

Bayesian inference as iterated random functions with applications to sequential inference in graphical models*

Arash A. Amini XuanLong Nguyen

August 7, 2018

Abstract

We propose a general formalism of iterated random functions with semigroup property, under which exact and approximate Bayesian posterior updates can be viewed as specific instances. A convergence theory for iterated random functions is presented. As an application of the general theory we analyze convergence behaviors of exact and approximate message-passing algorithms that arise in a sequential change point detection problem formulated via a latent variable directed graphical model. The sequential inference algorithm and its supporting theory are illustrated by simulated examples.

1 Introduction

The sequential posterior updates play a central role in many Bayesian inference procedures. As an example, in Bayesian inference one is interested in the posterior probability of variables of interest given the data observed sequentially up to a given time point. As a more specific example which provides the motivation for this work, in a sequential change point detection problem [1], the key quantity is the posterior probability that a change has occurred given the data observed up to present time. When the underlying probability model is complex, e.g., a large-scale graphical model, the calculation of such quantities in a fast and online manner is a formidable challenge. In such situations approximate inference methods are required – for graphical models, message-passing variational inference algorithms present a viable option [2, 3].

In this paper we propose to treat Bayesian inference in a complex model as a specific instance of an abstract system of iterated random functions (IRF), a concept that originally arises in the study of Markov chains and stochastic systems [4]. The key technical property of the proposed IRF formalism that enables the connection to Bayesian inference under conditionally independent sampling is the *semigroup* property, which shall be defined shortly in the sequel. It turns out that most exact and approximate Bayesian inference algorithms may be viewed as specific instances of an IRF system. The goal of this paper is to present a general convergence theory for the IRF with semigroup property. The theory is then applied to the analysis of exact and approximate message-passing inference algorithms, which arise in the context of distributed sequential change point problems using latent variable and directed graphical model as the underlying modeling framework.

We wish to note a growing literature on message-passing and sequential inference based on graphical modeling [5, 6, 7, 8]. On the other hand, convergence and error analysis of

*Part of this work is presented at the NIPS 2013 conference.

message-passing algorithms in graphical models is quite rare and challenging, especially for approximate algorithms, and they are typically confined to the specific form of belief propagation (sum-product) algorithm [9, 10, 11]. To the best of our knowledge, there is no existing work on the analysis of message-passing inference algorithms for calculating conditional (posterior) probabilities for latent random variables present in a graphical model. While such an analysis is a byproduct of this work, the viewpoint we put forward here that equates Bayesian posterior updates to a system of iterated random functions with semigroup property seems to be new and may be of general interest.

The paper is organized as follows. In Sections 2–3, we introduce the general IRF system and provide our main result on its convergence. The proof is deferred to Section 5. As an example of the application of the result, we will provide a convergence analysis for an approximate sequential inference algorithm for the problem of multiple change point detection using graphical models. The problem setup and the results are discussed in Section 4. An auxiliary result needed for the change point application is proved in Section 6 with some of the more technical aspects left to the appendices.

2 Bayesian posterior updates as iterated random functions

In this paper we shall restrict ourselves to multivariate distributions of binary random variables. To describe the general iteration, let $\mathcal{P}_d := \mathcal{P}(\{0, 1\}^d)$ be the space of probability measures on $\{0, 1\}^d$. The iteration under consideration recursively produces a random sequence of elements of \mathcal{P}_d , starting from some initial value. We think of \mathcal{P}_d as a subset of \mathbb{R}^{2^d} equipped with the ℓ_1 norm (that is, the total variation norm for discrete probability measures). To simplify, let $m := 2^d$, and for $x \in \mathcal{P}_d$, index its coordinates as $x = (x^0, \dots, x^{m-1})$. For $\boldsymbol{\theta} \in \mathbb{R}_+^m$, consider the function $q_{\boldsymbol{\theta}} : \mathcal{P}_d \rightarrow \mathcal{P}_d$, defined by

$$q_{\boldsymbol{\theta}}(x) := \frac{x \odot \boldsymbol{\theta}}{x^T \boldsymbol{\theta}} \quad (1)$$

where $x^T \boldsymbol{\theta} = \sum_i x^i \theta^i$ is the usual inner product on \mathbb{R}^m and $x \odot \boldsymbol{\theta}$ is pointwise multiplication with coordinates $[x \odot \boldsymbol{\theta}]^i := x^i \theta^i$, for $i = 0, 1, \dots, m-1$. This function models the prior-to-posterior update according to the Bayes rule. One can think of $\boldsymbol{\theta}$ as the likelihood and x as the prior distribution (or the posterior in the previous stage) and $q_{\boldsymbol{\theta}}(x)$ as the (new) posterior based on the two. The division by $x^T \boldsymbol{\theta}$ can be thought of as the division by the marginal to make a valid probability vector. (See Example 1 below.)

We consider the following general iteration

$$\begin{aligned} Q_n(x) &= q_{\boldsymbol{\theta}_n}(T(Q_{n-1}(x))), \quad n \geq 1, \\ Q_0(x) &= x, \end{aligned} \quad (2)$$

for some deterministic operator $T : \mathcal{P}_d \rightarrow \mathcal{P}_d$ and an i.i.d. random sequence $\{\boldsymbol{\theta}_n\}_{n \geq 1} \subset \mathbb{R}_+^m$. By changing operator T , one obtains different iterative algorithms.

Our goal is to find sufficient conditions on T and $\{\boldsymbol{\theta}_n\}$ for the convergence of the iteration to an extreme point of \mathcal{P}_d , which without loss of generality is taken to be $e^{(0)} := (1, 0, 0, \dots, 0)$. Standard techniques for proving the convergence of iterated random functions are usually based on showing some averaged-sense contraction property for the iteration function [4, 12, 13, 14], which in our case is $q_{\boldsymbol{\theta}_n}(T(\cdot))$. See [15] for a recent survey. These techniques are not applicable

to our problem since q_{θ_n} is not in general Lipschitz, in any suitable sense, precluding $q_{\theta_n}(T(\cdot))$ from satisfying the aforementioned conditions.

Instead, the functions $\{q_{\theta_n}\}$ have another property which can be exploited to prove convergence; namely, they form a semi-group under pointwise multiplication,

$$q_{\theta \odot \theta'} = q_{\theta} \circ q_{\theta'}, \quad \theta, \theta' \in \mathbb{R}_+^m, \quad (3)$$

where \circ denotes the composition of functions. If T is the identity, this property allows us to write $Q_n(x) = q_{\odot_{i=1}^n \theta_i}(x)$ — this is nothing but the Bayesian posterior update equation, under conditionally independent sampling, while modifying T results in an approximate Bayesian inference procedure. Since after suitable normalization, $\odot_{i=1}^n \theta_i$ concentrates around a deterministic quantity, by the i.i.d. assumption on $\{\theta_i\}$, this representation helps in determining the limit of $\{Q_n(x)\}$. The main result of this paper, summarized in Theorem 1, is that the same conclusions can be extended to general Lipschitz maps T having the desired fixed point.

3 General convergence theory

Consider a sequence $\{\theta_n\}_{n \geq 1} \subset \mathbb{R}_+^m$ of i.i.d. random elements, where $m = 2^d$. Let $\theta_n = (\theta_n^0, \theta_n^1, \dots, \theta_n^{m-1})$ with $\theta_n^0 = 1$ for all n , and

$$\theta_n^* := \max_{i=1,2,\dots,m-1} \theta_n^i. \quad (4)$$

The normalization $\theta_n^0 = 1$ is convenient for showing convergence to $e^{(0)}$. This is without loss of generality, since q_{θ} is invariant to scaling of θ , that is $q_{\theta} = q_{\beta\theta}$ for any $\beta > 0$.

Assume the sequence $\{\log \theta_n^*\}$ to be i.i.d. sub-Gaussian with mean $\leq -I_* < 0$ and sub-Gaussian norm $\leq \sigma_* \in (0, \infty)$. The sub-Gaussian norm can be taken to be the ψ_2 Orlicz norm (cf. [16, Section 2.2]), which we denote by $\|\cdot\|_{\psi_2}$. By definition $\|Y\|_{\psi_2} := \inf\{C > 0 : \mathbb{E}\psi_2(|Y|/C) \leq 1\}$ where $\psi_2(x) := e^{x^2} - 1$.

Let $\|\cdot\|$ denote the ℓ_1 norm on \mathbb{R}^m . Consider the sequence $\{Q_n(x)\}_{n \geq 0}$ defined in (2) based on $\{\theta_n\}$ as above, an initial point $x = (x^0, \dots, x^{m-1}) \in \mathcal{P}_d$ and a Lipschitz map $T : \mathcal{P}_d \rightarrow \mathcal{P}_d$. Let Lip_T denote the Lipschitz constant of T , that is $\text{Lip}_T := \sup_{x \neq y} \|T(x) - T(y)\| / \|x - y\|$.

Our main result regarding iteration (2) is the following.

Theorem 1. *Assume that $L := \text{Lip}_T \leq 1$ and that $e^{(0)}$ is a fixed point of T . Then, for all $n \geq 0$, and $\varepsilon > 0$,*

$$\|Q_n(x) - e^{(0)}\| \leq 2 \frac{1 - x^0}{x^0} (L e^{-I_* + \varepsilon})^n \quad (5)$$

with probability at least $1 - \exp(-cn\varepsilon^2/\sigma_*^2)$, for some absolute constant $c > 0$.

The proof of Theorem 1 is outlined in Section 5. Our main application of the theorem will be to the study of convergence of stopping rules for a distributed multiple change point problem endowed with latent variable graphical models. Before stating that problem, let us consider the classical (single) change point problem first, and show how the theorem can be applied to analyze the convergence of the optimal Bayes rule.

Example 1. In the classical Bayesian change point problem [1], one observes a sequence $\{X^1, X^2, X^3 \dots\}$ of independent data points whose distributions change at some random time λ . More precisely, given $\lambda = k$, X^1, X^2, \dots, X^{k-1} are distributed according to g , and X^{k+1}, X^{k+2}, \dots according to f . Here, f and g are densities with respect to some underlying measure. One also assumes a prior π on λ , usually taken to be geometric. The goal is to find a stopping rule τ which can predict λ based on the data points observed so far. It is well-known that a rule based on thresholding the posterior probability of λ is optimal (in a Neyman-Pearson sense). To be more specific, let $\mathbf{X}^n := (X^1, X^2, \dots, X^n)$ collect the data up to time n and let $\gamma^n[n] := \mathbb{P}(\lambda \leq n | \mathbf{X}^n)$ be the posterior probability of λ having occurred before (or at) time n . Then, the Shiriyayev rule

$$\tau := \inf\{n \in \mathbb{N} : \gamma^n[n] \geq 1 - \alpha\} \quad (6)$$

is known to asymptotically have the least expected delay, among all stopping rules with false alarm probability bounded by α .

Theorem 1 provides a way to quantify how fast the posterior $\gamma^n[n]$ approaches 1, once the change point has occurred, hence providing an estimate of the detection delay, even for finite number of samples. We should note that our approach here is somewhat independent of the classical techniques normally used for analyzing stopping rule (6). To cast the problem in the general framework of (2), let us introduce the binary variable $Z^n := 1\{\lambda \leq n\}$, where $1\{\cdot\}$ denotes the indicator of an event. Let Q_n be the (random) distribution of Z^n given \mathbf{X}^n , in other words,

$$Q_n := (\mathbb{P}(Z^n = 1 | \mathbf{X}^n), \mathbb{P}(Z^n = 0 | \mathbf{X}^n)).$$

Since $\gamma^n[n] = \mathbb{P}(Z = 1 | \mathbf{X}^n)$, convergence of $\gamma^n[n]$ to 1 is equivalent to the convergence of Q_n to $e^{(0)} = (1, 0)$. We have

$$P(Z^n | \mathbf{X}^n) \propto_{Z^n} P(Z^n, X^n | \mathbf{X}^{n-1}) = P(X^n | Z^n) P(Z^n | \mathbf{X}^{n-1}). \quad (7)$$

Note that $P(X^n | Z^n = 1) = f(X^n)$ and $P(X^n | Z^n = 0) = g(X^n)$. Let $\theta_n := (1, \frac{g(X^n)}{f(X^n)})$ and

$$\mathcal{R}_{n-1} := (\mathbb{P}(Z^n = 1 | \mathbf{X}^{n-1}), \mathbb{P}(Z^n = 0 | \mathbf{X}^{n-1})).$$

Then, (7) implies that Q_n can be obtained by pointwise multiplication of \mathcal{R}_{n-1} by $f(X^n)\theta_n$ and normalization to make a probability vector. Alternatively, we can multiply by θ_n , since the procedure is scale-invariant, that is, $Q_n = g\theta_n(\mathcal{R}_{n-1})$ using definition (1). It remains to express \mathcal{R}_{n-1} in terms of Q_{n-1} . This can be done by using the Bayes rule and the fact that $P(\mathbf{X}^{n-1} | \lambda = k)$ is the same for $k \in \{n, n+1, \dots\}$. In particular, after some algebra (see Appendix A), one arrives at

$$\gamma^{n-1}[n] = \frac{\pi(n)}{\pi[n-1]^c} + \frac{\pi[n]^c}{\pi[n-1]^c} \gamma^{n-1}[n-1], \quad (8)$$

where $\gamma^k[n] := \mathbb{P}(\lambda \leq n | \mathbf{X}^k)$, $\pi(n)$ is the prior on λ evaluated at time n , and $\pi[k]^c := \sum_{i=k+1}^{\infty} \pi(i)$. For the geometric prior with parameter $\rho \in [0, 1]$, we have $\pi(n) := (1 - \rho)^{n-1} \rho$ and $\pi[k]^c = \rho^k$. The above recursion then simplifies to $\gamma^{n-1}[n] = \rho + (1 - \rho)\gamma^{n-1}[n-1]$. Expressing in terms of \mathcal{R}_{n-1} and Q_{n-1} , the recursion reads

$$\mathcal{R}_{n-1} = T(Q_{n-1}), \quad \text{where } T\left(\begin{pmatrix} x_1 \\ x_0 \end{pmatrix}\right) = \rho \begin{pmatrix} 1 \\ 0 \end{pmatrix} + (1 - \rho) \begin{pmatrix} x_1 \\ x_0 \end{pmatrix}.$$

In other words, $T(x) = \rho \mathbf{e}^{(0)} + (1 - \rho)x$ for $x \in \mathcal{P}_2$.

Thus, we have shown that an iterative algorithm for computing $\gamma^n[n]$ (hence determining rule (6)), can be expressed in the form of (2) for appropriate choices of $\{\boldsymbol{\theta}_n\}$ and operator T . Note that T in this case is Lipschitz with constant $1 - \rho$ which is always guaranteed to be ≤ 1 .

We can now use Theorem 1 to analyze the convergence of $\gamma^n[n]$. Let us condition on $\lambda = k+1$, that is, we assume that the change point has occurred at time $k+1$. Then, the sequence $\{X^n\}_{n \geq k+1}$ is distributed according to f , and we have $\mathbb{E}\boldsymbol{\theta}_n^* = \int f \log \frac{g}{f} = -I$, where I is the KL divergence between densities f and g . Noting that $\|Q_n - \mathbf{e}^{(0)}\| = 2(1 - \gamma^n[n])$, we immediately obtain the following corollary.

Corollary 1. *Consider Example 1 and assume that $\log(g(X)/f(X))$, where $X \sim f$, is sub-Gaussian with sub-Gaussian norm $\leq \sigma$. Let $I := \int f \log \frac{f}{g}$. Then, conditioned on $\lambda = k+1$, we have for $n \geq 1$,*

$$|\gamma^{n+k}[n+k] - 1| \leq [(1 - \rho)e^{-I+\varepsilon}]^n \left(\frac{1}{\gamma^k[k]} - 1 \right)$$

with probability at least $1 - \exp(-cn\varepsilon^2/\sigma^2)$.

4 Multiple change point problem via latent variable graphical models

We now turn to our main application for Theorem 1, in the context of a multiple change point problem. In [17], graphical model formalism is used to extend the classical change point problem (cf. Example 1) to cases where multiple distributed latent change points are present. Throughout this section, we will use this setup which we now briefly sketch.

One starts with a network $G = (V, E)$ of d sensors or nodes, each associated with a change point λ_j . Each node j observes a private sequence of measurements $\mathbf{X}_j = (X_j^1, X_j^2, \dots)$ which undergoes a change in distribution at time λ_j , that is,

$$X_j^1, X_j^2, \dots, X_j^{k-1} \mid \lambda_j = k \stackrel{iid}{\sim} g_j, \quad X_j^k, X_j^{k+1}, \dots \mid \lambda_j = k \stackrel{iid}{\sim} f_j,$$

for densities g_j and f_j (w.r.t. some underlying measure). Each connected pair of nodes share an additional sequence of measurements. For example, if nodes s_1 and s_2 are connected, that is, $e = (s_1, s_2) \in E$, then they both observe $\mathbf{X}_e = (X_e^1, X_e^2, \dots)$. The shared sequence undergoes a change in distribution at some point depending on λ_{s_1} and λ_{s_2} . More specifically, it is assumed that the earlier of the two change points causes a change in the shared sequence, that is, the distribution of \mathbf{X}_e conditioned on $(\lambda_{s_1}, \lambda_{s_2})$ only depends on $\lambda_e := \lambda_{s_1} \wedge \lambda_{s_2}$, the minimum of the two, i.e.,

$$X_e^1, X_e^2, \dots, X_e^k \mid \lambda_e = k \stackrel{iid}{\sim} g_e, \quad X_e^{k+1}, X_e^{k+2}, \dots \mid \lambda_e = k \stackrel{iid}{\sim} f_e.$$

Letting $\lambda_* := \{\lambda_j\}_{j \in V}$ and $\mathbf{X}_*^n = \{\mathbf{X}_j^n, \mathbf{X}_e^n\}_{j \in V, e \in E}$, we can write the joint density of all random variables as

$$P(\lambda_*, \mathbf{X}_*^n) = \prod_{j \in V} \pi_j(\lambda_j) \prod_{j \in V} P(\mathbf{X}_j^n \mid \lambda_j) \prod_{e \in E} P(\mathbf{X}_e^n \mid \lambda_{s_1}, \lambda_{s_2}). \quad (9)$$

where π_j is the prior on λ_j , which we assume to be geometric with parameter ρ_j . Network G induces a graphical model [2] which encodes the factorization (9) of the joint density. (cf. Fig. 1)

Suppose now that each node j wants to detect its change point λ_j , with minimum expected delay, while maintaining a false alarm probability at most α . Inspired by the classical change point problem, one is interested in computing the posterior probability that the change point has occurred up to now, that is,

$$\gamma_j^n[n] := \mathbb{P}(\lambda_j \leq n \mid \mathbf{X}_*^n). \quad (10)$$

The difference with the classical setting is the conditioning is done on all the data in the network (up to time n). It is easy to verify that the natural stopping rule

$$\tau_j = \inf\{n \in \mathbb{N} : \gamma_j^n[n] \geq 1 - \alpha\}$$

satisfy the false alarm constraint. It has also been shown that this rule is asymptotically optimal in terms of expected detection delay. Moreover, an algorithm based on the well-known sum-product [2] has been proposed, which allows the nodes to compute their posterior probabilities 10 by message-passing. The algorithm is exact when G is a tree, and scales linearly in the number of nodes. More precisely, at time n , the computational complexity is $O(nd)$. The drawback is the linear dependence on n , which makes the algorithm practically infeasible if the change points model rare events (where n could grow large before detecting the change.)

In the next section, we propose an approximate message passing algorithm which has computational complexity $O(d)$, at each time step. This circumvents the drawback of the exact algorithm and allows for indefinite run times. We then show how the theory developed in Section 3 can be used to provide convergence guarantees for this approximate algorithm, as well as the exact one.

4.1 Fast approximate message-passing (MP)

We now turn to an approximate message-passing algorithm which, at each time step, has computational complexity $O(d)$. The derivation is similar to that used for the iterative algorithm in Example 1. Let us define binary variables

$$Z_j^n = 1\{\lambda_j \leq n\}, \quad Z_*^n = (Z_1^n, \dots, Z_d^n). \quad (11)$$

The idea is to compute $P(Z_*^n \mid \mathbf{X}_*^n)$ recursively based on $P(Z_*^{n-1} \mid \mathbf{X}_*^{n-1})$. By Bayes rule,

$$\begin{aligned} P(Z_*^n \mid \mathbf{X}_*^n) &\propto_{Z_*^n} P(Z_*^n, X_*^n \mid \mathbf{X}_*^{n-1}) = P(X_*^n \mid Z_*^n) P(Z_*^n \mid \mathbf{X}_*^{n-1}) \\ &= \left[\prod_{j \in V} P(X_j^n \mid Z_j^n) \prod_{\{i,j\} \in E} P(X_{ij}^n \mid Z_i^n, Z_j^n) \right] P(Z_*^n \mid \mathbf{X}_*^{n-1}), \end{aligned} \quad (12)$$

where we have used the fact that given Z_*^n, X_*^n is independent of \mathbf{X}_*^{n-1} . To simplify notation, let us extend the edge set to $\tilde{E} := E \cup \{\{j\} : j \in V\}$. This allows us to treat the private data of node j , i.e., \mathbf{X}_j , as shared data of a self-loop in the extended graph (V, \tilde{E}) . Let $u_e(z; \xi) := [g_e(\xi)]^{1-z} [f_e(\xi)]^z$ for $e \in \tilde{E}, z \in \{0, 1\}$. Then, for $i \neq j$,

$$P(X_j^n \mid Z_j^n) = u_j(Z_j^n; X_j^n), \quad P(X_{ij}^n \mid Z_i^n, Z_j^n) = u_{ij}(Z_i^n \vee Z_j^n; X_{ij}^n). \quad (13)$$

Algorithm 1 Message passing algorithm to compute approximate posteriors $\tilde{\gamma}_j^n[n]$ and $\tilde{\gamma}_{ij}^n[n]$

Initialize $\tilde{\gamma}_j^0[0] = 0$ for $j \in V$.

for all time $n \geq 1$ **do**

1. Compute $\tilde{\gamma}_j^{n-1}[n]$ based on $\tilde{\gamma}_j^{n-1}[n-1]$ using equation (15), for all $j \in V$.
2. Form the following joint distribution for $Z_*^n = (Z_1^n, \dots, Z_d^n)$,

$$\tilde{P}(Z_*^n | \mathbf{X}_*^n) = C \prod_{j \in V} u_j(Z_j^n; X_j^n) \prod_{\{i,j\} \in E} u_{ij}(Z_i^n \vee Z_j^n; X_{ij}^n) \prod_{j \in V} \nu(Z_j^n; \tilde{\gamma}_j^{n-1}[n]) \quad (16)$$

where $u_e(z; \xi) := [g_e(\xi)]^{1-z} [f_e(\xi)]^z$ for $e \in \tilde{E}$, and $\nu(z; \beta) := \beta^z (1 - \beta)^{1-z}$. The normalizing constant C is left undetermined at this point.

3. Invoke a message-passing algorithm (sum-product) on the joint distribution (16) to obtain marginal distributions $\tilde{P}(Z_j^n | \mathbf{X}_*^n)$, $j \in V$ and set $\tilde{\gamma}_j^n[n] = \tilde{P}(Z_j^n = 1 | \mathbf{X}_*^n)$.

(As a by-product of the message-passing, one also gets pair marginals $\tilde{P}(Z_i^n, Z_j^n | \mathbf{X}_*^n)$ and $\tilde{\gamma}_{ij}^n[n] := \tilde{P}(Z_i^n = 1 \text{ or } Z_j^n = 1 | \mathbf{X}_*^n)$ which are useful for constructing stopping rules for minimum of the two change points; see [17].)

end for

It remains to express $P(Z_*^n | \mathbf{X}_*^{n-1})$ in terms of $P(Z_*^{n-1} | \mathbf{X}_*^{n-1})$. It is possible to do this, exactly, at a cost of $O(2^{|V|})$. For brevity, we omit the exact expression. (See Lemma 1 for some details.) We term the algorithm that employs the exact relationship, the “exact algorithm”.

In practice, however, the exponential complexity makes the exact recursion of little use for large networks. To obtain a fast algorithm (i.e., $O(\text{poly}(d))$), we instead take a mean-field type approximation:

$$P(Z_*^n | \mathbf{X}_*^{n-1}) \approx \prod_{j \in V} P(Z_j^n | \mathbf{X}_*^{n-1}) = \prod_{j \in V} \nu(Z_j^n; \tilde{\gamma}_j^{n-1}[n]), \quad (14)$$

where $\nu(z; \beta) := \beta^z (1 - \beta)^{1-z}$. That is, we approximate a multivariate distribution by the product of its marginals. By an argument similar to that used to derive (8), we can obtain a recursion for the marginals,

$$\tilde{\gamma}_j^{n-1}[n] = \frac{\pi_j(n)}{\pi_j[n-1]^c} + \frac{\pi_j[n]^c}{\pi_j[n-1]^c} \tilde{\gamma}_j^{n-1}[n-1], \quad (15)$$

where we have used the notation introduced earlier in (8). Thus, at time n , the RHS of (14) is known based on values computed at time $n-1$ (with initial value $\tilde{\gamma}_j^0[0] = 0, j \in V$). Inserting this RHS into (12) in place of $P(Z_*^n | \mathbf{X}_*^{n-1})$, we obtain a graphical model in variables Z_*^n (instead of λ_*) which has the same form as (9) with $\nu(Z_j^n; \tilde{\gamma}_j^{n-1}[n])$ playing the role of the prior $\pi(\lambda_j)$.

In order to obtain the marginals $\tilde{\gamma}_j^n[n] = P(Z_j^n = 1 | \mathbf{X}_*^n)$ with respect to the approximate version of the joint distribution $P(Z_*^n, X_*^n | \mathbf{X}_*^{n-1})$, we need to marginalize out the latent variables Z_j^n 's, for which a standard sum-product algorithm can be applied (see [2, 3, 17]). The message update equations are similar to those in [17]; the difference is that the messages are now binary and do not grow in size with n . The approximate algorithm is summarized in Algorithm 1.

4.2 Convergence of MP algorithms

We now turn to the analysis of the approximate algorithm introduced in Section 4.1. In particular, we will look at the evolution of $\{\tilde{P}(Z_*^n|\mathbf{X}_*^n)\}_{n \in \mathbb{N}}$ as a sequence of probability distribution on $\{0, 1\}^d$. Here, \tilde{P} signifies that this sequence is an approximation. In order to make a meaningful comparison, we also look at the algorithm which computes the exact sequence $\{P(Z_*^n|\mathbf{X}_*^n)\}_{n \in \mathbb{N}}$, recursively. As mentioned before, this we will call the “exact algorithm”, the details of which are not of concern to us at this point (cf. Proposition 1 for these details.)

Recall that we take $\tilde{P}(Z_*^n|\mathbf{X}_*^n)$ and $P(Z_*^n|\mathbf{X}_*^n)$, as distributions for Z_*^n , to be elements of $\mathcal{P}_d \subset \mathbb{R}^m$. To make this correspondence formal and the notation simplified, we use the symbol \equiv as follows

$$\tilde{y}_n := \tilde{P}(Z_*^n|\mathbf{X}_*^n), \quad y_n := P(Z_*^n|\mathbf{X}_*^n) \quad (17)$$

where now $\tilde{y}_n, y_n \in \mathcal{P}_d$. Note that \tilde{y}_n and y_n are random elements of \mathcal{P}_d , due the randomness of \mathbf{X}_*^n . We have the following description.

Proposition 1. *The exact and approximate sequences, $\{y_n\}$ and $\{\tilde{y}_n\}$, follow general iteration (2) with the same random sequence $\{\theta_n\}$, but with different deterministic operators T , denoted respectively with T_{ex} and T_{ap} . T_{ex} is linear and given by a Markov transition kernel. T_{ap} is a polynomial map of degree d . Both maps are Lipschitz and we have*

$$\text{Lip}_{T_{\text{ex}}} \leq L_\rho := \left(1 - \prod_{j=1}^d \rho_j\right), \quad \text{Lip}_{T_{\text{ap}}} \leq K_\rho := \sum_{j=1}^d (1 - \rho_j). \quad (18)$$

Detailed descriptions of the sequence $\{\theta_n\}$ and the operators T_{ex} and T_{ap} , along with the proof of Proposition 1, are given in Section 6. As suggested by Theorem 1, a key assumption for the convergence of the approximate algorithm will be $K_\rho \leq 1$. In contrast, we always have $L_\rho \leq 1$.

Recall that $\{\lambda_j\}$ are the change points and their priors are geometric with parameters $\{\rho_j\}$. We analyze the algorithms, once all the change points have happened. More precisely, we condition on

$$\mathbb{M}_{n_0} := \{\max_j \lambda_j \leq n_0\}$$

for some $n_0 \in \mathbb{N}$. Then, one expects the (joint) posterior of Z_*^n to contract to the point $Z_j^\infty = 1$, for all $j \in V$. In the vectorial notation, we expect both $\{\tilde{y}_n\}$ and $\{y_n\}$ to converge to $\mathbf{e}^{(0)}$. Theorem 2 below quantifies this convergence in ℓ_1 norm (equivalently, total variation for measures).

Recall pre-change and post-change densities g_e and f_e , and let I_e denote their KL divergence, that is, $I_e := \int f_e \log(f_e/g_e)$. We will assume that

$$Y_e := \log(g_e(X)/f_e(X)) \quad \text{with} \quad X \sim f_e \quad (19)$$

is sub-Gaussian, for all $e \in \tilde{E}$, where \tilde{E} is extended edge notation introduced in Section 4.1. The choice $X \sim f_e$ is in accordance with conditioning on \mathbb{M}_{n_0} . Note that $\mathbb{E}Y_e = -I_e < 0$. We define

$$\sigma_{\max} := \max_{e \in \tilde{E}} \|Y_e\|_{\psi_2}, \quad I_{\min} := \min_{e \in \tilde{E}} I_e, \quad I_*(\kappa) := I_{\min} - \kappa \sigma_{\max} \sqrt{\log D..}$$

where $D := |V| + |E|$. The following is our main result regarding sequences (17) produced by the exact and approximate algorithms.

Theorem 2. *There exists an absolute constant $\kappa > 0$, such that if $I_*(\kappa) > 0$, the exact algorithm converges at least geometrically w.h.p., that is, for all $n \geq 1$,*

$$\|y_{n+n_0} - \mathbf{e}^{(0)}\| \leq 2 \frac{1 - y_{n_0}}{y_{n_0}} (L_\rho e^{-I_*(\kappa) + \varepsilon})^n \quad (20)$$

with probability at least $1 - \exp[-cn\varepsilon^2/(\sigma_{\max}^2 D^2 \log D)]$, conditioned on \mathbb{M}_{n_0} . If in addition, $K_\rho \leq 1$, the approximate algorithm also converges at least geometrically w.h.p., i.e., for all $n \geq 1$,

$$\|\tilde{y}_{n+n_0} - \mathbf{e}^{(0)}\| \leq 2 \frac{1 - \tilde{y}_{n_0}}{\tilde{y}_{n_0}} (K_\rho e^{-I_*(\kappa) + \varepsilon})^n \quad (21)$$

with the same (conditional) probability as the exact algorithm.

Proof. Proposition 1 and Theorem 1 provide all the ingredients for the proof. It remains to show that $\{\boldsymbol{\theta}_n\}_{n \geq n_0}$ as given in (41) satisfies the conditions of Theorem 1; namely, that $\{\log \boldsymbol{\theta}_n^*\}_{n \geq n_0}$ is i.i.d. sub-Gaussian. We work conditioned on the event $\mathbb{M}_{n_0} := \{\max_{j \in V} \lambda_j \leq n_0\}$, that is, we look at what happens to the iterations past all the change-points. Throughout this section, \mathbb{E} denotes conditional expectation given \mathbb{M}_{n_0} . Then, the fact that the sequence is i.i.d. follows immediately from the definition. Let us now focus on showing that $\log \boldsymbol{\theta}_{n_0}^*$ is sub-Gaussian with negative expectation. We can write

$$\log(\boldsymbol{\theta}_{n_0})_\ell = \sum_{e \in \tilde{E}} \nu_e^\ell Y_e$$

where \tilde{E} is the extended edge notation introduced in Section 4.1, $Y_e := \log[g_e(X_e^{n_0})/f_e(X_e^{n_0})]$, and $\nu_e^\ell \in \{0, 1\}$. Note that ν_e^ℓ is equal to either $1 - b_j(\ell)$ or $1 - b_i(\ell) \vee b_j(\ell)$. For $\ell \neq m - 1$, at least one of $\nu_e^\ell, e \in \tilde{E}$ is non-zero. From definition (4) and superscript to subscript index translation of (37), we have

$$\log \boldsymbol{\theta}_{n_0}^* = \max_{i=1,2,\dots,m-1} \log \boldsymbol{\theta}_{n_0}^i = \max_{\ell=0,1,\dots,m-2} \log(\boldsymbol{\theta}_{n_0})_\ell.$$

Let $\mathcal{V} \subset \{0, 1\}^{|\tilde{E}|}$ denote the set carved by $(\nu_e^\ell)_{e \in \tilde{E}}$ as ℓ takes the values $0, 1, \dots, m - 2$. We note that the all-zero vector does not belong to \mathcal{V} . Let $\nu = (\nu_e)_{e \in \tilde{E}}$ denote a generic point of $\{0, 1\}^{|\tilde{E}|}$. Then, we have

$$\log \boldsymbol{\theta}_{n_0}^* = \max_{\nu \in \mathcal{V}} \sum_{e \in \tilde{E}} \nu_e Y_e. \quad (22)$$

Note that $\mathbb{E} Y_e = \int f_e \log(g_e/f_e) = -I_e \leq -I_{\min}$. We can write

$$\begin{aligned} \mathbb{E} \log \boldsymbol{\theta}_{n_0}^* &\leq \mathbb{E} \left[\max_{\nu \in \mathcal{V}} \sum_{e \in \tilde{E}} \nu_e (Y_e - \mathbb{E} Y_e) \right] + \max_{\nu \in \mathcal{V}} \sum_{e \in \tilde{E}} \nu_e (\mathbb{E} Y_e) \\ &\leq \mathbb{E} \left[\max_{\nu \in \mathcal{V}} \sum_{e \in \tilde{E}} \nu_e |Y_e - \mathbb{E} Y_e| \right] + \max_{\nu \in \mathcal{V}} \sum_{e \in \tilde{E}} \nu_e (-I_{\min}). \end{aligned}$$

The second term above is equal to $-I_{\min}(\min_{\nu \in \mathcal{V}} \sum_{e \in \tilde{E}} \nu_e) = -I_{\min}$, due to the fact that at least one element of every $\nu \in \mathcal{V}$ is nonzero. Then, we have

$$\begin{aligned} \mathbb{E} \log \boldsymbol{\theta}_{n_0}^* &\leq \mathbb{E} \max_{\nu \in \mathcal{V}} \left[\left(\sum_{e \in \tilde{E}} \nu_e \right) \max_{e \in \tilde{E}} |Y_e - \mathbb{E}Y_e| \right] - I_{\min} \\ &\leq |\tilde{E}| \mathbb{E} \left(\max_{e \in \tilde{E}} |Y_e - \mathbb{E}Y_e| \right) - I_{\min} \end{aligned}$$

We know that $\|Y_e - \mathbb{E}Y_e\|_{\psi_2} \leq c\|Y_e\|_{\psi_2} \leq c\sigma_{\max}$, for some numerical constant $c > 0$. In addition by majorant characteristic of ψ_2 space (cf. [16, 18]),

$$\begin{aligned} \mathbb{E} \max_{e \in \tilde{E}} |Y_e - \mathbb{E}Y_e| &\leq C \sqrt{\log(1 + |\tilde{E}|)} \max_{e \in \tilde{E}} \|Y_e - \mathbb{E}Y_e\|_{\psi_2} \\ &\leq C' \sqrt{\log(1 + |\tilde{E}|)} \sigma_{\max}. \end{aligned}$$

Thus assuming $|\tilde{E}| \geq 2$, we have

$$\mathbb{E} \log \boldsymbol{\theta}_{n_0}^* \leq \kappa \sigma_{\max} \sqrt{\log |\tilde{E}|} - I_{\min} =: -I_*$$

for some absolute constant $\kappa > 0$, which is the desired bound on the expectation of $\log \boldsymbol{\theta}_{n_0}^*$.

To verify that $\log \boldsymbol{\theta}_{n_0}^*$ is sub-Gaussian, we use $|\max a_i| \leq \max |a_i|$ to write

$$|\log \boldsymbol{\theta}_{n_0}^*| \leq \max_{\nu \in \mathcal{V}} \sum_{e \in \tilde{E}} \nu_e |Y_e| \leq |\tilde{E}| \max_{e \in \tilde{E}} |Y_e|.$$

Since $\|\cdot\|_{\psi_2}$, as an Orlicz norm, is monotone (i.e., $|X| \leq |Y|$ implies $\|X\|_{\psi_2} \leq \|Y\|_{\psi_2}$ for any two random variables X and Y), we obtain

$$\begin{aligned} \|\log \boldsymbol{\theta}_{n_0}^*\|_{\psi_2} &\leq |\tilde{E}| \cdot \|\max_{e \in \tilde{E}} |Y_e|\|_{\psi_2} \\ &\leq C |\tilde{E}| \sqrt{\log |\tilde{E}|} \max_{e \in \tilde{E}} \|Y_e\|_{\psi_2} \leq C' \sigma_{\max} |\tilde{E}| \sqrt{\log |\tilde{E}|}, \end{aligned}$$

where the second inequality is again by the majorant character of ψ_2 . This completes the proof. \square

4.3 Simulation results

We present some simulation results to verify the effectiveness of the proposed approximation algorithm in estimating the posterior probabilities $\gamma_j^n[n]$. We consider a star graph on $d = 4$ nodes. This is the subgraph on nodes $\{1, 2, 3, 4\}$ in Fig. 1. Conditioned on the change points λ_* , all data sequences \mathbf{X}_* are assumed Gaussian with variance 1, pre-change mean 1 and post-change mean zero. All priors are geometric with $\rho_j = 0.1$. We note that higher values of ρ_j yield even faster convergence in the simulations, but we omit these figures due to space constraints. Fig. 1 illustrates typical examples of posterior paths $n \mapsto \gamma_j^n[n]$, for both the exact and approximate MP algorithms. One can observe that the approximate path often closely follows the exact one. In some cases, they might deviate for a while, but as suggested by Theorem 2, they approach one another quickly, once the change points have occurred.

From the theorem and triangle inequality, it follows that under $I_*(\kappa) > 0$ and $K_\rho \leq 1$, $\|y_n - \tilde{y}_n\|$ converges to zero, at least geometrically w.h.p. This gives some theoretical explanation for the good tracking behavior of approximate algorithm as observed in Fig. 1.

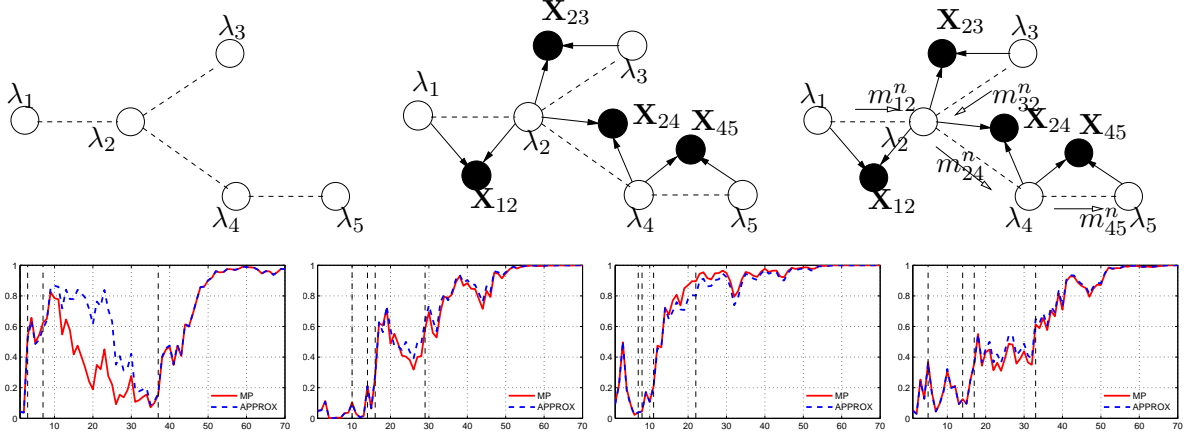


Figure 1: Top row illustrates a network (left), which induces a graphical model (middle). Right panel illustrates one stage of message-passing to compute posterior probabilities $\gamma_j^n[n]$. Bottom row illustrates typical examples of posterior paths, $n \mapsto \gamma_j^n[n]$, obtained by EXACT and approximate (APPROX) message passing, for the subgraph on nodes $\{1, 2, 3, 4\}$. The change points are designated with vertical dashed lines.

5 Proof of Theorem 1

For $x \in \mathbb{R}^m$ (including \mathcal{P}_d), we write $x = (x^0, \tilde{x})$ where $\tilde{x} = (x^1, \dots, x^{m-1})$. Recall that $\mathbf{e}^{(0)} = (1, 0, \dots, 0)$ and $\|x\| = \sum_{i=0}^{m-1} |x_i|$. For $x \in \mathcal{P}_d$, we have $1 - x^0 = \|\tilde{x}\|$, and

$$\|x - \mathbf{e}^{(0)}\| = \|(x^0 - 1, \tilde{x})\| = 1 - x^0 + \|\tilde{x}\| = 2(1 - x^0). \quad (23)$$

For $\boldsymbol{\theta} = (\boldsymbol{\theta}^0, \tilde{\boldsymbol{\theta}}) \in \mathbb{R}_+^m$, let

$$\boldsymbol{\theta}^* := \|\tilde{\boldsymbol{\theta}}\|_\infty = \max_{i=1, \dots, m-1} \boldsymbol{\theta}^i, \quad \boldsymbol{\theta}^\dagger := (\boldsymbol{\theta}^0, (\boldsymbol{\theta}^* L) \mathbf{1}_{m-1}) \in \mathbb{R}_+^m \quad (24)$$

where $\mathbf{1}_{m-1}$ is a vector in \mathbb{R}^{m-1} whose coordinates are all ones. We start by investigating how $\|q_{\boldsymbol{\theta}}(x) - \mathbf{e}^{(0)}\|$ varies as a function of $\|x - \mathbf{e}^{(0)}\|$.

Lemma 1. For $L \leq 1$, $\boldsymbol{\theta}^* > 0$, and $\boldsymbol{\theta}^0 = 1$,

$$N := \sup_{\substack{x, y \in \mathcal{P}_d, \\ \|x - \mathbf{e}^{(0)}\| \leq L \|y - \mathbf{e}^{(0)}\|}} \frac{\|q_{\boldsymbol{\theta}}(x) - \mathbf{e}^{(0)}\|}{\|q_{\boldsymbol{\theta}^\dagger}(y) - \mathbf{e}^{(0)}\|} = 1; \quad (25)$$

We prove Lemma 1 shortly in Section 5.1. Given the lemma, let us proceed to the proof of the theorem. Recall that $T : \mathcal{P}_d \rightarrow \mathcal{P}_d$ is an L -Lipschitz map, and that $\mathbf{e}^{(0)}$ is a fixed point of T , that is, $T(\mathbf{e}^{(0)}) = \mathbf{e}^{(0)}$. It follows that for any $x \in \mathcal{P}_d$, $\|T(x) - \mathbf{e}^{(0)}\| \leq L \|x - \mathbf{e}^{(0)}\|$. Applying Lemma 1, we get

$$\|q_{\boldsymbol{\theta}}(T(x)) - \mathbf{e}^{(0)}\| \leq \|q_{\boldsymbol{\theta}^\dagger}(x) - \mathbf{e}^{(0)}\| \quad (26)$$

for $\boldsymbol{\theta} \in \mathbb{R}_+^m$ with $\boldsymbol{\theta}^0 = 1$, and $x \in \mathcal{P}_d$. (This holds even if $\boldsymbol{\theta}^* = 0$ where both sides are zero.)

Recall the sequence $\{\boldsymbol{\theta}_n\}_{n \geq 1}$ used in defining functions $\{Q_n\}$ according to (2), and the assumption that $\boldsymbol{\theta}_n^0 = 1$, for all $n \geq 1$. Inequality (26) is key in allowing us to peel operator T , and

bring successive elements of $\{q_{\theta_n}\}$ together. Then, we can exploit the semi-group property (3) on adjacent elements of $\{q_{\theta_n}\}$.

To see this, for each θ_n , let θ_n^* and θ_n^\dagger be defined as in (24). Applying (26) with x replaced with $Q_{n-1}(x)$, and θ with θ_n , we can write

$$\begin{aligned} \|Q_n(x) - e^{(0)}\| &\leq \|q_{\theta_n^\dagger}(Q_{n-1}(x)) - e^{(0)}\| \quad (\text{by (26)}) \\ &= \|q_{\theta_n^\dagger}(q_{\theta_{n-1}}(T(Q_{n-2}(x)))) - e^{(0)}\| \\ &= \|q_{\theta_n^\dagger \circ \theta_{n-1}}(T(Q_{n-2}(x)))) - e^{(0)}\| \quad (\text{by semi-group property (3)}) \end{aligned}$$

We note that $(\theta_n^\dagger \circ \theta_{n-1})^* = L\theta_n^* \theta_{n-1}^*$ and

$$(\theta_n^\dagger \circ \theta_{n-1})^\dagger = (1, L(\theta_n^\dagger \circ \theta_{n-1})^* \mathbf{1}_{m-1}) = (1, L^2 \theta_n^* \theta_{n-1}^* \mathbf{1}_{m-1}).$$

Here, $*$ and \dagger act on a general vector in the sense of (24). Applying (26) once more, we get

$$\|Q_n(x) - e^{(0)}\| \leq \|q_{(1, L^2 \theta_n^* \theta_{n-1}^* \mathbf{1}_{m-1})}(Q_{n-2}(x)) - e^{(0)}\|.$$

The pattern is clear. Letting $\eta_n := L^n \prod_{k=1}^n \theta_k^*$, we obtain by induction

$$\|Q_n(x) - e^{(0)}\| \leq \|q_{(1, \eta_n \mathbf{1}_{m-1})}(Q_0(x)) - e^{(0)}\|. \quad (27)$$

Recall that $Q_0(x) := x$. Moreover,

$$\|q_{(1, \eta_n \mathbf{1}_{m-1})}(x) - e^{(0)}\| = 2(1 - [q_{(1, \eta_n \mathbf{1}_{m-1})}(x)]^0) = 2(1 - g_{\eta_n}(x^0)) \quad (28)$$

where the first inequality is by (23), and the second is easily verified by noting that all the elements of $(1, \eta_n \mathbf{1}_{m-1})$, except the first, are equal. Putting (27) and (28) together with the bound $1 - g_\theta(r) = \frac{\theta(1-r)}{r+\theta(1-r)} \leq \theta \frac{1-r}{r}$, which holds for $\theta > 0$ and $r \in (0, 1]$, we obtain $\|Q_n(x) - e^{(0)}\| \leq 2\eta_n \frac{1-x^0}{x^0}$. By sub-Gaussianity assumption on $\{\log \theta_k^*\}$, we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{k=1}^n \log \theta_k^* - \mathbb{E} \log \theta_1^* > \varepsilon\right) \leq \exp(-cn\varepsilon^2/\sigma_*^2), \quad (29)$$

for some absolute constant $c > 0$. (Recall that σ_* is an upper bound on the sub-Gaussian norm $\|\log \theta_1^*\|_{\psi_2}$.) On the complement of the event in 29, we have $\prod_{k=1}^n \theta_k^* \leq e^{n(-I_* + \varepsilon)}$, which completes the proof.

5.1 Proof of Lemma 1

We consider the simplest case first, namely $d = 2$. For $\theta \in \mathbb{R}_+$, let $g_\theta : [0, 1] \rightarrow [0, 1]$ be defined by

$$g_\theta(r) := \frac{r}{r + \theta(1-r)}. \quad (30)$$

This function completely describes q_θ when $d = 2$. More precisely, with $\theta = (1, \theta)$, one has $q_\theta(x) = (g_\theta(x^0), 1 - g_\theta(x^0))$. Note that $q_\theta(x)$ is close to e^0 iff $g_\theta(x^0)$ is close to 1. To simplify notation, let $\bar{r} := 1 - r$ for $r \in [0, 1]$. Similarly, let

$$\bar{g}_\theta(r) := 1 - g_\theta(r) = \frac{\theta \bar{r}}{1 - \bar{r} + \theta \bar{r}}. \quad (31)$$

The next lemma allows us to quantify how $|\bar{g}_\theta(r)|$ varies in terms of $|\bar{r}|$. Consider the following quantity

$$M_L(\theta, \gamma) := \sup \left\{ \frac{|\bar{g}_\theta(r)|}{|\bar{g}_\gamma(s)|} : \bar{r}, \bar{s} \in (0, 1], \bar{r} \leq L\bar{s} \right\}. \quad (32)$$

Lemma 2. *Assume that $L \leq 1$ and $\theta > 0$. Let $\varepsilon := 1 - \theta$ and $\gamma := 1 - \delta$. Then,*

$$M_L(\theta, \gamma) = \frac{\theta L}{|\gamma|} \max \left\{ 1, \left| \frac{1 - \delta}{1 - L\varepsilon} \right| \right\}. \quad (33)$$

In particular, for $\gamma = \theta L$, we have $M_L(\theta, \gamma) = 1$.

Proof. We can write

$$\begin{aligned} M_L(\theta, \gamma) &= \sup_{\bar{r}, \bar{s}} \left| \frac{\theta \bar{r}}{1 - \bar{r} + \theta \bar{r}} \frac{1 - \bar{s} + \gamma \bar{s}}{\gamma \bar{s}} \right| \\ &= \frac{\theta}{|\gamma|} \sup_{\bar{r}, \bar{s}} \left| \frac{\bar{r}}{\bar{s}} \cdot \frac{(\gamma - 1)\bar{s} + 1}{(\theta - 1)\bar{r} + 1} \right| \\ &= \frac{\theta}{|\gamma|} \sup_{\bar{r}, \bar{s}} \left| \frac{(\gamma - 1) + 1/\bar{s}}{(\theta - 1) + 1/\bar{r}} \right| \end{aligned}$$

Let $x = 1/\bar{r}$ and $z = \bar{r}/\bar{s}$. Then, the set $\{(\bar{r}, \bar{s}) : \bar{r}, \bar{s} \in (0, 1], \bar{r} \leq L\bar{s}\}$ corresponds to

$$\{(x, z) : x \geq 1, xz \geq 1, z \leq L\} = \{(x, z) : x \geq \frac{1}{L}, \frac{1}{x} \leq z \leq L\}$$

where in the second inequality, we used $L \leq 1$ and that $[\frac{1}{x}, L]$ is empty unless $x \geq \frac{1}{L}$. Letting $m(x, z) := (xz - \delta)/(x - \varepsilon)$, we obtain

$$M_L(\theta, \gamma) = \frac{\theta}{|\gamma|} \sup_{x \geq \frac{1}{L}, z \in [\frac{1}{x}, L]} |m(x, z)|$$

The function $m(x, z)$ is well-defined over the specified region (that is, finite-valued) since $\theta > 0$ implies $\varepsilon < 1$, hence $x - \varepsilon > 0$. For fixed $x \geq \frac{1}{L}$, the function $z \mapsto |m(x, z)|$ is convex, hence achieving its maximum over the convex set $[\frac{1}{x}, L]$, at one of the extreme points,

$$M_L(\theta, \gamma) = \frac{\theta}{|\gamma|} \sup_{x \geq \frac{1}{L}} \left[\max \left\{ |m(x, \frac{1}{x})|, |m(x, L)| \right\} \right]$$

Both $x \mapsto |m(x, \frac{1}{x})|$ and $x \mapsto |m(x, L)|$ are quasi-convex, hence their suprema over $[\frac{1}{L}, \infty)$ are obtained at one of the endpoints. Thus,

$$\begin{aligned} M_L(\theta, \gamma) &= \frac{\theta}{|\gamma|} \max \left\{ \sup_{x \geq \frac{1}{L}} \left| \frac{1 - \delta}{x - \varepsilon} \right|, \sup_{x \geq \frac{1}{L}} \left| \frac{xL - \delta}{x - \varepsilon} \right| \right\} \\ &= \frac{\theta}{|\gamma|} \max \left\{ \left| \frac{1 - \delta}{\frac{1}{L} - \varepsilon} \right|, 0, \left| \frac{L\frac{1}{L} - \delta}{\frac{1}{L} - \varepsilon} \right|, L \right\} \end{aligned}$$

which simplifies to (33).

For the special case, $\gamma = \theta L$, we first note that $L\theta/\gamma = g_{1/\theta}(L)$. Then, we have $M_L(\theta, \gamma) = \max\{1, g_{1/\theta}(L)\}$. Since $g_{1/\theta}(L) \in [0, 1]$, we get the desired result. \square

Let us now move to the case of general d . By (23), we have

$$N = \sup \left\{ \frac{1 - [q_{\boldsymbol{\theta}}(x)]^0}{1 - [q_{\boldsymbol{\theta}^\dagger}(y)]^0} : \bar{x}^0 \leq L\bar{y}^0, \|\tilde{x}\| = \bar{x}^0, \|\tilde{y}\| = \bar{y}^0 \right\}. \quad (34)$$

We are effectively optimizing over four variables x^0, y^0, \tilde{x} and \tilde{y} . Let us first optimize over \tilde{x} , fixing the other three. By definition (1), we have

$$\begin{aligned} \sup_{\tilde{x}: \|\tilde{x}\| = \bar{x}^0} \{1 - [q_{\boldsymbol{\theta}}(x)]^0\} &= \sup_{\tilde{x}: \|\tilde{x}\| = \bar{x}^0} \left\{ 1 - \frac{\boldsymbol{\theta}^0 x^0}{\boldsymbol{\theta}^0 x^0 + \boldsymbol{\theta}^T \tilde{x}} \right\} \\ &= 1 - \frac{\boldsymbol{\theta}^0 x^0}{\boldsymbol{\theta}^0 x^0 + \sup \{ \boldsymbol{\theta}^T \tilde{x} : \|\tilde{x}\| = \bar{x}^0 \}} = 1 - \frac{\boldsymbol{\theta}^0 x^0}{\boldsymbol{\theta}^0 x^0 + \|\tilde{\boldsymbol{\theta}}\|_\infty \bar{x}^0}, \end{aligned}$$

by the duality of ℓ_1 and ℓ_∞ norms. Recalling the definition (30), and using $\boldsymbol{\theta}^0 = 1$ and $\|\tilde{\boldsymbol{\theta}}\|_\infty = \boldsymbol{\theta}^*$, we have

$$\sup_{\tilde{x}: \|\tilde{x}\| = \bar{x}^0} \{1 - [q_{\boldsymbol{\theta}}(x)]^0\} = 1 - g_{\boldsymbol{\theta}^*}(x^0). \quad (35)$$

Next, we optimize over \tilde{y} . Let $\gamma^* := \boldsymbol{\theta}^* L$. We note for $\|\tilde{y}\| = \bar{y}^0$,

$$[q_{\boldsymbol{\theta}^\dagger}(y)]^0 = \frac{\boldsymbol{\theta}^0 y^0}{\boldsymbol{\theta}^0 y^0 + \gamma^* \mathbf{1}_{m-1}^T \tilde{y}} = \frac{\boldsymbol{\theta}^0 y^0}{\boldsymbol{\theta}^0 y^0 + \gamma^* \bar{y}^0} = g_{\gamma^*}(y^0)$$

where we have used $\mathbf{1}_{m-1}^T \tilde{y} = \|\tilde{y}\|$ and $\boldsymbol{\theta}^0 = 1$. In other words, we have shown

$$\sup_{\tilde{y}: \|\tilde{y}\| = \bar{y}^0} \{1 - [q_{\boldsymbol{\theta}^\dagger}(y)]^0\} = 1 - g_{\gamma^*}(y^0). \quad (36)$$

Substituting (35) and (36) in (34), and recalling the notation (31) and definition (32), we get

$$N = \sup \left\{ \frac{\bar{g}_{\boldsymbol{\theta}^*}(x^0)}{\bar{g}_{\gamma^*}(y^0)} : \bar{x}^0 \leq L\bar{y}^0 \right\} = M_L(\boldsymbol{\theta}^*, \gamma^*).$$

Applying Lemma 2 in the special case $\gamma^* = \boldsymbol{\theta}^* L$, we get $M_L(\boldsymbol{\theta}^*, \gamma^*) = 1$ which gives the desired result.

6 Proof of Proposition 1

We divide the proof into pieces with some of the more technical details deferred to the Appendix. We will need some extra notations for the indexing of coordinates of probability vectors in $\mathcal{P}_d = \mathcal{P}(\{0, 1\}^d)$. So far we have used superscripts to index the coordinates from left to right. It is sometimes convenient to use a complementary subscript indexing, by going from right to left. More specifically, for $x \in \mathcal{P}_d$, we write

$$x = (x^0, x^1, \dots, x^{m-1}) = (x_{m-1}, \dots, x_1, x_0) \quad (37)$$

so that $x_i = x^{m-1-i}$. We also interpret x_i as the value that x assigns to the binary representation¹ of i . Furthermore, for any $i = 0, \dots, m-1$, let

$$b_j(i) := j\text{th bit from the left in binary expansion of } i, \quad j \in [d], \quad (38)$$

so that the binary expansion of i is the string $b_1(i)b_2(i)\dots b_d(i)$.

Before starting the proof, let us give an explicit expression for the common sequence $\{\boldsymbol{\theta}_n\}$ used in the iterations of both the exact and approximate algorithms. Recall the notation $y_n := P(Z_*^n | \mathbf{X}_*^n)$ introduced in (17), in which $y_n \in \mathcal{P}_d$ is defined by looking at $P(Z_*^n | \mathbf{X}_*^n)$ as a random probability vector indexed by $Z_*^n \in \{0, 1\}^d$. Similarly, in view of (12), let

$$\mathbf{h}_n := \prod_{j \in V} P(X_j^n | Z_j^n) \prod_{\{i,j\} \in E} P(X_{ij}^n | Z_i^n, Z_j^n) \quad (39)$$

where the ingredients are given by (13). As before, in this expression, we are treating Z_*^n as indexing a random vector in \mathbb{R}_+^m . For $n \in \mathbb{N}$, let

$$\boldsymbol{\theta}_n := \frac{\mathbf{h}_n}{(\mathbf{h}_n)_{m-1}}, \quad (40)$$

where $(\mathbf{h}_n)_i$ denotes the i th entry of \mathbf{h}_n , using subscript indexing according to (37). In other words, to obtain $\boldsymbol{\theta}_n$, we normalize $\mathbf{h}_n = ((\mathbf{h}_n)_{m-1}, \dots, (\mathbf{h}_n)_0)$ by dividing it by its first entry. Using (13) and (38), we can write

$$(\boldsymbol{\theta}_n)_\ell = \prod_{j \in V} \left[\frac{g_j(X_j^n)}{f_j(X_j^n)} \right]^{1-b_j(\ell)} \prod_{\{i,j\} \in E} \left[\frac{g_{ij}(X_{ij}^n)}{f_{ij}(X_{ij}^n)} \right]^{1-b_j(\ell) \vee b_j(\ell)}, \quad (41)$$

where \vee denotes the maximum.

Recall that for $\rho \in [0, 1]$, we use the notation $\bar{\rho} := 1 - \rho$.

6.1 The approximate algorithm follows general iteration (2)

In order to avoid confusion with exact quantities, we will use a tilde to denote the posterior quantities produced by the approximate iteration. For example, (14) can be rewritten as an exact equality in terms of approximate quantities,

$$\tilde{P}(Z_*^n | \mathbf{X}_*^{n-1}) = \prod_{j \in V} \nu(Z_j^n; \tilde{\gamma}_j^{n-1}[n]) \quad (42)$$

We first note that recursion (15) is simplified for a geometric prior. We have $\pi_j(n) = \bar{\rho}_j^{n-1} \rho_j$ and $\pi_j[n]^c = \bar{\rho}_j^n$. Then, (15) for the approximate algorithm is

$$\tilde{\gamma}_j^{n-1}[n] = \rho_j + \bar{\rho}_j \tilde{\gamma}_j^{n-1}[n-1]. \quad (43)$$

Consider an operator \mathcal{R}_ρ on $\mathcal{P}_1 := \mathcal{P}(\{0, 1\})$ defined by

$$\mathcal{R}_\rho \left(\begin{pmatrix} x_1 \\ x_0 \end{pmatrix} \right) := \rho \begin{pmatrix} 1 \\ 0 \end{pmatrix} + (1 - \rho) \begin{pmatrix} x_1 \\ x_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + (1 - \rho) \begin{pmatrix} -x_0 \\ x_0 \end{pmatrix} \quad (44)$$

¹For example, for $d = 2$, $x = (x_3, x_2, x_1, x_0) = (x(\{(1, 1)\}), x(\{(1, 0)\}), x(\{(0, 1)\}), x(\{(0, 0)\}))$, where the multitude of parentheses is because in the RHS, we are treating x as a measure (i.e., a set-valued function) on all subsets of $\{0, 1\}^d$.

for any vector $x = (x_1, x_0) = (1 - x_0, x_0) \in \mathcal{P}_1$. (We are using the subscript indexing introduced in (37).)

Recall that $\mathcal{P}_d := \mathcal{P}(\{0, 1\}^d)$. Let $\mathcal{M}_j : \mathcal{P}_d \rightarrow \mathcal{P}_1$ be the j th marginalization operator, that is, an operator which produces the j -th marginal when applied to probability vector $y \in \mathcal{P}_d$. More explicitly,

$$[\mathcal{M}_j(y)]_1 := \sum_{i: b_j(i)=1} y_i. \quad (45)$$

(On the LHS, we are again using the subscript indexing.) For $z \in \mathcal{P}_r$ and $y \in \mathcal{P}_d$, let $z \otimes y \in \mathcal{P}_{r+d}$ be the probability vector corresponding to the product of z and y as measures. It is the usual tensor product if we think of z and y as vectors.

Now, let

$$\tilde{y}_n := \tilde{P}(Z_*^n | \mathbf{X}_*^n), \quad \text{and} \quad \tilde{w}_n := \tilde{P}(Z_*^n | \mathbf{X}_*^{n-1})$$

in the sense discussed in Section 4.2 leading to (17). In words, \tilde{y}_n is a vector in \mathcal{P}_d representing the estimate of the joint posterior of Z_*^n given \mathbf{X}_*^n , produced at the n -th step of the approximate algorithm. Similar interpretation holds for \tilde{w}_n .

Recall that $\tilde{\gamma}_j^n[n] = \tilde{P}(Z_j^n = 1 | \mathbf{X}_*^n)$ and $\tilde{\gamma}_j^{n-1}[n] = \tilde{P}(Z_j^n = 1 | \mathbf{X}_*^{n-1})$. In other words, $(\tilde{\gamma}_j^n[n], 1 - \tilde{\gamma}_j^n[n])$ is the j -th marginal of \tilde{y}_n , and $(\tilde{\gamma}_j^{n-1}[n], 1 - \tilde{\gamma}_j^{n-1}[n])$ is the j -th marginal of \tilde{w}_n . It follows from (43) and the definitions of \mathcal{R}_ρ and \mathcal{M}_j that

$$\mathcal{M}_j(\tilde{w}_n) = \mathcal{R}_{\rho_j}(\mathcal{M}_j(\tilde{y}_{n-1})).$$

On the other hand, (42) states that \tilde{w}_n is a product measure,

$$\tilde{w}_n = \otimes_{j=1}^d \mathcal{M}_j(\tilde{w}_n).$$

Combining the two, we get

$$\tilde{w}_n = \otimes_{j=1}^d [\mathcal{R}_{\rho_j}(\mathcal{M}_j(\tilde{y}_{n-1}))] =: T_{\text{ap}}(\tilde{y}_{n-1}). \quad (46)$$

It is easy to verify that each element of $T_{\text{ap}}(\tilde{y}_{n-1})$ as defined above is a polynomial of degree (at most) d in elements of \tilde{y}_{n-1} , with coefficients that depend only on $\{\rho_j\}$.

It remains to investigate how \tilde{w}_n produces \tilde{y}_n . Using (12), we observe that $\tilde{w}_n \equiv \tilde{P}(Z_*^n | \mathbf{X}_*^{n-1})$ is mapped to $\tilde{P}(Z_*^n, X_*^n | \mathbf{X}_*^{n-1})$ by a pointwise multiplication with \mathbf{h}_n as defined in (39). Since, $\tilde{y}_n \equiv \tilde{P}(Z_*^n | \mathbf{X}_*^n)$ is obtained from $\tilde{P}(Z_*^n, X_*^n | \mathbf{X}_*^{n-1})$ by a normalization over Z_*^n , we obtain

$$\tilde{y}_n = \frac{\tilde{w}_n \circ \mathbf{h}_n}{\tilde{w}_n^T \mathbf{h}_n} = \frac{\tilde{w}_n \circ \boldsymbol{\theta}_n}{\tilde{w}_n^T \boldsymbol{\theta}_n} = q_{\boldsymbol{\theta}_n}(\tilde{w}_n). \quad (47)$$

This completes the proof.

6.2 The exact algorithm follows general iteration (2)

Let

$$y_n := P(Z_*^n | \mathbf{X}_*^n), \quad \text{and} \quad w_n := P(Z_*^n | \mathbf{X}_*^{n-1})$$

be the posteriors produced by the exact algorithm. One observes that (47) holds with \tilde{w}_n replaced with w_n and \tilde{y}_n replaced with y_n . That is, $y_n = q_{\boldsymbol{\theta}_n}(w_n)$. The difference with the approximate algorithm is in updating w_n based on y_{n-1} . To derive this map, we need the following lemma. Recall that π_j is the prior on the j -th change point λ_j .

Lemma 3. Let $\mathcal{J} \subset [d]$ and consider collections of integers $\{k_j\}_{j \in \mathcal{J}}$ and $\{m_j\}_{j \in \mathcal{J}}$ in $\{n+1, n+2, \dots\}$. Then, we have

$$\frac{P(\lambda_j = k_j, j \in \mathcal{J} | \mathbf{X}_*^n)}{P(\lambda_j = m_j, j \in \mathcal{J} | \mathbf{X}_*^n)} = \prod_{j \in \mathcal{J}} \frac{\pi_j(k_j)}{\pi_j(m_j)}$$

Proof. This follows from Lemma 6 which implies $P(\lambda_j = k_j, j \in \mathcal{J} | \mathbf{X}_*^n)$ and $P(\lambda_j = m_j, j \in \mathcal{J} | \mathbf{X}_*^n)$ are equal for the collection of integers considered. \square

We note that both w_n and y_{n-1} are based on conditional probabilities, given \mathbf{X}_*^{n-1} , of events in terms of $\{\lambda_j\}$. Updating w_n based on y_{n-1} amounts to evaluating the values a fixed probability measure assigns to a collection of sets, based on the values it assigns to a different collection of sets. The particular nature of these sets and Lemma 3 allow this computation.

The formula has an algebraic structure. We work with polynomials of degree d , in indeterminate variables $\bar{\omega}$ and $\underline{\omega}$. We assume the product of $\bar{\omega}$ and $\underline{\omega}$ to be noncommutative. (That is, $\bar{\omega}\underline{\omega} \neq \underline{\omega}\bar{\omega}$.) Denote the space of such polynomials as \mathcal{X}_d . We think of $\bar{\omega}$ and $\underline{\omega}$ as digits 1 and 0, respectively. Then, a string consisting of $\bar{\omega}$ and $\underline{\omega}$ represents a binary number. Let $B(\cdot)$ be the map that produces this binary number given a string of $\bar{\omega}$ and $\underline{\omega}$. For example, $B(\bar{\omega}\underline{\omega}\bar{\omega}) = 101 \equiv 5$.

Let $L_{y_{n-1}}(\cdot)$ be a “linear” map defined on \mathcal{X}_d which maps a string s of $\bar{\omega}$ and $\underline{\omega}$ to $(y_{n-1})_{B(s)}$. This implies, for example,

$$L_{y_{n-1}}(2\bar{\omega}\underline{\omega}\bar{\omega} + 3\underline{\omega}\bar{\omega}\bar{\omega}) = 2(y_{n-1})_5 + 3(y_{n-1})_3.$$

Let

$$u_j^{(i)}(\bar{\omega}, \underline{\omega}) = \begin{cases} \bar{\rho}_j \underline{\omega}, & b_j(i) = 0 \\ \bar{\omega} + \rho_j \underline{\omega} & b_j(i) = 1. \end{cases} \quad (48)$$

The following lemma describes the rule mapping y_{n-1} to w_n .

Lemma 4. For $i = 0, \dots, m-1$,

$$(w_n)_i = L_{y_{n-1}}\left(u_1^{(i)}(\bar{\omega}, \underline{\omega}) u_2^{(i)}(\bar{\omega}, \underline{\omega}) \cdots u_d^{(i)}(\bar{\omega}, \underline{\omega})\right). \quad (49)$$

The sketch of the proof is given in Appendix B. To get a sense of what (49) means, consider the case $d = 2$. Then, for example,

$$\begin{aligned} (w_n)_2 &= L_{y_{n-1}}\left((\bar{\omega} + \rho_1 \underline{\omega})(\bar{\rho}_2 \underline{\omega})\right) = L_{y_{n-1}}\left(\bar{\rho}_2 \bar{\omega} \underline{\omega} + \rho_1 \bar{\rho}_2 \underline{\omega} \underline{\omega}\right) \\ &= \bar{\rho}_2 (y_{n-1})_2 + \rho_1 \bar{\rho}_2 (y_{n-1})_0. \end{aligned}$$

As can be seen from this example, (49) is a compact way of expressing a linear relation $w_n = T_{\text{ex}} y_{n-1}$, for some $m \times m$ matrix T_{ex} . For example, for $d = 2$, the matrix is given by

$$T_{\text{ex}} = \begin{pmatrix} 1 & \rho_2 & \rho_1 & \rho_1 \rho_2 \\ 0 & \bar{\rho}_2 & 0 & \rho_1 \bar{\rho}_2 \\ 0 & 0 & \bar{\rho}_1 & \bar{\rho}_1 \rho_2 \\ 0 & 0 & 0 & \bar{\rho}_1 \bar{\rho}_2 \end{pmatrix}. \quad (50)$$

This completes the proof.

6.3 Bounding Lipschitz constant of T_{ex}

Since T_{ex} is a Markov transition matrix, we have $\mathbf{1}_m^T T_{\text{ex}} = 0$. Note that our convention leads to the transpose of what is usually considered a Markov transition matrix. That is, columns of T_{ex} sum to 1 (not the rows). Based on Lemma 4, it is not hard to observe the following:

- The first column of T_{ex} is equal to $e^{(0)} := (1, 0, \dots, 0) \in \mathbb{R}^m$.
- The first row of T_{ex} consists of elements of the form $\prod_{j \in S} \rho_j$, for $S \subset [d]$. In particular, the first element of the first row is 1 (corresponding to $S = \emptyset$) while the last element is $\prod_{j=1}^d \rho_j$ (corresponding to $S = [d]$).

We will apply Lemma 5 of Appendix C to the linear map \tilde{F} given by $\tilde{F}(x) = T_{\text{ex}}x$ for $x \in \mathbb{R}^m$. The Jacobian of T_{ex} is constant and equal to T_{ex} . Applying Lemma 5 with $u(x) = (\prod_{j=1}^d \rho_j) e^{(0)}$ (independent of x), we obtain

$$\text{Lip}_{\tilde{F}} \leq \underbrace{\| T_{\text{ex}} - (\prod_{j=1}^d \rho_j) e^{(0)} \mathbf{1}_m^T \|_1}_{=: A}.$$

Note that $e^{(0)} \mathbf{1}_m^T$ is an $m \times m$ matrix with the first row being all ones, and the rest being all zeros. Thus, the matrix A coincides with T_{ex} outside the first row. Moreover, on the first row, where T_{ex} has entry $\prod_{j \in S} \rho_j$, A has entry $\prod_{j \in S} \rho_j - \prod_{j=1}^d \rho_j \geq 0$. That is, all the entries of A are nonnegative. Hence, the absolute column sums for A , are the same as its column sums. Furthermore, since all the columns of both T_{ex} and $e^{(0)} \mathbf{1}_m^T$ sum to one, we have $\|A\|_1 = \sum_i A_{ik} = 1 - \prod_{j=1}^d \rho_j$, for any k . This gives the desired bound on the Lipschitz constant. (It is not hard to verify that bound is sharp, that is, the Lipschitz constant is in fact equal to $1 - \prod_{j=1}^d \rho_j$.)

6.4 Bounding Lipschitz constant of T_{ap}

Recall the expression for T_{ap} given in (46). We will rewrite it as the composition of two functions. Recall that $m := 2^d$. Let $H : \mathbb{R}^d \rightarrow \mathbb{R}^m$ be defined as

$$H(u) := H(u_1, \dots, u_d) := \otimes_{j=1}^d \begin{pmatrix} u_j \\ 1 - u_j \end{pmatrix}$$

where \otimes is the (tensor) product of two measures defined in Section 6.1. Here, we use our convention (for embedding \mathcal{P}_d in \mathbb{R}^m) to treat the result of the tensor product as an element of \mathbb{R}^m . For example, for $d = 2$, $H(u_1, u_2) = (u_1 u_2, u_1(1 - u_2), (1 - u_1)u_2, (1 - u_1)(1 - u_2))$.

Also, let $K : \mathbb{R}^m \rightarrow \mathbb{R}^d$ be defined as

$$K(y) := \left(1 - \bar{\rho}_1[\mathcal{M}_1(y)]_0, \dots, 1 - \bar{\rho}_d[\mathcal{M}_d(y)]_0 \right)$$

where $[\mathcal{M}_j(y)]_0$ is the value assigned to 0 by the j th marginal of y . (Note that each marginal $\mathcal{M}_j(y)$ is a probability distribution on $\{0, 1\}$.) To simplify notation, we will also use

$$u_j(y) := 1 - \bar{\rho}_j[\mathcal{M}_j(y)]_0$$

so that $K(y) = (u_1(y), \dots, u_d(y))$. For example, for $d = 2$, with $y = (y_3, y_2, y_1, y_0)$, we have $u_1(y) = 1 - \bar{\rho}_1(y_1 + y_0)$ and $u_2(y) = 1 - \bar{\rho}_2(y_2 + y_0)$.

Recalling the definition (44) of \mathcal{R}_{ρ_j} , and (46), one observes that $H \circ K := H(K(\cdot))$ is an extension of T_{ap} to all of \mathbb{R}^m . In other words,

$$T_{\text{ap}} = H \circ K|_{\mathcal{P}_d}.$$

Thus, we can estimate the Lipschitz constant of T_{ap} by computing the Jacobian of $H \circ K$ and applying Lemma 5 of Appendix C. By chain rule, the Jacobian of the composition is the product of Jacobians. More precisely, $J_{H \circ K}(y) = J_H(u)J_K(y)$ with $u = K(y)$.

To compute $J_H(u) \in \mathbb{R}^{m \times d}$, first note that we can write the i th component of $H(u)$ as $[H(u)]_i = \prod_{k=1}^d u_k^{b_k(i)} (1 - u_k)^{1-b_k(i)}$ where $b_k(i)$ is the bit notation introduced in (38). It follows that

$$[J_H(u)]_{ij} = \partial_{u_j} [H(u)]_i = (-1)^{1-b_j(i)} \prod_{k \neq j} u_k^{b_k(i)} (1 - u_k)^{1-b_k(i)}$$

For $y \in \mathcal{P}_d$, we have $u = K(y) \in [0, 1]^d$, that is, both u_k and $1 - u_k$ are nonnegative for all $k \in [d]$. It is not then hard to verify that $\sum_{i=1}^m |[J_H(u)]_{ij}| = 2$, for all $j \in [d]$. That is, all the absolute column sums of J_H are equal to 2, which implies $\|J_H(u)\|_1 = 2$ for $u \in [0, 1]^d$.

Turning to $J_K(y) \in \mathbb{R}^{d \times m}$, we note that this is in fact a constant matrix, as K is an affine map. Using an expression similar to (45), we have

$$[J_K]_{j\ell} = \partial_{y_\ell} u_j = -\bar{\rho}_j \partial_{y_\ell} \left(\sum_{i: b_j(i)=0} y_i \right) = -\bar{\rho}_j (1 - b_j(\ell)).$$

In other words, the j -th row of J_K contains $-\bar{\rho}_j$ in columns ℓ with $b_j(\ell) = 0$, and is zero otherwise. For example, for $d = 3$ (and $m = 8$), we obtain

$$J_K = - \begin{pmatrix} 0 & 0 & 0 & 0 & \bar{\rho}_1 & \bar{\rho}_1 & \bar{\rho}_1 & \bar{\rho}_1 \\ 0 & 0 & \bar{\rho}_2 & \bar{\rho}_2 & 0 & 0 & \bar{\rho}_2 & \bar{\rho}_2 \\ 0 & \bar{\rho}_3 & 0 & \bar{\rho}_3 & 0 & \bar{\rho}_3 & 0 & \bar{\rho}_3 \end{pmatrix}$$

According to Lemma 5, it is possible to add a constant to each row of J_K and still obtain an upper bound on the Lipschitz constant of T_{ap} . We will add $\rho_j/2$ to each column in the j -th row. More precisely, let $\bar{r} := (\bar{\rho}_1, \dots, \bar{\rho}_d) \in \mathbb{R}^d$. Then, we consider $J_K + \frac{1}{2} \bar{r} \mathbf{1}_m^T$. For example, in the case of $d = 3$, we have

$$J_K + \frac{1}{2} \bar{r} \mathbf{1}_m^T = \frac{1}{2} \begin{pmatrix} \bar{\rho}_1 & \bar{\rho}_1 & \bar{\rho}_1 & \bar{\rho}_1 & -\bar{\rho}_1 & -\bar{\rho}_1 & -\bar{\rho}_1 & -\bar{\rho}_1 \\ \bar{\rho}_2 & \bar{\rho}_2 & -\bar{\rho}_2 & -\bar{\rho}_2 & \bar{\rho}_2 & \bar{\rho}_2 & -\bar{\rho}_2 & -\bar{\rho}_2 \\ \bar{\rho}_3 & -\bar{\rho}_3 & \bar{\rho}_3 & -\bar{\rho}_3 & \bar{\rho}_3 & -\bar{\rho}_3 & \bar{\rho}_3 & -\bar{\rho}_3 \end{pmatrix}.$$

It is easy to verify that the absolute column sum for each column of this new matrix equal to $\frac{1}{2} \sum_{j=1}^d \bar{\rho}_j$. That is, $\|J_K + \frac{1}{2} \bar{r} \mathbf{1}_m^T\|_1 = \frac{1}{2} \sum_{j=1}^d \bar{\rho}_j$.

We can now apply lemma 5 to obtain

$$\begin{aligned} \text{Lip}_{T_{\text{ap}}} &\leq \sup_{y \in \mathcal{P}_d} \|J_{H \circ K}(y) + \left(\frac{1}{2} J_H(u) \bar{r}\right) \mathbf{1}_m^T\|_1 \\ &= \sup_{y \in \mathcal{P}_d} \|J_H(u) [J_K + \frac{1}{2} \bar{r} \mathbf{1}_m^T]\|_1 \\ &\leq \sup_{y \in \mathcal{P}_d} \left\{ \|J_H(u)\|_1 \|J_K + \frac{1}{2} \bar{r} \mathbf{1}_m^T\|_1 \right\} = \sum_{j=1}^d \bar{\rho}_j \end{aligned}$$

where as before $u = K(y)$, and the last inequality follows by the sub-multiplicative property of $\|\cdot\|_1$. The proof is complete.

References

- [1] A. N. Shiryaev. *Optimal Stopping Rules*. Springer-Verlag, 1978.
- [2] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
- [3] M. I. Jordan. Graphical models. *Statistical Science*, 19:140–155, 2004.
- [4] P. Diaconis and D. Freedman. Iterated random functions. *SIAM Rev.*, 41(1):45–76, 1999.
- [5] O. P. Kreidl and A. Willsky. Inference with minimum communication: a decision-theoretic variational approach. In *NIPS*, 2007.
- [6] M. Cetin, L. Chen, J. W. Fisher III, A. Ihler, R. Moses, M. Wainwright, and A. Willsky. Distributed fusion in sensor networks: A graphical models perspective. *IEEE Signal Processing Magazine*, July:42–55, 2006.
- [7] X. Nguyen, A. A. Amini, and R. Rajagopal. Message-passing sequential detection of multiple change points in networks. In *ISIT*, 2012.
- [8] A. Frank, P. Smyth, and A. Ihler. A graphical model representation of the track-oriented multiple hypothesis tracker. In *Proceedings, IEEE Statistical Signal Processing (SSP)*. August 2012.
- [9] A. T. Ihler, J. W. Fisher III, and A. S. Willsky. Loopy belief propagation: Convergence and effects of message errors. *Journal of Machine Learning Research*, 6:905–936, May 2005.
- [10] Alexander Ihler. Accuracy bounds for belief propagation. In *Proceedings of UAI 2007*, July 2007.
- [11] T. G. Roosta, M. Wainwright, and S. S. Sastry. Convergence analysis of reweighted sum-product algorithms. *IEEE Trans. Signal Processing*, 56(9):4293–4305, 2008.
- [12] D. Steinsaltz. Locally contractive iterated function systems. *Ann. Probab.*, 27(4):1952–1979, 1999.
- [13] W. B. Wu and M. Woodroffe. A central limit theorem for iterated random functions. *J. Appl. Probab.*, 37(3):748–755, 2000.
- [14] W. B. Wu and X. Shao. Limit theorems for iterated random functions.. *J. Appl. Probab.*, 41(2):425–436, 2004.
- [15] Ö. Stenflo. A survey of average contractive iterated function systems. *J. Diff. Equa. and Appl.*, 18(8):1355–1380, 2012.
- [16] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

- [17] A. A. Amini and X. Nguyen. Sequential detection of multiple change points in networks: a graphical model approach. *IEEE Transactions on Information Theory*, 59(9):5824–5841, 2013.
- [18] Yu. V. Kozachenko V. V. Buldygin. *Metric characterization of random variables and random processes*. Amer. Math. Soc., 2000.

A Proof of (8)

Recall that $\pi(k) := \mathbb{P}(\lambda = k)$. Let $[n] := \{1, \dots, n\}$ and $[n-1]^c := \{n, n+1, \dots\}$. For $k, r \in [n-1]^c$, we have

$$\frac{\mathbb{P}(\lambda = k | \mathbf{X}^{n-1})}{\mathbb{P}(\lambda = r | \mathbf{X}^{n-1})} = \frac{P(\mathbf{X}^{n-1} | \lambda = k) \pi(k)}{P(\mathbf{X}^{n-1} | \lambda = r) \pi(r)} = \frac{\pi(k)}{\pi(r)} \quad (51)$$

since the function $k \mapsto P(\mathbf{X}^{n-1} | \lambda = k)$ is constant over $[n-1]^c$. In fact, $P(\mathbf{X}^{n-1} | \lambda = k) = \prod_{t=1}^{n-1} g(X^t)$ for all $k \geq n$. In (51), take $r = n$, and sum over $k \in [n-1]^c$ to obtain (after inversion)

$$\frac{\mathbb{P}(\lambda = n | \mathbf{X}^{n-1})}{\mathbb{P}(\lambda \in [n-1]^c | \mathbf{X}^{n-1})} = \frac{\pi(n)}{\pi[n-1]^c}.$$

For any subset $A \subset \mathbb{N} := \{1, 2, \dots\}$, let us use the notation $\gamma^{n-1}A := \mathbb{P}(\lambda \in A | \mathbf{X}^{n-1})$. Thus, we have shown $\gamma^{n-1}\{n\} = \frac{\pi(n)}{\pi[n-1]^c} \gamma^{n-1}[n-1]^c$.

From additivity of probability measures, we have

$$\gamma^{n-1}[n-1]^c = 1 - \gamma^{n-1}[n-1], \quad \gamma^{n-1}\{n\} = \gamma^{n-1}[n] - \gamma^{n-1}[n-1].$$

Substituting these in the earlier equation, we obtain

$$\gamma^{n-1}[n] = \frac{\pi(n)}{\pi[n-1]^c} + \left(1 - \frac{\pi(n)}{\pi[n-1]^c}\right) \gamma^{n-1}[n-1]$$

which is the desired result.

B Proof of Lemma 4

Let $\mathbb{N} := \{1, 2, \dots\}$ denote the set of natural numbers. Let $A := [n] := \{1, \dots, n\}$ and let A^c be the complement of A in \mathbb{N} , that is, $A^c = \{n+1, n+2, \dots\}$. Similarly, let $B = [n+1]$ and let $B^c = \{n+2, n+3, \dots\}$. We also let $b := \{n+1\}$. (These notations are local to this proof.)

For an index set $\mathcal{J} = \{i_1, \dots, i_r\} \subset d$, let $\gamma_{\mathcal{J}}^n$ denote the joint posterior of $\lambda_\ell, \ell \in \mathcal{J}$ given \mathbf{X}_*^n . More precisely, $\gamma_{\mathcal{J}}^n(E_1, \dots, E_r) = \mathbb{P}(\bigcap_{j=1}^r \{\lambda_{i_j} \in E_j\} | \mathbf{X}_*^n)$ for any collection E_1, \dots, E_r of subsets of \mathbb{N} . Let A° denote either A or A^c , and similarly for B° . We would like to compute quantities of the form $\gamma_{\mathcal{J}}^n(B^\circ, \dots, B^\circ)$ in terms of known quantities $\gamma_{\mathcal{J}}^n(A^\circ, \dots, A^\circ)$. For simplicity, we will drop superscript n from now on.

We will use $-$ and $+$ to denote set difference and disjoint union, respectively. For example, $B = A + b$ and $B^c = A^c - b$. We proceed in stages, by first finding probabilities of “sequences of A^c and b ”; we do this by an example. Consider $\gamma_{1234}(A^c, b, A^c, b)$. Applying Lemma 3, we have

$$\frac{\gamma_{1234}(A^c, b, A^c, b)}{\gamma_{1234}(b, b, b, b)} = \frac{\pi_1(A^c)}{\pi_1(b)} \frac{\pi_3(A^c)}{\pi_3(b)} = \frac{1}{\rho_1 \rho_3}.$$

Similarly,

$$\frac{\gamma_{1234}(A^c, A^c, A^c, A^c)}{\gamma_{1234}(b, b, b, b)} = \frac{1}{\rho_1 \rho_2 \rho_3 \rho_4}.$$

It follows that

$$\gamma_{1234}(A^c, b, A^c, b) = \rho_2 \rho_4 \gamma_{1234}(A^c, A^c, A^c, A^c),$$

which is the desired result, since the RHS is known. By induction, we have the following rule: The probability of a sequence of A^c and b is the probability of the sequence of all- A^c multiplied by “ ρ_i ”s associated with places of “ b ”s. We will later use a more compact notation: $A^c b A^c b = \rho_2 \rho_4 A^c A^c A^c A^c$, to express the same fact.

We turn to the case where we have a sequence of A^c and b and a single A . Consider, for example,

$$\begin{aligned} \gamma_{1234}(A^c, b, A, b) &= \gamma_{1234}(A^c, b, \mathbb{N}, b) - \gamma_{1234}(A^c, b, A^c, b) \\ &= \gamma_{124}(A^c, b, b) - \gamma_{1234}(A^c, b, A^c, b) \\ &= \rho_2 \rho_4 \gamma_{124}(A^c, A^c, A^c) - \rho_2 \rho_4 \gamma_{1234}(A^c, A^c, A^c, A^c) \\ &= \rho_2 \rho_4 \gamma_{1234}(A^c, A^c, A, A^c). \end{aligned}$$

where third equality follows by the rule regarding sequences of A^c and b . Thus, by induction, we can revise our rule to include the sequences with a single A : We proceed by replacing “ b ”s with A^c and multiplying by corresponding “ ρ_i ”s, leaving the A intact.

Now, consider a sequence with more than one A . For example,

$$\begin{aligned} \gamma_{1234}(A^c, b, A, A) &= \gamma_{1234}(A^c, b, A, \mathbb{N}) - \gamma_{1234}(A^c, b, A, A^c) \\ &= \gamma_{123}(A^c, b, A) - \gamma_{1234}(A^c, b, A, A^c) \end{aligned}$$

where both terms involve sequences with single A . Applying our rule to each term and combining the result as before, we get, in compact notation, $A^c b A A = \rho_2 A^c A^c A A$. Thus, by induction, our rule extends to sequences of A^c , b , and arbitrary number of “ A ”s: Replace “ b ”s with “ A^c ”s and scale appropriately, leaving “ A ”s intact.

We are now ready to obtain probabilities of a sequence of B s and B^c s. Consider the following example,

$$\begin{aligned} \gamma_{12}(B^c, B) &= \gamma_{12}(A^c - b, A + b) \\ &= \gamma_{12}(A^c - b, A) + \gamma_{12}(A^c - b, b) \\ &= \gamma_{12}(A^c, A) - \gamma_{12}(b, A) + \gamma_{12}(A^c, b) - \gamma_{12}(b, b), \end{aligned}$$

by finite additivity of probability measures. We can represent this identity in a compact form. $B^c B = (A^c - b)(A + b) = A^c A - bA + A^c b - bb$. Applying our rule, we obtain

$$\begin{aligned} B^c B &= A^c A - \rho_1 A^c A + \rho_2 A^c A^c - \rho_1 \rho_2 A^c A^c \\ &= (1 - \rho_1) A^c A + (1 - \rho_1) \rho_2 A^c A^c. \end{aligned}$$

This result can be obtained easier by replacing b in the first and the second sets of parentheses with $\rho_1 A^c$ and $\rho_2 A^c$, respectively, and following rules of a noncommutative associative algebra,

$$\begin{aligned} B^c B &= (A^c - b)(A + b) \\ &= (A^c - \rho_1 A^c)(A + \rho_2 A^c) = (1 - \rho_1) A^c (A + \rho_2 A^c) = \bar{\rho}_1 A^c A + \bar{\rho}_1 \rho_2 A^c A^c. \end{aligned}$$

Using this procedure, we can express the probability of any sequence of B and B^c in terms of sequences of A and A^c . As another example,

$$\begin{aligned}
B^c B B B^c &= (A^c - b)(A + b)(A + b)(A^c - b) \\
&= (A^c - \rho_1 A^c)(A + \rho_2 A^c)(A + \rho_3 A^c)(A^c - \rho_4 A^c) \\
&= (\bar{\rho}_1 A^c)(A + \rho_2 A^c)(A + \rho_3 A^c)(\bar{\rho}_4 A^c).
\end{aligned} \tag{52}$$

As before, the final expression is obtained by expanding. The general pattern is now clear and can be formally established by induction. The proof is complete. To link with the notation of the theorem, replace A^c with $\underline{\omega}$ and A with $\bar{\omega}$. The function $u^{(i)}$ defined in (48) replaces a set of parantheses, in derivations above, with the correct expression in terms of $\underline{\omega}$ and $\bar{\omega}$, depending on whether the set of parantheses contains a $+$ or a $-$ sign.

C Bounding the Lipschitz constant of a probability map

This appendix is devoted to a lemma which allows us to estimate the Lipschitz constant of a map $F : \mathcal{P} \rightarrow \mathcal{P}$, on a probability space \mathcal{P} , based on the Jacobian matrix of its extension. Here, $\mathcal{P} := \mathcal{P}_d := \mathcal{P}(\{0, 1\}^d)$ is considered to be a subset of \mathbb{R}^m where $m = 2^d$. For a C^1 function $\tilde{F} : U \rightarrow \mathbb{R}^m$ defined on some open subset U of \mathbb{R}^m , let $J_{\tilde{F}}$ denote its Jacobian matrix, i.e.,

$$J_{\tilde{F}} := (\partial_{x_j} \tilde{F}_i) \in \mathbb{R}^{m \times m}$$

where $\partial_{x_j} \tilde{F}_i$ is the partial derivative of the i -th component of \tilde{F} w.r.t. the its j -th variable.

For a square matrix A and $p \in [1, \infty]$, let $\|A\|_p$ denote its norm as an operator on ℓ_p , that is, $\|A\|_p := \sup_{\|x\|_p \leq 1} \|Ax\|_p$, where $\|\cdot\|_p$ is the vector ℓ_p norm. It is well-known that $\|A\|_1$ ($\|A\|_\infty$) is the maximum absolute column (row) sum of matrix A .

Recall that $\mathbf{1}_m \in \mathbb{R}^m$ denotes the all-ones vector.

Lemma 5. *Let U be an open subset of \mathbb{R}^m , containing \mathcal{P} . Let $\tilde{F} : U \rightarrow \mathbb{R}^m$ be a C^1 extension of $F : \mathcal{P} \rightarrow \mathcal{P}$, that is, $\tilde{F}|_{\mathcal{P}} = F$. Then, for any function $u : U \rightarrow \mathbb{R}^m$ with components in $L^1(U)$,*

$$\text{Lip}_{\tilde{F}} \leq \sup_{x \in \mathcal{P}} \|J_{\tilde{F}}(x) - u(x) \mathbf{1}_m^T\|_1. \tag{53}$$

Proof. Fix some $x, y \in \mathcal{P}$ and let $z_t := x + t(y - x)$ for $t \in [0, 1]$. For $v \in \mathbb{R}^m$, we have

$$\begin{aligned}
v^T (\tilde{F}(y) - \tilde{F}(x)) &= \int_0^1 v^T \frac{d}{dt} \tilde{F}(x + t(y - x)) dt \\
&= \int_0^1 v^T J_{\tilde{F}}(z_t)(y - x) dt \\
&= \int_0^1 v^T \underbrace{[J_{\tilde{F}}(z_t) - u(z_t) \mathbf{1}_m^T]}_{=: R_t^T} (y - x) dt
\end{aligned}$$

where the last line follows since $x, y \in \mathcal{P}$ implies $\mathbf{1}_m^T(y - x) = 0$. Using ℓ_1 - ℓ_∞ duality, we have

$$\begin{aligned} \|\tilde{F}(y) - \tilde{F}(x)\|_1 &= \sup_{\|v\|_\infty \leq 1} |v^T(\tilde{F}(y) - \tilde{F}(x))| \leq \int_0^1 \sup_{\|v\|_\infty \leq 1} |(R_t v)^T(y - x)| dt \\ &\leq \|y - x\|_1 \int_0^1 \sup_{\|v\|_\infty \leq 1} \|R_t v\|_\infty dt \\ &= \|y - x\|_1 \int_0^1 \|R_t\|_\infty dt. \end{aligned}$$

Let us denote the RHS of (53) by L . Since $z_t \in \mathcal{P}$ for all $t \in [0, 1]$, we have $\|R_t\|_\infty = \|R_t^T\|_1 \leq L$, for all $t \in [0, 1]$, which completes the proof. \square

D An auxiliary lemma

Here, we record the following ‘‘constancy’’ property of the likelihood for the graphical model (9). See [17, Lemma 3] for the proof.

Lemma 6. *Let $\{i_1, i_2, \dots, i_r\} \subset [d]$ be a distinct collection of indices. The function*

$$(k_1, k_2, \dots, k_r) \mapsto P(\mathbf{X}_*^n | \lambda_{i_1} = k_1, \lambda_{i_2} = k_2, \dots, \lambda_{i_r} = k_r)$$

is constant over $\{n + 1, n + 2, \dots\}^r$.