On temporal coherency of probabilistic models for audio-to-score alignment

Philippe Cuvillier

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE-ET-MARIE-CURIE

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Philippe CUVILLIER

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE-ET-MARIE-CURIE

Sujet de la thèse :

On temporal coherency of probabilistic models for audio-to-score alignment

à soutenir le 15 décembre 2016

devant le jury composé de :

Olivier Cappé
Arshia Cont
Yann Guédon
Sophie Mercier
Christopher Raphael

Philippe Cuvillier, *On temporal coherency of probabilistic models for audio-to-score alignment,* © December 2016

supervisor:
Arshia Cont

ABSTRACT

This thesis deals with automatic alignment of audio recordings with corresponding music scores. We study algorithmic solutions for this problem in the framework of probabilistic models which represent hidden evolution on the music score as stochastic process. We begin this work by investigating theoretical foundations of the design of such models. To do so, we undertake an axiomatic approach which is based on an application peculiarity: music scores provide nominal duration for each event, which is a hint for the actual and unknown duration. Thus, modeling this specific temporal structure through stochastic processes is our main problematic. We define temporal coherency as compliance with such prior information and refine this abstract notion by stating two criteria of coherency. Focusing on hidden semi-Markov models, we demonstrate that coherency is guaranteed by specific mathematical conditions on the probabilistic design and that fulfilling these prescriptions significantly improves precision of alignment algorithms. Such conditions are derived by combining two fields of mathematics, Lévy processes and total positivity of order 2. This is why the second part of this work is a theoretical investigation which extends existing results in the related literature.

RÉSUMÉ

Cette thèse porte sur l'alignement automatique d'un enregistrement audio avec la partition de musique correspondante. Nous adoptons une approche probabiliste et proposons une démarche théorique pour la modélisation algorithmique de ce problème d'alignement automatique. La question est de modéliser l'évolution temporelle des événements par des processus stochastiques. Notre démarche part d'une spécificité de l'alignement musical : une partition attribue à chaque événement une durée nominale, qui est une information *a priori* sur la durée probable d'occurrence de l'événement. La problématique qui nous occupe est celle de la modélisation probabiliste de cette information de durée. Nous définissons la notion de cohérence temporelle à travers plusieurs critères de cohérence que devrait respecter tout algorithme d'alignement musical. Ensuite, nous menons une démarche axiomatique autour du cas des modèles de semi-Markov cachés. Nous démontrons que ces critères sont respectés lorsque des conditions mathématiques particulières sont vérifiées par les lois *a priori* du modèle probabiliste de la partition. Ces conditions proviennent de deux domaines mathématiques jusqu'ici étrangers à la question de l'alignement : les processus de Lévy et la totale positivité d'ordre deux. De nouveaux résultats théoriques sont démontrés sur l'interrelation entre ces deux notions. En outre, les bienfaits pratiques de ces résultats théoriques sont démontrés expérimentalement sur des algorithmes d'alignement en temps réel.

TABLE OF CONTENTS

LIST OF SYMBOLS AND ACRONYMS

DTW	Dynamic Time Warping ... 27
MIR	Music Information Retrieval ... 24
HMM	Hidden Markov Model ... 34
HSMM	Hidden Semi-Markov Model ... 37
SSM	State-Space Model ... 28
i.d.d.	independent and identically distributed ... 57
$\gg$	dominated by ... 182
pmf	probability mass function ... 183
pdf	probability density function ... 183
cdf	cumulative distribution function ... 182
$*$	convolution product ... 183
$\star_{\mathbf{inf}}$	inf-convolution ... 63
$\mathcal{F}$	Legendre-Fenchel transform ... 66
supp	support ... 183
ess supp	essential support ... 183
$\overline{\text{mode}}$	largest mode ... 228
$\underline{\text{mode}}$	smallest mode ... 228
inf. div.	infinitely divisible ... 193
càdlàg	continuous from the right and limited from the left ... 187
DRHR	Decreasing Reverse Hazard Rate ... 207
IHR	Increasing Hazard Rate ... 207
LCAV	log-concave ... 206
MSU	Multiplicative Strongly Unimodal ... 233
M-IHR	Multiplicative Increasing Hazard Rate ... 233
M-DRHR	Multiplicative Decreasing Reverse Hazard Rate ... 233
hr	hazard rate order ... 210
lr	likelihood ratio order ... 210
rh	reverse hazard rate order ... 210
st	basic stochastic order ... 210
$\uparrow hr$	non-decreasing in hr order ... 210
$\uparrow lr$	non-decreasing in lr order ... 211

$\uparrow rh$	non-decreasing in rh order	211
$\uparrow st$	non-decreasing in st order	210
$\mathbf{TP}_2$	totally positive of order two	201
ULC	ultra-log-concave	163
$\mathbf{ULC}_k$	ultra-log-concave of order k	168
ULVX	ultra-log-convex	163
$cp(\lambda, F)$	compound Poisson distribution	195
$\mathcal{L}$	Laplace transform	237
Γ_λ	Poisson mixture	238
B_a	binomial thinning	157
$\mathcal{E}_\lambda$ **or** E_λ	exponential tilting	231

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Many signals are structured as time-contiguous *events* which generate specific observations, like sound, speech or text. Several areas of engineering aim at recognizing the sequence of events that generates the observed signal. In music, basic events may be notes, which are pitched sounds, and silences. The task of audio-to-score alignment consists of synchronizing an audio recording of some music piece with its corresponding music score. The original applicative motivation of this thesis is to improve the online alignment algorithm of a real-world system called `Antescofo`.

To recognize the sequence of events that generates an observed signal, probabilistic models are relevant when the event cannot be identified with certainty, but statistical relationships with observation are known. In particular, discrete State-Space Models assume that observation is stationary on time-intervals and identify each one as occupying a hidden state. This thesis deals with probabilistic models and focuses on one issue: modeling the hidden temporal evolution of position along a music score using stochastic processes. This issue is addressed by an axiomatic approach for score alignment algorithms and their design. We assert such algorithms should have *temporal coherency*, defined as compliance with available prior information — the music score in our applicative context. Conceiving an estimation algorithm firstly consists in choosing a specific kind of probabilistic models, and then tuning its free parameters appropriately. This leads to the first question that motivates this thesis.

Question 1. *How to design a probabilistic model that is temporally coherent with our prior information?*

Defining a probabilistic model is only half-way through the design of an inference algorithm. The other half is the way estimation of hidden quantities is performed out of the model. This is why an additional question also motivates this work.

Question 2. *Which estimation methods lead to a coherent algorithm?*

These questions are contextualized in section 2.5 within our main algorithmic framework: Hidden semi-Markov models (HSMM). To address such questions, we introduce two original criteria that define *coherent behaviors* we expect from alignment algorithms. Thus, the first part of this thesis deal with the two criteria and their applicative impacts. Both are investigated with the same methodology. First, we specify the coherency criterion after having explained its applicative motivations. Second, we formalize its statement for the given algorithmic context. Third, we investigate the mathematical conditions on the design of algorithm that fulfill this criterion. If characterizing all coherent choices is not possible, we look for necessary or sufficient conditions on parts of the criterion. Since behavior of probabilistic models also depends on the estimation method, separate investigations are done for the two methods we consider: Forward and Viterbi estimators.

The original inspiration for this work is the improvement of real-time algorithms to audio-to-score alignment. However, we do not cover all the different algorithmic aspects of automatic music alignment, and we do not compare performances to all existing state-of-the-art systems. To figure out such algorithmic improvements, we step away from most existing approaches in the literature of score alignment. We remark that current research efforts are mainly divided in two groups, according to two algorithmic aspects. The first group attempts to improve observation models, in other words the "ears" of algorithms. The idea is to get audio descriptors that *instantaneously* discriminate at best each music event. Such signal processing issue is not exactly related to our main problematic since we mainly care about the *temporal evolution* of a signal rather than its instantaneous description. Second group attempts to design machine learning methods to automatically optimize the alignment algorithm on real-world data. But for score alignment such approach has limitations. While learning algorithms needs a huge amount of training data with annotations, substantial datasets are lacking for this task. In addition, while training methods may easily learn parameters values, they are hardly able to learn a *structure*. We believe the peculiar temporal structure of music has to be properly modeled at the early design steps of algorithms. This is why our work focuses on probabilistic modeling instead of signal processing and machine learning techniques. Our main topic is modeling the belief structure on the temporal evolution of random and unobserved variables, that is the position along music scores.

1.1 SUMMARY OF CONTRIBUTIONS

Hereafter, we provide a summary of theoretical and applicative contributions of this work. This list of contributions is thematically sorted. The actual organization of the document is described afterwards.

First criterion: coherency with equivalent prior information

The first definition of coherency we consider is based on the case of equivalent scores: two different symbolic music scores may be musically identical and generate identical observations. The coherency criterion presented in chapter 3 basically asks that equivalent scores lead to equivalent algorithms. Our main contribution is to characterize all coherent semi-Markov chains thanks to the notion of infinite divisibility and *Lévy processes*. This result lies at the core of many subsequent topics of this thesis. Besides, it gives a strong prescription on the design of semi-Markovian duration models.

However, this characterization is only true with the Forward estimation method. At the opposite, we show that coherency is hardly achieved with Viterbi estimation: coherent duration models are too constrained to be of practical interest. Such result reveals the lack of coherency of such estimation for online alignment.

Second criterion: coherently modeling nominal durations of events

The second definition of coherency is the ability of an algorithm to properly account for prior information on durations of events that music scores contain. Alignment task consists in retrieving event durations. Even if actual durations are unknown, a music score gives a strong hint on the likely durations. Nevertheless, modeling these *nominal durations* provided by music scores is not straightforward and mostly relies on heuristics in the current state-of-the-art.

To theoretically assess coherency of algorithms, our criterion focuses on the special case of *non-discriminative observation*. Chapter 4 investigates coherency of online estimation algorithms. Explaining why the Viterbi estimator fails to be coherent in this setting is quite straightforward. On the contrary, discussing coherency of Forward estimation is much more involved. A key contribution is showing how the theory of *Total Positivity of order two* (TP_2) provides the right mathematical tools to formalize and address this problem. Several TP_2-related properties, like *reliability classes*, guarantees the existence of bounds on nominal durations such that chains are coherent. Such properties appear as relevant prescriptions for the design of semi-Markovian duration models. Afterwards, chapter 5 shows that Viterbi inference may be coherent for *offline* estimation. Similarly, some mathematical conditions guarantees the existence of bounds on nominal durations such that chains are coherent.

In addition, we show in the two cases that using Lévy processes — as suggested by previous criterion — significantly simplifies conditions of coherency for this new criterion even if this is not a necessary condition. Notably, numerically computing bounds on coherent nominal durations becomes tractable. This fact confirms that the two criteria and their respective prescriptions successfully combine together.

Relationship between discrete and continuous models for alignment

Another contribution of this thesis is a conceptual unification of two mainstream probabilistic approaches for audio-to-score alignment. First approach are discrete models of position like HMM, HSMM and related. Alternative approach is to model continuous position with Gaussian random walks. The two models are described in section 2.2 and their unification is obtained in section 3.3.

Whereas ordinary semi-Markov chains are inherently discrete thus non-continuous, we prove that coherent chains correspond to spatial discretizations of a continuous position model. But this kind of continuous process differs from random walks. We theoretically confront coherent semi-Markov chains (discrete position) and random walks (continuous position) through their respective probabilistic hypotheses. Such hypotheses turn out to be *mirrored*: one would recover a model by taking hypotheses of the other model and exchanging the two timelines, score position l and physical time t.

Afterwards, another aspect of unification is obtained in section 3.3.3 between discrete-time and continuous-time models. Indeed, mirroring our coherency criterion motivates another property: *compliance* between discrete and continuous time. We show that coherent semi-Markov chains and compliant random walks are perfectly mirrored: taking one model and exchanging score position l with physical time t would give exactly

the other model. Such a conceptual unification between discrete position, continuous position, discrete time, continuous time, is original in score alignment literature.

Study on Total Positivity of order two for Lévy processes

Combining prescriptions obtained from the two criteria prompt us to look for sub-classes of Lévy processes that exhibit specific TP_2 properties. This is why chapter 7 steps away from alignment and investigates the field of probability theory. A contribution of this thesis is to extend related results in such literature. Relationships between reliability classes and stochastic orders are established for Lévy processes and their first-passage-time processes. As proofs are built on approximations by Markov chains, some results are obtained for this more general class of process. In addition, several reliability classes are characterized for Lévy processes, and partial conditions are obtained for other classes. Finally, these general results are applied to a selection of common Lévy processes, leading to original results on some special mathematical functions.

Total positivity of order two: basics and applications

Since most of our results rely on TP_2 theory, gathering the standard tools of this outside literature has been necessary, and is done in appendix B. In order to rigorously apply these tools, several minor contributions on TP_2 fundamentals have been required. For instance, several proofs of standard results are provided when we could not find proofs with minimal assumptions. In addition, less standard distributional reliability classes are explicitly defined and characterized. Moreover, Poisson mixtures and its numerous preservation properties are systematically reviewed and a few original results are derived. The motivation for this review comes from some of our proof outlines, where this tool is used to construct approximations of any continuous processes by discrete processes.

In chapter 8, we provide a theoretical contribution of a different kind. There, we solve a few problems of the combinatorics literature by applying TP_2 theory. Conceptually, many TP_2-related results are about preservation of peculiar properties by peculiar linear operators. Translating adequately the problems in combinatorics within this framework leads to straightforward proofs and extensions of existing results. This contribution may be considered as a byproduct; such problems are *not* directly related to the original motivation of this work, but the mathematical background we have introduced so far provides all needed ingredients.

Applicative impact on semi-Markov-based alignment algorithms

The study on Total Positivity of Lévy processes provides several prescriptions on the design of alignment algorithms. Chapter 6 makes a survey of many durations models found in score alignment literature. It reveals that our results theoretically justify some common engineering heuristics, disqualify some other ones and brings about new engineering choices. In addition, comparative experiments of audio-to-score alignments are undertaken. They demonstrate the practical benefits in real-world applications of these theoretically grounded prescriptions.

The original applicative motivation of this thesis is to enhance the alignment algorithm of a real-world software called `Antescofo`. This goal has been reached as obtained results have been successfully implemented in `Antescofo` system which is used by hundreds of composers and artists worldwide. This work has led to significant improvement in practices of score following application in real-world concerts.

1.2 ORGANIZATION

Two parts can be distinguished in this manuscript. First part starts with chapter 2 and studies probabilistic models for score alignment through our two coherency criteria. Chapter 3 deals with first criterion and presents the discrete-continuous unification as a corollary. Chapter 4 and 5 are devoted to the second criterion. This part ends with chapter 6 which recapitulates the theoretically grounded prescriptions we have obtained, and empirically demonstrate their practical benefits with few experiments.

Second part is composed of chapters 7 and 8. This part contains the mathematical developments that have been motivated by score alignment, but these two chapters may be read independently of the previous ones since they only deals with probability theory and no longer with score alignment. However, their conclusions have been applied to obtain the design prescriptions for alignment algorithms.

Appendices contain a global presentation of all mathematical notions employed throughout the manuscript. Many of these notions are motivated one after the other in the main chapters, even though each chapter deals with few notions. Gathering all definitions and properties together emphasise their relations among themselves. Appendices essentially contain standard or at least already known material. But appendix B on Total Positivity contains a few original results, proofs and non-standard definitions. We have left these minor contributions next to known results on purpose, since focusing on them and their proofs would deviate attention from the global presentation.

2

BACKGROUND & MOTIVATION

This chapter introduces the general context and approaches to audio-to-score alignment, and attempts to make sense of common design difficulties. To do so, required background of alignment is introduced on section 2.1. Music-to-score alignment task is described and its underlying hypotheses are specified. Among many algorithmic approaches of alignment, this thesis exclusively focuses on probabilistic models. The different challenges facing the design of alignment algorithms are explained through an overall description of generative probabilistic models in section 2.2.

Then, we focus on the only challenge covered by this thesis: modeling the hidden evolution along scores with a stochastic process. In sections 2.3 and 2.4, typical choices of processes are reviewed parallel to their application in score alignment literature. While this thesis mainly focuses on discrete processes, continuous processes are also described for they have inspired part of the idea behind our scientific approach. Questions raised in introduction are refined for this specific framework in section 2.5.

2.1 APPLICATIVE MOTIVATIONS: MUSIC-TO-SCORE ALIGNMENT

Information Retrieval is a large domain whose goal is to develop algorithms to automatically understand the symbolic content of signals. Many natural phenomena exhibit a latent temporal structure and may be observed through a temporal signal, like music, speech or video. They are often structured as time-contiguous *events* which generate specific observations.

This work focuses on the symbolic content of audio signals. Many audio records are performances of some music score. Such score indicates all symbolic events like be notes (*i.e.*, pitched sounds), chords and silences that are allegedly played in the record. Audio-to-score alignment consists in temporally localizing the occurrence of each score event in the signal. As music scores is an *ad hoc* knowledge required for this task, such information is precisely defined in section 2.1.2. Then, a brief survey of existing algorithms in the literature is given in section 2.1.3.

2.1.1 TASK DESCRIPTION

Audio-to-score alignment consists of synchronizing an audio recording of a music piece with its music score. For this task, the algorithm *knows* a symbolic representation of the music score in advance. The recording is an audio-numerical signal which is assumed to be a performance of this particular score. “Aligning” is estimating *onset times* at which

each music event written on the score occurs in the signal. This task is illustrated by figure 2.1.

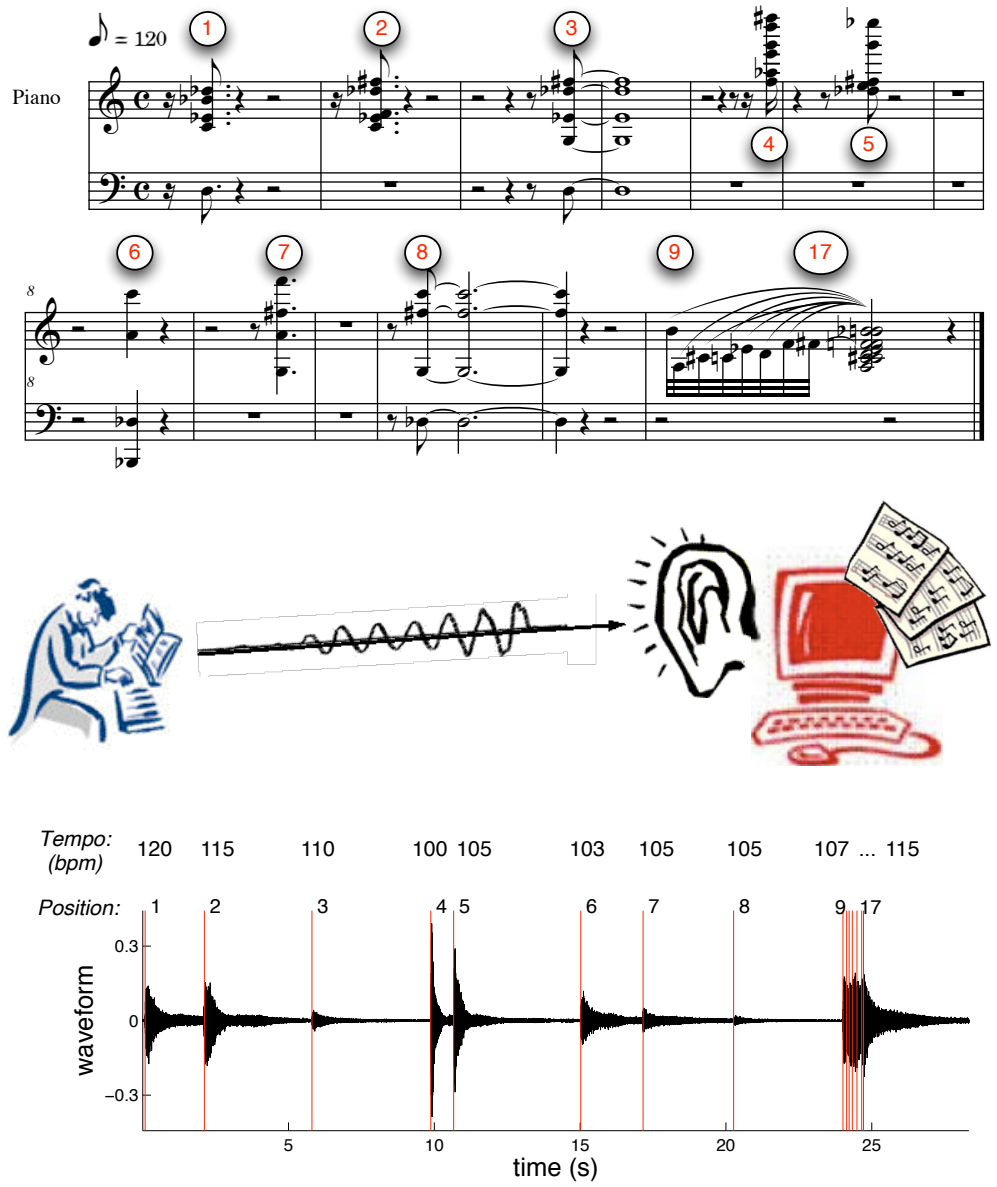


Figure 2.1: General scheme of audio-to-score alignment (courtesy of Cont (2010)). Red ticks on the audio signal waveforms indicate onset times for occurrences of symbolic events written in the music score.

From recognition to alignment

Many areas of engineering aim at recognizing the sequence of events that generates the observed signal. In *Music Information Retrieval* (MIR) domain where score alignment

belongs[1], popular *recognition* tasks are automatic transcription, chord identification, audio segmentation, change-point detection.

For our task, recognition boils down to *alignment* since the ordering between events is known. This prior knowledge on the signal is used as a *constraint* on likely occurrences[2]. Therefore, audio-to-score alignment aims conceptually at synchronizing an acoustic time-series (the audio signal) with a symbolic time-series (the music score). This definition is formalized by two assumptions.

Hypothesis 1. The symbolic events to recognize are totally ordered.

Hypothesis 2. A performance of the music score is strictly linear: it begins on the first event, then subsequent events occur with respect to its order.

Such hypotheses may be too restrictive for specific musical situations. First, music scores may have predetermined repeats or skips from and to specific event positions. A few approaches address this issue (Pardo and Birmingham, 2005; Montecchio and Cont, 2011). Second, performances might feature repeats and skips that are not expected on the score, for instance when the musician is rehearsing. Dealing with arbitrary jumps is still a very challenging extension of the base alignment task. Tackling the latter is rather rare but can be found in (Arzt and Widmer, 2010b; Nakamura et al., 2013). However, this thesis does not cover such cases and sticks to the two assumptions above.

Offline and online versions of alignments

Score alignment can be performed either offline or online.

- *Offline* algorithms make use of the whole signal and compute the alignment after its complete acquisition.
- *Online* versions align incrementally the part of the signal they have acquired, without "looking ahead" in the future.

When addressed online, audio-to-score alignment is also called *score following* (Vercoe and Puckette, 1985). Computing alignment in *real-time*, as the signal is being acquired, allows to synchronize computer actions with the human performer and to perform automatic accompaniment of live soloists. Thus, score following systems have since become the backbone of *mixed music* practices in computer music whose aim is the live association of human musicians with computers on stage. Indeed, such systems have largely widened the possibilities of mixed music and live interaction between humans musicians and electronic music.

`Antescofo` is a complete software solution for mixed music, score following, execution and specification of computer actions (Cont, 2008). Originally created at Ircam laboratory in 2007 by Arshia Cont and developed since by his team, it is now the state-of-the-art solution for each involved sub-tasks. The primary applicative motivation of

1. Score alignment is also related to others domains like automatic speech recognition. With music played by a singing voice, alignment could be done out of phonetic information and speech models. We have successfully used this approach in a side project (Gong et al., 2015).
2. Despite these hypothesis, alignment systems should be able to deal with local human errors or deviations form the original music score. In practice, state-of-the-art systems do manage errors of a moderate magnitude, thanks to a robust evolution model and natural inaccuracy of audio signals.

this thesis is to enhance the real-time alignment algorithm of `Antescofo` by studying system behavior in uncertain situations and proposing new models that improve robustness whenever they occur.

2.1.2 REPRESENTING MUSIC SCORES

Symbolic music scores considered in this work undergo three major assumptions or limitations.

First, a music score must be fully ordered accordingly to Hypothesis 1. This discards *open scores* in which optional repeats, forks or jumps are left to the performer's choice.

Second, a music score assigns a *nominal duration* to each event, that is prior information on their likely duration. Alignment consists in retrieving the actual duration during the performance as it is very likely to differ from the score's duration, even though this information gives a strong hint.

Third, alignment is restricted to *pitched events* and silent events. The latter are called *rests*. Pitched events are expected to produce harmonic sounds: they are composed of a superposition of harmonic series whose fundamental frequency is one the pitches indicated by the music score. Therefore, this approach of alignment not suitable for inharmonic instruments like drums and many percussions.

Apart from pitch and duration, we drop further information that are typically available in real world scores, such as instrumental formation, expressive annotations, or loudness marks. We do not consider ornaments (like grace notes) nor complex music event (like trills or glissandi) and refer to (Cont, 2010) for their treatment. In summary, a music score specification as prior information in our context can be defined as follows. Figure 2.2 depicts a toy example of music score.

Definition 2.1. A *score* is a totally ordered sequence of time-contiguous events.
Each event bears two kinds of information:

(i) its *nominal duration* l expressed in physical time,

(ii) a list of *pitches* (fundamental frequencies of each note).

Events with empty list are called rests, and other ones are called chords.

Event duration

We have assumed nominal durations l represents physical duration. However on real-world music sheets and digital music scores such as MusicXML files, each music note is written with a *relative duration* which is a rational multiple of some virtual duration called the *beat*. By convention, one beat is the base duration of a quarter note 𝅘𝅥. Then, a half note 𝅗𝅥 has relative duration 2, an eighth note 𝅘𝅥𝅮 has duration 1/2, and so for on. Rest events work similarly but with different symbols: 𝄽 = 1, 𝄼 = 2, 𝄾 = 1/2.

Converting relative duration (in beats) to physical duration (in seconds) requires knowledge of *tempo*. This quantity roughly measures the expected speed of the performance: doubling the tempo should make the performance twice faster. So we assume a *score tempo* is given by the score. For instance, the annotation 𝅘𝅥 = 60 BPM means that quarter note are expected to last 1 second.

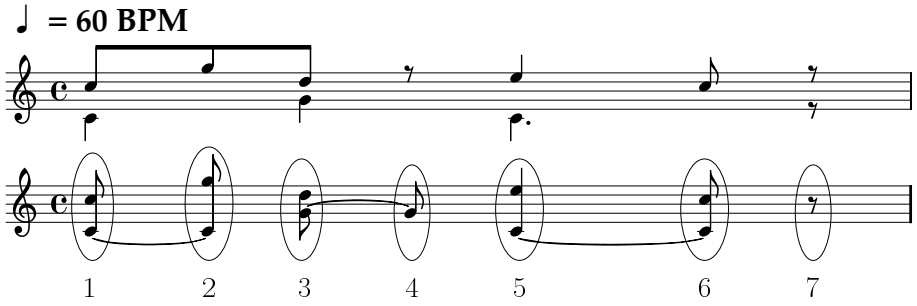


Figure 2.2: Example of chordification of a 2-voices polyphonic score (above). Nominal duration is 1 second for event 5 and 0.5 second for all other events. Several notes of the obtained chord sequence (below) are slurred between events, indicating that the original note has been divided into several chords.

Chordification

Alignment consists of specifying the event in the music score that is being played in the audio. This presumes that one and only on event is occurring at any time. However, music scores may be polyphonic, this means they feature several sequence of music events (called *voices*) that are performed in parallel.

Like almost all score alignment systems, we collapse polyphony into a single succession of chords and rests. A chord aggregates all simultaneously played pitches. A new chord is created whenever a new note begins in any voice. This *chordification* operation is illustrated in Figure 2.2. It is a simplifying assumption at is assumes perfect synchronization[3] between parallel voices which might not been achieved on interpretative purpose or involuntarily. In addition, it ignores the acoustic phenomena that happen during *unisons* when different instruments play the same pitch at the same time.

2.1.3 STATE-OF-THE-ART

Historically, the first score alignment systems have been simultaneously proposed by Vercoe (1984) and Dannenberg (1984). A survey of early research can be found in (Orio et al., 2003). Others interest surveys are (Joder, 2011) and (Arzt, 2008). We distinguish two main groups of algorithmic approaches for score alignment.

First group includes algorithms based on a cost function whose optimization can be solved efficiently. Most approaches are based on variants of a standard technique for aligning time series called *Dynamic Time Warping* (DTW). Although DTW is primarily suitable for offline alignments (Orio and Schwarz, 2001; Müller et al., 2006), online versions have been designed (Dixon, 2005; Arzt et al., 2008).

Second group are algorithms based on probabilistic models. Such approaches have several major benefits. First, they provide a probabilistic interpretation for the cost function to be optimised. Second, they have high applicative flexibility. From the same model, cost functions for both online and offline settings can be seamlessly derived, together with adapted training methods to learn the optimal parameters. The first

3. Handling asynchrony is still an ongoing challenge of score alignment (Devaney and Ellis, 2009).

probabilistic approach for score alignment is due to Grubb and Dannenberg (1994). We further distinguish two main categories [4] of probabilistic models.

Discrete state space: hidden state S is a discrete variable, like the sequence of events is. Hidden Markov Model is the prevalent approach for many recognition tasks (Raphael, 1999; Cano et al., 1999; Orio and Déchelle, 2001). Alternatives are based on more sophisticated graphical models. For instance, Hidden semi-Markov models have been successfully applied to online alignment (Cont, 2010). Aside from such *generative* models, only a few *discriminative* approaches have been proposed like Conditional Random Fields (Joder et al., 2010a; Joder, 2011).

Continuous state space: hidden state S is a continuous variable that represents position on the music score. In the score alignment literature, all approaches of this kind are based on a Gaussian random walk modelling (Montecchio and Orio, 2009; Montecchio and Cont, 2011; Otsuka et al., 2011; Duan and Pardo, 2011b; Korzeniowski et al., 2013). As exact inference is no longer tractable on continuous spaces, approximation methods like particular filtering must be employed.

This thesis exclusively focuses on generative probabilistic models as introduced in next section [5]. Among other modeling issues, the problematic about the choice of state space is raised in section 2.2.3 and the two categories introduced above are further described. Discrete state-space models are detailed in section 2.3 and continuous ones in section 2.4.

2.2 SCIENTIFIC BACKGROUND: GENERATIVE PROBABILISTIC MODELS

The application context of music-to-score alignment deals inherently with acoustic time-series as input (the audio signal), and symbolic time-series as prior (the music score). The goal is to infer the correspondence between the two incrementally (online version) or as a whole (offline version). In such applications, it is common practice to consider generative probabilistic models that describe the dynamics of the input signal as if it was generated by a state space representing the prior model. This section describes preliminaries on generative probabilistic models as used throughout this document.

2.2.1 STATE-SPACE MODELS AND BAYESIAN INFERENCE

State-Space Model (SSM) is a class of generative models thats deal with a time-series as input. Input is called *observation*[6] and is denoted $\mathbf{o}$. The goal of is to retrieve at

4. In between ly exceptions like the Hybrid Graphical Model of Raphael (2006), which stands apart as it combines discrete space for position with continuous space for tempo.
5. However, the methodology we will detail could be transposed to investigate the design of any alignment cost function.
6. In practice, we assume observation is discrete and it is a periodically sampled signal: $\mathbf{o} = (o_1, o_2, \ldots)$ and $t \in \mathbb{N}^*$.

each observation time t some unknown information. For our application, the unknown is *position*, that is the number of the currently occurring event in the event sequence.

The first key idea of SSM is to consider observation samples $\mathbf{o} = (o_1, o_2, \ldots)$ as realizations of a stochastic process $(O_t)_{n\in\mathbb{N}^*}$ called *observation process*. The second key idea is to assume that observations O are generated by an underlying stochastic process $S = (S_t)_t$ whose realizations $s_1, s_2, \ldots$ are not directly observable. This is why S is called the *hidden state process*. While the simplest model consists in defining the hidden state S_t as the unknown position, S_t can contain other variables — called *latent factors* — that help predicting observations O and future state realizations $S_{t+1}, S_{t+2}, \ldots$. Thus, the way the stochastic process S is designed represents our prior belief on the temporal evolution of events.

Remark. Formally, the time t that indexes both processes can be either continuous (indexed on a subset of $\mathbb{R}_+$) or discrete (indexed on a subset of $\mathbb{N}$). So continuous-time or discrete-time processes may be considered to model S and O. In practice, both are discrete and observation is periodically acquired. Its sampling rate drives inference rate.

To sum up with, a State-Space Model is defined as two stochastic processes (O, S). Given an observation $O = \mathbf{o}$, the goal is to retrieve the underlying hidden state-sequence $S = \mathbf{s}$ that has generated the observed sequence $\mathbf{o}$. This inverse problem-like task is referred to as *inference*. Its principle is to assign some likelihood to every hypotheses for the hidden state-sequence $(S_1, \ldots, S_t)$ by taking account available information at time t. Since the observed signal is acquired sequentially and causally, such information is partial observation sample $o_1^t = (o_1, \ldots, o_t)$.

To this aim, *Bayesian filtering* consists in computing probabilities of hidden state-sequences S_1^t conditionally to the realization $\{O_1^t = o_1^t\}$ of past and current observation. Such *posterior probabilities* are defined as $\mathbb{P}(S_1^t \mid O_1^t = o_1^t)$. Their computation relies on Bayes' theorem: for any couple of random events or vectors $\mathbf{X}, \mathbf{Y}$,

$$\mathbb{P}(\mathbf{X} \mid \mathbf{Y}) = \frac{\mathbb{P}(\mathbf{Y} \mid \mathbf{X})}{\mathbb{P}(\mathbf{Y})}\,\mathbb{P}(\mathbf{X})\,.$$

Take $\mathbf{X} = S_1^t$ and $\mathbf{Y} = O_1^t$. As the quantity $\mathbb{P}(\mathbf{Y})$ is numerically independent from the hidden state $\mathbf{X}$, it can be neglected. As a result, Bayesian filtering reads

$$\underbrace{\mathbb{P}\left(S_1^t \mid O_1^t\right)}_{\text{posterior}} \propto \underbrace{\mathbb{P}\left(O_1^t \mid S_1^t\right)}_{\text{likelihood}} \underbrace{\mathbb{P}\left(S_1^t\right)}_{\text{prior}}. \tag{2.1}$$

This formula sheds light on a very important principle that conceptually rules inference: posterior probabilities are a *compromise* between the likelihood of observations and the prior on evolution of hidden states.

MARKOVIAN HYPOTHESIS. All models we describe make the Markovian hypothesis on observation. The huge majority of approaches in the literature undertakes this assumption so that inference remains tractable and the statical model easy to define. It consists in assuming that conditionally on the current state S_t, the current observation O_t is independent from any other observation O_u or state S_u for which $u \neq t$. Conceptually, the hypothesis tells that one observation sample O_t is mostly determined

by the state S_t from which it is "emitted", and no further knowledge would help explaining this sample. Formally, the *Markovian hypothesis* is described by the following decomposition of the likelihood:

$$\forall t, u \in \mathbb{N}^*, \quad \mathbb{P}(O_t, \dots, O_{t+u} \mid S_t, \dots, S_{t+u}) = \prod_{v=t}^{t+u} \mathbb{P}(O_v \mid S_v). \tag{2.2}$$

Summing up, the design of a State-Space Model entails two statistical beliefs:

- a model on observation generated by the current hidden state: the instantaneous *likelihood* $\mathbb{P}(O_t \mid S_t)$,
- a model on evolution of the hidden state: the *prior* $\mathbb{P}(S_1, \dots, S_t)$.

2.2.2 OBSERVATION MODELS FOR MUSIC ALIGNMENT

Designing an original observation model is *not* the topic of this thesis. Our idea is rather to show how a good evolution model may enhance any algorithm. Nevertheless, this section illustrates what an observation model looks like. It describes a simplistic approach built on a template-based spectral similarity measure. As our applicative purpose is to improve `Antescofo` algorithm described in (Cont, 2010), we take the described observation model from there and use it later on for experiments. This model has been introduced by Raphael (2006) and is quite common in the alignment literature.

Audio signal processing

As usually in signal processing, an observation sample o_t is a *short-time descriptor* of the signal rather than a single audio sample.

- The signal is periodically cut into small segments $W_1, W_2, \dots$ called *frames*. Discrete time index t corresponds to the frame number and estimation is performed for each new frame W_t.
- The audio descriptor O_t is computed on W_t. A good descriptor should be able to discriminate between events.

Pitch-based observation model

Music scores are mainly composed of pitched notes and chords (*i.e.*, superposition of notes). So an observation model should eb able to discriminate different pitch contents. This is achieved by a short-time frequency representation that reveals the harmonic structure which differs between pitches. Thus, the first descriptor is chosen as the energy power spectrum: O_t denotes the energy of the discrete Fourier transform of a signal frame: $O_t := \mid \mathrm{DFT}[W_t] \mid^2$.

Then, for each state j, observation probabilities $\mathbb{P}(O_t \mid S_t = j)$ measure the similarity of this signal frame with a *template spectrum* T_j associated to j. Our model uses the following formula:

$$\mathbb{P}_{\text{pitch}}(O_t = o_t \mid S_t = j) \propto \exp(-\beta\, D_{\text{KL}}(o_t \| T_j)),$$

where β is some positive parameter and $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence between two vectors:

$$D_{\mathrm{KL}}(a\|b) \stackrel{\mathrm{def}}{=} \sum_k a_k \log \frac{a_k}{b_k}.$$

The symbol $\propto$ means that a normalization factor is dropped as it is independent from state j. Such similarity-based function has actually a rigorous generative interpretation — see (Raphael, 2006) or (Joder et al., 2011, Section III.B).

TEMPLATES CONSTRUCTION. The template T_j represents the power spectrum of the "ideal" sound produced by event j, according to the pitch content of j provided by the music score. As pitched audio signals are nearly periodic, their energy spectrum is expected to be constant over time. Therefore, such descriptor matches well the hypothesis of stationarity of observation process O.

Here, templates are constructed as the heuristic originally proposed by (Raphael, 2006). For a music note of fundamental frequency f_0, spectral energy should be mostly concentrated on partials $f_0, 2f_0, 3f_0, \ldots$ Here, we build the template of a single pitch f_0 as a mixture of K Gaussian peaks plus a background noise term that is uniform among frequencies:

$$T_{f_0}(f) \propto w_b + \sum_{k=1}^{K} w_k \mathcal{N}(f; k f_0, \sigma^2_{k,f_0}),$$

where f indexes frequency bins, w_k, w_b are positive weights and $\mathcal{N}(\cdot; \mu, \sigma^2)$ denotes the Gaussian density function with mean μ. All these values are parameters of the observation model. Figure 2.3 gives an illustration.

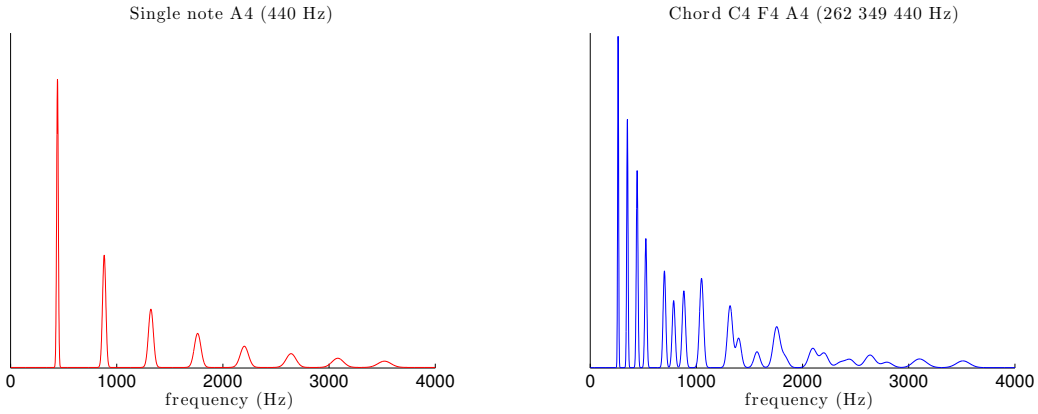


Figure 2.3: Illustration of two template spectra $T_{f_0}(f)$ for a single pitch and a chord of three pitches.

For a chord of multiples pitches $f_{j_1}, \ldots, f_{j_N}$, the template is defined as the mean of individual note templates:

$$T_j = \frac{1}{N} \sum_{k=1}^{N} T_{f_{j_k}}.$$

This additive model amounts to ignoring energy interferences between partials whose frequencies f_i are closed. The approximation would be valid for infinitely long periodic

signals but deteriorates as signal frame gets shorter. The equal weights $1/N$ suppose we do not know the energy balance between notes that compose a chord.

Energy-based observation

The pitch-based feature discriminated between events, it might not been efficient to discriminate silences from notes. To this aim, signal *loudness* is used. As described in references, a probability $P_{\text{rest}}(o_t)$ is computed based on the total energy of o_t.

Finally, the observation probabilities is defined as

$$\mathbb{P}(O_t = o_t \mid S_t = j) \overset{\text{def}}{=} \begin{cases} \mathbb{P}_{\text{pitch}}(o_t \mid j)\,(1 - P_{\text{rest}}(o_t)) & \text{if state } j \text{ is pitched,} \\ P_{\text{rest}}(o_t \mid j) & \text{if state } j \text{ is silence.} \end{cases}$$

Bibliographic remarks

The idea of a template-based similarity measures is very common in the literature of audio-to-score alignment. The observation model described above is also used in (Raphael, 2006; Montecchio and Orio, 2009; Montecchio and Cont, 2011; Joder et al., 2011). Common alternatives consist in (i) choosing a different spectral descriptor like semigram and chromagram — see (Joder, 2011, Section 2.2.1) for a review; (ii) choosing another divergence function or generative probabilities (Peeling et al., 2007; Duan and Pardo, 2011b). In most alignment algorithm, template construction is heuristic. Only a few approaches attempt to learn templates by supervised estimation of their parameters on annotated data (Joder et al., 2011; Korzeniowski and Widmer, 2013).

2.2.3 STATE SPACE MODELING OF MUSIC SCORES

Designing a State-Space Model begins with defining which unkown quantities are modeled by hidden state S, and choosing the adequate *state space* E that represents the set of all possible values for S. Basically, state space E represents the music score on which the audio signal is aligned. The practitioner has to choose an explicit procedure to build E out of any music score. The challenge is to properly model the two kinds of prior information a music score conveys: events are ordered and this ordering is known; events carry a nominal duration.

Discrete state space models

In a discrete ordered list of symbolic events, each one is labeled by its position on the list 1, 2, ... The simplest modeling consists in defining hidden state S as this event position. Such procedure induces a discrete state space $E = \{1, 2, \ldots\}$ where one event is modeled by one state.

A more elaborated idea consists in representing one event with several sub-states. This strategy is frequently used to make the model of observation more accurate. If observation O generated by an event is not stationary but exhibit several different phases, dividing the event into sub-states is interesting to define a different likelihood

for each phase. Then, state space is defined as union of all sub-states. Discrete state space are further described in section 2.3.

Continuous state space models

Among other applicative contexts of alignment, music scores has an outstanding property: they specify a nominal duration for each event. This property allows to identify each event through a continuous position variable, instead of the discrete label. Continuous states-space for score alignment are further described in section 3.3.1.

Latent variables

The minimum requirement for the hidden state S is to unambiguously identifies the quantity to be inferred (here, the event). However, a more sophisticated idea consists in adding *latent variables* into the state S. The Bayesian formula reveals that two kind of variables are potentially interesting. First, quantities that refine likelihood $\mathbb{P}(O_t \mid S_t)$ if the observation O emitted by each event. Second, quantities whose estimation improves prediction of future hidden states $S_{t+1}, S_{t+2}, \ldots$.

In score alignment, a popular latent variable is *tempo*, defined as the musical equivalent of speed. Indeed, it is common intuition that knowing the speed of a moving target should help predicting its future positions. Several algorithms (Raphael, 2006; Joder et al., 2011; Arzt and Widmer, 2010a) consider the augmented state space

$$E = \{\text{position} \times \text{tempo}\},$$

at the cost of designing a more involved statistical model for joint evolution of tempo and position. The present thesis does not cover tempo modeling nor any other latent variables. We exclusively focus on the basic problem of position modeling [7]. In experiments, whenever we mention the use of tempo decoding, it simply refers to the original tempo decoding model in `Antescofo` as described in (Cont, 2010).

2.3 DISCRETE STATE-SPACE MODELS

This section aims at introducing the main tool of our thesis: Hidden Semi-Markov Models (HSMMs). HSMMs are a generalization of HMMs which are far more popular in the scientific literature. So HMMs are introduced first in section 2.3.1 and their main drawback is explained: the duration model they offer is too constrained. Then, expanded state models are described and their lack of flexibility is underlined. This motivates using the HSMMs introduced in section 2.3.2. Even if semi-Markov models are more complex and far less common than Markov ones, they bring flexibility for nominal duration modeling. This turn out to be crucial for our theoretical study as well as our application performances.

7. Extending our approach for modeling tempo is however possible and left for future work.

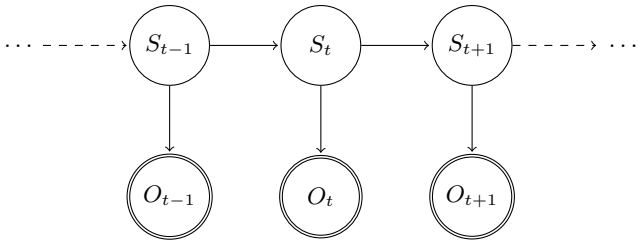


Figure 2.4: Probabilistic graphical model of a hidden Markov model (HMM). Arrows represent conditional dependencies between random variables. Double circles indicate observed variables.

Discrete space-space models are the most popular probabilistic approaches for alignment. To give a concrete illustration, a basic implementation for audio-to-score alignment drawn from the literature is described for every kind of model.

2.3.1 HIDDEN MARKOV MODELS

A *Hidden Markov Model* (HMM) is a State-Space Model (S, O) such that **(i)** the observation O checks the Markovian hypothesis; **(ii)** the hidden state S is a Markov chain. These two assumptions are equivalently represented by the graphical model depicted in figure 2.4.

A *Markov chain* is a discrete-time process defined on a countable state space E (typically $\{1, \dots, N\}, \mathbb{N}$ or $\mathbb{Z}$) and which checks the *Markov property*:

$$\mathbb{P}(X_{t+1} \mid X_1, \dots, X_t) = \mathbb{P}(X_{t+1} \mid X_t). \tag{2.3}$$

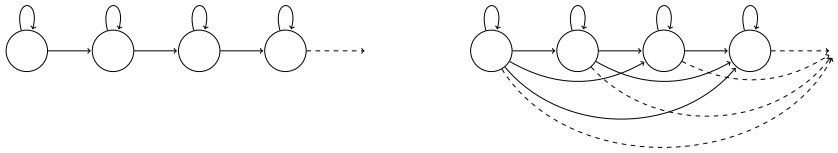
It means the evolution of a Markov process is *memoryless*: given current state S_t, future state S_{t+1} does not depend on the past $(S_1, \dots, S_{t-1})$. In other words, no further information besides knowing the current state S_t would help predicting future state S_{t+1}.

A Markov chain is called *time-homogeneous* if $\mathbb{P}(X_{t+1} \mid X_t)$ does not depend on t. Such a HMM is defined by its *transition matrix* $\mathbf{P} := (p_{i,j})_{i,j \in E}$ where $p_{i,j} := \mathbb{P}(X_{t+1} = j \mid X_t = i)$ are the *transition probabilities*, and its *initial distribution* $\pi := (\pi_i)_{i \in E}$ where $\pi_i := \mathbb{P}(X_1 = i)$. A more comprehensive presentation of discrete and continuous-time Markov processes is provided in the appendix A.2. We also refer to Cappé et al. (2005) for a survey on HMM.

INFERENCE ON HMM. HMM allows efficient computations of the posterior probabilities $\mathbb{P}(\cdot \mid S_1^t = s_1^t)$. As explained in appendix A.2.2, the Forward and Viterbi estimators may be computed with a linear complexity in time $O(t)$: computations take the form of a simple recursion over time t and therefore can be done on-line.

APPLICATION TO ALIGNMENT. The state space can be represented with an *automaton*, an oriented graph whose vertices are the possible states and whose edges $i \to j$ are weighted with $p_{i,j}$. Edges are represented only for allowed transitions (such that $p_{i,j} > 0$). In recognition tasks, the order of occurrence between events is not known, so

all transitions $i \to j$ are likely to be allowed. On the contrary, for alignment tasks the order is known. Such a prior information of ordering is modeled by using a *left-to-right* topology of automaton: $p_{i,j} = 0$ if $j < i$. In addition, assuming that no state can be skipped corresponds to the *linear* topology: $p_{i,j} = 0$ if $j \notin \{i, i+1\}$. Figure 2.5 illustrates both concepts. As we assume that the first event to occur is the first one of the score, initial probabilities are constrained to be $\pi_i = \delta_{1,i}$ in linear topology.



(a) linear chain (b) left-to-right chain

Figure 2.5: Examples of ordered HMM topologies. Arrows indicate the allowed transitions between states.

Linear HMM is the most widely used probabilistic model in audio-to-score alignment (Raphael, 1999; Cano et al., 1999; Orio and Déchelle, 2001; Cont, 2006; Montecchio and Orio, 2009). The simplest modelling of a music score consists in representing each of its symbolic event with one state, as in (Nakamura et al., 2013) for instance. Figure 2.6 illustrates this construction. State space is $E = \{1, 2, \ldots\}$ and hidden state S is defined as the event number. When modeling a music score with a linear HMM, the only choice up to the practitioner are self-transition probabilities $p_{j,j}$. Therefore, these parameters completely define the duration model of a linear HMM.

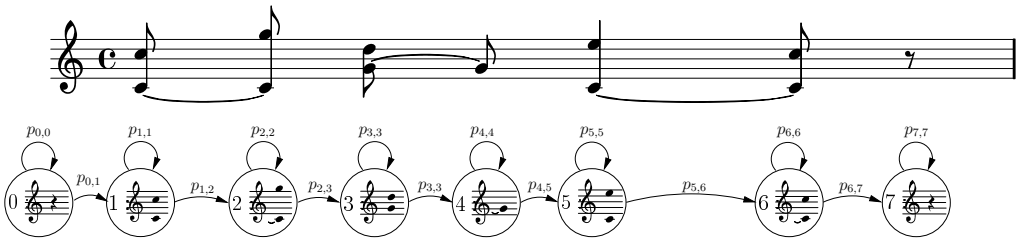


Figure 2.6: Modeling a music score by a linear HMM. Top: equivalent representation as sequence of events. Bottom: automaton of the corresponding state space with allowed transition probabilities and their parameter.

Markovian occupancy and expanded state Markov chains

Let L_i denotes the *occupancy* in state i, defined as the number of time steps spent by the hidden process S in the state i before leaving it. Implicitly, in a Markov chain the occupancy of each state implicitly obeys a geometric law parametrized by $p_{i,i}$,

$$\mathbb{P}(L_i = t) = (1 - p_{i,i})\, p_{i,i}^{t-1}.$$

Such restriction to geometric laws provides poor flexibility in modeling events of various lengths, especially for music signals. Even if one can adjust $p_{i,i}$, every geometric law

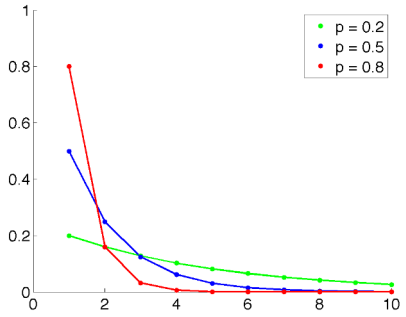


Figure 2.7: Graph of some geometric laws for different parameter values of p.

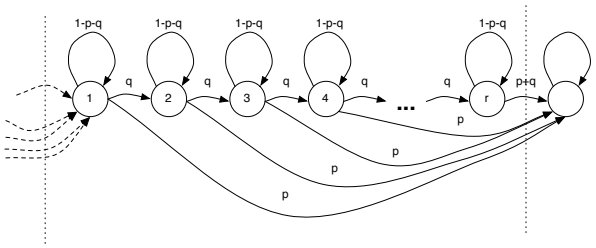


Figure 2.8: Example of a common topology of Markovian micro-states

favors shorter occupancies against long ones as Figure 2.7 reveals. Those laws poorly model long events such as music notes.

A more sophisticated strategy, called *Expanded state Markov chain*, consists in representing each symbolic event with more than one state. It is applied in most HMM-based audio-to-score alignment Raphael (1999); Montecchio and Orio (2009). The state space $E = \{1, \ldots, N\}$ is chosen as an arbitrary number N of *micro-states* and is partitioned into aggregates (aka. *macro-states*), where each aggregate represents one event:

$$E = \bigcup_{i=1,\ldots,J} C(i).$$

In the expanded setting, hidden state S is defined "sub-position" rather than event position. Each event i is represented by one *macro-state* $C(i)$, which is the union of its sub-states:

$$\{\text{event at time } t = i\} = \cup_{j \in C(i)} \{S_t = j\}.$$

Using micro-states having the same observation likelihood affects the occupancy distribution of events (Durbin et al., 1998, Section 3.4). Figure 2.8 depicts a common choice of topology. Combinations of Markovian micro-states could create very complex duration models, but this framework lacks flexibility. Evenf if the practitioner can tune transitions probabilities of sub-states, this is very difficult as the occupancy distribution of a combination of sub-states has no simple closed-form formula. Moreover, achieving interesting distributions may require an infinite number of sub-states. This is why we prefer focusing on HSMMs: they generalize the expanded state approach and give full expressiveness in duration model to the practitioner.

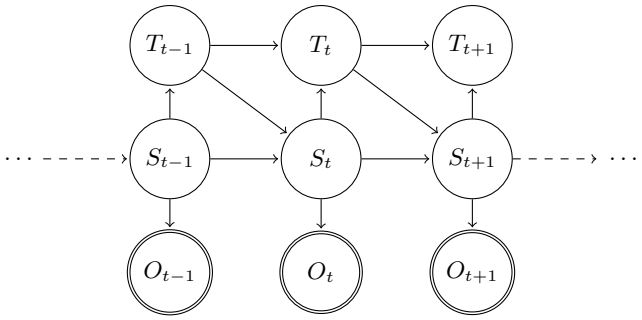


Figure 2.9: Probabilistic graphical model of a hidden semi-Markov model (HSMM). Arrows represent conditional dependencies between random variables. Double circles indicate observed variables.

2.3.2 HIDDEN SEMI-MARKOV MODELS

A Hidden Semi-Markov Model (HSMM) is a State-Space Model (S, O) such that **(i)** the observation O checks the Markovian hypothesis; **(ii)** the hidden state S is a semi-Markov chain. As already explained, the major drawback of Markov chains is the constraint of state occupancies: implicitly, occupancy distribution D_j of each state j is always a geometric law parametrized by the self-transition probability $p_{j,j}$. Semi-Markov chain are an extension of Markov chains that allows using explicit occupancy distribution D_j for each state j: the practitioner is free to choose *any* valid probability distribution on $\mathbb{N}$ (or on $\mathbb{R}_+$ for continuous-time chains). Consequently, a single semi-Markovian state can replace any aggregate of Markovian micro-states. Since setting D_j as a geometric distribution make this state Markovian, D_j may be seen as a generalization of Markovian self-transition probability $p_{j,j}$.

EQUIVALENT HMM DESCRIPTION. A semi-Markov chain S on E can be equivalently described by a Markov chain (S, T) on the augmented state space $E \times \mathbb{R}_+$, where T_t denotes the current occupancy time on the current state s such that $S_t = s$. This representation of HSMM is often called "explicit duration HMM" and its graphical model is depicted in Figure 2.9. Note that the sole semi-Markov process S is not memoryless and does not check the Markovian evolution property (equation (2.3)). However, the joint process (S, T) does:

$$\mathbb{P}\left((S_{t+1}, T_{t+1}) \mid (S_1, T_1), \ldots, (S_t, T_t)\right) = \mathbb{P}\left((S_{t+1}, T_{t+1}) \mid (S_t, T_t)\right).$$

As a result, inference can still be computed efficiently with recursive algorithms — refer to appendix A.3.3 or to (Guédon, 2003).

HSMMs have been announced as generalizations of expanded state HMMs. In truth the two models are theoretically equivalent. A semi-Markov chain can be represented by an expanded state Markov chain, but this has some drawbacks as the required number of micro-states might be infinite. We refer to (Guédon, 2005, Section 3) for a comprehensive discussion on this macro-state representation and its drawbacks.

APPLICATION TO MUSIC ALIGNMENT. Modeling hidden position on music score by a linear HSMM has been first done by Cont (2010). The modeling strategy is concep-

tually very simple: it suffices to represent each symbolic event with one semi-Markovian state. So the music score is represented by a chain with linear topology. Figure 2.10 illustrates this idea.

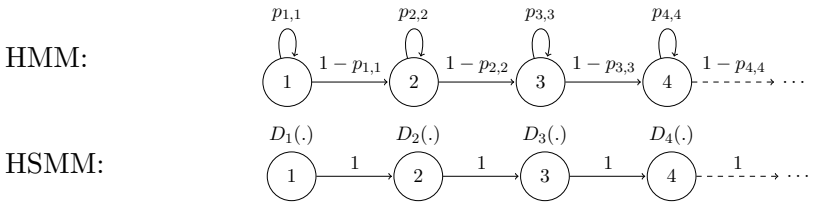


Figure 2.10: Comparison of linear graphical models for alignment and their parameters. Self-transitions probabilities $p_{j,j}$ of HMM are replaced by occupancy distributions $D_j(.)$ of HSMM.

As explained in appendix A.3.2, it is always possible to assume that a semi-Markovian state has no self-transition probability: $p_{j,j} = 0$. So in a linear HSMM, the duration model is completely defined by the set of all occupancy distributions $(D_j, j \in E)$, as they are the only free parameters left to the practitioner. But finding out the right distributions to model music events is a major design issue and is the main topic of this work.

2.4 CONTINUOUS STATE-SPACE MODELS

The main motivation of this thesis is the design of discrete models: how to encode the prior information of music scores with a HSMM? To get insights on discrete models, our idea it to seek inspiration in alternatives: continuous State-Space Models for score alignment.

Consequently, this section briefly describes such approaches. Section 2.4.1 explains that music events can be embedded in a continuous space, so that hidden state is chosen as a continuous variable. State-of-the art approaches model the evolution of this continuous position with Gaussian random walks, as described in section 2.4.2.

2.4.1 CONTINUOUS VERSUS DISCRETE REPRESENTATIONS

The Discrete models described in section 2.3 conveniently represent a list of symbolic events, as such information is inherently of discrete nature. However, music events carry additional information: besides being ordered and contiguous, events are associated to a *nominal duration*. Such information has an importance consequence: discrete events of music score may be embedded in a *continuous* half-line which represents the virtual time of music. Once this timeline is drawn, any music score is modeled by partitioning $\mathbb{R}_+$ with respect to the nominal durations $(l_1, l_2, \ldots)$ of its symbolic events. This means

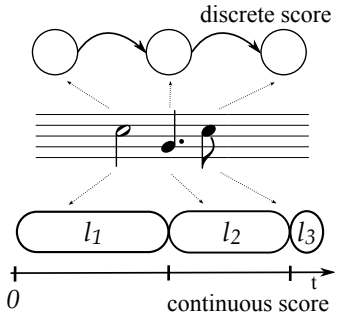


Figure 2.11: Two representations of a music score: sequence of discrete events (discrete position) or intervals of a virtual timeline (continuous position). Such intervals are determined by onset position (ticks).

an event j has explicitly an *onset position* $l_{1:j-1}$ (0 for the first event), a *nominal duration* l_j and an *offset position* $l_{1:j}$, where we use the notation

$$l_{1:j} \overset{\text{def}}{=} \sum_{i=1}^{j-1} l_i.$$

In other words, each event j is represented by its position interval $[l_{1:j-1}, l_{1:j})$. Figure 2.11 illustrates the two representations of a music score.

Consequently, the basic idea of a continuous state space is to directly model the *beat position* l, which is a real number in the continuous space $E = \mathbb{R}_+$ whose unit is the music beat. This differs from discrete models where score position is represented by event number j which is an integer in $E = \{1, 2, \ldots\}$.

As explained by Montecchio and Orio (2009), exploiting a continuous representation of the reference media unifies tasks of alignment onto symbolic sequences such as scores or continuous sequences such as other audio files. Hidden position l and its space $\mathbb{R}_+$ can either represent the symbolic time in a music score (measured in beats), or the physical time in another audio file (measured in seconds).

2.4.2 GAUSSIAN RANDOM WALKS

Continuous state spaces have already been suggested in diverse MIR applications as alternatives to discrete state-spaces like HSMMs. In audio-to-score alignment, all approaches we know have used *Gaussian random walks* in order to model beat position (Raphael, 2006; Montecchio and Orio, 2009; Montecchio and Cont, 2011; Duan and Pardo, 2011a,b; Otsuka et al., 2011) The probabilistic assumption is that continuous position $L = (L_n)_{n\in\mathbb{N}}$ evolves as a random walk along score timeline. Note that due to implementations issues, these models are discrete-time ($t \in \mathbb{N}$).

Let us explain this idea. The design of evolution on score should take into account prior information indicated on music score: ideally, the performance should perfectly respect score nominal durations. Modeling this prior information with continuous position L is straightforward. Ideally, its initial value is $L_0 = 0$ and its evolution obeys the following dynamics:

$$L_n = L_{n-1} + \tau\, \Delta T,$$

where ΔT denote the sampling period (in seconds) and τ denote the score tempo (in beats per seconds). Here, we have assumed the music score provides tempo value τ as it rules the conversion from beat position to physical time. We call *nominal performance* this ideal evolution which keeps the tempo constant.

The idea of modeling L by a random walk is to add random perturbations $W = (W_n)_{n\in\mathbb{N}}$ to nominal performance in order to model temporal deviations of actual performances. For Gaussian random walks, dynamics of L read:

$$L_n = L_{n-1} + \tau\,\Delta T + W_n, \tag{2.4}$$

where W is a Gaussian white noise: $W_n \sim \mathcal{N}(0, \sigma^2\Delta T)$ are i.i.d., centered Gaussian random variables with variance $\sigma^2\Delta T$. Figure 2.12 depicts a realization of this random walk L. Note that the process L is still Markovian, but contrary to discrete Markov chains it is supported on a continuous space. L is equivalently described by its state transition pdfs:

$$p(l_n \mid L_{n-1} = l_{n-1}) = \mathcal{N}(l_n; l_{n-1} + \tau\Delta T, \sigma^2\Delta T).$$

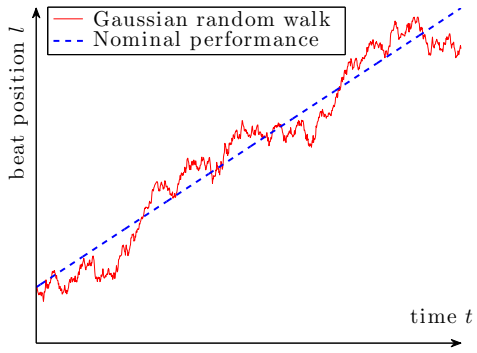


Figure 2.12: Numerical simulation of continuous position L modeled with a Gaussian random walk (equation (2.4) with $\tau = \sigma = 1$), and comparison with nominal performance.

FROM CONTINUOUS TO DISCRETE POSITION. We would like to highlight an important feature of random walk models: initially, continuous position L evolves along the music beat line $\mathbb{R}$ *independently* from music score. Then, choosing a given music score consists in dividing this line into position intervals $[l_{1:j-1}, l_{1:j})$ that represent each event j. Therefore, discrete position S is retrieved as *spatial discretization* of L between onset positions $\{0, l_1, l_1 + l_2, \ldots\}$:

$$\forall n \in \mathbb{N}, \qquad S_t = l_{1:j-1} \text{ for } j \in E \text{ such that } l_{1:j-1} \leq L_n < l_{1:j}.$$

PERSPECTIVES. This kind of model is further studied later on in section 3.3, where other interesting properties are discussed as well as some drawbacks. This is why the extension of this approach by general random walks is also suggested there.

2.5 PROBLEMATIC: DESIGNING COHERENT PROBABILISTIC MODELS

This section contextualizes the main problematic of this thesis, which is designing a hidden stochastic process that correctly model information on nominal duration of events. Taking into account this prior information in discrete models is a crucial and so far undermined question. Our investigation is primarily undertaken in the framework of hidden semi-Markov models (HSMM) as it provides explicit choice of duration model. Section 2.5.1 refines Question 1 (raised in chapter 1) in this context.

As explained in chapter 1, designing a probabilistic process is only one half of an inference algorithm. The other half is the way estimation of hidden quantities is performed out of the model. Section 2.5.2 refocuses Question 2 in our framework, with an emphasis on online estimation.

2.5.1 MODELING PRIOR INFORMATION OF DURATION

Our approach for alignment is to model every music score with a linear HSMM. In this framework, the question of duration modelling reduces to the design of D_j, which are prior distributions on the actual occupancy on each state j.

To tune such parameters, common approaches rely on statistical learning. This step consists in training an algorithm with some dataset so as to make the probabilistic fit data as best as possible. For instance, the Baum-Welch algorithm is an Expectation-Maximization (EM) method that (re-)estimates HMM parameters with the maximum likelihood criterion (Rabiner, 1989). This algorithm has been extended to HSMM by Guédon and Cocozza-Thivent (1990). But as the class of probability distributions has infinite dimension, such non-parametric estimation would require a prohibitive amount of training data. For this reason, most implementations assume that all occupancy distributions belongs to one of the standard parametric families of probabilities. Indeed, parametric estimation requires less training, generalizes better to new data, and is easy to implement with parametric versions of the HSMM Baum-Welch algorithm (Levinson, 1986; Mitchell and Jamieson, 1993). As reviewed by Yu (2010), popular choices of probability families are: Gaussian, Gamma, log-normal, Negative Binomial or Poisson laws. Whereas such choice is usually left as a secondary implementation question, this manuscript raises it as its main topic.

Question 1.1. *Which parametric families of probability distributions are coherent occupancy distributions?*

This work is based on the following assumption: *two events with identical nominal duration should get identical occupancy distributions*. This assumption behaves as a tying constraint between occupancy distributions D_j. So the duration model consists in a *set of durations* $L \subset \mathbb{R}_+$ and a *duration-indexed family* of probability distributions $(D_l)_{l \in L}$ such that for all state j, $l_j \in L$ and $D_j = D_{l_j}$, where l_j denotes the nominal duration of state j. This framework sharpens our problematic as follows.

Question 1.2. *Are there coherent mappings from nominal duration l to occupancy distribution D_l?*

Usually, common answers are grounded on two kind of reasons: computational efficiency, and heuristics inspired by *ad hoc* knowledge on the application. Many heuristics have been proposed in the Music Information Retrieval literature.[8]

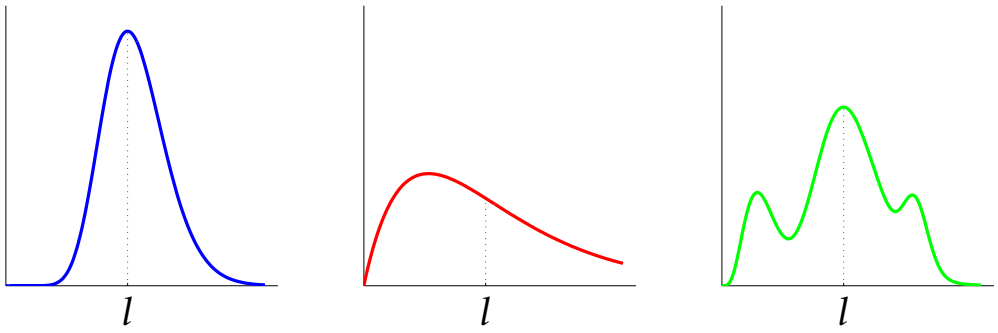


Figure 2.13: Three proposals of occupancy distribution D_l that model nominal length l.

Let us look at the three toy proposals for D_l on Figure 2.13. A common heuristic is to set the *mean* of D_l as l. Though it is fulfilled by the three proposals, only the first distribution looks relevant: it is unimodal (unlike the third one) and its *mode* also equals l (unlike the second one). So which statistical quantity of D_l should we map on l: mean, mode or both?

This work aims at providing theoretically grounded answers to such questions. Our approach consists in looking for some rationale to justify or disqualify some candidates among all possible ones. To do so, we make use of a specificity of our application: *musical events are associated with a nominal duration* — as explained in section 2.1.2. Although a few music alignment approaches like (Joder et al., 2010b) willingly discard this prior information, we believe encoding it in probabilistic models is crucial, motivating this work to get further insights.

2.5.2 CHOICE OF DECODING METHOD

As we will see in chapters 3 and 4, coherency of a probabilistic model is not decorrelated from other design choices such as the *estimation method* (also called *decoding method*). Indeed, the ultimate goal of alignment algorithms is to estimate the full sequence of hidden states $s_1^T = (s_1, \dots, s_T)$ which explains at best the observation sample o_1^T. But the notion of "best" must be specified first since it has no single and universal meaning. In the framework of probabilistic models, "best" is related to "most likely" and inference provides a mechanism to compute posterior probabilities $\mathbb{P}(\cdot \mid O_1, \dots, O_t)$. However, multiple definitions of "most likely" still compete.

Moreover, the way estimation is defined highly depends on the applicative setting. As explained in section 2.1.1, two kinds of alignment are usually carried on.

8. Suggestions of duration models in MIR literature are reviewed later on in section 6.1.2, so as to confront them with our results.

- *Offline* alignments make a single estimation at final time T once full observation sample o_1^T is known. So the full state-sequence s_1^T has to be estimated.
- *Online* alignments make sequential estimations at each intermediate time $t = 1 \dots T$ with partial observation sample o_1^t. At time t, only the current state s_t at time t needs to be estimated so as to incrementally recover the full path s_1^T.

This thesis mainly deals with online sequential alignment. In this setting, a decoding method consists of a classifying function[9] $\mathbf{C}(t) : E \to \mathbb{R}$, defined on the state space E, that is inferred at each time t conditionally to $\{O_1^t = o_1^t\}$. Then, current state is estimated as the *mode* of $\mathbf{C}(t)$, that is to say the state with highest function value:

$$\hat{s}_t = \mathrm{mode}[\mathbf{C}(t)] = \arg\max_{j \in E} C_j(t).$$

An additional goal of this work is to assess possible choices of estimators through their ability to account for nominal durations.. In this setting, Question 2 refines as follows.

Question 2.1. *Which decoding method is the most coherent for online estimation?*

Only the two most common methods are studied in this thesis: estimation of the most likely current *state* s_t, and estimation of the most likely partial *state-sequence* s_1^t, both with Maximum A Posteriori (MAP). But any alternative method could be examined with the same methodology.

Forward estimator. This method looks for current state with is marginally most likely:

$$\hat{s}_t \overset{\text{def}}{=} \arg\max_{s_t \in E} \mathbb{P}(S_t = s_t \mid O_1^t = o_1^t).$$

Forward inference consists in computing the following quantity for each time step t and state j:

$$f_j(t) \overset{\text{def}}{=} \mathbb{P}(S_t = j \mid O_1^t = o_1^t),$$

so this corresponds to estimator $C_j(t) = f_j(t)$.

Viterbi estimator. It looks for the end state $\hat{s}_t$ of the most likely path

$$\hat{s}_1^t \overset{\text{def}}{=} \arg\max_{s_1, \dots, s_t \in E} \mathbb{P}(S_1^t = s_1^t \mid O_1^t = o_1^t).$$

Viterbi inference consists in computing the following quantity for each time step t and state j:

$$\delta_j(t) \overset{\text{def}}{=} \max_{s_1, \dots, s_{t-1} \in E} \mathbb{P}(S_t = j, S_1^{t-1} = s_1^{t-1}, O_1^t = o_1^t),$$

so this correponds to estimator $C_j(t) = \delta_j(t)$.

Remark. Definitions of the estimators make no assumptions whether current state occupying hidden process S ends at or after current time t. Such a setting called *right-censored* estimation by Yu (2010) is not standard in all HSMM implementations. A common alternative is to assume that occupancy of current state always ends at t. This matter will be specifically examined in chapter 5.

9. In the literature, this function is usually called a *estimator*, a *classifier*, or a *decoder*.

Although both classifiers are defined for online alignment, in our terminology we associate the Viterbi estimator with state-sequence estimation, and the Forward estimator with state estimation. Indeed, the Viterbi classifier δ_j is also involved in offline estimation of the most likely state-sequence — further explanations are given in section 3.4. As a result, we also get insights on offline estimation while investigating online estimation.

SURVEY OF THE LITERATURE. Very few works discuss the issue of choosing an estimation method. Lember and Koloydenko (2014) undertake a theoretical discussion on offline estimation methods for HMM. In the MIR literature, the only substantial work we know is that of Raphael (1999), who introduces several non-conventional estimators for online alignment.

Joder (2011, Section 3.4.2) explains the effect of using several identical Markov states per event. Adding such micro-states does not really change the duration model for the Viterbi estimation, but it does for the Forward estimation[10]. Montecchio and Orio (2009); Orio and Déchelle (2001); Schwarz et al. (2004) explicitly prefer Forward to Viterbi. Orio and Déchelle (2001) support such choice with experimental evidence: *"the comparison of this decoding with Viterbi showed lower delay in detecting note changes and higher robustness to errors"*. Incidentally, chapter 4 will give further justifications to this claim. Raphael (1999) also chooses an estimator related to the Forward-Backward algorithm, which is the offline version of Forward estimation.

2.5.3 APPROACH

Generative modelling in our application context entails important design choices that have been discussed in this chapter. Despite their influence on the performances of any inference algorithm, they are severely underestimated in both the scientific and practical literature of score alignment and information retrieval. Machine Learning (ML) may optimize parameter values for given models but cannot decide for higher level choices that are left to the practitioner. Any ML method would requires at least one engineering choice: parametric methods requires a parametric family; iterative methods requires a first guess for the initialization. Existing approaches provide little insight on how to choose the "right" algorithmic design for a peculiar applicative context. The goal of this work is to provide theoretical backing and practical insight on each implementation choices, by studying their mathematical foundations and qualifying their practical impact on alignment algorithms.

To do so, this works carries on an axiomatic approach for the design of alignment algorithms. We introduce two original criteria that define how a *coherent* algorithm should behave. This abstract notion of coherency has to be understood as compliance with the available prior information — in our context, the music score — that algorithms have to correctly model. Chapter 3 investigates the first coherency criterion, which has been introduced in (Cuvillier and Cont, 2014). Chapters 4 and 5 deal with the second criterion which has been introduced in (Cuvillier, 2014). Both criteria are

10. This simple fact reveals the influence of estimation method on the effective duration model.

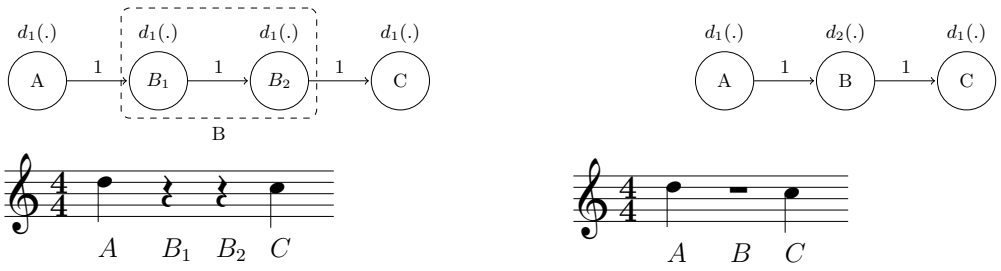
inspired by properties of continuous models of position. As we will see in the next chapters, not all known discrete models like HSMM would comply without specific conditions that our study will reveal. Such results either shed light on some popular *ad hoc* practices or provide new prescriptions about engineering choices. Those theoretically grounded properties equally improve performances of real-life alignment applications, as discussed in chapter 6 where comparative experiments are run.

3

COHERENT INFERENCE OF EQUIVALENT SEQUENCES

This chapter investigates our first criterion of coherency. Section 3.1 motivates and formalizes the criterion. In the reminder of the chapter, coherency of semi-Markov models is investigated for two estimation methods. Section 3.2 characterizes coherent models for the Forward estimator by means of the notions of infinite divisibility and Lévy processes. In addition, this result leads to an underlying model of continuous position. This is why section 3.3 compares this kind of process with the continuous models introduced in chapter 2 (section 2.4), after having extended this latter approach with general random walks. This provides a conceptual unification between the two approaches and a deeper understanding of the respective probabilistic hypotheses which lead the practitioner to choose either approach. Afterwards, section 3.4 goes back to the criterion and characterizes coherent semi-Markov chains for the Viterbi estimator by means of convex analysis. Coherent chains turn out to be too constrained to have a practical interest.

3.1 STATEMENT OF CRITERION OF COHERENCY



(a) Four events of duration 1.

(b) Three events of duration 1,2,1.

Figure 3.1: Two equivalent music scores and their graphical models. The long (𝅝) silence B aggregates the short (𝄽 𝄽) ones (B_1, B_2).

The criterion devised in this chapter is motivated by a peculiarity of our application domain. In music alignment, scores can feature *repeated events*: successive events that emit the same observation signal. Even more, an event may be rewritten into several events in another version of the music score. As a result, different real-life music scores may carry the same prior information. For instance, an event of nominal duration $l = 2$ is equivalent in terms of occupancy to two consequent events of nominal duration $l = 1$. As obvious as it might seem, this simple property is not necessarily true for probabilistic processes that model score position.

Figure 3.1 illustrates the concept with two toy sequences. The music score 3.1a informs of four events with equal nominal duration of 1. In the music score 3.1b one silence of duration 2 replaces the two ones of duration 1. Therefore, the two sequences would generate identical observations O since no physical signal could distinguish consecutive silences. The "one state per event" modeling procedure would map 3.1b to the 3 states graphical model (A, B, C) and 3.1a to the 4 states one (A, B_1, B_2, C), so state B carries the same prior as the aggregate (B_1, B_2).

Coherency would ask that equivalent music scores lead to equivalent inference algorithm. However, this cannot be *exactly* the case. Inference is carried on a state space E of 3 states for the score 3.1b, and of 4 states for score 3.1a. With such different state spaces, the inferred quantities cannot be numerically equivalent as they have not the same dimension. What we can ask is equivalent inference for the *other* states, *i.e.*, the ones that are not aggregated: such events are respectively $E \setminus \{B\}$ for score 3.1b and $E \setminus \{B_1, B_2\}$ for score 3.1a.

Coherency criterion 1. Aggregating a linear sub-chain of N events $j_1, \ldots, j_N$ with nominal duration $l_{j_1}, \ldots, l_{j_N}$ into a single event j of nominal duration $j = l_{j_1} + \ldots + l_{j_N}$ does not change the inferred quantities for the remaining states.

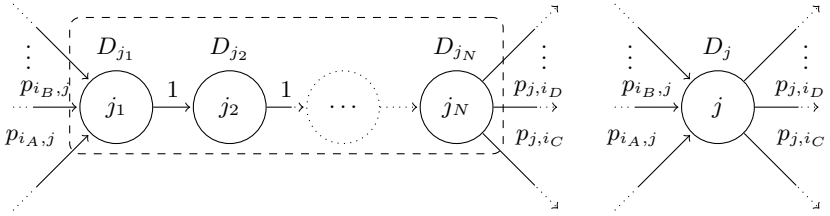
Intuitively, linear aggregates of events have the following interpretation: the duration spent in state j, called *occupancy*, is a random variable Occ_j whose law is D_j. The duration spent in an aggregate $(j, j+1)$ of two consecutive states is the sum of their individual durations: $Occ_{(j,j+1)} = Occ_j + Occ_{j+1}$. Since the two random variables are independent in a HSMM, the law of the aggregate is the convolution product of individual laws: $D_{(j,j+1)} = D_j * D_{j+1}$. According to this interpretation, the validity of the criterion 1 only depends on the occupancy distributions D_j, that is to say the prior duration model of the HSMM. Actually, the validity also depends on the choice of estimation method. This chapter focuses on the two methods introduced in section 2.5.2 for state estimation. Section 3.2 deals with the Forward estimator $f_j(t)$, whereas section 3.4 deals with the Viterbi estimator $\delta_j(t)$ — which is also involved in state-sequence estimation.

3.2 COHERENT STATE INFERENCE: FROM LINEAR SEMI-MARKOV CHAINS TO LÉVY PROCESSES

For state inference, the classifier $\mathbf{C}(t)$ is chosen as the Maximum A Posteriori (MAP) estimator $C_j(t) = f_j(t)$:

$$f_j(t) \stackrel{\text{def}}{=} P\Big(S_t = j \mid O_1^t = o_1^t\Big).$$

In the context of HMM/HSMM, it is also called the Forward estimator.



Chain A Chain B

Figure 3.2: Graphical model of a linear aggregate. State j in chain B (right) aggregates the linear sub-chain $(j_1, \dots, j_N)$ in chain A (left).

3.2.1 CHARACTERIZATION: INFINITELY DIVISIBLE DISTRIBUTIONS AND CONVOLUTION SEMIGROUPS

To study our coherency criterion, we generalize the toy example of Figure 3.1 and consider two semi-Markov chains. Chain A aggregates part of chain B. Their difference is illustrated in Figure 3.2.

Chain A: a chain with any topology on a state space E that features a linear sub-chain of states $j_1, \dots, j_N$ of nominal duration $l_{j_1}, \dots, l_{j_N} \in L$ and with identical observation probabilities $b := b_{j_1} = \dots = b_{j_N}$.

Chain B: the same chain except that states $j_1, \dots, j_N$ are replaced by a single state j of nominal duration $l_{j_1} + \dots + l_{j_N}$ and observation probabilities b.

With this formalization, it can be asserted that criterion 1 is fulfilled if and only if aggregating $(j_1, \dots, j_N)$ into j does not change the classifier values $f_k(t)$ between the two chains for all other states $k \notin \{j, j_1, \dots, j_N\}$.

DESCRIPTION OF THE EQUIVALENCE. The validity of criterion 1 mostly depends on occupancy distributions $(D_l)_{l \in L}$. So we say this family is *aggregate-coherent* for the Forward estimation if it leads to equivalent constructions. Additional conditions on other HSMM parameters are required to achieve equivalence. In particular, the Markovian hypothesis of conditional independence of the observation is an essential requirement. The following proposition provides all such mentioned conditions that ensure equivalence between chains A and B.

Proposition 3.1 (Coherent aggregates)**.** For the Forward state inference, a linear aggregate $(j_1, \dots, j_N)$ is equivalent to a single state j if:

1. states share the same observation probabilities:
   $b_j = b_{j_1} = \dots = b_{j_N}$,
2. observation model check the Markovian hypothesis 2.2,
3. initial probabilities of the aggregate check $\pi(j_k) = \pi(j)\, \delta_{1,k}$,
4. $(j_1, \dots, j_N)$ is a linear subchain and outer transition probabilities check
   $\forall i \neq j, \quad p_{i,j} = p_{i,j_1}$ and $p_{j,i} = p_{j_N,i}$,

5. occupancy distributions checks

$$D_j = D_{j_1} * \ldots * D_{j_N}.$$

Therefore, a family $(D_l)_{l\in L}$ of occupancy distributions is aggregate-coherent for the Forward inference if and only if:

$$\forall l_1, l_2 \in L, \qquad l_1 + l_2 \in L, \tag{3.1}$$

$$D_{l_1+l_2} = D_{l_1} * D_{l_2}. \tag{3.2}$$

Proof. We refer to forthcoming Proposition 3.5 as the proof is identical: tt suffices to replace maximizations max by summations $\sum$. □

Equation (3.1) is the definition of L being an *additive subsemigroup* of $\mathbb{R}_+$. This condition is the minimal one that ensures the construction of chain B is well-defined.

Equation (3.2) is the definition of $(D_l)_{l\in L}$ being a *convolution semigroup*[1] of probability distributions: the convolution product[2] $*$ is the internal operation. Indeed, convolution preserve probability measures: if D, E are such measure, so is $D * E$. Furthermore, convolution is an associative operator, so equivalence for the case $N = 2$ induces the general case $N \geq 2$.

CHARACTERIZATION OF COHERENT FAMILIES. The existence of aggregate-stable families depends on the structure of L, which is the set of nominal durations. In music scores, durations are usually written with rational subdivisions of a unit called a *beat*: $L \subset l_b\mathbb{Q}_+^*$, where l_b denotes the physical duration associated to a symbolic beat. However, the value l_b depends on the score tempo and on the sampling rate for discrete-time models.

Case 1: $L = l_0\mathbb{N}$. If a music score of finite length $\mathbf{l} = (l_1, \ldots, l_J)$, there always exists a base duration $l_0 > 0$ such that

$$\{l_1, \ldots, l_J\} \subset \{l_0, 2l_0, 3l_0, \ldots\} = l_0\mathbb{N}^* \subset L.$$

This base duration l_0 is called the *temporal atom* (or *tatum*) in musicology and in the MIR literature (Bilmes, 1993; Klapuri et al., 2006). In this case, aggregate-coherent families $(D_{nl_0})_{n\in\mathbb{N}}$ are characterized as convolution semigroups indexed on $\mathbb{N}$. To build such a semigroup, one can choose D_{l_0} as any valid probability measure D then compute its successive convolution powers:

$$\forall n \in \mathbb{N}^*, \qquad D_{nl_0} := \underbrace{D * D * \ldots * D}_{n \text{ times}}.$$

Case 2: $L = \mathbb{R}_+$ or at least $L \subset l_b\mathbb{Q}_+^*$. Indeed, the set of all possible music scores virtually contains all rational multiples of the pulse l_0. So one could ask at least that L contains all subdivisions of l_b:

$$l_b, l_b/2, l_b/3, \ldots \in L. \tag{3.3}$$

1. Convolution semigroups are further detailed in the appendix, see Definition A.18.
2. The convolution product of measures is defined in appendix A.1.1.1.

If condition (3.3) is true and the family is aggregate-coherent, then

$$\forall n \in \mathbb{N}^*, \quad D_{l_b} = \underbrace{D_{l_b/n} * D_{l_b/n} * \ldots * D_{l_b/n}}_{n \text{ times}}.$$

This latter property turns out to be the exact definition of D_{l_b} being an *infinitely divisible* probability distribution. Appendix A.4 gives all required mathematical background of infinite divisibility, including its relationship with *Lévy processes* and convolution semigroups. In particular, a key result is the equivalence between the three following statements:

- D_{l_b} is infinitely divisible and supported on $\mathbb{R}_+$.
- There exists a *continuous* convolution semigroup $(\tilde{D}_l)_{l\geq 0}$ supported on $\mathbb{R}_+$ such that $\tilde{D}_{l_b} = D_{l_b}$.
- There exists a $\mathbb{R}_+$-valued Lévy process $X = (X_l)_{l\geq 0}$ such that $\tilde{D}_l$ is the marginal law of X_l.

This result provides the characterization we are looking for. As soon as equations (3.2) and (3.3) holds for *one* value of l_b, then it holds for *all* values: there exists an aggregate-coherent family $(D_l)_{l\geq 0}$ indexed on the "full" set of lengths $L = \mathbb{R}_+$. And reciprocally, any continuous convolution semigroup supported on $\mathbb{R}_+$ provides a coherent family of this kind.

NULL-DURATION OCCUPANCIES. In theory, occupancy distributions are supported on $\mathbb{R}_+$ but should not have an atom at 0, *i.e.*, $D_l(\{0\}) = 0$. Indeed, the inverse condition would mean the hidden process is likely to spend a duration 0 on the state it enters. This is not a genuine occupancy but rather a skip to subsequent state. As a strictly linear topology does not allow such skips, atoms at 0 cannot be allowed. With continuous-time models, this is not a problem since many Lévy processes supported on $\mathbb{R}_+$ has no atom at 0. With discrete-time models, the situation is more involved as no Lévy process supported on $\mathbb{N}$ checks the requirement $D_l(\{0\}) = 0$. Fortunately we have figured out a way to overcome this limitation.

- A linear semi-Markov chain with occupancy distributions such that $D_j(\{0\}) > 0$ can be "normalized" into a left-to-right chain with zero-truncated occupancies such that $\tilde{D}_j(\{0\}) = 0$. This matter is further discussed in appendix A.3.2.
- Even if Proposition 3.1 is stated for linear aggregates, its result can be proved for left-to-right chains obtained after normalization of linear chains[3].

Criterion 1 tells that no strictly linear chain is fully coherent in discrete time: only left-to-right chains may be so. However in practice, moving from the linear to the left-to-right topology increases the computational complexity of inference from $O(J)$ to $O(J^2)$, where J is the number of states. To avoid this, we prescribe choosing a true convolution semigroup $(D_l)_{l>0}$ and truncating the values at 0 so as to keep the chain

3. We describe such left-to-right chains as "linear in a weak sense".

linear: $\tilde{D}_l := \frac{D_l(x)-D_l(\{0\})\delta_0}{1-D_l(\{0\})}$. The resulting family $(\tilde{D}_l)_{l\geq 0}$ fulfills equation (3.2) approximately. This approximation is rather poor with very small durations but becomes better as nominal durations l get longer. Indeed, $D_l(\{0\}) = (D_1(\{0\}))^l \underset{l\to\infty}{\to} 0$ for any convolution semigroup. This *"small duration problem"* will actually occur several times throughout the manuscript, for different reasons. It warns that inference on very short events might be hazardous.

CONCLUSION. A new result as outcome of the proposed coherency criterion is the promotion of infinitely divisible distributions as prior occupancy distributions of a HSMM. Such an idea has never been suggested in the literature of semi-Markov chains. Interestingly, most of the standard parametric laws that are employed in the literature *are* infinitely divisible distributions: Gamma, Negative Binomial, Log-normal, Gaussian, Exponential laws, and so on. So our criterion might give insights, and maybe theoretical grounds, to those practitioners choices.

3.2.2 RELATING POSITION TO OCCUPANCY

This section and the following one aim at interpreting the coherent semi-Markov chains we have characterized. To do so, we shall go back to the probabilistic foundations of this kind of model. We give a few general facts on linear models of position, then contextualize such facts to linear semi-Markov chains. Probabilistic models consists in considering that time spent on each state is random. From now one, assume the state space E has a linear topology.

Occupancy Occ_j of state $j \in E$ is defined as the random duration of j.

Onset T_j of j is defined as the random time at which j beings to occur. As state space is linear, the onset process $T = (T_j)_{j\in E}$ defines an *arrival time process*, whose interarrival times are successive occupancies Occ_j:

$$T_j = \sum_{k=1}^{j-1} Occ_k \qquad (T_0 = 0).$$

Discrete position $S = (S_t)_{t\geq 0}$ corresponds to the *counting process* associated to onset times $T = (T_j)_{j\in E}$. Counting processes are also called last-passage times, as they are defined as the last "symbolic time" j at which the arrival process stays on t:

$$S_t = \sup\{j \in E \mid T_j \leq t\} \qquad (0 \text{ if empty}).$$

Conversely, onset process T coincides with the left-continuous *first-passage times* associated to the discrete position process S. Occupancies are deduced as inter-onset times:

$$T_j = \inf\{t \geq 0 \mid S_t \geq j\}, \qquad \text{and } Occ_j = T_{j+1} - T_j.$$

Graphically, this identity tells that a realization of onset times T is a function which is the left-continuous inverse of the realization of S.

Case of linear semi-Markov chains

Within this setting, modeling discrete position S by a linear semi-Markov chain is equivalent to assuming that all occupancies are independent random variables $Occ_j \sim D_j$. Onset process T is an arrival process where interarrival times are independent but may not be identically distributed. Such processes are generalizations of *renewal processes* (Nelson, 1995, Chapter 6) which are a more usual notion in probability theory.

Now, let us focus on the marginal law of S_t which is called the *state distribution*. It is denoted $\mathbf{f}(t)$ and defines a probability distribution on E. For semi-Markov chains, state distributions are implicit and deduced from occupancy distributions D_j, which are the quantities to be explicitly chosen by the practitioner. For linear semi-Markov chains, the relationship between $\mathbf{f}(t) = (f_1(t), f_2(t), \dots)$ and D_j has a simple expression stated by next proposition.

Proposition 3.2 (State distributions of linear chains)**.** The state probabilities $f_j(t) = \mathbb{P}(S_t = j)$ of a linear semi-Markov chain $S = (S_t)_{t\geq 0}$ with occupancy distributions $(D_j)_{j\in E}$ are given by, for all $j \in E$, $t \geq 0$,

$$\begin{aligned} f_j(t) &= D_1 * \dots * D_{j-1}(t) - D_1 * \dots * D_j(t), \\ &= \overline{D_1 * \dots * D_j}(t^+) - \overline{D_1 * \dots * D_{j-1}}(t^+) \\ &\Big(= \overline{D_1 * \dots * D_j}(t+1) - \overline{D_1 * \dots * D_{j-1}}(t+1) \quad \text{if } D_j \text{ are discrete}\Big), \end{aligned}$$

where $\overline{D}$ denotes the survivor distribution of D. And in particular,

$$f_1(t) = 1 - D_1(t) = \overline{D}_1(t^+).$$

Remark. Last proposition assumes the process S starts out and "enters" the first state at time $t = 0$. Such convention is standard (Guédon, 2003, 2005). Moreover, the proposition holds for linear chains in a wide sense, for which occupancies may have null durations ($D_j(\{0\}) \geq 0$).

Proof. If topology is linear, S is the counting process associated to T. They are related by the standard identity (Gut, 2005, Equation 16.1),

$$\forall t \geq 0, j \in E, \qquad \{S_t \geq j\} = \{T_j \leq t\}.$$

Taking probabilities gives

$$\mathbb{P}(S_t \leq j) = \sum_{i\geq j} f_i(t) = \mathbb{P}(T_j \leq t).$$

As $T_j = \sum_{k=1}^{j-1} Occ_k$ and Occ_k are independent, T_j is distributed as $D_1 * \dots * D_{j-1}$ ($= \delta_0$ if $j = 1$). Together with the identity $D(t) + \overline{D}(t^+) = 1$, this gives

$$\sum_{i\geq j} f_i(t) = D_1 * \dots * D_{j-1}(t) = 1 - \overline{D_1 * \dots * D_{j-1}}(t^+),$$

and the result follows by computing differences with respect to j. □

3.2.3 INTERPRETING COHERENT SEMI-MARKOV CHAINS WITH CONTINUOUS PROCESSES

We assume each state j represented a timed-event with nominal duration l_j and nominal onset position $l_{1:j-1} := \sum_{k=1}^{j-1} l_k$. This means each event is associated to a nominal interval $[l_{1:j-1}, l_{1:j})$. So we can express all processes with these nominal quantities instead of event number.
Discrete position takes values on nominal onsets: $S_t = j \leftarrow l_{1:j-1}$.
Occupancies are indexed with nominal intervals: $Occ_j \leftarrow Occ_{[l_{1:j-1}, l_{1:j})}$.
Onsets times are indexed with nominal onsets: $T_j \leftarrow T_{l_{1:j-1}}$.
Once the set of score onsets $\mathbf{l} := (l_{1:j-1})_{j \in E}$ is chosen, the relationship between position and onset reads

$$T_{l_{1:j}} = \sum_{k=1}^{j} Occ_{[l_{1:j-1}, l_{1:j})}, \qquad S_t = \sup \{l \in \mathbf{l} \mid T_l \leq t\}.$$

In this setting of linear modeling of events, coherency criterion 1 may be rephrased as follows: the probabilistic representation of an event j only depends on the *nominal interval* $[l_{1:j-1}, l_{1:j})$ it occupies on the music score. As occupancies are determined by onset times as follows

$$Occ_{[l_{1:j-1}, l_{1:j})} = T_{l_{1:j-1}} - T_{l_{1:j}},$$

such idea is equivalent to the following statement.

Coherency criterion 1 (case of linear models, continuing from p. 48)**.** Onset time process $T = (T_l)_{l \in \mathbf{l}}$ is *independent from the score*, in the sense that T_l only depends on score onset value l and not on the whole score structure $\mathbf{l} = (l_{1:j-1})_{j \in E}$.

With general processes such as linear semi-Markov chains, onset process T highly depends on score structure. Indeed, for a given an onset value l and two different score, distribution of onset time T_l might differs as it depends on previous onsets times $l_j \leq l$. However in section 3.2.1, we have derived the following result: choosing all occupancies D_{l_j} in a continuous convolution semigroup $(D_l)_{l \geq 0}$ achieves this independence.

In addition, this condition leads to a new probabilistic construction of the semi-Markov chain S, based on a *continuous process*. The idea is to choose a non-negative Lévy process $T = (T_l)_{l \geq 0}$ whose distributions $(D_l)_{l \geq 0}$ gives occupancies. T is indexed by continuous half-axis $\mathbb{R}_+$ which represents the virtual timeline of beat positions l. Priorly to any music score, T gives a random function that maps virtual score time l to physical time[4] t. Then, any music score is represented by its list of onsets positions $\mathbf{l} := (l_{1:j-1})_{j \in E}$. There are two equivalent constructions of the semi-Markovian process S that models discrete position (with convention $S_t = l_{1:j-1}$ if current event number is j).

Construction 1. Sampling the fully-indexed process $(T_l)_{t \geq 0}$ on score onsets $\mathbf{l}$. Then, discrete position S is retrieved as described in section 3.2.2. S is the counting process associated to the sampled process $(T_l)_{l \in \mathbf{l}}$, defined as

$$S_t = \max\{l \in \mathbf{l} \in E \mid T_l \leq t\}.$$

4. Choosing a process T with discrete values provides a discrete-time model ($t \in \mathbb{N}$), whereas other choices lead to a continuous-time model ($t \in \mathbb{R}_+$).

Construction 2. "Inverting" the random time map T. This amounts to consider the *first-passage times* $L = (L_t)_{t\geq 0}$ associated to T, formally defined as right-continuous inverse function of T:

$$L_t = \inf\{l \geq 0 \mid T_l > t\}.$$

Such process L represents a continuous position whose evolution is independent from the chosen music score. Then, S is retrieved by spatial discretization of continuous position L between score onsets $\mathbf{l}$,

$$S_t = l_{1:j-1} \text{ for } j \in E \text{ such that } l_{1:j-1} \leq L_t < l_{1:j}.$$

The validity of last construction is ensured by the following proposition. It elaborates Proposition 3.2 for convolution semigroups.

Proposition 3.3. If the family of occupancy distributions $(D_l)_{l\geq 0}$ defines a convolution semigroup, then the state probability $f_j(t) = \mathbb{P}(S_t = j)$ of the linear semi-Markov chains S parametrized by nominal durations $(l_1, l_2, l_3, \ldots)$ is

$$f_j(t) = D_{l_{1:j-1}}(t) - D_{l_{1:j}}(t).$$

The state distribution $\mathbf{f}(t)$ coincides with spatial discretization of the distribution M_t of L_t between score onsets $\mathbf{l}$, where L_t is the first-passage time at threshold t associated to the Lévy process $T = (T_l)_{l\geq 0} \sim (D_l)_{l\geq 0}$,

$$f_j(t) = M_t[l_{1:j-1}, l_{1:j}).$$

Proof. Semigroup property and associativity of convolution product gives $\overline{D_{l_1} * \ldots * D_{l_j}} = \overline{D}_{l_1+\ldots+l_j}$. Combining this with Proposition 3.2 gives

$$f_j(t) = D_{l_{1:j-1}}(t) - D_{l_{1:j}}(t).$$

In addition, the cumulative distribution function of the first-passage time L_t is $M_t[l, \infty) = D_l(t)$. A proof of this elementary fact is postponed to Proposition 7.6 in chapter 7. This implies

$$M_t[l_{1:j-1}, l_{1:j}) = M_t[l_{1:j-1}, \infty) - M_t[l_{1:j}, \infty) = D_{l_{1:j-1}}(t) - D_{l_{1:j}}(t).$$

This formula does correspond to spatial discretization of L_t, since by definition of a probability distribution,

$$\mathbb{P}(l_{1:j-1} \leq L_t < l_{1:j}) = M_t[l_{1:j-1}, l_{1:j}).$$

□

Second construction is the most interesting one as it brings about a *continuous model* of position. It tells that coherent semi-Markov chains S are spatial discretizations of the first-passage times L associated to some non-negative Lévy process T. This means that our model for discrete position S is not fundamentally discrete: there exists a continuous position L from which S derives, and this continuous model is independent from the chosen music score. In addition, having a continuous position space simplify the relationship between onset and position:

- Continuous position L is the first-passage-time process of onset time T:

$$L_t = \inf \{l \geq 0 \mid T_l > t\} .$$

- Onset time T is the left-continuous first-passage-time process of position L:

$$T_l = \inf \{t \geq 0 \mid L_t \leq l\} .$$

In other words, a realization of T is a random function that maps virtual time (score position) l to physical time t, and L is the inverse mapping from t to l.

Besides this conceptual interest, this reasoning has important mathematical consequences. Indeed, some probabilistic properties are preserved through spatial discretization. If the Lévy process T has been chosen such that its first-passage times L bear one of these properties, last proposition ensures that all semi-Markov chains, obtained from all possible music scores $(l_j)_{j \in E}$, would inherit this property. This reasoning will be used several times in this chapter and in chapter 4. It also explains why chapter 7 later on is devoted to the theoretical study of first-passage times of Lévy processes, independently from their application to alignment.

3.3 UNIFICATION OF CONTINUOUS AND DISCRETE MODELS

Before studying coherency criterion 1 with another estimator, we raise an interesting remark that somehow unifies two different kinds of probabilistic models that have been used in MIR applications including audio-to-score alignment. Semi-Markov chains are discrete State-Space Models that represent discrete position. Section 3.2.1 suggests choosing occupancy distributions D_l as a convolution semigroup $(D_l)_{l \geq 0}$ of infinitely divisible distributions. As explained in section 3.2.3, such design provides the existence of a continuous position process L that gives back discrete position S when discretized between score onsets.

This conclusion is reminiscent of *continuous* State-Space Models that are alternatives to discrete State-Space Models. In particular, Gaussian random walks have been described in section 2.4.2 as explicit models of continuous position. Hereafter, section 3.3.1 further studies such processes and stresses out a particular property: compliance between discrete-time and continuous-time models. Then, it explains why Lévy processes exhaust all random walk models that achieve this compliance property. This suggests generalizing Gaussian model with random walks based on any Lévy process. In particular, Section 3.3.2 focuses on processes with non-decreasing paths as they respect the event ordering like semi-Markov models do.

Section 3.3.3 compares the two approaches, semi-Markov chains (discrete position) and random walks (continuous position), and emphasizes the differences in terms of probabilistic hypotheses that lead to either approaches. It shows how Lévy process conceptually unify the two approaches and rules out some of their probabilistic differences — but not all in general.

To finish with, section 3.3.4 goes further towards unification. It explains why the two approaches are strictly *equivalent* in one special case: the Poisson process. This conclusion is interesting as Poisson processes also appear as optimal choice for reasons of a different kind throughout the manuscript.

3.3.1 GENERALIZING CONTINUOUS POSITION MODELS WITH RANDOM WALKS

Gaussian Random walks have been suggested to model continuous position L on score timeline. As further explained in section 3.3.1, such model consists in assuming L has initial value $L_0 = 0$ and obeys the following dynamics:

$$L_n = L_{n-1} + \tau\,\Delta T + W_n, \tag{3.4}$$

where the random perturbations $W_n \sim \mathcal{N}(0, \sigma^2 \Delta T)$ are independent and identically distributed (i.d.d.), centered Gaussian random variables with variance $\sigma^2 \Delta T$.

Equation (3.4) describes dynamics of the discrete-time process $L = (L_n)_{n\in\mathbb{N}}$ associated to sampling time ΔT. Choosing Gaussian perturbations brings about an interesting property we would like to stress out: this discrete-time modeling is *compliant* with continuous-time modeling, in the sense there exists a continuous-time process $\tilde{L} = (\tilde{L}_t)_{t\geq 0}$ such that L coincides with the periodic discretization of $\tilde{L}$,

$$\exists \tilde{L} = (\tilde{L}_t)_{t\geq 0}, \qquad \forall \Delta T > 0, \qquad \forall n \in \mathbb{N}, \quad L_n = \tilde{L}_{n\Delta T}.$$

Gaussian random walks L are time-compliant and the corresponding $\tilde{L}$ is a continuous-time Gaussian process ruled by the following stochastic differential equation

$$\mathrm{d}\tilde{L}_t = \tau\,\mathrm{d}t + \sigma\,\mathrm{d}W_t,$$

where $W = (W_t)_{t\geq 0}$ is a *Brownian motion* (Sato, 1999, Chapter 1). Figure 3.3 depicts the Gaussian process $\tilde{L}$. Discretizing the latter equation would give back equation (3.4). This compliance with discretization explains the value $\sigma^2 \Delta T$ for the variance of W_n, as it would not hold for any other value. As a realization of $\tilde{L}$ is nowhere differentiable, its instantaneous speed is undefined. Tempo τ corresponds to the "mean" speed value called *drift*.

Generalization and characterization of time-compliant models

This discrete-time construction is generalized by choosing L as any random walk on $\mathbb{R}$.

Definition 3.4. A *random walk* $X = (X_n)_{n\in\mathbb{N}}$ is a discrete-time random process such that initial value X_0 equals 0 almost surely and increments $X_{n+1} - X_n$ are independent and stationary (*i.e.*, identically distributed).

Equivalently, one can choose any probability distribution D on $\mathbb{R}$. Consider $Y = (Y_n)_{n\in\mathbb{N}^*}$ being i.i.d. random variables such that $Y_n \sim D$. Then, $X_n := \sum_{k=1}^{n} Y_k$ defines a random walk X with increments distributed as D.

Now, we raise the question: which other random walks enjoy time compliance? Continuous-time versions of random walks are processes $X = (X_t)_{t\geq 0}$ with similar properties, formally defined as:

- null initial value: $X_0 = 0$ almost surely,
- independent increments: for any $0 \leq t_1 < t_2 < \ldots < t_n < \infty$, $X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \ldots, X_{t_n} - X_{t_{n-1}}$ are independent,
- stationary increments: for any $t, u, s \geq 0$, $X_{s+u} - X_s$ is equal in distribution to $X_{t+u} - X_t$.

It turns out that this class of processes **coincides** with Lévy processes. Refer to appendix A.4 for more details. In addition, the periodic sampling $(X_{n\Delta T})_{n\in\mathbb{N}}$ of a Lévy process always gives a random walk with infinitely divisible increments, for any sampling period $\Delta T > 0$. And reciprocally, any infinitely divisible distribution induces such a process. We conclude that a discrete-time model of random walk L is compliant with a continuous-time model if and only if the increments distribution D is infinitely divisible.

Summing up, the Gaussian random walk offers an explicit modelling of continuous position L. Such a model is generalized by one of the two following constructions:

Continuous time: choose $L = (L_t)_{t\geq 0}$ as a Lévy process.

Discrete time: choose $L = (L_n)_{n\in\mathbb{N}}$ as a random walk with infinitely divisible increments $L_{n+1} - L_n \sim D$.

These two methods are equivalent and exhaust the cases for which continuous-time and discrete-time models are compliant, in the sense we have defined above: $L_n \sim L_{n\Delta T}$.

3.3.2 CONTINUOUS POSITION MODELS WITH NON-DECREASING PATHS

The Gaussian processes introduced at the beginning has a conceptual problem. They do *not* have non-decreasing paths, as Figure 2.12 exhibits. Allowing such backwards moves has two conceptual pitfalls. First, they contradict the hypothesis that events occur with respect to the score ordering. Second, they allow beat position L with negative values, though this has no meaning since music scores start at beat position 0 wit their first event. Consequently, we suggest choosing L as a process with non-decreasing paths. For random walks or Lévy processes, the following three properties are equivalent: **(i)** having non-decreasing paths; **(ii)** having non-negative increments; **(iii)** being non-negative. So our prescription is to model continuous position L with a non-negative Lévy process or a random walk.

Relating onset times and occupancies

Assume L has non-decreasing paths. For such processes, as previously explained in section 3.2.2, onset times and positions are related through first-passage times. Let T_l denote the actual onset time of a (possibly virtual) event with onset position l. A realization of the process T is the inverse function of the realization of L, and vice versa. Figure 3.3 illustrates those processes L, T together with discrete position S.

- Onset time $T = (T_l)_{l\geq 0}$ is the left-continuous first-passage-time process of continuous position L,

$$T_l = \inf\{t \geq 0 \mid L_t \geq l\}. \tag{3.5}$$

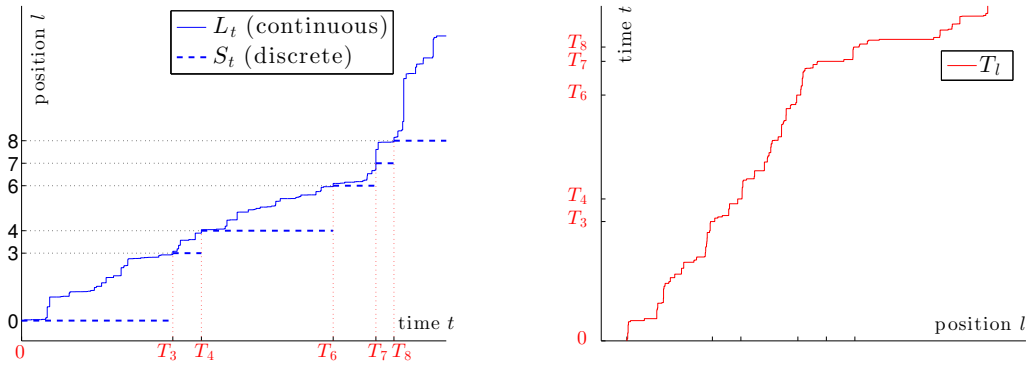


Figure 3.3: Realization of a continuous position $(L_t)_{t\geq 0}$ (a Gamma process) and its first-passage times $T = (T_l)_{l\geq 0}$. For a toy score of 6 onsets $(l_{1:j}) = (0, 3, 4, 6, 7, 8)$, discrete positions S_t are retrieved by spatial discretization. Onset times of events are retrieved as $T_{l_{1:j}}$.

- Continuous position $L = (L_t)_{t\geq 0}$ is the (right-continuous) first-passage-time process of onset time T,
$$L_t = \inf\{l \geq 0 \mid T_l > t\}.$$
- Then, take some event j whose nominal interval is $[l_{1:j}, l_{1:j-1})$. Its occupancy is deduced from onset times as
$$Occ_j = T_{l_{1:j}} - T_{l_{1:j-1}}.$$

Remark. For discrete-time models $L = (L_t)_{t\in\mathbb{N}}$, the result is still valid by considering the right-continuous extension of L: for all $t \geq 0$, $L_t := L_{\min\{n\in\mathbb{N} | n\geq t\}}$.

Random walks models for L are interesting as they *always* fulfill coherency criterion 1. Indeed, equation (3.5) shows the onset time T_j of a discrete event j only depends on onset position $l_{1:j-1}$, and not on the choice of score.

However, non-decreasing random walks has two probabilistic differences with semi-Markov chains. First, occupancies $(Occ_j)_{j\in E}$ are not independent (except for one random walk described later on in section 3.3.3). Second, the onset process T does not have stationary increments: occupancy Occ_j of an event depends not only of nominal duration l_j, but also of its onset position $l_{1:j-1}$. As a result, the same event of duration l_j does not get the same occupancy distribution at two different onset positions $a \neq b$, $T_{a+l_j} - T_a$ and $T_{b+l_j} - T_b$. Conceptually, this means that moving a music note along the score modifies the random time spent on this note.

3.3.3 COMPARISON OF CONTINUOUS AND DISCRETE POSITION PROCESSES

Discrete models based on semi-Markov chains and continuous models based on random walks are two mirrored points of view. We draw a parallel between their respective characteristics and underlying probabilistic hypotheses. We also summarize prescriptions we have figured out for each model and the benefits they bring.

With a general random walk:

- Position L is continuous and its evolution is independent from the score. Discrete position is retrieved through discretization of L between event score onsets $\mathbf{l} = \{0, l_1, l_1 + l_2, \ldots\}$.
- Position process L has independent increments and explicit probability distributions.
- Occupancies are correlated, non-stationary (in general) and have implicit distributions, induced from first-passage times T of position L by $Occ_j = T_{l_{1:j+1}} - T_{l_{1:j}}$. Occupancy of an event j depends on its nominal duration l_j and onset $l_{1:j-1}$.

Our prescription in section 3.3.1 is to choose an infinitely divisible distribution for position increments. Doing so makes such discrete-time modeling compliant with continuous-time: there exists a process $\tilde{L} = (\tilde{L}_t)_{t\geq 0}$ such that $L_n = \tilde{L}_{n\,\Delta T}$ for all $\Delta > 0$.

With a general linear-semi Markov chain:

- Position is discrete, and its evolution depends on the score.
- Occupancies Occ_j are independent and have explicit probability distributions D_j. Occupancy of an event j only depends on its nominal duration l_j.
- Discrete position S has correlated and non-stationary increments (in general) and implicit distributions, induced as first-passage times of onset time process $(T_j)_{j\in E}$ where $T_j = \sum_{k=1}^{j} Occ_k$.

Our prescription in section 3.2 is to choose all occupancies D_j in one convolution semigroup $(D_l)_{l\geq 0}$. Doing so makes such a discrete position model compliant with a continuous position, through the basic discretization operation: there exists a continuous random variable L_t which gives back discrete position S_t if discretized between $\{0, l_1, l_1 + l_2, \ldots\}$. This setting shares similarities with random walks: continuous position L evolves independently from the symbolic score; discrete position is obtained through spatial discretization. However a disparity remains discrete-time and continuous-time semi-Markov chains are *not* time-compliant. In general, periodic sampling $(S_{n\Delta T})_{n\in\mathbb{N}}$ of a semi-Markov chain $(S_t)_{t\geq 0}$ is no longer semi-Markovian. The only exception is when S is actually a Markov chain with exponentially distributed occupancies.

Unified constructions with Lévy processes

As a result, our prescriptions unify constructions of coherent semi-Markov chains (section 3.2) and time-compliant random walks (section 3.3.1). The common idea is to start from a random process X with independent and stationary increments. X provides the random mapping between onset times and beat positions or vice versa. Then, the two constructions differ by their probabilistic hypotheses and their interpretation of such process.

Semi-Markov chain: $X = (X_l)_{l\geq 0}$ is indexed on symbolic time l and maps it onto physical time t. The hypothesis is that *onset times* have independent and stationary increments.

Random walk: $X = (X_t)_{t\geq 0}$ is indexed on physical time t and maps it onto symbolic time l. The hypothesis is that *positions* have independent and stationary increments.

In both cases, the continuous and onset times processes are constructed *priorly* to any music scores. The continuous position (X itself or its first-passage times) evolves freely on the symbolic timeline of music, independently from music events. Afterwards, the music score only acts on discretization of continuous position.

3.3.4 EQUIVALENCE WITH POISSON / GAMMA PROCESS

Random walks and semi-Markov chains stem from different probabilistic hypotheses. In random walks, position increments are assumed to be independent. In semi-Markov chains, occupancies are so. However, there exists *one* special case where these two models coincide: the **Poisson process**. In this case, *both* occupancies and position increments are independent. It is implemented as follows:

Semi-Markov chain: choose the occupancy distributions D_l as Poisson laws of mean λl, $D_l \sim Po(\lambda l)$.

Random walk: choose the distribution of position increments as an exponential law of mean $1/\lambda$, $L_{n+1} - L_n \sim \mathcal{E}(\lambda)$.

Indeed, if $X = (X_l)_{l\geq 0}$ is a Poisson process $X_l \sim Po(\lambda l)$, a standard result states that its first-passage times $T = (T_n)_{n\in\mathbb{N}}$ define a discrete-time random walk with exponentially distributed increments $T_{n+1} - T_n \sim \mathcal{E}(\lambda)$.

Conversely, if $L = (L_n)_{n\in\mathbb{N}}$ is a discrete-time random walk with exponentially distributed increments $\mathcal{E}(\lambda)$, then its first-passage times $T = (T_t)_{t\geq 0}$ define a Poisson process[5].

As a result, the two implementations lead to the same probabilistic model. Continuous position $L = (L_n)_{n\in\mathbb{N}}$ has independent and stationary increments and is Gamma distributed: $L_n \sim \Gamma(n, 1/\lambda)$. Occupancies Occ_j are independent and Poisson distributed: $Occ_j \sim Po(\lambda l_j)$. They also are stationary as they only depend on duration l_j, not on the onset position $l_{1:j-1}$.

Remark. Consider Lévy process-based models where position has continuous and nonnegative increments. Among such models, the Poisson / Gamma model is the only one that simultaneously provides independent, stationary occupancies *and* independent, stationary position increments. Indeed, random walks with nonnegative increments are also called renewal processes. A classical result of renewal theory states that Poisson process is the only renewal process with independent and stationary increments — refer to (Nelson, 1995, Section 6.3.7). And similarly, we can show that Poisson process is the only non-decreasing Lévy process whose first-passage times define a random walk.

Remark. The Poisson / Gamma have discrete-time occupancies and continuous position, since the Poisson process is supported on $\mathbb{N}$ but indexed on $\mathbb{R}_+$. However, the exponential distribution of position increments turns out to be infinitely divisible. So the continuous position L could be embedded in a continuous-time Lévy process $\tilde{L}$, namely a *Gamma process* $\tilde{L}_t \sim \Gamma(t, 1/\lambda)$. Nevertheless, the continuous-time process $(\tilde{L}_t)_{t\geq 0}$ has not the same occupancies as the discrete-time random walk $(L_n)_{n\in\mathbb{N}}$: occupancies are independent for the latter but correlated for the former.

5. The two claims are classic results of renewal theory and can be found in many references such as (Nelson, 1995).

3.4 COHERENT STATE-SEQUENCE INFERENCE

So far we have investigated coherency of Forward state estimation and its consequences. This section repeats such study for Viterbi state estimation. Such alternative estimator if current state S_t consists in choosing classifier $\mathbf{C}(t)$ as the so-called Viterbi classifier $C_j(t) = \delta_j(t)$, with

$$\delta_j(t) \stackrel{\text{def}}{=} \max_{s_1,\dots,s_{t-1}\in E} \mathbb{P}(S_t = j, S_1^{t-1} = s_1^{t-1}, O_1^t = o_1^t).$$

Even if we are speaking of online state estimation, this classifier is also involved in offline estimation of state-sequence $(S_1, \dots S_t)$ as it corresponds to the MAP criterion on paths:

$$(S_1, \dots, S_t) \stackrel{\text{def}}{=} \operatorname*{arg\,max}_{(s_1,\dots,s_t)\in E^t} \mathbb{P}(S_1^t = s_1^t, O_1^t = o_1^t).$$

Indeed, state-sequence estimation can be recursively divided in two steps:

1. estimate end state S_t as $\arg\max_{j\in E} \delta_j(t)$.
2. estimate intermediate path $(S_1, \dots S_{t-1})$ by backtracking arg max during computations of $\delta_j(u)$, $u = 1 \dots t$.

3.4.1 PROPOSITION: CHARACTERIZING COHERENT DISTRIBUTIONS

The following proposition describes occupancy distributions that fulfill criterion 1 with the Viterbi estimator. It transposes Proposition 3.1 to such estimation. As a preliminary remark, we highlight that Viterbi estimator requires a regularity assumption on occupancy distributions D_j to be well-defined. All D_j must be either discrete (for discrete-time HSMMs) or absolutely continuous (for continuous-time). This is why this proposition deals with these two cases. Its proof is postponed to section 3.4.2.

Proposition 3.5. Let $(D_l)_{l\in L}$ be a family of probability distributions indexed by $L \subset \mathbb{R}_+$.

Assume each D_l is discrete and let d_l denote its pmf. The family is aggregate-coherent for the Viterbi estimator if and only if

$$\begin{aligned} \forall l_1, l_2 \in L, \qquad & l_1 + l_2 \in L, \\ & d_{l_1+l_2}(t) = \max_{u=0,\dots,t} d_{l_1}(u) d_{l_2}(t-u). \end{aligned}$$

Assume each D_l is absolutely continuous on $\mathbb{R}_+$ and let d_l denote its pdf. The family is aggregate-coherent for the Viterbi estimator if and only if

$$\begin{aligned} \forall l_1, l_2 \in L, \qquad & l_1 + l_2 \in L \\ & d_{l_1+l_2}(t) = \sup_{u\in[0,t]} d_{l_1}(u) d_{l_2}(t-u). \end{aligned}$$

This proposition highlights an operation called *inf-convolution* ($\star_{\inf}$) (Bauschke and Combettes, 2011, Chapter 12):

$$f_1 \star_{\inf} f_2(t) \stackrel{\text{def}}{=} \inf_{u\in\mathbb{R}} f_1(t-u) + f_2(u),$$

where $-\log f_i(x) = +\infty$ if $f_i(x) = 0$ or $f_i(x)$ is not defined. With this operation, conditions obtained in the two cases (discrete and continuous-time) can be summed up in a single one:

$$\forall l_1, l_2 \in L, \qquad l_1 + l_2 \in L, \\ [-\log d_{l_1}] \star_{\inf} [-\log d_{l_2}] = -\log d_{l_1+l_2}.$$

Remark.

- In discrete time, Proposition 3.5 is valid if $D_l(\{0\}) > 0$. In continuous time, it is only valid with strictly linear aggregates, *i.e.*, $D_l(\{0\}) = 0$.
- Like convolution $*$, inf-convolution $\star_{\inf}$ is an associative operation. This is why coherency of 2-state aggregates implies the result for N-state aggregates.
- The condition of Viterbi coherency is similar to Forward and their respective proofs are identical, expect that summation $\sum$ over possible durations is replaced by maximization sup.

Unlike convolution, inf-convolution of two probability densities does not always gives another probability density. The resulting density might not respect the normalization constraint as its integral might not equal 1. Therefore, conditions state in Proposition 3.5 for Viterbi aggregate-coherency are twofold:

Semigroup: $(-\log d_l)_{l\in L}$ is a semigroup of $[0,\infty]$-valued functions for inf-convolution $\star_{\inf}$.

Normalization: the integral/sum of each function d_l is equal to 1.

Hereafter, we show that no discrete-time family can fulfill both conditions. We also show that coherent families exist in continuous-time but have a severe limitation.

NONEXISTENCE IN DISCRETE-TIME. The normalization constraint turns out to be impossible to fulfill in discrete-time.

Proposition 3.6. No family of discrete probability mass functions $(d_l)_{l\in L}$ on $\mathbb{N}$ is aggregate-coherent for the Viterbi inference if $\{l, 2l\} \subset L$ for some length $l > 0$, except families composed of trivial distributions $\delta_a, a \in \mathbb{N}$.

Proof. The proof is straightforward since for any pmfs d_l, d_m,

$$\forall t \in \mathbb{N}, \quad \max_{u=0\ldots t} d_l(t-u)d_m(u) \le \sum_{u=0}^{t} d_l(t-u)d_m(u) = [d_1 * d_m](t).$$

The right-hand side quantity $d_l * d_m$ always defines a pmf. For the left-hand side to define a valid pmf, summing it over t must give 1. Therefore the inequality must be an equality:

$$\forall t \in \mathbb{N}, \quad \max_{u=0\ldots t} d_l(t-u)d_m(u) = \sum_{u=0\ldots t} d_l(t-u)d_m(u).$$

So all terms except one in the right-side must vanish. This can be true only if d_l or d_m is a trivial pmf δ_a for some integer a. □

EXISTENCE IN CONTINUOUS-TIME. It is worth investigating if non-existence of coherent families is an artifact of discrete-time modeling. Is the situation similar with *continuous-time* semi-Markov chains? The answer is no, coherent families of continuous-time probability densities do exist. Next proposition characterizes all these families. Even if they define a large semi-parametric class, all coherent families are highly constrained and bear the same drawback. This limitation is a consequence of the normalization constraint.

The characterization involves the notion of subadditivity.

Definition 3.7. A function $f : \mathbb{R}_+ \longrightarrow \mathbb{R}$ is said to be *subadditive* if

$$\forall x, y \in \mathbb{R}_+, \qquad f(x) + f(y) \geq f(x + y).$$

Proposition 3.8 (Aggregate-coherent families)**.** Assume there exists some base duration $l_0 > 0$ such that $\{l_0, 2l_0, 3l_0, \ldots\} \subset L$. The family of probability distributions $(D_l)_{l \in L}$ is aggregate-coherent for the Viterbi inference if and only if there exists an absolutely continuous distribution $\tilde{D}$ such that:

(i) distributions D_l are shifted versions of $\tilde{D}$: there exists $a \geq 0$ such that

$$\forall l \in l_0 \mathbb{Q}_+^* \cap L, \qquad D_l = \tilde{D}(\cdot - al),$$

(ii) there exists a real function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ which is subadditive, positive on $(0, \infty)$ with $\phi(0) = 0$, such that $\tilde{D}$ admits $\tilde{d}$ for pdf and

$$\forall t \geq 0, \qquad \tilde{d}(t) = e^{-\phi(t)}.$$

As the proof of this original result is somewhat lengthy, we postpone it to section 3.4.2. This result only describes occupancies D_l for l that are rational multiples of the base length l_0. Extending the description to all lengths $l \geq 0$ requires a further assumption of right-continuity: $D_{l+\epsilon}$ should become identical to D_l as ϵ goes to 0. This assumption is very natural as it guarantees inference does not vary drastically when one length l is replaced with an infinitely close value $l + \epsilon$.

Corollary 3.9. Let $(D_l)_{l \geq 0}$ be an aggregate-coherent family for the Viterbi inference (with $L = \mathbb{R}$). Assume this family is right-continuous at $l = 0$. *i.e.*, $\lim_{\epsilon \to 0^+} D_\epsilon = D_0$ in distribution. Then, there exists $a \geq 0$ such that

$$\forall l \geq 0, \qquad d_l(t) = d_0(t - al).$$

Conversely, any pdf d_0 such that $\phi = \log d_0$ fulfills the conditions of Proposition 3.8 induces a continuous Viterbi-coherent family on $L = \mathbb{R}_+$.

As a result, all Viterbi-coherent families are severely constrained. Necessarily, every d_l is a shifted version of the distribution d_0; d_l is supported on $[la, \infty)$ and reaches its maximum at la.
If $a = 0$, all densities d_l are identical: this is not acceptable as it would mean that nominal duration l_j of events do not influence the inference.
If $a \neq 0$, each path has to stay a minimum time la on each state of duration l. So a can be interpreted as the minimal allowed tempo. But d_l reaches its maximum at al, meaning this minimum tempo is also the most likely one.

3.4.2 PROOF OF PROPOSITIONS

Proof of Proposition 3.5

This section proves Proposition 3.5 that describes Viterbi aggregate-coherent families. For the sake of simplicity, we assume that

- $N = 2$
- both chains have linear topology,
- there are no null-duration occupancy, *i.e.*, $D_j(\{0\}) = 0$,
- D_j are discrete-time distributions.

However, the result holds without first three assumptions, and it can be extended to continuous-time chains by letting d_j denote the pdf of D_j instead of the pmf.

As further explained in appendix A.3.3, for linear chains Viterbi recursion reads:

$$\delta_j(t) = \max_{u<t} \delta^o_{j-1}(u)\overline{D}_j(t-u) \prod_{v=u+1}^{t} b_j(o_v),$$
$$\delta^o_j(t) = \max_{u<t} \delta^o_{j-1}(u)d_j(t-u) \prod_{v=u+1}^{t} b_j(o_v),$$

where d_j is the pmf and $\overline{D}_j$ is the survivor distribution of the occupancy distribution D_j of state j. Consider chains A and B as defined in section 3.2. We have to show that $\delta_k(t)$ are identical on both chains for $k \neq j$.

Step 1. If $k < j$, then $\delta_k(t)$ are identical in both chains as the paths ending in state k do not cross state j.

Step 2. If $k = j + 1$, in chain A, applying 2 times the recursion formula gives

$$\delta_{j+1}(T) = \max_{t<T} \delta^o_j(t)\overline{D}_{j+1}(T-t) \prod_{n=t+1}^{T} b_j(o_n),$$
$$\delta^o_j(t) = \max_{u<t} \delta^o_{j-1}(t-u)d_j(u) \prod_{n=t-u+1}^{t} b_j(o_n),$$

whereas in chain B, applying it 3 times gives

$$\delta_{j+1}(T) = \max_{t<T} \delta^o_{j_2}(t)\overline{D}_{j+1}(T-t) \prod_{n=t+1}^{T} b_j(o_n),$$
$$\delta^o_{j_2}(t) = \max_{v<t} \delta^o_{j_1}(t-v)d_{j_2}(v) \prod_{n=t-v+1}^{v} b_j(o_n),$$
$$\delta^o_{j_1}(t-v) = \max_{w<v} \delta^o_{j-1}(t-v-w)d_{j_1}(w) \prod_{n=t-v-w+1}^{t-v} b_j(o_n),$$

so,combining the two formulas above,

$$\delta^o_{j_2}(t) = \max_{v<t} \max_{w<v} d_{j_1}(w)d_{j_2}(v)\delta^o_{j-1}(t-(v+w)) \underbrace{\prod_{n=t-v-w+1}^{t-v} b_j(o_n) \prod_{n=t-v+1}^{t} b_j(o_n)}_{=\prod_{n=t-(v+w)+1}^{t} b_j(o_n)}$$

applying the change of variable $(v, w) \mapsto (v, u)$ with $u = v + w$

$$\delta^o_{j_2}(t) = \max_{u<t} \max_{v<u} d_{j_1}(u-v) d_{j_2}(v) \delta^o_{j-1}(t-u) \prod_{n=t-u+1}^{t} b_j(o_n)$$

$$\delta^o_j(t) = \max_{u<t} \left\{ \max_{v<u} d_{j_1}(u-v) d_{j_2}(v) \right\} \delta^o_{j-1}(t-u) \prod_{n=t-u+1}^{t} b_j(o_n).$$

The two quantities δ_{j+1} are equal if and only if the quantities δ^o_j (chain A) and $\delta^o_{j_2}$ (chain B) are equal. By equations above, this is equivalent to the condition:

$$\forall u \in \mathbb{N}, \quad d_j(u) = \max_{v<u} d_{j_1}(u-v) d_{j_2}(v). \tag{3.6}$$

Step 3. If δ_{j+1} are equal in both chains, by recursion, so are δ_{j+2}, then $\delta_{j+3}, \ldots$. So our reasoning is over: quantities δ_j are equal if and only if equation (3.6) is valid. This ends the proof of the proposition.

Proof of Proposition 3.8

This section proves Proposition 3.8 that characterize coherent chains for the Viterbi estimation. It requires a few tools from convex analysis. Indeed, inf-convolution $\star_{\inf}$ is strongly related with the Legendre-Fenchel transform $\mathcal{F}$. This operator transforms $\star_{\inf}$ into an addition $+$. We briefly introduce this operator and its basic properties, and refer to (Bauschke and Combettes, 2011) for the required background.

Definition 3.10 (Bauschke and Combettes 2011, Chapter 13)**.** A function $f : A \subset \mathbb{R} \longrightarrow]-\infty, +\infty]$ is said to be *proper* if there exists $x \in A$ such that $f(x) \neq \infty$.

The *Legendre-Fenchel transform* ($\mathcal{F}$) is the operator that associates to any proper function $f : A \subset \mathbb{R} \longrightarrow]-\infty, +\infty]$ the convex function $\mathcal{F}[f]$ defined on $\mathrm{conv}(A)$, the convex closure of A, by

$$\begin{array}{rccl} \mathcal{F}[f] : & \mathrm{conv}(A) & \longrightarrow &]-\infty, +\infty] \\ & x & \longmapsto & \sup_{t \in A} x\, t - f(t). \end{array}$$

Lemma 3.11 (Properties of $\mathcal{F}$, *ibid.*)**.** (i) For any proper function f, $\mathcal{F}\mathcal{F}f = \mathrm{cl}\,\mathrm{conv}\, f$, the closed convex envelope of f. It is the greatest closed convex function that minors f.

(ii) For any proper function f and positive number $a > 0$,

$$\mathcal{F}[a\, f](x) = a\, \mathcal{F}[f](x/a).$$

(iii) for any proper functions f, g,

$$\mathcal{F}[f \star_{\inf} g] = \mathcal{F}[f] + \mathcal{F}[g].$$

Proof of Proposition 3.8. [Claim (i)] Let $(D_l)_{l \in L}$ be a Viterbi aggregate-coherent family:

$$\forall l_1, l_2 \in L, \qquad -\log d_{l_1+l_2} = (-\log d_{l_1}) \star_{\inf} (-\log d_{l_2}). \tag{3.7}$$

The first key argument is provided by Legendre-Fenchel transform $\mathcal{F}$. Applying $\mathcal{F}$ to equation (3.7) and using claim (iii) of Lemma 3.11 gives

$$\forall l_1, l_2 \in L, \qquad \mathcal{F}[-\log d_{l_1+l_2}] = \mathcal{F}[-\log d_{l_1}] + \mathcal{F}[-\log d_{l_2}].$$

By hypothesis, $\{l_0, 2l_0, 3l_0, \ldots\} \subset L$ for some $l_0 > 0$. Assume for convenience that $l_0 = 1$. Then, last equality implies

$$\forall l \in \mathbb{N}, \quad \mathcal{F}[-\log d_l] = l\,\mathcal{F}[-\log d_1].$$

Define $c_l \overset{\text{def}}{=} \operatorname{cl}\operatorname{conv}[-\log d_l] = \mathcal{F}\mathcal{F}[-\log d_l]$. Applying again $\mathcal{F}$ and using claims (i-ii) of Lemma 3.11 gives

$$\forall l \in \mathbb{N}, t \in \mathbb{R}_+, \quad c_l(t) = l\,c_1(t/l). \tag{3.8}$$

The second key argument is a study of the supremum of d_l.

Step 1. For a general function f, $\mathcal{F}[f](0) = \sup_{\mathbb{R}} -f$ so $\mathcal{F}[-\log d_l](0) = \sup_{\mathbb{R}_+} \log d_l$. As in our case $\mathcal{F}[-\log d_l](0) = l\,\mathcal{F}[-\log d_1](0)$, we deduce that

$$\forall l \in \mathbb{N}, \quad \sup d_l = (\sup d_1)^l\,.$$

Step 2. By definition of convex envelope, $c_l \leq -\log d_l$. Combining this with equation (3.8),

$$\begin{aligned} \forall l \in \mathbb{N}, t \in \mathbb{R}_+, \quad l\,c_1(t/l) &\leq -\log d_l(t), \\ \exp -l c_1(t/l) &\geq d_l(t). \end{aligned}$$

Integrating this relationship on t gives

$$\forall l \in \mathbb{N}, \quad \int_{\mathbb{R}_+} (\exp -c_1(t/l))^l \,\mathrm{d}t \geq \int_{\mathbb{R}_+} d_l(t)\,\mathrm{d}t.$$

Applying the change of variable $t \leftarrow t/l$,

$$\forall l \in \mathbb{N}, \quad \int_{\mathbb{R}_+} l\,(\exp -c_1(t))^l \,\mathrm{d}t \geq \int_{\mathbb{R}_+} d_l(t)\,\mathrm{d}t.$$

Since we have assume that d_l are normalized probability densities, $\int d_l = 1$ for all l in L. Therefore,

$$\forall l \in \mathbb{N}, \quad \int_{\mathbb{R}_+} (\exp -c_1)^l \geq \frac{1}{l}.$$

Define the function $e_1 \overset{\text{def}}{=} \exp -c_1$ and l-norms $\|f\|_l \overset{\text{def}}{=} \left(\int |f|^l\right)^{\frac{1}{l}}$. Last equation reads:

$$\forall l \in \mathbb{N}, \quad \|e_1\|_l \geq l^{-\frac{1}{l}}.$$

It is well-known that as l goes to ∞, the l-norms $\|.\|_l$ converge to the essential supremum $\operatorname{ess\,sup}|.|$:

$$\operatorname{ess\,sup}|f| \overset{\text{def}}{=} \sup_{x \in \operatorname{ess\,supp}[f]} |f|(x).$$

So taking the limit in last equation gives $\operatorname{ess\,sup}|e_1| \geq 1$. Since the supremum is always greater than the essential supremum, $\sup|e_1| \geq \operatorname{ess\,sup}|e_1| \geq 1$. Since $e_1 \geq 0$, one obtains $\sup e_1 \geq 1$ and $\inf c_1 \leq 0$.

Step 3. In general, the infimum of a function and its closed convex envelope coincide: $\inf(\operatorname{cl}\operatorname{conv} f) = \inf f$. Indeed, the constant function $\tilde{f} \equiv \inf f$ is a closed convex minorant of f, $\tilde{f} \leq f$. So by definition of the greatest minorant, $\tilde{f} \leq \operatorname{cl}\operatorname{conv} f \leq f$. Taking the infimum in this inequality gives the result. In our case, this gives $\inf[-\log d_1] = \inf c_1$. As $\inf c_1 \leq 0$, $\sup d_1 \geq 1$ and

$$\forall l \in L, \quad \sup d_l \geq 1.$$

Step 4. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence such that $\lim_{n\to\infty} d_l(a_n) = \sup d_l$. Such a sequence exists by definition of the supremum. Equation (3.7) gives $d_{2l}(t+a) \geq d_l(t)d_l(a)$ for all $a, t \geq 0$. Integrating the relationship $d_{2l}(t + a_n) \geq d_l(t)d_l(a_n)$ gives $\int_{a_n}^{\infty} d_{2l} \geq d_l(a_n) \int_0^{\infty} d_l$. As d_l is a pdf, $\int_0^{\infty} d_l = 1$. As d_{2l} is a pdf, $\int_{a_n}^{\infty} d_{2l} \leq \int_0^{\infty} d_{2l} = 1$. So one obtains

$$\forall n \in \mathbb{N}, \quad 1 \geq d_l(a_n).$$

And therefore, $1 \geq \sup d_l$. As we already have the reverse inequality,

$$\forall l \in L, \quad \sup d_l = 1.$$

Step 5. Let l be any element in L such that $\{l, 2l, 3l, \ldots\} \subset L$. So we have $\lim_n d_l(a_n) = 1$. In addition, equation (3.7) implies

$$\forall t \geq 0, \quad d_{2l}(t + a_n) \geq d_l(t).$$

Integrating respectively on $[0, t]$ and (t, ∞) gives

$$\forall t \geq 0, \quad D_{2l}(t + a_n) \geq d_l(a_n)D_l(t), \qquad \overline{D}_{2l}(t + a_n) \geq d_l(a_n)\overline{D}_l(t).$$

As D_l, D_{2l} are absolutely continuous, the cdf and the survivor distributions are continuous functions and $D_{kl}(t) = 1 - \overline{D}_{kl}(t)$. Combining this identity with the two inequalities gives:

$$0 \leq D_{2l}(t + a_n) - d_l(a_n)D_l(t) \leq d_l(a_n) - 1,$$

and as $d_l(a_n) \to 1$, we conclude that $D_{2l}(t + a_n) \to D_l(t)$. By continuity of the cdf, this implies the existence of a limit $a \geq 0$ such that $a_n \to a$ and

$$\forall t \geq 0, \quad D_{2l}(t + a) = D_l(t). \tag{3.9}$$

The third argument is a study of the supports and the essential supports of d_l and d_{2l}. Consider $ES_l := \inf \operatorname{ess\,supp}[d_l]$ and $S_l := \inf \operatorname{supp}[d_l]$. Let us prove that $a = S_l$. Since $\lim_n d_l(a_n) = 1$, we can assume that $d_l(a_n) > 0$ for all $n \in \mathbb{N}$. Equation (3.7) implies $ES_{2l} \leq ES_l + S_l$. By definition of the support, there exists a sequence $(s_n)_{n\in\mathbb{N}}$ such that $d_l(s_n) > 0$ and $\lim_n s_n = S_l$. By definition of the essential support,

$$\forall \epsilon > 0, \quad \int_{ES_l-\epsilon}^{ES_l+\epsilon} d_l(t)\,\mathrm{d}t > 0.$$

Equation (3.7) gives $d_{2l}(s_n + t) \geq d_l(s_n)d_l(t)$. Integrating this relationship,

$$\begin{aligned} \forall \epsilon > 0, \quad \int_{ES_l-\epsilon}^{ES_l+\epsilon} d_{2l}(s_n + t)\,\mathrm{d}t &> d_l(s_n) \int_{ES_l-\epsilon}^{ES_l+\epsilon} d_l(t)\,\mathrm{d}t > 0, \\ \int_{ES_l+s_n-\epsilon}^{ES_l+s_n+\epsilon} d_{2l}(t)\,\mathrm{d}t &> 0, \end{aligned}$$

therefore $ES_l + s_n \in \operatorname{ess\,supp}[d_{2l}]$ and $ES_{2l} \leq ES_l + s_n$. Taking limit in n implies $ES_{2l} \leq ES_l + S_l$. In addition, equation (3.9) trivially implies $ES_{2l} = ES_l + a$. As a result, we have $a \leq S_l$. Since $d_l(a_n) > 0$, $a_n \in \operatorname{supp}[d_l]$ so $a_b \geq S_l$. Taking limit gives $a \geq S_l$. Therefore, we have proven that $a = S_l$.

Define shifted distribution $\tilde{D} := D_l(\cdot - a)$. Since $a = S_l$, $\tilde{D}$ is supported on $\mathbb{R}_+$.. In addition, we have just proved that $D_{2l}(\cdot + 2a) = \tilde{D}_l(\cdot)$. By induction, we can easily prove that for all rational numbers $q \in \mathbb{Q}_+^*$ (such that $ql \in L$),

$$D_{ql}(\cdot + qa) = \tilde{D}_l(\cdot),$$

which gives claim (i) of the proposition.

[Claim (ii)] In addition, $\tilde{D}$ is absolutely continuous. Since $a = S_l$, we have proven that $\sup \tilde{d} = 1 = \tilde{d}(0)$. Therefore its maximum is unique and $t > 0$ implies $-\log \tilde{d}(t) > 0$. To finish with, let us prove the subadditivity. Equation (3.7) implies

$$[-\log d_l(\cdot - a)] \star_{\inf} [\log d_l(\cdot - a)] = \log d_{2l}(\cdot - 2a).$$

In addition, $d_l(\cdot - a)$ and $d_{2l}(\cdot - 2a)$ are two admissible pdfs for $\tilde{D}$, so they must be almost everywhere equal to $\tilde{d}$. By defining $f := -\log \tilde{d}$, this reads

$$\forall t, u \geq 0, \quad f(t+u) \leq f(t) + f(u) \quad \text{a.e.},$$

which is the definition of ϕ being subadditive almost everywhere. As any function almost everywhere equal to $\tilde{d}$ is a valid pdf for $\tilde{D}$, one can choose $\tilde{d}$ such that ϕ is subadditive. □

4

COHERENT MODELING OF NOMINAL DURATIONS: STATE ESTIMATION

This chapter deals with a second criterion of time-coherency. We have devised this criterion from a peculiar situation we call *non-discriminative observation*. Section 4.1 introduces the motivation, states the coherency criterion and formalizes it. Then, coherency of semi-Markov models is investigated. The result highly depends on the chosen estimation method. Whereas chapter 5 focuses on state-sequence estimation, this chapter focuses on two methods of online state estimation described in section 2.5.2, namely Forward and Viterbi estimators.

Section 4.2 reveals the incoherency of Viterbi state estimation. Sections 4.3 and onwards deal with the Forward estimation. For this estimator, coherency can be achieved but not unconditionally. Characterizing *all* coherent chains seems to be impossible for this criterion, contrary to the previous criterion in chapter 3. This is why the investigation is much longer and builds out partial conditions. We explain how the theory of *total positivity of order two* (TP_2) provides the relevant tools we are looking for. Required mathematical background on this theory is detailed in appendix B. Step-by-step, we highlight a set of particular properties on occupancy distributions that help fulfilling the coherency criterion. Many TP_2 tools such as *ageing properties* and *reliability classes* of distributions play an important roles in other domains of applied probabilities, like statistics, reliability theory, information theory or actuarial sciences. This study highlights their relevance for our modeling problem as well.

Because state distributions of semi-Markov chains are implicit, their ageing properties might be hard to establish. In chapter 3, a different criterion suggests choosing occupancy distributions $(D_l)_{l\geq 0}$ as marginal distributions of a Lévy process. At first sight, such condition is unrelated to the second criterion as it does not appear as a necessary condition. However, section 4.6 explains how this condition greatly simplifies the discussion, as it relates ageing properties of the semi-Markov chains with their underlying Lévy process.

4.1 TIME-COHERENCY OF INFERENCE UNDER NON-DISCRIMINATIVE OBSERVATION

Our definition of time-coherency is inspired by the following realistic situation: on Figure 4.1, the music score features a long sequential repetition of the same note. Thus, a pitch-based observation model like ours would not help discriminating between two states of such sequence. Our coherency criterion is as follows: in absence of information coming form the observation, progression of the estimated states should exactly match

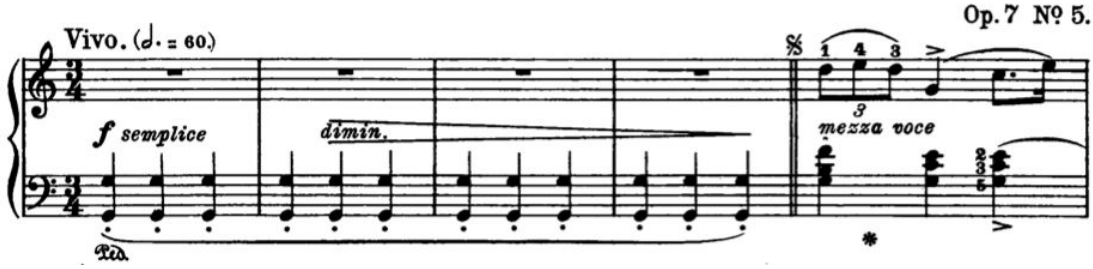


Figure 4.1: Music score of the Mazurka Op. 7 No. 5 by Frédéric Chopin. It begins with a long sequence of repeated events, *i.e.*, events with identical observation probabilities.

nominal durations indicated on the music score. We call *nominal performance* such a progression. Indeed, this information is the only available one.

Despite being obvious, this criterion is difficult to achieve with discrete models. We claim that without restrictions on their design, usual estimation algorithms are unlikely to respect it for any possible music score. To get insights on the problem, we choose to give it a mathematical formalization. The core of this chapter is to study how some theoretical hypotheses provide guarantees on the criterion validity. They can be used as prescriptions on the design of semi-Markov models for alignment.

Our base idea is to study inference with special case of music scores where *all* events share the same observation probabilities.

Definition 4.1. We refer to a model as having *non-discriminative observation* if all hidden states $j \in E$ have the same observation probability distribution: $b_1 = b_2 = \dots$.

In the case of probabilistic models, the non-discriminative assumption has a simple interpretation: inferring posterior probabilities of hidden state consists in propagating its *prior* probabilities.

Proposition 4.2. If the observation model is non-discriminative, then the posterior probabilities are equal to the prior probabilities:

$$\forall t \in \mathbb{N}^*, \quad \mathbb{P}(S_1, \dots, S_t \mid O_1, \dots, O_t) = \mathbb{P}(S_1, \dots, S_t).$$

Proof. Indeed, the Markovian assumption implies that $\mathbb{P}(O_1^t \mid S_1^t) = \prod_{u=1}^t \mathbb{P}(O_u \mid S_u) = \prod_{u=1}^t b_{S_u}(O_u)$. Assuming that $b_{S_u} = b$ gives $\mathbb{P}(O_1^t \mid S_1^t) = \prod_{u=1}^t b(O_u) = \mathbb{P}(O_1^t)$, so (S_t) and (O_t) are independent. □

With non-discriminative observation, inference becomes independent of observations. Therefore, coherency of an inference algorithm depends on only two design aspects:

- the prior evolution model $\mathbb{P}(S_1, \dots, S_t)$ of the hidden state process (S_t) — occupancy distributions D_j in our semi-Markovian setting,
- the estimation method that decodes hidden states $\hat{s}_1, \hat{s}_2, \dots$.

4.1.1 STATEMENT OF COHERENCY CRITERION

Our criterion is as follows: if the observation probabilities do not discriminate events, the successively estimated hidden states $S_1, S_2, \dots$ should match the nominal performance

indicated on the music score. In our context, a music score provides information on event ordering and nominal duration l_j of each event.

Coherency criterion 2. If observation is non-discriminative, estimation of hidden states $S_1, S_2, \ldots$ successively decodes states $1, 2, 3, \ldots$ at time steps $0, l_1, l_1 + l_2, \ldots$. In other words,

(i) states of the space E are estimated with respect to their ordering,

(ii) each state $j \in E$ is estimated for a duration which equals nominal duration l_j.

Remark. As this chapter exclusively deals with discrete-time models, the criterion formulation implicitly assumes that nominal durations l_j are integers. State space may be finite ($E = \{1, \ldots, N\}$) or infinite ($E = \mathbb{N}^*$). In addition, it accounts for the convention that first realization of the hidden process S_t starts at $t = 0$.

Denoting $\mathbf{C}(t)$ the chosen estimator, criterion 2 mathematically reads:

$$\forall j \in E \quad \text{mode}[\mathbf{C}(t)] = j \quad \Longleftrightarrow \quad 0 \leq t - (l_1 + \ldots + l_{j-1}) < l_j.$$

This chapter focuses exclusively on the two classifiers introduced in section 2.5.2. Next section deals with the Viterbi state estimation, while the remaining section deals with the Forward estimation.

4.2 INCOHERENCY OF VITERBI STATE ESTIMATION

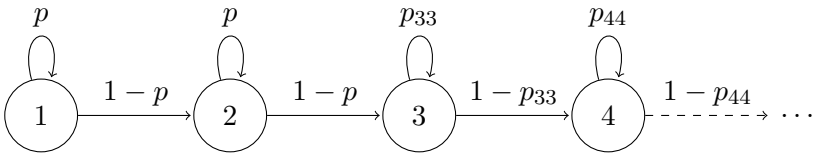


Figure 4.2: Example of a linear Markov chain with identical first two states: $p_{11} = p_{22} = p$.

This section investigates the validity of criterion 2 when estimation of current state S_t is performed online with the Viterbi classifier $C(t) = \delta(t)$, which has been introduced in section 2.5.2 and is further detailed in appendix A.2.2.

RESULT. The Viterbi classifier turns out to be *incoherent for state estimation*. Indeed, as time t grows, the estimated state $\hat{s}_t$ evolves with a fashion that is unaware of nominal durations. In the music alignment literature, we have found only two mentions about this situation of non-discriminative observation. Raphael (1999) describes the example of a linear chain where Markov states have all a self-transition probability of exactly 1/2. In this case, the Viterbi classifier gives all states the same probability. Joder (2011, Section 3.4.1) explains how on Markov chains the Viterbi classifier favors too much the state with the biggest self-transition probability. We give an example of this behavior before generalizing it.

CASE STUDY: FIRST TWO STATES. We illustrate how Viterbi current state estimation fails to be time-coherent with a toy example depicted in figure 4.2: a linear Markov chain with identical first two states. Let $\mathbf{S} = (S_1, \dots, S_{t+1})$ be an admissible path. If $\mathbf{S}$ ends at $S_{t+1} = 1$, then $\mathbb{P}(\mathbf{S}) = p^t$. If $S_{t+1} = 2$, then $\mathbb{P}(\mathbf{S}) = (1-p)p^{t-1}$. This gives $\delta_1(t) = p^{t-1}$, $\delta_2(t) = (1-p)p^{t-2} = (1-p)/p\delta_1(t)$. As a result, if $p > 1/2$ then state 1 is more likely than state 2 at all times for the Viterbi estimation, whereas if $p < 1/2$ state 2 is more likely than state 1 at all times $t > 1$. This simplistic is enough to reveal the lack of time-coherency: the parameters cannot be tuned to model a specific nominal duration l_1 and respect criterion 2.

RESULT FOR GENERAL MARKOV CHAINS. Next proposition states the asymptotic behavior of the Viterbi inference according to the self-transition probabilities p_{jj} for any linear Markov chain with non-discriminative observation. This behavior shows the lack of time-coherency: the estimation $\hat{s}_t$ either sticks on the same state, or either skips all states successively the faster it can.

Proposition 4.3. Let us consider a linear Markov chain on $E = \{1, \dots, J\}$ with $J \in \mathbb{N}^* \cup \{\infty\}$. Assume observation is non-discriminative.

(i) If $p_{jj} < 1/2$ for all $j \in E$ then for all time t the Viterbi inference decodes state t (and $J = \infty$):

$$\forall t \in \mathbb{N}^*, \quad \text{mode}[\delta(t+1)] = t.$$

(ii) Else, the Viterbi inference decodes the same state $j^* := \min\{j \in E \mid p_{jj} = \max_{i \in E} p_{ii}\}$ after some time t_0:

$$\exists t_0 \in \mathbb{N}, \quad \forall t \in \mathbb{N}, \quad t \geq t_0 \implies \text{mode}[\delta(t)] = j^*.$$

Proof. [(i)] Assume that $p_{jj} < 1/2$ for all $j \in E$. Then $1 - p_{jj} > p_{ji}$ for all $i, j \in E$. In other words, leaving a state is more likely that staying on it. Consequently, for all $j \in E$, $\delta_j(j+1) = \prod_{i=1}^{j}(1 - p_{ii}) > \delta_i(j+1)$ for all $i \neq j$. This gives the result with $t = j$.

[(ii)] Assume that $p_{ii} > 1/2$ for some $i \in E$. For linear Markov chains, the log-likelihood of a path is proportional to the time spent on each state j and to $\log p_{jj}$. So $\log \delta_j(t) \underset{t \to \infty}{\sim} t \log p_{jj}$. This leads to the result. □

Remark. The result still holds for any arbitrary initial probability π and left-to-right topologies. This result could be extended to linear semi-Markov chain by replacing $1 - p_{jj}$ with hazard rates $h_j(t)$, if we assume that all $h_j(t)$ converges to limits that are strictly different from $1/2$ as t goes to infinity.

4.3 COHERENCY OF FORWARD STATE ESTIMATION

This section investigates the validity of criterion 2 when estimation of the current state S_t is performed online with the Forward classifier $C(t) = \mathbf{f}(t)$, which has been

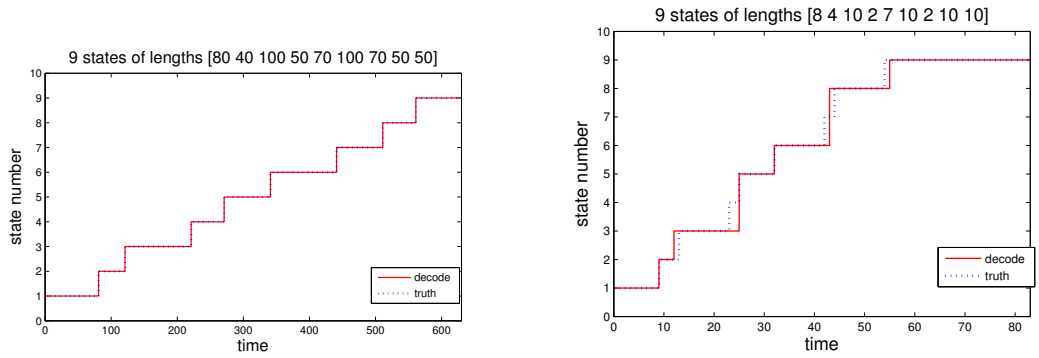


Figure 4.3: Illustration of a coherent (left) and an incoherent (right) linear semi-Markov chain. The nominal lengths l_j of their states are given on the top. Occupancy distributions are Poisson laws $D_l \sim Po(l)$. Full line shows evolution of mode$[\mathbf{f}(t)]$. Blue line shows coherent evolution with respect to criterion 2. For the chain on the right, two states are skipped and some nominal durations are not respected.

introduced in section 2.5.2 and is further detailed in appendix A.2.2. In this setting, criterion 2 reads

$$\forall j \in E, \quad \text{mode}[\mathbf{f}(t)] = j \quad \Longleftrightarrow \quad 0 < t - (l_1 + \ldots + l_{j-1}) \leq l_j.$$

Remember that Forward classifier is state probability distribution: $f_j(t) = \mathbb{P}(S_t = j)$. The semi-Markovian framework does not allow the practitioner to explicitly choose state distributions $\mathbf{f}(t) := (f_1(t), f_2(t), \ldots)$ so that they fulfill criterion 2. If the semi-Markov chain is linear, the relationship between $f_j(t)$ and D_j is straightforward as explained by Proposition 3.2 in previous chapter. We restate this result for it lies at the root of our study.

Proposition 4.4 (State probabilities of linear chains)**.** For a discrete-time and linear (in a wide sense) semi-Markov chain,

$$\begin{aligned} \forall j \in E, t \in \mathbb{N}^*, \qquad f_j(t-1) &= d_1 * \ldots * d_{j-1} * [\overline{D}_j - \delta](t) \\ &= \sum_{u=0}^{t-1} [d_1 * \ldots * d_{j-1}](u) \overline{D}_j(t-u). \end{aligned}$$

Remark. This study is undertaken exclusively for discrete-time HSMMs. For a discrete distribution D_j, d_j denote its probability mass function and $\overline{D}$ its survivor distribution. Both quantities are seen sequences supported on $\mathbb{N}$. Refer to appendix A.1 for more mathematical background.

One conclusion of the forthcoming study is that Forward inference *may* be time-coherent, but not unconditionally with all semi-Markov chains. Firstly, not all families of occupancy distributions $(D_l)_{l \in L}$ lead to a coherent inference. Secondly, given this family, not all configurations of state duration $(l_j)_{j \in E} = (l_1, l_2, \ldots)$ lead to coherency. It means that only certain music scores are fully coherent, whereas the other ones are only approximately coherent at best. This fact is illustrated by figure 4.3. Consequently, the methodology of this study is to look for sufficient or necessary conditions on occupancy distributions D_{l_j} and durations l_j that imply parts of criterion 2. We build our conditions step-by-step by introducing each mathematical tool when it is relevant for our discussion.

4.4 CASE STUDY: COHERENCY ON 2-STATE CHAIN

At initial time $t = 0$, state 1 is the most likely. Then, we expect that after some duration, state 2 becomes the most likely. So we begin our investigation a comparison between state probabilities f_1, f_2 of states 1 and 2. This case study illustrates the more general study that is undertaken in next section since it highlights the relevant mathematical tools.

4.4.1 INTRODUCTION AND DEFINITIONS

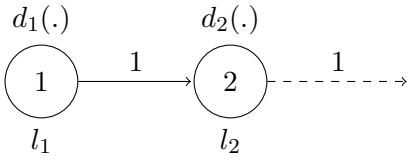


Figure 4.4: Graphical model of the first two states in a linear semi-Markov chain.

We begin with some notations that help formalizing the problem.

Definition 4.5. For two pmfs d_1, d_2, define the state probability ratio $\Delta(t; d_1, d_2)$ as

$$\Delta(t; d_1, d_2) \stackrel{\text{def}}{=} \frac{\sum_{u=0}^{t-1} d_1(u)\overline{D}_2(t-u)}{\overline{D}_1(t)}. \tag{4.1}$$

Define also the *turning point* $\mathbf{t}(d_1, d_2)$ as

$$\mathbf{t}(d_1, d_2) \stackrel{\text{def}}{=} \min\left[\{t \in \mathbb{N} \mid \Delta(t+1, d_1, d_2) > 1\} \cup \{+\infty\}\right].$$

When considering a family $(d_l)_{l\in L}$ such that $d_1 = d_{l_1}$ and $d_2 = d_{l_2}$, we use notations

$$\Delta(t; l_1, l_2) \stackrel{\text{def}}{=} \Delta(d_{l_1}, d_{l_2}), \qquad \mathbf{t}(l_1, l_2) \stackrel{\text{def}}{=} \mathbf{t}(d_{l_1}, d_{l_2}),$$

that define functions $\Delta : \mathbb{N} \times L \times L \to [0, \infty]$ and $\mathbf{t} : L \times L \to [0, \infty]$.

Remark. Δ is the probability ratio of state probabilities: $\Delta(t+1; d_1, d_2) = \frac{f_2(t)}{f_1(t)}$. Note the definition of Δ covers the case of occupancy distributions d_1 such that $d_1(0) \neq 0$. The turning point $\mathbf{t}$ is the first time where state 2 is decoded instead of state 1.

Now with such definitions, the coherency criterion 2 for the first two states reads as follows.

Coherency criterion 2 (case of 2-state chain, continuing from p. 73)**.**

$$\forall t \in \mathbb{N}, \qquad \Delta(t; l_1, l_2) \begin{cases} \leq 1 & \text{if } t < l_1 \\ > 1 & \text{if } l_1 < t \leq l_1 + l_2 \end{cases}$$

and in particular:

$$\mathbf{t}(l_1, l_2) = l_1.$$

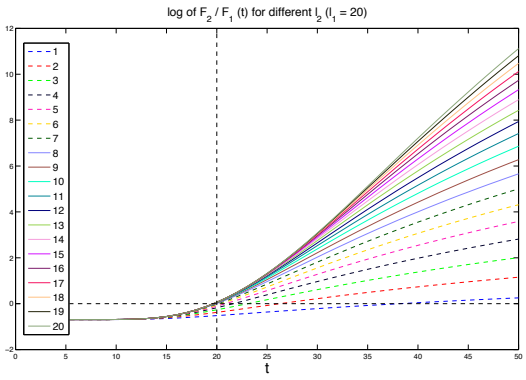


Figure 4.5: Numerical illustration of state probability ratios when occupancy distributions are Poisson laws $D_l \sim Po(l)$. Curves $t \mapsto \Delta(t, l_1, l_2)$ are sketched for $l_1 = 10$ and different values of l_2. They reveal a monotony in l_2 and t.

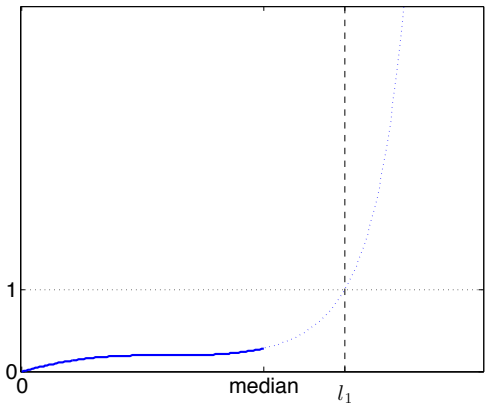


Figure 4.6: Behavior of $f_2(t)/f_1(t)$ for small durations t. The median bound guarantees state 2 is not decoded sooner than median$[d_1]$.

NUMERICAL EXAMPLES. To get insights on the validity of the criterion, we study it numerically with a peculiar choice of occupancy distributions d_l. We choose occupancy distributions $D_j = D_{l_j}$ as Poisson laws parametrized by nominal duration: $D_l \sim Po(l)$. Then, the idea is to compute several curves of ratios $\Delta(t, l_1, l_2)$ and numerically check whether they fulfill the coherency criterion 2. Figure 4.5 depicts the result. This example turns to be illuminating as it exhibits several interesting phenomena:

- the coherency criterion seems to be fulfilled when l_2 is large enough compared to l_1.
- the turning point $\mathbf{t}(l_1, l_2)$ is always greater than l_1.
- the probability ratios $\Delta(t, l_1, l_2)$ are monotone with respect to time t and duration of the second state l_2.

Now, our study aims at explaining such phenomena for more general families of occupancy distributions. As no explicit formula is available for Δ and $\mathbf{t}$ in the general case, we build up our study by proving partial conditions.

4.4.2 BOUND ON TIME: THE MEDIAN

We explain that whether this criterion is fulfilled depends on the *median* of occupancy distributions. It provides with a universal lower bound on the time before state 2 is decoded.

Definition 4.6. The *median* of a probability distribution d on $\mathbb{R}$ is defined by

$$\text{median}[d] \stackrel{\text{def}}{=} \max\{t \in \mathbb{R} \mid \overline{D}(t) \geq 1/2\}.$$

Proposition 4.7. Let D_1 be a probability distribution.
For all probability distributions D_2,

$$\mathbf{t}(d_1, d_2) \geq \text{median}[d_1].$$

Proof. Since the inequality $\overline{D}(t) \leq 1$ holds for all probability distribution d and time t, we can simply write

$$\sum_{u<t} d_{l_1}(u)\overline{D}_{l_2}(t-u) \leq \sum_{u<t} d_{l_1}(u) = 1 - \overline{D}_{l_1}(t)$$

Thus $f_2(t-1) - f_1(t-1) \leq 1 - 2\overline{D}_{l_1}(t)$, and $\overline{D}_{l_1}(t) \geq \frac{1}{2}$ implies $f_2(t-1) \leq f_1(t-1)$. □

Last proposition tells that state 1 is more likely for small times. This inequality guarantees that state 1 is more likely when $t < \text{median}[d_1]$. Figure 4.6 illustrates this first necessary condition of coherency. On addition, one can show this bound is tight in the sense that there exist distribution d_2 such that state 2 is strictly more likely at $t = \text{median}[d_1]$. As a conclusion, the coherency criterion requires that each distribution d_l has its *median equal to the nominal duration* l:

$$\text{Necessary condition:} \quad \forall l \in L, \quad \lceil l \rceil - 1 < \text{median}[d_l] - \leq \lceil l \lceil. \tag{4.2}$$

4.4.3 VALIDITY OF THE CRITERION AND STOCHASTIC ORDERS

The median condition in Equation (4.2) ensures that state 2 cannot be decoded before $t = l_1$. But some problematic situation could occur afterwards. For instance, the most likely state could oscillate between 1 and 2 shortly after time $t = l_1$. This situation is depicted in figure 4.7. A sufficient condition that prevents this situation is to require the probability ratio $\Delta(t+1) = \frac{f_2}{f_1}(t)$ increases over time t.

This condition is interesting as it is related to *stochastic orders*, which are important tools of a general theory called *total positivity of order two* (TP_2) — appendix B provides definitions and a full presentation. The actual relationship with orders turns out to be twofold. Firstly, the monotony of $\Delta(t)$ corresponds to the *likelihood ratio order* between state distributions $\mathbf{f}(t)$. This is more comprehensively explained later on for the N-state case. Secondly, the monotony is equivalent to the *hazard rate order* between occupancy distributions D_l of the chain, as next proposition explains.

Proposition 4.8. Let d_1, d_2 be two pmfs.
$\frac{f_2(t)}{f_1(t)}$ is a non-decreasing function of t if and only if $d_1 * d_2$ is greater than d_1 in the hazard rate ordering: $d_1 \underset{\text{hr}}{\leqslant} d_1 * d_2$.

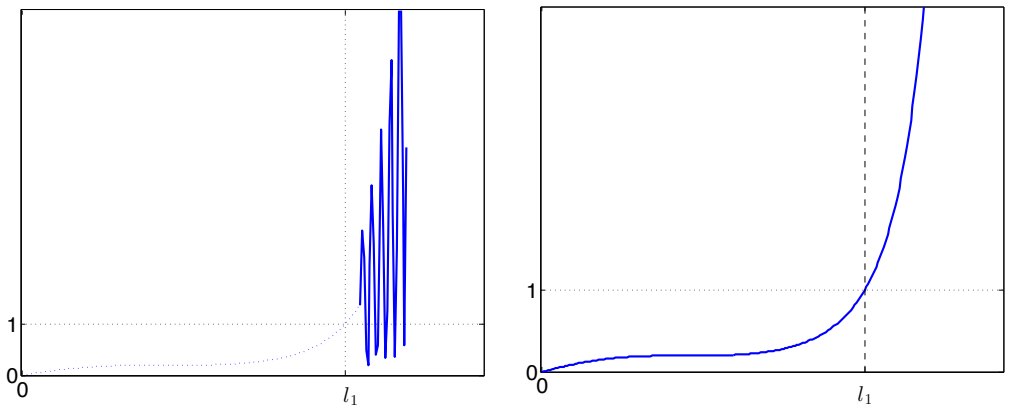


Figure 4.7: Left, example of a non-desirable curve for $t \mapsto f_1(t)/f_2(t)$ because the decoded state oscillates between 1 and 2. Right, example of a desirable curve: the monotony ensures the turning point $\mathbf{t}$ is unique.

This fact comes from the definition of *hr* order in appendix B.3.

Proof. The survivor distribution $\overline{D}$ associated to the convolution $d_1 * d_2$ is

$$\overline{D}(t) = \sum_{u=0}^{t-1} d_1(u)\overline{D}_2(t-u) + \overline{D}_1(t),$$

as explained in section A.1. Therefore,

$$\frac{\overline{D}(t)}{\overline{D}_1(t)} = \frac{\overline{D}_1(t) + \sum_{u=0}^{t-1} d_1(t-u)\overline{D}_2(u)}{\overline{D}_1(t)} = 1 + \frac{\sum_{u=0}^{t-1} d_1(t-u)\overline{D}_2(u)}{\overline{D}_1(t)} = 1 + \frac{f_2(t-1)}{f_1(t-1)}.$$

So $\frac{f_2(t)}{f_1(t)}$ is non-decreasing if and only if $\frac{\overline{D}(t^+)}{\overline{D}_1(t^+)}$ is, which means $d_1 \underset{\text{hr}}{\leqslant} d_1 * d_2$. □

TP_2 theory does not only provide stochastic orders. It also provides a distributional property that automatically ensures the *hr* ordering. Such property is called *Increasing Hazard Rate*. We only quote its definition here and refer to appendix B.2.2 for full background.

Definition 4.9. A discrete probability distribution d on $\mathbb{N}$ is said to be *discrete IHR* if its survivor distribution $\overline{D}(n) := \sum_{k \geq n} d(k)$ checks $\forall n \in \mathbb{N}, \overline{D}(n+1)^2 \geq \overline{D}(n)\overline{D}(n+2)$.

Next proposition is a basic result of TP_2 theory — see Proposition B.29. It underlines the interest of IHR distributions.

Proposition 4.10 (Sufficient conditions of ordering)**.** Let d_1 be a pmf on $\mathbb{N}$.
If d_1 is discrete IHR then $d_1 \underset{\text{hr}}{\leqslant} d_1 * d_2$ for *all* pmf d_2.

If probability ratio $\Delta(t)$ increases with t, then it converges to a limit. Next proposition provides an explicit expression which involves the Z-transform of probability distributions. It becomes easy to check if state 2 will be never decoded (when the limit is ≤ 1), or if it will be after some time.

Proposition 4.11 (Forward ratio monotony)**.** Let d_1, d_2 be two pmfs such that d_1 is discrete IHR. Then, $t \mapsto \Delta(t, d_1, d_2)$ is non-decreasing and converges to $Z[d_2](R_1) - 1 \in (0, \infty]$, where $R_1 \in (1, \infty]$ is the radius of convergence of $Z[d_1](z)$.

- If $Z[d_2](R_1) \leq 2$ then state 2 is never decoded: $\mathbf{t}(d_{l_1}, d_{l_2}) = \infty$.
- Else, $\mathbf{t}(d_{l_1}, d_{l_2}) < \infty$, state 1 is decoded if and only if $t < \mathbf{t}(d_{l_1}, d_{l_2})$ and state 2 is decoded if and only if $t \geq \mathbf{t}(d_{l_1}, d_{l_2})$.

Proof. Consider for all $t \in \mathbb{N}^*$ the functions $g_t : \mathbb{N} \to \mathbb{R}$ with

$$g_t(u) \overset{\text{def}}{=} \begin{cases} 0 & \text{if } u \leq t, \\ \frac{d_1(t-u)}{\overline{D}_1(t)} \overline{D}_2(u) & \text{if } u > t, \end{cases}$$

so that one has $\frac{f_2(t)}{f_1(t)} = \sum_{u=0}^{\infty} g_{t+1}(u)$.

If d_1 is discrete IHR, then proposition B.25 implies that for all $u \in \mathbb{N}$, $t \mapsto \frac{d_1(t-u)}{\overline{D}_1(t)}$ is non-decreasing. Moreover, one can show that such ratio converges to $(1 - p_1)\left(\frac{1}{p_1}\right)^u$ with $p_1 := 1/R_1$, and that $R_1 > 1$. So the sequence of functions $(g_t)_{t \in \mathbb{N}}$ converges monotonically to $u \mapsto (1 - p_1)\overline{D}_2(u)\left(\frac{1}{p_1}\right)^u$. We can apply the monotone convergence theorem to exchange limit and summation:

$$\lim_{t \to \infty} \sum_{u=0}^{\infty} g_t(u) = \sum_{u=0}^{\infty} \lim_{t \to \infty} g_t(u) = (1 - p_1) \sum_{u=1}^{t-1} \overline{D}_2(u) \left(\frac{1}{p_1}\right)^u,$$

using the Z-transform of survivor distribution $\overline{D}_2$,

$$= (1 - p_1)(Z[\overline{D}_2](1/p_1) - 1).$$

According to appendix A.1, this Z-transform equals

$$Z[\overline{D}](z) - 1 = \begin{cases} \frac{z}{1-z}(1 - Z[d](z)) & \text{if } z \neq 1, \\ \sum_{n=0}^{\infty} n\, d(n) = \mathbb{E}[d] & \text{if } z = 1, \end{cases}$$

so $(1 - p_1)Z[\overline{D}_2](1/p_1) = Z[d_2](1/p_1) - 1$. Finally,

$$\lim_{t \to \infty} \frac{f_2(t)}{f_1(t)} = Z[d_2](1/p_1) - 1.$$

This proves the first part of proposition. As D_1 is discrete IHR, it has been shown that $\frac{f_2(t)}{f_1(t)}$ increases with t. This proves the second part of the proposition. □

This section argues in favor of choosing occupancy distributions among the IHR class. We think this brings some justification to many common engineer choices like Poisson, Negative Binomial, Gamma, Normal laws which turn to be IHR.

$$\text{Sufficient condition:} \quad \forall l \in L, D_l \text{ is discrete IHR.}$$

4.4.4 BOUNDS ON EVENT DURATION

Previous results deal with two given distributions D_1, D_2. When $D_1 = D_{l_1}$, $D_2 = D_{l_2}$ belong to some duration-indexed family $(D_l)_{l\in L}$ we may wonder how the criterion is influenced by nominal durations l_1, l_2.

EXISTENCE OF BOUNDS. With mild conditions on occupancy distributions, we are able to characterize all configurations (l_1, l_2) that are time-coherent. Indeed, the monotony of probability ratio $\Delta(t; l_1, l_2)$ and turning point $\mathbf{t}(l_1, l_2)$ with respect to event durations l_1, l_2 can be related to stochastic orderings: $\Delta(t; d_{l_1}, d_{l_2})$ is non-decreasing with l_1 in the hazard rate ordering hr, and non-increasing with l_2 in the basic stochastic ordering st.

Proposition 4.12. Let d_1, d_2, d_A, d_B be pmfs.

$$d_A \underset{\text{st}}{\leqslant} d_B \quad \Longrightarrow \quad \Delta(t, d_1, d_A) \geq \Delta(t, d_1, d_B) \quad \Longrightarrow \quad \mathbf{t}(d_1, d_A) \leq \mathbf{t}(d_1, d_B).$$
$$d_A \underset{\text{hr}}{\leqslant} d_B \quad \Longrightarrow \quad \Delta(t, d_A, d_1) \geq \Delta(t, d_B, d_2) \quad \Longrightarrow \quad \mathbf{t}(d_A, d_2) \leq \mathbf{t}(d_B, d_2).$$

Proof. The proof is straightforward. Indeed, $d_A \underset{\text{st}}{\leqslant} d_B$ means that

$$\forall t \in \mathbb{N}, \forall u \leq t, \quad \overline{D}_A(t-u) \leqslant \overline{D}_B(t-u).$$

This implies

$$\forall t \in \mathbb{N}, \quad \sum_{u=0}^{t-1} \frac{d_1(u)}{\overline{D}_1(t)} \overline{D}_A(t-u) \leq \sum_{u=0}^{t-1} \frac{d_1(u)}{\overline{D}_1(t)} D_B(t-u).$$

Similarly, if $d_A \underset{\text{hr}}{\leqslant} d_B$, then lemma B.25 gives

$$\forall t \in \mathbb{N}, \forall u \leq t, \quad \frac{d_A(t-u)}{\overline{D}_A(t)} \leq \frac{d_B(t-u)}{\overline{D}_B(t)}.$$

Therefore,

$$\frac{d_A * \overline{D}_2}{\overline{D}_A}(t) \leq \frac{d_B * \overline{D}_2}{\overline{D}_B}(t).$$

□

Such monotony immediately provides the existence of bounds.

Proposition 4.13. Assume the family $(D_l)_{l\in L}$ is non-decreasing in basic ordering st.

(i) There exists two bounding functions $L_{\min}(.)$, $L_{\max}(.) \in [0, \infty]$ such that the 2-state version of criterion 2. is true if and only if

$$L_{\min}(l_1) < l_2 \leq L_{\max}(l_1).$$

(ii) If the median of d_{l_1} coincides with l_1, then $L_{\max}(l_1) = \infty$.

Proof. *[(i)]* The 2-state criterion is equivalent to

$$\begin{cases} \forall t \in (0, l_1], & \Delta(t; l_1, l_2) \leq 1, \\ \forall t \in [l_1 + 1, l_1 + l_2 + 1], & \Delta(t; l_1, l_2) > 1. \end{cases}$$

Last proposition implies that $l_2 \mapsto \Delta(t, l_1, l_2)$ are non-decreasing in l_2. This immediatly provides the existence of bounds.

[(ii)] If $\text{median}[d_{l_1}] = l_1$, proposition 4.7 implies that $\Delta(t, l_1, l_2) \leq 1$ holds for all $t \in (0, l_1]$ with no condition on l_2. This means that $L_{\max}(l_1) = \infty$. □

When the condition on the median is fulfilled, last proposition gives the existence of a lower bound on l_2: inference cannot be coherent unless state 2 is long enough compared to state 1.

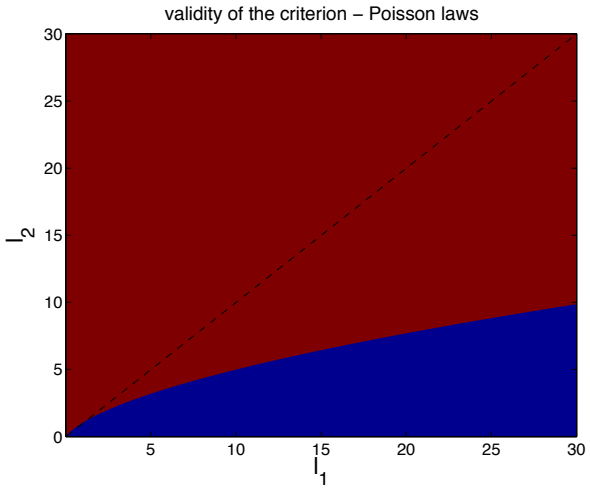


Figure 4.8: Validity of coherent nominal durations for Poisson laws $D_l \sim Po(l)$. Points (l_1, l_2) are inside the red area if they check criterion 2 and else in the blue area. The lower bound $L_{\min}(l_1)$ appears as the frontier.

NUMERICAL EXAMPLE. Figure 4.8 depicts numerical computation of bound $L_{\min}(l_1)$ when occupancy distributions are Poisson laws. This graph reveals remarkable properties. We wonder if they generalize to other convolution semigroups. Even in the Poisson case, an analytical proof is still an open question.

- is $L_{\min}(l_1)$ increasing?
- does $L_{\min}(l_1) \leq l_1$ hold?
- what is the asymptotic behavior of $L_{\min}(l_1)$ as $l_1 \to \infty$?

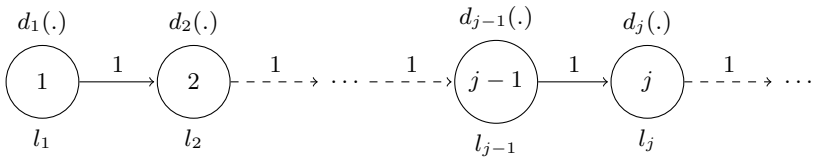


Figure 4.9: Graphical model of a general linear semi-Markov chain.

4.5 GENERALIZATION TO N-STATE CHAINS

Now, we try to extend the results obtained for the states 2 to any subsequent state j. We stick to the idea of studying probability ratios for all $j \in E$,

$$\frac{f_{j+1}(t)}{f_j(t)} = \frac{d_1 * \ldots * d_{j-1} * d_j * (\overline{D}_{j+1} - \delta)}{d_1 * \ldots * d_{j-1} * (\overline{D}_j - \delta)}.$$

4.5.1 BOUND ON TIME

Next proposition shows that the *median* of occupancy distribution d_1 gives a lower bound on the duration state 1 is decoded. It extends proposition 4.7 obtained for the 2-state case and has a similar proof. In addition, it shows that state distributions $\mathbf{f}(t)$ are non-increasing that at small times. Non-increasing distributions belongs to the class of *unimodal* distributions. Figure 4.10 illustrates both notions. Next section will develop how the notion of unimodality is conceptually interesting.

Proposition 4.14 (unimodality at small times)**.** For all $j \in E$,

$$t \leq \text{median}[d_1] \implies f_j(t) \leq f_1(t).$$

For all $j \in E$,

$$t \leq \text{median}[d_{j+1}] \implies f_{j+1}(t) \leq f_j(t).$$

As a consequence, $\mathbf{f}(t)$ is a non-increasing distribution on E for all $t \leq \min_{j \in E} \text{median}[d_j]$.

Proof. For $j > 1$, define $d_A = d_1 * \ldots * d_{j-1}$. Using the relationship $\overline{D}(n+1) = 1 - D(n)$ and the inequality $\overline{D}(n) \leq 1$, we get

$$\begin{aligned} f_j(n) &= \overline{d_A * d_j}(n+1) - \overline{d_A}(n+1) \\ &= D_A(n) - d_A * D_j(n), \end{aligned}$$

and

$$\begin{aligned} f_{j+1}(n+1) &= \overline{d_A * d_j * d_{j+1}}(n) - \overline{d_A * d_j}(n) \\ &\leq 1 - \overline{d_A \ldots d_j}(n) = d_A * D_j(n-1), \end{aligned}$$

therefore

$$\frac{f_j(n)}{f_{j+1}(n)} \geq \frac{D_A(n) - d_A * D_j(n)}{d_A * D_j(n)} = \frac{D_A(n)}{d_A * D_j(n)} - 1.$$

By definition, $n \leq \text{median}[d_j]$ implies $D_j(n-u) \leq 1/2$ for all $u = 0 \dots n$. So,

$$d_A * D_j(n) = \sum_{u=0}^{n} d_A(u) * D_j(n-u) \leq 1/2 \sum_{u=0}^{n} d_A(u) = D_A(n)/2.$$

Consequently, $\frac{D_A(n)}{d_A * D_j(n)} \geq 2$ and $f_j(n+1) \geq f_{j+1}(n+1)$. This proves the second claim. The first claim has the same proof as proposition 4.7. □

Remark. Here again, each bound is tight as it is reached for some choice of distribution D_j.

4.5.2 MONOTONE SEMI-MARKOV CHAINS

Previous section has illustrated the convenience of having a semi-Markov chain $S = (S_t)_{t\in\mathbb{N}}$ such that the ratio of first two states probabilities $f_2(t)/f_1(t)$ increases with time t, where $S_t \sim \mathbf{f}(t)$. Generalizing this idea, one could ask that the probability ratio $f_{j+1}(t)/f_j(t)$ of any pair of consecutive states $j, j+1 \in E$ increases with t. This turns out to be the exact definition of *monotone stochastic processes* in likelihood ratio ordering $\underset{\text{lr}}{\leqslant}$. Appendix B.3 provides full background on stochastic orders. We say that S is non-decreasing in the likelihood ratio ordering (lr-monotone for short) if $\mathbf{f}(t) \underset{\text{lr}}{\leqslant} \mathbf{f}(t+1)$ for all time $t \in \mathbb{N}$. Using chains with such monotony has two advantages:

1. Part (ii) of criterion 2 is easy to check numerically on a given chain: it requires a two comparisons per state, independently of final inference time T.
2. Part (i) of criterion 2 is unconditionally fulfilled: states are decoded with respect to their ordering,

Let us explain such claims. First, if a process X is lr-monotone then it fulfills criterion 2 if and only if

$$\forall j \in E, \quad \frac{f_{j+1}}{f_j}(l_1 + \ldots + l_j - 1) \leq 1 \quad \text{and} \quad \frac{f_{j+1}}{f_j}(l_1 + \ldots + l_j) > 1. \tag{4.3}$$

Such inequalities require 2 computations per state to be checked: each state only has to be compared to its successor.

The other claim comes from the following lemma whose proof is straightforward.

Lemma 4.15. If the family of distributions $(p_t)_{t\in I}$ is non-decreasing in the likelihood ratio ordering, then mode$[p_t]$ of p_t is non-decreasing with respect to t.

Such facts strongly promote lr-monotone semi-Markov chains as probabilistic model on the evolution of score position. Next two paragraphs gives supplementary reasons in favor of this property.

EXISTENCE OF BOUNDS ON EVENT DURATION. The 2-state case has revealed how stochastic orders guarantee the existence of bounds of coherency on nominal durations l_j. More specifically, the requirement is st-monotony between occupancy distributions D_l, as explained in Proposition 4.12. Obtaining the existence of such bounds in the general case requires an additional and much strogner hypothesis: lr-monotony of *all* semi-Markov chains $(l_1, l_2, l_3, \ldots)$ obtained out of $(D_l)_{l\in L}$.

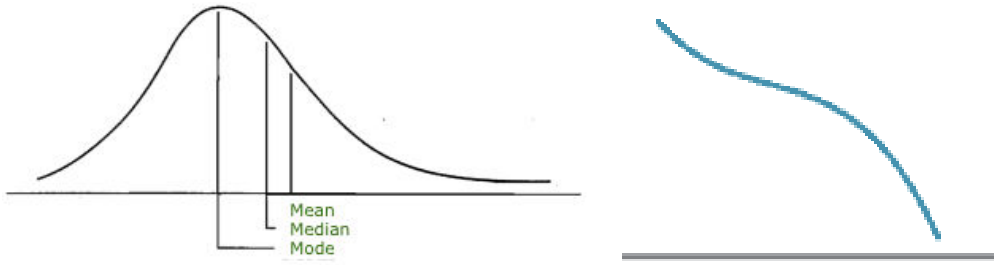


Figure 4.10: Left: an unimodal probability distribution. The *mode* is the position of the maximum. The probability decreases as we move from the mode. Right: a non-increasing probability distribution. The left-hand distribution is log-concave, while the right-hand side is not.

Proposition 4.16. Assume that

- the family $(D_l)_{l\in L}$ of occupancy distributions is *st*-monotone: if $l \leq m$, then $D_l \underset{\text{st}}{\leqslant} D_m$.
- all possible semi-Markov chains $(l_1, l_2, l_3, \ldots)$ built out of $(D_l)_{l\in L}$ are *lr*-monotone.

Then, there exists two $[0,\infty]$-valued bounding functions $L_{\min}, L_{\max}$ such that: the criterion 2 is true if and only if

$$\forall j \in E, \qquad L_{\min}(l_1,\ldots,l_j) < l_{j+1} \leq L_{\max}(l_1,\ldots,l_j).$$

Proof. For a chain that is *lr*-monotone, the criterion is equivalent to $\frac{f_{j+1}}{f_j}(l_{1:j}-1) \leq 1 < \frac{f_{j+1}}{f_j}(l_{1:j})$. Hypothesis of *lr* monotony tells the two ratios are non-increasing with respect to each quantity $\overline{D}_{l_{j+1}}(u)$. In addition, hypothesis of *st*-monotony exactly tells that such quantities are non-decreasing with l_{j+1}. So the proof becomes similar to Proposition 4.12. □

The existence of these bounds provides a procedure to obtain the set of all coherent chains, based on a pre-computation of bounds $L_{\min}, L_{\max}$. But numerically computing bounds for coherent J-state chains would require tabulating a J-dimension function. This rapidly becomes prohibitive as the total number of states J grows. Still this gives an interesting prescription on our design choices.

Sufficient condition: choose $(D_l)_{l\in L}$ such that semi-Markov chains are *lr*-monotone.

RELATIONSHIP WITH UNIMODALITY. Stochastic monotony induces a relationship with the notion of *unimodality*, which is illustrated in figure 4.10 and detailed in appendix B.4. Next lemma is a straightforward application of definitions.

Lemma 4.17. Assume the family of distributions $(P_t)_{t\in I}$ is an non-decreasing in the likelihood ratio ordering.
If $\cup_{t\in I} \operatorname{mode}[P_t] = E$, then all distributions P_t are unimodal.

Fulfilling criterion 2 implies that all states j in E are decoded at least one time. Owing to the lemma, this implies unimodality of state distributions $\mathbf{f}(t)$ at all time t. So this fact stresses out unimodality as necessary condition. It also extends the unimodal behavior at small times which has been obtained unconditionally in proposition 4.14.

Even if having unimodal state distributions is not required by our criterion, it has a very interesting interpretation: the farer a state j from the "nominal state" j_{true}, the lesser its prior probability is. This is compliant with the heuristic that bigger deviations from the nominal performance should be less likely than smaller ones.

SUFFICIENT CONDITION OF MONOTONY. So far this section has promoted semi-Markov chains that are monotone in the lr-ordering. So now, the question become: which design choices lead to lr-monotone chains? Unfortunately, characterizing all monotone semi-Markov chains seems to be out of reach. While hardly no results exist on semi-Markov chains, the monotony of Markov processes has been thoroughly studied for various stochastic orderings (Kijima, 1998). However all proofs rely on linear algebra with Markov transition matrices, so they cannot generalize to semi-Markov chains.

Next proposition is a first attempt in this direction. It tells log-concavity of occupancy distributions is a sufficient condition for the monotony. It is the only general result we have obtained holds for linear semi-Markov chains.[1] Its proof is straightforward thanks to basic tools of TP_2 theory. We just recall the definition of log-concavity and refer to appendix B.2.1 for more background on this notion.

Definition 4.18. A discrete distribution p on $\mathbb{N}$ is said to be discrete log-concave if it is nonnegative, has no internal zeros and checks: $\forall n \in \mathbb{N}^*, \quad p_{n-1}p_{n+1} \leq p_n^2$.

Next proposition is stated for discrete-time distributions only, but it also holds for continuous-time ones.

Proposition 4.19 (sufficient condition of monotony)**.** Let $S = (S_t)_{t\in\mathbb{N}}$ be a semi-Markov chain on $E \subset \mathbb{N}$. Assume S is linear (in a wide sense).
If all occupancy distributions D_j are discrete log-concave, then $(\mathbf{f}(t))_{t\in\mathbb{N}}$ is non-decreasing for the likelihood ratio ordering.

Proof. Let j be in E. Since d_j is discrete log-concave, it is also discrete IHR. This implies $\overline{D}_j - \delta \underset{\text{TP}}{\leqslant} d_j * (\overline{D}_{j+1} - \delta)$. In addition, one has

$$f_{j+1} = d_1 * \ldots * d_{j-1} * d_j * (\overline{D}_{j+1} - \delta),$$
$$f_j = d_1 * \ldots * d_{j-1} * (\overline{D}_j - \delta).$$

Proposition B.29 tells that convolution preserves log-concave distributions, so $d_1 * \ldots * d_{j-1}$ is log-concave. Therefore, Proposition B.41 implies that $f_j \underset{\text{TP}}{\leqslant} f_{j+1}$. □

The conclusion at this stage of the chapter is to choose occupancy distributions $(D_l)_{l\in L}$ as a st-monotone family of log-concave distributions.

Sufficient condition: $(D_l)_{l\in L}$ is st-monotone and for all $l \in L$, D_l is log-concave.

For its part, the study on criterion 1 in Chapter 3 has suggested choosing $(D_l)_{l\in L}$ as a convolution semigroup. At first sight this prescription is of a very different nature. Next section shows it greatly helps current discussion.

1. Later on in next section, wider characterizations are given in the case of convolution semigroups.

4.6 SPECIAL CASE OF LÉVY PROCESSES

This section shows how the discussion on criterion 2 carried on in previous section is greatly simplified if we assume that $(D_l)_{l\in L}$ is a convolution semigroup[2]. More specifically:

- A characterization of the lr-monotony of semi-Markov chains is obtained in section 4.6.2, along with other TP_2 properties.
- A limit result is derived in section 4.6.3: chains with infinitely many states cannot be coherent unless all states have the same nominal duration $l = l_1 = l_2 = \ldots$.
- As an illustration, the good behavior of two particular semigroups, the Poisson and the Negative Binomial laws, is shown in section 4.6.3.

4.6.1 REDUCTION OF N-STATE PROBLEM TO 3-STATE PROBLEM

If all occupancy distributions D_j belong to the same convolution semigroup $(D_l)_{l\in L}$, then *the N-state problem boils down to the 3-state problem*. Indeed, the semigroup property gives simple formulas for f_j and f_{j+1}:

$$\forall j \in E, \qquad \frac{f_{j+1}}{f_j}(t) = \frac{\overline{D}_{l_{1:j+1}} - \overline{D}_{l_{1:j}}}{\overline{D}_{l_{1:j}} - \overline{D}_{l_{1:j-1}}}(t+1),$$

with notation $l_{1:j} \overset{\text{def}}{=} l_1 + \ldots + l_j$. Thanks to this fact, studying quantities (f_j, f_{j+1}) in a general N-state linear chain boils down to comparing (f_2, f_3) in a 3-state linear chain where $l_1 = l_{1:j-1}, l_2 = l_j, l_3 = l_{j+1}$, as in such a chain one would have:

$$\frac{f_3}{f_2}(t) = \frac{\overline{D}_{l_1+l_2+l_3} - \overline{D}_{l_1+l_2}}{\overline{D}_{l_1+l_2} - \overline{D}_{l_1}}(t+1).$$

Figure 4.11 illustrates this equivalence.

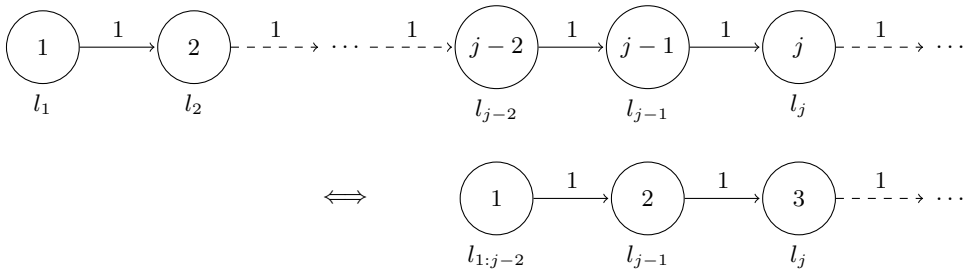


Figure 4.11: Equivalence of a N-state chain with a 3-state chain. Thanks to convolution stability, the state probabilities $f_{j-1}, f_j, \ldots$ depends only of $l_{1:j-2} = l_1 + \ldots + l_{j-2}$.

2. We recall that convolution semigroups $(D_l)_{l\in L}$ are exactly the marginal distributions of Lévy processes X, and that such distributions are infinitely divisible.

As a consequence, criterion 2 is fulfilled with a convolution semigroup $(D_l)_{l\in L}$ if and only if

$$\forall l_1, l_2, l_3 \in L, \quad \frac{\overline{D}_{l_1+l_2+l_3} - \overline{D}_{l_1+l_2}}{\overline{D}_{l_1+l_2} - \overline{D}_{l_1}}(t) \begin{cases} > 1 & \text{if } 0 < t - (l_1 + l_2) \leq l_3 \\ \leq 1 & \text{if } t - (l_1 + l_2) \leq 0. \end{cases} \tag{4.4}$$

To study the 3-state case, we extend the definitions used for the 2-state case.

Definition 4.20. For three probability distributions d_1, d_2, d_3, we define

$$\Delta(t; d_1, d_2, d_3) \stackrel{\text{def}}{=} \frac{d_1 * d_2 * [\overline{D}_3 - \delta]}{d_1 * [\overline{D}_2 - \delta]}(t).$$

When $d_j = d_{l_j}$, we again consider

$$\begin{aligned} \Delta(t; l_1, l_2, l_3) &\stackrel{\text{def}}{=} \Delta(t; d_{l_1}, d_{l_2}, d_{l_3}), \\ \mathbf{t}(l_1, l_2, l_3) &\stackrel{\text{def}}{=} \min\left[\{t \in \mathbb{N} \mid \Delta(t+1, l_1, l_2, l_3) > 1\} \cup \{\infty\}\right]. \end{aligned}$$

These 3-state definitions are coherent with 2-state definitions (4.1), in the sense that $\Delta(t; d_1, d_2) = \Delta(t; \delta_0, d_1, d_2)$, $\Delta(t; l_1, l_2) = \Delta(t; 0, l_1, l_2)$ and $\mathbf{t}(l_1, l_2) = \mathbf{t}(0, l_1, l_2)$.

4.6.2 VALIDITY OF THE CRITERION AND STOCHASTIC ORDERS

Previous section has stressed out the interest of monotone semi-Markov chains with respect to some stochastic orders. For linear (in a wide sense) semi-Markov chains made out of a convolution semigroup, we can easily characterize stochastic orders.

MONOTONY WITH RESPECT TO TIME t. This section exposes a strong result. For lr, hr, rh stochastic orders, all possible linear semi-Markov chains built out of a convolution semigroup $(D_l)_{l\geq 0}$ have the same monotony as a single family of distributions related to the corresponding Lévy process: its first-passage time measures. We recall its definition from section 3.2.2 in previous chapter.

Definition 4.21. Let $X = (X_l)_{l\geq 0}$ be a Lévy process. Assume X is supported on $K = \mathbb{R}_+$ or $\mathbb{N}$.

The *first-passage times* $T = (T_t)_{t\in K}$ are defined for all $t \in K$ as

$$T_t \stackrel{\text{def}}{=} \inf\{l > 0 \mid X_l > t\}.$$

The *first-passage measure* M_t is defined as the $(0, \infty)$-valued probability distribution of T_t.

Next proposition is the main result of this section. It holds for discrete-time as well as continuous-time processes.

Proposition 4.22. Let $(D_l)_{l\geq 0}$ be a convolution semigroup supported on $K = \mathbb{R}_+$ or $\mathbb{N}$. Let $M = (M_t)_{t\in K}$ denote the associated first-passage measures.
The following assertions are equivalent:

(i) M is lr-monotone.

(ii) For all sequence of positive durations $(l_1, l_2, l_3, \ldots)$, the linear semi-Markov is lr-monotone.

(iii) $t \mapsto \Delta(t, l_1, l_2, l_3)$ is non-decreasing for all $l_1, l_2, l_3 \in [0, \infty]$.

The following assertions are equivalent:

(i) M is lr-monotone.

(ii) For all sequence of positive durations $(l_1, l_2, l_3, \ldots) > 0$, the linear semi-Markov chain is rh-monotone

(iii) $t \mapsto \Delta(t, 0, l_1, l_2)$ is non-decreasing for all $l_1, l_2 \in [0, \infty]$.

The following assertions are equivalent:

(i) M is hr-monotone.

(ii) For all sequence of positive durations $(l_1, l_2, l_3, \ldots) > 0$, the linear semi-Markov chain is hr-monotone.

(iii) $t \mapsto \Delta(t, l_1, l_2, \infty)$ is non-decreasing for all $l_1, l_2 \in [0, \infty)$.

The proof is a straightforward consequence of two basic facts. First one is relationship between linear semi-Markov chains and first-passage times, as already explained in section 3.2.3. For the obtained linear semi-Markov chains, state distributions $\mathbf{f}(t)$ are spatial *discretizations* of the $[0, \infty)$-valued first-passage measures M_t. We state it again this fact which is explained in Proposition 3.3.

Proposition 4.23. Let $(\mathbf{f}(t))_{t\in\mathbb{N}}$ denote the state distributions of the linear semi-Markov chain $S = (S_t)_{t\in\mathbb{N}}$ characterized by event durations $(l_1, l_2, \ldots)$.

If all occupancy distributions D_j belong to the convolution semigroup $(D_l)_{l\geq 0}$ of a Lévy process $X = (X_l)_{l\geq 0}$, then

$$f_j(t) = M_t[l_{1:j-1}, l_{1:j}),$$

where $M = (M_t)_{t\geq 0}$ are the first-passage measures of X.

Second fact is preservation of TP_2 stochastic orders through any discretization of the probability distributions, as explained in the following lemma.

Lemma 4.24 (Shaked and Shanthikumar 2007, Theorem 1.C.5)**.** Let X, Y be two random variables on $\mathbb{R}$ and M, N denote their probability distributions.

$$M \underset{\mathrm{lr}}{\leqslant} N \iff [X \mid a \leqslant X < b] \underset{\mathrm{st}}{\leqslant} [Y \mid a \leqslant Y < b], \quad \text{for all } a \leq b.$$

$$M \underset{\mathrm{hr}}{\leqslant} N \iff [X \mid a \leqslant X] \underset{\mathrm{st}}{\leqslant} [Y \mid a \leqslant Y], \quad \text{for all } a \in \mathbb{R}.$$

$$M \underset{\mathrm{lrh}}{\leqslant} N \iff [X \mid a \geqslant X] \underset{\mathrm{st}}{\geqslant} [Y \mid a \geqslant Y], \quad \text{for all } a \in \mathbb{R}.$$

So now the proof of our main result is straightforward.

Proof of Proposition 4.22. [lr order] By definition,

$$\Delta(t+1, l_{1:j-1}, l_j, l_{j+1}) = \frac{M_t[l_{1:j}, l_{1:j+1})}{M_t[l_{1:j-1}, l_{1:j})}.$$

So the equivalence between (ii) and (iii) is immediate, as we can consider 3-state linear chains where l_1, l_2, l_3 take any positive values.

The equivalence between (i) and (iii) comes from a characterization of local order. Setting $a = l_1, b = l_1 + l_2 + l_3, x = l_1 + l_2$, (iii) is equivalent to: for all $a \leq x \leq b$, $\frac{M_t[a,x)}{M_t[a,b)}$ is non-decreasing with respect to t. Lemma 4.24 hereafter tells that this characterizes the lr order $(M_t)_{t\geq 0} \uparrow lr$, so it is equivalent to (i).

The proofs for other orders are similar. □

MONOTONY WITH RESPECT TO STATE DURATION l. Next proposition gives another interesting result: it characterizes the monotony of probability ratios f_{j+1}/f_j with respect to state durations l_1, l_2, l_3, using again first-passage measures.

Proposition 4.25. For all t, l_2, l_3, $\Delta(t, l_1, l_2, l_3)$ is increasing with respect to l_3. Moreover, the following assertions are equivalent:

(i) Every first-passage measure M_t is log-concave.

(ii) For all l_2, l_3, $\Delta(t; l_1, l_2, l_3)$ is non-increasing with respect to l_1

(iii) For all l_1, l_3, $\Delta(t; l_1, l_2, l_3)$ is non-increasing with respect to l_2

The following assertions are equivalent:

(i) Every first-passage measure M_t is IHR.

(ii) For all l_2, $\Delta(t; 0, l_1, l_2)$ is non-increasing with respect to l_1

The following assertions are equivalent:

(i) Every first-passage measure M_t is DRHR.

(ii) For all l_2, $\Delta(t; l_1, l_2, \infty)$ is non-increasing with respect to l_1

Remark. Each monotony result on Δ would imply the counterpart monotony of turning point $\mathbf{t}(l_1, l_2, l_3)$.

Sufficient condition: in short, this section promotes choosing occupancy distributions $(D_l)_{l\in L}$ as marginal laws of a Lévy process X whose first-passage-time process $T = (T_t)_{t\geq 0}$ is lr-monotone.

CONCLUSION. More generally, propositions 4.22 and 4.25 characterize semi-Markov chains S with first-passage times of X. Total positivity of Lévy processes and their first-passage times are studied later on in Chapter 7. In particular, an interesting result we obtain there is Proposition 7.10: the lr-monotony of $(T_t)_{t\geq 0}$ implies the log-concavity of every T_t. In other words, the monotony of $\Delta(t; l_1, l_2, l_3)$ with respect to time t implies its monotony with respect to event duration l_1, l_2 and l_3.

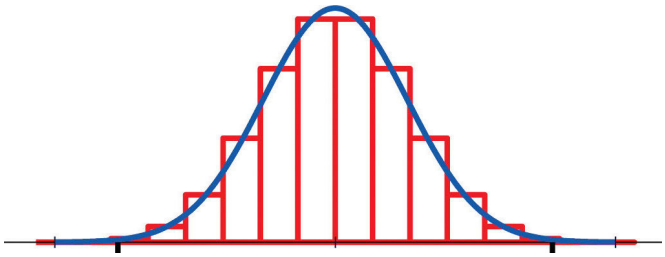


Figure 4.12: Periodic discretization of a continuous probability distribution. The discrete distribution inherits the log-concavity of the continuous one.

CASE OF CHAINS WITH IDENTICAL STATES. Proposition 4.25 above gives an additional information about the distributional properties of $\mathbf{f}(t)$. Indeed, Proposition 4.23 explains that state distributions $\mathbf{f}(t)$ are discretizations of first-passage measures M_t. Furthermore, a periodic discretization of a continuous probability measure preserves log-concavity, IHR, DRHR properties. Refer to appendix B.1 and see Figure 4.12 for an illustration.

Proposition 4.26. All state distributions $\mathbf{f}(t)$ of every linear semi-Markov chain with *identical* states $(l, l, l, \ldots)$ are discrete log-concave (resp. IHR, resp. DRHR) if and only if all first-passage measures M_t are log-concave (resp. IHR, resp. DRHR).

This result promotes chains with identical state durations.In next subsection, we again promote this case but with a different argument.

4.6.3 VALIDITY OF THE CRITERION ON INFINITELY LONG CHAINS

We now present an interesting phenomenon. As the chain becomes *infinitely long*, it can fulfill criterion 2 only if, asymptotically, all states have *identical duration*. Indeed, for a very long chain, a state that is strictly shorter than his predecessor would get skipped too early. Conversely, a state that is longer than his predecessor would get decoded for a too long duration.

Proposition 4.27. Let $(D_l)_{l\in L}$ be a non-degenerate convolution semigroup such that $\mu := \text{mean}[D_1] < \infty$ and $\text{Var}[D_1] < \infty$. As l_1 goes to $+\infty$, one has

$$\forall l_2, l_3 \in L, \quad \mathbf{t}(l_1, l_2, l_3) \underset{l_1\to\infty}{\sim} (l_1 + l_2)\mu C(l_2, l_3)$$

where $C(.,.)$ takes value $[0, \infty]$ and checks

$$C(l_2, l_3) \begin{cases} = 1 & \text{if } l_3 = l_3, \\ > 1 & \text{if } l_3 < l_2, \\ < 1 & \text{if } l_3 > l_2. \end{cases}$$

In the case $l_3 > l_2$, $C(l_2, l_3) > 0$ if and only if $d_{l_3}(0) + 1/d_{l_2}(0) > 2$. A sufficient condition is $l_2 > \ln 2 \approx 0.7$.

We are able to prove this result for discrete-time distributions, but we suspect it still holds with continuous-time ones. The proof comes from a straightforward application of a local limit theorem, also called Gnedenko's theorem.

Theorem 4.28 (Local limit theorem, Gnedenko and Kolmogorov, 1954)**.** Let d be a discrete pmf and $d^{*n} = d * \ldots * d$ denote the n-fold convolution of d. If d support minimum lattice is $\mathbb{N}$ and d has finite variance, then

$$\lim_{n\to\infty} \sup_{k\in\mathbb{N}} \left| \sqrt{n} d^{*n}(k) - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(k-n\mu)}{2n\sigma^2}} \right| = 0$$

where $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathrm{Var}[X]$ and $X \sim d$.

Proof of Proposition 4.27. We will prove this slightly more general result: there exists a unique $C(l_2, l_3) \in [0, \infty]$ such that

$$\forall t \in \mathbb{Z}, \quad \frac{\overline{D}_{l_1+l_2+l_3} - \overline{D}_{l_1+l_2}}{\overline{D}_{l_1+l_2} - \overline{D}_{l_1}}(C(l_2, l_3)l_1 + t) = 1.$$

Recall that

$$\frac{\overline{D}_{l_1+l_2+l_3} - \overline{D}_{l_1+l_2}}{\overline{D}_{l_1+l_2} - \overline{D}_{l_1}} = \frac{\overline{D}_{l_1+l_2+l_3} - \overline{D}_{l_1}}{\overline{D}_{l_1+l_2} - \overline{D}_{l_1}} - 1 = \frac{d_{l_1} * (\overline{D}_{l_2+l_3} - \delta_0)}{d_{l_1} * (\overline{D}_{l_2} - \delta_0)} - 1.$$

We begin with a small lemma. Let $(p_l)_{l\in L}$ be a convolution semigroup of pmfs, μ, σ denote the mean and standard deviation of p_1. For fixed t, k, applying Gnedenko's theorem with $n_= l\mu + t - k$ gives

$$\lim_{l\to\infty} \left| \sqrt{l} p_l(l\mu + t - k) - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-k)^2}{2\sigma^2 l}} \right| = 0,$$
$$\lim_{l\to\infty} \sup_{k\in\mathbb{N}} |\sqrt{l2\pi}\sigma p_l(l\mu + t - k) - 1| = 0.$$

therefore,

[Step 1.] Our proof relies on the so-called *Esscher transform* (Esscher, 1932), a.k.a. exponential tilting. Since $Z[d_l] = (Z[d_l])^l$, the Z-transforms of a convolution semigroup all have the same radius of convergence $R := R_{d_1} \geq 1$. Then, for all $\theta \in (0, R)$, $d_l^\theta(n) = d_l(n)\theta^l / Z[d_l](\theta)$ defines a new convolution semigroup of pmfs. Moreover, if $Z(\theta) := Z[d_1](\theta)$, then $\mu_\theta := \mathbb{E}[d_1^\theta] = Z'(\theta)$. Define also $\sigma_\theta^2 := \mathrm{Var}[d_1^\theta] < \infty$.

Then, using the simple identity $\sum_k f(k)g(n-k) = \theta^{-n} \sum_k f(k)\theta^k g(n-k)\theta^{n-k}$, one has for fixed $t \in \mathbb{Z}$ and $\theta \in (0, R)$,

$$\begin{aligned}
&\frac{d_l * (\overline{D}_{l_2+l_3} - \delta)}{d_l * (\overline{D}_{l_2} - \delta)}(\mu_\theta l + t) \\
&\qquad = \frac{\sum_{k=0}^{\mu_\theta l + t} \sqrt{l2\pi}\sigma_\theta d_l(l\mu + t - k) \frac{\theta^{l\mu+t-k}}{Z[d_l]} [\overline{D}_{l_2+l_3} - \delta](k)\theta^k}{\sum_{k=0}^{\mu l + t} \sqrt{l2\pi}\sigma d_l(l\mu + t - k) \frac{\theta^{l\mu+t-k}}{Z[d_l]} [\overline{D}_{l_2} - \delta](k)\theta^k} \\
&\qquad = \frac{\sum_{k=0}^{\infty} g_l(k)[\overline{D}_{l_2+l_3} - \delta](k)\theta^k}{\sum_{k=0}^{\infty} g_l(k)[\overline{D}_{l_2} - \delta](k)\theta^k}
\end{aligned}$$

with $g_l(k) := \sqrt{l2\pi}\sigma_\theta d_l^\theta(l\mu_\theta + t - k)$. Gnedenko's theorem gives $\lim_{l\to\infty} g_l \equiv 1$. This implies that $g_l(k) < M$ for some positive constant M. Since $\theta < R$, $M[\overline{D}_{l_2} - \delta](k)\theta^k$ is summable on $k \in \mathbb{N}$. So the theorem of dominated convergence allows limit-summation exchange:

$$\lim_{l\to\infty} \frac{d_l * (\overline{D}_{l_2+l_3} - \delta)}{d_l * (\overline{D}_{l_2} - \delta)}(\mu_\theta l + t) = \frac{\sum_{k=1}^\infty \overline{D}_{l_2+l_3}(k)\theta^k}{\sum_{k=1}^\infty \overline{D}_{l_2}(k)\theta^k} = \frac{Z[\overline{D}_{l_2+l_3}](\theta) - 1}{Z[\overline{D}_{l_2}](\theta) - 1}.$$

If $\theta = 1$, the ratio equals $l_3/l_2 + 1$. Else, it equals

$$\frac{1 - Z[d_{l_2+l_3}](\theta)}{1 - Z[d_{l_2}](\theta)} = \frac{1 - Z[d_{l_2}] + Z[d_{l_2}] - Z[d_{l_2} * d_{l_3}](\theta)}{1 - Z[d_{l_2}](\theta)} = \frac{Z[d_{l_2}] - Z[d_{l_2}](\theta)Z[d_{l_3}](\theta)}{1 - Z[d_{l_2}](\theta)} + 1.$$

As a conclusion,

$$\lim_{l_1\to\infty} \frac{\overline{D}_{l_1+l_2+l_3} - \overline{D}_{l_1+l_2}}{\overline{D}_{l_1+l_2} - \overline{D}_{l_1}}(\mu_\theta l_1 + t) = \begin{cases} \frac{1-(Z(\theta))^{l_3}}{(Z(\theta))^{-l_2}-1} & \text{if } \theta \neq 1, \\ \frac{l_3}{l_2} + 1 & \text{if } \theta = 1. \end{cases}$$

[Step 2.] It remains to set

$$C(l_2, l_3) \stackrel{\text{def}}{=} \mu_\theta/\mu_1 = Z'(\theta_{l_2,l_3})/Z(\theta_{l_2,l_3}),$$

for a well-chosen value of θ such that $\frac{1-(Z(\theta))^{l_3}}{(Z(\theta))^{-l_2}-1} = 1$. If such solution exist and is unique, it would ensure that $\mathbf{t}(l_1, l_2, l_3) \underset{l_1\to\infty}{\sim} C(l_2, l_3)\mu l_1$.

Let us look for solutions θ of

$$(Z(\theta))^{l_3} + (Z(\theta))^{-l_2} = 2.$$

Define $a := l_3, b := l_2$ and $f : x \mapsto x^a + x^{-b} - 2$. Then, $\lim_0 f = \lim_\infty f = \infty$. In addition, f reaches its unique minimum at $x_{a_b} = (b/a)^{1/(a+b)}$ and f is strictly monotone on $(0, x_{a,b})$ and $(x_{a,b}, \infty)$. Furthermore, $f(1) = 0$ and $f'(1) = a - b$. Now, three cases appear on (l_2, l_3).

[Case A.] If $l_2 = l_3$, then $x^{l_2}f(x) = x^{2l_2} + 1 - 2x^{l_2} = (x^{l_2} - 1)^2$ so $f(x) = 0$ has a unique solution $x = 1$. So $\theta = 1$ is the only solution.

[Case B.] If $l_3 > l_2$, then $f'(1) > 0$ and there exists a unique $x < 1$ such that $f(x) = 0$. As d_1 is not degenerate, Z is strictly increasing and maps $(0, 1]$ onto $(d_1(0), 1]$. If $(d_1(0))^{l_3} + (d_1(0))^{-l_2} \leq 2$, then there exists no solution θ: for all $c > 0$, one has $\lim_{l_1\to\infty} \frac{\overline{D}_{l_1+l_2+l_3} - \overline{D}_{l_1+l_2}}{\overline{D}_{l_1+l_2} - \overline{D}_{l_1}}(cl_1) > 1$, so $C(l_2, l_3) = 0$. Else there exists an unique $\theta_{l_2,l_3} \in (0, 1)$ such that $\frac{1-(Z(\theta))^{l_3}}{(Z(\theta))^{-l_2}-1} = 1$. As d_1 is not degenerate, Z is strictly convex and $\theta \mapsto \mu_\theta$ is strictly increasing, so $\mu_{\theta_{l_2,l_3}}/\mu_1 < 1$. Therefore,

$$C(l_2, l_3) \stackrel{\text{def}}{=} \mu_{\theta_{l_2,l_3}}/\mu < 1.$$

Moreover, $d_l(0) = (d_1(0))^l$ and the inequality $d_1(0) \geq e^{-1}$ holds for all infinitely divisible distribution pmf such that $\text{mean}[d_1] \leq 1$. So one have $(d_1(0))^{-l_2} \geq 2$ if $l_2 \geq \ln 2 \approx 0.7$. This condition ensures the existence of the solution θ.

[Case C.] If $l_3 < l_2$, else $f'(1) < 0$ and there exists a unique $x > 1$ such that $f(x) = 0$. $Z(\theta)$ is a one-to-one from $(0, 1/R)$ onto $(d_1(0), \gamma)$ where $\gamma := \lim_{z \to R^-} Z(z) > 1$ and $d_1(0) < 1$. If $f(\gamma) < 2$, there exists no solution θ: in such case, for all $c > 0$ one has $\lim_{l_1 \to \infty} \frac{\overline{D}_{l_1+l_2+l_3} - \overline{D}_{l_1+l_2}}{\overline{D}_{l_1+l_2} - \overline{D}_{l_1}}(cl_1) < 1$, therefore $C(l_2, l_3) = \infty$. Else, there exists an unique solution $\theta_{l_2,l_3} \in (1, 1/R)$ such that $\frac{1-(Z(\theta))^{l_3}}{(Z(\theta))^{-l_2}-1} = 1$. And as in previous case, $\theta_{l_2,l_3} > 1$ implies

$$C(l_2, l_3) \stackrel{\text{def}}{=} \mu_{\theta_{l_2,l_3}}/\mu > 1.$$

□

The conclusions of last proposition are twofold.

- Coherent semigroups with finite mean and finite variance must check

$$\forall l \in L, \qquad \text{mean}[D_l] = l.$$

- The only coherent chains are asymptotically composed of identical states. Bounds of coherency have following asymptotic behavior:

$$\lim_{l_1 \to \infty} L_{\min}(l_1, l_2) = \lim_{l_1 \to \infty} L_{\max}(l_1, l_2) = l_2.$$

Although last conclusion is negative, it ensures that tabulating the bounds of validity $L_{\min}(l_1, l_2)$, $L_{\max}(l_1, l_2)$ is possible even for extremely long values of l_1. As both quantities converge to l_2, one would not have to compute them numerically for all possible values.

CONSEQUENCE: RESHAPING THE INFERENCE METHOD. This fact suggests reshaping states that model the events to get states with identical nominal duration. To this aim, two procedures could be combined.

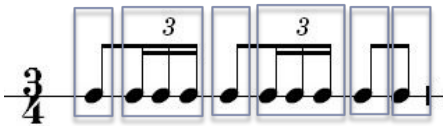


Figure 4.13: Reshaping state-space to increase coherency. Shortest notes are grouped into aggregates of equal nominal duration 1/2.

- *splitting events* in small duration intervals that would be approximately equal. This strategy is nevertheless limited: we have several times seen why the shorter the events, the worse the inference behaves.

- *aggregating events* in large time intervals that all have approximately equal time. Figure 4.13 illustrates such method.

In this setting, state estimation would be performed in two steps. First, the most likely aggregate is decoded. Second, the most likely state in this aggregate is selected. Such hierarchical method is not musically unjustified: estimation focuses on musical aggregates (beats, bars, group of bars) first, then focuses on the events that compose the most likely aggregate.

4.6.4 CASE STUDY: POISSON AND NEGATIVE BINOMIAL DISTRIBUTIONS

Suppose that all probability ratios $\Delta(t; l_1, l_2, l_3)$ are increasing with t. In this case, criterion 2 is true if and only if the turning point is coherently localized at $\mathbf{t}(l_1, l_2, l_3) = l_1 + l_2$. Proposition 4.27 tells this cannot be true as l_1 goes to infinity unless $l_2 = l_3$. But this is only a necessary condition as we only have $\mathbf{t}(l_1, l_2, l_3) \sim l_1 + l_2$. To get accurate results, recall from equation (4.4) that coherency criterion reads

$$\log \Delta(t + 1 + l_1 + l_2; l_1, l_2, l_3) \begin{cases} > 0 & \text{if } t \geq 0, \\ \leq 0 & \text{if } t < 0. \end{cases}$$

This suggests studying in the case $l_3 = l_2$ by considering shifted probability ratios $\Delta(t + 1 + l_1 + l_2; l_1, l_2, l_2)$. For $(D_l)_{l \geq 0}$ a convolution semigroup such that mean of D_l is l, it is known (due to the law of large numbers) that at fixed t,

$$\lim_{l_1 \to \infty} \Delta(t + 1 + l_1 + l_2; l_1, l_2, l_2) = 1.$$

We have run numerical computations with two common examples of semigroups, negative binomial laws and Poisson laws. Figure 4.14 depicts such curves of shifted ratios for one semigroup, but their shapes are very similar for other semigroups. Empirically the curves exhibit an interesting monotony with respect to l_1: the limit 1 is monotonically reached from above if $t > 0$, and form below if $t < 0$. Such behavior *automatically* implies the criterion.

We conjecture this empirical monotony is always true at least if durations l are integers and if l_1 is not smaller that l_2. We also wonder if other semigroups besides Poisson and Negative Binomial exhibit such behavior.

Conjecture 1. Assume $(D_l)_{l \geq 0}$ is a Poisson semigroup $D_l \sim Po(l)$. or a Negative Binomial semigroup $D_l \sim NB(l(1-p)/p, p)$ for some $p \in (0, 1)$.

Then, for all $l_1 \geq l_2 \geq 0$, shifted probability ratios $\Delta(t + 1 + l_1 + l_2; l_1, l_2, l_2)$ are non-increasing (resp. non-decreasing) with respect to l_1 if $t < 0$ (resp. $t \geq 0$).

As a consequence, every semi-Markov chains with identical states fulfills coherency criterion 2.

4.7 CONCLUSION

This chapter has introduced a second criterion to assess the coherency of score alignment algorithms. Investigating the coherency of online estimation with semi-Markov models

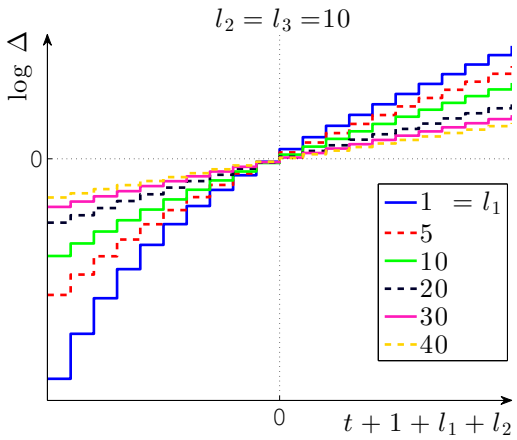


Figure 4.14: Numerical computations of shifted probability ratios with Negative Binomial laws $D_l \sim NB(l, 0.5)$. Curves $t \mapsto \Delta(t + l_1 + l_2 + 1; l_1, l_2, l_2)$ are sketched for different values of l_1. Turning points are coherent localized at $t - l_1 - l_2 - 1 = 0$, and curves exhibit monotonies in l_1 on both sides.

has stressed out the relevancy of several properties coming from the theory of total positivity: ageing properties like log-concavity or IHR; stochastic orderings like lr, hr, st. In addition, a connexion between the problematic of this chapter and the theory of Lévy processes has been highlighted: the same prescription as chapter 3 is retrieved with completely independent arguments.

More specifically, the prescriptions of this chapter is to choose occupancy distributions $(D_l)_{l \in L}$ among Lévy processes / convolution semigroups X that have the remarkable TP_2 properties. Two reliability classes are of interest. First, the Lévy process X itself should be monotone in the hr order. Second, its first-passage times T should be monotone in the lr order. This is why chapter 7 later on is devoted to mathematical investigation about total positivity of Lévy processes, of their first-passage distributions, and the relationship between both. But before, chapter 5 transposes this study of coherency to offline state-sequence estimation with general occupancy distributions.

5

COHERENT MODELING OF NOMINAL DURATIONS: STATE-SEQUENCE ESTIMATION

Previous chapter has dealt with online sequential alignment, where each hidden state S_t is individually estimated at occurrence time t. On the contrary, this chapter is about offline alignment, where the sequence (also called a *path*) of past hidden states $(S_1, \dots S_t)$ is estimated once at final time t. More specifically, it studies time-coherency of such backtracking estimation with respect to criterion 2. This procedure consists in estimating the state-sequence when its endstate S_t is given.

Section 5.1 describes backtracking estimation more formally, and define two estimators which are respectively called *right-censored* and *endtime constrained* Viterbi backtracking. As explained, such variants correspond to different probabilistic hypotheses. Section 5.1.2 adapts coherency criterion 2 to the two backtracking methods. This involves a discussion on the relevant definition of *nominal performance* in this specific estimation context. Afterwards, section 5.2 studies the coherency of the right-censored backtracking: mathematical prescriptions on occupancy distributions d_j are introduced, and the influence of nominal durations l_j on coherency is studied. Finally, section 5.3 studies the coherency of constrained endtime backtracking and provides similar results.

5.1 COHERENCY OF BACKTRACKING ESTIMATION

A *backtracking* procedure consists in estimating the whole past state-sequence $(S_1, \dots, S_T)$ that leads to a *given* endstate $S_T = \hat{s}_T$ at a *given* backtracking date T. Such a procedure is mostly employed to perform offline alignments. In this case, T is end time of observation $\mathbf{o}$. And $\hat{s}_T$ is usually set as the last state of the chain, accordingly with the hypothesis that all events occurred before the end time T. But backtracking is also useful in the online context, so as to retrieve accurate onset times of current and previous notes. In this case, T is current time and end state $\hat{s}_T$ is the output of online estimation.

5.1.1 DEFINITION OF BACKTRACKING METHODS

Several choices of decoding methods for the backtracked path coexist, as it is the case for online estimation. This chapters restrict to Maximum A Posteriori (MAP) estimation as it is the prevalent choice in the probabilistic literature.

THE TWO DEFINITIONS OF VITERBI BACKTRACKING. We call *Viterbi backtracking* the MAP path estimator: it decodes the path $s = (s_1, \dots s_T)$ with the best

posterior likelihood among admissible paths that respect the backtracking condition. In the context of HMMs, computing this estimate is straightforward with the standard Viterbi algorithm. However in the literature if HSMMs, two variants of the MAP estimator coexist and are equally employed[1]. They actually correspond to two different hypotheses on information provided by the backtracking condition. Definition 1 is the "rigorous" definition of MAP estimation. Definition 2 makes a simplifying assumption that is quite common in HSMM implementations: it assumes that hidden state S jumps on a different state at next time $T+1$. We interpret such variant as a different backtracking condition.

Definition 1: estimated path is the most likely path that fulfils backtracking condition $S_T = \hat{s}_T$,

$$(\hat{s}_1, \dots, \hat{s}_{T-1}) \in \underset{s_1,\dots,s_{T-1} \in E}{\arg\max}\, \mathbb{P}(S_T = \hat{s}_T, S_1^{T-1} = s_1^{T-1} \mid O_1^T = o_1^T).$$

This definition is called *right-censored* backtracking: no assumption is made whether current state $\hat{s}_T$ ends at or after last observation time $T+1$. The backtracked path corresponds to the arg max in the computation of the Viterbi classifier (with $j = \hat{s}_t$) defined in section A.3.3:

$$\delta_j(t) \overset{\text{def}}{=} \max_{s_1,\dots,s_{t-1}} \mathbb{P}(S_t = j, S_1^{t-1} = s_1^{t-1}, O_1^t = o_1^t).$$

Definition 2: estimated path is the most likely path that fulfils the alternative backtracking condition $S_T = \hat{s}_T, S_{T+1} \neq \hat{s}_T$,

$$(\hat{s}_1, \dots, \hat{s}_{T-1}) \in \underset{s_1,\dots,s_{T-1} \in E}{\arg\max}\, \mathbb{P}(S_{T+1} \neq \hat{s}_T, S_T = \hat{s}_T, S_1^{T-1} = s_1^{T-1}, O_1^T = o_1^T). \quad (5.1)$$

We call this definition *endtime constrained* backtracking: it forces current state $\hat{s}_T$ to end at time $T+1$. The backtracked path corresponds to the arg max in the computation of the *second* Viterbi classifier:

$$\delta_j^o(t) \overset{\text{def}}{=} \max_{s_1,\dots,s_{t-1}} \mathbb{P}(S_{t+1} \neq j, S_t = j, S_1^{t-1} = s_1^{t-1}, O_1^t = o_1^t).$$

In practice, backtracking is computed on any HSMM topology with the recursive Viterbi algorithm described in appendix A.3.3. Now, we introduce a few mathematical notations that contextualize backtracking to our two hypotheses: linear topology for alignment and non-discriminative observation.

Backtracking in a linear topology. A path s is defined as an arbitrary sequence of state $s_t \in E$ that represents a realization $(S_1, \dots, S_T)$ of the semi-Markov chain S between times 1 and T. The topology of the state-space E of S puts constraints on the feasible realizations of S. As this chapter exclusively deals with *linear* chains, admissible paths must start on initial state 1, then cross every intermediate state $2, 3, \dots$ in their

1. Appendix A.3.3 describes the two Viterbi algorithms. A review on these alternatives is also available in (Yu, 2010, Section 2.2.1).

given ordering and without skipping any. From now one, we assume the end state $\hat{s}_T$ is the $N+1$-th state, for some $N \in \mathbb{N}^*$. The set $\mathcal{P}$ of admissible paths s which end on state $S_t = N+1$ at time T is described as follows:

$$\mathcal{P} = \{s \in E^T \mid s_0 = 1, s_T = N+1, s_t \leq s_{t+1} \leq 1 + s_t\}.$$

Equivalently, each admissible path s is identified by the durations $u = (u_1, \dots, u_{N+1})$ it spends on each state $j = 1 \dots N+1$. There is one-to-one correspondence[2] between admissible paths $\mathcal{P}$ and admissible durations

$$\tilde{\mathcal{U}}_T := \left\{ (u_1, \dots, u_{N+1}) \in (\mathbb{N}^*)^{N+1} \mid \sum_{i=1}^{N+1} u_i = T+1 \right\}.$$

As the last duration is constrained the total duration T, $\tilde{\mathcal{U}}_T$ is in one-to-one correspondence with

$$\mathcal{U}_T := \left\{ (u_1, \dots, u_N) \in (\mathbb{N}^*)^{N} \mid \sum_{i=1}^{N} u_i < T \right\}.$$

Backtracking with non-discriminative observation. Our criterion is about the case of *non-discriminative observation* described in section 4.1. Under this hypothesis, the posterior likelihood equals prior probabilities and is independent from observation o:

$$\begin{aligned}
\delta_j(t) &= \mathbb{P}(S_t = j, S_1^{t-1} = s_1^{t-1}), \\
\delta_j^o(t) &= \mathbb{P}(S_{t+1} \neq j, S_t = j, S_1^{t-1} = s_1^{t-1}).
\end{aligned}$$

In the context of non-discriminative observation, the forthcoming study can be interpreted as an analysis of the prior path distributions of a semi-Markov chain. If additionally the chain has a linear topology, then the Viterbi quantities obey the following recursion:

$$\begin{aligned}
\delta_j(t) &= \max_{\substack{u, u_j \in \mathbb{N}^* \\ u + u_j = t}} \delta_{j-1}^o(u) \overline{D}_j(u_j), \\
\delta_j^o(t) &= \max_{\substack{u, u_j \in \mathbb{N}^* \\ u + u_j = t}} \delta_{j-1}^o(u) d_j(u_j),
\end{aligned}$$

where d_j is the occupancy pmf and $\overline{D}_j$ its survivor distribution. This leads to explicit formula:

$$\text{(Right-censored backtracking)} \qquad \delta_j(t) = \max_{\substack{u_1, \dots, u_j \in \mathbb{N}^* \\ u_1 + \dots + u_j = t}} d_1(u_1) d_2(u_2) \dots \overline{D}_j(u_j).$$

$$\text{(Constrained backtracking)} \qquad \delta_j^o(t) = \max_{\substack{u_1, \dots, u_j \in \mathbb{N}^* \\ u_1 + \dots + u_j = t}} d_1(u_1) d_2(u_2) \dots d_j(u_j).$$

5.1.2 STATEMENT OF COHERENCY CRITERION

Criterion coherency 2 has been introduced in the online alignement context, so it has to be adapted for backtracking. Its philosophy can be summarized as follows: in absence of

2. Correspondence from durations u to path s is given by $s_t = \min\{i \in \mathbb{N}^* \mid \sum_{j=1}^{i} u_i \geq t\}$.

discriminative observations, the estimated performance should be the *nominal performance*, namely the one that exactly respects available prior information. In the online estimation context, the music score is the only available information so the nominal performance is exactly as indicated by the score. However, for backtracking estimation, the backtracking condition brings further information. What is the exact definition of this *nominal path* in that context?

Definition of the nominal performance We define the nominal performance as the state-sequence that respects *all available information*: the nominal durations written on the music score, and also the backtracking condition. Defining nominal sequences rigorously is not straightforward and has to be carefully discussed. Indeed, this context brings two difficulties.

First, information brought by the backtracking condition *may contradict the music score*. This depends on backtracking duration T so several cases have to be distinguished.

Second, this backtracking condition depends on the chosen definition of backtracking. Here, we deal with the two variants introduced in section 5.1.1.

- For the constrained backtracking (definition 2.), the backtracking condition is

$$S_T = N + 1, \quad S_{T+1} \neq S_T.$$

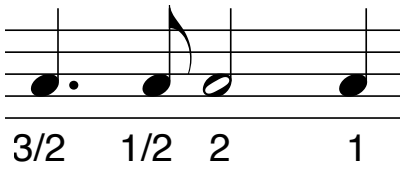
  This constraint reads "offset time of state $N + 1$ is exactly $T + 1$."
- For the right-censored backtracking (definition 1.), the backtracking condition is just

$$S_T = N + 1.$$

  This reads "offset time of state $N + 1$ is $\geq T + 1$, and onset time is $\leq T$". This does not tell if S_{T+1} stays or leaves the state, which leaves room for interpreting the likely offset time.

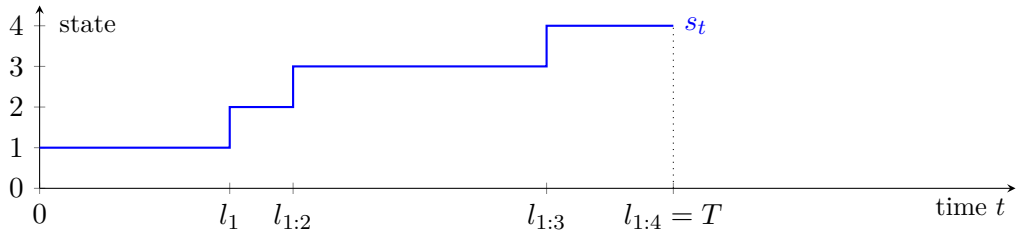
Therefore, the way nominal path is defined *differs* for each definition. In addition, we have to distinguish several cases on the backtracking duration T. Now, we go through all cases and state the definition of nominal performance for each. We also provide an illustration of nominal path, every time with the following toy score of $N + 1 = 4$:

$l_1 = 3/2$, $l_2 = 1/2$, $l_3 = 2$, $l_4 = 1$.



Suppose that $T = l_{1:N+1}$ where $N+1$ is the path end state $\hat{s}_T = N+1$. Interpreting this case is straightforward. Indeed, the constraint perfectly matches the score: $T + 1$ is the actual offset time of $N + 1$, and the offset times of previous states should be as in the score. So the nominal path is defined with durations u_i that are identical to the score:

$$\forall j = 1 \dots N + 1, \qquad u_j = l_j.$$

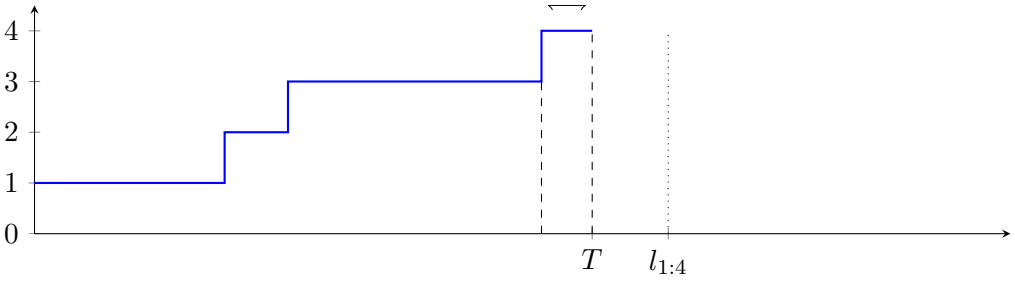


However, suppose that $T \neq l_{1:N+1}$. Contrary to the previous case, the nominal path has to be differently defined between the two backtracking methods. First, we state the definitions for each method and each case of backtracking duration T. Then, we give further justifications.

Definition 1: right-censored backtracking. We recall the constraint is $S_T = N + 1$.

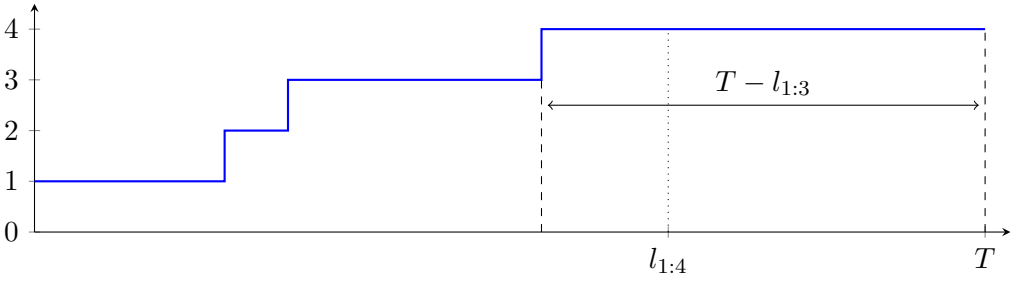
- If $l_{1:N} < T < l_{1:N+1}$, the constraint just says that state $N + 1$ is occurring between its score onset and offset times. This information still matches the score: $l_{1:N+1}$ is likely to be the actual offset time of state $N + 1$, it is simply not reached yet. The nominal path spends their nominal duration on intermediate states $j < N + 1$, and spends remaining duration on end state $N + 1$:

$$\forall j = 1 \dots N, \quad u_j = l_j \quad \text{and} \quad u_{N+1} = T - l_{1:N}.$$



- If $T > l_{1:N+1}$, the constraint contradicts the score: state $N + 1$ is still occurring later than its score offset time. We choose to define the nominal performance is defined as in previous case, so that all previous events durations match the score[3].

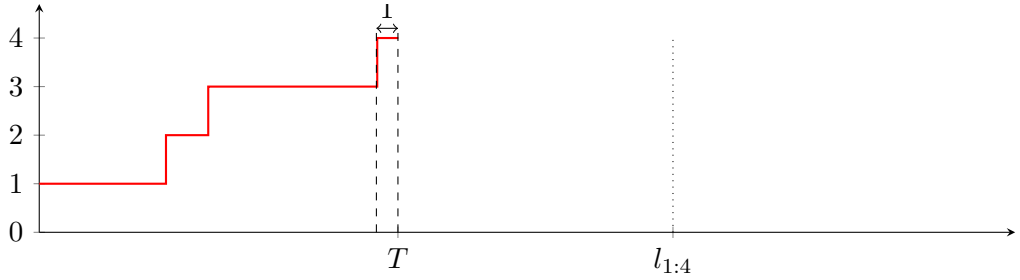
$$\forall j = 1 \dots N, \quad u_j = l_j \quad \text{and} \quad u_{N+1} = T - l_{1:N}.$$



3. In our opinion, this case is the only one where our definition of the nominal path is debatable. The *constant tempo* behavior defined hereafter could be an alternative.

- If $T < l_{1:N}$, the constraint contradicts the score: state $N+1$ is occurring sooner than its score onset time. The nominal path stays on $N+1$ the shortest possible duration (1 time step), and stays on each intermediate state a duration that is proportional to its nominal duration. This behavior called "constant tempo" is further detailed below.

$$\forall j = 1 \dots N, \quad u_j \propto l_j \quad \text{and} \quad u_{N+1} = 1.$$



Definition 2: constrained backtracking. We recall the constraint is $S_T = N+1, S_{T+1} \neq N+1$. This means: "the hidden process is on state $N+1$ at time T and leaves that state immediately afterwards".

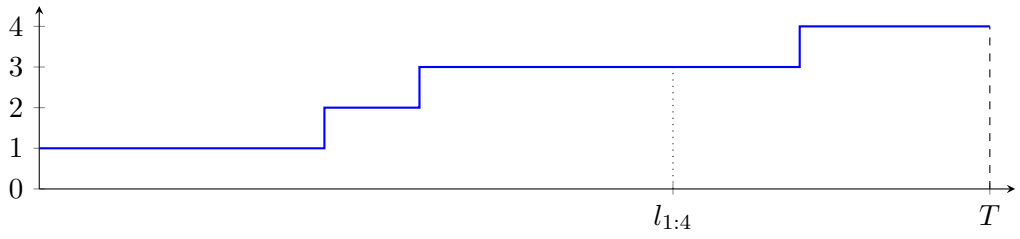
- If $T = l_{1:N+1}$, the constraint matches the score: state $N+1$ is ending at its score offset time. So the backtracked performance should be exactly the score performance:

$$\forall j = 1 \dots N+1, \qquad u_j = l_j.$$

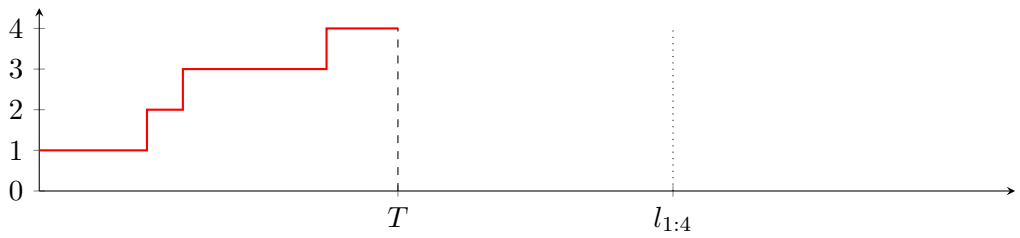
- If $T \neq l_{1:N+1}$, the constraint contradicts the score: state $N+1$ is ending sooner or later than its score offset time. Again, nominal path is defined with the constant tempo behavior:

$$\forall j = 1 \dots N+1, \quad u_j \propto l_j.$$

Next graph illustrates the case $T < l_{1:N+1}$.



Next graph illustrates the case $T > l_{1:N+1}$.



Note the two nominal performances differ from their respective counterpart in definition 1.

Justification of nominal paths When backtracking condition contradicts music score, defining the nominal performance is debatable. Our choices of definition are motivated by our applicative context. In music scores, nominal durations l are written as *relative durations* between events, which are translated into absolute values after a multiplication by the global *tempo*. This quantity roughly corresponds to the mean playing speed. If the backtracking duration contradicts the absolute nominal durations, we interpret this as a single change of tempo. Indeed, in absence of information coming from the observation, no intermediate changes of tempo could be noticed. So it is reasonable to assume the tempo globally differs from the score, but remains constant throughout the performance.

The idea is to apply the constant tempo behavior whenever an offset is known with certainty. With definition 2, the last offset time is surely known, whereas it is not with definition 1. In that latter case, the idea is to keep the nominal performance as close as possible to the score for the previous events.

Formulation of the coherency criterion With the linear topology depicted Figure 2.5, admissible paths necessarily starts from state 1 and goes to $N+1$ without skipping any state. So the first requirement of criterion 2 is unconditionally met. The criterion consists in asking that durations u^* of the backtracked path are equal to the durations of the nominal path, as defined above. Next two sections study each case separately.

5.2 COHERENCY OF RIGHT-CENSORED BACKTRACKING

Consider a given backtracking duration T and given chain of $N+1$ states with nominal durations $l_1, \dots, l_{N+1} \in L$ and $N \in \mathbb{N}^*$. For right-censored Viterbi backtracking, the optimal path u^* is given by

$$u^* \in \underset{\substack{u=(u_1,\dots,u_{N+1}) \\ u_1+u_2+\dots+u_{N+1}=T}}{\arg\max} d_{l_1}(u_1) d_{l_2}(u_2) \dots d_{l_N}(u_N) \overline{D}_{l_{N+1}}(u_{N+1}). \tag{5.2}$$

Two cases appear depending on the backtracking duration T is shorter or greater than nominal onset time of end state $l_{1:N} = \sum_{i=1}^{N} l_i$.

5.2.1 CASE OF SHORT DURATIONS T

Suppose $T \leq \sum_{i=1}^{N} l_i$. In this case, the coherency criterion 2 specializes to the following formulation.

Coherency criterion 2.1. Assume observation is non-discriminative. In the case $T \leq \sum_{i=1}^{N} l_i$, the backtracked path of duration T that goes from state 1 to $N+1$ spends on end state $N+1$ the shortest possible duration $u^*_{N+1} = 1$, and spends on each

intermediate state $j = 1 \dots N$ a duration that is proportional to the nominal duration $u^*_j \propto l_j$.

Note that the computation of the backtracking u^* takes the following recursive form:

$$u^*_{N+1} \in \underset{u_{N+1}=1\dots T-1}{\arg\max}\ \delta^o_N(T - u_{N+1}) D_{l_{N+1}}(u_{N+1}).$$

and intermediate durations $u^*_1, \dots, u^*_N$ are backtracked in the computation of $\delta^o_N(T - u^*_{N+1})$. This latter quantity corresponds to *constrained endtime* backtracking, which is the second method — studied later on in section 5.3. So in our case, coherency is equivalent to two claims:

1. the endstate duration is minimal: $u^*_{N+1} = 1$,
2. the constrained endtime backtracking with endstate N and duration $T - 1$ is coherent.

Claim 1. can be achieved with the very mild condition of unimodality — see appendix B.4.1 for definitions.

Proposition 5.1. Assume all d_l are discrete unimodal and $\overline{\text{mode}}[d_l] \leq l$. Then, the backtracked duration on the endstate is $u^*_{N+1} = 1$.

Proof. By contradiction, assume there exist $i_1 \leq N$ (say $i_1 = 1$) such that $u^*_1 > l_1 \geq \overline{\text{mode}}[d_{l_1}]$. By definition of unimodality,

$$d_{l_1}(u^*_1 - 1) > d_{l_1}(u^*_1).$$

As $\sum_{i=1}^N u^*_i \leq T = \sum l_i \leq \sum \overline{\text{mode}}[d_{l_i}]$, there exists $i_2 \neq i_1$ (say $i_2 = 2$) such that $u^*_2 \leq \overline{\text{mode}}[d_{l_2}]$. By unimodality again,

$$d_{l_2}(u^*_2 + 1) \geq d_{l_2}(u^*_2).$$

This contradicts the optimality of u^*, as exchanging (u^*_1, u^*_2) with $(u^*_1 - 1, u^*_2 + 1)$ gives a strictly better path.

As a result, one may assume that $u^*_i \leq \overline{\text{mode}}[d_{l_i}]$ for all $i \leq N$. By unimodality, $u_i \mapsto d_{l_i}(u_i)$ is non-decreasing on this region. As survivor distributions are non-increasing, $u_i \mapsto \overline{D}_{l_{N+1}}\left(T - \sum_j u_j\right)$ is non-decreasing too. So the optimum is reached at $T - \sum_{i=1}^N u^*_i = 1$, *i.e.*, $u^*_{N+1} = 1$. □

As for claim 2., we refer to section 5.3 which is devoted to constrained endtime backtracking. Among others results there, the *log-concavity* of d_l appears a necessary condition. In the present case, $T \leq \sum_{i=1}^N l_i$ so log-concavity of d_l is required only on $\{0, \dots, l\}$. This partial property corresponds to the *lower-half* log-concavity notion we introduce in appendix B.4.

5.2.2 CASE OF LONG DURATIONS T

Suppose $T > \sum_{i=1}^{N} l_i$. In this case, the coherency criterion 2 specializes to the following formulation.

Coherency criterion 2.2. Assume observation is non-discriminative. In the case $T > \sum_{i=1}^{N} l_i$, the backtracked path of duration T that goes from state 1 to $N+1$ spends on each intermediate state $j = 1 \dots N$ a duration equal to its nominal duration l_j, and spends on end state $N+1$ the remaining duration $T - \sum_{i=1}^{N} l_i$.

Formally the criterion reads:

$$(u_1^*, u_2^*, \dots, u_{N+1}^*) = \left(l_1, l_2, \dots, l_N, T - \sum_{i=1}^{N} l_i \right),$$

or equivalently, using the notation $L := \sum_{i=1}^{N} l_i$,

$$\forall u_1, \dots, u_N \in \mathbb{N}^*, \quad \sum_{i=1}^{N} u_i < T \implies \frac{D_{l_{N+1}}\left(T - \sum_{i=1}^{N} l_i\right)}{D_{l_{N+1}}\left(T - \sum_{i=1}^{N} u_i\right)} \geq \frac{d_{l_1}(u_1)}{d_{l_1}(l_1)} \frac{d_{l_2}(u_2)}{d_{l_2}(l_2)} \cdots \frac{d_{l_N}(u_N)}{d_{l_N}(l_N)}. \quad (5.3)$$

Next proposition explains how the optimal path is related to the *mode* of occupancy distributions d_l. Hereafter, the mode always refers to the greatest maximum,

$$\overline{\text{mode}}[d] \stackrel{\text{def}}{=} \sup\{t \in \mathbb{N} \mid \forall n \in \mathbb{N}, d(t) \geq d(n)\},$$

whereas $\underline{\text{mode}}$ refers to the smallest maximum.

Proposition 5.2. Let $u^* := (u_1^*, \dots, u_N^*)$ be the backtracked path given by equation (5.2).

(i) If $\underline{\text{mode}}[d_{l_i}] > l_i$ for some $i < N+1$, then equation (5.3) is always false for large enough T.

(ii) If $\text{mode}[d_l] = l$, then the optimal path u^* checks:

$$\sum_{i=1}^{N} u_i^* \geq \sum_{i=1}^{N} l_i.$$

(iii) If d_l is unimodal for all l, then for large enough T, the optimal path u^* checks:

$$\forall i = 1 \dots N, \quad u_i^* \geq \text{mode}[d_{l_i}].$$

Remark. Such results promote choosing d_l as unimodal distributions: the mode gives a lower bound of the time spent on each intermediate state. Besides, it tells that criterion 2.2 always fails unless

$$\forall l \in L, \quad \text{mode}[d_l] \leq l.$$

This condition used to be a sufficient condition in Proposition 2.2: now it appears as *necessary*.

Proof. *[(i)]* Define $l_j^* := \underline{\text{mode}}[d_{l_j}]$. Assume $l_j < l_j^*$ for some $j < N+1$. Define $L^* := l_j^* + \sum_{i=1, i\neq j}^{N} l_i$ so that $L < L^*$. As survivor distributions D are non-decreasing, $D_{l_{N+1}}(T-L)/D_{l_{N+1}}(T-L^*) \leq 1$. By definition of the smallest mode, one has $d_{l_j}(l_j^*)/d_{l_j}(l_j) > 1$.

Then, for any $T > L^*$, inequality (5.3) is broken by the path u such that $u_i = l_i$ for $i \notin \{j, N+1\}$ and $u_j = l_j^*$:

$$\frac{D_{l_{N+1}}(T-L)}{D_{l_{N+1}}(T-L^*)} \leq 1 < \frac{d_{l_j}(l_j)}{d_{l_j}(l_j^*)} = \frac{d_{l_j}(l_j*)\prod_{i\neq j}^{N} d_{l_i}(l_i)}{d_{l_j}(l_j)\prod_{i\neq j}^{N} d_{l_i}(l_i)}$$

[(ii)] Assume that l is the mode of d_l for all $l \in L$. Then inequality (5.3) always holds for u such that $\sum_{i=1}^{N} u_i \leq L$. Indeed, its the left-hand term is not lesser than 1, since $T - \sum_{i=1}^{N} u_i \geq T - L$ and survivor functions are non-increasing. And by definition of the mode, the right-hand term of inequality (5.3) is not greater than 1.

[(iii)] Finally, assume that for all $l \in L$, d_l is unimodal. Define $l_j^* := \text{mode}[d_{l_j}]$. Fix some admissible path $u := (u_1, \dots u_N, T - \sum_i u_i)$. By definition of unimodality and survivor distributions, the function $u_j \mapsto d_{l_j}(u_j)\overline{D}(T - u_j - \sum_{i\neq j} u_i)\prod_{i\neq j} d_{l_i}(u_i)$ is non-decreasing while $u_j \leq l_j^*$. As this argument works for any j, then necessarily, for any $T > \sum_{i=1}^{N} l_i^*$ the optimal path checks $\forall i \leq N, u_i \geq l_i^*$.

□

Next proposition explains how some further assumptions on $(d_l)_{l\in L}$ gives the existence of bounds on the validity of the criterion, and guarantees that tabulating such bounds is easy. Here again, *log-concavity* seems to be the key property.

Proposition 5.3. Assume that every distribution d_l is discrete log-concave and $\text{mode}[d_l] \leq l$.

(i) For all $N \in \mathbb{N}^*$, $l_1, \dots, l_N \in L$, there exists a bound $T_{l_1\dots,l_{N+1}} \in \mathbb{N}^* \cup \{\infty\}$ such that equation (5.3) holds if and only if

$$T \leq \sum_{i=1}^{N} l_i + T_{l_1\dots,l_{N+1}}.$$

(ii) In case $N = 1$, $T_{l,m}$ is given by the reciprocal function of the hazard rate h_m of d_m as:

$$T_{l,m} = \max\left\{t \in \mathbb{N}^* \mid h_m(t) \leq 1 - d_l(l+1)/d_l(l)\right\}.$$

(iii) In case $N > 1$,

$$T_{l_1\dots,l_{N+1}} = \min_{i=1\dots N} T_{l_i,l_{N+1}}.$$

(iv) Assume in addition that the family $(d_l)_{l\in L}$ is non-increasing in the hazard rate ordering. Then, every bound $T_{l,m}$ is a non-decreasing function of m.

Remark. Actually, all results remains valid if the occupancy distribution of the end state $d_{l_{N+1}}$ is only discrete IHR rather than log-concave.

Proof. Define $t^* := \max\left\{t \in \mathbb{N}^* \mid h_m(t) \leq 1 - d_l(l+1)/d_l(l)\right\}$. As log-concave distributions have increasing hazard rate,

$$\forall t \in \mathbb{N}^*, \quad t \leq t^* \quad \Longleftrightarrow \quad 1 - h_m(t) \geq d_l(l+1)/d_l(l).$$

The following two steps prove $(i-iii)$, while last step proves (iv).

Step 1. Assume $N = 1$ and define $l := l_1$, $m := l_2$. Let us prove that $t^* = T_{l,m}$. By log-concavity, $d_l(l+1+v)/d_l(l+v)$ is non-increasing with respect to $v \in \mathbb{N}$. Let $t \leq l + t^*$. One has

$$\forall v < t-l, \quad 1 - h_m(t-l-v) \geq \frac{d_l(l+1)}{d_l(l)} \geq \frac{d_l(l+1+v)}{d_l(l+v)}.$$

We recall that hazard rate is related to survivor distribution by $1-h(t) = \overline{D}(t+1)/\overline{D}(t)$. For any $u = l, \ldots, t-1$, multiplying this inequality side-by-side for $v = 1 \ldots u-l-1$ gives

$$\forall u = l \ldots t-1, \quad \frac{D_m(t-l)}{D_m(t-u)} \geq \left(\frac{d_l(l+1)}{d_l(l)}\right)^{u-l} \geq \frac{d_l(u)}{d_l(l)}. \tag{5.4}$$

This gives inequality (5.3) for all $u \geq l$. Since log-concavity implies unimodality, d_l is unimodal on $\mathbb{N}^*$. As $\mathrm{mode}[d_l] \leq l$ by assumption, proposition 5.2 gives the inequality for the remaining case $u < l$. This proves that $t_{l,m} \geq t^*$.

Reciprocally, if $t > l + t^*$, then by definition of t^* and of h_m:

$$\frac{\overline{D}_m(t-l)}{\overline{D}_m(t-l-1)} < \frac{d_l(l+1)}{d_l(l)},$$

so inequality (5.3) is violated with $(u_1, u_2) = (l_1+1, T-l_1-1)$. This ends the proof that $T_{l,m} = t^*$.

Step 2. Let N be any positive integer. Consider $i^* := \arg\min_{i=1\ldots N} T_{l_i, l_{N+1}}$. Firstly, we prove the existence and the $\leq$ inequality by showing that inequality (5.3) holds for all T such that $0 < t \leq T_{l_{i^*}, l_{N+1}}$ where $t := T - L$ and $L := \sum_{i=1}^N l_i$. Let T be such an integer. Since log-concavity implies unimodality, all d_l are unimodal on $\mathbb{N}^*$. By assumption, $\mathrm{mode}\, d_l \leq l$. So thanks to proposition 5.2, we only has to prove inequality (5.3) in the case: $\forall i \leq N, u_i \geq l_i$. So assume so.

By log-concavity, $u_i \geq l_i$ implies $\frac{d_{l_i}(u_i)}{d_{l_i}(l_i)} \leq \left(\frac{d_{l_i}(l_i+1)}{d_{l_i}(l_i)}\right)^{u_i - l_i}$.
By definition of $T_{l,m}$, $i^* \in \arg\max_{i=1\ldots N} \frac{d_{l_i}(l_i+1)}{d_{l_i}(l_i)}$. Combining both gives

$$\begin{aligned}\forall u_i \geq l_i, \quad \frac{d_{l_1}(u_1)}{d_{l_1}(l_1)} \cdots \frac{d_{l_N}(u_N)}{d_{l_N}(l_N)} &\leq \prod_{i=1}^N \left(\frac{d_{l_i}(l_i+1)}{d_{l_i}(l_i)}\right)^{u_i - l_i} \\ &\leq \left(\frac{d_{l_{i^*}}(l_{i^*}+1)}{d_{l_{i^*}}(l_{i^*})}\right)^{\sum_{i=1}^N (u_i - l_i)}.\end{aligned}$$

Define $U := \sum_{i=1}^N u_i$. In equation (5.3), U may take all values in $\{0, \ldots, T-1\}$. The case $U < L$ has been already proved. The remaining cases $U \geq L$ corresponds to this inequality:

$$\forall U = L \ldots T-1, \quad \frac{D_{l_{N+1}}(T-L)}{D_{l_{N+1}}(T-U)} \geq \left(\frac{d_{l_{i^*}}(l_{i^*}+1)}{d_{l_{i^*}}(l_{i^*})}\right)^{U-L}.$$

Using equation (5.4) above with $l = l_{i^*}, m = l_{N+1}$ proves this inequality is valid if T $(= t+L)$ is such that $0 < t \leq T_{l_{i^*}, l_{N+1}}$. This proves $T_{l_1 \ldots, l_N, l_{N+1}}$ exists and $T_{l_1 \ldots, l_N, l_{N+1}} \geq T_{l_{i^*}, l_{N+1}}$.

Secondly, we prove the $\leq$ inequality. If T is such that $t := T - L > T_{l_{i^*}, l_{N+1}}$, then

$$\frac{D_{l_{N+1}}(t)}{D_{l_{N+1}}(t-1)} > \frac{d_{l_{i^*}}(l_{i^*}+1)}{d_{l_{i^*}}(l_{i^*})} = \frac{d_{l_{i^*}}(l_{i^*}+1)\prod_{i\neq i^*}^{N} d_{l_i}(l_i)}{d_{l_{i^*}}(l_{i^*})\prod_{i\neq i^*}^{N} d_{l_i}(l_i)}$$

so inequality (5.3) is violated with the path $u = (u_1, \ldots, u_{l_{N+1}})$ such that $u_{i^*} = l_{i^*} + 1$, $u_{N+1} = t - 1$ and $u_i = l_i$. This proves that $T_{l_1 \ldots, l_N, l_{N+1}} \leq T_{l_{i^*}, l_{N+1}}$.

Step 3. $(D_l)_{l\in L}$ being hr-monotone means that the hazard rate h_m is non-increasing with m. The expression of $T_{l,m}$ clearly shows it has the reverse monotony of h_m with respect to m. □

PROPERTIES OF $T_{l,m}$. It would be interesting that coherency is valid when

$$l_{1:N} < T \leq l_{1:N+1}.$$

Indeed, this range of values for T is the one that does not contradict the information of music score about the nominal offset of $N+1$. With our notation, this reads $T_{l,l_{N+1}} \geq l_{N+1}$. Unfortunately, next proposition tells that for most distributions, this inequality is very unlikely to hold as soon as $l_{N+1} \leq l_j$ for some $j \leq N$.

Proposition 5.4 (bound on $T_{l,l}$)**.** If $h_l(l+1) > h_l(l)$, then

$$T_{l,l} \leq l.$$

Proof. d_l having a strictly increasing hazard rate at l reads $h_l(l+1) > h_l(l)$, or equivalently $\frac{\overline{D}(l+1)}{\overline{D}(l)} < \frac{d(l+1)}{d(l)}$. This violates inequality (5.3) with $T = 2l+1$. So $l + T_{l,l} < 2l+1$, which gives $T_{l,l} \leq l$. □

The condition is true if d_l is strictly IHR (at l), which is the case for almost all standard distributions that are log-concave. As a consequence, it is certain that if $l_{N+1} \leq l_j$, then the backtracking estimation cannot be coherent in the case $T > l_{1:N+1}$, which is the case where backtracking duration is strictly greater than nominal offset time of $N+1$.

Other analytic properties of $T_{l,m}$ would be interesting to investigate. We sum up a few questions which are still open:

- What is the monotony of $T_{l,m}$ with respect to to l?
- What is the magnitude of $T_{l,m}$ compared to m, l?
- What is the asymptotic behavior of $T_{l,m}$ as m goes to ∞?
- Are their some optimal family for this criterion? Can we compare $T_{l,m}$ between different families?

We have not achieved to obtain general result for the monotony of $T_{l,m}$ with respect to l. All we can do is proving that $T_{l,m}$ decreases with l for two special cases: Poisson and Negative Binomial distributions. Note that in both cases, parameters are chosen such that, for any integer $l \in \mathbb{N}^*$, the modes of d_l are exactly $\{l-1, l\}$.

Case of Poisson laws. Assume $D_l \sim Po(l)$. Then $(D_l)_{l\in\mathbb{N}^*}$ is a convolution semigroup. In addition, it is well-known that:

$$d_l(l+1)/d_l(l) = \frac{1}{e}\left(1+\frac{1}{l}\right)^l.$$

This quantity is increasing with l and maps $(0,1)$ onto itself, which proves the result. Indeed, by concavity of the function logarithm, the quantity $\frac{\log(1+x)}{x} = \frac{\log(1+x)-\log 1}{(1+x)-1}$ is non-decreasing with x.

Case of negative binomial laws. Assume $D_l \sim NB\left(1+\frac{1-p}{p}l, p\right)$. Then $(D_l)_{l\in\mathbb{N}^*}$ is *not* a convolution semigroup (but a delayed convolution semigroup). In addition, it is well-known that:

$$d_l(l+1)/d_l(l) = p + \frac{1-p}{\frac{1}{l}+1}$$

Again, this quantity is increasing with l and maps $(0,1)$ onto itself, which proves the result.

CASE OF CONVOLUTIONS SEMIGROUPS. Chapter 3 has raised the interest of choosing $D := (D_l)_{l\in L}$ as a convolution semigroup, for a very different coherency criterion. It turns out that here, again, there are some benefits of choosing D as a convolution semigroup, or at least an *additive* semigroup (*i.e.*, there exists a pmf $d_{l,m}$ such that $d_m = d_l * d_{l,m}$).

First, proposition 7.11 tells that for additive processes, the condition that all d_l are LCAV automatically implies the process is non-decreasing in the likelihood ratio order.

Second, if D is an additive semigroup and is unimodal, then it automatically has lower-half log-concavity under a mild assumption explained in Proposition 7.24. As explained in section 5.2.1, this property is interesting for time-coherency with short duration T.

Third, the asymptotic behavior of the bounds $T_{l,m}$ can be explicited.

Proposition 5.5. Assume $(d_l)_{l\in L}$ is a convolution semigroup, then

$$\lim_{m\to\infty} T_{l,m} = +\infty.$$

If in addition $\mathrm{mean}[d_1] = 1$ and d_1 has finite variance, then

$$\lim_{l\to\infty} T_{l,m} = 0.$$

Proof. If $\lim_{l\to\infty} D_l \equiv 1$, then the left-hand side of the inequality (5.3) converges to 1, whereas the right-hand side is always strictly greater than 1, so the inequality is true for all T.

If the mean of d_1 is 1, then mean of d_l is l. A consequence of the Central Limit Theorem is $\lim_{l\to\infty} D_l(l, l+1]/D_l(l-1, l] = 1$. This implies that the right-hand second converges to 1, whereas the left-hand side is strictly greater than 1. □

5.2.3 CONCLUSION

Our results justify why d_l should be chosed as *unimodal* distributions with mode located at l. In addition, choosing *log-concave* distributions gives a bound on backtracking duration T for which the coherency criterion is met, for a given set of nominal durations $(l_1)_{j=1\dots N}$. Bounds are very easy to compute and do not depend on the number of states N. However, such bounds on T are finite for almost all log-concave distributions. We wonder if coherency for arbitrary long durations T may be achieved with distributions that have a more complex shape.

Finally, having a non-decreasing family $(D_l)_{l\in L}$ in the hazard rate (hr) stochastic ordering provides bounds on nominal duration l_j: for a given duration T, the backtracked path is coherent if its end state is long enough.

5.3 COHERENCY OF CONSTRAINED ENDTIME BACKTRACKING

Consider a given backtracking duration T and a given chain of N states with nominal durations $l_1, \dots, l_N \in L$, where $N \in \mathbb{N}$ such that $N \geq 2$.
For constrained endtime Viterbi backtracking, the optimal path u^* is given by

$$u^* \in \underset{\substack{u=(u_1,\dots,u_N)\in(\mathbb{N}^*)^N \\ u_1+u_2+\dots+u_N=T}}{\arg\max} d_{l_1}(u_1)d_{l_2}(u_2)\dots d_{l_N}(u_N).$$

We recall the rationale of our coherency criterion. In music scores, the nominal durations l are written as *relative durations* between events. They are translated into absolute durations with a multiplication by the global tempo, which is roughly defined as the mean playing speed. If the backtracking duration deviates from the nominal duration, *i.e.*, $T \neq l_1 + \dots + l_j$, it is interpreted as a single change of the global tempo. In other words, our criterion coherency asks that the durations of each event in the nominal performance remains proportional to their relative duration.

Coherency criterion 2.3 (Criterion of constant tempo)**.** Assume observation is non-discriminative. For all time T and state N, the backtracked path that ends in N at time T and leaves N attikz $T+1$ assigns to each state $j = 1 \dots N$ a duration that is proportional to its nominal duration l_j.

Translating mathematically this statement is even simpler in continuous-time than an discrete-time. So exceptionally, both types of processes are considered in this section. Define $L := \sum_{j=1}^{N} l_j$.

- With continuous-time occupancy pdfs d_l, criterion 2.3 reads

$$\forall N \in \mathbb{N}^*, l_1, l_2, \dots, l_N \in L, u_j \in (0,\infty), \prod_{j=1}^{N} d_{l_j}(u_j) \leq \prod_{j=1}^{N} d_{l_j}\left(\frac{l_j}{L}U\right),$$

  for $U := \sum_{j=1}^{N} u_j$. This case is studied in section 5.3.1.

- With discrete-time occupancy pmfs d_l, criterion 2.3 reads

$$\forall N \in \mathbb{N}^*, l_1, l_2, \dots, l_N \in L, u_j \in \mathbb{N}^*, \quad \prod_{j=1}^{N} d_{l_j}(u_j) \leq \prod_{j=1}^{N} d_{l_j}(u_j^*), \tag{5.5}$$

  for some $(u_1^*, \dots, u_N^*) \in (\mathbb{N}^*)^N$ such that $u_j^* \approx \frac{l_j}{L} U$ for all j and $U := \sum_{j=1}^{N} u_j$. This case is devised in section 5.3.2. The discrete-time formalization is less straightforward because backtracked durations u_j^* must be positive integers, whereas expected durations $\frac{l_j}{L} U$ may not be so. Anyway one still expects $(u_1^*, \dots, u_N^*) \approx (\frac{l_1}{L} U, \dots, \frac{l_N}{L} U)$. But the relevant definition of such approximation $\approx$ has to be figured out.

5.3.1 CASE OF CONTINUOUS-TIME DISTRIBUTIONS

In the continuous-time case, we are able to completely characterize families that are coherent for criterion 2.3.

Theorem 5.6. A family $(d_l)_{l>0}$ of pdfs is coherent for criterion 2.3 if and only every d_l is log-concave on $(0, \infty)$ and

$$\forall l > 0, t > 0, \quad d_l(t) = \frac{1}{l \int [d_1(t)]^l \mathrm{d}t} [d_1(t/l)]^l.$$

Conversely, using this equation, any log-concave pdf d induces a coherent family $(d_l)_{l>0}$ of pdfs such that $d_1 = d$.

Proof. Step 1. Let us prove that d_l is log-concave for any $l > 0$. With $N = 2$ and $l_1 = l_2 = l$, equation 5.5 reads

$$\forall u = (u_1, u_2) \in (0, \infty)^2, \quad d_l(u_1) d_l(u_2) \leq \left[d_l \left(\frac{u_1 + u_2}{2} \right) \right]^2.$$

applying the logarithm gives

$$\forall u = (u_1, u_2) \in (0, \infty)^2, \quad \log d_l(u_1) + \log d_l(u_2) \leq 2 \log d_l \left(\frac{u_1 + u_2}{2} \right).$$

This equation means that the function $\log d_l$ are *midpoint concave*. As d_l is a pdf, it is a measurable function and so is $\log d_l$. It is well-known that for measurable functions, midpoint concavity is equivalent to concavity.

Step 2. As concave functions, $f_l = \log d_l$ admit left-derivatives $f'_{l,g}$ and right-derivatives $f'_{l,g}$ everywhere. Moreover $\lim_{t \to 0^+} f_l(0) = \log d_l(0^+)$ exists and is finite. Fix $N = 2, l_1 = 1, l_2 = l > 0$. The criterion reads that for all $U > 0$, the function $F_U : u \mapsto \log d_1(u) + \log d_l(U - u)$ reaches its maximum at $u = U/(1 + l)$, so that

$$U - u = Ul/(1 + l), \qquad u/(U - u) = 1/l + l.$$

As F_U is concave, this is equivalent to

$$\forall U > 0, \quad F'_{U,g} \left(U \frac{1}{1+l} \right) \leq 0, \qquad F_{U,d} \left(U \frac{1}{1+l} \right) \geq 0$$

$$\forall U > 0, \quad f'_{1,g} \left(U \frac{1}{1+l} \right) \leq f'_{l,g} \left(U \frac{1}{1+l} \right), \qquad f'_{1,d} \left(U \frac{1}{1+l} \right) \geq f'_{l,d} \left(U \frac{1}{1+l} \right).$$

Using the change of variables $U = u\frac{l+1}{l}$, this is equivalent to

$$\forall u > 0, \qquad f'_{1,g}(u/l) \leq f'_{l,g}(u), \qquad f'_{1,d}(u/l) \geq f'_{l,d}(u).$$

As log-concave functions are locally absolutely continuous on $(0, \infty)$, for any $t > 0$, integrating above equations on $u \in [0, t]$ gives

$$l(f_1(t/l) - f_1(0^+)) \leq f_l(t) - f_l(0), \qquad l(f_1(t/l) - f_1(0^+)) \geq f_l(t) - f_l(0^+),$$

therefore

$$\forall t > 0, \qquad l \log \frac{d_1(t/l)}{d_1(0^+)} = \log \frac{d_l(t)}{d_l(0^+)}.$$

and there is some $C_l \geq 0$ such that

$$\forall t > 0, \qquad d_l(t) = C_l d_1(t/l)^l. \tag{5.6}$$

As d_l is a pdf, $C_l > 0$ and $1/C_l = \int d_1(t/l)^l \mathrm{d}t = l \int d_1(t)^l \mathrm{d}t$. Since d_1 is log-concave, it has an exponential tail and so does $t \mapsto d_1(t)^l$. Therefore the above integrals exist.

Step 3. Reciprocally, let d_1 be a log-concave density and define $(d_l)_{l \in L}$ as in equation (5.6). The above lines prove that the family checks the condition for $N = 2$. With similar computations, we show that (d_l) checks the condition for any $N \geq 2$. For convenience we assume that d_1 is differentiable but the same reasoning holds if not. Define

$$f(u_2, \dots u_N) := \log d_{l_1}\left(U - \sum_{i=2}^{N} u_i\right) + \sum_{i=1}^{N} \log d_{l_i}(u_i).$$

As d_l are log-concave, f is concave. So it reaches its maximum at u^* if for all $j > 1$, $\partial_j f(u^*) = 0$. Since $\forall l > 0$, $\frac{d}{dt} \log d_l(t) = \frac{d}{dt} \log d_1(t\, l)$, one has $\partial_j f(u) = \log d_1(u_j\, l_j) - \log d_1((U - \sum_{i=2}^{N} u_i)\, l_1)$. These quantities all vanish if $\forall j > 1, u_j^*\, l_j = u_1^*\, l_1$ where $u_1^* \overset{\text{def}}{=} U - \sum_{i=2}^{N} u_i^*$. This implies $u_j^*/u_k^* = l_j/l_k$ for all j, k. As $\sum_{i=1}^{N} u_i^* = U$, the only possibility is $u_j^* = l_j U/L$. □

Remark. Assume $D_1 \sim \mathcal{N}(\mu, \sigma^2)$ (Gaussian law). The theorem would give $D_l \sim \mathcal{N}(l\mu, l\sigma^2)$: the mean and the variance would be proportional to l. Interestingly, this family $(D_l)_{l \in L}$ is a convolution semigroup of measures on $\mathbb{R}$. Since there are no other fully log-concave convolution semigroup on $\mathbb{R}$ or $\mathbb{R}_+$, it is the only convolution semigroup that checks the conditions of the theorem.

5.3.2 CASE OF DISCRETE-TIME DISTRIBUTIONS

The criterion is more involved in the discrete-time case. We have no characterization of coherent families of occupancy distributions, so we build up partial conditions.

LOG-CONCAVITY. In the continuous-time, log-concavity has been obtained as a necessary condition of the criterion. Actually, this implication still holds in discrete-time. Remember that log-concave distributions are unimodal.

Let us formalize discrete-time version of criterion 2.3 as

$$\forall N \in \mathbb{N}^*, l_1, l_2, \dots, l_N \in L, u_j \in \mathbb{N}^*, \quad \prod_{j=1}^{N} d_{l_j}(u_j) \leq \prod_{j=1}^{N} d_{l_j}(u_j^*), \tag{5.7}$$

for some $(u_1^*, \dots, u_N^*) \in (\mathbb{N}^*)^N$ such that $\sum_j u_j^* = U$ where $U := \sum_j u_j$

$$\forall j = 1 \dots N, \qquad |u_j^* - \frac{l_j}{L} U| < 1. \tag{5.8}$$

Proposition 5.7. If a family of pmfs $(d_l)_{l \in L}$ checks criterion 2.3 for $N = 2$, in the sense of equation (5.7) with condition (5.8), then all d_l are discrete log-concave.

Conversely, if d_l is discrete log-concave, then equation (5.7) with condition (5.8) is checked for all $N \geq 2$ if all states have equal length $l_1 = l_2 = \dots = l$.

Proof. For any $u_1, u_2 \in \mathbb{N}^*$, the only couples of integers (u_1^*, u_2^*) that checks $|u_j^* - \frac{l}{l+l}(u_1 + u_2)| < 1$ are $(u_1^*, u_2^*) = \left(\left\lfloor \frac{u_1+u_2}{2} \right\rfloor, \left\lceil \frac{u_1+u_2}{2} \right\rceil\right)$ and vice-versa.

[⇒] This is why in the case $N = 2$, equation (5.7) is equivalent to

$$\forall u_1, u_2 \in \mathbb{N}^*, \qquad d_l(u_1) d_l(u_2) \leq d_l\left(\left\lfloor \frac{u_1 + u_2}{2} \right\rfloor\right) d_l\left(\left\lceil \frac{u_1 + u_2}{2} \right\rceil\right).$$

This equation is the definition d_l being discrete *midpoint* log-concave. We can show this is equivalent to discrete log-concavity — see definition B.15 in the appendix..

Let $l > 0$, $n \in \mathbb{N}^*$. If $l_1 = l_2 = l$ and $u_1 = n, u_2 = n + 2$, then $U = 2n + 2$, $Ul_1/L = Ul_1/L = n + 1 \in \mathbb{N}^*$. Applying equation (5.7) with $N = 2$, gives

$$\forall n \in \mathbb{N}^*, \quad d_l(n) d_l(n + 2) \leq (d_l(n + 1))^2.$$

Since $d := d_l$ is nonnegative, it remains to prove it has no internal zero. By contradiction, assumes it has one. Then, there exists positive integers $2 + u_1 \leq u_2$ such that $d(u_1)d(u_2) > 0$ and $d(x) = 0$ for all $x \in \mathbb{N}^*$ such that $u_1 < x < u_2$.

Then, take $(u_1^*, u_2^*) = \left(\left\lfloor \frac{u_1+u_2}{2} \right\rfloor, \left\lceil \frac{u_1+u_2}{2} \right\rceil\right)$ or vice-versa. Since $u_2 - u_1 \leq 2$, one has $u_1 < u_1^* \leq u_2^* < u_2^*$, so $d(u_1^*)d(u_2^*) = 0$. This contradicts $d(u_1)d(u_2) \leq d(u_1^*)d(u_2^*)$.

[⇐] The sufficiency may be checked as done for the continuous case (theorem 5.6). It suffices to consider piecewise affine functions that extend d_l on $\mathbb{R}_+$. □

Equation (5.7) suggests the approximation condition $|u_j^* - \frac{l_j}{L} U| < 1$. It is possible to achieve such condition when $N = 2$ or when all l_j are equal. However, this condition is too restrictive in the general case. We believe it is possible to achieve the following approximation $|u_j^* - \frac{l_j}{L} U| < N - 1$. This latter condition coincides with the former one in the case $N = 2$. Therefore, we suggest a new formalization of coherency criterion 2.3:

$$\forall N \in \mathbb{N}^*, l_j \in L, u_j \in \mathbb{N}^*, \quad \prod_{j=1}^{N} d_{l_j}(u_j) \leq \prod_{j=1}^{N} d_{l_j}(u_j^*),$$

for some $(u_1^*, \dots, u_N^*) \in (\mathbb{N}^*)^N$ such that $\sum_j u_j^* = U$ with $U := \sum_j u_j$ and

$$\forall j = 1 \dots N, \qquad |u_j^* - \frac{l_j}{L} U| < N - 1. \tag{5.8 revisited}$$

We believe with this formulation, the N-state problem reduces to the 2-state one.

Conjecture 2. If a family of pmfs $(d_l)_{l \in L}$ checks equation (5.7) with condition (5.8 revisited) for $N = 2$ states, then it does for all $N \geq 2$.

A characterization of all families of distributions that checks the criterion is an open question. So far, we have found out that a well known family does.

COHERENCY OF POISSON DISTRIBUTIONS. Poisson distributions $Po(l)$ seem to fulfill criterion 2.3 for any $N \geq 2$. We are able to prove the case $N = 2$, which might be sufficient to imply the other cases. We wonder if it is the only family of discrete distributions, or at least the only discrete Lévy process that fulfills this criterion. It would prove that, for this criterion again, Poisson distributions are the ideal ones.

Proposition 5.8 (Conjecture)**.** The Poisson process $D_l \sim Po(l)$ satisfies equation (5.7) for any $N \geq 2$ with condition (5.8 revisited).

Now, we give the proof for the case $N = 2$.

Proposition 5.9. The Poisson process $d_l \sim Po(l)$ satisfies equation (5.7) for $N = 2$ with condition (5.8 revisited).

Proof. The pmf of the Poisson law $Po(\lambda)$ is $d_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!}$. Let $\lambda, \mu > 0$ and d_λ, d_μ denote their associated Poisson pmf. If $\mu = \lambda$, discrete log-concavity gives the result. So without restriction we can assume that $\lambda > \mu$.

Let t be a positive integer. Define $B : u \mapsto \log d_\lambda(t-u) + \log d_\mu(u)$ on $\{1, \dots, t-1\}$, $\hat{u} = \arg\max_{u=1,\dots,t-1} B(u)$ and $\tilde{u} \stackrel{\text{def}}{=} t\frac{\lambda}{\lambda+\mu}$. Equation (5.7) with $N = 2$ and $U = t$ holds if and only if $\hat{u} \in \{\lfloor \tilde{u} \rfloor, \lceil \tilde{u} \rceil\}$.

The proof comes from the properties of the function Γ. Remember that $\Gamma(n+1) = n!$ for all integer n, so B admits the following extension on the continuous interval $I = [1, t-1]$:

$$B(u) = -\lambda + (t-u)\log\lambda - \log\Gamma(t+1-u) - \mu + u\log\mu - \log\Gamma(u+1).$$

The resulting function B is strictly concave on I. Indeed, it is twice differentiable and

$$B'(u) = \log\frac{\mu}{\lambda} + \psi(t+1-u) - \psi(u+1),$$
$$B''(u) = -\psi'(t+1-u) - \psi'(u+1).$$

Since ψ is strictly increasing, $B''(u) < 0$ and B is strictly concave. The assumption $\lambda \geq \mu$ implies that $\psi(t+1-u^*) - \psi(t+1-u^*) = \log\lambda - \log\mu \geq 0$, so $t - u^* \geq u^*$, i.e. $u^* \in [1, t/2]$. So from now one, we restrict the study to this interval.

It is well known that $\psi(x+1) \approx \ln x$, so

$$\psi(t+1-u) - \psi(u+1) \approx \ln(t+1-u) - \ln u = \ln\frac{t-u}{u}.$$

This fact gives the intuition that $\log\frac{\mu}{\lambda} + \psi(t+1-u) - \psi(u+1) \approx 0$ holds if $\frac{t-u}{u} = \frac{\lambda}{\mu}$. We prove it rigorously with a sum-integral comparison, using the property that for all real number x and positive integer N, $\psi(x+N) - \psi(x) = \sum_{k=0}^{N-1}\frac{1}{x+k}$. So for all u such that $u < t-u$ one have $B'(u) = \ln\frac{\mu}{\lambda} + \sum_{k=u+1}^{t-u}\frac{1}{k}$.

We can prove the following inequality for all integers u, t such that $u < t-u$,

$$\sum_{k=u+1}^{t-u}\frac{1}{k} \leq \ln\frac{t-u}{u} \leq \sum_{k=u}^{t-u+1}\frac{1}{k},$$

in other words

$$B'(u) \leq \ln\frac{t-u}{u} - \ln\frac{\lambda}{\mu} < B'(u-1). \tag{5.9}$$

The proof relies on the so-called Euler-MacLaurin's formula (where $\{x\}$ is the fractional part of x)

$$\sum_{k=1}^{n} \frac{1}{k} - \ln n = \frac{1}{2n} + \frac{1}{2} + \int_1^n \frac{\{x\} - 1/2}{t^2} \mathrm{d}t.$$

Moreover, $u = \tilde{u}$ is the solution of $\ln \frac{t-u}{u} - \ln \frac{\lambda}{\mu} = 0$. Inequality (5.9) together with the monotony of B' implies that $B'(\lceil \tilde{u} \rceil) \leq 0$ and $B'(\lceil \tilde{u} \rceil - 1) \geq 0$. Since B is strictly concave, this implies that its discrete maximum $\hat{u}$ on $\{0, \ldots, t\}$ is equal to $\lceil \tilde{u} \rceil - 1$ or $\lceil \tilde{u} \rceil$. This gives exactly the result. □

5.3.3 JUSTIFICATIONS FOR COHERENT BEHAVIOR

Our coherency criterion is defined with a nominal performance that has the "constant tempo" behavior described in section 5.1.2. This choice may appear arbitrary. The goal of this section is to provide further justification.

Argument 1: identifiability/ Let us consider two states again $1, 2$ and their duration probabilities d_1, d_2. The Viterbi algorithm computes

$$\hat{u}(t) = \underset{u=1\ldots t-1}{\arg\max} \log d_1(u) + \log d_2(t-u), \tag{5.10}$$

and tells that duration spent on state 1 is $\hat{u}_1 = \hat{u}(t)$, and duration spent on state 2 is $\hat{u}_2 = t - \hat{u}(t)$.

Let us assume that $l_1 = l_2 = l$ so that $d_1 = d_2$. The function $(u_1, u_2) \mapsto d_1(u)d_1(t-u)$ is now symmetric and has two optima $(\hat{u}(t), t - \hat{u}(t)), (t - \hat{u}(t), \hat{u}(t))$. This coexistence of two distinct optima brings a *problem of identifiability*. The only solution that avoids such ambiguity would be $t - \hat{u}(t) = \hat{u}(t)$. This holds for all event time $t = 2v$ if and only if the distribution is log-concave.

As a conclusion, the constant tempo behavior that involves log-concave distribution is the only one that does not face this problem of identifiability. For all the other ones, this problem happens and gets worse as the number of consecutive states increases: distinct optimal solutions would increase too. One may argue that an *ad hoc* heuristic solves this problem of identifiability. For instance, let us present one that looks musically coherent. If $t > 2l$ then the tempo is slowing down; such tempo change would be monotonic if duration spent on state 2 is longer than state 1. So the coherent solution is only $(\hat{u}_1, \hat{u}_2)$ such that $\hat{u}_1 \geq \hat{u}_2$. If $t < l$ the situation is the converse: since the tempo speeds up, the coherent solution is $\hat{u}_1 \leq \hat{u}_2$. Nevertheless, our heuristic gives no clue when $t = 2l$. This case is still problematic for any behavior except the constant tempo one.

Argument 2: sensitivity. Even if the identification problem could be tackled with an heuristic, another problem would remain: the sensitivity to observation probabilities. So far we have assumed a non-discriminative observation $b_1 = b_2$. Now let us assume a very weakly informative observation: $b_1 = r\, b_2$ where r is a constant real number. If $r < 1$, observation favors state 1 whereas $r > 1$ favors state 2.

Assume the two states have the same occupancy distribution d. For a backtracking duration t, the optimal path $(\hat{u}_1(t), \hat{u}_2(t))$ for the constrained Viterbi backtracking is

$$\hat{u}_1(t) = \arg\max_{u=1\ldots t-1} \log d(u) + \log d(t-u) + u \log r,$$

and $\hat{u}_2(t) = t - \hat{u}_1(t)$. Assume d is not log-concave. Figure 5.1 illustrates the phenomenon with a distribution $d = d_1 = d_2$ that is neither log-concave nor log-convex, but very close to a log-concave function. In the case $r = 1$, there exists a t such that the function above has two distinct optima u^*, $t - u^*$ such that $u^* < t - u^*$. Now, assume the observation slightly favors state 2: $r = 1 - \epsilon$ with $0 < \epsilon \ll 1$. Then, it is easy to see that $\hat{u}_1(t) < \hat{u}_2(t)$, so $\hat{u}_1(t)$ is around u^*. Finally, assume observation slightly favors state 1: $r = 1 + \epsilon$. Then, $\hat{u}_1 > \hat{u}_2(t)$ and $\hat{u}_1(t)$ is around $t - u^*$. This simplistic case illustrates the sensitivity to observation: a slight difference in observation probabilities provokes a big difference in the estimation.

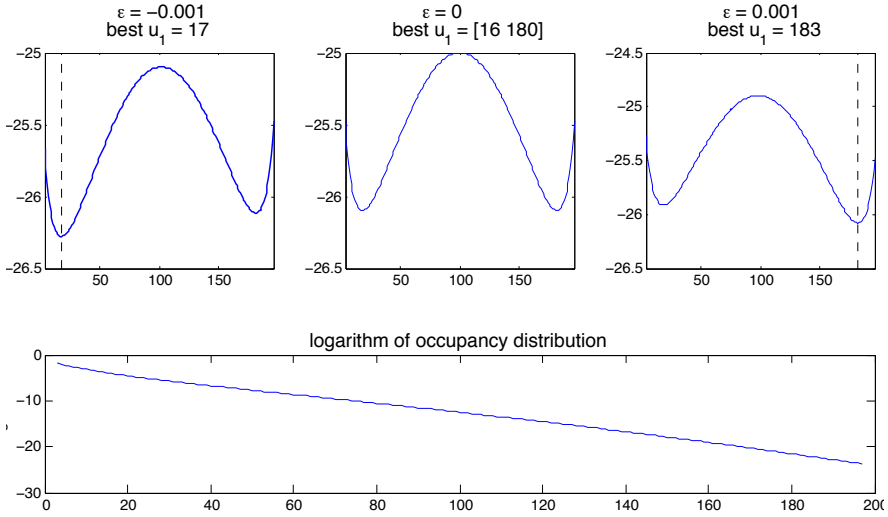


Figure 5.1: Illustration of the identifiability and sensitivity issues with a non-log-concave distribution. Two states $1, 2$ have the same occupancy distribution d. Below, graph of $\log d(t)$. Above, thicks indicate the backtracked duration $\hat{u}_1$ in three situations: $\epsilon = -0.001$ (left), $\epsilon = 0$ (middle), $\epsilon = +0.001$ (right). $\hat{u}_1$ dramatically changes between ± 0.001.

Example of incoherency: log-convex distributions. We show how two subclasses of distributions that are not log-concave are problematic for inference under non-discriminative observation.

Firstly, log-convex distributions lead to extreme incoherency. If some occupancy distribution d_l is log-convex, then

$$\forall t \in \mathbb{N}^*, \quad d_l(t-1)d_l(1) = d_l(1)d_l(t-1) = \max_{u=1\ldots t-1} d_l(t-u)d_l(u).$$

So the backtracking equation (5.10) has two solutions $(1, t-1)$ and $(t-1, 1)$. These solutions assign all duration t on one state and squeeze the duration of other states to

the smallest possible one. This completely contradicts the time-coherency. For instance, log-normal distributions have been suggested by Takeda et al. (2007). As those functions are log-convex, they are a wrong choice.

Secondly, backtracking with Markov chains give problems of identifiability. Markov states have geometric distributions. Those functions are both log-concave and log-convex, and one has:

$$\forall t, u \in \mathbb{N}^*, \quad d_l(t-u)d_l(u) = d_l(t-1)d_l(1).$$

Here, the backtracking equation (5.10) has $t-1$ different solutions: *any* duration u between 1 and $t-1$ is optimal for state 1. So with a HMM composed of identical states, inference purely relies on the observation model. This fact has been noticed by Raphael (1999, Section 5) and Joder (2011, Section 3.4.2). The discussion in this section gives it more generality.

5.3.4 CONCLUSION

This section brings prescriptions on the design of occupancy distributions D_l, when used for constrained backtracking estimation. The main ones are identical to those derived in section 5.2. First, D_l should be unimodal and with mode located at l. Second, D_l should be log-concave. In addition, this sections shows that achieving coherency is easier when all events are identical $l_1 = l_2 = \ldots$. Interestingly, the same conclusion has been drawn in section 4.6.3 for the coherency of state estimation.

Finally, this section shows that characterizing coherency distributions D_l is easy in continuous-time but still an open question in discrete-time. The Poisson distributions $Po(l)$ is the only family we know that is coherent for constrained backtracking estimation.

6

APPLICATION TO AUDIO-TO-SCORE ALIGNMENT AND EXPERIMENTS

The goal of this chapter is to bring theory into real-world practice. First, we gather the theoretical results of this thesis and confront them with the durations models so far proposed in score-alignment literature. In particular, we will discover a special choice of duration model that fulfills all desired conditions and seems to appear in no prior score-alignment literature. We thus adopt this fully coherent model to design our HSMM-based algorithm. Then, section 6.3 evaluates this model for online alignment by confronting the algorithm with an alternative model that is common but incoherent. Quantitative and qualitative evaluation exhibit the benefits of time-coherency.

6.1 SUMMARY OF RESULTS AND LITERATURE REVIEW

Through the notion of time-coherency, chapters 3, 4 and 5 have suggested several prescriptions on duration model, that is to say the occupancy distributions $(D_l)_{l\geq 0}$ that model the random time spent on each event. We provide a summary and emphasize their logical implications. Then, we survey all explicit duration models we have found in the literature and discuss their temporal coherency.

6.1.1 PRESCRIPTIONS ON DURATION MODEL

Table 6.1 recapitulates all prescribed properties for each criterion and each estimation method: "online" refers to Forward estimator and "offline" refers to Viterbi estimator. Appendix B.1 contains every mathematical definitions and required background. We recall that $\uparrow$ means non-decreasing for each considered stochastic order st, lr. Dashed arrows are the logical implications which holds if $(D_l)_{l\geq 0}$ is a convolution semigroup[1].

In chapter 1, we outlined our research within several questions where Question 1.2 asked specifically for prescription on statistical relations between nominal duration l and its associated probability distribution D_l. Merging above results suggests that l should be mean, median *and* mode of D_l. It is hard to find probability laws supported on $\mathbb{R}_+$ such that these three quantities would exactly coincide, but those with low variance could approach asymptotically such identity. It has also been remarked that the inequality mean $\geq$ median $\geq$ mode holds for the majority of standard probabilities (Basu and DasGupta, 1997).

1. Logical implications related to convolution semigroups and Lévy processes are results of forthcoming chapter 7.

	Online estimation	Offline estimation
Criterion 1	$(D_l)_{l\geq 0}$ convolution semigroup $\text{mean}[D_l] = l$	no coherent family
Criterion 2	$\text{median}[D_l] = l$	$\text{mode}[D_l] = l$

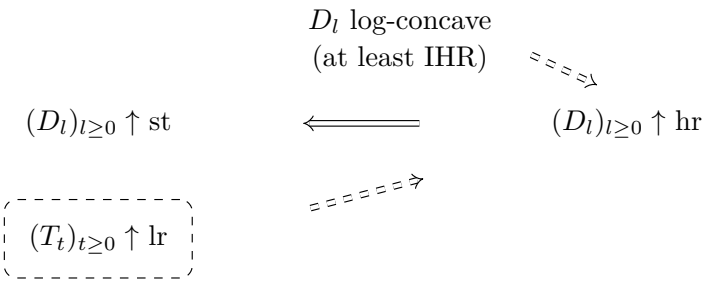


Table 6.1: Summary of obtained prescriptions on occupancy distributions D_l of a HSMM.

6.1.2 COMPARISON WITH DURATION MODELS IN THE LITERATURE

We confront our prescriptions on family $(D_l)_{l\in L}$ with a list of duration models found in the MIR literature. Refer to (Johnson et al., 1993) for definitions and properties of common probability laws.

- Nakamura et al. (2013) chooses geometric laws with mean fitted on l: $D_l \sim \mathcal{G}(1-1/l)$. The family $(D_l)_{l>1}$ respect criterion 2 but not criterion 1 as it is not a convolution semigroup.
- Cont (2010) chooses exponential laws with mean fitted on l: $D_l \sim \mathcal{E}(1/l)$. The family $(D_l)_{l>0}$ respect no criteria. It is not a convolution semigroup and the median of D_l is not l but $(l \ln 2)$.
- Takeda et al. (2007) choose log-normal distributions with constant shape parameter $\sigma > 0$ and log-scale fitted on l: $D_l \sim \ln\mathcal{N}(\ln l, \sigma)$. This family is not a convolution semigroup and no D_l is IHR.
- Joder et al. (2010b) use normal distributions with mean equal to l and standard deviation proportional to l: $D_l \sim \mathcal{N}(l, l^2\sigma^2)$ for some $\sigma > 0$. Raphael (2006) makes the same choice but sets the *variance* proportionally to l: $D_l \sim \mathcal{N}(l, l\sigma^2)$. Only the latter choice gives a convolution semigroup and it respects all prescriptions.
- Montecchio and Orio (2009) use negative binomial distributions with mean fitted on l: $D_l \sim NB(l(1-p)/p, p)$ for some $p \in (0,1)$. This defines a convolution semigroup that do respect our prescriptions for long enough l. This might explain why the authors do observe their inference working well with repeated events.

6.1.3 DURATION MODEL WITH BEST TIME-COHERENCY

Gaussian laws are a very popular choice of distribution. Our results also advocates such choice as $D_l \sim \mathcal{N}(l, l\sigma^2)$ matches all prescriptions. But Gaussian laws are continuous and supported on $(-\infty, \infty)$. So using them as occupancies of discrete-time HSMM involves truncation and discretization. The resulting distributions match prescriptions only approximately; for instance, they are no longer a convolution semigroup but still fulfill semigroup equation (3.2) approximately — and the longer the nominal length l, the better the approximation.

For discrete-time settings like our implementation, Poisson laws seem to be the best choice. They fulfill every criterion we have introduced to define coherency. Such distributions have already been suggested for semi-Markov models, but never for audio-to-score alignment. Our conclusion is that semi-Markovian states with Poisson laws should be the standard choice when modeling event for which a prior duration is available.

6.2 EVALUATION OF AUDIO-TO-SCORE ALIGNMENT

This goal of this chapter is to empirically assess influence of time-coherency on the precision of real-world score alignments. To do so, this section describes general background on score-alignment evaluation. Methodology is the standard procedure introduced by Cont et al. (2007).

6.2.1 ASSESSMENT METRICS

An alignment is a state-sequence $(S_1, S_2, \dots, S_T)$ where T is the observation length. Usually, the quality of alignments is measured on *onset times*. We recall that onset time t_i of event i is the first time this event occurs. Formally speaking, if estimated state-sequence is $(\hat{s}_1, \dots, \hat{s}_T)$, then estimated onset time of i is

$$t_i^e := \min\{t = 1, \dots, T \mid \hat{s}_t = i\} \qquad (\infty \text{ if the set is empty}).$$

To evaluate the precision of an alignment, the base quantity is the *onset error* e_i of each score event i. This quantity is defined as $e_i := |t_i^e - t_i^r|$, the absolute time lapse between the true onset time t_i^r as given by the ground truth and the estimated onset time t_i^e. Then, precision of alignments is quantitatively assessed with several *metrics* which are statistics computed on the set of all errors e_i of all alignments in dataset.

Misaligned rate An event is said to be misaligned if it has not been recognized close enough to its true value. This reads $e_i > \theta_e$ for some arbitrary threshold θ_e. *Misaligned rate* is the ratio of misaligned events over total events. We set $\theta_e = 300$ ms for our experiments.

Missed rate Events that exist in the score but are never reported by the alignment algorithm are called missed events. *Missed rate* is defined as the ratio of missed events

over total events. Note that missed rate is the limit case of misaligned rate as $\theta_e \to \infty$. This metric is meaningful for online alignments as sequential estimations may skip events. On the contrary, it is not well suited for offline alignments since they are usually constrained to go through each event.

Average imprecision The average imprecision μ_e is the mean of all absolute errors. While this assessment metric is usually chosen as the main measure of precision, an important drawback is its sensitivity to outliers. Another drawback is that μ_e cannot account for missed events and so would not penalize too much algorithms that skip many notes.

Median imprecision This is the median of all absolute errors (with $e_i = \infty$ for missed events). We think this metric is more suited one for quantitative evaluation of precision as it naturally discards outliers and does not depend on threshold value θ_e.

6.2.2 EVALUATION DATA MODEL

Evaluation requires triplets of audio record of a performance, a digital music score and a ground truth file that indicate onset times in the performance. In our audio-to-score task, running such evaluation has two main issues. First, publicly available databases are not very common and do not contain many records. Indeed, obtaining annotations is tedious as this task is usually done manually. There exists a international campaign called MIREX[2] that organizes yearly evaluation of several MIR tasks including score-alignment. Unfortunately, for this task the associated database is limited in size and musical diversity and part of annotations are not perfectly reliable.

Besides availability of data, we would like to emphasize a second issue with evaluation of automatic alignment. Most systems claim they do *audio-to-score* alignment, thought they are actually doing *audio-to-MIDI* alignment. Prior information differs between these two settings. The former requires a digital representation of music sheet, like a MusicXML file. The latter uses a MIDI file which is not faithful representation of music sheet. Although this difference is conceptually subtle, it might significantly influence results. Indeed, digital music score usually lack of some temporal information compared to MIDI files. A score file contains textual indications to performers that is not taken into account. In addition, it conveys implicit knowledge that is also discarded. On the contrary, a MIDI file encodes a typical interpretation. So it likely to contains more quantitative information on event timings than a digital score: for instance, abrupt tempo changes, progressive tempo slackening or accelerations, implicit rests and pauses. In practice, most evaluations are run with MIDI files as they are much more available than digital scores. We claim this situation does not always lead to a fair comparison with true audio-to-score setups like ours.

2. MIREX website: http://www.music-ir.org/mirex/wiki/MIREX_HOME.

6.3 EXPERIMENTS

This section describes a few experiments aiming at measuring the impact of *evolution model* on overall precision of automatic alignment. Does a time-coherent modeling of hidden position improve alignment algorithms in practice? To get insights we compare a coherent model with an incoherent model. This comparison is performed in two situations. First dataset is voluntarily composed of audio signals which are intrinsically unreliable. Second dataset feature a more balanced selection of diverse music pieces.

Our experimental approach is quite uncommon. A similar idea has been studied by Joder (2011, Section 4.4) in an online setting: HSMM are favorably compared to HMM. In the score-alignment literature, most experimental works compare different *observation model*, so as to find out the best audio features or assess the benefits of adding descriptors like onset detectors. We believe such trend is correlated with the predominance of offline over online algorithms. The former ones know where the alignment ends and when observation is over, whereas the latter ones have to constantly localize without such temporal clues. This may be why designing a good observation model is a more central preoccupation for offline approaches and ongoing research.

6.3.1 DESCRIPTION OF DATASETS

Two datasets are used for evaluation. We explain why they fit our needs even if content is limited in size. All audio records are lossless WAV files with sampling frequency 44 kHz. All digital music scores are accurate MusicXML files, thus preserving nominal durations and temporal indications in original music sheets. Ground truth files indicate true onset times.

Singing voice dataset Our first dataset is specific as it contains exclusively monophonic records of singing voice. Performances are excerpts of famous *chanson française* songs or opera arias. They are sung *a cappella* (without accompaniment) by two professional singers. Almost each song features two performances, one by a female singer and the other by a male signer. Table 6.2 lists each song. The dataset was recorded on purpose at Ircam laboratory for the needs of this experiment and a side project (Gong et al., 2015). Digital music scores have been encoded in MusicXML format[3], and such scores have been truncated accordingly for each performance. Annotations are handmade and double checked. They provide true onset time of each note written in the score. As usual, onset times of rest events are discarded for evaluation.

At first sight, this database should not be difficult for automatic alignment. For the selected music pieces, timing information provided by scores is faithful. Tempo remains globally constant along each performance and does not differ substantially from score tempo. Performances do not feature systematic temporal deviations that are not indicated by music scores. Nevertheless, singing voice is very challenging for

3. Audio files, MusicXML scores and annotations are available on request.

Title	Artist	Female	Male
Aria 'Belle Hermione' from *Cadmus et Hermione*	Lully, J.-B.		1:50
Aria 'Brillant Soleil' from *Les Indes galantes*	Rameau, J.-P.		0:49
Emmenez-moi	Aznavour, C.	2:09	1:00
Envole-moi	Goldman, J.-J.	1:46	1:08
Habanera from *Carmen*	Bizet, G.	1:16	1:23
J'irais où tu iras	Goldman, J.-J.	1:51	1:12
La Javanaise	Gainsbourg, S.	2:21	
Petite Marie	Cabrel, F.	0:58	1:00
Petite Marie (sung 2 tones lower)	Cabrel, F.	1:16	
Sensualité	Red, A.	2:05	0:51
Toute la musique	Hallyday, J.	0:53	0:53

Table 6.2: List of music pieces in the *Singing voice dataset*. Total length of these 18 performances is 27 minutes. Cases are left blank if there is no performance by the corresponding singer.

most alignment systems. Indeed, it is a prime example of *unreliable* signals. When a singer produces a note, the physical pitch mostly differs from the expected pitch written on score. This happens for many reasons such as natural vibrato of voice and interpretative intonations. As a result, observation model may be mislead. This trend is worsen by high timbral variability of voice: there is no universal observation model providing every aspect of vocal performances.

RWC sample Our second dataset is a subset of 26 pieces (2 hours of music) from the RWC[4] Classical Music Database. This set is listed in table 6.4. The RWC (Real World Computing) Music Database is a copyright-cleared music database built by the Real World Computing Partnership of Japan. The original Classical Music collection is a set of 50 pieces drawn from the common classical music repertoire. Records are complete and faithful performances of music scores.

Still, RWC database has two drawbacks. First, ground truth only contains partial annotations. Its does not provide the true onset time of each music event, but of each music beat (*i.e.*, onset positions $1, 2, 3, \ldots$). Second, original database does not contain genuine music scores but only MIDI files that are too messy and inaccurate to allow confident evaluations. This is why restricted experiments to a subset of 26 pieces for which we have found reliable MusicXML score files[5].

The goal of this data is to provide a fair overall evaluation of a score-alignment algorithm. The chosen music pieces bring about different challenges such as polyphony and features various instrumental groups, ranging from solo piano to symphonic orchestras. Many pieces exhibit extreme examples of temporal deviations that are not written in

4. RWC Music Database website: https://staff.aist.go.jp/m.goto/RWC-MDB/.
5. MusicXML score files are available on request.

the score. For instance, abrupt tempo changes of usual performances might be missing in music score because no quantified value of tempo has been written by the composer. An example of this is the Hungarian Dance by Brahms. In addition, many *fermata* (pauses) implicitly played in real-world performances are not indicated in scores by a long rest event. An example of this are the Twelve Variations by Mozart.

6.3.2 DESCRIPTION OF EXPERIMENTS

We compare performances of online audio-to-score alignment between two algorithms A and B. They are two variants of the HSMM-based `Antescofo` system designed by Cont (2008). Evolution model consists in modeling position with a linear semi-Markov chain as described in section 2.3.2. Each score event j with nominal duration l_j is mapped on a state j with occupancy distribution D_{l_j}. Observation model is as introduced in section 2.2.2 and further details are given in `Antescofo` reference articles (Cont, 2008, 2010).

The two algorithms have only *one* difference: the occupancy distributions D_l that model a event of nominal duration l.

Algorithm A: $D_l \sim \mathcal{G}(1 - 1/l)$. This means D_l is a geometric law with mean fitted on l.

Algorithm B: $D_l \sim Po(l)$. This means D_l is a Poisson law with mean fitted on l.

Algorithm A reduces to a simple HMM since occupancies have geometric laws. Algorithm B is a strict HSMM. Therefore, this comparison allows us to observe whether preferring HSMM to HMM is relevant and worth increased algorithmic complexity.

6.3.3 RESULTS AND DISCUSSIONS

Results of online alignments produced by the two algorithms are presented in table 6.3. They clearly indicate the superiority of HSMM model (algorithm B) over HMM model (algorithm A). Algorithm A is completely unable to manage the Singing voice dataset. It suffers from pitch fluctuations natural in voice that makes observations unreliable. Similar difficulties are observed with the RWC sample dataset where observation is inaccurate. Such bad results highlight the lack of temporal coherency of algorithm A.

On the contrary, algorithm B manages both datasets with fairly good precision. Its temporally coherent design makes it robust against blurred observations and outliers. This experiment shows how a simple modification of the evolution model may dramatically increase precision of automatic alignments. We would like to underline the fact that the observation model se use is *weak*. There is considerable room for its further perfection: spectral templates could be refined, additional descriptors indicating onset can be further employed, and more.

Further examination shows algorithm A output unstable alignments on both dataset. Many notes are missed and paths exhibit large fluctuations back and forth. Figure 6.1 shows a typical alignment sample demonstrating this phenomenon. On the contrary, alignment paths of algorithm B exhibit steady variations and very few backwards

moves. Such a behavior is highly desirable for real-world score following applications like `Antescofo`.

Singing voice dataset

	Misaligned rate	Missed rate	Average imprecision	Median imprecision
Algorithm A	91.5%	39.4%	3,286 ms	451 ms
Algorithm B	25.1%	11.4%	149 ms	68 ms

(∼3,400 annotated onsets)

RWC sample dataset

	Misaligned rate	Missed rate	Average imprecision	Median imprecision
Algorithm A	21.1%	3.2%	686 ms	129 ms
Algorithm B	6.7%	1.4%	133 ms	87 ms

(∼10,000 annotated onsets)

Table 6.3: Results of online alignment experiments described in section 6.3. Metrics are defined in section 6.2.1. Misaligned rate threshold is $\theta_e = 300$ ms.

6.4 CONCLUSION

This chapter has provided comparative experiments on online audio-to-score alignments. Results demonstrate the practical benefits in real-world applications of the theoretically grounded prescriptions we have found out. In addition, the reliable obtained results demonstrate that a coherent model of evolution may dramatically help a *weak* observation model like the one we have implemented. Indeed, this model is simplistic and does not employ external audio features, and the statistical description of the signal is all the more inaccurate as no learning step has been priorly done to optimize its parameters.

The coherent model we have suggested out has been implemented in the real-world alignment system called `Antescofo`. The scientific methodology only allows use to present quantitative results on annotated data. However, the numerous users of `Antescofo` all have reported qualitative improvements. The repertoire they have created for this software has offered us the unique possibility to observe in real-world situations the good performances of the improved algorithm. As a consequence of such additions, this repertoire has significantly increased in the past two years welcoming pieces involving highly polyphonic instruments, voice, and performing in concert halls world-wide without the presence of its inventors.

Index	Title	Composer	Length	Category
07	Brandenburg Concerto no.5 in D major, BWV.1050. 1st mvmt.	Bach, J. S.	10:01	Orchestra
11	Passacaglia and Fugue in C minor, BWV.582	Bach, J. S.	12:36	Organ
12	*The Musical Offering*, BWV.1079. Ricercare à 6	Bach, J. S.	6:27	Chamber
13	String Quartet no.19 in C major, K.465. 1st mvmt.	Mozart, W. A.	8:30	Chamber
22	Hungarian Dance no.5 in F$^{\sharp}$ minor	Brahms, J.	2:25	Pianos
	The Well-Tempered Clavier, Book I, BWV.846-847,	Bach, J. S.		Piano
25-A	no.1 in C major. Prelude		2:03	
25-B	*ibid.* Fugue		2:11	
25-C	no.2 in C minor. Prelude		1:31	
25-D	*ibid.* Fugue		1:54	
26	Piano Sonata in A major, K.331/300i. 1st mvmt.	Mozart, W. A.	10:00	Piano
27	Twelve Variations on "Ah vous dirai-je, Maman", K.265/300e	Mozart, W. A.	7:27	Piano
30	Nocturne no.2 in E$^{\flat}$ major, op.9-2	Chopin, F.	4:02	Solo
31	Étude in E major, op.10 no.3	Chopin, F.	4:16	Solo
32	Étude in F minor, op.25 no.2	Chopin, F.	1:49	Solo
33	Polonaise no.6 in A$^{\flat}$ major *Héroïque*, op.53	Chopin, F.	6:45	Piano
34	La campanella from *Grandes études de Paganini*, S.141	Liszt, F.	5:09	Piano
	Three Gymnopédies,	Satie, E.		Solo
35-A	no.1		3:49	
35-B	no.2		3:01	
35-C	no.3		2:46	
38	24 Caprices, op.1 no.24 in A minor	Paganini, N.	5:21	Violin
40	Méditation from *Thaïs*	Massenet, H.	5:06	Violin
42	Le Cygne from *Le Carnaval des Animaux*	Saint-Saëns, C.	2:31	Cello
43	Sicilienne op.78	Fauré, G.	4:09	Flute
44	The Flight of the Bumble Bee	Rimski-Korsakov	1:03	Flute
48	Der Lindenbaum from *Winterreise*, op.89/D.911	Schubert, F.	4:26	Vocal
49	Aria 'La donna è mobile...' from *Rigoletto*	Verdi, G.	2:11	Vocal

Table 6.4: List of music pieces in the *RWC sample dataset*. Total length of these 26 pieces is equal to 2 hours and 07 minutes.

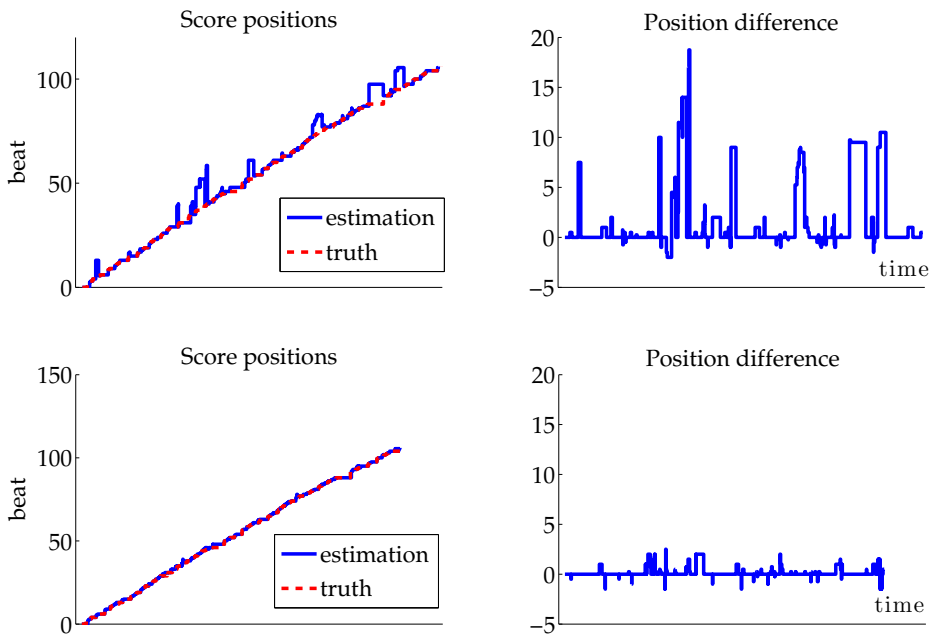


Figure 6.1: Automatic alignments of aria Belle Hermione by J. B. Lully. Top: algorithm A (HMM). Bottom: algorithm B (HSMM). True and estimated position are depicted on the left, and their difference is on the right.

7

LÉVY PROCESSES AND TOTAL POSITIVITY

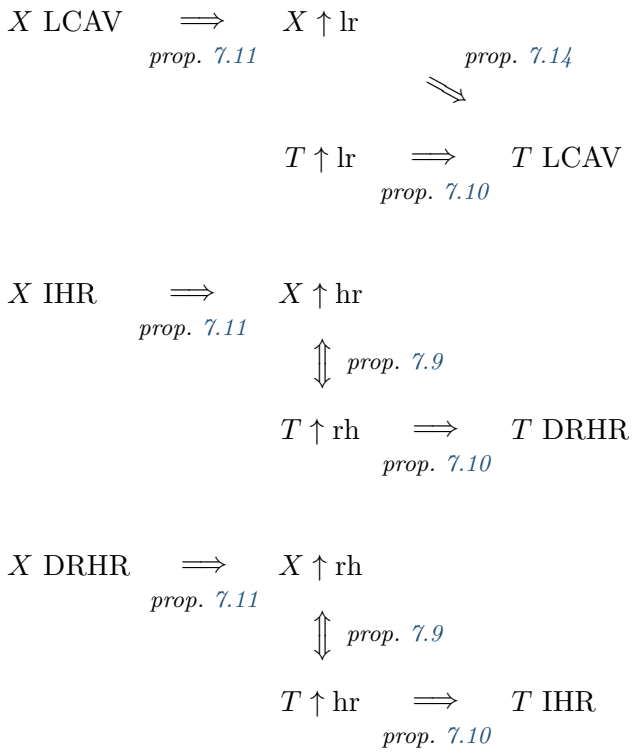
Chapter 3 exposed a relationship between linear semi-Markov chains and Lévy processes. If the occupancy distributions D_j of the semi-Markov chain S belong to a convolution semigroup, then the chain S is distributed as the first-passage times T of the underlying Lévy process. Moreover, chapter 4 drew the conclusion that a certain list of special properties are desirable for our modeling problem. Some of them are basic probabilistic properties, like mean, median, mode and unimodality. Other prescriptions are related to total positivity of order two (TP_2): distributional properties like log-concavity, IHR, DRHR reliability classes; stochastic orderings like lr, rh, hr orders. Those properties are either on the Lévy process X itself or its first-passage times T. For instance, section 4.6.2 promoted the lr-monotony of T, whereas section 4.4.3 promoted the hr-monotony of X.

As a result, our applicative study has raised a several theoretical. Now, we need to investigate which Lévy processes would bear those interesting properties. This is the goal of this chapter. Hereafter, section 7.1 begins with a clarification on the relationships between TP_2 properties of Lévy processes X and their first-passage times T. Then, section 7.2 focuses on the distributional properties of X itself. Finally, section 7.3 applies the obtained results to a selection of well-known Lévy processes, leading to new results on a few special functions.

7.1 RELATING LÉVY PROCESSES TO THEIR FIRST-PASSAGE TIMES

The main result of this section is summed up by the following schemes. Recall that $X = (X_l)_{l\geq 0}$ denotes a nonnegative Lévy process and $T = (T_t)_{t\geq 0}$ denotes its first passage times. Recall that $T_t = \inf\{l \geq 0 \mid X_l > t\}$. Their distributions are denoted $X_l \sim D_l$ and $T_t \sim M_t$.

Each scheme is about one stochastic order and its counterpart distributional property. It reveals logical implications that unconditionally hold between such properties. This is interesting has it diminishes the amount of work to achieve all our prescriptions. For instance, choosing a process X with log-concave distributions automatically provides the desired stochastic orders.



The schemes reveal a kind of "symmetry" between the Lévy process and the first-passage times process. Whereas the symmetry is exact for the hr, rh, it is not for the lr order: the lr monotony of X and T seems to be logically unrelated, although we show that both are sufficient condition for log-concavity of T.

Implications are detailed in separated propositions which are stated in next subsections. We begin with results that hold for more general classes of processes like Markov chains, and end with specific results for Lévy processes. Before this, we review existing results in literature.

7.1.1 REVIEW OF EXISTING RESULTS

Total positivity of first-passage times has been studied for Markov chains supported on $\mathbb{N}$ or its finite subsets. Most results are stated in (Kijima, 1998; Shaked and Shanthikumar, 1988; Bloch-Mercier, 2001; Shaked and Li, 2006). Related works are (Keilson, 1971; Shanthikumar, 1988; Brown and Chaganty, 1983). Sufficient conditions are obtained on the transition matrix (for discrete-time chains) or the infinitesimal generator (for continuous-time). These results can be specialized to Lévy processes supported on $\mathbb{N}$, since they are a particular kind of Markov chains. They provide sufficient conditions in terms of distributional properties on the Lévy measure.

Proposition 7.1 (Consequence of Bloch-Mercier, 2001, Proposition 2.3)**.** Let X be a non-decreasing Lévy process and M denotes its first-passage measures.

If the Lévy measure ν of X is discrete and non-decreasing on $\mathbb{N}^*$, then

$$M \uparrow \text{hr} \qquad \text{and} \qquad M \text{ is IHR.}$$

Using Markov chains approximations, results on Markov chains have been extended by Shaked and Shanthikumar (1988) to the bigger class of (non-decreasing) pure jump processes. This class contains continuous Lévy processes on $\mathbb{R}_+$.

Proposition 7.2 (Shaked and Shanthikumar, 1988, Proposition 5.1)**.** If the Lévy measure ν of X has a non-decreasing density on $(0, \infty)$,then

$$M \uparrow \text{hr} \qquad \text{and} \qquad M \text{ is IHR.}$$

Using a different approach based on TP_2 tools, the counterpart part result has been obtained by Kijima (1998).

Proposition 7.3 (Kijima, 1998, Theorem 4.1)**.** If the Lévy measure ν of X has a non-decreasing density on $(0, \infty)$,then

$$M \uparrow \text{hr} \qquad \text{and} \qquad M \text{ is IHR.}$$

The third counterpart result for the *lr* order can easily be deduced from (Shaked and Shanthikumar, 1988), though it is not explicitly written.

Proposition 7.4 (Consequence of Shaked and Shanthikumar, 1988, Theorem 1.2(iii))**.** If the Lévy measure ν of X is log-concave and supported $(0, \infty)$, then

$$M \uparrow \text{lr} \qquad \text{and} \qquad M \text{ is LCAV.}$$

Whereas the *hr- / rh*-monotonies of X and M are equivalent, the *lr*-monotony of M is not equivalent to the *lr*-monotony of X. Nevertheless, Keilson and Sumita (1982) have shown that log-concavity of Lévy measure implies both monotonies.

Proposition 7.5 (Keilson and Sumita, 1982)**.** If the Lévy measure ν of X is log-concave and supported on $(0, \infty)$ (or a discrete log-concave on $\mathbb{N}$), then

$$X \uparrow \text{lr}.$$

Such results consider stochastic ordering and its related distributional property as two unrelated consequences of the sufficient condition on Lévy measures. On the contrary, our main result sheds light on the logical implication between the two, and strictly extends sufficient conditions. Section 7.3 gives applications to several standard distributions. For instance, the result of Shaked and Shanthikumar (1988) proves the first-passage times of Gamma processes are IHR, whereas our result proves they are actually log-concave.

7.1.2 PROPOSAL FOR NON-DECREASING PROCESSES

For processes with non-decreasing sample paths, the relationship between the process itself and its first-passage times $T_t \sim M_t$ is simple.

Proposition 7.6 (First-passage measure of monotone processes)**.** Let $X = (X_l)_{l \geq 0}$ be a process on $\mathbb{R}_+$ and $(D_l)_{l \geq 0}$ denote its family of probability measures such that $X_l \sim D_l$.

Assume X has càdlàg and non-decreasing sample paths (almost surely). Assume that for $l, m, t \in \geq 0$, D_l and D_m have no atom at t.

Then, first-passage measure M_t is given by

$$\begin{aligned} M_t[l,m) &= D_l(t) - D_m(t) \qquad \text{, for all } m \geq l > 0 \\ &= \overline{D}_m(t^+) - \overline{D}_l(t^+) \\ &(= \overline{D}_m(t+1) - \overline{D}_l(t+1) \text{ if } X \text{ is supported on } \mathbb{N}). \end{aligned} \tag{7.1}$$

Its survivor distribution function is $\overline{M}_t(l) := M_t[l,\infty) = D_l(t)$.
Its cumulative distribution function is $M_t(l) := M_t(-\infty, l] = \overline{D}_l(t^+)$.
If the function $l \mapsto D_l(t)$ is absolutely continuous, M_t admits as probability density function $m_t(l) = \partial_l \overline{D}_l(t^+)$.

Proof. We use three results that can be found in (Veillette and Taqqu, 2010, Appendix A). First, as X is càdlàg and non-decreasing, one can prove that T is càdlàg and non-decreasing. Second,

$$\forall t, l \geq 0 \qquad \{X_l < t\} \subset \{T_t > l\} \subset \{X_l \leq t\}.$$

Therefore,

$$\forall t, l \geq 0 \qquad D_l(t^-) \leq \overline{M}_t(l^+) \leq D_l(t).$$

Third, fix $l \geq 0$. One can prove that for all $t \geq 0$ such that D_l as no atom at t (*i.e*, almost everywhere in $\mathbb{R}_+$), the relationship is valid:

$$\{T_t > l\} = \{X_l < t\},$$

which gives

$$\overline{M}_t(l^+) = D_l(t^-),$$

or equivalently,

$$M_t(l) = \overline{D}_l(t).$$

All remaining equalities may be deduced from this identity, and the fact that $l \mapsto D_l(t)$ is necessarily a right-continuous function. □

Remark 7.7. If X is supported on $\mathbb{N}$, then all atoms are on $\mathbb{N}$. So the formulas of last proposition are true for all $t \notin \mathbb{N}$. But necessarily, M_t is piecewise constant in t:

$$\forall t \geq 0, \qquad M_t = M_{\lfloor t \rfloor}.$$

Therefore, the formulas are true for $t \in \mathbb{N}$. In addition, it suffices to study $(T_t)_{t\in\mathbb{N}}$ and $(M_t)_{t\in\mathbb{N}}$.

Remark 7.8. If the process $(X_l)_{l\geq 0}$ has strictly increasing paths, one can show that T has strict increasing paths too and the formula is always valid.

Remark. In general, the set of atom of any probability distribution is at most countable. If X is a Lévy process with no drift, then the atoms of D_l are independent of l.

Then equivalence between the rh and hr monotony of X and T is straightforward.

Proposition 7.9. Let $X = (X_l)_{l\in I}$ be a non-negative stochastic process with (almost surely) càdlàg and non-decreasing paths, indexed on $I \subset \mathbb{R}$. Let $T = (T_t)_{t\in\mathbb{R}_+}$ denote its first-passage times $T_t \overset{\text{def}}{=} \inf\{l \in I \mid X_l \geq t\}$.

The following propositions are equivalent:

(i) T is non-decreasing in the hazard rate (hr) ordering,

(ii) X is non-decreasing in the reverse hazard rate (rh) ordering.

The following propositions are equivalent:

(i) T is non-decreasing in the reverse hazard rate ordering,

(ii) X is non-decreasing in the hazard rate ordering.

Proof. Assume first-passage measures M_t have no atom. We proof the first equivalence. Owing to previous proposition, claim (i) is equivalent to $\overline{D}_l(t^+)$ being TP_2 in (l,t) whereas claim (ii) is equivalent to $\overline{D}_l(t)$ being TP_2 in (t,l). As the total posivity relationship is reflexive and preserved by convergence in distribution, the two propositions are equivalent. The proof of the second equivalence is similar (with D instead of $\overline{D}$).

As the set of atoms of a probability measure is at most countable, and as TP_2 orders are preserved by convergence in distribution, a limit argument may be used to prove the general case. □

7.1.3 PROPOSAL FOR CONTINUOUS-TIME MARKOV CHAINS

We prove the implication between lr-monotony and log-concavity of first-passage measures M actually holds for more general Markov chains. Continuous-time Markov chains are defined in appendix A.2 together with the subclass of uniformizable chains. Remember that discrete Lévy processes (with no drift) belongs to such class of time-homogeneous Markov chains. Next proposition is one of the main results of this chapter. It provides the implication for Markov chains that have non-decreasing paths.

In addition, the proposition gives the counterparts for the hr and rh orders. Note the hr result extends (Bloch-Mercier, 2001, Proposition 2.3) in two ways: it has weaker conditions, and holds for time-inhomogeneous chains. In the latter case, the hr and the rh results each requires an additional hypothesis that is the opposite one of each other: they hold simultaneously only for time-homogeneous chains, like the lr result does.

Proposition 7.10. Let $X = (X_t)_{t\geq 0}$ be a uniformizable continuous-time Markov chain on $\mathbb{N}$, with infinitesimal generators $(\mathbf{Q}_t)_{t\geq 0}$. Let $M = (M_n)_{n\in\mathbb{N}}$ denote the first-passage measures of X.

1. Assume that:

   (i) X has non-decreasing paths: $q_t(i,j) = 0$ if $i > j$,

   (ii) the matrices $(\mathbf{Q}_t)_{t\geq 0}$ are stochastically non-increasing:

$$\forall t_1, t_2 \geq 0, \qquad t_1 \leq t_2 \quad \Longrightarrow \quad q_{t_1}(n,\cdot) \underset{\text{st}}{\geqslant} q_{t_2}(n,\cdot).$$

   If $(M_n)_{n\in\mathbb{N}}$ is non-decreasing for the reverse hazard rate ordering, then all M_n are DRHR.

2. Assume that:
   (i) X has non-decreasing paths: $q_t(i,j) = 0$ if $i > j$,
   (ii) the matrices $(\mathbf{Q}_t)_{t\geq 0}$ are stochastically non-decreasing:
   $$\forall t_1, t_2 \geq 0, \qquad t_1 \leq t_2 \quad \Longrightarrow \quad q_{t_1}(n,\cdot) \underset{\text{st}}{\leqslant} q_{t_2}(n,\cdot),$$
   (iii) the matrix $\mathbf{Q}_t$ is stochastically monotone for all $t \geq 0$:
   $$m \leq n \quad \Longrightarrow \quad q_t(m,\cdot) \underset{\text{st}}{\leqslant} q_t(n,\cdot).$$
   If $(M_n)_{n\in\mathbb{N}}$ is non-decreasing for the hazard rate ordering, then all M_n are IHR.
3. Assume that:
   (i) X has non-decreasing paths: $q_t(i,j) = 0$ if $i > j$,
   (ii) X is time-homogeneous: $\mathbf{Q}_t = \mathbf{Q}$ for all $t \geq 0$,
   (iii) the matrix $\mathbf{Q}$ is stochastically monotone:
   $$m \leq n \quad \Longrightarrow \quad q(m,\cdot) \underset{\text{st}}{\leqslant} q(n,\cdot).$$
   If $(M_n)_{n\in\mathbb{N}}$ is non-decreasing for the likelihood ratio (lr) ordering, then all M_n are log-concave.

Proof. Define the cumulative and survivor matrices:
$$Q_l(i,j) := \sum_{k\leq j} q_l(i,k), \qquad \overline{Q}_l(i,j) := \sum_{k\geq j} q_l(i,k).$$
Any infinitesimal generator checks
$$\forall n \in \mathbb{N}, \quad \sum_{k\in\mathbb{N}} q_t(n,k) = 0. \tag{7.2}$$

All proofs are based on the so-called *forward Kolmogorov equation*. Restricting to uniformizable Markov chains guarantees the existence of infinitesimal generators, the validity of forward Kolmogorov equation, and the unconditional validity of Proposition 7.6. With hypothesis (i), the Kolmogorov equation reads
$$\partial_l d_l(n) = \sum_{u=0}^{n} q_l(u,n) d_l(u).$$
Equation (7.2) and hypothesis (i) implies $\overline{Q}(n,n) = 0$. Summing the equality above over n gives
$$\partial_l \overline{D}_l(n) = \sum_{u=0}^{n-1} \overline{Q}_l(u,n) d_l(u).$$
Equation (7.2) implies $Q_t(u,n) = -\overline{Q}_t(u,n+1)$. Combining with $\partial_l D_l(n) = -\partial_l \overline{D}_l(n+1)$ gives
$$\partial_l D_l(n) = \sum_{u=0}^{n} Q_l(u,n) d_l(u).$$

Applying Abel's transformation gives

$$\partial_l D_l(n) = Q_l(n,n)D_l(n) + \sum_{u=0}^{n-1}[Q_l(u,n) - Q_l(u+1,n)]D_l(u).$$

[rh order] Assume $(M_n)_{n\in\mathbb{N}}$ is rh-monotone. This means for all $n, u \in \mathbb{N}$,

$$u \le n \quad \Longrightarrow \quad l \mapsto \overline{D}_l(u) \underset{\text{TP}}{\leqslant} l \mapsto \overline{D}_l(n).$$

Fix $n \in \mathbb{N}^*$. From the decomposition

$$\frac{d_l(u)}{\overline{D}_l(u)} = \frac{\overline{D}_l(u) - \overline{D}_l(u+1)}{\overline{D}_l(u)} = 1 - \frac{\overline{D}_l(u+1)}{\overline{D}_l(u)},$$

we obtain $u \in \mathbb{N}$, $l \mapsto d_l(u) \underset{\text{TP}}{\leqslant} l \mapsto \overline{D}_l(u)$. and by transitivity,

$$u \le n \implies \quad d_l(u) \underset{\text{TP}}{\leqslant} \overline{D}_l(n).$$

the Kolmogorov equation reads

$$\partial_l \overline{D}_l(n)/\overline{D}_l(n) = \sum_{u=1}^{n-1} \overline{Q}_l(n,u)\frac{d_l(u)}{\overline{D}_l(n)}.$$

By hypothesis (ii), $q_t(n,\cdot)$ is non-increasing in t for the *st* order. This reads $\overline{Q}_t(n,\cdot)$ being non-increasing in t. In addition, it is non-negative, and the TP-relationship above says $\frac{d_l(u)}{\overline{D}_l(n)}$ is non-increasing in l. As a sum of products of non-negative and non-increasing functions, the left-hand side quantity is non-increasing with l. As it equals $\partial_l \log M_n(l)$, we conclude that M_n is DRHR.

[hr order] Assume $(M_n)_{n\in\mathbb{N}}$ is hr-monotone. This means for all $u \le n \in \mathbb{N}$, $\frac{D_l(u)}{D_l(n)}$ is non-increasing in l. Since $M_0 = \delta_0$, M_0 is IHR. Let n be in $\mathbb{N}$. To prove M_{n+1} is IHR, take $l, t \ge 0$ and assume $l \le t$. it suffices to show $\partial_l \log \overline{M}_{n+1}(l) \ge \partial_l \log \overline{M}_{n+1}(t)$. The Kolmogorov equation reads

$$\partial_l D_l(n)/D_l(n) = \sum_{u=0}^{n} Q_l(u,n)d_l(u).$$

By hypothesis (ii), $q_l(u,\cdot) \underset{\text{st}}{\leqslant} q_t(u,\cdot)$. This implies $Q_l(u,n) \ge Q_t(u,n)$, so

$$\ge \sum_{u=0}^{n} Q_t(u,n)d_l(u).$$

Using Abel's transform,

$$\sum_{u=0}^{n} Q_t(u,n)d_l(u) = Q_t(n,n) + \sum_{u=0}^{n-1}[Q_t(u,n) - Q_t(u+1,n)]\frac{D_l(u)}{D_l(n)}.$$

The hr-monotony gives $\frac{D_l(u)}{D_l(n)} \ge \frac{D_t(u)}{D_t(n)}$. Hypothesis (iii) gives $q_t(u,\cdot) \underset{\text{st}}{\leqslant} q_t(u+1,\cdot)$ which implies $Q_t(u,n) - Q_t(u+1,n) \ge 0$. Therefore,

$$\partial_l D_l(n)/D_l(n) \ge Q_t(n,n) + \sum_{u=0}^{n-1}[Q_t(u,n) - Q_t(u+1,n)]\frac{D_t(u)}{D_t(n).}$$

With the same Abel's transform, we can see the right-hand side quantity equals $\partial_l D_t(n)/D_t(n) = \partial_l \log \overline{M}_n(t)$. Therefore,

$$\partial_l \log \overline{M}_n(l) \geq \partial_l \log \overline{M}_n(t).$$

[lr order] Assume $(M_n)_{n\in\mathbb{N}}$ is lr-monotone. As $m_{n+1}(l) = -\partial_l D_l(n)$, this means for all $u \leq n \in \mathbb{N}$, $\frac{\partial_l D_l(u)}{\partial_l D_l(n)}$ is non-increasing in l. As $M_0 = \delta_0$, M_0 is log-concave and the result is true for $n = 0$. Let n be in $\mathbb{N}$. We have to prove that $\partial_l \log m_{n+1}(l)$ is non-decreasing in l. Its expression is

$$\begin{aligned}\partial_l m_{n+1}(l) &= -\partial_l \partial_l D_l(n),\\ \partial_l \log m_{n+1}(l) &= -\partial_l \partial_l D_l(n)/\partial_l D_l(n).\end{aligned}$$

Kolmogorov equation reads (with hypothesis (ii) of time-homogeneity)

$$\partial_l D_l(n) = Q(n,n) + \sum_{u=0}^{n-1} [Q(u,n) - Q(u+1,n)] D_l(u).$$

Derivating further this equation gives

$$\begin{aligned}\partial_l \partial_l D_l(n) &= \sum_{u=0}^{n-1} [Q(u,n) - Q(u+1,n)] \partial_l D_l(u),\\ \partial_l \partial_l D_l(n)/D_l(n) &= \sum_{u=0}^{n-1} [Q(u,n) - Q(u+1,n)] \frac{\partial_l D_l(u)}{\partial_l D_l(n)}\end{aligned}$$

By hypothesis (iii), $q(u,\cdot) \underset{\text{st}}{\leqslant} q(u+1,\cdot)$. This reads $Q(u,n) - Q(u+1,n) \geq 0$. So the left-hand side is non-increasing in l as a sum of non-increasing functions. □

Remark. All conditions are fulfilled by Lévy processes on $\mathbb{N}$. In particular, condition 1.(iii) is true: since $q_t(n+1,\cdot) = q_t(n,\cdot+1)$, the infinitesimal generator is always stochastically monotone.

Remark. All results still hold for discrete-time Markov chains $(X_n)_{n\in\mathbb{N}}$, with the same hypotheses on transition matrices $\mathbf{P}_n$ instead of infinitesimal generators $\mathbf{Q}_t$.

7.1.4 PROPOSAL FOR LÉVY PROCESSES

Results on stochastic ordering of the Lévy process itself are specific to this class of process. However, result on first-passage-time process present "mirrored" logical are established by approximations with discrete Lévy processes and deduced from the general result on Markov chains in previous section.

Ageing Properties of Lévy process

Next proposition holds for the class of delayed additive process defined in appendix A.4.3, which is a little more general than Lévy processes. This proposition relates distributional properties to their corresponding stochastic orderings: when a subset of distributions have one distributional property, then automatically they get the corresponding stochastic order. The result is very illuminating, although almost very simple to prove thanks to the tools of total positivity.

Proposition 7.11. Let $D = (D_l)_{l\in I}$ be a delayed additive process of probability measures, indexed on $I \subset \mathbb{R}_+$. Assume its increments are supported on $\mathbb{R}_+$ (non-negative increments). Let L denote a subset of I.

(i) If the measures D_l are LCAV (or discrete-) for all $l \in L$, then $(D_l)_{l\in L}$ is increasing in the likelihood ratio order.
(ii) If the measures D_l are IHR (or discrete-) for all $l \in L$, then $(D_l)_{l\in L}$ is increasing in the hazard rate order.
(iii) If the measures D_l are DRHR (or discrete-) for all $l \in L$, then $(D_l)_{l\in L}$ is increasing in the reverse hazard rate order.

Proof. Let l, l' be in L and assume there exists $h \geq 0$ such that $l' = l+h$. By definition D of be being additive semigroup with non-negative increments, there exists a distribution $D_{l,h}$ supported on $\mathbb{R}_+$ such that $D_{l+h} = D_l * D_{l,h}$. This condition on the support means that $\delta_0 \underset{\text{lr}}{\leqslant} D_{l,h}$. Therefore, proposition B.41 in the appendix immediately gives;

(i) D_l being LCAV implies $D_l \underset{\text{lr}}{\leqslant} D_l * D_{l,h}$.
(ii) D_l being IHR implies $D_l \underset{\text{hr}}{\leqslant} D_l * D_{l,h}$.
(iii) D_l being DRHR implies $D_l \underset{\text{rh}}{\leqslant} D_l * D_{l,h}$.

And similarly in the discrete case. □

This proposition still holds if properties are *partially* fulfilled. Refer to appendix B.1 for the precise definitions of partial distributional properties.

Proposition 7.12. Let $D = (D_l)_{l\in I}$ be an additive process of finite measures, indexed on $I \subset \mathbb{R}_+$. Assume all D_l are supported on $\mathbb{R}_+$. Let L denote a subset of I.

(i) Let a be in $(0, \infty)$. If the measures D_l are LCAV (resp. IHR, DRHR) on $[0, a]$ for all $l \in L$, then $(D_l)_{l\in L}$ is non-decreasing in the likelihood ratio (resp. hazard rate, reverse hazard rate) order on $[0, a]$.
(ii) Let a be in $\mathbb{N}^*$. If the distributions d_l are discrete LCAV (resp. IHR, DRHR) on $\{0, \dots, a\}$ for all $l \in L$, then $(D_l)_{l\in L}$ is non-decreasing in the likelihood ratio (resp. hazard rate, reverse hazard rate) order on $\{0, \dots, a+1\}$.

Proof. As $D_{l+h} = D_{l,l+h} * D_l$ and all distributions are supported on $\mathbb{R}_+$, then the convolution only depends on their values on $[0, a]$. $\tilde{D}_{l+h} = D_{l,l+h} * \tilde{D}_l$ where $\tilde{D}_l$ is the restriction of D_l on any interval $[0, a]$ with $a > 0$. So the restrictions $(\tilde{D}_l)_{l\in I}$ still define an additive semigroup, and previous proposition applies.

Things are similar in the discrete case. Remember that D_l being discrete LCAV on $\{0, \dots, a\}$ means that sequences $(d_l(n))_{n\in\{0,\dots,a+1\}}$ are log-concave (and similarly for IHR, DRHR). □

Ageing Properties of First-Passage-Time Process

Next proposition is about first-passage processes. It relates stochastic orders to their corresponding distributional properties. Note the implication holds in a reverse fashion compared to the one that holds for Lévy processes (proposition 7.11).

Proposition 7.13. Let $M = (M_t)_{t\geq 0}$ be the first-passage measures of a non-negative Lévy process X.

(i) If M is non-decreasing in the lr ordering, then all M_t are log-concave.

(ii) If X is non-decreasing in the lr ordering, then all M_t are log-concave.

(iii) If X is non-decreasing in the rh ordering, then all M_t are IHR.

(iv) If X is non-decreasing in the hr ordering, then all M_t are DRHR.

Remark. The weaker stochastic orders rh and hr are equivalent between X and M, whereas the lr is not. However, either the lr-monotony of X or M implies the log-concavity of M.

When the Lévy process X is strictly increasing, it is absolutely continuous and X_l admits a density d_l. Let ν denote its Lévy measure. Its survivor distribution $\overline{\nu}(x) = \nu[x, \infty)$ is well-defined on $x \in (0, \infty)$. It has been proved by Meerschaert and Scheffler (2008, Theorem 3.1) and Vellaisamy and Kumar (2011, Theorem 2.1) that in such case, M_t are absolutely continuous and admits a density m_t that fulfills the following equation:

$$\forall l > 0, t > 0 \qquad m_t(l) = d_l * \overline{\nu}(t).$$

This equation corresponds to the Kolmogorov equation that holds for (almost all) Markov processes. Unfortunately, as the Lévy measure of such processes is infinite, this Kolmogorov equation cannot be further differentiated. As a result, the higher order derivatives of M_t cannot be expressed similar convolution-type formula. However, discrete Lévy processes with no drift do admit the same Kolmogorov equation, where densities are replaced by probability mass functions. Moreover, discrete Lévy measures are always finite, so higher derivatives can be also expressed with convolutions. Indeed, such Lévy processes are non-decreasing Markov chains on $\mathbb{N}$. Further details are given in Proposition A.26 in the appendix.

Thus, our strategy has two steps. First, proving the proposition for discrete Lévy processes with no drift. Then, generalizing the result using the powerful tool of *Poisson mixtures* Γ_λ. This operation maps every increasing Lévy process a discrete Lévy processes with no drift, while preserving properties related to total positivity — refer to appendix B.5 for full background on Poisson mixtures. The following schemes sum up this proof outline. Hereafter, X^λ denotes the discrete process obtained by Poisson mixture with X, and M^λ denotes the first-passage measures of X^λ (definitions are explained in Proposition 7.15).

Proof of Proposition 7.13.

Claim (i)

$$\begin{array}{ccc} (X_t)_{t\geq 0} \uparrow \mathrm{lr} & & M_t \text{ log-concave} \\ \Big\Updownarrow \text{ prop. B.80} & & \Big\Uparrow \text{ prop. 7.15, 4.} \\ \forall \lambda > 0, \quad (X_t^\lambda)_{t\geq 0} \uparrow \mathrm{lr} & \underset{\text{prop. 7.14}}{\Longrightarrow} & \forall \lambda > 0, \quad M_n^\lambda \text{ discrete log-concave} \end{array}$$

Claim (ii)

$$\begin{array}{ccc} (M_t)_{t\geq 0} \uparrow \text{lr} & & M_t \text{ log-concave} \\ \Big\Downarrow \text{ prop. 7.15, 3.} & & \Big\Uparrow \text{ prop. 7.15, 4.} \\ \forall \lambda > 0, \quad (M_n^\lambda)_{n\in\mathbb{N}} \uparrow \text{lr} & \underset{\text{prop. 7.10}}{\Longrightarrow} & \forall \lambda > 0, \quad M_n^\lambda \text{ discrete log-concave} \end{array}$$

Claim (iii-iv)

$$\begin{array}{ccc} (M_t)_{t\geq 0} \uparrow \text{hr (resp. rh)} & & M_t \text{ IHR (resp. DRHR)} \\ \Big\Updownarrow \text{ prop. 7.15, 1-2.} & & \Big\Uparrow \text{ prop. 7.15, 4.} \\ \forall \lambda > 0, \quad (M_n^\lambda)_{n\in\mathbb{N}} \uparrow \text{hr (resp. rh)} & \underset{\text{prop. 7.10}}{\Longrightarrow} & \forall \lambda > 0, \quad M_n^\lambda \text{ discrete IHR (resp. DRHR)} \end{array}$$

□

So now we only have to prove the intermediate implications. As discrete Lévy processes with no drift are time-homogeneous Markov chains, proposition 7.10 already proves some of them. The other ones are proved by forthcoming propositions 7.14 and 7.15.

Case of discrete-state Lévy processes Next proposition holds for discrete Lévy processes.

Proposition 7.14. Let $M = (M_t)_{t\in\mathbb{R}_+}$ be the first-passage measures of a delayed Lévy process X supported on $\mathbb{N}$.
If X is non-decreasing for the local ordering, then all M_t are log-concave.

Proof. As explained in section, a discrete Lévy process is necessarily a Compound Poisson process $X = (cp(l, F))_{l\geq 0}$. Its compounding measure F is supported on $\mathbb{N}$. Let f denote its pmf and let d_l denote the pmf of X_l. Proposition A.26 in the appendix gives the following forward Kolmogorov-like equation

$$\frac{\partial_l m_n(l)}{m_n(l)} = \frac{f * (\overline{F} - \delta) * d_l}{(\overline{F} - \delta) * d_l} - 1.$$

Assume $(D_l)_l$ is non-decreasing for the likelihood ratio ordering. Proposition 7.17 — stated in next section — tells that necessarily,

$$f \underset{\text{lr}}{\leqslant} f * f.$$

This implies $f \underset{\text{hr}}{\leqslant} f * f$, *i.e.*, $\overline{F} \underset{\text{TP}}{\leqslant} f * (\overline{F} - \delta)$. As $f(0) = 0 = [\overline{F} - \delta](0) = [f * (\overline{F} - \delta)](0)$, this implies

$$\overline{F} - \delta \underset{\text{TP}}{\leqslant} f * (\overline{F} - \delta).$$

Therefore, Lemma B.48 in the appendix can be applied. It directly gives that $\partial_l m_n(l)/m_n(l)$ is non-increasing with respect to l. This means that m_n is log-concave. □

Extension to continuous Lévy processes Now we extend previous result to continuous Lévy processes, using Poisson mixtures as in Proposition B.74 in appendix to get approximations by discrete Markov chains.

Proposition 7.15 (Discrete-continuous relationships)**.** Let $X = (X_l)_{l\geq 0}$ be a Lévy process on $\mathbb{R}_+$ and D_l be the distribution of X_l.
Let $T = (T_t)_{t\geq 0}$ denote the first-passage times of X.
Define $X^\lambda = (X_l^\lambda)_{l\geq 0}$ as the Poisson mixture:

$$X_l^\lambda \sim D_l^\lambda \overset{\text{def}}{=} \sum_{n\in\mathbb{N}} \Gamma_\lambda[\mathrm{d}D_l](n)\delta_{\frac{n}{\lambda}}.$$

Define $T^\lambda = (T_n^\lambda)_{n\in\mathbb{N}}$ as the first-passage times of X^λ.

1. The following propositions are equivalent:
   (i) T is non-decreasing in the hazard rate ordering.
   (ii) X is non-decreasing in the reverse hazard rate ordering.
   (iii) X^λ is non-decreasing in the reverse hazard rate ordering for all $\lambda > 0$.
   (iv) T^λ is non-decreasing in the hazard rate ordering for all $\lambda > 0$.

2. The following propositions are equivalent:
   (i) T is non-decreasing in the reverse hazard rate ordering.
   (ii) X is non-decreasing in the hazard rate ordering.
   (iii) X^λ is non-decreasing in the hazard rate ordering for all $\lambda > 0$.
   (iv) T^λ is non-decreasing in the reverse hazard rate ordering for all $\lambda > 0$.

3. If T is non-decreasing in the likelihood ratio ordering, then T^λ is non-decreasing in the likelihood ratio ordering, for all $\lambda > 0$.

4. If T_n^λ is log-concave (resp. IHR, DRHR) for all $\lambda > 0$ and $n \in \mathbb{N}$, then T_t is log-concave (resp. IHR, DRHR) for all $t > 0$.

Remark. unlike X and X^λ, the distribution of T^λ is the not a Poisson mixture corresponding to M. Actually, each T_n^λ distribution is a mixture of all T_t distributions. However, we are still able transfer some distributional properties and stochastic ordering between T and T^λ, thanks to tools from total positivity.

From now one, let M_t denote the measure of T_t, M_n^λ the measure of T_n^λ.

Proof. As X is a Lévy process on $\mathbb{R}$, its Poisson mixture X^λ is a Lévy process on $\mathbb{N}$. D_l the measure of X_l, D_l^λ the measure of X_l.

[Claims 1. and 2.] Proposition 7.9 gives (i) $\iff$ (ii) and (iii) $\iff$ (iv). The preservation properties of Poisson mixtures (Proposition B.80 in the appendix) gives (ii) $\iff$ (iii).

[Claim 3.] Let n be in $\mathbb{N}^*$. The survivor distribution of a Poisson mixture is given by Proposition B.73 in the appendix:

$$\overline{D}_l^\lambda(n) = \lambda \int_0^\infty e^{-\lambda t} \frac{\lambda^{n-1}}{(n-1)!} t^{n-1} \overline{D}_l(t)\,\mathrm{d}t.$$

As $M_n^\lambda(l) = \overline{D}_l^\lambda(n)$, M_n^λ is a mixture of M_t distributions:

$$M_n^\lambda(l) = \int_0^\infty e^{-\lambda t}\frac{\lambda^n}{\Gamma(n)}t^{n-1}M_t(l)\,\mathrm{d}t = \int_0^\infty M_t(l)g_n(t)\,\mathrm{d}t,$$

where the mixing distribution is a Gamma law $\mathcal{G}(n, 1/\lambda)$ whose pdf is

$$g_n(t) \overset{\text{def}}{=} e^{-\lambda t}\frac{\lambda^n}{\Gamma(n)}t^{n-1}.$$

It is known that $\mathcal{G}(n, 1/\lambda) \underset{\text{lr}}{<} \mathcal{G}(n+1, 1/\lambda)$. Indeed, their likelihood ratio equals $\frac{\lambda}{n+1}t$, so it increases with respect to t. Suppose M is non-decreasing in the likelihood ratio order. Therefore, we can apply Lemma B.42 (in the appendix) about mixtures to obtain $M_n^\lambda \underset{\text{lr}}{\leqslant} M_{n+1}^\lambda$.

[Claim 4.] For any t in $(0, \infty)$, it is known that the sequence of Gamma laws $\mathcal{G}(n, t/n)$ converges in distribution to the Dirac law δ_t supported on $\{t\}$. Since X is a Lévy process, all $(D_l)_{l\geq 0}$ have identical continuity points (Veillette and Taqqu, 2010, Section 2). Let t be such a continuity point. Then, Proposition 7.6 tells that

$$\forall l \geq 0, \quad M_t(l) = \overline{D}_l(t).$$

Therefore, for all $l \geq 0$, $|M_t(l)| \leq 1$ and t is a continuity point of $t \mapsto M_t(l)$. This means that the latter function is bounded, and continuous δ_t-almost everywhere since the support of δ_t is $\{t\}$. For all $l \geq 0$, the Portmanteau lemma implies that the mixture $M_n^{n/t}(l)$ converges to $\overline{D}_l(t)$. This proves that $(M_n^{n/t})_{n\in\mathbb{N}^*}$ converges in distribution to M_t. In the appendix, Proposition B.30 explains how log-concavity, IHR, DRHR properties are preserved by convergence in distribution.

As are result, we have proven the statement for any t that is a continuity point. To extend the statement on for all $t \in \mathbb{R}_+$, it suffices to remark that such continuity points are a dense set of $\mathbb{R}_+$ and $t \mapsto M_t$ is stochastically right-continuous in distribution since T is a càdlàg. So, again, preservation by convergence in distribution provides the result. □

7.1.5 SPECIAL CASE OF COMPOUND POISSON PROCESSES

This section presents further conditions for stochastic ordering of compound Poisson process. This subclass of Lévy processes is described in appendix A.4.2.1: it encompasses all discrete Lévy processes.

A compound Poisson process $(cp(l, D))_{l\geq 0}$ is a convolution semigroup indexed on $\mathbb{R}_+$ which is built out of some compounding measure D. As a probability measure, F induces a convolution semigroup $(D^{*n})_{n\in\mathbb{N}}$, indexed on the discrete set $\mathbb{N}$, that is composed of its convolution powers: $D^{*0} = \delta_0, \quad D^{*n+1} = D^{*n} * D$. Interestingly, if the discrete convolution semigroup (D^{*n}) admits some stochastic ordering, then this order is automatically transferred to the compound Poisson semigroup $(cp(l, D))_{l\geq 0}$. This fact is stated in (Keilson and Sumita, 1982, Corollary 4.9) for the *lr* ordering, but its proof outline holds for the other orderings.

Proposition 7.16 (Keilson and Sumita, 1982)**.** Let D be a probability measure supported on $\mathbb{R}_+$. Let $C \in \{lr, hr, rh\}$ be a stochastic ordering.

If $(D^{*n})_{n\in\mathbb{N}}$ is non-decreasing in C ordering, then the compound Poisson process $(cp(l, D))_{l\geq 0}$ is non-decreasing in C ordering.

Remark. Last result is a sufficient condition for the monotony of a Lévy process. Empirically, we observe an interesting phenomenon: for all standard discrete Lévy processes, monotony could be potentially explained by this condition. For instance, take a Negative Binomial semigroup $NB(r, p)$. The compounding distribution is the logarithmic law $f(n) \propto p^n/n$. Then, numerical computations make us conjecture that $f \underset{\text{lr}}{\leqslant} f * f \underset{\text{lr}}{\leqslant} f * f * f \underset{\text{lr}}{\leqslant} \ldots$. Proving this monotony seems to be much more difficult than proving the lr-monotony of $NB(r, p)$.

For instance, take a discrete stable semigroup with $\alpha \leq 1/2$. Its lr-monotony has been proved by Simon (2016) with complex analysis tools. The compounding distribution is a Sibuya law $f(n) = (-1)^n\binom{\alpha}{n}$. Again, numerical computations make us conjecture that $f \underset{\text{lr}}{\leqslant} f * f \ldots$, but we are unable to prove it.

This observation raises a question: is this sufficient condition also necessary? Next proposition only gives a partial answer for lr and hr orders. The reverse hazard rate (rh) order does not seem to imply any particular condition.

Proposition 7.17. Let D be a probability measure on $\mathbb{R}_+$.

If $(cp(\lambda, D))_{\lambda\geq 0}$ is non-increasing in the likelihood ratio order, then $D \underset{\text{lr}}{\leqslant} D * D$.

If $(cp(\lambda, D))_{\lambda\geq 0}$ is non-increasing in the hazard rate order, then $D \underset{\text{hr}}{\leqslant} D * D$.

Proof. Without restriction, one may assume that D is supported on $(0, \infty]$.

Define $D_l := cp(l, D)$ and $\tilde{D}_l := D_l - D_l(\{0\})\delta_0$. By assumption, $D_l \underset{\text{lr}}{\leqslant} D_{2l}$. Since $D_{2l} = D_l * D_l$, one can easily prove that

$$\tilde{D}_l \underset{\text{lr}}{\leqslant} \tilde{D}_l * \tilde{D}_l.$$

Proposition A.21 in the appendix tells that $\frac{1}{l}\tilde{D}_l$ converges in distribution D. As such convergence preserve stochastic ordering,, $D \underset{\text{lr}}{\leqslant} D * D$.

The proof for the hr order is similar, as $\frac{1}{l}\overline{\tilde{D}}_l(n)$ converges pointwise to $\overline{D}(.)$ on $(0, \infty)$. □

7.2 DISTRIBUTIONAL PROPERTIES OF LÉVY PROCESSES

In previous section, relationships are established between the TP_2 distributional properties of Lévy processes and TP_2 stochastic orders. Furthermore, the study led in previous chapters on linear semi-Markov chains has brought further motivations for distributional properties such like log-concavity and unimodality. This section is about such distributional properties, called reliability classes, of Lévy processes. The section begins

with a survey of all related results we have found in the literature. Then, it provides original results we have derived. Because some questions about the most interesting properties – log-concavity and unimodality – are still open, the study has led us to investigate other related reliability classes.

7.2.1 KNOWN PROPERTIES

All results of this section are standard and we be found in textbooks like (Sato, 1991).

LOG-CONCAVITY. Here, we look for Lévy processes X where all marginal distributions X_t, $t > 0$ share the property of log-concavity. Unfortunately the next proposition shows no Lévy process supported on $[0, \infty)$ is fully log-concave, except one, the Poisson process. Later on, Proposition 7.25 will give an extension and an alternative proof of this result.

Proposition 7.18 (Watanabe, 1991, Theorem 2)**.** Let X be a Lévy process.
X is fully log-concave on $\mathbb{R}$ if and only if it is a Gaussian process.
X is fully discrete log-concave on $\mathbb{N}$ if and only if it is a Poisson process.

Thus the Poisson semigroup is the *optimal semigroup for log-concavity*. No other convolution semigroup $(D_l)_{l\geq 0}$ can be fully log-concave. However, for many of them, it often happens that D_l is log-concave for some values l, but not all values. Next proposition gives a necessary condition: a discrete Lévy distribution D_l cannot be log-concave if l is too small. Such a situation confronts us with the "small durations problem": the desired property cannot be obtained under a minimal length l.

Proposition 7.19 (Johnson et al., 2013, Lemma 5.2)**.** If a discrete compound Poisson distribution $cp(F, \lambda)$ with compounding distribution F and intensity $\lambda \geq 0$ is discrete log-concave, then necessarily

$$\lambda \geq \frac{2f(2)}{f(1)^2},$$

where f is the pmf of F.

Proof. Let p denote the pmf of $cp(f, \lambda)$. Given the so-called Panjer's recursion — equation (A.5) in the appendix, $p(1)^2 \geq p(0)p(2)$ is equivalent to $(\lambda f(1)p(0))^2 \geq ((\lambda f(1))^2 p(0) + 2\lambda f(2)p(0))p(0)/2$, which is equivalent to $\lambda f(1)^2 \geq 2f(2)$. □

Next proposition is the most powerful sufficient condition that provides Lévy processes that are log-concave above this threshold.

Proposition 7.20 (Hansen, 1988, Theorem 1)**.** Let f be a pmf such that $(nf(n))_{n\in\mathbb{N}}$ is discrete log-concave. The compound Poisson distribution $cp(F, \lambda)$ is log-concave if and only if $f(1) > 0$ and $\lambda \geq \frac{2f(2)}{f(1)^2}$.

Proposition 7.20 is very powerful as it contains all infinitely divisible distributions we know for being log-concave.

UNIMODALITY. Unimodality of Lévy processes has been deeply studied, but characterizing fully unimodal semigroups remains an open question. Two majors achievements in this direction are two partial conditions. First one is unimodality of self-decomposable distributions, which forms an important subclass of infinitely divisible distributions. This sufficient condition is the strongest available one that gives fully unimodal semigroups. It is very powerful as most of commonly known infinitely divisible distributions are self-decomposable: Negative binomial, Poisson, Gaussian, Gamma laws.

Definition 7.21. An infinitely divisible distribution D is *self-decomposable* if its canonical measure k is unimodal with mode 0.

A discrete infinitely divisible distribution D supported on $\mathbb{N}$ is *discrete self-decomposable* if its canonical sequence $k := ((n+1)\, f(n+1))_{n\in\mathbb{N}}$ is non-increasing.

Remark. From the definition, it is obvious that a distribution is self-decomposable if and only if it can be embedded in a self-decomposable semigroup.

Theorem 7.22 (Watanabe, 1992, Theorems 1.1, 1.2)**.**
If a distribution is self-decomposable, then it is unimodal.
If a distribution is discrete self-decomposable, then it is discrete unimodal.

HALF LOG-CONCAVITY. Second major achievement is a necessary condition. Semigroups that are fully unimodal automatically get a *partial* form of log-concavity, introduced under the name of Yamazato property. This property is interesting as even if almost no convolution semigroup is fully log-concave, many semigroups are fully unimodal. In appendix B.4.2, we study more extensively this property which is not very standard. As explained there, we call it *lower-half log-concavity* since it corresponds to log-concavity of half of the distributions.

Definition 7.23. A measure F on $\mathbb{R}$ is said to be *lower-half log-concave* if one the following condition holds:

i. F is unimodal with mode $l_f \overset{\text{def}}{=} \inf \operatorname{supp}[F]$.

ii. F is unimodal with mode $a > l_f$, absolutely continuous, and admits a density f such that f is positive and log-concave on (l_f, a) and $f(a^-) \geq f(a^+)$.

A measure F on $\mathbb{Z}$ is said to be *discrete lower-half log-concave* if F is discrete unimodal with mode a and its pmf f is log-concave on $(-\infty, a) \cap \operatorname{supp}[F]$.

Theorem 7.24 (Watanabe, 1992, Theorem 1.1)**.**
If a non-negative Lévy process $X = (X_t)_{t\geq 0}$ is unimodal, then it is lower-half log-concave.
If a non-negative Lévy process $X = (X_t)_{t\geq 0}$ is discrete unimodal, then it is discrete lower-half log-concave.

Historically, property Y was established for self-decomposable processes by Yamazato (1978, Theorem 2). Watanabe extended the result to any unimodal process. Later on in section 7.2.2, we give an extension of this theorem and an alternative proof. Our proof is much simpler. It also reveals that this property does not rely on the infinite divisibility structure, but on another property that might hold for other kind of processes.

7.2.2 NEW RESULTS

While log-concavity is an important property, only a handful of Lévy processes are log-concave. We have seen that obtaining fully log-concave semigroups is hard. So now, we look for wider reliability classes of Lévy processes. This section contains our original results in this direction.

CHARACTERIZATION OF IHR AND DRHR PROCESSES. For instance, the inclusion LCAV $\subset$ IHR $\cap$ DRHR motivates the study of IHR and DRHR classes. We have achieved to characterize discrete and continuous Lévy processes that are fully IHR, or fully DRHR.

Proposition 7.25. Let X be a Lévy process on $\mathbb{R}$, and ν denote its Lévy measure.

a) X is discrete DRHR (on $\mathbb{Z}$) if and only if it has no Gaussian part, $\mathrm{supp}[\nu] \subset \{-1\} \cup \mathbb{N}$ and the pmf of ν is non-increasing on $\mathbb{N}^*$.

b) X is DRHR (on $\mathbb{R}$) if and only if X is spectrally positive (*i.e.,* $\mathrm{supp}[\nu] \subset (0, \infty)$) and ν is a concave measure on $(0, \infty)$ (*i.e.,* it admits a non-increasing density).

The IHR and DRHR properties are spatially symmetrical, as explained by the following lemma. Its proof is straightforward.

Lemma 7.26. Let M be a measure on $\mathbb{R}$. Define the reversed measure $\widetilde{M}$ by $\widetilde{M}([0, x]) \stackrel{\text{def}}{=} M([-x, 0])$. Note that $\widetilde{\widetilde{M}} = M$.

M is IHR on $\mathbb{R}$ if and only if $\tilde{M}$ is DRHR on $\mathbb{R}$.

M is discrete IHR on $\mathbb{Z}$ if and only if $\widetilde{M}$ is discrete DRHR on $\mathbb{Z}$.

X is a Lévy process with Lévy measure ν if and only if $\widetilde{X}$ is a Lévy process with Lévy measure $-\nu$.

Thanks to this lemma, we only have to prove the DRHR claim.

Proof of Proposition 7.25. [Necessary condition.] Let D_l denote the measure of X_l. Using arguments like (Steutel and van Harn, 2004, Proposition 4.19 (iv)), we can show the support of D_l includes the support of ν. Proposition A.21 tells that for all $x \in \mathbb{R}$:

$$\begin{aligned} \lim_{l \to 0} D_l(x) &= 0 && \text{if } x < 0, \\ \lim_{l \to 0} D_l(x) &= 1 && \text{if } x \geq 0, \end{aligned}$$

and for all continuity points x of ν:

$$\begin{aligned} \lim_{l \to 0} l D_l(x) &= \nu((-\infty, x]) = \nu(x) && \text{if } x < 0, \\ \lim_{l \to 0} l(D_l(x) - 1) &= \nu([x, \infty)) && \text{if } x > 0, \end{aligned}$$

This implies

$$\lim_{l \to 0} \frac{D_l(x+h) - D_l(x)}{D_l(x)} = \begin{cases} 0 & \text{if } 0 \leq x < x + h, \\ +\infty & \text{if } x < 0 \leq x + h, \\ \frac{\nu(x+h) - \nu(x)}{\nu(x)} & \text{if } x < x + h < 0 \text{ and if } \nu(x) \neq 0, \end{cases} \tag{7.3}$$

and in addition,

$$\forall x > 0, \quad \lim_{l \to 0} \frac{D_l(x+h) - D_l(x)}{l D_l(x)} = \nu(x+h) - \nu(x).$$

Case 1. Assume (D_l) are DRHR. This implies that $\frac{D_l(x+h)-D_l(x)}{D_l(x)}$ is non-increasing with respect to x on $\mathrm{supp}[D_l]$. This has two consequences. Firstly, in last identity the right-hand side quantity is non-increasing with respect to $x > 0$. This implies that ν is concave on $(0, \infty)$. Secondly, assume that $\inf \mathrm{supp}[\nu] < 0$. As there exists real numbers x, h such that $x < x + h < 0$, $\nu(x) > 0$, $\nu(x + h) > 0$ and therefore $\frac{\nu(x+h)-\nu(x)}{\nu(x)} < +\infty$. Equation (7.3) implies that for some $l > 0$, $\frac{D_l(x+h)-D_l(x)}{D_l(x)}$ cannot be non-decreasing with x.

Case 2. Suppose X is discrete DRHR on $\mathbb{Z}$. The asymptotic behavior written above it still valid with $x \in \mathbb{Z}^*$ and $h = 1$. As a consequence, ν is discrete concave (its pmf is non-increasing), and $\nu(n) = 0$ if $n < -1$.

[Sufficiency condition] Firstly, we prove it for discrete processes. Secondly, we use an approximation of continuous processes by discretization of the state space. *Case 2.* Let q be a nonnegative sequence on $\mathbb{Z}$ such that $\mathrm{supp}[q] \subset \{-1, 0, 1, 2, 3 \ldots\}$ and q is non-increasing on $\mathbb{N}^*$. For all $l \geq 0$, $\delta_0 + lq$ is discrete DRHR. Indeed, it is lower-half log-concave — see appendix B.4.2. Fix $l \geq 0$. As convolution preserves discrete DRHR distributions — see Proposition B.29 in the appendix — the n-fold convolution $q_n := (\delta_0 + l/nq)^{*n}$ is discrete DRHR for all $n \in \mathbb{N}^*$. Furthermore, $Z[q_n](z) = (1 + l/nZ[q])^n$ and $\lim_{n\to\infty} Z[q_n](z) = e^{lZ[q](z)}$. This means that q_n weakly converges to $X_l = cp(q, l)$. As weak convergence preserves the DRHR class, X_l is discrete DRHR. As such compound Poisson distributions exhaust discrete Lévy processes, this ends the proof in the discrete case.

Case 1. Now, we prove it when X is a general Lévy process supported on $\mathbb{R}$, using approximation with discrete Lévy processes X^h supported on $h\mathbb{Z}$. The drift does not change the DRHR property of X, so we can assume that X has no drift. As explained in Proposition B.29 in the appendix, convolution with a log-concave distribution such as Gaussian laws preserves the DRHR property, so we can assume that X has no Gaussian part.

Let h be in $(0, 1)$. We borrow the spatial discretization scheme of the Lévy measure ν from Mijatovic et al. (2014). Define

$$c^h \stackrel{\text{def}}{=} \int_{[0,h)} x^2 \nu(\mathrm{d}x)/2h^2, \qquad \mu^h \stackrel{\text{def}}{=} \sum_{n \in \mathbb{N}^*} n \int_{(h(n-1/2),h(n+1/2)]} \mathbf{1}_{[0,1]} \mathrm{d}\nu(h)$$

if ν is infinite, and $c^h \stackrel{\text{def}}{=} \mu^h \stackrel{\text{def}}{=} 0$ else. As a Lévy measure is non-decreasing, $c \geq 0$. Define the discrete measure ν^h by

$$\nu^h(\{hn\}) \stackrel{\text{def}}{=} \begin{cases} \nu((h(n-1/2), h(n+1/2)]) & \text{if } n > 1, \\ \nu((h(n-1/2), h(n+1/2)]) + c^h & \text{if } n = 1, \\ c^h + \mu^h & \text{if } n = -1, \\ -\sum_{k \in \mathbb{Z}^*} \nu^h(\{hk\}) & \text{, if } n = 0. \end{cases}$$

ν^h is a proper Lévy measure supported on $\{-h, 0, h, \ldots\}$. Let X^h be its corresponding Lévy process; it is compound Poisson and supported on $h\mathbb{Z}$. By hypothesis, ν is a concave measure. So $\nu((hn - h/2, hn + h/2])$ is non-increasing with respect to $n \in \mathbb{N}^*$. As $c \geq 0$, this implies that the discrete sequence $(\nu^h(hn))_{n\in\mathbb{Z}}$ is non-increasing on $\mathbb{N}^*$. So as shown in Case 1., X^h is discrete DRHR on $h\mathbb{Z}$.

The corresponding cumulant generating function of X_1^h is $\psi_h(u) = \sum_{n=-1}^{\infty}(e^{iuhn} - 1)\nu^h(hn)$. It is easy to prove the pointwise convergence of ψ^h to the cumulant generating function ψ of X_1 — see (Mijatovic et al., 2014, Remark 3.11(iii)). By Lévy continuity theorem, this implies the weak convergence of X_1^h to X_1 — refer to (Steutel and van Harn, 2004, Proposition 4.6).

Since the step h of the lattice $h\mathbb{Z}$ goes to 0, the (discrete) DRHR property is preserved through weak convergence as explained in Proposition B.30 in the appendix. As a result, X is DRHR on $\mathbb{R}$. □

Remark. Combining the two results gives that the Gaussian and Poisson processes are the only Lévy processes that are both IHR and DRHR. For general distributions, LCAV $\subset$ IHR $\cap$ DRHR. So our result gives an alternative proof of Proposition 7.18. It also shows this inclusion is an equality in case of Lévy processes. This fact has also been observed for continuous-time Markov chains by Kijima (1998, Corollary 3.1).

Remark. In the literature, Cai and Willmot (2005, Theorem 3.2) provides a sufficient condition for compound Poisson processes supported on $\mathbb{R}_+$. Our proposition extends this result to the wider class of Lévy processes and also gives a necessary condition.

From last proposition, we draw two meaningful conclusions.

- Non-negative Lévy processes that are unimodal are necessarily DRHR. In particular, self-decomposable processes are DRHR.
- Non-negative Lévy processes cannot be fully IHR except the Poisson process. The problem described for log-concavity by Proposition 7.19 also appears for IHR: non-decreasing Lévy processes $X = (X_l)$ cannot have IHR distributions at small duration l.

CHARACTERIZATION OF M-DRHR PROCESSES. Next proposition characterizes semigroups that have another property called M-DRHR. Refer to appendix B.4.3 for the mathematical background on this notion. The motivation of such result is twofold. On the one hand, unimodal processes are DRHR, and M-DRHR is a subclass of DRHR. On the other hand, self-decomposable processes are a very important subclass of unimodal processes. Our result relates both notions. It claims that self-decomposable processes are fully M-DRHR, and are the only kind of Lévy processes to be so. To prove it, we introduce an original method based on self-similar Markov processes.

Proposition 7.27. A non-negative Lévy process $X = (X_t)_{t\geq 0}$ is fully M-DRHR (resp. discrete-) if and only if X is self-decomposable (resp. discrete-).

In particular, every self-decomposable distribution is M-DRHR.

Our proof is built on a characterization of self-decomposability with additive processes having the property of self-similarity. This result was first derived by Sato (1991).

Refer to appendix A.4.3 for the definition additive processes. We formulate this characterization using scaling operators S_a.

Theorem 7.28 (Sato, 1999, Theorem 16.1)**.** Let S_a denote scaling of factor a, $S_a : D(\cdot) \mapsto D(a\cdot)$.

A probability measure D on $\mathbb{R}$ is self-decomposable if and only if the family $(S_{1/t}D)_{t\in[0,1]}$ defines an additive semigroup on $\mathbb{R}$.

In other words, a random variable X on $\mathbb{R}$ is self-decomposable if and only if $Y_1 \sim X$ for some additive process $(Y_t)_{t\geq 0}$ such that

$$\forall c > 0, \qquad (Y_{ct})_{t\geq 0} \overset{\mathrm{d}}{=} (c\, Y_t)_{t\geq 0}.$$

The process Y is said to be a 1-*self-similar process*.

This tool let us derive a very simple proof of our proposition.

Proof of Proposition 7.27, continuous case. Necessary condition. Let $X \sim (D_l)_{l\geq 0}$ be a Lévy process with Lévy measure ν. Assume X is fully M-DRHR. This implies each D_l admits a density d_l on $(0,\infty)$ such that $t \mapsto td_l(t)/D_l(t)$ is non-increasing. For a convolution semigroup (d_l) with Lévy measure ν, the small time behavior is

$$td_l(t)/D_l(t) \sim_{l\to 0} lt\nu(\mathrm{d}t),$$

for all $t \in (0,\infty)$ in the support of ν. This implies $t \mapsto t\nu(\mathrm{d}t)$ is non-increasing, which is the characterization of X being self-decomposable.

Sufficient condition. It suffices to show that any self-decomposable distribution D on $\mathbb{R}_+$ is M-DRHR. From the definition, it is immediate that scaling preserves this property. So all $S_{1/t}D$ are self-decomposable. So the family $Y = (S_{1/t}D)_{t\in[0,1]}$ is fully DRHR.

As D is self-decomposable, previous proposition tells that $(S_{1/t}D)_{t\in[0,1]}$ is an additive semigroup, has nonnegative increments and is fully DRHR. Consequently, Proposition 7.11 implies $(S_{1/t}D)_{t\in[0,1]}$ is rh-monotone. This is equivalent to D being M-DRHR. □

Next theorem defines discrete self-decomposability as Theorem 7.28 does for continuous distributions. The discrete counterpart of *scaling* $S_{1/a}$ is *binomial thinning* B_a, defined for all $a \in [0,1]$ – see definition 8.2 in next chapter for more details. In short, B_a is defined over Z-transforms of sequences as $B_a : P \mapsto P(az + 1 - a)$.

Theorem 7.29 (discrete counterpart)**.** Let B_a denote binomial thinning, for $a \in [0,1]$.

A probability measure D on $\mathbb{N}$ is discrete self-decomposable if and only if the family $(B_tD)_{t\in[0,1]}$ defines an additive semigroup on $\mathbb{N}$.

With this tool, we can easily end the proof of our Proposition 7.27.

Proof of Proposition 7.27, discrete case. From the definition, it is immediate to see that B_aD are discrete self-decomposable if D is so. This implies that $(B_tD)_{t\in[0,1]}$ is fully discrete DRHR. So the same proof as the continuous case may be used by replacing $S_{1/t}$ with B_t. □

Remark. Gathering propositions 7.25, 7.27 leads to an interesting scheme. For an infinitely divisible distribution D with Lévy measure ν,

$$\nu(x) \text{ is decreasing} \implies \frac{d(x)}{D(x)} \text{ is decreasing,}$$
$$x\,\nu(x) \text{ is decreasing} \implies x\,\frac{d(x)}{D(x)} \text{ is decreasing.}$$

As an open question, we wonder if a similar result could holds for $x^k\,\nu(x)$ with $k = 2, 3, \ldots$ or even other weighting functions.

ALTERNATIVE PROOF OF LCAV $\implies$ MSU. We show how our method is suitable for another result on MSU, a property which is stronger than M-DRHR. Whereas LCAV and MSU properties are unrelated for a general distribution, it has been shown for self-decomposable distributions that LCAV $\implies$ MSU, as next proposition explains.

Proposition 7.30 (Sato, 1999, Chapter 10)**.** Let D be a self-decomposable distribution on $\mathbb{R}_+$. If D is log-concave on $[0, a]$ for some $a > 0$, then D is MSU on $[0, a]$.

The discrete counterpart of this result cannot be found in the literature. Our method based on self-similar processes provides a straightforward proof.

Proposition 7.31 (discrete counterpart)**.** Let D be a discrete self-decomposable distribution on $\mathbb{N}$. If D is discrete log-concave on $\{0, \ldots, a\}$ for some $a \in \mathbb{N}$, then D is discrete MSU on $\{0, \ldots, a\}$.

Proof. Assume D discrete self-decomposable on $\mathbb{N}$ and log-concave on $I := \{0, \ldots, a\}$.

Step 1. Assume the pmf d is strictly log-concave on I. Let $B_t d$ denote the pmf of $B_t D$. As I is finite, there exists $t_0 < 1$ such that $d_t := B_t d$ is discrete log-concave on I, for all $t \in [t_0, 1]$. Indeed, the definition of binomial thinning immediately gives the continuity of all $t \mapsto B_t d(n)$, for $n \in \{0, \ldots, a+1\}$. As $d(n+1)^2 - d(n)d(n+2) > 0$ on this interval, and $B_1 d = d$, the continuity gives the existence of $t_0 < 1$ such that $d_t(n+1)^2 - d_t(n)d_t(n+2) \geq 0$ on this interval. So the family $(B_t d)_{t\in[t_0,1]}$ is discrete log-concave on $\{0, \ldots, a\}$. As d is discrete self-decomposable on $\mathbb{N}$, this family is an additive process with non-negative increments. By proposition 7.11, it is lr-monotone on I. Owing to Proposition B.67 in the appendix, since $t_0 < 1$ this is implies d being discrete MSU on I.

Step 2. If d is not strictly log-concave, define $d^\lambda = d * po_\lambda$, where po_λ is Poisson pmf of intensity $\lambda > 0$. As d, po^λ are self-decomposable and convolution preserve this property, so are the d^λ. Besides, as po_λ is strictly log-concave on $\mathbb{N}$ and d is log-concave on I, d^λ is strictly log-concave on I. So previous case applies: the d^λ are discrete MSU on I. Furthermore, one has pointwise convergence $\lim_{\lambda\to 0} d^\lambda = d$. As this convergence preserves the MSU property, d is discrete MSU on I. □

Remark 7.32. The proof is even more immediate if $a = \infty$, *i.e.*, d is log-concave on its whole support $\mathbb{N}$. Indeed, binomial thinning preserves full log-concavity — contrary to partial log-concavity. This fact is known under the name of Brenti's criterion (Brenti, 1989, Theorem 2.5.3). So if d is log-concave, $(B_t d)_{t\in[0,1]}$ is a discrete log-concave and additive semigroup on $\mathbb{N}$. Consequently, it is lr-monotone, and so d is MSU.

Remark. We have not managed to deduce the continuous case from the discrete case (or vice-versa). The problem is that partial log-concavity might not be preserved through the discretization schemes that preserve self-decomposability, like Poisson mixture or discretization of the Lévy measure.

ALTERNATIVE PROOF FOR HALF LOG-CONCAVITY. We suggest a slight extension of Theorem 7.24 and a different proof. This has several motivations.

- Whereas original formulation holds for nonnegative Lévy process, the following proposition holds for additive processes with nonnegative increments.
- Whereas original proof is suitable only for processes indexed on $\mathbb{R}_+$, our proof also works with processes indexed on $\mathbb{N}$.
- Our proof outline suggests the result could hold for wider class of Markov processes, although a systematic study of those is left as an open question.

Next proposition restates for more general additive processes. We begin with discrete-valued processes indexed on $\mathbb{N}$. In this case we are able to prove it when initial distribution is supported on $\mathbb{Z}$.

Proposition 7.33. Let $(d_n)_{n\in K}$ be a family of pmfs indexed on $K = \{0, \dots, N\}$ or $\mathbb{N}$. Define the additive process $(d_{0:j})_{j\in K}$ by $d_{0:j} := d_0 * \dots * d_j$. Assume that

(i) initial distribution D_0 is discrete lower-half log-concave and supported on $\mathbb{Z}$.
(ii) increments D_j are discrete DRHR and supported on $\mathbb{N}$, for all $j \in K$.
(iii) the additive process $d_{0:j}$ is unimodal, for all $j \in K$.
(iv) $\overline{\text{mode}}[d_{0:j+1}] \leq \overline{\text{mode}}[d_{0:j}] + 1$.

Then, the additive process $(d_{0:j})_{j\in K}$ is discrete lower-half log-concave.

Proof. The proof is simple induction on $j \in K$. Initialization is true as $d_{0:0} = d_0$ and assumption (i) tells it is lower-half log-concave. Let j be in K and assume $d_{0:j}$ is lower-half log-concave. Define $m := \overline{\text{mode}}[d_{0:j}]$ and $M := \{-\infty, \dots, m\}$. Define also $m' := \overline{\text{mode}}[d_{0:j+1}]$. Consider the first difference

$$\Delta f(n) \stackrel{\text{def}}{=} f(n) - f(n-1),$$

so that f is the cumulative distribution (cdf) associated to Δf.

By assumption (iii), $d_{0:j}$ is unimodal. Since m is its mode, this implies the sequence $\Delta d_{0:j}$ is non-negative. In addition, the induction assumption means $d_{0:j}$ is log-concave on M. This implies that $\Delta d_{0:j}(n)$ is discrete DRHR.

By assumption (ii), d_{j+1} is discrete DRHR.

In addition, since d_{j+1} is supported on $\mathbb{N}$, one has the following identity:

$$\forall n \in M, \quad \Delta d_{0:j+1}(n) = \Delta[d_{0:j} * d_{j+1}](n) = \sum_{k=0}^{\infty} \Delta d_{0:j}(n-k) d_{j+1}(k).$$

$\Delta d_{0:j}$ and d_{j+1} are discrete DRHR on M. Since d_{j+1} is supported on $\mathbb{N}$, the convolution product preserves such class, so $\Delta d_{0:j+1}$ is discrete DRHR on M. This means that $d_{0:j+1}$ is log-concave on M.

By assumption (iv), $d_{0:j+1}$ is unimodal and $m' \leq m+1$. It remains to prove that $d_{0:j+1}$ is log-concave on $\{-\infty, \dots, m'\}$ If $m' \leq m$, this is already done as we have proven that $d_{0:j+1}$ is log-concave on M. Else, $m' = m+1$, so it remains to prove that $d_{0:j+1}$ is log-concave at m'. Since m' is a mode of $d_{0:j+1}$, this is automatically true. So the induction is complete. □

Next, we state the result for additive processes with discrete nonnegative increments. The difference with previous proposition is that the additive process is now indexed on $\mathbb{R}_+$.

Proposition 7.34. Let $(D_l)_{l\geq 0}$ be an additive process. Assume that
(i) initial distribution D_0 is discrete lower-half log-concave and supported on $\mathbb{Z}$.
(ii) its increments are discrete DRHR and supported on $\mathbb{N}$.
(iii) the additive process D_l is discrete unimodal, for all $l \geq 0$.
Then, the additive process $(D_l)_{l\geq 0}$ is discrete lower-half log-concave.

Proof. The idea is to "sample" the process in order to apply previous proposition. By definition of an additive process indexed on $\mathbb{R}_+$, functions $l \mapsto d_l(n)$ are continuous in l for all $n \in \mathbb{N}$. This implies the $\mathbb{N}$-continuity in l of the greatest mode $l \mapsto \overline{\text{mode}}[d_l]$.

Fix $t > 0$. By continuity, there exists $N_0 \in \mathbb{N}$, $N \in \mathbb{N}^*$, and a strictly increasing sequence $(l_j)_{j=0\dots N}$ such that $l_0 = 0$, $l_N = t$, and $\overline{\text{mode}}[d_{l_j}] = N_0 + j$ for all $j < N$. As result, the "sampled" process $(D_{l_j})_{j=0\dots N}$ fulfills all hypotheses of Proposition 7.33. Therefore D_t is discrete lower-half log-concave. As this reasoning holds for all $t > 0$, the proof is over. □

Then, we move to additive process is supported *and* indexed on $\mathbb{R}_+$. In this case, we have to assume that initial distribution D_0 is supported on $\mathbb{R}_+$, but we suspect the result still holds if D_0 is supported on $\mathbb{R}$. However, this would require a completely different proof, since Poisson mixtures is only defined with nonnegative distributions.

Proposition 7.35. Let $(D_l)_{l\geq 0}$ be an additive process. Assume that
(i) initial distribution D_0 is lower-half log-concave and supported on $\mathbb{R}_+$.
(ii) its increments are DRHR and supported on $\mathbb{R}_+$.
(iii) the additive process D_l is unimodal, for all $l \geq 0$.
Then, the additive process $(D_l)_{l\geq 0}$ is lower-half log-concave.

Note that assumption (ii) was automatically true when the additive process is actually a Lévy process. But such assumption is required for such generalization of Theorem7.24 to additive processes.

Proof. The proof relies on Poisson mixtures Γ_λ and its numerous preservation properties. Appendix B.5 provides full background on this tool. In particular,

- D_l is lower-half log-concave if and only if $\Gamma_\lambda[D_l]$ if lower-half log-concave for all $\lambda > 0$ — see proposition B.81.
- For all $\lambda > 0$, $(\Gamma_\lambda[D_l])_{l\geq 0}$ is an additive process supported on $\mathbb{N}$ — see Proposition B.75.

So it only remains to justify that such discrete processes fulfill every hypothesis of Proposition 7.34.

(i) $\Gamma_\lambda[D_0]$ is discrete lower-half log-concave, since Poisson mixtures preserve this property.
(ii) the increments are discrete DRHR and supported on $\mathbb{N}$, since Poisson mixtures preserve this property.
(iii) $\Gamma_\lambda[D_l]$ is discrete unimodal, since Poisson mixtures preserve unimodality.

□

7.3 APPLICATIONS TO COMMON LÉVY PROCESSES

The results we have obtained so far are about large sub-classes of Lévy processes. To finish this chapter, we apply such general results to a selection of common Lévy processes. This leads to new results on a few special functions. More details on common Lévy processes are supplied by appendix A.4.4, see also (Steutel and van Harn, 2004).

7.3.1 LOG-CONCAVITY OF SPECIAL FUNCTIONS

GAMMA PROCESSES. A Gamma process is $X_l \sim \Gamma(l, \beta)$ is defined for any rate $\beta > 0$ by its cdf $D_l(t) = \gamma(k, \beta t)/\Gamma(k)$ where $\Gamma(\cdot)$ is the Gamma function and $\gamma(\cdot, \cdot)$ the upper incomplete Gamma function. It is known that Gamma processes have IHR first-passage times T_t (Shaked and Shanthikumar, 1988). As it is known that the Gamma process X is lr-monotone on I, our proposition 7.13 shows its first-passage times T_t are actually log-concave. This result does not seem trivial as the pdf m_t of T has a quite complex expression (Park and Padgett, 2005, Section 2.2):

$$m_t(x) = (\Psi(x) - \log(\beta t)) \frac{\gamma(x, \beta t)}{\Gamma(x)} + \frac{(\beta t)^x}{x^2 \Gamma(x)} \, {}_2F_2(x, x; x+1, x+1; -\beta t),$$

where Ψ is the digamma function (or logarithmic derivative of the Gamma function) and ${}_2F_2$ the hypergeometric function of order $(2, 2)$.

NEGATIVE BINOMIAL PROCESSES. The discrete counterpart of a Gamma process is a Negative Binomial process $X_l \sim NB(l, p)$ defined for any $p \in [0, 1)$ by its pmf $d_l(n) = \binom{n+r-1}{n}(1-p)^r p^n$. Similarly, as X is lr-monotone, we deduce its first-passage times T_t are log-concave.

However, either for the Gamma or the Negative Binomial, a question remains open. We do not know whether $(T_t)_{t \geq 0}$ is monotone in the likelihood ratio order on its full support $(0, \infty)$. We only know it holds on $(1, \infty)$.

7.3.2 CASE OF STABLE DISTRIBUTIONS

Stable distributions are the among the most important examples of Lévy processes. Our results prove new properties for nonnegative stable processes without drift. For such processes, first-passage measure are $\overline{M}_t(l) = D_l(t) = D_{1,\alpha}(tl^{-1/\alpha})$, so $m_t(l) = \frac{t}{\alpha} l^{-1-1/\alpha} d_{1,\alpha}(tl^{1/\alpha})$.

$M_{t,\alpha}$ is a t^α-scaling of the so-called *Mittag-Leffler distribution* $M_{1,\alpha}$ (Pillai, 1990). The Laplace transform of such a distribution is equal to $\mathbb{E}[e^{-sM_{1,\alpha}}] = E_\alpha(-s^\alpha)$, where E_α is the so-called *Mittag-Leffler function* (Mittag-Leffler, 1903)

$$E_\alpha(z) \overset{\text{def}}{=} \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + 1)}.$$

The following powerful result has been obtained recently by Simon (2016). Refer to appendix B.4.3 for definition and background on MSU, M-DRHR, M-IHR reliability classes.

Theorem 7.36 (Simon, 2016)**.** The following propositions are equivalent:

- $\alpha \leq 1/2$,
- $(M_{t,\alpha})_{t>0}$ is increasing in the likelihood ratio order,
- $(D_{l,\alpha})_{l>0}$ is increasing in the likelihood ratio order,
- $(M_{t,\alpha})_{t>0}$ are multiplicative strongly unimodal (MSU),
- $(D_{l,\alpha})_{l>0}$ are multiplicative strongly unimodal (MSU).

Distributional properties of first-passage measures $M_{t,\alpha}$ have been studied by Simon (2014). It is known they are unimodal and decreasing (their mode is 0) if and only if $\alpha \leq 1/2$. Our proposition 7.13 strengthens unimodality to log-concavity. We have not found such result in the literature, except for the case $\alpha = 1/2$ where an analytic expression is available: $M_{t,1/2}$ is a half centered-Gaussian.

Proposition 7.37. If $\alpha \leq 1/2$, then for all $t \geq 0$, then the Mittag-Leffler distribution $M_{t,\alpha}$ is log-concave.

For higher values of $\alpha \geq 1$, proposition 7.13 gives a weaker result.

Proposition 7.38. For all $\alpha \in (0, 1)$, stable distributions $(D_{l,\alpha})_{l>0}$ are M-DRHR and DRHR. Mittag-Leffer distributions $(M_{t,\alpha})_{t>0}$ are IHR and M-IHR.

The same results holds for discrete stable distributions. This can be easily proved using *Poisson mixtures* and its numerous preservation properties, as described in appendix B.5.

7.3.3 CASE OF CHI AND CHI-SQUARE LAWS

We explain why our result about M-DRHR distributions has an interesting consequence on a conjecture about the so-called *Marcum Q-function*. This special function defined for all $M > 0$, $a, b \geq 0$ by

$$Q_M(a, b) = \int_b^\infty x \left(\frac{x}{a}\right)^{M-1} \exp\left(-\frac{x^2 + a^2}{2}\right) I_{M-1}(ax)\, \mathrm{d}x.$$

Its distributional properties with respect to the three parameters M, a, b have been studied in great detailed in Sun et al. (2010). See the references therein for the related literature in information theory about this function, and its important applications in radar detection and communication. In this reference, some properties are left as

conjectures. Most have been solved by Yu (2011a). In some cases, the author of this preprint has borrowed results from Finner and Roters (1997), which are based on total positivity and probability theory. Indeed, as explained therein, the Q-function is related to two standard probability distributions: the χ (chi) law and the χ^2 (chi-square) laws.

EXISTING RESULT. Our matter is the following conjecture.

Conjecture 3 (Sun et al., 2010, Conjecture 2)**.** If $M \geq 1/2$, then the function $1 - Q_M(a, b)$ is log-concave in $b \in [0, \infty)$.

Two cases been solved by Yu (2011a): first when $M \geq M_0$ where $M_0 \approx 0.789776$, second when $M \geq 1/2$ and $a \leq 1$. So what remains open is the log-concavity for $M \in [1/2, M_0)$ and $a > 1$.

NEW RESULT. Now, we state an original result which extends the conjecture by suggesting the region of log-concavity of $1 - Q_M$ is actually greater. However, it does not prove the remaining part of the conjecture.

Proposition 7.39. If $M \geq a$; then the function $1 - Q_M(a, b)$ is log-concave in $b \in [0, \infty)$.

Actually, our proof gives the stronger fact: $1 - Q_M(a, e^b)$ is log-concave in $b \in \mathbb{R}$ for $M \geq a$. This implies the log-concavity of $1 - Q_M(a, b^\alpha)$ for any positive exponent $\alpha > 0$. We wonder if the condition $M \geq a$ is necessary for this stronger fact.

Proof. Our proof is based on the probabilistic interpretation if Q_M. Indeed, $b \mapsto 1 - Q_M(\sqrt{a}, \sqrt{b})$ is the cumulative distribution function of a chi-square law with $2M$ degrees of freedom and non-centrality parameter a. In other words, if $D \sim \chi^2_{2M}(a)$, then $D(t) = 1 - Q_M(\sqrt{a}, \sqrt{t})$ for all $t \geq 0$.

Now, let us prove $\chi^2_{2M}(a)$ distribution is self-decomposable if and only $M \geq a$. Its characteristic function is

$$\phi(z) = \frac{\exp\left(\frac{iat}{1-2it}\right)}{(1-2it)^M}.$$

So its Lévy measure is $\nu(\mathrm{d}t) = e^{-t/2}(\frac{a}{2}t + M)/t$. Indeed, $\chi^2_{2M}(a)$ is the convolution of two distributions: a central chi-square $\chi^2_{2M} = \chi^2_{2M}(0) = \Gamma(M, 1/2)$ (which is a Gamma distribution) and a chi-square with "zero" degrees of freedom $\chi^2_0(a)$: $\chi^2_{2M}(a) = \chi^2_0(a) * \chi^2_{2M}(0)$. The Lévy measure of $\Gamma(M, 1/2)$ is $\nu_1(\mathrm{d}t) := Me^{-t/2}/t$, whereas the Lévy measure of $\chi^2_0(a)$ is $\nu_2(\mathrm{d}t) := a/4 * e^{-t/2}$. One recover the Lévy measure of $\chi^2_{2M}(a)$ as $\nu = \nu_1 + \nu_2$.

The distribution $\chi^2_{2M}(a)$ is self-decomposable if and only if its Lévy measure is such that $k(t) := t\nu(\mathrm{d}t)$ is non-increasing in t. Here, $k(t) = e^{-t/2}(\frac{a}{2}t + M)$, so $k'(t) = -e^{-t/2}/2(a/2t + M - a)$ which is nonnegative if and only if

$$M \geq a.$$

When $\chi^2_{2M}(a)$ is self-decomposable, then by Proposition 7.27 it is M-DRHR. By definition of this property, this means that $b \mapsto D(e^b)$ is log-concave on $\mathbb{R}$, which implies in particular that $b \mapsto D(b^2) = 1 - Q_M(\sqrt{a}, b)$ is log-concave. □

8

FURTHER RESULTS ON TOTAL POSITIVITY

The idea of this section is to show how total positivity of order two, and the tools we have introduced so far, can efficiently solve problems of combinatorics. Each section is motivated by a specific problem found in the literature fo combinatorics. For each one, our idea is to deduce the problem from a preservation result of well-chosen distributional properties by well-chosen linear operators. *Note this chapter is unrelated to the main matter of the thesis.* Its main purpose is to give further applications of the theoretical results we have obtained so far.

8.1 PRESERVATION OF UNIMODALITY AND RELATED

This section deals with a problem of unimodality preservation raised in combinatorics. We show how the notion of half log-concavity (already used in section 7.2.1 of previous chapter and detailed in appendix B.4) is the right notion to address this problem, as it allows to prove and to extend existing results with elementary arguments.

8.1.1 INTRODUCTION & MOTIVATION

This section exclusively deals with sequences a of nonnegative numbers. They may be finite $a = (a_0, a_2, \dots, a_{N-1})$ or infinite $(a_n)_{n\in\mathbb{N}}$. A sequence is said to be unimodal if it is non-decreasing then non-increasing. Unimodality is ubiquitous in many branches of mathematics. In combinatorics, finding conditions ensuring unimodality or log-concavity is an important question. In this field, many results are stated with polynomials P since there is one-to-one correspondence between a real sequence a and its associated power series $P(z) = \sum_n a_n z^n$ (which is a polynomial if a is finite).

PRIOR RESULTS. This section is motivated by the three following results.

Result 1. If $P(z)$ has non-decreasing coefficients, then $P(z+t)$ has unimodal coefficients for all $t \geq 1$. This has been originally proved by Boros and Moll (1999) for $t = 1$, then conjectured by Alvarez et al. (2001) and proved by Wang and Yeh (2005) for any $t \geq 0$.

Result 2. If $P(z)$ has non-decreasing coefficients, then $P(z + t)$ have log-concave coefficients for all $t \geq 1$. This has been proved in two independent ways by Llamas and Martínez-Bernal (2010) and Chen et al. (2010).

Result 3. If $P(z)$ has log-concave coefficients, then so does $P(z + t)$ for all $t \geq 0$, see (Hoggar, 1974, Theorem 2) or (Brenti, 1989, Theorem 2.5.3).

CONTRIBUTIONS. The main contribution of this section is an original result which extends and unify the aforementioned ones in several ways.

- It widens the hypothesis of Result 1. on $P(z)$ that implies unimodality of $P(z+t)$ for all $t \geq 0$.
- It widens the hypothesis of Result 2. on $P(z)$ that implies log-concavity of $P(z+t)$ for $t \geq 1$, and refines the threshold value 1. These two results are stated as corollaries in section 8.1.5.
- It provides further information on the shape of the coefficients of $P(z+t)$ in the case $t \leq 1$.
- It reinterprets Result 1. and 2. as a preservation of some distributional property, like Result 3 asserts.

Consider the operator $T_t : P(z) \mapsto P(z+t)$. Results 1. and 3. tell that T_t preserve log-concavity and maps non-decreasing sequences to unimodal sequences. Our idea is to introduce two subclasses of unimodal sequences that we call lower- and upper-*half* log-concave. In parallel, we introduce the family of operators $T_{a,b} : P(z) \mapsto P(az+b)$ that generalizes T_t. Then, our result is as follows.

Proposition 8.1 (Main result)**.** Let $P(z)$ be a power series with nonnegative coefficients and positive radius of convergence.

- Suppose the coefficients of $P(z)$ are upper half-log-concave. Then, for all $a \geq 0$, $b \geq 0$ such that $a+b \geq 1$, the coefficients of $P(az+b)$ are upper half-log-concave.
- Suppose the coefficients of $P(z)$ are lower half-log-concave. Then, for all $a \geq 0$, $b \geq 0$ such that $a+b \leq 1$, the coefficients of $P(az+b)$ are lower half-log-concave.

In other words, any $T_{a,b}$ preserves lower half-log-concavity if $a+b \leq 1$, and upper half log-concavity if $a+b \geq 1$. This result divides the plane into two regions which intersects on the segment $a+b=1$.

To prove the result, we use tools from the theory of discrete probabilities. In this latter field, the operators $B_a : P(z) \mapsto P(az+1-a)$ have been considered under the names of *binomial thinning* (Steutel and van Harn, 1979), or *Rényi's thinning* (Rényi, 1957). Our proof relies one a characterization of log-concavity with operator $T_{a,b}$ and on the commutation between B_a and $T_{a,b}$. B_a is more interesting that T_t in terms of preservation, and consequently lie at the root of this work.

For three prior results, available proofs in (Chen et al., 2010; Llamas and Martínez-Bernal, 2010; Alvarez et al., 2001; Wang and Yeh, 2005) are be lengthy and involve heavy algebraic computations that are specific to the problem. Our proof provides a concise and interesting alternative that might generalize to other kind of operators.

8.1.2 PRELIMINARIES

Definition of operators

We introduce a two-parameter family of linear operators on sequences. It encompasses shift operators T_t, the binominal thinning B_a, but also some operators E_p that are

called exponential tiling or *Esscher transform* is the probabilistic literature (Esscher, 1932). The E_p operators are well-defined for any sequence on $\mathbb{N}$. On the contrary, $T_{a,b}P$ is defined only if the radius of convergence R of the power series $P(z)$ is such that $aR + b > 0$.

Definition 8.2. For all $a, b \geq 0$, the operator $T_{a,b}$ is defined on power series $P(z) \in \mathbb{R}[z]$ by

$$T_{a,b} : P(z) \longmapsto P(az + b).$$

Equivalently, $T_{a,b}$ is defined on sequences $(d(n))_{n\in I} \in (\mathbb{R})^I$, $I \subset \mathbb{N}$, by

$$T_{a,b}d(n) = \sum_{k=n}^{\infty} \binom{k}{n} d(k) a^n b^{k-n}.$$

The *binomial thinning* (B_a) operators are defined for all $a \in [0, 1]$ by $B_a \stackrel{\text{def}}{=} T_{a,1-a}$.

$$B_a p(n) = \sum_{k=n}^{\infty} \binom{k}{n} p(k) a^n (1-a)^{k-n} \quad \text{, for } n \in \mathbb{N},$$

$$B_a : P(z) \mapsto P(az + 1 - a).$$

The *exponential tilting* (E_p) operators are defined for all $p \in \mathbb{R}_+$ by $E_p \stackrel{\text{def}}{=} T_{p,0}$.

$$E_p d(n) = d(n)p^n, \text{ for } n \in \mathbb{N}, \qquad E_a : P(z) \mapsto P(az).$$

We also define $T_t \stackrel{\text{def}}{=} T_{1,t}$ and $T \stackrel{\text{def}}{=} T_{1,1}$.

$$Td(n) = \sum_{k=n}^{\infty} \binom{k}{n} d(k), \text{ for } n \in \mathbb{N}, \qquad T : P(z) \mapsto P(z + 1).$$

Remark. All $T_{a,b}$ preserve nonnegative sequences, but only B_a preserves distribution probabilities.

The following relationships will be useful later on. They tell that any operator $T_{a,b}$ is the composition of a binominal thinning B_a with some exponential tiling E_q. Their proof are immediate using power series $P(z)$.

Lemma 8.3. The following compositions formulas hold:

$$\forall a \geq 0, b > 0, \qquad T_{a,b} = B_{\frac{a}{a+b}} \circ E_{a+b} \qquad T_{a,b} = E_{\frac{a}{b}} \circ T \circ E_b,$$
$$\forall a \in [0, 1), \qquad B_a = E_{\frac{a}{1-a}} \circ T \circ E_{1-a}.$$

$T_{a,b}$ defines an affine semigroup: $\forall a, b, a', b' \geq 0$, $\quad T_{a',b'} \circ T_{a,b} = T_{a'a,b'a+b}$.
E_a defines a multiplicative group: $\forall a, a' \geq 0$, $\quad E_a \circ E_{a'} = E_{aa'}$.
B_a defines a multiplicative semigroup: $\forall a, a' \in [0, 1]$, $\quad E_a \circ B_{a'} = B_{aa'}$.

Definition of properties

The following properties are standard. Even if definitions and results focus one non-negative sequences, each could be extended to general measures on $\mathbb{R}$. This is done in appendix B.1 for log-concavity and B.4 for the other ones.

Definition 8.4. Let $d = (d(n))_{n\in I}$ be a real-valued sequence indexed on $I \subset \mathbb{N}$.

d is *non-decreasing* (resp. *non-increasing*) if $\operatorname{supp}[d]$ is an interval of $\mathbb{N}$, and
$\forall n, m \in I, \quad n \leq m \implies d(n) \leq$ (resp. $\geq$) $d(m)$.

d is *unimodal* if $\operatorname{supp}[d]$ is an interval of $\mathbb{N}$, and there exists $a \in I$ such that
$\forall n, m \in I, \quad n \leq m \leq a \implies d(n) \leq d(m)$ and
$\forall n, m \in I, \quad a \leq n \leq m \implies d(n) \geq d(m)$.
Such a is called a *mode* of d.

d is *log-concave* if d is nonnegative, $\operatorname{supp}[d]$ is an interval of $\mathbb{N}$ and
$\forall n \in \mathbb{N}, \quad d(n)d(n+2) \leq d(n+1)^2$ (with $d(n) = 0$ whenever $n \notin I$).

d is log-concave on an interval $[\![a, b]\!]$ if $(d(n))_{n\in[\![a-1,b+1]\!]}$ is log-concave.

Next two distributional properties are far less common in the literature. We have decided to call them of half-log-concavity. Lower half-log-concavity has been introduced by Watanabe (1992) under the name of *Yamazato property*, or *Y property*. On the contrary, the upper counterpart appears almost nowhere, except in the preprint (Yu, 2011a) where it is used without explicit name.

Definition 8.5. Let $d = (d(n))_{n\in\mathbb{I}}$ be a real-valued sequence indexed on $I \subset \mathbb{N}$.

d is *lower half-log-concave* if it is nonnegative, unimodal and has a mode a such that d is log-concave on $[\![0, a]\!]$.

d is *upper half-log-concave* if it is nonnegative, unimodal and has a mode a such that d is log-concave on $[\![a, \infty[\![$.

The following remarks are immediate:

- Log-concave sequences are unimodal.
- Non-increasing sequences are lower half-log-concave.
- Non-decreasing sequences are upper half-log-concave. This is why Proposition 8.1 extends, as announced, the sufficient condition of unimodality of (Boros and Moll, 1999) called Result 1. in the introduction.

Our fundamental remark is that (half) log-concavity can be easily characterized using exponential tilting E_p. Log-concave sequences are exactly the ones whose unimodality is preserved by E_p.

Lemma 8.6. Let d be a nonnegative sequence indexed on $I \subset \mathbb{N}$.
d is log-concave if and only if $E_p d$ is unimodal for all $p \geq 0$.
d is upper-half log-concave if and only if $E_p d$ is unimodal for all $p \geq 1$.
d is lower-half log-concave if and only if $E_p d$ is unimodal for all $p \in (0, 1]$.

Proof. This is a special case of Proposition B.55 in the appendix which proves the same result for the general measures on $\mathbb{R}$. □

Remark. In the literature, the first claim of lemma 8.6 appears in (Karlin, 1968) and (Yu, 2011b, Lemma 1). The other statements are original.

8.1.3 PRESERVATION OF UNIMODALITY

As a lemma for our main proposition, we prove that binomial thinning B_a preserves unimodality. Although this result appears nowhere in the literature, it is a straightforward corollary of existing results in probability theory. Indeed, B_a can be interpreted as a Markov chain. Therefore, results of Keilson and Kester (1978) about unimodality preservation in Markov chains can be applied.

Proposition 8.7. Binomial thinning B_a, for all $a \in [0,1]$, preserves unimodality.

Proof. B_a is a kernel operator whose infinite matrix $\mathbf{M}_a = (m_{i,j})_{i,j\in\mathbb{N}}$ is:

$$\forall k, n \in \mathbb{N}, \qquad m_{k,n} = \binom{k}{n} a^n (1-a)^{n-k}.$$

This means that the row vector of $B_a[d]$ is given by post-multiplication $\mathbf{d}\mathbf{M}_a$, where $\mathbf{d}$ is the row vector of d.

The matrices M_a are stochastic since B_a preserve probability distributions. Denoting $P_t = M_{-\log a}$, the family $(\mathbf{P}_t)_{t\geq 0}$ is a multiplicative semigroup of stochastic matrices: $\mathbf{P}_t\mathbf{P}_s = \mathbf{P}_{t+s}$. This means they are transition matrices of some time-homogeneous Markov chain X — background on Markov processes is provided by appendix A.2. In this case case, X is a birth-and-death process on $\mathbb{N}$, and more specifically a pure death process. The infinitesimal generator $\mathbf{Q} = (q_{ij})_{i,j\in\mathbb{N}}$ of $(\mathbf{P}_t)_{t\geq 0}$ is the lower triangular matrix given by:

$$\forall i, j \in \mathbb{N}, \quad q_{ij} = \begin{cases} -j & \text{if } i = j \\ j & \text{if } i = j+1 \\ 0 & \text{else} \end{cases}, \qquad \mathbf{Q} = \begin{pmatrix} 0 & & & & \\ 1 & -1 & & & \\ & 2 & -2 & & \\ & & 3 & -3 & \\ & & & \ddots & \ddots \end{pmatrix}.$$

This fact can easily established by differentiating the generating function $Z[B_a d](z) = Z[d](az+1-a)$ with respect to a. Note this latter computation is mentioned in Johnson (2007, Proposition 3.7), equation (10) therein, together with the semigroup interpretation of B_a.

Now, it suffices to apply a result of Keilson and Kester (1978) about unimodality in Markov processes. The authors call $\mathcal{M}_2^\dagger$ the set of stochastic matrices that preserve unimodality by postmultiplication. Theorem 4.7 therein characterizes birth-death processes in $\mathcal{M}_2^\dagger$ by the condition

$$\forall n \in \mathbb{N}, \quad \lambda_{n-1} - 2\lambda_n + \lambda_{n+1} = \mu_n - 2\mu_{n+1} + \mu_{n+2},$$

with notations $\lambda_n := q_{n,n+1}$ and $\mu_n := q_{n,n-1}$. In the case of $\mathbf{P}_t$, $\lambda_n = 0$ and $\mu_n = n$, so the condition is fulfilled. We conclude that every $\mathbf{P}_t$ preserves unimodality, which means that every B_a does so. □

8.1.4 PRESERVATION OF HALF LOG-CONCAVITY

Now, we show that exponential tiling and binomial thinning preserve partial log-concavity. The proof is almost immediate thanks to lemma 8.6 and the following commutation relationship.

Lemma 8.8. For all $p \geq 0$, $\alpha \in (0, 1]$, there exists $q \geq 0$, $\beta \in (0, 1]$ such that:

(i) $E_p \circ B_\alpha = B_\beta \circ E_q$,

(ii) $p \geq 1$ if and only if $q \geq 1$.

Proof. Using the relationship $T_{a',b} \circ T_{a,b} = T_{a'a,b'a+b}$, one has the result with $\beta = \frac{p\alpha}{p\alpha+1-a}$, $q = 1 + \alpha(p-1)$. Then, $\beta \in [0, 1]$ and $q - 1$ as the same sign as $p - 1$. Besides, $q \geq 0$ if and only if $p \geq 1 - 1/\alpha$. As $\alpha \in (0, 1]$ and $p \geq 0$, this trivially holds. □

Proposition 8.9. Binomial thinning B_a such that $a \in [0, 1]$ preserves upper-half log-concavity and lower-half log-concavity.
Exponential tiling E_p such that $p \geq 1$ preserves upper-half log-concavity.
Exponential tiling E_p such that $p \leq 1$ preserves lower-half log-concavity.

Proof. [E_p] This is a consequence of $E_p \circ E_q = E_{pq}$.

Let d be a upper-half log-concave distribution. Let q be a real number such that $q \geq 1$. For all $p \geq 1$, one has $pq \geq 1$ and $E_p[E_q d] = E_{pq}[d]$. By lemma 8.6, $E_{pq}d$ is unimodal. As this holds for all $p \geq 1$, $E_q d$ is upper-half log-concave.

The proof for the lower counterpart is similar.

[B_a] Let d be an upper-half log-concave distribution. Let $a \in (0, 1)$ and q be in $(1, \infty)$. Owing to lemma 8.6, it suffices to show that $E_q B_a d$ is unimodal.

According to lemma 8.8, there exists $p \geq 1$, $\beta \in (0, 1)$ such that $E_q B_a d = B_\beta[E_p d]$. Since d is upper-half log-concave and $p \geq 1$, $E_p d$ is unimodal. So owing to lemma 8.7, $B_\beta[E_p d]$ is unimodal.

The proof is similar if d is lower-half log-concave instead of upper-half. Let q be in $(0, 1)$. According to lemma 8.8, there exists $p \in (0, 1)$, $\beta \in (0, 1)$ such that $E_q B_a d = B_\beta[E_p d]$. The end of the proof is similar. □

Using the compositions formulas, we can easily extend last result to the whole family $T_{a,b}$ and thus end the proof of our main proposition.

Proof of Proposition 8.1. Let a, b be two positive numbers.
According to lemma 8.3, $T_{a,b} = B_\alpha \circ E_{a+b}$ with $\alpha = \frac{a}{a+b} \in [0, 1]$.

1. Assume that $a + b \geq 1$. Then E_{a+b} preserves upper-half concavity as well as B_α. By composition, so does $T_{a,b}$.

2. Assume that $a + b \leq 1$. Then E_{a+b} preserves lower-half concavity as well as B_α. By composition, so does $T_{a,b}$. □

8.1.5 APPLICATIONS

To finish with, we still have to justify what we have claimed in the introduction. The following corollary of Proposition 8.1 extends Result 1. and 2. by relaxing their respective hypothesis. The corollary also improves log-concavity threshold $t_0 = 1$ given by Result 2.

Corollary 8.10. Let $P(z)$ be a power series with nonnegative coefficients and positive radius of convergence.

(i) If $P(z)$ has upper-half log-concave coefficients, then $P(z+t)$ is unimodal for all $t \geq 0$.

(ii) If $P(z)$ has upper-half log-concave coefficients, then $P(z+t)$ is log-concave for all $t \geq t_0$, where $t_0 = \min\{t \in [0,1) \mid (1-t)[\log P]'(t) \leq 1\}$. This set is non-empty and $t_0 < 1$.

Proof. Suppose P has upper-half log-concave coefficients.

[Claim (i)] As $P(z+t) = T_{1,t}P$ and $1 + t \geq 1$ for all $t \geq 0$. Proposition 8.1 tells $P(z+t)$ has upper-half log-concave coefficients too, so necessarily unimodal.

[Claim (ii)] By Proposition 8.1 again, B_aP has upper-half log-concave coefficients for all $a \in (0,1]$ as B_a preserves this property. If the sequence B_aP is also non-increasing, then necessarily B_aP is fully log-concave.

As B_aP has unimodal coefficients $(u_0, u_1, \ldots)$, they are non-increasing if and only if $u_0 \geq u_1$. These values may be retrieved by derivating the polynomial: $u_0 = B_aP(0)$ and $u_1 = B_aP'(0)$. Since $B_aP(z) = P(az+1-a)$, $B_aP(0) = P(1-a)$, $B_aP'(z) = aP'(az+1-a)$, then $B_aP'(0) = aP'(1-a)$. The condition $u_0 \geq u_1$ is equivalent to $P(1-a) \geq aP'(1-a)$, which reads $1 \geq (1-t)[\log P]'(t)$ by setting $t = 1-a$. Now, since exponential tiling preserves log-concavity, $E_{1/a}B_aP = P(z+1-a) = P(z+t)$ has log-concave coefficients. And if it does for some t_0, it does for any $t \geq t_0$ since $T_{1,b}$ preserves log-concavity (see Result 3. or Proposition 8.11).

So it remains to prove there exists such a a in $(0,1]$. For $a = 0$, $aP'(1-a) = 0$ whereas $P(1-a)$ since P has nonnegative coefficients. $P(1-a) > 0$. So by continuity, there exists $a > 0$ such that the inequality is true and $P(z+t)$ has log-concave coefficients. □

Operators T_t do not preserve unimodal sequence, not even non-increasing ones. The following example illustrates this point and is due to Alvarez et al. (2000).
If $P(x) = 123 + 13x + 12x^2 + 11x^3 + 10x^4 + 8x^5$,
then $P(x+1) = 177 + 150x + 185x^2 + 131x^3 + 50x^4 + 8x^5$,
which is not unimodal as $177 > 150$ and $150 < 185$. However, our result shows T_t preserve the subclass of non-increasing, log-concave coefficients.

8.1.6 CONCLUSION & PERSPECTIVES

In this section, we have extended three existing results in combinatorics using tools of TP_2 theory. These results are sufficient conditions for two reliability classes, unimodality

and log-concavity. To unify these two classes, we have introduced a third class, half log-concavity, which lies exactly in between the two. Then, we have proven preservation of half log-concavity (Proposition 8.1) by the family of $T_{a,b}$ operators we have introduced. Finally, we have shown how this new result extends the existing ones (Corollary 8.10).

Looking at the literature, we have noticed that related works in the field of combinatorics mainly deal with operators T_t, whereas related works in the field of discrete probabilities mainly deal with operators B_a. By considering T_t and B_a as special cases of a more general family $T_{a,b}$ operators, our approach has conceptually unified both research efforts.

Since our approach is built on preservation of half log-concavity, an interesting perspective is to study if other related properties are preserved by operators $T_{a,b}$. This idea is a motivation for the next two sections of this chapter, where further preservation results of this kind are established.

8.2 PRESERVATION OF LOG-CONCAVITY AND RELATED

In this section, we review further distributional properties that are preserved by exponential tilting E_p, binomial thinning B_a, and more generally by all operators $T_{a,b}$. This section deals with classes related to log-concavity. We begin by quoting known results, then we state and prove original results. Similarly to previous section, our method is to encompass all results with a single one that states a preservation.

8.2.1 CASE OF LOG-CONCAVITY AND LOG-CONVEXITY

The fact that $T : P(z) \mapsto P(z+1)$ preserves log-concavity is known under the name of Brenti's criterion (Brenti, 1989, Theorem 2.5.3).

In the literature, this proposition appears as soon as (Karlin, 1968, Theorem 7.2). To understand the formulation of the theorem therein, it suffices to know that a sequence p is log-concave if and only if the function $p(n+k)$ is RR_2 in $(n,k) \in \mathbb{Z} \times \mathbb{Z}$.

Proposition 8.11 (Brenti's criterion)**.** The operators $T_{a,b}$ preserves log-concavity on $\mathbb{N}$ for all $a, b \geq 0$.

We can directly deduce this result from Proposition 8.1.

Proof. Log-concave distributions are exactly those which are upper- and lower-half log-concave. Proposition 8.1 tells B_a preserves both, so it also preserves log-concave distributions. In addition, Proposition B.55 immediately shows that exponential tilting E_p preserve log-concavity for all $p \geq 0$. Composing binomial thinning and exponential tilting proves the preservation for any operator $T_{a,b}$. □

The counterpart result for log-convexity is also true, even we have never seen it explicitly in the literature.

Definition 8.12. Let $d = (d(n))_{n\in\mathbb{N}}$ be a real-valued sequence indexed on $\mathbb{N}$.

d is *log-convex* on $\mathbb{N}$ if d is nonnegative, $I = [\![a, \infty[\![$ for some $a \in \mathbb{N}$ and $\forall n \in \mathbb{N}, \quad d(n)d(n+2) \geq d(n+1)^2$.

Remark. With our definition, an log-convex sequence must be supported on $\mathbb{N}$, unless it is degenerated and supported on $\{0\}$.

Proposition 8.13. The operators $T_{a,b}$ preserve log-convexity on $\mathbb{N}$ for all $a, b \geq 0$.

Several proofs of this result are possible. We left it as a corollary of Proposition 8.20 hereafter.

8.2.2 CASE OF ULTRA-LOG-CONCAVITY AND ULTRA-LOG-CONVEXITY

Ultra log-concavity has been introduced by Pemantle (2000) and Liggett (1997). It is much studied in information theory and probability theory (Johnson, 2007; Johnson et al., 2013).

The *reverse* counterpart of ultra-log-concavity is a less standard property. It is implicitly mentioned by (Yu, 2009a) in a result on infinitely divisible distributions. We call it ultra-log-convexity and proves as well its preservation.

Definition 8.14. Let $d = (d(n))_{n \in J}$ be a real-valued sequence on $J \subset \mathbb{N}$.

- d is said to be *ultra-log-concave* (ULC) if the sequence $(n!d(n))_{n \in \mathbb{N}}$ is log-concave on $\mathbb{N}$.
- d is said to be *ultra-log-convex* (ULVX) if the sequence $(n!d(n))_{n \in \mathbb{N}}$ is log-convex on $\mathbb{N}$.

Remark. With our definition, an ultra-log-convex sequence must be supported on $\mathbb{N}$, unless it is degenerated and supported on $\{0\}$. On the contrary, an ultra-log-concave might be supported on any interval of $\mathbb{N}$.

Proposition 8.15. The operators $T_{a,b}$, for all $a, b \geq 0$, preserve ultra-log-concavity and ultra-log-convexity.

The result on log-concavity has been already given in Johnson (2007, Proposition 3.7), and the proof therein would be suitable for the log-convexity counterpart. However, we highlight this proof contains a small flow. The author claims that if a differentiable function $f : [0, 1] \to \mathbb{R}$ checks $f(1) \leq 0$ and $f'(x) \geq 0$ whenever $f(x) = 0$, then f is non-positive. The function $x \mapsto -(x - \frac{1}{2})^3$ gives a counterexample to this claim; it would be true with an additional convexity condition like $f''(x) < 0$.

We give two alternative proofs which are direct in the sens that they are not based on differentiation along a semigroup. Note that Proposition 8.20 hereafter provides a third proof.

First proof. This proof relies on two well-known preservation results. Refer to Proposition B.29 in appendix for the first claim and to Davenport and Pólya (1949, Paragraph 2) for the second one.

- The convolution of log-concave sequences is log-concave.

- The weighted sum of log-convex (on $\mathbb{N}$) sequences is log-convex.

As any E_p trivially preserves the ULC and ULVX classes, it suffices to prove the preservations for T and distributions p such that Tp exists.

[ULC] Let p be a ULC sequence. Define $q(n) = p(n)n!$.

$$Tp(n) = \sum_{k=n}^{\infty} \binom{k}{n} p(k) = \sum_{k=0}^{\infty} \binom{k+n}{n} p(n+k),$$

$$n!Tp(n) = \sum_{k=0}^{\infty} \frac{(n+k)!}{k!} p(n+k) = \sum_{k=0}^{\infty} q(n+k)(r-k) = q * r(n),$$

where $r(n) = \frac{1}{(-n)!}$ if $n \leq 0$, $r(n) = 0$ if $n > 0$.

p being ULC means q is discrete log-concave. Besides, the ratio $r(-n+1)/r(-n) = 1/(-n)$ is increasing, so r is discrete log-concave. As convolution preserves discrete log-concavity, $(n!Tp(n))_{n\in\mathbb{N}}$ is log-concave. This means that Tp is ULC.

[ULVX] Let p be a ULVX sequence. Then, $q(n) = p(n)n!$ is log-convex.

$$n!Tp(n) = \sum_{k=0}^{\infty} \frac{(n+k)!}{k!} p(n+k) = \sum_{k=0}^{\infty} \frac{1}{k!} q_k(n),$$

where q_k is defined by $q_k(n) := q(n+k)$ if $n \geq 0$ and $q_k(n) = 0$ if $n < 0$.

As q is log-convex on $\mathbb{N}$, the shifted sequences q_k are log-convex on $\mathbb{N}$. As weighted summation preserves log-convexity, $(n!Tp(n))_{n\in\mathbb{N}}$ is log-convex. This means that Tp is ultra-log-convex. □

We suggest an alternative proof for it bears similarities with other preservation results in next sections.

Second proof. This proof relies on the fact that all linear operators $T_{a,b}$ are totally positive of order 2. They are kernel operators with matrix $t_{k,n} = \binom{n}{k} a^n b^{n-k}$. The TP_2 property of such binomial matrix is a standard and elementary result. Indeed,

$$t_{k+1,n}/t_{k,n} \propto k/(k+1-n),$$

and this quantity is non-decreasing with n.

Moreover, ultra log-concavity / convexity may be defined using the size-biaising operator M defined on any sequence d by

$$Md(n) \stackrel{\text{def}}{=} (n+1)d(n+1), \text{ for all } n \in \mathbb{N}.$$

Let us prove that d is ULC if and only if $d \underset{\text{TP}}{\leqslant} M[d]$, and d is ULVX if and only if $d \underset{\text{TP}}{\geqslant} M[d]$. Indeed, $(n+1)!d(n+1)/n!d(n) = (n+1)d(n+1)/d(n) = Md(n)/d(n)$. Since $\mathrm{supp}[d(.+1)] = \mathrm{supp}[d] - 1$, we have $\mathrm{supp}[Md] \subset \mathrm{supp}[d]$ if and only if

If $a = 0$, then $T_{a,b}d \propto \delta_0$ which is ULC and ULVX. So it remains to prove the results for $a \neq 0$. We do using the following commutation relationship: $T_{a,b} \circ M = aM \circ T_{a,b}$.

[ULC] Assume d is ULC. Then, $d \underset{\text{TP}}{\leqslant} M[d]$. As $T_{a,b}$ preserves TP-relationship, $T_{a,b}d \underset{\text{TP}}{\leqslant} T_{a,b}M[d]$. As $T_a, bM[d] = aMT_{a,b}[d]$ and $a \neq 0$, one obtains

$$T_{a,b}d \underset{\text{TP}}{\leqslant} M[T_{a,b}d],$$

which proves that $T_{a,b}d$ is ULC.

[ULVX] Assume instead that d is ULVX. The proof is similar by reversing $\underset{\text{TP}}{\leqslant}$ into $\underset{\text{TP}}{\geqslant}$. □

8.2.3 GENERALIZATION WITH RELATIVE LOG-CONCAVITY ORDER

Log-concavity, log-concavity and their convex counterpart can be conveniently defined in terms of the relative log-concavity. This (partial) ordering between discrete sequences has been introduced by Whitt (1985).

Definition 8.16. Let f and g be two sequences on $\mathbb{N}$. Then f is *log-concave relative to* g, written as $f \underset{\text{lc}}{\leqslant} g$, if

(i) f and g are nonnegative sequences with no internal zero,

(ii) $\text{supp}[f] \subset \text{supp}[g]$,

(iii) $f(n)/g(n)$ is log-concave on $n \in \text{supp}[f]$.

Remark 8.17. It is easy to check that relative log-concavity defines a partial semi-ordering among nonnegative sequences. In addition, $f \underset{\text{lc}}{=} g$ if and only if $E_\lambda f = cE_\mu g$ for some $\lambda, \mu, c > 0$. This means f and g only differs from an exponential tilting. In other words, exponential tilting does not affect relative log-concavity. One has $f \underset{\text{lc}}{\leqslant} g$ if and only if $E_\lambda f \underset{\text{lc}}{\leqslant} cE_\mu g$.

The goal of this section is to prove that $T_{a,b}$ operators preserve relative log-concavity order, in the special case where one sequence belongs to the so-called $(a, b, 0)$ class of distributions.

Definition 8.18. A sequence d on $\mathbb{N}$ is said to belong to the $(a, b, 0)$ class of sequences if it is nonnegative and there exists $a, b \in \mathbb{R}$ such that

$$\forall n \in \mathbb{N}^*, \quad d_n = d_{n-1}\left(a + \frac{b}{n}\right).$$

Only three kinds of nonnegative sequences belongs to the $(a, b, 0)$ class. These sequences corresponds to un-normalized and exponentially-tilted versions of three common probability distributions: Poisson, negative binomial and binomial laws. Binomial distributions are special because they have finite support whereas the other ones are infinite and supported on $\mathbb{N}$.

- Poisson sequences: $d(n) = \left(\frac{p^n}{n!}\right)_{n\in\mathbb{N}}$ for some $p \geq 0$.
- Negative binomial sequences: $d(n) = \left(p^n \binom{n+r-1}{r-1}\right)_{n\in\mathbb{N}}$ for some $p \geq 0$ and $r > 0$.
- Binomial sequences: $d(n) = \left(p^n \binom{m}{n}\right)_{n=0\ldots m}$ for some $p \geq 0$ and $m \in \mathbb{N}$.

The reason why $(a, b, 0)$ sequences are special is the following fact: on this class, the action of any operator $T_{a,b}$ reduces to an exponential tilting E_p. In other words, the operators $T_{a,b}$ does not change their relative log-concavity. We show this property characterizes the $(a, b, 0)$ class.

Lemma 8.19. Let d be a nonnegative sequence. The following propositions are equivalent:

(i) d belongs to the $(a, b, 0)$ class,

(ii) $\forall a, b > 0, \quad T_{a,b} d \underset{\text{lc}}{=} d,$

(iii) $\forall a, b \geq 0, \quad \exists p \geq 0, c > 0, \quad T_{a,b} d = c\, E_p d.$

Proof. [(i) $\implies$ (iii)]

Negative binomial sequence $d(n) = \binom{n+r-1}{r-1}$ has for power series $P(z) = (1-z)^{-r}$. Its radius of convergence is 1 and for all $t \in (0, 1)$

$$P(z+t) = (1-t-z)^{-r} = (1-t)^{-r} \left(1 - \frac{z}{1-t}\right)^{-r} = (1-t)^{-r}\, E_{\frac{1}{1-t}} P(z).$$

Poisson sequence $d(n) = 1/n!$ has for power series $P(z) = e^z$. Its radius of convergence is infinite and for all $t > 0$,

$$P(z+t) = e^{z+t} = e^t\, e^z = e^t\, P(z).$$

Binomial sequence $d(n) = \binom{m}{n}$ has for power series $P(z) = (z+1)^m$. Its radius of convergence is infinite and for all $t > 0$,

$$P(z+t) = (z+t+1)^m = (t+1)^m \left(\frac{z}{t+1} + 1\right) = (1+t)^m\, E_{\frac{1}{1+t}} P(z)$$

[(iii) $\implies$ (i)] Consider P being the Z-transform d. The $(a, b, 0)$ recursion is equivalent to

$$(b+1)P(z) = (z-a)P'(z). \tag{8.1}$$

Assume (iii). Then, for every $t \geq 0$, there exists functions $b, c : [0, \infty) \to (0, \infty)$ such that

$$c(t)P(z) = P(tz + b(t)).$$

Evaluating this at 1 gives $c(t) = P(t + b(t))$, so

$$\forall t \geq 0, \quad P(t + b(t))P(z) = P(tz + b(t)).$$

One can easily shows that the function b is differentiable and may be chosen so that $b(1) = 0$. So differentiating with respect to t gives

$$\forall t \geq 0, \quad P'(t + b(t))(1 + b'(t)))P(z) = (z + b'(t))P'(tz + b(t)).$$

Evaluating at $t = 1$ gives

$$P'(1)(1 + b'(1)))P(z) = (z + b'(1))P'(z),$$

which proves that P checks equation 8.1 that characterizes $(a, b, 0)$ recursion. □

Case of infinite sequences. Now, we show that relative log-concavity is preserved in the two ordering directions when one sequence is in $(a, b, 0)$ class and has infinite support. This means one distribution has to be is either Poisson or negative binomial.

Proposition 8.20. Let f, g be two sequences. Assume that g is either a Poisson or negative binomial sequence. Then,

$$\begin{aligned} f \underset{\text{lc}}{\leqslant} g \quad &\Longrightarrow \quad \forall a, b \geq 0, \quad T_{a,b} f \underset{\text{lc}}{\leqslant} g, \\ g \underset{\text{lc}}{\leqslant} f \quad &\Longrightarrow \quad \forall a, b \geq 0, \quad g \underset{\text{lc}}{\leqslant} T_{a,b} f. \end{aligned}$$

Relative log-concavity encompasses previous distributional properties. So this proposition generalizes previous preservation results, and gives an unified and alternative proof.

- Log-concavity is equivalent to relative log-concavity with geometric sequences (p^n), which are a special case of negative binomial. A sequence d is log-concave (resp. log-convex) on $\mathbb{N}$ iff for any $p > 0$, $d(n) \underset{\text{lc}}{\leqslant}$ (resp. $\underset{\text{lc}}{\geqslant}$) $(p^n)_{n\in\mathbb{N}}$, .
- Ultra-log-concavity is equivalent to relative log-concavity with Poisson sequences $(1/n!)$. A sequence d is ultra-log-concave (resp. ultra-log-convex) iff for any $p > 0$, $d(n) \underset{\text{lc}}{\leqslant}$ (resp. $\underset{\text{lc}}{\geqslant}$) $\left(\frac{1}{n!}\right)_{n\in\mathbb{N}}$.

However, comparison with negative binomial sequences has never been considered in the literature except in the case $r = 1$ which corresponds to geometric sequences. Thus, the corresponding result is original.

Our proof is based on two ingredients. First one is total positivity of order 2. As explained in appendix B.1.3, this notion is connected to the variation-diminution property and sign changes. We say that a sequence h has property S (on $\mathbb{N}$) if $h(n)$ has at most two sign changes on $n \in \mathbb{N}$ and, in the case of two changes, the sign pattern is $-, +, -$. On the one side, TP_2 operators preserve property S. On the other side, property S characterizes unimodal and log-concave sequences (Proposition B.57 in the appendix).

Lemma 8.21. The linear operators $T_{a,b}$ are TP_2, for all $a, b \geq 0$.

Consequently, if f is a sequence with property S, then $T_{a,b} f$ has property S.

Second is the following property of commutation between T and E_p. As this relationship holds for negative binomial and Poisson sequences but not for binomial ones, explains why the latter are not covered by the proposition.

Lemma 8.22. The Poisson and negative binomial sequences check the following property:

$$\forall b \geq 0, p \geq 0, \quad \exists q \geq 0, c > 0, \quad E_p \circ T_b[d] = c\, T_b \circ E_q[d]. \tag{8.2}$$

Proof. Let $b, q \geq 0$.

[Negative binomial] Here, $d(n) = \binom{n+r-1}{r-1}$. So $P(z) = (1-z)^{-r}$ and $T_b d$ is defined only if $b \leq 1$. First, $T_b \circ E_q = T_{q,q} = E_q \circ T_q$. So

$$T_b \circ E_q[d] = E_q T_{bq} d = (1-q)^{-r}\, E_q E_{1/(1-q)} d = (1-q)^{-r}\, E_{q/(1-q)}[d].$$

Conversely,

$$E_p \circ T_b[d] = (1-b)^{-r} E_p E_{1/(1-b)}[d] = (1-b)^{-r} E_{p/(1-b)}[d].$$

In addition, $q/(1-q) = 1/A$ if and only if $q = 1/(1+A)$. So setting $q = \frac{1}{1+(1-b)/p} \geq 0$ and $c = \left(\frac{1-q}{1-b}\right)^{-r} \geq 0$ gives equation (8.2).

[Poisson] Here $d(n) = 1/n!$. So $P(z) = e^z$ and $T_b d$ is defined for all $b \leq 0$. First,

$$T_b \circ E_q P(z) = T_b e^{qz} = e^{q(z+b} = e^{qz} e^{qb}.$$

Conversely,

$$E_p \circ T_b P(z) = E_p e^{z+b} = e^{pz+b} = e^{pz} e^b.$$

So setting $q = p \geq 0$ and $c = e^{b(q-1)} \geq 0$ gives the equation. □

Now, we are able to base our proof on the characterization of log-concavity with exponential tilting and sign changes – see lemma 8.6.

Proof of Proposition 8.20. As E_p does not affect the lc-order, $E_b f \underset{\text{lc}}{=} f$ and $T_{a,b} f \underset{\text{lc}}{=} E_{1/b} T_{1,b} f = T[E_b f]$. So $E_b f$ can be substituted to f as it checks the same hypotheses. As result, it enough to prove the proposition for T.

Let us prove that that Tg/Tf is log-concave. Thanks to lemma 8.6, it suffices to show that $E_p Tg - cTf$ has property S for all $c > 0, p > 0$.

Let c, p be two such constants. By hypothesis, there exists $q > 0, d > 0$ such that $E_p Tg = dTE_q g$. This gives $E_p Tg - cTf = T[dE_q g - cf]$.

The hypothesis $g \underset{\text{lc}}{\leqslant} f$ means that g/f is log-concave. By lemma 8.6 again, this implies that $dE_q g - cf$ has property S. As the operator T is totally positive, the sequence $T[dE_q g - cf]$ has also property S. This ends the proof. □

Ultra-log-concavity of order k. It remains to deal with the case of g being a binomial sequence $g(n) = \binom{k}{n}$ for some $k \in \mathbb{N}^*$. Next proposition shows that relative log-concavity with binomial sequences is preserved by $T_{a,b}$, but *only in one ordering direction.*

Proposition 8.23. Let f, g be two sequences. Assume g is a binomial sequence. Then,

$$f \underset{\text{lc}}{\leqslant} g \qquad \Longrightarrow \qquad \forall a, b \geq 0, \quad T_{a,b} f \underset{\text{lc}}{\leqslant} g.$$

Actually, this relative log-concavity with binomial sequences corresponds to the notion of ultra-log-concavity m introduced by Pemantle (2000) and Liggett (1997).

Definition 8.24 (Liggett, 1997)**.** Let k be in $\mathbb{N}^*$ and p be a real-valued sequence.

p is *ultra-log-concave of order k* (ULC_k) if p is supported on $\{0, \ldots, k\}$ and if $\left(p(n)/\binom{k}{n}\right)_{n=0,\ldots,k}$ is log-concave (with $p(n) = 0$ if undefined).

Looking at this definition, we see that p is ULC_k if and only if $p(n) \underset{\text{lc}}{\leqslant} \binom{k}{n}$, for any $k \in \mathbb{N}^*$. Therefore, Proposition 8.23 is strictly equivalent to preservation of all ULC_k classes.

Corollary 8.25 (of Proposition 8.23)**.** The operator $T_{a,b}$, for all $a, b \geq 0$, preserves ultra-log-concavity of any order $k \in \mathbb{N}^*$.

Proof of Proposition 8.23. As $T_{a,b}f \underset{\text{lc}}{=} E_{1/b}T_{1,b}f = T[E_b f]$ and $E_b f \underset{\text{lc}}{=} f$, f can be substituted by $E_b f$ and $T_{a,b}f$ by Tf. So, it suffices to prove the result for T.

As $\text{supp}[f] \subset \{0, \ldots, k\} = \text{supp}[g] = \text{supp}[Tg]$, it can be easily checked that $\text{supp}[Tf] \subset \text{supp}[Tg]$.

If $f \equiv 0$, then $Tf \equiv 0$ and the result is true. Else, $a := \min \text{supp}[f]$ exists. Assume $f(a+1) = 0$. Then f is concentrated on a single point.

Now, assume $f(a+1) \neq 0$. This implies $g(a+1) \neq 0$. As the sequence f/g is log-concave, it is non-increasing if and only if $f(a)/g(a) \geq f(a+1)/g(a+1)$. Substituting f by $E_r f$ with $r := \frac{f(a)g(a+1)}{f(a+1)g(a)}$ ensures the inequality holds and f still checks the hypotheses.

So, one can assume that f/g is non-increasing. So it is log-concave if it is upper-half log-concave. To show that, it suffices to prove that $cTf - E_p Tg$ has property S for all $p \in (0, 1]$. Let p be such a real number. Then, $1-p \geq 0$ and $E_p \circ T = T_{1/p} \circ E_p = T \circ T_{p,1-p}$. As the operator T is linear, $cTf - E_p Tg = T[cf - T_{p,1-p}g]$.

By hypothesis, $T_{p,1-p}g \underset{\text{lc}}{=} g \underset{\text{lc}}{\geq} f$. This means that the sequence $cf - T_{p,1-p}g$ has property S. By lemma 8.21, the total positivity of the kernel of T implies that $T_b[cf - T_{p,1-p}g]$ has property S too. This shows $cTf - E_p Tg$ has property S and this ends the proof. □

Remark. Denote ULC_k the class of distributions that ultra-log-concave of order k, and ULC_∞ the class of ultra-log-concave ones. The following inclusions are immediate:

$$\text{ULC}_1 \subset \text{ULC}_2 \subset \ldots \subset \text{ULC}_\infty \subset \text{LCAV}.$$

Remark. The reverse ordering direction is not preserved by $T_{a,b}$. It corresponds to *reverse* counterpart of ultra-log-concavity. Such notions has been considered by Chen and Gu (2009) but is far less standard in the literature. Actually, one can show that for any finite sequence d, $T_t[d]$ is always ULC_∞ for high enough values of t.

8.2.4 CONCLUSION & PERSPECTIVES

We started this section by considering two known results: preservation of log-concavity and ultra-log-concavity by binomial thinning B_a. We have shown how such results and other related ones may be formulated as preservation of relative log-concavity with distributions in $(a, b, 0)$ class (Proposition 8.20) Then, we have given a general roof that holds for all members $(a, b, 0)$. This has led to original results such as preservation of ultra-log-concavity of finite order (Proposition 8.23). As a perspective, we wonder if *lc*-order is preserved by convolution with may for more general sequences.

To achieve results, we have stressed out the importance of half log-concavity. Then, we have proven B_a operators preserve such property (Proposition 8.1). As operators B_a may be embedded in Markov chains, an interesting perspective is to study preservation of half log-concavity by more general Markov chains. In this direction, we would like to draw a parallel with Proposition 7.24 (in previous chapter) on unimodal Lévy processes: a Lévy process X with lower-half log-concave initial distribution X_0 preserves

this property, providing it is unimodal and has nonnegative and DRHR increments. The alternative proof we have established in section 7.2.2 for this result sheds light on the necessity of DRHR increments: this hypothesis ensures the process preserves DRHR property. Again, as Lévy processes are particular Markov chains, an interesting perspective would be to study DRHR preservation by general Markov chains. This perspective partially motivates next section where we prove that B_a has such preservation property.

8.3 PRESERVATION OF DRHR AND IHR PROPERTIES

This section studies further properties of exponential tiling E_p and binomial thinning B_a. Here, we prove that B_a preserves the two DRHR, IHR reliability classes and E_p preserves one of the two. The situation is identical to preservation of half log-concavity (Proposition 8.1 in previous section). Such result is not surprising given that half log-concave distributions are subclasses of DRHR or IHR.

Afterwards, section 8.3.2 relates DRHR preservation to a problem in combinatorics raised by Johnson and Goldschmidt (2006). We rephrase the results stated in this reference using stochastic orders, and establish a relationship with stochastic monotony of additive processes — which has been studied in chapter 7. This approach allows us to seamlessly generalize such results.

8.3.1 PRESERVATIONS RESULTS

We briefly recall definitions of DRHR and IHR reliability classes and refer to appendix B.1 for further background. As exponential tilting is well-defined for sequences on $\mathbb{Z}$, related results are stated for sequences on $\mathbb{Z}$. This is not the case of binomial thinning which is only defined on $\mathbb{N}$.

Definition 8.26. Let $d = (d(n))_{n\in I}$ be a real-valued sequence indexed on $I \subset \mathbb{Z}$.
d is *DRHR* if the cumulative sequence $D(n) := \sum_{k\leq n} d(k), n \in \mathbb{Z}$ is log-concave.
d is *IHR* if the survivor sequence $\overline{D}(n) := \sum_{k\geq n} d(k), n \in \mathbb{Z}$ is log-concave.

Next proposition states the preservation result.

Proposition 8.27. Binomial thinning B_a such that $a \in [0, 1]$ preserves discrete DRHR and IHR sequences on $\mathbb{N}$.
Exponential tiling E_p such that $p \leq 1$ preserves DRHR sequences on $\mathbb{Z}$.
Exponential tiling E_p such that $p \geq 1$ preserves IHR sequences on $\mathbb{Z}$.

For completeness, we rephrase previous result with the same formulation as Proposition 8.1.

Corollary 8.28 (corollary of Proposition 8.27)**.** Let $P(z)$ be a power series with nonnegative coefficients and positive radius of convergence.

(i) Suppose the coefficients of $P(z)$ are discrete IHR .Then, for all $a \geq 0, b \geq 0$ such that $a + b \geq 1$, the coefficients of $P(az + b)$ are discrete IHR.

(i) Suppose the coefficients of $P(z)$ are discrete DRHR. Then, for all $a \geq 0, b \geq 0$ such that $a + b \leq 1$, the coefficients of $P(az + b)$ are discrete DRHR.

Proof. [B_a, DRHR] Let d be a sequence on $\mathbb{N}$. By definition, its cdf checks $D = d * e_1$, where $e_1 = (1)_{n\in\mathbb{N}}$ is the constant sequence. So a distribution p is DRHR if and only if $p \overset{\text{TP}}{\geqslant} p * e_1$. Assume d is DRHR. Then, $d \overset{\text{TP}}{\geqslant} d * e_1$. Applying B_a gives

$$B_a d \overset{\text{TP}}{\geqslant} B_a[d * e_1].$$

If $a = 0$, $B_a d = \delta$ which is DRHR, so the result is true. So now, assume $a \neq 0$. As B_a commutes with convolution and $B_a e_1 = 1/a e_1$, one obtains

$$B_a d \overset{\text{TP}}{\geqslant} [B_a d] * e_1,$$

which means $B_a d$ is DRHR.

[B_a, IHR] By definition, the survivor distribution of D checks $\overline{D} = e_1 * [\delta - d]$. So a distribution p is IHR if and only if $p \overset{\text{TP}}{\geqslant} e_1 * [\delta - p]$. Assume d is IHR. Then, $d \overset{\text{TP}}{\geqslant} e_1 * [\delta - d]$. Applying B_a gives

$$B_a d \overset{\text{TP}}{\geqslant} B_a \left\{e_1 * [\delta - d]\right\}.$$

As B_a is linear, commutes with convolution and $B_a\delta = \delta$, one obtains (for $a \neq 0$)

$$B_a d \overset{\text{TP}}{\geqslant} e_1 * [\delta - B_a d],$$

which means $B_a d$ is IHR. *[E_p, IHR]* Assume $p \geq 1$. The survival sequence of d_p is $\overline{D}_p(n) = \sum_{k=n}^{\infty} d(k)p^k$. An Abel transform gives

$$\sum_{k=n}^{\infty} p^k d(n) = \frac{p-1}{p} \sum_{k=n+1} \overline{D}(k)p^k + \overline{D}(n)p^{n-1},$$

so the inverse of the hazard rate equals

$$\overline{\overline{D}_p(n)}d_p(n) = \frac{p-1}{p} \sum_{k=1}^{\infty} \frac{\overline{D}(n+k)}{d(n)} p^k + \frac{\overline{D}(n)}{d(n)} p^{-1}.$$

Assume d is IHR. This reads $d \overset{\text{TP}}{\geqslant} \overline{D} \overset{\text{TP}}{\geqslant} \overline{D}(. + k)$ for all $k \in \mathbb{N}$. So the quantity above in non-increasing for all $n \in \mathbb{N}$ and $E_p d$ is IHR. □

8.3.2 RELATIONSHIP WITH CONVOLUTION CLASSES

We show that preservation of DRHR sequences by exponential tiling (Proposition 8.27) is equivalent to a result of Johnson and Goldschmidt (2006, Section 4). The authors

have introduced subclasses of probability mass functions on $\mathbb{N}$ related to log-concavity. Let d be such a pmf:

$$\mathcal{C}_d \overset{\text{def}}{=} \{f \text{ pmf on } \mathbb{N} \mid f * d \text{ is discrete log-concave } \}.$$

We call *convolution classes* such sets. The authors have focused on the ones $\mathcal{C}_{\mathcal{G}(p)}$ related to geometric distributions: for $p \in [0,1)$, the $\mathcal{G}(p)$ pmf is $g_p(n) := (1-p)p^n, n \in \mathbb{N}$. They have proven an interesting result of monotone inclusion.

Proposition 8.29 (Johnson and Goldschmidt, 2006, Theorem 4.4)**.** For all $p, q \in [0,1)$,

$$p \leq q \Leftrightarrow \mathcal{C}_{\mathcal{G}(p)} \subset \mathcal{C}_{\mathcal{G}(q)}.$$

First, we extend the definition to sequences supported on $\mathbb{Z}$ and to any $p > 1$. To do so, we introduce the un-normalized geometric sequences e_p that are defined for all $p \geq 0$ by $e_p \overset{\text{def}}{=} (p^n)_{n\in\mathbb{N}}$.

$$\mathcal{C}_p \overset{\text{def}}{=} \{f \text{ sequence on } \mathbb{Z} \mid f * e_p \text{ is discrete log-concave } \}, \quad p \geq 0.$$

We remark that $\mathcal{C}_1$ coincides with the class of DRHR sequences. Indeed, $d * e_1(n) = \sum_{k\leq n} d(k)$ which is the cumulative sequence D of d. This fact explains the equivalence between Proposition 8.29 and DRHR preservation.

Proposition 8.30. Let $p, q > 0$. Then $E_p\mathcal{C}_q = \mathcal{C}_{pq}$, and in particular,

$$d \in \mathcal{C}_p \iff E_{1/p}d \text{ is discrete DRHR.}$$

Consequently, the following propositions are equivalent:

(monotone inclusion) $\quad \forall q, p > 0, p \leq q \iff \mathcal{C}_p \subset \mathcal{C}_q.$

(DRHR preservation) $\quad E_p d$ preserves DRHR sequences, for all $p \in [0,1]$.

Proof. Since E_p commutes with convolution, for all $q > 0$,

$$\begin{aligned} d * e_q \text{ LCAV} &\iff E_p[d * e_q] \text{ LCAV} &&\iff E_p d * E_p e_q \text{ LCAV} \\ &\iff E_p d * e_{eq} \text{ LCAV} \\ d \in \mathcal{C}_q &\iff E_p d \in \mathcal{C}_{pq}. \end{aligned}$$

Applying this identity with $pq = 1$ gives

$$d \in \mathcal{C}_p \iff E_{1/p}d \in \mathcal{C}_1 \iff E_{1/p}d \text{ is discrete DRHR.}$$

□

Relationship with stochastic orders

Second, we rephrase the original proof of Proposition 8.29 using stochastic orderings. This proof consists in combining the two following statements.

- $d * e_p$ is log-concave if and only if $d * e_p \overset{\text{TP}}{\geqslant} d$.

- If $d * e_p$ is log-concave, then $d * e_p \overset{\text{TP}}{\leqslant} d * e_q$ for all $q \geq p$.

Proposition 8.31. Let p be in $(0, 1)$.

If $d*e_{p_0}$ is log-concave, then the family $(d*e_q)_{q\in[p_0,1)}$ is log-concave and non-decreasing in the likelihood ratio order:

$$\forall p, q \in [p_0, \infty), \qquad p \leq q \quad \iff \quad d * e_p \overset{\text{TP}}{\geqslant} d * e_q.$$

If $d*e_{p_0}$ is log-convex, then the family $(d*e_q)_{q\in(0,p]}$ is log-convex and *non-increasing* in the likelihood ratio order:

$$\forall p, q \in (0, p_0], \qquad p \leq q \quad \iff \quad d * e_p \overset{\text{TP}}{\leqslant} d * e_q.$$

The proof can be deduce from the following lemma which is original and might have its own interest. It has a nice probabilistic formulation that is similar to (Shaked and Shanthikumar, 2007, Theorem 1.C.53) but provides a weaker ordering with a weaker assumption on X.

Lemma 8.32. Let $X \sim d$ be a discrete random variable on $\mathbb{N}$ and $Y \sim \mathcal{G}(p)$ with $p \in (0, 1)$. Assume X, Y are independent.

$X + Y$ is discrete log-concave (resp. log-convex) if and only if

$$\forall n, n' \in \mathbb{N}, \quad n \leq n' \implies [Y \mid X + Y = n] \underset{\text{st}}{\leqslant} (\text{resp.} \underset{\text{st}}{\geqslant}) [Y \mid X + Y = n'].$$

Proof. Let d denote the pmf of X, and g_p the pmf of Y. Then, $g_p(n) \propto p^n$. Note that the pmf of $X + Y$ is $d * g_p$, which is always supported on an interval $[\![a, \infty[\![$ for some $a \in \mathbb{N}$.

Let $\overline{H}_s$ denote the survivor distribution of $[Y \mid X + Y = s]$. Then, for all $s \geq a$, $n \in \mathbb{N}$,

$$\overline{H}_s(n) = \frac{\sum_{k=n}^{s} p^k d(s-k)}{\sum_{k=0}^{s} p^k d(s-k)},$$

so with a change of variable $k \leftarrow s - k$,

$$\overline{H}_s(n) = \frac{\sum_{k=0}^{s-n} p^{s-k} d(k)}{\sum_{k=0}^{s} p^{s-k} d(k)} = p^n \frac{\sum_{k=0}^{s-n} p^{s-n-k} d(k)}{\sum_{k=0}^{s} p^{s-k} d(k)}$$
$$\overline{H}_s(n) = p^n \frac{d * g_p(s-n)}{d * g_p(s)}.$$

By definition, $[Y \mid X + Y = s]$ is *st*-non-decreasing in $s \in \mathbb{N}$ if and only if for all $n \in \mathbb{N}$, the survivor distribution $\overline{H}_s(n)$ is non-decreasing with s, *i.e.*, $s \mapsto \frac{d*g_p(s-n)}{d*g_p(s)}$ is non-decreasing. As $d * g_p$ is supported on an interval, this is the definition of $d * g_p$ being log-concave.

Similarly, $[Y \mid X + Y = s]$ is *st*-non-increasing in $s \in \mathbb{N}$ if and only if $s \mapsto \frac{d*g_p(s-n)}{d*g_p(s)}$ is non-increasing. As $d * g_p$ is supported on an unbounded interval, this is the definition of $d * g_p$ being log-convex. □

Proof of Proposition 8.31. Consider the un-normalized geometric sequences $e_p = p^n$ so that $d * e_p = d * g_p/(1-p)$. The families $(d * g_p)_p$ and $(d * e_p)_p$ have the same stochastic monotony.

[log-concave] For each n, the quantity $d * e_p(n)$ are continuously differentiable with respect to p. In this case, Proposition B.33 in the appendix states that $(d * e_p)_p$ if *lr*-non-decreasing if and only if

$$\forall n, n' \in \mathbb{N}, \qquad n \leq n' \quad \Longrightarrow \quad \partial_p \log[d * e_p](n) \leq \partial_p \log[d * e_p](n').$$

This quantity has a probabilistic interpretation:

$$\begin{aligned} \partial_p \log[d * e_p](n) &= \frac{\sum_{k=0}^n k p^{k-1} d(n-k)}{\sum_{k=0}^n p^k d(n-k)} \\ &= \frac{1}{p} \frac{\sum_{k=0}^n k p^k d(n-k)}{\sum_{k=0}^n p^k d(n-k)}, \\ p\, \partial_p \log[d * e_p](n) &= \sum_{k=0}^n k \frac{p^k d(n-k)}{\sum_{k=0}^n p^k d(n-k)}, \end{aligned}$$

so with notations of Lemma 8.32,

$$p\, \partial_p \log[d * e_p](n) = \mathbb{E}[Y \mid X + Y = n].$$

Therefore, the inequality above is equivalent to the following one:

$$\forall n, n' \in \mathbb{N}^*, \quad n \leq n' \iff \mathbb{E}[Y \mid X + Y = n] \leq \mathbb{E}[Y \mid X + Y = n'].$$

This inequality is implied by the basic stochastic ordering:

$$\forall n, n' \in \mathbb{N}^*, \quad n \leq n' \iff [Y \mid X + Y = n] \underset{\text{st}}{\leqslant} [Y \mid X + Y = n'],$$

and according to Lemma 8.32, this ordering holds if $X + Y$ is discrete log-concave, in other words if $d * g_p$ is discrete log-concave.

[log-convex] The proof is similar by reversing $\leq$ to $\geq$. □

Generalization to other distributions

Third, we raise the question: does Proposition 8.29 holds for other distributions than geometric laws? Thanks to stochastic orders formulation, we are able to extend it to (un-normalized) Negative Binomial sequences $NB(r,p) = \left(p^n \binom{n}{n+r-1}\right)_{n \in \mathbb{N}}$, as they encompasses geometric sequences: $e_p = NB(1,p)$.

Proposition 8.33. Let $p \geq 0, r \geq 1$.

If $d * NB(r,p)$ is log-concave, then the family $(d * NB(r,q))_{q \geq p}$ is log-concave and non-increasing in the likelihood ratio order:

$$\forall p', q \geq p, \quad p' \leq q \implies d * NB(r,p') \underset{\text{lr}}{\leqslant} d * NB(r,q).$$

In particular,

$$\forall p, q \geq 0, r \geq 1, \quad p \leq q \implies \mathcal{C}_{NB(r,p)} \subset \mathcal{C}_{NB(r,q)}.$$

Our proof is built on infinite divisibility of Negative Binomial sequences. First, we need two lemmas on the structure of $\mathcal{C}_p$.

Lemma 8.34. For all $p > 0$, the $\mathcal{C}_p$ is stable by convolution.

Proof. Let f, g be in $\mathcal{C}_p$ for some $p > 0$. This means $E_{p^{-1}} f, E_{p^{-1}} g \in \mathcal{C}_1$. Proposition B.29 in the appendix states that convolution preserve DRHR ($= \mathcal{C}_1$) sequences, so $E_{p^{-1}} f * E_{p^{-1}} g \in \mathcal{C}_1$. Moreover expnential tiling commutes with convolution: $E_{p^{-1}} f * E_{p^{-1}} g = [f * g]$. Therefore, $E_{p^{-1}}[f * g] \in \mathcal{C}_1$, which means that $f * g \in \mathcal{C}_p$. □

Next lemma is about infinitely divisible distributions which belongs to $\mathcal{C}_p$. Refer to appendix A.4 for definitions of infinitely divisibility, convolution semigroups and related notions.

Lemma 8.35. Let $p > 0$ and f be a pmf on $\mathbb{N}^*$. A convolution semigroup $(cp(l, f))_{l \geq 0}$ on $\mathbb{N}$ belongs to $\mathcal{C}_p$ if and only if its compounding measure f checks $f(n+1) \leq p f(n)$ for all $n \in \mathbb{N}^*$.

Proof. $(cp(l, f))_{l \geq 0} \in \mathcal{C}_p$ if and only if $(E_{1/p} cp(l, f))_{l \geq 0}$ is discrete DRHR. In addition, one can prove that $E_{1/p} cp(l, f) = cp\left(l, E_{1/p} f\right)$. Therefore, Proposition 7.25 tells $\left(cp\left(l, E_{1/p} f\right)\right)_{l \geq 0}$ is discrete DRHR if and only if $E_{1/p} f$ is non-decreasing, which reads $f(n+1) \leq p f(n)$. □

Next corollary slightly extends a result of Hansen (1988, Theorem 5) and Yamazato (1982, Theorem 2). It also provides an alternative proof that does not require the theory of Stieltjes transforms and completely monotone functions.

Corollary 8.36. Let a be in $(0, 1)$. Suppose $r(n) = a^n + (n+1) \int_0^a y^n M(\mathrm{d}y)$ for some nonnegative measure M on $[0, a]$ such that $\sum r(n)/(n+1) < \infty$.

Then, the infinitely divisible distribution d_r whose canonical sequence is r is log-concave.

Proof. It suffices to remark that $d_r = g_a * cp(\tilde{r})$ with $\tilde{r}(n) = (n+1) \int_0^a y^n \mathrm{d}M(y)$. Its corresponding compounding measure is $\tilde{f}(n) = \int_0^a y^n \mathrm{d}M(y)$ for all $n \in \mathbb{N}$. It is easy to see that it checks $f(n+1)/f(n) \leq a$, so last proposition gives $cp(\tilde{r}) \in \mathcal{C}_a$. □

With these lemmas, the proof of Proposition 8.33 is straightforward.

Proof of Proposition 8.33. Let $q > p \geq 0$. For all $r \geq 0$, there exists a nonnegative sequence b^r on $\mathbb{N}$ such that $NB(r, p) * b^r = NB(r, q)$. Indeed, $NB(r, q)$ is infinitely divisible and its canonical sequence is $k_q(n) = r q^{n+1}$. In addition, $k_q(n) - k_p(n) \geq 0$. So choosing b^r as the infinitely divisible with canonical sequence $k(n) = r(q^{n+1} - p^{n+1})$ gives the result. It also shows that $(d * NB(r, q))_{q \geq p}$ is a delayed additive process with non-negative increments.

Now, assume that $r > 1$. $d * NB(r, p)$ being log-concave means that $d * NB(r-1, p) \in \mathcal{C}_p \subset \mathcal{C}_q$. As seen above, there exists an infinitely divisible distribution $b := b^{r-1}$ such that $NB(r-1, p) * b = NB(r-1, q)$. Indeed, its canonical distribution is $k(n) = (r-1)(q^{n+1} - p^{n+1}) = (n+1)(r-1) \int_p^q y^n \mathrm{d}y$.

First, corollary 8.36 applies with $\mathrm{M}(\mathrm{d}y) = \mathbf{1}_{[p,q]}(y)$ and gives $b \in \mathcal{C}_q$.

Second, Lemma 8.34 tells $\mathcal{C}_q$ is stable by convolution. So $d * NB(r-1,p) * b = d * NB(r-1,q) \in \mathcal{C}_q$. This means $d * NB(r-1,q) * e_q = d * NB(r,q)$ is log-concave. As a conclusion, $(d * NB(r,q))_{q \geq p}$ is a delayed additive process that is log-concave. Proposition 7.11 tells that such a family is lr-non-decreasing. □

8.3.3 CONCLUSION & PERSPECTIVES

In this section, we have rephrased and extended an existing result of combinatorics (stated as Proposition 8.29) by combining two different tools. First tool is total positivity. As in previous sections of this chapter, we have generalized the original result by stating a preservation of some reliability classes. In addition, we have also related the original result to stochastic orders. Second tool is the theory of infinitely divisible distributions, since we have remarked that geometric laws is of one such distributions.

We suggest two perspectives for further research in this area. Though all this section is about discrete distributions, the very same results could be obtained with continuous ones. Poisson mixtures (introduced in appendix B.5) would be the right tool to transfer such results from discrete to continuous distributions. In addition, the convolution classes of more general distributions could be studied. So far we moved from geometric distributions to negative binomial distributions. Next step would be to consider distributions in the so called *Generalized Negative Binomial Convolutions* class introduced by Bondesson (1992, Chapter 8).

8.4 CONCLUSION

This chapter has delved into the connection between discrete probability theory and combinatorics. Looking at TP_2 theory, we notice that many fundamental results are about preservation of some reliability classes by some linear operators. This chapter has extended this approach by establishing original results of this kind (Propositions 8.20, 8.25, 8.1, 8.27). In parallel, we have discussed how such approach provides original methods to solve problems in combinatorics (sections 8.3.2 and 8.3.2) and seamlessly extend existing results.

As a perspective, we wonder to what extend some properties of peculiar discrete sequences studied in combinatorics could be proved by such general results. Preservation of log-concavity and unimodality is an important research topic in the field of combinatorics. The results we have obtained in this chapter suggest extending this research effort to other reliability classes and contemplate further TP_2-related tools such as stochastic orders.

9

CONCLUSION & PERSPECTIVES

The goal of this work is to study the behavior of probabilistic models where time, duration and occupancy are of utmost importance during inference and decoding of signals. We focus our attention on audio-to-score alignment applications, whether online (real-time score following) or offline. This work has investigated theoretical foundations of the design of probabilistic models for audio-to-score alignment. For such application, our motivation is to figure out which stochastic processes would provide a "good" model of unknown quantity which is the position along the music score during a performance. Our mathematical studies are followed by prescriptions for probabilistic modeling of time and duration in such applications, and proven both theoretically and in their real-world use cases.

To address this question, we have undertaken an axiomatic approach based on an application peculiarity: music scores provide nominal duration for each event, which is a hint for the actual and unknown duration. We have introduced *temporal coherency* as compliance with such prior information on duration. Two original criteria refine this abstract notion by defining coherent behaviors we expect from alignment algorithms. Criterion 1 is consistency with equivalent music score (chapter 3); criterion 2 is consistency with nominal duration of events when observation is non-discriminative (chapters 4 and 5). Once such axioms are settled, the core of this work shows that coherency is theoretically guaranteed by specific mathematical conditions and that fulfilling these prescriptions does improve precision of alignment algorithms. The conditions we have established are about the probability distributions on the duration of individual events, and about the method for Bayesian estimation of alignments. To do so, we have drawn upon tools used in other fields of probabilistic literature. The main ingredient is the theory of *total positivity of order 2*, stochastic orderings and reliability classes. The second one is the theory of *Lévy processes* and infinitely divisible distributions. As a result, mixing both ideas has motivated an investigation of total positivity of Lévy processes (chapter 7). Original results have been obtained to extend the related literature.

The key achievement of this work is to have found out the relevant mathematical concepts to formalize our coherency criteria. Such initial step was the most laborious and hazardous part of this research effort. Subsequent steps were much easier since many of the results we need are direct applications of the relevant tools. At the beginning, we were not aware of the mere existence of many concepts we have used for this work. Such mathematical tools have been hardly applied to score alignment so far, not even mentioned in the literature of Bayesian inference. Thus, having established fruitful connections between our applicative domain and other theoretical fields is certainly a original feature of this work.

Interest of the Notion of Coherency

Our definition of coherency has been motivated by specific *temporal structure* of music scores. Implicitly, we have defined this structure as follows: each music event corresponds to a *region* of an *homogeneous continuum*; the prior trajectory is a motion at constant velocity along this continuous space of beat positions. Such a Newtonian vision conflicts with the prevalent idea that an event is a singularity in time. Indeed, every real-world event — apart from silences — contains borderline phenomena such as non-stationary observations at onsets and release that contradict the Newtonian vision. In addition, many music sounds exhibit perceptible phases (like the common Attack, Decay, Sustain, Release description) that discrete probabilistic models are able to represent with micro-states. Nevertheless, information on signal content and its potential heterogeneities are blurred when observation is *non-discriminative*. Our study has shown that the homogeneous continuum is the only coherent model under non-discriminative observation. But when observation exhibit latent phases, a still pending challenge is to properly model the discrete structure of symbolic events with homogeneous regions.

Applicative Perspectives

Our criteria of coherency could be useful for other algorithmic approaches to score alignment. This study has only dealt with generative models like HMM, HSMM and with two kinds of estimation methods. First, one could apply our methodology to assess any other estimation method. Second, temporal coherency of more sophisticated models proposed in the literature could be investigated. We especially think about probabilistic models on *tempo*, for which coherency would mean favoring constant tempo evolutions. Third, discriminative models like Conditional Random Fields (Joder et al., 2010a) are close to HSMM so our results could be easily extended to such models.

But our axiomatic approach is not limited to probabilistic approaches. We believe it can guide the design of popular alternatives for alignment such that DTW-based algorithms. We have shown this in a side project (Lajugie et al., 2016), where we successfully applied the second criterion to design a coherent regularization for a cost function solved by convex optimization techniques. This attempt paves the road for other applications of our base concept.

In a different direction, a very interesting challenge would be to combine the prescriptions for coherency we have derived with a machine learning algorithm in order to optimize occupancy distributions with training data. Coherency leads to semi-parametric constraints on the space of probability distributions. An interesting issue would be to integrate such constraints in an estimation algorithm of occupancy distributions. How to design this kind of estimators? Could they enjoy statistical properties like consistency? As estimation of semi-parametric models (Powell, 1994) is an important research domain in statistics, this investigation could start from there.

Theoretical Perspectives & Open Questions

Coherency with respect to event durations may be achieved only if they fit specific bounds. Many questions about analytical properties of the bounding functions remain

open, even in the special case of Poisson and Negative Binomial laws. In addition, another interesting question is the existence of an optimal family of probability laws with regard to these bounds. We wonder if TP_2 theory could provide a tool to *compare* bounds between different families.

Coherency is ensured by TP_2 properties on semi-Markov chains such as stochastic monotony of the chain itself or its first-passage-time process. Results have been obtained for linear topologies, but we wonder if similar results could be derived for more general topologies. We suggest further investigating the case of chains with identical states, at least with the left-to-right topology.

On the way to these applicative results, several open questions about Lévy processes have been met. Some of them are likely to remain open for some time,, such as characterization of log-concave or unimodal infinitely divisible distributions. This is why we call for a systematic investigation of alternative reliability classes. For instance, the Poisson mixtures provide a powerful tool to transfer properties from discrete distributions to general measures. We wonder to what extent existing results on continuous Lévy processes could be explained through such discrete approximations. In addition, some characteristics of the first-passage-time process have been derived from the Lévy process itself. We wonder if this principle holds for further TP_2 properties.

A

PROBABILITY THEORY

A.1 BASICS ON MEASURES AND PROBABILITY DISTRIBUTIONS

A.1.1 GENERAL DISTRIBUTIONS ON THE REAL LINE

In this manuscript, a measure is always (at least) a nonnegative measure defined on the Borel algebra of $\mathbb{R}$ (the set of Borel sets).

Definition A.1. Let A be a Borel subset of $\mathbb{R}$.

- A *Borel measure* on A is a $[0,\infty]$-valued function μ on the Borel subsets of A such that
  i. $\mu(\emptyset) = 0$,
  ii. μ is *sigma-additive*: for all countable family $(A_n)_{n\mathbb{N}}$ of pairwise disjoint Borel subsets of A,
  $$\mu\left(\bigcup_{n\in\mathbb{N}} E_k\right) = \sum_{n\in\mathbb{N}} \mu(E_k).$$
- A *Radon measure* on A is a Borel measure μ that is locally finite, *i.e.*, $\mu([a,b]) < \infty$ for all finite interval $[a,b] \subset A$.
- A *finite measure* on A is a Borel measure such that $\mu(A) < \infty$.
- A *probability measure* on A is a Borel measure such that $\mu(A) = 1$.

Remark. A finite measure is always a Radon measure. A finite measure on A may always be seen as a finite measure on $\mathbb{R}$ that vanishes outside A.

Next result explains the correspondence between measures and non-increasing functions.

Proposition A.2 (distribution function)**.** Let A be a Borel subset of $\mathbb{R}$.

A *distribution function* is a function $F : A \to \mathbb{R}$ that is non-decreasing and right-continuous.

For any Radon measure D on A, there exists a distribution function on A such that

$$\forall a, b \in \mathbb{R}, \qquad (a,b] \subset A \quad \Longrightarrow \quad D((a,b]) = F(b) - F(a).$$

Remark. The same notation D is used for the cdf and the measure itself. We also use the functional notation $D(.)$ for the distribution function.

Any distribution function uniquely determines a Radon measure. But for a Radon measure several choices of distribution functions exist. In case of of finite measures such as probabilities, there is an natural choice of non-negative distribution function.

Definition A.3. Let D be a finite measure on $\mathbb{R}$.

- The *cumulative distribution function* (cdf) of D is the distribution function $D : \mathbb{R} \to \mathbb{R}_+$ defined by

$$\forall x \in \mathbb{R}, \qquad D(x) \stackrel{\text{def}}{=} D((-\infty, x]).$$

- The *survivor distribution* of D is the function $\overline{D} : \mathbb{R} \to \mathbb{R}$ defined by

$$\forall x \in \mathbb{R}, \qquad \overline{D}(x) \stackrel{\text{def}}{=} D([x, +\infty)).$$

Remark. The same notation D is used for the cdf and the finite measure itself. This is quite usual as they are in one-to-one correspondence. We also use the functional notation $D(.)$ for the cdf.

A.1.1.1 *Regularity of measures*

Definition A.4. A measure F is said to atom at $a \in \mathbb{R}$ if $F(\{a\}) > 0$.

If F is a Radon measure, its atoms coincide with discontinuities of (any) distribution function $F(.)$. Indeed, $F(\{a\}) = F(a) - F(a^-)$. It is known that any Radon measure on $\mathbb{R}$ has at most a countable number of discontinuities.

Definition A.5. Let D, μ be two Borel measures on a Borel subset I of $\mathbb{R}$.
D is said to be *dominated by* ($\gg$) μ, written $\mu \gg D$, if one of the following equivalent conditions holds:

(i) for all Borel set $A \subset I$, $\mu(A) = 0 \implies D(A) = 0$.

(ii) there exists a nonnegative μ-measurable function $f : I \to \mathbb{R}$ such that for all Borel set $A \subset I$,

$$D(A) = \int_A d(t)\mathrm{d}\mu(t) = \int_I \mathbf{1}_A(t) d(t)\, \mathrm{d}\mu(t).$$

Such a function d is called a *density*, or *Radon-Nikodym derivative* of D with respect to μ. It is also denoted $\frac{\mathrm{d}D}{\mathrm{d}\mu}$. The equivalence between the two statements is called the Radon-Nikodym theorem.

In practice, usual measures on $\mathbb{R}$ are either absolutely continuous or discrete. Other ones are called *mixed* measures.

Definition A.6. Let P be a (Radon) measure on a Borel subset I of $\mathbb{R}$.

- P is said to be an *absolutely continuous* measure if one of the following equivalent conditions holds:
  (i) F is dominated by the Lebesgue measure λ (restricted on I).

(ii) the cdf F is a locally absolutely continuous function on I.

(iii) there exists a nonnegative Lebesgue-measurable function $f : \mathbb{R} \to \mathbb{R}_+$ such that for all Borel set $A \subset I$,

$$P(A) = \int_A f(t)\,\mathrm{d}t.$$

This equivalently reads $P(\mathrm{d}t) = f(t)\,\mathrm{d}t$.

Such function f is said to be a *probability density function* (pdf) of P.

- P is said to be a *discrete* measure if one of the following equivalent conditions holds:

(i) F is dominated by the counting measure on $\mathbb{Z}$, $\mu_{\mathbb{Z}} \stackrel{\text{def}}{=} \sum_{n\in\mathbb{Z}} \delta_n$.

(ii) $\mathrm{supp}[P] \subset \mathbb{Z}$.

In this case, the *i* (pmf) of P is the nonnegative sequence $p : \mathbb{Z} \to \mathbb{R}_+$ such that for all $n \in \mathbb{Z}$,

$$p(n) \stackrel{\text{def}}{=} P(\{n\}) = P(n) - P(n^-).$$

Remark. A pdf f is not unique: any nonnegative Lesbegue-measurable function f such that $f \equiv g$ almost everywhere is another valid pdf. However, the pmf is unique.

The support means different thing whether speaking about measures or functions.

Definition A.7. The *support* (supp) of a measure D on $\mathbb{R}$ is defined as

$$\mathrm{supp}[D] \stackrel{\text{def}}{=} \mathbb{R} \setminus \cup \{\Omega \subset \mathbb{R} \mid \Omega \text{ is open and } D(\Omega) = 0\}.$$

The *support* (supp) of a function $f : A \subset \mathbb{R} \to \mathbb{R}$ is the closed set (relatively to A) defined as

$$\mathrm{supp}[f] \stackrel{\text{def}}{=} \overline{\{x \in A \mid f(x) \neq 0\}}.$$

The *essential support* (ess supp) of a function $f : A \subset \mathbb{R} \to \mathbb{R}$ is the closed set (relatively to A) defined as

$$\mathrm{ess\,supp}[f] \stackrel{\text{def}}{=} \mathbb{R} \setminus \bigcup \{\Omega \subset A \,|\, \Omega \text{ is open, } f = 0 \text{ almost everywhere in } \Omega\}.$$

The notions are related as follows: if a measure D is absolutely continuous, then for any pdf d of D, $\mathrm{supp}[D] = \mathrm{ess\,supp}[d]$.

Convolution product

Definition A.8. Let μ, ν be two Radon measures on $\mathbb{R}$.

The *convolution product* $(*)$ is the measure $\mu * \nu$ defined for all Borel set $A \subset \mathbb{R}$ by

$$\mu * \nu(A) = \iint_{\mathbb{R}\times\mathbb{R}} \mathbf{1}_A(t+x)\,\mathrm{d}\mu(t)\mathrm{d}\nu(x).$$

If μ, ν are finite, then so is $\mu * \nu$.
Its cumulative distribution function is $\forall x \in \mathbb{R}, \quad \mu * \nu(x) = \int_{\mathbb{R}} \mu(x-t)\mathrm{d}\nu(t)$.
Its survivor function is $\forall x \in \mathbb{R}, \quad \overline{\mu * \nu}(x) = \int_{\mathbb{R}} \overline{\mu}(x-t)\mathrm{d}\nu(t)$.

Convolution preserve probability measures: if $\mu(\mathbb{R}) = \nu(\mathbb{R}) = 1$, then $\mu * \nu(\mathbb{R}) = 1$.

A.1.2 DISCRETE DISTRIBUTIONS

For general measures, the cdf and survivor distributions are defined on the whole line $\mathbb{R}$. This section focuses on discrete distributions supported on $\mathbb{N}$. In this case, all probabilistic quantities may be restricted to $\mathbb{N}$.

Distributions on $\mathbb{N}$ A s discrete distribution $p : \mathbb{N} \to [0, 1]$ defined as

$$p(n) \stackrel{\text{def}}{=} \mathbb{P}(\{n\}).$$

A pmf can be any nonnegative sequence on $\mathbb{N}$ such that $\sum_{n\in\mathbb{N}} p(n) = 1$.

The *cumulative distribution function* (cdf) of p is the distribution P on $\mathbb{N}$ defined as

$$P(n) = \sum_{k=1}^{n} p(k)$$

Thus P checks $p(n + 1) = P(n + 1) - P(n)$ so it is as a non-decreasing sequence such that $\lim_{n\to\infty} P(n) = 1$.

The *survivor distribution* of the distribution $\overline{P} : \mathbb{N} \to [0, 1]$ defined by

$$\overline{P}(n) = \sum_{k=n}^{\infty} p(k) = 1 - \sum_{k=0}^{n-1} p(k)$$

Thus $\overline{P}$ checks $p(n) = \overline{P}(n) - \overline{P}(n + 1)$ so it is as a non-increasing sequence such that $\lim_{n\to\infty} P(n) = 0$.

Remark A.9. Let $\mathbf{1}$ be the constant sequence on $\mathbb{N}$ that identically equals 1.
The cdf is related to the pmf by a convolution: $P = p * \mathbf{1}$.
The survivor distribution is related to the cdf by $\overline{P}(n) = 1 - P(n - 1)$, *i.e.*,
$\overline{P} = \mathbf{1} - \delta_1 * P = \delta_1 * \mathbf{1} * [\delta_0 - p]*$.

Z-transforms The Z-transform is a defined for any sequence on $\mathbb{N}$. It is called the probability-generating function when applied to discrete probabilities.

Definition A.10. Let p be a sequence supported on $\mathbb{N}$.

The *probability-generating function* (pgf) of p is the power series defined on a subset of $\mathbb{C}$ as the Z-transform of p

$$Z[p](z) \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} p(n)\, z^n.$$

The *radius of convergence* of p, denoted R_p, is the number in $[0, +\infty]$ such that the series converges for all complex numbers z with $|z| < R_p$ and diverges if $|z| > R_p$.

When p is a pmf, $Z[p](1) = \sum_n |p(n)| = 1$ so $R_p \geq 1$. Since $p \geq 0$, $Z[p](z)$ is non-decreasing on $[0, R_p[$ thus it has a (finite or not) limit at R_p^-.

Proposition A.11. The Z-transform of the cdf is given by

$$Z[P] = \sum_{n=0}^{\infty} P(n)\, z^n = \frac{1}{1 - z} Z[p](z),$$

and its radius of convergence is 1.

The Z-transform of the survivor distribution is given by

$$Z[\overline{P}] = \sum_{n=0}^{\infty} \overline{P}(n)\, z^n = \begin{cases} 1 + \frac{z}{1-z}(1 - Z[p](z)) & \text{if } z \neq 1, \\ \sum_{n=0}^{\infty} np(n) = 1 + \mathbb{E}[X] & \text{else,} \end{cases}$$

and its radius of convergence is the same one as $Z[p]$.

Proof. The Z-transform is linear and exchanges convolution with product. In addition, $Z[\delta_1](z) = z$ and $Z[\mathbf{1}](z) = 1/(1-z)$, where $\mathbf{1}$ is the constant distribution that equals 1. So it suffices to apply the Z-transform to the relationships $P = \mathbf{1} * p$ and $\overline{P} = \mathbf{1} - \delta_1 * P$. □

The following formulas are useful in practice.

Discrete convolution Let X, Y be two independent random variables on $\mathbb{N}$ and let p_1, p_2 be their respective pmf. The sum $X + Y$ is another random variable. Its pmf p is given by the discrete convolution product

$$p(t) = \sum_{u=0}^{t} p_1(u) p_2(t-u), \qquad \textit{i.e.,} \qquad p = p_1 * p_2,$$

its cdf is

$$P = P_1 * p_2 = p_1 * P_2,$$

its survivor distribution is

$$\overline{P}(t) = \overline{P}_1(t) + \sum_{u=0}^{t-1} p_1(u)\overline{P}_2(t-u) = \overline{P}_2(t) + \sum_{u=0}^{t-1} p_2(u)\overline{P}_1(t-u),$$

which reads

$$\overline{P} = \overline{P}_1 + p_1 * (\overline{P}_2 - \delta) = \overline{P}_2 + p_2 * (\overline{P}_1 - \delta.), \tag{A.1}$$

its Z-transform is

$$Z[p] = Z[p_1 * p_2] = Z[p_1]Z[p_2].$$

Proof. For the cdf, $P = \mathbf{1} * p = \mathbf{1} * p_1 * p_2$, so $P = (\mathbf{1} * p_1) * p_2 = p_1 * (\mathbf{1} * p_2)$.

For the survivor distribution, one can use the relationship $\overline{P}(n) = 1 - P(n-1)$. □

A.2 MARKOV PROCESSES

This section presents discrete-time and (a subclass of) continuous-time Markov processes, together with inference formulas for discrete-time HMM.

A.2.1 DEFINITIONS

We present discrete-time and (a subclass of) continuous-time Markov chains. In our terminology, a *Markov process* is a stochastic process that checks the Markov property, and Markov chains are the sub-class of Markov processes that are supported on a countable state-space.

DISCRETE-TIME MARKOV CHAINS.

Definition A.12. A *Markov process* is a discrete time-process $X = (X_n)_{n\in\mathbb{N}}$ with values in space E that checks the Markov property: for all $n \in \mathbb{N}$, $x_1, \dots, x_n \in S$,

$$\mathbb{P}(X_{n+1} = x \mid X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_{n+1} = x \mid X_n = x_n).$$

A Markov process is said to be *time-homogeneous* if $\mathbb{P}(X_{n+1} = x \mid X_n = x_n)$ does not depend on n.

A *Markov chain* is a Markov process on a countable state space E.

Any discrete-time Markov chain X is characterized by two quantities:

- the *initial distribution* π which is the marginal law of X_0:
  $\forall j \in E, \qquad \pi(j) \overset{\text{def}}{=} \mathbb{P}(X_0 = j)$.
- the *transition matrices* $(\mathbf{P}(n))_{n\in\mathbb{N}}$, defined as $\mathbf{P}(n) \overset{\text{def}}{=} (p_{i,j}(n))_{i,j\in E}$ with
  $\forall i, j \in E, \qquad p_{i,j}(n) \overset{\text{def}}{=} \mathbb{P}(X_{n+1} = j \mid X_n = i)$.

Remark. Any transition matrix $\mathbf{P}$ belongs to the class of *stochastic matrices*: each row $p_{i,\cdot}$ of such matrices defines a probability distribution on E. A time-homogeneous chain is characterized by π and a single *transition matrix* $\mathbf{P} = (p_{i,j})_{i,j\in E}$.

The state distribution $\mathbf{f}(n)$ is defined as the marginal law of X_n, that is to say $f_j(n) = \mathbb{P}(X_n = j)$. For any discrete-time Markov chain X, state distributions fulfill the following recursion:

$$\forall n \in \mathbb{N}, \quad \mathbf{f}(n+1) = \mathbf{f}(n)\mathbf{P}(n), \quad \textit{i.e.}, \quad \forall j \in E, \quad f_j(n+1) = \sum_{i\in E} f_i(n)p_{i,j}(n).$$

CONTINUOUS-TIME MARKOV CHAINS.

Definition A.13. A *continuous-time Markov process* is a stochastic process $X = (X_t)_{t\geq 0}$ with values in space E such that

i. X is stochastically continuous: $\forall t \geq 0, \delta > 0, \lim_{h\to 0^+} \mathbb{P}(|X_{t+h} - X_t| > \delta) = 0$.
ii. For any set of times $0 =: t_0 \leq t_1 < \ldots < t_{N+1}$ and states $(x_i)_{i=0}^{N+1} \in S^N$ with $N \in \mathbb{N}$,

$$\mathbb{P}\big(X_{t_{N+1}} = x_{N+1} \mid X_{t_i} = x_i, \forall i \leq N\big) = \mathbb{P}\big(X_{t_{N+1}} = x_{N+1} \mid X_{t_N} = x_N\big).$$

A Markov process is said to be *time-homogeneous* if $\mathbb{P}(X_{t+s} = x \mid X_t = y)$ does not depend on $t \geq 0$, for all $s \geq 0$ and $x, y \in S$.

In the whole manuscript, we implicitly restrict to the subclass of uniformizable Markov chains.

Definition A.14. A *Markov chain* is a Markov process on a countable state space E.
A Markov chain is said to be *uniformizable* if

i. (*stable*) the following limits exists in $(-\infty, \infty)$,

$$\forall i, j \in E, \qquad q_{i,j}(t) \overset{\text{def}}{=} \begin{cases} \lim_{\Delta T \to 0} \frac{\mathbb{P}(X_{t+\Delta T}=j | X_t=i)}{\Delta T}, & j \neq i \\ \lim_{\Delta T \to 0} \frac{\mathbb{P}(X_{t+\Delta T}=i | X_t=i)-1}{\Delta T}, & j = i, \end{cases}$$

ii. (*conservative*) $\sum_{j \in E} q_{i,j}(t) = 0$ for all $t \geq 0$, $i \in E$,

iii. $\inf_{i,j \in E} q_{i,i}(t) > -\infty$ for all $t \geq 0$.

A continuous-time Markov chain X is characterized by two quantities:

- the *initial distribution* π which is the marginal law of X_0:
  $\forall j \in E, \qquad \pi(j) \overset{\text{def}}{=} \mathbb{P}(X_0 = j)$.
- the *infinitesimal generators* $(\mathbf{Q}(t))_{t \geq 0}$ defined as $\mathbf{Q}(t) \overset{\text{def}}{=} (q_{i,j}(t))_{i,j \in E}$.

A time-homogeneous chain is characterized by π and a single *infinitesimal generator* $\mathbf{Q} = (q_{i,j})_{i,j \in E}$.

Remark. With the continuity assumption, a Markov chain may be chosen such that it has almost surely *continuous from the right and limited from the left* (càdlàg) sample paths. In addition, any infinitesimal generator $\mathbf{Q}$ fulfills the balance condition

$$\forall i \in E, \quad \sum_{j \in E} q_{i,j} = 0.$$

For a large class of continuous-time Markov chains that includes uniformizable ones, marginal state distributions $\mathbf{f}(t)$ fulfill the *forward Kolmogorov equation*:

$$\forall t \geq 0, \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{f}(t) = \mathbf{f}(t)\mathbf{Q}(t), \qquad \textit{i.e.}, \qquad \forall j \in E, \quad \frac{\mathrm{d}}{\mathrm{d}t} f_j(t) = \sum_{i \in E} f_i(t) q_{i,j}(t).$$

A.2.2 INFERENCE ALGORITHMS ON HMM

This section considers inference with discrete-time and homogeneous HMM (S, O), where O is the observed process and S is a hidden Markov chain. Refer to section 2.3.1 for a definition of HMM. When observation $O_0^T = o_0^T$ is known up to some total time $T \in \mathbb{N}$, inference consists in computing probabilistic quantities related to posterior probabilities. The following algorithms are very standard and may be found in (Rabiner, 1989). They provides efficient computations with complexity is $O(JT)$, where $J = |E|$ is the number of states and T is total time.

FORWARD INFERENCE. Forward inference aims at computing posterior state probabilities $f_j(t)$ defined as

$$f_j(t) \stackrel{\text{def}}{=} \mathbb{P}\Big(S_t = j \mid O_1^t = o_1^t\Big).$$

We also introduce the un-normalized quantities for they simplify recursion formulas,

$$\tilde{f}_j(t) \stackrel{\text{def}}{=} \mathbb{P}\Big(S_t = j, O_1^t = o_1^t\Big).$$

The Forward algorithm is a recursive computation of $\tilde{f}_j(t)$ over $t = 0, \dots, T$.

1. Initialization: $\tilde{f}_j(0) = \pi(j) b_j(o_0)$.
2. Recursion: $\tilde{f}_j(t) = \Big(\sum_{i \in E} p_{i,j} \tilde{f}_i(t-1)\Big)\, b_j(o_t)$.

Normalized posterior probabilities $f_j(t)$ are obtained afterwards as $f_j(t) = \frac{\tilde{f}_j(t)}{\sum_{i \in E} \tilde{f}_i(t)}$.

VITERBI INFERENCE. Viterbi inference aims at computing posterior state probabilities $\delta_j(t)$ defined as

$$\delta_j(t) \stackrel{\text{def}}{=} \max_{s_1, \dots, s_{t-1} \in E} \mathbb{P}\Big(s_t = j, S_1^{t-1} = s_t^{t-1}, O_1^t = o_1^t\Big).$$

The Viterbi algorithm is a recursive computation of $\tilde{\delta}_j(t)$ over $t = 0, \dots, T$.

1. Initialization: $\delta_j(0) = \pi(j) b_j(o_0)$.
2. Recursion: $\delta_j(t) = \Big(\max_{i \in E} p_{i,j} \tilde{\delta}_i(t-1)\Big)\, b_j(o_t)$.

Remark. Normalized posterior probabilities could be obtained as

$$\delta_j(t) \stackrel{\text{def}}{=} \mathbb{P}\Big(S_t = j \mid O_1^t = o_1^t\Big) = \frac{\tilde{\delta}_j(t)}{\sum_{i \in E} \tilde{f}_i(t)}.$$

A.3 SEMI-MARKOV PROCESSES

This section presents required background on continuous-time and discrete-time semi-Markov chains, together with inference formulas for discrete-time HSMM. The only contribution it contains is section A.3.2 where an explicit procedure called *normalization* is described to get rid of null-duration occupancies and self-transitions in any semi-Markov chain.

A.3.1 DEFINITIONS

Definition A.15. Consider a set of random variables $((X_n, T_n))_{n \in I}$ on $I \stackrel{\text{def}}{=} \{0, \dots, N\}$ or $I \stackrel{\text{def}}{=} \mathbb{N}$.

- $X = (X_n)_{n\in I}$ a discrete-time homogeneous Markov chain on a state space E.
- $T = (T_n)_{n\in I}$ is a set of $[0,\infty]$-valued random variables such that almost surely,

$$\sum_{n\in I} T_n = \infty.$$

- The following conditional independence holds

$$\begin{aligned}\forall n \in \{0,\dots,N-1\}, t \geq 0, i, j \in E,\\ \mathbb{P}(T_{n+1} \leq t, X_{n+1} = j | (X_0, T_0), \dots, (X_n = i, T_n))\\ = \mathbb{P}(T_{n+1} \leq t \mid X_{n+1} = j)\mathbb{P}(X_{n+1} = j | X_n = i).\end{aligned}$$

- $\mathbb{P}(T_n \leq t \mid X_n = j)$ does not depend on $n \in I$.

The *continuous-time semi-Markov chain* is the process $S = (S_t)_{t\geq 0}$ defined as

$$\forall t \geq 0, \quad S_t \overset{\text{def}}{=} X_n \text{ for } n \in I \text{ such that } \sum_{k=0}^{n-1} T_k \leq t < \sum_{k=0}^{n} T_k.$$

If all occupancy distributions D_j are $\mathbb{N}$-valued, then the semi-Markov chain is piecewise constant on $[n, n+1)$. In this case, we rather consider the *discrete-time semi-Markov chain* $S \overset{\text{def}}{=} (S_n)_{n\in\mathbb{N}}$.

Therefore, three quantities define a semi-Markov chain:

- initial distribution π such that $X_0 \sim \pi$.
- transition probabilities $p_{i,j} = \mathbb{P}(X_{n+1} = j | X_n = i)$.
- *occupancy distributions* D_j such that $[T_n \mid X_n = j] \sim D_j$.

To define a semi-Markov chain, each occupancy distribution $(D_j)_{j\in E}$ may be chosen as any probability measure on $[0,\infty]$ and transitions probabilities $\mathbf{P} = (p_{i,j})_{i,j\in E}$ as any stochastic matrix.

A.3.2 NORMALIZATION OF SEMI-MARKOV STATE SPACE

The practitioner is free to choose the transitions probabilities $p_{i,j}$ a any valid stochastic matrix and occupancy distributions D_j as any probability distribution supported on $[0,\infty]$. However, there always exists an equivalent[1] chain such that

- No self-transitions are allowed: $p_{j,j} = 0$ (except for absorbing states such that $p_{j,j} = 1$).
- No null-duration[2] occupancies are allowed: $D_j(\{0\}) = 0$.

1. Here, "equivalent" means the two processes are identical in distribution.
2. As the distribution D_j is supported on $[0,\infty]$, $D_j(\{0\}) = D_j(0)$ and this equals $d_j(0)$ if D_j is discrete and d_j denote its pmf.

When fulfilling these two conditions, transition probabilities and occupancy distributions have the probabilistic interpretation – except for absorbing states – stated by Guédon (2003, Section 2),

$$p_{i,j} = \mathbb{P}(S_t = j \mid S_t \neq i, S_{t^-} = i), \qquad D_j(u) = \mathbb{P}(S_{t+u} \neq j \mid S_t = j, S_{t^-} \neq j).$$

Obtaining this equivalent chain is achieved by modifying the original chain parameters according to Steps 1-5 that we describe right now. This result seems to be original.

Self-transitions. If a state j has positive self-transition probabilities $p_{j,j} > 0$, the process S may enter again in j after having spent a random duration D_j. This means occupancy in j is actually longer, so D_j is not the *true* occupancy distribution. So the following transformation has to be done for each state $j \in E$ such that $0 < p_{j,j} < 1$.

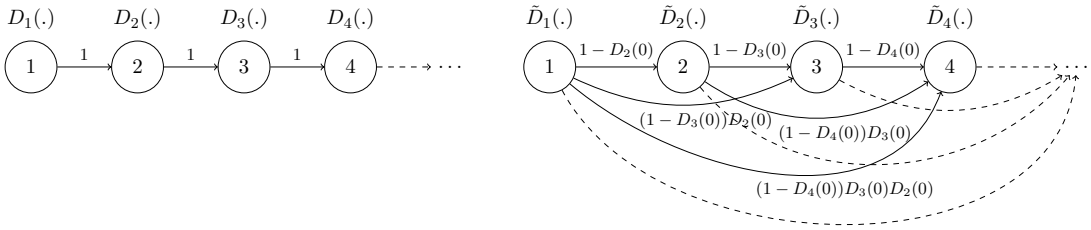
Step 1. Replacing occupancy distribution by $D_j \longleftarrow \sum_{n=0}^{\infty} \frac{(p_{j,j})^n}{1-p_{j,j}} D_j^{*n}$, where D_j^{*n} denotes the n-fold convolution of D_j.

Step 2. Replacing transition probabilities by $p_{j,i} \longleftarrow \frac{p_{j,i}}{1-p_{j,j}}$ (0 if $p_{j,j} = 1$) for $i \neq j$ and $p_{j,j} \longleftarrow 0$.

In the case $p_{j,j} = 1$, state j is absorbing. One may set $p_{j,j} = 0$ and $D_j = \delta_\infty$, but we do not so that $(p_{j,i})_{i \in E}$ always defines a valid probability distribution.

Null-duration occupancies. Semi-Markov chains allows choosing D_j as any probability distribution. However, $D_j(\{0\}) > 0$ would mean an occupancy of j is likely to have duration 0. This is dubious as entering a state j and spending a null duration is actually *not* entering on state j, but rather going to the subsequent state. Transforming D_j entails modifications of all other semi-Markov chains parameters: transition probabilities, but also initial probabilities.

Let us focus on the case of linear topology for it is easy. If null-duration occupancies are allowed, the equivalent chain topology is actually left-to-right. The following figures depics the original linear chain (left) and the equivalent left-to-right chain given by the transformation.



We call *linear in the weak sense* such left-to-right semi-Markov chains which stem from linear topology with null occupancies. They actually bear many similarities *strictly linear* for which null occupancies are not allowed ($D_j(0) = 0$). The relevant transformation has been described in (Wang and Xiao, 2006), and consists in three steps.

Step 3. Replacing transition probabilities $p_{i,j} = \delta_{i,i+1}$ for all $j > i$ by $p_{i,j} \longleftarrow (1 - D_j(0)) \prod_{k=i}^{j-1} D_k(0)$.

Step 4. Replacing initial probabilities $\pi(j) = \delta_{1,j}$ for all $j \in E$ by $\pi(j) \longleftarrow (1 - D_j(0)) \prod_{k=1}^{j-1} D_k(0)$.

Step 5. Truncating at 0 each occupancy distribution, which amounts to the substitution $D_j \longleftarrow \tilde{D}_j := \frac{D_j - D_j(0)\delta_0}{1 - D_j(0)}$.

Now, let us extend this transformation to other topologies. Let $\mathbf{P} := (p_{ij})_{i,j \in E}$ denote the transition probability matrix of S and $D_j(.)$ denote the cumulative distribution function (cdf). Define $\mathbf{D} := (D_j(0)\delta_{i,j})_{i,j \in E}$ as the $E \times E$ diagonal matrix whose diagonal values are probabilities of null occupancy. Let $\mathbf{I} := (\delta_{i,j})_{i,j \in E}$ denote the identity matrix of E. One may checks the transformation is equivalently described as follows:

$$\begin{array}{lllll} \text{Step 3.} & \mathbf{P} & \longleftarrow & \mathbf{P}\dfrac{\mathbf{I}-\mathbf{D}}{\mathbf{I}-\mathbf{DP}}, & \\ \text{Step 4.} & \pi & \longleftarrow & \pi\dfrac{\mathbf{I}-\mathbf{D}}{\mathbf{I}-\mathbf{DP}}, & \\ \text{Step 5.} & D_j(t) & \longleftarrow & \begin{cases} 0 & \text{if } t \leq 0, \\ \frac{D_j(t)}{1-D_j(0)} & \text{else.} \end{cases} & \end{array} \tag{A.2}$$

One can show this formulation is still valid for transforming a semi-Markov chain with *any* topology into a semi-Markov chain with non-null occupancies. This fact is out of the scope of thesis, but we have never found this transformation in the literature. So we left its proof as a future dedicated work.

Conjecture 4. For any chain topology, *i.e.*, for any transition matrix $\mathbf{P}$, the transformation described by equation (A.2) provides an equivalent semi-Markov chain thats has no null-duration occupancies.

A.3.3 INFERENCE ALGORITHMS ON HSMM

This section considers inference with discrete-time and homogeneous HSMM (S, O), where O is the observed process and S is a semi-Markov chain. Refer to section 2.3.2 for the definition of HSMM. When observation $O_0^T = o_0^T$ is known up to some total time $T \in \mathbb{N}$, inference consists in computing quantities related to posterior probabilities.

All following recursion formulas are standard and may be found in (Guédon, 2003). To be valid, they require D_j being the *true* occupancy distributions, in the sense that $D_j(\{0\}) = 0$ such that no null-duration occupancy is allowed. In case this assumption is false, the recursion has to be done on the equivalent HSMM given by the normalization procedure above.

RIGHT-CENSORED VERSUS SIMPLIFIED INFERENCE. As explained by Yu and Kobayashi (2003, Section 2.2.1), a simplifying assumption for computing posterior probabilities is to assume that at time t, the hidden process S is going to jumps on a different state at next time $t + 1$. This reads $S_{t+1} \neq S_t$. The "rigorous" inference should make no such assumptions. It is called *right-censored* inference and is detailed in (Guédon, 2003). Contrary to HMM, this difference matters for HSMM inference.

FORWARD INFERENCE. The Forward algorithm is a recursive computation to compute posteriors state probabilities $f_j(t)$, defined as

$$f_j(t) \stackrel{\text{def}}{=} \mathbb{P}_t\Big(S_t = j \mid O_1^t = o_1^t\Big).$$

While f_j corresponds to the right-censored inference, the simplifying assumption consists in using the following quantity

$$f_j^o(t) \stackrel{\text{def}}{=} \mathbb{P}_t\Big(S_{t+1} \neq j, S_t = j \mid O_1^t = o_1^t\Big).$$

We also introduce the the un-normalized quantities, as they make recursion formulas easier

$$\tilde{f}_j(t) \stackrel{\text{def}}{=} \mathbb{P}_t(S_t = j, O_1^t = o_1^t),$$
$$\tilde{f}_j^o(t) \stackrel{\text{def}}{=} \mathbb{P}_t(S_{t+1} \neq j, S_t = j, O_1^t = o_1^t).$$

The Forward algorithm provides efficient recursive computation of $\tilde{f}_j(t)$ over $t = 0, \ldots, T$.

$$\tilde{f}_j(t) = \overline{D}_j(t+1) \cdot b_j(o_0^t) \cdot \pi(j) + \sum_{u=1}^{t} \overline{D}_j(u) \cdot b_j(o_{t-u+1}^t) \cdot \sum_{i \neq j} p_{ij} \tilde{f}_i^o(t-u),$$

$$\tilde{f}_j^o(t) = d_j(t+1) \cdot b_j(o_0^t) \cdot \pi(j) + \sum_{u=1}^{t} d_j(u) \cdot b_j(o_{t-u+1}^t) \cdot \sum_{i \neq j} p_{ij} \tilde{f}_i^o(t-u).$$

Then, normalized quantities are obtained as

$$f_j(t) = \frac{\tilde{f}_j(t)}{\sum_{j \in E} \tilde{f}_j(t)}, \qquad f_j^o(t) = \frac{\tilde{f}_j^o(t)}{\sum_{j \in E} \tilde{f}_j(t)}.$$

Numerically, the difference between $\tilde{f}_j$ and $\tilde{f}_j^o$ amounts to replacing pmf d_j with survivor distribution $\overline{D}_j$. Note that the recursion could be directly expressed on normalized quantities f_j and f_j^o. This amounts to substitute $b_j(o_u)$ with $\tilde{b}_j(o_u) = b_j(o_u)/N_u$ in recursion formulas, where N_u is the normalizing factor $N_t \stackrel{\text{def}}{=} \mathbb{P}\Big(O_t = o_t \mid O_{t-1} = o_0^{t-1}\Big)$ introduced by Guédon (2003, Section 4.1). Indeed, one can show that

$$N_t = \sum_{j \in E} \tilde{f}_j(t) / \sum_{j \in E} \tilde{f}_j(t-1).$$

VITERBI INFERENCE. The Viterbi inference consists of estimating the most likely state-sequence $\hat{s}_t = (\hat{s}_t^1, \ldots, \hat{s}_t^t)$ conditionally to past observations o_1^t. As above, inference is computed with un-normalized quantities

$$\delta_j(t) = \max_{s_1, \ldots, s_{t-1}} \mathbb{P}(S_t = j, S_1^{t-1} = s_1^{t-1}, O_1^t = o_1^t),$$
$$\delta_j^o(t) = \max_{s_1, \ldots, s_{t-1}} \mathbb{P}(S_{t+1} \neq j, S_t = j, S_1^{t-1} = s_1^{t-1}, O_1^t = o_1^t).$$

The Viterbi algorithm is a dynamic programming method that provides efficient recursive computation of $\delta_j(t)$, $\delta_j^o(t)$ over $t = 0, \ldots, T$.

$$\delta_j(t) = \max\left[\overline{D}_j(t+1) \cdot b_j(o_0^t) \cdot \pi(j), \max_{u=1\ldots t}\left[\overline{D}_j(u) \cdot b_j(o_{t-u+1}^t) \cdot \max_{i \neq j}\{p_{ij}\delta_i^o(t-u)\}\right]\right],$$

$$\delta_j^o(t) = \max\left[d_j(t+1) \cdot b_j(o_0^t) \cdot \pi(j), \max_{u=1\ldots t}\left[d_j(u) \cdot b_j(o_{t-u+1}^t) \cdot \max_{i \neq j}\{p_{ij}\delta_i^o(t-u)\}\right]\right].$$

In the right-censored model the final state of the most likely state-sequence is defined by $\hat{s}_t^t = \arg\max_j \delta_j(t)$. The full sequence can be backtracked later on by storing the $\arg\max_{i \neq j}$ while computing $\delta_{\hat{s}_t}(t)$ and $\delta_j^o(u)$ for $u = 1 \dots t-1$.

Remark. The Viterbi and Forward quantities δ_j and $\tilde{f}_j$ have very similar expressions. Replacing all summations $\sum, +$ in the recursion of $\tilde{f}_j$ by max gives exactly δ_j.

A.4 INFINITELY DIVISIBLE DISTRIBUTIONS AND LÉVY PROCESSES

The material of this section is standard and comes from reference textbooks such as (Steutel and van Harn, 2004; Cont and Tankov, 2004; Sato, 1999). The only non-standard notions are additive processes introduced in section A.4.3.

A.4.1 GENERAL DEFINITIONS AND CHARACTERIZATIONS

Definition A.16. A probability measure μ on $\mathbb{R}$ is *infinitely divisible* (inf. div.) if, for any positive integer n, there is a probability measure $\mu_{1/n}$ on $\mathbb{R}$ such that

$$\mu = \underbrace{\mu_{1/n} * \dots * \mu_{1/n}}_{n \text{ times}}.$$

A real random variable measure X on is *infinitely divisible* if, for any positive integer n, there exists independent and identically distributed real random variables $X_{1/n,1}$, $X_{1/n,2} \dots$, $X_{1/n,n}$ such that

$$X = X_{1/n,1} + \dots + X_{1/n,n}.$$

The two definitions are equivalent: a random variable is inf. div. if and only if its probability measure is inf. div.

Another of view on infinitely divisibility is given by Lévy processes. Such process represents the motion of a point whose successive displacements are random and independent, and statistically identical over different time intervals of the same length. Thus, Lévy processes are the continuous-time analog of random walks.

Definition A.17. A stochastic process $X = (X_t)_{t \geq 0}$ on $\mathbb{R}$ is said to be a *Lévy process* if it satisfies the following properties.

Initial value. $X_0 = 0$ almost surely.

Independent of increments. For any $0 \leq t_1 < t_2 < \dots < t_n < \infty$, $X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \dots, X_{t_n} - X_{t_{n-1}}$ independent.

Stationary increments. For any $t, h \geq 0$, $X_{t+h} - X_t$ is distributed as X_h.

Stochastic continuity. For any $\epsilon > 0$ and $t \geq 0$, $\lim_{h \to 0} \mathbb{P}(|X_{t+h} - X_t| > \epsilon) = 0$.

If X is a Lévy process then one may construct a version of X that is càdlàg. This means that almost surely, its sample paths are right continuous and has left limits. So we suppose that all Lévy processes are càdlàg.

A third point of view on infinite divisibility is given by semigroups of probability measures where the group operation is the convolution product $*$.

Definition A.18. A family of probability measures $(\mu_l)_{l\in L}$ is said to be a *convolution semigroup* if

i. L is an additive subsemigroup of $\mathbb{R}_+$: for all $s,t \in L$, $s+t \in L$.

ii. for all $s,t \in L$, $\mu_s * \mu_t = \mu_{s+t}$.

A *continuous convolution semigroup* is a convolution semigroup indexed on $\mathbb{R}_+$ that is stochastically continuous in distribution: for all $t \geq 0$, $\mu_{t+\epsilon}$ converges in distribution to μ_t as $\epsilon \to 0$.

There is a one-to-one correspondence between infinitely divisible distributions, Lévy processes and continuous convolution semigroups. The marginal distributions of a Lévy process forms a convolution semigroup, and any infinitely divisible distribution may be embedded into a semigroup. This fact is explained in the following proposition.

Proposition A.19. The following propositions are equivalent.

(i) The probability measure μ is infinitely divisible.

(ii) There exists a Lévy process $X = (X_t)_{t\geq 0}$ such that X_1 is distributed as μ.

(iii) There exists a continuous convolution semigroup $(\mu_t)_{t\geq 0}$ such that $\mu_1 = \mu$.

In the two last cases, X_t is distributed as μ_t, and μ_t is the t-fold convolution of μ: for all $t \geq 0$, $\mu_t = \mu^{(*t)}$.

The fundamental theorem is the following characterization of infinitely divisible distributions.

Theorem A.20 (Lévy-Khintchine representation)**.** A probability measure μ on $\mathbb{R}$ is infinitely divisible if and only if there exists $a \in \mathbb{R}$, $\sigma \geq 0$, and a Radon measure ν on $\mathbb{R}\setminus\{0\}$ satisfying

$$\int_{\mathbb{R}} \min(1, x^2)\nu(x) < \infty,$$

such that for all $z \in \mathbb{R}$, $\int_{\mathbb{R}} e^{izx}\mu(\mathrm{d}x) = e^{\phi}(z)$ with

$$\phi(z) = -\frac{1}{2}\sigma^2 z^2 + aiz + \int_{\mathbb{R}\setminus\{0\}} \left(e^{izx} - 1 - izx\mathbf{1}_{|x|<1}\right)\nu(\mathrm{d}x). \tag{A.3}$$

In addition, the triplet (a, σ^2, ν) is uniquely determined by μ.

- ν is called the *Lévy measure* of μ.
- σ^2 is called the *Gaussian variance*.
- (a, σ^2, ν) is called the *characteristic triplet*.

The Lévy measure of a Lévy process $X = (X_l)_{l\geq 0}$ is defined as the Lévy measure of X_1. Indeed, the characteristic triplet of X_l is $(al, \sigma^2 l, \nu l)$.

The Lévy measure ν may be recovered from the convolution semigroup in the following way. Refer to the remark after (Theorem 4.4, Steutel and van Harn, 2004).

Proposition A.21. Let $(D_l)_{l\geq 0}$ be a continuous convolution semigroup. Let ν be the Lévy measure of D_1. For all continuity points x of ν,

$$\begin{aligned} \nu((-\infty, x]) &= \lim_{l\to 0} l D_l(x) && \text{if } x < 0, \\ \nu([x, \infty)) &= \lim_{l\to 0} l(D_l(x) - 1) && \text{if } x > 0. \end{aligned}$$

A.4.2 SPECIAL CASES OF LÉVY PROCESSES

A.4.2.1 *Compound Poisson processes*

Definition A.22. A *compound Poisson distribution* $(cp(\lambda, F))$ is a probability distribution $D = cp(\lambda, F)$ on $\mathbb{R}$ such that there exists $\lambda > 0$ and a probability distribution F on $\mathbb{R}$ such that

$$\forall x \in \mathbb{R}, \qquad D(x) = \sum_{n=0}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} F^{*n}(x),$$

or equivalently,

$$\phi_D(z) = \exp\left(\mathrm{i}z\lambda\left[\phi_F(z) - 1\right]\right),$$

where $\phi_D(z) \stackrel{\text{def}}{=} \int_{\mathbb{R}} e^{izx} D(\mathrm{d}x)$ and ϕ_F is defined similarly.
F is called the *compounding distribution* and λ is called the *intensity*.

A *compound Poisson process* is a process $X = (X_l)_{l\geq 0}$ such that $X \sim (cp(l, F))_{l\geq 0}$.

It is easy to see that a compound Poisson distribution is always infinitely divisible, and a compound Poisson process is always a Lévy process. Next proposition explains the reciprocal.

Proposition A.23. A Lévy process $X = (X_l)_{l\geq 0}$ is a compound Poisson process if and only if its Lévy measure is finite:

$$\nu\left(\mathbb{R} \setminus \{0\}\right) := \int_{\mathbb{R}\setminus\{0\}} \nu(\mathrm{d}x) < \infty.$$

In addition, $X_l \sim cp(\lambda l, F)$ with

$$\lambda \stackrel{\text{def}}{=} \int_{\mathbb{R}\setminus\{0\}} \nu(\mathrm{d}x) \qquad \text{and} \qquad F \stackrel{\text{def}}{=} \frac{1}{\lambda}\nu.$$

It is easy to see that $cp(\lambda, F)$ is discrete if and only if F is discrete. Last proposition shows that all discrete infinitely divisible distributions are compound Poisson.

A.4.2.2 *Nonnegative Lévy processes*

Nonnegative infinitely divisible distributions are related to non-decreasing Lévy processes, which are also called *subordinators*. Indeed, the following conditions are equivalent.

(i) D_t is supported on $\mathbb{R}_+$ for some $t > 0$.

(ii) D_t is supported on $\mathbb{R}_+$ for all $t \geq 0$.

(iii) X_t is almost surely nonnegative for some $t > 0$.

(iv) X_t is almost surely nonnegative for all $t \geq 0$.

(v) X has almost surely non-decreasing sample paths.

(vi) X has almost surely nonnegative increments $X_{t+h} - X_t$.

(vii) X has no Gaussian component, only positive jumps of finite variation and nonnegative drift. Equivalently, the characteristic triplet (a, σ, ν) of X (or X_1) checks the three following conditions

  (i) $\sigma = 0$.
  (ii) ν is supported on $\mathbb{R}_+$.
  (iii) $b \geq 0$ with $b := a - \int_{|x|\leq 1} x \, \mathrm{d}\nu(x)$.

Theorem A.24 (canonical representations)**.** A probability distribution μ is supported on $\mathbb{R}_+$ and infinitely divisible if and only if there exists $b \geq 0$ and a Radon measure K on $(0, \infty)$ satisfying

$$\int_{(1,\infty)} \frac{1}{x} \mathrm{d}K(x) < \infty,$$

such that for all $z \in \mathbb{R}$, $\int_{\mathbb{R}} e^{izx} \mu(\mathrm{d}x) = e^{\phi}(z)$ with

$$\phi(z) = izb + \int_{\mathbb{R}_+} \left(e^{izx} - 1\right) \frac{1}{x} \mathrm{d}K(x). \tag{A.4}$$

In addition, the couple (K, b) is uniquely determined by μ.

- K is called the *canonical measure* of μ.
- b is called the *drift*.

A.4.2.3 *Discrete Lévy processes on* $\mathbb{N}$

Theorem A.25 (canonical representation)**.** A probability distribution D on $\mathbb{N}$ is discrete infinitely divisible if and only if one the following conditions holds:

(i) D is infinitely divisible on $\mathbb{R}_+$, has no drift $b = 0$, and its canonical measure K is supported on $\mathbb{N}$ (or r equivalently, its Lévy measure ν is supported on $\mathbb{N}$).

(ii) There exists a nonnegative sequence on $\mathbb{N}$ $r = (r(n))_{n\in\mathbb{N}}$ satisfying

$$\sum_{n=0}^{\infty} \frac{r(n)}{n+1} < \infty,$$

such that the pmf d of D checks the *canonical equation* a.k.a. *Panjer's recursion*,

$$\forall n \in \mathbb{N}, \quad (n+1)d(n+1) = \sum_{k=0}^{n} d(k)r(n-k). \tag{A.5}$$

(iii) There exists a probability distribution F on $\mathbb{N}^*$ and a real number $\lambda \geq 0$ such that

$$\forall |z| \leq 1, \quad Z[d](z) = \exp(\lambda\,(Z[f](z) - 1)).$$

The quantities r and (λ, f) are uniquely determined by D.

- r is called the *canonical sequence*.
- for all $n \in \mathbb{N}$, $\quad r(n) = \lambda(n+1)f(n+1) = K(\{n+1\})$.
- for all $n \in \mathbb{N}^*$, $\quad f(n) = \lambda\nu(\{n\})$.

Therefore, any discrete infinitely divisible distribution D on $\mathbb{N}$ is necessarily a compound Poisson distribution $cp(\lambda, F)$ with compounding measure F supported on $\mathbb{N}^*$. For distributions we also use the notation

$$cp(r) \stackrel{\text{def}}{=} cp(\lambda, F).$$

Next proposition states that a discrete Lévy process is a particular kind of Markov chain whose Kolmogorov equation involves a convolution.

Proposition A.26 (Derivatives)**.** Let $D_l = cp(l, F)$ denote the compound Poisson distribution of intensity l and compounding measure F. Assume F is a discrete probability measure on $\mathbb{N}^*$ and let f denote its pmf. Then, for all $p \in \mathbb{N}^*$,

$$\begin{aligned}
\partial_{l^p}^p d_l &= d_l * \underbrace{(f - \delta_0) * \ldots * (f - \delta_0)}_{p \text{ times}}, \\
\partial_{l^p}^p D_l &= d_l * (F - 1) * \underbrace{(f - \delta_0) * \ldots * (f - \delta_0)}_{p-1 \text{ times}} = D_l * \underbrace{(f - \delta_0) * \ldots * (f - \delta_0)}_{p \text{ times}}, \\
\partial_{l^p}^p \overline{D}_l &= d_l * (\overline{F} - \delta_0) * \underbrace{(f - \delta_0) * \ldots * (f - \delta_0)}_{p-1 \text{ times}},
\end{aligned}$$

and in particular, for all $n \in \mathbb{N}$,

$$\begin{aligned}
\partial_l d_l(n) &= \sum_{u=1}^{n} f(u) d_l(n-u) - d_l(n), \\
\partial_l D_l(n) &= \sum_{u=1}^{n} F(u) d_l(n-u) - D_l(n) = \sum_{u=1}^{n} f(u) D_l(n-u) - D_l(n), \\
\partial_l \overline{D}_l(n) &= \sum_{u=1}^{n} \overline{F}(u) d_l(n-u).
\end{aligned}$$

Equivalently, let $(D_l)_{l \geq 0}$ denote the convolution semigroup of a Lévy process supported on $\mathbb{N}$. Let ν denote the Lévy measure, and $\overline{\nu}$ its survivor distribution restricted on $\mathbb{N}^*$. Then,

$$\forall l \geq 0, \qquad \partial_l \overline{D}_l = d_l * \overline{\nu}.$$

Equivalently, let $X = (X_l)_{l \geq 0}$ be a Lévy process supported on $\mathbb{N}$, and ν denote its Lévy measure. Then X is a continuous-time homogeneous Markov chain on $\mathbb{N}$ with initial distribution $\pi = \delta_0$ and infinitesimal generator

$$\mathbf{Q} = (q_{i,j})_{i \in \mathbb{N}, j \in \mathbb{N}} \qquad \text{with} \qquad q_{i,j} = \nu(\{i - j\}), \quad \forall i \neq j.$$

A.4.2.4 *Self-decomposable distributions*

Self-decomposability is one of the most important subclasses of inf. div. distributions. It is also called *class L* in the literature. We only quote the original definition of this class and its characterization by its Lévy measure. Refer to (Sato, 1991, Chapter 3) for a complete survey on this notion.

Definition A.27. Let F be a probability measure on $\mathbb{R}$. Let ϕ denote its characteristic function.
F is said to be *self-decomposable* if for all $\alpha \in [0,1]$, there exists a characteristic function ϕ_α of a probability measure on $\mathbb{R}$ such that

$$\forall t \in \mathbb{R}, \qquad \phi(t) = \phi(\alpha t)\phi_\alpha(t).$$

Equivalently, let X be a real random variable.
X is said to be self-decomposable if for all $\alpha \in [0,1]$, there exists a real random variable X_α such that

$$X = \alpha X' + X_\alpha,$$

where X' is distributed as X and X, X', X_α are independent.

Proposition A.28. A probability measure F on $\mathbb{R}_+$ is self-decomposable if and only if it is infinitely divisible and its canonical measure K is supported on $\mathbb{R}_+$ and non-increasing.

The discrete analog of self-decomposability has been introduced by Steutel and van Harn (1979).

Definition A.29. Let F be a discrete probability measure on $\mathbb{N}$. Let $P = Z[F]$ denote its Z-transform.
F is said to be *discrete self-decomposable* if for all $\alpha \in [0,1]$, there exists a Z-transform P_α of a discrete probability measure such that

$$\forall z \in (0, R_F), \qquad P(z) = P(1-\alpha+\alpha z)P_\alpha(z).$$

Proposition A.30. A probability measure F on $\mathbb{N}$ is discrete self-decomposable if and only if it is discrete infinitely divisible and its canonical sequence r is non-increasing.

A.4.3 DELAYED AND ADDITIVE PROCESSES

This section introduces slight generalizations of convolution semigroups.

Definition A.31. Let $(D_l)_{l\in L}$ be a family indexed on $L \subset \mathbb{R}_+$ of finite measures on $\mathbb{R}$.
$(D_l)_{l\in L}$ is said to be an *additive process* if

i. $D_0 = \delta_0$,

ii. for all $l, h \geq 0$, there exists a measure $D_{l,h}$ (called *increment*) such that

$$D_{l+h} = D_l * D_{l,h}.$$

An additive process is said to have *nonnegative increments* if for all $t, h \geq 0$, $D_{t,h}$ is supported on $\mathbb{R}_+$.

Now, we also define delayed processes. They consist is removing the assumption $D_0 = \delta_0$ to allow any arbitrary initial distribution.

Definition A.32. Let $(D_l)_{l\in L}$ be a family of measures.

$(D_l)_{l\in L}$ is a *delayed convolution semigroup* if there exists a finite measure D_0 on $\mathbb{R}$ and a convolution semigroup $(\tilde{D}_{l\in L})_{l\in L}$ such that $\forall l \in L, D_l = D_0 * \tilde{D}_l$.

$(D_l)_{l\in L}$ is a *delayed additive process* if there exists a finite measure D_0 on $\mathbb{R}$ and an additive process $(\tilde{D}_{l\in L})_{l\in L}$ such that $\forall l \in L, D_l = D_0 * \tilde{D}_l$.

A.4.4 EXAMPLES OF INFINITELY DIVISIBLE DISTRIBUTIONS

We list some examples of common infinitely divisible distributions. Further details are available in many textbooks like (Steutel and van Harn, 2004; Cont and Tankov, 2004; Sato, 1999; Johnson et al., 1993).

CONTINUOUS DISTRIBUTIONS. Except for Gaussian laws, all following distributions are supported on $\mathbb{R}_+$, have no drift nor Gaussian part.

Gaussian $\mathcal{N}(\mu, \sigma^2)$, with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$.
Its pdf is $d(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.
Its Lévy measure is $\nu \equiv 0$. Its Lévy triplet is $(\mu, \sigma, 0)$.

Gamma $\Gamma(k, \theta)$, with shape $k > 0$ and scale $\theta > 0$.
Its pdf is $d(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$.
Its cumulative distribution function is $D(x) = \frac{1}{\Gamma(x)}\gamma(t, \frac{x}{\theta})$.
Its characteristic function is $\phi(t) = (1 - \theta i t)^{-k}$.
Its Lévy measure is $\nu(\mathrm{d}t) = \frac{k}{t} e^{-\frac{x}{\theta}}$.
For $k = 0$, Gamma is defined as the degenerated distribution $\Gamma(0, \theta) = \delta_0$.

Exponential $\mathcal{E}(\lambda)$, with rate $\lambda > 0$.
Its pdf is $d(t) = \lambda e^{-\lambda t}$ on $\mathbb{R}_+$.
The exponential distribution is a special case of a Gamma function: $\mathcal{E}(\lambda) = \Gamma(1, \lambda)$.

Central chi-square χ^2_k, with $k > 0$.
Its probability density is $d(t) = \lambda e^{-\lambda t}$ on $\mathbb{R}_+$.
The exponential distribution is a special case of a Gamma function: $\mathcal{E}(\lambda) = \Gamma(1, \lambda)$.

Positive stable $S(\lambda, \alpha)$, with $\lambda > 0$ and $\alpha \in (0, 1)$. Its pdf has no tractable expression. Its Lévy measure is $\nu(\mathrm{d}t) = \frac{\lambda^{-1/\alpha}}{t^{1+\alpha}}$.

DISCRETE DISTRIBUTIONS. All distributions are supported on $\mathbb{N}$.

Poisson $Po(\lambda)$, with intensity $\lambda \geq 0$.
Its pmf is $p_n = e^{-\lambda}\frac{\lambda^n}{n!}$.
Their Lévy measure is the trivial distribution $f_n = \delta_1(n)$.
Both mean and variance of $Po(l)$ are equal to l.

Negative Binomial $NB(r,p)$, with rate $r \geq 0$ and failure rate $p \in [0,1)$. Its pmf if $d(n) = \binom{n+r-1}{n}(1-p)^r p^n$. Its Lévy measure is the *logarithmic distribution* $f_n = \frac{p^n}{-n\ln(1-p)}$.
The mean and variances are

$$\mathbb{E}[NB(r,p)] = r\frac{p}{1-p}, \quad \mathrm{Var}[NB(r,p)] = r\frac{p}{(1-p)^2}.$$

Geometric $\mathcal{G}(p)$, with failure rate $p \in [0,1)$.
Its pmf is $d(n) = (1-p)p^n$. The geometric distribution is a special case of negative binomial distribution: $\mathcal{G}(p) = NB(1,p)$.

Pòlya-Aeppli $PA(\lambda,p)$, with rate $\lambda > 0$ and $p \in [0,1)$. Its pmf is $p_0 = e^{-\lambda}$,

$$p_n = e^{-\lambda}\sum_{k=1}^{n}\binom{n-1}{k-1}\frac{(\lambda(1-p)/p)^k}{k!}.$$

Its Lévy measure is the (shifted) geometric distribution $f_n = (1-p)p^{n-1}$.

The mean and variances are

$$\mathbb{E}[PA(r,p)] = r\frac{1}{1-p}, \quad \mathrm{Var}[PA(r,p)] = r\frac{1+p}{(1-p)^2}.$$

A Pòlya-Aeppli law is very similar to a Negative Binomial laws but have a larger spread. They are known under alternative names, like geometric-Poisson laws.

B

TOTAL POSITIVITY AND STOCHASTIC ORDERINGS

B.1 TOTAL POSITIVITY OF ORDER 2 AND TP ORDERING

B.1.1 DEFINITIONS

Definition B.1. Let I, J be two subsets of $\mathbb{Z}$.

A matrix $\mathbf{A} = (a_{ij})_{i\in I, j\in J}$ is said to be *totally positive of order two* (TP_2), denoted by $\mathbf{A} \in \mathrm{TP}_2$, if all its entries are non-negative and all its 2×2 minors are non-negative, *i.e.*, $a_{ij}a_{kl} \geq a_{kj}a_{il}$ whenever $i \leq k$ and $j \leq l$.

A 2-variables function $f(x,\theta) : \mathcal{X}' \times \Theta' \mapsto$ is said to be *totally positive of order 2* (TP_2) in $(x,\theta) \in \mathcal{X} \times \Theta$ (where $\mathcal{X} \subset \mathcal{X}' \subset \mathbb{R}$ and $\Theta \subset \Theta' \subset \mathbb{R}$) if it is nonnegative and checks

$$\forall x_1, x_2 \in \mathcal{X}, \forall \theta_1, \theta_2 \in \Theta,$$
$$x_1 \leq x_2,\ \theta_1 \leq \theta_2 \quad \Longrightarrow \quad f(x_1,\theta_2)f(x_2,\theta_2) \geq f(x_1,\theta_2)f(x_2,\theta_1).$$

Remark B.2. The TP_2 property is symmetric. A function $f(x,\theta)$ is TP_2 in (x,θ) if and only if it is TP_2 in (θ,x).

Remark. TP2 functions are related to TP_2 matrices as follows. A function $f(x,\theta)$ is TP_2 in $(x,\theta) \in \mathcal{X} \times \Theta$ if and only if for all $N \in \mathbb{N}^*$, $x_1, \ldots, x_N \in \mathcal{X}$, $\theta_1, \ldots, \theta_N \in \Theta$ such that $x_1 \leq \ldots \leq x_N$ and $\theta_1 \leq \ldots \leq \theta_N$, the matrix $\mathbf{F} = (f(x_i,\theta_j))_{i,j=1\ldots N}$ is TP_2.

Definition B.3. Two real functions f_1, g_1 on $I_1, I_2 \subset \mathbb{R}$ are said to satisfy the TP order $f_1 \underset{\mathrm{TP}}{\leqslant} f_2$ if $f_i(x)$ is TP_2 in $(x,i) \in I_1 \cup I_2 \times \{1,2\}$ (with $f_i(x) = 0$ if $x \notin I_i$).

Remark. With this notation, a function $f(x,\theta)$ is TP_2 in $(x,\theta) \in \mathcal{X} \times \Theta$ if and only if partial functions check

$$\forall \theta_1, \theta_2 \in \Theta, \qquad \theta_1 \leq \theta_2 \quad \Longrightarrow \quad f(.,\theta_1) \underset{\mathrm{TP}}{\leqslant} f(.,\theta_2).$$

Remark. $f \underset{\mathrm{TP}}{\leq} g$ if and only if f, g are non-negative and g/f is non-decreasing on $\mathrm{supp}[f] \cup \mathrm{supp}[g]$ (with the convention $a/0 = \infty$ for $a > 0$).

We also define a slight generalization of TP order so as to deal with functions of arbitrary sign. This relationship has been introduced by Joag-dev et al. (1995) but does not define a partial ordering.

Definition B.4 (Joag-dev et al. 1995, Section 2)**.** Two real functions f, g on $I_1, I_2 \subset \mathbb{R}$ are said to satisfy the DP relationship $f \underset{\mathrm{DP}}{\leqslant} g$ if

- f is non-negative,
- for every $x \leq y$, $f(x)g(y) \geq f(y)g(x)$.

Remark. As f is non-negative, last condition is equivalent to $x \mapsto g(x)/f(x)$ being non-decreasing on $\operatorname{supp}[f] \cup \operatorname{supp}[g]$ (with the convention $a/0 = \operatorname{sign}[a]\infty$ for $a \neq 0$)

Remark. If g is nonnegative, then $f \underset{\text{DP}}{\leqslant} g$ is equivalent to $f \underset{\text{TP}}{\leqslant} g$.

It is useful to generalize matrices and functions with kernel operators.

Definition B.5. A *kernel operator* K is a linear operator defined on real functions as follows.

- Let $\mathcal{X}, \mathcal{Y}$ be two Borel subsets of $\mathbb{R}$.
- Let σ be a sigma-finite nonnegative Borel measure on $\mathcal{X}$.
- Let $K(x, y)$ be a real function of $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

For a function $f : \mathcal{X} \to \mathbb{R}$, $Kf : \mathcal{Y} \to \mathbb{R}$ is the function defined by

$$\forall y \in \mathcal{Y}, \quad Kf(y) \stackrel{\text{def}}{=} \int_{\mathcal{X}} K(x, y)h(x)\mathrm{d}\sigma(x).$$

K is only defined for functions f such that for all $y \in \mathcal{Y}$, the integral $\int_{\mathcal{X}} K(x, y)h(x)\mathrm{d}\sigma(x)$ is absolutely convergent.

K is said to be a *TP_2 operator* if and only if its kernel $K(\cdot, \cdot)$ is TP_2 in $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

A matrix $\mathbf{M}$ indexed on $I \times J \subset \mathbb{Z} \times \mathbb{Z}$ can be identified with a kernel operator M that maps sequences on I to sequences on J. In this case, σ is the counting measure on $\mathbb{Z}$. The mapping corresponds to pre-multiplication: the row vector associated to the sequence Md is $\mathbf{dM}$, where $\mathbf{d}$ is the row vector associated to d.

B.1.2 PRESERVATION OF TOTAL POSITIVITY

Basic composition formula is a simple algebraic formula that has tremendous consequences on total positivity and its preservation While it can be expressed with continuous integrals or discrete sums, the following formulation encompasses both cases.

Theorem B.6 (Basic composition formula, Karlin 1968, Chapter 1.2)**.** Let K, L, and M be Borel-measurable functions on $\mathbb{R} \times \mathbb{R}$ and let σ be a sigma-finite Borel measure on $\mathcal{Y} \subset \mathbb{R}$. Assume that for all x, z,

$$M(x, z) = \int_{\mathcal{Y}} K(x, y)L(y, z)\mathrm{d}\sigma(y),$$

where the integral is assumed to converge absolutely.

Then for all x_1, x_2, z_1, z_2,

$$\begin{vmatrix} M(x_1, z_1) & M(x_1, z_2) \\ M(x_2, z_1) & M(x_2, z_2) \end{vmatrix} = \iint\limits_{\substack{y_1, y_2 \in \mathcal{Y} \\ y_1 < y_2}} \begin{vmatrix} K(x_1, y_1) & K(x_1, y_2) \\ K(x_2, y_1) & K(x_2, y_2) \end{vmatrix} \begin{vmatrix} L(y_1, z_1) & L(y_1, z_2) \\ L(y_2, z_1) & L(y_2, z_2) \end{vmatrix} \mathrm{d}\sigma(y_1)\mathrm{d}\sigma(y_2).$$

TP$_2$ FUNCTIONS. A simple application of this formula has the following consequence on the preservation of TP$_2$ functions.

Corollary B.7. Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be three Borel subsets of $\mathbb{R}$. Let σ be a sigma-finite Borel measure on $\mathcal{Y}$. Let $K : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, $L : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}_+$ be two Borel-measurable functions. Define

$$M(x, z) \stackrel{\text{def}}{=} \int_{\mathcal{Y}} K(x, y) L(y, z) \mathrm{d}\sigma(y),$$

where all integrals are assumed to converge.

If $K(x, y)$ is TP$_2$ in $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $L(y, z)$ is TP$_2$ in $(y, z) \in \mathcal{Y} \times \mathcal{Z}$, then $M(x, z)$ is TP$_2$ in $(x, z) \in \mathcal{X} \times \mathcal{Z}$.

This result has three several interesting formulations.

TP$_2$ MATRICES. In the case of matrices, last result has a simple interpretation: the set of TP$_2$ matrices is closed under multiplication.

Corollary B.8. Let M, N be two (possibly infinite) matrices such that MN is well defined.

If M and N are TP$_2$ matrices, then so is MN.

Remark. For the matrix multiplication being defined, M and N must be respectively indexed on $I \times J$, $J \times K$, and all sums $\sum_{j \in J} m_{ij} n_{jk}$ must be absolutely convergent.

OPERATORS AND TP ORDER. The set of TP$_2$ operators is closed under composition.

Corollary B.9. Let M, N be two operators such that $M \circ N$ is well defined.

If M and N are TP$_2$ operators, then so is $M \circ N$.

PRESERVATION OF TP RELATIONSHIP. Next proposition is stated for general operators. It works in particular when M is a matrix.

Corollary B.10. Let f, g be two functions or sequences. Let M be a kernel operator whose domain contains f, g.

If M is a TP$_2$ operator, then

$$f \underset{\text{TP}}{\leqslant} g \qquad \Longrightarrow \qquad Mf \underset{\text{TP}}{\leqslant} Mg.$$

If M is a TP$_2$ operator, then

$$f \underset{\text{DP}}{\leqslant} g \qquad \Longrightarrow \qquad Mf \underset{\text{DP}}{\leqslant} Mg.$$

B.1.3 TP$_2$ AND VARIATION-DIMINISHING PROPERTY

Total positivity is related to preservation of number of sign changes. Refer to (Karlin, 1968, Section 1.3) for the general theory. Next proposition tells that TP$_2$ operator preserves functions with at most 2 sign changes. Here, the number of sign changes is counted discarding zero terms.

Theorem B.11 (Karlin 1968, Theorem 3.1, Chapter 1)**.** Let $\mathcal{X}, \mathcal{Y}$ be two Borel subsets of $\mathbb{R}$. Let σ be a sigma-finite nonnegative Borel measure on $\mathcal{X}$. Let $K(x,y)$ be a TP_2 function in $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Suppose $f : \mathcal{X} \to \mathbb{R}$ is a function such that $Kf(y) := \int_{\mathcal{X}} K(x,y)h(x)\mathrm{d}\sigma(x)$ is absolutely convergent.

If f has at most 2 sign changes on $\mathcal{X}$, then Kf has at most 2 sign changes on $\mathcal{Y}$. Moreover, if Kf has exactly 2 sign changes, then its sign pattern is the same as f.

B.2 TP_2 DISTRIBUTIONAL PROPERTIES

This section presents several properties for nonnegative functions, sequences and measures. In section B.2.1, log-concavity is introduced for functions and discrete log-concavity for sequences. Then, the two properties are generalized to measures. This notion is the most important one. Afterwards, the IHR and DRHR reliability classes are introduced in section B.2.2. Almost all definitions and properties are standard and can be found in many textbooks like (Shaked and Shanthikumar, 2007). The sets of probability measures that check one of such properties are also called *reliability classes*.

B.2.1 LOG-CONCAVITY

B.2.1.1 *Case of functions*

Definition B.12. A function $f : I \longrightarrow \mathbb{R}$ defined on $I \subset \mathbb{R}$ is said to be *log-concave* if

(i) f is nonnegative,

(ii) $\{x \in I \mid f(x) \neq 0\}$ is an interval of $\mathbb{R}$,

(iii) $\forall x, y \in \mathbb{R}, \forall \theta \in (0,1), \quad f(\theta x + (1-\theta)y) \geq f(x)^{\theta} f(y)^{\theta}$,

with the convention $f(x) = 0$ if $x \notin I$.

Remark. With these definition and convention, the property of log-concavity is *independent of the domain set* I. $f : I \to \mathbb{R}$ is log-concave if and only if $f : J \to \mathbb{R}$ is log-concave, for any set $J \subset \mathbb{R}$ such that $J \supset I$.

Remark. f being log-concave is equivalent to $f : \mathrm{supp}[f] \to \mathbb{R}$ being log-concave, which is equivalent to $\log f : \{x \in I \mid f(x) \neq 0\} \to \mathbb{R}$ being concave.

Next proposition is standard.

Proposition B.13 (regularity of log-concave functions)**.** Let f be a log-concave function, S denote its support, and $\mathring{S}$ the interior of S.

- f is continuous, right- and left- differentiable everywhere on $\mathring{S}$.
- f is almost everywhere differentiable on $\mathbb{R}$.
- f is locally absolutely continuous on $\mathring{S}$:

$$\forall a, b \in \mathring{S}, \quad f(b) - f(a) = \int_{[a,b]} f'_g = \int_{[a,b]} f'_d.$$

Many functions are log-concave only on a subset of their support. So for convenience we define partial log-concavity.

Definition B.14 (partial log-concavity)**.** Let J be an interval of $\mathbb{R}$ and I a subset of $\mathbb{R}$. A function $f : I \to \mathbb{R}$ is said to be *log-concave on* J if

(i) f is non-negative on J,

(ii) $\{x \in J \mid f(x) \neq 0\}$ is an interval of $\mathbb{R}$,

(iii) $\forall (x, y) \in J, \forall \theta \in (0, 1), \quad f(\theta x + (1-\theta)y) \geq f(x)^\theta f(y)^\theta$,

with the convention $f(x) = 0$ if $x \notin I$.

B.2.1.2 *Case of sequences*

Definition B.15. A discrete sequence $p = (p_n)_{n \in I}$ indexed on $I \subset \mathbb{Z}$ is said to be *discrete log-concave* if

(i) p is nonnegative,

(ii) p has no internal zero; in other words, the support of p is an interval of $\mathbb{Z}$,

(iii) $\forall n \in \mathbb{Z}, \quad p_n^2 \geq p_{n-1} p_{n+1}$,

with the convention $p_n = 0$ if $n \notin I$.

Remark B.16. We stress out the importance of condition (ii) in the definition of log-concavity we have chosen. Indeed, some authors defined log-concave sequences with conditions (i) and (iii) only. Sequences that fulfill the three conditions are also called Pólya Frequency sequences of order 2 (PF_2 sequences). The two notions do not coincide: for instance, the sequence $(1, 1, 0, 0, 1)$ checks the concavity inequalities but not the internal zeros condition. We refer to (Brenti, 1989, Chapter 2) for a discussion on this matter.

Remark B.17. With this definition, the property of log-concavity is *independent of the index set* I. $(p_n)_{n \in I}$ is log-concave if and only if $(p_n)_{n \in Z}$ is log-concave. It is also equivalent to $(p_n)_{n \in J}$ being log-concave for any $J \subset \mathbb{Z}$ such that $J \supset I$.

Many sequences does not satisfy the log-concavity inequality everywhere on their support. This is why we also define the partial notion.

Definition B.18 (partial log-concavity)**.** Let $[\![a, b]\!]$ be an interval of $\mathbb{Z}$, with $a, b \in \mathbb{Z}$. A sequence $p = (p_n)_{n \in I}$ indexed on $I \subset \mathbb{Z}$ is said to be *discrete log-concave on* $[\![a, b]\!]$ if

1. p is nonnegative on $[\![a-1, b+1]\!]$,
2. p has no internal zero on $[\![a-1, b+1]\!]$,
3. $\forall n \in [\![a, b]\!], \quad p_n^2 \geq p_{n-1} p_{n+1}$,

with the convention $p_n = 0$ if $n \notin I$.

This definition has straightforward extension to infinite intervals $[\![a, +\infty[\![$ or $]\!]-\infty, b]\!]$.

Remark. p being log-concave on $[\![a, b]\!]$ is not equivalent to the restriction $p_{|[\![a,b]\!]} \stackrel{\text{def}}{=} (p_a, p_{a+1}, \ldots, p_b)$ being log-concave. Actually it is equivalent to $p_{|[\![a-1,b+1]\!]}$ being log-concave. Moreover, a sequence p is log-concave if and only if it is log-concave on $\mathbb{Z}$.

B.2.1.3 *Relationship between log-concavity and total positivity*

Log-concavity has a connection with the concept of *total positivity* for which a whole body of literature has been devoted to — refer to (Karlin, 1968) for a survey. First relationship is the following characterization.

Proposition B.19. Let d be a discrete distribution on $\mathbb{Z}$.

$$\begin{aligned} d \text{ is discrete log-concave} &\iff \forall u \in \mathbb{N}, \quad d(.+u) \underset{\text{TP}}{\leqslant} d(.) \\ &\iff d(n-u) \text{ is } \mathrm{TP}_2 \text{ in } (n,u) \in \mathbb{Z}\times\mathbb{Z}. \end{aligned}$$

Let d be a function on $\mathbb{R}$.

$$\begin{aligned} d \text{ is log-concave} &\iff \forall u \in \mathbb{R}, \quad d(.+u) \underset{\text{TP}}{\leqslant} d(.) \\ &\iff d(n-u) \text{ is } \mathrm{TP}_2 \text{ in } (n,u) \in \mathbb{R}\times\mathbb{R}. \end{aligned}$$

Second, convolution with log-concave functions or distributions preserves TP ordering. This result is stated later on in Proposition B.41.

B.2.1.4 *Case of measures*

Log-concave measures are fundamental in many fields of probability theory. Such property several characterizations which are often used as alternative definitions.

Definition B.20. Let D be a locally finite measure on $I \subset \mathbb{R}$.

D is said to be *log-concave* (LCAV) if it is degenerated ($D = c\delta_a$, $a, c \in \mathbb{R}$), or absolutely continuous and admits a log-concave density.

D is said to be *discrete log-concave* if it is supported on $\mathbb{Z}$ and its probability mass function is log-concave.

Log-concave measures are an important subclass of *unimodal* measures. This matter is further discussed in section B.4.

Remark. A measure is both log-concave and discrete log-concave if and only if it degenerated.

Remark. Functions and sequences are particular cases of measures. All definitions are compliant. A function is log-concave if and only if

B.2.2 IHR AND DRHR

This section introduces the IHR and DRHR properties which are weaker than log-concavity. Even if such properties are only defined for measures in this section, functions and sequences are obtained as special cases. To begin with, let us recall definitions from section A.1.

The survivor distribution function is $\overline{D}(x) := D([x,\infty))$.

The cumulative distribution function is $D(x) := D((-\infty, x])$.

Another convention is to define the survivor distribution as $\overline{D}(x) = D((x,\infty))$ so that $\overline{D} = 1 - D$. Such alternative does not matter as all subsequent definitions and results are equivalent for both conventions.

Definition B.21. Let D be a finite nonnegative measure on $I \subset \mathbb{R}$.
D is said to be *Increasing Hazard Rate* (IHR) (resp. *discrete IHR*) if its survivor distribution function $\overline{D} : \mathbb{R} \mapsto \mathbb{R}$ (resp. $\mathbb{Z} \mapsto \mathbb{R}$) is a log-concave function (resp. sequence).
D is said to be *Decreasing Reverse Hazard Rate* (DRHR) (resp. *discrete DRHR*) if its cumulative distribution function $D : \mathbb{R} \mapsto \mathbb{R}$ (resp. $\mathbb{Z} \mapsto \mathbb{R}$) is a log-concave function (resp. sequence).

Remark. If a distribution d is supported on $\mathbb{N}$, we usually consider $\overline{D}$ or D as distributions restricted on $\mathbb{N}$. In the discrete case, such restriction does not alter the definitions provided the survivor distribution defined with $\overline{D}(x) = D([x, \infty))$.

Next proposition explains that IHR and DRHR properties are dual, in the sense that reversing the real line gives the other property.

Proposition B.22. A random variable X is (resp. discrete) DRHR if and only if $-X$ is (resp. discrete) IHR.

A measure D is (resp. discrete) DRHR if and only if the reflected measure $\tilde{D}$ is (resp. discrete) IHR, with $\tilde{D}[0, x] := D[-x, 0]$.

Next proposition in an immediate consequence of Proposition B.13. It tells that IHR or DRHR measure are absolutely continuous except maybe at one point.

Proposition B.23 (regularity)**.** An IHR measure D has at most one atom, at the right boundary of its support $\sup \operatorname{supp}[D]$ (no atom if infinite).

A DRHR measure D has at most one atom, at the left boundary of its support $\inf \operatorname{supp}[D]$ (no atom if infinite).

The IHR and DRHR properies have many characterizations which may be used as alternative definitions in the related literature. In probability theory, an important quantity is the hazard rate. Many textbooks introduce hazard rate first, IHR class then.

Definition B.24. Let D be a finite measure on $\mathbb{R}$. Let μ be a Radon measure on $\mathbb{R}$ that dominates D, and let d be a density of D.

The *hazard rate* h is defined by

$$\forall x \in \operatorname{supp}[\mu] \cap \{x \mid \overline{D}(x) \neq 0\}, \qquad h(x) \stackrel{\text{def}}{=} \frac{d(x)}{\overline{D}(x)}.$$

The *reversed hazard rate* h is defined by

$$\forall x \in \operatorname{supp}[\mu] \cap \{x \mid D(x) \neq 0\}, \qquad r(x) \stackrel{\text{def}}{=} \frac{d(x)}{D(x)}.$$

In practice, hazard is defined either for absolutely continuous measures (so h is a function on $\mathbb{R}$), or for discrete measures (h is sequence on $\mathbb{Z}$). In the former case, the definition of h depends on the chosen density.

Proposition B.25. Let D be a finite nonnegative measure on $\mathbb{R}$. Let $\overline{D}(.)$ denote its survivor distribution function and $D(.)$ its cumulative distribution function. Assume the measure D is discrete. Let d denote its pmf,

$$\begin{aligned}
D \text{ is discrete IHR} &\iff \text{hazard rate } h \text{ is non-decreasing (on } \mathbb{Z}) \\
&\iff \overline{D} \underset{\text{TP}}{\leqslant} d \\
&\iff \overline{D}(.+1) \underset{\text{TP}}{\leqslant} d \\
&\iff \forall u \in \mathbb{N}, \overline{D}(.+u) \underset{\text{TP}}{\leqslant} \overline{D} \underset{\text{TP}}{\leqslant} d(.). \\
D \text{ is discrete DRHR} &\iff \text{reverse hazard rate } r_d \text{ is non-increasing (on } \mathbb{Z}) \\
&\iff d \underset{\text{TP}}{\leqslant} D \\
&\iff d(.+1) \underset{\text{TP}}{\leqslant} D \\
&\iff \forall u \in \mathbb{N}, d(.+u) \underset{\text{TP}}{\leqslant} D(.+u) \underset{\text{TP}}{\leqslant} D(.).
\end{aligned}$$

Assume the measure is absolutely continuous. Let d denote a pdf, $\overline{D}$ denote its survivor distribution and D its cumulative distribution.

$$\begin{aligned}
D \text{ is IHR} &\iff \text{the hazard rate } h \text{ is non-decreasing (a.e. on } \mathbb{R}) \\
&\iff \overline{D} \underset{\text{TP}}{\leqslant} d \text{ (a.e.)} \\
&\iff \forall u \in \mathbb{R}_+, \overline{D}(.+u) \underset{\text{TP}}{\leqslant} \overline{D} \underset{\text{TP}}{\leqslant} d(.) \text{ (a.e.)}. \\
D \text{ is DRHR} &\iff \text{the reverse hazard rate } r_d \text{ is non-increasing (a.e. on } \mathbb{R}) \\
&\iff d \underset{\text{TP}}{\leqslant} D \text{ (a.e.)} \\
&\iff \forall u \in \mathbb{R}_+, d(.+u) \underset{\text{TP}}{\leqslant} D(.+u) \underset{\text{TP}}{\leqslant} D(.) \text{ (a.e.)}
\end{aligned}$$

Next proposition rephrase the previous one using shift operators T_a and the stochastic orders that will be introduced lateron on in section B.3. T_a is defined on random variables by $T_a X \overset{\text{def}}{=} X + a$ and on measures by $T_a D(\cdot) \overset{\text{def}}{=} D(\cdot - a)$. As a result, we get a characterization that sheds light on the close relationship between log-concavity, IHR and DRHR properties with lr, hr and rh stochastic orders.

Proposition B.26 (characterization with stochastic orders)**.** Let D be a finite measure on $\mathbb{R}$.
D is log-concave (resp. IHR, DRHR) if and only if $(T_a D)_{a \geq 0}$ is non-decreasing in the lr (resp. hr, rh) order.

Let D be a finite measure on $\mathbb{Z}$.
D is discrete log-concave (resp. IHR, DRHR) if and only if $(T_a D)_{a \geq 0}$ is non-decreasing in the lr (resp. hr, rh) order.

In other words, let X be a real random variable.

X is log-concave if and only if for all $a \geq 0$, $\quad X \underset{\text{lr}}{\leqslant} X + a$.

X is IHR if and only if for all $a \geq 0$, $\quad X \underset{\text{hr}}{\leqslant} X + a$.

X is DRHR if and only if for all $a \geq 0$, $\quad X \underset{\text{rh}}{\leqslant} X + a$.

Similarly to log-concavity, it is useful to define the partial counterpart of the IHR and DRHR properties.

Definition B.27 (partial properties)**.** A finite measure is said to be *IHR on* I, where $I \subset \mathbb{R}$, if it has no discontinuity on $\inf I$ and the survivor distribution function $\overline{D} : \mathbb{R} \to \mathbb{R}$ is log-concave on I.

A finite measure is said to be *discrete IHR on* I, where $I \subset \mathbb{Z}$, if the survivor distribution $\overline{D} : \mathbb{Z} \to \mathbb{R}$ is discrete log-concave on I.

Similarly, one defines the property of being DRHR/discrete DRHR on I (with $\sup I$ instead).

Equivalently, an absolutely continuous measure D is IHR on an interval $I \subset \mathbb{R}$ if and only its hazard rate is non-decreasing on I.

To finish with, next proposition states the logical implications between the different properties.

Proposition B.28. Log-concave distributions are IHR and DRHR.
Non-increasing distributions are DRHR.
Non-decreasing distributions are IHR.
Similar results hold for discrete distributions.

B.2.3 PRESERVATION RESULTS

Preservation of reliability classes is a fundamental aspect of TP$_2$ theory.

PRESERVATION BY CONVOLUTION. Log-concave, IHR, DRHR classes are closed under the convolution product. This result is among the most useful ones.

Proposition B.29 (Shaked and Shanthikumar, 2007)**.** Let F, G be two measures on $\mathbb{R}$.

- If F, G are log-concave, then $F * G$ is log-concave.
- If F, G are IHR, then $F * G$ is IHR.
- If F, G are DRHR, then $F * G$ is DRHR.

The same implications hold with discrete counterparts.

PRESERVATION BY LIMIT. All properties of this chapter are preserved under *convergence in distribution*. Besides, if a sequence of discrete measures converge to a continuous measure, then the discrete property is transformed into its continuous counterpart. This result is interesting for approximation of continuous measures by discrete ones.

Proposition B.30. Let C denote a property among { log-concavity, IHR, DRHR, unimodality }. Let $(F_n)_{n\in\mathbb{N}}$ be a sequence of measures that converges in distribution to a measure F.

Case 1. If all F_n have property C, then the limit F has property C.

Case 2. Assume F is discrete. If all F_n have discrete property C, then F has discrete property C.

Case 3. Assume each F_n is supported on $h_n\mathbb{Z}$ for some $h_n > 0$ and $\lim_n h_n = 0$. If all F_n have discrete property C (on $h_n\mathbb{Z}$ instead of $\mathbb{Z}$), then F has property C on $\mathbb{R}$.

B.3 STOCHASTIC ORDERINGS

Generalizing TP ordering between functions and sequences gives rise to several orderings between measures, called stochastic orders. Almost all definitions and results can be found textbooks like (Shaked and Shanthikumar, 2007). However, as minor contribution, this section provides original proofs for a few results. Indeed, for some results, every proof we have found in the literature makes assumptions on regularity of measures — they are usually assumed to be either discrete or absolutely continuous. This is why we have elaborated proofs that avoid such artificial assumptions.

B.3.1 DEFINITIONS AND CHARACTERIZATIONS

Definition B.31. Let F, G be two finite measures on $\mathbb{R}$. F is said to be smalled than G in the

- *basic stochastic order* (denoted $F \underset{\text{st}}{\leqslant} G$) if $\overline{F} \leq \overline{G}$, *i.e.*, for all $x \in \mathbb{R}$, $\overline{F}(x) \leq \overline{G}(x)$.
- *hazard rate order* (denoted $F \underset{\text{hr}}{\leqslant} G$) if $\overline{F} \underset{\text{TP}}{\leqslant} \overline{G}$.
- *reverse hazard rate order* (denoted $F \underset{\text{rh}}{\leqslant} G$) if $F \underset{\text{TP}}{\leqslant} G$.
- *likelihood ratio order* (denoted $F \underset{\text{lr}}{\leqslant} G$) if F, G are dominated by some measure μ and admits densities f, g such that $f \underset{\text{TP}}{\leqslant} g$.

Remark. The survivor distribution may either be defined with $\overline{F}(x) = F[x, \infty)$ or $\overline{F}(x) = F(x, \infty)$. The two conventions lead to equivalent definitions of *hr* order. Furthermore, for discrete distributions on $\mathbb{N}$, F and $\overline{F}$ may be considered as sequences on $\mathbb{N}$.

For convenience, we extend each definition to families of distributions that are indexed by a totally ordered set.

Definition B.32. A family of finite measures $(P_t)_{t\in I}$ indexed by $I \subset \mathbb{R}$ is said to be

- *non-decreasing in* st *order* ($\uparrow$ *st*) if one the following equivalent assertions holds:
  (i) $\forall t_1, t_2 \in I, \quad t_1 \leq t_2 \implies P_{t_1} \underset{\text{st}}{\leqslant} P_{t_2}$.
  (ii) $\forall t_1, t_2 \in I, \quad t_1 \leq t_2 \implies \overline{P}_{t_1} \leq \overline{P}_{t_2}$.
- *non-decreasing in* hr *order* ($\uparrow$ *hr*) if one the following equivalent assertions holds:

(i) $\forall t_1, t_2 \in I, \quad t_1 \leq t_2 \implies P_{t_1} \underset{\text{hr}}{\leqslant} P_{t_2}$.

(ii) $\overline{P}_t(x)$ is TP_2 in $t \in I$, $x \in \mathbb{R}$.

- *non-decreasing in* rh *order* ($\uparrow rh$) if one the following equivalent assertions holds:

(i) $\forall t_1, t_2 \in I, \quad t_1 \leq t_2 \implies P_{t_1} \underset{\text{rh}}{\leqslant} P_{t_2}$.

(ii) $P_t(x)$ is TP_2 in $t \in I$, $x \in \mathbb{R}$.

- *non-decreasing in* lr *order* ($\uparrow lr$) if one the following equivalent assertions holds:

(i) $\forall t_1, t_2 \in I, \quad t_1 \leq t_2 \implies P_{t_1} \underset{\text{lr}}{\leqslant} P_{t_2}$.

(ii) In case all P_t are dominated by some measure μ, they admit respective densities $p_t(.)$ such that $p_t(x)$ is TP_2 in $t \in I$, $x \in \mathbb{R}$.

The following characterizations are useful in case the family is smooth with respect to its parameter t.

Lemma B.33 (Lehmann, 1955, Lemma 1)**.** Let $(D_t)_{t\in I}$ be a family of finite measures indexed on an interval $I \subset \mathbb{R}$.

- If for all $x \in \mathbb{R}$, the functions $t \mapsto D_t(x)$ are continuously differentiable, then $(D_t)_{t\in I}$ is non-decreasing in rh order if and only if for all $t \in I$, $x \mapsto \partial_t \log D_t(x)$ is non-decreasing.

- If for all $x \in \mathbb{R}$, the functions $t \mapsto \overline{D}_t(x)$ are continuously differentiable, then $(D_t)_{t\in I}$ is non-decreasing in hr order if and only if for all $t \in I$, $x \mapsto \partial_t \log \overline{D}_t(x)$ is non-decreasing.

- Assume all D_t are dominated by some measure μ and admit respective densities d_t such that for all $x \in \text{supp}[\mu]$, the functions $t \mapsto d_t(x)$ are continuously differentiable. Then, $(D_t)_{t\in I}$ is non-decreasing in lr order if and only if for all $t \in I$, $x \mapsto \partial_t \log d_t(x)$ is non-decreasing μ-almost everywhere.

N.B. The convention $\log 0 = -\infty$ is used.

RELATIONSHIP BETWEEN ORDERS. lr order is stronger than hr and rh. These two orders are stronger than st.

The hr and rh orders are obtained one from the other by reflecting the real line. This is the same relationship as between IHR and DRHR properties (see Proposition B.22).

Proposition B.34. Let X, Y be two real random variables. Then,

$$X \underset{\text{hr}}{\leqslant} Y \qquad \iff \qquad -X \underset{\text{rh}}{\leqslant} -Y.$$

More generally, let D_1, D_2 be two measures. Define the reflected measures $\tilde{D}_i$ by $\tilde{D}_i[0, x] := D_i[-x, 0]$, $i \in \{1, 2\}$. Then,

$$D_1 \underset{\text{hr}}{\leqslant} D_2 \qquad \iff \qquad \tilde{D}_1 \underset{\text{rh}}{\leqslant} \tilde{D}_2.$$

CHARACTERIZATIONS OF *st* ORDER. The *st* order is defined for any finite measures but is mostly used to compare measures with equal mass. In this case, the two following characterizations may be used as definition.

Proposition B.35. Let F, G be two finite measures on $\mathbb{R}$. Assume that F and G have equal total mass, *i.e.*, $F(\mathbb{R}) = G(\mathbb{R}) < \infty$. Then the following propositions are equivalent:

(i) $F \underset{\text{st}}{\leqslant} G$.

(ii) For all $x \in \mathbb{R}$, $\overline{F}(x) \leq \overline{G}(x)$.

(iii) For all $x \in \mathbb{R}$, $F(x) \leq G(x)$.

CHARACTERIZATIONS OF hr/rh ORDERS. The TP_2 stochastic orders have many characterizations. The following one unifies hr, rh definitions with lr. Suppose a measure F is dominated by any arbitrary measure μ, written $\mu \gg F$. If F admits some function f as a density, we define with respect to μ the hazard rate $h_f(t) \overset{\text{def}}{=} f(t)/\overline{F}(t)$ for all $t \in \mathbb{R}$ such that $\overline{F}(t) > 0$ and reverse hazard rate $r_f(t) \overset{\text{def}}{=} f(t)/F(t)$ for all $t \in \mathbb{R}$ such that $F(t) > 0$.

Proposition B.36. Let F, G be two finite measures on $\mathbb{R}$.

- $F \underset{\text{hr}}{\leqslant} G$ if and only if for any nonnegative Radon measure μ that dominates F and G, they admit densities f, g w.r.t. μ such that $h_f(t) \geqslant h_g(t)$ for all $t \in \mathbb{R}$ such that $\overline{F}(t), \overline{G}(t) > 0$.
- $F \underset{\text{rh}}{\leqslant} G$ if and only if for any nonnegative Radon measure μ that dominates F and G, they admit densities f, g w.r.t. μ such that $r_f(t) \leqslant r_g(t)$ for all $t \in \mathbb{R}$ such that $F(t), G(t) > 0$.

The rigorous proof of this proposition requires the following strong result of measure theory.

Theorem B.37 (Besicovitch's derivation theorem, Leoni 2009, Theorem B.119)**.** Let μ, ν be two nonnegative Radon measures on a Borel set $A \subset \mathbb{R}$. If μ dominates ν, then there exists a Borel set $M \subset \operatorname{supp}[\mu]\mathbb{R}$ such that $\mu(M) = 0$ and

$$\forall x \in \mathbb{R} \setminus M, \quad \lim_{h \to 0^+} \frac{\nu[x-h, x+h]}{\mu[x-h, x+h]} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x) < \infty.$$

In other words, as h goes to 0, the functions $x \mapsto \nu[x-h, x+h]/\mu[x-h, x+h]$ pointwise converge μ-almost everywhere to the Radon-Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.

Remark. The validity of previous theorem comes from the following fact: by definition of the support, $x \in \operatorname{supp}[\mu]$ if and only if $\mu[x-h, x+h] > 0$ for all $h > 0$.

Proof of Proposition B.36. [rh order, $\Longrightarrow$] Assume $F \underset{\text{rh}}{\leqslant} G$ and $\mu \gg F, G$. Note this implies $\operatorname{supp}[F] \subset \operatorname{supp}[\mu]$. For all $x \in \operatorname{supp}[\mu]$ and $h > 0$,

$$F((x-h)^-)G(x+h) - F((x-h)^-)G((x-h)^-) \geq F(x+h)G((x-h)^-) - F((x-h)^-)G((x-h)^-).$$

By definition of the cdf, $F(x+h) - F((x-h)^-) = F[x-h, x+h]$. By definition of the support, $\mu[x-h, x+h] > 0$. Therefore,

$$\forall h > 0, \quad \frac{G[x-h,x+h]}{\mu[x-h,x+h]} F(x-h) \geq \frac{F[x-h,x+h]}{\mu[x-h,x+h]} G(x-h).$$

There exists a set $M \subset \text{supp}[\mu]$ such that $\mu(M) = 0$ and both ratios converge to the Radon-Nykodim derivatives $\tilde{f} = \frac{\mathrm{d}F}{\mathrm{d}\mu}$ and $\tilde{g} = \frac{\mathrm{d}G}{\mathrm{d}\mu}$. This implies

$$\forall x \in \mathbb{R} \setminus M, \quad \tilde{g}(x) F(x^-) \geq \tilde{f}(x) G(x^-).$$

As $\mu \gg F, G$, $F(M) = G(M) = 0$. So any function that coincides with $\tilde{f}$ (resp. $\tilde{g}$) on M is a valid density for F (resp. G). So the functions defined by

$$f(x) := \begin{cases} \tilde{f}(x) & \text{if } x \in M \\ G(x^-) & \text{if } x \notin M, \end{cases} \qquad g(x) := \begin{cases} \tilde{g}(x) & \text{if } x \in M \\ F(x^-) & \text{if } x \notin M, \end{cases}.$$

are valid densities for F, G and fulfill the following inequality everywhere:

$$\forall x \in \mathbb{R}, \quad g(x) F(x^-) \geq f(x) G(x^-).$$

Furthermore, by definition one has $F(\{x\}) = F(x) - F(x^-) = f(x)\mu(\{x\})$ and $G(\{x\}) = G(x) - G(x^-) = g(x)\mu(\{x\})$. So last inequality is equivalent to

$$\forall x \in \mathbb{R}, \quad g(x)[F(x) - f(x)\mu(\{x\})] \geq f(x)[G(x) + g(x)\mu(\{x\})],$$

which reduces to

$$\forall x \in \mathbb{R}, \quad g(x) F(x) \geq f(x) G(x).$$

For all $x \in \mathbb{R}$ such that $F(x)G(x) > 0$, one obtains $r_g(x) \geq r_f(x)$ by dividing last inequality by $F(x)G(x)$.

[rh order, $\Longleftarrow$]

[hr order] The proof is similar by reversing measures. □

CHARACTERIZATIONS OF lr ORDER. The definition of likelihood ratio order involves a third party dominating measure μ. On the contrary, next characterizations are intrinsic as they do not require external measures. First three characterizations are sometimes used as alternative definitions of lr orders. Last characterization is original.

Proposition B.38. Let F, G be two finite nonnegative measures on $\mathbb{R}$. The following propositions are equivalent to $F \underset{\text{lr}}{\leqslant} G$.

(i) $\begin{vmatrix} F(U) & G(U) \\ F(V) & G(V) \end{vmatrix} \geq 0$ for all Borel sets $U, V \subset \mathbb{R}$ such that $U \leq V$, *i.e.*, $\forall (u,v) \in U \times V, u \leq v$.

(ii) $\begin{vmatrix} F(U) & G(U) \\ F(V) & G(V) \end{vmatrix} \geq 0$ for all sets $U = (a,b), V = (c,d)$ such that $-\infty \leq a \leq c \leq \infty$ and $-\infty \leq b \leq d \leq \infty$.

(iii) $\begin{vmatrix} F(U) & G(U) \\ F(V) & G(V) \end{vmatrix} \geq 0$ for all sets $U = [a, b], V = [c, d]$ such that $a \leq c$ and $b \leq d$.

(iv) $F_{|\mathcal{G}}$ is absolutely continuous with respect to G and admits a density $\frac{\mathrm{d}F_{|\mathcal{G}}}{\mathrm{d}G}$ that is non-increasing, where $\mathcal{G} \stackrel{\text{def}}{=} [\inf \operatorname{supp}[G], +\infty)$ and $F_{|\mathcal{G}} \stackrel{\text{def}}{=} F(\cdot \cap \mathcal{G})$.

We provide a proof that makes no regularity assumptions, contrary to the proofs we have found in the literature.

*Proof. [*lr *order* $\implies$ *(i)]* Define $\mu = (F + G)/2$. Then μ is a Radon measure that dominates F, G. Indeed, for any set $A \subset \mathbb{R}$, $2\mu(A) \geq \max[F(A), G(A)]$ so $\mu(A) = 0$ implies $F(A) = G(A) = 0$. By the Radon-Nikodym theorem, this is equivalent to $\mu \gg F, G$. So by hypothesis, F, G admits densities f, g such that $f \underset{\text{TP}}{\leqslant} g$. So for all $u \leq v$, $g(v)f(u) - g(u)f(v) \geq 0$. Let U, V be two Borel sets such that $U \leqslant V$. Integrating this relationship gives

$$\iint_{u \in U, v \in V} [g(v)f(u) - g(u)f(v)]\mu(\mathrm{d}u)\mu(\mathrm{d}v) \geq 0,$$

which reduces to

$$G(V)f(U) \geq G(U)f(V).$$

[(i) $\implies$ *(ii)]* If $b \leq c$, $U \leq V$ and the proof is over.

Else, $U = (a, c] \cup (c, b)$ and $V = (c, b) \cup [b, d)$. The bilinearity of determinant gives

$$\begin{aligned}
\begin{vmatrix} F(U) & G(U) \\ F(V) & G(V) \end{vmatrix} &= \begin{vmatrix} F(a,c] + F(c,b) & G(a,c] + G(c,b) \\ F(c,d) & G(c,d) \end{vmatrix} \\
&= \begin{vmatrix} F(a,c] & G(a,c] \\ F(c,d) & G(c,d) \end{vmatrix} + \begin{vmatrix} F(c,b) & G(c,b) \\ F(c,b) + F[b,d) & G(c,b) + G[b,d) \end{vmatrix} \\
&= \begin{vmatrix} F(a,c] & G(a,c] \\ F(c,d) & G(c,d) \end{vmatrix} + 0 + \begin{vmatrix} F(c,b) & G(c,b) \\ F[b,d) & G[b,d) \end{vmatrix}.
\end{aligned}$$

As $(a, c] \leq (c, d)$ and $(c, b) \leq [b, d)$, the two right-hand side terms are nonnegative by assumption (i). Therefore, so does the left-hand side term.

[(ii) $\implies$ *(iii)]* Define $U = [a, b]$, $V = [c, d]$, $U_h = (a - h, b + h)$ and $V_h = (c - h, d + h)$. Assuming (i) implies $\begin{vmatrix} F(U_h) & G(U_h) \\ F(V_h) & G(V_h) \end{vmatrix} \geq 0$, for all h. Furthermore, $\bigcap_{h>0} U_h = U$ and $\bigcap_{h>0} V_h = V$. For any sigma-finite measure μ, $\lim_{h \to 0^+} \mu(U_h) = \mu(U)$ and $\lim_{h \to 0^+} \mu(V_h) = \mu(V)$. So taking limit in the inequality gives the result:

$$\begin{vmatrix} F(U) & G(U) \\ F(V) & G(V) \end{vmatrix} \geq 0.$$

[(iii) $\implies$ lr *order]* Let μ be some Radon measure such that $\mu \gg F, G$. Note that $\operatorname{supp}[F] \cup \operatorname{supp}[G] \subset \operatorname{supp}[\mu]$ For any $x, x' \in \operatorname{supp}[\mu]$ such that $x < x'$, claim (iii) gives

$$\begin{aligned}
&\forall h > 0, \quad F[x - h, x + h]G[x' - h, x' + h] \geq G[x - h, x + h]F[x' - h, x' + h], \\
&\forall h > 0, \quad \frac{F[x - h, x + h]}{\mu[x - h, x + h]} \frac{G[x' - h, x' + h]}{\mu[x' - h, x' + h]} \geq \frac{G[x - h, x + h]}{\mu[x - h, x + h]} \frac{F[x' - h, x' + h]}{\mu[x' - h, x' + h]}.
\end{aligned}$$

By taking the limit $h \to 0$, Besicovitch's theorem gives the existence of a set $M \subset \text{supp}[\mu]$ such that $F(M) = G(M) = 0$ and for all $x, x' \notin M$, $x < x'$ implies

$$f(x)g(x') \geq g(x)f(x'),$$

where f and g are densities for F, G. Redefining f, g by $f(x) = g(x) := 0$ for all $x \in M$ provides densities such that $f \underset{\text{TP}}{\leqslant} g$.

[(i) $\implies$ (iv)] Step 1. Define $m \overset{\text{def}}{=} \inf \mathcal{G}$. Then, $G((-\infty, x]) > 0$ for all $x > m$. Let U be a measurable subset of $\mathbb{R}$ such that $G(U) = 0$.

For all $\epsilon > 0$, $G(U \cap (m + \epsilon, \infty)) \leq G(U) = 0$ and $G((-\infty, m + \epsilon]) > 0$. Assumption (i) gives

$$G(U \cap [m + \epsilon, \infty))F((-\infty, m + \epsilon]) \geq G((-\infty, m + \epsilon])F(U \cap [m + \epsilon, \infty)) \geq 0.$$

As $G(U \cap [m+\epsilon, \infty)) = 0$ and $G((-\infty, m+\epsilon]) \neq 0$, this implies $F(U \cap [m+\epsilon, \infty)) = 0$. As the measure F is sigma-additive, $\lim_{\epsilon \to 0} F(U \cap [m + \epsilon, \infty)) = F(U \cap [m, \infty)) = F_{|\mathcal{G}}(U)$ as $\mathcal{G} = [m, \infty)$. This shows that $F_{|\mathcal{G}}(U) = 0$.

As a consequence, $F_{|\mathcal{G}}$ is absolutely continuous with respect to G.

Step 2. Now, let us prove that the Radon-Nikodym derivative $\frac{\mathrm{d}F_{|\mathcal{G}}}{\mathrm{d}G}$ is non-increasing G-almost everywhere. The proof relies on Besicovitch's differentiation theorem (Theorem B.37). For all $x \in \text{supp}[G], h > 0$, $G[x - h, x + h] > 0$ by definition of the support, so $R_h(x) := \frac{F[x-h,x+h]}{G[x-h,x+h]}$ is well-defined.

Let us show that (ii) implies $R_h(x)$ is non-increasing on $x \in \text{supp}[G]$. For any $U = [x - h, x + h]$, $V = [x' - h, x' + h]$ such that $x, x' \in \text{supp}[G]$ and $x \leq x'$, assuming (ii) implies

$$F[x - h, x + h]G[x' - h, x' + h] \geq F[x' - h, x' + h]G[x - h, x + h],$$

which reads

$$\frac{F[x - h, x + h]}{G[x - h, x + h]} \geq \frac{F[x' - h, x' + h]}{G[x' - h, x' + h]}. \tag{$\star$}$$

For all $x \in \text{supp}[G] \setminus \{m\}$, for small enough $h > 0$, $[x - h, x + h] \subset \mathcal{G}$ and therefore $R_h(x) = \frac{F_{|\mathcal{G}}[x-h,x+h]}{G[x-h,x+h]}$. Besicovitch's theorem implies that

$$\lim_{h \to 0^+} \frac{F_{|\mathcal{G}}[x - h, x + h]}{G[x - h, x + h]} = \frac{\mathrm{d}F_{|\mathcal{G}}}{\mathrm{d}G}(x) = \lim_{h \to 0^+} R_h(x),$$

for G-almost every $x \in \mathbb{R} \setminus \{m\}$. Combining this limit with monotony of R_h, the Radon-Nikodym derivative $\frac{\mathrm{d}F_{|\mathcal{G}}}{\mathrm{d}G}$ is non-increasing G-almost everywhere on $\mathbb{R} \setminus \{m\}$.

To finish the proof, we have to care about $\{m\}$. To do we, we distinguish two cases. If $G(\{m\}) = 0$, then the proof is over. Else, $m \in \text{supp}[G]$ and $\lim_{h \to 0^+} G[m - h, m + h] = G(\{m\}) > 0$. Besides, Besicovitch's theorem gives $\lim_{h \to 0^+} \frac{F_{|\mathcal{G}}[m-h,m+h]}{G[m-h,m+h]} = \frac{\mathrm{d}F_{|\mathcal{G}}}{\mathrm{d}G}(m)$. As the measure F is sigma-additive, $\lim_{h \to 0^+} F(x - h, x) = F(\emptyset) = 0$. So $\lim_{h \to 0^+} \frac{F[m-h,m)}{G[m-h,m+h]} = 0$.

Furthermore, by definition of $\mathcal{G}$,

$$\begin{aligned}\frac{F[m - h, m + h]}{G[m - h, m + h]} &= \frac{F[m - h, m)}{G[m - h, m + h]} + \frac{F_{|\mathcal{G}}[m, m + h)}{G[m - h, m + h]} \\ &= \frac{F[m - h, m)}{G[m - h, m + h]} + \frac{F_{|\mathcal{G}}[m - h, m + h]}{G[m - h, m + h]},\end{aligned}$$

so the inequality ($\star$) with $x = m$ reads.

$$\frac{F_{|\mathcal{G}}[m-h, m+h]}{G[m-h, m+h]} \geq R_h(x') - \frac{F[m-h, m)}{G[m-h, m+h]}.$$

Taking the limit $h \to 0$ in the inequality above gives $\frac{\mathrm{d}F_{|\mathcal{G}}}{\mathrm{d}G}(m) \geq \frac{\mathrm{d}F_{|\mathcal{G}}}{\mathrm{d}G}(x')$ for G-almost all $x' > m$. This ends the proof that $\frac{\mathrm{d}F_{|\mathcal{G}}}{\mathrm{d}G}$ is non-increasing G-almost everywhere.

Step 3. There exists an extending function $g : \mathbb{R} \to \mathbb{R}_+$ such that g is non-increasing and $\frac{\mathrm{d}F_{|\mathcal{G}}}{\mathrm{d}G}(x) = g(x)$ for G-almost every x. This means that g is a valid density. As it is non-increasing, the proof is done.

[(iv) $\implies$ lr *order]* Assuming (iv) provides existence a non-increasing function d such that $\frac{\mathrm{d}F_{|\mathcal{G}}}{\mathrm{d}G}(x) = d(x)$. Let μ be a measure such as $\mu \gg F, G$. Let $\tilde{f}$, $\tilde{g}$ denote respective densities of F and G with respect to μ. Define

$$f(x) \stackrel{\text{def}}{=} \begin{cases} \tilde{g}(x)d(x) & \text{if } x \in \mathcal{G}, \\ \tilde{f}(x) & \text{else,} \end{cases} \quad \text{and} \quad g(x) \stackrel{\text{def}}{=} \begin{cases} \tilde{g}(x) & \text{if } x \in \mathcal{G}, \\ 0 & \text{else.} \end{cases}$$

Then, by chain rule, f, g is a valid density for F, G with respect to μ. For all $x, y \in \mathbb{R}$ such that $x < y$,

$$f(x)g(y) - f(y)g(x) = \begin{cases} 0 & \text{if } x < y < m, \\ f(x)g(y) & \text{if } x < m \leq y, \\ \tilde{g}(x)\tilde{g}(y)(d(x) - d(y)) & \text{if } m \leq x < y, \end{cases}$$

so in all cases one has $f(x)g(y) - f(y)g(x) \geq 0$. This shows that $f \underset{\text{TP}}{\leqslant} g$ and proves the *lr* order. □

B.3.2 PRESERVATION OF STOCHASTIC ORDERS

PRESERVATION BY WEAK CONVERGENCE. The convergence in distribution preserves every stochastic order. This notion is more general than pointwise limits of functions or sequences.

Proposition B.39. Let C denote one of the stochastic orders { st, lr, hr, rh }.

If two sequences of finite positive measures F_n, G_n converge in distribution to measures F, G, then

$$\forall n \in \mathbb{N}, \quad F_n \underset{C}{\leqslant} G_n \qquad \Longrightarrow \qquad F \underset{C}{\leqslant} G.$$

Proofs available in the literature are usually given under the implicit assumption of *strong convergence*: $\lim_n F_n(A) = F(A)$ for all Borel set A. Convergence in distribution means this equality holds for Borel sets A that are continuity sets of F. This is strictly weaker than strong convergence if F is not absolutely continuous.

Proof. [st order] Our proof relies on characterization (ii) in Lemma B.44 of the *st* order. Let γ be a bounded, continuous, function that is non-decreasing on $\mathbb{R}$. Then, the

Portmanteau lemma gives $\lim_{n\to\infty}\int\gamma\mathrm{d}F_n=\int\gamma\mathrm{d}F$ and $\lim_{n\to\infty}\int\gamma\mathrm{d}G_n=\int\gamma\mathrm{d}G$. As $F_n \underset{C}{\leqslant} G_n$, $\int\gamma\mathrm{d}F_n\leq\int\gamma\mathrm{d}G_n$ and taking limits gives $\int\gamma\mathrm{d}F\leq\int\gamma\mathrm{d}G$.

[lr order] Our proof is based on characterization (ii) in Proposition B.38 of *lr* order. Let a,b,c,d be real numbers such that $a<b$, $c<d$, $a\leq c$ and $b\leq d$. For all $n\in\mathbb{N}$, since $F_n \underset{\mathrm{lr}}{\leqslant} G_n$, one has

$$F_n(c+h,d-h)G_n(a-h,b-h)\leq G_n(c+h,d-h)F_n(a-h,b-h),$$

and by positivity of measures,

$$F_n(c+h,d-h)G_n(a-h,b-h)\leq G_n[c+h,d-h]F_n[a-h,b-h],$$

therefore

$$\liminf_n F_n(c+h,d-h)G_n(a-h,b-h)\leq\limsup_n G_n[c+h,d-h]F_n[a-h,b-h].$$

By the Portmanteau lemma, since $[x,y]$ is closed and (x,y) is open,

$$\limsup_{n\to\infty} G_n[c+h,d-h]F_n[a-h,b-h]\leq G[c+h,d-h]F[a-h,b-h],\liminf_{n\to\infty} F_n(c+h,d-h)G_n(a-h,b-h)$$

Therefore,

$$F(c+h,d-h)G(a-h,b-h)\leq G[c+h,d-h]F[a-h,b-h],$$

and this inequality is valid for all $h>0$. In addition, by sigma additivity of F

$$\lim_{h\to0^+}F(c+h,d-h)=F(c,d)=\lim_{h\to0^+}F[c+h,d-h],$$

and similarly for G. Taking limits in previous inequality, one obtains

$$F(c,d)G(a,b)\leq G(c,d)F(a,b),$$

which gives back characterization (ii) Proposition B.38.

[rh order] We have $\lim_{n\to\infty}F_n(x)=F(x)$ for all continuity point x of F, and similarly for G. As measures have at most countable discontinuities, there exists a dense subset I of $\mathbb{R}$ such that

$$\forall x\in I,\quad\lim_{n\to\infty}F_n(x)=F(x)\qquad\text{and}\quad\lim_{n\to\infty}G_n(x)=G(x).$$

Let x,y be in $\mathbb{R}$ such that $x<y$. By density, there exists two sequences $(x_k)_{k\in\mathbb{N}},(y_k)_{k\in\mathbb{N}}$ in I that converge to x,y and such that $x\leq x_k<y\leq y_k$. By assumption $F_n \underset{\mathrm{rh}}{\leqslant} G_n$,

$$\forall k,n\in\mathbb{N},\quad F_n(x_k)G_n(y_k)\geq G_n(x_k)F_n(y_k).$$

As n goes to ∞, the weak convergence implies

$$\forall k\in\mathbb{N},\quad F(x_k)G(y_k)\geq G(x_k)F(y_k).$$

As cumulative distributions functions are right-continuous, $\lim_{k\to\infty}F(x_k)=F(x)$ and similarly for G. As k goes to ∞, one obtains the *rh* order:

$$F(x)G(y)\geq G(x)F(y).$$

[hr order] The proof is similar by reversing measures. □

PRESERVATION BY CONVOLUTION. Log-concave functions and sequences are intimately related to TP ordering: their convolution preserves TP ordering. This result plays a key role in the theory of total positivity of order 2.

Proposition B.40 (Preservation of TP order)**.** Let f, g, h be three non-negative functions on $\mathbb{R}$ (resp. sequences on $\mathbb{Z}$).
If f is (resp. discrete) log-concave, then

$$g \overset{\text{TP}}{\leqslant} h \qquad \Longrightarrow \qquad f * g \overset{\text{TP}}{\leqslant} f * h.$$

This proposition extends to measures and other stochastic orders. Conceptually, each stochastic order lr, hr, rh is the counterpart of a TP_2 distributional property: LCAV, IHR, DRHR. This parallelism has been revealed by Proposition B.26. Next proposition reinforces it by showing shows each TP_2 property preserves its corresponding stochastic order. The three claims can be found as Theorem 1.C.9, Lemma 1.B.3. and Lemma 1.B.44 of (Shaked and Shanthikumar, 2007).

Proposition B.41. Let F, G, H be three probability measure on $\mathbb{R}$.
If F is log-concave (or, discrete log-concave and G, H are both discrete) then

$$G \underset{\text{lr}}{\leqslant} H \qquad \Longrightarrow \qquad F * G \underset{\text{lr}}{\leqslant} F * H.$$

If F is IHR (or, discrete IHR and G, H are both discrete) then

$$G \underset{\text{hr}}{\leqslant} H \qquad \Longrightarrow \qquad F * G \underset{\text{hr}}{\leqslant} F * H.$$

If F is DRHR (or, discrete DRHR and G, H are both discrete) then

$$G \underset{\text{rh}}{\leqslant} H \qquad \Longrightarrow \qquad F * G \underset{\text{rh}}{\leqslant} F * H.$$

Usual proofs available in the literature make regularity assumptions on measures. For the lr order, the proof given by Shaked and Shanthikumar (2007, Theorem 1.C.9) assumes that all distributions are absolutely continuous. The proof for discrete distributions is also given in may textbooks. However, we have not found any explicit proof in the case of mixed distributions. For the hr and rh results, the proofs are given by Lynch et al. (1987, Corollary 2.3) also assume absolute continuity, Shanthikumar (1988, Lemma 2.1) is valid for discrete measures. Therefore, we provide a proof that rules out such assumptions.

Proof. This result is a corollary of forthcoming Proposition B.42. Indeed, convolution may be expressed as mixture: for all Borel set A,

$$F * G(A) = \int_{\mathbb{R}} G(A - \theta) \mathrm{d}F(\theta) = \int_{\Theta} G_\theta(A) \mathrm{d}F(\theta),$$

with $\Theta := \mathbb{R}$ and shifted measures $G_\theta(\cdot) := G(\cdot - \theta)$ for all $\theta \in \Theta$.

[lr order] Assume G is log-concave. Then, Proposition B.26 proves that $(G_\theta)_{\theta \in \Theta} \uparrow lr$. So Proposition B.42 (i) gives the result.

[hr order] Assume G is IHR. Then, $(G_\theta)_{\theta\in\Theta} \uparrow hr$. In addition, as a cdf is right-continuous, $\theta \mapsto G(x-\theta)$ is a left-continuous function for all $x \in \mathbb{R}$. So Proposition B.42 (ii) can be applied and gives the result.

[rh order] Assume G is DRHR. Then, $(G_\theta)_{\theta\in\Theta} \uparrow rh$. In addition, as a cdf is right-continuous, $\theta \mapsto G(x-\theta)$ is a left-continuous function for all $x \in \mathbb{R}$. So Proposition B.42 (iii) gives the result. □

PRESERVATION BY COMPOSITION. The following proposition shows that stochastic orders are preserved by mixtures if both mixing and compounding distributions bear the same ordering. As this result is about measures, it generalizes the basic composition formula and other preservation results for TP_2 functions and sequences that are stated in section B.1.

Proposition B.42. Consider a family of measures $(G_\theta)_{\theta\in\Theta}$ indexed on a measurable set $\Theta \subset \mathbb{R}$. Let F_1 and F_2 be two measures supported on Θ. Define the mixed measures H_i for $i = 1, 2$ by

$$H_i \overset{\text{def}}{=} \int_\Theta G_\theta \mathrm{d}F_i(\theta).$$

(i) If $F_1 \underset{\text{lr}}{\leqslant} F_2$ and $G_\theta \underset{\text{lr}}{\leqslant} G_{\theta'}$ whenever $\theta \leqslant \theta'$, then $H_1 \underset{\text{lr}}{\leqslant} H_2$.

Assume[1] in addition that for all $x \in \mathbb{R}$, $\theta \mapsto G_\theta(x)$ is a right-continuous (or left-continuous) function of Θ.

(ii) If $F_1 \underset{\text{hr}}{\leqslant} F_2$ and $G_\theta \underset{\text{hr}}{\leqslant} G_{\theta'}$ whenever $\theta \leqslant \theta'$, then $H_1 \underset{\text{hr}}{\leqslant} H_2$.

(iii) If $F_1 \underset{\text{rh}}{\leqslant} F_2$ and $G_\theta \underset{\text{rh}}{\leqslant} G_{\theta'}$ whenever $\theta \leqslant \theta'$, then $H_1 \underset{\text{rh}}{\leqslant} H_2$.

For the *lr* order, this result is proved by Shaked and Shanthikumar (2007, Theorem 1.C.17) under the assumption that G_θ and F_i are absolutely continuous, and it is proved by Keilson and Sumita (1982, Theorem 4.8) if G_θ and F_i are discrete. For the *hr* and *rh* orders, proofs given by Lynch et al. (1987) assume all measures are absolutely continuous and $G_\theta(x)$ are absolutely continuous functions with respect to θ. We extend such proofs to the general case. This is not straightforward for *hr* and *rh* orders. We begin with a general lemma about convergence of mixtures.

Lemma B.43 (convergence of mixtures)**.** Let Θ be a Borel-measurable subset of $\mathbb{R}$, and F be a probability measure supported on Θ. Let $(G^n)_{n\in\mathbb{N}}$ be a sequence of families $G^n := (G^n_\theta)_{\theta\in\Theta}$ of probability measures supported on $\mathbb{R}$. Assume that the mixed probability measures H_n exist for all $n \in \mathbb{N}$, where

$$H_n \overset{\text{def}}{=} \int_\Theta G^n_\theta \mathrm{d}F(\theta).$$

If for all $\theta \in \Theta$, G^n_θ converge in distribution to some probability measure G_θ, then H_n converges in distribution to the probability measure H given by

$$H \overset{\text{def}}{=} \int_\Theta G_\theta \mathrm{d}F(\theta).$$

1. Our smoothness assumption is very minimalistic since in practice, right-continuity is always fulfilled by usual continuous-time processes such as càdlàg ones, and unconditionally fulfilled by discrete-time processes. Still, we wonder if such assumption could be removed.

Remark. For the mixture to be well-defined, at least we have to assume that $\theta \mapsto G_\theta^n(x)$ is a Borel-measurable function for all $x \in \mathbb{R}$. Then, H_n exists and its expression is defined by Lebesgue-Stieltjes integration. In addition, $\theta \mapsto G_\theta(x)$ is automatically a Borel-measurable function as a pointwise limit of such functions, so H is defined as well.

Proof. The tricky part is about continuity points: for some θ and x values on might have $\lim_n G_\theta^n(x) \neq F_\theta(x)$. So it is not possible to prove it directly using the dominated convergence theorem. But here, H_n and H are probability measures. So we can base our proof on the Portmanteau lemma: a sequence $(M_n)_n$ of probability measures converges in distribution to a probability measure M if and only if $\liminf_n M_n(U) \geq M(U)$ for all open sets U of $\mathbb{R}$.

Let U be such an open set. Portmanteau lemma gives:

$$\forall \theta \in \Theta, \quad \liminf_{n\to\infty} G_\theta^n(U) \geq G_\theta(U).$$

By assumption, $\theta \mapsto G_\theta^n(U)$ are measurable functions. Therefore, $\theta \mapsto \liminf_n G_\theta^n(U)$ is measurable and its Lebesgue-Stieltjes integral exists. By positivity of integration,

$$\int_\Theta \liminf_{n\to\infty} G_\theta^n(U) \mathrm{d}F(\theta) \geq \int_\Theta G_\theta(U) \mathrm{d}F(\theta) = H(U).$$

Moreover, by Fatou's lemma,

$$\liminf_{n\to\infty} \int_\Theta G_\theta^n(U) \mathrm{d}F(\theta) \geq \int_\Theta \liminf_{n\to\infty} G_\theta^n(U) \mathrm{d}F(\theta).$$

Combining the two inequalities gives

$$\liminf_{\sigma\to 0} H_n(U) \geq H(U).$$

As this holds for any open set U, Portmanteau lemma tells that $(H_n)_n$ converges in distribution to H. □

Proof of Proposition B.42. Remark. As *lr*, *hr* and *rh* orders imply the *st* order, for all $x \in, \theta \mapsto G_\theta(x)$ is non-increasing. As monotone functions are Borel-measurable, this is enough to guarantee the existence of the mixture.

[lr order] Assume that $F_1 \underset{\text{lr}}{\leqslant} F_2$. Define $\mu := F_1 + F_2$. As already proved, $\mu \gg F_1, F_2$. Therefore, F_i admits densities f_i such that $f_1 \underset{\text{TP}}{\leqslant} f_2$. Let A_1, A_2 be two Borelian subsets of $\mathbb{R}$ such that $A_1 \leq A_2$. For $i, j \in \{1, 2\}$, Then $H_j(A_i) = \int G_y(A) \mathrm{d}F_j(y) = \int G_y(A_i) f_j(y) \mathrm{d}\mu(y)$. So the basic composition formula gives

$$\begin{vmatrix} H_1(A_1) & H_1(A_2) \\ H_2(A_1) & H_2(A_2) \end{vmatrix} = \iint_{s<t} \begin{vmatrix} G_s(A_1) & G_s(A_2) \\ G_t(A_1) & G_t(A_2) \end{vmatrix} \begin{vmatrix} f_1(s) & f_1(t) \\ f_2(s) & f_2(t) \end{vmatrix} \mathrm{d}\mu(s) \mathrm{d}\mu(t).$$

As $f_1 \underset{\text{TP}}{\leqslant} f_2$, the third determinant is nonnegative. By assumption $G_s \underset{\text{lr}}{\leqslant} G_t$ for all $s < t$, the second determinant is nonnegative. As the measure μ is nonnegative, the left-hand side is nonnegative, which proves *lr* order.

[hr order] We only deal with *hr* order since *rh* is similar.

Case 1. The result has been proved by Lynch et al. (1987) under three assumptions:

(i) measures G_θ are absolutely continuous,
(ii) measures F_i are absolutely continuous,
(iii) Θ is an interval and $\theta \mapsto G_\theta(x)$ is an absolutely continuous function for all $x \in \mathbb{R}$.

Step 2. Now, we forget assumption (i). The idea is to approximate G_θ by smooth measures. Consider for all $\sigma > 0$ and $\theta \in \Theta$ the measure

$$G_\theta^\sigma := G_\theta * \mathcal{N}(0, \sigma),$$

where $\mathcal{N}(0, \sigma)$ denotes the Gaussian measure of mean 0 and standard deviation σ.

Then, G_θ^σ are absolutely continuous. Consider the corresponding mixtures $H_i^\sigma := \int_\Theta G_\theta^\sigma \mathrm{d}F_i(\theta)$, $i = 1, 2$. As measures $\mathcal{N}(0, \sigma)$ are log-concave, they preserve stochastic orders (see proposition B.41): so $G_\theta^\sigma \underset{\text{hr}}{\leqslant} G_{\theta'}^\sigma$ whenever $\theta \leqslant \theta'$. So one can apply the previous case to obtain $H_1^\sigma \underset{\text{hr}}{\leqslant} H_2^\sigma$.

In addition, $G_\theta^\sigma \underset{\sigma \to 0}{\to} G_\theta$ in distribution. So Lemma B.43 proves that $H_i^\sigma \underset{\sigma \to 0}{\to} H_i$ in distribution. As stochastic orderings are preserved by convergence in distribution, this implies $H_1 \underset{\text{hr}}{\leqslant} H_2$.

Case 3. Now, we forget assumption (ii). The idea is to smooth F_i. Consider for all $\sigma > 0$ and $i = 1, 2$ the probability measures

$$F_i^\sigma := F_i * \mathcal{N}(0, \sigma).$$

Then F_i^σ are absolutely continuous and $F_i^\sigma \underset{\sigma \to 0}{\to} F_i$ in distribution.

To define mixtures with F_i^σ, one has to extend the family $(G_\theta)_{\theta \in \Theta}$ to the whole real line: $\Theta \leftarrow \mathbb{R}$. Define for all $\theta \in \mathbb{R}$, the distribution function G_θ as

$$G_\theta(x) \overset{\text{def}}{=} \begin{cases} \sup_{\theta' \in \Theta, \theta' \geq \theta} G_{\theta'}(x) & \text{if } \theta < \sup \Theta, \\ \inf_{\theta' \in \Theta} G_{\theta'}(x) & \text{else.} \end{cases} \tag{B.1}$$

The *lr / hr / rh* orders all imply that $\theta \mapsto G_\theta(x)$ is non-increasing. So G_θ coincides with the original one for all $\theta \in \Theta$. In addition, $G = (G_\theta)_{\theta \in \mathbb{R}}$ defines a family of measures that has the same stochastic monotonies ($\uparrow$ *hr* in our case). And as F_i are supported on Θ one still has

$$H_i(x) = \int_{\mathbb{R}} G_\theta(x) \mathrm{d}F_i(\theta) = \int_\Theta G_\theta(x) \mathrm{d}F_i(\theta).$$

Consider the smoothed mixtures

$$\forall x \in \mathbb{R}, \quad H_i^\sigma(x) := \int_{\mathbb{R}} G_\theta(x) \mathrm{d}F_i^\sigma.$$

Since F_i^σ are absolutely continuous, the previous case can be applied and

$$\forall \sigma > 0, \quad H_1^\sigma \underset{\text{hr}}{\leqslant} H_2^\sigma.$$

By assumption, (iii) $\theta \mapsto G_\theta(x)$ are continuous functions on $\mathbb{R}$. Since F_i^σ converges in distribution to F_i, the Portmanteau lemma gives

$$\forall x \in \mathbb{R}, \quad \lim_{\sigma \to 0} H_i^\sigma(x) = H_i(x),$$

which in turns gives the convergence of H_i^σ to H_i in distribution. As stochastic orderings are preserved under convergence in distribution, we conclude

$$H_1 \underset{\text{hr}}{\leqslant} H_2.$$

Case 4. Now, we forget assumption (iii) partially. Θ can any subset of $\mathbb{R}$. Assume $\theta \mapsto G_\theta(x)$ are left-continuous for all $x \in \mathbb{R}$ (as functions defined on Θ). The idea is to smooth the family G. To do so, we work with the extended family $(G_\theta)_{\theta\in\mathbb{R}}$ defined by equation (B.1). One can show that the extended family is still left-continuous. Let us smooth the extended family G by defining for $\theta, x \in \mathbb{R}$, $h > 0$,

$$G_\theta^h(x) \overset{\text{def}}{=} \frac{1}{h}\int_0^h G_{\theta-u}(x)\mathrm{d}u. \tag{B.2}$$

First, for all $h > 0$, this provides a family $G^h = (G_\theta^h)_{\theta\in\mathbb{R}}$ such that for all $x \in \mathbb{R}$, the function $\theta \mapsto G_\theta^h(x)$ is continuously differentiable (therefore absolutely continuous).

Second, mixtures G_θ^h have G^h has the same stochastic monotony as G. Indeed, the mixing measures are translated uniform distributions $U_\theta \sim \mathcal{U}(\theta, \theta + h)$, which are absolutely continuous measures. Since $\mathcal{U}(0, h)$ is a log-concave measure, And the family $(U_\theta)_{\theta\in\mathbb{R}}$ is non-increasing in the likelihood ratio order. In addition, $\theta \mapsto G_\theta^h(x)$ is an absolutely continuous function on $\mathbb{R}$. Consequently, we can apply the first case on G^h and the corresponding mixtures $H_i^h := \int_{\mathbb{R}} G_\theta^h \mathrm{d}F(\theta)$ to obtain

$$\forall h > 0, \qquad H_1^h \underset{\text{hr}}{\leqslant} H_2^h.$$

Third, for all $x \in \mathbb{R}$, $\lim_{h\to 0^+} G_\theta^h = G_{\theta^-}$ in distribution. Since we have assumed the left-continuity of G_θ in θ, this implies convergence in distribution $\lim_{h\to 0^+} G_\theta^h = G_\theta$. So Lemma B.43 gives the convergence of distribution $\lim_{h\to 0^+} H_i^h = H_i$. As such convergence preserves stochastic orders,

$$H_1 \underset{\text{hr}}{\leqslant} H_2.$$

Case 4.bis Assume that $\theta \mapsto G_\theta(x)$ is right-continuous instead of left-continuous. The same proof outline can be used with two modifications. First, extend the family $(G_\theta)_{\theta\in\Theta}$ on $\theta \in \mathbb{R}$ as follows instead of equation (B.1),

$$G_\theta(x) \overset{\text{def}}{=} \begin{cases} \inf_{\theta'\in\Theta, \theta'\le\theta} G_{\theta'}(x) & \text{if } \theta > \inf\Theta, \\ \sup_{\theta'\in\Theta} G_{\theta'}(x) & \text{else.} \end{cases} \tag{B.3}$$

Second, smooth the extended family G_θ as follows instead of equation (B.2),

$$G_\theta^h \overset{\text{def}}{=} \frac{1}{h}\int_0^h G_{\theta+u}\mathrm{d}u.$$

□

B.3.3 MOMENT INEQUALITY FORMULAS

Stochastic orders provides inequalities involving moments, that is to say quantities of the type $\mathbb{E}[f(X)] = \int f(x)\mathrm{d}P_X(x)$. Proposition B.47 provides such inequalities. Stating and proving such result in its full generality is the main purpose of this section.

PRELIMINARIES. Our idea is to consider weighted distributions. To do so, a couple of results are required. Next lemma is a well-known characterization of the basic stochastic order.

Lemma B.44 (Lehmann and Romano 2005, Lemma 3.4.2)**.** Let F_1, F_2 be two finite measures on $\mathbb{R}$. The following assertions are equivalent.

(i) $F_1 \underset{\text{st}}{\leqslant} F_2$ and $F_1(\mathbb{R}) = F_2(\mathbb{R})$.

(ii) $\int_{\mathbb{R}} \gamma(x)\mathrm{d}F_1(x) \leq \int_{\mathbb{R}} \gamma(x)\mathrm{d}F_2(x)$ for all function $\gamma : \mathbb{R} \mapsto \mathbb{R}$ that is bounded, continuous and non-decreasing on $\mathbb{R}$.

(iii) $\int_{\mathbb{R}} \gamma(x)\mathrm{d}F_1(x) \leq \int_{\mathbb{R}} \gamma(x)\mathrm{d}F_2(x)$ for all function $\gamma : \mathbb{R} \mapsto [-\infty, \infty]$ that is non-decreasing F_1- and F_2-almost everywhere and such that the integrals are absolutely convergent.

In other words, let X_1, X_2 be two real random variables. If $X_1 \underset{\text{st}}{\leqslant} X_2$, then $\mathbb{E}[\gamma(X_1)] \leq \mathbb{E}[\gamma(X_2)]$ for all non-decreasing function γ such that $\mathbb{E}\left[|\gamma(X_1)|\right], \mathbb{E}\left[|\gamma(X_2)|\right] < \infty$.

The most common proof relies on an approximation of γ with piece-wise constant functions. We give a different proof that relies on Lebesgue-Stieltjes integration by parts, and on approximation with continuous functions – we state such standard approximation in the following lemma without giving a proof.

Lemma B.45. Let F_1, F_2 be two measures on $I \subset \mathbb{R}$. Let $f : J \subset \mathbb{R} \to \mathbb{R}$ be a function such that

- f is integrable with respect to F_1 and F_2,
- f is non-decreasing F_1-, F_2- almost everywhere.

Then, there exists a sequence of functions $(f_n)_{n\in\mathbb{N}}$ of functions $f_n : \mathbb{R} \to \mathbb{R}$ such that

- f_n is bounded, continuous, non-decreasing on $\mathbb{R}$,
- f_n pointwise converges to f on $\mathbb{R}$: $\forall x \in \mathbb{R}, \lim_{n\to\infty} f_n(x)$.abcdefghijklmnopqrstuvw
- f_n is integrable with respect to F_1 and F_2,
- for all $i \in \{1, 2\}$, $\lim_{n\to\infty} \int |f_n - f|\mathrm{d}F_i = 0$.

Proof of Lemma B.44. [(i) $\implies$ (ii)] Define $M := F_1(\mathbb{R}) = F_2(\mathbb{R})$. Assume that γ is bounded, continuous and non-decreasing. Stieltjes integration by parts gives for all $a, b \in \mathbb{R}$,

$$\int_{[a,b]} \gamma(x)\mathrm{d}F_i(x) = \gamma(a)\overline{F}_i(a) - \gamma(b)F_i(b) + \int_{[a,b]} \overline{F}_i(x)\mathrm{d}\gamma(x)$$

Since $\lim_{a\to-\infty} \overline{F}_i(a) = M$, $\lim_{b\to\infty} \overline{F}_i(b) = 0$, $\lim_{b\to\infty} \gamma(b) = \sup \gamma$ and $\lim_{a\to-\infty} \gamma(a) = \inf \gamma > -\infty$ and γ is F_i-integrable, then $\overline{F}_i$ is γ-integrable and

$$\int_{\mathbb{R}} \gamma(x)\mathrm{d}F_i(x) = M \inf \gamma + \int_{\mathbb{R}} \overline{F}_i(x)\mathrm{d}\gamma(x)$$

$F_1 \underset{\text{st}}{\leqslant} F_2$ and $\overline{F}_1(x) \leq \overline{F}_2(x)$. Since γ is non-decreasing, the positivity of integral gives $\int \overline{F}_1(x)\mathrm{d}\gamma(x) \geq \int \overline{F}_2(x)\mathrm{d}\gamma(x)$ and this proves the result for bounded functions γ: $\int_{\mathbb{R}} \gamma(x)\mathrm{d}F_1(x) \leq \int_{\mathbb{R}} \gamma(x)\mathrm{d}F_2(x)$.

[(ii) $\implies$ (iii)] There exists a Borel set $I \subset \mathbb{R}$ such that $F_i(I) = F_i(\mathbb{R})$, γ is non-increasing on I and $\int_{\mathbb{R}} \gamma \mathrm{d}F_i = \int_I \gamma \mathrm{d}F_i$.

Note that I cannot be empty (unless $F_1 = F_2 = 0$), so $M = \sup I$ exists in $\mathbb{R} \cup \{\infty\}$. Define $\tilde{\gamma} : \mathbb{R} \to \mathbb{R}$ by

$$\tilde{\gamma}(x) = \begin{cases} \inf_{\substack{y \in I \\ y \geq x}} \gamma(y) & \text{if } x < M \\ \sup_{y \in I} \gamma(y) & \text{if } x \geq M \end{cases}.$$

Then $\tilde{\gamma}$ is non-increasing on $\mathbb{R}$, so it is measurable. Besides, it coincides with γ on I, so it is integrable on I and $\int_I \gamma \mathrm{d}F_i = \int_I \tilde{\gamma} \mathrm{d}F_i = \int_{\mathbb{R}} \tilde{\gamma} \mathrm{d}F_i$.

According to Lemma B.45, there exists a sequence $(\gamma_n)_{n \in \mathbb{N}}$ of bounded, continuous, non-increasing functions such that $\lim_{n \to \infty} \int \gamma_n \mathrm{d}F_i = \int \gamma \mathrm{d}F_i$ for all $i \in \{1, 2\}$. Since each $|\gamma_n|$ is bounded, last argument applies for all n and passing to the limit gives the result.

[(iii) $\implies$ (i)] Let x be in $\mathbb{R}$. The function $\gamma_x = \mathbf{1}_{[x,\infty]}$ is non-decreasing and checks $\int \gamma_x \mathrm{d}F_i = \overline{F}_i(x)$, so (iii) implies $\overline{F}_1(x) \leq \overline{F}_2(x)$. This is the definition of $\overline{F}_1 \underset{\text{st}}{\leqslant} \overline{F}_2$. In addition, choosing $\gamma \equiv 1$ and $\gamma \equiv -1$ respectively gives $F_1(\mathbb{R}) \leq F_2(\mathbb{R})$, $-F_1(\mathbb{R}) \geq -F_1(\mathbb{R})$, which implies $F_1(\mathbb{R}) = F_2(\mathbb{R})$. □

Next lemma is a preservation result of stochastic orders between weighted distributions.

Lemma B.46 (Shaked and Shanthikumar 2007, Example 1.B.23)**.** Let F, G be two finite measures on $\mathbb{R}$. Let $w : \mathbb{R} \mapsto \mathbb{R}_+$ be a nonnegative Borel-measurable function such that $\int_{\mathbb{R}} w\mathrm{d}F$ and $\int_{\mathbb{R}} w\mathrm{d}G$ exist and are non-vanishing. Define the *weighted measure* F^w as the measure whose cdf is

$$F^w(t) \overset{\text{def}}{=} \frac{\int_{(-\infty,t]} w(x)\mathrm{d}F(x)}{\int_{(-\infty,\infty)} w(x)\mathrm{d}F(x)}.$$

and define G^w similarly.

(i) If $F \underset{\text{lr}}{\leqslant} G$, then $F^w \underset{\text{lr}}{\leqslant} G^w$.

(ii) If $F \underset{\text{hr}}{\leqslant} G$ and w is non-decreasing, then $F^w \underset{\text{hr}}{\leqslant} G^w$.

(iii) If $F \underset{\text{rh}}{\leqslant} G$ and w is non-increasing, then $F^w \underset{\text{rh}}{\leqslant} G^w$.

Again, available proofs in the literature only deal with the case of absolutely continuous or discrete probability measures. We provide an original proof that does not require regularity assumptions.

Proof. [(i)] The *lr* order is straightforward. Indeed, the restriction of G on $I = \mathrm{supp}[G]$ is dominated by F and admits a non-decreasing density $d = \frac{\mathrm{d}G}{\mathrm{d}F}$.

It is easy to see that $\mathrm{supp}[G^w] \subset I$. Then, the Radon-Nikodym theorem implies that and G^w is dominated by F^w on I and admits $d = \frac{\mathrm{d}G^w}{\mathrm{d}F^w}$ as a density. Indeed, for all measurable set $A \subset I$,

$$\int_A \mathrm{d}G^w(x) = \int_A w(x)\mathrm{d}G(x) = \int_A w(x)d(x)\mathrm{d}F(x) = \int_A d(x)\mathrm{d}F^w(x).$$

so d is a valid density. As it is non-decreasing on I, one has $F^w \underset{\mathrm{lr}}{\leqslant} G^w$.

[(ii)] Assume w is non-decreasing. The order $F^w \underset{\mathrm{hr}}{\leqslant} G^w$ is true if and only if for all $x \leq y$,

$$\frac{F^w[x,\infty)}{F^w[y,\infty)} \geq \frac{G^w[x,\infty)}{G^w[y,\infty)} \quad \text{or equivalently} \quad \frac{F^w[x,y)}{F^w[y,\infty)} \geq \frac{G^w[x,y)}{G^w[y,\infty)}.$$

Case 1. Assume that w is left-continuous. As it monotone, it has right limits. As $\overline{F}$ is left-continuous and has right limits, by Stieltjes integration by parts, the left side equals

$$\begin{aligned}\frac{\int_{[x,y)} w(x)\mathrm{d}F(u)}{\int_{[y,\infty)} w(x)\mathrm{d}F(u)} &= \frac{-w(y)\overline{F}(y) + w(x)\overline{F}(x) + \int_{(x,y]} \overline{F}(u^-)\mathrm{d}w(u)}{\int_y^\infty w(x)\mathrm{d}F(u)} \\ &= \frac{\overline{F}(y)}{\int_y^\infty w(x)\mathrm{d}F(u)}\left(-w(y) + w(x)\frac{\overline{F}(x)}{\overline{F}(y)} + \int_x^y \frac{\overline{F}(u^-)}{\overline{F}(y)}\mathrm{d}w(u)\right) \\ &= \frac{1}{\mathbb{E}[w(X) \mid X > y]}\left(-w(y) + w(x)\frac{\overline{F}(x)}{\overline{F}(y)} + \int_x^y \frac{\overline{F}(u^-)}{\overline{F}(y)}\mathrm{d}w(u)\right).\end{aligned}$$

As $X \underset{\mathrm{hr}}{\leqslant} Y$, the inequality $\frac{\overline{F}(v^-)}{\overline{F}(y)} \geq \frac{\overline{G}(v)}{\overline{G}(y)}$ holds for all $v \leq y$. As w is non-decreasing, integrating this relationship gives

$$\int_x^y \frac{\overline{F}(u^-)}{\overline{F}(y)}\mathrm{d}w(u) \geq \int_x^y \frac{\overline{G}(u^-)}{\overline{G}(y)}\mathrm{d}w(u).$$

As w is non-negative, one also have $w(x)\frac{\overline{F}(x)}{\overline{F}(y)} \geq w(x)\frac{\overline{G}(x)}{\overline{G}(y)}$. So the inequality is obtained for the terms inside the bracket.

As $X \underset{\mathrm{hr}}{\leqslant} Y$, then $[X \mid X > y] \underset{\mathrm{st}}{\leqslant} [Y \mid Y > y]$ (see Lemma 4.24). As w is non-decreasing, Lemma B.44 gives $\mathbb{E}[w(X) \mid X > y] \leq \mathbb{E}[w(Y) \mid X > y]$. Since the bracket is non-negative, one can combine the two inequalities to obtain the result.

Case 2. In the general case, thanks to Lemma B.45 one can approximate the monotone function w by a sequence $(w_n)_{n\in\mathbb{N}}$ of continuous functions such that for all Borel set $A \subset \mathbb{R}$,

$$F^{w_n}(A) = \int_A w_n(x)\mathrm{d}F(x) \underset{n\to\infty}{\longrightarrow} \int_A w(x)\mathrm{d}F(x) = F^w(A).$$

This shows that F^{w_n} converges in distribution to F^w, and similarly so does G^{w_n} with G^w. As $F^{w_n} \underset{\mathrm{hr}}{\leqslant} G^{w_n}$ and stochastic order is preserved through convergence in distribution, we conclude that $F^w \underset{\mathrm{hr}}{\leqslant} G^w$.

[(iii)] The proof is similar, and can be deduced from the previous one by reversing measures. □

INEQUALITY FORMULAS. Next proposition gives an inequality result but with different sets of hypotheses, which correspond to the three TP_2 stochastic orderings. The result on *lr* order is stated in (Bickel and Lehmann, 1975, Lemma 2). The *hr*, *rh* ones in (Joag-dev et al., 1995, Theorem 2) and the remark below. We give original proofs based on preservation of stochastic orderings by weighting. Our motivation is to remove the regularity assumptions made on measures in all available proofs in the literature including (Bickel and Lehmann, 1975; Joag-dev et al., 1995; Shaked and Shanthikumar, 2007).

Proposition B.47. Let G_1, G_2 be two finite measures on $\mathbb{R}$, and $\alpha, \beta : I \mapsto \mathbb{R}$ be two functions on $I = \mathrm{supp}[G_1] \cup \mathrm{supp}[G_2]$.

Assume

- $\int \beta(x)\mathrm{d}G_i(x) < \infty$, $\int \alpha(x)\mathrm{d}G_i(x) < \infty$ for $i = 1, 2$,
- $\beta \underset{\mathrm{DP}}{\leqslant} \alpha$,
- and one of the three following condition holds:
  - i. $G_1 \underset{\mathrm{lr}}{\leqslant} G_2$,
  - ii. $G_1 \underset{\mathrm{hr}}{\leqslant} G_2$ and β is non-decreasing on I,
  - iii. $G_1 \underset{\mathrm{rh}}{\leqslant} G_2$ and β is non-increasing on I.

Then,

$$\left(\int \alpha(x)\mathrm{d}G_1(x)\right)\left(\int \beta(x)\mathrm{d}G_2(x)\right) \leq \left(\int \alpha(x)\mathrm{d}G_2(x)\right)\left(\int \beta(x)\mathrm{d}G_1(x)\right).$$

Proof. Introducing the weighted distributions G_i^β such that $\mathrm{d}G_i^\beta(x) \overset{\mathrm{def}}{=} \beta(x)\mathrm{d}G_i(x)/(\int \beta \mathrm{d}G_i)$ for $i = 1, 2$ and $\gamma(x) \overset{\mathrm{def}}{=} \alpha(x)/\beta(x)$ for all $x \in I$ (with convention $\alpha(x)/0 = \mathrm{sign}[\alpha(x)]\infty$), the inequality is equivalent to

$$\int \gamma(x) g_1^\beta(x)\mathrm{d}x \leq \int \gamma(x) g_2^\beta(x)\mathrm{d}x. \tag{B.4}$$

[i.] Assume $G_1 \underset{\mathrm{lr}}{\leqslant} G_2$. Lemma B.46 gives $G_1^\beta \underset{\mathrm{lr}}{\leqslant} G_2^\beta$, which implies $G_1^\beta \underset{\mathrm{st}}{\leqslant} G_2^\beta$. As γ is non-decreasing on I, it is non-increasing G_1- and G_2-almost everywhere. So the characterization of the *st* order (see Lemma B.44) gives the inequality (B.4).

[ii.] Assume $G_1 \underset{\mathrm{hr}}{\leqslant} G_2$ and β is non-decreasing. Lemma B.46 gives $G_1^\beta \underset{\mathrm{hr}}{\leqslant} G_2^\beta$, which implies $G_1^\beta \underset{\mathrm{st}}{\leqslant} G_2^\beta$. So the same conclusion as above holds.

[iii.] Assume $G_1 \underset{\mathrm{rh}}{\leqslant} G_2$ and β is non-increasing. Lemma B.46 gives $G_1^\beta \underset{\mathrm{rh}}{\leqslant} G_2^\beta$, which implies $G_1^\beta \underset{\mathrm{st}}{\leqslant} G_2^\beta$. So the same conclusion as above holds. □

We end with a useful corollary.

Lemma B.48. Let L be an ordered set, α, β and $(d_l)_{l \in L}$ be nonnegative density functions on $I \subset \mathbb{R}_+$ or sequences on $I \subset \mathbb{N}$.

Assume that

- $\beta \underset{\text{DP}}{\leqslant} \alpha$,
- $d_l * \beta, d_l * \alpha$ are well defined and $[d_l * \beta](t) > 0$ for all $l \in L, t \in I$,
- and one of the following conditions is true:
  1. the family $(d_l)_{l\in L}$ is non-decreasing in the likelihood ratio ordering,
  2. the family $(d_l)_{l\in L}$ is non-decreasing in the reverse hazard rate ordering and β is non-decreasing on $\operatorname{supp}[d_l]$.

Then, for all $t \in I$, $l \mapsto \frac{d_l * \alpha}{d_l * \beta}(t)$ is non-increasing.

Proof. For convenience we prove only the discrete case $I = \mathbb{N}$, as the other one is similar. Let N be in $\mathbb{N}$, l_1, l_2 be in L and assume $l_1 < l_2$. Define $\mathbf{w} = (\beta(N), \beta(N-1), \dots, \beta(0))$ and $\mathbf{z} = (-\alpha(N), -\alpha(N-1), \dots, -\alpha(0))$. Then $\mathbf{w} \underset{\text{DP}}{\leqslant} \mathbf{z}$. Define also $\mathbf{g}_i = (d_{l_i}(0), d_{l_i}(1), \dots, d_{l_i}(N))$ for $i = 1, 2$. Condition 1. implies that $g_1 \underset{\text{lr}}{\leqslant} g_2$ so Proposition B.47 (i) applies. Moreover, condition 2. implies that β is non-decreasing and that $g_1 \underset{\text{rh}}{\leqslant} g_2$, since the cdfs of g_i and d_{l_i} coincide on $1 \dots N$. So proposition B.47 (ii) applies. In both cases one obtains

$$(\mathbf{g}_1 \cdot \mathbf{z})(\mathbf{g}_2 \cdot \mathbf{w}) \leq (\mathbf{g}_2 \cdot \mathbf{z})(\mathbf{g}_1 \cdot \mathbf{w}).$$

This reads

$$[d_{l_1} * \alpha](N)[d_{l_2} * \beta](N) \geq [d_{l_2} * \alpha](N)[d_{l_1} * \beta](N),$$

or equivalently,

$$\frac{[d_{l_1} * \alpha](N)}{[d_{l_1} * \beta](N)} \geq \frac{[d_{l_2} * \alpha](N)}{[d_{l_2} * \beta](N)}.$$

□

B.4 FURTHER DISTRIBUTIONAL PROPERTIES

This section presents other distributional properties that are closely related to total positivity. In section B.4.1, we review unimodality which a very common property in many probabilistic contexts. Then, we introduce a subclass of unimodal distributions that we call *half log-concavity*. Since this notion is quite uncommon in the literature, we present a rigorous introduction to this notion. A minor contribution of this section is to provide a novel and powerful characterization of this property.

B.4.1 UNIMODALITY

All material of this section is standard and may be found in (Bertin et al., 1997).

Definition B.49.
- A function $F : I \subset \mathbb{R} \to \mathbb{R}$ is said to be *unimodal* if there exists an interval S of $\mathbb{R}$ and $a \in \mathbb{R} \cup \{-\infty, \infty\}$ such that f vanishes outside S, f is non-increasing on $(-\infty, a) \cap S$ and non-decreasing on $(a, \infty) \cap S$.

  Such a is said to be a *mode* of F.

- A sequence $F = (F(n))_{n \in I \subset \mathbb{Z}}$ is said to be *discrete unimodal* if there exists an interval S of $\mathbb{Z}$ and $a \in \mathbb{Z} \cup \{-\infty, \infty\}$ such that f vanishes outside S, f is non-increasing on $]\!]-\infty, a]\!] \cap S$ and non-decreasing on $[\![a, \infty[\![\cap S$.
- A measure F on $\mathbb{R}$ is said to be *unimodal* if there exists $a \in \mathbb{R} \cup \{-\infty, \infty\}$ such that
$$F(\mathrm{d}t) = F(\{a\})\delta_a + f(t)\mathrm{d}t,$$
where $f : \mathbb{R} \to \mathbb{R}$ is an unimodal function with mode a.
- A measure F on $\mathbb{Z}$ is said to be *discrete unimodal* its pmf is discrete unimodal.

Remark. The definition above uses the convention $F(x) = 0$ for $x \notin I$. It can be proved that unimodality is independent of the domain I.

Remark. The mode of a function/sequence/measure is not unique in general, but mode$[F]$ is always a closed interval of $\mathbb{R}$ or $\mathbb{Z}$. We also define the *smallest mode* ($\underline{\text{mode}}$) and the *largest mode* ($\overline{\text{mode}}$) by
$$\underline{\text{mode}}[F] \stackrel{\text{def}}{=} \inf \text{mode}[F],$$
$$\overline{\text{mode}}[F] \stackrel{\text{def}}{=} \sup \text{mode}[F].$$

It is easy to see that an unimodal measure has at most one atom, and if it does this atom is its unique mode.

The following characterization of unimodality may be used as alternative definition.

Proposition B.50. A measure F is (resp. discrete-) unimodal with mode $a \in \mathbb{R} \cup \{-\infty, \infty\}$ if and only if its cumulative distribution function F is (resp. discrete-) convex on $(-\infty, a)$ and (resp. discrete-) concave on (a, ∞).

Remark B.51. A unimodal measure F have infinitely many densities, but only a few of them are unimodal. However, concave/convex functions always admit a right-continuous derivative. Therefore, one may show that F is unimodal at a if and only if any right-continuous density of $F - F(\{a\})\delta_a$ is a unimodal function.

STRONG UNIMODALITY. It is easy to see that log-concave measures are unimodal functions. In general, the convolution of two unimodal distribution is *not* unimodal. However, one can prove it does if at least one measure is log-concave.

Proposition B.52. Let F, G be two measures / sequences / functions on $\mathbb{R}$.
If F is log-concave and G is unimodal, then $F * G$ is unimodal.
If F is discrete log-concave and G is discrete unimodal, then $F * G$ is discrete unimodal.

A measure whose convolution preserves unimodality is said to be strongly unimodal. The fact that the two notions coincides is an important and non-trivial theorem. As a consequence, strong unimodality is often used a synonym for log-concavity.

Theorem B.53 (Ibragimov, 1956)**.** A measure is log-concave if and only if it is *strongly unimodal*: its convolution with any unimodal measure is unimodal.

A measure is discrete log-concave if and only if it is *discrete strongly unimodal*: its convolution with any discrete unimodal measure is discrete unimodal.

MONOTONE MEASURES. Monotone measures are an important subclass of unimodal measures.

Definition B.54. A measure is said to be *non-increasing* (resp. *non-decreasing*) if it is unimodal with a mode located on $\inf \operatorname{supp}[F]$ (resp. $\sup \operatorname{supp}[F]$).

A measure is said to be discrete *non-increasing* (resp. *non-decreasing*) if it is discrete unimodal with a mode located on $\inf \operatorname{supp}[F]$ (resp. $\sup \operatorname{supp}[F]$).

B.4.2 HALF LOG-CONCAVITY

This section introduces two distributional properties that we have decided to call *half* log-concavity. It provides an original characterization with exponential functions (Proposition B.57). Lower-half log-concavity is called *Yamazato property* by Watanabe (1992) and has been introduced by Yamazato (1978). Upper-half log-concavity appears in Lemma 1 of the preprint (Yu, 2011a) but without specific name. Both classes are subclasses of unimodal measures that strictly contains monotone measures.

Definition B.55. Let $f : I \subset \mathbb{R} \to \mathbb{R}$ be a function.

- f is said to be *lower-half log-concave* if
  - (i) f is unimodal with mode $a \in \mathbb{R} \cup \{-\infty, \infty\}$,
  - (ii) f is log-concave on its ascending phase $(-\infty, a)$ if $a \neq -\infty$,
  - (iii) its density f checks $f(a^-) \geq f(a^+)$ if $a \in \mathbb{R}$.
- f is said to be *upper-half log-concave* if
  - (i) f is unimodal with mode $a \in \mathbb{R} \cup \{-\infty, \infty\}$,
  - (ii) f is log-concave on its declining phase $(-\infty, a)$ if $a \neq -\infty$,
  - (iii) its density f checks $f(a^+) \geq f(a^-)$ if $a \in \mathbb{R}$.

Let F be a measure on $I \subset \mathbb{R}$.

- F is said to be *lower-half log-concave* if one of the following conditions holds
  - (i) F is a non-increasing measure,
  - (ii) F is absolutely continuous and admits a lower-half log-concave density.
- F is said to be *upper-half log-concave* if one of the following conditions holds
  - (i) F is a non-decreasing measure,
  - (ii) F is absolutely continuous and admits a upper-half log-concave density.

Let F be a measure on $I \subset \mathbb{Z}$.

- F is said to be *discrete lower-half log-concave* if its pmf f is unimodal and admits a mode $a \in \mathbb{Z} \cup \{-\infty, \infty\}$ such that f is log-concave on $]\!] - \infty, a [\![$.
- F is said to be *discrete upper-half log-concave* if its pmf f is unimodal and admits a mode $a \in \mathbb{Z} \cup \{-\infty, \infty\}$ such that f is log-concave on $]\!] a, \infty [\![$.

Remark B.56. The two half log-concavity are spatially symmetric, as the IHR adnd DRHR properties are: a measure M is lower-half log-concave if and only if the reversed measure $\tilde{M}$ is upper-half log-concave.

Remark. Half-log-concave measures may have an atom if they are monotone. Else, they are absolutely continuous.

CHARACTERIZATION WITH SIGN CHANGES. Unimodality of f can be characterized through the number of crossings with constant functions. Such idea can be traced back to (Karlin, 1968) and new proposition extends it to log-concavity by considering crossing with exponential functions. The part of this result about half log-concavity result is original.

A function or sequence f is said to have property S on $I \subset \mathbb{R}$ if it has at most two sign changes on I, and if the sign pattern is $-$, $+$, $-$ in the case of two changes (the number of sign changes is counted discarding zero terms).

Proposition B.57. Let $f : J \subset \mathbb{R} \mapsto \mathbb{R}$ be a function (resp. a sequence). Let I be the smallest interval of $\mathbb{R}$ (resp. of $\mathbb{Z}$) such that $J \subset I$. Define $f(x) = 0$ if $x \notin J$.

(i) f is (resp. discrete) unimodal if and only if $f(x) - c$ has property S on I for all $c \in \mathbb{R}$.

(ii) F is (resp. discrete) log-concave if and only if $f(x) - ce^{-\lambda x}$ has property S on I for all $c \in \mathbb{R}$, $\lambda \in \mathbb{R}$.

(iii) F is (resp. discrete) lower-half log-concave if and only if $f(x) - ce^{-\lambda x}$ has property S on I for all $c \in \mathbb{R}$, $\lambda \leq 0$.

(iv) F is (resp. discrete) upper-half log-concave if and only if $f(x) - ce^{-\lambda x}$ has property S on I for all $c \in \mathbb{R}$, $\lambda \geq 0$.

Proof. [(i), unimodality] Assume f is not unimodal. There exists of $x < y < z$ such that $f(x) > f(y)$ and $f(z) > f(y)$. Take $c = (f(y) + \min(f(x), f(z)))/2$ such that $f(x) > c > f(y)$ and $f(x) > c > f(z)$. So $f - c$ has not property S: it has more than two sign changes, or exactly two sign changes with sign pattern $+, -, +$.

Reciprocally, assume f is unimodal. Then, $f - c$ is non-decreasing on $(-\infty, a) \cap I$ and has at most one sign change there with pattern $-$, $+$. In addition, $f - c$ is non-increasing on $(a, \infty) \cap I$ and has at most sign change there with pattern $+$, $-$. So globally on I, $f - c$ has property S.

[(ii), log-concavity] This has been proved in (Karlin, 1968, Propositions 3.1 and 3.2), see also (Yu, 2011b).

[(iii), lower, $\Longrightarrow$] Assume $\lambda < 0$. Let us prove that $g(x) := e^{\lambda x} f(x)$ is unimodal. By assumption, f is log-concave on $(-\infty, a)$. So by claim (ii), $g(x)$ is unimodal on this interval, so it is non-increasing then non-decreasing. By definition of unimodality, f is non-increasing on (a, ∞). As $\lambda < 0$, g is non-increasing too on this interval By definition of half-log-concavity, $g(a^-) \geq g(a^+)$. This suffices to show that g is unimodal on the whole real line.

[(iii), lower, $\Longleftarrow$] First, setting $\lambda = 0$ and using claim (i) proves that f is unimodal. If f is non-decreasing, then f is lower-half log-concave. Else, there exists $a \in (-\infty, \infty]$ such that f is non-decreasing on $(-\infty, a)$ and non identically vanishing on this interval.

Second, let us show that f is log-concave on this interval. Consider the restriction of f on this interval, denoted $g := f_{|(-\infty,a)}$. For all $\lambda > 0$, since g is non-decreasing, $e^{\lambda x} g(x)$ is non-decreasing and therefore unimodal. For all $\lambda < 0$, $e^{\lambda x} f(x)$ is unimodal by assumption, therefore its restriction $e^{\lambda x} g(x)$ is unimodal too. So claim (ii) applies and proves that g is log-concave.

Third, let us show that $f(a^-) \geq f(a^+)$ (in case $a < \infty$). By contradiction, assume that $f(a^-) < f(a^+)$. As f is non-identically vanishing and nonnegative on $(-\infty, a)$ (since it is log-concave), there exists $b < a$ such that $f(b) > 0$. Choose $\lambda < 0$ such that $e^{\lambda b} f(b) > e^{\lambda a} f(a^-)$. As $e^{\lambda a} f(a^+) > e^{\lambda a} f(a^-)$, this contradicts $e^{\lambda x} f(x)$ being unimodal. So we have proven that f is lower-half log-concave.

[(iv), upper] The proof is similar and can be deduced by spatial reversion $x \leftarrow -x$. □

CHARACTERIZATION WITH EXPONENTIAL TILTING. Next proposition extends the previous one to general measures. It characterizes half-log-concavity using *exponential tilting* ($\mathcal{E}_\lambda$ or E_λ) which is also called *Esscher transform* (Esscher, 1932).

Definition B.58. The exponential tilting operators $\mathcal{E}_\lambda$ are defined for all $\lambda \in \mathbb{R}$.

$\mathcal{E}_\lambda$ maps a measure F on $A \subset \mathbb{R}$ onto the measure $\mathcal{E}_\lambda F$ on A defined by

$$\forall t \in A, \qquad \mathcal{E}_\lambda F(\mathrm{d}t) \stackrel{\text{def}}{=} e^{\lambda t} F(\mathrm{d}t).$$

$\mathcal{E}_\lambda$ maps a function (or a sequence) $F : I \subset \mathbb{R} \to \mathbb{R}$ onto the function $\mathcal{E}_\lambda F : I \to \mathbb{R}$ defined by

$$\forall t \in I, \qquad \mathcal{E}_\lambda F(t) \stackrel{\text{def}}{=} e^{\lambda t} F(t).$$

Proposition B.59. Let F be a measure, function or sequence on $I \subset \mathbb{R}$.

(i) F is lower-half log-concave if and only if $\mathcal{E}_\lambda F$ is unimodal for all $\lambda \leq 0$.

(ii) F is upper-half log-concave if and only if $\mathcal{E}_\lambda F$ is unimodal for all $\lambda \geq 0$.

(iii) F is log-concave if and only if $\mathcal{E}_\lambda F$ is unimodal for all $\lambda \in \mathbb{R}$.

Similar equivalences hold for the discrete counterparts and sequences on $I \subset \mathbb{Z}$.

Proof. We only have to prove claim (i). Indeed, it is easily proven that spatial reversion commutes with exponential tilting as follows: $\widetilde{\mathcal{E}_\lambda F} = \mathcal{E}_{-\lambda} \tilde{F}$. So thanks to spatial reversion, the first two claims are equivalent. In addition, their combination gives last claim due to Proposition B.60

[$\Longrightarrow$] It is immediate as F has a (resp. discrete-) lower-half log-concave density (resp. pmf) f and $\mathcal{E}_\lambda F(\mathrm{d}t) = e^{\lambda t} f(t) \mathrm{d}t$. As $e^{\lambda t} > 0$, the sign changes of $e^{\lambda t} f(t) - c$ are identical to $f(t) - ce^{-\lambda t}$, which proves the result thanks to Proposition B.57.

[$\Longleftarrow$] Assume $\mathcal{E}_\lambda F$ is unimodal for all $\lambda \geq 0$.

Case 1. Assume F has an atom at a. For all $\lambda \leq 0$, $\mathcal{E}_\lambda F$ has an atom at a. By hypothesis, $\mathcal{E}_\lambda F$ is unimodal, so necessarily its unique mode is a. As $F = \mathcal{E}_0 F$, it is unimodal with mode a. Let f be a right-continuou unimodal density of the unimodal

measure $\hat{F} \overset{\text{def}}{=} F - F(\{a\})\delta_a$. Assume there exists $b < a$ such that $f(b) > 0$. Let us show it leads to a contradiction. As f is non-decreasing on $(-\infty, a)$, then $\infty > f(c) \geq f(b) > 0$ with $c := (a+b)/2$. Choose λ such as $e^{\lambda c} f(c) < e^{\lambda} f(b)$. Then $\lambda > 0$ and

$$\mathcal{E}_\lambda f(c) < \mathcal{E}_\lambda f(b).$$

As $\mathcal{E}_\lambda f$ is a valid density of $\mathcal{E}_\lambda \hat{F}$ and is right-continuous, Remark B.51 tells it has to be unimodal with mode a. But the inequality proves it cannot be non-decreasing on $(b, c) \subset (-\infty, a)$. This contradicts $\mathcal{E}_\lambda F$ having its unique mode at a. So we have proven that f vanishes on $(-\infty, a)$. This proves that F is non-decreasing, hence lower-half log-concave by definition.

Case 2. Assume F is absolutely continuous and let f be a right-continuous density of F. As $\mathcal{E}_\lambda f$ is also right-continuous density of $\mathcal{E}_\lambda F$, it has to be unimodal for all $\lambda \leq 0$. Proposition B.57 implies f is lower-half log-concave.

□

RELATIONSHIP WITH TP_2 PROPERTIES. Unimodality implies *partial* DRHR, IHR properties. Half log-concavity implies one full property among the two.

Proposition B.60. The following results also holds for discrete counterparts.

(i) Let F be an unimodal measure with mode $a \in \mathbb{R}$.
F is IHR on (a, ∞).
F is DRHR on $(-\infty, a)$.

(ii) Let F be a discrete unimodal measure with mode $a \in \mathbb{Z}$.
F is discrete IHR on $]\!] - \infty, a]\!]$.
F is discrete DRHR on $[\![a, \infty[\![$.
F is discrete log-concave at a.

(iii) A lower-half log-concave measure is DRHR. In particular, a non-increasing measure is DRHR.

(iv) A upper-half log-concave measure is IHR. In particular, a non-decreasing measure is IHR.

(v) A measure is log-concave if and only if it is lower-half and upper-half log-concave.

B.4.3 MULTIPLICATIVE STRONG UNIMODALITY AND RELATED

Log-concavity (strong unimodality), IHR and DRHR classes are all characterized in Proposition B.26 by stochastic monotony under shifts $D(\cdot) \mapsto D(\cdot - a)$. For a random variable X, shift corresponds to an addition $X + a$. This section introduces the multiplicative counterparts of these three reliability classes: MSU, M-IHR, M-DRHR. The base idea is to replace addition $X + a$ by multiplication aX. Whereas MSU property is quite standard, the two others ones are far less common in the literature and there is no consensus on their name. We have decided to call them M-IHR and M-DRHR in order to emphasize the parallel with IHR and DRHR classes.

Afterwards, this section introduces the discrete counterparts of the three properties. A contribution it to derive two original characterizations which mimic the case of continuous. Since there are no consensus on the definition of such discrete properties, this similarity brings justification to our choice.

CONTINUOUS DISTRIBUTIONS. Our definition of MSU follows (Simon, 2011). Note it is slightly more restrictive than other common definitions like Cuculescu and Theodorescu (1998). All statements about continuous properties can be found in the two references.

Definition B.61. Let F be a finite nonnegative measure on $[0, \infty)$. F is said to be

- *Multiplicative Strongly Unimodal* (MSU) if absolutely continuous and admits a density f on $[0, \infty)$ such that $x \mapsto f(e^x)$ is log-concave on $\mathbb{R}$.
- *Multiplicative Increasing Hazard Rate* (M-IHR) if $x \mapsto \overline{F}(e^x)$ is log-concave on $\mathbb{R}$.
- *Multiplicative Decreasing Reverse Hazard Rate* (M-DRHR) if $x \mapsto F(e^x)$ is log-concave on $\mathbb{R}$.

The following characterization relates the three distributional properties to stochastic orders. It mimics the characterization of log-concavity, IHR, DRHR properties (Proposition B.26). Shifting $F(\cdot - a)$ is replaced by scaling $S_a F$.

Proposition B.62. Let F be a probability measure on $[0, \infty)$ and S_a denote the scaling operator.

- F is MSU if and only if $(S_a F)_{a \in (0,1]}$ is non-decreasing in the likelihood ratio ordering.
- F is M-IHR if and only if $(S_a F)_{a \in (0,1]}$ is non-decreasing in the hazard rate ordering.
- F is M-DRHR if and only if $(S_a F)_{a \in (0,1]}$ is non-decreasing in the reverse hazard rate ordering.

In other words, let X be a real random variable.

- X is MSU if and only if for all $a \in (0, 1]$, $aX \underset{\text{lr}}{\leqslant} X$.
- X is M-IHR if and only if for all $a \in (0, 1]$, $aX \underset{\text{hr}}{\leqslant} X$.
- X is M-DRHR if and only if for all $a \in (0, 1]$, $aX \underset{\text{rh}}{\leqslant} X$.

The result is still valid by replacing $(0, 1]$ with $(1, \infty)$, $(0, \infty)$ or any interval that strictly contains 1.

The following proposition is another characterization that will conveniently extend to discrete properties. Note that a M-DRHR, M-IHR distribution may have an atom at the boundary of it support.

Proposition B.63. Let F be a finite measure on $[0,\infty)$. Assume F is absolutely continuous.

- F is MSU if and only if F admits on a density f that is almost everywhere differentiable and such that one of the following equivalent conditions holds:
  (i) $x \mapsto x\frac{f'(x)}{f(x)}$ is non-increasing on $\{x \geq 0 \mid f(x) \neq 0\}$.
  (ii) $f(x) \underset{\text{DP}}{\leqslant} -xf'(x)$.
- F is M-IHR if and only if F is absolutely continuous on $(-\infty, \sup \text{supp}[F])$ admits a density f on such that one of the following equivalent conditions holds:
  (i) $x \mapsto x\frac{f(x)}{\overline{F}(x)}$ is non-decreasing on $\{x \geq 0 \mid \overline{F}(x) \neq 0\}$.
  (ii) $\overline{F}(x) \underset{\text{TP}}{\leqslant} xf(x)$.
- F is M-DRHR if and only if F is absolutely continuous on $(\inf \text{supp}[F], \infty)$ and admits a density f such that one of the following equivalent conditions holds:
  (i) $x \mapsto x\frac{f(x)}{F(x)}$ is non-increasing on $\{x \geq 0 \mid F(x) \neq 0\}$.
  (ii) $xf(x) \underset{\text{TP}}{\leqslant} F(x)$.

DISCRETE DISTRIBUTIONS. We introduce the discrete counterparts of the MSU, M-IHR, M-DRHR classes. These discrete properties are even less common in the literature and there is no consensus on their definition. The versions we choose are motivated by discretizing quantities introduced in Proposition B.63. Our version of the definition of discrete M-IHR we get is identical to (Banciu and Mirchandani, 2013). Our version for discrete M-DRHR matches (Veres-Ferrer and Pavía, 2012), but not (Veres-Ferrer and Pavía, 2016) where the corresponding ratio (called *elasticity*) is rather defined as $nf(n)/F(n)$.

Definition B.64. Let F be a discrete measure and f denote its pmf.

- F is discrete MSU if $n \mapsto \frac{(n+1)f(n+1)-nf(n)}{f(n)}$ is non-increasing on $\text{supp}[f(.)]$.
- F is discrete M-IHR if $n \mapsto \frac{nf(n)}{\overline{F}(n)}$ is non-decreasing on $\text{supp}[\overline{F}(.)]$.
- F is discrete M-DRHR if $n \mapsto \frac{(n+1)f(n+1)}{F(n)}$ is non-increasing on $\text{supp}[F(.)]$.

We give two original characterizations of the three discrete properties. The first one requires a couple of linear operators. More background on binomial thinning is provided in Section 8.1.2, Chapter 8, refer to Definition 8.2.

Definition B.65. The size-biasing operator M is defined on discrete sequences d on $\mathbb{Z}$ by $Md(n) \overset{\text{def}}{=} (n+1)d(n+1)$.

The forward difference operator Δ is defined on discrete sequences d on $\mathbb{Z}$ by $\Delta d(n) \overset{\text{def}}{=} d(n+1) - d(n)$.

Binominal thinning B_a is defined for all $a \in [0,1]$ on discrete sequences on $\mathbb{N}$ by $B_a d(n) \overset{\text{def}}{=} \sum_{k=n}^{\infty} d(k)a^k(1-a)^{n-k}$.

Proposition B.66. Let F denote a discrete distribution on $\mathbb{N}$.

- F is discrete MSU if and only if $f \underset{\text{DP}}{\leqslant} -\Delta M f$.
- F is discrete M-IHR if and only if $\overline{F} \underset{\text{TP}}{\leqslant} Mf$.
- F is discrete M-DRHR if and only if $Mf \underset{\text{TP}}{\leqslant} F$.

Next proposition characterizes the three properties with stochastic ordering, similarly to Proposition B.62. The discrete counterpart of scaling S_a is binomial thinning B_a, defined for $a \in [0,1]$. Contrary to S_a, the proof that binomial thinning B_a preserve MSU, M-IHR, M-DRHR properties is non-trivial.

Proposition B.67. Let F denote a discrete distribution on $\mathbb{N}$.

- Binomial thinning B_a preserves discrete MSU, M-DRHR, M-IHR classes.
- F is discrete MSU $\iff$ $(B_a F)_{a\in(0,1]}$ is non-decreasing in the likelihood ratio ordering.
- F is discrete M-DRHR $\iff$ $(B_a F)_{a\in(0,1]}$ is non-decreasing in the reverse hazard rate ordering.
- F is discrete M-IHR $\iff$ $(B_a F)_{a\in(0,1]}$ is non-decreasing in the hazard rate ordering.

The conclusions are still valid if $(0,1]$ is replaced by any non-empty interval $(a_0,1]$ with $a_0 < 1$.

Proof. Preliminaries. The operators M, B_a commute in the sense $M \circ B_a = a(B_a \circ M)$. Indeed, M is equivalently defined by derivative of the Z-transform, $Z[Mf](z) = Z[f]'(z)$. *And*B_a is equivalently defined by $Z[B_a g](z) = Z[g](az + 1 - a)$. So $Z[M \circ B_a f] = aZ[f]'(az + 1 - a) = aZ[B_a \circ Mf]$.

Similarly, the operators Δ, B_a commute in the sense $B_a \circ \Delta = a(\Delta \circ B_a)$. Indeed, Δ is equivalently defined as $Z[\Delta g](z) = (z-1)Z[g](z)$. Let g be any sequence,

$$\begin{aligned} Z[B_a \circ \Delta g](z) &= a(z-1)Z[g](1-a+az) = a(z-1)Z[g](1-a+az) \\ &= a(z-1)Z[B_a g](z) = aZ[\Delta \circ B_a g](z) = Z[a\Delta \circ B_a g](z), \end{aligned}$$

which proves the identity.

Step 1. Since $B_0 f = \delta$ which is MSU, M-IHR and M-DRHR, assume $a \neq 0$.

[Preservation of MSU] The pmf f is MSU if and only if $f \underset{\text{DP}}{\leqslant} -\Delta \circ Mf$. As B_a is a linear TP_2 operator, it preserves DP ordering: $B_a f \underset{\text{DP}}{\leqslant} -B_a \circ \Delta \circ Mf$. Since $B_a \circ \Delta \circ M = \Delta \circ [aB_a \circ M] = \Delta \circ M \circ B_a$, this gives $[B_a f] \underset{\text{DP}}{\leqslant} -\Delta \circ M[B_a F]$. So $B_a f$ is MSU.

[Preservation of M-DRHR] A pmf f is M-DRHR if and only if $Mf \underset{\text{TP}}{\leqslant} f * \mathbf{1}$, where $\mathbf{1} := (1)_{n\in\mathbb{N}}$ denote the constant sequence equal to 1 on $\mathbb{N}$. Indeed, the cumulative sequence F can be written as $F = f * \mathbf{1}$. Since B_a is TP_2 operator, it preserves TP ordering:

$$B_a Mf \underset{\text{TP}}{\leqslant} B_a[f * \mathbf{1}].$$

First, $B_a \circ Mf = aM \circ B_a f$. Second, it can be readily checked that B_a commutes with convolution and $B_a\mathbf{1} = (1/a)\mathbf{1}$. As $a \neq 0$, one obtains

$$M[B_a f] \underset{\text{TP}}{\leqslant} [B_a f] * \mathbf{1}.$$

This proves that $B_a f$ is M-DRHR.

[Preservation of M-IHR] A pmf f is M-IHR if and only if $Mf \underset{\text{TP}}{\geqslant} \mathbf{1} * [\delta - f]$. Indeed, the ratio $r(n) := nf(n)/\overline{F}(n)$ is non-negative and vanishes at $n = 0$, it is non-decreasing on $\text{supp}[\overline{F}(.)]$ if and only if $(n+1)f(n+1)/\overline{F}(n+1)$ does on $\text{supp}[\overline{F}(.+1)]$. In addition, the survivor sequence $\overline{F}$ checks $\overline{F}(n+1) = 1 - F(n) = \mathbf{1}(n) - f * \mathbf{1}(n) = \mathbf{1} * [\delta - f](n)$ for all $n \in \mathbb{N}$.

Assume f is M-IHR. Since B_a is TP_2 operator, it preserves TP ordering:

$$B_a Mf \underset{\text{TP}}{\geqslant} B_a[\mathbf{1} * [\delta - f](n)].$$

First, $B_a \circ Mf = aM \circ B_a f$. Second, $B_a[\mathbf{1}] = 1/a\mathbf{1}$, $B_a[\delta_0]\delta_0$, B_a is linear commutes with convolution. This gives $B_a[\mathbf{1} * [\delta - f]] = 1/a\mathbf{1} * [\delta - B_a f]$. Therefore,

$$M[B_a f] \underset{\text{TP}}{\geqslant} \mathbf{1} * [\delta - B_a f],$$

which proves that $B_a f$ is M-IHR.

Step 2. Define $I := (a_0, 1]$ for some $a_0 < 1$. Let us prove that $(B_a F)_{a \in I} \uparrow lr$ is equivalent to F being MSU. Let f_a denote the pmf of $B_a F$. The proof is a simple application on Proposition B.33: if the functions $a \mapsto f_a(n)$ are continuously differentiable on I (with ∂_a at $a = 1$ being the left-hand side derivative), then a family (f_a) is non-decreasing in the lr, hr or rh ordering if and only if the for all $a \in I$, the respective functions $\partial_a \log f_a(n)$, $\partial_a \log \overline{F}_a(n)$, $\partial_a \log F_a(n)$ are non-decreasing with respect to n. In the present case, $Z[f_a](z) = P(az + 1 - a)$, so

$$\forall |z| < R_{f_a}, \quad \partial_a Z[f_a](z) = a(z-1)Z[f_a](z) = Z[\partial_a f_a](z).$$

This proves for all $a \in (0, 1]$ that

$$\forall n \in \text{supp}[f_a], \quad \partial_a \log f_a(n) = \frac{\partial_a f_a(n)}{f_a(n)} = -\frac{1}{a}\frac{(n+1)f_a(n+1) - nf_a(n)}{f_a(n)}.$$

So $n \mapsto \partial_a \log f_a(n)$ being non-decreasing for all $a \in I$ is clearly equivalent to $B_a F$ being MSU for all $a \in I$. And due to preservation by B_a, this latter assertion is equivalent to F being MSU.

By summations over n, one obtains

$$\forall a \in (0, 1], \forall n \in \text{supp}[\overline{F}_a], \quad \partial_a \log \overline{F}_a(n) = \frac{1}{a} n f_a(n)/\overline{F}_a(n),$$
$$\forall a \in (0, 1], \forall n \in \text{supp}[F_a], \quad \partial_a \log F_a(n) = -\frac{1}{a}(n+1) f_a(n+1)/F_a(n),$$

so the same argument gives the two other equivalences for M-DRHR and M-IHR. □

RELATIONSHIP BETWEEN PROPERTIES. The following relationships hold for continuous and discrete properties. In general, there is no relationship between MSU and log-concavity, except for the case of non-increasing distributions.

Proposition B.68. The same implications hold with continuous counterparts.

$$\begin{aligned} f \text{ is discrete MSU} &\implies f \text{ is discrete M-IHR and discrete M-DRHR} \\ f \text{ is discrete M-DRHR} &\implies f \text{ is discrete DRHR} \\ f \text{ is discrete M-IHR} &\impliedby f \text{ is discrete IHR} \\ f \text{ is discrete LCAV and non-increasing} &\implies f \text{ is discrete MSU} \end{aligned}$$

Proof. *[1.]* It suffices to know that the lr order implies rh and hr orders.

[2.] F being discrete M-DRHR means that $nf(n)/F(n)$ is non-increasing. This trivially implies that $f(n)/F(n)$ is non-increasing.

[3.] F being discrete IHR means that $f(n)/F(n)$ is non-decreasing. This trivially implies that $nf(n)/\overline{F}(n)$ is non-decreasing.

[4.] $-\frac{(n+1)f(n+1-nf(n))}{f(n)} = (n+1)\left(1-\frac{f(n+1)}{f(n)}\right) - 1$. f being non-increasing means that $1 - f(n+1)/f(n)$ is nonnegative, and f being discrete log-concave means that $f(n+1)/f(n)$ is non-increasing. So $1-f(n+1)/f(n)$ is non-decreasing and nonnegative as $(n+1)$ is, so multiplying both gives that $-\frac{(n+1)f(n+1-nf(n))}{f(n)}$ is non-decreasing, which gives f being discrete MSU. □

B.5 POISSON MIXTURES AND TOTAL POSITIVITY

For many TP_2-related properties, Poisson mixtures are very powerful tool to relate the definition for discrete measures and the one for continuous measures. The two main interests of this transform is commuting with many operations on measures, such as convolution, and mapping many TP_2 properties to their discrete counterpart. In addition, Poisson mixing may be considered as a special procedure of discretization of continuous distributions on $\mathbb{R}_+$.

Most definitions and results in this section are known material. Refer to (Steutel and van Harn, 2004, Chapter 6, Section 6) for instance. But this section also contains original material or at least results we could not find explicitly in the literature. This section also provides several alternative proofs that are all based on preservation results.

B.5.1 DEFINITION

Poisson mixtures are based on the Laplace-Stieltjes transform which extends the standard Laplace transform to measures. Hereafter, we use the notation $\overline{\mathbb{R}} \stackrel{\text{def}}{=} \mathbb{R} \cup \{\infty\}$.

Definition B.69 (*Laplace transform* ($\mathcal{L}$))**.**

- Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a Lebesgue-measurable function.
  - The *abscissa of convergence* $\mathrm{abs}[f] \in \overline{\mathbb{R}}$ is defined as $\mathrm{abs}[f] \stackrel{\text{def}}{=} \inf\{\lambda \in \overline{\mathbb{R}} \mid t \mapsto e^{-\lambda t} f(t) \in L^1(\mathbb{R}_+)\}$.
  - The *Laplace-Stieltjes transform* of f such that $\mathrm{abs}[f] < \infty$ is the real-valued function $\mathcal{L}f$ defined for all $\lambda > \mathrm{abs}[f]$ as

$$\mathcal{L}F(\lambda) \stackrel{\text{def}}{=} \int_0^\infty e^{-\lambda t} f(t) \mathrm{d}t.$$

- Let F be a nonnegative measure on $\mathbb{R}_+$.
  - The *abscissa of convergence* $\mathrm{abs}[F] \in \overline{\mathbb{R}}$ is defined as $\mathrm{abs}[F] \stackrel{\text{def}}{=} \inf\{\lambda \in \overline{\mathbb{R}} \mid t \mapsto e^{-\lambda t} \in L^1(\mathbb{R}_+, F)\}$.
  - The *Laplace-Stieltjes transform* of F such that $\mathrm{abs}[F] < \infty$ is the real-valued function $\mathcal{L}F$ defined for all $\lambda > \mathrm{abs}[f]$ by

$$\begin{aligned}\mathcal{L}F(\lambda) &\stackrel{\text{def}}{=} \int_{[0,\infty)} e^{-\lambda t} \mathrm{d}F(t) \\ &= F(0) + \int_{(0,\infty)} e^{-\lambda t} \mathrm{d}F(t).\end{aligned}$$

Definition B.70 (*Poisson mixture* (Γ_λ))**.**

- Let F be a measure on $\mathbb{R}_+$ such that $\mathrm{abs}[F] < \infty$. For all $\lambda > \max(\mathrm{abs}[F], 0)$, the *Poisson mixture* $\Gamma_\lambda[\mathrm{d}F]$ is the distribution on $\mathbb{N}$ defined by

$$\begin{aligned}\forall n \in \mathbb{N}, \quad \Gamma_\lambda[\mathrm{d}F](n) &\stackrel{\text{def}}{=} \frac{(-\lambda)^n}{n!} \frac{\mathrm{d}^n}{\mathrm{d}\lambda^n} \mathcal{L}F(\lambda) \\ &= F(0)\delta_0(n) + \int_{(0,\infty)} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \mathrm{d}F(t).\end{aligned} \tag{B.5}$$

- Let $f : \mathbb{R}_+ \mapsto \mathbb{R}$ be a measurable function such that $\mathrm{abs}[f] < \infty$. For all $\lambda > \max(\mathrm{abs}[f], 0)$, the *Poisson mixture* $\Gamma_\lambda[f]$ is the distribution on $\mathbb{N}$ defined by

$$\forall n \in \mathbb{N}, \quad \Gamma_\lambda[f](n) \stackrel{\text{def}}{=} \int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t} f(t) \mathrm{d}t.$$

Poisson mixture can be alternatively described using Laplace and Z-transforms. As Proposition B.82 explains later on, this fact stresses out the interpretation of the λ parameter of Γ_λ as a scale factor.

Proposition B.71. Let F be a measure such that $\lambda_F := \max(0, \mathrm{abs}[F]) < \infty$. Let $\mathcal{L}F$ denote its Laplace-Stieltjes transform. Then, the Z-transform of the sequence $\Gamma_\lambda[\mathrm{d}F]$ equals,

$$\forall \lambda > \lambda_F, \quad \forall |z| < 1 - \frac{\lambda_F}{\lambda}, \quad Z[\Gamma_\lambda[\mathrm{d}F]](z) = \mathcal{L}F(\lambda(1-z)).$$

Remark.
- All definitions are consistent between functions and measures. If F admits a density f, then $\mathrm{abs}[f] = \mathrm{abs}[F]$ and $\Gamma_\lambda[\mathrm{d}F] = \Gamma_\lambda[f]$.

- The definition of $\mathcal{L}$ and Γ_λ uses Lebesgue integration for functions and Lebesgue-Stieltjes integration for measures.
- Previous definition involves two standard result about Laplace transform. First, for all distribution/function F, the function $\mathcal{L}F$ is analytical on $(\mathrm{abs}[F], \infty)$. Second, for all $\lambda > \mathrm{abs}[f]$ and $n \in \mathbb{N}$, $\lim_{t\to\infty} e^{-\lambda t} t^n F(t) = 0$ and $t \mapsto e^{-\lambda t} t^n F(t) \in L^1(\mathbb{R}_+)$.
- If F is a bounded function or a finite measure such as a probability distribution, then $\mathrm{abs}[F] \leq 0$.

PRESERVATION OF PROBABILITY MEASURES. The main interesting Poisson mixtures is transforming any probability measure on $\mathbb{R}_+$ into a *discrete* probability measure on $\mathbb{N}$. Indeed, $p_\lambda(n) := e^\lambda \lambda^n/(n!)$ is the pmf of Poisson distribution $Po(\lambda)$ with intensity λ. Thus $\Gamma_\lambda[\mathrm{d}F]$ is a mixture of Poisson distributions with F (more precisely, $S_\lambda F$) as compounding distribution:

$$\Gamma_\lambda[\mathrm{d}F](n) = \int_{\mathbb{R}_+} p_{\lambda t}(n) \mathrm{d}F(t).$$

Next proposition gives more generality: Poisson mixture map nonnegative measure to nonnegative sequences, and it preserves the total mass of the measure. In addition, the resulting sequence has always full support.

Proposition B.72. (i) If F is a nonnegative measure on $\mathbb{R}_+$, then

$$\sum_{n=0}^{\infty} \Gamma_\lambda[\mathrm{d}F](n) = F(\mathbb{R}_+) \in \mathbb{R} \cup \{\infty\}.$$

(ii) If F is a nonnegative measure on $\mathbb{R}_+$,

$$\forall n \in \mathbb{N}, \quad \Gamma_\lambda[\mathrm{d}F](n) \geq 0$$
$$> 0 \text{ if } F \text{ is not proportional to } \delta_0.$$

In other words, the Poisson mixture of $F \not\propto \delta_0$ has full support, $\mathrm{supp}[\Gamma_\lambda[\mathrm{d}F]] = \mathbb{N}$.

(iii) If f is a nonnegative function on $\mathbb{R}_+$, then

$$\forall n \in \mathbb{N}, \quad \Gamma_\lambda[f](n) \geq 0$$
$$> 0 \text{ if } f \text{ does not vanish almost everywhere.}$$

Proof. [(i)] The identity $e^z = \sum_{n=0}^\infty z^n/n!$ implies

$$\begin{aligned}\sum_{n=0}^{\infty} \Gamma_\lambda[\mathrm{d}F](n) &= F(0) + \int_{(0,\infty)} \mathrm{d}F(t) \left[e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} \right] \\ &= F(0) + \int_{(0,\infty)} \mathrm{d}F(t) = F([0,\infty)) = 1.\end{aligned}$$

[(ii-iii)] Since, $e^\lambda \lambda^n > 0$ on $(0, \infty)$, this is immediate by positivity of integration. □

Surprisingly, the cumulative and survivor distributions of a Poisson mixture are also (scaled) Poisson mixtures.

Proposition B.73. Let F be a probability measure on $[0,\infty)$ and $\lambda > 0$.
Then, $p(n) := \Gamma_\lambda[\mathrm{d}F](n)$ defines a probability mass function on $\mathbb{N}$.
Its cumulative distribution is $P(n) = \lambda\,\Gamma_\lambda[F](n) = \Gamma_1[F(\cdot\lambda)](n)$.
Its survivor distribution is
$$\overline{P}(n) = \begin{cases} \lambda\,\Gamma_\lambda\left[\overline{F}\right](n-1) & \text{if } n \in \mathbb{N}^*, \\ 1 & \text{if } n = 0. \end{cases}$$

Proof. [Claim 1] If F is a probability measure, $F \geq 0$ and $F(\mathbb{R}) = 1$. So $p := \Gamma_\lambda[\mathrm{d}F]$ is a nonnegative sequence whose sum equals 1. So it defines a valid pmf.

[Claim 2] P is the only sequence that checks $P(0) = p(0)$ and the recursion $\forall n \in \mathbb{N}^*, P(n) - P(n-1) = p(n)$.

Integration by parts in the Stieltjes sense (see (Fristedt and Gray, 1997, Chapter 4, Proposition 19)) gives
$$\int_{-\infty}^{\infty} \phi(t)\mathrm{d}F(t) = \phi(0) + \int_{(0,\infty)} (1 - F(t))\phi'(t)\mathrm{d}t$$
for a differentiable function ϕ with bounded variation. For $n \in \mathbb{N}^*$, applying it to $\phi(t) = e^{-\lambda t}t^n$, $\phi'(t) = e^{-\lambda t}(-\lambda t^n + nt^{n-1})$ gives,
$$\int_{-\infty}^{\infty} \phi(t)\mathrm{d}F(t) = \phi(0) + \int_0^{\infty} (1 - F(t))\phi'(t)\mathrm{d}t$$
as $\lim_\infty \phi = 0$,
$$= \phi(0) + \int_0^{\infty} \phi'(t)\mathrm{d}t - \int_0^{\infty} F(t)\phi'(t)\mathrm{d}t$$
$$\int_{-\infty}^{\infty} e^{-\lambda t}\frac{(\lambda t)^n}{n!}\mathrm{d}F(t) = -\int_0^{\infty} F(t)e^{-\lambda t}\lambda\left[\frac{(\lambda t)^{n-1}}{(n-1)!} - \frac{(\lambda t)^n}{n!}\right]\mathrm{d}t,$$

which proves the recursion since it reads
$$\Gamma_\lambda[\mathrm{d}F](n) = \lambda\Gamma_\lambda[F](n) - \lambda\Gamma_\lambda[F](n-1).$$
For the initialization $n = 0$, $\phi(t) = e^{-\lambda t}$ and $\phi'(t) = -\lambda e^{-\lambda t}$. So the integration by parts gives
$$\Gamma_\lambda[\mathrm{d}F](0) = \lambda\Gamma_\lambda[\overline{F}](0),$$
and this ends proving the second claim.

[Claim 3] for all $n \in \mathbb{N}^*$, the discontinuities of a monotone function like $\overline{F}$ are at most countable, so the functions $\overline{F}(.^-)$ and $\overline{F}(.)$ are almost everywhere equal and this implies
$$\Gamma_\lambda[\overline{F}](n) = \int_{\mathbb{R}_+} e^{-\lambda t}\frac{(\lambda t)^n}{n!}\overline{F}(t)\mathrm{d}t = \int_{\mathbb{R}_+} e^{-\lambda t}\frac{(\lambda t)^n}{n!}\overline{F}(t^-)\mathrm{d}t.$$
Therefore,
$$\lambda\Gamma_\lambda[\overline{F}](n) + \lambda\Gamma_\lambda[F](n) = \int_{\mathbb{R}_+} e^{-t}\frac{t^n}{n!}\overline{F}(t)\mathrm{d}t + \int_{\mathbb{R}_+} e^{-t}\frac{t^n}{n!}F(t)\mathrm{d}t$$
$$= \int_{\mathbb{R}_+} e^{-t}\frac{t^n}{n!}\left(\overline{F}(t^-) + F(t)\right)\mathrm{d}t,$$

and since $\overline{F}(t^-)+F(t)=1$ holds for all time t and any F by definition of the continuous survivor distribution,

$$= \frac{\int_{\mathbb{R}_+} e^{-t}\, t^{(n+1)-1}\mathrm{d}t}{n!},$$

and by definition of the Gamma function Γ,

$$= \frac{\Gamma(n+1)}{\Gamma(n+1)} = 1.$$

By definition of the discrete survivor distribution, $P(n+1) = 1-\overline{P}(n)$ for all $n \in \mathbb{N}^*$. So $\overline{P}(n) = 1-\lambda\Gamma_\lambda[F](n-1) = \lambda\Gamma_\lambda[\overline{F}](n-1)$. Besides, as p is supported on $\mathbb{N}$, $\overline{P}(0) = \sum_{k=0}^{\infty} p(k) = 1$. This proves the last claim. □

Next proposition is of major interest. It explains how Poisson mixture provides with a method to approximate any measure by discrete measures. By considering $1/\lambda$ as the discretization spatial step, the mixtures converge to the original distribution as this step goes to 0.

Proposition B.74 (discrete approximation)**.** Let F be a probability measure on $\mathbb{R}_+$. For all $\lambda > 0$, let F^λ be the measure defined by

$$F^\lambda \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} \Gamma_\lambda[\mathrm{d}F](n)\, \delta_{\frac{n}{\lambda}}.$$

Then, F^λ is a discrete probability measure supported on $\lambda^{-1}\mathbb{N}$, and F^λ converge to F in distribution as $\lambda \to \infty$.

Let X be a nonnegative random variable such that $X \sim F$. Let $N = (N(t)_{t\geq 0}$ be a Poisson process such that $N(t) \sim Po(t)$. Then, $X^\lambda \stackrel{\text{def}}{=} \lambda^{-1} N(\lambda X)$ defines a $\lambda^{-1}\mathbb{N}$-valued random variable such that $X^\lambda \sim F^\lambda$.

Proof. The Laplace-Stieltjes transforms of the finite measures F_λ are defined for all $t > 0$ and equal to:

$$\begin{aligned}\mathcal{L}F_\lambda(t) &= \sum_{n=0}^{\infty} \Gamma_\lambda[\mathrm{d}F](n) e^{-\frac{t}{\lambda}n} \\ &= \sum_{n=0}^{\infty} \frac{1}{n!}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathcal{L}F(\lambda)(-\lambda e^{-\frac{t}{\lambda}})^n\end{aligned}$$

Since F is a finite measure, $\mathcal{L}F$ is analytical on $(0,\infty)$: $\mathcal{L}F(z) = \sum_{n=0}^{\infty} \frac{1}{n!}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathcal{L}F(z_0)(z-z_0)^n$ for all $z, z_0 > 0$. In particular,

$$\mathcal{L}F_\lambda(t) = \mathcal{L}F(\lambda - \lambda e^{-\frac{t}{\lambda}})$$

Since $1-e^{-\frac{t}{\lambda}} \stackrel{\lambda\to\infty}{\sim} \frac{t}{\lambda}$, $\lim_{\lambda\to\infty} \mathcal{L}F_\lambda(t) = \mathcal{L}F(t)$ for all $t > 0$. By Lévy's continuity theorem, this shows that $\lim_{\lambda\to\infty} F_\lambda = F$ vaguely. Since $F_\lambda(\mathbb{R}_+) = F_\lambda(\mathbb{R}_+)$, this convergence holds in distribution. □

PRESERVATION OF INFINITE DIVISIBILITY. Next proposition is also of major interest. Poisson mixture preserves convolution together with addition. This makes Poisson mixture stand out of other discretization methods.

Proposition B.75 (preservation of convolution)**.** For all probability measures F, G, and real number $c \in \mathbb{R}$, then

$$\Gamma_\lambda[\mathrm{d}(F + cG)] = \Gamma_\lambda[\mathrm{d}F] + c\Gamma_\lambda[\mathrm{d}G],$$

$$\Gamma_\lambda[\mathrm{d}(F * G)] = \Gamma_\lambda[\mathrm{d}F] * \Gamma_\lambda[\mathrm{d}G].$$

This simple result has two fundamental consequences. First, Poisson mixtures of preserves infinite divisibility. The relationship is all the more stronger as Poisson mixtures preserve many interesting subclasses of infinitely divisible distributions. This is summarized in the following scheme.

Γ_λ : probabilities on $\mathbb{R}_+$	$\longrightarrow$	discrete probabilities on $\mathbb{N}$
inf. div. on $\mathbb{R}_+$	$\longrightarrow$	discrete inf. div. on $\mathbb{N}$
self-decomposable on $\mathbb{R}_+$	$\longrightarrow$	discrete self-decomposable on $\mathbb{N}$
stable on $\mathbb{R}_+$	$\underset{\longleftarrow}{\longrightarrow}$	discrete stable on $\mathbb{N}$

All common discrete inf. div. distributions are actually Poisson mixture. Nevertheless, none of the mappings are surjective (except the last one). This means there exists discrete inf. div. distributions which are *not* Poisson mixtures. Next propositions give further explanations. A deeper discussion on the matter may be found in (Puri and Goldie, 1979) and Steutel and van Harn (2004).

Proposition B.76. Let F be a probability measure on $\mathbb{R}_+$.
F is infinitely divisible on $\mathbb{R}_+$ if and only if $\Gamma_\lambda[\mathrm{d}F]$ is discrete infinitely divisible on $\mathbb{N}$ for all $\lambda > 0$.
F is self-decomposable on $\mathbb{R}_+$ if and only if $\Gamma_\lambda[\mathrm{d}F]$ is discrete self-decomposable on $\mathbb{N}$ for all $\lambda > 0$.

More specifically, Poisson mixture maps canonical measures to canonical sequences and absorbs the drift component. Note that this relationship works with canonical measures but on Lévy measures. Indeed, the Lévy measure of an infinitely divisible Poisson mixture is not necessarily a Poisson mixture, as noticed in (Puri and Goldie, 1979, Remark 1). In addition, the Poisson mixture may be not defined on infinite Lévy measures.

Proposition B.77. Let F be an infinitely divisible on $\mathbb{R}_+$, K denote the canonical measure of F and $a \geq 0$ its drift component.
$\Gamma_\lambda[\mathrm{d}F]$ is discrete infinitely divisible and its canonical sequence is $\Gamma_\lambda[\mathrm{d}K] + (\lambda a)\delta_0$.
Reciprocally, let D be a discrete infinitely divisible distribution supported on $\mathbb{N}$. Then D is a Poisson mixture if and only if its canonical sequence r is a Poisson mixture.

A second approximation is that Poisson mixtures preserve the class of Lévy processes or additive processes. Next propositions is about random variables, but similar results hold for convolution semigroups of distributions. In addition, the same result holds for delayed Lévy and delayed additive processes.

Proposition B.78. Let $X = (X_l)_{l\geq 0}$ be a process and $X^\lambda = (X_l^\lambda)_{\geq 0}$ be as described in Proposition B.74.
If X is a nonnegative Lévy process, then X^λ is a discrete Lévy process.
If X is an nonnegative additive process, then X^λ is a discrete additive process.

As a conclusion, any inf. div. distribution on $\mathbb{R}_+$ can be approximated by a sequence of discrete inf. div. distributions. Any nonnegative Lévy process can be approximated by discrete Lévy processes. This provides a powerful tool that allows transferring results obtained on discrete distributions to continuous ones.

EXAMPLES OF POISSON MIXTURES.

- Poisson mixture maps constant functions $c \in \mathbb{R}$ to constant sequences, $\Gamma_\lambda : c \mapsto c/\lambda$
- Poisson mixture maps exponential distributions $\mathcal{E}(\nu)$ to geometric distributions $\mathcal{G}\left(\frac{\lambda}{\lambda+\nu}\right)$, for $\nu > -\lambda$, $\Gamma_\lambda : \nu e^{-\nu x} \mapsto \frac{\nu}{\lambda+\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^n$.
- Poisson mixture maps deterministic distributions δ_a $(a \geq 0)$ to Poisson distributions, $\Gamma_\lambda : \delta_a \mapsto Po(\lambda a)$
- In particular, Poisson mixture preserves δ_0, $\Gamma_\lambda : \delta_0 \mapsto \delta_0$

B.5.2 POISSON MIXTURE AND TOTAL POSITIVITY

Preservation of TP_2 properties Poisson mixtures preserves TP_2 distributional properties, as well as TP_2 stochastic orders. Results are stated for general measures but also work with functions.

Proposition B.79 (Preservation of TP_2 properties)**.** Let F be a nonnegative measure on $[0, \infty)$. Let λ_0 be any real number such that $\lambda_0 > \max(\mathrm{abs}[F], 0)$.

F is log-concave $\iff$ $\Gamma_\lambda[\mathrm{d}F]$ is discrete log-concave for all $\lambda > \lambda_0$.
F is IHR $\iff$ $\Gamma_\lambda[\mathrm{d}F]$ is discrete IHR for all $\lambda > \lambda_0$.
F is DRHR $\iff$ $\Gamma_\lambda[\mathrm{d}F]$ is discrete DRHR for all $\lambda > \lambda_0$.

The log-concavity part is obtained by Block and Savits (1980, Corollary 3.7) under the assumption that F admits a continuous density on $(0, \infty)$. The IHR part is obtained by Vinogradov (1974). The DRHR part is obtained by Nanda and Sengupta (2005, Theorem 3.3).

Proposition B.80 (Preservation of TP_2 orders)**.** Let $F, G : [0,\infty) \to \mathbb{R}$ be two functions. Denote $\lambda_0 := \max(\mathrm{abs}[F], \mathrm{abs}[G])$ and assume $\lambda_0 < \infty$.

$f \leqslant g \implies \Gamma_\lambda[f] \leqslant \Gamma_\lambda[g]$ for all $\lambda > \lambda_0$.

$f \underset{\mathrm{DP}}{\leqslant} g \implies \Gamma_\lambda[f] \underset{\mathrm{DP}}{\leqslant} \Gamma_\lambda[g]$ for all $\lambda > \lambda_0$.

$f \underset{\mathrm{TP}}{\leqslant} g \implies \Gamma_\lambda[f] \underset{\mathrm{TP}}{\leqslant} \Gamma_\lambda[g]$ for all $\lambda > \lambda_0$.

Let F, G be two finite nonnegative distributions on $[0,\infty)$. For the following stochastic orders $C \in \{\mathrm{st}, \mathrm{lr}, \mathrm{hr}, \mathrm{rh}\}$,

$F \underset{C}{\leqslant} G \iff \Gamma_\lambda[\mathrm{d}F] \underset{C}{\leqslant} \Gamma_\lambda[\mathrm{d}G]$ for all $\lambda > \lambda_0$.

Proof. [$\leqslant$ order] It follows from the linearity and positivity of Γ_λ operators.

[DP order] Let $n \in \mathbb{N}$. $\Gamma_\lambda[f](n+1)\Gamma_\lambda[g](n) \leq \Gamma_\lambda[f](n)\Gamma_\lambda[g](n+1)$ is equivalent to

$$\int_{\mathbb{R}_+} t^{n+1}e^{-t}f(t)\mathrm{d}t \int_{\mathbb{R}_+} t^n e^{-t}g(t)\mathrm{d}t \leq \int_{\mathbb{R}_+} t^n e^{-t}f(t)\mathrm{d}t \int_{\mathbb{R}_+} t^{n+1}e^{-t}g(t)\mathrm{d}t.$$

Let us denote $g_1(t) = t^{n+1}e^{-t}$, $g_2(t) = t^n e^{-t}$, $\beta = f$, $\alpha = g$. We have $g_1 \underset{\mathrm{TP}}{\leqslant} g_2$ since $g_2(t)/g_1(t) = t$ which is increasing. If $f \underset{\mathrm{DP}}{\leqslant} g$, then $f \geqslant 0$ so the first claim gives $\Gamma_\lambda[f](n) \geq 0$. Furthermore, Proposition B.44 gives the inequality above for all $n \in \mathbb{N}$. This inequality implies $\Gamma_\lambda[f] \underset{\mathrm{DP}}{\leqslant} \Gamma_\lambda[g]$.

[TP order] If $f \underset{\mathrm{TP}}{\leqslant} g$, then second claim gives $f \underset{\mathrm{DP}}{\leqslant} g$ In addition $g \geqslant 0$, so $\Gamma_\lambda[g](n) \geq 0$. Therefore, $f \underset{\mathrm{TP}}{\leqslant} g$.

[st order, $\implies$] $F \underset{\mathrm{st}}{\leqslant} G$ is equivalent to $F(.) \geqslant G(.)$. Let $\lambda > \lambda_0$. First claim implies $\Gamma_\lambda[F](.) \geqslant \Gamma_\lambda[G](.)$. Furthermore, $\lambda\Gamma_\lambda[F], \lambda\Gamma_\lambda[G]$ are respective cdfs of $\Gamma_\lambda[\mathrm{d}F], \Gamma_\lambda[\mathrm{d}G]$. As $\lambda \neq 0$, one obtains $\Gamma_\lambda[\mathrm{d}F] \leqslant \Gamma_\lambda[\mathrm{d}G]$.

[lr order, $\implies$] Assume $F \underset{\mathrm{lr}}{\overset{\mathrm{st}}{\leqslant}} G$. Similarly to the TP order, $\Gamma_\lambda[\mathrm{d}F] \underset{\mathrm{TP}}{\leqslant} \Gamma_\lambda[\mathrm{d}G]$ is equivalent to

$$\int_{\mathbb{R}_+} t^{n+1}e^{-t}\mathrm{d}F(t) \int_{\mathbb{R}_+} t^n e^{-t}\mathrm{d}G(t) \leq \int_{\mathbb{R}_+} t^n e^{-t}f(t)\mathrm{d}F(t) \int_{\mathbb{R}_+} t^{n+1}e^{-t}\mathrm{d}G(t).$$

Again, Proposition B.44 gives this inequality above for all $n \in \mathbb{N}$.

[rh order, $\implies$] $F \underset{\mathrm{rh}}{\leqslant} G$ is equivalent to $F \underset{\mathrm{TP}}{\leqslant} G$. Similarly, the TP claim gives $\lambda\Gamma_\lambda[F] \underset{\mathrm{TP}}{\leqslant} \lambda\Gamma_\lambda[G]$, which means $\Gamma_\lambda[\mathrm{d}F] \underset{\mathrm{rh}}{\leqslant} \Gamma_\lambda[\mathrm{d}G]$.

[hr order, $\implies$] $F \underset{\mathrm{hr}}{\leqslant} G$ is equivalent to $\overline{F} \underset{\mathrm{TP}}{\leqslant} \overline{G}$. Similarly, the TP claim gives $\Gamma_\lambda[\overline{F}] \underset{\mathrm{TP}}{\leqslant} \Gamma_\lambda[\mathrm{d}G]$. Similarly, $\lambda\Gamma_\lambda[\overline{F}], \lambda\Gamma_\lambda[\overline{G}]$ is the survivor distribution of $\Gamma_\lambda[\mathrm{d}F], \Gamma_\lambda[\mathrm{d}G]$ respectively (with the alternative definition of the survivor distribution: $\overline{D} = 1 - D$). This suffices to show that $\Gamma_\lambda[\mathrm{d}F] \underset{\mathrm{hr}}{\leqslant} \Gamma_\lambda[\mathrm{d}G]$ for all $\lambda > 0$.

[Reciprocals] Assume $\Gamma_\lambda[\mathrm{d}F] \underset{C}{\leqslant} \Gamma_\lambda[\mathrm{d}G]$ holds for all $\lambda > 0$. Define F^λ, G^λ as in Proposition B.74. Then, $F^\lambda \underset{C}{\leqslant} G^\lambda$ and $\lim_{\lambda\to\infty} F^\lambda = F$, $\lim_{\lambda\to\infty} G^\lambda = G$ in distribution. Since such convergence preserves every stochastic order C (se Proposition B.30), this implies $F \underset{C}{\leqslant} G$. □

Preservation of unimodality and related Poisson mixtures preserve additional properties related to unimodality. Next proposition sums up three preservation results that exists in the literature and states an original result (claim (iv)).

Proposition B.81 (Preservation of distributional properties)**.** Let F be a nonnegative measure on $\mathbb{R}_+$ such that $\mathrm{abs}[F] < \infty$. Let λ_0 be any real number such that $\lambda_0 > \max(\mathrm{abs}[F], 0)$.

(i) F is unimodal if and only if $\Gamma_\lambda[\mathrm{d}F]$ is discrete unimodal for all $\lambda > \lambda_0$. (Holgate, 1970)

(ii) F is non-increasing if and only if $\Gamma_\lambda[\mathrm{d}F]$ is discrete non-increasing for all $\lambda > \lambda_0$. (Forst, 1979, Lemma 1)

(iii) F is lower-half log-concave if and only if $\Gamma_\lambda[\mathrm{d}F]$ is discrete lower-half log-concave for all $\lambda > \lambda_0$. (Watanabe, 1992, Lemma 3.2)

(iv) F is upper-half log-concave if and only if $\Gamma_\lambda[\mathrm{d}F]$ is discrete upper-half log-concave for all $\lambda > \lambda_0$.

We give an alternative proof for claim (iii) based on exponential tilting $\mathcal{E}_a$. Indeed, we have achieved in Proposition B.59 to characterize half log-concavity with unimodality preservation by $\mathcal{E}_a$ operators. This method also proves claim (iv) in the same manner.

Proof. Recall that $\mathcal{E}_\lambda$ is defined for all $a \in \mathbb{R}$ by $\mathcal{E}_a F(\mathrm{d}t) = e^{at} F(\mathrm{d}t)$. From Proposition B.59, we know that
F is (resp. discrete) upper-half log-concave if and only if $\mathcal{E}_a F$ is (resp. discrete) unimodal for all $a > 0$.
From proposition B.85, we know that

$$\forall \lambda, a \in \mathbb{R}_+, \quad a > \lambda > \lambda_0 \implies \Gamma_\lambda[\mathrm{d}(\mathcal{E}_a F)] = \mathcal{E}_{\log \frac{\lambda}{\lambda - a}} \Gamma_{\lambda - a}[\mathrm{d}F].$$

Let $\tilde{\lambda} > 0$, $\tilde{a} \in \mathbb{R}$. Then,

$$\begin{cases} \tilde{\lambda} = \lambda - a \\ \tilde{a} = \log \lambda/(\lambda - a) \end{cases} \iff \begin{cases} \lambda = e^{\tilde{a}} \tilde{\lambda} \\ a = \tilde{\lambda}(e^{\tilde{a}} - 1), \end{cases} \tag{B.6}$$

and it is immediate to see that $a \leq 0$ if and only if $\tilde{a} \leq 0$, and $\tilde{\lambda} > 0$ if and only if $\lambda > a$.

[lower, $\impliedby$] Fix $a \leq 0$. For all $\lambda > \max(\lambda_0, a)$, $\Gamma_{\lambda - a}[\mathrm{d}F]$ is discrete lower-half log-concave. In addition, $\log \lambda/(\lambda - a) \leq 0$. Owing to Proposition B.59, $\mathcal{E}_{\log \lambda/(\lambda-a)} \Gamma_{\lambda-a}[\mathrm{d}F]$ is discrete unimodal. So we have discrete unimodality of $\Gamma_\lambda[\mathrm{d}(\mathcal{E}_a F)]$ for all $\lambda > \max(\lambda_0, a)$. Owing to claim (i), $\mathcal{E}_a F$ is unimodal. Since this reasoning holds for any $a \leq 0$, this characterizes F being lower-half log-concave.

[lower, $\implies$] Fix $\tilde{\lambda} > \lambda_0$. Fix $\tilde{a} \leq 0$. Take (a, λ) as in equation (B.6). Then, as $a \leq 0$, $\mathcal{E}_a F$ is unimodal. Owing to claim (i), $\Gamma_\lambda[\mathrm{d}(\mathcal{E}_a F)]$ is discrete unimodal. As it equals $\mathcal{E}_{\tilde{a}} \Gamma_{\tilde{\lambda}}[\mathrm{d}F]$, the latter is discrete unimodal. Since this reasoning holds for any $\tilde{a} \leq 0$, this characterizes $\Gamma_{\tilde{\lambda}}[\mathrm{d}F]$ being discrete lower-half log-concave.

[upper] The proof is similar by reversing $a, \tilde{a} \leq 0$, with $a, \tilde{a} \geq 0$. □

B.5.3 NEW PRESERVATION PROPERTIES

We end with a few related results we have not found in the literature. First, Poisson mixtures exchanges scaling S_a with binomial thinning B_a, proving the latter is truly the discrete counterpart of the former.

Proposition B.82. If S_λ denotes the scaling operator $S_\lambda : D(\cdot) \mapsto D(\lambda\cdot)$, then

$$\forall \lambda > 0, \qquad \Gamma_\lambda = \Gamma_1 \circ S_\lambda.$$

Poisson mixture operators map (continuous) scaling S_a to (discrete) binomial thinning B_a,

$$\forall a \in (0,1], b > 0, \qquad B_a \circ \Gamma_b = \Gamma_{ab} = \Gamma_b \circ S_a.$$

Proof. Let P denote the Z-transform of a measure $\Gamma_1[\mathrm{d}F]$. Let L, L_a denote the Laplace transforms of F, $S_a F$. Then, $L_a(s) = L(sa)$. The Z-transform of $B_a\Gamma_1[\mathrm{d}F]$ is $P(az + 1 - a) = L(1 - (az + 1 - a)) = L(a(1 - z)) = L_a(1 - z)$, which is the Z-transform of $\Gamma_1[\mathrm{d}(S_a F)]$. □

Second, it is known that Poisson mixtures exchange log-concavity, IHR, DRHR properties with their discrete counterparts. We prove this preservation still holds for the multiplicative properties MSU, M-IHR, M-DRHR.

Proposition B.83. Let F be a finite nonnegative measure on $[0, \infty)$.

F is MSU $\iff$ $\Gamma_\lambda[\mathrm{d}F]$ is discrete MSU for all $\lambda > 0$.

F is M-IHR $\iff$ $\Gamma_\lambda[\mathrm{d}F]$ is discrete M-IHR for all $\lambda > 0$.

F is DRHR $\iff$ $\Gamma_\lambda[\mathrm{d}F]$ is discrete M-DRHR for all $\lambda > 0$.

Proof. The proof is a mere combination of Proposition B.82 with Proposition B.62 which characterizes MSU properties using stochastic orders.

[MSU] We show that both propositions are equivalent to $(\Gamma_a[\mathrm{d}F])_{a>0}$ being nondecreasing in lr order (denoted $\uparrow lr$).

$$F\text{is MSU} \iff (S_a F)_{a>0} \uparrow lr,$$

as Poisson mixtures preserve stochastic orders (Proposition B.80),

$$\iff \forall b > 0, \quad (\Gamma_b[\mathrm{d}(S_a F)])_{a>0} \uparrow lr,$$

and as $\Gamma_b[\mathrm{d}(S_a F)] = \Gamma_{ba}[\mathrm{d}F]$,

$$\iff (\Gamma_a[\mathrm{d}F])_{a>0} \uparrow lr.$$

Let p_λ denote the pmf $p_\lambda(n) := \Gamma_\lambda[\mathrm{d}F](n)$. Then,

$$(B_a p)_{a\in(0,1]} = (\Gamma_a[\mathrm{d}F])_{a\in(0,\lambda]}.$$

By Proposition B.67,

$$\begin{aligned} \forall \lambda > 0, p_\lambda \text{ is MSU} \quad &\iff \quad \forall \lambda > 0, (B_a p_\lambda)_{a\in(0,1]} \uparrow lr \\ &\iff \quad \forall \lambda > 0, (\Gamma_a[\mathrm{d}F])_{a\in(0,\lambda]} \uparrow lr \\ &\iff \quad (\Gamma_a[\mathrm{d}F])_{a>0} \uparrow lr. \end{aligned}$$

[M-DRHR, M-IHR] Replace the lr order by the rh or hr ones. □

Now, we study two examples of measure weightings that are also preserved by Poisson mixtures: weighting by exponentials functions and by linear functions.

Definition B.84. The *size-biasing* operator M is defined

- on measures F on $\mathbb{R}$ by $MF(\mathrm{d}x) \stackrel{\text{def}}{=} xF(\mathrm{d}x)$,
- on discrete sequence d on $\mathbb{N}$ by $Md(n) \stackrel{\text{def}}{=} (n+1)d(n+1)$.

Exponential tilting operators $\mathcal{E}_a$, for all $a \in \mathbb{R}$, are defined on measures by $\mathcal{E}_a F(\mathrm{d}x) \stackrel{\text{def}}{=} e^{ax}F(\mathrm{d}x)$ and on sequences similarly.

Remark. Note that size-biasing of sequences is *not* compliant with size-biasing of discrete .

Proposition B.85. Let F be a measure on $\mathbb{R}_+$ such that $\mathrm{abs}[F] < \infty$.

(i) $\mathrm{abs}[MF] \leq \mathrm{abs}[F]$ and for all $\lambda > \mathrm{abs}[F]$,

$$\begin{aligned}\Gamma_\lambda[\mathrm{d}(MF)](n) &= \lambda^{-1}(n+1)\Gamma_\lambda[\mathrm{d}F](n+1) \quad \text{for all } n \in \mathbb{N},\\ \Gamma_\lambda[\mathrm{d}(MF)] &= \lambda^{-1}M\left[\Gamma_\lambda[\mathrm{d}F]\right],\\ \Gamma_1 \circ M &= M \circ \Gamma_1.\end{aligned}$$

(ii) $\mathrm{abs}[\mathcal{E}_a F] = \mathrm{abs}[F] - a$ and for all $\lambda > \max(0, \mathrm{abs}[F]) + a$,

$$\begin{aligned}\Gamma_\lambda[\mathrm{d}(\mathcal{E}_a F)](n) &= \left(\frac{\lambda}{\lambda - a}\right)^n \Gamma_{\lambda-a}[\mathrm{d}F](n) \quad \text{for all } n \in \mathbb{N},\\ \Gamma_\lambda[\mathrm{d}(\mathcal{E}_a F)] &= \mathcal{E}_{\log\frac{\lambda}{\lambda-a}}\Gamma_{\lambda-a}[\mathrm{d}F](n),\\ \Gamma_\lambda \circ \mathcal{E}_a F &= \mathcal{E}_{\log\frac{\lambda}{\lambda-a}} \circ \Gamma_{\lambda-a}.\end{aligned}$$

Proof. [(i)] By definition of Γ_λ, for all $n \in \mathbb{N}$,

$$\begin{aligned}\Gamma_\lambda[\mathrm{d}(MF)](n) &= \int_{\mathbb{R}_+} e^{-\lambda}\frac{1}{n!}\lambda^n x^n x \mathrm{d}F(x)\\ &= \int_{\mathbb{R}_+} e^{-\lambda}\frac{n+1}{(n+1)!}\lambda^{-1}\lambda^{n+1}x^{n+1}\mathrm{d}F(x)\\ &= \lambda^{-1}(n+1)\int_{\mathbb{R}_+} e^{-\lambda}\frac{1}{(n+1)!}\lambda^{n+1}x^{n+1}\mathrm{d}F(x)\\ &= \lambda^{-1}(n+1)\Gamma_\lambda[\mathrm{d}F](n+1).\end{aligned}$$

[(ii)] Consider $a \in \mathbb{R}$ and $\lambda > \mathrm{abs}[F] + a$. Define $G = \mathcal{E}_a F$. Then

$$\begin{aligned}\Gamma_\lambda[\mathrm{d}G](n) &= \int \lambda^n e^{-\lambda t}\frac{t^n}{n!}\mathrm{d}G(t)\\ &= \int \lambda^n e^{-(\lambda-a)t}\frac{t^n}{n!}\mathrm{d}F(t)\\ &= \left(\frac{\lambda}{\lambda-a}\right)^n \int (\lambda-a)^n e^{-(\lambda-a)t}\frac{t^n}{n!}\mathrm{d}F(t)\\ &= \left(\frac{\lambda}{\lambda-a}\right)^n \Gamma_{\lambda-a}[f](n).\end{aligned}$$

□

Last result justifies our definition of size-biasing M. To ensure Poisson mixtures preserve size-biasing, this operator has to be defined differently for discrete sequences: $Mf(n) = (n+1)f(n+1)$ instead of $nf(n)$.

Interestingly, last result has an immediate consequence about the shape of Poisson mixtures: any of them must be "less concave" than a Poisson distribution. Refer to chapter 8 where log-convexity is introduced in Definition 8.12 and relative concavity is explained. Next proposition has been stated for absolutely continuous mixing measures by Misra et al. (2003, Lemma 3.1(a)) and Yu (2009b, Section 2). Just by combining previous results, we obtain an alternative proof that does not make such assumptions.

Proposition B.86. Let F be a nonnegative measure on $\mathbb{R}_+$. Assume that F is not proportional to δ_0. Assume $\lambda_0 := \max(\mathrm{abs}[F], 0) < \infty$.

For all $\lambda > \lambda_0$, the sequence $n!\Gamma_\lambda[\mathrm{d}F](n)$ is discrete log-convex on $\mathbb{N}$. In other words, $\Gamma_\lambda(\mathrm{d}F)$ is less concave than Poisson distributions

Proof. Define $p(n) := \Gamma_\lambda[\mathrm{d}F](n)$. Since $F \not\propto \delta_0$, by Proposition B.72, $\mathrm{supp}[p] = \mathbb{N}$.

By definition, $MF(\mathrm{d}x) = F(\mathrm{d}x)x$, so $\mathrm{d}F$ is dominated by F and admits a density $\frac{\mathrm{d}MF}{\mathrm{d}F}(x) = x$ that is increasing on $\mathbb{R}_+$. This proves that

$$F \underset{\mathrm{lr}}{\leqslant} MF.$$

Owing to Proposition B.80, Poisson mixtures preserve lr order. So

$$\Gamma_\lambda[\mathrm{d}F] \underset{\mathrm{lr}}{\leqslant} \Gamma_\lambda[\mathrm{d}MF].$$

Since $\Gamma_\lambda[\mathrm{d}MF](n) = (n+1)\Gamma_\lambda[\mathrm{d}F](n+1)$ and $\mathrm{supp}[p] = \mathbb{N}$, $(n+1)p(n+1)/p(n)$ is non-decreasing on $\mathbb{N}$. As this ratio equals $(n+1)!p(n+1)/(n!p(n))$ and $q(n) := n!p(n) > 0$ for all $n \in \mathbb{N}$, this means the sequence q is discrete log-convex. □

This result is interesting as it mirrors a another one on infinite divisible distributions: if such a distribution is discrete log-concave, then it is less concave than any Poisson distribution.

BIBLIOGRAPHY

Alvarez, J., Amadis, M., Boros, G., Karp, D., Moll, V. H., and Rosales, L. (2001). An extension of a criterion for unimodality. *The Electronic Journal of Combinatorics [electronic only]*, 8(1):R30. (Cited on pages 155 and 156.)

Alvarez, J., Amadis, M., and Rosales, L. (2000). Unimodality and log-concavity of polynomials. Technical report, The Summer Institute in Mathematics for Undergraduates (SIMU). (Cited on page 161.)

Arzt, A. (2008). Score following with dynamic time warping: an automatic page-turner. Master's thesis, Vienna University of Technology, Vienna, Austria. (Cited on page 27.)

Arzt, A. and Widmer, G. (2010a). Simple tempo models for real-time music tracking. In Serra, X., editor, *Proceedings of the Seventh Sound and Music Computing Conference (SMC 2010), Barcelona, Spain, July 21–24, 2010*. (Cited on page 33.)

Arzt, A. and Widmer, G. (2010b). Towards effective 'any-time' music tracking. In Ågotnes, T., editor, *Proceedings of the Fifth Starting AI Researchers' Symposium (STAIRS 2010), Lisbon, Portugal, August, 16–20, 2010*, volume 222 of *Frontiers in Artificial Intelligence and Applications*, pages 24–36. IOS Press. (Cited on page 25.)

Arzt, A., Widmer, G., and Dixon, S. (2008). Automatic page turning for musicians via real-time machine listening. In Ghallab, M., Spyropoulos, C. D., Fakotakis, N., and Avouris, N. M., editors, *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI 2008), Patras, Greece, July 21–25, 2008*, volume 178 of *Frontiers in Artificial Intelligence and Applications*, pages 241–245. IOS Press. (Cited on page 27.)

Banciu, M. and Mirchandani, P. (2013). Technical note – New results concerning probability distributions with increasing generalized failure rates. *Operations Research*, 61(4):925–931. (Cited on page 234.)

Basu, S. and DasGupta, A. (1997). The mean, median, and mode of unimodal distributions: A characterization. *Theory of Probability & Its Applications*, 41(2):210–223. (Cited on page 119.)

Bauschke, H. H. and Combettes, P. L. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, NY, USA. (Cited on pages 63 and 66.)

Bertin, E. M. J., Cuculescu, I., and Theodorescu, R. (1997). *Unimodality of Probability Measures*. Mathematics and its Applications. Springer Netherlands. (Cited on page 227.)

Bickel, P. J. and Lehmann, E. L. (1975). Descriptive statistics for nonparametric models II. Location. *The Annals of Statistics*, 3(5):pp. 1045–1069. (Cited on page 226.)

Bilmes, J. A. (1993). Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, USA. (Cited on page 50.)

Bloch-Mercier, S. (2001). Monotone Markov processes with respect to the reversed hazard rate ordering: An application to reliability. *Journal of Applied Probability*, 38(1):pp. 195–208. (Cited on pages 130 and 133.)

Block, H. W. and Savits, T. H. (1980). Laplace transforms for classes of life distributions. *The Annals of Probability*, 8(3):pp. 465–474. (Cited on page 243.)

Bondesson, L. (1992). *Generalized Gamma Convolutions and Related Classes of Distributions and Densities*. Lecture Notes in Statistics. Springer, New York, NY, USA. (Cited on page 176.)

Boros, G. and Moll, V. H. (1999). A criterion for unimodality. *The Electronic Journal of Combinatorics [electronic only]*, 6(1):R10. (Cited on pages 155 and 158.)

Brenti, F. (1989). *Unimodal, Log-Concave and Pólya Frequency Sequences in Combinatorics*. Memoirs of the AMS Series. American Mathematical Society. (Cited on pages 149, 155, 162, and 205.)

Brown, M. and Chaganty, N. R. (1983). On the first passage time distribution for a class of Markov chains. *The Annals of Probability*, 11(4):1000–1008. (Cited on page 130.)

Cai, J. and Willmot, G. E. (2005). Monotonicity and aging properties of random sums. *Statistics & Probability Letters*, 73(4):381–392. (Cited on page 147.)

Cano, P., Loscos, A., and Bonada, J. (1999). Score-performance matching using HMMs. In *Proceedings of the 1999 International Computer Music Conference (ICMC 1999), Beijing, China, October 22–27, 1999*, pages 441–444. Michigan Publishing. (Cited on pages 28 and 35.)

Cappé, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer-Verlag, New York, NY, USA. (Cited on page 34.)

Chen, W. Y., Yang, A. L., and Zhou, E. L. (2010). Ratio monotonicity of polynomials derived from nondecreasing sequences. *The Electronic Journal of Combinatorics [electronic only]*, 17(1):R37. (Cited on pages 155 and 156.)

Chen, W. Y. C. and Gu, C. C. Y. (2009). The reverse ultra log-concavity of the Boros-Moll polynomials. *Proceedings of the American Mathematical Society*, 137(12):3991–3998. (Cited on page 169.)

Cont, A. (2006). Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMs. In Duhamel, P. and Vandendorpe, L., editors, *Proceedings of the IEEE International Conference on Acoustics*

Speech and Signal Processing (ICASSP 2006), Toulouse, France, May 14–19, 2006, pages 245–248. IEEE. (Cited on page 35.)

Cont, A. (2008). Antescofo: Anticipatory synchronization and control of interactive parameters in computer music. In *Proceedings of the 2008 International Computer Music Conference (ICMC 2008), Belfast, Ireland, August 24–29, 2008*, pages 33–40. Michigan Publishing. (Cited on pages 25 and 125.)

Cont, A. (2010). A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):974–987. (Cited on pages 24, 26, 28, 30, 33, 37, 120, and 125.)

Cont, A., Schwarz, D., Schnell, N., and Raphael, C. (2007). Evaluation of real-time audio-to-score alignment. In Dixon, S., Bainbridge, D., and Typke, R., editors, *Proceedings of the 8th International Society for Music Information Retrieval (ISMIR 2007), Vienna, Austria, September 23–27, 2007*, pages 315–316. Austrian Computer Society. (Cited on page 121.)

Cont, R. and Tankov, P. (2004). *Financial Modelling with Jump Processes*. Financial Mathematics Series. Chapman & Hall/CRC Press, Boca Raton, FL, USA, 1st edition. (Cited on pages 193 and 199.)

Cuculescu, I. and Theodorescu, R. (1998). Multiplicative strong unimodality. *Australian & New Zealand Journal of Statistics*, 40(2):205–214. (Cited on page 233.)

Cuvillier, P. (2014). Time-coherency of Bayesian priors on transient semi-Markov chains for audio-to-score alignment. In Mohammad-Djafari, A. and Barbaresco, F., editors, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014), Amboise, France, 21–26 September, 2014*, volume 1641 of *AIP Conference Proceedings*, pages 255–262. AIP. (Cited on page 44.)

Cuvillier, P. and Cont, A. (2014). Coherent time modeling of semi-Markov models with application to real-time audio-to-score alignment. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2014), Reims, France, September 21–24, 2014*. IEEE. (Cited on page 44.)

Dannenberg, R. B. (1984). An on-line algorithm for real-time accompaniment. In Wessel, D., editor, *Proceedings of the 1984 International Computer Music Conference (ICMC 1984), Paris, France, October 19–23, 1984*, pages 193–198. Michigan Publishing. (Cited on page 27.)

Davenport, H. and Pólya, G. (1949). On the product of two power series. *Canadian Journal of Mathematics*, 1(1):1–5. (Cited on page 163.)

Devaney, J. and Ellis, D. P. W. (2009). Handling asynchrony in audio-score alignment. In Scavone, G., Verfaille, V., and da Silva, A., editors, *Proceedings of the 2009 International Computer Music Conference (ICMC 2009), Montreal, Quebec, Canada, August 16–21, 2009*, pages 29–32. Michigan Publishing. (Cited on page 27.)

Dixon, S. (2005). An on-line time warping algorithm for tracking musical performances. In Kaelbling, L. P. and Saffiotti, A., editors, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05), Edinburgh, Scotland, UK, July 30 – August 5, 2005*, pages 1727–1728. Morgan Kaufmann. (Cited on page 27.)

Duan, Z. and Pardo, B. (2011a). Soundprism: An online system for score-informed source separation of music audio. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1205–1215. (Cited on page 39.)

Duan, Z. and Pardo, B. (2011b). A state space model for online polyphonic audio-score alignment. In Tichavsky, P. and Chambers, J. A., editors, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), Prague, Czech Republic, May 22–27, 2011*, pages 197–200. IEEE. (Cited on pages 28, 32, and 39.)

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. (Cited on page 36.)

Esscher, F. (1932). On the probability function in the collective theory of risk. *Scandinavian Actuarial Journal*, 15(3):175–195. (Cited on pages 92, 157, and 231.)

Finner, H. and Roters, M. (1997). Log-concavity and inequalities for chi-square, F and beta distributions with applications in multiple comparisons. *Statistica Sinica*, 7(3):771–787. (Cited on page 154.)

Forst, G. (1979). A characterization of self-decomposable probabilities on the half-line. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 49(3):349–352. (Cited on page 245.)

Fristedt, B. E. and Gray, L. F. (1997). *A Modern Approach to Probability Theory*. Birkhäuser, Boston, MA, USA. (Cited on page 240.)

Gnedenko, B. V. and Kolmogorov, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables*. Translated from the Russian. Addison-Wesley. (Cited on page 92.)

Gong, R., Cuvillier, P., Obin, N., and Cont, A. (2015). Real-time audio-to-score alignment of singing voice based on melody and lyric information. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015), Dresden, Germany, September 6–10, 2015*, pages 3312–3316. ISCA. (Cited on pages 25 and 123.)

Grubb, L. and Dannenberg, R. B. (1994). Automated accompaniment of musical ensembles. In Hayes-Roth, B. and Korf, R. E., editors, *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94), Seattle, WA, USA, July 31 – August 4, 1994*, volume 1, pages 94–99. AAAI Press / The MIT Press. (Cited on page 28.)

Guédon, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12(3):604–639. (Cited on pages 37, 53, 190, 191, and 192.)

Guédon, Y. (2005). Hidden hybrid Markov/semi-Markov chains. *Computational Statistics and Data Analysis*, 49(3):663–688. (Cited on pages 37 and 53.)

Guédon, Y. and Cocozza-Thivent, C. (1990). Explicit state occupancy modelling by hidden semi-Markov models: application of Derin's scheme. *Computer Speech & Language*, 4(2):167–192. (Cited on page 41.)

Gut, A. (2005). *Probability: A Graduate Course*. Springer Texts in Statistics. Springer-Verlag, New York, NY, USA. (Cited on page 53.)

Hansen, B. G. (1988). On log-concave and log-convex infinitely divisible sequences and densities. *The Annals of Probability*, 16(4):1832–1839. (Cited on pages 143 and 175.)

Hoggar, S. G. (1974). Chromatic polynomials and logarithmic concavity. *Journal of Combinatorial Theory, Series B*, 16(3):248–254. (Cited on page 155.)

Holgate, P. (1970). The modality of some compound Poisson distributions. *Biometrika*, 57(3):666–667. (Cited on page 245.)

Ibragimov, I. A. (1956). On the composition of unimodal distributions. *Theory of Probability & Its Applications*, 1(2):255–260. (Cited on page 228.)

Joag-dev, K., Kochar, S., and Proschan, F. (1995). A general composition theorem and its applications to certain partial orderings of distributions. *Statistics & Probability Letters*, 22(2):111–119. (Cited on pages 201 and 226.)

Joder, C. (2011). *Alignement temporel musique-sur-partition par modèles graphiques discriminatifs* [Audio-to-Score Temporal Alignment with Discriminative Graphical Models]. PhD thesis, Télécom ParisTech, Paris, France. (Cited on pages 27, 28, 32, 44, 73, 117, and 123.)

Joder, C., Essid, S., and Richard, G. (2010a). A conditional random field viewpoint of symbolic audio-to-score matching. In Bimbo, A. D., Chang, S., and Smeulders, A. W. M., editors, *Proceedings of the 18th International Conference on Multimedia 2010 (MM 10), Firenze, Italy, October 25–29, 2010*, pages 871–874. ACM. (Cited on pages 28 and 178.)

Joder, C., Essid, S., and Richard, G. (2010b). An improved hierarchical approach for music-to-symbolic score alignment. In Downie, J. S. and Veltkamp, R. C., editors, *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010), Utrecht, Netherlands, August 9–13, 2010*, pages 39–45. (Cited on pages 42 and 120.)

Joder, C., Essid, S., and Richard, G. (2011). Hidden discrete tempo model: A tempo-aware timing model for audio-to-score alignment. In Tichavsky, P. and Chambers, J. A., editors, *Proceedings of the IEEE International Conference on Acoustics, Speech,*

and Signal Processing (ICASSP 2011), Prague, Czech Republic, May 22–27, 2011, pages 397–400. IEEE. (Cited on pages 31, 32, and 33.)

Johnson, N. L., Kotz, S., and Kemp, A. W. (1993). *Univariate Discrete Distributions*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition. (Cited on pages 120 and 199.)

Johnson, O. (2007). Log-concavity and the maximum entropy property of the Poisson distribution. *Stochastic Processes and their Applications*, 117(6):791–802. (Cited on pages 159 and 163.)

Johnson, O. and Goldschmidt, C. (2006). Preservation of log-concavity on summation. *ESAIM: Probability and Statistics*, 10:206–215. (Cited on pages 170, 171, and 172.)

Johnson, O., Kontoyiannis, I., and Madiman, M. (2013). Log-concavity, ultra-log-concavity, and a maximum entropy property of discrete compound Poisson measures. *Discrete Applied Mathematics*, 161(9):1232–1250. Jubilee Conference on Discrete Mathematics. (Cited on pages 143 and 163.)

Karlin, S. (1968). *Total positivity*, volume I. Stanford University Press, Stanford, CA, USA. (Cited on pages 158, 162, 202, 203, 204, 206, and 230.)

Keilson, J. (1971). Log-concavity and log-convexity in passage time densities of diffusion and birth-death processes. *Journal of Applied Probability*, 8(2):391–398. (Cited on page 130.)

Keilson, J. and Kester, A. (1978). Unimodality preservation in Markov chains. *Stochastic Processes and their Applications*, 7(2):179–190. (Cited on page 159.)

Keilson, J. and Sumita, U. (1982). Uniform stochastic ordering and related inequalities. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 10(3):181–198. (Cited on pages 131, 141, 142, and 219.)

Kijima, M. (1998). Hazard rate and reversed hazard rate monotonicities in continuous-time Markov chains. *Journal of Applied Probability*, 35(3):pp. 545–556. (Cited on pages 86, 130, 131, and 147.)

Klapuri, A., Eronen, A. J., and Astola, J. (2006). Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):342–355. (Cited on page 50.)

Korzeniowski, F., Krebs, F., Arzt, A., and Widmer, G. (2013). Tracking rests and tempo changes: Improved score following with particle filters. In *Proceedings of the 39th International Computer Music Conference (ICMC 2013), Perth, Australia, August 12–16, 2013*, pages 93–99. Michigan Publishing. (Cited on page 28.)

Korzeniowski, F. and Widmer, G. (2013). Refined spectral template models for score following. In Bresin, R., editor, *Proceedings of the Sound and Music Computing Conference 2013 (SMC 2013), Stockholm, Sweden, 30 July – 3 August, 2013*, pages 376–382. Logos Verlag. (Cited on page 32.)

Lajugie, R., Bojanowski, P., Cuvillier, P., Arlot, S., and Bach, F. R. (2016). A weakly-supervised discriminative model for audio-to-score alignment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), Shanghai, China, March 20–25, 2016*, pages 2484–2488. IEEE. (Cited on page 178.)

Lehmann, E. L. (1955). Ordered families of distributions. *The Annals of Mathematical Statistics*, 26(3):399–419. (Cited on page 211.)

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, NY, USA, 3rd edition. (Cited on page 223.)

Lember, J. and Koloydenko, A. A. (2014). Bridging Viterbi and posterior decoding: a generalized risk approach to hidden path inference based on hidden Markov models. *Journal of Machine Learning Research*, 15(1):1–58. (Cited on page 44.)

Leoni, G. (2009). *A First Course in Sobolev Spaces*. Graduate studies in mathematics. American Mathematical Society. (Cited on page 212.)

Levinson, S. E. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1(1):29–45. (Cited on page 41.)

Liggett, T. M. (1997). Ultra logconcave sequences and negative dependence. *Journal of Combinatorial Theory, Series A*, 79(2):315–325. (Cited on pages 163 and 168.)

Llamas, A. and Martínez-Bernal, J. (2010). Nested log-concavity. *Communications in Algebra*, 38(5):1968–1981. (Cited on pages 155 and 156.)

Lynch, J., Mimmack, G., and Proschan, F. (1987). Uniform stochastic orderings and total positivity. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 15(1):63–69. (Cited on pages 218, 219, and 220.)

Meerschaert, M. M. and Scheffler, H.-P. (2008). Triangular array limits for continuous time random walks. *Stochastic Processes and their Applications*, 118(9):1606–1633. (Cited on page 138.)

Mijatovic, A., Vidmar, M., and Jacka, S. (2014). Markov chain approximations for transition densities of Lévy processes. *Electronic Journal of Probability*, 19(7):1–37. (Cited on pages 146 and 147.)

Misra, N., Singh, H., and Harner, E. J. (2003). Stochastic comparisons of Poisson and binomial random variables with their mixtures. *Statistics & Probability Letters*, 65(4):279–290. (Cited on page 248.)

Mitchell, C. D. and Jamieson, L. H. (1993). Modeling duration in a hidden Markov model with the exponential family. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93), Minneapolis, Minnesota, USA, April 27–30, 1993*, volume 2, pages 331–334. (Cited on page 41.)

Mittag-Leffler, M. G. (1903). Sur la nouvelle fonction $E_\alpha(x)$. *Comptes Rendus de l'Académie des sciences de Paris*, 137:554–558. (Cited on page 153.)

Montecchio, N. and Cont, A. (2011). A unified approach to real time audio-to-score and audio-to-audio alignment using sequential Montecarlo inference techniques. In Tichavsky, P. and Chambers, J. A., editors, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), Prague, Czech Republic, May 22–27, 2011*, pages 193–196. IEEE. (Cited on pages 25, 28, 32, and 39.)

Montecchio, N. and Orio, N. (2009). A discrete filter bank approach to audio to score matching for polyphonic music. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009), Kobe, Japan, October 26–30, 2009*, pages 495–500. (Cited on pages 28, 32, 35, 36, 39, 44, and 120.)

Müller, M., Mattes, H., and Kurth, F. (2006). An efficient multiscale approach to audio synchronization. In *Proceedings of the 7th International Society for Music Information Retrieval (ISMIR 2006), Victoria, Canada, October 8–12, 2006*, pages 192–197. (Cited on page 27.)

Nakamura, T., Nakamura, E., and Sagayama, S. (2013). Acoustic score following to musical performance with errors and arbitrary repeats and skips for automatic accompaniment. In Bresin, R., editor, *Proceedings of the Sound and Music Computing Conference 2013 (SMC 2013), Stockholm, Sweden, 30 July – 3 August, 2013*, pages 299–304. Logos Verlag. (Cited on pages 25, 35, and 120.)

Nanda, A. K. and Sengupta, D. (2005). Discrete life distributions with decreasing reversed hazard. *Sankhyā: The Indian Journal of Statistics (2003-2007)*, 67(1):106–125. (Cited on page 243.)

Nelson, R. (1995). *Probability, Stochastic Processes, and Queueing Theory: The Mathematics of Computer Performance Modeling*. Springer-Verlag, New York, NY, USA. (Cited on pages 53 and 61.)

Orio, N. and Déchelle, F. (2001). Score following using spectral analysis and hidden Markov models. In Schloss, A., Dannenberg, R., and Driessen, P., editors, *Proceedings of the 2001 International Computer Music Conference (ICMC 2001), Havana, Cuba, September 17–22, 2001*, pages 151–154. Michigan Publishing. (Cited on pages 28, 35, and 44.)

Orio, N. and Schwarz, D. (2001). Alignment of monophonic and polyphonic music to a score. In Schloss, A., Dannenberg, R., and Driessen, P., editors, *Proceedings of the 2001 International Computer Music Conference (ICMC 2001), Havana, Cuba, September 17–22, 2001*, pages 129–132. Michigan Publishing. (Cited on page 27.)

Orio, N., Serge, L., Schwarz, D., and Schnell, N. (2003). Score following: State of the art and new developments. In Thibault, F., editor, *Proceedings of the 2003 International Conference on New Interfaces for Musical Expression (NIME-03), Montreal, Canada, May 22–24, 2003*, pages 36–41. Faculty of Music, McGill University. (Cited on page 27.)

Otsuka, T., Nakadai, K., Takahashi, T., Ogata, T., and Okuno, H. G. (2011). Real-time audio-to-score alignment using particle filter for coplayer music robots. *EURASIP Journal on Advances in Signal Processing*, 2011(2):1–13. (Cited on pages 28 and 39.)

Pardo, B. and Birmingham, W. P. (2005). Modeling form for on-line following of musical performances. In Veloso, M. M. and Kambhampati, S., editors, *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05) and the Seventeenth Innovative Applications of Artificial Intelligence Conference (IAAI-05), Pittsburgh, Pennsylvania, USA, July 9–13, 2005*, pages 1018–1023. AAAI Press / The MIT Press. (Cited on page 25.)

Park, C. and Padgett, W. J. (2005). Accelerated degradation models for failure based on geometric Brownian motion and gamma processes. *Lifetime Data Analysis*, 11(4):511–527. (Cited on page 152.)

Peeling, P. H., Cemgil, A. T., and Godsill, S. J. (2007). A probabilistic framework for matching music representations. In Dixon, S., Bainbridge, D., and Typke, R., editors, *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23-27, 2007*, pages 267–272. Austrian Computer Society. (Cited on page 32.)

Pemantle, R. (2000). Towards a theory of negative dependence. *Journal of Mathematical Physics*, 41(3):1371–1390. (Cited on pages 163 and 168.)

Pillai, R. N. (1990). On Mittag-Leffler functions and related distributions. *Annals of the Institute of Statistical Mathematics*, 42(1):157–161. (Cited on page 153.)

Powell, J. L. (1994). Estimation of semiparametric models. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 4, chapter 41, pages 2443–2521. Elsevier. (Cited on page 178.)

Puri, P. S. and Goldie, C. M. (1979). Poisson mixtures and quasi-infinite divisibility of distributions. *Journal of Applied Probability*, 16(1):138–153. (Cited on page 242.)

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286. (Cited on pages 41 and 187.)

Raphael, C. (1999). Automatic segmentation of acoustic musical signals using hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):360–370. (Cited on pages 28, 35, 36, 44, 73, and 117.)

Raphael, C. (2006). Aligning music audio with symbolic scores using a hybrid graphical model. *Machine Learning*, 65(2):389–409. (Cited on pages 28, 30, 31, 32, 33, 39, and 120.)

Rényi, A. (1957). A characterization of Poisson processes. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 1:519–527. in Hungarian. (Cited on page 156.)

Sato, K.-i. (1991). Self-similar processes with independent increments. *Probability Theory and Related Fields*, 89(3):285–300. (Cited on pages 143, 147, and 198.)

Sato, K.-i. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1st edition. (Cited on pages 57, 148, 149, 193, and 199.)

Schwarz, D., Orio, N., and Schnell, N. (2004). Robust polyphonic Midi score following with hidden Markov models. In Tzanetakis, G., Essl, G., and Leider, C., editors, *Proceedings of the 2004 International Computer Music Conference (ICMC 2004), Miami, Florida, USA, November 1–6, 2004*, pages 442–445. Michigan Publishing. (Cited on page 44.)

Shaked, M. and Li, H. (2006). Aging first-passage times. In Kotz, S., Balakrishnan, N., Read, C. B., and Vidakovic, B., editors, *Encyclopedia of Statistical Sciences*, volume 1, pages 60–67. John Wiley & Sons, Inc., 2nd edition. (Cited on page 130.)

Shaked, M. and Shanthikumar, J. G. (1988). On the first-passage times of pure jump processes. *Journal of Applied Probability*, 25(3):501–509. (Cited on pages 130, 131, and 152.)

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer-Verlag, New York, NY, USA, 1st edition. (Cited on pages 89, 173, 204, 209, 210, 218, 219, 224, and 226.)

Shanthikumar, J. G. (1988). DFR property of first-passage times and its preservation under geometric compounding. *The Annals of Probability*, 16(1):397–406. (Cited on pages 130 and 218.)

Simon, T. (2011). Multiplicative strong unimodality for positive stable laws. *Proceedings of the American Mathematical Society*, 139(7):2587–2595. (Cited on page 233.)

Simon, T. (2014). Comparing Fréchet and positive stable laws. *Electronic Journal of Probability*, 19(16):1–25. (Cited on page 153.)

Simon, T. (2016). Total positivity in stable semigroups. *Constructive Approximation*, 44(1):103–120. (Cited on pages 142 and 153.)

Steutel, F. W. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 7(5):893–899. (Cited on pages 156 and 198.)

Steutel, F. W. and van Harn, K. (2004). *Infinite Divisibility of Probability Distributions on the Real Line*. Pure and Applied Mathematics. Marcel Dekker. (Cited on pages 145, 147, 152, 193, 195, 199, 237, and 242.)

Sun, Y., Baricz, A., and Zhou, S. (2010). On the monotonicity, log-concavity, and tight bounds of the generalized Marcum and Nuttall Q-functions. *IEEE Transactions on Information Theory*, 56(3):1166–1186. (Cited on pages 153 and 154.)

Takeda, H., Nishimoto, T., and Sagayama, S. (2007). Rhythm and tempo analysis toward automatic music transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), Honolulu, Hawaii, USA, April 15–20, 2007*, pages 1317–1320. IEEE. (Cited on pages 117 and 120.)

Veillette, M. and Taqqu, M. S. (2010). Using differential equations to obtain joint moments of first-passage times of increasing Lévy processes. *Statistics & Probability Letters*, 80(7–8):697–705. (Cited on pages 132 and 141.)

Vellaisamy, P. and Kumar, A. (2011). First-exit times of an inverse Gaussian process. *ArXiv e-prints*. (Cited on page 138.)

Vercoe, B. (1984). The synthetic performer in the context of live performance. In Wessel, D., editor, *Proceedings of the 1984 International Computer Music Conference (ICMC 1984), Paris, France, October 19–23, 1984*, pages 199–200. Michigan Publishing. (Cited on page 27.)

Vercoe, B. and Puckette, M. (1985). Synthetic rehearsal: Training the synthetic performer. In Truax, B., editor, *Proceedings of the 1985 International Computer Music Conference (ICMC 1985), Burnaby, BC, Canada, August 19–22, 1985*, pages 275–278. Michigan Publishing. (Cited on page 25.)

Veres-Ferrer, E. J. and Pavía, J. M. (2012). La elasticidad: una nueva herramienta para caracterizar distribuciones de probabilidad. *Rect: Revista Electrónica de Comunicaciones y Trabajos de ASEPUMA*, 13:145–158. (Cited on page 234.)

Veres-Ferrer, E. J. and Pavía, J. M. (2016). The elasticity function of a discrete random variable and its properties. *Communications in Statistics – Theory and Methods*. (Cited on page 234.)

Vinogradov, O. P. (1974). The definition of distribution functions with increasing hazard rate in terms of the Laplace transform. *Theory of Probability & Its Applications*, 18(4):811–813. (Cited on page 243.)

Wang, Y. and Yeh, Y.-N. (2005). Proof of a conjecture on unimodality. *European Journal of Combinatorics*, 26(5):617–627. (Cited on pages 155 and 156.)

Wang, Z.-y. and Xiao, X. (2006). Duration-distribution-based HMM for speech recognition. *Frontiers of Electrical and Electronic Engineering in China*, 1(1):26–30. (Cited on page 190.)

Watanabe, T. (1991). On the strong unimodality of Lévy processes. *Nagoya Mathematical Journal*, 121:195–199. (Cited on page 143.)

Watanabe, T. (1992). On Yamazato's property of unimodal one-sided Lévy processes. *Kodai Mathematical Journal*, 15(1):50–64. (Cited on pages 144, 158, 229, and 245.)

Whitt, W. (1985). Uniform conditional variability ordering of probability distributions. *Journal of Applied Probability*, 22(3):619–633. (Cited on page 165.)

Yamazato, M. (1978). Unimodality of infinitely divisible distribution functions of class L. *The Annals of Probability*, 6(4):523–531. (Cited on pages 144 and 229.)

Yamazato, M. (1982). On strongly unimodal infinitely divisible distributions. *The Annals of Probability*, 10(3):589–601. (Cited on page 175.)

Yu, S.-Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, 174(2):215–243. (Cited on pages 41, 43, and 98.)

Yu, S.-Z. and Kobayashi, H. (2003). A hidden semi-Markov model with missing data and multiple observation sequences for mobility tracking. *Signal Processing*, 83(2):235–250. (Cited on page 191.)

Yu, Y. (2009a). On the entropy of compound distributions on nonnegative integers. *IEEE Transactions on Information Theory*, 55(8):3645–3650. (Cited on page 163.)

Yu, Y. (2009b). Stochastic ordering of exponential family distributions and their mixtures. *Journal of Applied Probability*, 46(1):244–254. (Cited on page 248.)

Yu, Y. (2011a). On log-concavity of the generalized Marcum Q function. *ArXiv e-prints*. (Cited on pages 154, 158, and 229.)

Yu, Y. (2011b). On normal variance-mean mixtures. *ArXiv e-prints*. (Cited on pages 158 and 230.)