DISSERTATIONES
MATHEMATICAE
UNIVERSITATIS
TARTUENSIS
136

JOONAS SOVA

Pairwise Markov Models

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
136

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
136

JOONAS SOVA

Pairwise Markov Models

UNIVERSITY OF TARTU
Press

Institute of Mathematics and Statistics, Faculty of Science and Technology, University of Tartu, Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in mathematical statistics on 18th of June, 2021, by the Council of the Institute of Mathematics and Statistics, Faculty of Science and Technology, University of Tartu.

Supervisor:

Prof. Jüri Lember
Institute of Mathematics and Statistics
University of Tartu
Estonia

Opponents:

Prof. Pavel Chigansky
Faculty of Social Sciences
Hebrew University of Jerusalem
Isreal

Prof. Evgeny Verbitskiy
Mathematical Institute
Leiden University
Netherlands

The defense will take place at 25th of August, 2021, at 14:15 in Narva mnt 18-1007, Tartu, Estonia.

ISSN 1024-4212
ISBN 978-9949-03-663-9 (print)
ISBN 978-9949-03-664-6 (pdf)

University of Tartu Press
http://www.tyk.ee

Contents

Acknowledgments

My deepest gratitude goes to my supervisor Jüri Lember. Without his immense dedication and passion for science, tireless work ethics and endless patience as a teacher this dissertation would have never been possible. I would also like to thank The Institute of Mathematics and Statistics of Tartu University for providing the funding that allowed to complete this work. Finally I would like to thank my family, and my colleagues both from academia and industry for their support and encouragement.

Publications

This dissertation is based on the following articles:

[I] J. Lember and J. Sova. "Existence of infinite Viterbi path for pairwise Markov models". *Stochastic Processes and their Applications* 130.3 (2020), pp. 1388–1425

[II] J. Lember and J. Sova. "Regenerativity of viterbi process for pairwise markov models". *Journal of Theoretical Probability* 34.1 (2021), pp. 1–33

[III] J. Lember and J. Sova. "Exponential forgetting of smoothing distributions for pairwise Markov models". *Electronic Journal of Probability* (2021), to appear

The author's contribution in all three papers is in working jointly with the co-author to develop the theory and write the text for publication. Publication of this dissertation has been supported by the Estonian Research Council grant PRG865.

Introduction

Markovian latent variable models are a great success story of modern statistics. Nowadays there is increasing prevalence of data where the classic assumptions of independence cannot be assumed, and so the classic statistical methodology fails to be effective. In contrast, the Markovian latent variable models have been shown to provide efficient and highly adaptable methodologies for dealing with various types of statistical problems related to complex and inter-dependent data. Here we explore a wide class of such models, namely the "pairwise Markov models" (PMM's). PMM is simply defined as a latent variable model where the latent or hidden layer and observed layer both constitute a Markov chain. As such, the PMM encompasses several well-known and widely applied models, like hidden Markov models, autoregressive switching models, hidden Markov models with dependent noise and many more.

The purpose of this thesis is to give an overview of the key results in the three papers listed above, all of which investigate certain aspects of the PMM. It is often the danger in papers of technical nature that the main driving ideas are overshadowed by the prevalence of technical minutiae. Here my main goal is to present the key ideas and results of the three papers as accessibly as possible, while "hiding" the more technical aspects of the proofs as well as some of the results which are less significant in terms of scientific contribution or novelty.

In Paper I the main subject of interest is the Viterbi path – the maximum likelihood estimate of the hidden layer of the PMM. The question of the stability of the Viterbi path is non-trivial, because adding a single observation to our sample can in theory change the whole path estimate. We study the asymptotic path-convergence of the Viterbi classifier on several levels of abstraction – starting from the general PMM up to examples of specific parametrized models. In particular, we prove that under general conditions it is possible to extend the Viterbi path estimate to infinity. This in turn enables to ensure the existence of the infinite Viterbi encoding of the observation sequence, or what is termed the "Viterbi process".

Paper II continues the study of the stability of the Viterbi classifier beyond the question of path-convergence itself. More specifically, we show that based on concepts developed in Paper I it is possible to construct a series of regeneration times for the PMM such that the Viterbi process depends on the observations up to each regeneration time only. This in turn enables to derive strong laws of large numbers and central limit theorems for the Viterbi classifier.

The subject matter of Paper III departs from the previous two papers and is no longer related to the Viterbi estimation. Rather, we study certain forgetting properties of the smoothing probabilities of the PMM. The main novelty here is the condition which ensures the exponential forgetting rate for the smoothing probabilities. In fact, we demonstrate through several examples how this condition is much more lenient than several other known conditions for similar purposes in the setting of finite hidden state space. Interestingly, this same condition is also prominent in Paper I for ensuring the existence of the Viterbi process, even though its application there is completely different.

The structure of the thesis is as follows: Section 1 gives some background information on the pairwise Markov models, including hidden Markov models; Section 2 introduces the notation and the theoretical framework that is used

throughout the thesis; Section 3 gives some background information on the Viterbi estimation and the Viterbi classifier; Section 4 introduces several necessary concepts and definitions that are used for the study of the stability of the Viterbi classifier in Papers I and II; and Sections 5-7 give the summaries of the key results in Papers I-III.

1 Pairwise Markov models

Hidden Markov models (HMM) have been called "one of the most successful statistical modeling ideas that have came up in the last fifty years" [1]. Indeed, the applications of such models in different fields have been so wide-ranging, that I will not attempt to list them here. The appeal of HMM's for data researchers can be attributed to several factors. On the one hand, based on the overall Markovian structure of the HMM, several estimation methods have been developed that suit the needs of various types of statistical problems. Examples of such methods are the Baum-Welch algorithm (also known as the Expectation Minimization or the EM-algorirthm), forward-backward algorithm, Viterbi algorithm, and so on. On the other hand the observed process of the HMM (without the latent layer) is not generally Markovian and can have a highly complex and rich dependence structure. Therefore, in today's data-driven world HMM's have become increasingly relevant as a data inference tool in situations where the observed process cannot be assumed to follow the assumptions of classic statistics such as independence or even the Markovian structure.

Simply put, an HMM can be viewed as a Markov chain with some "added noise". More rigorously, we can consider a Markov chain $Y = \{Y_k\}$ taking values in a finite state space. For each k the conditional distribution of the observation X_k given $Y_k = i$ is determined by a density f_i, called the *emission density*. Requiring that, for each k, X_k is conditionally independent of all other random variables given Y_k, the law of the observation process $X = \{X_k\}$ is now fully described by the distribution of Y_1, transition matrix of Y and densities f_i. For some sample size n, $(X_1, \ldots, X_n)$ is the observed sample, while the sequence $(Y_1, \ldots, Y_n)$ is latent or in other words hidden from the observer, hence the term "hidden Markov model". Data inference is done based on the observed sample $(X_1, \ldots, X_n)$ only. As stated, the sample $(X_1, \ldots, X_n)$ does not generally follow the i.i.d. law nor is it a Markov chain, and can have a rather complex dependence structure. However, conditionally given $(Y_1, \ldots, Y_n)$ the sample elements $X_1, \ldots, X_n$ are independent of each other. Moreover, assuming that the hidden chain follows a stationary distribution $\pi(i)$, the distribution of each X_k can be easily expressed as $P(X_k \in A) = \sum_i \pi(i) \int_A f_i(x)\,\mathrm{d}x$. (Here we assumed that that X_k are continuous random variables, but in the discrete case the integral would simply be replaced by a sum.) Thus when Y is stationary, then the marginal distributions of the X-process are simply given by mixtures of the densities f_i over the stationary distribution π.

It is also helpful to think of the HMM in terms of stochastic representation as follows. For each state i let $\xi_k(i)$ be random variables having density f_i, and assume that all $\xi_k(i)$ independent of each other and independent of the hidden chain Y. Then X_k can be defined through the stochastic equation

$$X_k = \xi_k(Y_k).$$

For comparison, we can consider a more general model

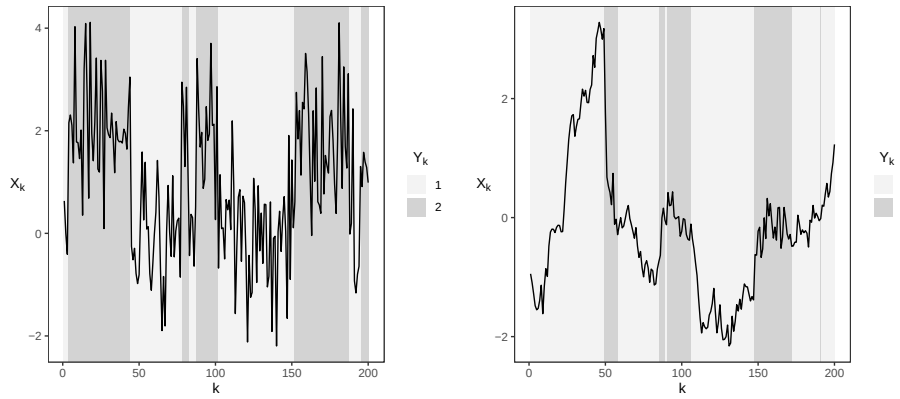
$$X_k = \alpha(Y_k)X_{k-1} + \xi_k(Y_k). \tag{1}$$

Where $\alpha(i)$ are constants. This type of model is called a "Markov switching model" or "autoregressive switching model" in literature [1, 2, 3, 4]. Here we refer to this model as "linear Markov switching model" owing to the linear link in (1). This model is different from HMM in that it allows conditional dependence between X_{k-1} and X_k given Y_k, and becomes HMM in the special case where $\alpha(i)$ are all zero. In both the HMM and linear Markov switching model, the variables $\xi_k(i)$ can be considered as the "random noise" in the sense that they contain all the randomness of the X-process beyond the hidden chain Y. In particular, when $\xi_k(i)$ are constants, then the X-process simply becomes the (non-random) encoding of the discrete Markov chain Y.

To investigate the difference of the two models further, I have generated a sample from both of them. Figure 1a displays 200 simulated observations from an HMM with a two-state hidden chain. The transition matrix of the hidden Markov chain is taken to be symmetric: the transition probability of maintaining the same state is 0.95 and switching the sate is 0.05. The emission densities are taken to be normal with both standard deviations equal to 1 and mean values for the hidden states 1 and 2 equal to 0 and 2, respectively. The hidden sates 1 and 2 are respectively indicated by the light gray and darker gray background color. As can be seen from the figure, the observations from the HMM inherit their overall discrete step-wise structure from the hidden chain. Cutting out the dark gray areas would result in an i.i.d. standard normal sample, and similarly cutting out the lighter areas would leave a i.i.d. normal sample with mean 2.

Figure 1a presents 200 observations from a two-state linear Markov switching model. The hidden chain is generated just like in the HMM-case with the 0.95 probability of maintaining the same state and 0.05 probability of switching the sate. The random noise $\xi_k(i)$ are taken to be normal with mean 0 and standard deviation $\frac{1}{4}$ for both states. The parameters $\alpha(i)$ are taken to be 1.01 and 0.5 for $i = 1$ and $i = 2$, respectively. (Here the model and its parameters are specified in a way that will ensure the overall stability of the observations in a certain sense. We will touch more on the stability conditions of the linear Markov switching model later.) As we can see, the behavior of the linear switching model is very different from that of the HMM. When the hidden state is 1, this model ensures that the observations have a tendency to autoregressively move away from 0, either to positive or negative direction. On the other hand, if the hidden state is 2, the observation start to move back towards the 0 value.

As we have seen, two different types of Markovian latent variable models can exhibit a very different behavior. However, these two models are only special cases of a much vaster class of latent variable models, namely the pairwise Markov models (PMM's). A PMM is simply any model such that the pairs of observations and hidden states constitute a Markov chain. If we follow the notation introduced above for the observation process $\{X_k\}$ and hidden process $\{Y_k\}$, then saying that the model is PMM simply means that $\{(X_k, Y_k)\}$ is a Markov chain. A PMM thus retains the general Markovian structure of both the HMM and the linear Markov switching model. This is a very useful property, since it means that many of the estimation tools used for HMM, such as the Viterbi algorithm or the EM-algorithm, can be implemented for this model as



(a) (b)

Figure 1: Simulations from an HMM (a) and linear Markov switching model (b)

well. On the other hand, unlike the HMM and the linear Markov switching model, for PMM the hidden process $\{Y_k\}$ is no longer necessarily a Markov chain. It can be shown, however, that conditionally given $\{X_k\}$, $\{Y_k\}$ is a non-homogeneous Markov chain, and vice versa.

The term "pairwise Markov chain" we have adopted from Pieczynski et al. who have used it in a series of papers to study such models [5, 6, 7, 8, 9]. We use this term to emphasize the much more general nature of this model when compared to a simple HMM. It should be noted, however, that this distinction is not always so clear in the scientific literature, and the term "hidden Markov model" is sometimes applied more generally than in the classic sense used here. In the next section we will introduce the exact theoretical framework and notation that applies to all three papers, as well as some definitions commonly used in the study of general-state Markov chains.

2 Theoretical framework and notation

As mentioned, we consider a two-dimensional Markov chain

$$\{(X_k, Y_k)\}_{k\geq 1} = ((X_1, Y_1), (X_2, Y_2), \ldots).$$

The process $X = \{X_k\}_{k\geq 1}$ is called the *observation process*, and its elements take values from the observation-space $\mathcal{X}$. We assume that $\mathcal{X}$ is Polish (separable completely metrizable) and equipped with its Borel σ-field $\mathcal{B}(\mathcal{X})$. The process $Y = \{X_k\}_{k\geq 1}$ is the *hidden* or *latent process*, and its elements take values from a finite state-space $\mathcal{Y} = \{1, 2, \ldots, |\mathcal{Y}|\}$. Both X and Y are defined on the probability space $(\Omega, \mathcal{F}, P)$

We denote $Z_k = (X_k, Y_k)$ for each k, $Z = \{Z_k\}_{k\geq 1}$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Thus the pairwise Markov chain Z is taking values from the product space $\mathcal{Z}$. We equip $\mathcal{Z}$ with product topology $\tau \times 2^{\mathcal{Y}}$, where τ denotes the topology of $\mathcal{X}$ and $2^{\mathcal{Y}}$ denotes the discrete topology on $\mathcal{Y}$. Furthermore, $\mathcal{Z}$ is equipped with its Borel σ-field $\mathcal{B}(\mathcal{Z}) = \mathcal{B}(\mathcal{X}) \otimes 2^{\mathcal{Y}}$, which is the smallest σ-field containing sets of the form $A \times B$, where $A \in \mathcal{B}(\mathcal{X})$ and $B \in 2^{\mathcal{Y}}$.

Let now μ be a σ-finite measure on $\mathcal{B}(\mathcal{X})$, and let c be the counting measure on $2^{\mathcal{Y}}$. We assume that the transition kernel of Z admits a density $q(z'|z)$ with respect to measure $\mu \times c$. This means that the transition kernel of Z expresses as

$$P(Z_2 \in C|Z_1 = z') = \int_C q(z|z')\, \mu \times c(\mathrm{d}z), \quad z' \in \mathcal{Z}, \quad C \in \mathcal{B}(\mathcal{Z}). \tag{2}$$

Here, the mapping

$$q\colon \mathcal{Z}^2 \to [0, \infty), \quad (z, z') \mapsto q(z|z')$$

is a measurable non-negative function such that for each $z' \in \mathcal{Z}$ the function $z \mapsto q(z|z')$ is a probability density function with respect to product measure $\mu \times c$. Since every $C \in \mathcal{B}(\mathcal{Z})$ is of the form $C = \cup_{j\in\mathcal{Y}} A_j \times \{j\}$ for some $A_j \in \mathcal{B}(\mathcal{X})$, then taking $(x', i) = z'$, (2) can be rewritten as

$$\begin{aligned} P(Z_2 \in C|Z_1 = z') &= \sum_{j\in\mathcal{Y}} P(X_2 \in A_j, Y_2 = j|X_1 = x', Y_1 = i) \\ &= \sum_{j\in\mathcal{Y}} \int_{A_j} q(x, j|x', i)\, \mu(\mathrm{d}x). \end{aligned}$$

We also assume that Z_1 has density with respect to product measure $\mu \times c$. Then, for every n, the random vector $Z_{1:n}$ has a density with respect to the measure $(\mu \times c)^n$. For every vector $(a_1, \ldots, a_n)$ we shall adopt the notation $a_{1:n}$. With a slight abuse of notation the letter p will be used to denote the various joint and conditional densities. Thus $p(z_k) = p(x_k, y_k)$ is the density of Z_k determined at $z_k = (x_k, y_k)$, $p(z_{1:n}) = p(z_1) \prod_{k=2}^n q(z_k|z_{k-1})$ is the density of $Z_{1:n}$ determined at $z_{1:n}$, $p(z_{2:n}|z_1) = \prod_{k=2}^n q(z_k|z_{k-1})$ stands for the conditional density and so on. Sometimes it is convenient to use other symbols besides x_k, y_k, z_k as the arguments of some density; in that case we indicate the corresponding probability law using the equality sign, for example

$$p(x_{2:n}, y_{2:n}|x_1 = x, y_1 = i) = q(x_2, y_2|x, i) \prod_{k=3}^n q(x_k, y_k|x_{k-1}, y_{k-1}), \quad n \geq 3.$$

The notation $P_z(\cdot)$ will represent the probability measure, when the initial distribution of Z is the Dirac measure on $z \in \mathcal{Z}$ (i.e. $P_z(A) = P(A|Z_1 = z)$). Likewise, if $T = g(Z)$ for any measurable mapping $g\colon \mathcal{Z}^\infty \to \mathbb{R}$, then $\mathbb{E}_z[T]$ denotes the expectation of T conditioned on $\{Z_1 = z\}$. See [10] for more details on the construction of the probability space, above probability measures and the conditional expectation for the general state Markov chain.

Classification of pairwise Markov models Returning now to the HMM-case, we can see that PMM becomes HMM under the following conditions: if $p(y_2|x_1, y_1)$ does not depend on x_1, and $p(x_2|y_2, x_1, y_1)$ does not depend on neither x_1 nor y_1. Indeed, in that case, denoting

$$p_{ij} = p(y_2 = j|y_1 = i), \quad f_j(x) = p(x_2 = x|y_2 = j),$$

the transition kernel density factorizes into

$$\begin{aligned} q(x, j|x', i) &= p(x_2 = x|y_2 = j, x_1 = x', y_1 = i)p(y_2 = j|x_1 = x', y_1 = i) \\ &= p_{ij} f_j(x). \end{aligned}$$

Here the density functions f_j (with respect to μ) are again the emission densities, like they were introduced in the first subsection, and p_{ij} are transition probabilities of the hidden Markov chain Y. The dependence structure of the HMM is represented in Figure 2a. The arrows in the figure represent a simplest possible scheme by which the HMM can be generated. We can imagine that we have simulated a pair (X_k, Y_k) from our specified HMM, and we now need to simulate the next pair (X_{k+1}, Y_{k+1}). Since Y is a Markov chain, Y_{k+1} can be simulated based on the transition matrix (p_{ij}) and value of Y_k alone – this is represented in the graph by a single arrow from Y_k to Y_{k+1}. Likewise, once we have simulated the value of Y_{k+1}, the value of X_{k+1} can be simulated based on Y_{k+1} only, which is represented by another arrow from Y_{k+1} to X_{k+1}. Indeed, as we mentioned previously, the conditional distribution of X_{k+1} given $Y_{k+1} = i$ has the density f_i.

If $p(y_2|x_1, y_1)$ does not depend on x_1, and $p(x_2|y_2, x_1, y_1)$ does not depend on y_1, then we call Z a *Markov switching model*. Thus HMM's constitute a subclass of Markov switching models. In case of Markov switching model, denoting

$$f_j(x|x') = p(x_2 = x|y_2 = j, x_1 = x'),$$

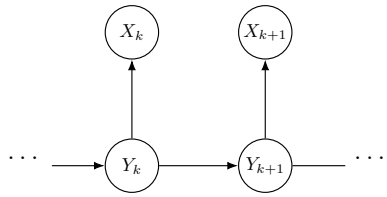
the transition kernel density becomes

$$q(x, j|x', i) = p_{ij} f_j(x|x').$$

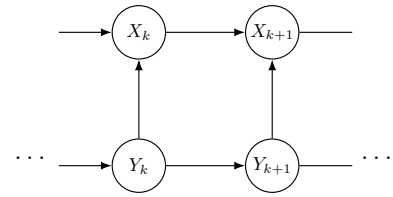
Again, it is not difficult to confirm that in for the Markov switching model, just like in case of HMM, Y is also a homogeneous Markov chain with transition matrix (p_{ij}). The linear Markov switching model (1) from which we simulated observations earlier is a special case of this type model with $f_j(x|x') = h_j(x - \alpha(j)x')$, where h_j denote the densities of the random noise variables $\xi_k(j)$ and μ is Lebesgue measure. In the case of Markov switching model, it is no longer possible to simulate the variable X_{k+1} based on the Y_{k+1} variable only. Indeed, to simulate X_{k+1}, the value of the previous observation X_k is also required. In particular, the conditional density of X_{k+1} given $X_k = x'$ and $Y_{k+1} = j$ is given by $f_j(\cdot|x')$. This dynamic is represented in Figure 2b, where there is an additional arrow pointing from X_k to X_{k+1} when compared to the HMM-case.

In the most general case of the PMM, however, the simulation scheme for the Markov switching model may not apply either anymore. Figure 2c shows one of the possible ways how Y_{k+1} and X_{k+1} can be simulated in this case. We can see that according to this scheme, Y_{k+1} is simulated based on X_k and Y_k. Then the next observation X_{k+1} is simulated based on all three variables Y_k, X_k and Y_{k+1}. An alternative approach would be to generate X_{k+1} before Y_{k+1}, in which case the arrow between X_{k+1} and Y_{k+1} would be flipped. In this general case, Y may no longer be a Markov chain. It is not difficult to see, however, that conditionally given $Y_{1:n}$, $X_{1:n}$ is always a (generally non-homogeneous) Markov chain and vice versa.

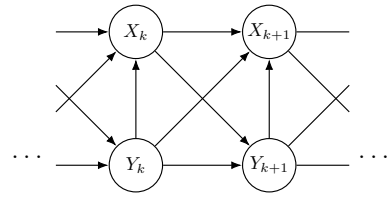
Harris chains In all three papers we construct some set of vectors $B \subseteq \mathcal{Z}^r$, and then for our proofs to work we need the Markov chain Z to return to B



(a) Hidden Markov model (HMM)



(b) Markov switching model



(c) Pairwise Markov model (PMM)

Figure 2: Directed dependence graphs of different types of PMM's

infinitely often a.s. In other words, we want the particular constructed set B to satisfy $P(Z \in B \text{ i.o.}) = 1$, where we denote

$$\{Z \in B \text{ i.o.}\} = \left\{ \bigcap_{k=1}^{\infty} \bigcup_{l=k}^{\infty} \{Z_{l:l+r-1} \in B\} \right\}.$$

Of course, if Z is ergodic in the sense of ergodic theory and $P(Z_{1:r} \in B) > 0$, then indeed $P(Z \in B \text{ i.o.}) = 1$ according to Birkhoff's ergodic theorem. However, the notion of ergodicity in case of general state space Markov chains is a rather abstract one, so we have relied on the theory of Harris recurrent Markov chains instead. The advantage of this theory is that it has a set of well-developed and powerful tools for deriving concrete yet general stability conditions for any specific model. We shall now introduce some key terms from the theory of Harris reccurent chains which will be used in the three papers. For a much more comprehensive overview of Harris chains see [10].

Markov chain Z is called *φ-irreducible* for some σ-finite measure φ on $\mathcal{B}(\mathcal{Z})$, if $\varphi(A) > 0$ implies $\sum_{k=2}^{\infty} P_z(Z_k \in A) > 0$ for all $z \in \mathcal{Z}$. If Z is φ-irreducible, then there exists [10, Prop. 4.2.2.] a *maximal irreducibility measure* ψ in the sense that for any other irreducibility measure φ the measure ψ dominates φ, $\psi \succ \varphi$. The symbol ψ will be reserved to denote the maximal irreducibility measure of Z. Chain Z is called *Harris recurrent* when it is ψ-irreducible and $\psi(A) > 0$ implies $P_z(Z_k \in A \text{ i.o.}) = 1$ for all $z \in \mathcal{Z}$. Note that if Z is Harris recurrent, then Z returns infinitely often a.s. to any set $A \in \mathcal{B}(\mathcal{Z})$ satisfying $\psi(A) > 0$. However, our goal was to characterize the infinite recurrence of vector sets, not single-element sets. Indeed, there are some technical details that need to be worked out before one can deduce from Harris recurrence the recurrence of some vector set. This has essentially been done in the proof of [I, Prop. 2.2] which links the infinite recurrence of single-element sets to that of vector sets. Similarly, [III, Prop. 2.1] characterizes the ψ-irreducibility and Harris recurrence of the overlapping r-block Markov chain $(Z_{1:r}, Z_{2:r+1}, \ldots)$ through the ψ-irreducibility and Harris recurrence of Z itself.

We have demonstrated in Figure 1a how the behavior of HMM is largely governed by its discrete hidden chain Y. Thus, it is not surprising that any HMM with irreducible hidden chain is ψ-irreducible and Harris recurrent. Here the maximal irreducibility measure ψ is defined by

$$\psi(A \times \{i\}) = \mu(A \cap G_i), \quad A \in \mathcal{B}(\mathcal{X}), \quad i \in \mathcal{Y},$$

where we denote $G_i = \{x \in \mathcal{X} \mid f_i(x) > 0\}$. This is a rather trivial consequence of the theory of Harris recurrent Markov chains, but the formal proof is given under [I, Lem. A.2]. For more complex models proving Harris recurrence might be more difficult and the conditions need to be derived for each model separately. For example, it can be shown that for the linear Markov switching model (1) with Gaussian noise the sufficient conditions for Harris recurrence are that Y is irreducible and $\max_{i \in \mathcal{Y}} \sum_{j \in \mathcal{Y}} p_{ij} |\alpha(j)| < 1$ (see [I, Lem. 4.3]). Thus the model used for the simulations in Figure 1b is Harris recurrent, because in this case

$$\max_{i \in \{1,2\}} \sum_{j \in \{1,2\}} p_{ij} |\alpha(j)| = 0.95 \cdot 1.01 + 0.05 \cdot 0.5 = 0.9845.$$

The maximal irreducibility measure for model (1) with Gaussian noise is $\mu \times c$, where μ is Lebesgue measure. Therefore, for the particular model in the example X will enter any interval infinitely many times with probability one.

In the next two sections we shall introduce the Viterbi classifier and all the related concepts that are used in Papers I and II. This content is not relevant to Paper III which does not deal with the Viterbi estimation.

3 Viterbi classifier

In many practical applications of an HMM or PMM, the goal of the data analysis is to estimate the hidden path $Y_{1:n}$ based on the observations $X_{1:n}(\omega) = x_{1:n}$. This is referred to as the *segmentation problem*. The most popular estimate is probably the path with maximum likelihood, defined by

$$v(x_{1:n}) = \arg\max_{y_{1:n}} p(y_{1:n}, x_{1:n}).$$

The mapping $v \colon \mathcal{X}^n \to \mathcal{Y}^n$ is called the Viterbi classifier and the estimate $v(X_{1:n})$ is referred to as Viterbi path or alignment (also maximum a posteriori path or alignment).

The Viterbi classifier maximizes the probability of estimating the whole hidden sequence correctly, that is

$$P(Y_{1:n} = v(X_{1:n})) = \sup_g P(Y_{1:n} = g(X_{1:n})), \tag{3}$$

where the supremum is taken over all measurable mappings of the form $g \colon \mathcal{X}^n \to \mathcal{Y}^n$. Indeed, for any same-length vectors a and b, let $\mathbb{I}(a = b)$ denote 1 if $a = b$ and 0 otherwise. For any classifier g we have

$$\begin{aligned} P(Y_{1:n} = v(X_{1:n})) &= \int_{\mathcal{X}^n} \sum_{y_{1:n}} \mathbb{I}(y_{1:n} = v(x_{1:n})) \cdot p(y_{1:n}, x_{1:n}) \, \mu^n(\mathrm{d}x_{1:n}) \\ &\geq \int_{\mathcal{X}^n} \sum_{y_{1:n}} \mathbb{I}(y_{1:n} = g(x_{1:n})) \cdot p(y_{1:n}, x_{1:n}) \, \mu^n(\mathrm{d}x_{1:n}) \\ &= P(Y_{1:n} = g(X_{1:n})). \end{aligned}$$

Viterbi algorithm The notion of the Viterbi path would not have any practical relevance, if there was no way to calculate it from the observed data. Note that the number of possible hidden paths is $|\mathcal{Y}|^n$, a quantity that grows exponentially in n. Therefore it is not possible to apply the brute force algorithm to directly calculate the Viterbi path for any reasonably sized sample. Luckily, there is a well-known dynamic programming algorithm – called the *Viterbi algorithm* – which calculates the Viterbi path in linear time with respect to n. This algorithm utilizes the Markov property in a rather simple and straightforward way to obtain the path with maximum likelihood. In its standard form the algorithm moves from the beginning of the observed sequence to the end, calculating for each state j and each position k the maximum possible likelihood up to k while also remembering the state that leads to the maximum likelihood value. Then the algorithm backtracks from the end to the beginning again to construct the Viterbi path based on the memorized states.

More formally, at $k = 1$ the algorithm calculates the values $\delta_1(j) = p(x_1, y_1 = j)$, and then for each $k = 2, \ldots, n$ and $j \in \mathcal{Y}$ it calculates

$$\delta_k(j) = \max_{i \in \mathcal{Y}} \delta_{k-1}(i) q(x_k, j | x_{k-1}, i), \tag{4}$$

$$\gamma_k(j) = \arg\max_{i \in \mathcal{Y}} \delta_{k-1}(i) q(x_k, j | x_{k-1}, i). \tag{5}$$

Thus $\delta_k(j) = \delta_{k-1}(\gamma_k(j)) q(x_k, j | x_{k-1}, \gamma_k(j))$. Once the final $\delta_n(j)$ and $\gamma_n(j)$ have been calculated, the maximum likelihood is given by $\max_{j \in \mathcal{Y}} \delta_n(j)$ and the Viterbi path $(v_1, \ldots, v_n)$ can be constructed by backtracking as follows:

$$\begin{aligned}
v_n &= \arg\max_{j \in \mathcal{Y}} \delta_n(j), \\
v_{n-1} &= \gamma_n(v_n), \\
v_{n-2} &= \gamma_{n-1}(v_{n-1}), \\
&\vdots \\
v_1 &= \gamma_2(v_2).
\end{aligned}$$

Note that the algorithm only relies on the Markov property of Z, and therefore can be utilized with any PMM regardless of the specific model.

There might be many paths which achieve the maximal likelihood, so the Viterbi path is not necessarily unique. Note, however, that if each application of the arg max function is based on some fixed ordering on $\mathcal{Y}$, then the Viterbi algorithm chooses the colexicographically maximal path based on the same ordering. That is, if there are several paths achieving the maximum likelihood, the algorithm will choose the colexicographically first one. In the above algorithm, the procedure runs from the first index to the last, and then returns to the first one. Alternatively, one can reverse the algorithm, so that procedure starts and ends with the last index n. In that case the natural tie-breaking scheme will not be colexicographical any more, but lexicographical. The reversed algorithm is essentially symmetrical to the one above, but its precise description is given for the sake of completeness in Appendix A.

Decision theoretic analysis To gain a better understanding of the Viterbi classifier, it is also useful investigate it from the viewpoint of decision theory. In

that framework a loss function is assigned, which penalizes the estimate based on how badly it misses the correct value of the estimand. The risk function is then defined as the expected value of the loss function, and the best classifier with respect to the specified loss function is the one that minimizes its risk. Thus, the choice of the loss function determines the optimal estimate in terms of its risk. For the Viterbi classifier, the loss of classifying the true path $y_{1:n}$ as $y'_{1:n}$ is simply defined as

$$L_v(y_{1:n}, y'_{1:n}) = \mathbb{I}(y_{1:n} \neq y'_{1:n}),$$

where $\mathbb{I}(a \neq b)$ denotes 1 if $a \neq b$ and 0 otherwise. In other words the loss is 0 only if it is absolutely correct and 1 otherwise. Then, equivalently to (3), the Viterbi classifier achieves the smallest risk over all classifiers:

$$\begin{aligned}\mathbb{E}L_v(Y_{1:n}, v(X_{1:n})) &= \inf_g \mathbb{E}L_v(Y_{1:n}, g(X_{1:n})) \\ &= 1 - \sup_g P(Y_{1:n} = g(X_{1:n})).\end{aligned}$$

The loss function L_v could be criticized for being overly absolutist. For example, the estimate which is different from the true path in only one position but correct in all the other ones would be seen as a very good one by anyone, but for the loss function L_v it is as bad as getting all positions wrong. From that perspective, the loss function that better conforms to practical reality is the *pointwise loss* defined by

$$L_p(y_{1:n}, y'_{1:n}) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}(y_k \neq y'_k). \tag{6}$$

This function simply counts the average number of misclassified positions, assigning loss of 1 only when all positions are misclassified and loss of 0 if none are. The classifier that has the minimal expected pointwise loss (i.e. pointwise risk), is called the *pointwise a posteriori* (PMAP) classifier. The PMAP classifier determines each position of the whole path individually, so that at each position the probability of having the correct state is maximal. However, because the PMAP classifier is only concerned with each position locally, it is susceptible of producing path estimates with very small overall probability, or indeed with zero probability. In terms of their loss functions, the Viterbi classifier is the polar opposite of the PMAP classifier: the former being concerned only with the whole path globally while the latter is only concerned with each state locally. It is possible to compromise between these two estimators by maximizing the probabilities of correctly classifying all pairs, all triplets, etc. See [11, 12] for the description of the dynamic programming algorithms for different types of classifiers and their risk-based analysis.

Despite their potential drawbacks, the Viterbi and PMAP classifiers remain the overwhelming favorites for estimating the hidden path among data analysts. This popularity can largely be attributed to the simplicity, intuitive appeal and ease of implementation of both estimators.

4 Infinite Viterbi path

To get a better understanding of the Viterbi classifier, it is useful to study its behavior when the sample size n goes to infinity. For example, in the previous

section we criticized its loss function for assigning constant loss regardless of the number of misclassified states. However, this does not necessarily imply that the Viterbi classifier is bad at estimating states at single positions. In reality, its behavior will vary from model to model, and in some instances its pointwise risk may be close to the optimal one achieved by the PMAP classifier. To investigate this further, we would like to study the limit

$$\lim_n L_p(Y_{1:n}, v(X_{1:n})) \tag{7}$$

where L_p is the pointwise loss function defined in (6). If such a limit exists and is a constant, it would quantify the overall pointwise misclassification rate of the Viterbi classifier. This rate could then be compared to the analogous rate of the PMAP classifier – again, assuming that it exists –, and would thereby give a sense of how far the Viterbi classifier is from the optimal PMAP classifier in terms of its ability to correctly classify states at individual positions. The limit (7) would depend on the specific model, but could be estimated through simulations for each model. It turns out that such a limit does exist a.s. under general conditions (in particular, it is implied by [II, Th. 4]), but it takes quite a lot of preparatory work to arrive to that point.

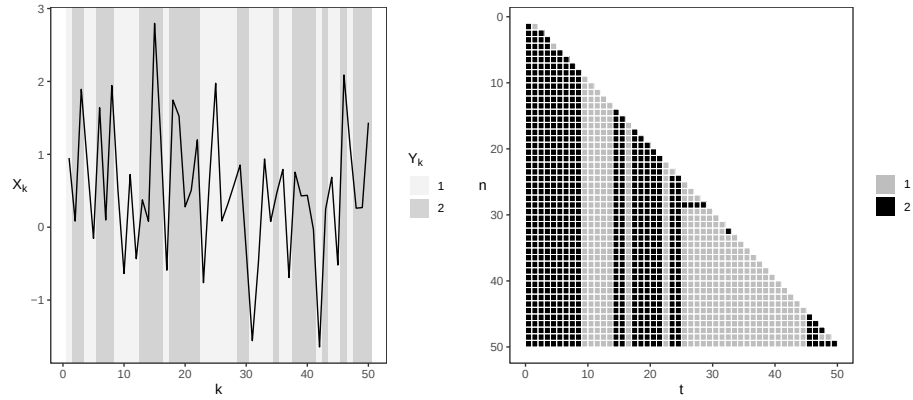
Indeed, the asymptotic study of the Viterbi path is non-trivial by the fact that adding a single observation to our sample can theoretically change the path estimate at any position. More formally, it is not necessarily the case that $v(x_{1:n})$ is the same vector than the n first elements of $v(x_{1:n+1})$. Intuitively, adding a single element to the end of our observation sequence should not affect the front part of our path estimate in any significant way, and if it does, this would generally be viewed as a pathological behavior of the model. Fortunately, in practice such pathological behavior usually does not occur, and the front part of the Viterbi path stabilizes rather quickly as the size of n grows. To illustrate this phenomenon, I have simulated 50 observations from an HMM. The transition matrix for the hidden Markov chain Y was taken to be symmetric, with 0.6 probability of maintaining the same state and with 0.4 probability of switching the state. The emission densities were taken to be normal with both standard deviations equal to 1 and mean values for the states 1 and 2 equal to 0 and 1, respectively. The observations from the HMM along with the hidden states are displayed in Figure 3a. Figure 3b displayes the corresponding Viterbi paths $v(x_{1:n})$ over $n = 2, \ldots, 50$. We can see that while there are some fluctuations of the Viterbi estimates on some positions as n increases, those are all localized to the end part of the Viterbi paths. The remaining front part quickly stabilizes into a fixed pattern.

Thus there is empirical evidence to support the idea that for some models the first t elements of the Viterbi path will stay the same under any sufficiently large sample size n. If this is true of any t, then the infinite Viterbi path $v_{1:\infty}$ can be defined as follows. In the below definition $v(x_{1:n})_{1:t}$ are the first t elements of the n-elemental vector $v(x_{1:n})$.

Definition 1. *Let $x_{1:\infty}$ be a realization of X. The sequence $v_{1:\infty} \in \mathcal{Y}^\infty$ is called the* infinite Viterbi path of $x_{1:\infty}$ *if for any $t \geq 1$ there exists $m(t) \geq t$ such that*

$$v(x_{1:n})_{1:t} = v_{1:t}, \quad \forall n \geq m(t).$$

The goal of Paper I is to prove that under general conditions infinite Viterbi path exists for almost every realization of X. Conversely, there are models



(a) (b)

Figure 3: Simulations from an HMM (a) and corresponding Viterbi paths with increasing n (b)

where there is no such path for almost any realization of X. Below is a simple example of such a 2-state HMM.

Example 1. Consider a two-state HMM with emission densities being equal, $f_1 = f_2$. Let the transition matrix of the hidden chain be

$$\begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{pmatrix} \frac{1}{10} & \frac{9}{10} \\ \frac{8}{10} & \frac{2}{10} \end{pmatrix} \end{array}.$$

Note that because the emission densities are equal, then X and Y are independent. Thus $p(x_{1:n}, y_{1:n}) = p(x_{1:n})p(y_{1:n})$ and the Viterbi path is the one that maximizes the probability $p(y_{1:n})$. Let the initial distribution of the hidden chain be $(\frac{49}{100}, \frac{51}{100})$, so that there is a slightly higher probability that the chain will start with 2 rather than 1. Observe now that the Viterbi path for any sample size n will be either 1212... or 2121... Indeed, the probability of switching state is always larger than maintaining one, so the Viterbi path must always alternate between 1 and 2. We can express the likelihoods of both possible paths for $n \geq 3$ as

$$p(y_{1:n} = 1212...) = \begin{cases} \frac{49}{100} \cdot \frac{9}{10} \cdot \left(\frac{8}{10} \cdot \frac{9}{10}\right)^{(n-2)/2}, & \text{if } n \text{ is even} \\ \frac{49}{100} \cdot \left(\frac{8}{10} \cdot \frac{9}{10}\right)^{(n-1)/2}, & \text{if } n \text{ is odd} \end{cases}$$

and

$$p(y_{1:n} = 2121...) = \begin{cases} \frac{51}{100} \cdot \frac{8}{10} \cdot \left(\frac{8}{10} \cdot \frac{9}{10}\right)^{(n-2)/2}, & \text{if } n \text{ is even} \\ \frac{51}{100} \cdot \left(\frac{8}{10} \cdot \frac{9}{10}\right)^{(n-1)/2}, & \text{if } n \text{ is odd} \end{cases}.$$

Therefore, because $\frac{49}{100} \cdot \frac{9}{10} > \frac{51}{100} \cdot \frac{8}{10}$, the Viterbi paths will be 1212... and 2121... for the even and odd n, respectively. This shows that there is no infinite Viterbi path for any realization of X.

Note that if we had used here the stationary distribution $\pi(i)$ for the initial distribution, the example would not have worked quite as well, because by the equality $\pi(1)\frac{9}{10} = \pi(2)\frac{8}{10}$ the likelihoods for the paths 1212... and 2121... would have been equal for the even n. Further, because $\pi(2) > \pi(1)$, then for odd n the Viterbi path would always be 2121..., so in that case the existence of the infinite Viterbi path would have depended on the tie-breaking scheme of the Viterbi classifier. In particular, under colexicographic scheme induced by ordering $2 \succ 1$ the infinite Viterbi path would not exist, but under the reverse ordering $1 \succ 2$ it would.

In the above example X was independent of Y, which is clearly not a realistic scenario for data analysis. Below is a different HMM example where X does depend on Y and, furthermore, the initial distribution can be chosen to be stationary regardless of the tie-breaking scheme. In this example the hidden chain is also irreducible and aperiodic. It is known that any HMM with irreducible, stationary and aperiodic hidden chain is ergodic (see e.g. [13, 14, 15]), hence this example demonstrates how infinite Viterbi path may fail to exist even for models with very stable probabilistic behavior. This in turn further underlines the need for special theory for dealing with the long-run behavior of the Viterbi classifier.

Example 2. Consider a 4-state HMM with the observation space $\mathbb{R}$ and the following transition matrix for the hidden chain:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} \frac{3}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{pmatrix} \end{array}.$$

Take emission densities as follows: f_1 and f_2 are both uniform on the interval $[0, 1]$, f_3 is uniform on interval $[0, \frac{1}{4}]$ and f_4 is uniform on interval $[\frac{3}{4}, 1]$. Hence $f_1 = f_2 = \mathbb{I}_{[0,1]}$, $f_3 = 4 \cdot \mathbb{I}_{[0,\frac{1}{4}]}$ and $f_4 = 4 \cdot \mathbb{I}_{[\frac{3}{4},1]}$, where $\mathbb{I}_A$ denotes the indicator function on set A. For the sake of elegance, we set the initial distribution of Y to be the stationary distribution, which can be calculated to be $(\frac{4}{10}, \frac{4}{10}, \frac{1}{10}, \frac{1}{10})$. Hence most of the time the hidden chain will spend in state space $\{1, 2\}$, but occasionally it will make a detour to the space $\{3, 4\}$.

Note that moving from state 1 to 2 is only possible through state 3, and moving from state 2 to 1 is only possible through state 4. Also note that the Viterbi path will never go *through* states 3 and 4, because staying in either state 1 or 2 will always a greater likelihood. Indeed, for all $x_{1:3} \in [0, 1]^3$ and all $y_{1:3} \in \{(1, 1, 1), (2, 2, 2)\}$ we have $p(x_{2:3}, y_{2:3}|x_1, y_1) = \frac{9}{16}$, while on the other hand for all $x_{1:3} \in \mathbb{R}^3$ and for all $y_{1:3} \in \mathcal{Y} \times \{3, 4\} \times \mathcal{Y}$ we have $p(x_{2:3}, y_{2:3}|x_1, y_1) \leq 4 \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{2} < \frac{9}{16}$.

It is possible, however, for the last element of the Viterbi path to enter the state space $\{3, 4\}$. Indeed, note that the single-step likelihoods for transitioning

from states 1 and 2 express as

$$p_{1j}f_j(x) = \begin{cases} 1, & \text{if } j = 3 \text{ and } x \in [0, \frac{1}{4}] \\ \frac{3}{4}, & \text{if } j = 1 \text{ and } x \in [0, 1] \\ 0, & \text{otherwise} \end{cases}$$

and

$$p_{2j}f_j(x) = \begin{cases} 1, & \text{if } j = 4 \text{ and } x \in [\frac{3}{4}, 1] \\ \frac{3}{4}, & \text{if } j = 2 \text{ and } x \in [0, 1] \\ 0, & \text{otherwise} \end{cases},$$

and so this implies that the last element of the Viterbi path $v(x_{1:n})$ will be 3 if $x_n \in [0, \frac{1}{4}]$, and 4 if $x_n \in [\frac{3}{4}, 1]$.

In conclusion, assuming for the sake of concreteness a colexicographic ordering scheme based on the ordering $1 \succ 2 \succ \ldots$, we have that the whole Viterbi path expresses as

$$v(x_{1:n}) = \begin{cases} 11...13, & \text{if } x_n \in [0, \frac{1}{4}] \\ 22...24, & \text{if } x_n \in [\frac{3}{4}, 1] \\ 11...11, & \text{otherwise} \end{cases}.$$

Because almost every observation sequence $x_{1:\infty}$ goes through intervals $[0, \frac{1}{4}]$ and $[\frac{3}{4}, 1]$ infinitely many times, it follows that for any fixed position k and for increasing $n = k, k+1, \ldots$ the k^{th} element of the Viterbi path $v(x_{1:n})$ will alternate between 1 and 2 infinitely often. Thus the infinite Viterbi path does not exist for almost any realization of the observation process.

Nodes and barriers The above examples show that the infinite Viterbi path may not exist for every model. We shall now turn our attention to the other direction to try to understand when it does exist. For every $n \geq 2$ and $i, j \in \mathcal{Y}$ denote

$$p_{ij}(x_{1:n}) = \max_{y_{1:n}\colon y_1 = i, y_n = j} p(x_{2:n}, y_{2:n} | x_1, y_1). \tag{8}$$

Next, let us fix the observation sequence $x_{1:\infty}$ and denote for all $k \geq 1$ and $i \in \mathcal{Y}$

$$\delta_k(i) = \max_{y_{1:k}\colon y_k = i} p(x_{1:k}, y_{1:k}).$$

This notation is consistent with the same notation used in the description of the Viterbi algorithm in (4). According to its definition the existence of the infinite Viterbi path means that for every time t, there exists a time $m \geq t$ such that the first t elements of $v(x_{1:n})$ are fixed as soon as $n \geq m$. The following is a sufficient condition m to be such a time: for every two states $j, s \in \mathcal{Y}$ and for some $i \in \mathcal{Y}$,

$$\delta_t(i)p_{ij}(x_{t:m}) \geq \delta_t(s)p_{sj}(x_{t:m}), \tag{9}$$

Indeed, there might be several states besides i satisfying (9), but the ties can always be broken in favor of the Viterbi path passing state i at position t, so

that

$$v(x_{1:n})_{1:t} = \underset{y_{1:t}\,:\,y_t=i}{\arg\max}\, p(x_{1:t}, y_{1:t}), \quad \forall n \geq m. \tag{10}$$

In other words, the ties can be always be broken so that the first t elements of the Viterbi path remain constant for every sample size $n \geq m$. However, the tie-breaking scheme that achieves this is not generally (co)lexicographic[1] – the natural tie-breaking scheme of the Viterbi algorithm. Therefore it is more practical to consider the following slightly strengthened version of the above condition: for every two states $j, s \in \mathcal{Y}$ the inequality (9) is strict for any j and $s \neq i$ for which the left side of the inequality is positive. This latter condition will ensure that (10) always holds under any (co)lexigographic ordering scheme. These observations have been combined into

Definition 2. *Let $x_{1:m}$ be a vector of observations. If inequalities (9) hold for any pair of states j and s, then the time t is called an* i-node of order $r = m - t$. *Time t is called a* strong i-node of order r, *if it is an i-node of order r, and the inequality (9) is strict for any j and $s \neq i$ for which the left side of the inequality is positive. We call t a node of order r if for some i, it is an i-node of order r.*

Suppose now that there exists an infinite sequence of i-nodes $u_1 < u_2 < \cdots$. We call two nodes u_{k-1} and u_k *separated*, if $u_k \geq u_{k-1} + r$. If the nodes u_{k-1} and u_k are not separated or not strong, then it might not be possible to break the ties in favor of i at both nodes – see [II, Ex. 4]. However, since from an unseparated sequence of nodes it is always possible to pick a separated subsequence, then there is no loss of generality in assuming that $u_1 < u_2 < \cdots$ are all separated, and in that case the infinite Viterbi path can be constructed *piecewise* as follows. Take

$$v_{1:u_1} = \underset{y_{1:u_1}\,:\,y_{u_1}=i}{\arg\max}\, p(x_{1:u_1}, y_{1:u_1})$$

and for all $k \geq 2$ take

$$v_{u_{k-1}:u_k} = \underset{y_{u_{k-1}:u_k}\,:\,y_{u_k-1}=y_{u_k}=i}{\arg\max}\, p(x_{u_{k-1}+1:u_k}, y_{u_{k-1}+1:u_1} | x_{u_{k-1}}, y_{u_{k-1}}).$$

Denote now $u(n) = \max\{u_k \leq n - r \mid k \geq 1\}$ and define the Viterbi path up to sample size n as

$$(v_{1:u(n)}, \underset{y_{u(n)+1:n}}{\arg\max}\, p(x_{u(n)+1:n}, y_{u(n)+1:n} | x_{u(n)}, y_{u(n)} = i)). \tag{11}$$

By the assumption that the nodes $u_1 < u_2 < \cdots$ are all separated and by the definition of a node, this path is well-defined as the one with maximal likelihood. Since for any n the first $u(n)$ elements of the Viterbi path are $v_{1:u(n)}$, and since $\lim_n u(n) = \infty$, it follows immediately that $v_{1:\infty}$ is the infinite Viterbi path of $x_{1:\infty}$.

If the nodes $u_1 < u_2 < \cdots$ are not strong, then the path (11) will not necessarily be (co)lexicographically first among all the paths with maximal likelihood. Therefore the piecewise construction will not be generally in alignment

[1] Here and henceforth we use the adjective "(co)lexicographic" for an ordering which is either lexicographic or colexicographic.

with the natural ordering of the Viterbi algorithm. On the other hand, if the nodes $u_1 < u_2 < \cdots$ are all strong (not necessarily separated) and all arg max functions are based on a (co)lexicographic ordering scheme, then it can be easily verified that (11) is the (co)lexicographically first one among all paths with maximal likelihood based on the same ordering scheme. For that reason we would like to work with strong nodes only. Fortunately, the requirement for strong nodes vs. simply nodes does not turn out to be restrictive. Indeed, this should not be surprising, considering that the difference between them is merely in the strictness of the equality (9).

Whether a time t is a node of order r or not depends in general on the sequence $x_{1:t+r}$. Sometimes, however, there is some small block of observations that guarantees the existence of a node regardless of the other observations. The following example, which is based on [I, Ex. 3], illustrates this.

Example 3. Suppose that there exists a state $i \in \mathcal{Y}$ such that for any triplet $u, j, s \in \mathcal{Y}$

$$q(x_t, i|x_{t-1}, u)q(x_{t+1}, j|x_t, i) \geq q(x_t, s|x_{t-1}, u)q(x_{t+1}, j|x_t, s). \tag{12}$$

Then for all $j, s \in \mathcal{Y}$

$$\begin{aligned}\delta_t(i)q(x_{t+1}, j|x_t, i) &= \max_u \delta_{t-1}(u)q(x_t, i|x_{t-1}, u)q(x_{t+1}, j|x_t, i) \\ &\geq \max_u \delta_{t-1}(u)q(x_t, s|x_{t-1}, u)q(x_{t+1}, j|x_t, s) \\ &= \delta_t(s)q(x_{t+1}, j|x_t, s),\end{aligned}$$

and so t is an i-node of order 1. Whether (12) holds or not, depends on triplet (x_{t-1}, x_t, x_{t+1}). In case of HMM, (12) is equivalent to

$$p_{ui}f_i(x_t) \cdot p_{ij} \geq p_{us}f_s(x_t) \cdot p_{sj}. \tag{13}$$

For a more concrete example, consider a 2-state HMM with both emission densities f_1 and f_2 being some continuous densities which are positive on the interval $[0, 1]$ and zero everywhere else. Also, let the transition matrix of hidden chain be symmetric with probability of maintaining the state equal to $p \in (\frac{1}{2}, 1)$ and the probability of switching the state equal to $1 - p$. Taking for the sake of concreteness $i = 1$, the inequalities (13) hold for all $u, s, j \in \{1, 2\}$ and some $x = x_t \in [0, 1]$ if and only if

$$\max_{x \in [0,1]} \frac{f_1(x)}{f_2(x)} \geq \frac{p^2}{(1-p)^2}. \tag{14}$$

The inequality (14) imposes a rather strict requirement on the maximal ratio of the emission densities. For example, if $p = \frac{9}{10}$, it requires that this ratio must be at least as great as 81, which is quite extreme. It is therefore evident that the inequalities (12) do not necessarily yield general or useful conditions for specific models.

While using a sequence of three observations to obtain a node might not be the most fruitful approach, the general concept of a sequence of observations generating a node is still a useful one. This concept is captured in the following

Definition 3. *Given $i \in \mathcal{Y}$, $b_{1:M}$ is called a* (strong) i-barrier of order r and length M, *if for any $x_{1:\infty} \in \mathcal{X}^\infty$ and $m \geq M$ satisfying $x_{m-M+1:m} = b_{1:M}$, $m - r$ is a (strong) i-node of order r.*

Hence, if (12) holds, then the triplet (x_{t-1}, x_t, x_{t+1}) is an i-barrier of order 1 and length 3. The advantage of working with the concept of barriers rather than simply nodes is that it provides a straightforward mathematical avenue to ensure the existence of infinitely many nodes. Indeed, if the observation sequence contains infinitely many i-barriers of order r, then there must exist an infinite sequence of i-nodes of order r, and so the infinite Viterbi path must exist by the piecewise construction.

Viterbi process The notion of the infinite Viterbi path for a single realization of the observation process X can be naturally extended to the infinite Viterbi path of the process X itself. This extension is called the *Viterbi process*. Formally this process is defined as follows. Let $V = \{V_k\}_{k\geq 1}$ be some random process taking values from state space $\mathcal{Y}$. We assume that V is defined on the same probability space as Z, namely $(\Omega, \mathcal{F}, P)$.

Definition 4. *The process V is called the* Viterbi process *if there exists a set $\Omega' \in \mathcal{F}$ such that $P(\Omega') = 1$ and for all $\omega \in \Omega'$ the sequence $V(\omega)$ is the infinite Viterbi path of $X(\omega)$.*

For each $\omega \in \Omega$ the infinite Viterbi path of $X(\omega)$ is well-defined by Definition 1, if it exists. However, the set $\{\omega \in \Omega | V(\omega)$ is the infinite Viterbi path of $X(\omega)\}$ might not be a measurable one, so the above definition simply requires that the complement of this set must be within some event with zero probability. Suppose now that there exists a set $\mathcal{X}^* \subseteq \mathcal{X}^M$ such that each of its element is a strong i-barrier of order r, and $P(X \in \mathcal{X}^*$ i.o.$) = 1$. Let $\Omega' = \{X \in \mathcal{X}^*$ i.o.$\}$. Because the observation sequence $X(\omega)$ contains infinitely many strong i-nodes of order r for all $\omega \in \Omega'$, it is now straightforward to verify the existence of the Viterbi process V.

Indeed, formally for each $k \geq 1$ the random variable V_k can be defined as follows. Take $T(k) = \min\{m - r \,|\, X_{m-M+1:m} \in \mathcal{X}^*, m \geq k + r\}$. Thus $T(k) \geq k$, and we have by definition of a barrier that $T(k)(\omega)$ is a strong i-node of order r for all $\omega \in \Omega'$. For each k, the random variable V_k is then defined as the k^{th} element of the random-length vector $v(X_{1:T(k)+r})$. We already know from the piecewise construction of the Viterbi path that for all $\omega \in \Omega'$ the Viterbi path up to $T(n)(\omega)$ is fixed for all $n \geq T(k)(\omega) + r$, and so V_k is well-defined as the k^{th} element of the Viterbi process given that it is a measurable random variable.

To verify that V_k can indeed chosen to be measurable note that assuming (co)lexicographic ordering scheme based on some fixed ordering on $\mathcal{Y}$, we have that the mapping

$$x_{1:n} \mapsto \arg\max_{y_{1:n}} p(x_{1:n}, y_{1:n}) \tag{15}$$

is measurable for all $n \geq 1$. It follows that for any $i \in \mathcal{Y}$,

$$\{V_k = i\} \cap \Omega' = \cup_{n \geq k+r} \{V_k = i, T(k) + r = n\} \cap \Omega' \in \mathcal{F}.$$

Finally, to make sure that V_k is formally a mapping on the whole space Ω, define $V_k(\omega) = 1$ for all $\omega \in \Omega'^c$. Since then $\{V_k = i\} \cap \Omega'^c$ is equal to Ω'^c if $i = 1$ and

to $\emptyset$ otherwise, we have

$$\{V_k = i\} = \big(\{V_k = i\} \cap \Omega'\big) \cup \big(\{V_k = i\} \cap \Omega'^c\big) \in \mathcal{F}.$$

Thus V_k is a well-defined random variable.

When the barriers in $\mathcal{X}^*$ are not strong, then the construction of the Viterbi process is also possible, but is slightly more complicated. In that case the tie-breaking for the mapping $X_{1:n} \mapsto v(X_{1:n})$ is no longer fixed and will depend on the position of nodes prior to n and thereby also on the random sequence $X_{1:n}$. Even so, it is easy to show based on the piecewise construction of the infinite Viterbi path that the Viterbi process still exists. However, the corresponding ordering scheme will not be (co)lexicographic anymore and therefore will not align with the ordering of the Viterbi algorithm.

The goal of Paper I is to find practical and general conditions for the existence of the barrier set $\mathcal{X}^*$. In Paper II it is proved that under general conditions on the barrier set $\mathcal{X}^*$, there exist regeneration times for the Markov chain Z which are also nodes of fixed order for almost every realization of X. These regeneration times break the Z-process into i.i.d. cycles, and because up to each regeneration time the Viterbi path is fixed for sufficiently large sample size, then – as argued in Section 6 below – SLLN and CLT type results apply for the Viterbi classifier. Thus the Viterbi process does not only ensure the overall path-stability of the Viterbi estimation, but is also a useful tool for obtaining related asymptotic convergence properties.

History of the problem To the best of our knowledge, prior to Paper I the existence of Viterbi process has been proven for HMM's only. The first such results were obtained in 2002 by Caliebe and Rösler [16] (see also [17]) who essentially define the concept of nodes and prove the existence of infinitely many nodes under rather restrictive assumptions like (13). A much more general treatment of the HMM-case was given in 2010 by Lember and Koloydenko [18] who introduce the general definitions of nodes and barriers, as are given above, and prove the existence of the Viterbi process under broad conditions. The basic ideas behind the Paper I are the same as in [18], but applying these ideas to the general PMM is far from straightforward due to the much more complex nature of the model. Indeed, from our general barrier construction theorem in Paper I we were able to strengthen the HMM-result in [18]. This strengthened result will be presented in the next section along with the discussion of its conditions. In [19] Lember and Koloydenko also consider specifically the 2-state ergodic HMM, and prove that for such model the Viterbi process always exists if there is essential difference between the two emission densities.

In Papers I-III we are only considering the case when the state space $\mathcal{Y}$ is finite, because that is the most suitable and fruitful framework for our ideas. When the state space $\mathcal{Y}$ is continuous, then the existence infinite Viterbi path as defined here becomes too restrictive, so it is more generally defined in terms of convergence of the path estimate. In [20] Chigansky and Ritov study such convergence and prove the existence of the limiting Viterbi process under certain restrictive conditions, such as the log-concavity of the emission and transition densities. More recently, Whiteley et al. [21] also study the existence of such limiting process under a different set of conditions. The latter paper is more motivated by the computational aspects and the scalability of the Viterbi path approximation via parallelization.

5 Summary of Paper I

As mentioned, the goal of Paper I is to construct a set $\mathcal{X}^*$ which consists of barriers and satisfies $P(X \in \mathcal{X}^* \text{ i.o.}) = 1$. On the one hand we want the conditions which guarantee the existence of such a set be as general as possible, but on the other hand we would like them to be easily verifiable for specific models. However, because we are dealing with such a general framework, we were not able to derive simple conditions which are immediately applicable to any model. Rather, in Paper I we build a general theory which applies to all types of models, and then through examples demonstrate how this theory can be relatively easily applied to specific cases. This theory consists of two barrier construction theorems – Theorem 1 below which applies to any PMM, and Theorem 2 which applies to PMM's with lower semi-continuous transition densities.

The main intuition behind our theory can be easily illustrated by the following example. Suppose $\mathcal{X} = \mathbb{R}$ and there exists an interval $I \subset \mathbb{R}$ such that $q(x, 1|x', 1)(1-\epsilon) > q(x, j|x', i)$ for some $\epsilon > 0$, all $(i, j) \in \mathcal{Y}^2 \setminus \{(1, 1)\}$ and all $x, x' \in I$. Consider the set $A(k) = I^k = I \times \cdots \times I$ (k times). When a block from $A(k)$ occurs within the observation sequence, we have that the conditional likelihood of the hidden path going through only state 1 at the corresponding positions is greater than never visiting a 1 by a factor of $1/(1-\epsilon)^{k-1}$:

$$p(x_{2:k}, y_{2:k} = (1, \ldots, 1)|x_1, y_1 = 1)(1-\epsilon)^{k-1} > \max_{y_{1:k} \in (\mathcal{Y}\setminus\{1\})^k} p(x_{2:k}, y_{2:k}|x_1, y_1), \quad \forall x_{1:k} \in A(k).$$

Thus the intuition is that the Viterbi path must go through 1 when a block of observations from $A(k)$ occurs, if we take k sufficiently large. However, this intuition is not entirely correct, and no matter how large the value of k is, it is not necessarily the case that $A(k)$ is a barrier set. Nevertheless, the general idea of concatenating k cycles of the interval I is still a useful one, because it turns out that under general conditions it is possible to construct the "edges" for the set $A(k)$ which will ensure that the $A(k)$ is the "center part" of the barrier set.

More formally, for any set A consisting of vectors of length $n > 1$ we adopt the following notation:

$$A_{(k,l)} = \{a_{k:l} \mid a_{1:n} \in A\}, \quad 1 \le k \le l \le n,$$
$$A_{(k)} = \{a_k \mid a_{1:n} \in A\}, \quad 1 \le k \le n.$$

Thus $A_{(k)}$ is the image of the k-th projection on A. Recall the definition of $p_{ij}(\cdot)$ from (8), and for any $n \ge 2$, define

$$\mathcal{Y}^+(x_{1:n}) = \{(i, j) \mid p_{ij}(x_{1:n}) > 0\}. \tag{16}$$

Note that

$$\mathcal{Y}^+(x_{1:n})_{(1)} = \{i \mid \exists j(i) \text{ such that } p_{ij}(x_{1:n}) > 0\}$$
$$\text{and} \quad \mathcal{Y}^+(x_{1:n})_{(2)} = \{j \mid \exists i(j) \text{ such that } p_{ij}(x_{1:n}) > 0\}.$$

Also observe that if $i \in \mathcal{Y}^+(x_{1:n})_{(1)}$ and $j \in \mathcal{Y}^+(x_{1:n})_{(2)}$, then it is not necessarily the case that $(i, j) \in \mathcal{Y}^+(x_{1:n})$. Our first barrier set construction theorem can now be stated as follows.

Theorem 1 [I, Th. 2.1]. *Assume the following.*

(V1) There exists $N \geq 2$, $n_1 < \cdots < n_{2N+2}$, set $\mathcal{X}^* \subseteq \mathcal{X}^{n_{2N+2}}$ and $\epsilon > 0$ such for all $k = 1, \ldots, 2N$ and all $\mathbf{x} \in \mathcal{X}^*_{(n_k, n_{k+1})}$

$$p_{11}(\mathbf{x}) \geq p_{i1}(\mathbf{x}), \quad \forall i \in \mathcal{Y} \setminus \{1\}, \tag{17}$$
$$p_{11}(\mathbf{x}) \geq p_{1j}(\mathbf{x}), \quad \forall j \in \mathcal{Y} \setminus \{1\}, \tag{18}$$
$$p_{11}(\mathbf{x})(1-\epsilon) > p_{ij}(\mathbf{x}), \quad \forall i, j \in \mathcal{Y} \setminus \{1\}. \tag{19}$$

(V2) There exist constants $0 < \delta \leq \Delta < \infty$ such that

$$p_{ij}(\mathbf{x}) \leq \Delta, \quad \forall i, j \in \mathcal{Y}, \quad \forall \mathbf{x} \in \mathcal{X}^*_{(1,n_1)}, \quad \forall \mathbf{x} \in \mathcal{X}^*_{(n_{2N+1}, n_{2N+2})},$$
$$\mathcal{Y}^+(\mathbf{x}) \neq \emptyset \quad \textit{and} \quad p_{i1}(\mathbf{x}) \geq \delta, \quad \forall i \in \mathcal{Y}^+(\mathbf{x})_{(1)}, \quad \forall \mathbf{x} \in \mathcal{X}^*_{(1,n_1)},$$
$$\mathcal{Y}^+(\mathbf{x}) \neq \emptyset \quad \textit{and} \quad p_{1j}(\mathbf{x}) \geq \delta, \quad \forall j \in \mathcal{Y}^+(\mathbf{x})_{(2)}, \quad \forall \mathbf{x} \in \mathcal{X}^*_{(n_{2N+1}, n_{2N+2})}.$$

(V3) It holds

$$\frac{\Delta}{\delta}(1-\epsilon)^N < 1.$$

Then $\mathcal{X}^*$ consists of 1-barriers of order $n_{2N+2} - n_{N+1}$. Furthermore, if (V1) holds with either inequalities (17) or inequalities (18) being strict, then the barriers are strong.

We shall now make a few notes on the conditions of the theorem. First, here and elsewhere in this section we are for the sake of concreteness mostly considering 1-barriers only, as opposed to general i-barriers. Obviously, up to equivalence of label switching there is no difference, and the above theorem holds when we switch 1 with some other fixed state from $\mathcal{Y}$.

Second, the bold $\mathbf{x}$ here is an observation sequence whose length is determined by the length of the vector set it originates from. Hence there is no ambiguity in the definitions of $\mathcal{Y}^+(\mathbf{x})$ and $p_{ij}(\mathbf{x})$, since those operators are defined on any domain of the form $\mathcal{X}^n$, where $n \geq 2$.

Third, the condition (V1) sets the requirements for the "center part" of the barrier set $\mathcal{X}^*$. This center part – namely, the set $\mathcal{X}^*_{(n_1, n_{2N+1})}$ – consists of $2N$ cycles, and for each of those cycles there is a requirement that the inequalities (17), (18) and (19) hold for each observation sequence $\mathbf{x}$ from the cycle. In the theorem the center part is given as generally as possible, so the cycle lengths $n_{k+1} - n_k + 1$ can be different for $k = 1, \ldots, 2N$. In practice, however, the cycle lengths can usually all be taken to be equal. Further, it is typically most useful to take all the cycle lengths equal to 2, because in that case $p_{ij}(\mathbf{x})$ becomes the transition kernel density, thereby allowing to tie (V1) directly to the parameters of the specific model.

Fourth, in the above theorem the center part $\mathcal{X}^*_{(n_1, n_{2N+1})}$ need not have the structure of the product set. However, in practice it is usually most convenient to find a set $B \subseteq \mathcal{X}^r$ for some $r \geq 2$ such that $B = B_{(1,r-1)} \times B_{(1)}$ and (17), (18) and (19) holds for all $\mathbf{x} \in B$. Then the center part can be constructed as the product set $B_{(1,r-1)} \times \cdots \times B_{(1,r-1)}$ ($2N$ times). In that case, assuming that also (V2) holds, N will be independent of the constants ϵ, δ and Δ, and so regardless

of the specific values of those constants, N can always be taken so large that (V3) holds. For example, above we briefly considered the case where $\mathcal{X} = \mathbb{R}$ and there exists an interval $I \subset \mathbb{R}$ such that $q(x, 1|x', 1)(1-\epsilon) > q(x, j|x', i)$ for some $\epsilon > 0$, all $(i, j) \in \mathcal{Y}^2 \setminus \{(1, 1)\}$ and all $x, x' \in I$. Then for any $N \geq 1$, the set I^{2N} is the center part with $2N$ cycles of length 2 which satisfies (V1).

Finally, note that the condition (V2) is only concerned with the "edges" of the barrier set, namely the sets $\mathcal{X}^*_{(1,n_1)}$ and $\mathcal{X}^*_{(n_{2N+1},n_{2N+2})}$. There is overlap with the edges and the center part on the sets $\mathcal{X}^*_{(n_1)}$ and $\mathcal{X}^*_{(n_{2N+1})}$, and so the conditions (V1) and (V2) are tied together in this way.

Theorem 1 does not by itself guarantee the existence of the Viterbi process, because it does not say that X will go through the barrier set $\mathcal{X}^*$ infinitely often. This problem is addressed by the following lemma. Recall the definitions of ψ-irreducibility and Harris recurrence introduced earlier.

Lemma 1 [I, Lem. 2.1]**.** *Let $\mathcal{X}^* \subseteq \mathcal{X}^M$ satisfy (V1) and (V2) and let Z be Harris recurrent. Moreover, assume that there exists $i \in \mathcal{Y}$ such that $i \in \mathcal{Y}^+(\mathbf{x})_{(1)}$ for every $\mathbf{x} \in \mathcal{X}^*_{(1,n_1)}$. If $\mu^{M-1}(\{x_{2:M} \mid x_{1:M} \in \mathcal{X}^*\}) > 0$ for all $x_1 \in \mathcal{X}^*_{(1)}$ and $\psi(\mathcal{X}^*_{(1)} \times \{i\}) > 0$, then $P(X \in \mathcal{X}^* \text{ i.o.}) = 1$.*

Here, the condition for the positive $\mu^{M-1}(\{x_{2:M} \mid x_{1:M} \in \mathcal{X}^*\})$ for all $x_1 \in \mathcal{X}^*_{(1)}$ is a very natural one. In particular, this holds if the set $\mathcal{X}^*$ has positive μ^M measure and has the product structure $\mathcal{X}^* = \mathcal{X}^*_{(1)} \times \mathcal{X}^*_{(2,M)}$. The requirement for having a fixed $i \in \mathcal{Y}^+(\mathbf{x})_{(1)}$ for every $\mathbf{x} \in \mathcal{X}^*_{(1,n_1)}$ is not a restrictive one; indeed if this does not hold for the whole set $\mathcal{X}^*$, one can specify a subset where it does hold and consider that set as the new barrier set. The condition $\psi(\mathcal{X}^*_{(1)} \times \{i\}) > 0$ is a natural one for ψ-irreducible chains, and can be easily verified for most models.

Hidden Marko model Prior to Paper I the most general conditions for the existence of the Viterbi process for the HMM-case were obtained the by Lember and Koloydenko in [18]. It turns out that not only can this result be replicated using Theorem 1 and Lemma 1, but in some aspects even significant improvements can be made. For the HMM-case we have the following result. Define $p_{\cdot j} = \max_{i \in \mathcal{Y}} p_{ij}$ and $G_j = \{x \in \mathcal{X} \mid f_j(x) > 0\}$.

Corollary 1 [I, Cor. 4.1]**.** *Assume that Z is an HMM satisfying the following conditions.*

(i) For each state $j \in \mathcal{Y}$

$$\mu\left(\left\{x \in \mathcal{X} \mid p_{\cdot j} f_j(x) > \max_{i \in \mathcal{Y} \setminus \{j\}} p_{\cdot i} f_i(x)\right\}\right) > 0.$$

(ii) Markov chain Y is irreducible, and there exists a set $C \subseteq \mathcal{Y}$ such that

$$\mu\left[(\cap_{i \in C} G_i) \setminus (\cup_{i \notin C} G_i)\right] > 0 \tag{20}$$

and the sub-stochastic matrix $\mathbb{P}_C = (p_{ij})_{i,j \in C}$ is irreducible and aperiodic.

Then for some $i^* \in \mathcal{Y}$ there exists a set $\mathcal{X}^*$ consisting of strong i^*-barriers of fixed order and satisfying $P(X \in \mathcal{X}^* \text{ i.o.}) = 1$.

Observe that the conditions of the corollary are invariant up to μ-equivalence of the emission densities $\{f_j\}$. In other words, if the transition matrix (p_{ij}) is fixed, and $\{f_j\}$ and $\{f'_j\}$ are μ-equivalent emission densities (i.e. $f_j(x) = f'_j(x)$ for μ-a.e. $x \in \mathcal{X}$ and all $j \in \mathcal{Y}$), then the conditions hold for $\{f_j\}$ if and only if they hold for $\{f'_j\}$.

The condition (i) is used to construct the center part of the barrier set such that (V1) holds. The requirement of the stochastic $\mathbb{P}_C$ in (ii) to be irreducible and aperiodic is known to be equivalent to it being *regular*, which means that there exists a $k \geq 1$ such the matrix $\mathbb{P}_C^k$ consists of only positive elements. This property of being regular in conjunction with (20) is used to construct the edges of the barrier set which satisfy (V2). The set C may consist of a single element, i.e. $C = \{i\}$ for some state i; in that case the requirement for $\mathbb{P}_C$ means that p_{ii} must be positive. The set C can also be the whole state space $\mathcal{Y}$; in that case we define $\cup_{i \notin \mathcal{Y}} G_i = \emptyset$, so that the condition (20) becomes $\mu(\cap_{i \in \mathcal{Y}} G_i) > 0$. If $f_i(x)$ are all positive everywhere on $\mathcal{X}$, then (20) holds only if $C = \mathcal{Y}$, and, consequently, (ii) holds if and only if the Markov chain Y is irreducible and aperiodic. If Y is irreducible and there exists a set $C \subseteq \mathcal{Y}$ such that the sub-stochastic matrix $\mathbb{P}_C = (p_{ij})_{i,j \in C}$ is irreducible and aperiodic, then it can be easily shown that Y must be aperiodic too; thus (ii) implies that Y is aperiodic.

Corollary 1 is stronger than the result of [18] in the following ways. Firstly, and perhaps most significantly, Corollary 1 ensures the existence of infinitely many strong nodes, whereas in [18] the nodes are not proven to be strong (recall that non-strong nodes do not align with the natural (co)lexicographic ordering scheme of the Viterbi algorithm and are therefore undesirable).

Second, in [18] instead of (20) the following condition is used:

$$\mu\left(\cap_{i \in C} G_i\right) > 0 \quad \text{and} \quad \mu\left[\left(\cap_{i \in C} G_i\right) \cap \left(\cup_{i \notin C} G_i\right)\right] = 0. \tag{21}$$

Note that this is a significantly stricter requirement. As an example we can consider the case when $|\mathcal{Y}| = 2$ and $C = \{1\}$. Then (21) states that G_1 and G_2 must not overlap except on a μ-null set. This would mean that the HMM ceases to be hidden, because the hidden states would be revealed by their emissions. Conversely, (20) would simply mean that there is some μ-positive set where G_1 and G_2 do not overlap – a much less restrictive requirement.

Finally, in [18] it is assumed that Y is stationary. This is used to ensure the infinite recurrence of barriers through the ergodicity of Z. Here we have relied on the theory of Harris chains instead, so this assumption is not needed.

Lower semi-continuous transition densities As we saw, Theorem 1 can be effectively used to derive broad conditions for the existence of the Viterbi process in case of HMM. However, for more complex models proving conditions (V1)-(V3) might be more challenging. In particular, the interplay between conditions (V1) and (V2) poses technical difficulties which depending on the specific model range from mildly inconvenient to painfully arduous. In our second barrier construction theorem we seek to overcome this problem by considering a different set of conditions which imply (V1)-(V3). More specifically, we shall assume that the transition kernel densities are lower semi-continuous and that there exists a cyclic center part of the barrier set as given in (V1). The main achievement of this theorem is then the construction of the edges of the barrier set under a general condition which is independent of the center part.

A point $z^* \in \mathcal{Z}$ is called *reachable* if for every open neighborhood O of z^*,

$$\sum_{k=2}^{\infty} P_z(Z_k \in O) > 0, \quad \forall z \in \mathcal{Z}.$$

For ψ-irreducible Z, the point z^* is reachable if and only if it belongs to the support of ψ [10, Lem. 6.1.4]. Since we have equipped space $\mathcal{Z}$ with product topology $\tau \times 2^{\mathcal{Y}}$, where τ denotes the topology of $\mathcal{X}$ and $2^{\mathcal{Y}}$ denotes the discrete topology on $\mathcal{Y}$, the above-stated definition is in fact equivalent to the following: point $(x, i) \in \mathcal{Z}$ is called reachable, if for every open neighborhood O of x,

$$\sum_{k=2}^{\infty} P_z(Z_k \in O \times \{i\}) > 0, \quad \forall z \in \mathcal{Z}.$$

Recall that a measure is called *strictly positive*, if it assigns a positive measure to every non-empty open set. The second barrier construction theorem can now be stated as follows.

Theorem 2 [I, Th. 3.1]**.** *Assume the following.*

(LV1) For arbitrary number of cycles $2N \geq 4$ there exists an open center part of a barrier set $\mathcal{X}^*_{(n_1,n_{2N+1})}$ satisfying (V1). We assume that both set $\mathcal{X}^*_{(n_1)}$ and parameter ϵ of (V1) are independent of N, and there exists a compact set $K \subseteq \mathcal{X}$, which is independent of N, such that $\mathcal{X}^*_{(n_{2N+1})}$ is contained in K for each N. Furthermore, we assume that there exists $x^* \in \mathcal{X}^*_{(n_1)}$ such that $(x^*, 1)$ is reachable.

(LV2) There exists an open set $E \subseteq \mathcal{X}^q$, $q \geq 2$, such that $\mathcal{Y}^+ = \mathcal{Y}^+(\mathbf{x})$ is the same for every $\mathbf{x} \in E$ and satisfies the following property: $(i, j) \in \mathcal{Y}^+$ for every $i \in \mathcal{Y}^+_{(1)}$ and $j \in \mathcal{Y}^+_{(2)}$. Furthermore, we assume that there exists a reachable point in $E_{(1)} \times \mathcal{Y}^+_{(1)}$.

Let μ be strictly positive and let for every pair of states $i, j \in \mathcal{Y}$ function $(x, x') \mapsto q(x, i|x', j)$ be lower semi-continuous and bounded. There exists a barrier set $\mathcal{X}^*$ consisting of 1-barriers of fixed order. Moreover, if Z is Harris recurrent, then $P(X \in \mathcal{X}^* \text{ i.o.}) = 1$. If for each N the center part of barrier set satisfies (V1) with either (17) or (18) being strict, then the barriers are strong.

Thus (LV1) requires the existence of a center part of a barrier set satisfying (V1) for any number of cycles $2N \geq 4$, such that the last projection $\mathcal{X}^*_{(n_{2N+1})}$ of the center part is contained in the compact set K and the first projection $\mathcal{X}^*_{(n_1)}$ is the same for every N. Usually in practice the easiest way to prove this is to construct an open cycle $B \subseteq \mathcal{X}^r$, $r \geq 2$, such that $B = B_{(1,r-1)} \times B_{(1)}$ and (17), (18) and (19) hold for all $\mathbf{x} \in B$. In that case for any $N \geq 2$ the center part can be constructed as $B_{(1,r-1)} \times \cdots B_{(1,r-1)}$ ($2N$ times), and then for (LV1) to hold one needs to only ensure that $B_{(1)}$ is contained in a compact set and contains an x^* such that $(x^*, 1)$ is reachable.

The condition (LV2) is used in the construct the edges for the barrier set which satisfy (V2). Since the corresponding constants $0 < \delta < \Delta < \infty$ can be shown to depend on the set E, compact set K and the set $\mathcal{X}^*_{(n_1)}$ only, then the number of cycles $2N$ in (LV1) can always be taken so large that (V3) holds. The

condition (LV2) is not a restrictive one and is usually relatively easy to verify for specific models. Essentially the same condition is also central to our results in Paper III. However, the general subject matter of Paper III is very different from Paper I, and so the fact that the same condition can be effectively applied in both areas of research could potentially indicate its wider significance. We shall also see in Section 7 that this condition ensures in a way the geometric ergodicity of the conditional signal process $P(Y_{1:n} \in \cdot | X_{1:n})$, so this gives some insight into its broader relevance in the study of PMM's.

Discrete observation space Suppose now that $\mathcal{X}$ is s discrete (finite or countable), and Z is an irreducible and recurrent Markov chain on some subspace $\mathcal{Z}' \subseteq \mathcal{Z}$ (not necessarily product space). In the discrete topology every function is continuous and so Theorem 2 applies. For discrete irreducible Markov chain on state space $\mathcal{Z}'$ the maximal irreducibility measure ψ is the counting measure on $\mathcal{Z}'$ and Harris recurrence is equivalent to the usual recurrence. However, in order to apply Theorem 2, we should define the Markov chain Z on the product space $\mathcal{Z}$. This product space might not be equal to $\mathcal{Z}'$, in which case the transition matrix should be formally extended. It is easy to do so in a way that the chain remains ψ-irreducible and Harris recurrent, where ψ is the counting measure on $\mathcal{Z}'$. Then every element of $\mathcal{Z}'$ is reachable and the following result can be easily derived from Theorem 2.

Corollary 2 [I, Cor. 4.2]**.** *Let $\mathcal{X}$ be discrete and let Z be an irreducible and recurrent Markov chain on state-space $\mathcal{Z}' \subseteq \mathcal{Z}$. Then the following conditions ensure that there exists a barrier set $\mathcal{X}^*$ consisting of a strong 1-barrier of fixed order and satisfying $P(X \in \mathcal{X}^* \text{ i.o.}) = 1$.*

(i) For some $n \geq 1$ there exists $x_{1:n} \in \mathcal{X}^n$ such that $(x_1, 1) \in \mathcal{Z}'$, and defining $\mathbf{x} = (x_{1:n}, x_1)$, we have

$$p_{11}(\mathbf{x}) \geq p_{i1}(\mathbf{x}) \quad \forall i \in \mathcal{Y} \setminus \{1\}, \tag{22}$$
$$p_{11}(\mathbf{x}) \geq p_{1j}(\mathbf{x}) \quad \forall j \in \mathcal{Y} \setminus \{1\}, \tag{23}$$
$$p_{11}(\mathbf{x}) > p_{ij}(\mathbf{x}) \quad \forall i, j \in \mathcal{Y} \setminus \{1\},$$

where either inequalities (22) *or inequalities* (23) *are strict.*

(ii) There exists $q \geq 2$ and a sequence $x'_{1:q} = \mathbf{x}' \in \mathcal{X}^q$ such that $(x'_1, i') \in \mathcal{Z}'$ for some $i' \in \mathcal{Y}^+(\mathbf{x}')_{(1)}$, and $(i, j) \in \mathcal{Y}^+(\mathbf{x}')$ for every $i \in \mathcal{Y}^+(\mathbf{x}')_{(1)}$ and $j \in \mathcal{Y}^+(\mathbf{x}')_{(2)}$.

Proof. Indeed, (i) implies that the center part of the barrier set required by (LV1) can be defined as $\{(x_{1:n}, \ldots, x_{1:n})\}$ ($2N$ times). The condition (ii) immediately implies (LV2) with $E = \{\mathbf{x}'\}$. □

Paper I contains further discussion on the condition (ii), as well as a demonstration of applying the above corollary to a specific 2-state and 2-observation model.

Gaussian linear switching model There seems to be no way to extrapolate from (LV1) and (LV2) simple conditions on the transition kernel density which apply to any PMM. Therefore for each specific model (LV1) and (LV2) need

to be verified separately. In addition, usually there is more than one way to construct the center part of the barrier set, and different approaches will result in different sets of conditions. Generally the easiest way is to use the transition kernel density itself so that the cycle length is 2. We shall now demonstrate such approach on the linear switching model with Gaussian noise. In Paper I a more general treatment is given for any lower semi-continuous noise densities on d-dimensional Euclidean space. Nonetheless, the approach in Paper I is similarly based on cycle length 2, and the main idea for constructing the center part can be easily illustrated on the 1-dimensional Gaussian model.

Recall that the linear Markov switching model is based on the stochastic equation $X_k = \alpha(Y_k)X_{k-1} + \xi_k(Y_k)$, where Y is a homogeneous Markov chain on $\mathcal{Y}$, $\alpha(i)$ are constants and $\xi_k(i)$ are noise variables. Here we assume that $\mathcal{X} = \mathbb{R}$ and the noise variables $\xi_k(i)$ are Gaussian with respective means μ_i and variances σ_i^2. Thus μ is Lebesgue measure, which is obviously strictly positive. Denoting with h_i the Gaussian density function of $\xi_k(i)$ we have that the transition kernel density for the model expresses as $q(x, j|x', i) = p_{ij}h_j(x - \alpha(j)x')$, where p_{ij} are the transition probabilities of the hidden Markov chain Y. Assume now that Y is irreducible. Since the Gaussian densities h_i are positive on the whole real line, we have that the model is $\mu \times c$-irreducible, where c denotes the counting measure on $\mathcal{Y}$. It follows that every element in $\mathcal{Z} = \mathbb{R} \times \mathcal{Y}$ is reachable. Further, if Y is also aperiodic, then denoting with $\mathbb{P} = (p_{ij})$ the transition matrix of Y, we have that there exists $k \geq 1$ such that $\mathbb{P}^k$ contains only positive elements. Thus $\mathcal{Y}^+(\mathbf{x}) = \mathcal{Y} \times \mathcal{Y}$ for every $\mathbf{x} \in \mathbb{R}^k$, and so it follows that (LV2) must hold with $E = \mathbb{R}^k$.

Assume now that p_{11} dominates its column in the transition matrix $\mathbb{P}$, i.e. $p_{11} = \max_{i \in \mathcal{Y}} p_{i1}$. Then

$$p_{11}h_1(x - \alpha(1)x') \geq p_{i1}h_1(x - \alpha(1)x') \quad \forall x, x' \in \mathbb{R}, \quad \forall i \in \mathcal{Y}. \tag{24}$$

Next, suppose there exists a single $x^* \in \mathbb{R}$ such that

$$p_{11}h_1(x^* - \alpha(1)x^*) > p_{ij}h_j(x^* - \alpha(j)x^*) \quad \forall i \in \mathcal{Y}, \quad \forall j \in \mathcal{Y} \setminus \{1\}. \tag{25}$$

By continuity of the Gaussian densities there must exist an $\epsilon > 0$ and an open interval B which contains x^* such that

$$p_{11}h_1(x - \alpha(1)x')(1 - \epsilon) > p_{ij}h_j(x - \alpha(j)x') \\ \forall i \in \mathcal{Y}, \quad \forall j \in \mathcal{Y} \setminus \{1\}, \quad \forall x, x' \in B. \tag{26}$$

Thus for every $N \geq 2$ the center part of the barrier set can be constructed as $B \times \cdots \times B$ ($2N$ times). By (24) and (26) this center part satisfies the inequalities (17), (18) and (19) of (V1), and so (LV1) holds. Also note that the inequalities (18) are strict, so the prerequisite for the strong barriers is satisfied.

Thus to prove (LV1) it remains to ensure the existence of x^* satisfying (25). This value may not always exist, because the weighted Gaussian densities $p_{ij}h_j$ on the right side of (25) may combine to fully cover the weighted density $p_{11}h_1$ on the left side. Nevertheless, more often than not it does exist, and sufficient conditions for this are exhibited in the following result. A generalized version of the same result on the observation space $\mathbb{R}^d$ is given in [I, Cor. 4.3].

Corollary 3. *Let Z be the Gaussian linear Markov switching model. If the following conditions are fulfilled, then there exist a barrier set $\mathcal{X}^*$ consisting of strong 1-barriers of fixed order and satisfying $P(X \in \mathcal{X}^*$ i.o.$) = 1$:*

(i) Y *is irreducible;*

(ii) $p_{11} = \max_{i \in \mathcal{Y}} p_{i1}$*;*

(iii) $\alpha(1) \neq 1$*, and for all* $i \in \mathcal{Y}$ *and* $j \in \mathcal{Y} \setminus \{1\}$ *either* $p_{ij} = 0$ *or*

$$[\alpha(1)\mu_j - \alpha(j)\mu_1]^2 > -2\sigma_j^2 \ln\left(\frac{p_{11}\sigma_j}{p_{ij}\sigma_1}\right); \tag{27}$$

(iv) $\max_{i \in \mathcal{Y}} \sum_{j \in \mathcal{Y}} p_{ij}|\alpha(j)| < 1$.

Proof. Take $x^* = \mu_1/(1-\alpha(1))$. Then $x^* - \alpha(1)x^* = \mu_1$, which maximizes h_1. Pick now $i \in \mathcal{Y}$ and $j \in \mathcal{Y} \setminus \{1\}$, and note that by (ii) $p_{11} > 0$, and so when $p_{ij} = 0$, then inequality (25) holds trivially. On the other hand, if $p_{ij} > 0$, then substituting the value $\mu_1/(1-\alpha(1))$ into x^* in (25) yields (27). Thus (LV1) must hold. By (ii) $p_{11} > 0$, which together with (i) implies that Y is irreducible aperiodic; thus (LV2) also holds. Recall that (iv) ensures that the model is Harris recurrent, so the statement now follows from Theorem 2. □

Note that when $(p_{11}\sigma_j)/(p_{ij}\sigma_1) > 1$, then the right hand side of (27) becomes negative, and so the inequality will hold regardless of the other values.

The conditions of Corollary 3 are elegant, easy to check and hold for many models used in practice. However, they do have one drawback. Namely, the condition (ii) is somewhat restrictive, requiring up to equivalence of label switching that there exists at least one diagonal entry of the transition matrix $\mathbb{P}$ which dominates its column. While true for many models, we would like to also have sufficient conditions for the existence of the Viterbi process when this is not the case. In Paper I grouping together the elements of Z is suggested to overcome this issue. For example, if $\{Z_k\}_{k \geq 1}$ is the Gaussian linear Markov switching model on space $\mathbb{R} \times \mathcal{Y}$, then the pairs $\{(Z_{2k-1}, Z_{2k})\}_{k \geq 1}$ can be shown to be Gaussian linear Markov switching model on $\mathbb{R}^2 \times \mathcal{Y}^2$. Because in Paper I the conditions (i)-(iv) are given generally for d-dimensional observation space, one can check if (ii) holds for the paired model when it does not hold for Z itself.

An alternative approach is to try to use cycle length that is greater than 2. For example, the cycle length 3 could be employed as follows. Suppose there exists $s \in \mathcal{Y}$ and $x_1^*, x_2^* \in \mathbb{R}$ such that

$$\begin{aligned} p_{1s}h_s(x_2^* - \alpha(s)x_1^*) \cdot p_{s1}h_1(x_1^* - \alpha(1)x_2^*) > \\ p_{ij}h_j(x_2^* - \alpha(j)x_1^*) \cdot p_{jk}h_k(x_1^* - \alpha(k)x_2^*), \\ \forall (i,k) \in \mathcal{Y}^2 \setminus \{(1,1)\}, \quad \forall j \in \mathcal{Y}. \end{aligned} \tag{28}$$

This means that the inequalities (17), (18) and (19) of (V1) all hold strictly for $\mathbf{x} = (x_1^*, x_2^*, x_1^*)$. Then the continuity argument again ensures the existence of $\epsilon > 0$ and open neighborhood $B \subset \mathbb{R}^2$ of (x_1^*, x_2^*) so that the center part of the barrier set which satisfies (LV1) can be constructed as $B \times \cdots \times B$ ($2N$ times). To optimize for (28) one may want to find x_1^* and x_2^* so that the left side of the inequality is maximal. This would involve equating the 2 partial derivatives of the left side of (28) with 0, and solving the resulting linear equations. Note that the necessary condition for (28) is that the transition matrix element p_{1s} strictly dominates its column. Consequently, for $s = 1$ this is approach would be almost equivalent to the one of Corollary 3, except the condition (ii) would be

replaced with a slightly stronger requirement that p_{11} must dominate its column *strictly*. In general, however, s need not be 1, and so we have gotten rid of the condition that a diagonal element of the transition matrix of Y must dominate its column. The drawback of this approach is that the resulting conditions for the parameters of the Gaussian densities will be much more arithmetically involved than (iii) when $s \neq 1$, so the final result will not be as elegant as Corollary 3.

6 Summary of Paper II

The existence of the Viterbi process shows that the Viterbi path converges to a fixed state as the sample size n grows. This justifies the use of the Viterbi classifier for estimating the hidden path $Y_{1:n}$. In Paper II we take a step further, and prove a general theorem which allows to deduce SLLN and CLT type results on the Viterbi classifier. The idea behind the theorem is to construct regeneration times $\{S_k\}_{k\geq 1}$ which break the Markov chain Z into i.i.d. cycles and also coincide with the nodes. The standard approach for creating regeneration times for any general state Markov chain is the so called *splitting method*. The main achievement of Paper II is showing that under general conditions the splitting method can be employed in a way which ensures that the resulting regeneration times are also nodes.

More formally, for any sequence $u = (u_k)_{k\geq 1}$ and a sequence of times $s = (s_k)_{k\geq 1}$, $1 \leq s_1 < s_2, \ldots$, a *shift operator* θ_t, $t \geq 1$, is defined by $\theta_t(u, s) = ((u_k)_{k\geq t}, (s_k - t + 1)_{k\geq n(t)})$, where $n(t) = \min\{n \mid s_n \geq t\}$. A process $\{U_k\}_{k\geq 1}$ is called *regenerative* [22, 23] (in the classic sense), if there exists a sequence of random times $S = \{S_k\}_{k\geq 1}$, $1 \leq S_1 < S_2 < \ldots$, called *regeneration times*, such that for each $n \geq 1$

$$\theta_{S_n}(U, S) \stackrel{d}{=} \theta_{S_1}(U, S),$$
$$\theta_{S_n}(U, S) \text{ is independent of } (\{U_k\}_{k=1}^{S_n-1}, S_1, \ldots, S_n).$$

Here $\stackrel{d}{=}$ denotes equality by distribution. Random variables $S_k - S_{k-1}$ are called *inter-regeneration times*. Typically one is interested in the case where inter-regeneration times have finite mean, i.e. $\mathbb{E}[S_2 - S_1] < \infty$. For any $A \in \mathcal{B}(\mathcal{Z})$ let τ_A denote the number of time-steps for Z to reach A after time 1:

$$\tau_A = \min\{n \geq 1 \mid Z_{n+1} \in A\}.$$

Throughout this section we will be dealing with strong nodes only, and hence we will assume that the Viterbi classifier follows some (co)lexicographic ordering scheme. The main theorem of Paper II now reads as follows.

Theorem 3 [II, Th. 3]**.** *Assume the following.*

(R1) There exists a set $\mathcal{X}^* \subseteq \mathcal{X}^M$ satisfying the following conditions:

- *(i) $M \geq 3$ and $\mathcal{X}^*$ consists of strong 1-barriers of order $r \in \{1, \ldots, M-2\}$;*
- *(ii) there exist state $i_0 \in \mathcal{Y}$, such that for every $z \in \mathcal{Z}$*

$$P_z\left(Z_{1:M-r-1} \in \widetilde{\mathcal{Z}} \text{ i.o.}\right) = 1,$$

where we define $\widetilde{\mathcal{Z}} = \mathcal{X}^*_{(1,M-r-1)} \times (\mathcal{Y}^{M-r-2} \times \{i_0\})$;

(iii) it holds $\mathcal{X}^* = \mathcal{X}^*_{(1,M-r-1)} \times \mathcal{X}^*_{(M-r,M)}$;

(iv) there exists $i_1 \in \mathcal{Y}$ *such that for all* $x \in \mathcal{X}^*_{(M-r)}$

$$P(X_{M-r+1:M} \in \mathcal{X}^*_{(M-r+1,M)} | Z_{M-r} = (x, i_1)) > 0.$$

(R2) Denote $\mathcal{Z}_0 = \mathcal{X}^*_{(M-r-1)} \times \{i_0\}$ *and* $\mathcal{Z}_1 = \mathcal{X}^*_{(M-r)} \times \{i_1\}$. *There exists a probability measure* ν *on* $\mathcal{B}(\mathcal{Z})$ *and* $\beta \in (0,1)$ *such that* $\nu(\mathcal{Z}_1) = 1$ *and*

$$P_z(Z_2 \in A) \geq \beta \nu(A), \quad \forall z \in \mathcal{Z}_0, \quad \forall A \in \mathcal{B}(\mathcal{Z}).$$

(R3) It holds $\sup_{z \in \mathcal{Z}_0} \mathbb{E}_z[\tau_{\mathcal{Z}_0}] < \infty$.

Then there exists a sequence of regeneration times $\{S_k\}_{k \geq 1}$ *for Markov chain* Z *such that* $E[S_2 - S_1] < \infty$, *and* $P(X_{S_k - M + r + 1 : S_k + r} \in \mathcal{X}^*) = 1$ *for all* $k \geq 1$.

The conditions (R1) and (R2) imply the classic constraint for employing the splitting method: i.e. that there exists a set $\mathcal{Z}_0 \subseteq \mathcal{Z}$ such that $P(Z \in \mathcal{Z}_0 \text{ i.o.}) = 1$ and that there exists a probability measure ν on $\mathcal{B}(\mathcal{Z})$ such that for some $\beta > 0$ the Markov kernel $P_z(Z_2 \in \cdot)$ of Z dominates the measure $\beta\nu$ for all $z \in \mathcal{Z}_0$. In addition, the conditions (R1) and (R2) also specify the existence of the barrier set $\mathcal{X}^*$ as well as some non-restrictive technical requirements for the barrier set and its relation to the measure ν and set $\mathcal{Z}_0$. The condition (R3) is solely used to ensure that the resulting inter-regeneration times have finite mean, i.e. $\mathbb{E}[S_2 - S_1] < \infty$.

The theorem states the existence of regeneration times $\{S_k\}_{k \geq 1}$ such that for all $k \geq 1$ the portion $X_{S_k - M + r + 1 : S_k + r}$ of the observation sequence is almost surely contained in the barrier set $\mathcal{X}^*$. Firstly, note that this immediately implies that the observation process goes through $\mathcal{X}^*$ infinitely many times, and so the Viterbi process $V = \{V_k\}_{k \geq 1}$ must exist. Secondly, by the definition of a barrier all the regeneration times are almost surely nodes, that is, for each $k \geq 1$ the time $S_k(\omega)$ is a strong r-order node of the observation sequence $X(\omega)$ for almost all ω (for all $\omega \in \cap_{k \geq 1} \{X_{S_k - M + r + 1 : S_k + r} \in \mathcal{X}^*\}$, to be more specific). The relevance of this for the asymptotics of the Viterbi classifier requires further explanation, but first, let us discuss how the conditions (R1)-(R3) can be verified for specific models.

If Z is HMM then (R1)-(R3) hold when the conditions of Corollary 1 in Section 5 hold. This result is proven in [II, Lem. 3] for the sake of completeness, but is not in itself that surprising, because the regenerative property of the Viterbi process for HMM was already known from [18].

More interesting is the case when Z is not HMM. It turns out that, coincidentally, lower semi-continuity of the transition kernel density becomes useful again for proving (R2). In particular, below lemma ties (R1)-(R3) to the conditions of the Theorem 2 in Section 5. A set $A \in \mathcal{B}(\mathcal{Z})$ is called *regular* when Z is ψ-irreducible, if for all B satisfying $\psi(B) > 0$, $\sup_{z \in A} \mathbb{E}_z[\tau_B] < \infty$.

Lemma 2 [II, Lem. 1]**.** *Assume the following:* μ *is strictly positive, function* $(x, x') \mapsto q(x, j | x', i)$ *is lower semi-continuous and bounded for all* $i, j \in \mathcal{Y}$, Z *is Harris recurrent, (LV1) holds with either inequalities (17) or (18) of (V1) being strict, and (LV2) holds. Then the resulting barrier set* $\mathcal{X}^*$ *of Theorem 2 satisfies (R1)-(R2). Further, if the set* $\mathcal{X}^*_{(n_{N+1}-1)} \times \mathcal{Y}$ *is regular, then (R3) also holds.*

The lemma tells us that once we have already confirmed the existence of the set $\mathcal{X}^*$ consisting of strong barriers through Theorem 2, then we do not need to worry about proving (R1) and (R2) anymore, because those are automatically fulfilled. Further, to ensure (R3) one only needs to make sure that the center part of the barrier set is such that $\mathcal{X}^*_{(n_{N+1}-1)} \times \mathcal{Y}$ is regular. This is usually easily verifiable for specific models, and the theoretical tools that ensure the Harris recurrence of the chain Z often also give a characterization of the regular sets. For example, the condition $\max_{i\in\mathcal{Y}} \sum_{j\in\mathcal{Y}} p_{ij}|\alpha(j)| < 1$ we used earlier together with irreducibility of Y to ensure the Harris recurrence of the Gaussian linear switching model also guarantees that every compact set in $\mathcal{B}(\mathcal{Z})$ is regular (this is implied by [II, Lem. 7]). Since for the Gaussian linear Markov swithcing model we constructed the center part based on bounded intervals, then it is clear that it satisfies (R3) without any further conditions. When $\mathcal{X}$ is finite then obviously all subsets of $\mathcal{Z}$ are regular. When $\mathcal{X}$ is discrete but countable, then positive recurrence of Z ensures that every finite subset of $\mathcal{Z}$ is regular; in particular, in Corollary 2 positive recurrence need to be additionally assumed to ensure that the center part of the barrier set satisfies (R3).

Asymptotics of the Viterbi classifier In Paper II the main regeneration theorem ([II, Th. 3]) differs from Theorem 3 above, and states that (R1)-(R3) imply the existence of the regeneration times $\{S_k\}$ for the whole dual process $(Z, V) = \{(Z_k, V_k)\}_{k\geq 1}$. However, this is a mistake in Paper II, and such regeneration times cannot be inferred from (R1)-(R3). In Theorem 3 above this mistake has been fixed, and $\{S_k\}$ are no longer claimed to be the regeneration times for the whole dual process (Z, V), but only for Markov chain Z itself. However, the usefulness of this restated theorem for understanding the asymptotic properties of the Viterbi classifier is not immediately obvious, and requires further explanation.

Assume (R1)-(R3). As noted, the regeneration times $\{S_k\}$ coincide with the strong 1-nodes. Thus $V_{S_k} = 1$ a.s. for all $k \geq 1$. Moreover, note that by the piecewise structure of the Viterbi process, we have that

$$V_{S_{k-1}:S_k} \text{ depends on } X_{S_{k-1}:S_k} \text{ only for any } k \geq 2. \tag{29}$$

This does not imply – as falsely claimed in Paper II – that $\{S_k\}$ are regeneration times for the Viterbi process V, because $V_{S_{k-1}:S_k-1}$ depends on X_{S_k} and so the pieces $V_{S_{k-1}:S_k}$ and $V_{S_k:S_{k+1}}$ might be dependent. Nevertheless, we are still able to make deductions on the asymptotic behavior of the Viterbi classifier using the following argument. Denote

$$\eta_k = ((Z_{S_k}, V_{S_k}), (Z_{S_k+1}, V_{S_k+1}) \ldots, (Z_{S_{k+1}}, V_{S_{k+1}})), \quad k \geq 1.$$

By Theorem 3 $\{S_k\}$ are regeneration times for Z, and so by (29) we have that the stationary random process $\{\eta_k\}$ is *1-dependent*, that is, $\eta_{1:k-1}$ is independent of $\eta_{k+1:\infty}$ for all $k \geq 2$. This means that the standard SLLN and CLT results for m-dependent stationary random processes can be applied to $\{\eta_k\}$ (see e.g. [24, 25]). Because by Theorem 3 the inter-regeneration times $S_k - S_{k-1}$ have finite mean, it is now in turn possible to derive SLLN and CLT type results for the dual process $(Z, V) = \{(Z_k, V_k)\}$ (the effect of the front part $((Z_1, V_1), \ldots, (Z_{S_1-1}, V_{S_1-1}))$ will vanish as n goes to infinity). Further, since up to each regeneration time S_k the Viterbi path will remain fixed for sufficiently large sample size, and the

effect of the remaining non-fixed portion of the Viterbi path will disappear as n grows, one can under general conditions extend these results to the Viterbi classifier $v(X_{1:n})$ itself.

A version of an SLLN for the Viterbi classifier is given in the theorem [II, Th. 4]. This theorem was used to study asymptotic properties of the Viterbi training algorithm through the convergence of certain empirical measures. In particular, it is known that the Viterbi training algorithm is biased, but the limit of these empirical measures is useful for quantifying and estimating the magnitude of this bias for specific models. The theorem's proof – while straightforward – is mistakenly based on the assumption that the dual process (Z, V) is regenerative under (R1)-(R3), but this error can be easily fixed by relying instead on the 1-dependence argument above.

Regenerativity of the Viterbi process As mentioned, (R1)-(R3) do not generally imply that $\{S_k\}$ are the regeneration times for the dual process (Z, V) due to the dependence between $V_{S_{k-1}:S_k-1}$ and X_{S_k}, but as we saw above, this is not an obstacle for inferring asymptotic convergence results for the Viterbi classifier. Nevertheless, in many situations one can in fact make the stronger statement that whole dual process (Z, V) is regenerative. For instance, this is true for any HMM. Indeed, recall that by definition the transition kernel density of an HMM expresses as $p_{ij}f_j(x)$, where p_{ij} are the transition probabilities of the hidden Markov chain Y and f_j are the emission densities. Thus, for any sample size n, the conditional joint likelihood of $x_{1:n}$ and $y_{1:n}$ expresses as

$$p(x_{2:n}, y_{2:n}|x_1, y_1) = \prod_{k=2}^{n} p_{y_{k-1}y_k} f_k(x_k).$$

The path which maximizes the above expression over all paths $y_{1:n}$ satisfying $y_n = 1$ does not depend on x_n, and so it follows that $V_{S_{k-1}:S_k-1}$ is independent of X_{S_k}, which we needed to show.

The regenerative property of (Z, V) can also be ensured when $\mathcal{X}^*$ consists of a single vector, which can always be assumed without loss of generality when $\mathcal{X}$ is finite and is typically also true for countably large $\mathcal{X}$. Indeed, suppose $\mathcal{X}^* = \{x^*_{1:M}\}$, where $x^*_{1:M}$ is a barrier of order r. Thus X_{S_k} are all equal to the constant value of x^*_{M-r}, and so $V_{S_{k-1}:S_k-1}$ is always trivially independent of X_{S_k}.

Beyond HMM and discrete $\mathcal{X}$ the situation is more complex, but the regenerative property of (Z, V) still holds for many models. More specifically, recall from Section 5 that the most straightforward way to to construct the center part of the barrier set, is to have a set $B \subseteq \mathcal{X}^q$ for some $q \geq 2$ such that $B = B_{(1,q-1)} \times B_{(1)}$ and (17), (18) and (19) of (V1) hold for all $\mathbf{x} \in B$. Then the center part can be constructed as the product set $B_{(1,q-1)} \times \cdots \times B_{(1,q-1)}$ ($2N$ times) consisting of $2N$ cycles of length q. Usually in practice the set B is also constructed so that

$$v_{1:q} = \underset{y_{1:q}\colon\, y_1 = y_q = 1}{\arg\max}\; p(x_{2:q}, y_{2:q}|x_1, y_1) \text{ is the same for all } x_{1:q} \in B. \tag{30}$$

Suppose now that $\mathcal{X}^*$ is a barrier set containing such a center part, consisting of strong 1-barriers of order r and satisfying (V1)-(V3). In that case we can always ensure that $\mathcal{X}^*$ consists of also 1-barriers of order $r - q + 1$. Indeed, if that is

not the case, we can simply increase the number of cycles in the center part by two (i.e increase N by 1), and then one can easily verify from the statement of Theorem 1 that it must be so for the new $\mathcal{X}^*$. Next, assume that $\mathcal{X}^*$ satisfies (R1)-(R3); then the times $\{S_k\}$ are (random) strong 1-nodes of order r, but in addition also the times $\{S_k - q + 1\}$ are strong 1-nodes of order $r + q - 1$. Thus it follows from (30) and the piecewise structure of the Viterbi process that $V_{S_k-q+1:S_k}$ are constant and equal to $v_{1:q} = (1, v_{2:q-1}, 1)$, and that $V_{S_{k-1}:S_k-q}$ depends on $X_{S_{k-1}:S_k-q+1}$ only. Thus $V_{S_{k-1}:S_k-1}$ is independent of X_{S_k}, which is what we needed to show to ensure that (Z, V) is regenerative. In particular, this regenerative property holds for the Gaussian linear switching model under the conditions of Corollary 3.

7 Summary of Paper III

In Paper III the main subject matter is the *conditional signal process*, i.e. the process $Y_{1:n}$ conditioned on $X_{1:n}$. More specifically, we study the random distributions

$$P(Y_t \in \cdot | X_{s:n}), \tag{31}$$

where $s \leq t \leq n \leq \infty$. Formally, for each $A \subseteq \mathcal{Y}$, $P(Y_t \in A | X_{s:n})$ is defined as the conditional expectation $\mathbb{E}[\mathbb{I}_A(Y_t)|X_{s:n}]$, where $\mathbb{I}_A$ denotes the indicator function on A. Traditionally, when $t < n$, the probabilities (31) are called the *smoothing probabilities*, and when $t = n$, they are called the *filtering probabilities*. Note that when t increases together with n such that $t \leq n$, then intuitively for most well-behaving models we should expect the difference between (31) and $P(Y_t \in \cdot | X_{1:n})$ to disappear, because the effect of $X_{1:s-1}$ will vanish. This is called the *forgetting property* of the smoothing/filtering probabilities. The main result of Paper III states that this difference will decrease exponentially under general conditions. More specifically, we prove that there exists an $\alpha \in (0, 1)$ such that for all $s \leq t \leq n \leq \infty$,

$$\|P(Y_t \in \cdot | X_{1:n}) - P(Y_t \in \cdot | X_{s:n})\|_{TV} \leq C_s \alpha^{t-s}, \quad \text{a.s.}, \tag{32}$$

where C_s is a $\sigma(X_{s:\infty})$-measurable random variable and $\|\cdot\|_{TV}$ denotes the total variation norm. Thus the decrease of total variation difference between the two measures is exponential in t with a random coefficient C_s.

Because of its high relevance in the theoretical study of HMM's, extensive research has been done on this type of forgetting property under various conditions. Our paper stands apart in two aspects. Firstly, our paper applies to any PMM, not just HMM's. Secondly, rather than considering potentially uncountable hidden state space, we only look at the case when $\mathcal{Y}$ is finite. The special focus on finite $\mathcal{Y}$ allows us to exploit the discrete Markovian structure of the process $Y_{1:n}$ conditioned on $X_{1:n}$. This in turn enables us to derive more general conditions for the exponential forgetting than would have been possible for continuous $\mathcal{Y}$. Indeed, in Paper III we compare our assumptions to several existing conditions in the literature (adapted to finite $\mathcal{Y}$), and demonstrate how they can easily be shown to imply essentially as lenient if not strictly more lenient conditions.

In their most general form, our main assumptions can succinctly be formulated as follows. Recall the definition of the operator $\mathcal{Y}^+(\cdot)$ from (16).

(S1) There exists a set $E \subseteq \mathcal{X}^q$, $q \geq 2$, such that $\mathcal{Y}^+ = \mathcal{Y}^+(\mathbf{x}) \neq \emptyset$ is the same for all $\mathbf{x} \in E$ and satisfies the following property: $(i,j) \in \mathcal{Y}^+$ for every $i \in \mathcal{Y}^+_{(1)}$ and $j \in \mathcal{Y}^+_{(2)}$.

(S2) Chain Z is ψ-irreducible, with $\psi(E_{(1)} \times \mathcal{Y}^+_{(1)}) > 0$. Furthermore,

$$\mu^{q-1}(\{x_{2:q} \mid x_{1:q} \in E\}) > 0$$

for all $x_1 \in E_{(1)}$.

Note that these conditions are very similar to condition (LV2) of Theorem 2, but instead of relying on reachable points and open sets, we are requiring ψ-irreducibility of Z. (In fact, it is not difficult to verify that (S1)-(S2) are more lenient than (LV2) when Z is ψ-irreducible and μ is strictly positive, both of which are assumptions of Theorem 2). We shall return to the parallels with Paper I shortly within the context of HMM's, but first let us discuss the intuition behind conditions (S1)-(S2).

The intuitive meaning of (S1) is fairly simple, because it can be considered as the "irreducibility" and "aperiodicity" of the conditional signal process as follows. Suppose we have an inhomogeneous Markov chain $Y' = \{Y'_t\}_{t\geq 1}$, with $\mathcal{Y}_t$ being the finite state space of Y'_t. The canonical concepts of irreducibility and aperiodicity are not defined for such a Markov chain, but a natural generalization would be as follows: for every time t, there exists a time $n > t$ such that

$$P(Y'_n = j | Y'_t = i) > 0 \text{ for all } i \in \mathcal{Y}_t \text{ and } j \in \mathcal{Y}_n. \tag{33}$$

If Y' is homogeneous, then this property implies that Y' is irreducible and aperiodic, hence also geometrically ergodic. When we fix $n > t$ and define

$$\mathcal{Y}^+ = \{(i,j) \mid i \in \mathcal{Y}_t, j \in \mathcal{Y}_n, P(Y'_n = j | Y'_t = i) > 0\},$$

then (33) means that $(i,j) \in \mathcal{Y}^+$ for all $i \in \mathcal{Y}^+_{(1)} = \mathcal{Y}_t$ and $j \in \mathcal{Y}^+_{(2)} = \mathcal{Y}_n$. Conditionally given $X_{1:n}$, $Y_{1:n}$ is a non-homogeneous Markov chain, and so the assumption (S1) applies the above idea to conditional signal process $P(Y_{1:n} \in \cdot | X_{1:n} = x_{1:n})$. Indeed, note that for every $\mathbf{x} \in \mathcal{X}^q$, $q \geq 2$, we have

$$\mathcal{Y}^+(\mathbf{x}) = \{(i,j) \mid P(Y_{t+q-1} = j | Y_t = i, X_{t:t+q-1} = \mathbf{x}) > 0\}$$

and so (S1) states the following: there exists a set $E \subseteq \mathcal{X}^q$, $q \geq 2$, such that for all $\mathbf{x} \in E$ the set $\mathcal{Y}^+ = \mathcal{Y}^+(\mathbf{x})$ is the same and – analogously to (33) –,

$$P(Y_{t+q-1} = j | Y_t = i, X_{t:t+q-1} = \mathbf{x}) > 0 \text{ for all } i \in \mathcal{Y}^+_{(1)} \text{ and } j \in \mathcal{Y}^+_{(2)}.$$

The condition (S2) can be shown together with Harris recurrence of Z to imply that X enters E infinitely often a.s., and this together with (S1) is used to derive our main result below.

The main theorem The main result of Paper III can now be stated as follows. Instead of the total variation distance (32), we consider more generally the random distributions $P(Y_{t:\infty} \in \cdot | X_{s:n})$ and its total variation distance from $P(Y_{t:\infty} \in \cdot | X_{l:n})$, where $l \leq s \leq t$. Both of these distributions are defined on the cylindrical σ-algebra of $\mathcal{Y}^\infty$. Markov chain Z is called *positive*, if its transition

kernel admits an invariant measure, that is, if there exists a probability measure π on $\mathcal{B}(\mathcal{Z})$ such that $\int_{\mathcal{Z}} P_z(Z_2 \in A)\,\pi(\mathrm{d}z) = \pi(A)$ for all $A \in \mathcal{B}(\mathcal{Z})$. Thus if π is the invariant measure of Z and also the distribution of Z_1, then Z becomes stationary, although we shall not make this assumption. In Paper III the below result is derived as a corollary, but since it is in fact more general then the theorem it is derived from ([III, Th. 3.1]), we state it here as the main theorem.

Theorem 4 [III, Cor. 3.1]. *Assume (S1)-(S2) and let Z be Harris recurrent.*

(i) Then for all $s \geq l \geq 1$

$$\lim_{t\to\infty} \sup_{n\geq t} \|P(Y_{t:\infty} \in \cdot|X_{l:n}) - P(Y_{t:\infty} \in \cdot|X_{s:n})\|_{\mathrm{TV}} = 0, \quad a.s. \tag{34}$$

(ii) If Z is positive, then there exists a constant $\alpha \in (0,1)$ such that the following holds: for every $s \geq 1$ there exist a $\sigma(X_{s:\infty})$-measurable random variable $C_s < \infty$ such that for all $t \geq s \geq l \geq 1$

$$\sup_{n\geq t} \|P(Y_{t:\infty} \in \cdot|X_{l:n}) - P(Y_{t:\infty} \in \cdot|X_{s:n})\|_{\mathrm{TV}} \leq C_s\alpha^{t-s}, \quad a.s. \tag{35}$$

Part (i) of the theorem ensures the forgetting of the smoothing probabilities for any Harris chains satisfying (S1)-(S2), but of main interest is the part (ii) which ensures that the convergence has the exponential rate. The exact definitions of the convergence rate α and the random coefficient C_s can be traced from the proofs in Paper III. Note that both α and C_s are independent of the specific choice of l (as long as $l \leq s$). The additional constraint of (ii) that Z must be positive is not restrictive and can usually be easily verified using the standard tools from the theory of Harris chains. As argued in Paper III, we can also go with n to the infinity, that is, (34) implies

$$\|P(Y_{t:\infty} \in \cdot|X_{l:\infty}) - P(Y_{t:\infty} \in \cdot|X_{s:\infty})\|_{\mathrm{TV}} \underset{t}{\to} 0, \quad \text{a.s.}$$

and (35) implies

$$\|P(Y_{t:\infty} \in \cdot|X_{l:\infty}) - P(Y_{t:\infty} \in \cdot|X_{s:\infty})\|_{\mathrm{TV}} \leq C_s\alpha^{t-s}, \quad \text{a.s.}$$

The main outline for the proof of Theorem 4 follows along the same lines as the proof of a similar HMM-result by Lember in [26]. Namely, the elements of $Y_{1:n}$ are grouped into appropriate blocks and then (S1)-(S2) are applied to obtain an upper bound to the Dobrushin coefficients of the blocked conditional signal process. However, the generalization from HMM to PMM is far from trivial, and several technical issues need to be overcome to obtain the generalized theorem above. Further, as we argue in our paper, even in the special HMM-case, conditions (S1)-(S2) are significantly more lenient than the ones used for the HMM-result of [26].

In Paper III also the forgetting properties for the two-sided stationary extension of Markov chain Z are considered, but these largely analogous to the one-sided case above. The results for two-sided forgetting are mainly motivated by the asymptotic study of the PMAP-classifier, which was mentioned in Section 3.

Hidden Markov model We essentially saw already in Section 5 that the Gaussian linear Markov switching model satisfies (S1)-(S2) when the hidden Markov chain Y is irreducible and aperiodic. More general treatment of the linear Markov switching model is given in Paper III. Here we shall focus on the implications of conditions (S1)-(S2) on HMM. Recall the condition (ii) in Corollary 1: Markov chain Y is irreducible, and there exists a set $C \subseteq \mathcal{Y}$ such that

$$\mu\left[(\cap_{i\in C} G_i) \setminus (\cup_{i\notin C} G_i)\right] > 0 \tag{36}$$

and the sub-stochastic matrix $\mathbb{P}_C = (p_{ij})_{i,j\in C}$ is irreducible and aperiodic. Recall that here $G_i = \{x | f_i(x) > 0\}$. In Paper III this condition is called the "cluster condition", and here the parallels with Paper I re-emerge, because the same condition is in fact sufficient for (S1)-(S2) to hold. Indeed, by the irreducibility and aperiodicity of the substochastic matrix $\mathbb{P}_C$ there must exist $k \geq 1$ such that $\mathbb{P}_C^k$ consists of only positive elements. Defining $\mathcal{Y}_C = \{i \mid p_{ij} > 0, j \in C\}$ and taking

$$E = (\cup_{i\in\mathcal{Y}_C} G_i) \times \left[(\cap_{i\in C} G_i) \setminus (\cup_{i\notin C} G_i)\right]^{k+1},$$

we have that (S1) holds with $\mathcal{Y}^+ = \mathcal{Y}_C \times C$. Observe that $E_{(1)} = \cup_{i\in\mathcal{Y}_C} G_i$, and recall that any HMM with irreducible Y is ψ-irreducible, where

$$\psi(A \times \{i\}) = \mu(A \cap G_i), \quad A \in \mathcal{B}(\mathcal{X}), \quad i \in \mathcal{Y}.$$

Since $\psi(E_{(1)} \times \mathcal{Y}^+_{(1)}) = \mu\big(\cup_{i\in\mathcal{Y}_C} G_i\big) > 0$, we see that by (36) the condition (S2) is also satisfied. Because any HMM with irreducible Y is positive Harris, then the cluster condition immediately implies the exponential forgetting rate of (35). In Paper III a more detailed analysis of the cluster condition is given together with its comparison with several other existing conditions from the literature which are similarly used to ensure the exponential decay of the smoothing probabilities. It turns out that the cluster condition can be shown to be essentially as lenient or strictly more lenient than most other existing conditions adapted to the finite state space.

While very useful, the cluster condition is not necessary for (S1)-(S2), and alternative conditions can be explored. For example in [III, Prop. 4.1] it is proved that when Y is irreducible, then having a row with only positive entries in the transition matrix $\mathbb{P} = (p_{ij})$ of Y is sufficient for (S1)-(S2) to hold. This condition does not assume anything about the emission densities f_i and is very easy to verify directly from the transition matrix $\mathbb{P}$. At the same time, it is not comparable with the cluster condition, because the latter can be satisfied with a zero in every row. Returning again to the parallels with Paper I, it is possible to show that the statement of Corollary 1 (about the existence of the Viterbi process) still holds, if one replaces the cluster condition (ii) with this same assumption that Y is irreducible and its transition matrix contains a row with non-zero entries.

Concluding remarks

Our approach for proving the existence of the Viterbi process is straightforward, and is based on the simple observation that whenever a sequence of observations – called a barrier – occurs in the observation sequence which fixes a Viterbi

estimate at some position, then also the Viterbi path is fixed up to the same position. Thus, when a barrier occurs in the observation sequence infinitely many times, then the Viterbi process can be constructed piecewise. The cyclical construction of the barrier set is then used to tie the existence of the Viterbi process back to the transition kernel density. The non-trivial part of our theory in Paper I is providing the sufficient conditions for such a cyclical construction in the two barrier-construction theorems.

Indeed, as Example 2 demonstrates, the Viterbi process may fail to exist even for an ergodic HMM, and so we know that at least some additional conditions are required. We believe that our barrier-based approach provides a robust and effective method for obtaining such conditions. Another benefit of this approach is that it enables to apply the regeneration-based analysis to prove SLLN and CLT type results for the Viterbi classifier. In this way, the existence of the Viterbi process not only ensures the overall path-stability of the Viterbi estimation, but also becomes a useful tool for obtaining related asymptotic convergence results. A barrier set is not necessary for the existence of the Viterbi process (see [I, Ex. 2]), and so there might be alternative approaches to our barrier-based theory. However, it is unclear if such an approach would be as useful for the asymptotic analysis of the Viterbi classifier (regeneration-based or otherwise).

Disregarding the technical details, it can be said that essentially the same condition which ensures the existence of the cyclical barrier set in case of lower semi-continuous transition densities in Paper I also ensures the exponential forgetting rate of smoothing probabilities in Paper III. This is interesting, because the way this condition is used in the proofs is completely different between both papers. Thus there is reason to believe that this condition could play a more significant role in the study of pairwise Markov models. Some explanation for this was given in Section 7 where we saw that this condition ensures in a way the geometric ergodicity of the conditional signal process.

In all three papers we assume that the hidden state space is finite, and this assumption is relied upon heavily in our proofs. It is generally understood that the leap from finite to continuous hidden state space is non-trivial, so the special focus on the finite space is more than justified. This is not to say that some of the ideas featured in Papers I-III cannot be used in the continuous case, but a separate treatment would be needed for that, and it is unlikely that the resulting conditions would be as broad as they are in the finite case.

Appendix A Reversed Viterbi algorithm

The reversed Viterbi algorithm applied to observations $x_{1:n}$ can be described as follows. Take $\delta_n^*(i) \equiv 1$, and calculate for all $k = n-1, \ldots, 1$

$$\begin{aligned}
\delta_k^*(i) &= \max_{j \in \mathcal{Y}} q(x_{k+1}, j | x_k, i) \delta_{k+1}^*(j), \\
\gamma_k^*(i) &= \arg\max_{j \in \mathcal{Y}} q(x_{k+1}, j | x_k, i) \delta_{k+1}^*(j).
\end{aligned}$$

Thus $\delta_k^*(i) = q(x_{k+1}, \gamma_k^*(i) | x_k, i) \delta_{k+1}^*(\gamma_k^*(i))$. Now the Viterbi path $v_{1:n} = v(x_{1:n})$ can be calculated as follows:

$$\begin{aligned}
v_1 &= \arg\max_{i \in \mathcal{Y}} p(x_1, y_1 = i) \delta_1^*(i), \\
v_2 &= \gamma_1^*(v_1), \\
v_3 &= \gamma_2^*(v_2), \\
&\vdots \\
v_n &= \gamma_{n-1}^*(v_{n-1}).
\end{aligned}$$

If each arg max is applied based on the same ordering on $\mathcal{Y}$, then this algorithm returns the lexicographically first path (among all paths with maximal likelihood) based on the same ordering.

References

[I] J. Lember and J. Sova. "Existence of infinite Viterbi path for pairwise Markov models". *Stochastic Processes and their Applications* 130.3 (2020), pp. 1388–1425.

[II] J. Lember and J. Sova. "Regenerativity of viterbi process for pairwise markov models". *Journal of Theoretical Probability* 34.1 (2021), pp. 1–33.

[III] J. Lember and J. Sova. "Exponential forgetting of smoothing distributions for pairwise Markov models". *Electronic Journal of Probability* (2021), to appear.

[1] O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer, 2005.

[2] J. D. Hamilton. "A new approach to the economic analysis of nonstationary time series and the business cycle". *Econometrica: Journal of the Econometric Society* (1989), pp. 357–384.

[3] J. D. Hamilton. "Analysis of time series subject to changes in regime". *Journal of econometrics* 45.1-2 (1990), pp. 39–70.

[4] J. D. Hamilton. "Regime switching models". *Macroeconometrics and time series analysis*. Springer, 2010, pp. 202–209.

[5] W. Pieczynski. "Pairwise Markov chains". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.5 (2003), pp. 634–639.

[6] S. Derrode and W. Piecynski. "Signal and image segmentation using pairwise Markov chains". *IEEE Transactions on Signal Processing* 52.9 (2004), pp. 2477–2489.

[7] S. Derrode and W. Piecynski. "Unsupervised data classification using pairwise Markov chains with automatic copula selection". *Computational Statistics and Data Analysis* 63 (2013), pp. 81–98.

[8] P. Lanchantin, J. Lapuyade-Lahorgue, and W. Pieczynski. "Unsupervised segmentation of randomly switching data hidden with non-Gaussian correlated noise". *Signal Processing* 91.2 (2011), pp. 163–175.

[9] W. Pieczynski. "Pairwise markov chains". *IEEE Transactions on pattern analysis and machine intelligence* 25.5 (2003), pp. 634–639.

[10] S. P. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2009.

[11] J. Lember and A. A. Koloydenko. "Bridging Viterbi and posterior decoding: a generalized risk approach to hidden path inference based on hidden Markov models". *The Journal of Machine Learning Research* 15.1 (2014), pp. 1–58.

[12] J. Lember, K. Kuljus, and A. Koloydenko. "Theory of segmentation". *Hidden Markov Models, Theory and Applications* (2011), pp. 51–84.

[13] Y. Ephraim and N. Merhav. "Hidden markov processes". *IEEE Transactions on information theory* 48.6 (2002), pp. 1518–1569.

[14] V. Genon-Catalot, T. Jeantheau, C. Larédo, et al. "Stochastic volatility models as hidden Markov models and statistical applications". *Bernoulli* 6.6 (2000), pp. 1051–1079.

[15] B. G. Leroux. "Maximum-likelihood estimation for hidden Markov models". *Stochastic processes and their applications* 40.1 (1992), pp. 127–143.

[16] A. Caliebe and U. Rösler. "Convergence of the maximum a posteriori path estimator in hidden Markov models". *IEEE Transactions on Information Theory* 48.7 (2002), pp. 1750–1758.

[17] A. Caliebe. "Properties of the maximum a posteriori path estimator in hidden Markov models". *IEEE Transactions on Information Theory* 52.1 (2006), pp. 41–51.

[18] J. Lember and A. Koloydenko. "A constructive proof of the existence of Viterbi processes". *IEEE Transactions on Information Theory* 56.4 (2010), pp. 2017–2033.

[19] A. Koloydenko and J. Lember. "Infinite Viterbi alignments in the two state hidden Markov models". *Acta et Commentationes Universitatis Tartuensis de Mathematica* 12 (2008), pp. 109–124.

[20] P. Chigansky and Y. Ritov. "On the Viterbi process with continuous state space". *Bernoulli* 17.2 (2011), pp. 609–627.

[21] N. Whiteley, M. W. Jones, and A. P. Domanski. "The Viterbi process, decay-convexity and parallelized maximum a-posteriori estimation". *arXiv preprint arXiv:1810.04115* (2018).

[22] V. V. Kalashnikov. *Topics on regenerative processes*. CRC Press, 1994.

[23] H. Thorisson. *Coupling, stationarity, and regeneration*. Vol. 200. 0. Springer New York, 2000.

[24] W. Hoeffding, H. Robbins, et al. "The central limit theorem for dependent random variables". *Duke Mathematical Journal* 15.3 (1948), pp. 773–780.

[25] J. P. Romano and M. Wolf. "A more general central limit theorem for m-dependent random variables with unbounded m". *Statistics & probability letters* 47.2 (2000), pp. 115–124.

[26] J. Lember. "On approximation of smoothing probabilities for hidden Markov models". *Statistics & probability letters* 81.2 (2011), pp. 310–316.

[27] J. Lember, H. Matzinger, J. Sova, and F. Zucca. "Lower bounds for moments of global scores of pairwise Markov chains". *Stochastic Processes and their Applications* 128.5 (2018), pp. 1678–1710.

Eestikeelne sisukokkuvõte

Paariviisi Markovi ahelad

Varjatud muutujatega Markovi mudelid on kaasaegse statistika suur edulugu. Tänapäeval on üha suurem vajadus analüüsida keerulise struktuuriga andmeid, mis ei järgi klassikalise statistika eeldusi, nagu valimi liikmete sõltumatus ja sama jaotus, ning mille puhul klassikalise statistika meetodid jäävad ebaefektiivseks. Teisalt varjatud muutujatega Markovi mudelid võimaldavad rakendada erinevaid kergesti kohandatavaid meetodeid analüüsimaks keerulisi omavahel sõltuvaid andmeid. Käesolev doktoritöö uurib laia selliste mudelite klassi nimega „paariviisi Markovi mudelid“ (PMM). PMM on lihtsalt mistahes varjatud muutujatega mudel, mille puhul varjatud ehk latentne kiht ning vaatluste kiht koos moodustavad Markovi ahela. PMM hõlmab väga laia mudelite hulka, kuid enimtuntud ja praktikas kõige rohkem rakendatud on kindlasti varjatud Markovi mudel (ingl. k. hidden Markov model). Viimase näol on tegemist PMM-i erijuhuga, mille puhul vaatlused sõltuvad üksteisest ainult läbi mudeli varjatud kihi.

Käesolev doktoritöö annab ülevaate kolmest artiklist, mis kõik käsitlevad PMM-ide teatud aspekte. Tihti on PMM-ide rakendamise eesmärk hinnata vaatluste põhjal mudeli varjatud kihti. Tõenäoliselt praktikas enimlevinud meetod selleks on kuulus “Viterbi algoritm”. See algoritm leiab valimi suuruse suhtes lineaarse ajaga PMM-i varjatud kihile suurima tõepära hinnangu. Vastavat hinnangut tuntakse ka kui „Viterbi joondust“ või „Viterbi rada“ (ingl. k. Viterbi alignment või Viterbi path). Viterbi joondus maksimeerib tõenäosust, et kogu varjatud kihi hinnang on õige. Samas Viterbi joondus üldiselt ei maksimeeri keskmist õigesti hinnatud elementide arvu – seda teeb nn PMAP-hinnang (ingl. k. pointwise maximum a posteriori estimate).

Artiklite I ja II kesksekseks teemaks on Viterbi joonduse stohhastiline stabiilsus. Viterbi joonduse käitumine vaatluste arvu kasvades ei ole triviaalne küsimus, kuna iga element Viterbi joonduses sõltub üldiselt kõikidest vaatlustest. Seega ühe vaatluse juurde lisamine võib mõjutada ka sellele vaatlusele eelnevaid Viterbi joonduse elemente. Samas praktikas see mõju näib olevat üpris lokaliseeritud mudeli lõpuosasse ning vaatluste arvu kasvades Viterbi joonduse eesosa üldjuhul stabiliseerub kiiresti fikseeritud seisundisse. Artiklis I anname sellele fenomenile teoreetilise põhjenduse: nimelt tõestame, et üldistel tingimustel vaatluste arvu kasvades Viterbi joondus koondub piirprotsessiks, mida nimetame „Viterbi protsessiks“. Tõestuse idee põhineb tähelepanekul, et vaatluste jadas võib esineda blokke, mis tagavad, et Viterbi joondus kuni selle blokini on fikseeritud. Me nimetame selliseid blokke „barjäärideks“ ning tõestame, et teatud tingimustel vaatluste jada sisaldub tõenäosusega 1 lõpmatult palju barjääre, tagades omakorda Viterbi protsessi olemasolu.

Artiklis II arendatakse artikli I teooriat edasi ning uuritakse Viterbi joonduse asümptootilisi omadusi. Isegi kui Viterbi protsessi olemasolu on garanteeritud, on selle protsessi tõenäosuslik struktuur väga keeruline, ning seega edasised järeldused Viterbi joonduse asümptootika kohta ei ole ilmsed. Näitame, et üldistel tingimustel on võimalik konstrueerida nn „regeneratsiooniajad“ (ingl. k. regeneration times), mis jagavad PMM-i sõltumatuteks tükkideks ning samas langevad kokku ka barjääridega. See omakorda võimaldab tuletada Viterbi protsessile suurte arvude seadusi ja tsentraalseid piirteoreemi. Standardne teh-

nika Markovi mudelite puhul regeneratsiooniaegade konstrueerimiseks on nn lõhestusmeetod (ingl. k. splitting method). Artikli II põhitulemus näitab, et teatud tingimustel saab lõhestusmeetodit rakendada nii, et vastavad regeneratsiooniajad langevad kokku barjääridega.

Artikli III temaatika erineb kahest eelmisest ning käsitleb PMM-i varjatud kihi jaotust tinglikkustatuna üle vaatluste. Täpsemalt uurime PMM-i silumistõenäosusi (ingl. k. smoothing probabilities) ja näitame, et need teatud mõttes „unustavad“ vaatluste jada esimesed väärtused. Tõestame, et selle unustamise kiirus on teatud tingimustel juhusliku koefitsiendi suhtes eksponentsiaalne. Silumistõenäosuste unustusomadused on varjatud muutujatega Markovi mudelite teoorias keskse tähtsusega ning seetõttu on sellele teemale pühendatud mitmeid teadusartikleid. Näitame, et lõpliku arvuga varjatud seisundite kontekstis on meie tingimused sisuliselt sama leebed või rangemalt leebemad kui teised kirjanduses esinevad tingimused.

Publications

Curriculum Vitae (English)

Personal information

Name	Joonas Sova
Date of birth	June 1st , 1987
Citizenship	Estonian
E-mail	joonas.sova@gmail.com
Phone	+372 56 509 777
Residence	Tartu, Estonia

Education

2015-...	Mathematical Statistics, Ph.D, *University of Tartu*
2013-2015	Mathematical Statistics, M.Sc, *University of Tartu*
2009-2013	Mathematical Statistics, B.Sc, *University of Tartu*

Employment

2018-...	Statistical programmer, *IQVIA*
2011-2013	Freelancer translator
2010-2011	Cook's assistant

List of publications

J. Lember, H. Matzinger, J. Sova, and F. Zucca. "Lower bounds for moments of global scores of pairwise Markov chains". *Stochastic Processes and their Applications* 128.5 (2018), pp. 1678–1710

J. Lember and J. Sova. "Existence of infinite Viterbi path for pairwise Markov models". *Stochastic Processes and their Applications* 130.3 (2020), pp. 1388–1425

J. Lember and J. Sova. "Regenerativity of viterbi process for pairwise markov models". *Journal of Theoretical Probability* 34.1 (2021), pp. 1–33

J. Lember and J. Sova. "Exponential forgetting of smoothing distributions for pairwise Markov models". *Electronic Journal of Probability* (2021), to appear

Curriculum Vitae (eesti keeles)

Isiklik teave

Nimi	Joonas Sova
Sünniaeg	01.06.1987
Kodakondsus	eestlane
E-mail	joonas.sova@gmail.com
Telefon	+372 56 509 777
Elukoht	Tartu, Estonia

Haridus

2015-...	matemaatiline statistika, Ph.D, *Tartu Ülikool*
2013-2015	matemaatiline statistika, M.Sc, *Tartu Ülikool*
2009-2013	matemaatiline statistika, B.Sc, *Tartu Ülikool*

Töökogemus

2018-...	statistiline programmeerija, *IQVIA*
2011-2013	vabakutseline tõlkija
2010-2011	koka abi

Publikatsioonid

J. Lember, H. Matzinger, J. Sova ja F. Zucca. "Lower bounds for moments of global scores of pairwise Markov chains". *Stochastic Processes and their Applications* 128.5 (2018), lk. 1678–1710

J. Lember ja J. Sova. "Existence of infinite Viterbi path for pairwise Markov models". *Stochastic Processes and their Applications* 130.3 (2020), lk. 1388–1425

J. Lember ja J. Sova. "Regenerativity of viterbi process for pairwise markov models". *Journal of Theoretical Probability* 34.1 (2021), lk. 1–33

J. Lember ja J. Sova. "Exponential forgetting of smoothing distributions for pairwise Markov models". *Electronic Journal of Probability* (2021), to appear

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. **Mati Heinloo.** The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu, 1991, 23 p.
2. **Boris Komrakov.** Primitive actions and the Sophus Lie problem. Tartu, 1991, 14 p.
3. **Jaak Heinloo.** Phenomenological (continuum) theory of turbulence. Tartu, 1992, 47 p.
4. **Ants Tauts.** Infinite formulae in intuitionistic logic of higher order. Tartu, 1992, 15 p.
5. **Tarmo Soomere.** Kinetic theory of Rossby waves. Tartu, 1992, 32 p.
6. **Jüri Majak.** Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu, 1992, 32 p.
7. **Ants Aasma.** Matrix transformations of summability and absolute summability fields of matrix methods. Tartu, 1993, 32 p.
8. **Helle Hein.** Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu, 1993, 28 p.
9. **Toomas Kiho.** Study of optimality of iterated Lavrentiev method and its generalizations. Tartu, 1994, 23 p.
10. **Arne Kokk.** Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu, 1995, 165 p.
11. **Toomas Lepikult.** Automated calculation of dynamically loaded rigid-plastic structures. Tartu, 1995, 93 p, (in Russian).
12. **Sander Hannus.** Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74 p, (in Russian).
13. **Sergei Tupailo.** Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu, 1996, 134 p.
14. **Enno Saks.** Analysis and optimization of elastic-plastic shafts in torsion. Tartu, 1996, 96 p.
15. **Valdis Laan.** Pullbacks and flatness properties of acts. Tartu, 1999, 90 p.
16. **Märt Põldvere.** Subspaces of Banach spaces having Phelps' uniqueness property. Tartu, 1999, 74 p.
17. **Jelena Ausekle.** Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu, 1999, 72 p.
18. **Krista Fischer.** Structural mean models for analyzing the effect of compliance in clinical trials. Tartu, 1999, 124 p.
19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
20. **Jüri Lember.** Consistency of empirical k-centres. Tartu, 1999, 148 p.
21. **Ella Puman.** Optimization of plastic conical shells. Tartu, 2000, 102 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.

23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
25. **Maria Zeltser.** Investigation of double sequence spaces by soft and hard analitical methods. Tartu, 2001, 154 p.
26. **Ernst Tungel.** Optimization of plastic spherical shells. Tartu, 2001, 90 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 p.
28. **Rainis Haller.** *M(r,s)*-inequalities. Tartu, 2002, 78 p.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
30. Töö kaitsmata.
31. **Mart Abel.** Structure of Gelfand-Mazur algebras. Tartu, 2003. 94 p.
32. **Vladimir Kuchmei.** Affine completeness of some ockham algebras. Tartu, 2003. 100 p.
33. **Olga Dunajeva.** Asymptotic matrix methods in statistical inference problems. Tartu 2003. 78 p.
34. **Mare Tarang.** Stability of the spline collocation method for volterra integro-differential equations. Tartu 2004. 90 p.
35. **Tatjana Nahtman.** Permutation invariance and reparameterizations in linear models. Tartu 2004. 91 p.
36. **Märt Möls.** Linear mixed models with equivalent predictors. Tartu 2004. 70 p.
37. **Kristiina Hakk.** Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 137 p.
38. **Meelis Käärik.** Fitting sets to probability distributions. Tartu 2005, 90 p.
39. **Inga Parts.** Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140 p.
40. **Natalia Saealle.** Convergence and summability with speed of functional series. Tartu 2005, 91 p.
41. **Tanel Kaart.** The reliability of linear mixed models in genetic studies. Tartu 2006, 124 p.
42. **Kadre Torn.** Shear and bending response of inelastic structures to dynamic load. Tartu 2006, 142 p.
43. **Kristel Mikkor.** Uniform factorisation for compact subsets of Banach spaces of operators. Tartu 2006, 72 p.
44. **Darja Saveljeva.** Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 117 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
46. **Annely Mürk.** Optimization of inelastic plates with cracks. Tartu 2006. 137 p.
47. **Annemai Raidjõe.** Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97 p.
48. **Olga Panova.** Real Gelfand-Mazur algebras. Tartu 2006, 82 p.

49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
50. **Margus Pihlak.** Approximation of multivariate distribution functions. Tartu 2007, 82 p.
51. **Ene Käärik.** Handling dropouts in repeated measurements using copulas. Tartu 2007, 99 p.
52. **Artur Sepp.** Affine models in mathematical finance: an analytical approach. Tartu 2007, 147 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
54. **Kaja Sõstra.** Restriction estimator for domains. Tartu 2007, 104 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
57. **Evely Leetma.** Solution of smoothing problems with obstacles. Tartu 2009, 81 p.
58. **Ants Kaasik.** Estimating ruin probabilities in the Cramér-Lundberg model with heavy-tailed claims. Tartu 2009, 139 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
60. **Indrek Zolk.** The commuting bounded approximation property of Banach spaces. Tartu 2010, 107 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
63. **Marek Kolk.** Piecewise Polynomial Collocation for Volterra Integral Equations with Singularities. Tartu 2010, 134 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
65. **Larissa Roots.** Free vibrations of stepped cylindrical shells containing cracks. Tartu 2010, 94 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
68. **Olga Liivapuu.** Graded q-differential algebras and algebraic models in noncommutative geometry. Tartu 2011, 112 p.
69. **Aleksei Lissitsin.** Convex approximation properties of Banach spaces. Tartu 2011, 107 p.
70. **Lauri Tart.** Morita equivalence of partially ordered semigroups. Tartu 2011, 101 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.

72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.
74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
75. **Nadežda Bazunova.** Differential calculus $d^3 = 0$ on binary and ternary associative algebras. Tartu 2011, 99 p.
76. **Natalja Lepik.** Estimation of domains under restrictions built upon generalized regression and synthetic estimators. Tartu 2011, 133 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
80. **Marje Johanson.** $M(r, s)$-ideals of compact operators. Tartu 2012, 103 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
82. **Vitali Retšnoi.** Vector fields and Lie group representations. Tartu 2012, 108 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
85. **Erge Ideon**. Rational spline collocation for boundary value problems. Tartu, 2013, 111 p.
86. **Esta Kägo.** Natural vibrations of elastic stepped plates with cracks. Tartu, 2013, 114 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
88. **Boriss Vlassov.** Optimization of stepped plates in the case of smooth yield surfaces. Tartu, 2013, 104 p.
89. **Elina Safiulina.** Parallel and semiparallel space-like submanifolds of low dimension in pseudo-Euclidean space. Tartu, 2013, 85 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
93. **Kerli Orav-Puurand.** Central Part Interpolation Schemes for Weakly Singular Integral Equations. Tartu, 2014, 109 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multiparty computation. Tartu, 2015, 201 p.

95. **Kaido Lätt.** Singular fractional differential equations and cordial Volterra integral operators. Tartu, 2015, 93 p.
96. **Oleg Košik.** Categorical equivalence in algebra. Tartu, 2015, 84 p.
97. **Kati Ain.** Compactness and null sequences defined by ℓ_p spaces. Tartu, 2015, 90 p.
98. **Helle Hallik.** Rational spline histopolation. Tartu, 2015, 100 p.
99. **Johann Langemets.** Geometrical structure in diameter 2 Banach spaces. Tartu, 2015, 132 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
105. **Md Raknuzzaman.** Noncommutative Galois Extension Approach to Ternary Grassmann Algebra and Graded q-Differential Algebra. Tartu, 2016, 110 p.
106. **Alexander Liyvapuu.** Natural vibrations of elastic stepped arches with cracks. Tartu, 2016, 110 p.
107. **Julia Polikarpus.** Elastic plastic analysis and optimization of axisymmetric plates. Tartu, 2016, 114 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.
113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
115. **Tiina Kraav.** Stability of elastic stepped beams with cracks. Tartu, 2017, 126 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.

117. **Silja Veidenberg.** Lifting bounded approximation properties from Banach spaces to their dual spaces. Tartu, 2017, 112 p.
118. **Liivika Tee.** Stochastic Chain-Ladder Methods in Non-Life Insurance. Tartu, 2017, 110 p.
119. **Ülo Reimaa.** Non-unital Morita equivalence in a bicategorical setting. Tartu, 2017, 86 p.
120. **Rauni Lillemets.** Generating Systems of Sets and Sequences. Tartu, 2017, 181 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.
123. **Kaur Lumiste.** Improving accuracy of survey estimators by using auxiliary information in data collection and estimation stages. Tartu, 2018, 112 p.
124. **Paul Tammo.** Closed maximal regular one-sided ideals in topological algebras. Tartu, 2018, 112 p.
125. **Mart Kals.** Computational and statistical methods for DNA sequencing data analysis and applications in the Estonian Biobank cohort. Tartu, 2018, 174 p.
126. **Annika Krutto.** Empirical Cumulant Function Based Parameter Estimation in Stable Distributions. Tartu, 2019, 140 p.
127. **Kristi Läll.** Risk scores and their predictive ability for common complex diseases. Tartu, 2019, 118 p.
128. **Gul Wali Shah.** Spline approximations. Tartu, 2019, 85 p.
129. **Mikk Vikerpuur.** Numerical solution of fractional differential equations. Tartu, 2019, 125 p.
130. **Priit Lätt.** Induced 3-Lie superalgebras and their applications in superspace. Tartu, 2020, 114 p.
131. **Sumaira Rehman.** Fast and quasi-fast solvers for weakly singular Fredholm integral equation of the second kind. Tartu, 2020, 105 p.
132. **Rihhard Nadel.** Big slices of the unit ball in Banach spaces. Tartu, 2020, 109 p.
133. **Katriin Pirk.** Diametral diameter two properties, Daugavet-, and Δ-points in Banach spaces. Tartu, 2020, 106 p.
134. **Zahra Alijani**. Fuzzy integral equations of the second kind. Tartu, 2020, 103 p.
135. **Hina Arif.** Stability analysis of stepped nanobeams with defects. Tartu, 2021, 165 p.