Forgetting Properties of Finite-State Reciprocal Processes

Riley Bruce-Doust

A thesis presented for the degree of
Master of Philosophy

School of Electrical and Electronic Engineering
University of Adelaide
September 2017

Contents

Abstract

Reciprocal chains (RC) are a class of discrete-index, finite-state stochastic process having the non-causal generalisation of the Markov property, where rather than the future being conditionally independent of the past given the present, intervals are conditionally independent of their complement given their endpoints. RCs are more powerful models than Markov chains (MC) but are n times more complex to process with their associated smoothing algorithm for number of states n, that is $\mathcal{O}(n^3T)$ for fixed interval smoothing an interval of length T. In this thesis it was established that RCs have a forgetting property - geometric decay in dependence between separated variables: it was found that the dependence matrix between two variables in an RC has a form expressible in terms of element-wise product of matrix products. The theory of forgetting in matrix products using Birkhoff's contraction coefficient, was extended to this case. It was shown that because MCs are a special case of RCs, arising under particular boundary conditions, forgetting property means the distributions of the RC that are far from the boundary condition are well approximated by those of MC models. It was shown that the forgetting property extends to the RC fixed interval smoothing algorithm so that close approximation occurs by estimates from the MC interval smoother for variables far from the boundary. These results would imply that RCs were not computationally efficient models for intervals that are long with respect to the forgetting rate, however an approximate interval smoothing algorithm was developed for RCs which is a modified form of the MC algorithm (the forward-backward algorithm) and is of comparable complexity to it. The forgetting theory was used to bound the error in the approximation, which is small on the long intervals for which the RC was inefficient for exact estimation.

Declaration

I, Riley Bruce-Doust, certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree. I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signature ..... Date .....

Acknowledgement

We thank Professor Lang White for theoretical and practical instruction and Mr. George Stamatescu for constructive discussion.

Chapter 1

Introduction

A random process is a set of random variables on an ordered index - usually representing time - used to model some evolving signal with a random component. Questions that can be answered with process models include prediction, where the observed history of the process informs the future; and estimation of a 'hidden' process, in which the signal is not directly observed but a (possibly future) history of observations statistically correlated with it are combined according to the process model to give the posterior probability distribution (estimate) of the process' current value. Random process can be classed by the rule by which variables relate to others as a function of their relative positions on the index. Ubiquitous in many application areas, is the class of models with the Markov property - that the 'future' of the process, given the present value, is conditionally independent of the process' past. Markovianity is akin to memorylessness; there can be defined a 'state', the value of which contains everything relevant about the past. A discrete-index Markov process is called a Markov chain (MC). In this thesis use of 'chain' for models will imply discrete-variable as well as discrete-index. A random walk on the integers is an example of an MC: at each time the walker draws its new state from a probability distribution that depends on the present state, for example moving to either neighbouring integer with equal chance. One reason for the ubiquity of the Markov assumption is that this recursive dependence structure and the splitting of the data into a past and future history, allows fast evaluation of marginals and estimates by efficient matrix-vector calculations. It also happens that if a random walker can occupy only a finite set, after a large number of steps the position of the walker's position will not 'say anything' about where the walker started. In this way the recursive structure and limited state space of finite-state MCs means that as time elapses, the process' probability distribution, i.e. predictions and estimates, become independent of past distributions at a geometric rate. Such 'forgetting' is actually a property of non-negative matrix products. In MCs this property is desirable as it means error due to an incorrect initialisation is limited to the early part of the sequence, and furthermore, that posterior distributions of hidden Markov chain (HMC) variables can be obtained which approach the best possible estimate using only 'recent' (past and future) observations, meaning faster estimation still.

A more recently conceived class is reciprocal processes, which generalise the Markov property to be two-sided. The reciprocal property requires only that given the values of the endpoints of an interval, the interior region is conditionally independent of the outside. Markov models are reciprocal but reciprocal models are generally not Markov. A discrete-index, discrete-state reciprocal model has been called a reciprocal chain (RC). The RC counterpart to the random walk example involves three adjacent times where the present state is drawn from a distribution given the immediate past and future position. Reciprocals are thus not recursive with time but by suitable arithmetic, the models can be 'run forward', i.e. used to predict the future distribution given the past. Reciprocals can be used in place of a Markov model when modelling a fixed interval. Unlike a Markov model, the reciprocal interval's endpoints are allowed to be dependent given the interior values. The forward-running RC walk could be a walker that still takes steps only to neighbouring integers but over the length of some interval, induces a different relationship than would result from Markov walking. For this reason RCs have recently been proposed as useful for modelling a simple behavioural notion of 'destination awareness' of moving targets, particularly ground vehicle tracking by radar, where target position evolves on a finite state space such as network of road junctions. Markovian (memorylessness) dynamics on a network can constrain movement to be between connected areas but cannot capture the tendency of targets to proceed purposely through the network, rather than doubling back on their path - without this constraint, a tracking algorithm is more easily detached from its target onto another vehicle crossing the target's path. A target can be modelled as being 'aware' of a destination by conditioning a random walk (Markovian) model to take a particular value at a future time T. To model proceeding between varied origins and destinations with a single process model requires a model that is non-Markov. In fact 'pinning' (conditioning) a Markov model to take values at the endpoints of an interval and then creating a mixture of the pinned processes according to a joint distribution on the endpoints results in a (generally non-Markov) reciprocal model. Such a reciprocal model is referred to as being generated from a Markov base. Two models which share a base share 'reciprocal dynamics' (the three-time transitions) and the 'difference' between the models, that manifests in the amount of error from using one to model the other, is proportional to the difference between the boundary conditions. The reciprocal model is Markov if the joint endpoint distribution applied matches the relationship implied by the base model. An algorithm has been developed for interval smoothing a hidden RC (HRC) (estimation of the distributions of all variables n an interval, using all observations received during the interval) so that RC models can be deployed in applications. RC models are n times more complex (where n is the number of states), and computationally costly to estimate per time step, than MC models.

The forgetting property associated with Markov chains is an example of a more general notion that can be called locality where, in a model of a collection of variables,

variables that are distant in some sense weakly influence each other. In MCs the relevant distance is index (usually time) distance, hence the term forgetting. For (time) process models, 'forgetting' could be defined in the strictest sense to mean geometric decay with index distance, however we will use the terms forgetting and locality, somewhat interchangeably, to mean there is some distance that can be defined for a model with which the influence between variables decays geometrically. An example of models that have locality/forgetting in this more general sense is the class of continuous variable (such as Gaussian) reciprocal processes. In these, the dependence between variables is a linear combination of a geometric decay with index distance, a term decaying with the length of the interval complement. This effectively means the influence, for example of the boundary condition on the distribution at time t, decays geometrically with $\min(t, T-t)$. In finite variable models, the presence of a rule of conditional independence given nearest-neighbours (like the reciprocal property) does not guarantee any forgetting - two-index (e.g. 2-dimensional spatial) models can have 'long-range' (never decaying) dependence. For RCs, there is no theory. The question of whether RCs have locality matters because on the one hand, if the purpose of using an RC model in an application is to model a walker that moves from initial states to likely end-states over an interval, then if there are variables whose distributions are unaffected by either the source or the destination, the model is not efficient, in some sense. On the other hand a locality property can be desirable: estimation algorithms involve marginalisation of joint distributions, which generally is combinatorial, so approximations are desirable to make estimation tractable/fast, and a locality property is associated with the success of the approximations. RC model structure means estimation is not combinatorial, only $\mathcal{O}(n^3T)$ is required, but n may be large so we still would like to know if we could do better. A locality/forgetting property raises its own question for RCs that are generated from a Markovian base model. Since the a Markovian model results from certain cases of endpoint choice, forgetting the boundary condition would imply distributions in common with, thus well approximated by an MC model. This would mean for employing an RC smoother on data, while the beginning / end relationship is now well modelled, the estimates for what could be many variables in between are inefficiently calculated by the RC model. However since as mentioned, the locality property is conducive to approximate estimation, the MC result matching hints at how this calculation could be reduced.

In this thesis we treat finite-state RCs that are constructed from a set of MC base transitions. It can be shown (as we discuss in the Appendix) that base transitions can be found for RCs under mild conditions, but the case of most application interest, and that we assume, is where these transitions are directly available. We confirm that finite-state RCs have the same form of locality as the Gaussian ones: we find that the dependence matrix between two variables in an RC has a form expressible in terms of an element-wise product of two matrix products. We are able to extend the properties of the matrix product theory (Birkhoff's contraction coefficient) to this case. The locality

means, in terms of a time index advancing, that a reciprocal walker on a long interval will forget its initial position, until it nears the end of the interval where it will start to tend toward a destination that is compatible with the initial position, according to the 'long-term transition' contained in the end-point joint distribution. The rate of decay is the base model's forgetting rate. The impact of an incorrect initial state specification is limited, not as in Markov chains to the early part of the interval, but to near both ends of the interval. So too is the impact of incorrectly specifying the entire boundary condition. However, this possibly desirable trait is overshadowed by the confirmation that, the distributions, can be well approximated by those of base MC model on a long enough interval. The MC model can be initialised with the RC's initial distribution, and as the index increases, the MC distributions diverge from those of the RC when approaching the end of the interval, where the RC is again influenced by the initial distribution and the MC is not. When we check how the dynamics (that is, conditional distributions given some past values) of an RC compare to the MC base and find that being far from the boundary is sufficient but not necessary for compatibility with a Markov model. The conditional distribution at some time given a set of past variables, reduces to the distribution given the latest variable (i.e. the Markov property) if the length of the complement, with respect to the interval, of the index separation between the present and the earliest past variable is long relative to the forgetting rate. It is known that a reciprocal model on an infinite interval is Markov. This result confirms that an RC on an finite interval can behave in a 'Markovian' like way within a local region.

The results listed so far would seem to suggest RCs only model the additional information they contain with respect to MCs efficiently when the interval length is on the order of the time in which the base model forgets its initialisation (to whichever level considered significant), however we are able to achieve a resolution in the estimation context. We first confirm that locality results extend to the HRC estimation algorithm by deriving a matrix product form for the algorithm and showing that the estimates in the mid-interval approach those due to the HMC estimation algorithm using the base transitions. The forgetting rate is the HMC algorithm's forgetting rate (with the base transitions). Using the locality property and reflecting the idea of RCs being locally Markov, we find that approximate estimation of HRCs by the a modified version of the HMC interval smoothing algorithm (the forward-backward algorithm), which treats the late part of the interval as near the interval's start, and vice-versa, can produce accurate estimates precisely in the case that the interval is long with respect to the forgetting rate. The conclusion of this thesis is that RC models can be used to efficiently model not only short intervals (with respect to the forgetting rate) but a long interval, and its end point relationship, with comparable computational complexity to the MC models that cannot model the end point relationship.

1.1 Thesis Work

The contents of the thesis is as follows, and we consider the labelled lemmas/theorems etc. to be its main contributions:

Chapter 2 A review of relevant results in process models, estimation, and forgetting properties, with technical details needed for the new work.

Chapter 3 This chapter contains results about RC models. Lemma 3.2 extends the properties of the Birkhoff contraction coefficient to give a bound under the element-wise product. Theorem 3.1 finds an RC's variable interdependence is bounded by an expression which is the the sum of a decay with index distance terms and a term decaying with the length of the interval complement. Theorem 3.2 finds that the marginals of an RC model become better approximated by those of the base MC model with distance from the end of the interval. Theorem 3.3 finds the marginals of an RC in the middle process, forget the entire boundary condition. Theorem 3.4 bounds the difference between the conditional marginal of an RC model given one past variable and given several, as a function the complement of the distance from the first variable in the set to the marginalised variable, with respect to the interval length. Numerical examples are provided which show the bounds are tight.

Chapter 4 This chapter contains results about forgetting in the HRC estimation context. We show that the HRC estimation algorithm can be put into a matrix form similar to that of the RC itself. Theorem 4.1 shows that with increasing distance from the endpoints, estimates of the HRC and HMC algorithms become similar at a geometric rate. An algorithm for approximate fixed-point smoothing (the posterior distribution of one variable given all observations on an interval) of HRCs is derived, and also an algorithm for fixed-interval smoothing an HRC, which has complexity on the order of the HMC smoother. Bounds on error (relative to the exact smoother distributions) are given and demonstrated numerically.

Appendix An appendix is included which discusses further related work that can be applied to reciprocal models. It is demonstrated how an existing theorem implies that sets of base model transitions exist for all positive RCs, and an existing approximate estimation algorithm which can be applied to RCs is compared to this thesis' algorithms.

1.2 Bibliographic Note

- A paper in preparation deals with the material of Chapters 3 and 4.

- G. Stamatescu and L. B. White and R. Bruce-Doust, Track Extraction with Hidden Reciprocal Chains. IEEE Transactions on Automatic Control. 2017 (accepted 18 Aug.) DOI: 10.1109/TAC.2017.2741919. Simulates the application of RC models in tracking algorithms. This paper involves some work outside the forgetting theme of this thesis. Some of its findings are discussed in the background chapter.

Chapter 2

Background

This chapter is a combined technical background and literature review. Our domain is stochastic process models, and since these models are so often (in the signal processing context) employed for the purpose of process estimation, we first in Section 1 set the problem of Bayesian discrete-index process estimation, and use this to introduce the common Markov assumption and hidden Markov models (HMM) and finite-state hidden Markov chains (HMC) which are the natural comparison to (hidden) reciprocal chains (HRC). We recount some of the properties of Markov models and their estimation algorithms that stem from the model structure - fast marginalisation (by matrix-vector multiplication) and geometric rate forgetting. We give a detailed derivation of the forgetting property proof as the method is one we will extend to new models. We then touch on how estimation is applied in the tracking problem to show how tracking's requirements can strain the capability of models utilising the Markov assumption. In the Section 2 we recount the nature of the reciprocal chain, a non-Markov signal model that can be used in place of MCs, including being generated by adding a long term information to a given MC model, but which require their own more computationally expensive estimation algorithm. We mention where and why RC models are starting to be employed in tracking applications to redress Markov models' limitations. We then discuss how the question of the forgetting behaviour of RCs motivates this work.

2.1 (Markov) Process Models and Estimation

This section introduces process estimation, but specifically addresses (hidden) Markov processes, because not only is the Markov assumption ubiquitous therefore forming the 'standard practice', but the difference in the outputs (distributions and estimates) between RC and MC models of the same scenario is one of the main themes of this thesis. Further, the desirable properties of MCs we describe are what we will contrast to and then try to extend to, in a way, the new models. We set the problems of process estimation, specifically filtering and smoothing, motivating the Markov property signal

model structure that facilitates fast algorithms, specifically the forward and forward-backward algorithms. We show how in the finite-state case, both MC models and the their estimation algorithms are parametrised by matrices and that their outputs usually involve fast matrix-vector calculations. We show the further consequence of matrix product operations that is the forgetting property, and how this means only a subset of an HMC's observations are required to produce an approximate posterior distribution which approaches the exact (all observations) distribution. This section concludes with a brief discussion of the extension of estimation that is the tracking problem, within which certain scenarios motivate the use of *non*-Markov models.

2.1.1 Bayesian Process Estimation

It is easy to call to mind scenarios from various signal processing domains where it is desirable to find the value of a signal that cannot be directly measured. In process estimation, measurements correlated with the signal of interest are combined over time according to the signal model to give a posterior probability distribution of the process' current state. Consider the following scenario: Let there be some time varying signal X_t, which at a set of times $t \in \{0, ..., T\}$ (where T may be infinity), is 'indirectly observed', that is, random variables Y_t, statistically dependent on the signal are sampled. We will refer to the values taken $Y_t = y_t$ as observations. We outline the typical estimation questions for this scenario. The following methods apply generally to the cases that X_t and Y_t are either discrete or continuous variables but we use summation notation informally in anticipation of our specific finite-state focus.

2.1.1.1 (HMM) Filtering

The desired quantity in Bayesian filtering is the posterior or *conditional marginal* distribution of the signal's state at the current time given the set of all measurements taken up to the present. and Letting $\mathbf{P}(x_t)$ stand for $\mathbf{P}(X_t = x_t)$, (for notational convenience, and as it is not so important in the finite state case, we dispense with a formal probabilistic framework so in general $\mathbf{P}$ stands for the relevant probability function in the given context) define

$$\hat{\pi}_t(x_t) = \mathbf{P}(x_t | \{y_{0:t}\}),$$

for $t \geq 0$ where $\{y_{0:t}\}$ stands for $\{y_0, y_1, ...y_t\}$. We will use the hat notation to denote quantities that are conditioned on observations, though not all quantities will be probability distributions. Following [1] Section 15.2, we first define the joint probability of a state and the received measurements,

$$\alpha_t(x_t) = \mathbf{P}(x_t, \{y_{0:t}\}) \tag{2.1}$$

for $t \geq 0$. Evaluating (2.1) by total probability involves summing out the conditional probabilities of all sequences up to t, equivalent to marginalisation of the joint probability $\mathbf{P}(X_0, ..., X_t, y_0, ..., y_t)$, which is computationally impractical as it increases with

t in the absence of assumptions about how the data is generated from the process and how the process depends on itself. The following introduces the conditional independence assumptions that define the structure the *hidden Markov model* of signals X_t and Y_t. Note that expanding (2.1) by total probability and the probability chain rule, for $t \geq 1$,

$$\begin{aligned}\alpha_t(x_t) &= \sum_{x_{t-1}} \mathbf{P}(x_t, x_{t-1}, \{y_{0:t}\}) \\ &= \sum_{x_{t-1}} \mathbf{P}(y_t|x_t, x_{t-1}, \{y_{0:t-1}\})\mathbf{P}(x_t|x_{t-1}, \{y_{0:t-1}\}) \\ &\qquad \mathbf{P}(x_{t-1}, \{y_{0:t-1}\}),\end{aligned} \tag{2.2}$$

Let it be assumed that data is generated according to the rule, in [1] called the *emission rule*

$$\mathbf{P}(y_t|\{x_{0:T}, y_{0:T}\}) = \mathbf{P}(y_t|x_t). \tag{2.3}$$

This holds when, for example $Y_t = X_t + E_t$ where E_t is a 'noise' random variable. Most generally the 'observation' at t may give a likelihood function over the signal space directly. Let there also be a signal model or process model

$$\mathbf{P}(X_t|x_{0:t-1}) = \mathbf{P}(X_t|x_{t-1}) \tag{2.4}$$

This is the Markov assumption, and we will elaborate conditions for its suitability as we go on. With it (2.2) reduces to the recursion

$$\alpha_t(x_t) = \mathbf{P}(y_t|x_t) \sum_{x_{t-1}} \mathbf{P}(x_t|x_{t-1})\alpha_{t-1}(x_{t-1}) \tag{2.5}$$

with $\alpha_0(x_0) = \mathbf{P}(y_t|x_0)\mathbf{P}(x_0)$. The algorithm for filtering an HMM, repeated application of (2.5), is called the *forward algorithm*, which has a constant amount of calculation per step (see the next subsection). For this reason and because data needs to only be incorporated once as it arrives, Markov models and the hidden Markov paradigm is popular for processing *on-line* or real-time signals in many fields such as communications, control systems and aerospace/military sensing. The popular Kalman filter (see for example [2]) is a special case of this framework but in continuous state domain. Note that, from (2.1)

$$\mathbf{P}(\{y_{0:t}\}) = \sum_{x_t} \alpha_t \tag{2.6}$$

gives the probability of the observed data according to the model. The quantities $\hat{\pi}_t$ are available simply by normalisation if necessary,

$$\hat{\pi}_t(x_t) = \alpha_t(x_t) / \sum_{x_t} \alpha_t(x_t). \tag{2.7}$$

2.1.1.2 (HMM) Fixed-Interval Smoothing

If some interval of data has been received observations from 'future history' of the process can be incorporated to get more accurate estimates. More precisely if the data is available on some fixed interval $\{0, ..., T\}$, let there be the joint probability

$$\gamma_t(x_t) = \mathbf{P}(x_t, \{y_{0:T}\}). \tag{2.8}$$

Noting that

$$\gamma_t(x_t) = \mathbf{P}(x_t, \{y_{0:t}\}, \{y_{t+1:T}\}), \tag{2.9}$$

then by the emission rule and Markov property, the past and future observations are conditionally independent given the process value, so that

$$\mathbf{P}(x_t, \{y_{0:t}\}, \{y_{t+1:T}\}) = \mathbf{P}(\mathbf{P}(x_t, \{y_{0:t}\})\mathbf{P}(\{y_{t+1:T}\}|x_t). \tag{2.10}$$

The first term of the LHS is $\alpha_t(x_t)$ and define

$$\beta_t(x_t) = \mathbf{P}(y_{t+1:T}|x_t)$$

which by similar steps to the forward case, can be shown to reduce to the backward recursion:

$$\beta_t(x_t) = \sum_{x_{t+1}} \mathbf{P}(x_{t+1}|x_t)\mathbf{P}(y_{t+1}|x_{t+1})\beta_{t+1}(x_{t+1}) \tag{2.11}$$

So for each $t \in \{0, ..., T\}$,

$$\gamma_t(x_t) = \alpha_t(x_t)\beta_t(x_t), \tag{2.12}$$

which each are produced by a recursive 'pass' that sums out the model's either past or future variables. The algorithm is called *forward-backward*. The algorithm has often been rediscovered in various fields but for a development in the HMM context see [3]. The fact that two recursive passes gives the posterior distributions of the entire interval's variables makes the algorithm and its HMM framework appealing in many applications, so that even for batch-processed intervals of data where there is no for recursion to incorporating observations in real time, the Markov paradigm is common.

Note that *point smoothing* is the problem of obtaining *one* variables estimate given all available observations.

2.1.2 Markov Chains

The previous section introduced the hidden Markov model. The signal model itself is a Markov model, which has its own applications beside process estimation, often to answer what a long term distribution of an evolving quantity looks like, but also in

for example search planning, to predict the distribution at a specific future time after a sighting. [4, 5] The Markovian paradigm is widely adopted - the state-space models from aerospace and control engineering are implicitly Markov by there being a definable 'state': a variable that represent for a given time everything that determines the future of the variables of interest, up to the random component. The Markov property was mentioned earlier but we need a more formal definition: a process is Markov when for any indices $\{r_1, ...r_n\}, < s < t$

$$\mathbf{P}(X_t|x_s, \{x_{r_1}, ..., x_{r_n}\}) = \mathbf{P}(X_t|x_s) \tag{2.13}$$

For a process on a finite index set, this separation implies (2.4), and that the joint probability of a specific sequence of values factorises as

$$\mathbf{P}(x_t, x_{t-1}, ..., x_0) = \mathbf{P}(x_t|x_{t-1})\mathbf{P}(x_{t-1}|x_{t-2})...\mathbf{P}(x_1|x_0)\mathbf{P}(x_0) \tag{2.14}$$

A finite-index, finite-state Markov model in this thesis will be called a Markov chain. That is, where X_t takes values on some finite set with n elements. It follows from (2.14) that an MC may then be parametrised by the set of next-neighbour transition rules or *transitions*, that are $n \times n$ A_t for $t \in \{0 : T-1\}$: where

$$A^{(t)}_{j,i} = \mathbf{P}(X_{t+1} = j|X_t = i), \tag{2.15}$$

where A is *stochastic*, meaning $\sum_j A^{(t)}_{j,i} = 1$. (A note about notation: finite quantities (matrices and vectors) which have both variable indices (times) and element indices, will be represented with subscripts for the time indices, unless the the element indices need to be referenced, where the time indices become parenthesised superscripts as in the above). Define π_0 so that $(\pi^{(0)}_j = \mathbf{P}(X_0 = j)$. Factorising the model into these terms means that the total probability of a given sequence can easily be assessed. A sequence from the model can be generated or *realised* by drawing values sequentially according to distributions $\mathbf{P}(X_t|x_{t-1})$. In the finite-state case these are the columns of A_t in (2.15). The transition rule(s), the parameters, that are the entries of A_t may originate from some theory or, if the modeller has access to example sequences the parameters can be estimated by counting the relative number of transitions to a state given the past time's state. A further note on Markov transitions: define a general two-point transition (of any type of process) $\Phi_{t,s}$ such that

$$\Phi^{(t,s)}_{j,i} = \mathbf{P}(X_t = j|X_s = i). \tag{2.16}$$

The Chapman-Kolmogorov property of Markov transitions (see [6] p. 531) is that for non-adjacent events, e.g. for $r < s < t$

$$\mathbf{P}(x_t|x_r) = \sum_{x_s} \mathbf{P}(x_t|x_s)\mathbf{P}(x_s|x_r).$$

This corresponds to the transition matrix property $\Phi_{t,r} = \Phi_{t,s}\Phi_{s,r}$. By this rule MC general two-time transitions can therefore be generated from the neighbour (parametrising) transitions. Define the transition $\mathrm{F}_{t,s}$ as specifically a matrix composed of an appropriately ordered product chain of the parametrising matrices,

$$\mathrm{F}_{t,s} = \begin{cases} \mathrm{A}_{t-1}\mathrm{A}_{t-2}...\mathrm{A}_{s+1}\mathrm{A}_s & s < t \\ \mathrm{I}_n & s = t, \end{cases} \tag{2.17}$$

For Markov chains $\Phi_{t,s} = \mathrm{F}_{t,s}$ but this property doesn't hold in general, i.e. for non-Markovian models.

2.1.3 MCs/HMCs and Matrix-Vector Operations

As it was in the hidden case, of most interest in signal processing with Markov models are the marginal distributions π_t for $t \geq 0$ where $\pi_i^{(t)} = \mathbf{P}(x_t = i)$. The marginal is obtained by summing the sequence joint over the other (past) variables of the rest of the process, which can be subsumed into the evaluation of the matrix product as in (2.17) so that, as long as the initial distribution π_0 is specified,

$$\pi_t = \Phi_{t,0}\pi_0 = \mathrm{F}_{t,0}\pi_0. \tag{2.18}$$

When the state-space is finite, as in an HMC, the filtered estimate given by the recursions of (2.5) has a similar matrix form: define the $n \times 1$ vector $\bar{c}_t$ to be the vector of likelihood of state values given the observation at time t. In the case of there being a finite set of possible observations, $\bar{c}_t$ is the i'th row of an $n \times m$ matrix C_t where

$$C_{i,j}^{(t)} = \mathbf{P}(Y_t = i | X_t = j). \tag{2.19}$$

Let c_t be a diagonal matrix $\mathrm{diag}(\bar{c}_t)$, and let an *update matrix* for $t \geq 0$ be $U_t = c_{t+1}A_t$. Whereas we defined a general quantity $\alpha_t(x_t)$ in (2.1), we define specifically for the finite-state, forward algorithm operation, an $n \times 1$ vector quantity α_t^{MC} about which we shall make some qualifications shortly: let

$$\alpha_j^{(t)MC} = \mathbf{P}(X_t = j | \{y_0, ..., y_t\}),$$

and by recursive application of 2.5,

$$\alpha_t^{MC} = U_{t-1}U_{t-2}...U_0 c_0 \pi_0. \tag{2.20}$$

Thus, while we previously defined the ordered product $\mathrm{F}_{t,s} = A_{t-1}, ..., A_s$, define for $t > s$

$$\hat{\mathrm{F}}_{t,s} = U_{t-1}...U_{s+1}U_s \tag{2.21}$$

(let $\hat{\mathrm{F}}_{t,t} = I_n$) such that $\alpha_t^{MC} = \hat{\mathrm{F}}_{t,s}\hat{\pi}_s$ where $\hat{\pi}_s = c_s\pi_s$. Likewise, define β_t^{MC} such that

$$\beta_j^{(t)MC} = \mathbf{P}(\{y_{t+1}, ..., y_T\}|X_t = j),$$

and similarly it holds that

$$\beta_t^{MC} = U_t' U_{t+1}' ... U_T' \mathbb{1}, \tag{2.22}$$

so that $\beta_t^{MC} = \hat{\mathrm{F}}_{T,t}' \mathbb{1}$, where $\mathbb{1}$ is an $n \times 1$ vector of all ones. The matrices $\mathrm{F}_{t,0}$ and $\hat{\mathrm{F}}_{t,0}$ are useful for analysis, however neither should be explicitly evaluated in the practice of marginalising/estimating because the series of square-matrix multiplications at cost $\mathcal{O}(n^3)$ per index increment can be reduced to $\mathcal{O}(n^2)$ by proceeding right to left in the expanded product by matrix-vector multiplications. This is what it means to say that the forward-backward algorithm is 'fast', that the influence of the entire rest of the model can be into two vectors, and further that their evaluation can be organised into repeated recursions of computationally cheap matrix-vector products. Note that because in filtering and fixed interval smoothing, the quantities of interest are the relative posterior probabilities of the state values, and the absolute values can be miniscule, to avoid numerical underflow, quantities like $\alpha_t^{MC}, \beta_t^{MC}$ are typically normalised at each recursion. For notational convenience we will continue to use these symbols when normalisation can be assumed, unless stated otherwise.

There is a third aspect to HMC estimation being fast, however which is that the properties of matrix products mean that the whole set of observations may not even be needed. This is forgetting, which begins with the model itself, with the implication that the initialisation doesn't necessarily affect, or rather, necessarily doesn't affect, later marginals.

2.1.4 Markov/Hidden Markov Chain Forgetting

For an MC model to be *initialised* at the nominal initial time $t = 0$ is by some marginal probability,unconditional of the transition rules, π_0. It is a property of homogeneous Markov chains (for which $A_t = A$ for any stochastic matrix so that $F_{t,0} = A^t$), following from the well known Perron-Frobenius theorem, (see for example [7]) that (informally)

$$\lim_{t\to\infty} A^t = \pi\mathbb{1}' \tag{2.23}$$

where π is the stationary distribution of the MC, being the unique solution of $A\,\pi = \pi$, and the rate of convergence is geometric. Since $\pi\mathbb{1}'\pi_0 = \pi\sum\pi_0 = \pi$, (2.23) implies

$$\lim_{t\to\infty} \pi_t = \pi,$$

i.e. the marginal converges (geometrically) to the stationary distribution. Such a Markov chain thus has a marginal which becomes independent of the choice of initial

distribution and depends only on the the transition rule, in other words the *dynamics.* A less well known result is that the tendency of products of matrices to become 'dyadic' is general to non-negative matrices, so non-homogeneous MCs and also the estimator's update matrix product also 'go to' rank 1. Non-negative matrix chains go to rank 1 in an element-wise sense, i.e. for any sequence of matrices used to form a transition $\mathrm{F}_{t,s}$, there is some rank 1 matrix $V_{t,s}$ and $|F_{i,j}^{(t,s)} - V_{i,j}^{(t,s)}| \to 0$ geometrically with $(t-s)$. [7] For a transition to actually be rank 1 would trivially imply that the posterior marginal is independent of the prior: let $F_{t,s} = \pi(t,s)\mathbb{1}'$ for some probability vector $\pi(t,s)$, then

$$\pi_t = F_{t,s}\pi_s = \pi(t,s)\mathbb{1}'\pi_s = \pi(t,s).$$

The forgetting property of non-homogeneous MCs and the HMC estimation algorithm is therefore shown implicitly by proving that chains of matrices go to rank 1, and explicitly by bounding the difference between posterior distributions as a function of the product length. Both aspects are relevant to our subsequent results so we recount results relating to distribution convergence as well as matrix entries.

Birkhoff's Contraction Coefficient The non-negative product result requires setting convergence in a certain (pseudo-) metric space rather than a simple normed space which sufficed for the homogeneous MC case. The following is due to Birkhoff ([8], see [7]). Let x and y be two vectors of finite length, with positive elements denoted x_i, y_i, The Hilbert projective pseudo-metric [1]

$$d(x,y) = \max_{i,j} \ln \left[\frac{x_i \cdot y_j}{x_j \cdot y_i}\right] \geq 0. \tag{2.24}$$

reflects how close two vectors are to being scalar multiples, in other words to being 'aligned'. Note $d(x,y) = 0 \Leftrightarrow x = c \cdot y$ for some positive scalar c. The pseudo-metric satisfies a contraction property

$$d(Py, Px) < d(x,y)$$

for any positive matrix square matrix P and positive vectors x, y (positive throughout refers to element-wise positive).

$$\sup_{x,y>0} \frac{d(Px,Py)}{d(x,y)} = \tanh \frac{\phi(P)}{4} \equiv \tau_B(P) \tag{2.25}$$

is Birkhoff's contraction coefficient where, letting P_i denote the i'th column of P [9]

$$\phi(P) = \sup_{x,y>0} d(Px,Py) = \max_{i,j} d(P_i, P_j). \tag{2.26}$$

[1] A pseudo-metric $d(.,.)$ satisfies the usual axioms of a metric apart from the property that $d(x,y) = 0 \Rightarrow x = y$.

For the explicit derivation of τ_B, see [7]. It follows from the above and the triangle inequality of $d(.,.)$

$$\tau_B(P) < 1$$
$$\tau_B(PQ) \leq \tau_B(P)\tau_B(Q) \tag{2.27}$$
$$\tau_B(P) = 0 \Leftrightarrow P = vw' \tag{2.28}$$

for positive vectors v, w. We will later need a property of τ_B which can be shown from the definition: for any diagonal matrices D_1, D_2

$$\tau_B(D_1 A D_2) = \tau_B(A). \tag{2.29}$$

Consider a Markov chain with transitions as in (2.17). For initalisations π_0^p, π_0^q, let π_t^p, π_t^q be the marginals given the respective initialisation. Then by (2.18), (2.25) and (2.26)

$$d(\pi_t^p, \pi_t^q) \leq \tau_B(\mathrm{F}_{t,0}) \cdot d(\pi_0^p, \pi_0^q) \leq \phi(\mathrm{F}_{t,0}) \tag{2.30}$$

and by (2.27),

$$\tau_B(\mathrm{F}_{t,0}) \leq \prod_{\tau=0}^{t-1} \tau_B(A_\tau).$$

The conditions to ensure contraction are looser than described here, as we will discuss in subsection 2.1.5, but for now let A_t be positive for $t \geq 0$ and let there be a $\lambda < 1$ such that

$$\tau_B(A_t) \leq \lambda, \ t \geq 0 \tag{2.31}$$

This can be guaranteed when, for example there exists a minimum entry condition

$$A_{i,j}^{(t)} \geq \gamma \Rightarrow \tau_B(A_t) \leq \lambda \equiv (1-\gamma)/(1+\gamma) < 1, \tag{2.32}$$

which follows from the definitions of $d(.,.)$ and τ_B. Under the condition (2.31), there is uniformly geometric contraction of τ_B at rate at least λ, giving the Markov chain forgetting rule:

Theorem 2.1. *For an MC model generated from one-step transitions $A_t, \ t = \{1, ..., T\}$ where $\tau_B(A_t) \leq \lambda < 1$ then*

$$\tau_B(\Phi_{t,s}) \leq \lambda^{t-s}. \tag{2.33}$$

More directly applicably is the corollary, due to [10]

Corollary 2.1.1.

$$||\pi_t^p - \pi_t^q||_\infty \leq C\lambda^t \tag{2.34}$$

for some positive constant C.

Proof. By (2.30) and Theorem 2.1

$$d(\pi_t^p, \pi_t^q) \leq C\lambda^t, \tag{2.35}$$

where C can be chosen as either $d(\pi_0^p, \pi_0^q)$ or $\ln(1/\gamma^2)/\lambda$ (by the bound of $\phi(A_t)$), whichever gives the closer bound. For probability vectors, tending to 'alignment' means tending to equality: For $v, w \in (0,1]^n$,

$$\begin{aligned} d(v,w) &= \ln \max_k \left(\frac{v_k}{w_k}\right) + \ln \max_k \left(\frac{w_k}{v_k}\right) \\ &\geq \max_k |\ln v_k - \ln w_k| \\ &\geq \max_k |v_k - w_k| = ||v - w||_\infty, \end{aligned} \tag{2.36}$$

using the fact that ln has derivative ≥ 1 on $(0,1]$. By equivalence of norms on finite-length vectors the result is therefore given for any vector norm. □

In our results we will tend to use the infinite norm to show contractions because of the equivalence to other norms and as other 'probability distances' that can be defined are implied to go to zero as the infinity norm does. (2.36) will be useful for vector quantities that are not probability distributions but are normalised

Matrix Entry Results Considering only stochastic matrices (with columns summing to one) contraction results are available for the matrix entries themselves, which we will need. A contraction coefficient for stochastic matrix P is

$$\tau_1(P) = \frac{1}{2} \max_{i,j} \sum_k |p_{k,i} - p_{k,j}|,$$

define also

$$\rho(P) = \max_i \max_{j,k} |P_{i,j} - P_{i,k}|. \tag{2.37}$$

It holds ([7] Theorem 3.13) that [2]

$$\rho(P) \leq \tau_1(P) \leq \tau_B(P). \tag{2.38}$$

Backward Result The implicit counterpart to the fact that as time increases, marginals become independent of past states, is that to evaluate a that a marginal at t, given information from increasingly far in the past, converges to a fixed value. Under repeated backward multiplication (in our configuration, left-multiplication) by transitions, those transitions cease to affect the output marginal: Let $\pi_{t|t-\Delta}$ be

$$\pi_{t|t-\Delta} = \mathrm{F}_{t,t-\Delta}\pi_{t-\Delta} \tag{2.39}$$

[2] In [7], $\rho(P)$ is denoted $\alpha(P)$.

it holds from (2.33) and (2.18)

$$||\pi_{t|t-\Delta} - \pi_{t|t-\Delta-1}||_\infty \leq \tau_B(\mathrm{F}_{t,t-\Delta}) \leq C\lambda^\Delta. \tag{2.40}$$

Informally, this is a contraction implying a there is a fixed vector π_t, and $\mathrm{F}_{t,t-\Delta}$ approaches $\pi_t \mathbb{1}'$. We will be more explicit on what the backward result means for matrix entries within Chapter 3.

2.1.4.1 HMC Forgetting

The same forgetting theory applies to the product of update matrices that give the filter estimate so that for an estimate informed by Δ past observations $\alpha^{t|t-\Delta} = \hat{\mathrm{F}}_{t,t-\Delta}\mathbb{1}$,

$$||\alpha^{t|t-\Delta} - \alpha^{t|t-\infty}||_\infty \leq C\bar{\lambda}^\Delta$$

where $\bar{\lambda}$ is the forgetting rate of the estimator, which depends on the observation sequence - we will discuss dealing with suitable rate evaluation in the next subsection. Informally the same argument applies to the recursions that give β_T^{MC} so that for an HMC model with $\gamma^{t|t-\Delta,\ldots,t+\Delta}$ being the fixed-point smoothed estimate with 2Δ adjacent observations

$$||\gamma^{t|t-\Delta,\ldots,t+\Delta} - \gamma^{t|-\infty,\ldots,\infty}||_\infty \leq C\bar{\lambda}^\Delta \tag{2.41}$$

The formal derivation of (2.41) is given in [11], but also follows from results in Chapter 4 of this work. The significance of (2.41) is that an estimate which approaches the exact (all observations) estimate can be obtained with limited observations, which can reduce calculation of a posterior distribution. (The problem of obtaining an estimate for a single variable given a set of observations is called point smoothing.)

2.1.5 Forgetting Rates and Example Model Scenario

Forgetting Rate - Model We gave the rate as λ, the maximum coefficient of matrices in the set A_t for $t \geq 0$, however this bound may not be tight enough in practice (for bounding error etc.), furthermore, the forgetting property extends to a broader class of transitions which are of practically importance, that have coefficient equal to one, specifically 'primitive' transitions, which means A is non-negative but there is a τ such that A^τ is positive.[3] The forgetting theory then applies normally at the time scale of epochs of size τ. In general, evaluating the coefficient for a transition of larger τ', gives a tighter bound, $\tau_B(A^{\tau'})^{t/\tau'}$. The asymptotic rate of decay of $\tau_B(A^t)$ with t is the magnitude of the second eigenvalue of A, however since reciprocal processes occupy finite-length intervals, the intermediate behaviour which can be numerically evaluated is more relevant. Numerical evaluation of a matrix τ_B is of complexity $\mathcal{O}(n^4)$, which is

[3] The property of primitivity can equally be applied to a non-homogeneous set of transitions that when τ transitions are combined in any order, the overall transition is positive.

not so slow as to prevent numerical exploration of the practical bound in the 'offline' analysis. In subsequent results we use λ in the theoretical bounds to refer to whichever rate is preferred.

Estimate Forgetting Rate It is stated in [11], the forgetting rate of the estimator is no slower than that of the signal model, which parallels a result for Kalman filters. A simulation study is conducted in [11] suggesting that under the condition of informative measurements (i.e. the likelihood of the state values given some measurement is not uniform), the rate is strictly faster than that of the signal model (the cases are all . We find that the faster rate can actually be proved for the $n = 2$ case. We include this proof the Appendix, but we proceed assuming that this holds generally. The rate of forgetting in the estimation algorithm is dependent on the data model, and the observation sequence itself, so as discussed in [11] simulations can be to performed to obtain average rates. Assuming the estimator forgetting is no slower than the model, the model rate can be used as an absolute worst case rate for the estimator. As seen in the example of the next paragraph, the rate of forgetting in the estimator is typically fairly well grouped between different realisations, therefore the approach we take in this thesis to evaluating estimator forgetting rate $\bar{\lambda}$ for a model is to average $\tau_B(\hat{\mathrm{F}}_{T,0})$ for 30 observation sequence realisations, at cost of $\sim 30n^4$.

Example Signal Model We draw an example transition matrix, (we will consider a homogeneous process) from the context of the target tracking application domain considered in the next section, to show the rate of contraction for a prediction marginal vs. the bounds, and demonstrate estimate contraction rates. We will use the transitions from the example in demonstration of our results in subsequent chapters. Consider model of a simplified urban road network, letting road segments that a vehicle might occupy be states, linked in a two-dimensional square lattice with edge length $\sqrt{n}$ (there are n states). A vehicle, after an increment of time, can have transitioned to a connected road segment (in the 4-connected sense) or remain on the same segment so that as a general rule (letting i, j be the indices of the lattice)

$$\mathbf{P}\left\{X_t = (j,i) | x_{t-1}\right\} = \begin{cases} p, & \text{if } x_{t-1} = (j,i) \\ \frac{1-p}{4}, & \text{if } x_{t-1} = (j, i \pm 1), (j \pm 1, i) \\ 0, & \text{otherwise.} \end{cases} \tag{2.42}$$

for some p which models how likely it is that the target has not had time to move out of its state in the time-step. (2.42) applies unless the prior state is on the edge of the lattice where, in this example, the probability is redistributed over the existing neighbour states. (2.42) is encoded into a transition matrix $\bar{A}$ after choosing some single index for the states, giving an $n \times n$ matrix. The transition is clearly not a positive matrix, but it is 'primitive' - that is there is a τ such that $\bar{A}^\tau$ is positive. We can see directly for this specific scenario that the transition $\mathrm{F}_{t,0}$ will not be positive

until the walker has been able to take the shortest passage between the furthest states. One corner to another requires $\tau = 2(n-1)$. (Note that [12] gives that for a set of transitions that is primitive, meaning when multiplied in any order, become positive, then τ cannot be greater that $2^n - 2$.) If we take strictly steps of this transition we can see, as in Fig. 2.1 (a), there is no forgetting until τ. To better assess the representative forgetting rate, a transition $A = \bar{A}^\tau$ is created, and its forgetting assessed as a function of it's powers which are 'epochs' relative to the original transition. We numerically evaluate λ from A and plot the bound due to it, showing that even for this positive transition, the bound is not tight. To calculate a tighter bound at the expense of calculation we numerically evaluate $\lambda_2 = \tau_B(A^2)^{1/2}, \lambda_4 = \tau_B(A^4)^{1/4}$, those transitions requiring one and two matrix multiplications respectively to obtain.

Example Observation Model To demonstrate the estimator forgetting rate we synthesise example observation sequences, and construct the chain of update matrices that form $\hat{\mathrm{F}}_{t,0}$. To do so, state sequences are realised from the model, and from each realisation a sequence of observations is drawn from the observation model. The example observation model emulates 'signal plus noise'- the noise a is circular Gaussian variable with variance $\sigma = 1$ (1 lattice step). From each sequence the set of state likelihoods given observation y_t, $\bar{c}_t$, are used to form chain $\hat{\mathrm{F}}_{t,0}$, whose coefficient is numerically evaluated as shown in Figure 2.1 (c) (one observation per time-step) and (d) (one observation per epoch). An illustration of an observation sequence generated by this model is shown in Figure 4.2 (a) (Chapter 4).

2.1.6 Target Tracking by Estimation

Tracking's Two Sub-Problems Consider that in the aforementioned estimation schemes (either filtering or smoothing), all of the data/observations received are assumed to originate from the process of interest. It is conceivable that data might arrive not from the modelled process (such as a moving target) but as detections of irrelevant objects (so called clutter), or additionally in the case of multiple present targets, from the wrong target. Obtaining state estimates under these conditions is the tracking problem, which thus has two simultaneous and interdependent sub-problems: deciding whether/which observations come from a target (assignment) and estimation of the target state(s). Tracking is generally a real time problem meaning the assignment decision has to be made as data arrives, but since the assignment sub-problem is optimally solved over an interval by applying the filter according to all possible assignment combinations and seeing which produces the highest likelihood of observations given the target model, various schemes for tough tracking cases use 'delayed decision' where a sliding window is batch processed as an interval to find the assignment, according to which the filter (optionally with a smoothing pass) then gives the state estimates. In [13], for example, it is shown that single target tracking with clutter (false alarm returns) can be treated in such a way as a case of the interval estimation problem with

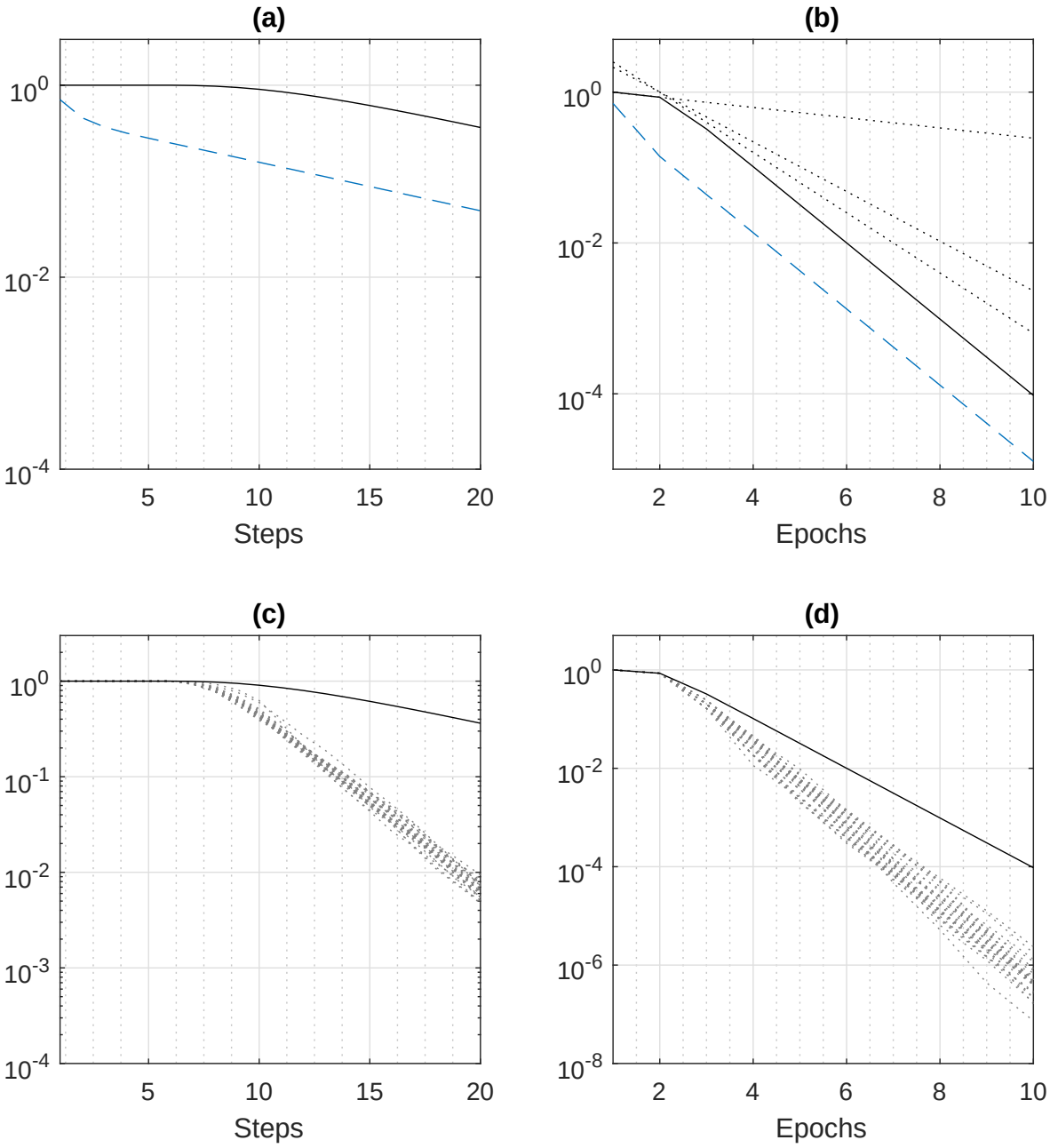

Figure 2.1: Marginal and estimate forgetting with primitive transition (4-connected on a lattice of side-length 4). The black line in each plot is numerical $\tau_B(\mathrm{F}_{t,0})$, In (a),(c) the transition is $\bar{A}$, in (b),(d) it is A. In (a) and (b), the blue line the infinity norm between marginals at t given different initial states (the most distant pair, opposing corners). In (b) the dotted lines are bounds due to λ (loosest), λ_2 and λ_4 (tightest). In (c),(d) dotted (grey) lines are numerical $\tau_B(\hat{\mathrm{F}}_{t,0})$ corresponding to 30 example sets of $\{y_0, ...y_T\}$. Note that (d) reflects receiving one observation every epoch.

an extended observation model.

Motion Models and Memory Tracking is historically an aerospace application, where targets move in free space and their state includes velocity variables, but in ground vehicle tracking by radar [14, 15] and submarine tracking systems [4], targets are subject to highly nonlinear constraints so that their motion is best modelled by a random walk on finite states such as road segments or cells of a spatial grid. In this case the memorylessness of a Markov model can cause large errors, for example, if a segment has two reachable neighbours (such as along a road) the model will equally predict the neighbour from which the target arrived. [16] Futhermore, unlike aerospace tracking, where targets that *cross paths*, from the sensors point of view, are often able to be distinguished by their velocity, the finite state model does not differentiate targets which momentarily share a grid cell or segment. This leads to potential swap overs in the multiple target case as being simultaneously on the same stretch of road means both targets subsequent movements *are* fully compatible with each others' memoryless motion model. Recently a body of work is starting to build around the employment of a finite-state model which can address these limitations with a relatively low amount of additional complexity. These are reciprocal models, which are the subject of the next section.

2.2 Reciprocal Process Models and Estimation

We recount reciprocal process models, including the situation of creating a reciprocal model 'based on' some Markov model, the algorithm for smoothing processes modelled by a reciprocal chain, and the forgetting property question that motivates this thesis' work.

2.2.1 Reciprocal Processes

A process is reciprocal if it admits the conditional independence property that for indices $0 \leq a < b \leq T$, and t, s where $t \in (a, b)$, $s \in [0, a) \cup (b, T]$, it holds that

$$\mathbf{P}(X_t, X_s | x_a, x_b) = \mathbf{P}(X_t | x_a, x_b)\mathbf{P}(X_s | x_a, x_b). \tag{2.43}$$

Since (2.43) is equivalent to the property

$$\mathbf{P}(X_t | x_s, x_a, x_b, x_u) = \mathbf{P}(X_t | x_a, x_b) \tag{2.44}$$

for any $s < a < t < b < u$, the term 'reciprocal property' refers to either (2.43) or (2.44). (It is easy to see that (2.44) implies (2.43), for the proof of the converse see [17] Lemma 1.) The Markov property implies reciprocal but not the converse. Specification

of a reciprocal process is by the functions that are three-point 'transitions' (like those of a Markov model, but a set of distributions given a past and future value)

$$\mathbf{P}\{X_t|x_a, x_b\} \quad [0 \leq a < t < b \leq T] \tag{2.45}$$

but because by the reciprocal property, X_0 and X_T are not conditionally independent given the internal variables (consider $a = 1, b = T - 1$ in (2.43)), the initialising distribution that must be given is the joint boundary relationship $\mathbf{P}\{X_0, X_T\}$, It follows that in the finite-index case, for $t \in \{0, ..., T\}$

$$\mathbf{P}(X_t|\{x_0, ...x_{t-1}, x_{t+1}, .., x_T\}) = \mathbf{P}(X_t|x_{t-1}, x_{t+1})$$

and so the transitions that can be specified to parametrise a finite-index, finite state process are each variable's dependence on its nearest neighbours: an RC model on $\{0, ..., T\}$ is specified by the set, for $t = \{1, ..., T - 1\}$

$$\begin{aligned} Q_{i,j,k}(t) &= \mathbf{P}(X_t = j|X_{t-1} = i, X_{t+1} = k) \\ \Pi_{i,j} &= \mathbf{P}(X_0 = j, X_T = i). \end{aligned} \tag{2.46}$$

where Q is an array such that for any i, k $\sum_j Q_{i,j,k} = 1$. Like Markov transitions, the Q can be counted from example sequences, or stem from theory including being generated from other transitions, as we will show in subsection 2.2.1.1. It is not possible to realise a sequence directly from the nearest-neighbour transitions, as each requires two neighbours to already be drawn. Two things are needed, general (i.e. not neighbouring) transitions and some schedule to move through the sequence. General transitions can, as in Markov chains, be calculated from the parametrising ones:

Working with general relations With Markov models the Chapman-Kolmogorov relation, corresponding to the semigroup property of transitions, allowing use of the matrix product to build any 'transition' recursively from the next-neighbour ones. Reciprocal transitions must obey what [18] calls the Schrodinger-Jamison relations: For $0 \leq s < t < u < v \leq T$

$$\sum_{x_u} \mathbf{P}(x_u|x_s, x_v) = 1 \tag{2.47}$$

$$\begin{aligned} &\mathbf{P}(X_u|X_s, X_v)\mathbf{P}(X_t|X_s, X_u) \\ &= \mathbf{P}(X_t|X_s, X_v)\mathbf{P}(X_u|X_t, X_v) \end{aligned} \tag{2.48}$$

where (2.48) says that the joint probability of two events given two events outside them, $\mathbf{P}(X_t, X_u|X_s, X_v)$, must be able to be be factorised consistently into either the LHS or RHS of (2.48). This can be proved by the reciprocal property. Relation (2.48) is the means to get general three time transitions, for example rearranging (2.48), a longer range transition is given

$$\mathbf{P}(X_t|X_s, X_v) = \frac{\mathbf{P}(X_u|X_t, X_v)}{\mathbf{P}(X_t|X_s, X_u)}\mathbf{P}(X_u|X_s, X_v) \tag{2.49}$$

where the rightmost term can be seen not to depend on t so that it acts as a normalising constant for vectors making up the new three point transition obtained from two nearest neighbour ones.

Bridge Decomposition It was mentioned that to generate a sequence some schedule is required to move through the sequence. One such follows from the result of [19] that a reciprocal process conditioned on taking a value at its endpoint is then Markov. The 'bridge decomposition', converted to the finite-state case in [20], allows sequences to be realised and is also used in the RC estimation algorithm. Decomposition of the process into 'bridges' is by the factorisation

$$\mathbf{P}(x_0, x_1, ...x_T) = \mathbf{P}(x_0, x_T) \prod_{t=1}^{T-1} \mathbf{P}(x_t | x_{t-1}, x_T). \tag{2.50}$$

The terms $\mathbf{P}(x_t|x_{t-1}, x_T)$ form processes having Markov transitions denoted, B_t^k for the bridge pinned to state k at T so that

$$B_{j,i}^{(t),k} = \mathbf{P}(X_{t+1} = j | X_t = i, X_T = k)$$

for $t = \{0, ..., T-2\}$. There is no 'final transition' as the variable X_T is not part of this conditioned process. (This factorisation has the additional benefit of being able to be interpreted as a collection of Markov models each pinned to 'arrive' at some 'destination' at time T). Realisations of sequences are achieved with the bridge decomposition by drawing the final state first, from π_T where $\pi_k^{(T)} = \sum_j \Pi_{k,j}$. The first state is then drawn from the conditional initial distribution π_0^k where

$$\pi_i^{(0),k} = \mathbf{P}(X_0 = i | X_T = k) = \frac{\mathbf{P}(X_0 = i, X_t = k)}{\mathbf{P}(X_t = k)} = \frac{\Pi_{k,i}^{(T,0)}}{\sum_j \Pi_{k,j}^{(T,0)}},$$

before the other variables are drawn in the usual Markov manner by the transitions B_t^k. With this decomposition the marginal is the weighted sum

$$\pi_t = \sum_k (B_t^k B_{t-1}^k ... B_0^k \pi_0^k) \mathbf{P}(x_T = k). \tag{2.51}$$

Reference [20] describes the procedure to evaluate B_t^k from Q_t: Note that $B_{i,j}^{(T-2),k} = Q_{i,j,k}^{(T-1)}$, then proceeding recursively backward with t, B_t^k can be evaluated by (2.49),

$$B_{j,i}^{(t),k} = \frac{1}{z_{j,i}} \frac{Q_{i,j,\ell}^{(t+1)}}{B_{j,i}^{(t+1),\ell}} \tag{2.52}$$

for some ℓ (the result is independent of ℓ. where

$$z_{j,i} = \sum_{m=1}^{N} \frac{Q_{i,m,\ell}^{(t+1)}}{B_{m,i}^{(t+1),\ell}}.$$

It is easier to visualise the dynamics of a reciprocal process in the case that the reciprocal model is one derived from some Markovian model's transitions and an additional end point distribution. This is also the case of most current application interest. This case was the original inception of reciprocal models and we will recount some of that development as it gives some insight into the relationship between Markov and reciprocal models:

2.2.1.1 RC with Markovian Base

Schrodinger Bridge In the 1930s, Erwin Schrodinger [21] attempted to answer the following question in the context of modelling diffusions: For a density of particles observed at some time, and at some future time where the second distribution is different to what was expected according to the theoretical diffusion dynamics, what was the most likely dynamics of the empirical process? Diffusion is assumed to be Markovian so the theoretical diffusion dynamics is specified in two-time transitions rules but since the final distribution is different, the two-time transition rule of the empirical process necessarily differed from the theoretical one. What Schrodinger intuited was that the most likely dynamics - what in modern terms is the 'closest' model (in the Kullback-Leibler divergence sense (see the next section)) that would produce the observed final distribution, resulted from holding invariant between the theoretical and empirical dynamics not the two-time diffusion rule but the three-time rule for the intermediate distribution given a past and a future value, and factorising the process not like (2.14) but as in (2.50). The three-point transition rule was obtained from the theoretical diffusion's Markov transition rule: for $r < s < t$, the Markov property and chain rule give that the joint distribution of the three events is given by

$$\mathbf{P}(x_t|x_s)\mathbf{P}(x_s|x_r)\mathbf{P}(x_r). \tag{2.53}$$

Bayes rule can be applied so that the conditional distribution of the middle variable is

$$\mathbf{P}(x_s|x_r, x_t) = \frac{\mathbf{P}(x_t|x_s)\mathbf{P}(x_s|x_r)}{\sum_{x'_s} \mathbf{P}(x_t|X_s = x'_s)\mathbf{P}(X_s = j|x_r).} \tag{2.54}$$

By the Markov semigroup property the denominator is simply $\mathbf{P}(x_t|x_r)$. Schrodinger's interest was the case where the end point joint is such that the process is still Markovian, however in the following year S. Bernstein [22] noticed that this was not always the case, and a new class called Bernstein or reciprocal process was defined.

RC with Markovian Base For an RC generated from a set of 'base' Markov transitions A_t $0 \geq t < T$, the general three-point transitions are at obtained more easily than the recursion (2.52). Bridge transitions are a particular case of (2.54),

$$\begin{aligned} B^k_{j,i}(t) &= \mathbf{P}(x_{t+1} = j|x_t = i, x_T = k) \\ &= (A_t)_{j,i}\frac{(\mathrm{F}_{T,t+1})_{k,j}}{(\mathrm{F}_{T,t})_{k,i}} \end{aligned} \tag{2.55}$$

Relation Between RC and MC Models In this thesis we will give various results about the difference between single-point distributions (marginals) of variables due to RC and MC models. Here we recount what is currently established about the quantification of the overall difference between the models. The Kullback-Leibler divergence (see [23]) is an information theory based quantify that measured the difference between two probability distributions. More or less divergence is reflected in quantities of interest like the error due to using one distribution to model data of which another distribution is true. The divergence of two distributions on the same space, with elements indexed by i, $P(i)$ and $Q(i)$, the divergence is defined as

$$D_{KL}(P||Q) = \sum_i P_i \ln \frac{P_i}{Q_i} \tag{2.56}$$

In the case of process models, i indexes possible sequences. Consider that due to the bridge factorisation (2.50), a reciprocal formed from some base model undergoes a cancellation in (2.56) so that the divergence between an RC and its base MC model reduces to a function of the two joint endpoint distributions [23]

$$D_{KL}(RC||MC) = \sum_{i,j} \Pi_{i,j} \ln \frac{\Pi_{i,j}^{(T,0)}}{\mathrm{F}_{i,j}^{(T,0)} \cdot (\pi_0^{MC})_j} \tag{2.57}$$

π_0^{MC} however is not part of the base model itself, so it can be set to match the initial marginal implied by the RC joint. Factorising the RC's endpoint joint consider 'long-term transition' $\Phi_{T,0} = \mathbf{P}(x_T|x_0)$ implied by the RC joint so that $\Pi_{T,0} = \Phi_{T,0}\mathrm{diag}(\pi_0)$. Setting $\pi_0^{MC} = \pi_0$,

$$D_{KL}(RC||MC) = \sum_{i,j} \Pi_{i,j} \ln \frac{\Phi_{T,0_{i,j}}}{(\mathrm{F}_{T,0})_{i,j}}. \tag{2.58}$$

In [13] is is evidenced in simulation studies that the time-average relative error in an estimation task due to using an MC base to model RC data vs. the matched RC model, is increasing with this term, as is not surprising. What is not established by any existing demonstration or theory, is where (in time) this error is distributed, this will be addressed by the results of this thesis.

2.2.2 Hidden Reciprocal Chain Estimation

A hidden reciprocal chain (HRC) is the reciprocal equivalent of the hidden Markov chain has the same emission rule as a HMC. An reciprocal model cannot be exactly smoothed using the forward-backward algorithm because under the reciprocal property the sets of observations in the past and future of a time are not conditionally independent given the signal value, as they are by the Markov property. Other algorithms for the posterior distribution have been developed, in continuous state [24] developed a

smoother which uses (like forward-backward) a double-sweep of the interval, however the recursions are more complex. The finite-state HRC smoother was derived in [20] using the bridge decomposition. The forward-backward algorithm is applied to each (Markovian) 'bridge' conditional process, and then Bayesian recombination (mixing) of the conditioned processes gives the posterior distributions. Define

$$\alpha_j^{(t),k} = \mathbf{P}(X_t = j, \{y_0, ..., y_t\}, X_T = k),$$

Let $U_t^k = c_{t+1} B_t^k$, then

$$\alpha_t^k = U_t^k U_{t-1}^k ... U_0^k \hat{\pi}_0^k \tag{2.59}$$

where $\hat{\pi}_0^k = c_0 \pi_0^k$ and $\pi_i^{(0),k} = \Phi_{k,i}^{(T,0)}$. Likewise

$$\beta_t^k = U_t'^k U_{t+1}'^k ... U_{T-1}'^k \mathbb{1}$$

By Bayes:

$$\gamma_j^t = \sum_{k=1}^{N} \alpha_j^{(t),k} \beta_j^{(t),k} \pi_k^{(T)}. \tag{2.60}$$

2.2.2.1 Complexity

For interval smoothing, due to performing the $\mathcal{O}(n^2 T)$ forward-backward algorithm n times, the HRC interval smoothing algorithm is $\mathcal{O}(n^3 T)$ per time-step. Note that for applications in target tracking, signal models for the motion of targets may have large state spaces, such as 50,000 in the Markov model used in the system of [4], so an additional power of n merits investigation into both the significance of the RC model vs. an MC and any opportunities to reduce calculation.

2.2.3 Example Signal Model

In answer to the need outlined in Section 2.1.6, there is a nascent effort to implement and analyse the performance of tracking algorithms utilising reciprocal chain models. [13, 25–28] Typically an RC model is built by starting with assumed MC transitions (i.e. a random walk on some state space) and applying a chosen $\Pi_{T,0}$ that better reflects the long term behaviour of targets than the $\Pi_{T,0}$ induced by the MC transitions. Consider as motivation for this, that for a simple random walk such as the example given in Section 2.1.5, after an interval of time, the most likely position of the target is its original state, due to the unbiasedness of the transition steps (see the central limit theorem applied to a random walk). This is however not not in accordance with a purposeful (real) walker. An RC can model such purpose by asserting a long-term relationship $\Pi_{T,0}$ that is more 'spread out' than that due to the random walk. Example implementations of this concept are where a modeller specifies a $\Pi_{T,0}$ where far

apart source-destination pairs are explicitly more likely than they are in $\mathrm{F}_{T,0}$, biasing targets to move away from the their origin, or where the modeller chooses a $\Pi_{T,0}$ that is closer to rank 1 than $\mathrm{F}_{T,0}$, so that over the finite interval, the target must move in a such a way that its final distribution is independent of its origin. With such models, as demonstrated in [13], more likelihood is attributed by the estimator /assignment algorithm to paths that proceed through the state-space, rather frequently than doubling back. Figure 2.2 (b) demonstrates targets trajectories generated by such an example model, compared to those generated by a Markov model in (a). The transition rule of (a) is rule 2.42, with parameter p set very low, inducing the target to transition between states regularly. Model (b) is an RC generated from the transitions of (a), with a non-Markov $\Pi_{T,0}$. For illustrative purposes an exaggerated $\Pi_{T,0}$ was engineered for (b) such that source-destination pairs which are most unlikely according to the long term behaviour of (a) (i.e. the longest paths) were made the most likely by setting

$$\Pi_{j,i}^{(T,0)} = \frac{1}{Z}\frac{1}{\mathrm{F}_{j,i}^{(T,0)}},$$

with Z the normalising factor. The Markov trajectories of Figure 2.2 (a) can be seen to double back frequently, and the distribution of destinations can be seen to group around the origin state. The pertinent features of Figure 2.2 (b) are firstly, that the reciprocal model's trajectories infrequently double-back over a series of steps, and secondly the that the distribution of path end-states (destinations) for paths departing from the top-left corner state, can be seen to be grouped around the bottom-right corner, whereas the top-right originating paths typically end at the bottom-left. This behaviour cannot be captured by a Markov model. The model's dynamics are such that trajectories between different source-destination pairs 'cross', in other words paths can occupy the same state mid-trajectory. In a Markov model, the future depends only on the present state and so the subsequent trajectory must be independent of origin state.

2.2.4 The Reciprocal Chain Forgetting Question

For Markov processes the question of forgetting concerned the decay with which a variable's value or distribution at one time affected the distribution at future times. Of particular interest was the influence of the initialisation, which in the Markov case is the same question because the initialisation is a distribution on the variable's value at nominal start time. Non-Markov processes cannot be said to be initialised by a distribution on the variable, because the variable is not the 'state' (the quantity that contains all information from the past). [29] In the RC case, the model is properly initialised by the end-points joint distribution. Nevertheless it is desirable to maintain the consistency with MC concepts such as initial distribution when utilising RCs as process models where there is a meaningful order of and progression of the index: We

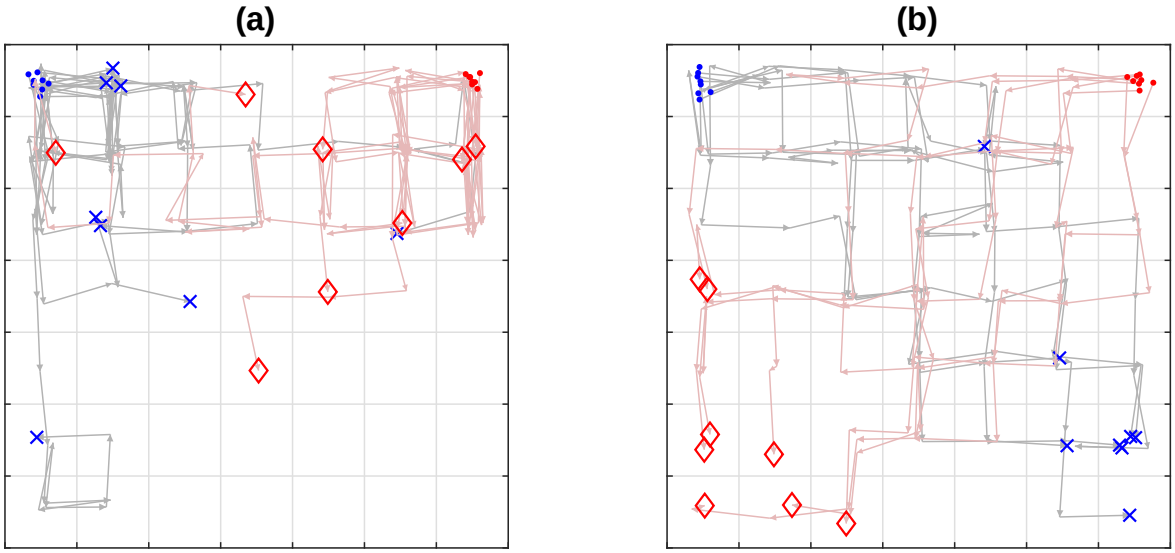

Figure 2.2: Best viewed in colour. Selected realisations from MC and RC models, visualised in 2-dimensional layout, where grid cells represent state values. Only the trajectories originating in two states are shown. The model generating the trajectories in (a) is an MC with transitions given by rule 2.42, $n = 49$, $p = 0.005$. The model in (b) is an RC generated from the transitions of (a), and a non-Markov $\Pi_{T,0}$ (described in the body text). Origins (i.e. x_0) marked with point marker. Destinations (i.e. x_T) are marked according to originating state - from the top-left state, crosses, from the top-right diamonds. Note that a realisation is a set of T=14 state values chosen from set $\{1, ..., 49\}$, however random 'jitter' offsets have been applied to target positions within the cells purely to aid visualisation of transitions.

would like to be able to ask of the model, if the walker is here or there at a particular time, to what extent is its distribution at other times affected? This is not obvious from the bridge decomposition. We will see in the next chapter that we can organise the RC model to get the two-point transition so that a past distribution goes into the model for a future distribution to be predicted. The MC forgetting question was answered by bounding Birkhoff's contraction coefficient of the two-point transition as a function of distance on the index. In asking the same question of an RC we will necessarily get a different answer - the MC transition over the interval $[0, T]$ must have coefficient bounded by λ^T, but for an RC the long-range transition's coefficient can be arbitrarily prescribed. What happens along the way? This question matters because as outlined in the introduction, if the purpose of using an RC model in an application is to model a walker that moves from an initial state to likely end-states over an interval, then if there are marginals that are unaffected by either the source or the destination, the model is not efficient, in some sense. At the same time a forgetting property can allow approximate estimates with only 'local' data to be accurate.

Forgetting in the MC sense is one example of the more general notion of locality. An

example of models that have forgetting in the more general sense is continuous variable (such as Gaussian) reciprocal processes: It is known that there are some parallels between constant variance processes on continuous variables and finite-state processes. In continuous variable processes, the magnitude of the autocorrelation, defined for a stationary, zero-mean, unit variance process as $R(\tau) = \mathbf{E}(x_t x_{t+\tau})$, is an influence measure. For a Markovian constant variance process, such as the AR(1) on finite index, and the Ornstein-Uhlenbeck process on continuous index, the autocorrelation function is geometric decay, showing the association between Markov structure with constrained dynamics/state, and geometric forgetting. Autocorrelation functions of constant variance RPs were established firstly for the continuous index case in [17, 30, 31], and separately for the discrete index case in [32], where in each case, the autocorrelation has the general form

$$R(\tau) = Ae^{-\alpha\tau} - (1-A)e^{\alpha\tau}, \tag{2.61}$$

where $\alpha > 0$ and A are constants. Different choices of boundary condition (co-variance of x_0, x_T) lead to different forms of the correlation function through A. The special Markovian case ($A = 1$) has $R(t) = e^{-\alpha t}$. 'Periodic' boundary condition $x_T = x_0$ (on interval to $[0, T]$), gives a process whose correlation is a hyperbolic cosine function:

$$R(\tau) = \frac{e^{-\alpha\tau} + e^{-\alpha(T-\tau)}}{1 + e^{-\alpha T}}. \tag{2.62}$$

The 'anti-cyclic' ($x_T = -x_0$) condition leads to a hyperbolic sin correlation:

$$R(\tau) = \frac{e^{-\alpha\tau} - e^{-\alpha(T-\tau)}}{1 - e^{-\alpha T}} \tag{2.63}$$

The form of the correlation function is between the superposition and cancellation of a decay and a growth, depending on the boundary condition. Note that the magnitude of (2.63) and in fact (2.61) in general, is no greater than (2.62) so (2.62) serves also as the *bound* on magnitude of $R(\tau)$. In (2.62) $R(0) = R(T) = 1$, but the function drops to a minimum at $T/2$ of

$$R(T/2) = \frac{1}{1 + e^{-\alpha T}} e^{-\alpha \frac{T}{2}}.$$

So while a variable can be highly correlated with another that is distant on the index, a bound limits the influence on other variables - variable pairs are able to be found with arbitrarily small influence of one on another as long as the sequence interval is long enough, which is no less a requirement in MCs. Tf x_0 and x_T are defined together as the boundary (variable), then with distance from it, i.e. $\min(t, T-t)$ for x_t, there is geometric forgetting of its influence, by

$$C(e^{-\alpha\tau} + e^{\alpha(T-\tau)}) \leq 2Ce^{-\alpha \min(\tau, T-\tau)}.$$

The effect of asymptotically enlarging the interval is that the influence of the boundary value goes geometrically to zero.

However the presence of a conditional independence (separation) rule like the reciprocal property does not guarantee even this kind of forgetting/locality. We note that reciprocal processes are Markov random fields (MRFs), and finite-state MRFs can have a different behaviour. Markov random fields are a generalisation of the Markov property, expanding the notion of the separating 'present' to some neighbouring set, such as on a 2-dimensional lattice, the 4-connected neighbours. In MRFs there can be multiple 'paths between' variables which a single-index model with the Markov property doesn't have. In certain multi-path finite-state MRFs, for example, the 2-dimensional lattice, there is a critical strength of coupling between neighbours above which the correlation function is not geometric decay but an inverse power law and for some configurations not even necessarily decaying to zero but a constant value, so that a value can influence the values of the whole rest of the model, including the boundary, even if is at infinite distance. This is referred to as long-range dependence but in a difference sense to what we have used it for RCs. In this sense it means all-range dependence and is the opposite of locality. We note that long-range dependence is specifically a finite-state phenomenon - an MRF with the same lattice structure but continuous (Gaussian) variables does not have it. [33] RCs are finite-state MRFs with a one-dimensional index [34] but for which there are multiple paths between variables, namely two, by the fact that the reciprocal property does not separate sets of variables from each other given one variable's value, but given two. There is some further discussion of the implications of RCs being MRFs in Appendix A, but given these ambiguous precedents, the first question of forgetting for finite-state reciprocal chains, is whether they can have long-range dependence in the sense of MRFs or is there a geometric bound in the sense of the continuous variable reciprocal models, in which case we could refer to RC models as a way to model long-range dependence without 'long-range dependence', and thus being amenable to some local approximation etc.

Chapter 3

Forgetting Properties of Reciprocal Chains

In this chapter we address two questions about the model itself, whether RCs have a forgetting property in the sense of the constant-variance reciprocal processes; and if so, for an RC that is based on some Markov model, given that the Markov model itself is a special case of that RC with a particular endpoints distribution, are the marginals and dynamics of the RC on the region away from the end times convergent to those of the base model? In Section 1 we answer the first question by showing that that RC forgetting is geometric in the sense that it is in constant variance reciprocal models, by deriving the bound on the Birkhoff coefficient of the two-point transition between any two variables in the model. We give an explicit counterpart for RCs to the theorem of Markov chains that bounds the difference between a marginal at time t given two different initial values. In Section 2 we address the second question using the conceptual division of the boundary distribution into a long-range endpoints relationship (transition) and an initial marginal. We bound the distance between the marginals of a reciprocal process and those of its Markov base when the Markov process is initialised with the same initial marginal as the RC (the models thus differ by their long-range transition). We show that the difference decays geometrically with distance of t from T. Then simply by considering the Markov forgetting theorem, we find that marginals far from both initial and final times are insensitive to all aspects of the initialisation. Finally we investigate the dynamics (conditional behaviour) of the model and show that the conditional distribution at some time given a set of past variables, reduces to the distribution given the latest variable (i.e. the Markov property) if the first variable in the set is relatively 'close' to the time in questions, that is, if the length of the relative complement of their index separation on the interval is long relative to the forgetting rate.

3.1 RC Forgetting Behaviour

In this section we answer whether RC forgetting is like RPs. As well as the fundamental implications for the nature of the model and its estimation as outlined, the question is motivated some contextual considerations from the modelling applications which entertain the possibility of long-memory (not forgetting) in RCs created from an MC when particular boundary distribution is applied. For example, a Markov model's transition $\mathrm{F}_{T,0}$ entries define which 'journeys' (pairs of initial and final value) are more or less likely. In the case of a simple random walker this can be proportional to how many distinct paths go between the pair. When constructing a RC by imposing an arbitrary joint distribution $\Pi_{T,0}$ the modeller can make unlikely pairs likely (as is the case when applying RCs to make walkers *'move along'*) leading to a smaller number of viable paths overall. Does the reduced diversity of paths inhibit forgetting? Note in this case there is a 'tension' between the imposed and 'natural' dynamics. Alternately the MRF case of long-range (all-range) dependence was related to the boundary values being all the same (at the same time as the neighbor relationships preferred neighbours to have the same values), so the $x_T = x_0$ case is also under suspicion. Rather than tension, here the boundary and neighbour relationships reinforce each other. Aside from upper bounding to show forgetting, a secondary question for RCs created from MCs is how *forcing forgetting*, also related to the attempt to model walkers that moving through the space, works. This is the case of applying a long-range transition that has $\tau_B = 0$, even if the base Markov model's dynamics whose transition over the interval has little or no forgetting. We wonder by what function τ_B decays from 1 to 0.

3.1.1 RC Two-Point Transitions

Note that the marginal distributions, π_0 and π_T, and 'long-range' transitions $\Phi_{T,0}$, $\Phi_{0,T}$ are prescribed by endpoint distribution $\Pi_{T,0}$,

$$\Phi^{(T,0)}_{m,h} = \frac{\Pi^{(T,0)}_{m,h}}{\pi^{(0)}_h}, \qquad \Phi^{(0,T)}_{h,m} = \frac{\Pi^{(T,0)}_{h,m}}{\pi^{(T)}_m}\ , \tag{3.1}$$

where

$$\pi^{(0)}_h = \sum_{m=1}^{n} \Pi^{(T,0)}_{m,h}, \qquad \pi^{(T)}_k = \sum_{m=1}^{n} \Pi^{(T,0)}_{k,m}\ .$$

We will derive the general *two-point transitions* for an RC generated from an MC base model. Note that these quantities are not Markov transitions, i.e. they don't, in general, satisfy the composition property $\Phi_{t,u}\,\Phi_{u,s} = \Phi_{t,s}$ for $s < u < t$.

Lemma 3.1. *Let $\{x_t : t = 0, \ldots, T\}$ be an RC generated from base Markov transitions*

$F_{t,s},\ t,s=0,\ldots,T$, and end-points distribution $\Pi_{T,0}$. Define the $n\times n$ matrix

$$\Psi^{(T,0)}_{k,h} = \frac{\Pi^{(T,0)}_{k,h}}{F^{(T,0)}_{k,h}}\ . \tag{3.2}$$

Then for $0 \le s < t \le T$,

$$\Phi_{t,s} = \left(F_{t,s} \circ \left(F'_{T,t}\,\Psi_{T,0}\,F'_{s,0}\right)\right)\ \left(\operatorname{diag}(\pi_s)\right)^{-1}. \tag{3.3}$$

Proof. By total probability

$$\Phi^{(t,s)}_{j,i} = \frac{\mathbf{P}(X_t = j, X_s = i)}{\mathbf{P}(X_s = i)}$$

Using the rule for RC variable joint distribution (2.48), the numerator is (let $t > s$ for this case)

$$\begin{aligned}
\Pi^{(t,s)}_{j,i} &= \sum_{k=1}^{n}\sum_{\ell=1}^{n} Q^{(s,t,T)}_{j,i,k}\ Q^{(0,s,T)}_{\ell,j,k}\Pi^{(T,0)}_{k,\ell} &(3.4)\\
&= \sum_{k=1}^{n}\sum_{\ell=1}^{n} \frac{\mathrm{F}^{(t,s)}_{j,i}\mathrm{F}^{(T,t)}_{k,j}}{\mathrm{F}^{(T,s)}_{k,i}}\ \frac{\mathrm{F}^{(s,0)}_{i,h}\mathrm{F}^{(T,s)}_{k,i}}{\mathrm{F}^{(T,0)}_{k,h}}\Pi^{(T,0)}_{k,\ell}\\
&= \mathrm{F}^{(t,s)}_{j,i}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\mathrm{F}^{(s,0)}_{i,h}\Psi^{(T,0)}_{k,h}\mathrm{F}^{(T,t)}_{k,j}\\
&= \left(\mathrm{F}_{t,s}\right)_{j,i}\ \left(\mathrm{F}'_{T,t}\,\Psi_{T,0}\,\mathrm{F}'_{s,0}\right)_{j,i}\ , &(3.5)
\end{aligned}$$

dividing by the marginal yields the result. It is readily verified (by simplification of the above argument), that the result holds for $s = 0$ and/or $t = T$ using the definition that $\mathrm{F}_{u,u} = I_n$ for all u. □

A time-reversed transition, that is $\Phi_{s,t},\ s < t$, simply results from normalising the transpose of the two-point joint

$$\Phi_{s,t} = \mathrm{F}'_{t,s} \circ \left(\mathrm{F}_{s,0}\,\Psi'_{T,0}\,\mathrm{F}_{T,t}\right)\ \left(\operatorname{diag}(\pi_t)\right)^{-1} \tag{3.6}$$

The two-point transition matrix thus involves two matrix chains, one which represents the inner variables on the interval between t and s, one which represents the variables in the complement, and the boundary condition. Given that each matrix is a chain, each must have a decay of its Birkhoff coefficient so that it approaches rank 1 entry-wise. The matrices interact by entry-wise product - there is a matrix rank condition, $\operatorname{rank}(A \circ B) \le \operatorname{rank}(A)\operatorname{rank}(B)$, so if both factors are actually rank 1, so must $\Phi_{t,s}$ be also. But when either is not rank 1, what happens? Until now there is no result that says what the Birkhoff coefficient of a Hadamard product is in terms of $\tau_B(.)$ of the factor matrices. Intuition suggests that two matrices with small $\tau_B(.)$ will not Hadamard-multiply to result in a large coefficient but a proof is needed so we derive this here:

3.1.2 Birkhoff's Coefficient Under Hadamard Product

We introduce now a new property of $\tau_B(.)$ – its behaviour under the elementwise (Hadamard) product.

Lemma 3.2. *For two positive matrices P and Q,*

$$\frac{|\tau_B(P) - \tau_B(Q)|}{1 - \tau_B(P)\tau_B(Q)} \leq \tau_B(P \circ Q) \leq \frac{\tau_B(P) + \tau_B(Q)}{1 + \tau_B(P)\tau_B(Q)} \tag{3.7}$$

Proof. Let $B = P \circ Q$. Let $b_{i,k}$ denote the i, k entry of B, then (2.24) and (2.26) can be expressed together as

$$\phi(B) = \max_{ijk\ell} \ln \frac{b_{ik} b_{j\ell}}{b_{jk} b_{i\ell}} \tag{3.8}$$

Note that since the maximand in (3.8) is the log of a ratio, minimisation of the same quantity is by inverting the ratio, thus

$$\min_{ijk\ell} \ln \frac{b_{ik} b_{j\ell}}{b_{jk} b_{i\ell}} = -\phi(B) \tag{3.9}$$

Since $\tau_B(B) = \tanh(\phi(B)/4)$ and tanh is strictly increasing, to upper/lower bound $\tau_B(B)$ it suffices to upper/lower bound $\phi(B)$. Expanding the right hand side of (3.8) gives

$$\begin{aligned} \phi(B) &= \max_{ijk\ell} \ln \frac{p_{ik} q_{ik} \cdot p_{j\ell} q_{j\ell}}{p_{jk} q_{jk} \cdot p_{i\ell} q_{i\ell}} \\ &= \max_{ijk\ell} \left[\ln \frac{p_{ik} p_{b\ell}}{p_{jk} p_{i\ell}} + \ln \frac{q_{ik} q_{j\ell}}{q_{jk} q_{i\ell}} \right] \end{aligned} \tag{3.10}$$

Immediately from the fact that for general functions

$$\max_x (f(x) + g(x)) \leq \max_x (f(x)) + \max_x (g(x)),$$

it holds that

$$\phi(B) \leq \phi(P) + \phi(Q). \tag{3.11}$$

For the lower bound we now need to consider that

$$\max_x (f(x) + g(x)) \geq \min \left[\max_x (f(x)) + \min_x (g(x)),\ \min_x (f(x)) + \max_x (g(x)) \right],$$

which gives

$$\phi(B) \geq |\phi(P) - \phi(Q)|. \tag{3.12}$$

Returning to the upper bound, substituting (3.11) into the τ_B definition (2.25) gives

$$\tau_B(B) \leq \tanh(\phi(P)/4 + \phi(Q)/4). \tag{3.13}$$

The hyperbolic tangent addition formula, then gives

$$\tau_B(B) \leq \frac{\tanh(\phi(P)/4) + \tanh(\phi(Q)/4)}{1 + \tanh(\phi(P)/4)\tanh(\phi(Q)/4)} \tag{3.14}$$

Using the τ_B definition again gives the upper bound in (3.7). Likewise the lower bound follows from (3.12) with the same substitutions. □

We can replace the coefficients of the factor matrices with their upper bounds and still suitably upper bound the element-wise product's coefficient:

Lemma 3.3. *For positive matrices P, Q with $\tau_B(P) \leq p$, $\tau_B(Q) \leq q$, and $B = P \circ Q$, $\tau_B(B) \leq \frac{p+q}{1+pq}$*

Proof. The gradient of the tangent addition formula is always positive for arguments $0 \leq p', q' \leq 1$

$$\Delta(\tau_B(B)) = \left[\frac{1-p'^2}{(p'q'+1)^2}, \frac{1-q'^2}{(p'q'+1)^2}\right] > 0. \tag{3.15}$$

□

So we find that small-coefficient matrices approximately add their coefficients, staying small, but also note that matrices with more significant τ_B are able to cancel to small values. Regarding attainment of the bounds, attainment of the upper bound is by coincidence of the maximising arguments in (3.8) between the factor matrices. For example, sufficient, but not necessary, is $P = Q$, such that for p=$\tau_B(P)$, $\tau_B(P \circ P) = 2p/(1+p^2)$. Attainment of the lower bound is by coincidence of the maximising arguments in (3.8) of one factor with the minimising arguments of the other, and further condition that the set of arguments maximise (3.8) for the (element-wise) product B. This occurs in the following result which gives the condition for non-rank 1 matrices' elementwise product to be rank 1 :

Lemma 3.4. *For positive matrices P, Q with $\tau_B(P) = p > 0$, $\tau_B(Q) = q > 0$, $\tau_B(P \circ Q) = 0 \Leftrightarrow q_{ik} = 1/p_{ik} \cdot \alpha_i \cdot \beta_k$ for positive vectors α, β.*

Proof. If the right hand side holds then $P \circ Q = \alpha\beta'$ and $\tau_B = 0$. Conversely, letting $B = P \circ Q$, $\tau_B(B) = 0 \Rightarrow \phi(B) = 0 \Rightarrow \frac{b_{ik}b_{j\ell}}{b_{jk}b_{i\ell}} = 1 \;\forall\; ijk\ell \Rightarrow p_{ik}q_{ik} = \alpha_i\beta_k \;\forall i, k.$ □

3.1.3 RC Forgetting Bound

We have established the form of $\Phi^{RC}_{t,s}$ and we can now use the new property of τ_B under Hadamard product, in combination with the rule for chains, to evaluate its dependence bound.

Theorem 3.1. *For an RC model generated from Markov transitions $F_{t,s}$, $t, s \in \{0, ..., T\}$, for whose next-neighbour transitions A_t, $\tau_B(A_t) \leq \lambda < 1$, let $m = \tau_B(\Psi_{T,0})$ then*

$$\tau_B(\Phi^{RC}_{t,s}) \leq \lambda^{t-s} + m\lambda^{T-(t-s)}$$

Proof. Applying Lemma 3.2, to forward (3.3) or backward (3.6) transitions, noting there is no effect on τ_B of the diagonal normalising matrix due to property (2.29),

$$\tau_B(\Phi^{RC}_{t,s}) \leq \frac{\tau_B(\mathrm{F}_{t,s}) + \tau_B(\mathrm{F}'_{T,t}\Psi_{T,0}\mathrm{F}'_{s,0})}{1 + \tau_B(\mathrm{F}_{t,s})\tau_B(\mathrm{F}'_{T,t}\Psi_{T,0}\mathrm{F}'_{s,0})} \tag{3.16}$$

$$\leq \frac{\lambda^{t-s} + m\lambda^{T-t}\lambda^s}{1 + m\lambda^{T-t}\lambda^s\lambda^{t-s}} \tag{3.17}$$

where by Lemma 3.3, we have replaced the terms in the RHS with their bounds given by the matrix product chain restult from the background chapter, Theorem 2.1. (The RHS denominator which simplifies to $1 + m\lambda^T$ is bounded above 1, we let it equal 1 for simplicity). □

Note that (3.17), in the $m = 1$ case, for $\tau = t - s$ reduces to

$$\tau_B(\Phi^{RC}_{t,t+\tau}) \leq \frac{\lambda^\tau + \lambda^{T-\tau}}{1 + \lambda^T},$$

which is the same overall bound as the continuous variable autocorrelation bound (2.62).

We are able to make a direct counterpart to the MC result of the difference between marginals given different initial distributions, corollary 2.1.1 (which includes the special case of initial values) using the separation of $\Pi_{T,0}$ into initial marginal and transition:

Corollary 3.1.1. *Let there be an RC model as above, initialised by distinct boundary-distribution choices Π^p, Π^q, which have a common 'long-range transition' $\Phi_{T,0}$ but differ in initial state distribution such that each $\Pi^i = \Phi_{T,0}\mathrm{diag}(\pi^i_0)$ for $i = p, q$. Let π^p_t, π^q_t be the unconditional marginals at t given the respective initialisations, then*

$$||\pi^p_t - \pi^q_t|| \leq C\lambda^t + mC\lambda^{(T-t)}$$

for positive constant C.

The proof follows from the coefficient bound in the same manner as in the Markov case given for corollary 2.1.1. The mid-point marginal at $t = T/2$ is bounded by $2C\lambda^{T/2}$. Thus the main conclusion of this section is reached, the mid-process marginals *always* forget x_0 (or π_0) and x_T (π_T) for sufficient T, regardless of boundary condition. The boundary condition still affects details of the forgetting behaviour, as we will discuss in the next paragraph. Then follows examples of bounds compared to the numerical value of $\tau_B(\Phi_{t,0})$ for some RCs with common generating process and different boundary conditions in subsection 2.1.5.

Boundary Condition Note that the coefficient m is $\tau_B(\Psi_{T,0})$ and not $\tau_B(\Pi_{T,0})$. Before we address the forgetting of the BC entirely, in Section 3.2, we will comment about the effect of the choice of boundary condition on the marginal forgetting behaviour. The choice of $\Pi_{0,T}$ firstly results in a value m and secondly impacts the behaviour of the dependence function below the bound. We will look into the second effect through examples in the next subsection. Regarding m, this term's magnitude results from the relationship between the long-range transition imposed by the boundary condition and that due to the base dynamics. $\Psi_{T,0}$ is itself the element-wise product of $\Pi_{T,0}$ and a matrix whose entries are the inverse of the entries of $\mathrm{F}_{T,0}$ (this matrix necessarily has Birkhoff coefficient $\tau_B(\mathrm{F}_{T,0}) \leq \lambda^T$), letting $p = \tau_B(\Pi)$, then by Lemma 3.2:

$$\frac{|p - \tau_B(\mathrm{F}_{T,0})|}{1 - p \cdot \tau_B(\mathrm{F}_{T,0})} \leq m \leq \frac{p + \tau_B(\mathrm{F}_{T,0})}{1 + p \cdot \tau_B(\mathrm{F}_{T,0})},$$

So we note, firstly that for any $\Pi_{T,0}$ that is not positive, e.g a diagonal matrix, $p = 1$, $m = 0$. Secondly $p = 0$ does not give $m = 0$, rather $p = 0 \Rightarrow m = \tau_B(\mathrm{F}_{T,0}) \leq \lambda^T$. Rather, Lemma 3.4 thus asserts the condition to attain $m = 0$, by which (if we take m's magnitude as reflecting the degree of non-Markovianity of the model) we find agreement with the result of [35] that a reciprocal process is only Markov when the boundary distribution has form $\Pi_{T,0} = \mathrm{diag}(\nu_T)\mathrm{F}_{T,0}\mathrm{diag}(\nu_0)$, for positive vectors ν_0, ν_T.

3.1.4 Numerical Examples

To illustrate that the mid-process forgetting edge values is geometric (in the sense established) regardless of boundary choice we first evaluate the bound, numerical value of τ_B, and norm between marginals of processes given different initial states, for systems representing the motivating cases mentioned in the section introduction. The base Markov model for these RCs is the road network example from the background, 'epoch' version. As in the background example, the bound resulting from $\lambda = \tau_B(A)$ is not be tight enough to be useful, so also demonstrated are bounds using λ_2, λ_4. The first boundary condition case is simply $x_T = x_0$ such that $\Phi_{T,0} = I_n$. Note that $m = 1$ by the fact that $\Pi_{T,0}$ is not positive. The second case represents the application motivated idea of encouraging targets to move across the space - a $\Phi_{T,0}$ is created where unlikely initial-final value pairs (according to the base transitions) become likely by taking the

elementwise inverse of the (positive) transition A and normalising. Note that here $\tau_B(\Phi_{T,0}) = \tau_B(A) = \lambda \approx 1$, so the bound functions are essentially the same as for the first example. Figure 3.1 (a) and (b) show that for both these cases the bounds hold, and the mid-process forgets the initial state. In (b) we see the middle forgets even more because in this case the two terms of the bound cancel rather than add. (This is like reminiscent of the hyperbolic sin function from the anticyclic boundary case in continuous variable RPs.) A look at example realisations shows this is because this Π makes many paths occupy the same state around $T/2$. The limited number of paths mentioned in the section introduction results here in more, not less forgetting.

The third example relates to the secondary question of trying to make faster forgetting: $\Pi_{T,0}$ is the outer product of the stationary vector of the transition, $\pi\pi'$, thus $\Phi_{T,0}$ is normalised to $\pi\mathbb{1}'$. We see in (c) that rather than the forgetting being at an increased rate, the RC function matches the MC function (compare to Figure 2.1) (b) until very late on the interval when the plot ends, as it's value at $T = 0$. This is due to the lower bound in (3.7). Since $p = 0$, $m = \lambda^T$, and the term in the RC transition that modifies its behavior from the base model has $\tau_B()$ only of comparable size near T. We can deduce that such a configuration does not 'speed up' forgetting, rather there is a normal decay until a collapse toward the end.

The last example shows the limitations of the bound. In this case the transition matrix is the road network example single-step primitive transition $\bar{A}$, i.e. the x-axis is single time steps not the epochs. Compare to Figure 2.1) (a). Here, while $F_{T,0}$ is positive, $\tau_B(F_{T,0}) \approx 1$. For the RC, $\Phi_{T,0}$ is reused from the previous example, having rank 1, $\tau_B = 0$. because all the MC chain terms in the bound are ≈ 1, the best bound on m is ≈ 1, and likewise $\lambda^\tau \approx 1$, $\tau \in \{0, ..., T\}$, giving the transition bound close to unity for all distances on this interval, whereas the specific behaviour of $\tau_B(\Phi_{T,0})$ is shown in Figure 3.1 (d), to be decay with not a particularly predictable function.

3.1.5 Discussion

The conclusion of this section, is that an RC must forget π_0 in the middle, and likewise π_T. We cannot choose a boundary condition transition that so that we there is not geometric forgetting, essentially with $\min(t, T - t)$ as in continuous variable models, thus RCs have a locality property. The marginals in the mid process are insensitive to the initial/final state but in the next section we establish what the marginals there actually are.

3.2 Forgetting in RCs Restores Markovianity

From the last section we learned that there's no choice of boundary condition that will prevent the mid-process from forgetting the values, or marginals, at the end points. Presuming for the moment that this also means the mid-process forgets all aspects of the entire boundary distribution, recall from the background discussion that for

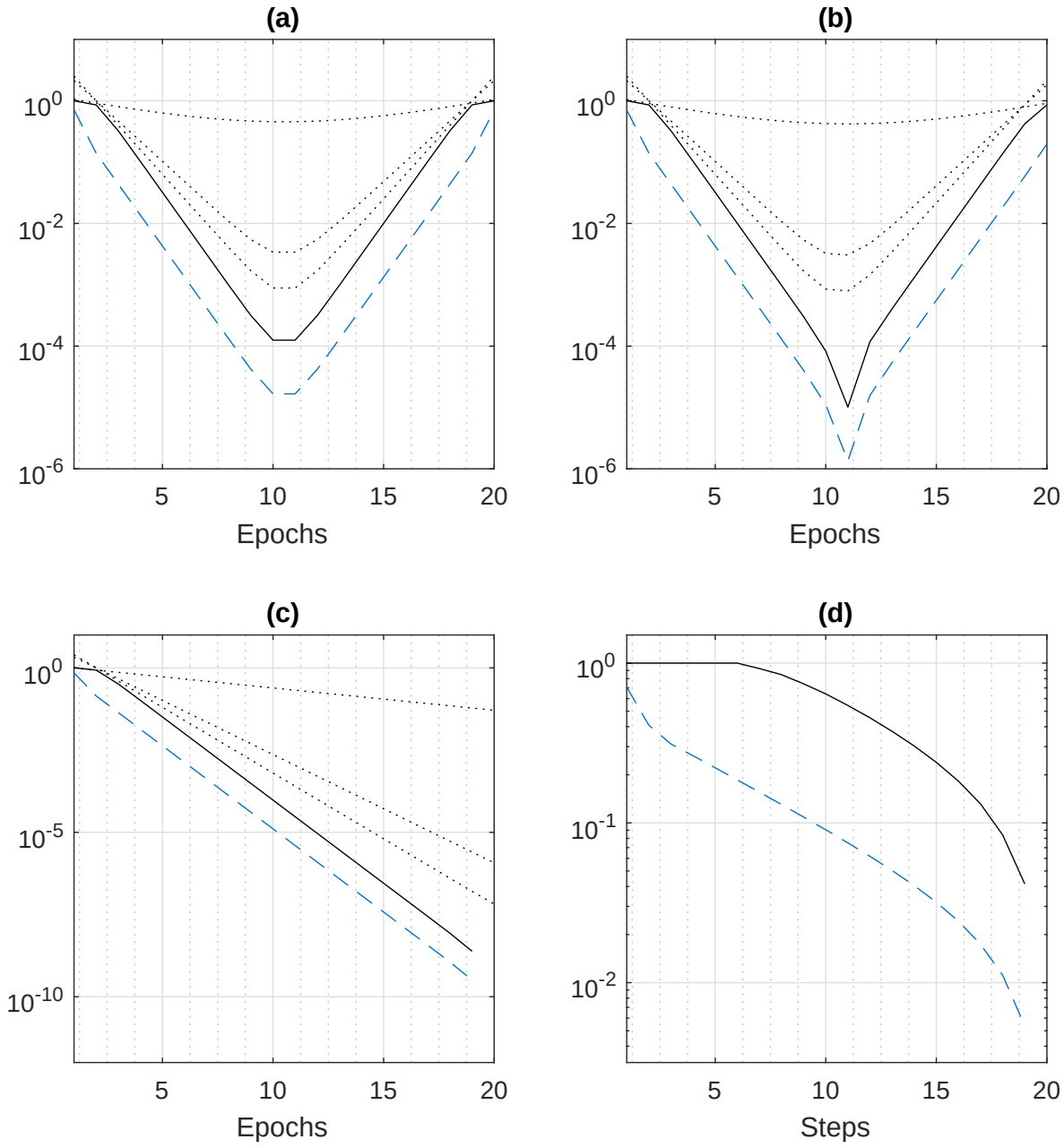

Figure 3.1: Solid line numerical $\tau_B(\Phi^{RC}_{t,0})$, dotted lines coefficient bounds with $\lambda, \lambda_2, \lambda_4$, dashed (blue) line the infinity norm between marginals at t given different initial states. See the body text for descriptions of scenarios pictured.

RCs that are based on an MC model, the Markovian base model itself results due to a special case of the boundary condition. Since to 'forget' the boundary condition means approaching behaviour shared with all BC cases including that one, forgetting the endpoints should result in matching the base model in some sense. As outlined

in the introduction, if the mid-process matches the Markovian base, for example in terms of marginals and more importantly dynamics, i.e. conditional properties, then an applying an RC to an interval which is long with respect to the forgetting rate would more often that not produce the same marginals as the simpler MC model. However when it came to estimation, would presumably mean the faster Markov estimation algorithm could be substituted on that part. We first check the marginals situation, where we find an MC model (which can be initialised with the same marginal as the RC) distributions need only diverge from those of the RC when approaching the end of the interval. The confirmation of forgetting the entire boundary by marginals mid sequence follows as a consequence of this result and the MC forgetting property. Investigating the dynamics we find it is not forgetting the boundary, but just considering 'local' sets of variables that gives Markov conditional dynamics, though not necessarily with the same transitions as the base.

3.2.1 RC Marginals Go to MC Marginals

Theorem 3.2. *Let there be an MC with transitions $F_{t,s}$ composed of A_t, $t, s \in \{0, ..., T\}$, where $\tau_B(A_t) \leq \lambda < 1$, and further entry condition $F^{(T,0)}_{j,i} \geq \gamma > 0$. Let there be an RC based on $F_{t,s}$ with joint endpoint distribution $\Pi_{T,0}$. Let the intial marginal of the RC, $\pi_0 = \Pi'_{T,0}\mathbb{1}$ be the intitialisation of the MC model. Let π^{MC}_t and π^{RC}_t be marginals at t according to of the respective models. Then*

$$||\pi^{MC}_t - \pi^{RC}_t||_\infty \leq C\lambda^{(T-t)}$$

for positive constant C.

Proof. Note $\pi^{RC}_t = \Phi^{RC}_{t,0}\pi_0$. From the $s = 0$ case of Lemma 3.1,

$$\Phi^{RC}_{t,0} = \left(\mathrm{F}_{t,0} \circ \left(\mathrm{F}'_{T,t}\,\Psi_{T,0}\right)\right)\,\left(\mathrm{diag}\,(\pi_0)\right)^{-1}.$$

Define

$$G_t = \mathrm{F}'_{T,t}\frac{\Phi_{T,0}}{\mathrm{F}_{T,0}},$$

recalling $\Phi_{T,0}$'s definition (3.1), so that $\Phi^{RC}_{t,0} = \mathrm{F}_{t,0} \circ G_t$. We show that the entries of $G_t \to 1$ with the distance of t from the final time. Note that

$$\begin{aligned} G^{(t)}_{i,j} &= \sum_k (\mathrm{F}_{T,t})_{k,i}\frac{\Phi^{(T,0)}_{k,j}}{(\mathrm{F}_{T,0})_{k,j}} \\ &= \sum_k \frac{(\mathrm{F}_{T,t})_{k,i}}{(\mathrm{F}_{T,0})_{k,j}}\Phi^{(T,0)}_{k,j} \end{aligned} \tag{3.18}$$

The 'backward' forgetting result of stochastic chains gives that $\mathrm{F}_{T,0} \to \mathrm{F}_{T,t}$ in the elementwise sense with $(T-t)$. Following [7] (Theorems 4.17 and 4.19): By the definition of $\rho(P)$, (2.37) and (2.38),

$$|F_{i,j}^{(T,t)} - F_{i,k}^{(T,t)}| \leq \tau_B(\mathrm{F}_{T,t}),\ \forall i,j,k. \tag{3.19}$$

It also holds that

$$|F_{i,j}^{(T,t)} - F_{i,h}^{(T,0)}| \leq \tau_B(\mathrm{F}_{T,t}),\ \forall i,j,h, \tag{3.20}$$

where (3.20) is obtained from (3.19) by considering $\mathrm{F}_{T,0} = \mathrm{F}_{T,t}\mathrm{F}_{t,0}$ such that

$$F_{i,h}^{(T,0)} = \sum_k F_{i,k}^{(T,t)} F_{k,h}^{(t,0)}$$

and for $\epsilon = \rho(\mathrm{F}_{T,t})$

$$\sum_k (F_{i,j}^{(T,t)} - \epsilon) F_{k,h}^{(t,0)} \leq \sum_k F_{i,k}^{(T,t)} F_{k,h}^{(t,0)} \leq \sum_k (F_{i,j}^{(T,t)} + \epsilon) F_{k,h}^{(t,0)}.$$

For all k,i,j, then

$$\left| 1 - \frac{(\mathrm{F}_{T,t})_{k,i}}{(\mathrm{F}_{T,0})_{k,j}} \right| = \left| \frac{(\mathrm{F}_{T,t})_{k,i} - (\mathrm{F}_{T,0})_{k,j}}{(\mathrm{F}_{T,0})_{k,j}} \right| \leq \frac{\tau_B(\mathrm{F}_{T,t})}{\gamma} \leq C\lambda^{(T-t)}.$$

Letting (3.18) then be written as

$$\begin{aligned} G_{i,j}^{(t)} &= \sum_k \left(1 + \epsilon_t(i,j,k)\right) \Phi_{k,j}^{(T,0)} \\ &= \sum_k \Phi_{k,j}^{(T,0)} + \sum_k \epsilon_t(i,j,k) \Phi_{k,j}^{(T,0)} \\ &= 1 + \epsilon_t(i,j), \end{aligned}$$

then

$$\begin{aligned} \pi_j^{(t),RC} &= \sum_i \mathrm{F}_{j,i}^{(t,0)} G_{j,i}^{(t)} \pi_i^{(0)} \\ &= \sum_i \mathrm{F}_{j,i}^{(t,0)} \pi_i^{(0)} + \sum_i \epsilon_t(j,i) \pi_i^{(0)} \\ &= (\pi_t^{MC})_j + \epsilon_t(j) \end{aligned}$$

where

$$|\epsilon_t(j)| \leq |\epsilon_t(j,i)| \leq |\epsilon_t(j,i,k)| \leq C\lambda^{(T-t)}.$$

□

For an example see Figure 3.2 (a). RCs are inherently bidirectional so we could ask whether this result can go backward in time, with a comparison to an MC which shares the final marginal with the RC. The problem is that an MC cannot be arbitrarily assigned a final distribution. There is such a thing as a backward-running MC, where this is possible, but the backward MC achieves the final marginal by modifying its transitions, so such an MC cannot be considered to be the 'base' model to which we wanted to compare. This is not to say that an MC cannot be created which matches the RCs marginals. In fact a model with entirely independent variables at each time can be made to match an RC's marginals, but such a model will not return anything like the RC's predictions or estimates. To do this a model needs to match the RC's dynamics, or conditional behaviour, which is what we will soon investigate. First however, for completeness we show that the middle-process (as opposed to the early-to-mid process from the previous result) forgets all aspects of the boundary condition.

3.2.2 Forgetting the Boundary in the Middle

Theorem 3.3. *For an RC model as above, initialised with distinct boundary-distribution choices $\Pi^p_{T,0}, \Pi^q_{T,0}$, let ${}^p\pi_t, {}^q\pi_t$ be marginals given the respective initialisations, then*

$$||{}^p\pi_t - {}^q\pi_t||_\infty \leq C_1\lambda^t + C_2\lambda^{(T-t)}$$

Proof. Let ${}^i\pi^{MC}_t$ be the marginal of an MC with the base transitions and the initial marginal corresponding to $\Pi^i_{T,0}$, then by Theorem 3.2,

$$||{}^i\pi_t - {}^i\pi^{MC}_t||_\infty \leq C_0\lambda^{(T-t)}$$

for positive constant C_0. Given the Markov forgetting result corollary 2.1.1 however,

$$||{}^p\pi^{MC}_t - {}^q\pi^{MC}_t||_\infty \leq C_1\lambda^t.$$

By the triangle inequality of norms

$$\begin{aligned} ||{}^p\pi_t - {}^q\pi_t||_\infty &\leq ||{}^p\pi_t - {}^p\pi^{MC}_t||_\infty + ||{}^p\pi^{MC}_t - {}^q\pi^{MC}_t||_\infty + ||{}^q\pi^{MC}_t - {}^q\pi_t||_\infty \\ &\leq C_1\lambda^t + C_2\lambda^{(T-t)} \end{aligned}$$

for $C_2 = 2C_0$. □

For an example see Figure 3.2 (b).

3.2.3 Markovianity Condition

Theorem 3.2 shows that RC marginals go to that of its Markov base away from T, but marginals are not enough to show that the Markov property attains. To show that an RC model 'goes to' a Markovian one on a long interval requires demonstration

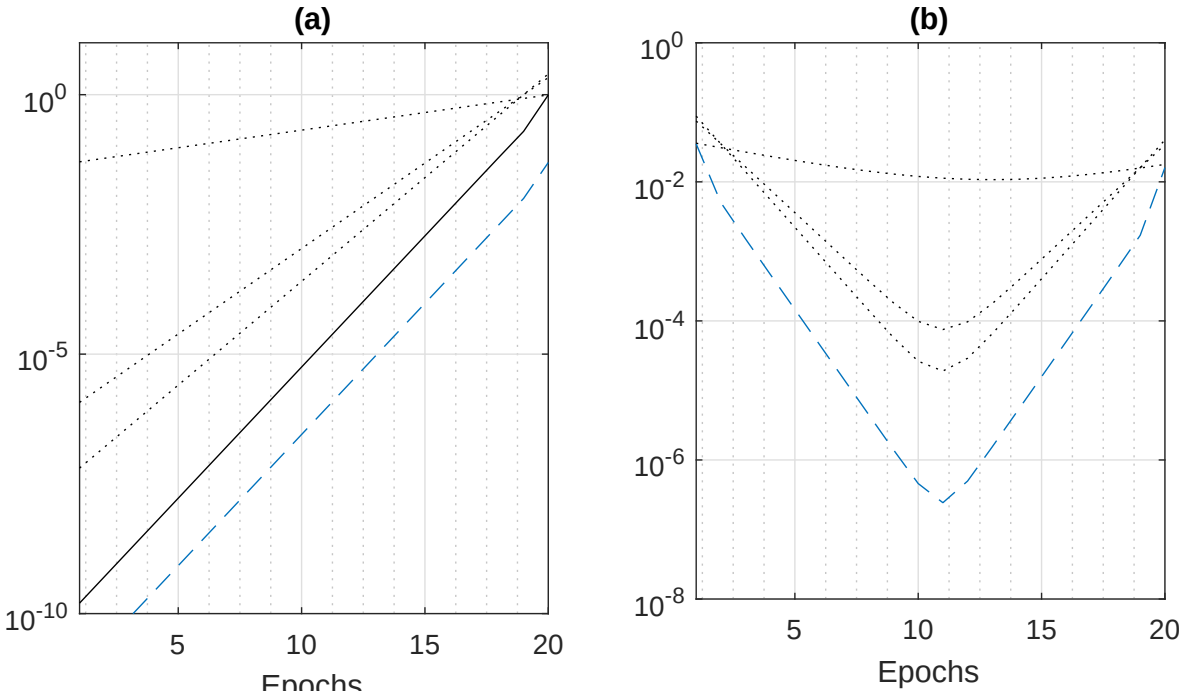

Figure 3.2: (a): Dashed (blue) line infinity norm between RC marginal and marginal of MC with the base transitions and initial marginal of the RC. Solid line numerical $\tau_B(\mathrm{F}_{T,t})$. Dotted lines $\lambda^{(T-t)}$ for $\lambda, \lambda_2, \lambda_4$. (b): Dashed (blue) line infinity norm between marginals of RCs with common transitions, different $\Pi_{T,0}$. Dotted lines $C_1\lambda^t + C_2\lambda^{(T-t)}$ for $\lambda, \lambda_2, \lambda_4$, where C_1 used is the infinity norm between the initial marginals of the $\Pi_{T,0}$s and C_2 likewise of the final marginals.

that the future conditioned on the past requires only the immediate past. There is a so far unmentioned third type of 'transition' that belongs to RCs, which reduces in MCs to the two-point. Consider, in the case of a Markov process, for any indices $\{r_0, ...r_n\}, < s < t$

$$\mathbf{P}(X_t|x_s, \{x_{r_1}, ..., x_{r_n}\}) = \mathbf{P}(X_t|x_s) \tag{3.21}$$

but for a reciprocal process, the reciprocal property separates X_t from $X_{r_n}, ..., X_{r_2}$ by x_s and x_{r_0}, but nothing else, meaning

$$\mathbf{P}(X_t|x_s, \{x_{r_0}, ..., x_{r_n}\}) = \mathbf{P}(X_t|x_s, x_{r_0}). \tag{3.22}$$

The latest value and the earliest, of a set of past variables, define the evolution of the future. Considering predictions made by a model given a set of observations, this property sets RCs as a contrast to, for example second-order Markov models, a type of model where the present and closest-past variable inform the future. To answer the question of this subsection, let r denote the earliest realized or observed variable (i.e. r_0) of a set and s the latest. Letting t_i stand for $\mathbf{P}(x_t = i)$, the distribution given the set is $\mathbf{P}(t_i|s_j, r_i)$ which has a form in terms of matrices:

Lemma 3.5.

$$\mathbf{P}(t_\ell|s_j, r_i) = c_{j,i}\ (F_{t,s})_{\ell,j}(F'_{T,t}\Psi_{T,0}F'_{r,0})_{\ell,i} \tag{3.23}$$

Proof.

$$\mathbf{P}(t_\ell|s_j, r_i) = \frac{\mathbf{P}(t_\ell, s_j, r_i)}{\mathbf{P}(s_j, r_i)} \tag{3.24}$$

The denominator $\Pi^{(s,r)}_{j,i}$ is available from (3.5), and the numerator by

$$\mathbf{P}(t_\ell, s_j, r_i) = \Pi^{(t,r)}_{\ell,i} Q^{(t,s,r)}_{\ell,j,i}$$

so that in expanded form,

$$\begin{aligned}\mathbf{P}(t_\ell|s_j, r_i,) &= \frac{(\mathrm{F}_{t,r} \circ \mathrm{F}'_{T,t}\Psi_{T,0}\mathrm{F}'_{r,0})_{\ell,i}}{(\mathrm{F}_{s,r} \circ \mathrm{F}'_{T,t}\Psi_{T,0}\mathrm{F}'_{s,0})_{j,i}} \frac{\mathrm{F}^{(t,s)}_{\ell,j}\mathrm{F}^{(s,r)}_{j,i}}{\mathrm{F}^{(t,r)}_{\ell,i}} \\ &= \frac{(\mathrm{F}'_{T,t}\Psi_{T,0}\mathrm{F}'_{r,0})_{\ell,i}}{(\mathrm{F}'_{T,t}\Psi_{T,0}\mathrm{F}'_{s,0})_{j,i}} \mathrm{F}^{(t,s)}_{\ell,j}\end{aligned} \tag{3.25}$$

which gives (3.23) with $c_{j,i} = (\mathrm{F}'_{T,t}\Psi_{T,0}\mathrm{F}'_{s,0})_{j,i}$, not dependent on ℓ, acting as normalising constant. □

The RC two-point transition can now be seen to be the special case $s = r$ of (3.23).

Investigating how this transition compares to a Markov transition rule (note, not necessarily the base model's transition rule), we find it is not actually necessary to be far from the boundary to converge:

Theorem 3.4. *For and $r < s < t \leq T$ on an RC with base as above, with additional condition that transitions and boundary are such that $(F'_{T,t}\Psi_{T,0}F'_{s,0})_{j,i} \geq \gamma > 0,\ \forall s,t,j,i$. Let $\pi^{t|s,r}$ be the conditional marginal at t, given definite (but unspecified) values taken by the process at indices s and r. Let the conditional marginal given a definite value taken only at index s be $\pi^{t|s}$. Then*

$$||\pi^{t|s,r} - \pi^{t|s}||_\infty \leq C \cdot m\lambda^{(T-(t-r))}$$

for positive constant C.

Proof. Let

$$H^{(t,r)}_{j,i} = \frac{(\mathrm{F}'_{T,t}\Psi_{T,0}\mathrm{F}'_{r,0})_{\ell,i}}{(\mathrm{F}'_{T,t}\Psi_{T,0}\mathrm{F}'_{s,0})_{j,i}}$$

so that from (3.25), $\mathbf{P}(t_\ell|s_j,r_i,) = \mathrm{F}^{(t,s)}_{j,i}H^{(t,r)}_{j,i}$ noting that ℓ is not mentioned (let ℓ be any index). Noting that $\mathrm{F}'_{s,0} = \mathrm{F}'_{r,0}\mathrm{F}'_{s,r}$, by the same argument as Theorem 3.2,

$$|1 - h^{(t,r)}_{j,i}| \leq \tau_B(\mathrm{F}'_{T,t}\Psi_{T,0}\mathrm{F}'_{r,0}) \leq \frac{m\lambda^{(T-t+r)}}{\gamma}$$

As in the previous proof, this limits the marginal difference. □

So it is not necessary to be away from the boundary, just that the length of the complement of distance between r and t is long relative to the forgetting rate. Or, noting that the m term is now present ($m = \tau_B(\mathbf{M})$), that the boundary condition's long range transition is close to the base model's. Consider that to make a prediction by an RC model, given a set of values (such as sightings of a target), in contrast to alternative non-Markov models such as n'th order Markov for which further past variables are less relevant, the further into the past is an early 'sighting' given to an RC model, the more relevant.

3.2.4 Discussion

We have shown the model forgets, that it has locality in the sense of a distance $\min(t-s, T-(t-s))$, and that considering a 'local' set of variables, the forgetting behaviour of an RC means these variables have similar behaviour to a Markov model. With respect to the concept of moving target modelling the implication of the results of Section 2 is that if an RC is intended to be used to induce a tendency to proceed purposefully, the interval length should be short relative to the forgetting rate. On too long an interval, an RC walker will be not much less likely to 'double back' on its recent position than an MC walker. However, in tracking applications where the relative position at the start and end of the interval is important for distinguishing targets, an RC model is useful even while the estimates in the mid-interval will be the same as the MC counterpart.

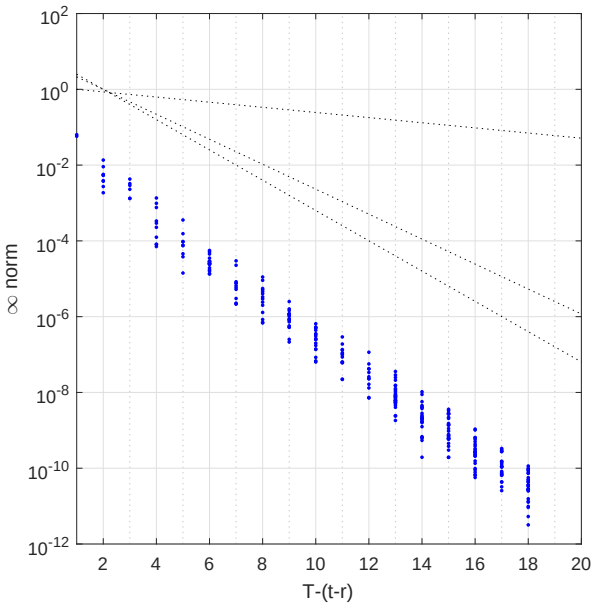

Figure 3.3: Scatter plot of infinity norm between $\pi^{t|s,r}, \pi^{t|s}$ for randomly chosen r, s, t and indices $X_t = i, x_s = j, x_r = k$. Dotted lines $\lambda^{T-(t-r)}$ for $\lambda, \lambda_2, \lambda_4$.

Before drawing conclusions it is necessary to show that the properties described carry over to the estimation algorithm the way the Markov model's properties carry to its algorithm, We address this in the next chapter, and thereby find that some of the concerns about interval length can be addressed by changing the algorithm.

Chapter 4

Forgetting Properties of Reciprocal Chain Estimators

In this chapter we address the forgetting/locality property of the HRC smoothing algorithm, and find its practical application in an approximate smoothing algorithm. In Section 1 we show that the interval smoothing algorithm from [20] has a matrix element-wise product form, with the same structure as that of the model, so that the same concept of locality applies. We develop the result that (smoothed) estimates of variables away from the boundary forget the boundary distribution and are therefore well approximated by those of the MC algorithm (forward-backward) for variables far from the boundary. In Section 2 we exploit the locality property and fact of RCs having 'Markov like' local structure evidenced by Theorem 3.4, to attain an approximate interval smoothing algorithm for RCs that achieves estimation speed-up at small cost in accuracy.

4.1 Forgetting of Observations in HRCs

In this section, by finding the matrix form of the HRC estimation algorithm we establish that the estimator forgetting behaviour is like that of the model, and give a formalisation of the approach of the HRC algorithm's estimates to those of the HMC algorithm (using the base transitions) for variables further from the interval ends.

4.1.1 RC Smoother Matrix Chain Forms

We will now show that the smoothed posterior distribution γ_t^{RC} of an HRC given by the algorithm of [20] (recounted in Section 2.2.2) has a matrix product/element-wise product form, again working with the case of an RC which is based on known/given MC transitions.

Lemma 4.1. *With $\hat{F}_{t,s}$ defined as in* (2.21) *and γ_t^{RC} as in* (2.60), *$\Phi_{T,0}$ as in* (3.1) *and $\hat{\pi}_s = c_s \pi_s$,*

$$\gamma_t^{RC} = (\hat{F}_{t,s} \circ \hat{F}'_{T,t} \hat{\Psi}_{T,0} \hat{F}'_{s,0}) \hat{\pi}_s. \tag{4.1}$$

Proof. Expanding (2.59),

$$\alpha_t^k = c_t B_{t-1}^k ... c_1 B_0^k c_0 \pi_0^k. \tag{4.2}$$

Let d_t^k be a diagonal matrix formed from the k'th column of the matrix $\mathrm{F}_{T,t}$. Note that B_t^k (see (2.55)) can be represented

$$B_t^k = d_{t+1}^k \; A_t \; (d_t^k)^{-1}. \tag{4.3}$$

In substituting (4.3) into the chain of (4.2), note the cancellation that occurs between adjacent increments so that for some t

$$\begin{aligned} B_t^k \; c_t \; B_{t-1}^k \; c_{t-1} &= d_{t+1}^k \; A_t \; (d_t^k)^{-1} \; c_t \; d_t^k \; A_{t-1} \; (d_{t-1}^k)^{-1} \; c_{t-1} \\ &= d_{t+1}^k \; A_t \; c_t \; A_{t-1} \; c_{t-1} \; (d_{t-1}^k)^{-1} \end{aligned}$$

by commutativity of diagonal matrices, so that the chain of (4.2) simplifies to

$$\alpha_t^k = d_t^k \; \hat{\mathrm{F}}_{t,0} \; (d_0^k)^{-1} \; \hat{\pi}_0^k$$

Expanding the matrix product to a sum

$$\begin{aligned} \alpha_j^k(t) &= \sum_{i=1}^N \mathrm{F}_{k,j}^{(T,t)} \hat{\mathrm{F}}_{j,i}^{(t,0)} / \mathrm{F}_{k,i}^{(T,0)} \hat{\pi}_i^{(0),k} \\ &= \sum_{i=1}^N \mathrm{F}_{k,j}^{(T,t)} \hat{\mathrm{F}}_{j,i}^{(t,0)} \hat{\Phi}_{k,i}^{(T,0)} \end{aligned}$$

for $\hat{\Phi}_{T,0} = \Phi_{T,0} c_0$. Meanwhile

$$\beta_t^k = B_t'^k c_{t+1} ... B_{T-1}'^k c_T \mathbb{1}$$

but note that

$$\begin{aligned} B_{j,i}^{(T-1),k} &= A_{j,i}^{(T-1)} \frac{(I_n)_{k,j}}{A_{k,i}^{(T-1)}} \\ B_{T-1}^k &= \mathbf{e}_k \mathbb{1}' \end{aligned}$$

where $\mathbf{e}_k$ is a vector of zeros except k'th entry 1, so that using the substitution of (4.3)

$$\begin{aligned} \beta_j^{(t),k} &= (d_t^k)^{-1} A_t \; d_{t+1}^k \; c_{t+1} \; (d_{t+1}^k)^{-1} \; A_{t+1} \; d_{t+1}^k ... \mathbb{1} \mathbf{e}_k' c_T \mathbb{1} \\ &= \frac{\hat{\mathrm{F}}_{k,j}^{(T,t)}}{\mathrm{F}_{k,j}^{(T,t)}}. \end{aligned}$$

Combining as in (2.60),

$$\begin{aligned}\gamma_j(t) &= \sum_{k=1}^{N}\left(\sum_{i=1}^{N}\mathrm{F}_{k,j}^{(T,t)}\hat{\mathrm{F}}_{j,i}^{(t,0)}\hat{\Phi}_{k,i}^{(T,0)}\right)\frac{\hat{\mathrm{F}}_{k,j}^{(t,0)}}{\mathrm{F}_{k,j}^{(T,t)}}(\pi_T)_k \\ &= \sum_{k=1}^{N}\left(\sum_{i=1}^{N}\hat{\mathrm{F}}_{j,i}^{(t,0)}\hat{\Phi}_{k,i}^{(T,0)}\right)\hat{\mathrm{F}}_{k,j}^{(T,t)}(\pi_T)_k \\ &= \sum_{i=1}^{N}\hat{\mathrm{F}}_{j,i}^{(t,0)}\sum_{k=1}^{N}\hat{\mathrm{F}}_{k,j}^{(T,t)}\hat{\Phi}_{k,i}^{(T,0)} \\ &= \sum_{i=1}^{N}\hat{\mathrm{F}}_{j,i}^{(t,0)}(\hat{\mathrm{F}}_{T,t}'\hat{\Phi}_{T,0})_{j,i}(\pi_0)i.\end{aligned}$$

Dividing c_0 from $\hat{\Phi}_{T,0}$ and forming $\hat{\pi}_0$ by $A\,\mathrm{diag}(x)y = A(y \circ x)$, we attain

$$\gamma_t^{RC} = (\hat{\mathrm{F}}_{t,0} \circ \hat{\mathrm{F}}_{T,t}'\Phi_{T,0})\hat{\pi}_0. \tag{4.4}$$

Let $\hat{\Psi}_{T,0} = \Psi_{T,0}c_0$. Subsuming $\hat{\pi}_0$ into $\hat{\Psi}_{T,0}$ so that

$$\gamma_t^{RC} = (\hat{\mathrm{F}}_{t,0} \circ \hat{\mathrm{F}}_{T,t}'\hat{\Psi}_{T,0})\mathbb{1}$$

and rearranging by the transformation

$$(A \circ B)x = \mathrm{diag}(B'A) \cdot x, \tag{4.5}$$

it holds that

$$\gamma_t^{RC} = \mathrm{diag}(\hat{\mathrm{F}}_{t,0}\hat{\Psi}_{T,0}'\hat{\mathrm{F}}_{T,t}). \tag{4.6}$$

We can reverse (4.5), splitting the chain at any point to get the general dependence matrix for an estimate on a particular observation, giving the result. □

From (4.1) it can be seen that the estimates mirror the model result, with the influence of observation y_s (via c_s in $\hat{\pi}_s$) on estimate γ_t^{RC} being bounded by $C\bar{\lambda}^{t-s} + C\bar{\lambda}^{T-(t-s)}$, analogously to Theorem 3.1, with $\bar{\lambda}$ the estimator forgetting rate. This immediately suggests that an approximate point smoothing algorithm for HRCs could be designed to use only a local subset of observations, and reduce $\sim 2n^3T$ operations to $\sim 2n^3\Delta$ for some adequate Δ. However when we address approximate estimation in Section 2 we will find that the reduction in complexity when switching from exact to approximate 'local' estimation of HRCs is more dramatic than it was for HMCs. Before addressing alternative algorithms, we will show that when smoothed by the HRC algorithm, estimates of variables away from the boundary forget the boundary distribution. (Note first from the proof above, the form (4.6) of γ_t^{RC}, which facilitates subsequent results.)

4.1.2 Estimates Forget the Boundary

The (4.1) form of γ_t^{RC} given by Lemma 4.1 shows that estimates will forget distant data. In this subsection we formalise the intuitive notion that estimates far from the boundary also forget the boundary distribution. That is, when an HRC is smoothed by the algorithm provided with the same base transitions but two different initialisations, the posterior distributions of variables increasingly far from the boundary become increasingly the same at a geometric rate. The result implies the error due to an initialisation mismatched to the system generating the trajectories, is limited to the interval edges. For this and the rest of this chapter's results we will benefit from defining some vectors, whose entries do not necessarily have probabilistic meaning: For a positive matrix B, define positive vectors $\alpha(B), \beta(B)$

$$\alpha(B)_j = \sum_{i=1}^{n} B_{j,i}, \qquad \beta(B)_i = \sum_{j=1}^{n} B_{i,j}. \tag{4.7}$$

Note that $\alpha(B) = B\mathbb{1},\ \beta(B) = B'\mathbb{1}$. Then define

$$\alpha_t^{RC} = \alpha(\hat{\mathrm{F}}_{t,0}\hat{\Psi}'_{T,0}\hat{\mathrm{F}}_{T,t}) \ = \hat{\mathrm{F}}_{t,0}\hat{\Psi}'_{T,0}\hat{\mathrm{F}}_{T,t}\mathbb{1} \tag{4.8}$$

$$\beta_t^{RC} = \beta(\hat{\mathrm{F}}_{t,0}\hat{\Psi}'_{T,0}\hat{\mathrm{F}}_{T,t}) \ = \hat{\mathrm{F}}'_{T,t}\hat{\Psi}_{T,0}\hat{\mathrm{F}}'_{t,0}\mathbb{1}. \tag{4.9}$$

(For notational convenience we henceforth assume these quantities to be normalised without explicit indication.) These quantities can be cheaply evaluated in right-to-left order by T matrix-vector multiplications, therefore having $\mathcal{O}(n^2T)$ complexity, but are not necessarily meaningful quantities for a general HRC model. For the special case of an RC which is Markov i.e. where $\Pi = \hat{\mathrm{F}}_{T,0}\mathrm{diag}(\pi_0)$ for some π_0, such that $\Psi = \pi_0\mathbb{1}'$, we find

$$\begin{aligned}\alpha_t^{RC} &= \hat{\mathrm{F}}_{t,0}\pi_0\mathbb{1}'\hat{\mathrm{F}}'_{T,t}\mathbb{1} \\ &= \hat{\mathrm{F}}'_{T,t}\pi_0 \\ &= \alpha_t^{MC}\end{aligned}$$

for α_t^{MC} defined in (2.1), i.e. the forward pass of the forward-backward algorithm. Likewise

$$\begin{aligned}\beta_t^{RC} &= \hat{\mathrm{F}}'_{T,t}\mathbb{1}\pi_0\hat{\mathrm{F}}'_{t,0}\mathbb{1} \\ &= \hat{\mathrm{F}}'_{T,t}\mathbb{1} \\ &= \beta_t^{MC}\end{aligned}$$

and by (4.6), since the matrix involved is rank 1, it's diagonal $\gamma_t^{RC} = \alpha_t^{MC} \circ \beta_t^{MC}$. Thus in the MC case γ_t^{RC} coincides with that due to the MC algorithm, and γ_t^{RC} could be evaluated in $\mathcal{O}(n^2T)$ time by forward-backward algorithm. That γ_t in the Markovian

case is the element-wise product of two vectors reflects the split of the observation set into an independent past and future that facilitates the use of fast matrix-vector multiplications to get the posterior distributions. In non-Markov RCs, the matrix within (4.6) is not rank 1 and to obtain its diagonal, the full matrix must be evaluated by matrix-matrix multiplication.

To show that the posterior distributions from the estimator forget the initialisation, we develop the argument that the diagonal of the matrix in (4.6) approaches the element-wise product of two vectors when the interval is long, and that these vectors forget the boundary far from it. We will require some lemmas. The following lemma bounds the distance between vectors that are element-wise products:

Lemma 4.2. *For positive vectors x_1, x_2, y_1, y_2 and $z_1 = x_1 \circ y_1,\ z_2 = x_2 \circ y_2$*

$$d(z_1, z_2) \leq d(x_1, x_2) + d(y_1, y_2) \tag{4.10}$$

Proof. From the definition of $d(.)$ (2.24),

$$\begin{aligned}
d(z_1, z_2) &= \max_{i,j} \ln \left[\frac{z_i^{(1)} \cdot z_j^{(2)}}{z_j^{(1)} \cdot z_i^{(2)}} \right] \\
&= \max_i \ln \frac{z_i^{(1)}}{z_i^{(2)}} \ + \ \max_i \ln \frac{z_i^{(2)}}{z_i^{(1)}} \\
&= \max_i \ln \left[\frac{x_i^{(1)} y_i^{(1)}}{x_i^{(2)} y_i^{(2)}} \right] \ + \ \max_i \ln \left[\frac{x_i^{(2)} y_i^{(2)}}{x_i^{(1)} y_i^{(1)}} \right] \\
&\leq \max_i \ln \frac{x_i^{(1)}}{x_i^{(2)}} + \max_i \ln \frac{y_i^{(1)}}{y_i^{(2)}} \ + \ \max_i \ln \frac{x_i^{(2)}}{x_i^{(1)}} + \max_i \ln \frac{y_i^{(2)}}{y_i^{(1)}} \\
&\leq d(x_1, x_2) + d(y_1, y_2)
\end{aligned}$$

□

When a matrix is rank 1, the vector formed from it's diagonal is an element-wise product of two vectors. For a non-rank 1 matrix, the following lemma bounds the difference between a matrix' diagonal and this product as a function of τ_B. This lemma assumes normalisation of relevant quantities.

Lemma 4.3. *For positive matrix B, where $tr(B) = \sum_i B_i i \leq 1$*

$$||\mathrm{diag}(B) - \alpha(B) \circ \beta(B)||_\infty \leq \tau_B(B)$$

Proof. Consider the stochastic matrix $A = B \cdot \mathrm{diag}(\beta(B))^{-1}$. Note $\tau_B(A) = \tau_B(B)$. By property (2.29). By the definition of $\rho(A)$ (2.37) and bound (2.38)

$$\max_{ijk} |A_{i,j} - A_{i,k}| \leq \tau_B(A)$$

so that or a vector π, defined

$$\pi_j = \frac{1}{n}\sum_{i=1}^{n} A_{j,i},$$

it holds that

$$||\pi\mathbb{1}' - A||_\infty \leq \rho(A)$$

and follows that

$$||\pi\mathbb{1}'\mathrm{diag}(\beta(B)) - A\beta(B)||_\infty \leq \rho(A)\max_j(\beta(B)_j)$$

giving the result by $\max_j(\beta(B)_j) \leq 1$. □

The theorem follows:

Theorem 4.1. *Let there be HRC models with distinct boundary distributions ${}^p\Pi_{T,0}$,${}^q\Pi_{T,0}$, and let ${}^p\gamma_t$,${}^q\gamma_t$ denote the estimates due to the HRC smoothing algorithm supplied with the respective models. Let the HRCs have base model transitions and observation models in common, with estimate forgetting rate $\bar{\lambda}$. Let ${}^p\Psi$,${}^q\Psi$ be the matrices associated with the respective boundary conditions as in* (3.2) *and let ${}^pm = \tau_B({}^p\Psi)$,${}^q m = \tau_B({}^q\Psi)$. Then*

$$||\ {}^p\gamma_t - \ {}^q\gamma_t||_\infty \leq C_1\bar{\lambda}^t + C_1\bar{\lambda}^{T-t} + ({}^pm + {}^q m)\bar{\lambda}^T \tag{4.11}$$

for some constants C_1, C_2.

Proof. Consider vectors, $\psi = \hat{\Psi}'_{T,0}\hat{\mathrm{F}}_{T,t}\mathbb{1}$, $\bar{\psi} = \hat{\Psi}_{T,0}\hat{\mathrm{F}}'_{t,0}\mathbb{1}$ so that

$$\alpha_t^{RC} = \hat{\mathrm{F}}_{t,0}\psi, \tag{4.12}$$
$$\beta_t^{RC} = \hat{\mathrm{F}}'_{T,t}\bar{\psi}. \tag{4.13}$$

Particular boundary condition i of ${}^i\Pi_{T,0}$ gives the associated ${}^i\Psi$, and thus a pair ${}^i\psi$,${}^i\bar{\psi}$, finally ${}^i\alpha_t^{RC}$, ${}^i\beta_t^{RC}$. Then by (2.30),(4.12),(4.13),

$$d({}^p\alpha_t^{RC},\ {}^q\alpha_t^{RC}) \leq C_1\bar{\lambda}^t \tag{4.14}$$
$$d({}^p\beta_t^{RC},\ {}^q\beta_t^{RC}) \leq C_2\bar{\lambda}^{(T-t)} \tag{4.15}$$

From Lemma 4.3 , for $i = p, q$ (dropping the 'RC' superscript for now)

$${}^i\gamma_t = \ {}^i\alpha_t \circ \ {}^i\beta_t + \ {}^i\epsilon_t$$

for a vector ${}^i\epsilon_t$, $||{}^i\epsilon_t||_\infty \leq \ {}^im \cdot \bar{\lambda}^T$ so that

$${}^p\gamma_t - \ {}^q\gamma_t = \ {}^p\alpha_t \circ \ {}^p\beta_t + \ {}^p\epsilon_t - \ ({}^q\alpha_t \circ \ {}^q\beta_t + \ {}^q\epsilon_t)$$

and

$$||^p\gamma_t - {}^q\gamma_t||_\infty \leq ||\ {}^p\alpha_t \circ\ {}^p\beta_t - {}^q\alpha_t \circ\ {}^q\beta_t||_\infty + ({}^pm +^q m)\bar{\lambda}^T \tag{4.16}$$

By (4.14),(4.15) and Lemma 4.2

$$d({}^p\alpha_t \circ\ {}^p\beta_t,\ {}^q\alpha_t \circ\ {}^q\beta_t) \leq C_1\bar{\lambda}^t + C_2\bar{\lambda}^{(T-t)}. \tag{4.17}$$

Letting the quantities be normalised the first term in the RHS of (4.16) is bounded by the LHS of (4.17). □

Theorem 4.1 is illustrated in Figure 4.1. The example system is the road network system from Section 2.2.3 with $\Pi_{T,0}$ inducing targets to cross the area, as in Figure 2.2 (b). The observation model is as described in Section 2.1.5. In (a) numerical evaluations of the signal model and estimation forgetting rates are depicted. In (b) two models are used to smooth 400 observation sequences, the RC that generated the data (i.e. matched model), and a model whose $\Pi_{T,0}$ is mismatched. The error, defined as the infinity norm difference between γ_t from the models, is shown. A bound on the error according to Theorem 4.1 is given with the rate $\bar{\lambda}$ determined from (a).

4.1.3 HMC vs. HRC Derived Estimates

A smoother with a mismatched initialisation includes cases where the generating distribution is an RC but the smoother is given an initialisation that makes its model Markov. So Theorem 4.1 implies a bound on the error that will result from naively using the forward-backward algorithm on an HRC as if it were an HMC (using the base RC's base transitions as the MC model). In fact the mismatched model in (b) is an RC with a Markovian $\Pi_{T,0}$. The same (mismatched) posterior distributions could therefore be obtained with either the HRC or HMC algorithms. We could propose to substitute the HMC algorithm's estimates for the middle variables of long RC processes to save calculation. However in the next section we show that by exploiting 'locally Markov' structure, successful approximation of an HRC by an HMC-like algorithm can be extended to the entire interval. (Note that the numerically evaluated error in Figure 4.1 is much less than the bound because the Markovian $\Pi_{T,0}$ is far from the most mismatched initialisation that can be applied.)

4.2 Fast Approximate Smoother for RCs

In this section we show how HRCs can be approximately smoothed with recursive matrix-vector calculations. First, an approximate HRC point smoother is introduced which treats the end part of the interval as adjacent to the beginning, and passes recursive updates over this junction. A regime for interval smoothing an HRC is introduced which exploits 'approximate redundancy' in the calculations of the point

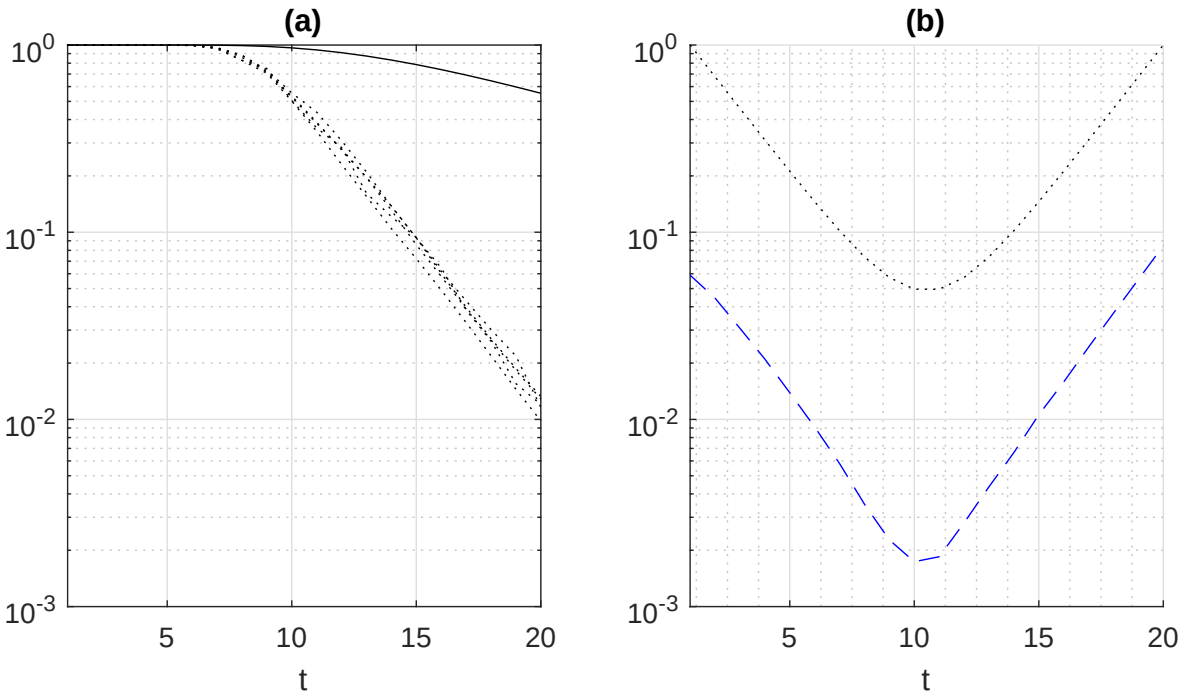

Figure 4.1: Smoothing by matched and mismatched model. System as in Figure 2.2 (b) with $n = 16$, $p = 0.5$. Observation model described in Section 2.1.5. (a) Rates of forgetting. Solid line model, dotted lines estimator forgetting rate for 30 example sequences. (b) Error by mismatched model. Dotted line: bound from Theorem 4.1 with rate attained from (a). (Blue) dashed line-infinity norm between smoothed posterior distributions due to matched and mismatched (Markov) model.

smoother, so that posterior distributions are given which are a second approximation to the exact ones, and thus have some additional error to the approximate point smoothed distributions, but via far fewer calculations.

4.2.1 Point Smoother

For special Markovian case the calculation of the vector γ_t^{RC} was split into two vectors by the matrix of (4.6) being rank 1, which allowed interval smoothing with forward-backward by matrix-vector multiplication with complexity $\mathcal{O}(n^2T)$, whereas the non-Markovian case the matrix within (4.6) is not rank 1 and to obtain its diagonal, the full matrix must be evaluated by matrix-matrix multiplication with complexity $\mathcal{O}(n^3T)$. However considering that (for any t), the matrix in (4.6) is a chain of T matrices, if the interval is long with respect to the estimator forgetting rate, Lemma 4.3 gives that the diagonal, γ_t^{RC}, is close to the product of two vectors. Specifically, the vectors $\alpha_t^{RC}, \beta_t^{RC}$, which are obtainable by matrix-vector multiplications. Define the approximate poste-

rior point distribution

$$\bar{\gamma}_t = \alpha_t^{RC} \circ \beta_t^{RC}, \tag{4.18}$$

then by Lemma 4.3

$$||\bar{\gamma}_t^{RC} - \gamma_t^{RC}||_\infty \leq m\bar{\lambda}^T. \tag{4.19}$$

An algorithm for approximate point smoothing an HRC then is to simply evaluate $\alpha_t^{RC}, \beta_t^{RC}$ recursively by (4.8), (4.9) and then $\bar{\gamma}_t^{RC}$ by (4.18). This procedure resembles the forward-backward algorithm, but now treating X_T and X_0 as adjacent variables, and treating $\Psi_{T,0}$ as a transition despite it's not being a stochastic matrix. While HRC point smoothing this method is fast, estimating the distributions of an interval of variables by this method is not proposed to be practical: In the HMC forward-backward algorithm, interval smoothing is effectively repreated point smoothing because of redundancy between the set of distributions. That is, for each t,

$$\alpha_{t+1}^{MC} = U_t \; \alpha_t^{MC},$$

etc. However in the HRC case, it can be seen from (4.8) that

$$\alpha_{t+1}^{RC} \neq B \; \alpha_t^{RC}$$

for any matrix B. Without this redundancy, interval smoothing by $\bar{\gamma}_t$ would involve repeating the process for each variables in the interval and an overall complexity $\mathcal{O}(n^2T^2)$. In the next subsection, investigation of how much extra error results from recursing anyway, results in a fast second-approximation algorithm suitable for interval smoothing.

4.2.2 Interval Smoothing Regime

We show that vectors can be obtained which approximate the approximate point smoother's component vectors, without unacceptably higher error. Consider for an index t and increment τ, such that $t + \tau < T$, attempting to approximate $\alpha_{t+\tau}^{RC}$ by recursion from α_t^{RC}. From (4.8), the point smoother vector we hope to attain for $t + \tau$ is

$$\alpha_{t+\tau}^{RC} = \hat{\mathrm{F}}_{t+\tau,t}\hat{\mathrm{F}}_{t,0}\Psi_{T,0}\hat{\mathrm{F}}_{T,t+\tau}\mathbb{1} \tag{4.20}$$

while α_t^{RC} can be expanded

$$\begin{aligned}\alpha_t^{RC} &= \hat{\mathrm{F}}_{t,0}\Psi_{T,0}\hat{\mathrm{F}}_{T,t}\mathbb{1} \\ &= \hat{\mathrm{F}}_{t,0}\Psi_{T,0}\hat{\mathrm{F}}_{T,t+\tau}\hat{\mathrm{F}}_{t+\tau,t}\mathbb{1}\end{aligned}$$

The presence of the term $\hat{\mathrm{F}}_{t+\tau,t}$ already, within α_t^{RC}, prevents $\alpha_{t+\tau}^{RC}$ being attainable from α_t^{RC} by a multiplication by $\hat{\mathrm{F}}_{t+\tau,t}$. However, defining the approximation

$$\begin{aligned}\bar{\alpha}_{t+\tau|t}^{RC} &= \hat{\mathrm{F}}_{t+\tau,t} \cdot \alpha_t^{RC} \\ &= \hat{\mathrm{F}}_{t+\tau,t} \cdot \hat{\mathrm{F}}_{t,0} \cdot \Psi_{T,0} \cdot \hat{\mathrm{F}}_{T,t}\mathbb{1} \\ &= \hat{\mathrm{F}}_{t+\tau,t} \cdot \hat{\mathrm{F}}_{t,0} \cdot \Psi_{T,0} \cdot \hat{\mathrm{F}}_{T,t+\tau} \cdot \hat{\mathrm{F}}_{t+\tau,t}\mathbb{1} \qquad (4.21)\end{aligned}$$

and comparing (4.20) and (4.21), it can be seen that

$$\begin{aligned}d(\bar{\alpha}_{t+\tau|t}^{RC}, \alpha_{t+\tau}^{RC}) &\leq \tau_B(\hat{\mathrm{F}}_{t+\tau,t} \cdot \hat{\mathrm{F}}_{t,0} \cdot \Psi_{T,0} \cdot \hat{\mathrm{F}}_{T,t+\tau}) \qquad (4.22) \\ &\leq m\bar{\lambda}^T \qquad (4.23)\end{aligned}$$

which is independent of increment τ. In other words there is some additional error, by recursing in the manner of a forward pass, but no accumulation of that error to get a set of $\bar{\alpha}^{RC}$. The above argument applies equally to backward recursion used to attain an equivalently defined $\bar{\beta}_{t-\tau}^{RC}$. Since the error in these terms is independent of τ, define

$$\bar{\bar{\gamma}}_t = \bar{\alpha}_t^{RC} \circ \bar{\beta}_t^{RC}, \qquad (4.24)$$

where $\bar{\alpha}_t^{RC} = \bar{\alpha}_{t|t-\tau}^{RC}$ for some τ, likewise $\bar{\beta}_t^{RC}$. Then by Lemma 4.3, (4.23) and Lemma 4.2:

$$||\bar{\bar{\gamma}}_t - \gamma_t^{RC}||_\infty \leq 3m\bar{\lambda}^T. \qquad (4.25)$$

Define then an algorithm to approximately interval smooth an HRC:

Algorithm 1: Fast Approximate HRC Interval Smoother

```
Evaluate α_0^RC by (4.8);
ᾱ_0^RC ← α_0^RC;
for (t = 1; t ≤ T; t ← t + 1) do
    ᾱ_t^RC ← U_{t−1} ᾱ_{t−1}^RC;
end
Evaluate β_T^RC by (4.9);
β̄_T^RC ← β_T^RC;
for (t = T − 1; t ≥ 0; t ← t − 1) do
    β̄_t^RC ← U'_{t+1} β̄_{t+1}^RC;
end
for (t = 0; t ≤ T; t ← t + 1) do
    ⎓γ_t ← ᾱ_t^RC ∘ β̄_t^RC;
end
```

The error in this approximate algorithm's posterior distributions is bounded by (4.25).

4.2.2.1 Complexity

Evaluating $\alpha_0^{RC}, \beta_T^{RC}$ each require $\sim n^2T$ Each subsequent recursive pass requires another T matrix-vector multiplications so that the total calculations of this scheme for the entire interval's estimates then is $\sim 4Tn^2$ vs $\sim 2Tn^2$ for interval smoothing an HMC with forward backward. As mentioned n may be large in applications, so an accurate approximation with complexity $\mathcal{O}(n^2T)$ is significant. Timing results are given in the following example problem, and in Figure 4.3 (b), timing is compared for exact and approximate interval smoothing algorithms vs. state-space size n.

4.2.3 Approximate Estimation Example

To demonstrate the applicability of the proposed algorithm, it is used to smooth an example HRC. The system tested is again the road network example as in Section 2.2.3, and Figures 2.2 (b) and 4.1. A sample trajectory from a road network with its associated sequence of observations is shown in Figure 4.2 (a). Note this is not the exact system used for the simulation study, the illustrated target travels through a larger state space in a shorter time ($n = 25,\ T = 10$) for visual simplicity. The smoothed system parameters are $n = 16,\ T = 20$. For Figure 4.2 (b), 400 observation sequences were smoothed by the approximate point and fast approximate interval HRC smoothers. Error was measured as the infinity norm distance between these posterior distributions and the exact posterior distributions obtained by the 'bridge-decomposition' algorithm. The error is charted along with bounds (4.19) and (4.25), with $\bar{\lambda}$ derived from the data of Figure 4.1 (a). The error resultant from naively using the Markov-assuming forward-backward smoother (with the base transitions), as in Figure 4.1 is shown for comparison. The CPU time for the algorithms to smooth 400 sequences was as follows: HRC exact smoother 20.9s, approximate point smoother 45.0s, naive forward-backward 1.0s, fast approximate interval smoother, 1.9s. The fast approximate interval smoother is of comparable speed to the MC with vastly less error. In Figure 4.3 (a), the speed of computation for the exact, naive HMC and proposed algorithm are compared for smoothing systems as in this example but with varying n. In the next subsection the general relationship of error between the methods are briefly explored.

4.2.4 Discussion

In Section 1 of this chapter it was shown that naively applying the forward-backward algorithm to an HRC as if it were an HMC (with the base transitions) produces estimates that are increasingly close to those due to the HRC smoother, the further the variable estimated is from the boundary. Considering an HRC model constructed from base transitions and a $\Pi_{T,0}$ applied over some required interval length T - when T is long relative to the forgetting rate, more variables are far from the boundary and the time-average error is low, however the end time relationship, which can be important for applications such as tracking that aim to keep track of multiple trajectories, must

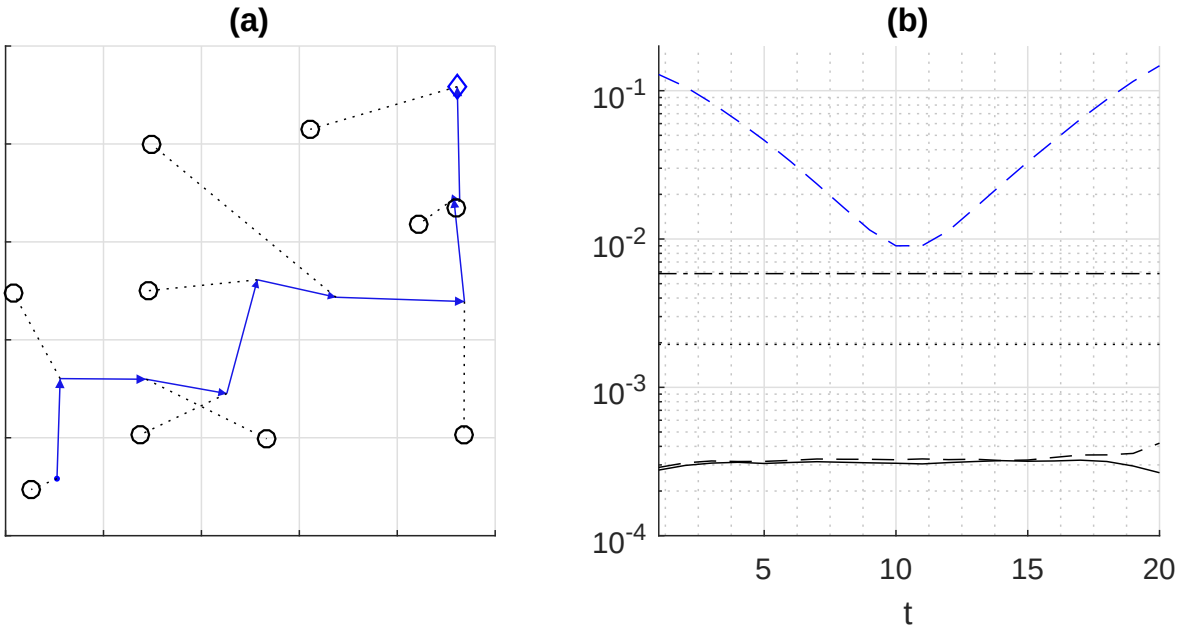

Figure 4.2: Example smoothing problem. (a) An illustration of an example trajectory and observation sequence. (b) Infinity norm error in approximate (slow) point smoother $\bar{\gamma}_t$ (solid line) and fast interval smoother $\bar{\bar{\gamma}}_t$ (dashed line). Dotted line is point-smoother error bound $m\bar{\lambda}^T$, dot-dash line interval smoother error bound $3m\bar{\lambda}^T$. Dashed blue line error due naive forward-backward smoothing as in Figure 4.1.

still be modelled. The HRC model on the long interval can therefore be considered beneficial but inefficient. The proposed fast approximate interval smoothing algorithm error bound is geometric with the interval length. The approximation can thus produce accurate estimates precisely in the case that it is needed. Figure 4.3 compares the error due to applying naive HMC forward-backward (which decreases roughly linearly with T) and due to the proposed algorithm (geometric).

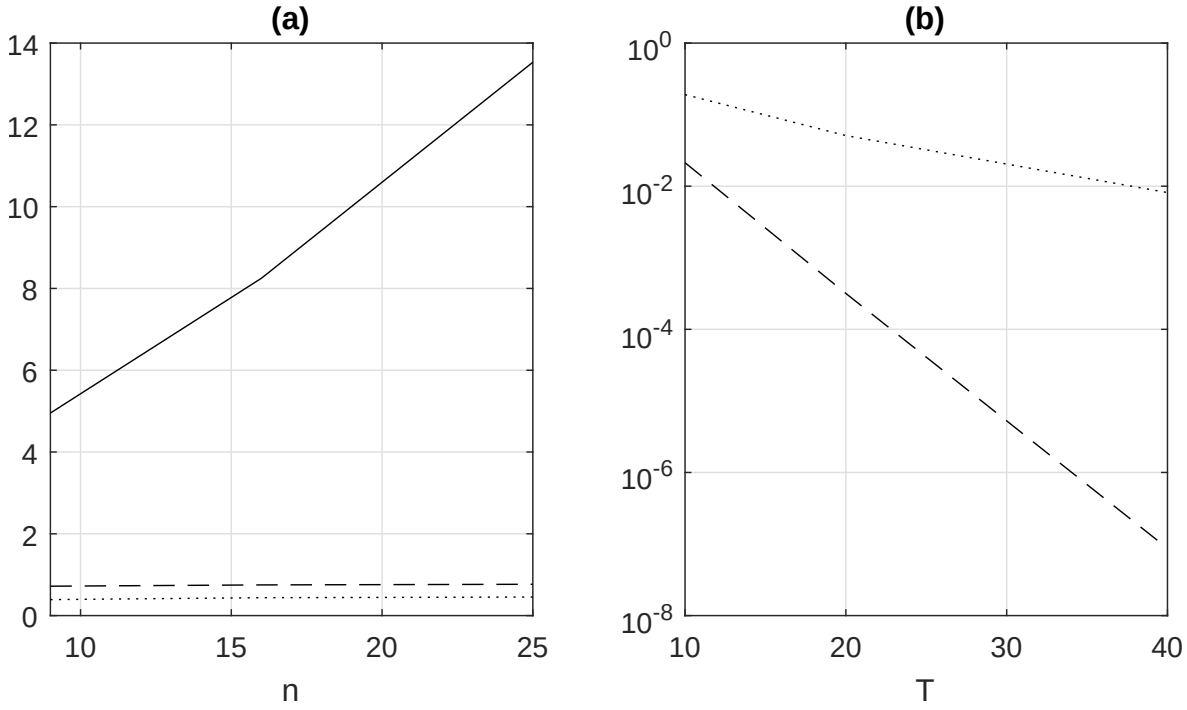

Figure 4.3: (a) CPU time (s) to interval smooth 400 observations sequences generated by a system model as in example of section 4.2.3 but with varied state size n. By the exact HRC smoother (solid line), forward-backward HMC smoother (dotted line) and fast approximate HRC interval smoother (dashed line). (b) Error (average infinity norm distance between exact posterior distributions and approximations) from interval smoothing sequences from example of Section 4.2.3 but with varied interval length T ($n = 16$), by the naive forward-backward HMC smoother (dotted line) and fast approximate HRC interval smoother (dashed line).

Appendix A

Further Related Work - Markov Random Fields

The theory of reciprocal models was given rigorous probabilistic treatment in the 1970s in [35], since then, though drawing on even earlier models, some theory has been developed that deals with probabilistic models not limited to time-indexed processes, such as multiple finite-indices (lattices), or generally, collections of variables. In this appendix we review some more current related work, for further insight into RCs and looking for overlap between this theory and our results. No results seem equivalent to ours, the closest is that the approximate estimation algorithm 'loopy belief propagation' (LBP) almost corresponds to the point smoother of Section 4.2. We show however that for RCs specifically, our algorithms improve on LBP.

A Markov random field (MRF) (see [36]) is a set of variables with a more general Markov property than the one given in (2.4). Instead of the future separated from the past by the present, a variable is separated from the rest of the set by some 'neighbouring' set. The reciprocal property can be seen to be an example. The joint distribution of the variables and be expressed in terms of conditional probabilities, for example as in (2.50), however, the conditional independence structure implied by the separation rules induces a graph over the variables (the 'graphical model'). Following [37], assuming that all probabilities are nonzero, the Hammersley-Clifford theorem [38] guarantees that the probability distribution will also factorise into a product of functions of the maximal cliques of the graph, where cliques are 'all-to-all' connected sub-graphs. Marginal distributions can be found from this parametrisation by summing out the functions over all other variables. Given the positivity condition, the two representations can be converted into each other. There is however some established theory about forgetting/locality in models parametrised by the clique functions, and because there are some schemes/algorithms for exact and approximate estimation of general models parametrised that way.

In this thesis we have given results about RCs that are generated from some MC transitions. In his rigorous treatment of reciprocal models in [35], B. Jamison at-

tempted to answer the question of whether all reciprocal models that can be specified are generated from some base Markov transitions. His conclusion was not definitive, and a finite-index, finite-state example was actually given as a counter-example. In Section 1 of this appendix we will show, by showing the conversion between RC parametrisations by clique functions and transitions, that for a finite-index, finite-state RC with positive three-point transitions *on interval of size* $T > 2$, a base process always exists. In Section 2 we will discuss whether the known results forgetting in MRFs cover our own results, and in Section 3 we will compare an MRF estimation algorithms applied to an RC by the conversion to clique function form, with our algorithms.

A.1 RCs and Clique Functions

To find the graph for a model, its set of variables are initially assumed to be all mutually connected, the conditional independence rules then specify which connections are culled. For example, an MC where each variable's past and future are separated results in a singly connected line (including the observation variables present in the HMC, each variable X_t has a 'branch' which is Y_t). Note that the graph does not depend on the the content of the MRF's parametrising conditional distributions ('transitions' in the MC and RC terminology). In the MC case all cliques consist of two variables ($\{X_{t+1}, X_t\}$ for $t \geq 1$) so the model is called 'pairwise' and each clique function is some $n \times n$ matrix. These are therefore like the transitions, but clique functions are not limited to have column sums equal to 1, or other probabilistic characteristics (only to be positive). A positive RC naturally has a graphical model. Its form was first presented as a specific finding in [39] however this work asserts the condition of the RC having cyclic boundary condition, which is unnecessary. Applying the reciprocal property to culling connections, each variable is left connected to its neighbours (as in the MC) but as the reciprocal property does not separate X_0 and X_T, these are connected resulting in a graph that resembles a ring. It can thus immediately be seen that the distance, mentioned throughout the thesis starting from the background section 2.2.4, $\min(t-s, T-(t-s))$, with which an RC's variables were shown to forget, is graph distance on the RC's graph. The structure of the RC graph is like that of an MC with an additional function between X_0, X_T. The graph is still pairwise but not 'singly-connected' (one path between any two variables) as was the case for MCs.

A.1.1 Most RCs Have a Markovian Base Model

In this subsection we use the clique functions form of an RC to ground some of our RCs material in some established theory, specifically that $\Psi_{T,0}$ defined in 3.2 is a potential function (hence our use of the Ψ symbol heretofore), and to show that a set of Markov transitions that serves as the base model can always be found for a positive RC on $T > 2$. It will be useful to first address specifying a Markov chain by potential functions.

Markov Chain Potential Functions For an MC on $\{0, ..., T\}$, the conditional independence structure and the Hammersley-Clifford theorem gives that

$$\mathbf{P}(X_0 = x_0, X_1 = x_1, ..., X_T = x_T) = \frac{1}{Z} \prod_{t=0}^{T-1} \Psi_{X_{t+1}, X_t}(x_{t+1}, x_t),$$

for normalisation constant Z and a set of functions. Let $\Psi_{t+1,t}$ stand for a function between $\{X_{t+1}, X_t\}$ and $\Psi_{j,i}^{(t+1,t)}$ for its entries. An example set of functions for an MC can be obtained from it's transitions factorisation (2.14),

$$\Psi_{1,0}{=}\mathbf{P}(x_1|x_0)\mathbf{P}(x_0),\ \Psi_{t+1,t}{=}\mathbf{P}(x_{t+1}|x_t)$$

for $t \geq 1$. In this case the Ψ have probabilistic meanings, and columns summing to 1 etc. Clique functions can be arbitrary positive functions, however so an implication of the Hammersley-Clifford theorem is that any functions should be able to specified and still result in meaningful MC transitions: Consider, for arbitrary functions

$$\pi_j^{(0)} \ = \ \mathbf{P}(X_0 = j) = \sum_{i_1=1,...,i_T=1} \Psi_{i_1,j}^{(1,0)} \Psi_{i_2,i_1}^{(2,1)} ... \Psi_{i_T,i_{T-1}}^{(T,T-1)}$$

which can be organised

$$= \sum_{i_1=1} \Psi_{i_1,j}^{(1,0)} \sum_{i_2=1} \Psi_{i_2,i_1}^{(2,1)} ... \sum_{i_T=1} \Psi_{i_T,i_{T-1}}^{(T,T-1)}. \tag{A.1}$$

Define positive vectors h_t where

$$h_j^{(t)} = \sum_{i_{t+1}=1} \Psi_{i_{t+1},j}^{(t+1,t)} h_{i_{t+1}}^{(t+1)},$$

and $h_j^{(T-1)} = \sum_{i_T=1} \Psi_{i_T,j}^{(T,T-1)}$. Then there is a set of Markov transitions $A_t, t \in \{0, ..., T-1\}$ given by

$$\begin{aligned} A_{\ell,j}^{(t)} &= \mathbf{P}(X_{t+1} = \ell | X_t = j) \\ &= \frac{\mathbf{P}(X_{t+1} = \ell, X_t = j)}{\mathbf{P}(X_t = j)} \\ &= \Psi_{\ell,j}^{(t+1,t)} \frac{h_\ell^{(t+1)}}{h_j^{(t)}}. \end{aligned}$$

Note that $h_0 = \pi_0$.

Reciprocal Chain Potential Functions Moving on to RCs we see from the model that for a positive RC on $T > 2$,

$$\mathbf{P}(X_0 = x_0, ..., X_T = x_T) = \frac{1}{Z}\Psi_{X_T,X_0}(x_T, x_0)\prod_{t=0}^{T-1}\Psi_{X_{t+1},X_t}(x_{t+1}, x_t),$$

This means any RC that can be specified can be represented by this set of arbitrary functions. To see base transitions can be found for any such RC, consider the subset of functions other than $\Psi_{T,0}$. Setting $\Psi_{T,0}$ be any rank 1 matrix vw' for positive vectors, let

$$\begin{aligned}\Psi'_{1,0} &= \Psi'_{1,0}\text{diag}(w)\\ \Psi'_{T,T-1} &= \text{diag}(w)\Psi'_{T,T-1}.\end{aligned}$$

Carrying out the calculations to find Markov transitions for the functions above, gives a set that serves as a base model for that RC, from which three-point transitions can be constructed and an arbitrary $\Psi_{T,0}$ or $\Pi_{T,0}$ can be applied. Note that the counter example where the transitions could not be found in [35] was postivie but on interval of length $T = 3$. In this case the reciprocal property does not separate any variables and the maximum clique size is not 2 but 3. This is the only case where the above development does not apply.

It will be useful for comparing estimation algorithms to show the relation between $\Psi_{T,0}$ and $\Pi_{T,0}$.

$$\begin{aligned}\Pi^{(T,0)}_{k,h} &= \Psi^{(T,0)}_{k,h}\sum_{i_1=1,...,i_{T-1}=1}\Psi^{(1,0)}_{i_1,j}\Psi^{(2,1)}_{i_2,i_1}...\Psi^{(T,T-1)}_{k,i_{T-1}}\\ &= \Psi^{(T,0)}_{k,h}\sum_{i_1=1}\Psi^{(1,0)}_{i_1,j}\sum_{i_2=1}\Psi^{(2,1)}_{i_2,i_1}...\sum_{i_T=1}\Psi^{(T,T-1)}_{k,i_{T-1}}\\ &= \Psi^{(T,0)}_{k,h}\text{F}^{(T,0)}_{k,h}.\end{aligned}$$

So for a nominated set of base transitions and joint $\Pi_{T,0}$, the potential function $\Psi_{T,0}$ is given by

$$\Psi^{(T,0)}_{k,h} = \frac{\Pi^{(T,0)}_{k,h}}{\text{F}^{(T,0)}_{k,h}}.$$

Base Model From Three-point Transitions We have shown how a set of transitions can be found for an RC specified by clique functions, but it is more likely in the signal processing context that a model will be specified by its conditional distributions (transitions). Although all application works to date (e.g. [13, 28]) deal with processes where the base (two-point) transitions are known *a priori*, in the case where example sequences are available and the counting of relative frequencies is used to obtain joint and conditional distributions, the three-point transitions and end-point joint will be

available but the base model transitions not explicitly. In this case it seems like iteration may be required to find a complete set of base transitions, however this should not be necessary to implement the forgetting analysis in with the methods given in this thesis, due to the fact that the 'bridge transitions' of (2.50), are able to be counted from example sequences, and are base transitions for the model on $t \in \{0, ..., T-2\}$. While the last transition is not avaiable, the bridge set has the same forgetting rate as all sets of base transitions for a given RC do. This fact is due to all transitions being common up to column and row multiplications which cancel between adjacent transitions (as can be seen for example in (4.3).)

A.2 Locality/Forgetting in MRFs

The Hammersley-Clifford theorem means that MRFs such as RCs can be specified by the potential functions as well as conditional (transitions). A model that is therefore by this equivalent to an $n = 2$ RC is the 1-dimensional Ising model of magnetic spins on a lattice (see [40]), which is specified by the relationships between nearest-neighbours on a single index, and the condition that the first and last variable on the interval have the same value. This model has geometric forgetting like the RC [41]. There is a 2-dimensional version (the square lattice Ising model, see also [40]) which famously has a 'phase change' behaviour which means that for certain nearest-neighbour clique function values, the variable values on the boundary determine the distribution of all the other variables in the mode, even for infinite lattice size. In other words, this model does not necessarily have forgetting. Because the 1-D Ising has forgetting and whereas the 2-D has long-range dependence behaviour, it can be asked whether a general rule exists about models which would automatically show that 1-dimensional models (which are nearest-neighbour connected) necessarily have locality. There is a test called Dobrushin's uniqueness condition see [42], which is a candidate for this role. It is not within our scope to summarise Dobrushin's condition, except to say that it involves requirements on the strength of influence and importantly the number of neighbours a variables has. Applying a test of Dobrushin's condition to an RC converted to its clique function form, may have provided the general conclusion of 3.1, that RCs have locality, but our work is useful for giving specific bounds, for example on the norm between marginals, which would not be provided by Dobrushin's condition.

A.3 MRF Estimation

The algorithms for exact estimation of MRFs parametrised by clique functions amount to reorganisation of the otherwise combinatorial summation of the rest of the model, using the structure to avoid redundant calculation. The optimal (exact) RC smoothing algorithm of [20], by the bridge decomposition, corresponds to one variant, the clique-tree algorithm [43]. Since our work has dealt with approximation, we consider an

approximate algorithm for MRFs which would appear to have a similar approach:

A.3.1 MRF Approximate Estimation

In the background chapter section 2.1.1 we described the forward-backward algorithm, which in light of the clique functon form of an MC can be seen to be an instance of organising the summing out of potential functions When a model is singly connected the procedure, which from a single variable's point of view consists of receiving 'messages' α_t, β_t and passing them on after adding the local observation, gives the exact posterior. Thus the general form of forward-backward, applicable to singly connected models, is called message passing or belief propagation. When the graph has loops, it has been proposed to run message passing anyway, multiply counting the model's observation, until the local messages either convergence or not. This algorithm is called loopy belief propagation (LBP). The RC graph has (is) a single loop as described. So LBP can be applied to RCs as was first pointed out in [44], treating Gaussian RCs. [37, 45, 46] are the significant works relating to performance of LBP on a single loop, where the two main questions are convergence and whether accuracy of the resultant estimate relative to the exact. An example in [37] is analysed which is equivalent to a clique function parametrised, $n = 2$ RC. The argument for convergence of LBP on the single loop is similar to showing forgetting. In fact the general condition for convergence of LBP is given in [42] to be Dobrushin's condition - the same as the locality condition.

A.3.2 Loopy Belief Propagation vs. RC Approximate Algorithms

[37] features the matrix (4.6), and the author acknowledges that the diagonal of that matrix is the exact (interval smoothed) estimate. However the premise of LBP is different to that of the algorithms given in Chapter 4. Where the aim of our algorithms was for forgetting to allow the replacement of matrix-matrix calculations with matrix-vector, without multiply counting and observations, LBP anticipates that forgetting will allow convergence of messages, and multiply counting all of the observations will 'cancel out' to the correct estimate. In work such as [37] this can be shown not to occur even on the single loop, the estimates converge, but not to the exact estimate. This raises the question of whether it would be better for LBP to stop counting after incorporating all observations once, rather than converging. LBP run on an RC, stopped at T passes so that each observation was counted once by each direction of message pass, almost corresponds to the point smoother algorithm of Chapter 4 (LBP (over) counts one additional observation). As is shown in Figure A.1, allowing LBP to converge, increases the error from the single-counting case. It is reasonable to conclude that loopy belief propagation should not be employed for approximate interval smoothing of RCs, instead either stopped version, i.e. the point smoother of section Chapter 4 (applied to each time) for best accuracy, (n^2T^2 calculations) or the RC approximate

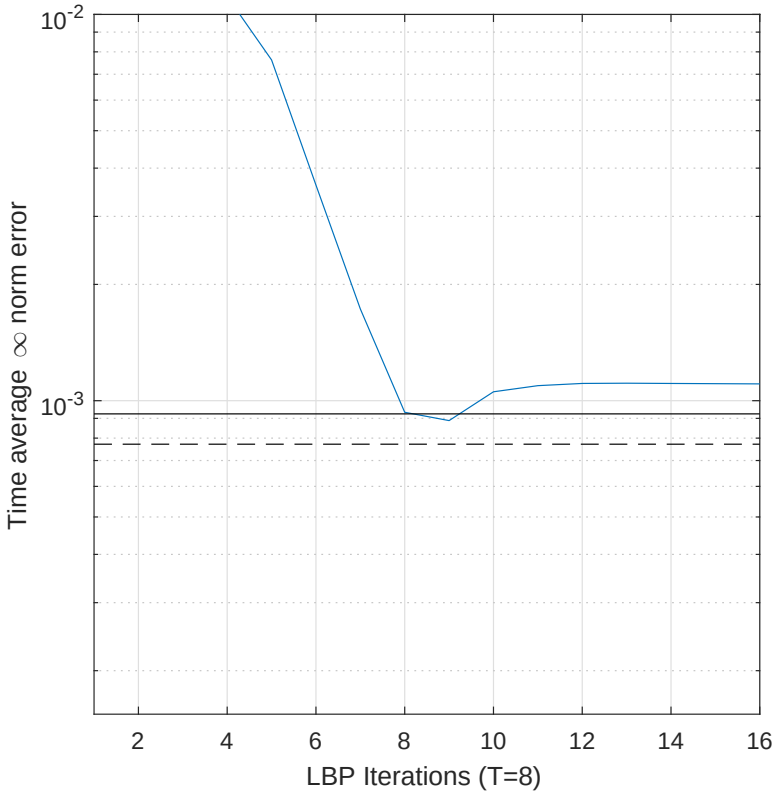

Figure A.1: Comparison of error (infinity norm between approximate and exact estimate) due to point smoother (dashes line) and interval smoother (solid line) of this work, and loopy belief propagation (blue line), vs. number of message passes of belief propagation. The model is that of (b) from Figure 3.1, on a shorter interval of $T = 8$.

interval smoothing regime ($4n^2T$ calculations) for best efficiency.

Appendix B

Proof of Faster Forgetting in Two-State Filter

In the $n = 2$, symmetric transition homogeneous MC it is possible to prove that the inclusion of any 'informative' observations, that is, where the observation is not equally likely to have produced by either state value, increases the forgetting rate of the estimation algorithm with respect to the model itself, i.e. $\tau_B(\hat{\mathrm{F}}_{t,s}) < \tau_B(\mathrm{F}_{t,s})$ as t increases.

Proof. Any symmetric 2x2 transition can be expressed

$$A = \frac{1}{2}\begin{bmatrix} 1+\beta & 1-\beta \\ 1-\beta & 1+\beta \end{bmatrix}$$

$-1 \leq \beta \leq 1$. Without loss of generality let $0 \leq \beta \leq 1$, then

$$A * A = \frac{1}{2}\begin{bmatrix} 1+\beta^2 & 1-\beta^2 \\ 1-\beta^2 & 1+\beta^2 \end{bmatrix},$$

Let diagonal matrix c have positive diagonal values c_1, c_2. Since incorporating observations replaces some instance of $A * A$ with $A * c * A$ we can show

$$\begin{aligned} A * c * A &= \frac{1}{2}\frac{1}{2}\begin{bmatrix} 1+\beta & 1-\beta \\ 1-\beta & 1+\beta \end{bmatrix}\begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}\begin{bmatrix} 1+\beta & 1-\beta \\ 1-\beta & 1+\beta \end{bmatrix} \\ &= \frac{c_1+c_2}{4}\begin{bmatrix} (1+\beta^2)+2\beta\frac{c_1-c_2}{c_1+c_2} & 1-\beta^2 \\ 1-\beta^2 & (1+\beta^2)+2\beta\frac{c_2-c_1}{c_1+c_2} \end{bmatrix} \\ &= \frac{c_1+c_2}{4}\begin{bmatrix} (1+\beta^2)+\epsilon & 1-\beta^2 \\ 1-\beta^2 & (1+\beta^2)-\epsilon \end{bmatrix} \end{aligned}$$

so that

$$\begin{aligned} \phi(A * c * A) &= \ln \frac{(1+\beta^2)+\epsilon}{1-\beta^2}\frac{(1+\beta^2)-\epsilon}{1-\beta^2} \\ &= \ln \frac{(1+\beta^2)^2-\epsilon^2}{(1-\beta^2)^2} \leq \ln \frac{(1+\beta^2)^2}{(1-\beta^2)^2} = \phi(A * A) \end{aligned}$$

τ_B is monotonic increasing with ϕ. The two ϕ are only equal when $\epsilon = 2\beta\frac{c_1-c_2}{c_1+c_2} = 0$, i.e. $c_1 = c_2$. □

Bibliography

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice Hall Press, 3rd ed., 2009.

[2] B. Anderson and J. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

[3] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, pp. 257–286, 1989.

[4] L. D. Stone, T. L. Corwin, and C. A. Barlow, *Bayesian Multiple Target Tracking*. Norwood, MA, USA: Artech House, Inc., 1st ed., 1999.

[5] S. Davey, N. Gordon, I. Holland, M. Rutten, and J. Williams, *Bayesian Methods in the Search for MH370*. Springer, 2016.

[6] A. Papoulis and S. Pillai, *Probability, random variables, and stochastic processes*. McGraw-Hill electrical and electronic engineering series, McGraw-Hill, 2002.

[7] E. Seneta, *Non-negative matrices and Markov chains*. Springer Science & Business Media, 1981.

[8] G. Birkhoff, "Extensions of Jentzsch's theorem," *Transactions of the American Mathematical Society*, pp. 219–227, 1957.

[9] S. Gaubert and Z. Qu, "Dobrushins ergodicity coefficient for markov operators on cones," *Integral Equations and Operator Theory*, vol. 81, no. 1, pp. 127–150, 2015.

[10] V. Climenhaga, "The PerronFrobenius theorem and the Hilbert metric," https://www.math.uh.edu/ climenha/blog-posts/hilbert-metric-2.pdf.

[11] L. Shue, B. Anderson, and S. Dey, "Exponential stability of filters and smoothers for hidden Markov models," *Signal Processing, IEEE Transactions on*, vol. 46, no. 8, pp. 2180–2194, 1998.

[12] J. E. Cohen and P. H. Sellers, "Sets of nonnegative matrices with positive inhomogeneous products," *Linear Algebra and its Applications*, vol. 47, pp. 185–192, 1982.

[13] G. Stamatescu, L. B. White, and R. Bruce-Doust, "Track extraction with hidden reciprocal chains," *IEEE Transactions on Automatic Control*, vol. PP, no. 99, pp. 1–1, 2017 DOI: 10.1109/TAC.2017.2741919.

[14] T. Kirubarajan, Y. Bar-Shalom, K. R. Pattipati, and I. Kadar, "Ground target tracking with variable structure imm estimator," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 36, pp. 26–46, Jan. 2000.

[15] B. Pannetier, K. Benameur, V. Nimier, and M. Rombaut, "VS-IMM using road map information for a ground target tracking," in *Proc. 7th Int. Conf. Information Fusion*, vol. 1, pp. 8 pp.–, July 2005.

[16] C.-C. Ke, J. G. Herrero, and J. Llinas, "Comparative analysis of alternative ground target tracking techniques," in *Proc. Third Int. Conf. Information Fusion FUSION 2000*, vol. 2, pp. WEB5/3–WEB510 vol.2, July 2000.

[17] B. Jamison, "Reciprocal processes: The stationary Gaussian case," *The Annals of Mathematical Statistics*, pp. 1624–1630, 1970.

[18] A. J. Krener, "Realizations of reciprocal processes," in *Modelling and Adaptive Control*, pp. 159–174, Springer, 1988.

[19] B. Jamison, "The Markov processes of Schroedinger," *Probability Theory and Related Fields*, vol. 32, no. 4, pp. 323–331, 1975.

[20] L. B. White and F. Carravetta, "Optimal smoothing for finite state hidden reciprocal processes," *Automatic Control, IEEE Transactions on*, vol. 56, no. 9, pp. 2156–2161, 2011.

[21] E. Schrödinger, *Über die Umkehrung der naturgesetze*. Verlag Akademie der wissenschaften in kommission bei Walter de Gruyter u. Company, 1931.

[22] S. Bernstein, "Sur les liaisons entre les grandeurs aleatoires," 1932.

[23] P. Dai Pra, "A stochastic control approach to reciprocal diffusion processes," *Applied Mathematics & Optimization*, vol. 23, no. 1, pp. 313–329, 1991.

[24] B. C. Levy, R. Frezza, and A. J. Krener, "Modeling and estimation of discrete-time Gaussian reciprocal processes," *IEEE Transactions on Automatic Control*, vol. 35, no. 9, pp. 1013–1023, 1990.

[25] G. Stamatescu, A. Dick, and L. B. White, "Multi-camera tracking of intelligent targets with hidden reciprocal chains," in *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*, pp. 1–8, IEEE, 2015.

[26] M. Fanaswala, V. Krishnamurthy, and L. White, "Destination-aware target tracking via syntactic signal processing," in *Proc. Speech and Signal Processing (ICASSP) 2011 IEEE Int. Conf. Acoustics*, pp. 3692–3695, May 2011.

[27] M. Fanaswala and V. Krishnamurthy, "Detection of anomalous trajectory patterns in target tracking via stochastic context-free grammars and reciprocal process models," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 1, pp. 76–90, 2013.

[28] M. Fanaswala and V. Krishnamurthy, "Spatiotemporal trajectory models for met-alevel target tracking," *IEEE Aerospace and Electronic Systems Magazine*, vol. 30, pp. 16–31, Jan. 2015.

[29] N. van Kampen, "Remarks on Non-Markov Processes," *Brazilian Journal of Physics*, vol. 28, pp. 90 – 96, 06 1998.

[30] S. C. Chay, "On quasi-Markov random fields," *Journal of Multivariate Analysis*, vol. 2, no. 1, pp. 14–76, 1972.

[31] J.-P. Carmichael, J.-C. Massé, and R. Theodorescu, "Processus gaussiens stationnaires réciproques sur un intervalle," *CR Acad. Sci. Paris Sér. I Math*, vol. 295, no. 3, pp. 291–293, 1982.

[32] B. C. Levy and A. Ferrante, "Characterization of stationary discrete-time Gaussian reciprocal processes over a finite interval," *SIAM journal on matrix analysis and applications*, vol. 24, no. 2, pp. 334–355, 2002.

[33] J. M. Moura and N. Balram, "Recursive structure of noncausal Gauss-Markov random fields," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 334–354, 1992.

[34] B. C. Levy and A. J. Krener, "Dynamics and kinematics of reciprocal diffusions," *Journal of Mathematical Physics*, vol. 34, no. 5, pp. 1846–1875, 1993.

[35] B. Jamison, "Reciprocal processes," *Probability Theory and Related Fields*, vol. 30, no. 1, pp. 65–86, 1974.

[36] R. Kindermann and J. L. Snell, *Markov random fields and their applications*. Providence, R.I. : American Mathematical Society, 1980.

[37] Y. Weiss, "Correctness of local probability propagation in graphical models with loops," *Neural Comput.*, vol. 12, pp. 1–41, Jan. 2000.

[38] P. Clifford, "Markov random fields in statistics," *Disorder in physical systems: A volume in honour of John M. Hammersley*, pp. 19–32, 1990.

[39] F. P. Carli, "Modeling and estimation of discrete-time reciprocal processes via probabilistic graphical models," *arXiv preprint arXiv:1603.04419*, 2016.

[40] R. J. Baxter, *Exactly solved models in statistical mechanics*. Academic Press, 1982.

[41] I. M. Elfadel, *From random fields to networks*. PhD thesis, Massachusetts Institute of Technology, Dept. of Mechanical Engineering, 1993.

[42] S. C. Tatikonda and M. I. Jordan, "Loopy belief propagation and Gibbs measures," in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, (San Francisco, CA, USA), pp. 493–500, Morgan Kaufmann Publishers Inc., 2002.

[43] S. L. Lauritzen, *Graphical models*. Clarendon Press, 1996.

[44] F. P. Carli, "On the geometry of message passing algorithms for gaussian reciprocal processes," in *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pp. 4572–4577, IEEE, 2016.

[45] J. M. Mooij and H. J. Kappen, "Sufficient conditions for convergence of loopy belief propagation," in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, (Arlington, Virginia, United States), pp. 396–403, AUAI Press, 2005.

[46] A. T. Ihler, J. Iii, and A. S. Willsky, "Loopy belief propagation: Convergence and effects of message errors," in *Journal of Machine Learning Research*, pp. 905–936, 2005.