PhD Dissertation

Limit Theorems for Harmonizable Lévy-driven Process
and Analysis of Sequential Medical Data

AARHUS
UNIVERSITY
DEPARTMENT OF MATHEMATICS

Thorbjørn Ø.B. Grønbæk
September 2019

Limit Theorems for Harmonizable Lévy-driven Processes
and
Analysis of Sequential Medical Data

PhD Dissertation by Thorbjørn Ø.B. Grønbæk

Department of Mathematics, Aarhus University
Ny Munkegade 118, 8000 Aarhus C, Denmark.

Main supervisor Associate Professor Andreas Basse-O'Connor

Co-supervisors Associate Professor Jan Pedersen
Associate Professor Lars Nørvang Andersen
Professor Kim Mouridsen

Submitted September 4, 2019

Contents

Preface

This dissertation is the culmination of my PhD studies at the Department of Mathematics, Aarhus University, from August 1, 2015 until September 4th, 2019. The studies were performed under the supervision of Andreas Basse-O'Connor (main supervisor), Jan Pedersen, Kim Mouridsen and Lars Nørvang Andersen and were fully funded by Andreas' grant "Time-wise behavior of fractional processes" (DFF-4002-00003) from the Danish Council of Independent Research. The dissertation consists of four papers, divided into two parts.

Paper A Andreas Basse-O'Connor, Thorbjørn Grønbæk and Mark Podolskij. Local asymptotic self-similarity for heavy tailed harmonizable fractional Lévy motions. *Submitted*.

Paper B Andreas Basse-O'Connor, Thorbjørn Grønbæk and Mark Podolskij. Limit theory for quadratic variation of harmonizable Lévy-driven processes. *Working paper*.

Paper C Thorbjørn Grønbæk, Lars N. Andersen and Kim Mouridsen. Integration of sequential EHR information using mathematical representations. *Working paper*.

Paper D Thorbjørn Grønbæk. Analysis of medication sequences for sepsis patients. *Draft*.

In all of the above papers, I have made major contributions in both the research and writing phase. The papers A-B were completed during the first two years of my PhD and constitute the first (Part I) of my PhD. The main topic is limit theorems for harmonizable Lévy-driven processes and the common context of Papers A-B is discussed in Chapter 1. The main results of the papers A-B were included in my progress report, for which I received the degree Master of Science in Statistics as part of a 4+4 PhD structure (combined M.Sc. and Ph.D. studies). Parts of both papers were written during the last 2 years of my PhD studies.

The papers C-D were written during the last 2 year of my PhD and constitute the second (Part II) of my PhD. Paper C is the primary article for part II of my PhD. Paper D correspond to the work I completed during my exchange visit to Stanford University and was written in late summer 2019 and may appear a bit unpolished. A brief introduction is provided in Chapter 4 and technical supplementary material is provided in Chapter 5.

This dissertation marks the end of my PhD studies at Aarhus University and it has been a truly unique, challenging and rewarding experience. It is often said that a dissertation could not have been completed without the support of certain people – I now realize just how true such a statement may be. My path

through the PhD has been certainly not been a straight line, but sometimes felt like an increasingly complex curve. I would like to highlight Lars Nørvang Andersen and Kim Mouridsen for stepping up as advisors midway through my PhD – my sincere thanks for this. In particular, I would like to express my gratitude to Lars for always having his door open for advice and our discussions.

During my studies, I had the pleasure to visit Professor Daniel Rubin at the Department of Biomedical Informatics at Stanford University. I am very grateful for this and also wish to thank Imon, Alfiia, Armin, Jason, Arturo, Matthew, Steve and Blaine for engaging discussions, cozy lunches and good times.

I was lucky to share my PhD journey at Aarhus University with many people. I would like to thank Victor, Mikkel, Simon Bang, Mathias, Julie, Patrick, Claudio, Mads, Mark, Jevgenijs and many others for enduring my talkative nature, keeping up my good spirits and for the recurring discussions of the weather and prices in the canteen. Lunch will feel different without you and although I never liked Staff Lounge coffee, I certainly enjoyed the conversation while waiting for it to brew.

To my friends and family, thank you for always being there. You remind me that theorems, proofs, code and quantitative analysis is nice but life is about shared laughter and memories. In December 2019, a little star was born, when my girlfriend gave birth to our daughter Annabell – she should know that her smiles makes me feel more complete than ever. To my girlfriend Cecilie, thank you for laughing when I try to find my keys for the 100th time, being patient when I work long hours, challenging me when I appear too confident, forgiving me when I forget to buy milk and rye bread (again) and ultimately for your love. It means the world to me.

Throughout these studies, I have always felt, and had, the full support of Andreas and Jan, who have always kept my best interest in mind – even if I did not know my own best interest. This thesis would never have been without the two of them. Thanks does not express my gratitude towards them. Our shared experiences has become part of who I am and I will miss their company, discussions, jokes, laughter and advice. Jan and Andreas exemplified to me values of excellence, patience, openness and kindness. If I succeed in living these values half as well as Jan and Andreas does, I will have done well. With these words, I dedicate this thesis to them.

Thorbjørn Øystein Brynimin Grønbæk
Aarhus, September 2019

Summary

This dissertation consists of two parts. The first part concerns the study of a class of complex-valued stochastic processes called harmonizable processes. Our main topic of interest will be on limit theorems for α-stable harmonizable processes. We define a class of harmonizable fractional Lévy processes and give a practical existence criterion for their these. Next, we study the local behavior of Lévy-driven harmonizable processes and show that if the driving Lévy process has heavy tails then they are locally self-similar with tangent process harmonizable fractional stable motion. Previously it has been shown that a class harmonizable processes where the driving Lévy process has all moments, are locally self-similar with tangent process fractional Brownian motion and thus this gives a more complete picture. We show a limit theorem for the quadratic variation of a class of harmonizable processes towards a non-degenerate limit. It is of note that harmonizable fractional stable motion is neither mixing, weakly mixing nor ergodic and thus the above results constitutes a new contribution to limit theory for non-ergodic processes. The results implies that harmonizable fractional stable motion cannot be a semimartingale for $H < 1/2$.

The second part is focused on the analysis of sequential medical data through two applications. We study electronic health records which are at the center of health care systems, foremost documenting the patient history but secondly attributing credit and responsibility to the entity performing each task. It is therefore of utmost importance that each event is correctly registered. However, for clinical treatment in a stressed hospital environment, time spent registering an event equates time not spent diagnosing/treating patients. This discrepancy of priorities, between registration and actual treatment, appears to be a natural precondition, but if registration of an event could assist correct diagnosis then both goals would align. This is the aim in both our two applications. The first application is sequential semantic meaning of events in electronic health records, gathered for a specific cohort of patients at Regionshospitalet Silkeborg, Denmark. We show that it is possible to automatically incorporate semantic meaning into a numerical vector representation by analyzing the records using an algorithm called Skip-Gram. An interesting visualization is provided, which shows that the algorithm automatically identifies events groups. To the best of our knowledge, this utilization of the Skip-Gram method to incorporate sequential semantic meaning in medical data is new to the literature – previous studies focused primarily on free-text reports. In the second application, medication orders for sepsis patients is the central topic. The medication logs are extremely short which complicates statistical analysis. Initially, we study whether a new sepsis alert system alters the treatment behaviour using a group variable. We show that the most com-

mon medications and frequencies of these are the same for each group. From this, we conclude that the treatment sequences are unchanged between the groups and proceed to study prediction of the next treatment package and the graph of treatment packages on larger dataset of sepsis patients. Finally, we describe our experiences and thoughts on sequential medical data and several prospective ideas are discussed.

Resumé

Denne afhandling består af to dele. Den første del omhandler studier af en klasse af stokastiske processer med komplekser værdier kaldet harmoniske processer. Vi er særligt interesseret i grænseværdisætninger for α-stabile harmoniske processer. We definerer klassen af harmoniske fraktionale Lévy processer og giver et praktisk eksistenskriterie for disse. Hernæst studerer vi deres lokale opførsel og under antagelsen om α-stabile halefordelingen af den drivende Lévy proces, viser vi at harmoniske fraktionale Lévy processer er lokalt selv-similære med tangent proces den harmoniske fraktionale stabile bevægelse. Dette sættes i kontekst af et resultat fra litteraturen vedrørende selv-similaritet for harmoniske fraktional Lévy processer, hvor den drivende Lévy process har alle momenter. Vi viser også at den kvadratiske variation for en klasse af harmoniske processer konvergerer mod en ikke-degenereret stokastisk variabel. Dette resultat er en interessant tilføjelse til grænseværdisætninger for ikke-ergodiske processer, idet den harmoniske fraktionale stabile bevægelse ikke er ergodisk. Som konsekvens af konvergens kan vi konkludere at den harmoniske fraktionale stabile bevægelse ikke er en semimartingal for $H < 1/2$.

Den anden del omhandler analysen af sekventielle sundhedsdata. Vi studerer elektroniske patientjournaler, som er centrale objekter i sundhedssystemer, først og fremmest ved at dokumentere patienthistorikken, men dernæst ved at give kredit og ansvar til sundhedsaktøren som udfører opgaven/hændelsen. Det er derfor særlig vigtigt at hver enkelt hændelse bliver korrekt registreret. Men for klinisk behandling i et stresset hospitalsvæsen, er tid brugt på registrering af en hændelse lig tid ikke brugt på at diagnosticere og behandle patienter. Denne diskrepans mellem prioriteter, kan forekomme at være en naturlov, men hvis registreringen af en hændelse assistere diagnosticering, så vil målet for både registrering og behandlingen være opnået. Dette er formålet med vores to studier af sekventielle sundhedsdata. I første datasæt studerer vi sekventiel semantisk mening for hændelser i elektroniske patientjournaler for en specifik patientgruppe fra Regionshospitalet Silkeborg, Danmark. Vi viser at det er muligt at integrere semantisk mening automatisk i numerisk vektorrepræsentationer ved at analysere journalerne med en en algoritme kaldet Skip-Gram. Vi laver en interessant visualisering, som viser at algoritmen automatisk finder hændelsesforløb. Efter vores bedste overbevisning at denne måde at udnytte Skip-Gram til at finde sekventielt semantisk mening er ny i litteraturen – andre studier har typisk fokuseret på fri-tekst rapporter. I det andet datasæt er det centrale emne medicinjournaler for sepsis patienter . Medicinjournalerne er ekstremt korte hvilket komplicerer statistisk analyse. Initialt studerer vi hvorvidt et nyt sepsis alarmsystem ændrer behandlingsforløbet baseret på en gruppevariabel. Vi viser at de hyppigste behandlingspakker (og medicin)

samt deres frekvens er ens for de to grupper. Fra dette, konkluderer vi at behandlingsforløbene er uændret på tværs af gruppevariablen og vi mangler de to grupper med at et større datasæt (uden en gruppevariabel) med sepsis patienter. Til sidst beskriver vi vores erfaringer og tanker om udnyttelsen af sekventielt sundhedsdata og flere ideer diskuteres.

Part I

Limit Theorems for Harmonizable Lévy-driven Processes

CHAPTER 1

Introduction to harmonizable processes

This chapter starts with a brief motivation for stochastic processes followed by a background brush-up on stationary processes, moving average and spectral representation. Then, we discuss the intersection of moving averages and spectral representations and exemplify this through fractional Brownian motion. Fractional Brownian motion is subsequently generalized to the α-stable case and in this case two different processes arise, namely harmonizable fractional stable motion and linear fractional stable motion. We present new results on the local asymptotical self-similarity and quadratic variation for the former and draw perspectives to known results on the latter.

1.1 Stochastic processes

"Random" behavior or phenomena occurs frequent in daily life, for example in stock prices, weather forecasts, particle movement, decision-making, imprecision of measurements and many other examples. A stochastic process, $(X_t)_{t\in T}$, is a collection of random variables (or vectors) indexed by T ("time"), exemplified in Figure 1.1. For a statistician, stochastic processes are typically

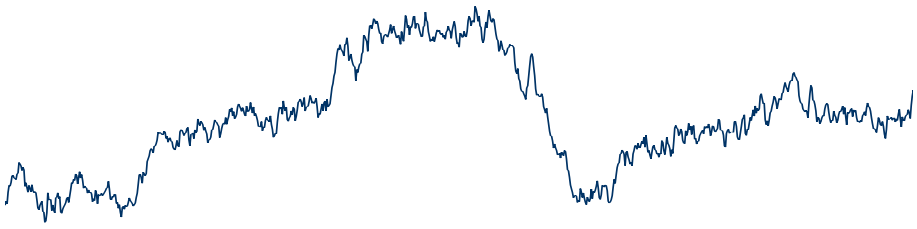


Figure 1.1: A simulated stochastic process (Brownian motion).

applied to the study of properties which are invariant across T, allowing the observer to perform statistical inference. In probability theory, such properties are commonly referred to as *stationarity*.

1.1.1 Background

We start out with a few basic definitions for stochastic processes. Formally, a stochastic process $(X_t)_{t\in\mathbb{R}}$ is *stationary*, if

$$(X_t)_{t\in\mathbb{R}} \overset{d}{=} (X_{t+h})_{t\in\mathbb{R}}, \quad \forall\, h \in \mathbb{R}, \tag{1.1}$$

where $\overset{d}{=}$ denotes equality in finite-dimensional distributions. Similarly, a stochastic process $(X_t)_{t\in\mathbb{R}}$ has *stationary increments* if

$$(X_{t+h} - X_h)_{t\in\mathbb{R}} \overset{d}{=} (X_t - X_0)_{t\in\mathbb{R}}, \quad \forall\, h \in \mathbb{R}. \tag{1.2}$$

In other words, $(X_{t+h} - X_h)_{t\in\mathbb{R}}$ constitute a stationary process for all $h \in \mathbb{R}$. If the process itself is not stationary, then the increments, log-increments or other functionals are often stationary. A stochastic process $\mathbf{X}$ is *weakly stationary* if $X_t \in L^2(\mathbb{P})$, $\mathbb{E}[X_t] = 0$, $\mathrm{Cov}(X_{t+h}, X_h) = \mathrm{Cov}(X_t, X_0)$ for all t, $h \in \mathbb{R}$ and $t \mapsto \gamma_X(t) \coloneqq \mathrm{Cov}(X_t, X_0)$ is a continuous function. In particular if $\mathbf{X}$ is stationary and $\mathbb{E}[X_0^2] < \infty$ (second moment) then $\mathbf{X}$ is weakly stationary.

A stochastic process X has independent increments if for any $m \in \mathbb{N}$ and any sequence $t_0 < t_1 < \cdots < t_m$, it holds that the random variables (or vectors)

$$X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \ldots, X_{t_m} - X_{t_{m-1}} \tag{1.3}$$

are independent. These processes are common for many of the stochastic processes studied in the literature. A common example of a process with stationary independent increments is the so-called Brownian motion.

Example 1.1.1 (Brownian motion)**.** A real-valued stochastic process $\mathbf{B} = (B_t)_{t\in\mathbb{R}}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ is called a Brownian motion (without drift) if it satisfies that

- $t \mapsto B_t(\omega)$ is a continuous function for all $\omega \in \Omega$.
- $B_0 = 0$.
- $\mathbf{B}$ has stationary independent increments.
- B_t has a normal distribution with mean zero and variance $\sigma^2|t|$ for all $t \in \mathbb{R}$.

If $\sigma^2 = 1$ it is called a standard Brownian motion and a drift μ can be added in the mean as $\mu \cdot t$ to achieve a Brownian motion with drift μ.

The Brownian motion is an example of the larger class of Lévy processes which again have stationary independent increments.

Example 1.1.2 (Lévy processes)**.** An $\mathbb{R}^d$-valued stochastic process $\mathbf{L} = (L_t)_{t\in\mathbb{R}}$ is a Lévy process on $\mathbb{R}$ if it satisfies that

- $t \mapsto L_t(\omega)$ is a càdlàg function for all $\omega \in \Omega$

- **L** has stationary independent increments.
- $L_0 = 0$.

A Lévy process induces an infinitely divisible independently scattered random measure (see [12]) through

$$\Lambda_L((s,t]) = L_t - L_s$$

which may be extended to $\Lambda_L(A) = \int \mathbb{1}_A \, dL$ for $A \in \mathcal{B}_b(\mathbb{R})$ (bounded Borel sets) using standard methodology. Consequently, Lévy processes can be used to define stochastic integrals as discussed in [12], and these may yield many different stationary processes (or processes with stationary increments).

Example 1.1.3 (Lévy-driven moving averages)**.** Let $\mathbf{L} = (L_t)_{t\in\mathbb{R}}$ denote a $\mathbb{R}$-valued Lévy process and let $\tilde{\phi}, \phi : \mathbb{R} \to \mathbb{R}$ denote measurable functions which are zero for $t \leq 0$. The stochastic process $\mathbf{Y} = (Y_t)_{t\in\mathbb{R}}$ defined by

$$Y_t = \int_{-\infty}^{t} \phi(t-s)\, dL_s$$

is then a stationary process, provided that the integral exists in the sense of [12]. Similarly, the stochastic process $\mathbf{Z} = (Z_t)_{t\in\mathbb{R}}$ defined by

$$Z_t = \int_{-\infty}^{t} [\phi(t-s) - \tilde{\phi}(-s)]\, dL_s$$

has stationary increments, provided that the integral exists. The stochastic processes defined in the previous example, **Y** and **Z**, are often referred to as continuous-time (Lévy-driven) moving averages, respectively stationary increments (Lévy-driven) moving averages. This definition may be extended in a natural way to functions $\phi : \mathbb{R} \to \mathbb{R}^{k\times d}$ and $\mathbb{R}^d$-valued Lévy processes **L**, resulting in **Y** and **Z** with values in $\mathbb{R}^k$.

A particular interesting property in relation to Brownian motion is *selfsimilarity*. A stochastic process $(X_t)_{t\in\mathbb{R}}$ is *self-similar* with index H if

$$(X_{ct})_{t\in\mathbb{R}} \overset{d}{=} c^H (X_t)_{t\in\mathbb{R}}, \quad \text{for } c > 0,$$

where $\overset{d}{=}$ denotes equality in finite-dimensional distributions. The Brownian motion is self-similar with index $H = 1/2$ and its generalization fractional Brownian motion in Example 1.2.1 is self-similar with index $0 < H < 1$. The following Example 1.1.4 and Theorem 1.1.5 underlines self-similarity and stationary as related and central properties in the study of stochastic processes.

Example 1.1.4 (Self-similarity and limits)**.** Self-similarity arises as a property of limit distributions. Namely that if for a stochastic process $\mathbf{X} = (X_t)_{t\in\mathbb{R}}$ it holds that

$$\left(\frac{X_{ct} - g(c)}{f(c)}\right) \xrightarrow[c\to\infty]{d} Y_t, \tag{1.4}$$

for a real-valued function $g : \mathbb{R} \to \mathbb{R}$ and a strictly positive, real-valued increasing function f with $f(c) \to \infty$ as $c \to \infty$, this implies that the process $\mathbf{Y} = (Y_t)_{t\in\mathbb{R}}$ is self-similar. Here $\xrightarrow[c\to\infty]{d}$ denotes convergence in distribution. Conversely if a process $\mathbf{Y}$ is self-similar, then $\mathbf{Y}$ occurs as such a limit in equatiion (1.4) for some process $(X_t)_{t\in\mathbb{R}}$. For the full results, see [10].

In addition, self-similarity has a direct connection to stationarity, namely through its Lamperti transform, equation (1.5), as stated in the following theorem.

Theorem 1.1.5 (Lamperti, [10])**.** *If $(X_t)_{t\geq 0}$ is a stationary process which is continuous in probability and we define*

$$Y_t = t^H X_{\log t}, \quad t > 0 \tag{1.5}$$

then $(Y_t)_{t\geq 0}$ is self-similar with index H. Conversely, every nontrivial self-similar process with $Y_0 = 0$ is obtained this way from a stationary process $(X_t)_{t\geq 0}$.

Thus stationarity and self-similarity are central properties in the study of stochastic processes and whenever we have a stationary process, the Lamperti transform in equation (1.5) yields a self-similar process.

1.1.2 Moving average representation

The class of Lévy-driven moving averages is a subclass of the larger class of moving averages. The Wold-Karhunen representation, see [1, 8], states that any *weakly stationary* stochastic process $\mathbf{X}$ may be represented by

$$X_t = \int_{\mathbb{R}} \phi(t-s)\, d\xi_s + V_t, \quad \forall\, t \in \mathbb{R} \tag{1.6}$$

where $\mathbf{V} = (V_t)_{t\in\mathbb{R}}$ is process which is measurable wrt. its ultimate past $\mathcal{V}_{-\infty}$, e.g. $\mathbf{V}$ is measurable wrt. $\mathcal{V}_{-\infty} := \bigcap_{t\in\mathbb{R}} \overline{\operatorname{span}}(V_s \colon s \leq t)$ ($\overline{\operatorname{span}}$ denotes the L^2 closure of the linear span), ξ_X is a weakly stationary stochastic processes with orthogonal increments and ϕ is a Lebesgue square-integrable determinisitic function. An application of Bochner's theorem yields a finite positive measure F_X, often called the *spectral measure* of $\mathbf{X}$, such that

$$\mathcal{F}[F_X](u) = \gamma_X(u) := \mathbb{E}[X_h X_0], \tag{1.7}$$

where $\mathcal{F}[F_X](u) = \int_{\mathbb{R}} e^{ius} F_X(ds)$ denotes the Fourier transform of the measure F_X. Let $F_X'(x)$ denote the absolutely continuous part of F_X. If F_X satisfies the condition

$$\int_{\mathbb{R}} \frac{|\log F_X'(x)|}{1+x^2}\, dx < \infty$$

then $\phi(t) \equiv 0$ for all $t \leq 0$, it will result in

$$X_t = \int_{-\infty}^{t} \phi(t-s)\, d\xi_s + V_t, \quad \forall\, t \in \mathbb{R},$$

a so-called backward moving average (e.g. depending only on the past). If the measure F_X is absolutely continuous wrt. the Lebesgue measure (cf. Theorem 4.1 in [1]), then $\mathbf{X}$ has a *moving average* representation,

$$X_t = \int_{\mathbb{R}} \psi(t-u)\, \xi_X(ds), \tag{1.8}$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is a Lebesgue square-integrable deterministic function and ξ_X is a weakly stationary stochastic process with mean zero and orthogonal increments. This corresponds to $\mathbf{V} \equiv 0$ in the Wold-Karhunen representation (1.6).

1.1.3 Spectral representation

In this section we introduce spectral representations and harmonizable processes (of which harmonizable fractional stable motion from Example 1.2.2 is an element) and relate this definition to weak stationarity. By a *harmonizable* process $\mathbf{Z} = (Z_t)_{t\in\mathbb{R}}$, we understand a process defined by

$$Z_t = \int_R e^{its} M(ds)$$

where M denotes a complex-valued random measure. A process $\mathbf{Z} = (Z_t)_{t\in\mathbb{R}}$ has harmonizable increments if

$$Z_{t+h} - Z_h = \int_{\mathbb{R}} \frac{e^{i(t+h)s} - e^{ihs}}{is} M(ds).$$

where M is a complex-valued random measure. In alignment with the nomenclature in the literature, a process with harmonizable increments will be named a harmonizable process despite the confusing double naming. In each instance, we will try to fully write out the representation of the process to avoid this confusion. A common choice for M is an isotropic infinitely divisible independently scattered random measure, see equation (2.10), but it may also involve dependency across time, as seen in the harmonizable representation of fractional Brownian motion in Example 1.2.1. The subclass of harmonizable processes of the form

$$X_{t+h} - X_h = \int \frac{e^{i(t+h)s} - e^{ihs}}{is} g(s)\, dL_s, \quad (t \in \mathbb{R}) \tag{1.9}$$

where L is a complex-valued isotropic Lévy process and g is a complex-valued deterministic function, we shall call *Lévy-driven harmonizable processes*. The increments of harmonizable fractional stable motion from Example 1.2.2 is an example of Lévy-driven harmonizable process. The existence of Lévy-driven harmonizable process can be checked by combining the result in Theorem 2.3.3 with Lemma 2.4.2. Theorem 2.3.3 directly connects stationarity (or stationary increments) with isotropy for harmonizable processes $\mathbf{X}$ with infinitely divisible independently scattered random measure – a well-known result in

discrete time from [16]. If we drop independently scattered, it is possible to obtain harmonizable processes which are stationary but the measure M is not isotropic. $\tilde{B}$ in the harmonizable representation of fractional Brownian motion is an example of this.

If the increments of a stochastic process $\mathbf{X} = (X_t)_{t\in\mathbb{R}}$ are weakly stationary, we may always express it using its *spectral representation* (or harmonizable) (cf. [2, 17]), given by

$$X_{t+h} - X_h = \int_{\mathbb{R}} \frac{e^{i(t+h)s} - e^{ihs}}{is} \Lambda_X(ds), \quad (t \in \mathbb{R}) \tag{1.10}$$

where the stochastic process Λ_X is complex-valued, square-integrable, continuous in L^2 and has orthogonal increments, i.e. $\mathbb{E}[(\Lambda_X(v)-\Lambda_X(u))\overline{(\Lambda_X(t)-\Lambda_X(s))}] = 0$ for $u \le v \le s \le t$. The process Λ_X is often called the spectral process of X, due to the direct similarity with the Fourier transformation for deterministic functions. We will use the terms spectral representation and harmonizable representation interchangeably as appears common in the literature.

The *control measure* (or spectral measure/distribution) of Λ_X, denoted F, is defined by

$$F((a,b]) \coloneqq \mathbb{E}[|\Lambda_X(a) - \Lambda_X(b)|^2] \tag{1.11}$$

The function F_X is equivalent to the one found in the Wold-Karhunen representation in equation (1.6), provided both representations exists.

1.2 Intersection between moving averages and harmonizable processes

An important interpretation of both the moving average and spectral representation is that they separate time-dependency and stochasticity of the process into a deterministic function ψ and a random measure ξ. The deterministic kernel function carries the time-dependency, whereas the random measure has stationary and orthogonal (uncorrelated) increments. This motivates analyzing ψ and ξ_X separately to determine their effect on the resulting process.

Lévy processes with second moment satisfy the criteria for both types of representations and consequently have both a spectral representation and a moving average representation. This is not necessarily an improvement, since Λ_L for a Lévy process L would satisfy weaker assumptions than L itself, leading to less structure on the stochastic part of the representation. As a quick example, the increments of Brownian motion $\mathbf{B}$ may thus be written using its *spectral representation*

$$B_{t+h} - B_t = \int_{\mathbb{R}} \frac{e^{i(t+h)s} - e^{its}}{is} \Lambda_B(ds), \quad (t \in \mathbb{R})$$

where Λ_B is the spectral process of $\mathbf{B}$. $\mathbf{B}$ can also easily be written in a stationary increments moving average form, namely

$$B_t - B_0 = \int_{-\infty}^{t} [\mathbb{1}_{(-\infty,0]}(t-s) - \mathbb{1}_{(-\infty,0]}(-s)]\, dB_s = \int_{\mathbb{R}} \mathbb{1}_{(0,t]}(s)\, dB_s.$$

In this case, however, the moving average / spectral representation does not improve amount of known structure on the stochastic component (Λ_B or $\mathbf{B}$).

Example 1.2.1 (Fractional Brownian motion, [11])**.** Fractional Brownian motion (fBm) was studied in [11], and generalizes the self-similarity property of Brownian motion. It is a stochastic process with mean-zero, normally distributed marginals, stationary increments and covariance given by

$$\mathrm{Cov}(X_t, X_s) = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |s-t|^{2H}) \quad \forall\, s,t \in R.$$

Note that we no longer have independent increments unless the similarity index is $H = 1/2$ (in which case, fBm equals Brownian motion). It was established in [11] that fBm is the *only* centered H-self-similar Gaussian process with stationary increments. Fractional Brownian motion has both a moving average representation and a spectral representation (see Chapter 7 of [15]), namely

$$B_t^H = C_1(H) \int_{-\infty}^{t} (t-s)_+^{H-1/2} - (-s)_+^{H-1/2}\, dB_s, \quad t \in \mathbb{R}, \tag{1.12}$$

and

$$B_t^H = C_2(H) \int_{\mathbb{R}} \frac{e^{its}-1}{is} |s|^{-H-1/2+1}\, \tilde{B}(dx), \quad t \in \mathbb{R}. \tag{1.13}$$

Here $(B_s)_{s\in\mathbb{R}}$ is a Brownian motion and $\tilde{B}$ is an complex-valued Gaussian measure defined by $\tilde{B} = B^{(1)} + iB^{(2)}$, where $B^{(1)}$ and $B^{(2)}$ are independent Gaussian random measures, satisfying $B^{(1)}(A) = B^{(1)}(-A)$, $B^{(2)}(A) = -B^{(2)}(A)$ for any Borel set A of finite Lebesgue measure. Both representations yield a real-valued Gaussian H-selfsimilar process with stationary increments and thus by [11] the same process.

A natural mathematical question is whether the equality between the *moving average* and *spectral representation* can be generalized to non-Gaussian self-similar processes? A first step in this direction was to use α-stable Lévy processes as integrators in (1.12)-(1.13), where the Gaussian process is the special case $\alpha = 2$. This idea was pursued in [6] and generalizes the representations above to α-stable Lévy processes.

Example 1.2.2 (Fractional stable motions, [6])**.** *Linear fractional stable motion* (lfsm) is analog to (1.12) and defined as

$$\begin{aligned} L_{\alpha,H}(t) = &\int_{\mathbb{R}} a\Big((t-s)_+^{H-1/\alpha} - (-s)_+^{H-1/\alpha}\Big)\, dM_s^\alpha \\ &+ \int_{\mathbb{R}} b\Big((t-s)_-^{H-1/\alpha} - (-s)_-^{H-1/\alpha}\Big)\, dM_s^\alpha, \quad (t \in \mathbb{R}) \end{aligned}$$

where $(M_s^\alpha)_{s\in\mathbb{R}}$ is a (symmetric) α-stable Lévy process on $\mathbb{R}$, $a, b \in \mathbb{R}$ such that $|a| + |b| > 0$, $0 < \alpha < 2$, $0 < H < 1$, $H \neq 1/\alpha$. The complex-valued *harmonizable fractional stable motion* (hfsm) is similarly analog to (1.13) and defined as

$$C_{\alpha,H}(t) = \int_{\mathbb{R}} \frac{(e^{its}-1)}{is}\Big(a(s_+)^{-H+1-1/\alpha} + b(s_-)^{-H+1-1/\alpha}\Big)\, d\tilde{M}_s, \quad t \in \mathbb{R},$$

where $\tilde{M}$ denotes a complex-valued isotropic α-stable Lévy process and the parameter space is $0 < \alpha < 2$, $0 < H < 1$, $a \geq 0$, $b \geq 0$, where $a + b > 0$. We refer to Chapter 7 of [15] for detailed definitions of these processes.

Both the *lfsm* and the real part of hfsm are real-valued H-selfsimilar α-stable processes with stationary increments, however contrary to the Gaussian case in [11], we no longer have find that such properties implies uniqueness of the process. Indeed, the main finding of [6] is the non-trivial result that hfsm and lfsm are different processes. This raises the question as to which extent the lfsm $(X_t)_{t\in\mathbb{R}}$ and hfsm $(Z_t)_{t\in\mathbb{R}}$ are related? This is the starting point of our research.

Comparison of properties for lfsm and hfsm has been done extensively in the literature, e.g. codifference in [15] and path properties in [9], where lfsm and hfsm behave differently. Many papers have also been investigating the structure of stationary α-stable processes, e.g. the characterisation result [14], stating a α-stable stationary process can be decomposed into a moving average part, a harmonizable part and a process of a third kind. This reveals that by studying moving averages and harmonizables processes, we will, in fact, be studying a large subclass of the stationary stable processes.

1.3 Our contributions

Self-similarity is studied in [4] for a class of Lévy-driven harmonizable fields $(Y_t)_{t\in\mathbb{R}^d}$, where it is crucially assumed that all moments of their driving Lévy process exist. They obtain the following *local asymptotical selfsimilarity*

$$\left(\frac{Y_{u+\varepsilon t} - Y_u}{\varepsilon^H}\right) \xrightarrow[\varepsilon\to 0_+]{d} c_0(B_H(t))_{t\in\mathbb{R}^d},$$

where $(B_H(t))_{t\in\mathbb{R}}$ is the fractional Brownian motion, c_0 is a constant given by the moments for the Lévy measure and the convergence is in a strong functional distributional sense. To clarify, a stochastic proces $(Y_t)_{t\in\mathbb{R}}$ is *locally asymptotically selfsimilar* (lass), with index $h(u)$ at point u, if

$$\lim_{\varepsilon\to 0_+}\left(\frac{Y_{u+\varepsilon t} - Y_u}{\varepsilon^{h(u)}}\right) \overset{d}{=} (T_u(t))_{t\in\mathbb{R}},$$

where $(T_u(t))_{t\in\mathbb{R}^d}$ is a non-degenerate stochastic process. The process $(T_u(t))_{t\in\mathbb{R}^d}$ is called the *tangent process* at point u. We will often refer to this as *locally selfsimilar*.

In Chapter 2, we study local self-similarity for a class of harmonizable processes, namely the following harmonizable fractional Lévy motions.

Definition 1.3.1. *A stochastic process $(X_t)_{t\in\mathbb{R}}$ is called a harmonizable fractional Lévy motion (hflm) with parameters $(\alpha, H) \in \mathbb{R}^2$ if*

$$X_t = \int_{\mathbb{R}} \frac{e^{its} - 1}{is}\Big(a(s_+)^{-H-1/\alpha+1} + b(s_-)^{-H-1/\alpha+1}\Big)\,dL_s, \quad t \in \mathbb{R} \tag{1.14}$$

where **L** *is a isotropic (same as rotational invariance, see Definition 2.3.1) complex-valued Lévy process, and* $a, b \in \mathbb{R}$.

Theorem 2.3.3 and Theorem 2.4.1 give explicit criteria for existence of hflm when **L** is isotropic. We shall need the following assumption to state our result.

Assumption (A): *Suppose that L is a rotationally invariant complex-valued Lévy process without Gaussian component and let* ν *denote its Lévy measure. We assume that* ν *is absolutely continuous with respect to the two dimensional Lebesgue measure with a density* $f : \mathbb{R}^2 \to \mathbb{R}_+$ *satisfying*

$$\begin{aligned} f(x) &\sim c_0 \|x\|^{-2-\alpha} \quad \text{as } \|x\| \to \infty, \\ f(x) &\leq C\|x\|^{-2-\alpha} \quad \text{for all } x \in \mathbb{R}^2, \end{aligned}$$

where $c_0, C > 0$ *and* $0 < \alpha < 2$.

In this case, we have the following theorem.

Theorem 1.3.2. *Let* **X** *be harmonizable fractional Lévy motion with parameters* (α, H) *which satisfies* ***(A)****. Then it holds that* **X** *is locally self-similar with index H and tangent process harmonizable fractional stable motion, i.e. that*

$$\left(\frac{X_{u+\epsilon t} - X_u}{\epsilon^H}\right)_{t\in\mathbb{R}} \xrightarrow[\epsilon \downarrow 0]{d} (\mathcal{C}_{\alpha,H}(t))_{t\in\mathbb{R}}, \tag{1.15}$$

where the convergence is in finite-dimensional distribution and the limit $\mathcal{C}_{\alpha,H}(t)$ *is a harmonizable fractional stable motion with parameters* (α, H).

In contrast to [4], we drop the assumption of all moments and instead assume heavy tailed behaviour on the Lévy process. To summarize [4] and our result – with all moments we obtain a (moving average/harmonizable) selfsimilar Gaussian tangent process, whereas a lack of moments yields a harmonizable selfsimilar α-stable tangent process. An alternative formulation is that the small-scale behaviour of harmonizable fractional Lévy motion is approximately Gaussian if all moments exists and α-stable if the Lévy process is heavy-tailed.

In Chapter 3, we aim to study the limit theory for the power variation of harmonizable processes, i.e. processes of the form (1.10), and in particular hfsm. Recall that low-frequency power variation (or p-variation) of a process $(X_t)_{t\in\mathbb{R}}$ is defined as

$$V(p)_n := \sum_{t=1}^{n} |X_t|^p$$

Similarly, the high-frequency p-variation is simply defined as

$$V_p(t, n) = \sum_{i=1}^{\lfloor tn \rfloor} |X_{i/n} - X_{(i-1)//n}|^p \tag{1.16}$$

For self-similar processes, the low- and high-frequency settings can be related by self-similarity. This limit theory would allow comparison to the moving

averages studied in [3] (which includes lfsm) and yield a deeper understanding of the connection between kernel function structure and the corresponding limit theory.

The special case ($p = 2$), called the quadratic variation, is often the first non-linear functional to be studied and has received particular interest. It measures variability of the process $\mathbf{X}$ and is given by

$$\sum_{i=1}^{n} \left(X_{i/n} - X_{(i-1)/n} \right)^2$$

in the high-frequency setting and is used to estimate parameters in many stochastic processes and diffusions. In the case of Brownian motion $\mathbf{B}$, it is well-known that

$$\sum_{i=1}^{n} (B_{i/n} - B_{(i-1)/n})^2 \to \sigma^2 \text{ a.s.}, \quad \text{as } n \to \infty.$$

We present our main result on quadratic variation ($p = 2$) for harmonizable Lèvy-driven processes in the low-frequency setting.

Theorem 1.3.3. *Let $X = (X_t)_{t \in \mathbb{R}}$ denote a harmonizable Lévy-driven process driven by a complex-valued isotropic Lévy process $L = L^1 + iL^2$. Then it holds that*

$$\frac{1}{n} \sum_{t=1}^{n} |X_t|^2 \xrightarrow{\mathbb{P}} U_0,$$

where U_0 is an infinitely divisible variable of the form

$$U_0 = \int_{\mathbb{R}} |g(s)|^2 \, d\left([L^1] + [L^2] \right)_s,$$

where $[L^1]$ and $[L^2]$ denotes the quadratic variation of the Lévy process L^1 respectively L^2.

For a sketch of the proof, we refer to Section 1.3.1. It relies on the chosen power $p = 2$ and the multiple integration theory developed in [7]. We expect that the result holds for general harmonizable processes $(Z_t)_{t \in \mathbb{R}}$ as well.

However, hfsm is neither mixing nor a semi-martingale (to the best of our knowledge). We believe the latter observation to be well-known but we have been unable to find a reference for it. Theorem 1.3.3 yields the following corollary for hfsm which partially answers this.

Corollary 1.3.4. *Harmonizable fractional stable motion $\mathbf{X}$ is not a semi-martingale for $H < 1/2$.*

This aligns exactly with the fractional Brownian motion results from [13]. The cases $H = 1/2$ and $H > 1/2$ appear open. In particular, whether $H = 1/2$ yields a semi-martingale appears to be an interesting open problem since [6] shows that once α is introduced as another "parameter" in the analogs for

the moving average and harmonizable representations of fractional Brownian motion, the two processes are different.

The local self-similarity and power variation of stationary increment Lévy-driven moving averages (including lfsm) was studied in the article [3]. They show that *lfsm* arises as a tangent process of certain moving averages, e.g.

$$\left(\frac{Y_{u+\varepsilon t}-Y_u}{\varepsilon^H}\right)\xrightarrow{\mathbb{P}}(L_{\alpha,H}(t))_{t\in\mathbb{R}},\quad \text{as } \varepsilon\downarrow 0_+, \tag{1.17}$$

where $L_{\alpha,H}$ denotes *lfsm* and under the assumption that the Lévy measure of the driving Lévy process has regularly varying tail behavior with index $0<\alpha<2$. The stronger convergence in probability for moving averages in (1.17) enables the authors [3] to transfer results regarding the power variation of lfsm to a larger class of stationary increments Lévy-driven moving averages in the infill asymptotic setting (fixed time horizon and the number of observations converge to infinity). The weaker convergence towards *hfsm* in Theorem 1.3.2 does not allow us to transfer results on the power variation for *hfsm* to harmonizable fractional Lévy motion. This led us to focus on the quadratic variation ($p=2$) which has been studied extensively in stochastic integration theory to model volatility.

In comparison to Theorem 1.3.3, [3] obtained the following result for *lfsm* (in the case where $H+1/\alpha>0$)

$$n^{-2/\alpha}\sum_{t=1}^{n}(Y_t-Y_{t-1})^2\xrightarrow{\mathcal{L}-s}\tilde{U}, \tag{1.18}$$

where $\xrightarrow{\mathcal{L}-s}$ refers to stable convergence in law. The stable central limit theorem entails that an *i.i.d.* sequence of symmetric $\alpha/2$-stable random variables $(Y_i)_{i\in\mathbb{N}}$, that

$$n^{-2/\alpha}\sum_{t=1}^{n}Y_i\xrightarrow{d}Y,$$

where Y is a symmetric $\alpha/2$-stable random variable. The tail distributions of $(X_t-X_{t-1})^2$ for both lfsm and hfsm behaves as a power law with index $\alpha/2$. Comparing the *i.i.d.* case to equation (1.18), we see that the normalization rates are the same whereas the normalization rate is different for hfsm . This may be related to α-stable moving averages being *weakly mixing*, whereas α-stable harmonizable processes are never, see [5]. Weakly mixing can be thought of as a type of asymptotic independence, and thus figuratively speaking harmonizable processes contain much more memory. This relates intuitively to the integration area $\mathbb{R}$ of harmonizable processes, whereas for moving averages we write the following

$$X_{t_2}-X_{t_1}=\int_0^{t_1}\phi(t_2-s)-\phi(t_1-s)\,d\xi_X(ds)+\int_{t_1}^{t_2}\phi(t_2-s)\,d\xi_X(ds).$$

Assuming that ϕ is decreasing, e.g. $\phi(t)\to 0$ as $t\to\infty$, the above for fixed $t_1\le t_2$ (heuristically) approximates

$$X_{t_2}-X_{t_1}\approx\int_{t_1}^{t_2}\phi(t_2-s)\,d\xi_X(ds)-\int_0^{t_1}\phi(t_1-s)\,d\xi_X(ds),$$

as $|t_1 - t_2| \to \infty$. Note that in this case the two integrals are independent which is very useful for proving limit theorems. The stronger normalization rate of the lfsm compared to the hfsm may be motivated by lfsm having a smaller Hölder index of continuity and thus necessitates stronger normalization to "smooth" its paths. However, due to the absence of H in the normalization of hfsm, path properties do not fully explain the difference in normalization rates.

1.3.1 Proof sketch of main result

In this section, we sketch the proof of Theorem 1.3.3. It motivates why the theory of multiple stochastic integration is needed. Recall that we aim to prove the convergence of

$$n^a \sum_{t=1}^{n} |X_t|^2 \to Y, \quad \text{for } a < 0,$$

to a non-trivial limit Y and wish to determine both the convergence rate a and the resulting limit Y. Consider the stochastic process $(X_t(s))_{s\in\mathbb{R}}$ defined by

$$X_t(s) = \int_{-\infty}^{s} e^{itu} g(u)\, dL_u, \quad s \in \mathbb{R} \tag{1.19}$$

where $g : \mathbb{R} \to \mathbb{R}$ is some deterministic complex-valued function and L is a complex-valued Lévy process. Then $(X_t(s))_{s\in\mathbb{R}}$ is a semi-martingale on $\mathbb{R}$, $X_t = X_t(\infty)$ and we may apply integration by parts for semi-martingales to obtain

$$|X_t|^2 = \int_{\mathbb{R}} X_t(s_-)\, d\overline{X_t}(s) + \int_{\mathbb{R}} \overline{X_t}(s_-)\, dX_t(s) + [X, \overline{X}]_\infty = V_t + U_0,$$

where

$$U_0 := [X, \overline{X}]_\infty = \int_{\mathbb{R}} |e^{its} g(s)|^2\, d([L^1] + [L^2])_s,$$
$$V_t := 2\mathfrak{Re}\left(\int_{\mathbb{R}} \int_{-\infty}^{s_-} e^{its} g(s) e^{-itu} \overline{g(u)}\, d\overline{L}_u\, dL_s \right).$$

The next step is to study each of these terms separately. Observe that U_0 has no dependence on t and will thus be an invariant component of each term in the sum for the p-variation. Hence, we have that

$$\frac{1}{n} \sum_{t=1}^{n} |X_t|^2 \xrightarrow{\mathbb{P}} U_0, \quad \text{if} \quad \frac{1}{n} \sum_{t=1}^{n} 2\mathfrak{Re}\left(\int_{\mathbb{R}} X_t(s)\, d\overline{X_t}(s) \right) \xrightarrow{\mathbb{P}} 0.$$

At this point, the theory of multiple stochastic integrals is needed. The key observation for the convergence towards zero is actually very simple. First observe that we may identify

$$\int_{\mathbb{R}} \left(\int_{-\infty}^{s_-} f(u,s)\, dL^1_u \right) dL^2_s = \int_{\mathbb{R}} \int_{\mathbb{R}} f(s,u) \mathbb{1}_{[s>u]}\, dL^1_u\, dL^2_s,$$

where the former is a semi-martingale integral and the latter is interpreted as a double stochastic integral. Continuity and linearity of multiple integrals can thus be used to study convergence. By linearity, we have that

$$\frac{1}{n}\sum_{t=1}^{n} V_t = 2\Re\mathfrak{e}\Big(\int_{\mathbb{R}}\int_{\mathbb{R}} \tilde{f}_n(s,u)\, d\overline{L}_u\, dL_s\Big),$$

where our kernel function is given as

$$\tilde{f}_n(s,u) := g(s)\overline{g(u)}\frac{1}{n}\sum_{t=1}^{n} e^{it(s-u)}\mathbb{1}_{[s>u]}$$

This kernel function converges to zero as n tends to infinity and continuity of the multiple stochastic integral implies that the proof of Theorem 1.3.3 is complete.

References

[1] Ole E Barndorff-Nielsen and Andreas Basse-O'Connor. "Quasi Ornstein–Uhlenbeck processes". *Bernoulli* 17.3 (2011), 916–941.

[2] Andreas Basse-O'Connor, Svend-Erik Graversen and Jan Pedersen. "Multiparameter processes with stationary increments: Spectral representation and integration". *Electronic Journal of Probability* 17 (2012).

[3] Andreas Basse-O'Connor, Raphaël Lachièze-Rey and Mark Podolskij. "Power variation for a class of stationary increments Lévy driven moving averages". *Ann. Probab.* 45.6B (2017), 4477–4528.

[4] Albert Benassi, Serge Cohen and Jacques Istas. "Identification and properties of real harmonizable fractional Lévy motions". *Bernoulli* 8.1 (2002), 97–115.

[5] Stamatis Cambanis, Clyde D Hardin and Aleksander Weron. "Ergodic properties of stationary stable processes". *Stochastic Processes and their Applications* 24.1 (1987), 1–18.

[6] Stamatis Cambanis and Makoto Maejima. "Two classes of self-similar stable processes with stationary increments". *Stochastic Process. Appl.* 32.2 (1989), 305–329.

[7] O. Kallenberg and J. Szulga. "Multiple integration with respect to Poisson and Lévy processes". *Probab. Theory Related Fields* 83.1-2 (1989), 101–134.

[8] Kari Karhunen. "Über die Struktur stationärer zufälliger Funktionen". *Arkiv för Matematik* 1.2 (1950), 141–160.

[9] Norio Kôno and Makoto Maejima. "Hölder continuity of sample paths of some self-similar stable processes". *Tokyo J. Math.* 14.1 (1991), 93–100.

[10] John Lamperti. "Semi-stable stochastic processes". *Transactions of the American mathematical Society* 104.1 (1962), 62–78.

[11] Benoit B. Mandelbrot and John W. Van Ness. "Fractional Brownian motions, fractional noises and applications". *SIAM Rev.* 10 (1968), 422–437.

[12] Balram S. Rajput and Jan Rosiński. "Spectral representations of infinitely divisible processes". *Probab. Theory Related Fields* 82.3 (1989), 451–487.

[13] L Chris G Rogers. "Arbitrage with fractional Brownian motion". *Mathematical Finance* 7.1 (1997), 95–105.

[14] Jan Rosiński. "On the structure of stationary stable processes". *Ann. Probab.* 23.3 (1995), 1163–1187.

[15] Gennady Samorodnitsky and Murad S. Taqqu. *Stable non-Gaussian random processes*. Stochastic Modeling. Stochastic models with infinite variance. Chapman & Hall, New York, 1994.

[16] Kazimierz Urbanik. "Random measures and harmonizable sequences". *Studia Mathematica* 31.1 (1968), 61–88.

[17] Akiva M Yaglom. *An Introduction to the Theory of Stationary Random Functions*. Courier Corporation, 2004.

Chapter 2

Local asymptotic self-similarity for heavy tailed harmonizable fractional Lévy motions

Andreas Basse-O'Connor, Thorbjørn Grønbæk and Mark Podolskij
Department of Mathematics, Aarhus University

Submitted

Abstract

In this work we characterize the local asymptotic self-similarity of harmonizable fractional Lévy motions in the heavy tailed case. The corresponding tangent process is shown to be the harmonizable fractional stable motion. In addition, we provide sufficient conditions for existence of harmonizable fractional Lévy motions.

Keywords: local asymptotic self-similarity; harmonizable processes; fractional processes; spectral representations.

2.1 Introduction

The class of self-similar stochastic processes plays a key role in probability theory as they appear in some of the most fundamental limit theorems, see [8], and in modeling they are used in geophysics, hydrology, turbulence and economics, see [17] for numerous references. This class of stochastic processes are invariant in distribution under suitable time and space scaling, that is, a stochastic process $(X_t)_{t\in\mathbb{R}}$ is called self-similar with index $H \in \mathbb{R}$ if for all $c > 0$ the two processes $(X_{ct})_{t\in\mathbb{R}}$ and $(c^H X_t)_{t\in\mathbb{R}}$ equals in finite dimensional distributions. The only self-similar centered Gaussian process with stationary increments is the fractional Brownian motion (up to scaling), which is a centered Gaussian process $(X_t)_{t\in\mathbb{R}}$ with $X_0 = 0$ a.s. and covariance function

$$\mathrm{Cov}(X_t, X_s) = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |s-t|^{2H}) \qquad \text{for all } s,t \in \mathbb{R}, \tag{2.1}$$

where $H \in (0,1)$. The fractional Brownian motion has a *moving average representation* and a *harmonizable representation*, and both lead to the same process (defined by (2.1)), for further details see Subsection 2.2. However for non-Gaussian processes their moving average and harmonizable representations are very different, see e.g. [5] and [14] for the case of α-stable processes. Only a very specific class of processes are exact self-similar, but a much larger class of processes behaves locally as a self-similar processes - this is already seen within the class of Lévy processes.

A stochastic process $(X_t)_{t\in\mathbb{R}}$ is said to be *locally asymptotically self-similar* if there exists a number $H \in \mathbb{R}$ and a non-degenerate process $(T_t)_{t\in\mathbb{R}}$ such that

$$\left(\frac{X_{\epsilon t}}{\epsilon^H}\right)_{t\in\mathbb{R}} \xrightarrow[\epsilon\to 0_+]{d} (T_t)_{t\in\mathbb{R}}, \tag{2.2}$$

where $\xrightarrow{d}$ denotes converege in finite dimensional distributions. The process $T = (T_t)_{t\in\mathbb{R}}$ is called the tangent process of X, and by (2.2), T is necessarily self-similar. Local self-similarity means that at small time-scales the stohastic process $(X_t)_{t\in\mathbb{R}}$ is approximately self-similar and may be approximated by its tangent process. This property was introduced to provide a more flexible modeling framework compared to global self-similarity. For applications, it has been used to study the behaviour of flows, see [6] and [15], and for showing high frequency asymptotic results, see [2] or [3].

Moving average fractional Lévy motions: Starting from the moving average representation of the fractional Brownian motion, [9] has, among many others, studied fractional Lévy processes defined as

$$X_t = \int_{-\infty}^{t} \Big((t-s)_+^\beta - (-s)_+^\beta \Big) dL_s, \qquad t \in \mathbb{R}, \tag{2.3}$$

where $\beta \in (0, 1/2)$ and $(L_t)_{t\in\mathbb{R}}$ is a centered Lévy process with finite second moment. Throughout this paper $x_+ := \max\{x, 0\}$ and $x_- := -\min\{x, 0\}$ denote the positive and negative parts of any number $x \in \mathbb{R}$.

In the following, we will call such processes for *moving averages fractional Lévy motions* to distinct them from their harmonizable counterpart. Under a regular variation assumption on the Lévy measure of L near zero, [9] shows that a moving average fractional Lévy motion is never self-similar, but it is locally asymptotically self-similar with tangent process the linear fractional stable motion, which is a process of the form (2.3) with L being an α-stable Lévy process, cf. [5] and Theorems 4.4 and 4.5 of [9].

Harmonizable fractional Lévy motions: Next we define the class of harmonizable fractional Lévy motions which includes the harmonizable fractional stable motion introduced in [5].

Definition 2.1.1. *A stochastic process $(X_t)_{t\in\mathbb{R}}$ is called a harmonizable fractional Lévy motion with parameters $(\alpha, H) \in \mathbb{R}^2$ if*

$$X_t = \int_{\mathbb{R}} \frac{e^{its} - 1}{is} \Big(a(s_+)^{-H-1/\alpha+1} + b(s_-)^{-H-1/\alpha+1} \Big) dL_s, \quad t \in \mathbb{R} \tag{2.4}$$

where L is a rotationally invariant complex-valued Lévy process, and $a, b \in \mathbb{R}$.

The (over) parametrization in Definition 2.1.1 is chosen due to our forthcoming Assumption (A). In fact under Assumption (A) below, the H parameter in Definition 2.1.1 turns out to be exactly the number H in the definition of local asymptotic self-similarity. From Theorem 2.4.1, below, it follows that the harmonizable fractional Lévy motions have stationary increments and rotational invariant distributions. Furthermore, we give concrete conditions for existence of the harmonizable fractional Lévy motion on (α, H) and the Lévy measure of L.

In [4], local asymptotic self-similarity is studied for a slightly different class of harmonizable fractional motions under the assumption that all moments are finite, e.g. the Lévy measure ν of the Lévy process L satisfies that

$$\int_{|x|>1} |x|^p \, \nu(dx) < \infty \qquad \text{for all } p > 0. \tag{2.5}$$

Their result is the following:

Theorem 2.1.2 (Benassi, Cohen and Istas)**.** *Let $(X_t)_{t\in\mathbb{R}}$ denote a harmonizable fractional Lévy motion as in Definition 2.3 of [4] satisfying the moment condition (2.5). Then the process X is locally asymptotically self-similar with index H and tangent process the fractional Brownian motion, that is,*

$$\left(\frac{X_{\epsilon t}}{\epsilon^H}\right)_{t\in\mathbb{R}} \xrightarrow[\epsilon\to 0_+]{d} (c_0 B_t^H)_{t\in\mathbb{R}}. \tag{2.6}$$

where $(B_t^H)_{t\in\mathbb{R}}$ is a fractional Brownian motion with Hurst index H and c_0 is a suitable constant.

The main aim of this work is to characterize the local asymptotic self-similarity of the harmonizable fractional Lévy motion when L has heavy tails, violating the moment condition (2.5). The methods of [4] rely heavily on power series expansion of the characteristic function which is only available under the assumption (2.5). Instead of this assumption, we consider the case where the Lévy measure ν is regular varying in the following sense.

Assumption (A): *Suppose that L is a rotationally invariant complex-valued Lévy process without Gaussian component and let ν denote its Lévy measure. We assume that ν is absolutely continuous with respect to the two dimensional Lebesgue measure with a density $f : \mathbb{R}^2 \to \mathbb{R}_+$ satisfying*

$$\begin{aligned} f(x) &\sim c_0 \|x\|^{-2-\alpha} \quad \text{as } \|x\| \to \infty, \\ f(x) &\le C\|x\|^{-2-\alpha} \quad \text{for all } x \in \mathbb{R}^2, \end{aligned}$$

where $c_0, C > 0$ and $0 < \alpha < 2$.

The following theorem, which is the main result of this paper, characterizes the local asymptotic self-similarity of harmonizable fractional Lévy motions in the heavy-tailed case, and additional provides an existence result for them.

Theorem 2.1.3. *Let $(\alpha, H) \in (0,2) \times (0,1)$ and suppose that Assumption (A) is satisfied. Then the harmonizable fractional Lévy motion $(X_t)_{t\in\mathbb{R}}$, defined in* (2.4)*, is well-defined and it is locally asymptotically self-similar with index H and tangent process the harmonizable fractional stable motion, that is,*

$$\left(\frac{X_{\epsilon t}}{\epsilon^H}\right)_{t\in\mathbb{R}} \xrightarrow[\epsilon\to 0_+]{d} (C_t)_{t\in\mathbb{R}}, \tag{2.7}$$

where the convergence is in finite dimensional distributions and $(C_t)_{t\in\mathbb{R}}$ denotes a harmonizable fractional stable motion with parameters (α, H), which is defined in (2.4) *with L being a complex-valued rotationally invariant α-stable Lévy process.*

The choice of constants for $(C_t)_{t\in\mathbb{R}}$ can be found by examining the proof. We note that the tangent process in Theorem 2.1.3 differs from the tangent processes appearing in Theorem 2.1.2 and Theorem 4.5 of [9]. From this we infer that it is the behaviour of the Lévy measure of L close to zero which dominates in the moving average setting, whereas it is the behaviour of the Lévy measure of L far away from zero which dominates in the harmonizable setting. The structure of the paper is as follows: Section 2.2 explains the role played by harmonizable processes within the class of stationary processes. Section 2.3 introduces complex random measures, their integration and provide existence criterias for harmonizable processes. Finally, at the end of the last section, we present the proof of Theorem 2.1.3.

2.2 Background on harmonizable processes

Stationary processes are one of the main classes of stochastic processes. For stationary, centered Gaussian processes, it is well-known that every L^2-continuous process $(X_t)_{t\in\mathbb{R}}$ has a *harmonizable representation* of the form

$$X_t = \int_{\mathbb{R}} e^{its} M(ds), \quad t \in \mathbb{R}, \tag{2.8}$$

for some complex-valued Gaussian random measure M defined on $\mathbb{R}$. Furthermore, a rather large class of these processes have, in addition, a *moving average representation*, that is, a representation of the form

$$X_t = \int_{\mathbb{R}} g(t-s)\, dB_s, \quad t \in \mathbb{R}, \tag{2.9}$$

where g is a deterministic function and $(B_t)_{t\in\mathbb{R}}$ is a two-sided real-valued Brownian motion. (Note that, the Brownian motion may be viewed as a shift-invariant Gaussian random measure.) Indeed, the class of Gaussian processes having a moving average representation corresponds exactly to those processes with absolute continuous spectral measure μ. Recall that the spectral measure μ is given by $\mu(A) = \mathbb{E}[|M(A)|^2]$ for $A \in \mathcal{B}(\mathbb{R})$, where M is given in (2.8). These classical results can be found in e.g. [7] or [18].

The only centered Gaussian self-similar process with stationary increments is the fractional Brownian motion $(B_t^H)_{t\in\mathbb{R}}$ with Hurst index $H \in (0,1)$, and

as already mentioned in the introduction, this process has the following two representations

$$B_t^H = \int_{\mathbb{R}} \Big((t-s)_+^{H-1/2} - (-s)_+^{H-1/2} \Big) dB_s, \quad \text{("moving average representation")}$$
$$B_t^H = \int_{\mathbb{R}} \frac{e^{its}-1}{is} |s|^{-H-1/2+1} M(ds), \qquad \text{("harmonizable representation")},$$

which yields the same process in distribution, see Chapter 7.2 of [14] for further details. Hence, the fractional Gaussian noise $(B_n^H - B_{n-1}^H)_{n\in\mathbb{Z}}$ has both a harmonizable, (2.8), and a moving average, (2.9), representation. For comparison we will discuss the structure of stationary α-stable processes with $\alpha \in (0,2)$ in the following.

In sharp contrast to the Gaussian situation the class of α-stable stationary increments self-similar processes, $\alpha \in (0,2)$, is huge, and is far from being understood by now. However, two natural generalizations of the fractional Brownian motion to the α-stable setting are proposed in [5] generalizing the fractional Brownian motion to α-stabe processes by replacing the driving Gaussian random measure with an α-stable random measure in its moving average and harmonizable representations. This leads to the *harmonizable fractional stable motion* $(X_t)_{t\in\mathbb{R}}$, which is defined as

$$X_t = \int_{\mathbb{R}} \frac{e^{its}-1}{is} \Big(a(s_+)^{-H-1/\alpha+1} + b(s_-)^{-H-1/\alpha+1} \Big) dL_s, \quad t \in \mathbb{R},$$

where $(L_t)_{t\in\mathbb{R}}$ is a two-sided, complex-valued, α-stable, rotationally invariant Lévy process, and to the *linear fractional stable motion* $(X_t)_{t\in\mathbb{R}}$, which is defined as

$$X_t = \int_{\mathbb{R}} a\Big((t-s)_+^{H-1/\alpha} - (-s)_+^{H-1/\alpha} \Big) + b\Big((t-s)_-^{H-1/\alpha} - (-s)_-^{H-1/\alpha} \Big) dL_s,$$

where $(L_t)_{t\in\mathbb{R}}$ is a two-sided, real-valued, α-stable, symmetric Lévy process. Notice that corresponding noise processes $(X_n - X_{n-1})_{n\in\mathbb{Z}}$ for the linear and harmonizable fractional stable motions are moving averages and harmonizable processes, respectively.

Indeed, the Gaussian assumption is crucial for the above equality between the harmonizable and moving average representations to hold, as it turns out that harmonizable fractional stable motion and linear fractional stable motion as quite different processes, cf. [5] and [14]. The seminal paper [11] shows that every stationary α-stable process has a *unique* decomposition into a (mixed) moving average component, a harmonizable component and a process of the "third kind", which does not admit moving average nor harmonizable components. The class of mixed moving averages may be viewed as the class of processes having the least memory, whereas class of harmonizable processes is the class having the largest degree of memory, and the processes of the third kind are in between. These facts come from ergodic consideration, see the introduction of [12] for more details, and are also illustrated by the fact that

moving averages are always mixing and harmonizable processes are never ergodic nor mixing. Hence by studying moving averages and harmonizable processes, we are examining the two extremes of stationary α-stable processes.

Thus the comparison of results on local asymptotical self-similarity in the introduction between linear fractional stable motions and harmonizable fractional stable motions are, in fact, a comparison between α-stable self-similar stationary increments processes with the least memory and with the most memory. This encircles the local asymptotical behaviour of general α-stable self-similar processes with stationary increments.

2.3 Preliminaries on complex stochastic integration theory

All random variables and processes will be defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A real-valued stochastic variable X is symmetric α-stable ($S\alpha S$) if for some $\alpha \in (0, 2]$, the characteristic function of X satisfies

$$\mathbb{E}[\exp\{itX\}] = \exp(-\sigma^\alpha |t|^\alpha), \qquad \text{for all } t \in \mathbb{R},$$

for some parameter $\sigma > 0$ called the scale parameter. If $\alpha = 2$, then X has a centered Gaussian distribution and σ^2 is the variance of X. Rotationally invariant random variables and processes are defined as follows:

Definition 2.3.1. *A complex-valued random variable X is* rotationally invariant *if*

$$e^{i\theta} X \overset{d}{=} X, \quad \textit{for all } \theta \in [0, 2\pi), \tag{2.10}$$

where $\overset{d}{=}$ denotes equality in distribution. Similarly, a complex-valued stochastic process $(X_t)_{t\in T}$ is rotationally invariant *if every complex linear combination is rotationally invariant, e.g. $\sum_{n=1}^{N} z_n X_{t_n}$ is rotationally invariant.*

Rotational invariance is called isotropy in some references but due to the ambiguity of isotropy we chose to use rotational invariance, cf. the discussion in Example 1.1.6 of [13]. A complex-valued process can equivalently be regarded as a $\mathbb{R}^2$-valued random variable, in which case rotational invariance is invariance in distribution wrt. rotation matrices. We will with some ambiguity switch between the $\mathbb{C}$ and $\mathbb{R}^2$. From the definition it is immediate that a rotationally invariant random variable $X = X_1 + iX_2$ is symmetric and furthermore if it is infinitely divisible, then X_1 and X_2 share the same Lévy measure ν. Let $\mathcal{B}(\mathbb{R})$ denote the Borel sets on $\mathbb{R}$, $\mathcal{B}_b(\mathbb{R})$ the bounded Borel sets on $\mathbb{R}$ and $L^0_{\mathbb{C}}(\Omega)$ the complex-valued random variables. For completeness, we define complex-valued infinitely divisible random measures and state well known stochastic integration results, cf. [16] and [10].

Definition 2.3.2 (Complex-valued random measure)**.** *A complex-valued random measure is by definition a complex-valued set function*

$$M : \mathcal{B}_b(\mathbb{R}) \to L^0_{\mathbb{C}}(\Omega),$$

such that for disjoints sets $A_1, A_2, \ldots \in \mathcal{B}_b(\mathbb{R})$, the complex-valued random variables

$$M(A_1), M(A_2), \ldots$$

are independent and infinitely divisible, and if $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{B}_b(\mathbb{R})$ then

$$M\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} M(A_n) \quad a.s.,$$

where the series converges almost surely.

Given a complex-valued random measure M we can find a σ-finite deterministic measure λ on $\mathbb{R}$ such that $\lambda(A_n) \to 0$ implies $M(A_n) \to 0$ in probability. We call λ a control measure for the random measure M. Letting $\nu_A(\cdot)$ denote the Lévy measure of $M(A)$, we can then apply Proposition 2.4 of [10] to obtain a decomposition such that

$$F(A \times B) := \nu_A(B) = \int_{\mathbb{R}} \int_{\mathbb{R}^2} \mathbb{1}_{A\times B}(s,x)\, \rho(s,dx)\lambda(ds),$$

where $\{\rho(s,dx)\}_{s\in\mathbb{R}}$ denotes a family of Lévy measures on $\mathbb{R}^2$. For the rest of the paper, we shall use the notation

$$K(\theta,s) := \int_{\mathbb{R}^2} \Big[e^{i\langle\theta,x\rangle} - 1 - \mathbb{1}_{\{\|x\|\le 1\}}\langle\theta,x\rangle\Big]\rho(s,dx), \quad (\theta,s) \in \mathbb{R}^2 \times \mathbb{R}. \tag{2.11}$$

A simple complex-valued function $f : \mathbb{R} \to \mathbb{C}$ is a function of the (canonical) form

$$f(s) = \sum_{j=1}^{n} z_j \mathbb{1}_{A_j}, \tag{2.12}$$

where $n \in \mathbb{N}$, $z_1, \ldots, z_n$ are complex numbers and $A_1, \ldots, A_n$ are disjoint sets from $\mathcal{B}_b(\mathbb{R})$. For a simple function f, of the form (2.12), and $A \in \mathcal{B}(\mathbb{R})$ we define

$$\int_A f\, dM = \sum_{j=1}^{n} z_j M(A \cap A_j).$$

A (general) measurable function $f : \mathbb{R} \to \mathbb{C}$ is said to be M-integrable, if there exists a sequence of simple function $\{f_n\}_{n\in\mathbb{N}}$ such that

(i) $f_n \to f$, λ-almost surely.

(ii) for every $A \in \mathcal{B}(\mathbb{R})$, the sequence $\{\int_A f_n\, dM\}_{n\in\mathbb{N}}$ converges in probability, as $n \to \infty$.

In the affirmative case, we define

$$\int_A f\, dM := \mathbb{P} - \lim_{n\to\infty} \int_A f_n\, dM,$$

where $\{f_n\}$ satisfies (i) and (ii) and $\mathbb{P}-\lim$ denotes limit in probability. It can be shown that this definition does not depend on the approximating sequence $\{f_n\}$. For further details on stochastic integration theory we refer to [10], [13], [14] and [16]. In the following $\mathfrak{Re}(z), \mathfrak{Im}(z)$ denotes real, respectively imaginary, part of a complex number z.

Theorem 2.3.3.
(a): Let $f : \mathbb{R} \to \mathbb{C}$ be a measurable function. Write $f = f_1 + if_2$. Then f is M-integrable if the following condition hold true

$$\int_{\mathbb{R}} \int_{\mathbb{R}^2} \min\left(1, \left[\|\Big(f_1(s)+f_2(s),\ f_1(s)-f_2(s)\Big)\|^2\right]\|x\|^2\right)\rho(s,dx)\lambda(ds) < \infty,$$

and, in the affirmative case, the characteristic function of $\int_{\mathbb{R}} f\, dM$ is given by

$$\mathbb{E}\left[\exp\left(i\left\{\theta_1 \mathfrak{Re}\left(\int_{\mathbb{R}} f\, dM\right) + \theta_2 \mathfrak{Im}\left(\int_{\mathbb{R}} f\, dM\right)\right\}\right)\right]$$
$$= \exp\left(\int_{\mathbb{R}} K\Big(\theta_1 f_1(s) + \theta_2 f_2(s),\ \theta_2 f_1(s) - \theta_1 f_2(s)\Big), s\Big)\lambda(ds)\right).$$

(b): Suppose $f_1, \ldots, f_n$ are M-integrable. The joint characteristic function is given by

$$\mathbb{E}\left[\exp\left\{i\sum_{j=1}^{n}\left(\theta_j^{(1)}\mathfrak{Re}\left(\int f_j\, dM\right) + \theta_j^{(2)}\mathfrak{Im}\left(\int f_j\, dM\right)\right)\right\}\right]$$
$$= \exp\left(\int_{\mathbb{R}} K\left(\sum_{j=1}^{n}\theta_j^{(1)} f_{j,1} + \theta_j^{(2)} f_{j,2}, \sum_{j=1}^{n}\theta_j^{(2)} f_{j,1} - \theta_j^{(1)} f_{j,2}, s\right)\lambda(ds)\right).$$

(c): Let $M = M^{(1)} + iM^{(2)}$ be a rotationally invariant complex-valued random measure and let $f : \mathbb{R} \to \mathbb{C}$ be a measurable function. Then the following integrals exists simultaneously and are equal in distribution:

$$\int_{\mathbb{R}} f\, dM \overset{d}{=} \int_{\mathbb{R}} \|f\|\, dM = \int_{\mathbb{R}} \|f\|\, dM^{(1)} + i\int_{\mathbb{R}} \|f\|\, dM^{(2)}. \tag{2.13}$$

Proof. (a) follows from the same steps as Theorem 2.7 in [10] using complex-valued functions instead. (b) follows by the same steps as in the proof for Proposition 6.2.1(iii) of [14]. (c) follows by closely examining the results and arguments in [16]. □

Often it is easier to think of the complex-valued stochastic integral as

$$\int_{\mathbb{R}} f\, dM_s = \int_{\mathbb{R}} (f_1 + if_2)\, d(M^{(1)} + iM^{(2)})$$
$$= \int_{\mathbb{R}} f_1\, dM^{(1)} - \int_{\mathbb{R}} f_2\, dM^{(2)} + i\left(\int_{\mathbb{R}} f_1\, dM^{(2)} + \int_{\mathbb{R}} f_2\, dM^{(1)}\right)$$

and show existence for each of the above four integrals separately (this is a more strict existence criterion). As a consequence of (*c*), it is also necessary to prove existence of all of these four integrals, when M is a rotationally invariant random measure.

2.4 Existence and properties of harmonizable fractional Lévy motions

Recall that a harmonizable fractional Lévy motion $(X_t)_{t\in\mathbb{R}}$ is defined by

$$X_t = \int_{\mathbb{R}} \frac{e^{its}-1}{is}\Big(a(s_+)^{-H-1/\alpha+1} + b(s_-)^{-H-1/\alpha+1}\Big)\,dL_s, \quad t\in\mathbb{R},$$

where L is a rotational invariant complex-valued Lévy process. Our next result gives a general existence criterion for harmonizable fractional Lévy motions together with some properties.

Theorem 2.4.1. *Let $(L_t)_{t\in\mathbb{R}}$ be a complex-valued rotational invariant Lévy process without Gaussian component. The harmonizable fractional Lévy motion $(X_t)_{t\in\mathbb{R}}$, defined in Definition 2.1.1, with parameters $(\alpha, H)\in(0,2)\times(0,1)$ exists if both of the following (a)–(b) are satisfied:*

1. $\int_{|x|>1} |x|^{\frac{1}{H+1/\alpha}}\,\nu_R(dx) < \infty$,
2. $\int_{|x|\le 1} |x|^{\frac{1}{H+1/\alpha-1}}\,\nu_R(dx) < \infty$,

where ν_R denotes the Lévy measure of the real-part of $(L_t)_{t\in\mathbb{R}}$. Furthermore, if X exists, then it has stationary increments, rotational invariant distribution and the characteristic function is given by

$$\mathbb{E}\Big[\exp\Big\{i\Big\langle\theta,\Big(\mathfrak{Re}(X_t),\mathrm{Im}(X_t)\Big)\Big\rangle\Big\}\Big] = \exp\bigg(\int_{\mathbb{R}} K\Big(\theta_1 f_1(s)+\theta_2 f_2(s),\ \theta_2 f_1(s)-\theta_1 f_2(s), s\Big)\,\lambda(ds)\bigg),$$

for all $\theta=(\theta_1,\theta_2)\in\mathbb{R}^2$, where K is given by (2.11).

To prove Theorem 2.4.1 we will first show the following lemma. In this result, and in the following, we will write $f(t)\sim g(t)$ as $t\to a$ for real-valued functions f and g, if $\lim_{t\to a}(f(t)/g(t)) = c$ for some constant $c\neq 0$.

Lemma 2.4.2. *Let L be a real-valued symmetric Lévy process without Gaussian component and Lévy measure ν. Let $f:\mathbb{R}\to\mathbb{R}$ be a measurable function, bounded on $[-1,1]^c$, and satisfying*

$$|f(s)|\sim|s|^{\beta} \text{ as } s\to 0 \qquad \text{and} \qquad |f(s)|\sim|s|^{-\gamma} \text{ as } |s|\to\infty, \tag{2.14}$$

for some $\beta\le 0$ and $\gamma>0$. Then the stochastic integral $\int f\,dL$ exists if and only if the following two conditions (a) and (b) are satisfied:

1. *$\gamma > 1/2$ and the following condition hold true*

$$\int_{|x|>1} |x|^{\frac{1}{\gamma}}\,\nu(dx) < \infty.$$

2. *We have that*

$$\int_{|x|\leq 1} |x|^{\frac{1}{-\beta}}\, \nu(dx) < \infty.$$

If $\sim$ in Lemma 2.4.2 is replaced by $f(s) = O(|s|^{\beta})$ as $s \to 0$, or $f(s) = O(|s|^{-\gamma})$ as $|s| \to \infty$, the criteria for existence of the integral $\int f\, dL$ remain sufficient. Note that if $\beta > -1/2$, the second criterion holds for any Lévy measure.

Proof of Lemma 2.4.2. Writing out the conditions in Theorem 2.7 of [10] and observing that these are increasing in the function f, it suffices to study these conditions for a function $g(s) \coloneqq \mathbb{1}_{[-1,1]}(s)|s|^{\beta} + \mathbb{1}_{[-1,1]^c}(s)|s|^{-\gamma}$. Recall that the general condition for existence of $\int g\, dL$ is given by

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \min(1, |xg(s)|^2)\, \nu(dx)\, \lambda(ds) < \infty, \tag{2.15}$$

where ν denotes the Lévy measure of L and λ denotes the Lebesgue measure. Divide this condition into the following four areas,

$$\begin{aligned}
A_{11} &= \{(s,x) \in \mathbb{R}\times\mathbb{R} : |s| \leq 1,\ |x| \leq 1\},\\
A_{12} &= \{(s,x) \in \mathbb{R}\times\mathbb{R} : |s| \leq 1,\ |x| > 1\},\\
A_{21} &= \{(s,x) \in \mathbb{R}\times\mathbb{R} : |s| > 1,\ |x| \leq 1\},\\
A_{22} &= \{(s,x) \in \mathbb{R}\times\mathbb{R} : |s| > 1,\ |x| > 1\}.
\end{aligned}$$

The monotonicity of g on these sets can then be used to simplify the condition in (2.15) into 1 and 2. We first consider A_{22} and let $x \in [-1,1]^c$ be given. Divide the inner integral into

$$\begin{aligned}
&\int_{\{|s|>1\}\cap\{|s|>|x|^{1/\gamma}\}} |x|^2|s|^{-2\gamma}\, \lambda(ds) + \int_{\{|s|>1\}\cap\{|s|\leq|x|^{1/\gamma}\}} 1\, \lambda(ds)\\
&\quad = |x|^2 \int_{|x|^{1/\gamma}}^{\infty} |s|^{-2\gamma}\, \lambda(ds) + 2\lambda((1, |x|^{1/\gamma}]),\\
&\quad = |x|^2 \left[\frac{2}{-2\gamma+1} s^{-2\gamma+1}\right]_{|x|^{1/\gamma}}^{\infty} + 2(|x|^{1/\gamma} - 1)\\
&\quad = |x|^2 |x|^{-2+1/\gamma} \frac{-2}{-2\gamma+1} + 2(|x|^{1/\gamma} - 1) = 3|x|^{1/\gamma} - 2,
\end{aligned}$$

where we have used that $\gamma > 1/2$ to ensure the finiteness of the integral and afterwards that $\frac{2}{-2\gamma+1} < 0$. Inserting the derived into the original criterion on the set A_{22}, we get that

$$\int_{|x|>1} \left[3|x|^{1/\gamma} - 2\right] \nu(dx) < \infty.$$

Since the area $|x| > 1$ is of finite ν-measure, this reduces to

$$\int_{|x|>1} |x|^{1/\gamma}\, \nu(dx) < \infty,$$

which is one of the stated criterions. For A_{11}, let $x \in [-1,1]$ be given and assume that $\beta < 0$. The inner integral can in this case be written as

$$\begin{aligned}
&\int_{\{|s|\le 1\}\cap\{|s|^\beta \le |x|^{-1}\}} |xg(s)|^2 \, \lambda(ds) + \int_{\{|s|\le 1\}\cap\{|s|^\beta > |x|^{-1}\}} \lambda(ds) \\
&= |x|^2 \int_{1\ge|s|\ge|x|^{-1/\beta}} |s|^{2\beta} \, \lambda(ds) + \int_{\{|s|\le 1\}\cap\{|s|\le|x|^{-1/\beta}\}} \lambda(ds) \\
&= |x|^2 \frac{2}{2\beta+1}\Big[s^{2\beta+1}\Big]_{|x|^{-1/\beta}}^{1} + 2\lambda([0,|x|^{-1/\beta}]).
\end{aligned}$$

Inserting this into the outer integral we obtain

$$\int_{|x|\le 1} \left(|x|^2 \frac{2}{2\beta+1}\Big[s^{2\beta+1}\Big]_{|x|^{-1/\beta}}^{1} + 2\lambda([0,|x|^{-1/\beta}]) \right) \nu(dx)$$

which reduces to the second condition by applying the definition of a Lévy measure. For $\beta = 0$, the proof is trivial. For A_{12}, let $x \in [-1,1]^c$ be given. We can again rewrite the inner integral into

$$\begin{aligned}
&\int_{\{|s|\le 1\}\cap\{|s|^{-\gamma} \le |x|^{-1}\}} |x|^2 |s|^{-2\gamma} \lambda(ds) + \int_{\{|s|\le 1\}\cap\{|s|^{-\gamma} > |x|^{-1}\}} \lambda(ds) \\
&= \int_{\{|s|\le 1\}\cap\{|s|\ge|x|^{1/\gamma}\}} |x|^2 |s|^{-2\gamma} \, \lambda(ds) + \int_{\{|s|\le 1\}\cap\{|s|<|x|^{1/\gamma}\}} \lambda(ds) \\
&= 0 + \lambda([0,1]),
\end{aligned}$$

where we used that $|x| > 1$. Inserting this into the outer integral reduces to a trivial condition for Lévy measures. For the last area, A_{21}, let $x \in [-1,1]$ be given. In this case the condition similarly reduces to

$$\int_{|x|\le 1} |x|^2 \, \nu(dx) < \infty,$$

which is trivial. This concludes the proof. □

Proof of Theorem 2.4.1. Let f denote the integrand of the harmonizable fractional Lévy motion. Observe that f is bounded on $[-1,1]^c$, and

$$f(s) = O(|s|^{-H-1/\alpha}) \text{ as } |s| \to \infty, \quad \text{and} \quad f(s) = O(|s|^{1-H-1/\alpha}) \text{ as } s \to 0.$$

The existence criteria now follows by Lemma 2.4.2. The stationary increments follows by a straightforward extension of Theorem 4.1 in [16] to continuous time, see also Theorem 6.5.1 in [14] for the stable case. The isotropic distribution follows immediately from (c) in Theorem 2.3.3. □

We are now ready to complete the proof of our main result.

Proof of Theorem 2.1.3. We study the characteristic function of the finite dimensional distributions for the left-hand side of (2.7) and show convergence towards the characteristic function of harmonizable fractional stable motion.

The characteristic function for the finite dimensional distribution of (2.7) is given by Theorem 2.3.3. For $(\theta_j^{(1)}, \theta_j^{(2)}) \in \mathbb{R}^2$ for $j = 1, \dots, n$, we have that

$$\begin{aligned} A_\epsilon &:= \log \mathbb{E}\left[\exp\left(i \sum_{j=1}^{n} \left[\theta_j^{(1)} \frac{\mathfrak{Re}(X(\epsilon t_j))}{\epsilon^H} + \theta_j^{(2)} \frac{\mathfrak{Im}(X(\epsilon t_j))}{\epsilon^H} \right] \right)\right] \\ &= \int_{\mathbb{R}} \psi\Bigg(\epsilon^{-H} \sum_{j=1}^{n} \theta_j^{(1)} f_{\epsilon t_j,1}(s) + \epsilon^{-H} \sum_{j=1}^{n} \theta_j^{(2)} f_{\epsilon t_j,2}(s), \\ &\qquad \epsilon^{-H} \sum_{j=1}^{n} \theta_j^{(2)} f_{\epsilon t_j,1}(s) - \epsilon^{-H} \sum_{j=1}^{n} \theta_j^{(1)} f_{\epsilon t_j,2}(s) \Bigg) ds, \end{aligned} \tag{2.16}$$

where $f_{\epsilon t_j,1}, f_{\epsilon t_j,2}$ denotes the real, respectively imaginary, part of integrand $f_{\epsilon t_j}$ for $X_{\epsilon t_j}$ and with $z = (z_1, z_2)$

$$\psi(z_1, z_2) := \int_{\mathbb{R}^2} \left[e^{i\langle z,x\rangle} - 1 - \mathbb{1}_{\{\|x\| \le 1\}}(x) \langle z, x\rangle \right] \nu(dx)$$

Writing $u = \epsilon s$, we substitute the ϵ out of the time index of f and obtain

$$f_{\epsilon t}(s) = \frac{e^{i\epsilon t s} - 1}{is} \left(a(s_+)^{-H-1/\alpha+1} + b(s_-)^{-H-1/\alpha+1} \right) = f_t(u) \epsilon^{H+1/\alpha}.$$

Making the substitution $u = \epsilon s$ in equation (2.16) thus yields that

$$\begin{aligned} A_\epsilon = \int_{\mathbb{R}} \psi\Bigg(&\epsilon^{H+1/\alpha-H} \left(\sum_{j=1}^{n} \theta_j^{(1)} f_{t_j,1}(u) + \sum_{j=1}^{n} \theta_j^{(2)} f_{t_j,2}(u) \right), \\ &\epsilon^{H+1/\alpha-H} \left(\sum_{j=1}^{n} \theta_j^{(2)} f_{t_j,1}(u) - \sum_{j=1}^{n} \theta_j^{(1)} f_{t_j,2}(u) \right) \Bigg) \epsilon^{-1} \, du. \end{aligned}$$

To simplify notation, define $g_{\theta,t}(u) \in \mathbb{R}^2$ by

$$\left(\left(\sum_{j=1}^{n} \theta_j^{(1)} f_{t_j,1}(u) + \sum_{j=1}^{n} \theta_j^{(2)} f_{t_j,2}(u) \right), \left(\sum_{j=1}^{n} \theta_j^{(2)} f_{t_j,1}(u) - \sum_{j=1}^{n} \theta_j^{(1)} f_{t_j,2}(u) \right) \right), \tag{2.17}$$

and let $k(z, x) := e^{i\langle z,x\rangle} - 1 - i\mathbb{1}_{D^c}(x)\langle z, x\rangle$. Inserting the defined notation, this implies that we may rewrite the characteristic function to

$$\begin{aligned} A_\epsilon &= \int_{\mathbb{R}} \int_{\mathbb{R}^2} k(\epsilon^{1/\alpha} g_{\theta,t}(u), x) \, \nu(dx) \epsilon^{-1} \, du \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}^2} k(\epsilon^{1/\alpha} g_{\theta,t}(u), x) \, f(x) \, dx \, \epsilon^{-1} \, du \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}^2} k(g_{\theta,t}(u), x) f(\epsilon^{-1/\alpha} x) \epsilon^{2(-1/\alpha)} \, dx \, d\epsilon^{-1} \, du, \end{aligned} \tag{2.18}$$

where we used $\nu(dx) = f(x)\,dx$ and a simple scaling of parameters in $\mathbb{R}^2$. The next step is to show pointwise convergence of the integrand as $\epsilon \to 0_+$. After

this we apply the dominated convergence theorem to insert the found limit under the integral. We postpone the argument for dominated convergence theorem until the end of this proof. Assumption (A) on the Lévy measure ν gives us that for every $\delta > 0$ we can find $R_\delta > 0$ such that

$$1 - \delta \leq \frac{f(x)}{\|x\|^{-2-\alpha}} \leq 1 + \delta, \quad \text{for } \|x\| \geq R_\delta.$$

Fix $x \in \mathbb{R}^2 \setminus \{0\}$ and $u \in \mathbb{R}$. For every $\delta > 0$ we can choose ϵ sufficiently small such that $\|\epsilon^{-1/\alpha} x\| > R_\delta$, which implies that

$$1 - \delta \leq \frac{f(\epsilon^{-1/\alpha} x) e^{-2/\alpha - 1}}{\|x\|^{-2-\alpha}} = \frac{f(\epsilon^{-1/\alpha} x)}{\|\epsilon^{-1/\alpha} x\|^{-2-\alpha}} \leq 1 + \delta.$$

Thus in the limit we find that

$$\lim_{\epsilon \downarrow 0} f(\epsilon^{-1/\alpha} x) \epsilon^{-2/\alpha - 1} = \|x\|^{-2-\alpha},$$

and hence,

$$\lim_{\epsilon \downarrow 0} k(g_{\theta,t}(u), x) f(\epsilon^{-1/\alpha} x) \epsilon^{-2/\alpha - 1} = \|x\|^{-2-\alpha} k(g_{\theta,t}(u), x).$$

This finishes the proof of pointwise convergence for f_ϵ as $\epsilon \to 0$. Applying the dominated convergence theorem we find that

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \log \mathbb{E}\left[\exp\left(i \sum_{j=1}^{n} \left[\theta_j^{(1)} \frac{\mathfrak{Re}(Y(\epsilon t_j))}{\epsilon^K} + \theta_j^{(2)} \frac{\mathfrak{Im}(Y(\epsilon t_j))}{\epsilon^K}\right]\right)\right] \\
&\quad = \lim_{\epsilon \downarrow 0} \int_{\mathbb{R}} \int_{\mathbb{R}^2} k(g_{\theta,t}(u), x) f(e^{-1/\alpha} x) \epsilon^{-2/\alpha - 1} \, dx \, du \\
&\quad = \int_{\mathbb{R}} \int_{\mathbb{R}^2} k(g_{\theta,t}(u), x) \|x\|^{-2-\alpha} \, dx \, du. \qquad (2.19)
\end{aligned}$$

The book [1], p. 37, identifies $\|x\|^{-2-\alpha}$ in (2.19) as the Lévy measure of a rotationally invariant two-dimensional α-stable Lévy process. We can continue our derivations in polar coordinates and observe that the inner integral may be rewritten as

$$\begin{aligned}
&\int_{\mathbb{R}^2} k(g_{\theta,t}(u), x) \|x\|^{-2-\alpha} \, dx \\
&= \int_0^{2\pi} \int_0^{\infty} k(g_{\theta,t}(u), r(\cos(s), \sin(s))) \, r^{-1-\alpha} \, dr \, ds \\
&= \int_0^{2\pi} -c_0 |\langle g_{\theta,t}(u), \big(\cos(s), \sin(s)\big)\rangle|^\alpha \, ds.
\end{aligned}$$

Here we used the following result, which follows by substituting $z = yr$,

$$-c_0 |y|^\alpha = \int_0^{\infty} \Big[\exp(iyr) - 1 - i \mathbb{1}_{\{\|r\| \leq 1\}}(x) yr\Big] r^{-1-\alpha} \, dr,$$

where $c_0 := \int_0^\infty [\cos(r) - 1] r^{-1-\alpha}\, dr$. Write $g_{\theta,t}(u) = \|g_{\theta,t}(u)\|(\cos(\kappa_u), \sin(\kappa_u))$ in polar form for some κ_u. Inserting this notation and applying a standard trigonometric rule, we obtain

$$\begin{aligned}
&\int_0^{2\pi} -c_0 |\langle g_{\theta,t}(u), \big(\cos(s), \sin(s)\big)\rangle|^\alpha \, ds. \\
&= -c_0 \|g_{\theta,t}(u)\|^\alpha \int_0^{2\pi} |\langle (\cos(\kappa_u), \sin(\kappa_u)), (\cos(s), \sin(s))\rangle|^\alpha \, ds \\
&= -c_0 \|g_{\theta,t}(u)\|^\alpha \int_0^{2\pi} |\cos(\kappa_u)\cos(s) + \sin(\kappa_u)\sin(s)|^\alpha \, ds \\
&= -c_0 \|g_{\theta,t}(u)\|^\alpha \int_0^{2\pi} |\cos(s - \kappa_u)|^\alpha \, ds = -c_0 \|g_{\theta,t}(u)\|^\alpha c_1,
\end{aligned}$$

where $c_1 = \int_0^{2\pi} |\cos(s)|^\alpha \, ds$. Inserting this into (2.19), we identify the characteristic function as

$$\exp\left(-c_0 c_1 \int_{\mathbb{R}} \|g_{\theta,t}(u)\|^\alpha \, ds\right)$$

which is the characteristic function of harmonizable fractional stable motion stated in Theorem 6.3.4 of [14] and on p. 359 of the same book when we insert $g_{\theta,t}$ (up to a scaling factor). Thus all that remains is the argument for dominated convergence theorem in equation (2.19). By assumption there exists a $C > 0$ such that $f(x) \le C\|x\|^{-2-\alpha}$ for all $x \in \mathbb{R}$. This implies that

$$f(\epsilon^{-1/\alpha} x)\epsilon^{-2/\alpha - 1} \le C\|\epsilon^{-1/\alpha} x\|^{-2-\alpha} \epsilon^{-2/\alpha - 1} = C\|x\|^{-2-\alpha}.$$

Thus a good candidate for a dominating (integrable) function would be

$$F(x, u) = \|(g_{\theta,t}(u), x)\| C \|x\|^{-2-\alpha}.$$

From classical theory of Lévy measures, we know that

$$|k(g_{\theta,t}(u), x)| \le 1 \wedge \Big[\|g_{\theta,t}(u)\|^2 \|x\|^2\Big],$$

which implies that

$$F(x, u) \le C\Big(\|x\|^{-2-\alpha} \wedge \Big[\|g_{\theta,t}(u)\|^2 \|x\|^{-\alpha}\Big]\Big).$$

By changing to polar coordinates we obtain that (the constant changes from line to line)

$$\begin{aligned}
&\int_{\mathbb{R}} \int_{\mathbb{R}^2} C\Big(\|x\|^{-2-\alpha} \wedge \Big[\|g_{\theta,H,t}(u)\|^2 \|x\|^{-\alpha}\Big]\Big) \, dx \, du \\
&= \int_{\mathbb{R}} \int_0^{2\pi} \int_0^\infty C\Big(r^{-2-\alpha} \wedge \Big[\|g_{\theta,H,t}(u)\|^2 r^{-\alpha}\Big]\Big) r \, dr \, d\psi \, du \\
&\le \sum_{j=1}^n C \int_{\mathbb{R}} \int_0^\infty \Big(r^{-1-\alpha} \wedge \Big[\|f_{t_j}(u)\|^2 r^{-\alpha+1}\Big]\Big) \, dr \, du,
\end{aligned}$$

where f_t denotes the integrand of the harmonizable fractional Lévy motion at time t. This is exactly the criterion for the existence of the stochastic integral $\int |f_t| d\tilde{L}_s$ wrt. an α-stable real-valued Lévy process $\tilde{L}$. By the choice of $(\alpha, H) \in (0,2)\times(0,1)$ such an integral exists by the existence of the harmonizable fractional stable motion for these parameters. This concludes the argument for dominated convergence and hence the proof. □

References

[1] David Applebaum. *Lévy processes and stochastic calculus*. Cambridge University Press, 2009.

[2] Jean-Marc Bardet and Donatas Surgailis. "Nonparametric estimation of the local Hurst function of multifractional Gaussian processes". *Stochastic Process. Appl.* 123.3 (2013), 1004–1045.

[3] Andreas Basse-O'Connor, Raphaël Lachièze-Rey and Mark Podolskij. "Power variation for a class of stationary increments Lévy driven moving averages". *Ann. Probab.* 45.6B (2017), 4477–4528.

[4] Albert Benassi, Serge Cohen and Jacques Istas. "Identification and properties of real harmonizable fractional Lévy motions". *Bernoulli* 8.1 (2002), 97–115.

[5] Stamatis Cambanis and Makoto Maejima. "Two classes of self-similar stable processes with stationary increments". *Stochastic Process. Appl.* 32.2 (1989), 305–329.

[6] Mark E. Crovella and Azer Bestavros. "Self-similarity in World Wide Web Traffic: Evidence and Possible Causes". *SIGMETRICS Perform. Eval. Rev.* 24.1 (May 1996), 160–169.

[7] Joseph L. Doob. *Stochastic Processes*. Vol. 7. Wiley New York, 1953.

[8] John Lamperti. "Semi-stable stochastic processes". *Transactions of the American mathematical Society* 104.1 (1962), 62–78.

[9] Tina Marquardt. "Fractional Lévy processes with an application to long memory moving average processes". *Bernoulli* 12.6 (2006), 1099–1126.

[10] Balram S. Rajput and Jan Rosiński. "Spectral representations of infinitely divisible processes". *Probab. Theory Related Fields* 82.3 (1989), 451–487.

[11] Jan Rosiński. "On the structure of stationary stable processes". *Ann. Probab.* 23.3 (1995), 1163–1187.

[12] Jan Rosiński and Gennady Samorodnitsky. "Classes of mixing stable processes". *Bernoulli* 2.4 (1996), 365–377.

[13] Gennady Samorodnitsky. *Stochastic Processes and Long Range Dependence*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2016.

[14] Gennady Samorodnitsky and Murad S. Taqqu. *Stable non-Gaussian random processes*. Stochastic Modeling. Stochastic models with infinite variance. Chapman & Hall, New York, 1994.

[15] Stilian Stoev, Murad S Taqqu, Cheolwoo Park, George Michailidis and JS Marron. "LASS: a tool for the local analysis of self-similarity". *Computational statistics & data analysis* 50.9 (2006), 2447–2471.

[16] Kazimierz Urbanik. "Random measures and harmonizable sequences". *Studia Mathematica* 31.1 (1968), 61–88.

[17] Walter Willinger, Murad S Taqqu and Ashok Erramilli. "A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks". *Stochastic networks: Theory and applications* (1996), 339–366.

[18] Akiva M Yaglom. *An Introduction to the Theory of Stationary Random Functions*. Courier Corporation, 2004.

Chapter 3

Limit theory for quadratic variation of harmonizable Lévy-driven processes

Andreas Basse-O'Connor, Thorbjørn Grønbæk and Mark Podolskij
Department of Mathematics, Aarhus University

Working paper

3.1 Background

The classical model of Brownian motion (Bm) was first mentioned by Robert Brown in the 1820's and mathematically defined at the beginning of the twentieth century. Ease of use and desirable properties (continuity, stationary independent increments) lead to widespread applications in many scientific fields. Fractional Brownian motion generalizes the self-similarity property of Brownian motion as mentioned in Chapter 1. There exists several stochastic integral representations of fractional Brownian motion, but we shall emphasize two of them, namely the *moving average representation*

$$B_t^H = C_1(H) \int_{-\infty}^t (t-s)_+^{H-1/2} - (-s)_+^{H-1/2} \, dB_s, \quad t \in \mathbb{R}, \tag{3.1}$$

and the *harmonizable representation*

$$B_t^H = C_2(H) \int_{\mathbb{R}} \frac{e^{its}-1}{is} |s|^{-H+1/2} \tilde{B}(dx), \quad t \in \mathbb{R}. \tag{3.2}$$

where C_1 and C_2 are constants depending only on H, $(B_s)_{s\in\mathbb{R}}$ is a Brownian motion and finally $\tilde{B}$ is an complex-valued Gaussian measure defined by $\tilde{B} = B^{(1)} + iB^{(2)}$, where $B^{(1)}$ and $B^{(2)}$ are independent real-valued Gaussian random measures, satisfying $B^{(1)}(A) = B^{(1)}(-A)$, $B^{(2)}(A) = -B^{(2)}(A)$ for any Borel set A of finite Lebesgue measure. Both representations yield a real-valued Gaussian H-selfsimilar process with stationary increments and thus the same process by [13].

Similarly, recall their analogs for α-stable Lévy processes. The *linear fractional stable motion* (lfsm) which is analog to (3.1) and defined as

$$L_{\alpha,H}(t) = \int_{\mathbb{R}} a\Big((t-s)_+^{H-1/\alpha} - (-s)_+^{H-1/\alpha}\Big) dM_s^\alpha + \int_{\mathbb{R}} b\Big((t-s)_-^{H-1/\alpha} - (-s)_-^{H-1/\alpha}\Big) dM_s^\alpha, \quad t \in \mathbb{R}, \tag{3.3}$$

where $(M_s^\alpha)_{s\in\mathbb{R}}$ is a (symmetric) α-stable Lévy process on $\mathbb{R}$, $a, b \in \mathbb{R}$ such that $|a| + |b| > 0$, $0 < \alpha < 2$, $0 < H < 1$, $H \neq 1/\alpha$. The complex-valued *harmonizable fractional stable motion* (hfsm) is similarly analog to (3.2) and defined as

$$C_{\alpha,H}(t) = \int_{\mathbb{R}} \frac{(e^{its}-1)}{is}\Big(a(s_+)^{-H+1-1/\alpha} + b(s_-)^{-H+1-1/\alpha}\Big) d\tilde{M}_s, \quad t \in \mathbb{R}, \tag{3.4}$$

where $\tilde{M}$ denotes a complex-valued isotropic α-stable Lévy process and the parameter space is $0 < \alpha < 2$, $0 < H < 1$, $a \geq 0$, $b \geq 0$, where $a + b > 0$. We refer to Chapter 7 of [19] for detailed definitions of both processes. As mentioned in the introduction chapter, they are *different* processes, contrary to the Gaussian case. This raised the question how the lfsm and (possibly real-valued) hfsm are related. Recall that the low-frequency power variation (or p-variation) of a process $(X_t)_{t\in\mathbb{R}}$ is defined as

$$V(p)_n := \sum_{t=1}^{n} |X_t|^p.$$

Similarly, the high-frequency p-variation is simply defined as

$$V_p(t,n) = \sum_{i=1}^{\lfloor tn \rfloor} |X_{i/n} - X_{(i-1)/n}|^p. \tag{3.5}$$

For self-similar processes (hfsm and lfsm), the low- and high-frequency settings are related through the self-similarity. The limit theory of the power variation would allow comparison to the moving averages (which includes lfsm) studied in [2] and yield a deeper understanding of the connection between kernel function structure and the corresponding limit theory. The special case $p = 2$ is called the quadratic variation. It is a measure of variability for a process $\mathbf{X}$ and is given by

$$\sum_{i=1}^{n} \Big(X_{i/n} - X_{(i-1)/n}\Big)^2$$

in the high-frequency setting and is used to estimate parameters in many stochastic processes and diffusions. In the case of Brownian motion $\mathbf{B}$, it is well-known that

$$\sum_{i=1}^{n} (B_{i/n} - B_{(i-1)/n})^2 \to \sigma^2 \text{ a.s.}, \quad \text{as } n \to \infty.$$

The above convergence extends to the larger class of semi-martingale. However for non-semimartingales we may need an additional normalization term. Indeed for the standard fractional Brownian motion $\mathbf{B}^H$, [5] show that

$$\frac{1}{n-1}\sum_{i=1}^{n} n^{2H}\left(B^H_{i/n} - B^H_{(i-1)/n}\right)^2 \to 1 \text{ a.s.}$$

Observe that this does not equal the empirical volatility unless $H = 1/2$. Fractional Brownian motion is not a semi-martingale as shown in [16] but it is *mixing* as mentioned in [4] and utilized for lfsm in [2]. However this does not mean that studying the quadratic variation is not useful. In fact, [5] show that a functional of the quadratic variation of fBm is the best estimator of H with the asymptotically smallest variance. As another example, a stock price process $\mathbf{X}$ may be described by a diffusion, e.g.

$$dX_t = \mu_t dt + \sigma_t \, dB_t,$$

where $\mathbf{B}$ is a Brownian motion and μ_t and σ_t are suitable processes such this an Itô diffusion. Itô diffusions are semi-martingales, a class which serve as the building block of price processes in mathematical finance. It is often central in mathematical finance to estimate the volatility σ and determine whether we take on a lot risk (or volatility) as compared to the expected payoff. This estimation is commonly done through the realized volatility. Power variation, and in particular quadratic variation, of general semimartingales is studied in [1], where they obtain that

$$V_p(t,n) \xrightarrow[n\to\infty]{\mathcal{L}-s} \int_0^t \sigma_s \, ds,$$

where $\xrightarrow{\mathcal{L}\text{-}s}$ denotes convergence in stable law, see [1]. This theory can be utilized to perform statistical inference by estimating parameters of the underlying semi-martingale model.

In this paper, we study the quadratic variation of certain harmonizable processes and show first-order convergence towards a stochastic variable. As a corollary, we study how this affects the semi-martingale property of harmonizable fractional stable motion.

3.2 Results

We apply the limit theory for the power variation of harmonizable process. Recall that by a *harmonizable process*, we understand a process of the form

$$Z_t = \int_{\mathbb{R}} e^{its} M(ds),$$

where M is a complex-valued random measure. Similarly, a process $\mathbf{Z}$ has harmonizable increments if

$$Z_{t+h} - Z_h = \int_R [e^{i(t+h)s} - e^{ihs}] M(ds)$$

where M is a complex-valued random measure. For definition and existence of certain subclasses of harmonizable processes we refer to Section 2.3. Recall that a Lévy-driven harmonizable process is of the form

$$X_t = \int_{\mathbb{R}} e^{its} g(s)\, dL_s, \tag{3.6}$$

where L is a complex-valued isotropic Lévy process and $g : \mathbb{R} \to \mathbb{C}$ is a deterministic function. We present our main result.

Theorem 3.2.1. *Let $\mathbf{X} = (X_t)_{t\in\mathbb{R}}$ denote a Lévy-driven harmonizable process generated by a complex-valued isotropic Lévy process $L = L^1 + iL^2$. Then it holds that*

$$\frac{1}{n}\sum_{t=1}^{n} |X_t|^2 \xrightarrow{\mathbb{P}} U_0, \tag{3.7}$$

where U_0 is an infinitely divisible variable of the form

$$U_0 = \int_{\mathbb{R}} |g(s)|^2\, d\Big([L^1] + [L^2]\Big)_s,$$

where $[L^1]$, and $[L^2]$, denotes the quadratic variation of the Lévy process L^1, respectively L^2.

The proof is postponed to later in this chapter. It relies on the chosen power $p = 2$ and multiple integration theory developed in [9, 10, 20]. We expect the result holds for general harmonizable processes $(Z_t)_{t\in\mathbb{R}}$ as well, but additional theory is needed to generalize it.

Note that the increments of hfsm is a harmonizable process. The kth order increments of a stochastic process is defined as

$$\Delta_{t,k} X := \sum_{j=0}^{k} (-1)^j \binom{k}{j} X_{t-j}, \quad t \geq k,$$

and observe that $\Delta_{t,1} = X_t - X_{t-1}$ and $\Delta_{t,2} = X_t - 2X_{t-1} + X_{t-2}$. Elaborating on the above, we may observe that the kth order increments of a harmonizable process

$$\Delta_{t,k} X = \int_{\mathbb{R}} e^{its} \underbrace{(1 - e^{-is})^k g(s)}_{\tilde{g}(s)}\, dL_s, \quad t \geq k.$$

is once again a harmonizable process and thus the result in Theorem 3.2.1 will apply to these processes as well.

Hrmonizable fractional stable motion is *not* a semi-martingale (to the best of our knowledge) nor is it mixing. In fact, Theorem 1.3.3 yields the following corollary for hfsm on whether it is a semi-martingale. We believe this to be well-known but we have been unable to find a reference for it.

Corollary 3.2.2. *Harmonizable fractional stable motion $\mathbf{X}$ is not a semimartingale for $H < 1/2$.*

To prove this, observe that

$$\sum_{t=1}^{n}|X_{t/n}-X_{(t-1)/n}|^2 \overset{d}{=} n^{-2H}\sum_{t=1}^{n}|X_t-X_{t-1}|^2 \xrightarrow[n\to\infty]{\text{Theorem 1.3.3}} \begin{cases} U_0, & \text{if } H=1/2.\\ 0, & \text{if } H>1/2.\\ \infty, & \text{if } H<1/2. \end{cases}$$

This follows along the known results for fractional Brownian motion in [16]. The cases $H = 1/2$ and $H > 1/2$ appear to be open.

The previous section gave the background and context for the main result in the current section. Section 3.3 contains further results and the proof of the main result. We include Section 3.4 as supplementary material on multiple stochastic integrals.

3.3 Proofs and further results

In this section we aim to pave the way for the proof of Theorem 3.2.1. To do this, we prove some results of separate interest and multiple integration theory. Finally at the end of this section, we assemble our observations to prove Theorem 3.2.1 in short order. The proof relies heavily on the following two results.

Theorem 3.3.1. *Let $(X_t)_{t\in\mathbb{R}}$ denote a Lévy-driven harmonizable process as in* (3.6)*. Then it holds that*

$$|X_t|^2 = U_0 + V_t,$$

where U_0 is defined in Theorem 3.2.1 and V_t is a multiple stochastic integral of the form

$$V_t = 2\mathfrak{Re}\Big(\int_{\mathbb{R}}\int_{-\infty}^{s_-} e^{its}g(s)e^{-itu}g(u)\,d\overline{L}_u\,dL_s\Big),$$

where $\mathfrak{Re}$ denotes the real part of a complex number.

In this Theorem, it is implied that both U_0 and V_t exist under no further assumptions than the existence of the process $\mathbf{X} = (X_t)_{t\in\mathbb{R}}$. A *symmetric* $\mathbb{R}^2$-valued Lévy process $\mathbf{L} = (L_t^1, L_t^2)_{t\in\mathbb{R}}$ satisfies that $-L \overset{d}{=} L$.

The next step is to show a connection between multiple (stochastic) integrals and stochastic integrals, understood as in the semimartingale theory of [8].

Proposition 3.3.2. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a measurable function vanishing on the diagonal with $f(s,t) = 0$ for all $s \geq t$. Let $L = (L^1, L^2)$ be a symmetric two-dimensional Lévy process and suppose the multiple integral of f wrt. L exists. Then the integrals below exists simultaneously and are equal:*

$$\int_{\mathbb{R}}\int_{\mathbb{R}} f(s,u)\mathbb{1}_{\{s>u\}}\,dL_u^1\,dL_s^2 = \int_{\mathbb{R}}\int_{-\infty}^{s_-} f(s,u)\,dL_u^1\,dL_s^2, \tag{3.8}$$

where the left-hand side denotes a multiple integral and the right-hand side denotes stochastic integrals (with the existence of these also being implied).

Proposition 3.3.2 is used to prove Theorem 3.3.1, presented at the start of this section.

Proof of Theorem 3.3.1. We study a Lévy-driven harmonizable process $(X_t)_{t\in\mathbb{R}}$, e.g.

$$X_t := \int_{\mathbb{R}} e^{its} g(s)\, dL_s.$$

Let $f_t = e^{its}g(s)$ denote the kernel function of $(X_t)_{t\in\mathbb{R}}$. We may identify the interval $[-\infty,\infty]$ with $[0,1]$ via a suitable mapping which allows us to interpret the integral as a semimartingale. Applying integration by parts to the complex-valued semimartingales

$$\begin{aligned}
&|X_t|^2 = X_t\overline{X}_t \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{s_-} e^{itu} g(u)\, dL_u\right) d\overline{X}_s + \int_{-\infty}^{\infty}\left(\int_{-\infty}^{s_-} e^{-itu}\overline{g(u)}\, d\overline{L}_u\right) dX_s + [X_t,\overline{X}_t]_\infty \\
&= \int_{\mathbb{R}}\left(\int_{-\infty}^{s_-} e^{itu} g(u)\, dL_u\right) d\overline{f_t(s)\Delta L_s} + \int_{\mathbb{R}}\left(\int_{-\infty}^{s_-} e^{-itu}\overline{g(u)}\, d\overline{L}_u\right) d(f_t(s)\Delta L_s) \\
&\quad + [X_t,\overline{X}_t]_\infty,
\end{aligned}$$

where the multiple integrals are defined through Definition 3.4.10. Due to the semimartingale property, it follows that

$$\begin{aligned}
[X_t,\overline{X}_t]_\infty &= \int_{\mathbb{R}} e^{its} g(s) e^{-its}\overline{g(s)}\, d[L,\overline{L}]_s = \int_{\mathbb{R}} |g(s)|^2\, d[L,\overline{L}]_s \\
&= \int_{\mathbb{R}} |g(s)|^2\, d([L^1]+[L^2])_s
\end{aligned}$$

Inserting the definition of X_t and applying Proposition 3.3.2 (this gives the existence of the terms), we may recognize the two terms

$$\int_{\mathbb{R}}\left(\int_{-\infty}^{s_-} e^{itu} g(u)\, dL_u\right) d\overline{f_t(s)\Delta L_s} + \int_{\mathbb{R}}\left(\int_{-\infty}^{s_-} e^{-itu}\overline{g(u)}\, d\overline{L}_u\right) d(f_t(s)\Delta L_s)$$

as multiple stochastic integrals in the sense of Definition 3.4.10. Existence of all these integrals follows by applying linearity, then Proposition 3.3.2 and finally Lemma 3.4.4. Thus we have identified that

$$|X_t|^2 = U_0 + V_t,$$

where U_0 is an infinitely divisible random variable and V_t is a multiple stochastic integral of the form

$$V_t = 2\mathfrak{Re}\Big(\int_{\mathbb{R}}\int_{-\infty}^{s_-} e^{its} g(s) e^{-itu}\overline{g(u)}\, d\overline{L}_u\, dL_s\Big)$$

□

We finally have all the pieces needed to complete the proof of Theorem 3.2.1.

Proof of Theorem 3.2.1. Collecting the previous observations in Theorem 3.3.1 we see that

$$\frac{1}{n}\sum_{t=1}^{n}|X_t|^2 = U_0 + \frac{1}{n}\sum_{t=1}^{n} V_t.$$

Thus all that remains is to show that

$$\frac{1}{n}\sum_{t=1}^{n} V_t \xrightarrow{\mathbb{P}} 0.$$

By linearity of the multiple integral we have

$$\frac{1}{n}\sum_{t=1}^{n} V_t = 2\Re\mathfrak{e}\Big(\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{1}{n}\sum_{t=1}^{n}\tilde{f}_t(s,u)\,d\overline{L}_u\,dL_s\Big),$$

where

$$f_t(s,u) := e^{its}g(s)e^{-itu}\overline{g(u)}\mathbb{1}_{\{s>u\}}.$$

Lemma 3.3.3 implies that

$$\frac{1}{n}\sum_{t=1}^{n}\tilde{f}_t(s,u) \to 0, \quad \text{and} \quad \Big|\frac{1}{n}\sum_{t=1}^{n}\tilde{f}_t(s,u)\Big| \le |g(s)g(u)|\mathbb{1}_{\{s>u\}}.$$

Due to continuity of multiple integrals, it suffices to show that the multiple integral of $|g(s)g(u)\mathbb{1}_{\{s>u\}}|$ exists. However this follows from Lemma 3.4.4 and concludes the proof of Theorem 3.2.1. □

The following lemma provides the remaining necessary observations for the proof of Theorem 3.2.1.

Lemma 3.3.3 (Properties of kernel function)**.** *Our kernel function in the multiple integral is given as*

$$\frac{1}{n}\sum_{t=1}^{n}\tilde{f}_t(s,u) = \mathbb{1}_{\{s>u\}}g(s)\overline{g(u)}\frac{1}{n}\sum_{t=1}^{n}e^{its}e^{-itu}.$$

We observe the following properties

(i)

$$\left|\frac{1}{n}\sum_{t=1}^{n}e^{it(u-s)}\right| \le 2,$$

for all s, u, n. This implies the following inequality

$$\left|\frac{1}{n}\sum_{t=1}^{n}\tilde{f}_t(s,u)\right| \le 2\big|g(s)\overline{g(u)}\big| = 2|g(u)||g(s)|.$$

where the right-hand side does not depend on n or t.

(ii) We have the following pointwise convergence

$$\frac{1}{n}\sum_{t=1}^{n}\tilde{f}_t(s,u) \xrightarrow[n\to\infty]{} 0,$$

except on a Lebesgue null set.

Proof of Lemma 3.3.3. The first item (i) in the lemma follows immediately from the triangular inequality. The last statement follows the observation, that

$$\sum_{t=0}^{n}(e^{ix})^t = \frac{1-e^{ix(n+1)}}{1-e^{ix}}, \tag{3.9}$$

for $x \neq 0$ modulus 2π. We prove this by induction in n. Assume $n = 0$. Then clearly the equation holds because

$$e^{i0x} = 1 = \frac{1-e^{ix}}{1-e^{ix}}.$$

Assume equation (3.9) holds for n. Then for $n+1$ we have that

$$\begin{aligned}\sum_{t=0}^{n+1}(e^{ix})^t =& e^{ix(n+1)} + \frac{1-e^{ix(n+1)}}{1-e^{ix}} = \frac{e^{ix(n+1)}(1-e^{ix})+1-e^{ix(n+1)}}{1-e^{ix}} \\ =& \frac{1-e^{ix(n+2)}}{1-e^{ix}}\end{aligned}$$

and the proof of equation (3.9) is complete. This observation implies that

$$\left|\frac{1}{n}\sum_{t=1}^{n}e^{it(s-u)}\right| \leq \frac{1}{n}\left(\left|\frac{1-e^{i(n+1)(s-u)}}{1-e^{i(s-u)}}\right|+1\right) \leq \frac{1}{n}\left(\frac{2}{1-e^{i(s-u)}}+1\right).$$

Letting n tend to infinity, we see that this converges to 0, under the condition that $(s-u) \neq 0$ modulus 2π. The excluded set of points is given as

$$\{(s,u)\in\mathbb{R}^2 \,|\, (s-u) = 0 \text{ modulus } 2\pi\} = \bigcup_{k\in\mathbb{Z}}(D+2k\pi),$$

which is a Lebesgue null-set (D denotes the diagonal in $\mathbb{R}^2$). This implies that

$$\left|\frac{1}{n}\sum_{t=1}^{n}\tilde{f}_t(s,u)\right| = |g(s)\overline{g}(u)|\left|\frac{1}{n}\sum_{t=1}^{n}e^{it(s-u)}\right| \to 0 \quad \text{Leb } a.s.$$

□

3.4 Multiple integration theory

In this section, we aim to define and discuss the theory of multiple stochastic integrals. It is meant as a supplementary reading and the only original results

are a few straightforward connections to the stochastic integration theory for semimartingales.

Multiple integrals wrt. α-stable and symmetric Lévy processes has been studied extensively by many authors in the 1980's and 1990's. An incomplete list includes [3, 11, 14, 17, 18] and references therein. The culmination of this research appears to the book [12] for predictable integrands and slightly prior to this, the papers [9, 20]. Recently, the book [10] based on [9, 20] has been published and this will serve as the main reference in the present exposition. Today, this theory is still relevant and exemplified in the topic Malliavian Calculus which utilizes multiple integration, albeit with a modified approach. Our main application of this theory will be for double stochastic integrals. The chief goal is to properly define

$$\int_{\mathbb{R}} \int_{\mathbb{R}} f \, dX_1 \, dX_2, \tag{3.10}$$

where $\mathbf{X} = (X_t)_{t\in\mathbb{R}}$ is an $\mathbb{R}^d$-valued symmetric Lévy process, and a suitable measurable function $f : \mathbb{R}^2 \to \mathbb{R}$. However, as the theory is essentially the same, we shall instead introduce multiple stochastic integration theory of arbitrary order d, that is

$$\underbrace{\int_{\mathbb{R}} \cdots \int_{\mathbb{R}}}_{d \text{ times}} f \, \mathrm{d}\xi_1 \cdots d\xi_d, \tag{3.11}$$

where ξ_j are point processes. We will highlight simplifications and intuition whenever the order $d = 2$ allows it, as seen in Remark 3.4.3 and Lemma 3.4.4.

In subsection 3.4.1, we define multiple stochastic integrals for classical Poisson random measures as it is the essential building block for multiple integration theory for Lévy processes. In the next Section 3.4.2 we define multiple integrals for symmetric Poisson random measures and Lévy processes as a result of this.

3.4.1 Multiple integration for Poisson random measures

We may write this point process using its *atomic* representation (see Chapter 2 in [10]), given by

$$\xi = \sum_n \delta_{\tau_n}, \tag{3.12}$$

where (τ_n) is a sequence of distinct S-valued random variables. By a M-marked point process ξ on S, we shall mean a simple point process on the product space $S \times M$, such that $\xi(\{s\} \times M) \leq 1$ for all $s \in S$. In other words, the process must only jump once per time point. In the setting of a $\mathbb{R}^d$-valued Lévy process $\mathbf{L} = (L_t)_{t\in\mathbb{R}}$, $S = \mathbb{R}$ is the time and $M = \mathbb{R}^d \setminus \{0\}$ is the set of possible jump values. Furthermore, such a process is Poisson if and only if the intensity measure of ξ, denoted ν, satisfies that $\nu(\{s\} \times M) = 0$ for all $s \in S$. Or in other words, the probability that the process jumps at a given time-point s is zero for all s.

We consider multiple integrals of the form

$$\xi_1 \cdots \xi_d f = \int_{S\times M} \cdots \int_{S\times M} f(t_1,\ldots,t_d)\xi_1(t_1)\cdots\xi_d(t_d)$$

where $\xi_1,\ldots,\xi_d$ are independent copies of M-marked point process ξ on S with independent increments and intensity $\mathbb{E}[\xi] = \nu$, and $f \geq 0$ is a measurable function on $\bar{S}^d = (S\times M)^d$ which vanishes on the diagonal (henceforth referred to as *non-diagonal*). The case $\xi_k = \xi$ for all k is also treated and is denoted $\xi^d f$ instead of $\xi_1 \cdots \xi_d f$. The integral should be understood as the multiple sums (possibly infinite) wrt. ξ

$$\sum_{j_1} \cdots \sum_{j_d} f(\tau_{j_1},\ldots,\tau_{j_d}),$$

where $\tau_{j_i} \in (S\times M)$ are the atoms of the measure ξ_j from equation (3.12) for all i. Heuristically, the reader should understand these two cases as the minimal and maximal dependency scenarios. Furthermore, as described in the introduction of [9], most results can be extended by simple projection to general Poisson point process (with dependency). Interestingly, it turns out that existence of the integral in (3.24) for the two cases is simultaneous due to the following theorem.

Theorem 3.4.1 (Kallenberg, [10])**.** *For any non-diagonal, measurable function $f \geq 0$ on $\bar{S}^d$, we have*

(i) $\xi_1 \cdots \xi_d f < \infty$ *a.s.* $\iff \xi^d f < \infty$ *a.s.*

(ii) $\xi_1 \cdots \xi_d f_n \xrightarrow{P} 0 \iff \xi^d f_n \xrightarrow{P} 0.$

(iii) $\xi_1 \cdots \xi_d f_n \xrightarrow{P} \xi_1 \cdots \xi_d f < \infty \implies \xi^d f_n \xrightarrow{P} \xi^d f.$

This does not give explicit criteria for the existence. To formulate explicit criteria for existence we will define the following notation. These technicalities may be simplified way for the double stochastic integral, $d = 2$, in Remark 3.4.3.

Given a measurable function $f \geq 0$ on $\bar{S}^J$, where $m = |J| < \infty$, we recursively define functions $f_1,\ldots f_m \geq 0$ on $\bar{S}^J$ by

$$f_1 = f \wedge 1, \quad f_{k+1} := f_k \prod_{|I|=k} \mathbb{1}_{\{\nu^I f_k \leq 1\}}$$

where the product extends over all sets $I \subset J$ with $|I| = k$, and ν^I denotes integration in the arguments indexed by I, so that $\nu^I f_k$ becomes a measurable function of the remaining arguments indexed by $J \setminus I$. The notation $\nu^I f$ denotes the integral of j in the coordinates corresponding to I. The next step is to recursively define classes $\mathcal{C}_d$ of measurable functions $f \geq 0$ on $\bar{S}^d$. First set $\mathcal{C}_0 = \{0,1\}$ (the constant functions), and assume $\mathcal{C}_k$ to be known for all $k < d$. We define the classes $\mathcal{C}_d$ recursively by the functions satisfying that

1. $\nu^{d-k}\mathbb{1}_{\{\nu^J f_k = \infty\}} = 0$

2. $\mathbb{1}_{\{\nu^J f_k > 1\}} \in \mathcal{C}_{d-k}$

for all index subsets J of $\{0,1\}^d$ with $k = |J| > 0$.

Theorem 3.4.2 (Kallenberg, [10]). *For any measurable function $f \geq 0$ on $\bar{S}^d$, we have*

$$P(\xi_1 \cdots \xi_d f < \infty) = \mathbb{1}_{\{f \in \mathcal{C}_d\}}$$

and similarly for $\xi^d f$, when f is non-diagonal.

Remark 3.4.3 (Double stochastic integral). Through this example let ξ denote a simple point process on $\mathbb{R}^2$ with intensity λ (not necessarily the Lebesgue measure in this example) and let $f \geq 0$ denote a measurable function on $\mathbb{R}^2$. Define the functions

$$f'(x) := \lambda(f(x,\cdot) \wedge 1), \quad f''(y) := \lambda(f(\cdot,y) \wedge 1),$$

e.g. $\nu^J f$ using the above notation. Corollary 6.5 in [9] shows that the double stochastic integral of a deterministic function f exists if and only if the following criterions hold true

(i) $\lambda(\{f' \vee f'' = \infty\}) = 0$,

(ii) $\lambda(\{f' \vee f'' > 1\}) < \infty$,

(iii)

$$\int_{\mathbb{R}} \int_{\mathbb{R}} [f(x,y) \wedge 1] \mathbb{1}_{\{f'(x) \leq 1\}} \mathbb{1}_{\{f''(y) \leq 1\}} \lambda(dx)\lambda(dy) < \infty.$$

The double integral is simply defined as a double (stochastic) sum, i.e.

$$\sum_n \sum_m f(T_n^1, T_n^2).$$

Item (i) is easily interpreted as the inner sum being finite (regardless of the order of integration). Next, consider the integrals

$$\int f' \, d\xi \quad \text{and} \quad \int f'' \, d\xi.$$

These integrals are finite if and only if $\lambda(f' \wedge 1) < \infty$ and $\lambda(f'' \wedge 1) < \infty$.

This allows us to split this criterion into two sets for both f' and f'', i.e.

$$\int f' \wedge 1 \, d\lambda = \int f' \mathbb{1}_{\{f' \leq 1\}} \, d\lambda + \int \mathbb{1}_{\{f' > 1\}} \, d\lambda.$$

The existence of the second integral for f' and f'' follows by item (ii). The final criterion deals appears odd, but actually treats the existence of $\int f' \mathbb{1}_{\{f' \leq 1\}} \, d\xi$. Indeed, observe that

$$\begin{aligned} &\int_{\mathbb{R}} \int_{\mathbb{R}} [f(x,y) \wedge 1] \mathbb{1}_{\{f'(x) \leq 1\}} \mathbb{1}_{\{f''(y) \leq 1\}} \, \lambda(dx)\lambda(dy) \\ &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} [f(x,y) \wedge 1] \mathbb{1}_{\{f'(x) \leq 1\}} \lambda(dx) \right] \mathbb{1}_{\{f''(y) \leq 1\}} \lambda(dy) \end{aligned}$$

If the inner bracket did not contain the restriction to $\{f' \leq 1\}$, it would be exactly the integrability criterion for $\int f'' \mathbb{1}_{\{f'' \leq 1\}} \, d\lambda$. Thus we observe that the following implication holds true

$$\int f' \, d\xi < \infty \quad a.s. \implies \int \int f \, d\xi \, d\xi < \infty \quad a.s.$$

Although the converse statement would be a beautiful result, namely that the double integral exists if and only $\int f' \wedge 1 \lambda$ is finite, we have been unable to prove it so far. The only observation needed for equivalence is whether

$$\int_{\mathbb{R}} \left[\int_{\mathbb{R}} [f(x,y) \wedge 1] \mathbb{1}_{\{f'(x) > 1\}} \lambda(dx) \right] \mathbb{1}_{\{f''(y) \leq 1\}} \lambda(dy) < \infty,$$

given items (i)-(iii). If it holds, the double stochastic integrals exists if and only if "double" integrating deterministically is finite – a very beautiful criterion.

A natural attribute of a multiple integral is that the existence of the single stochastic integral $\int g(s) \, dL_s$ implies the existence of $\int \int \tilde{g}(s,u) \, dL_u \, dL_s$ in the special case $\tilde{g}(s,u) = g(s)g(u)\mathbb{1}_{\{s \neq u\}}$.

Lemma 3.4.4. *Let $f : \mathbb{R} \to \mathbb{R}$ be a measurable function such that the (single) stochastic integral $\int f \, dL_s$ exists. Set $\tilde{f}(s,u) = f(s)f(u)\mathbb{1}_{\{s \neq u\}}$. Then it holds that*

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \tilde{f}(s,u) \, dL_u \, dL_s$$

exists.

Proof. Existence criteria for the double stochastic integral of a function $f : \mathbb{R}^2 \to \mathbb{R}$ are given in Remark 3.4.3. The criteria depend on the function obtained by integrating out one of the pairs (s,x) or (u,y) while the other pair remains fixed, i.e.

$$f'(s,x) := \int_{\mathbb{R}} \int_{\mathbb{R}} (\tilde{f}(s,u)xy)^2 \wedge 1 \, \nu(dy) \, \lambda(du),$$
$$f''(u,y) := \int_{\mathbb{R}} \int_{\mathbb{R}} (\tilde{f}(s,u)xy)^2 \wedge 1 \, \nu(dx) \, \lambda(ds).$$

The definition of $\tilde{f}$ implies $f' = f''$. Writing f' with the notation of $z = f(s)x$, we see that

$$g(z) := \int_{\mathbb{R}} \int_{\mathbb{R}} (zf(u)y)^2 \wedge 1 \, \nu(dy) \, \lambda(du). \tag{3.13}$$

Observe that $g(z) = f'(s,x)$, and is only a function of z through $|z| = |f(s)x|$. Monotone Convergence implies

$$g(z) \xrightarrow[z \to \infty]{} \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{1}_{\{f \neq 0\}}(u) \, \nu(dy) \lambda(du) = \nu(\mathbb{R}) \lambda(\{f \neq 0\}). \tag{3.14}$$

The inequality $|a \wedge 1 - b \wedge 1| \leq |(a-b) \wedge 1|$ for $a, b \geq 0$ implies that $g(z)$ is a continuous function of z, since

$$|g(z_1) - g(z_2)| \leq \int \int |(z_1 - z_2)f(u)y|^2 \wedge 1 \, \nu(dy) \, \lambda(du) \xrightarrow[|z_1 - z_2| \to 0]{} 0,$$

where the right-hand-side is finite due to the existence of $\int f\,dL$. It is useful to divide into two cases now, namely

(I) $\nu(\mathbb{R}) = \infty$ or $\lambda(\{f \neq 0\}) = \infty$.

(II) $\nu(\mathbb{R}) < \infty$ and $\lambda(\{f \neq 0\}) < \infty$.

We assume $\nu(\mathbb{R}) > 0$ and $\lambda(\{f \neq 0\}) > 0$, as the contrary cases are trivial. First we consider case (I). In this case $g(z)$ will converge to infinity. Due to this and continuity of g, there exists $K > 0$, such that

$$|f(s)x| = |z| \in [0, K] \quad \Longleftrightarrow \quad f'(s,x) = g(z) \leq 1. \tag{3.15}$$

Thus, we see that

$$\Big\{(s,x) \in \mathbb{R}^2 \,\Big|\, f'(s,x) > 1\Big\} = \Big\{(s,x) \in \mathbb{R}^2 \,\Big|\, |f(s)x| > K\Big\}. \tag{3.16}$$

From the existence of the stochastic integral $\int f\,dL_s$, we know that

$$[\lambda \otimes \nu]\Big(\{(s,x) \in \mathbb{R}^2 \,|\, |f(s)x| > 1\}\Big) < \infty,$$

where $\lambda \otimes \nu$ denotes the product measure between λ and ν. Observe that if $K \geq 1$, then

$$[\lambda \otimes \nu]\Big(\{(s,x) \in \mathbb{R}^2 \,|\, |f(s)x| > K\}\Big) \leq [\lambda \otimes \nu]\Big(\{(s,x) \in \mathbb{R}^2 \,|\, |f(s)x| > 1\}\Big) < \infty.$$

If contrarily $K < 1$, it suffices to show

$$[\lambda \otimes \nu](\{(s,x) \in \mathbb{R}^2 \,|\, K < |f(s)x| < 1\}) < \infty.$$

Once again due to the existence of $\int f\,dL$, we know that

$$\begin{aligned}
\infty > &\int_{\mathbb{R}} \int_{\mathbb{R}} (f(s)x)^2 \mathbb{1}_{\{|xf(s)|<1\}}\, \nu(dx)\, \lambda(ds) \\
&\geq \int_{\mathbb{R}} \int_{\mathbb{R}} K^2 \mathbb{1}_{\{K<|xf(s)|<1\}}\, \nu(dx)\, \lambda(ds) \\
&= K^2 [\lambda \otimes \nu](\{(s,x) \in \mathbb{R}^2 \,|\, K < |f(s)x| < 1\}).
\end{aligned}$$

In both cases, we have verified that the set in equation (3.16) is of finite $\lambda \otimes \nu$-measure, which show that item (i) from Remark 3.4.3 holds. The second existence criterion is item (iii) from Remark 3.4.3, i.e.

$$\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \Big[[f(s)f(u)xy]^2 \wedge 1\Big] \mathbb{1}_{\{f'(s,x)\leq 1\}} \mathbb{1}_{\{f'(u,y)\leq 1\}}\, \nu(dx)\, \lambda(ds)\, \nu(dy)\, \lambda(ds) < \infty. \tag{3.17}$$

By (3.15) we may rewrite this as

$$\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \Big[[f(s)f(u)xy]^2 \wedge 1\Big] \mathbb{1}_{\{|f(s)x|\leq K\}} \mathbb{1}_{\{|f(u)y|\leq K\}}\, \nu(dx)\, \lambda(ds)\, \nu(dy)\, \lambda(ds),$$

for some $K > 0$. Divide this integral into the following two subsets A_1 and A_2 of $\mathbb{R}^4$,

$$A_1 := \{|f(s)f(u)xy| \le 1\} \cap \Big(\{|f(s)x| \le K\} \times \{|f(u)y| \le K\}\Big),$$
$$A_2 := \{|f(s)f(u)xy| > 1\} \cap \Big(\{|f(s)x| \le K\} \times \{|f(u)y| \le K\}\Big),$$

where $\times$ denotes the Cartesian product of the sets. On the set A_2, the integrand $[f(s)f(u)xy]^2 \wedge 1 \equiv 1$. Moreover, A_2 is empty for $K < 1$. Observe that

$$\begin{aligned} A_2 &\subseteq \{K^{-1} \le |f(s)x| \le K\} \times \{K^{-1} \le |f(u)y| \le K\} \\ &\subseteq \{K^{-1} \le |f(s)x|\} \times \{K^{-1} \le |f(u)y|\}, \end{aligned}$$

which we saw was a set of finite $\lambda \otimes \nu$-measure in the proof for the first criterion. This completes the finiteness of the integral on the set A_2. For the set $A_1 \subseteq \mathbb{R}^4$, we observe

$$A_1 \subseteq \{|f(s)x| \le K\} \times \{|f(u)y| \le K\}$$

Hence we obtain

$$\begin{aligned} &\int_{\mathbb{R}^2} \int_{\mathbb{R}^2} [f(s)f(u)xy]^2 \mathbb{1}_{A_1} \, \nu(dx)\, \lambda(ds)\, \nu(dy)\, \lambda(ds) \\ &\le \int_{\mathbb{R}^2} (f(s)x)^2 \mathbb{1}_{\{|f(s)x| \le K\}} \, \nu(dx)\, \lambda(ds) \cdot \int_{\mathbb{R}^2} (f(u)y)^2 \mathbb{1}_{\{|f(u)y| \le K\}} \, \nu(dy)\, \lambda(du). \end{aligned}$$

Now assume that $K \le 1$. Then the finiteness of the above follows immediately from the existence of the single stochastic integral. If instead $K \ge 1$ we only need to study the part

$$\int_{\mathbb{R}^2} (f(s)x)^2 \mathbb{1}_{\{1 \le |f(s)x| \le K\}} \, \nu(dx)\, \lambda(ds) \le K^2 \int_{\mathbb{R}^2} \mathbb{1}_{\{|f(s)x| \ge 1\}} \, \nu(dx)\, \lambda(ds),$$

and finiteness once again immediately follows from the existence of the single stochastic integral. This complete the proof in the case (I). The proof in the case (II) is much simpler and therefore omitted. □

Proof of Proposition 3.3.2. We assume the multiple integral of f wrt. L exists. By definition of the domain, this implies

$$\sum_{t \in \mathbb{R}} \sum_{s \le t} f^2(s,t) (\Delta L_s^1)^2 (\Delta L_t^2)^2 < \infty, \quad a.s. \tag{3.18}$$

The multiple integral of f is understood as the multiple integral of the function

$$(s,t,x,y) \mapsto f(s,t)xy \quad wrt. \quad \mu_L,$$

where μ_L denotes the (symmetric) *PRM* on $\mathbb{R} \times \mathbb{R}^2$ induced by L with known intensity measure $\lambda \otimes \nu$. Let μ_1, respectively μ_2, denote the jump measure of L_1, respectively L_2. Write $f(s,t)xy = g(s,t,x,y) + h(s,t,x,y)$ where

$$g(s,t,x,y) := (f(s,t)xy)\mathbb{1}_{[-1,1]}(f(s,t)xy), \qquad (L^2\text{-case})$$
$$h(s,t,x,y) := (f(s,t)xy)\mathbb{1}_{[-1,1]^c}(f(s,t)xy), \qquad (\text{Finite-variation case})$$

In general, we could choose any cut-off ϵ for the jumps and show that it is a finite variation (i.e. stochastic) integral. Equation (3.18) entails that

$$\infty > (\mu_1\mu_2)(h^2) = \sum_{t\in\mathbb{R}}\sum_{s\leq t} f^2(s,t)(\Delta L_s^1)^2(\Delta L_s^2)^2 \mathbb{1}_{[-1,1]^c}(f(s,t)(\Delta L_s^1)(\Delta L_t^2))$$

Being a countable sum (due to càdlàg sample paths) of numbers greater than 1 or 0's, the sum must have finitely terms. We shall call this the finite variation case. This implies that both sums are finite almost surely. Thus the integral equals

$$\begin{aligned}&\sum_{t\in\mathbb{R}}\sum_{s\leq t} f(s,t)(\Delta L_s^1)(\Delta L_t^2)\mathbb{1}_{[-1,1]^c}(f(s,t)(\Delta L_s^1)(\Delta L_t^2))\\ =&\sum_{t\in\mathbb{R}}\Big(\sum_{s\leq t} f(s,t)\Delta L_s^1 \mathbb{1}_{[-1,1]^c}(f(s,t)(\Delta L_s^1)(\Delta L_t^2))\Big)\Delta L_t^2.\end{aligned}$$

which is a finite variation integral, i.e. stochastic integral. To complete the proof in the finite variation case, we need to show the process inside the parenthesis is predictable. First of all, we may recognize it as a process of the form

$$H_t = \int_{\mathbb{R}} f(s,t)\,dL_s = \int_{-\infty}^{t} f(s,t)\,dL_s,$$

where $f(s,t) = 0$ for $s \geq t$. Such a process is clearly adapted. We wish to show that this process is predictable. In Theorem 3 $(iii)-(iv)$ of [6] we may replace adapted and progressively measurable with adapted and predictable, since the approximating sequence in the paper is in fact predictable. Due to this, it suffices to show that

$$\mathbb{R} \ni t \mapsto f(\cdot,t) \in M,$$

is a measurable mapping, where M denotes the Musielak-Orlicz space defined in [15]. The proof of this is technical and we omit it. Next, we turn our attention to the function g defined by

$$g(s,t,x,y) \coloneqq (f(s,t)xy)\mathbb{1}_{[-1,1]}(f(s,t)xy).$$

We will refer to this as the L^2-case. By assumption, we know that

$$\infty > \sum_{t\in\mathbb{R}}\sum_{s\leq t} f(s,t)^2(\Delta L_s^1)^2(\Delta L_t^2)^2\mathbb{1}_{[-1,1]}(f(s,t)(\Delta L_s^1)(\Delta L_t^2)).$$

Define the localization $(T_n)_{n\geq 1}$ by

$$T_n = \inf\left\{T\in\mathbb{R}\,\middle|\,\sum_{t\leq T}\sum_{s\leq t} g^2(s,t,\Delta L_s^1,\Delta L_t^2) > n\right\}.$$

This localization is well-defined due to equation (3.18) and continuity of the multiple integral. The localized process will only make jumps of size less than 1 due to the definition of g. To prove existence of the stochastic integral, we need to show

$$\mathbb{E}\left[\int_{t=-\infty}^{T_n}\int_{\mathbb{R}} (g(\cdot,t,\cdot,y) * \mu_{t_-}^1)^2 \nu^2(dy)\,dt\right] < \infty,$$

where ν^2 denotes the Lévy measure of L^2. The existence criterion is taken from [8]. Observe that

$$\begin{aligned}
&\mathbb{E}[\int_{t=-\infty}^{T_n}\int_{\mathbb{R}}(g(\cdot,t,\cdot,y)*\mu^1_{t_-})^2\nu^2(dy)\,dt]\\
=&\int_{\mathbb{R}}\int_{\mathbb{R}}\mathbb{E}\Big[(g(\cdot,t,\cdot,y)*\mu^1_{(t\wedge T_n)_-})^2\Big]\nu^2(dy)\,dt\\
=&\mathbb{E}\left[\int_{-\infty}^{T_n}\int_{\mathbb{R}}\sum_{s\le t}g^2(s,t,\Delta L^1_s,y)\nu^2(dy)\,dt\right]\\
=&\mathbb{E}\left[\sum_{t\le T_n}\sum_{s\le t}g^2(s,t,\Delta L^1_s,\Delta L^2_t)\right]\le n+1<\infty,
\end{aligned}$$

where the first equality uses Tonelli, the second equality uses the compensator of the integral $(g(\cdot,t,\cdot,y)*\mu^1_{t_-})$ and the last equality uses the definition of the predictable compensator. This implies that the process $g(\cdot,t,\cdot,y)*\mu^1_t \in G_{loc}(\mu^2)$ in the notation of [8] and thus the stochastic integral of it can be defined.

Define the function

$$g_m(s,t,x,y) := g(s,t,x,y)\mathbb{1}_{\{|f(s,t)xy|\ge\frac{1}{m}\}}$$

and observe that $g_m \to g$ pointwise and $|g_m|\le|g|$. The pointwise convergence and domination for multiple integrals implies that

$$\int_{\mathbb{R}}\int_{\mathbb{R}} g_m\,dL^1\,dL^2 \xrightarrow{\mathbb{P}} \int_{\mathbb{R}}\int_{\mathbb{R}} g\,dL^1\,dL^2.$$

For g_m the finite-variation part of our argument implies that

$$\int_{\mathbb{R}}\int_{\mathbb{R}} g_m\,dL^1\,dL^2 = \int_{\mathbb{R}}\left[\int_{\mathbb{R}} g_m\,dL^1\right]dL^2$$

where the left-hand side is a multiple integral and the right-hand side is a stochastic integral. All that remains is to show that

$$\int_{\mathbb{R}}\left[\int_{\mathbb{R}} g_m\,dL^1\right]dL^2 \xrightarrow{\mathbb{P}} \int_{\mathbb{R}}\left[\int_{\mathbb{R}} g\,dL^1\right]dL^2, \tag{3.19}$$

which by uniqueness of limits in probability will imply

$$\int_{\mathbb{R}}\int_{\mathbb{R}} g\,dL^1\,dL^2 = \int_{\mathbb{R}}\left[\int_{\mathbb{R}} g\,dL^1\right]dL^2.$$

Observe first that above we saw that both g and hence g_m are locally integrable. Thus $g-g_m$ is locally integrable and dominated by g, e.g.

$$\int_{\mathbb{R}}\int_{\mathbb{R}}\mathbb{E}\Big[(\{g-g_m\}(\cdot,t,\cdot,y)*\mu^1_{(t\wedge T_n)_-})^2\Big]\nu^2(dy)\,dt\to 0.$$

This implies equation (3.19) and the proof is complete. □

The development of the multiple integration theory relies heavily on the ability to reduce a PRM on a complete separable metric space to the unit-rate Poisson process on $\mathbb{R}$. A generalization to the theory was also mentioned in the introduction of this section, where it is said that, we may generalize to general Poisson processes using simple projections. These two observations allow the authors of [9, 10] to prove their results in a simplistic manner and still create general multiple integration theory. The remainder of this section will explain the first observation. Kallenberg & Szulga [9] states (without proof) that

> "... it will be seen how any multiple integral $X_1 \dots X_d f$ with respect to a positive or symmetric Lévy process $X = (X_1, \dots, X_d)$ in $\mathbb{R}^d$ may be written in the form $\eta^d g$ or $\tilde{\xi}^d h$, respectively, where η is a Poisson process on $S = \mathbb{R}_+ \times (\mathbb{R}^d \setminus \{0\})$, while $\tilde{\xi}$ is a suitably defined symmetric version of η, and where g and h are suitable functions on S^d associated with f. By a Borel isomorphism, we may then reduce the discussion to the case of unit rate Poisson processes on $\mathbb{R}_+$. More generally, any result in the latter context which involves only Poisson and Lebesgue integrals can be easily extended, via a measure isomorphism (cf. Halmos (1950), p. 173), to the case of Poisson processes ξ on an arbitrary measurable space S, such that $(S, \mathbb{E}[\xi])$ is separable. "
>
> — [9] on page 102

In order to perform this reduction, let us recall what defines a PRM with intensity μ.

Definition 3.4.5. *Let $(X, \mathcal{E}, \mu)$ denote a σ-finite measure space. A Poisson random measure Π with control measure μ is a family of random variables $\{\Pi(A)\}_{A \in \mathcal{E}}$ defined on a probability space $(\Omega, \mathcal{F}, P)$, such that*

(i) $\Pi(A)$ is a Poisson random variable with rate $\mu(A)$ for all $A \in \mathcal{E}$.

(ii) For disjoint sets $A_1, A_2, \dots, A_n \in \mathcal{E}$, the random variables $\Pi(A_1), \Pi(A_2), \dots, \Pi(A_n)$ are independent.

(iii) The mapping $A \mapsto \Pi_{(\omega)}(A)$ is a measure on $(S, \mathcal{E})$ for all $\omega \in \Omega$.

In the following, λ will denote the Lebesgue measure on $\mathbb{R}$. The following theorem is central to this simplification.

Theorem 3.4.6 (Ito (1984), [7])**.** *Let S be a complete separable metric space. Then every regular probability measure P on S is standard.*

We shall omit the exact definition of "standard" but mention the parts we utilize and avoid some technicalities. Theorem 3.4.6 implies there exists a bijective map $T : S \to \mathbb{R}$ such that

$$P(A) = \lambda(T(A)), \quad A \in \mathcal{B}(S). \tag{3.20}$$

This theorem allows us to reduce several questions regarding measure spaces to the unit interval.

Remark 3.4.7 (Reduction to unit-rate Poisson process)**.** Given a σ-finite measure μ on $(S, \mathcal{B}(S))$ and using Theorem 3.4.6, we may always find probability measure P and a Radon-Nikodym derivative h on the same space , such that

$$\mu = h\,dP = h\,d(\lambda \circ T),$$

where the last equality is due to (3.20). Thus we may and do define an integral of f with respect to μ through

$$\int_S f\,d\mu := \int_S f\,d(\lambda \circ T) = \int_{\mathbb{R}} f(T^{-1}(x))h(T^{-1}(x))\,\lambda(dx). \tag{3.21}$$

Thus we may reduce the question of integration f wrt. μ to an associated function $\tilde{f} := f(T^{-1})h(T^{-1})$ with respect to the Lebesgue measure λ. In other words, once we know how to integrate functions with respect to Lebesgue measure, we can generalize this to a large class of spaces. Indeed, in this case we may define the stochastic integral of $f : S \to \mathbb{R}_+$ wrt. Π_S as

$$\int_S f\,d\Pi_S := \int_{\mathbb{R}} f(T^{-1}(x))h(T^{-1}(x))\,\Pi(dx). \tag{3.22}$$

The control measure is easily checked to be

$$\begin{aligned}\mathbb{E}\left[\int_S f\,d\Pi_S\right] &= \mathbb{E}\left[\int_{\mathbb{R}} f(T^{-1}(x))h(T^{-1}(x))\,d\Pi\right] \\ = \int_{\mathbb{R}} f(T^{-1}(x))h(T^{-1}(x))\,d\lambda &= \int_S f\,h\,d(\lambda \circ T) = \int f\,d\mu.\end{aligned}$$

3.4.2 Multiple integration for symmetric PRM and Lévy processes

In this section we aim define multiple integrals for symmetric Poisson random measures, e.g. a "symmetrized" Poisson process with values ± 1 instead of just one. This allows us to extend multiple integrals to $\mathbb{R}^d$-valued Lévy processes which induces a PRM on $\mathbb{R} \times \mathbb{R}^d$. Let η denote a PRM with intensity ν with atomic representation $\xi = \sum_n \delta_{\tau_n}$, see Chapter 2 in [10]. The symmetrization $\tilde{\eta}$ of η is defined as

$$\tilde{\eta}(A) = \sum_n \sigma_n \delta_{\tau_n}(A), \quad A \in \mathcal{B}_b(\mathbb{R}_+ \times \mathbb{R}^d) \tag{3.23}$$

where $(\sigma_n)_{n\in\mathbb{N}}$ is a sequence of independent random signs, independent of the original PRM η. In the case of a symmetric real-valued Lévy process as integrand, we may also use the original signs of X to create $\tilde{\eta}$. The main existence criteria for defining multiple integrals for symmetric PRMs is given in the following theorem.

Theorem 3.4.8 (Kallenberg, [10])**.** *Let ξ and $\xi_1, \ldots, \xi_d$ be simple point processes on S with symmetrizations $\tilde{\xi}$ and $\tilde{\xi}_1, \ldots, \tilde{\xi}_d$ generated by independent sign sequences σ and $\sigma_1, \ldots, \sigma_d$ and fix a measurable function $f : S^d \to \mathbb{R}$. Then*

(i) the integral $\tilde{\xi}_1 \cdots \tilde{\xi}_d f$ exists if and only if $\xi_1 \cdots \xi_d f^2 < \infty$ a.s., and similarly $\tilde{\xi}^d f$ and $\xi^d f^2$, when f is symmetric and non-diagonal.

(ii) the following representations hold a.s., whenever either side exists:

$$\tilde{\xi}_1 \cdots \tilde{\xi}_d f = (\sigma_1 \dots \sigma_d) f(\xi_1 \cdots \xi_d), \quad \tilde{\xi}^d f = \sigma^d f(\xi^d).$$

However, we are mainly interested in multiple integrals for symmetric Lévy process, e.g.

$$X_1 \cdots X_d f = \int \cdots \int f(t_1, \dots, t_d) X(dt_1) \cdots X(dt_d). \tag{3.24}$$

where $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_d)$ denote an $\mathbb{R}^d$-valued symmetric Lévy process of pure-jump type. Any pure-jump Lévy process X induces a PRM η through $\Delta X_t = X_t - X_{t-}$ by

$$\eta = \sum_t \delta_{(t, \Delta X_t)} = \sum_n \delta_{(T_n, \Delta X_{T_n})}$$

on $\mathbb{R} \times \mathbb{R}^d \setminus \{0\}$, where $(T_n)_{n \in \mathbb{N}}$ denotes the jump times of $\mathbf{X}$. It can be seen that η has intensity measure $\lambda \otimes \nu$, where ν denotes the Lévy measure of X, λ is the Lebesgue measure on $\mathbb{R}$ and $\otimes$ denotes the product measure. Let f denote a measurable function on $\mathbb{R}^d$ vanishing on the diagonal (two or more coordinates agree), and define the operators L and L' on such a function by

$$Lf(t_1, \dots, t_d; x_1, \dots, x_d) \coloneqq x_1 \dots, f(t_1, \dots, t_d)$$
$$L'f(t_1, \dots, t_d; x_{ij}, i, j \le d) \coloneqq x_{11} \dots x_{dd} f(t_1, \dots, t_d)$$

We define the multiple integral of $\mathbf{X}$ as

$$X_1 \cdots X_d f \coloneqq \tilde{\eta}^d(L'f) = \tilde{\eta}_1 \cdots \tilde{\eta}_d(Lf), \tag{3.25}$$

where $\tilde{\eta}$ denotes the symmetrized version of η and η_j denotes the PRM generated by X^j. In other words, by defining integrals for symmetric Poisson random measures and subsequently defining using η. In this case, we have the following special case of Theorem 3.4.8, which is the main tool in the proof of Theorem 3.2.1.

Corollary 3.4.9 (Kallenberg, [10])**.** *Let $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_d)$ be a symmetric, purely discontinuous Lévy process in $\mathbb{R}^d$ and let η denote its induced PRM. The set function in* (3.27) *extends a.s. uniquely to a linear operator $X_1 \cdots X_d f$, on the domain $\mathcal{D}_X$ of measurable non-diagonal functions f on $\mathbb{R}^d$ with $\eta(Lf)^2 < \infty$ a.s.. In the affirmative case, we have that*

$$X_1 \cdots X_d f \coloneqq \tilde{\eta}^d(Lf) = \tilde{\eta}_1 \cdots \tilde{\eta}_d(Lf) \quad a.s. \tag{3.26}$$

The class $\mathcal{D}_X$ has the following properties

1. *linearity, i.e. if $f, g \in \mathcal{D}_X$ and $\alpha, \beta \in \mathbb{R} \implies \alpha f + \beta g \in \mathcal{D}_X$*
2. *solid, i.e. if $|f| \le |g|$, $g \in \mathcal{D}_X \implies f \in \mathcal{D}_X$*

3. $\mathcal{D}_X$ *is continuous, i.e.*

$$f_n \longrightarrow f,\ |f_n| \le g \in \mathcal{D}_X \quad \Longrightarrow \quad X_1 \cdots X_d f_n \xrightarrow{\mathbb{P}} X_1 \cdots X_d f.$$

Let Γ *be another operator satisfying the above properties and*

$$\Gamma(I_1 \times \cdots I_d) := \prod_{i \le d} \Big(X_i(t_i) - X_i(s_i) \Big), \tag{3.27}$$

where $I_j = (s_j, t_j]$ *are disjoint intervals for all* j*. Then the domain of* Γ*, satisfies that* $\mathcal{D}_\Gamma \subseteq \mathcal{D}_X$*. In other words,* $\mathcal{D}_X$ *is the largest domain for which a multiple stochastic integral may be defined.*

As we will be working with $\mathbb{C}$-valued (or $\mathbb{R}^2$-valued) Lévy processes, we will need to define the multiple integral for it. It is straightforwardly defined by linearity and existence of all integrals as seen in the following definition. Note that for isotropic Lévy processes many of the existence criterion are simplified due to Theorem 2.3.3.

Definition 3.4.10 (Complex multiple integral)**.** *Let* $f = f_1 + i f_2 : \mathbb{R}^2 \to C$ *be a complex-valued function and let* $L = L^1 + iL^2$ *be an isotropic complex-valued Lévy process. Let* $\overline{L}$ *denote the complex conjugate of* L*. We define the double stochastic integal of a complex-valued function* f *wrt.* $(L, \overline{L})$ *whenever all the integrals in equation 3.28 exists. In the affirmative case, we define the multiple stochastic integral of* f *wrt.* $(L, \overline{L})$ *by*

$$\begin{aligned} &\int_{\mathbb{R}} \int_{\mathbb{R}} f(s,u)\, dL_u\, d\overline{L}_s \\ &= \int_{\mathbb{R}} f_1(s,u) + i f_2(s,u)\, d(L^1 + iL^2)_u\, d(L^1 - iL^2)_s \\ &:= \sum_{j,k,m=1}^{2} p(j,k,m) \int_{\mathbb{R}} \int_{\mathbb{R}} f_m(s,u)\, dL_u^j\, dL_s^k, \end{aligned} \tag{3.28}$$

where the function p *is defined by*

$$p(j,k,m) = i^{j-1+m-1+3(k-1)}.$$

References

[1] Ole E Barndorff–Nielsen, Svend Erik Graversen, Jean Jacod, Mark Podolskij and Neil Shephard. "A central limit theorem for realised power and bipower variations of continuous semimartingales". *From stochastic calculus to mathematical finance*. Springer, 2006, 33–68.

[2] Andreas Basse-O'Connor, Raphaël Lachièze-Rey and Mark Podolskij. "Power variation for a class of stationary increments Lévy driven moving averages". *Ann. Probab.* 45.6B (2017), 4477–4528.

[3] Stamatis Cambanis, Jan Rosinski and Wojbor A Woyczynski. "Convergence of quadratic forms in p-stable random variables and θp-radonifying operators". *The Annals of Probability* (1985), 885–897.

[4] Patrick Cheridito. "Mixed fractional Brownian motion". *Bernoulli* 7.6 (2001), 913–934.

[5] Jean-François Coeurjolly. "Estimating the parameters of a fractional Brownian motion by discrete variations of its sample paths". *Statistical Inference for stochastic processes* 4.2 (2001), 199–227.

[6] Donald L Cohn. "Measurable choice of limit points and the existence of separable and measurable processes". *Probability Theory and Related Fields* 22.2 (1972), 161–165.

[7] Kiyosi Itô. *An Introduction to Probability Theory*. Cambridge University Press, 1984.

[8] Jean Jacod and Albert N Shiryaev. *Limit theorems for stochastic processes*. Vol. 288. Springer Science & Business Media, 2013.

[9] O. Kallenberg and J. Szulga. "Multiple integration with respect to Poisson and Lévy processes". *Probab. Theory Related Fields* 83.1-2 (1989), 101–134.

[10] Olav Kallenberg. *Random measures, theory and applications*. Vol. 77. Springer, 2017.

[11] Stanislaw Kwapien and Wojbor A Woyczynski. "Double stochastic integrals, random quadratic forms and random series in Orlicz spaces". *The Annals of Probability* (1987), 1072–1096.

[12] Stanislaw Kwapien and Wojbor A Woyczynski. *Random series and stochastic integrals*. Birkhäuser, 1992.

[13] Benoit B. Mandelbrot and John W. Van Ness. "Fractional Brownian motions, fractional noises and applications". *SIAM Rev.* 10 (1968), 422–437.

[14] Terry R McConnell and Murad S Taqqu. "Decoupling inequalities for multilinear forms in independent symmetric random variables". *The Annals of Probability* (1986), 943–954.

[15] Balram S. Rajput and Jan Rosiński. "Spectral representations of infinitely divisible processes". *Probab. Theory Related Fields* 82.3 (1989), 451–487.

[16] L Chris G Rogers. "Arbitrage with fractional Brownian motion". *Mathematical Finance* 7.1 (1997), 95–105.

[17] Jan Rosinski and WA Woyczynski. "On Itô stochastic integration with respect to p-stable motion: Inner clock, integrability of sample paths, double and multiple integrals". *The Annals of Probability* 14.1 (1986), 271–286.

[18] Gennady Samorodnitsky and Jerzy Szulga. "An asymptotic evaluation of the tail of a multiple symmetric α-stable integral". *The Annals of Probability* (1989), 1503–1520.

[19] Gennady Samorodnitsky and Murad S. Taqqu. *Stable non-Gaussian random processes*. Stochastic Modeling. Stochastic models with infinite variance. Chapman & Hall, New York, 1994.

[20] Jerzy Szulga. "Multiple stochastic integrals with respect to symmetric infinitely divisible random measures". *The Annals of Probability* (1991), 1145–1156.

Part II

Analysis of Sequential Medical Data

Chapter 4

Introduction to sequential medical data

In this part of the thesis, study sequential medical data through two applications. The first concerns electronic health records for patients under suspicion of a serious illness, who have been referred to Diagnostisk Center (diagnostic unit), Silkeborg Regionshospital. The second dataset originates from Stanford Health Care and includes data recorded by a newly implemented sepsis alert system. Common to both applications is the study of how a conceptual underlying graph affects sequentially ordered categorical data and how inference on this graph can be used clinical decision support (automated recommendations, predictions). The general purpose of both application is to do statistical inference on this underlying graph.

In the current chapter, we introduce the thematics topics for the Chapters 5-7. Chapter 5 explains many topics used throughout Chapters 6 and 7 and serves as a technical supplement. In Chapter 6, we study the use of embeddings to introduce (sequential) semantic meaning into vectors representation of events from an electronic health record. The resulting vectors are visualized using t-distributed Stochastic Neighborhood Embedding (see [1]) and show the ability to associate events from the same treatment package with each other. In Chapter 7, we analyze the medication logs following a triggered alert from an automatic sepsis alert system. The aim is to provide a prioritized list of recommended antibiotics and study whether automated alert systems change clinical decisions.

4.1 Sequential medical data

The central object for both datasets is sequential categorical data, i.e. an electronic health record or a medication log. Mathematically, we write this as observing a sequence s, consisting of items, which belong to a finite space of possible items $\mathcal{I}$, and write the sequence as

$$s = (i_1, i_2, \ldots, i_n), \tag{4.1}$$

for some $n \in \mathbb{N}$, denoting the length of the sequence s. In the present chapter and Chapter 5, the term item refers to the value of an entry in a sequence (or

log/electronic health record). In Chapter 6, item denotes an event/encounter from an electronic health record. In Chapter 7, item refers to a treatment package from a 24 hour medication log.

The observed sequences consists of items, e.g. categorical data, which by nature poses a few structural challenges. As described in Section 5.2, we encode categorical data using the so-called one-hot encoding which results in very high-dimensional unit vectors. In linear models (and indeed in many other models) each dimension is given its own parameter. The problem is that we observe many explanatory categorical variables but comparatively few response variables – and this may lead to over-fitting. In our application in Chapter 6, we focus on incorporating and analyzing semantic meaning, and thus overfitting will not be our primary concern. However, we do obtain a lower-dimensional representation of the categorical variables, but only by utilizing the high-dimensional encoding – thus it is unclear whether this improves the over-fitting problem.

4.1.1 Sequential semantic meaning

Semantic meaning refers to the inherent concept or information a word intends to convey. The word "door" conveys the concept of a door, and humans are generally very adept at understanding the semantic concept of a door – we may recognize many different objects as doors through our conceptual understanding of a door. We also know that doors are positioned in walls and give or deny access to rooms. This conceptual understanding and all its implications illustrates the semantic meaning of a concept. Similarly, the position of an item in a sequence (e.g. an event in an electronic health record) conveys sequential semantic meaning to the reader, portraying information from previous entries and inducing information on future entries. This defines the sequential semantic meaning of an item.

However, semantic meaning is not conveyed by naively representing items as levels of a categorical variable. An example of this is Pamol and Panodil, two light sedatives, that may have each their level of a categorical variable, but have common usage patterns (e.g. occurs frequently in the same concepts interchangeably). Without any additional information, these will be entered as separate variables in the sequence. The two events, Pamol and Panodil, are semantically interchangeable, and hence we would prefer a model which constructs semantic meaning to identify this. A different example of semantic meaning is a specific medication for arthritis may often be given with another supplementary medication – a semantic fact that we would like to identify and incorporate in the statistical analysis.

A formalized example of semantic meaning

Suppose that i_{500} in sequence A and i_{3500} in sequence B are interchangeable – but we do not know this. We assign no information to the actual index numbers 500 and 3500. However, if $i_{497}, i_{498}, i_{499}$ and $i_{3497}, i_{3498}, i_{3499}$ are the same (or

nearly the same), we want to infer that i_{500} and i_{3500} are interchangeable (provided that we observe such an occurrence several times throughout the sequences). Similarly, item i_{300} may induce that $(i_t)_{303\leq t\leq 305}$ takes a certain value (or heighten the probability). In this case, we want to infer the semantic relation portrayed by these relations.

4.1.2 Mathematical framework for sequential semantic meaning

The above example of semantic meaning leads to the *context* of an item i_t in a sequence of item $s = (i_1, i_2, \ldots, i_N)$. The context of an entry t is an index subset I with values near t and defines the set of indexes which affect t. One may choose a context definition freely based on the application and we present a few typical choices in Section 5.2.

4.2 Our contributions: Two sequential medical datasets

Common to the two projects with sequential medical data is the study of an underlying graph which connects the items with each other in a manner unknown to us. In Chapter 6, we seek to incorporate these graph relations into a vector representation of each item to obtain a "dense" vector representation with the graph relation embedded in it. Similarly, in Chapter 7, we study whether the graph for two groups are the same by comparing edges, vertices and frequencies. Finally, by weighting the graph edges with their frequency (e.g. probability), we predict the next item based the maximal frequency.

4.2.1 Embedding of sequential semantic meaning

The embedding of items into vector representations is studied in [2, 3] which introduce the neural network called Skip-Gram. Their application is specific for embedding words into vectors with semantic meaning, but their methodology may be applied to the general concept of ordered sequences as in Section 5.1.

In Chapter 6, we apply the algorithm Skip-Gram to items in an electronic health record. Our dataset originates from Silkeborg Regionshospital, Denmark, and consist of 169 electronic health records for 169 patients. The patient cohort was chosen from a pool of patients labeled "Suspicion of Serious Ilness" (SSI) (Danish: Mistanke om Alvorlig Sygdom), which are known to be difficult to diagnose. The label SSI is given by the patients' general practitioner who provides a referral to the diagnostic unit (Danish: Diagnostisk Center) at Silkeborg Regionshospital. Overall, the patients represent a complex diagnostic problem, as they are often multi-sick, i.e. suffering from several concurrent ailments and have specifically been referred by the general practitioner for a more precise diagnostic elucidation. The dataset consists of 178 thousand electronic health records entries which is much less than related studies but what was available in a Danish setting and may be used as a proof of concept.

The underlying hypothesis is that items affect the occurrence of other items and thus carry a semantic meaning in the context of other items. The Skip-

Gram algorithm aims to induce sequential semantic meaning of electronic health records into embeddings of events/treatments based on their sequential order. The t-SNE visualization of the embeddings in Figure 6.4 reveal intriguing relations between events and successfully show that the embedding holds some semantic meaning which is directly interpretable. It is possible to identify groups of related events and by consultation with a medical doctor, these groups represent standard treatment packages ordered in the clinical workflow – the algorithm discovers these relations without prior knowledge of them. This is a promising result for automatic detection and incorporation of sequential semantic meaning.

To quantify the quality of the Skip-Gram vector representations, we evaluate the clustering algorithm k-means' ability to rediscover select annotated groups from Figure 6.4, using the vectors representations as inputs. We compare the performance Skip-Gram representation to a benchmark of a Markov Chain on three different classification tasks. The Markov Chain slightly outperforms the Skip-Gram vector representation on two of the tasks, which may be explained by its more local behavior and the dataset characteristics. However as input in a more complex model (Recurrent Neural Network; RNN) we show that the Skip-Gram representations outperforms the Markov Chain in next-event prediction. To which extent this is due to the Skip-Gram representations or the RNN is difficult to determine, but it does show promising results for the use of vector representations from Skip-Gram as inputs in other statistical models.

4.2.2 Analyzing medication logs for sepsis patients

Sepsis is a life-threatening condition which occurs primarily in the hospital settings and has a high mortality rate. Although not completely understood, it has been closely associated bad hygiene standards, weak immune systems and implanted foreign bodies (e.g. for fixation of broken bones).

In Chapter 7, we analyze the 24 hour time window following an alert (or registration) of potential sepsis. We have two datasets on sepsis which both originate from Stanford Health Care. The first dataset concerns testing a new automatic rule-based sepsis alert system, which during a trial period would registers alerts and for half of the registered alerts, an alert was forwarded to a doctors pager. For the second dataset, we have general sepsis alert registrations, but no group variable, and we study the 24 hour time window following the registration.

The initial goal of the project was search for differences in the medications given to each group in the first dataset – however the analysis showed no major differences in the frequency of common states, the possible order of the medications (through the transitions in a Markov Chain), and visualizing both graphs revealed no major differences. Hence, we concluded that the alert system did not appear to result in a treatment-altering behavior. Following this conclusion, we merged the first dataset with a second, much larger, dataset.

Our main objective for the merged dataset is to predict the next medication using a Markov Chain. A Markov Chain may appear as a simple tool to model a complex decision process, but the sequences are on average 3 entries long and hence more advanced methods were not considered suitable. We report our prediction accuracy on a test set and the result appears reasonable, given the semi-large set of medications and the simple estimation procedure of a Markov Chain.

4.2.3 Conclusion, experiences and thoughts

In the future, we believe that studies of sequential medical data need to incorporate both categorical variables and quantitative measurements to obtain enough signal that a proper, reliable and precise prediction can be made. This could for example suggest a prioritized list of recommended medications. It would require integration of many categorical variables with quantitative measurements, but this appears necessary to move beyond idea and the proof of concept stage to an actual clinical decision support system. Alternatively, much larger datasets (think all national registrations of sepsis) would be another avenue of obtaining more signal. Most probably, this approach would not change the fact that sepsis medication sequences typically consist of 2-7 entries. Hence, this does not appear to be the way forwards. A deeper and more comprehensive analysis of the graph generated by the medications is another avenue which would benefit significantly from more data – the idea of studying sub-graphs for patient sub-groups in collaboration with medical doctors may be fruitful for further hypothesis generation on the causes and factors of sepsis.

References

[1] Laurens van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE”. *Journal of Machine Learning Research* (2008), 2579–2605.

[2] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. “Efficient estimation of word representations in vector space” (2013). arXiv: 1301.3781.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. “Distributed representations of words and phrases and their compositionality”. *Advances in Neural Information Processing Systems*. 2013, 3111–3119.

Chapter 5

Technical introduction to embeddings

This chapter contains supplementary technical material to the chapters 4, 6, and 7. Apart from the first two sections which are interconnected, any section in this chapter can be read almost independently and serves as a quick lookup for the reader. The aim is to make the thesis as self-contained as possible.

In Section 5.1 we formally define the data structures that we study and in Section 5.2 we explain how the data structures is encoded as mathematical vector representation and how we define the context of an event.

5.1 Preliminaries

In this section, embeddings are introduced, both in terms of the presumptions we make on observed data and the corresponding mathematical details, inspired by [5]. We use this general setup in both Chapter 6 and 7. We observe an unordered collection of sequences, $\mathcal{S}$, e.g.

$$\mathcal{S} = [s_1, s_2, \ldots, s_p],$$

for some $p \in \mathbb{N}$. Each sequence s_j is an ordered set of itemsets (typically ordered by time or sequential order) and is denoted by

$$s_j = (X_1^j, X_2^j, \ldots, X_{n_j}^j),$$

for some $n_j \in \mathbb{N}$ which denotes the number of entries in the sequence. The length of a sequence $s = (X_1, \ldots, X_{n_s})$ is defined by $l(s) = |X_1| + \ldots + |X_{n_s}|$, where $|\cdot|$ denotes the number of elements. An itemset X_k^j is an ordered set of items, e.g.

$$X_k^j = \{i_1, i_2, \ldots, i_{m_{j,k}}\}$$

for some $m_{j,k} \in \mathbb{N}$ which denotes the number of items in the set. The ordering of the items is fixed prior to analysis and not relevant to us. An item i belongs to a sequence $s = (X_1, X_2, \ldots, X_n)$ if

$$i \in s \quad \Longleftrightarrow \quad \exists\, k \in \{1, 2, \ldots, n_s\} : \{i\} \subseteq X_k.$$

A sequence $s_a = (Y_1, Y_2, \ldots, Y_{n_a})$ is *contained* in another sequence $s_b = (X_1, X_2, \ldots, X_{n_b})$ if there is a strictly increasing sequence of integers $(m_j)_{1 \leq j \leq n_a}$ such that

$$Y_1 \subseteq X_{m_1}, Y_2 \subseteq X_{m_2}, \ldots, Y_{n_a} \subseteq B_{m_{n_a}}. \tag{5.1}$$

In the affirmative case, s_a is called a *subsequence* of s_b and we denote this by $s_a \sqsubseteq s_b$. We define the set of all *unique* items in the collection $\mathcal{S}$ by

$$\mathcal{I} \coloneqq \left\{ i \;\middle|\; \exists s_j \in \mathcal{S}, \text{ such that there exists } k\colon\ i \in X_k^j \right\}. \tag{5.2}$$

The itemsets and the order of itemsets define the sequence and hence the sequences

$$s_a \coloneqq (\{a\}, \{b\}, \{c\}), \quad s_b \coloneqq (\{a\}, \{c\}, \{b\})$$

are not equal, provided the items were letters in the alphabet.

Examples of this structure could be a corpora of documents (sequences) with items being words, or a database of electronic health records with record entry names being items. In these applications, the itemsets X only contain a single item. The underlying presumption is that the sequential structure defines the purpose and meaning of each itemset – in natural language processing this is called the Distributional Hypothesis [8]. Exactly how we utilize the sequential structure to interpret the Distributional Hypothesis to produce quality embeddings, is a *modeling* question. `Word2vec`, introduced in [14] and computationally enhanced in [15], contains two models, Skip-Gram and CBOW, grounded on the Distributional Hypothesis. Another model is a Markov Chain.

The framework above is used in Sequential Pattern Mining (SPM). In our application of SPM, we analyse treatment packages but due to practical data collection issues we only consider itemsets with a single item. For a good introduction to the field of SPM, we refer to [5]. An example of general sequential databases is grocery shopping: Each customer corresponds to a sequence of purchases and each purchase (or basket) consists of groceries, and each grocery is an item in the above terminology. A central point of analysis in this field is the relation and co-occurrence of items, which can be used for recommendation of additional sales items.

5.2 Encodings and embeddings

In this section, we define embeddings and encodings and discuss their functionality in the analysis of sequences of categorical items (as in Section 5.1). We begin with the following open definitions. These definitions are used in Chapter 6 and we define them for clarification and lookup.

An *encoding* is a mapping $\mathcal{E} : \mathcal{I} \mapsto \mathbb{R}^K$, where $\mathcal{I}$ denotes the set of unique items from Section 5.1 and $K = |\mathcal{I}|$. Thus an encoding does not change the dimensionality. An *embedding* is mapping $\mathcal{E} : \mathcal{T}_1 \to \mathcal{T}_2$, from a space $\mathcal{T}_1$ to another space $\mathcal{T}_2$ such that $\dim(\mathcal{T}_1) \gg \dim(\mathcal{T}_2)$, where $\gg$ denotes much greater than. Clearly, an embedding reduces the dimensionality and typically $\mathcal{T}_1 = \mathbb{R}^{K_1}$ and $\mathcal{T}_2 = \mathbb{R}^{K_2}$, where $K_1 \gg K_2$.

To make embeddings widely applicable and independent of human bias, we do not utilize *any* domain knowledge prior to fitting the embeddings apart from the sequential structure. Therefore, we may as well observe a sequence of (uninterpretable) numbers, i.e.

$$\begin{aligned} s_1 &= (1,4,5,1,7,2,5,315,986,10), \\ s_2 &= (3,1,4,7,1,5,560,2), \\ &\vdots \\ s_p &= (7,2,34,67,23,1050). \end{aligned} \tag{5.3}$$

In other words, for each item $e \in \mathcal{I}$, we associate a unique index number in $\{1,2,\ldots,|\mathcal{I}|\}$. This one-to-one association is an encoding and is sometimes also referred to as *label encoding*.

We proceed to encode each item number in (5.3), as a unit vector in $\mathbb{R}^{\mathcal{I}}$, i.e.

$$\mathcal{I} \ni i \mapsto \mathbf{e}_i \in \mathbb{R}^{\mathcal{I}},$$

where

$$(\mathbf{e}_i)_k = \begin{cases} 1, & \text{if } i = k \\ 0, & \text{otherwise.} \end{cases}$$

This mapping is called a *one-hot encoding*, since all entries are "cold" (or more precisely, zero) except the ith entry which is 1 and is also implemented in scikit-learn [18]. Using this processs, we may represent a sequence $s = (i_1, i_2, \ldots, i_{n_s})$ as

$$s = \left(e_{i_1}, e_{i_2}, \ldots, e_{i_{n_s}}\right) \tag{5.4}$$

where the subscript i_k are understood as the number encoding for the item i_k from equation (5.3).

We are at the starting point in the process of learning to infer the meaning of each number by their sequential placement, i.e. according to the Distributional Hypothesis, "learning the language" for the collection of sequences. Recall that $\mathcal{I}$ the denotes the set of unique items in the collection $\mathcal{S}$. The one-hot encoding is an encoding procedure, i.e. it does not change the inherent dimensionality but it does create a mathematical vector representation (although it is the most naive representation). Most of the time, label encoding and one-hot encoding are both performed in succession. In this case the procedure of applying both is often still referred to as one-hot encoding.

The *context* of an entry, denoted $\mathcal{C}$, is a user-defined set of points surrounding an entry. It is usually independent of the index (apart from corner cases at the beginning and end of a sequence). In Skip-gram, the context of an entry i_j, $\mathcal{C}$, is defined as the window of size C around the entry, e.g.

$$i_1, i_2, \ldots, \underbrace{i_{k-C}, i_{k-C+1}, \ldots, i_k, i_{k+1}, \ldots, i_{k+C}}_{\text{window of size } C \text{ around } k \text{ (excluding } i_k)}, i_{k+C+1}, \ldots. \tag{5.5}$$

In a first order Markov Chain, the context of an entry is merely defined as prior entry, i.e.

$$\ldots, \underbrace{i_{j-1}}_{\text{context of } i_j}, i_j, i_{j+1}, \ldots,$$

whereas a kth order Markov Chain has the context

$$\ldots, i_{j-k-1}, \underbrace{i_{j-k}, \ldots, i_{j-1}}_{\text{context of } i_j}, i_j, i_{j+1}, \ldots,$$

Naturally, the choice of context heavily influences the embedding.

5.3 Neural network architectures

Neural networks is a current hot topic, and here we define a couple neural network architectures used in Chapter 6. We describe feed-forward fully-connected neural networks and recurrent neural networks in some detail. In Section 5.4 we elaborate on the optimization of feed-forward neural networks but we refrain from presenting "back-propagation through time" for recurrent neural networks, as it requires significant notation and further technical details outside the scope of our usage.

A neural network has no strict definition and should simple be understood as an (extensive) composition of mappings which incorporates some parameters. It is most easily understood as a feed-forward fully-connected neural network, of which linear regression is a special case. We describe this in the next section.

5.3.1 Feed-forward fully-connected neural networks

A feed-forward fully-connected neural network with K layers is a composition of mappings such that the following recurrent relation holds true

$$z^{[k]} := a^{[k]}(W^{[k]} z^{[k-1]}) \quad \text{for } 1 \leq k \leq K, \tag{5.6}$$

where $a^{[k]}$ is an activation function, $W^{[k]}$ is a matrix of suitable dimension, $z^{[0]} := x$ is the initial input vector and $z^{[k]}$ is the output vector of the kth composition (see Figure 5.1 for an illustration of a single hidden layer fully-connected feed-forward neural network). An activation function is applied element-wise (on each entry of a vector) and can be any function, as long as it is (nearly) differentiable – typical choices are tanh or $\sigma(x) := 1/(1+\exp(-x))$. We omit the bias term, since it may be incorporated into the vector of each layer (explained in Chapter 5 of [1]). In this way, the jth entry of z_k is given by

$$z_j^{[k]} = a^{[k]}\Big((W^{[k]} z^{[k-1]})_j\Big) = a^{[k]}\Big(\sum_m W_{jm}^{[k]} z_m^{[k-1]}\Big), \tag{5.7}$$

where $W_{jm}^{[k]}$ denotes the (j,m)th entry in $W^{[k]}$ and similarly for the vector $z^{[k-1]}$. Each composition f_k is called the kth layer. The "inner" layers, that is the layers

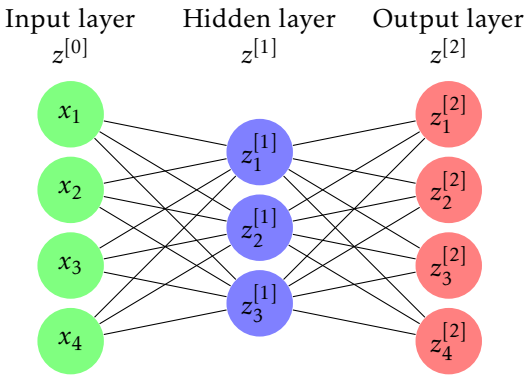


Figure 5.1: A feed-forward fully-connected neural networks architecture.

with $1 \leq k \leq K-1$, are called hidden layers. The term feed-forward refers to the fact that inputs in a layer only depend on past layers (and not on future layers). 'Fully-connected' refers to the fact that the value of $(z^{[k]})_j$ is (possibly) affected by every value of the previous layers, as seen in equation (5.7) and Figure 5.1.

Feed-forward neural networks are often used with many layers to construct highly non-linear functions and they are often difficult to interpret due to the composite output (Skip-Gram is somewhat an exception, see Section 5.5).

5.3.2 Recurrent neural networks

A recurrent neural network (RNN) is an neural network architecture used primarily to model sequential data inputs (contrary to feed-forward neural networks which are not suitable for this). For example, words in a text documents or speech (in machine translation, language models) with input of the form

$$x = (x_1, x_2, \ldots, x_T)$$

and often (depending on use case and specific choice of network architecture) with associated response variables $(h_1, h_2, \ldots, h_T)$. A recurrent neural network is defined through the central concept of a cell. It is often illustrated as in Figure 5.2. Recurrent neural networks are optimized using "back-propagation

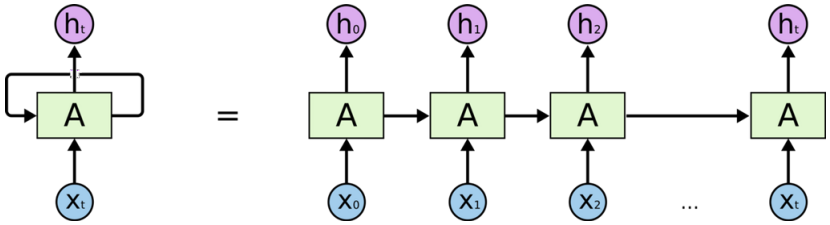


Figure 5.2: Unfolding of a recurrent neural network.

through time" or variants thereof. In Chapter 6, we mainly use RNN as a quick

reference for a temporal model, capable of accepting high-dimensional vector of highly varying values in sequential order and producing predictions.

5.4 Loss function, gradient descent and back-propagation

Neural networks contain many parameters and compositions which allows them to fit highly nonlinear data surfaces. Gradient descent and back-propagation is often used to train this parameters. The first, gradient descent is the method used to optimize the weights of a composition of mappings, and the step-wise process utilized in feed-forward neural networks to pass errors backwards, is called back-propagation (see [22]). To explain gradient descent, consider an arbitrary finite composition of mappings

$$x \mapsto \hat{y}(x) \coloneqq (f_K \circ f_{K-1} \circ \cdots \circ f_1)(x) = f_K(f_{K-1}(\ldots(f_1(x)))) \tag{5.8}$$

where $x \in \mathbb{R}^m$ denotes an input vector and $(f_i)_{i \leq K}$ denotes a family of mappings from $\mathbb{R}^{m_{i-1}} \to \mathbb{R}^{m_i}$. The result $\hat{y}(x)$ is the estimate of our target variable (henceforth referred to as the estimate or output). The target variable can either be a observed label or a regression value (as in supervised learning), or a target value formulated through an optimization criterion (as in unsupervised learning). An example of the latter is k-means, used in Chapter 6.

5.4.1 Loss and cost function

A *loss function* quantifies the error between the estimate and the target variable, i.e.

$$\mathcal{L}(y, \hat{y}),$$

where $\hat{y}$ is the output of the mapping and y is the true output. A good loss function quantifies different types of errors in a desirable way, given the application. Thus there is no universal good choice for loss function. Common choices are squared error, $\mathcal{L}(y, \hat{y}) = \frac{1}{2}\sum_i (y_i - \hat{y}_i)^2$ for regression (fitting a numerical value) and cross-entropy $\mathcal{L}(y, \hat{y}) = -\sum_i y_i log(\hat{y}_i)$ for classification/prediction. The choice depends on domain, target output and fit choices (for example heavily penalizing large values). Given a loss function $\mathcal{L}$, the cost function is given as

$$C = \frac{1}{N}\sum_{j=1}^{N} \mathcal{L}(y_j, \hat{y}_j),$$

where N denotes the number of samples, and hence the cost function is simply the average loss function. Caution is advised, since the names "loss-function" and "cost function" are often used synonymously in the literature. We choose to distinguish between the two, as we later optimize the weights based on the cost function. The formulation of "optimizing the loss function" is confusing, as it may refer to changing the loss function to a different choice or optimizing it with gradient descent.

5.4.2 Gradient descent

In this section, we use gradient descent to adjust the weights in the composition of mappings in equation (5.8) with the objective of minimizing the cost function. We start with the cost function and view it as a function of an arbitrary parameter z. The linearity of differentation implies that

$$\frac{\partial C}{\partial z} = \sum_{j=1}^{p} \frac{\partial \mathcal{L}(\mathbf{y}_j, \hat{\mathbf{y}}_j)}{\partial z}.$$

Thus to decrease the notational load, we omit the sum and just study derivatives of the mapping

$$\mathbf{x} \longrightarrow \mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}(\mathbf{x})) = \mathcal{L}(\mathbf{y}, (f_K \circ f_{K-1} \circ \cdots \circ f_1)(\mathbf{x})).$$

Define $z^{[0]} := x$ and $z^{[k]} := f_k(z^{[k-1]})$ for $1 \leq k \leq K$. Suppose $f_j(x) = a(Wx)$ for some matrix $W = \{w_{ij}\}$ of suitable dimension and activation function a (but it could be any mapping which introduces some parameter z for which it is meaningful to differentiate the loss function). Provided that $1 \leq j < K$, the chain rules implies that for a fixed w_{ij},

$$\frac{\partial \mathcal{L}(\mathbf{y}_j, \hat{\mathbf{y}}_j(w_{ij}))}{\partial w_{ij}} = \frac{\partial L}{\partial a^{[K]}} \frac{\partial a^{[K]}}{\partial a^{[K-1]}} \cdots \frac{\partial a^{[j]}}{\partial w_{ij}},$$

where the derivatives/fractions are the Jacobian matrices. The structure of feed-forward neural networks allows us to compute each of these Jacobian matrices in a step-wise procedure called back-propagation which we elaborate on in the next subsection. Let $w_{ij}^{(0)}$ denote the initial value of w_{ij}. The parameter update at iteration t of gradient descent is given by

$$w_{ij}^{(t)} = w_{ij}^{(t-1)} - \alpha \frac{\partial \mathcal{L}(\mathbf{y}_j, \hat{\mathbf{y}}_j)}{\partial w_{ij}},$$

where α denotes the learning rate (a hyperparameter, see Section 5.5.4). The process of "back-propagating" errors (or derivatives) stepwise towards the input is called backpropagation (see [22]) and is the general method used to optimize feed-forward neural networks with multiple layers. Certain optimization algorithms may modify how this optimization is performed.

5.4.3 Back-propagation

In this section, we describe back-propagation as a stepwise process of updating the weights in a feed-forward neural network. We lean heavily on the excellent explanation provided in Chapter 5 of the book [1]. Consider the composition of mappings (or layers) from Subsection 5.3.1, i.e.

$$\mathbf{z}^{[j]} := a^{[j]}(\mathbf{W}^{[j]}\mathbf{z}^{[j-1]}), \quad \text{for } 1 \leq j \leq K,$$

where $a^{[j]}$ is an activation function, $\mathbf{W}^{[j]}$ is a matrix of suitable dimension and $z^{[j]}$ is a vector with $z^{[0]} = x$, where x is the input vector of the feed-forward

neural network, and finally $z^{[K]}$ is the output of the feed-forward neural network (i.e. the result of the last mapping). Define the notation ($\mathbf{h}$ should not be interpreted as hidden unit – we merely needed more notation to simplify calculations later)

$$\mathbf{h}^{[j]} = \mathbf{W}^{[j]}\mathbf{z}^{[j-1]}, \quad 1 \leq j \leq K.$$

To measure the error, we use a loss function $\mathcal{L}(\mathbf{y}, \mathbf{z}^{[K]})$ and the first goal is to compute the derivative

$$\frac{\partial \mathcal{L}(\mathbf{y}, \mathbf{z}^{[K]})}{\partial w_{ij}^{[K]}},$$

where $w_{ij}^{[K]}$ is the (i,j)'th entry of the matrix $\mathbf{W}^{[K]}$. Since $w_{ij}^{[K]}$ only enters in the ith coordinate of $\mathbf{z}^{[K]}$ by equation (5.7) and combine this with the chain rule, we may write this as

$$\frac{\partial \mathcal{L}(\mathbf{y}, z^{[K]})}{\partial w_{ij}^{[K]}} = \frac{\partial \mathcal{L}(y, \mathbf{z}^{[K]})}{\partial \mathbf{z}_i^{[K]}} \frac{\partial a^{[K]}}{\partial \mathbf{h}_i^{[K]}} \frac{\partial \mathbf{h}_i^{[K]}}{\partial w_{ij}^{[K]}}.$$

Define for each j the notation $\delta_j^{[K]}$ by

$$\delta_j^{[K]} := \frac{\partial \mathcal{L}(y, \mathbf{z}^{[K]})}{\partial \mathbf{z}_i^{[K]}} \frac{\partial a^{[K]}}{\partial \mathbf{h}_i^{[K]}},$$

and refer to $\delta_i^{[K]}$ as the errors in the K'th layer. Observe that

$$\frac{\partial \mathbf{h}_i^{[K]}}{\partial w_{ij}^{[K]}} = z_j^{[K-1]},$$

and thus the overall derivative becomes

$$\frac{\partial \mathcal{L}(\mathbf{y}, \mathbf{z}^{[K]})}{\partial w_{ij}^{[K]}} = \delta_i^{[K]} z_j^{[K-1]}.$$

Similarly, define $\delta_j^{[K-1]}$ for the layer $[K-1]$ by

$$\delta_j^{[K-1]} := \frac{\partial \mathcal{L}}{\partial h_j^{[K-1]}}.$$

We may rewrite this derivative using the chain rule

$$\delta_j = \left(\frac{\partial \mathcal{L}}{\partial \mathbf{h}^{[K]}}\right)^T \frac{\partial \mathbf{h}^{[K]}}{\partial h_j^{[K-1]}} = \sum_k \frac{\partial \mathcal{L}}{\partial h_k^{[K]}} \frac{\partial h_k^{[K]}}{\partial h_j^{[K-1]}}$$

from which we may simplify the notation to

$$\delta_j^{[K-1]} = \sum_k \delta_k^{[K]} w_{kj} \frac{\partial a^{[K-1]}}{\partial h_j^{[K-1]}} = \partial h_j^{[K-1]} \sum_k \delta_k^{[K]} w_{kj}^{[K]}.$$

This relation holds for any given $\delta_j^{[s]}$ for $1 \leq s \leq K$ by iteration. The procedure of calculating the δ through the previously calculated δs is called *back-propagation*. The take-away is that we may compute $\delta_j^{[s]}$ through the previously calculated δ's (by iterating this procedure backwards) and obtain the desired derivatives for gradient descent through the formula

$$\frac{\partial \mathcal{L}}{\partial w_{ij}^{[s]}} = \frac{\partial \mathcal{L}}{\partial \mathbf{h}_i^{[s]}} \frac{\partial \mathbf{h}_i^{[s]}}{\partial w_{ij}^{[s]}} = \delta_i^{[s]} z_j^{[s-1]},$$

where $w_{ij}^{[s]}$ denotes the (i,j)'th entry of $\mathbf{W}^{[s]}$ in the sth layer.

5.5 Skip-Gram

Skip-Gram was introduced as a method for natural language processing in [14] with supplementary computational optimization techniques in [15]. Its application is natural language processing but the methodology is applicable to the sequential structure described in Section 5.1. In Chapter 6, we apply Skip-Gram to obtain word embeddings/representations of events in an electronic health record.

Skip-Gram is a single hidden layer, feed-forward fully-connected neural network and thus may be optimized using back-propagation through the network layers. However, the neural network for Skip-Gram is very simple and we instead choose to compute the derivatives of the composition of mappings directly in order to explicitly understand the update of each parameter.

We observe a collection of sequences $\mathcal{S}$ as in Section 5.1 which we assume are one-hot encoded, i.e. of the form in equation (5.4). A central part of the Skip-Gram method is the creation of the inputs. This is described in detail in the next section.

5.5.1 Skip-Gram input generation

In this section, we describe the input generation of Skip-Gram, given as sequence s. The main point here is that the sampling distribution inside the context window is a hyper-parameter for which several choices are available. We use a uniform distribution over the context window, but the original article [14] suggests that given an entry i_k to more frequently sample the neighboring entries compared to the entries near the edge of context window.

The input generation procedure is automatic and only depends on a set of hyper-parameters – hence Skip-Gram is an unsupervised learning method which generates its own input and output with the training performed as a supervised learning method utilizing the generated input/output. Suppose we have a sequence s, given by

$$s = (i_1, i_2, \ldots, i_k). \tag{5.9}$$

In Skip-Gram, the context of an entry i_j, denoted $\mathcal{C}$, is defined in equation (5.5).

We will call the inputs items (x, y) for a pair, as it corresponds to an item x and an item y in the context of x. Creating the pairs is straightforward, disregarding the corner cases of starting and endpoints of the sequence and is illustrated in Algorithm 1. For the beginning and end of the sequence, we "wrap" around the sequence. That is, for an entry with index less than window size, the left-most part of the window is small than the right – we fix this by enlarging the left-part of the context to include the last entries of the sequence – wrap around). This is done to avoid overestimating (and hence sampling too many) pairs at the beginning of the sequence. Note that this introduces synthetic/false pairings but due to the average length of our sequence and a window size of 10, we deem this to be negligible. The same is done for the end of the sequence. We recently realized that a preferable method would be to decrease the sampling frequency of entries at the beginning and end of the sequence.

Skip-Gram was originally introduced with the creation several pairs for each input and a single parse through the sequence, but this is nearly equivalent to passing through the dataset several times (and thus obtaining several pairs for each input). We perform the latter, and this small change does allow for re-sampling of the same output word, which may put a bit more emphasis on frequent word pairings.

According to the discussion in Section 5.2, to each item, we create the label encoding, e.g. a 1-1 mapping to a unique index. Next, to each index we perform one-hot encoding to create a 1-1 mapping to a unit vector in $\mathbb{R}^{|\mathcal{I}|}$. For the practical map composition of Skip-Gram, we understand x as an input vector, which is simply the label combined with one-hot encoding of the event i, resulting in a unit vector $\mathbf{e}_x$ in $\mathbb{R}^{|\mathcal{I}|}$, and similarly for y.

Algorithm 1: SkipGram pairing generator(sequence)

```
Result:
list sequence ;                                   /* or a vector */
int sequenceEnd = length(sequence);
vector 𝒟 = sampling distribution ;          /* hyperparameter, a
 probability vector */
int windowSize ;                              /* a hyperparameters */
int returnList = list() ;                  /* initialize empty list */
for index in sequenceIndexes do
    if index ≤ windowSize then
        wrappedContext = Context-wrap around (see text);
        sampledEntry = ample from wrapped context;
    else if sequenceEnd - index then
        wrappedContext = Context-wrap around (see text);
        Sample from wrapped context;
    else
        sampledEntry = sample an entry from the window around index
          according to 𝒟;
        pair = (sequence[index],sampledEntry) ;     /* a vector or a
          tuple */
        returnList.append(pair);
    end
end
return returnList ;                               /* a list of pairs */
```

5.5.2 Skip-Gram mapping

The Skip-Gram composition of mapping is studied in this section. Skip-Gram is merely a composition of mapping having some input $\mathbf{x}$ and true output class $\mathbf{y}$. Here, $\mathbf{x}$ and $\mathbf{y}$ is a pair from the pairing strategy in Section 5.5.1 and both are high-dimensional unit vectors in $\mathbb{R}^{|\mathcal{I}|}$. The definition of Skip-Gram is simple and contains only a few mapping. These are given as

$$\begin{aligned} f_1 &: \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}^D, \quad f_1(\mathbf{x}) = W\mathbf{x} \\ f_2 &: \mathbb{R}^D \to \mathbb{R}^{|\mathcal{I}|}, \quad f_2(\mathbf{z}) = W'\mathbf{z} \\ f_3 &: \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}^{|\mathcal{I}|}, \quad f_3(\mathbf{z}) = \text{softmax}(\mathbf{z}), \end{aligned}$$

with the composition

$$\mathbf{x} \mapsto \hat{\mathbf{y}}(\mathbf{x}) = f_3(f_2(f_1(\mathbf{x}))) = W'W\mathbf{x},$$

where W is an $|\mathcal{I}| \times D$-dimensional matrix of weights (or, paramters), W' is a $D \times |\mathcal{I}|$-dimensional matrix of weights (or parameters), D denotes the embedding dimension (a hyperparameter, see Section 5.5.4) and finally softmax denotes

the function

$$\text{softmax} : \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}^{|\mathcal{I}|}, \quad \text{softmax}(z) = \left(\frac{\exp(z_j)}{\sum_k \exp(z_k)} \right)_{j=1,\dots,|\mathcal{I}|}.$$

Further simplications can be made on the Skip-Gram mapping, since $\mathbf{x}$ and $\mathbf{y}$ are unit vectors. Indeed, suppose that $\mathbf{x} = \mathbf{e}_I$ and $\mathbf{y} = \mathbf{e}_O$, where the I stands for input vector and the O for output. Let $\mathbf{w}_i$ denote the ith column of $\mathbf{W}$ and let $\mathbf{w}'_i$ denote the ith row of $\mathbf{W}'$. Since x is a unit vector, the function f_1 yields the Ith column of $\mathbf{W}$ as can be seen by the following computation

$$\begin{bmatrix} — & \mathbf{w}'_1 & — \\ & \vdots & \\ — & \mathbf{w}'_{|\mathcal{I}|} & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_{|\mathcal{I}|} \\ | & | & & | \end{bmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{bmatrix} — & \mathbf{w}'_1 & — \\ & \vdots & \\ — & \mathbf{w}'_{|\mathcal{I}|} & — \end{bmatrix} \begin{pmatrix} | \\ \mathbf{w}_I \\ | \end{pmatrix} = \begin{pmatrix} \mathbf{w}'_1 \mathbf{w}_I \\ \mathbf{w}'_2 \mathbf{w}_I \\ \vdots \\ \mathbf{w}'_{|\mathcal{I}|} \mathbf{w}_I \end{pmatrix},$$

where the bars indicate in which direction the vectors extends (e.g. row or column vector). Notice that the final matrix is simply a vector of inner products between $\mathbf{w}'_j$ and $\mathbf{w}_I$. Thus the jth entry in $\hat{y}(x)$ is given by

$$\hat{y}_j(x) = \frac{\exp(\mathbf{w}'_j \mathbf{w}_I)}{\sum_k \exp(\mathbf{w}'_k \mathbf{w}_I)}, \quad j \in 1, \dots, |\mathcal{I}|. \tag{5.10}$$

This is interpreted as an estimate of the probability that y is the context word, given the input word x, e.g.

$$\mathbb{P}(y \mid x) \leftarrow \hat{\mathbf{y}}_j(x).$$

To measure the error of mapping, we use the cross-entropy loss function, given by

$$\mathcal{L}(y, p) = -\sum_{j=1}^{|\mathcal{I}|} y_i \log p_i.$$

for a probability vector y (true distribution) and another probability vector p (candidate/estimated distribution) – we interpret the probability vector as a distribution. Since the vector y is simply $\mathbf{e}_O$, this reduces to

$$\begin{aligned} \mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}(x)) &= -\sum_{j=1}^{|\mathcal{I}|} y_j \log \hat{y}_j(x) = -\log \frac{\exp(\mathbf{w}'_O \mathbf{w}_I)}{\sum_k \exp(\mathbf{w}'_k \mathbf{w}_k)} \\ &= \log\Big(\sum_k \exp(\mathbf{w}'_k \mathbf{w}_I) \Big) - \mathbf{w}'_O \mathbf{w}_I. \end{aligned} \tag{5.11}$$

Using this loss function, we obtain the following total cost function

$$\begin{aligned} C &= -\frac{1}{N}\sum_{j=1}^{N}\mathcal{L}(y,\hat{y}(x)) \\ &= \frac{1}{N}\sum_{j=1}^{N}\Big\{\log\Big(\sum_{k}\exp\big(\mathbf{w}'_k[\mathbf{w}_I]_j\big)\Big)-\big[\mathbf{w}'_O\big]_j[\mathbf{w}_I]_j\Big\}, \end{aligned} \tag{5.12}$$

where the sum over N training samples, and $[\mathbf{w}_I]_j$ denotes true input class of jth sample ($x_j = e_I$) and similarly for $y_j = e_O$. It is common to update the weights after each (or a batch) of samples instead of the total cost function as above. We adopt this approach as well. This modification of the update strategy is mentioned in [22] as being effective at avoiding getting stuck in local minima. It is sometimes referred to as *stochastic* gradient descent. The only remaining step is to optimize the cost function which will be described in Section 5.5.3.

5.5.3 Skip-Gram gradient descent

In this section we describe the procedure for updating the weights in Skip-Gram, i.e. the matrices $\mathbf{W}$ and $\mathbf{W}'$. This is done by minimizing the cost function in equation (5.12). Note that the cost function is later modified by Negative Sampling from Section 5.5.5. Let z denote an entry (or, weight/parameter) from $\mathbf{W}$ or $\mathbf{W}'$ (for a general mapping, any trainable (free) weight/parameter in the mapping). The goal is to compute

$$\frac{\partial\mathcal{L}(\mathbf{y},\hat{\mathbf{y}}(x))}{\partial z},$$

with $\hat{\mathbf{y}}$ defined in equation 5.10 and $\mathbf{y}$ is the one-hot encoded vector of the true label. We proceed to update z at iteration t of the training procedure according to the equation

$$z^{(t)} = z^{(t-1)} - \alpha\frac{\partial\mathcal{L}(y,\hat{y}(x))}{\partial z},$$

where $z^{(t)}$ denotes value of z at the start of iteration t and α denotes a learning rate, see Section 5.5.4 and subsection 5.4. Note that we must first compute all such derivatives and then simultaneously update all z.

Let $z = (\mathbf{w}_i)_j$ denote the jth entry in the ith column of $\mathbf{W}$ (also commonly known as $\mathbf{W}_{ji}$) and consider the derivative of the following mapping from equation (5.11)

$$(\mathbf{w}_i)_j \mapsto \mathcal{L}(\mathbf{y},\hat{\mathbf{y}}(x)) = \log\Big(\sum_{k}\exp(\mathbf{w}'_k\mathbf{w}_I)\Big)-\mathbf{w}'_O\mathbf{w}_I,$$

for a fixed pair input/output vectors $x = e_I$ and $y = e_O$ from the sampling methodology in Section 5.5.1 and where y_O, $\hat{y}_O(x)$ denotes the Oth entry of the

probability vectors. We may divide this into whether $i = I$ or not, and obtain

$$\frac{\partial \mathcal{L}}{\partial (\mathbf{w}_i)_j} = \begin{cases} \frac{\sum_k (\mathbf{w}'_k)_j \exp(\mathbf{w}'_k \mathbf{w}_I)}{\sum_m \exp(\mathbf{w}'_m \mathbf{w}_I)} - (\mathbf{w}'_O)_j, & \text{if } i = I. \\ 0, & \text{if } i \neq I. \end{cases}$$

Observe that this will only update $\mathbf{w}_I$, Ith column of $\mathbf{W}$ corresponding to the input item. Written as a vector, the above gradient is simply

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}_I} &= \sum_k \left[\frac{\exp(\mathbf{w}'_k \mathbf{w}_I)}{\sum_m \exp(\mathbf{w}'_m \mathbf{w}_I)} \right] \mathbf{w}'_k - \mathbf{w}'_O \\ &= \underbrace{\left(\frac{\exp(\mathbf{w}'_O \mathbf{w}_I)}{\sum_m \exp(\mathbf{w}'_m \mathbf{w}_I)} - 1 \right)}_{\in(-1,0)} \mathbf{w}'_O + \sum_{k \neq O} \left[\frac{\exp(\mathbf{w}'_k \mathbf{w}_I)}{\sum_m \exp(\mathbf{w}'_m \mathbf{w}_I)} \right] \mathbf{w}_k, \qquad (5.13) \\ \frac{\partial L}{\partial \mathbf{w}_{\neq I}} &= 0, \end{aligned}$$

which results in the update

$$\mathbf{w}_I^{(t)} = \mathbf{w}_I^{(t-1)} - \alpha \left(\frac{\sum_k \mathbf{w}'_k \exp(\mathbf{w}'_k \mathbf{w}_I)}{\sum_k \exp(\mathbf{w}'_k \mathbf{w}_I)} - \left(\frac{\exp(\mathbf{w}'_k \mathbf{w}_I)}{\sum_m \exp(\mathbf{w}'_m \mathbf{w}_I)} - 1 \right) \mathbf{w}'_O \right). \qquad (5.14)$$

Note that this update adds a bit of $\mathbf{w}'_O$ from $\mathbf{w}_I$ while subtracting a bit of $\mathbf{w}'_{k \neq O}$. The effect is that the inner product $\langle \mathbf{w}'_O, \mathbf{w}_I \rangle$ is increased while it is decreased for $\mathbf{w}'_{k \neq O}$. For the probability estimate $\hat{\mathbf{y}}$, this results in increased $\hat{y}_O$ while it decreases $\hat{y}_{k \neq O}$. This makes intuitively sense, i.e. we observe the pair (x, y) and the update increases the estimated probability of observing the pair.

Similarly, let $z = (\mathbf{w}'_i)_j$ (also known as $\mathbf{W}'_{ij}$). We consider the mapping

$$(\mathbf{w}'_i)_j \mapsto \mathcal{L}(y, \hat{y}(x)) = -y_O \log \hat{y}_O(x) = \log \Big(\sum_k \exp(\mathbf{w}'_k \mathbf{w}_I) \Big) - \mathbf{w}'_O \mathbf{w}_I,$$

and wish to compute its derivative with respect to $(\mathbf{w}'_i)_j$. Once again, we may divide into the cases whether $O = i$ or not, and obtain

$$\frac{\partial L}{\partial (\mathbf{w}'_i)_j} = (\mathbf{w}_I)_j \left(\frac{\exp(\mathbf{w}'_O \mathbf{w}_I)}{\sum_k \exp(\mathbf{w}'_k \mathbf{w}_I)} - \mathbb{1}_{\{i=O\}} \right).$$

This results in the following parameter update

$$(\mathbf{w}'_i)_j^{(t)} = (\mathbf{w}'_i)_j^{(t-1)} - \alpha (\mathbf{w}_I)_j \left(\frac{\exp(\mathbf{w}'_O \mathbf{w}_I)}{\sum_k \exp(\mathbf{w}'_k \mathbf{w}_I)} - \mathbb{1}_{\{i=O\}} \right).$$

Written as a vector, this corresponds to

$$\mathbf{w}_i'^{(t)} = \mathbf{w}_i'^{(t-1)} - \alpha (\mathbf{w}_I) \left(\frac{\exp(\mathbf{w}'_O \mathbf{w}_I)}{\sum_k \exp(\mathbf{w}'_k \mathbf{w}_I)} - \mathbb{1}_{\{i=O\}} \right). \qquad (5.15)$$

Observe that we always update all of $\mathbf{W}'$ by this. This update adds a bit of $\mathbf{w}_I$ to $\mathbf{w}'_O$ and subtracts a bit of $\mathbf{w}_I$ from $\mathbf{w}'_{k \neq O}$. This increases the inner products $\langle \mathbf{w}_I, \mathbf{w}'_O \rangle$, while decreasing $\langle \mathbf{w}_I, \mathbf{w}'_{k \neq O} \rangle$. Thus the estimated probability $\hat{y}_O$ is once again increased and $\hat{y}_{k \neq O}$ is decreased.

5.5.4 Skip-Gram hyper-parameters

Skip-Gram contains a multitude of hyper-parameters, all of which affect the cost function. Hence to obtain a good result, it is important to study each of them and how they affect both the cost function and the interpretation of the model. Overall, we have the following hyper-parameters along their effect described briefly in the parenthesis

1. Embedding dimension (affects the dimension of the embeddings).
2. Negative sampling (from [15], both if used and the amount of negative samples).
3. Window size for the context (affects the possible set of pairings).
4. Window sampling distribution (affects the frequency of pairings).
5. Cut-off level (remove rare entries and homogenize – but at the cost of throwing away/masking rare events).
6. Cut-off technique (replace entry by a standard token in the sequence or completely remove entry from the sequence).
7. Noise distribution for Negative Sampling (affects the frequency of items with which items are used as negative samples).

Furthermore, initialization of weight matrices is not a hyper-parameter that affects the obtained representations but it may nonetheless be useful for faster convergence of the algorithm.

5.5.5 Negative Sampling

Negative Sampling was introduced in [15] as a computational optimization for Skip-Gram. The paper introduces Negative Sampling as the procedure performed by optimizing the following cost function,

$$\sigma(\mathbf{w}'_O \mathbf{w}_I) + \sum_{i=1}^{k} \mathbb{E}_{\mathbf{w}_i \sim P_n(\mathbf{w})}[\log \sigma(-\mathbf{w}_i \mathbf{w}_I)], \tag{5.16}$$

where $\sigma(x) = 1/(1+\exp(-x))$, k denotes the number of negative samples, $P_n(\mathbf{w})$ denotes a noise distribution on all items and $\mathbf{w}_i$ is the vector representation of an item i sampled from the noise distribution P_n. The goal of equation (5.16) is to train the model to distinguish between the actual output and the noise distribution – as described in the original methodology paper [7] on noise-contrastive estimation, of which Negative Sampling is a special case. It describes the training task as learning to distinguish between the target word (item) $\mathbf{w}_O$ and the noise samples $(\mathbf{w}_i)_{1 \leq i \leq k}$. The paper [6] describes this change in the cost function in further detail and observe that Negative Sampling modifies the original training objective in equation (5.11) by introducing a different objective than original Skip-Gram.

Define the set of all item and context pairs (w, w_c) that occur in the text, denoted D, by

$$D := \{(w, w_c) \,|\, (w, w_c) \text{ is a possible pair from the pairing generators}\}.$$

Another way of formulating equation (5.16) is to introduce a random variable Z which indicates whether the pair (w, w_c) is the true output ($Z = 1$) or a sample from the noise distribution ($Z = 0$). In other words, we seek to find the optimal set of parameter $\theta = (\mathbf{W}, \mathbf{W}')$ such that the following joint probability is maximized

$$\arg\max_{\theta} \mathbb{P}(Z = 1 \,|\, w, w_c) \prod_{i=1}^{k} \mathbb{P}(Z = 0 \,|\, w, w_i).$$

which for all samples results in optimizing

$$\arg\max_{\theta} \prod_{(w,w_c)\in D} \left[\mathbb{P}(Z = 1 \,|\, w, w_c) \prod_{i=1}^{k} \mathbb{P}(Z = 0 \,|\, w, w_i) \right].$$

Optimizing this, is the same as optimizing the log, which results in

$$\arg\max_{\theta} \sum_{(w,w_c)\in D} \left[\log \mathbb{P}(Z = 1 \,|\, w, w_c) + \sum_{i=1}^{k} \log \mathbb{P}(Z = 0 \,|\, w, w_i) \right]$$

The negative samples $(w_i)_{1\leq i\leq k}$ are drawn according to a noise distribution P_n. Most commonly, this distribution utilizes the frequency of item i, denoted $f(i)$ and defined as

$$f(i) = \frac{\text{count}(i)}{\sum_{i=1}^{p} n_i},$$

where n_i denotes the length of sequence s_i, and count(i) denotes the number of times item i occurs across all sequences. Note that this is a distribution across i and thus could potentially be used as a noise distribution. The authors [15] suggest a modified version of this, defined as

$$\mathbb{P}_n(i) = \frac{f(i)^{3/4}}{\sum_{i=1}^{|\mathcal{I}|} f(i)^{3/4}}.$$

and thus sample the k negative samples according to $\mathbb{P}_n$. In this case, we utilize the observation from natural language processing that the occurrence frequency of words approximately follows a log-uniform distribution (see [19] for some description and history of the name), given by

$$f(i) \sim \exp(U(0, \log(|\mathcal{I}|))),$$

where $U(0, |\mathcal{I}|)$ denotes the uniform distribution on $0, |\mathcal{I}|$. We draw k negative samples using this noise distribution for the above equation (5.16). We did not test different choices of noise distribution although recent experimental evidence [2] suggests that other choices of noise distribution may be preferable for non-natural language processing tasks.

5.6 Stochastic Neighborhood Embeddings

The technique t-distributed Stochastic Neighborhood Embedding, introduced in [13] and commonly known as t-SNE, visualizes high-dimensional data in a 2 or 3-dimensional space, and has quickly become popular. Our aim for the present section is to attempt a strictly mathematical explanation of t-SNE whenever it is possible, with the goal of showing how the mathematical theory relates to the practical interpretations. Note that t-SNE is inherently stochastic (by the random initialization) and thus each visualization is different from the other. We use t-SNE to visualize the Skip-Gram vector representations and proceed to identify several treatment packages in the visualization.

Suppose we observe a collection of data points in a high-dimensional space

$$\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}, \quad \text{where } \mathbf{x}_i \in \mathbb{R}^m$$

for some large m (in the context of Section 5.2, m is the embedding dimension $\mathcal{E}$). We presume that points close to each other are similar. To visualize these high-dimensional points we seek corresponding points

$$\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\} \subseteq \mathbb{R}^v, v \in \{2, 3\}$$

such that

$$\mathbf{y}_i \text{ and } \mathbf{y}_j \text{ are similar} \quad \Longleftrightarrow \quad \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are similar.} \tag{5.17}$$

This is the fundamental problem statement for visualization techniques. The (chosen) visualization technique defines the similarity measure and how to compute/optimize it. To name a few methods, we mention Isomap [27], Stochastic Neighborhood Embedding [10], Locally Linear Embedding [21] and classical PCA [17]. A large review of visualization techniques is available in [12].

5.6.1 SNE

We start our explanation of t-SNE, by describing its predecessor, Stochastic Neighborhood Embedding (SNE). SNE is a (stochastic) optimization technique for finding high quality pairings $(\mathbf{y}_i)_{i\in\{1,\dots,n\}}$ such that the idea in equation (5.17) holds true. A metric is a (potential very rough) mathematical way of quantifying similarity between points. A very simple measure of similarity is the Euclidean norm, d_{EU}, given by

$$\begin{aligned} d_{\text{EU}}(\mathbf{x_1}, \mathbf{x_2}) &= \sqrt{\|\mathbf{x_1} - \mathbf{x_2}\|^2} \\ &:= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 \dots + (x_{1m} - x_{2m})^2}. \end{aligned} \tag{5.18}$$

To refine our measure of similarity, we define for $i, j \in \{1, 2, \dots n\}$

$$p_{j|i} := \begin{cases} \dfrac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}, & \text{if } j \neq i, \\ 0, & \text{if } j = i. \end{cases} \tag{5.19}$$

Here σ_i^2 is some parameter acting as the variance of the Gaussian distribution and it will be specified later. Contrary to d_{EU}, this similarity measure between i and j is no longer symmetric, i.e. $p_{j|i} \neq p_{i|j}$ in general. Note $p_{j|i}$ is simply a chosen quantification to measure the similarity between the point $\mathbf{x}_j$ with the given point $\mathbf{x}_i$, but we can assign the following probabilistic interpretation to it: For each i, let P_i denote the Gaussian distribution with center $\mathbf{x}_i$ and variance σ_i^2, i.e.

$$P_i := N_m(\mathbf{x}_i, \sigma_i^2 \mathbf{I}_m), \tag{5.20}$$

where $\mathbf{I}_m$ denotes the m-dimensional identity matrix, and N_m denotes the m-dimensional normal distribution. The value $p_{j|i}$ is proportional to

$$p_{j|i} \propto f_{N_m(\mathbf{x}_i, \sigma_i^2 I_m)}(\mathbf{x}_j), \tag{5.21}$$

where f denotes the density of $N_m(x_i, \sigma_i^2 I_m)$. The intuitive interpretation is that $p_{j|i}$ denotes the conditional probability that $\mathbf{x}_i$ would pick $\mathbf{x}_j$ as its neighbor under a Gaussian distribution centered at $\mathbf{x}_i$ with variance σ_i^2. Similarly, we define

$$q_{j|i} := \begin{cases} \dfrac{\exp(-\|\mathbf{y}_i - \mathbf{y}_j\|^2)}{\sum_{k \neq i} \exp(-\|\mathbf{y}_i - \mathbf{y}_k\|^2)}, & \text{if } j \neq i \\ 0, & \text{if } j = i \end{cases} \tag{5.22}$$

i.e. the conditional probability that $\mathbf{y}_i$ would pick $\mathbf{y}_j$ as its neighbor under a Gaussian distribution centered around $\mathbf{y}_i$ with variance $1/\sqrt{2}$. We shall denote this Gaussian distribution by Q_i. Note that we do not include a variance parameter σ in $q_{j|i}$, as we did for $p_{j|i}$. Stochastic Neighborhood Embedding is performing optimally if these two conditional distributions match. To measure the match, let P and Q denote probability distribution on some discrete space $\mathcal{X}$ (e.g. finite or countable space). The Kullback-Leibler divergence, D_{KL}, is defined as

$$D_{\mathrm{KL}}(P \parallel Q) := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}. \tag{5.23}$$

The cost function is defined as the Kullback-Leibler divergence over all data points

$$C = \sum_i D_{\mathrm{KL}}(P_i \parallel Q_i) = \sum_i \sum_{j \neq i} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}. \tag{5.24}$$

Gradient descent can be used to optimize (5.24) with respect to $\mathbf{y}_i$ and it can be shown that

$$\frac{\delta C}{\delta \mathbf{y}_i} = 2 \sum_j (\mathbf{y}_i - \mathbf{y}_j)(p_{i|j} - q_{i|j} + p_{j|i} - q_{j|i}). \tag{5.25}$$

Several modifications to the optimization procedure for GD are treated in Section 5.6.2. We will now describe the choice of the parameter σ. Let us first examine how σ interacts with the value $p_{j|i}$ when the distance changes. For fixed i, j, we may write

$$p_{j|i} = \frac{\exp(-d_{ij}/2\sigma^2)}{\sum_{k \neq i} \exp(-d_{ik}/2\sigma^2)}, \tag{5.26}$$

where $d_{ik} = \|\mathbf{x}_i - \mathbf{x}_k\|^2$ for $k \neq i$. If we let σ tend to infinity, then

$$p_{j|i} \xrightarrow[\sigma\to\infty]{} \frac{1}{p-1},$$

i.e. each point $\mathbf{x}_j$ is weighted uniformly and thus

$$(p_{j|i})_{j\neq i} \xrightarrow[\sigma\to\infty]{} \text{uniform}(\{1,\ldots,i-1,i+1,\ldots,p\}). \tag{5.27}$$

Thus, for large σ, we put approximately equal emphasis on every point $\mathbf{x}_k$, regardless of their distances d_{ik} to $\mathbf{x}_i$. Observe that

$$p_{j|i} = \frac{\exp(-d_{ij}/2\sigma^2)}{\sum_{k\neq i}\exp(-d_{ik}/2\sigma^2)} = \frac{1}{1+\sum_{k\neq\{i,j\}}\exp((d_{ij}-d_{ik})/2\sigma^2)},$$

where we assume that $d_{ik} > 0$ for all $k \neq i$ and d_{ik} are all different. Let $d = \min(\{d_{i1},\ldots,d_{i(i-1)},d_{i(i+1)},\ldots,d_{ip}\})$. If $d_{ij} = d$, then

$$(d_{ij} - d_{ik}) < 0 \quad \forall k \neq \{i,j\} \tag{5.28}$$

and hence

$$p_{j|i} \xrightarrow[\sigma\to 0]{} \frac{1}{1+0} = 1. \tag{5.29}$$

If $d_{ij} \neq d$, then

$$\exists k \in \{1,2,\ldots,p\} \setminus \{i,j\} : d_{ik} < d_{ij} \tag{5.30}$$

and hence the term in the denominator

$$\sum_{k\neq\{i,j\}} \exp((d_{ij}-d_{ik})/2\sigma^2)$$

converges to infinity as σ tends to zero. Consequently,

$$(p_{j|i})_{j\neq i} \xrightarrow[\sigma\to 0]{\mathscr{D}} \delta_{\min(\{d_{i1},\ldots,d_{i(i-1)},d_{i(i+1)},\ldots,d_{ip}\})},$$

where δ_a denotes the Dirac measure in a. The heuristic interpretation of this is that for very small σ, the probability mass is (nearly) concentrated in a single point – the nearest point, provided the data points are different.

We now analyze the case where σ takes some value between 0 and ∞, i.e. cases between complete certainty and uninformed random guessing. We assume that there exists a subset K of $\{1,2,\ldots,p\} \setminus \{i\}$ and some boundary ϵ such that

$$\forall\, k \in K : d_{ik}/2\sigma^2 \in [0,\epsilon] \tag{5.31}$$

$$\forall\, k \notin K : d_{ik}/2\sigma^2 \gg \epsilon. \tag{5.32}$$

This is implies that $\exp(-d_{ik}/2\sigma^2)$ only "matters" in the value of $p_{j|i}$ if $k \in K$. Importantly, note how σ directly affects this boundary, e.g. we could formulate this equivalently as

$$\forall\, k \in K : d_{ik} \in [0, 2\sigma^2]$$

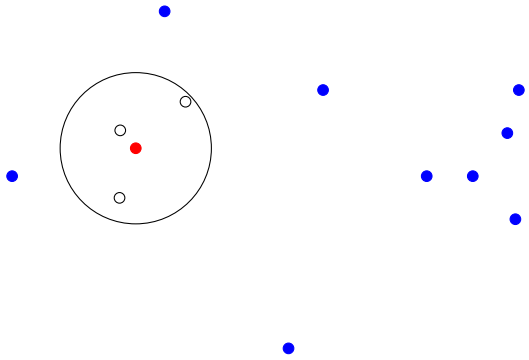


Figure 5.3: Example of a separating neighborhood around the red point.

and

$$\forall\, k \notin K : d_{ik} \gg 2\sigma^2,$$

where we for simplicity of notation assume that $\epsilon = 1$. In practical terms, we draw a suitable sized circle (with radius σ) around $\mathbf{x}_i$ such that the circle separates the space to nearby points and others, as visualized in Figure 5.3. It is evident that other choices of σ may also lead to other reasonable separations – thus a qualified estimation of σ is critical. Inserting this into equation (5.19), if $j \in K$ then

$$\frac{\exp(-d_{ij}/2\sigma^2)}{\sum_{k\neq i}\exp(-d_{ik}/2\sigma^2)} \approx \frac{\exp(-d_{ij}/2\sigma^2)}{\sum_{k\in K}\exp(-d_{ik}/2\sigma^2)}$$

and similarly if $j \notin K$,

$$\frac{\exp(-d_{ij}/2\sigma^2)}{\sum_{k\neq i}\exp(-d_{ik}/2\sigma^2)} \approx 0$$

which results in

$$p_{j|i} \approx \begin{cases} \dfrac{\exp(-d_{ij}/2\sigma^2)}{\sum_{k\in K}\exp(-d_{ik}/2\sigma^2)}, & \text{if } j \in K \\ 0, & \text{if } j \notin K. \end{cases}$$

Hopefully, at this point, it is clear that σ directly affects the set K, e.g. the set of nearby neighbors. Thus a good choice of σ can be found by estimating the set of effective neighbors.

To estimate the number of effective neighbors, we use the *perplexity* of a (discrete) distribution P on a discrete space $\mathcal{X}$, defined as

$$\text{Perp}(P) := e^{H(P)} = e^{-\sum_{x\in\mathcal{X}} p(x)\ln p(x)} = \prod_x \frac{1}{p(x)^{p(x)}}, \tag{5.33}$$

where $H(P)$ denotes the *entropy* of a probability distribution, defined as

$$H(P) := -\sum_{x\in\mathcal{X}} p(x)\ln p(x), \tag{5.34}$$

where ln denotes the natural logarithm[1]. We will try to explain entropy conceptually but add a bit of math to draw a few interesting parallels. The entropy is a measure of how much information that is contained in your probability distribution concerning a future value. A high value corresponds to little information and a low value corresponds a lot of information. It can be shown that the entropy is the only measure of information which satisfies certain desirable criteria, see [24]. Our two limiting distributions for σ, uniform and δ, maximizes, respectively minimizes, entropy, e.g.

$$H(\text{uniform}) = \ln(p-1), \quad \text{and} \quad H(\delta) = 0,$$

with the convention that $0 \cdot \log 0 = 0$. Moreover, since the uniform distribution on the set $\{1, 2, \ldots, U\}$ maximizes the entropy with the value $\ln U$, we provide the following alternative conceptual understanding of entropy. Let the entropy of a distribution P (or a random variable with a distribution P) be $H(P)$ and asumme for simplicity that $H(P)$ is integer-valued. We may find $U > 0$, such that the uniform distribution on $\{1, \ldots, U\}$ satisfies that

$$H(\text{uniform}) = H(P) = \ln(U).$$

Provided that $H(P)$ is a measure of the information provided by a random sample from the distribution P, we observe that $H(P)$ corresponds to the information provided by throwing a $\ln(K)$-sided dice. This suggests interpreting the entropy as estimating log of the effective number of classes. Using this interpretation, the perplexity estimates the effective number of distinct values as defined in equation (5.33). Finally, for SNE this implies that perplexity may be interpreted as an estimate of the effective number of neighbors.

The SNE algorithm is run with a user-defined perplexity $\mathscr{P}$, and for each i, σ_i is found such that

$$\text{Perp}(P_i) = \mathscr{P}.$$

In the practical implementation, σ_i is merely found once $|\mathscr{P} - \text{Perp}(P_i)|$ is below a very small threshold (for example $1e$-5 in scikit-learn implementation, [18]).

5.6.2 Training and optimizing SNE

An unfortunate attribute of unmodified gradient descent is a tendency to become stuck in saddle points (also known as local minima). To negate this tendency, the authors "anneal noise" to gradient descent, which is described in some detail in [13]. A search through the literature did not reveal an exact formulation/procedure for "annealing noise" but the following rough summarization may be helpful.

Annealing noise: Suitable Gaussian noise is added to the gradient during training. The added Gaussian noise is controlled by a rate which is large at the beginning of optimization and then decays to zero.

[1] The perplexity and entropy can be formulated with any logarithmic base number b. We use the natural logarithm due to equation (5.19). Note that the perplexity is invariant of the logarithmic base number.

However the procedure of annealing noise was not used in t-SNE, and they state the following caveat:

> "... in the early stage of the optimization, Gaussian noise is added to the map points after each iteration. Gradually reducing the variance of this noise performs a type of simulated annealing that helps the optimization to escape from poor local minima in the cost function. If the variance of the noise changes very slowly at the critical point at which the global structure of the map starts to form, SNE tends to find maps with better global organization. Unfortunately, this requires sensible choice of the initial amount of Gaussian noise and the rate at which it decays. Moreover, these choice interact with the amount of momentum and the step size that are employed in the gradient descent. It is therefore common to run the optimization several times on a dataset to find appropriate values for the parameters. In this respect, SNE is inferior to ..."
>
> — [13] on page 2583

To initialize the optimization procedure of equation (5.25), the initialization for $\mathcal{Y}_0$ is drawn from an isotropic Gaussian distribution with small variance centered around the origin, e.g. an n-dimensional Gaussian distribution N_n with mean zero and $\mathrm{Var}(N_n) = \sigma^2 I_n$ for some small $\sigma > 0$, e.g.

$$\mathcal{Y}^{(0)} \sim N_n.$$

These initialized values are updated using momentum according to the following scheme

$$\mathcal{Y}^{(t)} := \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right), \tag{5.35}$$

where η denotes the learning rate, $\mathcal{Y}^{(t)}$ denotes the solution at iteration t, and $\alpha(t)$ denotes the momentum at iteration t. Momentum modifies the weight update by allowing it to depend on its most recent values, effectively accumulating a "momentum". The advantage of this procedure appears to be faster convergence and better avoidance of local minima (due to the momentum). We refer to [26] for a discussion on this.

5.6.3 t-SNE

The t-SNE method was motivated by certain deficiencies of SNE, namely difficulty of optimization, the "crowding problem" and "outlier problem" as described in Section 3.1-3.2 of [13].

We first present the changes for P, the distribution on the high-dimensional datapoints, and subsequently present the outlier problem attached to this. In t-SNE, instead of $p_{j|i}$ which is non-symmetrical, we symmetrize it by

$$p_{ij} := \frac{p_{j|i} + p_{i|j}}{2n}, \tag{5.36}$$

where $p_{j|i}$ originates from equation (5.19) in SNE. Unlike $p_{j|i}$ in equation (5.19), this is no longer a probability distribution over the index j, but instead

$$\begin{aligned}\sum_{k\neq l} p_{kl} &\coloneqq \sum_k \sum_{l:\, l\neq k} p_{kl} = \sum_k \sum_{l\neq k} \frac{p_{k|l} + p_{l|k}}{2n} \\ &= \frac{1}{2n}\Big(\sum_k \sum_{l:\, l\neq k} p_{k|l} + \sum_k \sum_{l:\, l\neq k} p_{l|k}\Big) = \frac{1}{2n}\Big(\sum_l \sum_{k:\, k\neq l} p_{k|l} + \sum_k 1\Big) = \frac{n+n}{2n} = 1,\end{aligned}$$

i.e. a probability distribution over all pairs of distinct indexes. The definition in equation (5.36) is not the natural generalization to a probability distribution over all pairs of distinct indexes, which would be

$$p_{ij} \coloneqq \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2)}{\sum_{k\neq l}(-\|\mathbf{x}_k - \mathbf{x}_l\|^2)}, \quad i \neq j. \tag{5.37}$$

The problem with this definition, compared to (5.36), is outliers, i.e. if $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ is large for all j. To illustrate this, we similarly define

$$q_{ij} = \frac{\exp(-\|\mathbf{y}_i - \mathbf{y}_j\|^2)}{\sum_{k\neq l}\exp(-\|\mathbf{y}_k - \mathbf{y}_l\|^2)},$$

and set the cost function as the Kullback-Leibler divergence between P and Q over all distinct indices, i.e.

$$C_{\text{sym}} = D_{\text{KL}}(P \parallel Q) = \sum_{k\neq l} p_{kl} \log \frac{p_{kl}}{q_{kl}} = \sum_{k\neq l} p_{kl} \log p_{kl} - p_{kl} \log q_{kl}, \tag{5.38}$$

This choice for definition of q and p corresponds to the method called symmetric SNE, and is indicated by subscript sym in the above cost-function. For symmetric SNE (with p_{ij} as equation (5.37)), the gradient of the cost function turns out to be

$$\frac{\delta C_{\text{sym}}}{\delta \mathbf{y}_i} = 4 \sum_{j\neq i} (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j), \tag{5.39}$$

using calculations very similar to Appendix A in [13]. For an outlier $\mathbf{x}_i$, p_{ij} is very small for all j and hence the gradient wrt. $\mathbf{y}_i$ will approximately be

$$\frac{\delta C_{\text{sym}}}{\delta \mathbf{y}_i} = 4 \sum_{j\neq i} -q_{ij}(\mathbf{y}_i - \mathbf{y}_j).$$

Observe that this is unaffected by the high-dimensional distribution P. Note that with p_{ij} defined as in equation (5.36), the gradient is unaffected by the change in q_{ij}, and thus yields the same gradient as C_{sym} of its respective cost function

$$\frac{\delta C}{\delta \mathbf{y}_i} = 4 \sum_{j\neq i} (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j)$$

The term $\sum_j p_{i|j}/2n = 1/2n$ in (5.36) guarantees that $\sum_j p_{ij} > \frac{1}{2n}$ and consequently that $\mathbf{y}_i$ always makes a substantial contribution to the gradient in equation (5.35).

The major change from symmetric SNE to t-SNE is in the low-dimensional distribution q which in symmetric SNE is given as

$$q_{ij} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_k \sum_{l \neq k} \exp(-\|y_k - y_l\|^2)},$$

but in t-SNE is given as

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|y_k - y_l\|^2)^{-1}}. \tag{5.40}$$

Observe that this is proportional to the Students t-distribution with 1 degree of freedom, e.g.

$$q_{ij} \propto f_{t(1)}(d_{ij}).$$

This results in the gradient of the cost function for t-SNE being given by

$$\frac{\delta C}{\delta \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}.$$

Crowding problem

The main reason to change q_{ij} is the so-called "crowding problem" which comes from empirical observations.

> "... A related problem is the very different distribution of pairwise distances in the two spaces(red. high- and low-dimensional spaces). The volumne of a sphere centered on datapoint i scales as r^m, where r is the radius and m is the dimensionality of the sphere. So if datapoints are approximately uniformly distributed in the region around i on the ten-dimensional manifold (red. in a higher dimensional space), and we try to model the distances from i to the other datapoints in the two-dimensional map, we get the following 'crowding problem': the area of the two-dimensional map that is available to accommodate moderately distant datapoints will not be nearly large enough compared with the area available to accommodate nearby datapoints. Hence, if we want to model the small distances accurately in the map, most of the points that are at a moderate distance from datapoint i will have to be placed much too far away in the two-dimensional map. In SNE, the spring connecting datapoint i to each of these too-distant map points will thus exert a very small attractive force. Although these attractive forces are very small, the very large number of such forces crushes together the points in the center of the map, which specific to SNE, but that it also occurs in other local techniques for multidimensional scaling such as Sammon mapping."
>
> — [13] on page 2585

We give the following interpretation of the crowding problem. The gist of the statement is that the dimensionality differences between $\mathbf{x}$ and $\mathbf{y}$ causes a problem for dimensionality reduction techniques (such as SNE and t-SNE) to properly reflect moderately distanced points and distinguish between clusters

in the low-dimensional representation, as the volume area of moderately distanced points in the lower dimensional space is much smaller the high dimensional space. t-SNE tackles this problem by imposing a heavier tail distribution on the low-dimensional points. The heavier tail distribution decreases the (proportional) weight of moderately distanced points by giving more weight to distant points and similar weight to nearby points – as compared to SNE. Empirical verification of this property is also provided in [13].

5.7 Classifiers and the kernel method

We utilize three different classifiers in the model evaluation step in Chapter 6 to classify the data source of each item/event. In this section, we give a short description of the methodology for each classifier. As with any classifiers, we have a dataset of pairs $(x_i, y_i)_{i\in 1,\dots,|\mathcal{I}|}$[2], where x_i denotes the input vector and y_i the true class label for sample i. The goal is to find a function $C\colon \mathbb{R}^D \to \{1,\dots,\mathcal{I}|\}$, which utilizes some parameter set θ, such that

$$\arg\max_{\theta} \sum_{i=1}^{|\mathcal{I}|} \log \mathbb{P}(y_i \mid x_i; \theta), \tag{5.41}$$

is maximized. Many different methods of estimating the above probability can be used and in this section we aim to describe three of them, namely Support Vector classifier (SVC), Random Forest classifier (RFC) and k-Nearest Neighbor classifier (k-NN classifier). Near the end of this section, we describe the Rand index, a method for comparing partitions (and hence classifiers) which may have different class size.

5.7.1 Random Forest Classifier

Random forest classifier originates from decision trees and consists of "bagging"(or bootstrap aggregating) many decision trees together into a "forest" and thus creating a classifier with smaller variance than the individual decision trees. A random forest has a few major components, namely

1. Resampling strategy to resample from the data (the "random" of random forest, typically bootstrap).
2. Splitting strategy for growing each decision tree on a resampled dataset (the same for each tree).
3. Combination strategy of the resulting decision tress into one classifier.

The algorithm is straightforward (once the inner parts of resampling and splitting strategy is understood), i.e. we fix a forest size B (note that B conveniently matches bootstrap *and* bagging).

[2] We may in general have K observations and may want to train a classifier on these. However we stick to $|\mathcal{I}|$, since it fits our application in Chapter 6.

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $(x_i^b, y_i^b)_{i\in 1,\dots,|\mathcal{I}|}$ with replacement from the empirical distribution $(x_i, y_i)_{i\in 1,\dots,|\mathcal{I}|}$.
   (b) Grow a random-forest tree T_b, by recursively applying the following steps as part of the splitting strategy of a decision tree until the minimum node size $n_m in$ is achieved.
      i. Select m variables at random from the D input variables in $(x) = (x_1, x_2, \dots, x_D)$, where x denotes a generic input vector.
      ii. Pick the best split according to the splitting strategy (see equation (5.45)).
      iii. Split the tree into two branches according to the above best split.

Furthermore, random forests utilizes the concept of out-of-bag classification, i.e. for an observation (x_i, y_i) the predicted class of the sample is computed by averaging over the trees in which the observation *did not* occur.

We discuss "random-forest trees" in the next subsection. This explanation is heavily inspired by [9].

Classification trees and random-forest trees

A decision tree is illustrated in Figure 5.4. The tree in Figure 5.4 illustrates a partition of the input space into regions, for example the region

$$R = \{(x_1, x_2, x_3) \,|\, x_3 > 0.5, x_1 < 6, x_2 < 2.7\}, \tag{5.42}$$

is obtained by following the left-most branches until the end-leaf. In this way, we obtain a partition of the observations. The goal of classification trees is to obtain a "pure" of the data, such that each end-leaf only (or mostly) contains only one class (hence it is "pure").

We first need some notation to explain classification trees. A classification tree is a partition of the input space into disjoint regions according to simple boolean rules on the input variables. In other words, the classifier of a decision tree is given by

$$C(x) = \sum_{j=1}^{m} \mathbb{1}_{\{x\in R_j\}} \arg\max_k p_{j,k}, \tag{5.43}$$

where $p_{j,k}$ denotes the probability of the class k in the region R_j, and $(R_j)_{1\leq j\leq m}$ denotes the partition regions/areas. This is estimated by counting the frequency of class k among the observations in region R_j. The optimal tree is found by optimizing across all possible boolean partitions/splits of the dataset, i.e.

$$\arg\min_{m\in\mathbb{N}, R_1,\dots,R_m} \sum_{i=1}^{|\mathcal{I}|} \mathcal{L}(y_i, f(x_i)),$$

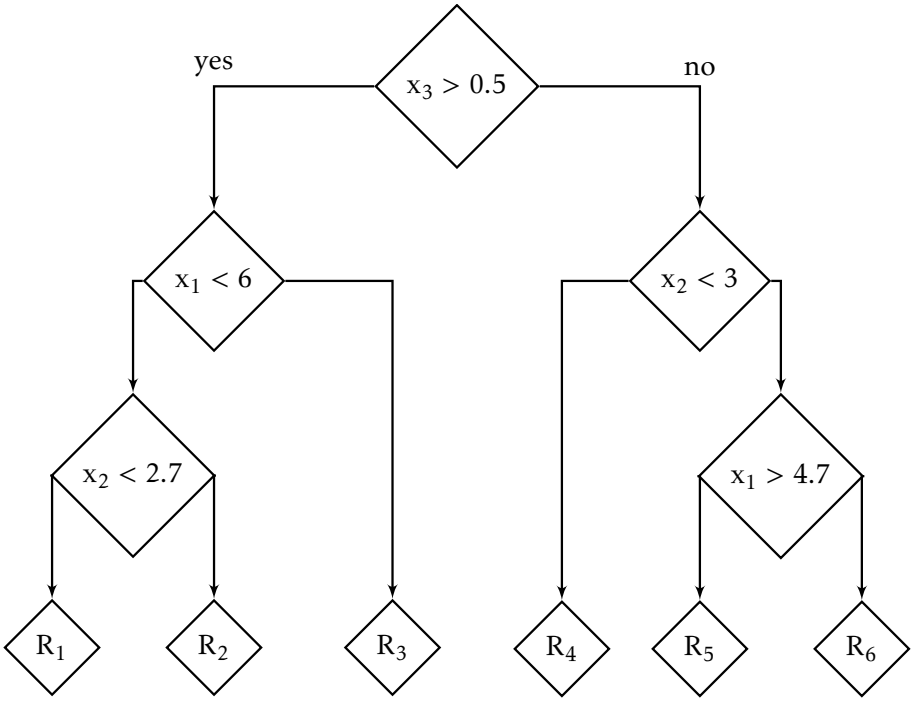


Figure 5.4: An illustration of a decision tree.

where $\mathcal{L}$ denotes a function measuring the impurity/error between y_i, the estimated class $f(x_i)$ (typically a loss or impurity function) and the regions $(R_i)_{1\leq i\leq m}$ are of the form in equation (5.44)[3]. Note that this problem is computationally infeasible as each choice may be permuted. Instead, we pursue a greedy algorithm, in terms of maximizing the split at each step.

As our next step, we elaborate on the splitting strategy of the tree and later return to the pruning strategy of merging nodes in the tree. Define the regions $R_1(j,s)$ and $R_2(j,s)$ by

$$R_1(j,s) = \{X \mid X_j \leq s\} \quad \text{and} \quad R_2(j,s) = \{X \mid X_j \leq s\}. \tag{5.44}$$

Let $\text{imp}(t)$ denote the impurity of node t (representing a region R_t) in the tree (Gini index of heterogeneity or the cross-entropy are common choices) calculated on the estimated probability of the node, namely

$$\hat{p}_{t,k} = \frac{1}{N_t} \sum_{x_i \in R_\text{t}} \mathbb{1}_{\{y_i = k\}}, \quad k \in 1,\ldots,K$$

where $N_t = |\{x_i \in R_t\}|$ and K denotes the number of classes. Note that since we are still growing the tree, the region R_t and m are merely intermediate parameters for the currently grown tree (we are not necessarily finished with

[3]We suppress the constraint that the minimum node size hyper-parameter entails on m. The goal is simply to emphasize that this computation is unfeasible.

the splitting procedure yet). Thus they do not correspond to the final regions in equation (5.43). The goal is to find a pair (j, s) such that the changes in impurity is maximized, e.g.

$$\Delta\text{imp}((j,s),t) = \text{imp}(t) - p_{\text{t}_{\text{left}}} \cdot \text{imp}(\text{t}_{\text{left}}) - p_{t_{right}} \cdot \text{imp}(\text{t}_{\text{right}}) \tag{5.45}$$

where $p_{\text{t}_{\text{left}}}$ and $p_{\text{t}_{\text{right}}}$ denotes the probability of the left node t_{left}, respectively right node t_{right}, created the split pair (j, s). This procedure is then repeated for each resulting node until a stopping criterion is achieved (e.g. the node is pure or the number of samples in the region R_{t} is below a set threshold, or maximal tree depth, or the improvement in impurity is below a specified minimum).

Pruning the decision tree

Following the completion of the full tree, it is common to "prune" the resulting tree, i.e. collapse some of its internal (non-terminal) nodes. Consider a tree T with terminal nodes/end-leafs numerated $1, \ldots, m$ in accordance with the final regions $R_1, \ldots, R_m$ and let, as previously

$$\begin{aligned} N_t &= |\{x_i \in R_t\}|, \\ \hat{p}_{t,k} &= \frac{1}{N_t} \sum_{x_i \in R_t} \mathbb{1}_{\{y_i = k\}}, \quad k = 1, \ldots, K. \end{aligned}$$

Denote the node impurity for terminal node/end-lead t in the tree T by

$$Q_t(T) = \sum_{k=1}^{K} \hat{p}_{t,k}(1 - \hat{p}_{t,k}).$$

During the "pruning" of the tree, the goal is to optimize the cost-complexity function

$$C_\eta = \sum_{t=1}^{m} N_t Q_t(T) + \eta m,$$

where η is a tuning parameter. In other words, we seek to optimize the node impurity while penalizing trees with a large number of regions m. In general, we seek a smaller tree $T_\eta \subseteq T_{\text{orig}}$, where T_{orig} denotes the original tree obtained from the splitting procedure.

Random-forest tree

The random-forest tree merely differs by adding the additional step. Instead of splitting based on all variables, we instead sample m of the D input variables in $(x_i) = (x_{1,i}, x_{2,i}, \ldots, x_{D,i})$ and compute the splitting costs for these according to the equation (5.45).

Resampling strategy

The resampling strategy used is commonly bootstrap, but other sampling strategies, such as subsampling [4] can be used.

5.7.2 k-nearest neighbor Classifier

The k-nearest neighbor classifier is refreshingly simple. For all data points $(x_i, y_i)_{i\in 1,\dots,|\mathcal{I}|}$ we compute the (Euclidean) distance matrix D of dimension $|\mathcal{I}| \times |\mathcal{I}|$ between all possible pairs (x_i, x_i'). We then classify a datapoint x_i according to the majority vote of its k nearest neighbors $x_{(1)},\dots,x_{(k)}$ (the k datapoints which have smallest distance to x_i), e.g.

$$\arg\max_c \sum_{j=1}^{k} \mathbb{1}_{\{y_{(j)}==c\}},$$

where $y_{(j)}$ denotes the class of $x_{(j)}$. Clearly, this yields a (explicitly) local classifier, but through a method which depends on the distances between (x_i) and not the actual inputs of x_i – contrary to Random Forest Classifier from Subsection 5.7.1. We mention that several optimization and modification schemes can be employed to the k-nearest neighbor method but we did utilize these.

5.7.3 Support Vector Classifier (SVC)

In this section, we briefly introduce Support Vector Classifier (SVC) and describe roughly how it works without going into the deeper mathematical details – for this we refer to [9] with their references and the original work [3]. The overall goal is to find the optimal separating hyperplane for a two-class system (we later describe how this can be extended to multiple classes) and we assume that such perfect (linear) separation of the two classes is available. To moderate this procedure, "slack" variables are introduced which allow exceptions from perfect separation and produce what is instead called *soft margins*. The method the use of slack variables to constrain their influence on the obtained hyperplanes. The optimal support vector classifier can be found using classical convex optimization techniques.

The multi-class case of SVC is commonly solved by using several binary classifiers. We use the standard implementation of SVC from scikit-learn [18] which uses a *one-versus-one* strategy for this. The one-versus-one strategy builds a classifier for each possible pair of classes – with n classes, we obtain $n(n-1)/2$ classifiers. Given an observation $\mathbf{x}$, each binary classifier $\hat{G}_{ij}(x)$ then predicts either class i or class j. For each class, we count the number of binary support vector classifiers which places x into it, i.e. the binary classifiers "votes" for its predicted class. Finally, we define our one-vs-one classifier of $\mathbf{x}$ by the class which received the most votes from our binary classifiers.

5.7.4 Kernel method

The kernel method is a common tool used to allow some statistical learning techniques to fit nonlinear and high-dimensional manifolds. We utilize it on the vector representations from Skip-Gram to see whether their (kernel) transformation can be separated.

In this description of the kernel method, we rely heavily on [23] and Chapter 14 in [16]. The kernel trick is a general principle which may be applied to algorithm which can be formulated through inner products $\langle \mathbf{x}, \mathbf{x}' \rangle$. An example of this is k-nearest neighbor, where we observe

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \langle \mathbf{x}_i, \mathbf{x}_i \rangle + \langle \mathbf{x}_j, \mathbf{x}_j \rangle - 2\langle \mathbf{x}_i, \mathbf{x}_j \rangle. \tag{5.46}$$

Formally, we define a kernel function as a mapping $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with no further restrictions. The kernel is often symmetric, i.e. $\kappa(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x}', \mathbf{x})$, and non-negative, i.e. $\kappa(\mathbf{x}, \mathbf{x}') \geq 0$.

The kernel trick

In this subsection, we describe the kernel trick. Let the observations $[\mathbf{x}_1, \ldots, \mathbf{x}_N]$ be from some space $\mathcal{X}$. Assume that there exists some mapping $\phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is an inner product space. Define the kernel κ by

$$\kappa(\mathbf{x}, \mathbf{x}') := \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle. \tag{5.47}$$

By applying Mercer's theorem we may choose the kernel κ (provided that it satisfies certain assumptions) because it induces a mapping ϕ. Mercer's theorem states that there exists a function $\phi : \mathcal{X} \to \mathbb{R}$ so that

$$\kappa(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{V}},$$

provided that κ is a symmetric and positive definite. The key step is then to replace each occurrence of the inner product $\langle \mathbf{x}, \mathbf{x}' \rangle$ with $\kappa(\mathbf{x}, \mathbf{x}')$ in the formulation of the algorithm/fit function (such as in equation (5.46)). The trick here is that we may choose our kernel freely, as long as it satisfies the conditions of Mercer's theorem. Thus $\mathcal{H}$ is potentially infinitely dimensional inner product space and we do not need the exact form of ϕ due to Mercer's theorem which provides the existence of ϕ given κ. Note that not the kernel trick does not apply to all methods and algorithms – an exception is k-means, but the kernel trick applies to the related k-medoids.

Examples of kernel functions satisfying the conditions of Mercer's theorem are

1. Radial basis function (RBF), defined as

   $$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2} \right),$$

   where σ^2 is a positive constant.

2. Linear kernel, i.e.

   $$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}' = \sum_i x_i x_i' \tag{5.48}$$

3. Homogeneous polynomial kernels, defined as

   $$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}')^d,$$

   where $d \in \mathbb{N}$ and T denotes transpose (see Chapter 5 of [25]).

Many examples of kernels are available in Chapter 14 of [16].

5.7.5 Rand index

In this subsection, we describe the Rand index and adjusted Rand index, where the former was introduced in [20] and the latter in [11]. The purpose of the Rand index is to compare two partitions (or classifiers/clustering algorithms) with each other, and we use the Rand Index in Chapter 6 to compare a clustering algorithm with the true labels.

We lean heavily on the explanations from [16]. Consider two partitions $U = \{u_1, \dots, u_R\}$ and $V = \{v_1, \dots, v_C\}$ of N data points $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$. Define the following,

1. TP (True positives) is number of data points pairs $(\mathbf{x}_i, \mathbf{x}_j)$, $i \neq j$, for which it holds that

$$\exists r \in \{1, \dots, R\} \colon x_i, x_j \in u_r \quad \text{and} \quad \exists c \in \{1, \dots, C\} \colon x_i, x_j \in v_c,$$

   i.e. the pairs for which both U and V place them in the same cluster.

2. TN (True negatives) is the number of data point pairs $(\mathbf{x}_i, \mathbf{x}_j)$, $i \neq j$, for which it holds that

$$\nexists r \in \{1, \dots, R\} \colon x_i, x_j \in u_r \quad \text{and} \quad \nexists c \in \{1, \dots, C\} \colon x_i, x_j \in v_c,$$

   i.e. the pairs for which both U and V place them in different clusters.

3. FN (False negatives) is the number of data point pairs $(\mathbf{x}_i, \mathbf{x}_j)$, $i \neq j$,, for which it holds that

$$\nexists r \in \{1, \dots, R\} \colon x_i, x_j \in u_r \quad \text{and} \quad \exists c \in \{1, \dots, C\} \colon x_i, x_j \in v_c,$$

   i.e. the pairs for which they are in different clusters in U but in the same cluster in V.

4. FP (False positives) is the number of data point pairs $(\mathbf{x}_i, \mathbf{x}_j)$, $i \neq j$,, for which it holds that

$$\exists r \in \{1, \dots, R\} \colon x_i, x_j \in u_r \quad \text{and} \quad \nexists c \in \{1, \dots, C\} \colon x_i, x_j \in v_c,$$

Finally, the Rand index R is defined by

$$\mathrm{R} \coloneqq \frac{TP + TN}{TP + FP + FN + TN}. \tag{5.49}$$

The lower bound for the Rand index is 0 and the upper bound is 1 – the former indicating zero overlap between two partitions and the latter indicating perfect overlap.

The adjusted Rand index is best explained given a contingency table as visualized in Table 5.1. The values n_{rc} in the table denotes the number of data

		Partition V				
	Class	v_1	v_2	$\cdots$	v_C	Sums
Partition U	u_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1\cdot}$
	u_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	u_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	$n_{R\cdot}$
	Sums	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot C}$	$n_{\cdot\cdot} = n$

Table 5.1: A standard contingency table for two partitions U and V. Table taken from [11].

points $\mathbf{x}$ which are clustered into u_r and v_c. The adjusted Rand index (ARI) is given by

$$\text{ARI} := \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{n_{i\cdot}}{2}\sum j\binom{n_{\cdot j}}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{n_{i\cdot}}{2} + \sum j\binom{n_{\cdot j}}{2}] - [\sum_i \binom{n_{i\cdot}}{2}\sum j\binom{n_{\cdot j}}{2}]/\binom{n}{2}},$$

where the undefined notation can be found in Table 5.1. The adjusted rand index takes values in $[-1, 1]$, with 0 corresponding to expected value of a hyper-geometric distribution given the clusters sizes , see [11].

An advantage of the Rand index is that it allows comparison between differently sized partitions. This is useful if given a reference class label and attempting to find a partition (or classifier) such that clusters are pure (in the sense that they consist of a single class) – and which may require more clusters than reference classes.

References

[1] Christopher M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] H. Caselles-Dupré, Florian Lesaint and Jimena Royo-Letelier. "Word2vec applied to recommendation: Hyperparameters matter". *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM. 2018, 352–356.

[3] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". *Machine learning* 20.3 (1995), 273–297.

[4] Roxane Duroux and Erwan Scornet. "Impact of subsampling and pruning on random forests". *arXiv:1603.04261* (2016).

[5] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, Yun Sing Koh and Rincy Thomas. "A survey of sequential pattern mining". *Data Science and Pattern Recognition* 1.1 (2017), 54–77.

[6] Yoav Goldberg and Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method" (2014). arXiv: 1402.3722.

[7] Michael Gutmann and Aapo Hyvärinen. "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models". *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010, 297–304.

[8] Zellig S Harris. "Distributional structure". *Word* 10.2-3 (1954), 146–162.

[9] Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

[10] Geoffrey E Hinton and Sam T Roweis. "Stochastic neighbor embedding". *Advances in neural information processing systems*. 2003, 857–864.

[11] Lawrence Hubert and Phipps Arabie. "Comparing partitions". *Journal of classification* 2.1 (1985), 193–218.

[12] John A Lee and Michel Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[13] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". *Journal of Machine Learning Research* (2008), 2579–2605.

[14] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. "Efficient estimation of word representations in vector space" (2013). arXiv: 1301.3781.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. "Distributed representations of words and phrases and their compositionality". *Advances in Neural Information Processing Systems*. 2013, 3111–3119.

[16] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[17] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), 559–572.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[19] David MW Powers. "Applications and explanations of Zipf's law". *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*. Association for Computational Linguistics. 1998, 151–160.

[20] William M Rand. "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical association* 66.336 (1971), 846–850.

[21] Sam T Roweis and Lawrence K Saul. "Nonlinear dimensionality reduction by locally linear embedding". *Science* 290.5500 (2000), 2323–2326.

[22] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. "Learning representations by back-propagating errors". *Nature* 323.6088 (1986), 533.

[23] Bernhard Schölkopf and Alex Smola. "Support vector machines and kernel algorithms". *Encyclopedia of Biostatistics* 8 (2002), 5328–5335.

[24] Claude Elwood Shannon. "A mathematical theory of communication". *Bell system technical journal* 27.3 (1948), 379–423.

[25] Alex J Smola, Zoltan L Ovari and Robert C Williamson. "Regularization with dot-product kernels". *Advances in neural information processing systems*. 2001, 308–314.

[26] Ilya Sutskever, James Martens, George Dahl and Geoffrey Hinton. "On the importance of initialization and momentum in deep learning". *International conference on machine learning*. 2013, 1139–1147.

[27] Joshua B Tenenbaum, Vin De Silva and John C Langford. "A global geometric framework for nonlinear dimensionality reduction". *Science* 290.5500 (2000), 2319–2323.

Chapter 6

Integration of sequential EHR information using mathematical representations

Thorbjørn Grønbæk[1,2], Lars N. Andersen[1], Kim Mouridsen[2]
[1]Department of Mathematics, Aarhus University
[2]Center for Functionally Integrative Neuroscience and MINDLab,
Department of Clinical Medicine, Aarhus University [1]

Working paper.

Keywords: electronic health records, unsupervised feature learning, preprocessing, data integration, neural networks

Abstract

Background and objective: Electronic health records store information about encounters and events between health care systems and patients, and represent a rapidly growing source for medical data analysis. A problematic attribute of most health care data is heterogeneity and the curse of dimensionality from transforming qualitative observations. In this study we aim to transform the encounters and events to computationally efficient mathematical representations by utilizing electronic health records displayment treatment trajectories.
Methods: Mathematical representation are obtained with Skip-Gram from natural language processing and probabilistic distributions from Markov Chains. We compare their encoding quality by inserting these representation and distribution estimates as inputs in unsupervised learning tasks. The clustering of central input groups are compared.
Results: We obtain computationally efficient representations which preserves the relational structure between encounters and reduce dimensionality significantly. We show Skip-Gram outperforms Markov Chains as inputs in un-

[1]Email adresses: {larsa, thorbjoern}@math.au.dk, kim@cfin.au.dk

supervised clustering techniques and subsequent identification of treatment packages.
Conclusion and perspective: The embedding of EHR using Skip-Gram yields an effective and intriguing representation of events. This method reduces dimensionality by clustering treatment packages in a mathematical representation suitable for other data analysis problems, effectively homogenizing the dataset. Using this to merge identified treatment packages into a single EHR entry could decrease data heterogeneity significantly.

6.1 Introduction

A fundamental problem for the successful incorporation of EHR data into clinical workflows is data heterogeneity. This occurs across multiple fronts: different data sources with differing standards, duplicate entries, incorrect registration but also numerical measurements and free-text observations (the latter two often called structured, respectively unstructured data). This heterogeneity has made a unified analysis, encompassing the entire patient history and characteristics unfeasible [8, 24]. Historically, this has limited model development to (subjective) expert choices of class variables, where such choices have often been based on hard-earned experience through series of studies. The aim has often been isolating the effects of a few variables and study their consequences with respect to a specific ailment. This approach is effective in finding causal relations between the disease but may fail to understand the specific patient unless sufficiently many class variables are included (which again may lead to less interpretable results). Simultaneous integration of diverse data sources is thus a key challenge for medical data analysis.

The past 10 years has seen major breakthroughs in the analysis of previously unassailable text data using machine learning, artificial intelligence and natural language processing, see [13, 14, 17] and their references. Combining this with modern countries endeavoring on vast collection strategies for electronic health care data ([28]), the analysis of EHRs is expanding and changing rapidly both in terms of the available data and methodology. The flexibility of machine learning models and ability to handle diverse data formats show promising results and may help to solve the limited scope cherry-picking class variables using expert identification which might divert a fair statistical analysis. However, many initiatives are still developing and data collection practices are still converging towards national guidelines.

Part of the integration problem is the treatment of all (unique) actions and qualitative observations as class variables, resulting in an enormous dimensionality problem. Solving this requires immense amounts of data – much more than is even available in the era of "big data", and if there is sufficient data, often the practical usefulness of the model is limited [19] – to qualify a lot of data as relevant. Furthermore, in comparisons to other application areas of machine learning, such as Internet text mining, speech recognition and automatic translation, we cannot easily get more data. Access is limited to due

privacy concerns and quickly generating health care data is not an option (as compared to text on web pages or speech recognition). These traits complicates the integration of health care data.

To tackle the integration problem and solve the sub-problem of utilizing free-text sources in medical reports, several studies have adopted a technique called `word2vec` from natural language processing to extract knowledge from free-text radiology reports and other medical text-sources. The seminal work [16] introduced the `word2vec` algorithms consisting of two techniques, CBOW and Skip-Gram, which uses unsupervised learning to obtain semantically meaningful latent word embeddings by training a simple neural network. Their methodology is not restricted to word embeddings but instead assumes that each word is defined by its longitudinal position to other words, a conjecture first introduced in natural language processing as the Distributional Hypothesis [9]. In mathematical terms, this translates into assuming a sequential probabilistic relation between words in a text document (or equivalently in a sequence of discrete states).

In the clinical field, one of the first adoptions of `word2vec` was in [20], to learn concept representations of medical terms based on medical text-databases such as Pub Med. This inspired many authors to use `word2vec`-related methods as a preprocessing step see for example [1–7]. Several of these papers adopt modifications to the `word2vec`-method, either in the data cleaning pipeline [1] prior to `word2vec` application or in neural network architecture [3]. A fundamentally different approach using nonnegative restricted Boltzmann machines is pursued in [31]. Common for all of these studies, the success of the embedding procedure is evaluated on a subsequent classification task, independent of the `word2vec` model.

6.1.1 Contributions of this work

In this study, we aim to solve this integration and dimensionality problem by applying `word2vec`, through utilization of the sequential structure in electronic health records. We examine the obtained representations using clustering techniques to quantify the clustering performance on predetermined event groups (e.g. dentist-related treatment, a specific blood sample package etc.) and verify their strength and sensitivity with respect to our hyperparameters. We successfully identify clusters of events which may serve as a dimensionality and noise reduction technique for patient trajectory analysis. Compared to the above-mentioned studies, which are mostly based on datasets of the magnitude 100 thousand patients and 5+ million of EHR entries, our dataset consists of only 169 patients with a total of 178 thousand event entries. This serves to prove `word2vec` may be successful even for smaller datasets.

The next section provides a description of the dataset and modeling methods. In the results section, we report and discuss our findings and finally in the final section provide a conclusion on our findings.

6.2 Dataset description

We start by providing a practical dataset description and afterwards provide a formal description. The purpose of the formal description is straightforward, it provides an abstract formulation of the framework, containing only the necessary assumptions in order to simplify and clarify the structural assumptions on the dataset. The aim is to make it easy to understand and replicate our results.

6.2.1 Specific dataset description

Our dataset consists of electronic health records (EHRs) from 169 patients. These were collected at Diagnostisk Center (DC), Regionshospitalet Silkeborg, the diagnostic unit at the regional hospital in Silkeborg, Denmark. Our patients are typically chronically ill and often suffer from chronic pulmonary disease and several other ailments, including high blood pressure and high cholesterol – and are often under treatment for these. It is thus important to perceive these patients as being multi-sick, e.g. suffering from several diseases/ailments concurrently, complicating coherent symptoms and successful diagnosis. Adding to this, doctors personalize treatment with their professional, but subjective, judgment which further increase data heterogeneity. For each patient, the EHR is comprised of entries from a number of diverse source (e.g. blood sample, general practitioner visits, medications, surgeries) and as such the data appears highly heterogeneous.

Patient ID	Date (YYYY-MM-DD)	Data source	Eventname
6	2010-12-18	Pharmacy	Mandolgin
6	2014-11-29	Blood Sample	P-Natrium
15	2015-09-05	Procedure	EKG
39	2016-09-02	Pharmacy	Hjerdyl
⋮	⋮	⋮	⋮

Table 6.1: EHR sample entries.

The records are divided into three groups according to final diagnosis from DC, namely lung cancer, colon cancer and arthritis. In this setting, the arthritis patients act as a control group. EHR entries from two years prior to DC referral and until seven days prior to DC referral are included to emulate the diagnostic time frame of a general practitioner. The EHR entries consists of the following variables: anonymous patient ID, data type, event and event date as illustrated in Table 6.1. The data type variable reflects the database from which the event in the EHR was pulled from. This could for example be pharmaceutical database or blood sample database. We note many events occur in batches - for example a single blood sample may be used to perform eight blood tests which results in eight entries in the electronic health record.

The timestamps in each patient record have been pushed by a random time to de-identify the data, but preserve the sequential and temporal structure. A number of summary statistics for the dataset is presented in Table 6.2-6.4.

Level	Feature	Group	Count
Patient-statistics	No. of patients		168
		Arthritis	74
		Colon cancer	38
		Lung cancer	56
Event-level statistics	No. of. unique events		1141
	Most common event count		4441
	Average event count		82

Table 6.2: SSI dataset statistics.

Feature	Group	Value
Total no.		93405
	Arthritis	36088
	Colon cancer	20799
	Lung cancer	36518
Max. no. in one EHR		4021
Min. no. in one EHR		36
Average no. of entries		556
	Arthritis	488
	Colon cancer	547
	Lung cancer	652

Table 6.3: Entry-level statistics.

6.2.2 Formal dataset description

Formally, the dataset can be described in the following way. We consider an unordered collection of sequences e.g.

$$\mathcal{S} = [s_1, s_2, \ldots, s_p],$$

for some $p \in \mathbb{N}$. Each sequence s_j is an ordered set, consisting of items (i.e. entry name or symbol) and is denoted by

$$s_j = \left(i_1, i_2, \ldots, i_{n_j}\right),$$

Data source	No. of events	No. of. entries
Total	1141	94305
Blood sample	320	40883
GP-related activity	174	19766
Pharmacy/Prescription	436	16765
Procedure	177	8473
Radiology	30	378
Hospitalization	2	127
Ambulant	1	1145
Unknown	1	5868

Table 6.4: Data source distribution.

for some $n_j \in \mathbb{N}$ and $i_j \in \mathcal{I}$, where $\mathcal{I}$ denotes the set of all *unique* items in the database $\mathcal{S}$ given by

$$\mathcal{I} := \left\{ i \,\middle|\, \exists\, j \in \{1, 2, \ldots, p\} : i \in s_j \right\}.$$

Note that an item belongs to a sequence $s = \left(i_1, i_2, \ldots, i_{n_s}\right)$ (consisting of n_s items) if

$$i \in s \quad \Longleftrightarrow \quad \exists\, k \in \{1, 2, \ldots, n_s\} : i = i_k.$$

The definition of $\mathcal{I}$ allows us to pair each symbol with a unique numerical symbol ID in the natural numbers. Thus henceforth an item or event may also refer to its corresponding ID. The items and their order define the sequence and hence the sequences

$$s_a := (i_1, i_2, i_3), \quad s_b := (i_2, i_1, i_3)$$

are not equal, provided that $i_1 \neq i_2$. Examples of this structure could be a corpora of documents (sequences) with items being words, or a database of electronic health records with recorded entry names being items. In the context of Section 6.2.1, the collection of sequences correspond to the datasets of 169 electronic health records (sequences), consisting of events (items) ordered by their time & date. The underlying presumption is that the sequential structure and local context (as defined by the ordering) defines the purpose and meaning of each item – in natural language processing this is called the Distributional Hypothesis [9]. Given Table 6.1, the sequence for patient 6, start as

$$s_6 = (\text{Mandolgin, P-Natrium}, \ldots).$$

Assuming that Mandolgin, respectively P-Natrium, are given the event ID 1, respectively event ID 2, the sequence would start as

$$s_6 = (1, 2, \ldots).$$

As added auxiliary information in our dataset, we know the data source for each item/event. To formalize this, let DS denote the space of possible data source (in our case 8 different data sources as shown in Figure 6.4) and observe that

$$\forall\ i\ \exists!\ d \in \mathrm{DS} : \text{data source of } i \text{ is } d,$$

where $\exists!\ d$ denotes there exists a unique d. To ease the notation, we will Section 6.4.3 simply refer to d as $\mathrm{DS}(i)$. Additionally, for each sequence $s \in \mathcal{S}$ we know the resulting diagnosis $\mathrm{Diag}(s) \in \{\text{arthritis, colon cancer, lung cancer}\}$ which is relevant in Section 6.4.5.

Exactly how we utilize the sequential structure in Section 6.2.2 is a modeling question, one possible choice is the embeddings from `Word2vec`, [16]. Embeddings arises in natural language processing and concern encodings, e.g. mathematical representations of words (or items). The roughest possible encoding is the one-hot encoding where each unique symbols is represented by a $|\mathcal{I}|$-dimensional standard unit vector, i.e. let y denote an item with item ID k

$$y = (0, 0, \ldots, 0, \underset{\substack{\uparrow \\ k\text{th index}}}{1}, 0, \ldots, 0) \in \mathbb{R}^{|\mathcal{I}|}.$$

In natural language processing, this illustrates how we would understand words, if we were only able to form sentences consisting of a single word with no relational structure surrounding it (i.e. a sentence of multiple words).

6.2.3 Workflow description

The ecosystem, or workflow, for the data analysis is shown in Figure 6.1. It illustrates how we first clean the raw records through cut-off application (masking each event which occurs below a certain threshold to a dummy value 0), then apply `word2vec` Skip-Gram to obtain event embeddings. These embeddings are then fed into some classification tasks and the performance can be measured. The main point of embeddings is that they are trained prior to training the classifier, allowing a stronger starting point than inputs which have not been trained prior to classification.

6.3 Methods

Event embeddings have been introduced in the previous section and overall concern the mathematical encoding of qualitative variables, in our cases events, into numerical vectors. We aim to model the event meanings by their sequential placement in EHRs. This is done using two techniques, namely Skip-Gram and Markov Chains, which each find an encoding based on the sequential order of events. Note that both encodings do not necessarily optimize with respect to the same goal and the "encodings" are of different dimensions. Nonetheless they both aim at representing the relationship between states through their sequential structure.

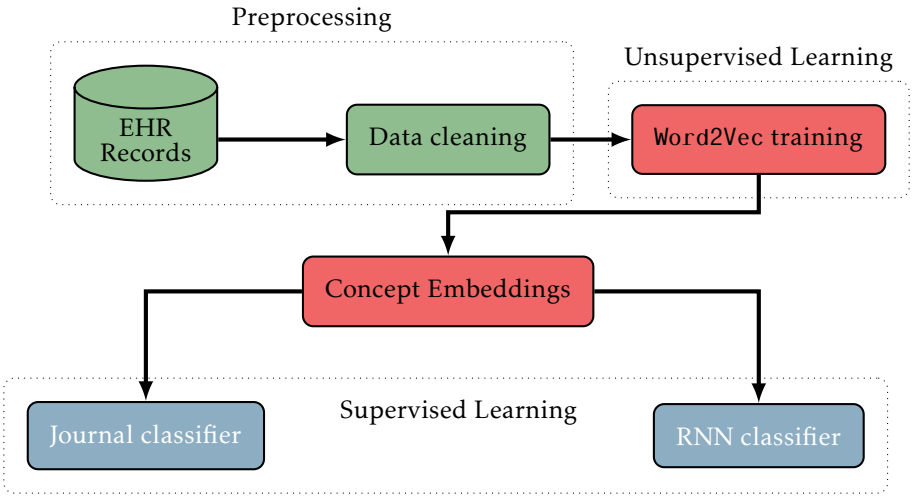


Figure 6.1: Model development pipeline describing the workflow, starting with a dataset and ending with a prediction task

6.3.1 `Word2vec`: continuous Skip-Gram

Continuous Skip-Gram was introduced in [16] as an unsupervised learning method to obtain semantically meaningful word representations in natural language processing, with additional computational methods in [18]. The objective of the Skip-Gram algorithm is to estimate the weight matrix $\mathbf{W}$ and $\mathbf{W}'$ in equation (6.1). Here each row $\mathbf{W}_i$ gives an event representation for the event with ID i. The neural network for Skip-Gram is single hidden layer fully-connected feedforward neural network, where the input is a one-hot encoding of the event-ID and the output is an estimated event within the context size C of the input event. The context around the entry i_k in a sequence, is defined as the window of size W around the entry, e.g.

$$i_1, i_2, \ldots, \underbrace{i_{k-W}, i_{k-W+1}, \ldots, i_k, i_{k+1}, \ldots, i_{k+W}}_{\text{window of size } W \text{ around } k \text{ (apart from } i_k)}, i_{k+W+1}, \ldots,$$

The true output event is sampled from the indices within context size C, according to a uniform distribution. The Skip-Gram dataflow graph with input x can be written as the composition of the following mappings

$$x \mapsto (\mathbf{W}x) \mapsto \mathbf{W}'(\mathbf{W}x) \mapsto \operatorname{softmax}(\mathbf{W}'(\mathbf{W}x)) := \hat{y} \in \mathbb{R}^V, \tag{6.1}$$

where $\mathbf{W}$ and $\mathbf{W}'$, denotes matrices of unknown trainable parameters (or weights) with size $D \times V$, respectively $V \times D$, where D denotes the (chosen) embedding dimension and V, the vocabulary size, denotes the number of unique events after the cutoff (hence $V \leq |\mathcal{I}|$). The function *softmax* is defined on $z \in \mathbb{R}^V$ as

$$\mathbb{R}^V \ni \operatorname{softmax}(z) := \left(\frac{\exp(z_j)}{\sum_{k=1}^{V} \exp(z_k)} \right)_{j=1,\ldots,V}.$$

Let y denote the one-hot encoding true output event with ID k. We measure the similarity between y and $\hat{y}$ with the cross-entropy loss function[2] and this results in

$$\mathcal{L}(y, \hat{y}) = -\sum_{j=1}^{V} y_j \log \hat{y}_j = \hat{y}_k - \log\Big(\sum_{j=1}^{V} \exp(\hat{y}_j)\Big), \tag{6.2}$$

Backpropagation, introduced in [26], is used to optimize the parameters in (6.2). Post-training, we normalized each event representation by its Euclidean norm, as is standard in the literature. The additional modifications to Skip-Gram, Hierarchical Softmax and Negative Sampling, was added in [18]. We only used experimented with Negative Sampling and a standard softmax (i.e. no computational optimization). The purpose of Hierarchical Softmax is computational simplification (approximation to the softmax function) which was unnecessary for us due to our small dataset.

In reference to the findings in [29], we experimented with normalized and unnormalized representations/rows, but we did not notice any systematic differences between normalized and unnormalized vectors.

Parameter fine-tuning in Skip-Gram

The loss function of Skip-Gram changes whenever the number of negative samples is adjusted and hence test/training error cannot be used as a performance criteria to set the number of negative samples. Parameter fine-tuning was thus done using grid-search and visualizing the embedding using t-SNE, see [15]. t-SNE is a visualization technique used to illustrate similarities between high-dimensional points in a low-dimensional setting (2- or 3-dimensional) and has gained significant popularity since its introduction. Before the analysis, a cutoff threshold was applied in the following way, if the event occurs between the cutoff, we replace the event with a generic event token Unknown event as seen in Figure 6.4. The alternative was to completely remove such events, but this way maintains the sequential spacing between events and avoid synthetic pairings arising from removals. Notable hyperparameters of Skip-Gram are are

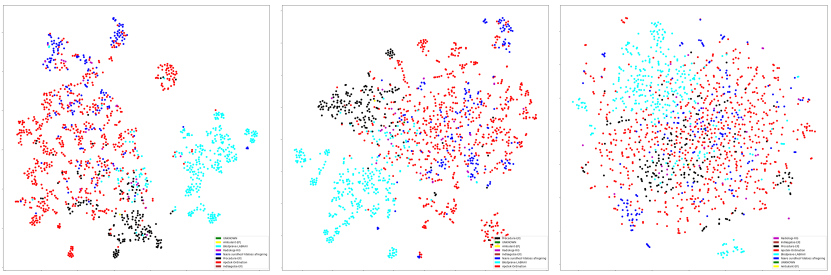
- #negative samples, embedding dimension, cutoff threshold, window-size.

Our primary visualizations and analysis are based on

- #negatives samples = 12, embedding dimension $\in \{10, 20\}$, censoring cutoff = 5.

If the dataset is very heterogeneous like ours, censoring cutoff can heavily homogenize the results (though with the risk of oversimplifying). If the timestamps are unreliable, e.g. 10 observations at 10:24 on the same day (typical for medical observations), then the window-size can be chosen to take this into account.

[2]The one-hot encoding y is viewed as a degenerated probability distribution for the cross-entropy definition.



(a) Embed. dim. 5 **(b)** Embed. dim. 10 **(c)** Embed. dim. 50

Figure 6.2: t-SNE visualizations of Skip-Gram event representations with colouring based on data source variable.

Dimensionality reduction of Skip-Gram

Skip-Gram is often introduced understood as a dimensionality reduction technique, e.g. it reduces the (one-hot encoding) dimension of data to a chosen embedding dimension. However, given that we may choose *any* embedding dimension, a natural question is:

Which embedding dimension is the best?

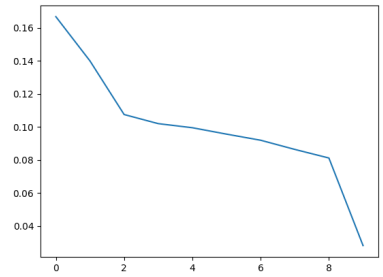
Clearly, this question depends on the definition of best. We note however, that the effect of the embedding dimension can be drastic. This is illustrated in Figure 6.2 which shows the effect of changing embedding dimension through t-SNE ([15]) visualizations. Traditionally, dimensionality reduction can be evaluated using (kernel) Principal Component Analysis (k-PCA) on the obtained weight representations $(\mathbf{W}_i)_{i\in 1,\dots,V}$, see [22] and [30]. k-PCA aims to capture (inherent) lower-dimensionality in a higher-dimensional space and the *kernel* allows fitting to certain non-linear lower-dimensional manifolds.

We test this using grid-search across linear, polynomial and Gaussian kernels and embedding dimension 5, 10, 15, 20, 50 and 100. Figure 6.3 shows the eigenvalues decay for some of these runs. The initial drop-off point consists of a single eigenvalue and is explained by the average. Overall, we observe that eigenvalues are generally of the same magnitude without a sharp cut-off point which is the common measure to identify a lower dimensionality. This indicates data cannot be represented through a lower-dimensional manifold. This tells us two things, either our dataset consists almost mostly of noise with no inherent lower dimensionality, or Skip-Gram is not doing dimensionality reduction in the traditional sense.

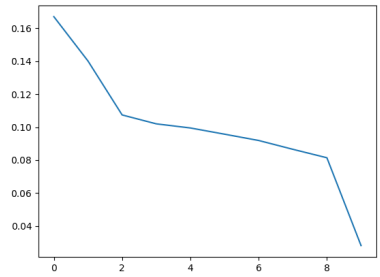
6.3.2 Markov Chains

Markov Chain are classical in mathematical modeling, explicitly constructed to model probabilistic transitions between discrete states, e.g.

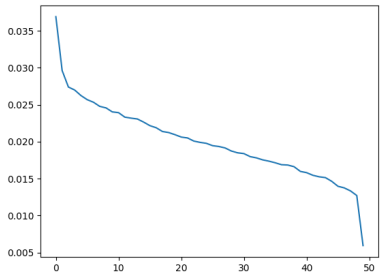
$$\cdots \to i_j \to i_{j+1} \to \cdots$$



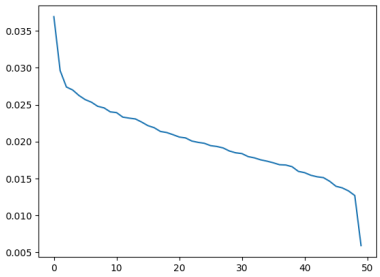
(a) Linear kernel. Embedding dimension 10



(b) Polynomial kernel. Embedding dimension 10



(c) Radial Basis kernel. Embedding dimension 50



(d) Linear kernel. Embedding dimension 50

Figure 6.3: Plots of eigenvalues for different kernels and embedding sizes

where X_t belong to some discrete set of events. For more theory, we refer the reader to [25]. As an alternative to the event representation obtained using Skip-Gram, we estimate the $V \times V$ transition matrix $\mathbf{P}$ of a simple one-step (or order one) Markov Chain and use the rows of P as event representations. In a 1-order Markov Chains, the context of an entry is simple and merely defined as prior entry, i.e.

$$\ldots, \underbrace{i_{j-1}}_{\text{context of } i_j}, i_j, i_{j+1}, \ldots.$$

An event representation from a Markov Chain refers to the ith row $\mathbf{P}_i$ of $\mathbf{P}$, e.g. the estimated conditional distribution given current event equals i. This estimation procedure may lead to imprecise estimates for rare transitions due to our dataset size. The nature of Markov Chains lends itself to effective estimation of next-event, provided the event space is small. This is different from Skip-Gram which utilizes a context to pair events and thus is trained to predict the probability of state b inside the context window of state a. This difference shows up again in Section 6.4.4.

6.3.3 Implementation of methods

We train Skip-Gram and fit a Markov Chain on a train/test split of 60/40 percent across patients. Due to the size of our dataset, we have chosen a relatively large test set to avoid over-fitting. The code-implementation of Skip-Gram was done using the open source machine learning library Tensorflow in Python 3.5. The implementation of Markov Chains is straightforward and was implemented using the library Numpy in Python 3.5. The classification algorithms were implemented using the default configuration of the algorithm in the scientific computing library [23] for Python. We refer to their documentation for implementation details as well as [10] for mathematical details.

6.4 Results

In this study, we explore the ability of Skip-Gram to cluster medical events and support data analysis in clinical decision support systems (CDSS). We started by finding the event representations, that is, the weight matrix W in Skip-Gram and transition matrix P for the Markov Chain.

We evaluate the representations using the following criteria

(1) t-SNE-visualization of representation with annotated event groups.

(2) Unsupervised clustering to rediscover the annotated event groups from (1).

(3) Data source classification based on representation and the data source class.

(4) Next-event prediction, given the current event.

(5) Diagnosis classification, given the averaged event representation of the patients electronic health record.

We elaborate further on the evaluation criteria in their respective subsections.

6.4.1 Visualization of Skip-Gram

We visualize our Skip-Gram embedding using t-SNE obtained with embedding size 10 in Figure 6.4, where each point in the plot refers to a unique event. In the figure, post-visualization, we have identified and marked a number of categories and we see that the figure demonstrates how the Skip-Gram representations cluster related events together, and in this sense captures an equivalent of a "semantic meaning" for the EHRs.

Practically, we observe a large portion of events in a scattered cloud which we identify as consisting mostly of events performed/prescribed by a general practitioner (GP). Various treatment packages, dental treatments, mammography screening, arthritis-related activities and blood circulation treatment surround the larger cloud of GP-related events and illustrate the notion that these treatment packages are subgraph in the patient trajectory and supports

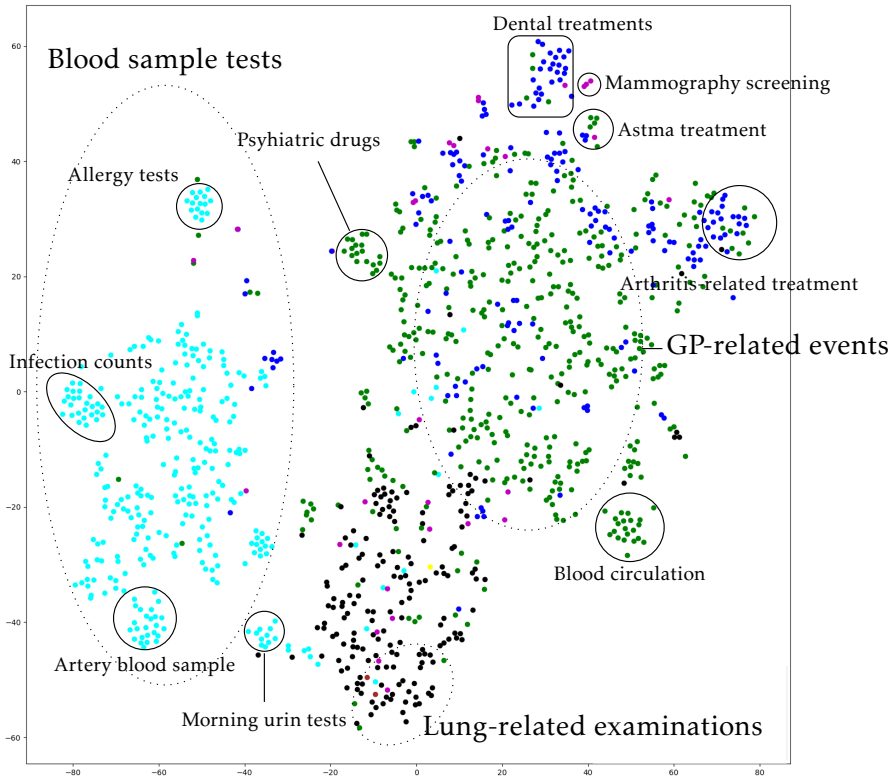


Figure 6.4: t-SNE visualization of Skip-Gram event representations. The coloring is based on the data source and the event name for each point is known. The marked groups have been identified post-visualization and illustrates the effectiveness of Skip-Gram in identifying clusters of events. The drawn lines is directly related groups and the dotted lines are conceptually related groups. Read more on this in the text.

our understanding of the general practitioner as being the hub for referrals. Central in the plot, at the bottom, a scattered grouping is colored in mostly black and green. This cloud consists several types of events with no single common denominator other than examinations of multi-sick patients, with the lower half of this cloud consists of lung-related examinations connected with a reduced lung functionality but this is also closely connected to dietary recommendations. The upper parts of the cloud appear to be Parkinson's disease examinations but certain events obfuscate this hypothesis, leading us to believe it is a more general group of psychiatric/neurological evaluation(s). In conclusion, it appears to mainly consists of directly diagnosis-related treatments and few treatment-related events. To the left, a large teal cloud consist almost entirely of blood sample tests. Within this group, we observe several

subgroups such as a group for tests of allergies, artery blood sample test package and several other groups. It is particularly evident here that blood sample tests are performed in packages. Some of these are exemplified as markings in Figure 6.4. The structure of the entire plot in Figure 6.4, indicates many events are related to each other (the large green/black cloud), whereas certain groups of events happen mostly in connection with each other, illustrated as isolated groups and exhibiting a treatment-subgraph.

6.4.2 Unsupervised Clustering

Unsupervised clustering technique can be used to "rediscover" an underlying class structure. In Subsection 6.4.1, we identified several clusters of treatment packages in a t-SNE visualization of Skip-Gram embeddings. We perform this analysis to objectively quantify the derived clusters and avoid visual bias. To test this quantitatively, we apply unsupervised clustering techniques (k-means and k-medoids, see [10]) to the Skip-Gram representations $(W_i)_{i\leq V}$ and verify whether the algorithm successfully creates "clean" clusters, e.g. the same data source variables in a cluster (even for large k) and whether it recreates clusters identified in Figure 6.4. The standard F_1-score is not applicable here, since k-means may need more clusters than there are data source variable values to fully describe the data. Instead, we use the adjusted Rand Index [12], which is a measure of the agreements between two different partitions of the data points and can be used across different partition sizes (e.g. different number fo classes). The adjusted Rand Index is one for a perfect partition match, and has expected value zero for a random cluster (e.g. random cluster assignemnt). However, one should exercise caution in merely studying the value of the adjusted Rand Index, as it scales inversely with the number of clusters k, as evidenced in Table 6.5. In other words, we compare the partition/clustering from k-means/medoids with the partition from the data source variable (true label). Let $\mathcal{I}^* \subseteq \mathcal{I}$ denote an annotated cluster from Figure 6.4 and let $(K_j)_{1\leq j\leq k}$ denote the jth cluster from k-means or k-medoids. Define cluster accuracy for $\mathcal{C}$ by

$$\mathrm{CA}_{\mathcal{I}^*} = \frac{\max_{1\leq j\leq k} |K_j \cap \mathcal{I}^*|}{|\mathcal{I}^*|},$$

e.g. the most common cluster K_j for the events in $\mathcal{C}$ divided by the total cluster size. Hyperparameter search was done with $k \in \{5, 6, \ldots, 150\}$ but for simplicity we only display $k \in \{5, 10, 20\}$ as the results extrapolate. The "kernel trick" (see p. 488 in [21]) was used on k-medoids with linear, polynomial and Gaussian kernels, however we subsequently omitted the three kernels, as they had exactly the same performance. The results are shown in Table 6.5. To account for this, we compute a 95-% confidence interval obtained via bootstrapping and study whether the Rand Index belongs to this. Table 6.5 shows the adjusted Rand index is firmly outside the bootstrap interval, signifying that our unsupervised clustering technique is able to clusters point into their correct data source class and much stronger than random guessing. For the predetermined

	k-means		k-medoids	
n	Bootstrap interval	Rand index	Bootstrap interval	Rand index
5	(-0.0025,0.0040)	0.36	(-0.0028,0.0040)	0.31
10	(-0.0024, 0.0031)	0.22	(-0.0024,0.0026)	0.24
20	(-0.0016, 0.0020)	0.11	(-0.0020,0.0022)	0.13

Table 6.5: Unsupervised clustering performance.

groups of events, as annotated in Figure 6.4, the performance of k-means is shown in Table 6.6. The accuracy is measured as the percentage of the cluster within the same unsupervised cluster from k-means. We observe that k-means successfully clusters several of the annotated groups to the same clusters.

$k = 20$		k-means	k-medoids
Event Cluster	No. of events	Cluster accuracy	
Allergy tests	16	1.0	1.0
Artery blood sample	24	0.96	0.96
Arthritis-related treatment	35	0.8	0.77
Urin tests	12	0.67	0.75

Table 6.6: Unsupervised clustering of select subgroups.

6.4.3 Data source classification

Events tend to occur in batches, as mentioned in the dataset description, and in particular close to events of the same *data source*, see Figure 6.1 for sample entries and Figure 6.4 for summary statistics. This motivates the following. We train three classifier: k-nearest neighbors classifier (k-NN), Random Forest Classifier (RFC) and Support Vector Machines (SVM). For details on the classifiers, we refer [10]. Denoting the classifier by C, our prediction is given by

$$C(x) = \hat{y},$$

where x denotes the event representation from Skip-Gram and $\hat{y} \in \mathbb{R}^V$ is the predicted data source from the classifier. For the SVM, the "kernel trick" (see [21]) was used with linear, polynomial and radial basis kernels but neither showed any performance improvement over the other, hence for simplicity we just report the results of a linear kernel. We measure the performance by the accuracy, i.e.

$$C_{\text{AC}} = \frac{1}{\#\text{samples}} \sum_j \mathbb{1}_{\{C(x_j)=y_j\}}$$

For all of the above classification methods, the 5-fold cross-validation accuracy was computed and is stated in Table 6.7. These results apply across a wide range

Table 6.7: Data source classifier accuracy

	Skip-Gram			Markov Chain
	k-NN	RFC	SVM	Next-event DS
CV accuracy	0.77	0.77	0.76	0.87

of embedding dimensions, with negligible changes in accuracy. We observe that the classification significantly outperforms naive guessing (which at best would guess the most common class every time, resulting in accuracy of 0.38). This shows that our event representations contains strong relational information and despite not being trained for it, it is capable of finding structure that was not part of the original input (only directly, due to batches). Knowing that events tends to occurs in batches, a more direct and local approach to guessing the data source could be more straightforward. For example given an event x, our predicted data source is given by

$$C(x) = \text{DS}(\arg\max_j (\mathbf{P}_x)_j),$$

where DS(·) denotes the data source of the event with id ·, i.e. the data source of the most likely next event. This result in an accuracy of 0.87 and outperforms the other methods. This is not surprising since the Markov Chain is specifically trained to predict the next event (and thus indirectly, the next data type), and aligns nicely with the batch occurrence of events. For comparison, Skip-Gram is trained with a larger context window (window size 10) and thus is not as "local" as Markov Chains.

6.4.4 Next-event prediction comparison

We evaluate the performance of the representation by testing its ability to predict the next event, e.g. the accuracy of next-event prediction. The large event space (~1100 unique events) implies that these predictions cannot have a high accuracy. For the Markov Chain, the prediction for event y_{i+1} given y_i is simply

$$\hat{y}_{i+1} = \arg\max_i \mathbf{P}_i. \tag{6.3}$$

For Skip-Gram, we use two methods due to the training goal of Skip-Gram as discussed in Section 6.3.2. In the first method, the next-event prediction is the event in closest proximity of the average contextual event representation, e.g. the average of the event representations. This corresponds to the Continuous Bag-of-Words (CBOW) algorithm from the `word2vec` paper, e.g.

$$\widehat{\text{event}}_{i+1} = \arg\max \text{softmax}\left(\mathbf{W}^T(\mathbf{W}')^T \frac{1}{|C|}\sum_{j\in C} x_j\right),$$

where C denotes the context size and x_j denotes the jth event in the context.

In the second method, a recurrent neural network (RNN; [11, 27]) is trained on the embedded patient trajectories to perform next-event prediction. We use a long short-term memory cell for the RNN (same as [2]), where the cell input is the Skip-Gram event representation. Every output of the RNN is turned into an appropriately dimension probability distribution, using a fully connected feed-forward neural network with softmax activation function.

	CBOW	Markov Chain	RNN
Test accuracy	0.01	0.18	0.28

Table 6.8: Next-event accuracy

The CBOW algorithm performs badly and falls victim to the same problem as mentioned at the end of Section 6.4.4, namely mostly predicting the average event with an extremely small diversity of predicted events. Nonetheless, the same Skip-Gram representations used in a RNN outperforms the Markov Chain.

6.4.5 Diagnosis classification

Each electronic health record correspond to a patient who has visited the Diagnostic Center, and received a diagnosis (lung cancer, colon cancer or arthritis). To draw comparison to the related article mentioned in the introduction, we also evaluate a classification task.

Similar to [1], we train a k-NN classifier on patient journals through journal average representations (e.g. average of word vectors). We fit the classifier on a train/test split of 60/40 percent across patients and report the test set accuracy. Our test set accuracy using journal averages does not amount to more than

	Normalized journal vectors
Test set accuracy	0.45

Table 6.9: Accuracy for the diagnosis prediction.

random guessing, considering that 44 percent of all patients belong to the arthritis group. Evidently, our classification performance is poor compared directly to other results, e.g. [2] and [1], but comparatively our dataset is smaller and much more heterogeneous which may explain the difference. Noteworthy, classification performance may depend on the events which are unique to a certain class, hence effectively acting as a classifier group (see visualizations in [1] and [15]). Comparatively, this is not the case for our dataset – data heterogeneity implies that no common, or even uncommon, event(s) can effectively be attributed to a single diagnosis.

6.5 Conclusion

In this study, we benchmarked the ability of Skip-Gram to obtain latent event representations with Markov Chains. We successfully identify concept groups in Figure 6.4, showing that Skip-Gram effectively and autonomously clusters treatment regimens together without prior knowledge of treatment relations other than the sequential structure. In addition, visualization reveals information on the large-scale structure of treatment trajectories, successfully identifying larger regions, e.g. GP-related events and blood samples. The results are often so clear, that even a non-specialist is able to recognize the similarities produced with this method. We conclude that Skip-Gram is capable of identifying subgraphs in the patient trajectory and associating events. We believe that this could serve as inspiration to standardize treatment packages into a single EHR entry, simplifying patient trajectory and pathway analysis and decreasing data heterogeneity significantly.

The lower-dimensional mathematical representation allows us to integrate the data with other numerical observations while the sequential structure mitigates the information loss in connection with the dimensionality restriction. This, in turn, would greatly enhance model robustness and lower variance.

For the future, we aim to establish whether this the methodology could be utilized as dimensionality reduction in the following two-fold iterative manner. First, reduce dimensionality and heterogeneity by identifying sub-graphs using clustering techniques on initial `Word2Vec` representation. Manually inspection with domain knowledge to group treatment packages into a single entry (requires reasonably sized event spaces). Secondly, reapply `Word2Vec` representation on the reduced to obtain mathematically rich relational representations.

Acknowledgements

The authors would like to thank Peter Vedsted, Professor at Institute for Public Health, Aarhus University and Andrew Bolas, Data Manager at Regionshospital Silkeborg for providing the dataset.

Appendix

Research data for this article.

> Due to the sensitive nature of electronic health records, raw data is protected health information and possibly personally identifiable and hence must remain confidential and cannot not be shared.
>
> 🔒 Data not available / The data that has been used is confidential.

Funding: This work was supported in part by a Sapere Aude grant from the Independent Research Fund Denmark (danish: Det Frie Forskningsråd).

References

[1] Imon Banerjee, Matthew C Chen, Matthew P Lungren and Daniel L Rubin. "Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort". *Journal of biomedical informatics* 77 (2018), 11–20.

[2] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart and Jimeng Sun. "Doctor ai: Predicting clinical events via recurrent neural networks". *Machine Learning for Healthcare Conference*. 2016, 301–318.

[3] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo and Jimeng Sun. "Multi-layer representation learning for medical concepts". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, 1495–1504.

[4] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart and Jimeng Sun. "GRAM: Graph-based attention model for healthcare representation learning". *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, 787–795.

[5] Edward Choi, Andy Schuetz, Walter F Stewart and Jimeng Sun. "Medical concept representation learning from electronic health records and its application on heart failure prediction". *arXiv:1602.03686* (2016).

[6] Edward Choi, Andy Schuetz, Walter F Stewart and Jimeng Sun. "Using recurrent neural network models for early detection of heart failure onset". *Journal of the American Medical Informatics Association* 24.2 (2016), 361–370.

[7] Youngduck Choi, Chill Yi-I Chiu and David Sontag. "Learning low-dimensional representations of medical concepts". *AMIA Summits on Translational Science Proceedings* 2016 (2016), 41.

[8] Keith Feldman, Louis Faust, Xian Wu, Chao Huang and Nitesh V Chawla. "Beyond volume: The impact of complex healthcare data on the machine learning pipeline". *Towards Integrative Machine Learning and Knowledge Extraction*. Springer, 2017, 150–169.

[9] Zellig S Harris. "Distributional structure". *Word* 10.2-3 (1954), 146–162.

[10] Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

[11] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". *Neural computation* 9.8 (1997), 1735–1780.

[12] Lawrence Hubert and Phipps Arabie. "Comparing partitions". *Journal of classification* 2.1 (1985), 193–218.

[13] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". *Advances in neural information processing systems*. 2012, 1097–1105.

[14] Honglak Lee, Roger Grosse, Rajesh Ranganath and Andrew Y Ng. "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations". *Proceedings of the 26th annual international conference on machine learning*. ACM. 2009, 609–616.

[15] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". *Journal of Machine Learning Research* (2008), 2579–2605.

[16] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. "Efficient estimation of word representations in vector space" (2013). arXiv: 1301.3781.

[17] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ and Sanjeev Khudanpur. "Recurrent neural network based language model". *Eleventh Annual Conference of the International Speech Communication Association*. 2010.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. "Distributed representations of words and phrases and their compositionality". *Advances in Neural Information Processing Systems*. 2013, 3111–3119.

[19] William Dwight Miller, Kimngan Nguyen, Sitaram Vangala and Erin Dowling. "Clinicians can independently predict 30-day hospital readmissions as well as the LACE index". *BMC health services research* 18.1 (2018), 32.

[20] José Antonio Minarro-Giménez, Oscar Marin-Alonso and M. Samwald. "Exploring the application of deep learning techniques on medical text corpora." *Studies in health technology and informatics* 205 (2014), 584–588.

[21] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[22] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), 559–572.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[24] John H Phan, Chang F Quo, Chihwen Cheng and May Dongmei Wang. "Multiscale integration of-omic, imaging, and clinical data in biomedical informatics". *IEEE reviews in Biomedical Engineering* 5 (2012), 74–87.

[25] Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.

[26] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. "Learning representations by back-propagating errors". *Nature* 323.6088 (1986), 533.

[27] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. "Learning representations by back-propagating errors". *Cognitive modeling* 5.3 (1988), 1.

[28] Ruby Sahney and Mukesh Sharma. "Electronic health records: A general overview". *Current Medicine Research and Practice* 8.2 (2018), 67–70.

[29] Adriaan M. J. Schakel and Benjamin J. Wilson. "Measuring Word Significance using Distributed Representations of Words". *CoRR* (2015). arXiv: 1508.02297.

[30] Bernhard Schölkopf, Alexander Smola and Klaus-Robert Müller. "Kernel principal component analysis". *International Conference on Artificial Neural Networks*. Springer. 1997, 583–588.

[31] Truyen Tran, Tu Dinh Nguyen, Dinh Phung and Svetha Venkatesh. "Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM)". *Journal of Biomedical Informatics* 54 (2015), 96–105.

Chapter 7

Analysis of medication sequences for sepsis patients

Thorbjørn Grønbæk

Draft

7.1 Introduction

A central belief in every person is that we should not get "more" sick by attending the hospital or a similar health care institution. Sepsis, an infection in the blood, however, occurs primarily in the hospital setting for already sick patients. It has been shown to often be connected to insufficient hygiene precautions and implanted foreign bodies. Sepsis is be a life-threatening condition with a 30-80 percent mortality rate, depending on the circumstances, and every precaution should be taken to prevent it from occurring and to treat it effectively when it does. It may be difficult to diagnose as symptoms may be few, if any, or overlap with other diseases so that the patient is already in a serious sepsis condition when diagnosed. Both a timely diagnosis and effective treatment are needed, as fast response time may drastically reduce the mortality rate.

In this work, we study the treatment of sepsis, which primarily includes antibiotics and intravenous fluids. Unfortunately, we have only limited knowledge on the effectiveness of a given antibiotic for a specific patient, despite the alarming mortality rate. Most treatment is currently based on a general prioritized list of effective medications contrary to custom medication choices based on patient attributes.

We study two datasets of medication orders for patients admitted to Stanford Health Care. In the first dataset, we study the medication order for patients for whom an alert for sepsis is registered when admitted. This dataset is divided into two groups during a trial period – in one group an alert is simply registered and in the other group it is registered and sent to a doctors pager. It is not confirmed in the dataset whether the patients actually had sepsis or not

– thus we will also be analyzing false positives.

We analyze how the alert system to study affects the treatment. Further details on the alert system and data collection process are presented in Section 7.2.

In the second, much larger, dataset, we also study the medication ordering and the graph of treatment packages (introduced in the next section). This dataset constitute a single group as it had no alert registration. Since we conclude in our analysis on the first dataset that sending the alert to a doctors pager does not alter the treatment, we decided to merge the first dataset with the second dataset to form a merged dataset.

7.2 Data collection and datasets

The data consists of 24 hour time window medication logs following a sepsis alert (or registration for the second dataset). The alerts and registrations includes false positives. The log stores medications and intravenous fluids provided to the patient during the treatment of sepsis (or possibly no treatment, if no sepsis was diagnosed). A sample record is displayed in Table 7.1. We shall

Patient ID	Date (YYYY-MM-DD HH:MM)	Medication ID	Medication name
6	2010-12-18 08:13	143	Terazosin
6	2010-12-18 08:37	37	Pravastatin
6	2010-12-18 11:00	14	Metformin
6	2010-12-18 14:00	17	Metoprolol

Table 7.1: Synthetic sample log for a single alert occurring with timestamp 2010-12-18 6:37.

henceforth refer to a 24 hour time window with medications and intravenous fluids as a *sequence*. As described below on noisy timestamps, it is necessary to truncate timestamps into the nearest hour and thus the resulting sequence obtained from Table 7.1 is given by

$$(\{143, 37\}, \{14\}, \{17\}), \tag{7.1}$$

where we use the notation of Subsection 5.1. We shall call the itemsets (as explained in Subsection 5.1) for treatment packages, e.g. $\{143, 37\}$ and $\{14\}$ are *treatment packages*. Both datasets have been preprocessed for errors, and thus the represented characteristics are obtained after cleaning. The cleaning included removing patients which had no medication entries at all in the dataset (only 2 patients) and removing alerts which had both an active alert and a control alert within the designed 24 hour window of observation – thus rendering it unclear whether to assign the alert to the active or control group. These two patients were not included in the merged dataset.

Noisy timestamps

Timestamps are unfortunately noisy and imprecise – this is a part of the data collection pipeline, as we use the timestamp on which the data is entered (other timestamp choices are not always available). As an example of this, a patient had an alert trigged at 7pm in the evening and the next entry in the entire patient log (not just a medication entry) was at 1pm the following day, where around 20 treatments/events were logged. Clearly, we cannot have 20 treatments/events within a single minute in the real world, but in the logged world, this occurs somewhat regularly. This limits reliability of the given sequential ordering of entries. Of course, this is an extreme case, but it serves to show that timestamps are noisy (thus necessarily the sequential order is noisy). To mitigate this, we instead study 'treatment packages', i.e. we merge all the given medications within a single hour into a single treatment package (item) to avoid relying too heavily on the noisy timestamp. This means that the sequences are truncated and may contain at most 24 entries (1 for each hour) and that the space of possible treatments packages is increased (by combining different medications at different hour slots). This truncation aligns with the clinical reality, namely that a treatment strategy is often chosen rather than a single medication. Note that if the hour contains only a single or zero medications, the given treatment package is the same as the given medication.

7.2.1 Two datasets

The first dataset studied in this paper originates from Stanford Health Care and concerns testing a clinical alert system for sepsis. Previously, no system had been used to warn of sepsis risk/symptoms and thus as an introductory study, an alert system was set into place where half of the alerts was sent to doctor pagers (active group) and the other half simply logged as a control group. We shall refer to these groups as active and control to align with the medical literature. The alert system itself was simple and only based on value thresholds for certain observables acting as alert triggers. The setup of the alert system is not further specified.

The goal was to determine whether sending the alert had an actual treatment-altering effect. The characteristics of the first dataset is described for both groups in Table 7.2 and Table 7.3. Note that number of sequences does not correspond to number of patients as a patient may contribute with several alerts, as long as the 24 hour time windows do not overlap.

The second dataset contains logged sepsis registrations but without division into groups as in the first dataset. Medications and intravenous fluids are registered using the same procedure as for the first dataset. As mentioned in the introduction, we decided to merge the second dataset with the first dataset by dropping the active/control group variable for the first dataset. This results in the merged dataset.

The merged dataset consists of the two datasets and these have been cross-checked for duplicate patients and sepsis registrations. The characteristics of

Level	Variable	Amount
Dataset	No. of sequences	737
	No. of empty sequences	55
Sequence	Shortest sequence	1
	Longest sequence	18
	Average sequence length	3.6
	Median sequence length	3
	No. of unique treatment packages	1061

Table 7.2: Statistics for the active group of the first dataset.

Level	Variable	Amount
Dataset	No. of sequences	685
	No. of empty sequences	50
Sequence	Shortest sequence	1
	Longest sequence	17
	Average sequence length	3.5
	Median sequence length	3
	No. of unique treatment packages	947

Table 7.3: Statistics for the control group of the first dataset.

the merged dataset is described in Table 7.4. The goal for the analysis of the merged dataset is to visualize the graph to look for rare sub-graphs and predict the next medication using a Markov Chain.

It is evident from the average sequence length that we cannot expect a deep analysis, when we often only have 2 or 3 medications during the 24 hour time window. Clearly, a Markov Chain with 16971 unique states and $\approx 3 \times 14833 \approx 44500$ entries is not enough to properly estimate the transition probabilities for all states. Nonetheless, we still wish to fit a Markov Chain to establish a benchmark. Thus it is necessary to setup a further data processing step, in which we filter out infrequent entries to obtain fewer states. This is done using Sequential Pattern Mining.

7.3 Methods

Overall our data consist of qualitative variables (medications/intravenous fluids) and we aim to model them through a Markov Chain. The initial data sequences are, however, too short and diverse to obtain a strong model fit. To mitigate this, we use Sequential Pattern Mining to find the frequent subsequences of treatment packages and filter the sequences using these.

Level	Variable	Amount
Dataset	No. of sequences	14833
	No. of empty sequences	4907
Sequence	Shortest sequence	1
	Longest sequence	19
	Average sequence length	2.9
	Median sequence length	2
	No. of unique treatment packages	16971

Table 7.4: Merged dataset statistics.

7.3.1 Sequential Pattern Mining

In this section, we introduce Sequential Pattern Mining (SPM) for which we observe that our datasets fits the framework of subsection 5.1. Sequential Pattern Mining (SPM) is a classical method from computer science to find frequent subsequences in a sequential database using effective search algorithms. Recall that given a sequential database $\mathcal{S}$, $\mathcal{I}$ defines the set of unique items in the dataset. In substantially larger databases than ours, it is unfeasible to count the frequency of all possible subsequences and this necessitates effective search algorithms. We mention PrefixSpan [3] and CM-Spam [1] which we tested and used from the Java-library SPMF [2].

Table 7.4 establishes that $|\mathcal{S}| = 14833$ (including empty sequence) and $|\mathcal{I}| = 16971$ through a total of 14833 sequences with an average length of approximately 3. To filter these sequence, we need to define when a subsequence is frequent. The frequency (or support) of a subsequence s_a is defined by

$$f(s_a) = |\{s \in \mathcal{S} \mid s_a \sqsubseteq s\}|, \tag{7.2}$$

where $\sqsubseteq$ is defined in Section 5.1. We introduce a hyper-parameter $\mathcal{P}$ called the minimum support, which defines a threshold. A subsequence s_a is *frequent* if $f(s_a) \geq \mathcal{P}$.

Note that treatment package is an item in the terminology of Section 5.1. Thus in particular for each treatment package t, the above definition is equivalent to

$$f(t) = \frac{|\{s \in \mathcal{S} \mid t \in s\}|}{|\mathcal{S}|},$$

e.g. the number of sequences that t occurs in. This definition aligns with the above as $f(\{t\}) = f(t)$. Observe that a subsequence s_a of length 2 can only be frequent if each item (treatment package) in it is frequent itself.

This definition of frequent is analog to the cut-off level from the Skip-Gram algorithm used in Chapter 6, though we use the frequency acrosss sequences instead of entries.

Parallels to Markov Chains

Sequential Pattern Mining has a direct connection to Markov Chains (MC). Firstly, if the itemsets are of size one, this corresponds exactly to studying Markov Chains as itemsets can be identified with states. In this special case, frequent itemsets in SPM corresponds to subsets of the state space in the Markov Chain, for which it holds that there exists a path through the states and this path occurs in at least $\mathcal{P}$ percent of the observed sequences.

Secondly, for the general case, SPM always corresponds to a MC with the following state space M

$$M = \{0,1\}^{|\mathcal{I}|}$$

and the itemset i can be written as the state i, given by

$$i = \underbrace{(0,1,0,0,\ldots,0,1,0,1)}_{|\mathcal{I}| \text{ entries}}$$

where 1 in entry j of state i corresponds to item j being in the itemset i. Technically it is required of the state space M that at least one coordinate is 1, though the NO-MED token in next chapter could correspond to the null-state (all zeros). This leads to a large state space of the MC as a given itemset may contain all unique items $\mathcal{I}$. The transition matrix of this MC is poorly estimated due to the state space size. But we may still search for frequent subsequence, which does not rely on the transition matrix. With this connection established, write a candidate subsequence as $s = (Y_1, Y_2, \ldots, Y_n)$. This candidate subsequence may be formulated as a sequence of states in the MC. It occurs in the Markov Chain if there exists a sequence $\tilde{s} = (X_1, X_2, \ldots, X_m)$ of states X_1 such that

$$\exists m_1 < m_2 < \ldots < m_n \colon \forall i \forall k \in Y_i \exists X_{m_i} \colon 1 = (X_{m_i})_k, \tag{7.3}$$

where $(X_{m_i})_k$ denotes the kth entry of the state X_{m_i}. Observe that this formulation is parallel to the subsequence formulation in subsection 5.1.

Filtering using SPM

We proceed to filter the sequences by removing treatment packages which occurs below the threshold $\mathcal{P}$ and collecting the resulting sequences which consist of only frequent treatment packages – we have thus modified the sequences. Note that frequent subsequences of length 2 or greater are not used for filtering. The resulting dataset is described in Table 7.5. Note that we have removed the empty sequences from the dataset, but we mention them as they are a substantial part of the original dataset. Moreover, note that the number of unique treatment package correspond to subsequences of length 1. Furthermore, this dataset still includes sequence of length 1, which cannot be used for Markov Chains (as we need k prior state(s) for a k-order Markov Chain and $k \geq 1$).

Level	Variable	Amount
Dataset	No. of sequences	7475
	No. of empty sequences	0
Sequence	Shortest sequence	1
	Longest sequence	9
	Average sequence length	2.04
	Median sequence length	2
	No. of unique treatment packages	31
	No of unique subsequences with length > 1	6

Table 7.5: Filtered merged dataset statistics.

7.3.2 Markov Chains

The central question given a sepsis patient is the treatment/medication strategy. We used a Markov Chain to predict the next medication, given a prior medication, on the filtered dataset from Section 7.3.1. Due to the average length of the sequences, it is only feasible to use a first-order Markov Chain. We had to filter away sequences of length 1, in order to obtain an initial state. This results in dropping another 3306 sequence, ultimately yielding the dataset described in Table 7.6. For the theoretical foundations of Markov Chains, we

Level	Variable	Amount
Dataset	No. of sequences	4169
Sequence	Shortest sequence	2
	Longest sequence	9
	Average sequence length	2.87
	Median sequence length	2

Table 7.6: Markov adapted merged dataset statistics.

refer to [4], but in general we mention that the transition probabilities are estimated by the observed frequencies. We chose to give longer sequences more weight, by counting frequencies by number of occurrences in total across all sequences, divided by the total number of occurrences. We did this to obtain more diversified sequences and give weight to longer sequences, where more clinical decision information may be contained.

7.4 Results

In this section we describe the results of our analysis. In Subsection 7.4.1, we analyze whether the alert system alters the treatment strategy for sepsis

patients. We test whether the inclusion of a *NO-MED* token in empty hour slots may lead to better analysis of medications gaps in Subsection 7.4.2. Finally, in the Subsection 7.4.3 we test whether a Markov Chain can be used to predict the next medication.

7.4.1 Graph analysis for control and active

In this subsection, we analyze the differences in the treatment packages between the two groups of the first dataset. Table 7.2 and Table 7.3 describe summary statistics of the dataset. We observe that these do not differ significantly from each other. The 9 most common medications (treatment packages of size 1) are plotted with their frequency for the active group in Figure 7.1 and the control group in Figure 7.2, where the group 0 is the accumulated frequency of all other remaining medications (among the frequent medications). These plots shows that the frequent medications for both groups are very similar and the only moderate difference is the cumulative frequency of the remaining states. As we note later, this encouraged us to study semi-rare treatment packages, as the common states are similar. Further comparison of

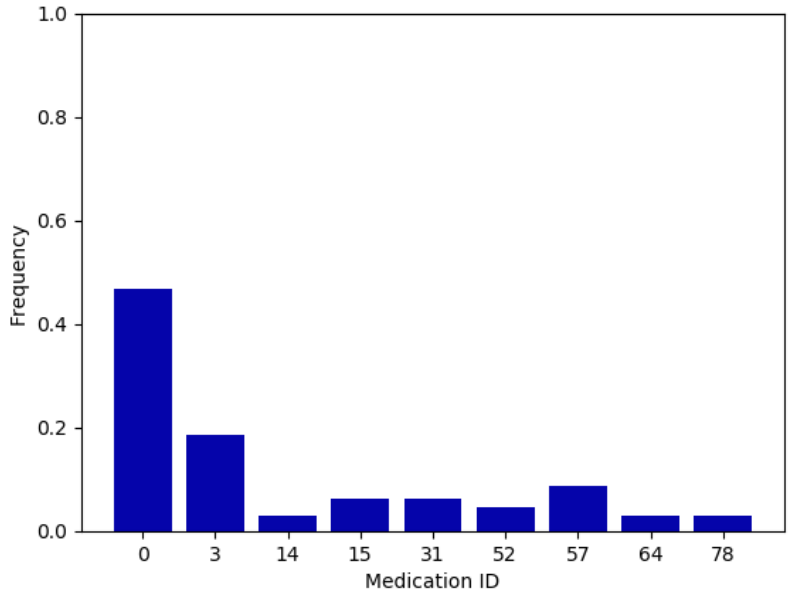


Figure 7.1: Barplot for 9 most frequent medications of the active group. The value 0 correspond to the cumulative frequency of the other medications.

nodes, edges, frequency of edges in the graph revealed no major differences between the active and control group. Finally, we studied the graph created by the transition matrix of both groups. A visualization is shown in Figure 7.3 based on the merged dataset, but the graphs were too sparse to make sense for the two groups in the first dataset. Each node (red dot) represented a treatment package and the node sizes in Figure 7.3 are scaled with the inverse frequency

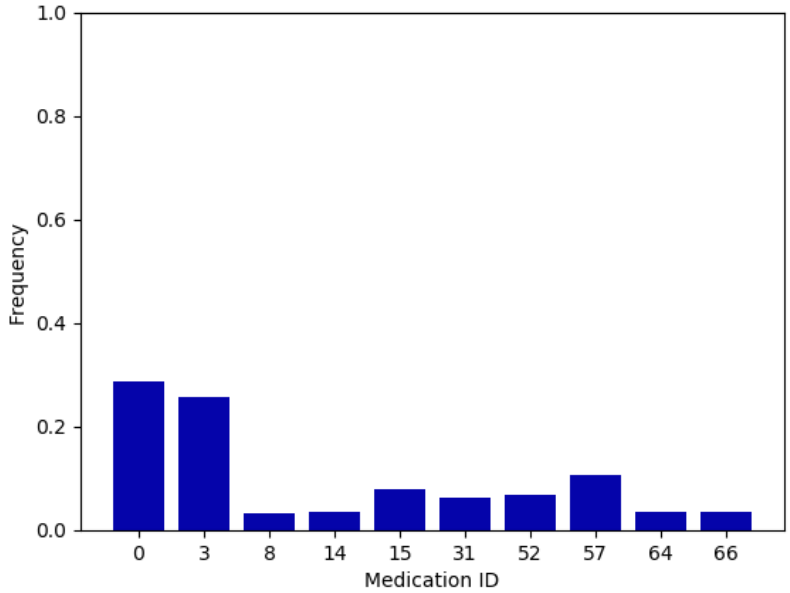


Figure 7.2: Barplot for 9 most frequent medications of the control group. The value 0 correspond to the cumulative frequency of the other medications.

of the treatment packages. A line is drawn between two nodes if the transition from one to the other occurs more than once. The clinicians found this visualization particularly helpful and were interested in (semi-rare) subgraphs of the visualized graph. This led to exploration of sub-graphs – or in other words rare treatment package patterns which are used infrequently. We observed no major differences between the groups and we decided to merge the groups for further analyses.

7.4.2 No medication token

We included a *NO-MED* token in each hour slot of the 24 hour time window after an alert to research whether the number of *NO-MED* tokens could be an attribute separating the control and active group. The sample entry from Table 7.1 would thus become,

$$(\{NO-MED\},\{NO-MED\},\{143,37\},\{NO-MED\},\{NO-MED\},\{14\},\\ \{NO-MED\},\{NO-MED\},\{17\},\{NO-MED\},\ldots,\{NO-MED\}),$$

where the vector has 24 entries (one for each hour from the alert time.) The idea was that the number of *NO-MED* between entries could be a predictor. Unfortunately, it introduceleadd *NO-MED* as the most common entry and made the model increasingly hard to interpret. The inclusion seemed to overcomplicate the interpretation of both the frequent itemsets and the Markov Chain predictions tasks and thus we decided to exclude it. However, for a different model or larger dataset, it could be worthwhile to revisit this idea.

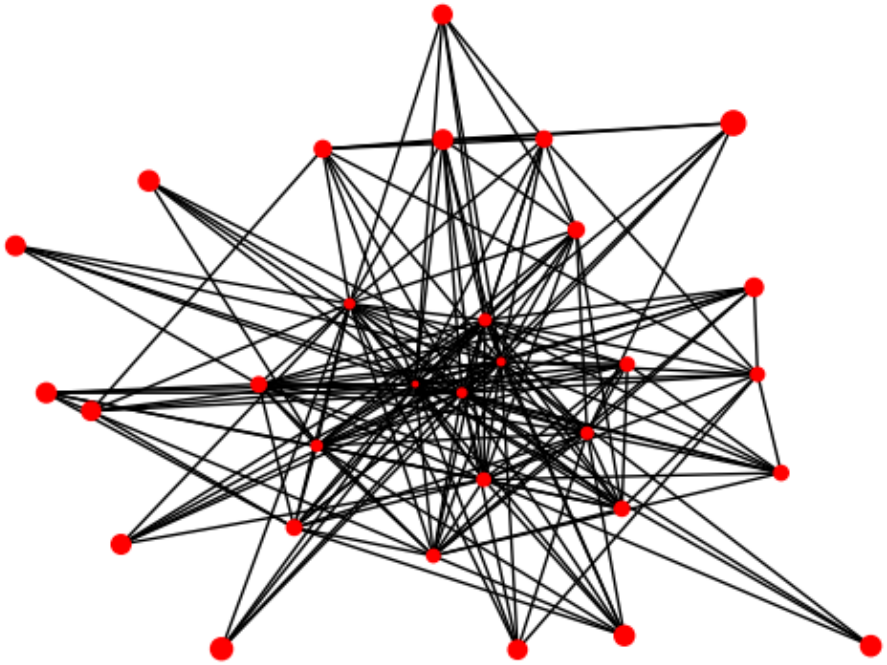


Figure 7.3: Network visualization with node size equal to inverted frequency of the treatment package.

7.4.3 Medication prediction

On the merged dataset, the goal was to predict the next medication and we model this using a Markov Chain on the merged filtered dataset adapted to a Markov Chain, described in Table 7.6. Note that the difference from the merged filtered dataset to the merged filtered dataset adapted to a Markov Chain is that we have removed sequences of length 1.

Given a prior even x, a Markov Chain estimates the conditional probability $\mathbb{P}(y \mid x)$ by the estimate $\hat{p}$, given by

$$\hat{p} = P_x$$

where P denotes the transition matrix of the Markov Chain, and P_x its xth row. Naturally, the prediction of the next state is simply

$$\hat{y} = \underset{j=1,2,\ldots,37}{\arg\max}\, P_x.$$

We split the dataset in Table 7.6 into a training and test set split of 80/20 percent. We obtain a test accuracy of 0.24, which appears reasonable as a Markov Chain is a simple model and we have a 31 possible states.

Dataset	Accuracy
Training set	0.24
Test set	0.24

Table 7.7: Markov Chain prediction accuracy.

7.5 Conclusion and perspectives

In this study, we initially aimed to analyze the effect of an alert system for sepsis as to whether it altered the treatment. The initial data cleaning procedure revealed major problems in the amount of data and we had to filter away many entries to obtain data suitable for analyzing the frequency of treatments and predicting the next medication using a Markov Chain. The results from our analysis points in the direction of no treatment altering effect, but we believe that a larger study may be warranted to study semi-rare subgraphs and discover alternate treatment pathways.

Later on, we received the second dataset, which was substantially larger than the first, but it did not include the group variable. This led us to focus on prediction of the next medication using a Markov Chain but we still needed to filter the dataset to obtain data suitable for statistical analysis. Most likely, a Markov Chain is not an optimal choice of a model. However, with extremely short sequences (average length around 3) it seemed infeasible to attempt more advanced models.

At this point, the main goal of the project shifted away from prediction and towards feature learning and visualization, which helps doctors to suggest further hypotheses using their domain-knowledge and helps the statistician select more clinically relevant problems and methodology.

In the future, we would like to study the application of a word embedding model to the problem, similar to Chapter 6, but we would need a larger dataset with much longer sequences. This could be obtained by including non-medical events from the the electronic health record of the patient. We would use the embedding model to transform the input from qualitative variables to numerical vectors in a lower-dimensional space (but retaining sufficient semantic meaning). We would then visualize these vectors using t-SNE and use/feed these transformed vectors (and thus transformed treatment sequence) as inputs into a recurrent neural network as we did in Chapter 6.

Another possibility is to include other patient attributes. One could hypothesize that the sub-graph of patients with a certain attribute (old, young, tall, short etc.) could lead to treatment patterns which are more frequent for the sub-group than the general cohort.

Chapter acknowledgements

The author would like to extend his warm thanks to Professor Daniel Rubin at Stanford University and Research Scientist Imon Banerjee at Stanford (now Assistant Professor at Emory University) for their hospitality and engaging discussions during his stay at Stanford University, Department of Biomedical Informatics.

References

[1] Philippe Fournier-Viger, Antonio Gomariz, Manuel Campos and Rincy Thomas. "Fast vertical mining of sequential patterns using co-occurrence information". *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2014, 40–52.

[2] Philippe Fournier-Viger, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Cheng-Wei Wu and Vincent S Tseng. "SPMF: a Java open-source pattern mining library". *The Journal of Machine Learning Research* 15.1 (2014), 3389–3393.

[3] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal and Mei-Chun Hsu. "Mining sequential patterns by pattern-growth: The prefixspan approach". *IEEE Transactions on knowledge and data engineering* 16.11 (2004), 1424–1440.

[4] Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.