \right) \frac{e^{\frac{4\beta}{n} (j^2 - k^2)}}{\left( \frac{n}{2} + k \right)}. 
\]

Let us begin by rewriting the ratio of the two binomial coefficients as

\[
\frac{\binom{n}{j+k}}{\binom{n}{j+k+1}} = \prod_{i=0}^{j-1} \frac{n - k - i}{n + k + i + 1},
\]

(A.3)

Then we note that for any of the values of the triple $(i, j, k)$ involved in the calculations

\[
0 \leq \frac{i}{n + k + 1} \leq \frac{i}{n - k} \leq 1.
\]

In what follows we make use of the next two easy lemmas.
Lemma A.1 For $x \in [0, 1]$,
\[
(1 - x) \frac{1}{1 + x} \leq e^{-2x}.
\]

Proof. We carry out the proof via term by term comparison.

\[
\frac{1}{1 + x} = 1 - x + x^2 - x^3 + \ldots,
\]

\[
(1 - x) \frac{1}{1 + x} = 1 - 2x + \sum_{n \geq 1} (2x^{2n} - 2x^{2n+1}),
\]

\[
e^{-2x} = 1 - 2x + \sum_{n \geq 1} \frac{(2x)^{2n}}{(2n)!} - \frac{(2x)^{2n+1}}{(2n + 1)!}.
\]

Then, we only need to show that \( \frac{(2x)^{2n}}{(2n)!} - \frac{(2x)^{2n+1}}{(2n + 1)!} \geq 2x^2 - 2x^{2n+1} \), or equivalently \( \frac{2^n}{2^n} - \frac{2^{n+1}}{2^n} \geq 2x^2 - 2x^{2n+1} \). The latter inequality holds for \( x \in [0, 1] \) since \( \frac{2^n}{2^n} - 2 \geq \frac{2^{n+1}}{2^n} - 2 \). \( \square \)

Lemma A.2 For \( 0 \leq y \leq x \leq 1 \),
\[
(1 - x) \frac{1}{1 + y} \leq e^{-x-y}.
\]

Proof. If \( x = y \) the result is true from Lemma A.1. If \( x \neq y \) then there exists \( 0 \leq \delta \leq 1 \) such that \( x = y + \delta \). Thus, the thesis is equivalent to

\[
(1 - y - \delta) \frac{1}{1 + y} \leq e^{-2y-\delta}.
\]

By Lemma A.1 and the easy inequality chain \( \frac{x}{1+y} \geq \delta e^{-y} \geq \delta e^{-2y} \),
\[
(1 - y) \frac{1}{1 + y} - \frac{\delta}{1 + y} \leq e^{-2y}(1 - \delta) \leq e^{-2y-\delta}.
\]

\( \square \)

We now bound \((A.3)\) by Lemma A.2 to obtain

\[
\left( \frac{n}{k} \right)^{j-k} \left( \frac{n}{k} \right)^{j-k} \leq \left( \frac{n}{k} \right)^{j-k} e^{-\sum_{i=0}^{k-1} \left[ \frac{k}{n} + \frac{k-1}{n} \right]},
\]

\[
= \left( \frac{n}{k} \right)^{j-k} e^{\frac{2(j-k)^2}{n^2 + \frac{2}{n^2}}} e^{\frac{-2(j-k)^2}{n^2 + \frac{2}{n^2}}} \leq \left( \frac{n}{k} \right)^{j-k} e^{\frac{-2(j-k)^2}{n^2 + \frac{2}{n^2}}}. \tag{A.4}
\]
Thus, for \( \frac{1}{2} \sqrt{\frac{n}{1-\beta}} \leq k \leq \frac{n}{2} - \log n \) and \( n \) sufficiently large,

\[
\sum_{j=k}^{\frac{n}{2}} \left( \frac{n}{2} + j \right) e^{\frac{4k}{n} (j^2 - k^2)} \leq \sum_{j=k}^{\frac{n}{2}} \left( \frac{n}{2} - k \right) e^{n \left( \frac{2}{n} \right) - \frac{4k}{n} (j^2 - k^2)} \leq \sum_{l=0}^{2 - k} \left( \frac{n}{2} - k \right) e^{4k l} \left( 1 + O \left( \log^{-1} n \right) \right),
\]

\[
\leq \sum_{l=0}^{\infty} \left( \frac{n}{2} - k \right) e^{4k l} \left( 1 + O \left( \log^{-1} n \right) \right),
\]

\[
= \left( \frac{n}{2} + k \right) \left( 1 + O \left( \log^{-1} n \right) \right),
\]

\[
\leq \frac{n}{2} \left( 1 - e^{4k} \right) + k \left( 1 + e^{4k} \right),
\]

\[
\leq \frac{n}{2} + k + \frac{2}{2(1-\beta)k} \left( 1 + O \left( \log^{-1} n \right) \right).
\]

For \( \frac{n}{2} - \log n \leq k \leq \frac{n}{2} \), we continue the inequality chain above from (A.5).

We get

\[
\sum_{l=0}^{n-k} \left( \frac{n}{2} - k \right) e^{4k l} \left( 1 + O \left( \log^{-1} n \right) \right),
\]

where \( c_1 \) is a suitable constant and \( c_2 = e^2 c_1 \). Therefore,

\[
\mathbb{E} \left[ c_n^1 \right] \leq \frac{n}{2(1-\beta)} \left[ \sum_{k=\frac{1}{2} \sqrt{\frac{n}{1-\beta}}}^{n-\log n} \frac{1 + e^{4k l}}{k} \right] \left( 1 + O \left( \log^{-1} n \right) \right)
\]

\[
+ \sum_{k=\frac{1}{2} \log n}^{\frac{n}{2} + k} \frac{n}{2} \left( 1 + e^{4k l} \right) \sum_{l=0}^{n-k} \left( c_2 \frac{\log n}{n} \right),
\]

\[
= \frac{1}{2(1-\beta)} n \log n + O(n).
\]

and

\[
\mathbb{E} \left[ c_n^1 - c_n^0 \right] \leq \frac{n}{2(1-\beta)} \left[ \sum_{k=\frac{1}{2} \sqrt{\frac{n}{1-\beta}}}^{\frac{n}{2} \sqrt{\frac{n}{1-\beta}}} \frac{1 + e^{4k l}}{k} \right] + O(n),
\]

\[
= (1 + \log \theta) O(n).
\]

From previous computations,

\[
\mathbb{E} \left[ c_{k-1} \right] \leq \sqrt{e} \frac{1 + e^{4k l}}{2(1-\beta) k} n \]

(A.6)

since

\[
e^{n \left( \frac{2}{n} \right) - \frac{4k}{n} (j^2 - k^2)} \leq \sqrt{e}.
\]
Then, by summation,
\[ \mathbb{E} [\hat{\zeta}_{k+1} - \hat{\zeta}_{k-1}] \leq \frac{\sqrt{\epsilon}}{2(1 - \beta)} n \log \left(1 + \frac{1}{k}\right) + O(n). \]  
(A.7)

From (A.2), using (A.4) and (A.6)-(A.7),
\[ \text{Var}[\hat{\zeta}_{k-k-1}] \leq \frac{n}{n^2 + k} \left(1 + e^{4\beta k}(k-1)\right) \times \]
\[ \times \sum_{j=k}^{\infty} \mathbb{E} [\hat{\zeta}_{j-k-1}] \left(\frac{n}{n^2} \right) \left(\frac{n}{n^2 + k} e^{4\beta k} \right)^j \]
\[ \leq \frac{c n^2}{n^2 + k} \sum_{j=0}^{\infty} j \left(\frac{n}{n^2 + k} e^{4\beta k} \right)^j \]
\[ \leq \frac{c n^2}{n^2 + k} \sum_{j=0}^{\infty} j \left(\frac{n}{n^2 + k} e^{4\beta k} \right)^j \]
\[ \leq \frac{c n^2}{n^2 + k} \left(\frac{n}{n^2 + k} e^{4\beta k} \right)^{2} \]
\[ \leq c \frac{n^3}{k^2}, \]

where \( c = c(\beta) \) is a constant which depends on \( \beta \) only. Therefore,
\[ \text{Var}[\hat{\zeta}_{n+1}] = \sum_{k=1}^{n} \text{Var}[\hat{\zeta}_{k-k-1}] \leq O(n^2). \]

Finally, let us bound \( \mathbb{E} [\hat{\zeta}_{n+1}] \) from below. From (A.1) and (A.3),
\[ \mathbb{E} [\zeta_{n+1}] \geq \sum_{j=k}^{n} \frac{n}{n^2 + k} \left(1 + e^{4\beta k}(k-1)\right) \sum_{j=k}^{k+\sqrt{n}} \left(\frac{n}{n^2 + k} e^{4\beta k} \right)^j \times \]
\[ \times \prod_{i=0}^{j-k-1} \left(1 - \frac{i}{j-2-k}\right) \left(\frac{1}{1 + \frac{i}{j-2-k}} \right) e^{4\beta k}(j^2 - k^2) \]

Then,
\[ \sum_{j=k}^{k+\sqrt{n}} \left(\frac{n}{n^2 + k} e^{4\beta k} \right)^j \]
\[ \geq \sum_{j=k}^{k+\sqrt{n}} \left(\frac{n}{n^2 + k} e^{4\beta k} \right)^j \left(\prod_{i=0}^{j-k-1} \left(1 - \frac{i}{j-2-k}\right) e^{4\beta k}(j^2 - k^2) \right) \]
\[ \geq \sum_{j=k}^{k+\sqrt{n}} \left(\frac{n}{n^2 + k} e^{4\beta k} \right)^j \left(\prod_{i=0}^{j-k-1} e^{4\beta k}(j^2 - k^2) \right) \]
\[ \geq e^{4\beta k}(j^2 - k^2) \varepsilon_1. \]  
(A.8)

where \( \varepsilon_1 \) tends to 0 exponentially fast in \( n \).
Remark A.1. The error $\varepsilon_1$ gives a negligible contribution to $\mathbb{E}[\hat{c}_n^1]$ being exponentially small, for this reason we will henceforth drop it.

The RHS in (A.8) can be rewritten as

$$k + \sum_{j=0}^{\log \log \log n} \left( \frac{n-k}{2} \right) \left( -\frac{2}{n} \right)^{j-1} \sum_{j=0}^{n-1} \left( \frac{n-k}{2} \right) \left( -\frac{2}{n} \right)^{j-1} \frac{2k}{n^2} \left( j^2 - k^2 \right)$$

$$= \sum_{l=0}^{\sqrt{n}} \left( \frac{n-k}{2} \right) \left( -2 \right)^{l-1} \sum_{j=0}^{n-1} \left( \frac{n-k}{2} \right) \left( -2 \right)^{l-1} \frac{2k}{n} \left( 1 + \varepsilon_2 \right), \quad (A.9)$$

where $\varepsilon_2 = o(n^{-1})$. Let now $\varphi = -\frac{2k}{n^2} + \frac{2}{n} - \frac{2k}{n^2}$. Then, (A.9) can be rewritten as

$$= \sum_{l=0}^{\sqrt{n}} \left( \frac{n-k}{2} \right) \left( -2 \right)^{l-1} \sum_{j=0}^{n-1} \left( \frac{n-k}{2} \right) \left( -2 \right)^{l-1} \frac{2k}{n} \left( 1 + \varepsilon_2 \right)$$

$$= \sum_{l=0}^{\sqrt{n}} \left( \frac{n-k}{2} \right) \left( -2 \right)^{l-1} \sum_{j=0}^{n-1} \left( \frac{n-k}{2} \right) \left( -2 \right)^{l-1} \frac{2k}{n} \left( 1 + \varepsilon_2 \right), \quad (A.9)$$

where $\varepsilon = O\left( \log^{-2}(\log n) \right)$ and $\varepsilon_3 = O\left( n^{-\frac{1}{2}} \log \log n \right)$.

Therefore,

$$\mathbb{E}[\hat{c}_n^1] \geq (1 - \varepsilon) \sum_{l=0}^{\sqrt{n}} \left( \frac{n-k}{2} \right) \left( 1 + e^{4\beta k} \frac{1}{n} \right) \sum_{l=0}^{\sqrt{n}} \left( \frac{n-k}{2} \right) \left( 1 + e^{4\beta k} \frac{1}{n} \right) \left( 1 - \varepsilon \right)$$

$$\geq (1 - \varepsilon) \sum_{k=1/2 \log n / \sqrt{n}}^{\sqrt{n}} \left( \frac{n-k}{2} \right) \left( 1 + e^{4\beta k} \frac{1}{n} \right) \times$$

$$\times \left[ 1 - \left( \frac{n-k}{2} \right) \left( 1 + e^{4\beta k} \frac{1}{n} \right) \right]$$

$$\geq (1 - \gamma)(1 - \varepsilon) \sum_{k=1/2 \log n / \sqrt{n}}^{\sqrt{n}} \left( \frac{n-k}{2} \right) \left( 2 + O\left( \log^{-1} n \right) \right), \quad (A.10)$$

where

$$\gamma = \left[ 1 - \frac{\log n}{\sqrt{n}(1 - \beta)} \right] \left( 1 + 2\beta \frac{\log n}{\sqrt{n}(1 - \beta)} + O\left( \log^2 n \right) \right),$$

and

$$\Gamma = \frac{n-k}{2} \left( \frac{-4\beta k}{n} + O\left( \log^{-2} n \right) \right) + k \left( 2 + 4\beta k \right) + O\left( \log^{-2} n \right) + 2.$$
Since $\gamma$ can be rewritten as

$$\gamma = \left[1 - \frac{2(1 - \beta) \log n}{\sqrt{n(1 - \beta)}} + O\left(\frac{\log^2 n}{n}\right)\right]^{1 + \frac{\sqrt{n}}{\log \log n}},$$

we see that $\gamma$ asymptotically tends to 0 for $n \to \infty$. The RHS in (A.10) now becomes

$$(1 - \gamma)(1 - \epsilon) \sum_{\frac{n}{\log n}}^{\frac{n}{\log n}} \frac{2n \left(1 + O\left(\log^{-1} n\right)\right)}{\frac{1}{2} \log n \sqrt{\frac{n}{\log n}}},$$

and we see that

$$\mathbb{E}\left[\epsilon_n \right] \geq \frac{1}{2(1 - \beta)} n \log n$$

to leading order in $n$. 
In this appendix we present in full details the estimates for $\mathbb{E} \left[ \zeta_1^n \right]$ and $\text{Var} \left[ \zeta_1^n \right]$ we have used to apply Theorem 3.4 to the random walk presented in Section 4.2. From formulas (2.43)–(2.47),

$$
\mathbb{E} \left[ \zeta_1^n \right] = \sum_{k=n+1}^{n} \mathbb{E} \left[ \zeta_{k \rightarrow k-1} \right],
$$

$$
= \sum_{k=n+1}^{n} \frac{2n}{k} \sum_{m=k}^{n} \frac{\pi_n(m)}{\pi_n(k)},
$$

where $\zeta_{k \rightarrow k-1}$ is the first time the chain visits $k-1$ after visiting $k$. By (4.6) and reversibility,

$$
\phi(k) = \sum_{m=k}^{n} \frac{\pi_n(m)}{\pi_n(k)},
$$

$$
= \sum_{m=k}^{n} \frac{k}{m} 2^{k-m},
$$

$$
\approx k 2^k \int_{-n \log 2}^{-k \log 2} \frac{e^t}{t} \, dt.
$$

Using the properties of the exponential integral,

$$
\phi(k) = \frac{1}{\log 2} - \frac{k}{n \log 2} 2^{(k-n)} + O \left( \frac{1}{k} \right),
$$

and therefore,

$$
\mathbb{E} \left[ \zeta_1^n \right] = \frac{2(1 - \varepsilon)}{\log 2} n \log n + O \left( n^{1-\varepsilon} \right).
$$

Similarly, for sufficiently large $n$,

$$
\mathbb{E} \left[ \zeta_1^n - \zeta_0^n \right] = \frac{n^{1-\varepsilon} n^{2\varepsilon-1}}{\log 2} 2n \phi(k),
$$

$$
= \frac{2n^{2\varepsilon}}{\log 2} \log \theta + O \left( n^{\varepsilon} \log \theta \right). \quad (B.1)
$$

From (B.1) we see that for $n$ sufficiently large, $\mathbb{E} \left[ \zeta_1^n - \zeta_0^n \right] = O \left( n^{2\varepsilon} \right)$. To compute $\text{Var} \left[ \zeta_1^n \right]$ we use the following formulas

$$
\text{Var} \left[ \zeta_1^n \right] = \sum_{k=n+1}^{n} \text{Var} \left[ \zeta_{k \rightarrow k-1} \right],
$$

$$
\text{Var} \left[ \zeta_{k \rightarrow k-1} \right] = \frac{2n}{k} \sum_{m=k}^{n} \left( 2 \mathbb{E} \left[ \zeta_{m \rightarrow k-1} \right] - \mathbb{E} \left[ \zeta_{k \rightarrow k-1} \right] \right) \frac{\pi_n(m)}{\pi_n(k)} \mathbb{E} \left[ \zeta_{k \rightarrow k-1} \right].
$$
Then we estimate the sum from below by its first term, i.e.,

\[
\text{Var} [\zeta_{k \to k-1}] \geq \left( \frac{2n}{k} - 1 \right) \mathbb{E} [\zeta_{k \to k-1}] = \left( \frac{2n}{k} - 1 \right) \frac{2n}{k} \phi(k),
\]

(\text{B.2})

and from above as

\[
\text{Var}[\zeta_{k \to k-1}] \leq \frac{4n}{k} \sum_{m=k}^{n} \mathbb{E} [\zeta_{m \to k-1}] \frac{\pi_n(m)}{\pi_n(k)},
\]

\[
\leq \frac{c n^2}{k} \sum_{m=k}^{n} \log \left( \frac{m}{k} \right) \frac{\pi_n(m)}{\pi_n(k)},
\]

\[
= c n^2 \sum_{m=k}^{n} \frac{2^{k-m}}{m} \log \left( \frac{m}{k} \right),
\]

\[
= c n^2 \sum_{j=0}^{n-k} \frac{2^{-j}}{k+j} \log \left( 1 + \frac{j}{k} \right),
\]

\[
\leq \frac{c n^2}{k^2} \sum_{j=0}^{\infty} j 2^{-j}.
\]

(\text{B.3})

From (\text{B.2}) and (\text{B.3}) we see that \text{Var}[\zeta_{k \to k-1}] = O \left( \frac{n^2}{k^2} \right). Therefore,

\[
\text{Var}[\zeta^1_n] = O \left( n^{2-\epsilon} \right).
\]


[SES06] SESAR. Air Transport Framework. The Performance Target – D2. [https://www.eurocontrol.int/esar/gallery/content/public/docs/DLM-0667-001-02-00a.pdf](https://www.eurocontrol.int/esar/gallery/content/public/docs/DLM-0667-001-02-00a.pdf), 2006.


ABOUT THE AUTHOR

Carlo Lancia was born in July 30, 1984 in Italy. He completed secondary school in 2002 at Liceo Scientifico Statale “F. Severi”, Frosinone.

In the autumn of 2002 he began his studies in Mathematical Engineering, at the University of Rome “Tor Vergata”, from which he graduated cum laude with a bachelor’s degree in 2006. In 2005 Carlo participated to the Erasmus programme, studying for six months in Ghent, Belgium. In the spring of 2009 he graduated cum laude from the University of Rome “Tor Vergata” with a master’s degree in Mathematical Engineering.

In November 2009 Carlo enrolled in the PhD school of Mathematics at the University of Rome “Tor Vergata”. Under the supervision of Prof. Benedetto Scoppola, he worked on the cutoff phenomenon for finite Markov chains. In the spring of 2011 he presented some preliminary results at EURANDOM, Eindhoven. In that occasion he met Francesca R. Nardi, assistant professor at Technische Universiteit Eindhoven, with whom he started a fruitful collaboration. In November 2012 Carlo moved to Eindhoven to complete his PhD. The results of his research are collected in this essay. They were presented in Rome, Cortina d’Ampezzo, Eindhoven, and Dublin.

Starting December 2013, Carlo will work as a postdoctoral researcher at the University of Rome “Tor Vergata” within ProGA, a SESAR WP-E research project coordinated by the National Aerospace Laboratory of the Netherlands in partnership with Deep Blue from Italy and Onera from France.
COLOPHON

This document was typeset with \LaTeX using the typographical look-and-feel classicthesis developed by André Miede. The style was inspired by Robert Bringhurst’s seminal book on typography “The Elements of Typographic Style” [Bri08]. Hermann Zapf’s Palatino and Euler type faces are used, the “typewriter” text is typeset in Bera Mono.

classicthesis is available for both \LaTeX and \LyX at the web address http://code.google.com/p/classicthesis/

The word cloud appearing on the cover was realised using the web service worlde.net. The image on the frontispiece is an adaptation of the image Roman hotel shutters by the Flickr user blazej.wojtyla. The original image is licensed under the Creative Commons Attribution-NonCommercial 2.0 Generic License.

This dissertation was printed by Wöhrmann Print Service, Zutphen, The Netherlands.

Final Version as of October 20, 2013 (classicthesis version 4.1).

This PhD has Super Cow Powers.