Journal of Machine Learning Research 11 (2010) 2745-2783 Submitted 5/10; Revised 9/10; Published 10/10

Mean Field Variational Approximation for Continuous-Time Bayesian Networks[*]

Ido Cohn[†] IDO_COHN@CS.HUJI.AC.IL
Tal El-Hay[†] TALE@CS.HUJI.AC.IL
Nir Friedman NIR@CS.HUJI.AC.IL
School of Computer Science and Engineering
The Hebrew University
Jerusalem 91904, Israel

Raz Kupferman RAZ@MATH.HUJI.AC.IL
Institute of Mathematics
The Hebrew University
Jerusalem 91904, Israel

Editor: Manfred Opper

Abstract

Continuous-time Bayesian networks is a natural structured representation language for multi-component stochastic processes that evolve continuously over time. Despite the compact representation provided by this language, inference in such models is intractable even in relatively simple structured networks. We introduce a mean field variational approximation in which we use a product of *inhomogeneous* Markov processes to approximate a joint distribution over trajectories. This variational approach leads to a globally consistent distribution, which can be efficiently queried. Additionally, it provides a lower bound on the probability of observations, thus making it attractive for learning tasks. Here we describe the theoretical foundations for the approximation, an efficient implementation that exploits the wide range of highly optimized ordinary differential equations (ODE) solvers, experimentally explore characterizations of processes for which this approximation is suitable, and show applications to a large-scale real-world inference problem.

Keywords: continuous time Markov processes, continuous time Bayesian networks, variational approximations, mean field approximation

1. Introduction

Many real-life processes can be naturally thought of as evolving continuously in time. Examples cover a diverse range, starting with classical and modern physics, but also including robotics (Ng et al., 2005), computer networks (Simma et al., 2008), social networks (Fan and Shelton, 2009), gene expression (Lipshtat et al., 2005), biological evolution (El-Hay et al., 2006), and ecological systems (Opper and Sanguinetti, 2007). A joint characteristic of all above examples is that they are complex systems composed of multiple components (e.g., many servers in a server farm and multiple residues in a protein sequence). To realistically model such processes and use them in

[*] A preliminary version of this paper appeared in the Proceedings of the Twenty Fifth Conference on Uncertainty in Artificial Intelligence, 2009 (UAI 09).

[†] These authors contributed equally.

making sensible predictions we need to learn how to reason about systems that are composed of multiple components and evolve continuously in time.

Generally, when an evolving system is modeled with sufficient detail, its evolution in time is Markovian; meaning that its future state it determined by its present state—whether in a deterministic or random sense—independently of its past states. A traditional approach to modeling a multi-component Markovian process is to discretize the entire time interval into regular time slices of fixed length and represent its evolution using a *Dynamic Bayesian network,* which compactly represents probabilistic transitions between consecutive time slices (Dean and Kanazawa, 1989; Murphy, 2002; Koller and Friedman, 2009). However, as thoroughly explained in Nodelman et al. (2003), discretization of a time interval often leads either to modeling inaccuracies or to an unnecessary computational overhead. Therefore, in recent years there is a growing interest in modeling and reasoning about multi-component stochastic processes in continuous time (Nodelman et al., 2002; Ng et al., 2005; Rajaram et al., 2005; Gopalratnam et al., 2005; Opper and Sanguinetti, 2007; Archambeau et al., 2007; Simma et al., 2008).

In this paper we focus on *continuous-time Markov processes* having a discrete product state space $S = S_1 \times S_2 \times \cdots \times S_D$, where D is the number of components and the size of each S_i is finite. The dynamics of such processes that are also *time-homogeneous* can be determined by a single rate matrix whose entries encode transition rates among states. However, as the size of the state space is exponential in the number of components so does the size of the transition matrix. *Continuous-time Bayesian networks* (CTBNs) provide an elegant and compact representation language for multi-component processes that have a sparse pattern of interactions (Nodelman et al., 2002). Such patterns are encoded in CTBNs using a directed graph whose nodes represent components and edges represent direct influences among them. The instantaneous dynamics of each component depends only on the state of its parents in the graph, allowing a representation whose size scales linearly with the number of components and exponentially only with the indegree of the nodes of the graph.

Inference in multi-component temporal models is a notoriously hard problem (Koller and Friedman, 2009). Similar to the situation in discrete time processes, inference in CTBNs is exponential in the number of components, even with sparse interactions (Nodelman et al., 2002). Thus, we have to resort to approximate inference methods. The recent literature has adapted several strategies from discrete graphical models to CTBNs in a manner that attempts to exploit the continuous-time representation, thereby avoiding the drawbacks of discretizing the model.

One class of approximations includes sampling-based approaches, where Fan and Shelton (2008) introduce a likelihood-weighted sampling scheme, and more recently El-Hay et al. (2008) introduce a Gibbs-sampling procedure. The complexity of the Gibbs sampling procedure has been shown to naturally adapt to the rate of each individual component. Additionally it yields more accurate answers with the investment of additional computation. However, it is hard to bound the required time in advance, tune the stopping criteria, or estimate the error of the approximation.

An alternative class of approximations is based on *variational principles*. Recently, Nodelman et al. (2005b) and Saria et al. (2007) introduced an *Expectation Propagation* approach, which can be roughly described as a local message passing scheme, where each message describes the dynamics of a single component over an interval. This message passing procedure can be efficient. Moreover it can automatically refine the number of intervals according to the complexity of the underlying system. Nonetheless, it does suffer from several caveats. On the formal level, the approximation has no convergence guarantees. Second, upon convergence, the computed marginals do not necessarily form a globally consistent distribution. Third, it is restricted to approximations in the form of piecewise-homogeneous messages on each interval. Thus, the refinement of the number of intervals depends on the fit of such homogeneous approximations to the target process. Finally, the approximation of Nodelman *et al* does not provide a provable approximation on the likelihood of the observation—a crucial component in learning procedures.

Here, we develop an alternative variational approximation, which provides a different tradeoff. We use the strategy of structured variational approximations in graphical models (Jordan et al., 1999), and specifically the variational approach of Opper and Sanguinetti (2007) for approximate inference in latent Markov Jump Processes, a related class of models (see below for more elaborate comparison). The resulting procedure approximates the posterior distribution of the CTBN as a product of independent components, each of which is an inhomogeneous continuous-time Markov process. We introduce a novel representation that is both natural and allows numerically stable computations. By using this representation, we derive an iterative variational procedure that employs passing information between neighboring components as well as solving a small set of differential equations (ODEs) in each iteration. The latter allows us to employ highly optimized standard ODE solvers in the implementation. Such solvers use an adaptive step size, which as we show is more efficient than any fixed time interval approximation.

We finally describe how to extend the proposed procedure to branching processes and particularly to models of molecular evolution, which describe historical dynamics of biological sequences that employ many interacting components. Our experiments on this domain demonstrate that our procedure provides a good approximation both for the likelihood of the evidence and for the expected sufficient statistics. In particular, the approximation provides a lower-bound on the likelihood, and thus is attractive for use in learning.

The paper is organized as follows: In Section 2 we review continuous-time models and inference problems in such models. Section 3 introduces a general variational principle for inference using a novel parameterization. In Section 4 we apply this principle to a family of factored representations and show how to find an optimal approximation within this family. Section 5 discusses related work. Section 6 gives an initial evaluation. Section 7 presents branching process and further experiments, and Section 8 discusses our results.

2. Foundations

CTBNs are based on the framework of *continuous-time Markov processes (CTMPs)*. In this section we begin by briefly describing CTMPs. See, for example, Gardiner (2004) and Chung (1960) for a thorough introduction. Next we review the semantics of CTBNs. We then discuss inference problems in CTBNs and the challenges they pose.

2.1 Continuous Time Markov Processes

A *continuous-time stochastic process with state space S* is an uncountable collection of S-valued random variables $\{X^{(t)} : t \geq 0\}$ where $X^{(t)}$ describes the state of the system at time t. Systems with multiple components are described by state spaces that are Cartesian products of spaces, S_i, each representing the state of a single component. In this paper we consider a D-component stochastic process $X^{(t)} = (X_1^{(t)}, \dots, X_D^{(t)})$ with state space $S = S_1 \times S_2 \times \dots \times S_D$, where each S_i is finite. The states in S are denoted by vectors, $x = (x_1, \dots, x_D)$.

A *continuous-time Markov process* is a continuous-time stochastic process in which the joint distribution of every finite subset of random variables $X^{(t_0)}, X^{(t_1)}, \ldots, X^{(t_K)}$, where $t_0 < t_1 < \cdots < t_K$, satisfies the conditional independence property, also known as the Markov property:

$$\Pr(X^{(t_K)} = x_K | X^{(t_{K-1})} = x_{K-1}, \ldots, X^{(t_0)} = x_0) = \Pr(X^{(t_K)} = x_K | X^{(t_{K-1})} = x_{K-1}).$$

In simple terms, the knowledge of the state of the system at a certain time make its states at later times independent of its states at former times. In that case the distribution of the process is fully determined by the conditional probabilities of random variable pairs $\Pr(X^{(t+s)} = y | X^{(s)} = x)$, namely, by the probability that the process is in state y at time $t+s$ given that is was in state x at time s, for all $0 \leq s < t$ and $x, y \in S$. A CTMP is called *time homogeneous* if these conditional probabilities do not depend on s but only on the length of the time interval t, thus, the distribution of the process is determined by the *Markov transition functions*,

$$p_{x,y}(t) \equiv \Pr(X^{(t+s)} = y | X^{(s)} = x), \qquad \text{for all } x, y \in S \text{ and } t \geq 0,$$

which for every fixed t can be viewed as the entries of a stochastic matrix indexed by states x and y.

Under mild assumptions on the Markov transition functions $p_{x,y}(t)$, these functions are differentiable. Their derivatives at $t = 0$,

$$q_{x,y} = \lim_{t \to 0^+} \frac{p_{x,y}(t) - \boldsymbol{1}_{x=y}}{t},$$

are the entries of the *rate matrix* $\mathbb{Q}$, where $\boldsymbol{1}$ is the indicator function. This rate matrix describes the infinitesimal transition probabilities,

$$p_{x,y}(h) = \boldsymbol{1}_{x=y} + q_{x,y} h + o(h), \tag{1}$$

where $o(\cdot)$ means decay to zero faster than its argument, that is $\lim_{h \downarrow 0} \frac{o(h)}{h} = 0$. Note that the off-diagonal entries of $\mathbb{Q}$ are non-negative, whereas each of its rows sums up to zero, namely,

$$q_{x,x} = -\sum_{y \neq x} q_{x,y}.$$

The derivative of the Markov transition function for t other than 0 satisfies the so-called *forward*, or *master equation,*

$$\frac{d}{dt} p_{x,y}(t) = \sum_z q_{z,y} p_{x,z}(t). \tag{2}$$

A similar characterization for the time-dependent probability distribution, $p(t)$, whose entries are defined by

$$p_x(t) = \Pr(X^{(t)} = x), \qquad x \in S,$$

is obtained by multiplying the Markov transition function by entries of the initial distribution $p(0)$ and marginalizing, resulting in

$$\frac{d}{dt} p = p\mathbb{Q}. \tag{3}$$

The solution of this ODE is

$$p(t) = p(0) \exp(t\mathbb{Q}),$$

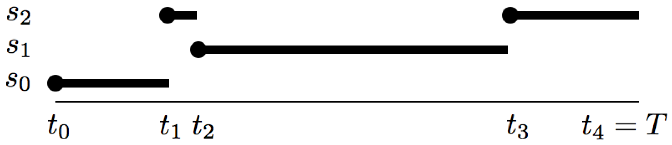


Figure 1: An example of a CTMP trajectory: The process starts at state $x_1 = s_0$, transitions to $x_2 = s_2$ at t_1, to $x_3 = s_1$ at t_2, and finally to $x_4 = s_2$ at t_3.

where $\exp(t\mathbb{Q})$ is a matrix exponential, defined for any square matrix $\mathbb{A}$ by the Taylor series,

$$\exp(\mathbb{A}) = \mathbf{I} + \sum_{k=1}^{\infty} \frac{\mathbb{A}^k}{k!} \quad .$$

Applying this solution to the initial condition $p_{x'}(0) = \mathbf{1}_{x=x'}$, we can express the Markov transition function $p_{x,y}(t)$ using the rate matrix $\mathbb{Q}$ as

$$p_{x,y}(t) = [\exp(t\mathbb{Q})]_{x,y}. \tag{4}$$

Although a CTMP is an uncountable collection of random variables (the state of the system at every time t), a *trajectory* σ of $\{X^{(t)}\}_{t\geq 0}$ over a time interval $[0,T]$ can be characterized by a finite number of transitions K, a sequence of states $(x_0, x_1, \ldots, x_K)$ and a sequence of transition times $(t_0 = 0, t_1, \ldots, t_K, t_{K+1} = T)$. We denote by $\sigma(t)$ the state at time t, that is, $\sigma(t) = x_k$ for $t_k \leq t < t_{k+1}$. Figure 1 illustrates such a trajectory.

2.2 Multi-component Representation - Continuous-Time Bayesian Networks

Equation (4) indicates that the distribution of a homogeneous Markov process is fully determined by an initial distribution and a single rate matrix $\mathbb{Q}$. However, since the number of states in a D-component Markov Process is exponential in D, an explicit representation of this transition matrix is often infeasible. *Continuous-time Bayesian networks* are a compact representation of Markov processes that satisfy two assumptions. First it is assumed that only one component can change at a time, thus transition rates involving simultaneous changes of two or more components are zero. Second, the transition rate of each component i depends only on the state of some subset of components denoted $\mathbf{Pa}_i \subseteq \{1,\ldots,D\} \setminus \{i\}$ and on its own state. This dependency is represented using a directed graph, where the nodes are indexed by $\{1,\ldots,D\}$ and the parent nodes of i are $\mathbf{Pa}_i$ (Nodelman et al., 2002). With each component i we then associate a conditional rate matrix $\mathbb{Q}^{i|\mathbf{Pa}_i}_{\cdot|u_i}$ for each state u_i of $\mathbf{Pa}_i$. The off-diagonal entries $q^{i|\mathbf{Pa}_i}_{x_i,y_i|u_i}$ represent the rate at which X_i transitions from state x_i to state y_i given that its parents are in state u_i. The diagonal entries are $q^{i|\mathbf{Pa}_i}_{x_i,x_i|u_i} = -\sum_{y_i \neq x_i} q^{i|\mathbf{Pa}_i}_{x_i,y_i|u_i}$, ensuring that each row in each conditional rate matrix sums up to zero. The dynamics of $X^{(t)}$ are defined by a rate matrix $\mathbb{Q}$ with entries $q_{x,y}$, which combines the conditional rate matrices as follows:

$$q_{x,y} = \begin{cases} q^{i|\mathbf{Pa}_i}_{x_i,y_i|u_i} & \delta(x,y) = \{i\} \\ \sum_i q^{i|\mathbf{Pa}_i}_{x_i,x_i|u_i} & x = y \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\delta(x,y) = \{j | x_j \neq y_j\}$ denotes the set of components in which x differs from y.

To have another perspective on CTBN's, we may consider a discrete-time approximation of the process. Let h be a sampling interval. The subset of random variables $\{X_{t_k} : k \geq 0\}$, where $t_k = kh$, is a discrete-time Markov process over a D-dimensional state-space. *Dynamic Bayesian networks (DBNs)* provide a compact modeling language for such processes, namely the conditional distribution of a DBN $\mathrm{P}_h(X^{(t_k+1)}|X^{(t_k)})$ is factorized into a product of conditional distributions of $X_i^{(t_{k+1})}$ given the state of a subset of $X^{(t_k)} \cup X^{(t_{k+1})}$. When h is sufficiently small, the CTBN can be approximated by a DBN whose parameters depend on the rate matrix $\mathbb{Q}$ of the CTBN ,

$$\mathrm{P}_h(X^{(t_{k+1})} = y | X^{(t_k)} = x) = \prod_{i=1}^{D} \mathrm{P}_h(X_i^{(t_{k+1})} = y_i | X_i^{(t_k)} = x_i, U^{(t_k)} = u_i), \tag{6}$$

where

$$\mathrm{P}_h(X_i^{(t_{k+1})} = y_i | X_i^{(t_k)} = x_i, U^{(t_k)} = u_i) = \mathbf{1}_{x_i = y_i} + q^{i|\mathbf{Pa}_i}_{x_i,y_i|u_i} h. \tag{7}$$

Each such term is the local conditional probability that $X_i^{(t_{k+1})} = y_i$ given the state of X_i and U_i at time t_k. These are valid conditional distributions, because they are non-negative and are normalized, that is

$$\sum_{y_i \in S_i} \left(\mathbf{1}_{x_i = y_i} + q^{i|\mathbf{Pa}_i}_{x_i,y_i|u_i} h \right) = 1$$

for every x_i and u_i. Note that in this discretized process, transition probabilities involving changes in more than one component are $o(h)$, as in the CTBN. Moreover, using Equations (1) and (5) we observe that

$$\Pr(X^{(t_k+1)} = y | X^{(t_k)} = x) = \mathrm{P}_h(X^{(t_k+1)} = y | X^{(t_k)} = x) + o(h).$$

(See Appendix A for detailed derivations). Therefore, the CTBN and the approximating DBN are asymptotically equivalent as $h \to 0$.

Example 1 An example of a multi-component process is the *dynamic Ising model*, which corresponds to a CTBN in which every component can be in one of two states, -1 or $+1$, and each component prefers to be in the same state as its neighbor. These models are governed by two parameters: a *coupling parameter* β (it is the inverse temperature in physical models, which determines the strength of the coupling between two neighboring components), and a *rate parameter* τ, which determines the propensity of each component to change its state. Low values of β correspond to weak coupling (high temperature). More formally, we define the conditional rate matrices as

$$q^{i|\mathbf{Pa}_i}_{x_i,y_i|u_i} = \tau \left(1 + e^{-2y_i\beta \sum_{j \in \mathbf{Pa}_i} x_j} \right)^{-1}$$

where $x_j \in \{-1, 1\}$. This model is derived by plugging the Ising grid to *Continuous-Time Markov Networks*, which are the undirected counterparts of CTBNs (El-Hay et al., 2006).

Consider a two component Ising model whose structure and corresponding DBN are shown in Figure 2. This system is symmetric, that is, the conditional rate matrices are identical for $i \in \{1, 2\}$. As an example, for a specific choice of β and τ we have:

$$\mathbb{Q}^{i|\mathbf{Pa}_i}_{\cdot|-1} = \begin{array}{c|cc} & - & + \\ \hline - & -1 & 1 \\ + & 10 & -10 \end{array} \qquad \mathbb{Q}^{i|\mathbf{Pa}_i}_{\cdot|+1} = \begin{array}{c|cc} & - & + \\ \hline - & -10 & 10 \\ + & 1 & -1 \end{array}$$

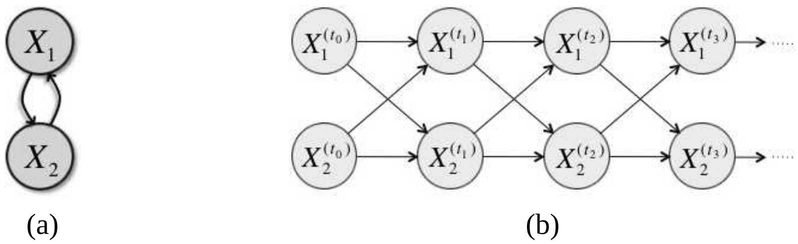


Figure 2: Two representations of a two binary component dynamic process. *(a)* The associated CTBN. *(b)* The DBN corresponding to the CTBN in (a). The models are equivalent when $h \to 0$.

The local conditional distributions of the DBN can be directly inferred from Equation (7). For example

$$\mathrm{P}_h(X_1^{(t_{k+1})} = 1 | X_1^{(t_k)} = -1, X_2^{(t_k)} = 1) = 10h.$$

Here, in both components the conditional rates are higher for transitions into states that are identical to the state of their parent. Therefore, the two components have a disposition of being in the same state. To support this intuition, we examine the amalgamated rate matrix:

$$\mathbb{Q} \;=\; \begin{array}{c|cccc} & \text{-}\text{-} & \text{-}\text{+} & \text{+}\text{-} & \text{+}\text{+} \\ \hline \text{-}\text{-} & -2 & 1 & 1 & 0 \\ \text{-}\text{+} & 10 & -20 & 0 & 10 \\ \text{+}\text{-} & 10 & 0 & -20 & 10 \\ \text{+}\text{+} & 0 & 1 & 1 & -2. \end{array}$$

Clearly, transition rates into states in which both components have the same value is higher. Higher transitions rate imply higher transition probabilities, for example:

$$\begin{aligned} p_{\text{-+},\text{--}}(h) &= 10h + o(h), \\ p_{\text{--},\text{-+}}(h) &= h + o(h). \end{aligned}$$

Thus the probability of transitions into a coherent state is much higher than into an incoherent state. ∎

2.3 Inference in Continuous-time Markov Processes

Our setting is as follows: we receive evidence of the states of several or all components within a time interval $[0,T]$. The two possible types of evidence that may be given are continuous evidence, where we know the state of a subset $U \subseteq X$ continuously over some sub-interval $[t_1,t_2] \subseteq [0,T]$, and point evidence of the state of U at some point $t \in [0,T]$. For convenience we restrict our treatment to a time interval $[0,T]$ with full end-point evidence $X^{(0)} = e_0$ and $X^{(T)} = e_T$. We shall discuss the more general case in Section 5.

Given a continuous-time Bayesian network and evidence of the above type we would like to evaluate the likelihood of the evidence, $\Pr(e_0, e_T; \mathbb{Q})$ and to compute pointwise posterior probabilities of various events (e.g., $\Pr(U^{(t)} = u | e_0, e_T)$ for some $U \subseteq X$). Another class of queries are conditional expectations of statistics that involve entire trajectories of the process. Two important examples for queries are the *sufficient statistics* required for learning. These statistics are the amount of time in which X_i is in state x_i and $\mathbf{Pa}_i$ are in state u_i, and the number of transitions that X_i underwent from x_i to y_i while its parents were in state u_i (Nodelman et al., 2003). We denote these statistics by $T^i_{x_i|u_i}$ and $M^i_{x_i,y_i|u_i}$ respectively. For example, in the trajectory of the univariate process in Figure 1, we have $T_{s_2} = t_2 - t_1 + t_4 - t_3$ and $M_{s_0,s_2} = 1$.

Exact calculation of these values is usually a computationally intractable task. For instance, calculation of marginals requires first calculating the pointwise distribution over X using a forward-backward like calculation:

$$\Pr(X^{(t)} = x|e_0, e_T) = \frac{p_{e_0,x}(t)\, p_{x,e_T}(T-t)}{p_{e_0,e_T}(T)}, \tag{8}$$

and then marginalizing

$$\Pr(U^{(t)} = u|e_0, e_T) = \sum_{x\setminus u} \Pr(X^{(t)} = x|e_0, e_T),$$

where $p_{x,y}(t) = [\exp(t\mathbb{Q})]_{x,y}$, and the size of $\mathbb{Q}$ is exponential in the number of components. Moreover, calculating expected residence times and expected number of transitions involves integration over the time interval of these quantities (Nodelman et al., 2005a):

$$\begin{aligned}
\mathbf{E}\left[T_x\right] &= \frac{1}{p_{e_0,e_T}(T)} \int_0^T p_{e_0,x}(t)\, p_{x,e_T}(T-t)dt, \\
\mathbf{E}\left[M_{x,y}\right] &= \frac{1}{p_{e_0,e_T}(T)} \int_0^T p_{e_0,x}(t)\, q_{x,y}\, p_{y,e_T}(T-t)dt \quad .
\end{aligned}$$

These make this approach infeasible beyond a modest number of components, hence we have to resort to approximations.

3. Variational Principle for Continuous-Time Markov Processes

Variational approximations to structured models aim to approximate a complex distribution by a simpler one, which allows efficient inference. This problem can be viewed as an optimization problem: given a specific model and evidence, find the “best” approximation within a given class of simpler distributions. In this setting the inference is posed as a constrained optimization problem, where the constraints ensure that the parameters correspond to valid distributions consistent with the evidence. Specifically, the optimization problem is constructed by defining a lower bound to the log-likelihood of the evidence, where the gap between the bound and the true likelihood is the divergence of between the approximation and the true posterior. While the resulting problem is generally intractable, it enables us to derive approximate algorithms by approximating either the functional or the constrains that define the set of valid distributions. In this section we define the lower-bound functional in terms of a general continuous-time Markov process (that is, without assuming any network structure). Here we aim at defining a lower bound on $\ln P_{\mathbb{Q}}(e_T|e_0)$ as well as to approximating the posterior probability $P_{\mathbb{Q}}(\cdot \mid e_0, e_T)$, where $P_{\mathbb{Q}}$ is the distribution of the Markov process whose instantaneous rate-matrix is $\mathbb{Q}$. We start by examining the structure of the posterior and introducing an appropriate parameterization.

Recall that the distribution of a time-homogeneous Markov process is characterized by the conditional transition probabilities $p_{x,y}(t)$, which in turn is fully redetermined by the constant rate matrix $\mathbb{Q}$. It is not hard to see that whenever the prior distribution of a stochastic process is that of a homogeneous Markov process with rate matrix $\mathbb{Q}$, then the posterior $P_{\mathbb{Q}}(\cdot|e_0, e_T)$ is also a Markov process, albeit generally not a homogeneous one. The distribution of a continuous-time Markov processes that is not homogeneous in time is determined by conditional transition probabilities, $p_{x,y}(s, s+t)$, which depend explicitly on both initial and final times. These transition probabilities can be expressed by means of a time-dependent matrix-valued function, $\mathbb{R}(t)$, which describes instantaneous transition rates. The connection between the time-dependent rate matrix $\mathbb{R}(t)$ and the transition probabilities, $p_{x,y}(s, s+t)$ is established by the master equation,

$$\frac{d}{dt} p_{x,y}(s, s+t) = \sum_z r_{z,y}(s+t) p_{x,z}(s, s+t),$$

where $r_{z,y}(t)$ are the entries of $\mathbb{R}(t)$. This equation is a generalization of Equation (2) for inhomogeneous processes. As in the homogeneous case, it leads to a master equation for the time-dependent probability distribution,

$$\frac{d}{dt} p_x(t) = \sum_y r_{y,x}(t) p_y(t),$$

thereby generalizing Equation (3).

By the above discussion, it follows that the posterior process can be represented by a time-dependent rate matrix $\mathbb{R}(t)$. More precisely, writing the posterior transition probability using basic properties of conditional probabilities and the definition of the Markov transition function gives

$$P_{\mathbb{Q}}(X^{(t+h)} = y | X^{(t)} = x, X^{(T)} = e_T) = \frac{p_{x,y}(h) p_{y,e_T}(T - t + h)}{p_{x,e_T}(T - t)}.$$

Taking the limit $h \to 0$ we obtain the instantaneous transition rate of the posterior process

$$r_{x,y}(t) = \lim_{h \to 0} \frac{P_{\mathbb{Q}}(X^{(t+h)} = y | X^{(t)} = x, X^{(T)} = e_T)}{h} = q_{x,y} \cdot \frac{p_{y,e_T}(T-t)}{p_{x,e_T}(T-t)}. \tag{9}$$

This representation, although natural, proves problematic in the framework of deterministic evidence because as t approaches T the transition rate into the observed state tends to infinity. In particular, when $x \neq e_T$ and $y = e_T$, the posterior transition rate is $q_{x,e_T} \cdot \frac{p_{e_T,e_T}(T-t)}{p_{x,e_T}(T-t)}$. This term diverges as $t \to T$, because the numerator approaches 1 while the denominator approaches 0. We therefore consider an alternative parameterization for this inhomogeneous process that is more suitable for variational approximations.

3.1 Marginal Density Representation

Let Pr be the distribution of a Markov process, generally not time homogeneous. We define a family of functions:

$$\begin{aligned} \mu_x(t) &= \Pr(X^{(t)} = x), \\ \gamma_{x,y}(t) &= \lim_{h \downarrow 0} \frac{\Pr(X^{(t)} = x, X^{(t+h)} = y)}{h}, \; x \neq y. \end{aligned} \tag{10}$$

The function $\mu_x(t)$ is the marginal probability that $X^{(t)} = x$. The function $\gamma_{x,y}(t)$ is the probability density that X transitions from state x to y at time t. Note that this parameter is not a transition rate, but rather a product of a point-wise probability with the point-wise transition rate of the distribution, that is, the entries of the time-dependent rate matrix of an equivalent process can be defined by

$$r_{x,y}(t) = \begin{cases} \frac{\gamma_{x,y}(t)}{\mu_x(t)} & \mu_x(t) > 0, \\ 0 & \mu_x(t) = 0. \end{cases} \tag{11}$$

Hence, unlike the (inhomogeneous) rate matrix at time t, $\gamma_{x,y}(t)$ takes into account the probability of being in state x and not only the rate of transitions.

We aim to use the family of functions μ and γ as a representation of the posterior process. To do so, we need to characterize the set of constraints that these functions satisfy. We begin by constraining the marginals $\mu_x(t)$ to be valid distributions that is, $0 \leq \mu_x(t) \leq 1$ and $\sum_x \mu_x(t) = 1$. A similar constraint on the pairwise distributions implies that $\gamma_{x,y}(t) \geq 0$ for $x \neq y$. Next, we infer additional constraints from consistency properties between distributions over pairs of variables and their uni-variate marginals. Specifically, Equation (10) implies that for $x \neq y$

$$\Pr(X^{(t)} = x, X^{(t+h)} = y) = \gamma_{x,y}(t)\, h + o(h). \tag{12}$$

Plugging this identity into the consistency constraint

$$\mu_x(t) = \Pr(X^{(t)} = x) = \sum_y \Pr(X^{(t)} = x, X^{(t+h)} = y),$$

defining

$$\gamma_{x,x}(t) = -\sum_{y \neq x} \gamma_{x,y}(t)$$

and rearranging, we obtain

$$\Pr(X^{(t)} = x, X^{(t+h)} = y) = \mathbf{1}_{x=y}\mu_x(t) + \gamma_{x,y}(t)\, h + o(h), \tag{13}$$

which unlike (12) is valid for all x, y. Marginalizing (13) with respect to the second variable,

$$\Pr(X^{(t+h)} = x) = \sum_y \Pr(X^{(t)} = y, X^{(t+h)} = x),$$

we obtain a forward update rule for the uni-variate marginals

$$\mu_x(t+h) = \mu_x(t) + h \sum_y \gamma_{y,x}(t) + o(h).$$

Rearranging terms and taking the limit $h \to 0$ gives a differential equation for $\mu_x(t)$,

$$\frac{d}{dt}\mu_x(t) = \sum_y \gamma_{y,x}(t).$$

Finally, whenever $\mu_x(t) = 0$ we have $\Pr(X^{(t)} = x, X^{(t+h)} = y) = 0$, implying in that case that $\gamma_{x,y}(t) = 0$. Based on these observations we define:

Definition 1 A family $\eta = \{\mu_x(t), \gamma_{x,y}(t) : 0 \leq t \leq T\}$ of functions is a *Markov-consistent density set* if the following constraints are fulfilled:

$$\begin{aligned}
\mu_x(t) &\geq 0, \quad \sum_x \mu_x(0) = 1, \\
\gamma_{x,y}(t) &\geq 0 \qquad \forall y \neq x, \\
\gamma_{x,x}(t) &= -\sum_{y \neq x} \gamma_{x,y}(t), \\
\frac{d}{dt}\mu_x(t) &= \sum_y \gamma_{y,x}(t),
\end{aligned}$$

and $\gamma_{x,y}(t) = 0$ whenever $\mu_x(t) = 0$. We denote by $\mathcal{M}$ the set of all Markov-consistent densities. ∎

Using standard arguments we can show that there exists a correspondence between (generally inhomogeneous) Markov processes and density sets η. Specifically, given η, we construct a process by defining an inhomogeneous rate matrix $\mathbb{R}(t)$ whose entries are defined in Equation (11) and prove the following:

Lemma 2 *Let $\eta = \{\mu_x(t), \gamma_{x,y}(t) : 0 \leq t \leq T\}$. If $\eta \in \mathcal{M}$, then there exists a continuous-time Markov process* Pr *for which μ_x and $\gamma_{x,y}$ satisfy (10) for every t in the right-open interval [0,T).*

Proof See appendix B ■

The converse is also true: for every integrable inhomogeneous rate matrix $\mathbb{R}(t)$ the corresponding marginal density set is defined by $\frac{d}{dt}\mu_x(t) = \sum_y r_{y,x}(t)\mu_y(t)$ and $\gamma_{x,y}(t) = \mu_x(t) r_{x,y}(t)$. The processes we are interested in, however, have additional structure, as they correspond to the posterior distribution of a time-homogeneous process with end-point evidence. In that case, multiplying Equation (9) by $\mu_x(t)$ gives

$$\gamma_{x,y}(t) = \mu_x(t) \cdot q_{x,y} \cdot \frac{p_{y,e_T}(T-t)}{p_{x,e_T}(T-t)}. \tag{14}$$

Plugging in Equation (8) we obtain

$$\gamma_{x,y}(t) = \frac{p_{e_0,x}(t) \cdot q_{x,y} \cdot p_{y,e_T}(T-t)}{p_{e_0,e_T(T)}},$$

which is zero when $y \neq e_T$ and $t = T$. This additional structure implies that we should only consider a subset of $\mathcal{M}$. Specifically the representation η corresponding to the posterior distribution $P_{\mathbb{Q}}(\cdot|e_0, e_T)$ satisfies $\mu_x(0) = \mathbf{1}_{x=e_0}$, $\mu_x(T) = \mathbf{1}_{x=e_T}$, $\gamma_{x,y}(0) = 0$ for all $x \neq e_0$ and $\gamma_{x,y}(T) = 0$ for all $y \neq e_T$. We denote by $\mathcal{M}_e \subset \mathcal{M}$ the subset that contains Markov-consistent density sets satisfying these constraints. This analysis suggests that for every homogeneous rate matrix and point evidence e there is a member in $\mathcal{M}_e$ that corresponds to the posterior process. Thus, from now on we restrict our attention to density sets from $\mathcal{M}_e$.

3.2 Variational Principle

The marginal density representation allows us to state the variational principle for continuous processes, which closely tracks similar principles for discrete processes. Specifically, we define a functional of functions that are constrained to be density sets from $\mathcal{M}_e$. The maximum over this set is the log-likelihood of the evidence and is attained for a density set that represents the posterior distribution. This formulation will serve as a basis for the mean-field approximation, which is introduced in the next section.

Define a *free energy functional,*

$$\mathcal{F}(\eta;\mathbb{Q}) = \mathcal{E}(\eta;\mathbb{Q}) + \mathcal{H}(\eta),$$

which, as we will see, measures the quality of η as an approximation of $P_{\mathbb{Q}}(\cdot|e)$. (For succinctness, we will assume that the evidence e is clear from the context.) The two terms in the functional are the *average energy,*

$$\mathcal{E}(\eta;\mathbb{Q}) = \int_0^T \sum_x \left[\mu_x(t) q_{x,x} + \sum_{y\neq x} \gamma_{x,y}(t) \ln q_{x,y} \right] dt,$$

and the *entropy,*

$$\mathcal{H}(\eta) = \int_0^T \sum_x \sum_{y\neq x} \gamma_{x,y}(t)[1 + \ln \mu_x(t) - \ln \gamma_{x,y}(t)]dt.$$

The following theorem establishes the relation of this functional to the Kullback-Leibler (KL) divergence and the likelihood of the evidence, and thus allows us to cast the variational inference into an optimization problem.

Theorem 3 *Let $\mathbb{Q}$ be a rate matrix, $e = (e_0, e_T)$ be states of X, and $\eta \in \mathcal{M}_e$. Then*

$$\mathcal{F}(\eta;\mathbb{Q}) = \ln P_{\mathbb{Q}}(e_T|e_0) - \boldsymbol{D}(P_\eta \| P_{\mathbb{Q}}(\cdot|e))$$

where P_η is the distribution corresponding to η and $\boldsymbol{D}(P_\eta \| P_{\mathbb{Q}}(\cdot|e))$ is the KL divergence between the two processes.

We conclude from the non-negativity of the KL divergence that the energy functional $\mathcal{F}(\eta;\mathbb{Q})$ is a lower bound of the log-likelihood of the evidence. The closer the approximation to the target posterior, the tighter the bound. Moreover, since the KL divergence is zero if and only if the two distributions are equal almost everywhere, finding the maximizer of this free energy is equivalent to finding the posterior distribution from which answers to different queries can be efficiently computed.

3.3 Proof of Theorem 3

We begin by examining properties of distributions of inhomogeneous Markov processes. Let $X^{(t)}$ be an inhomogeneous Markov process with rate matrix $\mathbb{R}(t)$. As in the homogeneous case, a trajectory σ of $\{X^{(t)}\}_{t\geq 0}$ over a time interval $[0,T]$ can be characterized by a finite number of transitions K, a sequence of states $(x_0, x_1, \ldots, x_K)$ and a sequence of transition times $(t_0 = 0, t_1, \ldots, t_K, t_{K+1} = T)$. We denote by Σ the set of all trajectories of $X^{[0,T]}$. The distribution over Σ can be characterized by a collection of random variables that consists of the number of transitions κ, a sequence of states $(\chi_0, \ldots, \chi_\kappa)$ and transition times $(\tau_1, \ldots, \tau_\kappa)$. Note that the number of random variables that characterize the trajectory is by itself a random variable. The density $f_{\mathbb{R}}$ of a trajectory $\sigma = \{K, x_0, \ldots, x_K, t_1, \ldots, t_K\}$ is the derivative of the joint distribution with respect to transition times, that is,

$$f_{\mathbb{R}}(\sigma) = \frac{\partial^K}{\partial t_1 \cdots \partial t_K} P_{\mathbb{R}}(\kappa = K, \chi_0 = x_0, \ldots, \chi_K = x_K, \tau_1 \leq t_1, \ldots, \tau_K \leq t_K),$$

which is given by

$$f_{\mathbb{R}}(\sigma) = p_{x_0}(0) \cdot \prod_{k=0}^{K-1} \left[e^{\int_{t_k}^{t_{k+1}} r_{x_k,x_k}(t)dt} r_{x_k,x_{k+1}}(t_{k+1}) \right] \cdot e^{\int_{t_K}^{t_{K+1}} r_{x_K,x_K}(t)dt}.$$

For example, in case $\mathbb{R}(t) = \mathbb{Q}$ is a homogeneous rate matrix this equation reduces to

$$f_{\mathbb{Q}}(\sigma) = p_{x_0}(0) \cdot \prod_{k=0}^{K-1} \left[e^{q_{x_k,x_k}(t_{k+1}-t_k)} q_{x_k,x_{k+1}} \right] \cdot e^{q_{x_K,x_K}(t_{K+1}-t_K)}.$$

The expectation of a random variable $\psi(\sigma)$ is an infinite sum (because one has to account for all possible numbers of transitions) of finite dimensional integrals,

$$\mathbf{E}_{f_{\mathbb{Q}}}[\psi] \equiv \int_{\Sigma} f_{\mathbb{R}}(\sigma)\psi(\sigma)d\sigma \equiv \sum_{K=0}^{\infty} \sum_{x_0} \cdots \sum_{x_K} \int_0^T \int_0^{t_K} \cdots \int_0^{t_2} f_{\mathbb{R}}(\sigma)\psi(\sigma)dt_1 \cdots dt_K.$$

The *KL-divergence* between two densities that correspond to two inhomogeneous Markov processes with rate matrices $\mathbb{R}(t)$ and $\mathbb{S}(t)$ is

$$\boldsymbol{D}(f_{\mathbb{R}} \| f_{\mathbb{S}}) = \int_{\Sigma} f_{\mathbb{R}}(\sigma) \ln \frac{f_{\mathbb{R}}(\sigma)}{f_{\mathbb{S}}(\sigma)} d\sigma \ . \tag{15}$$

We will use the convention $0 \ln 0 = 0$ and assume the support of $f_{\mathbb{S}}$ is contained in the support of $f_{\mathbb{R}}$. That is $f_{\mathbb{R}}(\sigma) = 0$ whenever $f_{\mathbb{S}}(\sigma) = 0$. The KL-divergence satisfies $\boldsymbol{D}(f_{\mathbb{R}} \| f_{\mathbb{S}}) \geq 0$ and is exactly zero if and only if $f_{\mathbb{R}} = f_{\mathbb{S}}$ almost everywhere (Kullback and Leibler, 1951).

Let $\eta \in \mathcal{M}_e$ be a marginal density set consistent with e. As we have seen, this density set corresponds to a Markov process with rate matrix $\mathbb{R}(t)$ whose entries are defined by Equation (11), hence we identify $f_\eta \equiv f_{\mathbb{R}}$.

Given evidence e on some event we denote $f_{\mathbb{Q}}(\sigma, e) \equiv f_{\mathbb{Q}}(\sigma) \cdot \boldsymbol{1}_{\sigma \models e}$, and note that

$$P_{\mathbb{Q}}(e) = \int_{\{\sigma:\sigma \models e\}} f_{\mathbb{Q}}(\sigma)d\sigma = \int_{\Sigma} f_{\mathbb{Q}}(\sigma, e)d\sigma \ ,$$

where $\sigma \models e$ is a predicate which is true if σ is consistent with the evidence. The density function of the posterior distribution $P_{\mathbb{Q}}(\cdot|e)$ satisfies $f_{\mathbb{S}}(\sigma) = \frac{f_{\mathbb{Q}}(\sigma,e)}{P_{\mathbb{Q}}(e)}$ where $\mathbb{S}(t)$ is the time-dependent rate matrix that corresponds to the posterior process.

Manipulating (15), we get

$$\boldsymbol{D}(f_\eta \| f_{\mathbb{S}}) = \int_{\Sigma} f_\eta(\sigma) \ln f_\eta(\sigma)d\sigma - \int_{\Sigma} f_\eta(\sigma) \ln f_{\mathbb{S}}(\sigma)d\sigma \equiv \mathbf{E}_{f_\eta}[\ln f_\eta(\sigma)] - \mathbf{E}_{f_\eta}[\ln f_{\mathbb{S}}(\sigma)].$$

Replacing $\ln f_{\mathbb{S}}(\sigma)$ by $\ln f_{\mathbb{Q}}(\sigma, e) - \ln P_{\mathbb{Q}}(e)$ and applying simple arithmetic operations gives

$$\ln P_{\mathbb{Q}}(e) = \mathbf{E}_{f_\eta}[\ln f_{\mathbb{Q}}(\sigma, e)] - \mathbf{E}_{f_\eta}[\ln f_\eta(\sigma)] + \boldsymbol{D}(f_\eta \| f_{\mathbb{S}}).$$

The crux of the proof is in showing that the expectations in the right-hand side satisfy

$$\mathbf{E}_{f_\eta}[\ln f_{\mathbb{Q}}(\sigma, e)] = \mathcal{E}(\eta; \mathbb{Q}),$$

and

$$-\mathbf{E}_{f_\eta}\left[\ln f_\eta(\sigma)\right] = \mathcal{H}(\eta),$$

implying that $\mathcal{F}(\eta;\mathbb{Q})$ is a lower bound on the log-probability of evidence with equality if and only if $f_\eta = f_{\mathbb{Q}}$ almost everywhere.

To prove these identities for the energy and entropy, we treat trajectories as functions $\sigma : \mathcal{R} \to \mathcal{R}$ where $\mathcal{R}$ is the set of real numbers by denoting $\sigma(t) \equiv X^{(t)}(\sigma)$—the state of the system at time t. Using this notation we introduce two lemmas that allow us to replace integration over a set of trajectories by a one dimensional integral, which is defined over a time variable. The first result handles expectations of functions that depend on specific states:

Lemma 4 *Let $\psi : S \times \mathcal{R} \to \mathcal{R}$ be a function, then*

$$\mathbf{E}_{f_\eta}\left[\int_0^T \psi(\sigma(t),t)dt\right] = \int_0^T \sum_x \mu_x(t)\psi(x,t)dt.$$

Proof See Appendix C.1 ■

As an example, by setting $\psi(x',t) = \mathbf{1}_{x'=x}$ we obtain that the expected residence time in state x is $\mathbf{E}_{f_\eta}[T_x] = \int_0^T \mu_x(t)dt$. The second result handles expectations of functions that depend on transitions between states:

Lemma 5 *Let $\psi(x,y,t)$ be a function from $S \times S \times \mathcal{R}$ to $\mathcal{R}$ that is continuous with respect to t and satisfies $\psi(x,x,t) = 0$, $\forall x, \forall t$ then*

$$\mathbf{E}_{f_\eta}\left[\sum_{k=1}^{K^\sigma} \psi(x_{k-1}^\sigma, x_k^\sigma, t_k^\sigma)\right] = \int_0^T \sum_x \sum_{y \neq x} \gamma_{x,y}(t)\psi(x,y,t)dt,$$

where the superscript σ stresses that K^σ, x_k^σ and t_k^σ are associated with a specific trajectory σ.

Proof See Appendix C.2 ■

Continuing the example of the previous lemma, here by setting $\psi(x',y',t) = \mathbf{1}_{x'=x}\mathbf{1}_{y'=y}\mathbf{1}_{x\neq y}$ the sums within the left hand expectation become the number of transitions in a trajectory σ. Thus, we obtain that the expected number of transitions from x to y is $\mathbf{E}_f[M_{x,y}] = \int_0^T \gamma_{x,y}(t)dt$.

We now use these lemmas to compute the expectations involved in the energy functional. Suppose $e = \{e_0, e_T\}$ is a pair of point evidence and $\eta \in \mathcal{M}_e$. Applying these lemmas with $\psi(x,t) = q_{x,x}$ and $\psi(x,y,t) = \mathbf{1}_{x\neq y} \cdot \ln q_{x,y}$ gives

$$\mathbf{E}_{f_\eta}\left[\ln f_{\mathbb{Q}}(\sigma,e)\right] = \int_0^T \sum_x \left[\mu_x(t) q_{x,x}(t) + \sum_{y\neq x} \gamma_{x,y}(t) \ln q_{x,y}(t)\right] dt\ .$$

Similarly, setting $\psi(x,t) = r_{x,x}(t)$ and $\psi(x,y,t) = \mathbf{1}_{x\neq y} \cdot \ln r_{x,y}(t)$, where $\mathbb{R}(t)$ is defined in Equation (11), we obtain

$$-\mathbf{E}_{f_\eta}\left[\ln f_\eta(\sigma,e)\right] = -\int_0^T \sum_x \left[\mu_x(t)\frac{\gamma_{x,x}(t)}{\mu_x(t)} + \sum_{y\neq x} \gamma_{x,y}(t) \ln \frac{\gamma_{x,y}(t)}{\mu_x(t)}\right] dt = \mathcal{H}(\eta)\ .$$

4. Factored Approximation

The variational principle we discussed is based on a representation that is as complex as the original process—the number of functions $\gamma_{x,y}(t)$ we consider is equal to the size of the original rate matrix $\mathbb{Q}$. To get a tractable inference procedure we make additional simplifying assumptions on the approximating distribution.

Given a D-component process we consider approximations that factor into products of independent processes. More precisely, we define $\mathcal{M}_e^i$ to be the continuous Markov-consistent density sets over the component X_i, that are consistent with the evidence on X_i at times 0 and T. Given a collection of density sets $\eta^1,\ldots,\eta^D$ for the different components, the product density set $\eta = \eta^1 \times \cdots \times \eta^D$ is defined as

$$\begin{aligned}
\mu_x(t) &= \prod_i \mu^i_{x_i}(t), \\
\gamma_{x,y}(t) &= \begin{cases} \gamma^i_{x_i,y_i}(t)\mu_x^{\setminus i}(t) & \delta(x,y)=\{i\} \\ \sum_i \gamma^i_{x_i,x_i}(t)\mu_x^{\setminus i}(t) & x=y \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

where $\mu_x^{\setminus i}(t) = \prod_{j\neq i}\mu^j_{x_j}(t)$ is the joint distribution at time t of all the components other than the i'th (it is not hard to see that if $\eta^i \in \mathcal{M}_e^i$ for all i, then $\eta \in \mathcal{M}_e$). We define the set $\mathcal{M}_e^F$ to contain all factored density sets. From now on we assume that $\eta = \eta^1 \times \cdots \times \eta^D \in \mathcal{M}_e^F$.

Assuming that $\mathbb{Q}$ is defined by a CTBN, and that η is a factored density set, we can rewrite

$$\mathcal{E}(\eta;\mathbb{Q}) = \sum_i \int_0^T \sum_{x_i} \left[\mu^i_{x_i}(t)\mathbf{E}_{\mu^{\setminus i}(t)}\left[q_{x_i,x_i|U_i}\right] + \sum_{x_i,y_i\neq x_i} \gamma^i_{x_i,y_i}(t)\mathbf{E}_{\mu^{\setminus i}(t)}\left[\ln q_{x_i,y_i|U_i}\right]\right] dt$$

(see derivations in Appendix D). Similarly, the entropy term factors as

$$\mathcal{H}(\eta) = \sum_i \mathcal{H}(\eta^i) \ \ .$$

Note that terms such as $\mathbf{E}_{\mu^{\setminus i}(t)}\left[q_{x_i,x_i|U_i}\right]$ involve only $\mu^j(t)$ for $j \in \mathbf{Pa}_i$, because $\mathbf{E}_{\mu^{\setminus i}(t)}\left[f(U_i)\right] = \sum_{u_i}\mu_{u_i}(t)f(u_i)$. Therefore, this decomposition involves only local terms that either include the i'th component, or include the i'th component and its parents in the CTBN defining $\mathbb{Q}$.

To make the factored nature of the approximation explicit in the notation, we write henceforth,

$$\mathcal{F}(\eta;\mathbb{Q}) = \tilde{\mathcal{F}}(\eta^1,\ldots,\eta^D;\mathbb{Q}).$$

4.1 Fixed Point Characterization

The factored form of the functional and the independence between the different η^i allows optimization by *block ascent*, optimizing the functional with respect to each parameter set in turn. To do so, we should solve the following optimization problem:

> Fixing i, and given $\eta^1,\ldots,\eta^{i-1},\eta^{i+1},\ldots,\eta^D$, in $\mathcal{M}_e^1,\ldots\mathcal{M}_e^{i-1},\mathcal{M}_e^{i+1},\ldots,\mathcal{M}_e^D$, respectively, find
>
> $$\arg\max_{\eta^i\in\mathcal{M}_e^i} \tilde{\mathcal{F}}(\eta^1,\ldots,\eta^D;\mathbb{Q}) \ \ .$$

If for all i, we have a $\mu^i \in \mathcal{M}_e^i$, which is a solution to this optimization problem with respect to each component, then we have a (local) stationary point of the energy functional within $\mathcal{M}_e^F$.

To solve this optimization problem, we define a Lagrangian, which includes the constraints in the form of Definition 1. These constraints are to be enforced in a continuous fashion, and so the Lagrange multipliers $\lambda^i_{x_i}(t)$ are continuous functions of t as well. The Lagrangian is a functional of the functions $\mu^i_{x_i}(t)$, $\gamma^i_{x_i,y_i}(t)$ and $\lambda^i_{x_i}(t)$, and takes the following form

$$\mathcal{L} = \tilde{\mathcal{F}}(\eta;\mathbb{Q}) - \sum_{i=1}^{D} \int_0^T \lambda^i_{x_i}(t) \left(\frac{d}{dt}\mu^i_{x_i}(t) - \sum_{y_i} \gamma^i_{x_i,y_i}(t) \right) dt \quad .$$

A necessary condition for the optimality of a density set η is the existence of λ such that (η,λ) is a *stationary point* of the Lagrangian. A stationary point of a functional satisfies the *Euler-Lagrange* equations, namely the *functional derivatives* with respect to μ, γ and λ vanish (see Appendix E for a brief review). The detailed derivation of the resulting equations is in Appendix F. Writing these equations in explicit form, we get a fixed point characterization of the solution in term of the following set of ODEs:

$$\begin{aligned} \frac{d}{dt}\mu^i_{x_i}(t) &= \sum_{y_i \neq x_i} \left(\gamma^i_{y_i,x_i}(t) - \gamma^i_{x_i,y_i}(t) \right), \\ \frac{d}{dt}\rho^i_{x_i}(t) &= -\rho^i_{x_i}(t) \left(\bar{q}^{\,i}_{x_i,x_i}(t) + \psi^i_{x_i}(t) \right) - \sum_{y_i \neq x_i} \rho^i_{y_i}(t) \tilde{q}^i_{x_i,y_i}(t) \end{aligned} \tag{16}$$

along with the following algebraic constraint

$$\rho^i_{x_i}(t)\gamma^i_{x_i,y_i}(t) = \mu^i_{x_i}(t)\tilde{q}^i_{x_i,y_i}(t)\rho^i_{y_i}(t),\ x_i \neq y_i \tag{17}$$

where ρ^i are the exponents of the Lagrange multipliers λ_i. In these equations we use the following shorthand notations for the average rates

$$\begin{aligned} \bar{q}^{\,i}_{x_i,x_i}(t) &= \mathbf{E}_{\mu^{\setminus i}(t)} \left[q^{i|\mathbf{Pa}_i}_{x_i,x_i|U_i} \right], \\ \bar{q}^{\,i}_{x_i,x_i|x_j}(t) &= \mathbf{E}_{\mu^{\setminus i}(t)} \left[q^{i|\mathbf{Pa}_i}_{x_i,x_i|U_i} \mid x_j \right], \end{aligned}$$

where $\mu^{\setminus i}(t)$ is the product distribution of $\mu^1(t),\ldots,\mu^{i-1}(t),\mu^{i+1}(t),\ldots,\mu^D(t)$. Similarly, we have the following shorthand notations for the geometrically-averaged rates,

$$\begin{aligned} \tilde{q}^i_{x_i,y_i}(t) &= \exp\left\{ \mathbf{E}_{\mu^{\setminus i}(t)} \left[\ln q^{i|\mathbf{Pa}_i}_{x_i,y_i|U_i} \right] \right\}, \\ \tilde{q}^i_{x_i,y_i|x_j}(t) &= \exp\left\{ \mathbf{E}_{\mu^{\setminus i}(t)} \left[\ln q^{i|\mathbf{Pa}_i}_{x_i,y_i|U_i} \mid x_j \right] \right\} \quad . \end{aligned}$$

The last auxiliary term is

$$\psi^i_{x_i}(t) = \sum_{j \in Children_i} \sum_{x_j} \left[\mu^j_{x_j}(t)\bar{q}^{\,j}_{x_j,x_j|x_i}(t) + \sum_{x_j \neq y_j} \gamma^j_{x_j,y_j}(t) \ln \tilde{q}^j_{x_j,y_j|x_i}(t) \right] \quad .$$

To uniquely solve the two differential Equations (16) for $\mu^i_{x_i}(t)$ and $\rho^i_{x_i}(t)$ we need to set boundary conditions. The boundary condition for $\mu^i_{x_i}$ is defined explicitly in $\mathcal{M}_e^F$ as

$$\mu^i_{x_i}(0) = \mathbf{1}_{x_i = e_{i,0}} \ . \tag{18}$$

The boundary condition at T is slightly more involved. The constraints in $\mathcal{M}_e^F$ imply that $\mu^i_{x_i}(T) = \mathbf{1}_{x_i=e_{i,T}}$. As stated in Section 3.1, we have that $\gamma^i_{e_{i,T},x_i}(T) = 0$ when $x_i \neq e_{i,T}$. Plugging these values into (17), and assuming that all elements of $\mathbb{Q}^{i|\mathbf{Pa}_i}$ are non-zero we get that $\rho_{x_i}(T) = 0$ for all $x_i \neq e_{i,T}$ (It might be possible to use a weaker condition that $\mathbb{Q}$ is irreducible). In addition, we notice that $\rho_{e_{i,T}}(T) \neq 0$, for otherwise the whole system of equations for ρ will collapse to 0. Finally, notice that the solution of (16,17) for μ^i and γ^i is insensitive to the multiplication of ρ^i by a constant. Thus, we can arbitrarily set $\rho_{e_{i,T}}(T) = 1$, and get the boundary condition

$$\rho^i_{x_i}(T) = \mathbf{1}_{x_i=e_{i,T}}. \tag{19}$$

Putting it all together we obtain a characterization of stationary points of the functional as stated in the following theorem:

Theorem 6 *$\eta^i \in \mathcal{M}_e^i$ is a stationary point (e.g., local maxima) of $\tilde{\mathcal{F}}(\eta^1,\ldots,\eta^D;\mathbb{Q})$ subject to the constraints of Definition 1 if and only if it satisfies (16–19).*

Proof see Appendix F ■

It is straightforward to extend this result to show that at a maximum with respect to all the component densities, this fixed-point characterization must hold for all components simultaneously.

Example 2 Consider the case of a single component, for which our procedure should be exact, as no simplifying assumptions are made on the density set. In that case, the averaged rates $\bar{q}^{\,i}$ and the geometrically-averaged rates $\tilde{q}^i$ both reduce to the unaveraged rates q, and $\psi \equiv 0$. Thus, the system of equations to be solved is

$$\frac{d}{dt}\mu_x(t) = \sum_{y \neq x}(\gamma_{y,x}(t) - \gamma_{x,y}(t)),$$
$$\frac{d}{dt}\rho_x(t) = -\sum_y q_{x,y}\rho_y(t),$$

along with the algebraic equation

$$\rho_x(t)\gamma_{x,y}(t) = \mu_x(t)q_{x,y}\rho_y(t), \qquad y \neq x.$$

These equations have a simple intuitive interpretation. First, the backward propagation rule for ρ_x implies that

$$\rho_x(t) = \Pr(e_T|X^{(t)} = x).$$

To prove this identity, we recall the notation $p_{x,y}(h) \equiv \Pr(X^{(t+h)} = y|X^{(t)} = x)$ and write the discretized propagation rule

$$\Pr(e_T|X^{(t)} = x) = \sum_y p_{x,y}(h) \cdot \Pr(e_T|X^{(t+h)} = y) \ .$$

Using the definition of q (Equation 1), rearranging, dividing by h and taking the limit $h \to 0$ gives

$$\frac{d}{dt}\Pr(e_T|X^{(t)} = x) = -\sum_y q_{x,y} \cdot \Pr(e_T|X^{(t)} = y),$$

which is identical to the differential equation for ρ. Second, dividing the above algebraic equation by $\rho_x(t)$ whenever it is greater than zero we obtain

$$\gamma_{x,y}(t) = \mu_x(t) q_{x,y} \frac{\rho_y(t)}{\rho_x(t)}. \tag{20}$$

Thus, we reconstructed Equation (14).

This analysis suggest that this system of ODEs is similar to forward-backward propagation, except that unlike classical forward propagation, here the forward propagation already takes into account the backward messages to directly compute the posterior. Given this interpretation, it is clear that integrating $\rho_x(t)$ from T to 0 followed by integrating $\mu_x(t)$ from 0 to T computes the exact posterior of the processes.

This interpretation of $\rho_x(t)$ also allows us to understand the role of $\gamma_{x,y}(t)$. Equation (20) suggests that the instantaneous rate combines the original rate with the relative likelihood of the evidence at T given y and x. If y is much more likely to lead to the final state, then the rates are biased toward y. Conversely, if y is unlikely to lead to the evidence the rate of transitions to it are lower. This observation also explains why the forward propagation of μ_x will reach the observed $\mu_x(T)$ even though we did not impose it explicitly. ▮

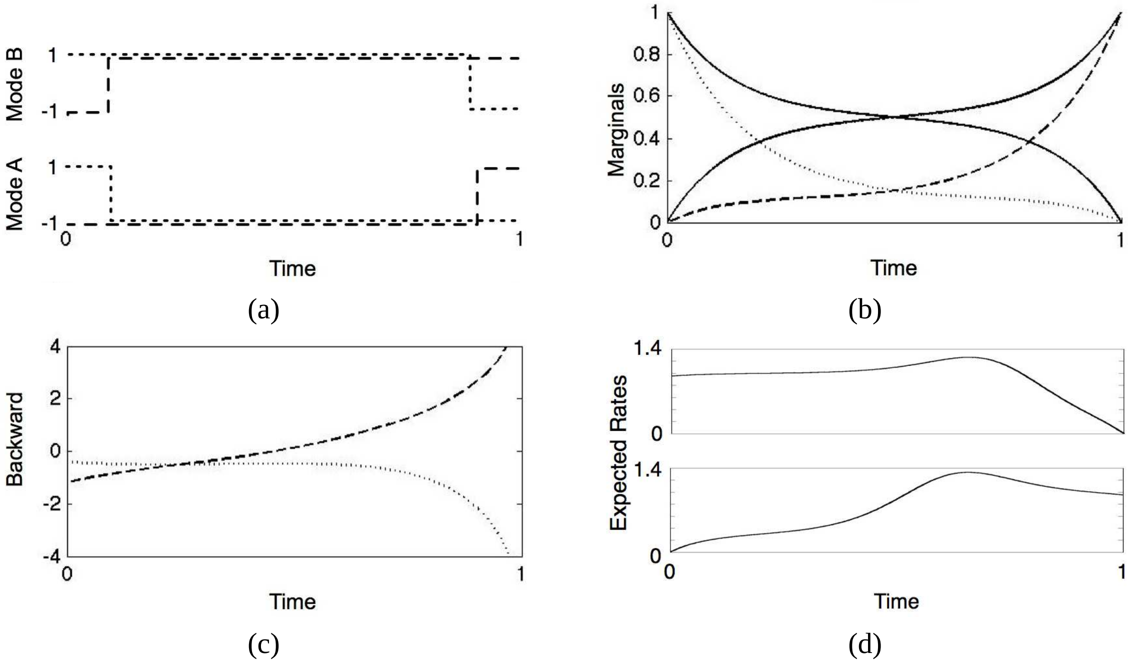
Example 3 Let us return to the two-component Ising chain in Example 1 with initial state $X_1^{(0)} = -1$ and $X_2^{(0)} = 1$, and a reversed state at the final time, $X_1^{(T)} = 1$ and $X_2^{(T)} = -1$. For a large value of β, this evidence is unlikely as at both end points the components are in a undesired configurations. The exact posterior is one that assigns higher probabilities to trajectories where one of the components switches relatively fast to match the other, and then toward the end of the interval, they separate to match the evidence. Since the model is symmetric, these trajectories are either ones in which both components are most of the time in state -1, or ones where both are most of the time in state 1 (Figure 3(a)). Due to symmetry, the marginal probability of each component is around 0.5 throughout most of the interval. The variational approximation cannot capture the dependency between the two components, and thus converges to one of two local maxima, corresponding to the two potential subsets of trajectories (Figure 3(b)). Examining the value of ρ^i, we see that close to the end of the interval they bias the instantaneous rates significantly. For example, as t approaches 1, $\rho_1^1(t)/\rho_{-1}^1(t)$ approaches infinity and so does the instantaneous rate $\gamma_{-1,1}^1(t)/\mu_{-1}^1(t)$, thereby forcing X_1 to switch to state 1 (Figure 3(c)).

This example also allows to examine the implications of modeling the posterior by inhomogeneous Markov processes. In principle, we might have used as an approximation Markov processes with homogeneous rates, and conditioned on the evidence. To examine whether our approximation behaves in this manner, we notice that in the single component case we have

$$q_{x,y} = \frac{\rho_x(t)\gamma_{x,y}(t)}{\rho_y(t)\mu_x(t)},$$

which should be constant.

Consider the analogous quantity in the multi-component case: $\tilde{q}^i_{x_i,y_i}(t)$, the geometric average of the rate of X_i, given the probability of parents state. Not surprisingly, this is exactly a mean field approximation, where the influence of interacting components is approximated by their average influence. Since the distribution of the parents (in the two-component system, the other component)



(a) (b) (c) (d)

Figure 3: Numerical results for the two-component Ising chain described in Example 3 where the first component starts in state -1 and ends at time $T = 1$ in state 1. The second component has the opposite behavior. *(a)* Two likely trajectories depicting the two modes of the model. *(b)* Exact (solid) and approximate (dashed/dotted) marginals $\mu^i_1(t)$. *(c)* The log ratio $\log \rho^i_1(t)/\rho^i_{-1}(t)$. *(d)* The expected rates $\tilde{q}^1_{1,-1}(t)$ and $\tilde{q}^1_{-1,1}(t)$ of component X_1 of the Ising chain in Example 1. We can notice that the averaged rates are highly non-constant, and so cannot be approximated well with a constant rate matrix.

changes in time, these rates change continuously, especially near the end of the time interval. This suggests that a piecewise homogeneous approximation cannot capture the dynamics without a loss in accuracy. As expected in a dynamic process, we can see in Figure 3(d) that the inhomogeneous transition rates are very erratic. In particular, the rates of X_1 spike at the transition point selected by the mean field approximation. This can be interpreted as putting most of the weight of the distribution on trajectories which transition from state -1 to 1 at that point. ∎

4.2 Optimization Procedure

If $\mathbb{Q}$ is irreducible, then $\rho^i_{x_i}$ and $\mu^i_{x_i}$ are non-zero throughout the open interval $(0,T)$. As a result, we can solve (17) to express $\gamma^i_{x_i,y_i}$ as a function of μ^i and ρ^i, thus eliminating it from (16) to get evolution equations solely in terms of μ^i and ρ^i. Abstracting the details, we obtain a set of ODEs of the form

$$\frac{d}{dt}\rho^i(t) = \alpha(\rho^i(t), \mu^{\setminus i}(t)) \qquad \rho^i(T) = \text{given},$$
$$\frac{d}{dt}\mu^i(t) = \beta(\mu^i(t), \rho^i(t), \mu^{\setminus i}(t)) \quad \mu^i(0) = \text{given}.$$

where α and β are defined by the right-hand side of the differential equations (16). Since the evolution of ρ^i does not depend on μ^i, we can integrate backward from time T to solve for ρ^i. Then, integrating forward from time 0, we compute μ^i. After performing a single iteration of backward-forward integration, we obtain a solution that satisfies the fixed-point equation (16) for the i'th component (this is not surprising once we have identified our procedure to be a variation of a standard forward-backward algorithm for a single component). Such a solution will be a local maximum of the functional w.r.t. to η^i (reaching a local minimum or a saddle point requires very specific initialization points).

This suggests that we can use the standard procedure of asynchronous updates, where we update each component in a round-robin fashion. Since each of these single-component updates converges in one backward-forward step, and since it reaches a local maximum, each step improves the value of the free energy over the previous one. As the free energy functional is bounded by the probability of the evidence, this procedure will always converge, and the rate of the free energy increase can be used to test for convergence.

Potentially, there can be many scheduling possibilities. In our implementation the update scheduling is simply random. A better choice would be to update the component which would maximally increase the value of the functional in that iteration. This idea is similar to the scheduling of Elidan et al. (2006), who approximate the change in the beliefs by bounding the *residuals* of the messages, which give an approximation of the benefit of updating each component.

Another issue is the initialization of this procedure. Since the iteration on the i'th component depends on $\mu^{\setminus i}$, we need to initialize μ by some legal assignment. To do so, we create a fictional rate matrix $\tilde{\mathbb{Q}}_i$ for each component and initialize μ^i to be the posterior of the process given the evidence $e_{i,0}$ and $e_{i,T}$. As a reasonable initial guess, we choose at random one of the conditional rates $\mathbb{Q}^{i|u_i}$ using some random assignment u_i to determine the fictional rate matrix.

The general optimization procedure is summarized in the following algorithm:

```
For each i, initialize $\mu^i$ using some legal marginal function.
while not converged do
    1. Pick a component $i \in \{1,\ldots,D\}$.
    2. Update $\rho^i(t)$ by solving the $\rho^i$ backward differential equation in (16).
    3. Update $\mu^i(t)$ and $\gamma^i(t)$ by solving the $\mu^i$ forward differential equation in (16) and
       using the algebraic equation in (17).
end
```

Algorithm 1: Mean field approximation in continuous-time Bayesian networks

4.3 Exploiting Continuous-Time Representation

The continuous-time update equations allow us to use standard ODE methods with an adaptive step size (here we use the Runge-Kutta-Fehlberg (4,5) method). At the price of some overhead, these procedures automatically tune the trade-off between error and time granularity. Moreover, this overhead is usually negligible compared to the saving in computation time, because adaptive integration can be more efficient than *any* fixed step size integration by an order of magnitude (Press et al., 2007).

To further save computations, we note that while standard integration methods involve only initial boundary conditions at $t = 0$, the solution of μ^i is also known at $t = T$. Therefore, we stop the adaptive integration when $\mu^i(t) \approx \mu^i(T)$ and t is close enough to T. This modification reduces the number of computed points significantly because the derivative of μ^i tends to grow near the boundary, resulting in a smaller step size.

The adaptive solver selects different time points for the evaluation of each component. Therefore, updates of η^i require access to marginal density sets of neighboring components at time points that differ from their evaluation points. To allow efficient interpolation, we use a piecewise linear approximation of η whose boundary points are determined by the evaluation points that are chosen by the adaptive integrator.

5. Perspectives and Related Work

Variational approximations for different types of continuous-time processes have been recently proposed. Examples include systems with discrete hidden components (Opper and Sanguinetti, 2007); continuous-state processes (Archambeau et al., 2007); hybrid models involving both discrete and continuous-time components (Sanguinetti et al., 2009; Opper and Sanguinetti, 2010); and spatiotemporal processes (Ruttor and Opper, 2010; Dewar et al., 2010). All these models assume noisy observations in a finite number of time points. In this work we focus on structured discrete-state processes with noiseless evidence.

Our approach is motivated by results of Opper and Sanguinetti (2007) who developed a variational principle for a related model. Their model is similar to an HMM, in which the hidden chain is a continuous-time Markov process and there are (noisy) observations at discrete points along the process. They describe a variational principle and discuss the form of the functional when the approximation is a product of independent processes. There are two main differences between the setting of Opper and Sanguinetti and ours. First, we show how to exploit the structure of the target CTBN to reduce the complexity of the approximation. These simplifications imply that the update of the i'th process depends only on its Markov blanket in the CTBN, allowing us to develop efficient approximations for large models. Second, and more importantly, the structure of the evidence in our setting is quite different, as we assume deterministic evidence at the end of intervals. This setting typically leads to a posterior Markov process in which the instantaneous rates used by Opper and Sanguinetti diverge toward the end point—the rates of transition into the observed state go to infinity, leading to numerical problems at the end points. We circumvent this problem by using the marginal density representation which is much more stable numerically.

Taking the general perspective of Wainwright and Jordan (2008), the representation of the distribution uses the natural sufficient statistics. In the case of a continuous-time Markov process, the sufficient statistics are T_x, the time spent in state x, and $M_{x,y}$, the number of transitions from state x to y. In a discrete-time model, we can capture the statistics for every random variable. In a continuous-time model, however, we need to consider the time derivative of the statistics. Indeed, as shown in Section 3.3 we have

$$\frac{d}{dt}\mathbf{E}\left[T_x(t)\right] = \mu_x(t) \quad \text{and} \quad \frac{d}{dt}\mathbf{E}\left[M_{x,y}(t)\right] \quad = \gamma_{x,y}(t).$$

Thus, our marginal density sets η provide what we consider a natural formulation for variational approaches to continuous-time Markov processes.

Our presentation focused on evidence at two ends of an interval. Our formulation easily extends to deal with more elaborate types of evidence: (1) If we do not observe the initial state of the i'th component, we can set $\mu^i_x(0)$ to be the prior probability of $X^{(0)} = x$. Similarly, if we do not observe X_i at time T, we set $\rho^i_x(T) = 1$ as initial data for the backward step. (2) In a CTBN where one (or more) components are fully observed throughout some interval, we simply set μ^i for these components to be a distribution that assigns all the probability mass to the observed trajectory. Similarly, if we observe different components at different times, we may update each component on a different time interval. Consequently, maintaining for each component a marginal distribution μ^i throughout the interval of interest, we can update the other ones using their evidence patterns.

6. Evaluation on Ising Chains

To gain better insight into the quality of our procedure, we performed numerical tests on models that challenge the approximation. Specifically, we use Ising chains with the parameterization introduced in Example 1, where we explore regimes defined by the degree of coupling between the components (the parameter β) and the rate of transitions (the parameter τ). We evaluate the error in two ways. The first is by the difference between the true log-likelihood and our estimate. The second is by the average relative error in the estimate of different expected sufficient statistics defined by $\sum_j |\hat{\theta}_j - \theta_j|/\theta_j$, where θ_j is exact value of the j'th expected sufficient statistics and $\hat{\theta}_j$ is the approximation. To obtain a stable estimate the average is taken over all $\theta_j > 0.05 \max_{j'} \theta_{j'}$.

Applying our procedure on an Ising chain with 8 components, for which we can still perform exact inference, we evaluated the relative error for different choices of β and τ. The evidence in this experiment is $e_0 = \{+,+,+,+,+,+,-,-\}$, $T = 0.64$ and $e_T = \{-,-,-,+,+,+,+,+\}$. As shown in Figure 4(a), the error is larger when τ and β are large. In the case of a weak coupling (small β), the posterior is almost factored, and our approximation is accurate. In models with few transitions (small τ), most of the mass of the posterior is concentrated on a few canonical "types" of trajectories that can be captured by the approximation (as in Example 3). At high transition rates, the components tend to transition often, and in a coordinated manner, which leads to a posterior that is hard to approximate by a product distribution. Moreover, the resulting free energy landscape is rough with many local maxima. Examining the error in likelihood estimates (Figure 4(b),(c)) we see a similar trend.

Next, we examine the run time of our approximation when using fairly standard ODE solver with few optimizations and tunings. The run time is dominated by the time needed to perform the backward-forward integration when updating a single component, and by the number of such updates until convergence. Examining the run time for different choices of β and τ (Figure 5), we see that the run time of our procedure scales linearly with the number of components in the chain. The differences among the different curves suggest that the runtime is affected by the choice of parameters, which in turn affect the smoothness of the posterior density sets.

7. Evaluation on Branching Processes

The above-mentioned experimental results indicate that our approximation is accurate when reasoning about weakly-coupled components, or about time intervals involving few transitions (low transition rates). Unfortunately, in many domains we face strongly-coupled components. For example, we are interested in modeling the evolution of biological sequences (DNA, RNA, and proteins).

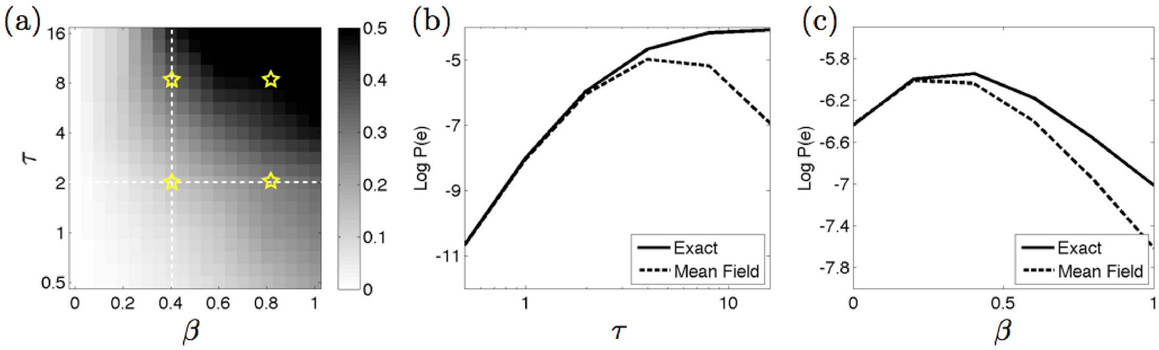


Figure 4: *(a)* Relative error as a function of the coupling parameter β (*x*-axis) and transition rates τ (*y*-axis) for an 8-component Ising chain. *(b)* Comparison of true vs. estimated likelihood as a function of the rate parameter τ. *(c)* Comparison of true vs. likelihood as a function of the coupling parameter β.

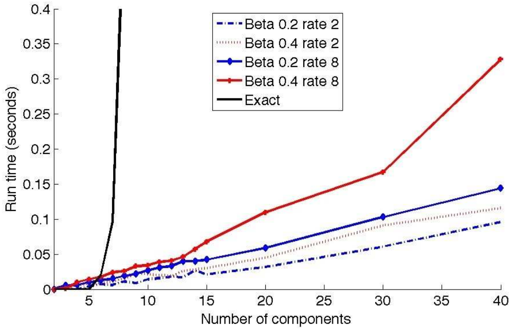


Figure 5: Evaluation of the run time of the approximation versus the run time of exact inference as a function of the number of components.

In such systems, we have a *phylogenetic tree* that represents the branching process that leads to current day sequences (see Figure 6).

It is common in sequence evolution to model this process as a continuous-time Markov process over a tree (Felsenstein, 2004). More precisely, the evolution along each branch is a standard continuous-time Markov process, and branching is modeled by a replication, after which each replica evolves independently along its sub-branch. Common applications are forced to assume that each character in the sequence evolves independently of the other.

In some situations, assuming an independent evolution of each character is highly unreasonable. Consider the evolution of an RNA sequence that folds onto itself to form a functional structure, as in Figure 7(a). This folding is mediated by complementary base-pairing (A-U, C-G, etc) that stabilizes the structure. During evolution, we expect to see compensatory mutations. That is, if a *A* changes into *C* then its based-paired *U* will change into a *G* soon thereafter. To capture such

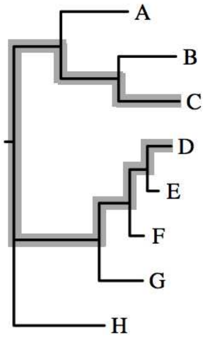


Figure 6: An example of a phylogenetic tree. Branch lengths denote time intervals between events. The interval used for the comparison with non-branching processes is highlighted.

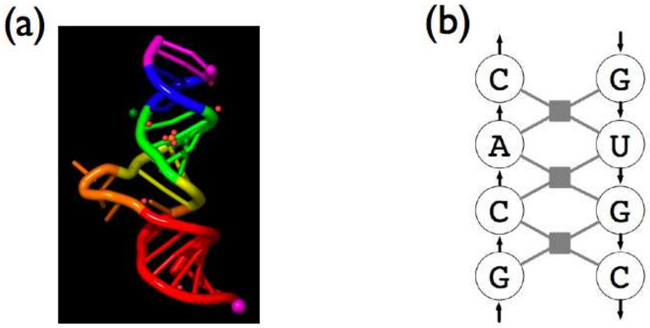


Figure 7: *(a)* Structure of an RNA molecule. The 3 dimensional structure dictates the dependencies between the different positions. *(b)* The form of the energy function for encoding RNA folding, superimposed on a fragment of a folded structure; each gray box denotes a term that involves four nucleotides.

coordinated changes, we need to consider the joint evolution of the different characters. In the case of RNA structure, the stability of the structure is determined by *stacking potentials* that measure the stability of two adjacent pairs of interacting nucleotides. Thus, if we consider a factor network to represent the energy of a fold, it will have structure as shown in Figure 7(b). We can convert this factor graph into a CTBN using procedures that consider the energy function as a fitness criteria in evolution (El-Hay et al., 2006; Yu and Thorne, 2006). Unfortunately, inference in such models suffers from computational blowup, and so the few studies that deal with it explicitly resort to sampling procedures (Yu and Thorne, 2006).

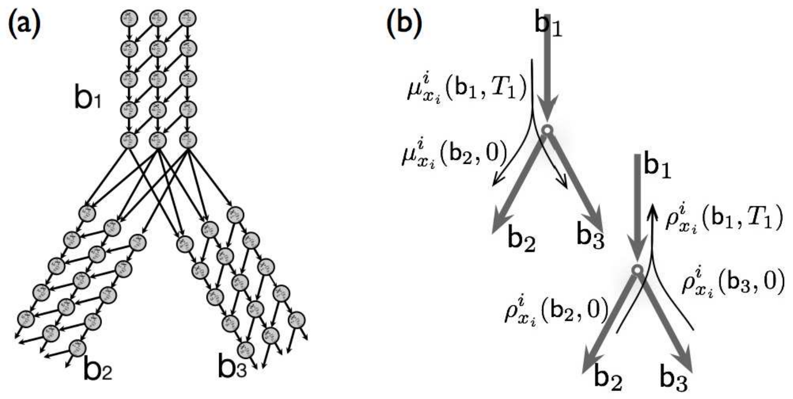

(a)
b1
b2
b3
(b)
b1
$\mu^i_{x_i}(\mathsf{b}_1, T_1)$
$\mu^i_{x_i}(\mathsf{b}_2, 0)$
b2
b3
b1
$\rho^i_{x_i}(\mathsf{b}_1, T_1)$
$\rho^i_{x_i}(\mathsf{b}_3, 0)$
$\rho^i_{x_i}(\mathsf{b}_2, 0)$
b2
b3

Figure 8: Structure of the branching process. *(a)* The discretized CTBN underlying the process in an intersection. *(b)* Illustration of the ODE updates on a directed tree, updating $\rho^i(t)$ backwards using (21) and $\mu^i(t)$ forwards using (22).

7.1 Representation

To consider phylogenetic trees, we should take a common approach in evolutionary analysis, in which inference of the tree topology and branch lengths is performed separately from inference of sequence dynamics. Thus, we need to extend our framework to deal with branching processes, where the branching points are fixed and known. In a linear-time model, we view the process as a map from $[0,T]$ into random variables $X^{(t)}$. In the case of a tree, we view the process as a map from a point $\mathsf{t} = \langle \mathsf{b}, t\rangle$ on a tree $\mathcal{T}$ (defined by branch b and the time t within it) into a random variable $X^{(\mathsf{t})}$. Similarly, we generalize the definition of the Markov-consistent density set η to include functions on trees. We define continuity of functions on trees in the obvious manner.

To gain intuition on this process we return to the discrete case, where our branching process can be viewed as a branching of the Dynamic Bayesian Network from one branch to two separate branches at the vertex, as in Figure 8(a).

7.2 Inference on Trees

The variational approximation on trees is thus similar to the one on intervals. Within each branch, we deal with the same update formulas as in linear time. We denote by $\mu^i_{x_i}(\mathsf{b}, t)$ and $\rho^i_{x_i}(\mathsf{b}, t)$ the messages computed on branch b at time t. The only changes occur at vertices, where we cannot use the Euler-Lagrange equations (Appendix E), therefore we must derive the propagation equations using a different method.

The following proposition establishes the update equations for the parameters $\mu^i(t)$ and $\rho^i(t)$ at the vertices, as depicted in Figure 8(b):

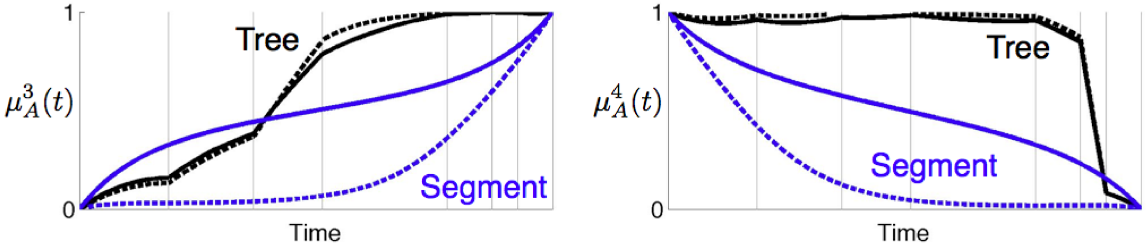


Figure 9: Comparison of exact vs. approximate inference along the highlighted path from C to D in the tree of Figure 6 with and without additional evidence at other leaves. In the latter case the problem is equivalent to inference on a linear segment. Exact marginals are shown in solid lines, whereas approximate marginals are in dashed lines. The horizontal gray lines indicate branch points along the path. Notice that evidence at the leaves result in discontinuities of the derivatives at such points. The two panels show two different components.

Proposition 7 *Given a vertex T with an incoming branch* b_1 *and two outgoing branches* $\mathsf{b}_2, \mathsf{b}_3$*. The following are the correct updates for our parameters $\mu^i_{x_i}(t)$ and $\rho^i_{x_i}(t)$:*

$$\rho^i_{x_i}(\mathsf{b}_1, T) = \rho^i_{x_i}(\mathsf{b}_2, 0)\rho^i_{x_i}(\mathsf{b}_3, 0), \tag{21}$$

$$\mu^i_{x_i}(\mathsf{b}_k, 0) = \mu^i_{x_i}(\mathsf{b}_1, T) \qquad k = 2,3. \tag{22}$$

Proof See Appendix G ■

Using Proposition 7 we can set the updates of the different parameters in the branching process according to (21–22). In the backward propagation of ρ^i, the value at the end of b_1 is the product of the values at the start of the two outgoing branches. This is the natural operation when we recall the interpretation of ρ^i as the probability of the downstream evidence given the current state (which is its exact meaning in a single component process): the downstream evidence of b_2 is independent of the downstream evidence of b_3, given the state of the process at the vertex $\langle \mathsf{b}_1, T \rangle$. The forward propagation of μ^i simply uses the value at the end of the incoming branch as initial value for the outgoing branches.

When switching to trees, we essentially increase the amount of evidence about intermediate states. Consider for example the tree of Figure 6 with an Ising chain model (as in the previous subsection). We can view the span from C to D as an interval with evidence at its end. When we add evidence at the tip of other branches we gain more information about intermediate points between C and D. Even though this evidence can represent evolution from these intermediate points, they do change our information state about them. To evaluate the impact of these changes on our approximation, we considered the tree of Figure 6, and compared it to inference in the backbone between C and D (Figure 4). Comparing the true marginal to the approximate one along the main backbone (see Figure 9) we see a major difference in the quality of the approximation. The evidence in the tree leads to a much tighter approximation of the marginal distribution. A more systematic comparison (Figure 10) demonstrates that the additional evidence reduces the magnitude of the error throughout the parameter space.

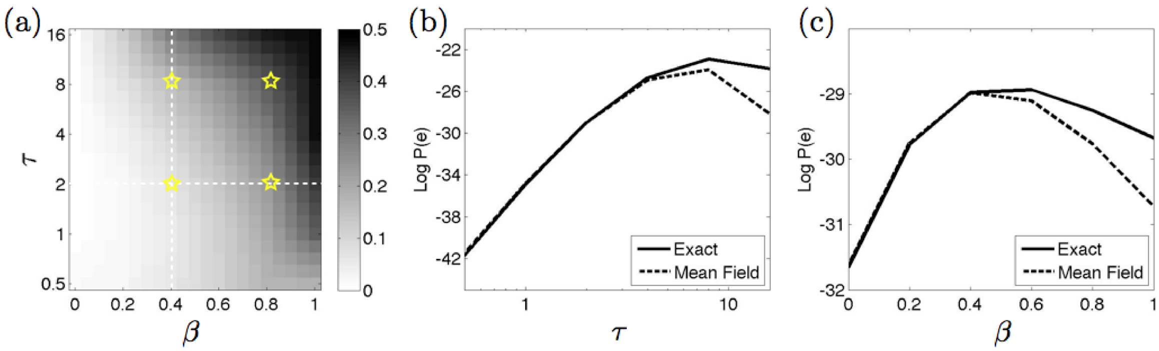


Figure 10: *(a)* Evaluation of the relative error in expected sufficient statistics for an Ising chain in branching-time; compare to Figure 4(a). *(b),(c)* Evaluation of the estimated likelihood on a tree w.r.t. the rate τ and coupling β; compare to Figure 4(b),(c).

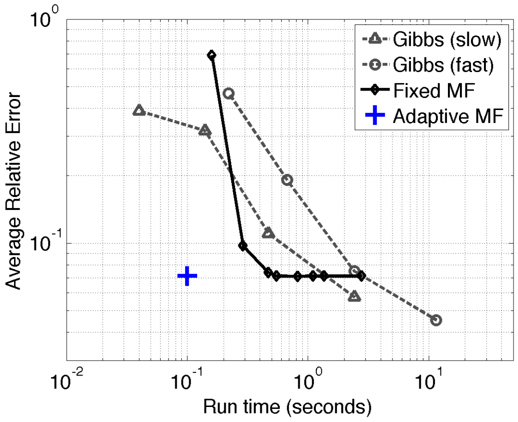


Figure 11: Evaluation of the run time vs. accuracy trade-off for several choices of parameters for mean field and Gibbs sampling on the branching process of Figure 6.

Similarly to mean-field, the Gibbs sampling procedure for CTBNs (El-Hay et al., 2008) can also be extended to deal with branching processes. Comparing our method to the Gibbs sampling procedure we see (Figure 11) that the faster mean field approach dominates the Gibbs procedure over short run times. However, as opposed to mean field, the Gibbs procedure is asymptotically unbiased, and with longer run times it ultimately prevails. This evaluation also shows that the adaptive integration procedure in our methods strikes a better trade-off than using a fixed time granularity integration.

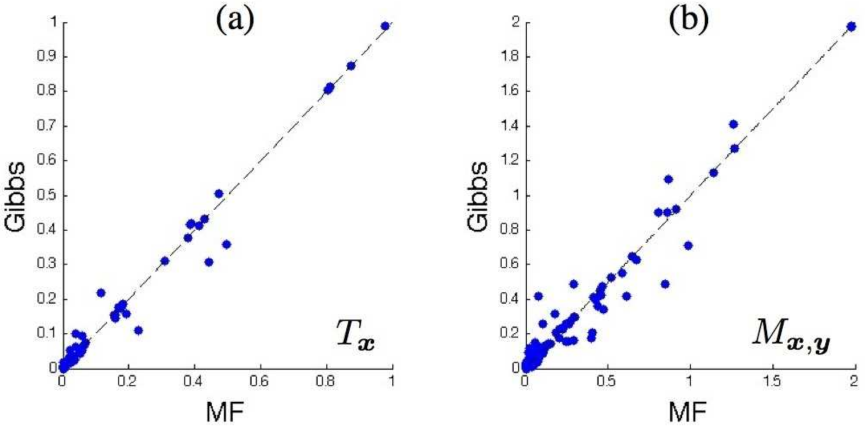


Figure 12: Comparison of estimates of expected sufficient statistics in the evolution of 18 interacting nucleotides, using a realistic model of RNA evolution. Each point is an expected value of: (a) residence time in a specific state of a component and its parents; (b) number of transition between two states. The x-axis is the estimate by the variational procedure, whereas the y-axis is the estimate by Gibbs sampling.

As a more demanding test, we applied our inference procedure to a model similar to the one introduced by Yu and Thorne (2006) for a stem of 18 interacting RNA nucleotides in 8 species in the phylogeny of Figure 6. In this model the transition rate between two sequences that differ in a single nucleotide depends on difference between their folding energy. Specifically, the transition rate from sequence x to sequence y is given by

$$q_{x,y} = 1.6\left(1+e^{E_{\text{fold}}(y)-E_{\text{fold}}(x)}\right)^{-1}, \qquad |\delta(x,y)| = 1,$$

where E_{fold} is the folding energy of the sequence. This equation implies that transition rates are increasing monotonically with the reduction of the folding energy. Hence, this model tends to evolve into low energy states. The folding energy in turn is a sum of local stacking energies, involving quadruples of nucleotides as described by the factors in Figure 7. Denoting the subset of positions contained in each quadruple by D_k, the energy is

$$E_{\text{fold}}(x) = \sum_k E^k_{\text{fold}}(x|_{D_k}),$$

where $x|_{D_k}$ is the subset of nucleotides that belong factor k. This model is equivalent to a CTBN in which the parents of each components are the other components that share the same factors. This property follows from the fact that for any pair x and y, where $\delta(x,y) = \{i\}$, the difference between the energies of these two sequences depends only on the factors that contain i.

We compared our estimate of the expected sufficient statistics of this model to these obtained by the Gibbs sampling procedure. The Gibbs sampling estimates were chosen by running the procedure with an increasing computation time until there was no significant change in the results. The final estimates was obtained using 5000 burn-in rounds, 10000 number of samples and 100 rounds between two consecutive samples. The results, shown in Figure 12, demonstrate that over all the two approximate inference procedures are in good agreement about the value of the expected sufficient statistics.

8. Discussion

In this paper we formulate a general variational principle for continuous-time Markov processes (by reformulating and extending the one proposed by Opper and Sanguinetti, 2007), and use it to derive an efficient procedure for inference in CTBNs. In this mean field approximation, we use a product of independent inhomogeneous processes to approximate the multi-component posterior.

Our procedure enjoys the same benefits encountered in discrete-time mean field procedure (Jordan et al., 1999): it provides a lower-bound on the likelihood of the evidence and its run time scales linearly with the number of components. Using asynchronous updates it is guaranteed to converge, and the approximation represents a consistent joint distribution. It also suffers from expected shortcomings: the functional has multiple local maxima, it cannot capture complex interactions in the posterior (Example 3). By using a time-inhomogeneous representation our approximation does capture complex patterns in the temporal progression of the marginal distribution of each component. Importantly, the continuous-time parameterization enables straightforward implementation using standard ODE integration packages that automatically tune the trade-off between time granularity and approximation quality. We show how it is extended to perform inference on phylogenetic trees, where the posterior distribution is directly affected by several evidence points, and show that it provides fairly accurate answers in the context of a real application (Figure 12).

A key development is the introduction of marginal density sets. Using this representation we reformulate and extend the variational principle proposed by Opper and Sanguinetti (2007) , which incorporates a different inhomogeneous representation. This modification allows handling direct evidence of the state of the process, as in the case of CTBNs, while keeping the representation of the approximation bounded. The extension of this principle to CTBNs follows by exploiting their networks structure. This adaptation of continuously inhomogeneous representations to CTBNs increases the flexibility of the approximation relative to the piecewise homogeneous representation of Saria et al. (2007) and, somewhat surprisingly, also significantly simplifies the resulting formulation.

The proposed representation is natural in the sense that its functions are the time-derivatives of the expected sufficient statistics that we are willing to evaluate. Hence, once finding the optimal value of the lower bound, calculating these expectations is straightforward. This representation is analogous to mean parameters which have proved powerful in variational approximations of exponential families over finite random variable sets (Wainwright and Jordan, 2008).

We believe that even in cases where evidence is indirect and noisy, the marginal density representation should comprise smoother functions than posterior rates. Intuitively, in the presence of a noisy observation the posterior probability of some state x can be very small. In such cases, the posterior transition rate form x into a state that better explains the observation might tend to a large quantity. This reasoning suggests that marginal density representations should be better handled by adaptive numerical integration algorithms. An interesting direction would be to test this conjecture empirically.

A possible extension is using our variational procedure to generate the initial distribution for the Gibbs sampling procedure and thus skip the initial burn-in phase and produce accurate samples. Another attractive aspect of this new variational approximation is its potential use for learning model parameters from data. It can be easily combined with the EM procedure for CTBNs (Nodelman et al., 2005a) to obtain a Variational-EM procedure for CTBNs, which monotonically increases the likelihood by alternating between steps that improve the approximation η (the updates discussed here) and steps that improve the model parameters θ (an M-step Nodelman et al., 2005a). Finally, marginal density sets are a particularly suitable representation for adapting richer representations such as Bethe, Kikuchi and convex approximations to non-homogeneous versions (El-Hay et al., 2010). Further work in that direction should allow bridging the gap in the wealth of inference techniques between finite domain models and continuous-time models.

Acknowledgments

We thank Ofer Meshi, Menachem Fromer and the anonymous reviewers for helpful remarks on previous versions of the manuscript. This research was supported in part by a grant from the Israel Science Foundation. Tal El-Hay is supported by the Eshkol fellowship from the Israeli Ministry of Science.

Appendix A. The Relation Between CTBNs and DBNs

In this section we show that the DBN construction of Equations (6-7) is such that as h approaches 0, the distribution P_h approaches Pr. To show this, it suffice to show that

$$\lim_{h\to 0} \frac{P_h(X^{(t_{k+1})} = y | X^{(t_k)} = x) - \boldsymbol{1}_{x=y}}{h} = q_{x,y} \quad .$$

We ensured this condition holds component-wise, and now need to show that this leads to global consistency.

Plugging Equation (7) into Equation (6), the transition probability of the DBN is

$$P_h(X^{(t_{k+1})} = y | X^{(t_k)} = x) = \prod_i \left(\boldsymbol{1}_{x_i=y_i} + q^{i|\mathbf{Pa}_i}_{x_i,y_i|u_i} \cdot h \right) \quad .$$

Since we consider the limit as h approaches 0, any term that involves h^d with $d > 1$ is irrelevant. And thus, we can limit our attention to the constant terms and terms linear in h. Expanding the product gives

$$P_h(X^{(t_{k+1})} = y | X^{(t_k)} = x) = \prod_i \boldsymbol{1}_{x_i=y_i} + \sum_i q^{i|\mathbf{Pa}_i}_{x_i,y_i|u_i} \cdot h \prod_{j\neq i} \boldsymbol{1}_{x_j=y_j} + o(h) \quad .$$

Now, $\prod_i \boldsymbol{1}_{x_i=y_i} = \boldsymbol{1}_{x=y}$. Moreover, it is easy to verify that

$$q_{x,y} = \sum_i q^{i|\mathbf{Pa}_i}_{x_i,y_i|u_i} \prod_{j\neq i} \boldsymbol{1}_{x_j=y_j} \quad .$$

Thus,

$$P_h(X^{(t_{k+1})} = y | X^{(t_k)} = x) = \boldsymbol{1}_{x=y} + q_{x,y}h + o(h),$$

proving the required condition.

Appendix B. Marginal Density Sets and Markov Processes - Proof of Lemma 2

Proof Given η, we define the *inhomogeneous rate matrix* $\mathbb{R}(t)$ as in Equation (11). $\mathbb{R}(t)$ is a valid rate matrix because its off-diagonals are non-negative as they are the quotient of two non-negative functions, and because applying the requirement on $\gamma_{x,x}(t)$ in Definition 1

$$r_{x,x}(t) = \frac{\gamma_{x,x}(t)}{\mu_x(t)} = -\frac{\sum_{y\neq x}\gamma_{x,y}(t)}{\mu_x(t)} = -\sum_{y\neq x} r_{x,y}(t) \ \ ,$$

we see that $\mathbb{R}(t)$'s diagonals are negative and the rows sum to 0. We can use these rates with the initial value $\mu_x(0)$ to construct the Markov process P_η from the forward master equation

$$\frac{d}{dt}P_\eta(X^{(t)}=x) = \sum_y P_\eta(X^{(t)}=y) r_{y,x}(t) \ ,$$

and

$$P_\eta(X^{(0)}) = \mu(0) \ \ .$$

To conclude the proof we show that P_η and the marginal density set satisfy (10). First, from Definition 1 it follows that $\mu(t)$ is the solution to the master equation of $P_\eta(X^{(t)})$, because the initial values match at $t=0$ and the time-derivatives of the two functions are identical. Thus

$$P_\eta(X^{(t)}=x) = \mu_x(t) \ \ .$$

Next, the equivalence of the joint probability densities can be proved:

$$\begin{aligned}
\lim_{h\to 0}\frac{\Pr(X^{(t)}=x, X^{(t+h)}=y)}{h} &= \lim_{h\to 0}\frac{\mu_x(t)\Pr(X^{(t+h)}=y|\Pr(X^{(t)}=x)}{h} \\
&= \lim_{h\to 0}\frac{\mu_x(t) r_{x,y}(t) h}{h} \\
&= \mu_x(t) r_{x,y}(t) \ .
\end{aligned}$$

From the definition of $r_{x,y}(t)$ and the fact that $\gamma_{x,y}(t)=0$ whenever $\mu_x(t)=0$, it follows that $\mu_x(t)r_{x,y}(t)$ is exactly $\gamma_{x,y}(t)$ ■

Appendix C. Expectations in Inhomogeneous Processes

This section includes the proofs of the lemmas used in the proof of the variational lower bound theorem.

C.1 Expectations of Functions of States - Proof of Lemma 4

Proof Changing the order of integration we obtain

$$\mathbf{E}_{f_\eta}\left[\int_0^T \psi(\sigma(t),t)dt\right] \equiv \int_\Sigma f_\eta(\sigma)\int_0^T \psi(\sigma(t),t)dtd\sigma \ \ = \ \ \int_0^T\int_\Sigma f_\eta(\sigma)\cdot\psi(\sigma(t),t)\,d\sigma dt \ .$$

For each $t \in T$ we decompose the inner integral according to possible states at that time:

$$\begin{aligned}\int_\Sigma f_\eta(\sigma)\cdot\psi(\sigma(t),t)\,d\sigma &= \sum_x \int_\Sigma f_\eta(\sigma)\cdot \mathbf{1}_{\sigma(t)=x}\cdot\psi(x,t)\,d\sigma \\ &= \sum_x \psi(x,t)\int_\Sigma f_\eta(\sigma)\cdot\mathbf{1}_{\sigma(t)=x}\,d\sigma \\ &= \sum_x \psi(x,t)\mu_x(t)\ .\end{aligned}$$

■

C.2 Expectations of Functions of Transitions - Proof of Lemma 5

Proof Given a trajectory σ there exists a small enough $h>0$ such that for every transition and for every $t\in(t_k-h,t_k)$ we have $\sigma(t)=x_{k-1}$ and $\sigma(t+h)=x_k$. In that case we can rewrite the sum in the expectation term as

$$\begin{aligned}\sum_{k=1}^{K^\sigma}\psi(x^\sigma_{k-1},x^\sigma_k,t^\sigma_k) &= \sum_{k=1}^{K^\sigma}\frac{1}{h}\int_{t_k-h}^{t_k}\psi(\sigma(t),\sigma(t+h),t)dt+\frac{o(h)}{h} \\ &= \frac{1}{h}\int_0^{T-h}\psi(\sigma(t),\sigma(t+h),t)dt+\frac{o(h)}{h}\ ,\end{aligned}$$

where the first equality follows from continuity and the second one from the requirement that $\psi(x,x,t)=0$. Taking the limit $h\to 0$ and using this requirement again gives

$$\sum_{k=1}^{K^\sigma}\psi(x^\sigma_{k-1},x^\sigma_k,t^\sigma_k)=\frac{d}{ds}\left[\int_0^T\psi(\sigma(t),\sigma(t+s),t)dt\right]_{s=0}\ .$$

Taking expectation we obtain

$$\begin{aligned}&\int_\Sigma f(\sigma)\frac{d}{ds}\left[\int_0^T\psi(\sigma(t),\sigma(t+s),t)dt\right]_{s=0}d\sigma \\ &= \int_\Sigma f(\sigma)\frac{d}{ds}\left[\int_0^T\sum_x\sum_{y\neq x}\psi(x,y,t)\,\mathbf{1}_{\sigma(t)=x}\mathbf{1}_{\sigma(t+s)=y}\,dt\right]_{s=0}d\sigma \\ &= \frac{d}{ds}\left[\int_0^T\sum_x\sum_{y\neq x}\psi(x,y,t)\int_\Sigma f(\sigma)\,\mathbf{1}_{\sigma(t)=x}\mathbf{1}_{\sigma(t+s)=y}d\sigma\,dt\right]_{s=0}\ .\end{aligned}$$

The inner integral in the last term is a joint probability

$$\int_\Sigma f(\sigma)\,\mathbf{1}_{\sigma(t)=x}\mathbf{1}_{\sigma(t+s)=y}d\sigma=\Pr(X^{(t)}=x,X^{(t+s)}=y)\ .$$

Switching the order of integration and differentiation and using

$$\frac{d}{ds}\Pr(X^{(t)}=x,X^{(t+s)}=y)\Big|_{s=0}=\gamma_{xy}(t),\quad x\neq y,$$

gives the desired result. ■

Appendix D. Proof of the Factored Representation of the Energy Functional

Proof We begin with the definition of the average energy

$$\begin{aligned}\mathcal{E}(\eta;\mathbb{Q}) &= \int_0^T \sum_x \left[\mu_x(t)q_{x,x} + \sum_{y\neq x} \gamma_{x,y}(t)\ln q_{x,y}\right]dt \\ &= \int_0^T \sum_x \left[\mu_x(t)q_{x,x} + \sum_i \sum_{y_i\neq x_i} \gamma^i_{x_i,y_i}(t)\mu^{\setminus i}(t)\ln q_{x,y}\right]dt\ .\end{aligned}$$

where the equality stems from the observation that the only states y that may have $\gamma_{x,y}(t) > 0$, are those with $\delta(x,y) \leq 1$ (all the rest are 0). Thus, the enumeration over all possible states collapses into an enumeration over all components i and all states $y_i \neq x_i$. Due to the fact that we are only considering transitions in single components, we may replace the global joint density $\gamma_{x,y}$ with $\gamma^i_{x_i,y_i} \cdot \mu^{\setminus i}(t)$, as per definition.

Using (5), we can decompose the transition rates $q_{x,x}$ and $q_{x,y}$ to get

$$\begin{aligned}\mathcal{E}(\eta;\mathbb{Q}) &= \sum_i \int_0^T \sum_x \left[\mu_x(t)q_{x_i,x_i|u_i} + \sum_{y_i\neq x_i} \gamma^i_{x_i,y_i}(t)\mu^{\setminus i}(t)\ln q_{x_i,y_i|u_i}\right]dt \\ &= \sum_i \int_0^T \sum_{x_i} \left[\mu^i_{x_i}(t)\sum_{x\setminus i}\mu^{\setminus i}_{x\setminus i}(t)q_{x_i,x_i|u_i} + \sum_{y_i\neq x_i} \gamma^i_{x_i,y_i}(t)\mu^{\setminus i}_{x\setminus i}(t)\ln q_{x_i,y_i|u_i}\right]dt\ \ .\end{aligned}$$

To get to the last equality we use the factorization of $\mu(t)$ as a product of $\mu^i(t)$ with $\mu^{\setminus i}(t)$ and the reordering of the summation. Next we simply write the previous sum as an expectation over $X\setminus i$

$$\mathcal{E}(\eta;\mathbb{Q}) = \sum_i \int_0^T \sum_{x_i} \mu^i_{x_i}(t)\mathbf{E}_{\mu^{\setminus i}(t)}\left[q_{x_i,x_i|U_i}\right] + \sum_i \int_0^T \sum_{y_i\neq x_i} \gamma^i_{x_i,y_i}(t)\mathbf{E}_{\mu^{\setminus i}(t)}\left[\ln q_{x_i,y_i|U_i}\right]dt\ ,$$

which concludes the proof.

Turning to the entropy-like term we have:

$$\begin{aligned}\mathcal{H}(\eta) &= \int_0^T \sum_x \sum_{y\neq x} \gamma_{x,y}(t)[1+\ln\mu_x(t) - \ln\gamma_{x,y}(t)]dt \\ &= \sum_i \int_0^T \sum_x \sum_{y_i\neq x_i} \mu^{\setminus i}(t)\gamma_{x_i,y_i}(t)[1+\sum_i \ln\mu^i_{x_i}(t) - \ln\gamma_{x_i,y_i}(t) - \sum_{j\neq i}\ln\mu_{x_j}(t)]dt \\ &= \sum_i \int_0^T \sum_{x_i} \sum_{y_i\neq x_i} \gamma_{x_i,y_i}(t)[1+\ln\mu^i_{x_i}(t) - \ln\gamma_{x_i,y_i}(t)]dt \\ &= \sum_i \mathcal{H}(\eta^i)\ \ ,\end{aligned}$$

where, the first equality is definition of $\mathcal{H}$. The second one follows from the definition of the factored density set. The third one is obtained by algebraic manipulation and the last one is again the definition of $\mathcal{H}$. ■

Appendix E. Euler-Lagrange Equations

The problem of finding the fixed points of *functionals* whose arguments are continuous functions comes from the field of *Calculus of variations*. We briefly review the usage Euler-Lagrange equation for solving optimization problems over functionals. Additional information can be found in Gelfand and Fomin (1963).

A functional is a mapping from a vector space to its underlying field. In our case the functional is the Lagrangian introduced in Section 4, which is an integral over real-valued functions, and the underlying field is the real numbers.

Given a functional over a normed space of continuously differentiable real functions of the form

$$I[y] = \int_a^b f(t, y(t), y'(t))dt$$

where $y'(t)$ is the time-derivative of the function $y(t)$, we would like to find a function $y(t)$ that minimizes (or in our case maximizes) the functional subject to $y(a) = y_a$ and $y(b) = y_b$. In the simplest case, when there are no additional constraints, a necessary condition for y to be a local optimum is that y is a *stationary point*. Roughly, a stationary point is a function y, where $I[y]$ is insensitive to small variations in y. That is, given a function $h(t)$ where $h(a) = 0$ and $h(b) = 0$, the change of the functional $I[y+h] - I[y]$ is small relative to the norm of h. For $y(t)$ to be a stationary point, it must satisfy the *Euler-Lagrange* equations (Gelfand and Fomin, 1963)

$$\frac{\partial}{\partial y} f(t, y(t), y'(t)) - \frac{d}{dt}\left(\frac{\partial}{\partial y'} f(t, y(t), y'(t))\right) = 0 \ . \tag{23}$$

In this paper we have additional constraints describing the time derivative of μ. The generalization of the Euler-Lagrange equations to that case is straightforward. Denoting the subsidiary constraints by $g(t, y(t), y'(t)) = 0$, we simply replace $f(t, y, y')$ by $f(t, y, y') - \lambda(t)g(t, y, y')$ in Equation 23.

An example for the use of this equation is in the following proof.

Appendix F. Stationary Points of the Lagrangian - Proof of Theorem 6

Proof For convenience, we begin by rewriting the Lagrangian in explicit form: $\mathcal{L} = \int_0^T f(y(t), y'(t))dt$ where $y(t) = \langle \mu(t), \gamma(t), \lambda(t) \rangle$ is a concatenation of the parameters and Lagrange multiplier and

$$\begin{aligned} f(y(t), y'(t)) \quad = \quad & \sum_{i=1}^{D} \sum_{x_i} \Bigg[\mu^i_{x_i}(t) \mathbf{E}_{\mu^{\setminus i}(t)} \left[q_{x_i, x_i | U_i} \right] + \sum_{y_i \neq x_i} \gamma^i_{x_i, y_i}(t) \mathbf{E}_{\mu^{\setminus i}(t)} \left[\ln q_{x_i, y_i | U_i} \right] \\ & + \sum_{y_i \neq x_i} \gamma_{x_i y_i} \left[1 + \ln \mu^i_{x_i}(t) - \ln \gamma^i_{x_i y_i}(t) \right] - \lambda^i_{x_i}(t) \left(\frac{d}{dt} \mu^i_{x_i}(t) - \sum_{y_i} \gamma^i_{x_i y_i}(t) \right) \Bigg] \ . \end{aligned}$$

The Euler-Lagrange equations of the Lagrangian define its stationary points w.r.t. the parameters of each component $\mu^i(t)$, $\gamma^i(t)$ and $\lambda^i(t)$.

First, we take the partial derivatives of f w.r.t to $\mu^i_{x_i}(t)$ as well as $\frac{d}{dt}\mu^i_{x_i}(t)$ and plug them into Equation 23. We start by handling the energy terms. These terms involve expectations in the form

$\mathbf{E}_{\mu\backslash j(t)}\left[g(U_j)\right] = \sum_{u_j}\mu_{u_j}(t)g(u_j)$. The parameter $\mu^i_{x_i}(t)$ appears in these terms only when i is a parent of j and u_j is consistent with x_i. In that case $\frac{\partial}{\partial\mu^i_{x_i}}\mu_{u_j} = \mu_{u_j}/\mu^i_{x_i}$. Thus,

$$\frac{\partial}{\partial\mu^i_{x_i}}\mathbf{E}_{\mu\backslash j(t)}\left[g(U_j)\right] = \mathbf{E}_{\mu\backslash j(t)}\left[g(U_j) \mid x_i\right]\cdot\delta_{j\in Children_i}$$

Recalling the definitions of the averaged rates

$$\bar{q}^{\,i}_{x_i,x_i|x_j}(t) = \mathbf{E}_{\mu\backslash i(t)}\left[q^{i|\mathbf{Pa}_i}_{x_i,x_i|U_i} \mid x_j\right]$$

and

$$\tilde{q}^{i}_{x_i,y_i|x_j}(t) = \exp\left\{\mathbf{E}_{\mu\backslash i(t)}\left[\ln q^{i|\mathbf{Pa}_i}_{x_i,y_i|U_i} \mid x_j\right]\right\}$$

we obtain

$$\frac{\partial}{\partial\mu^i_{x_j}}\mathbf{E}_{\mu\backslash j(t)}\left[q^j_{x_j,x_j|U_j}\right] = \delta_{j\in Children_i}\bar{q}^{\,j}_{x_j,x_j|x_i}(t)$$

and

$$\frac{\partial}{\partial\mu^i_{x_j}}\mathbf{E}_{\mu\backslash j(t)}\left[\ln q^j_{x_j,x_j|U_j}\right] = \delta_{j\in Children_i}\ln\tilde{q}^{j}_{x_j,x_j|x_i}(t).$$

Therefore the derivative of the sum over $j \neq i$ of the energy terms is

$$\psi^i_{x_i}(t) \equiv \sum_{j\in Children_i}\sum_{x_j}\left[\mu^j_{x_j}(t)\bar{q}^{\,j}_{x_j,x_j|x_i}(t) + \sum_{x_j\neq y_j}\gamma^j_{x_j,y_j}(t)\ln\tilde{q}^{j}_{x_j,y_j|x_i}(t)\right] \quad .$$

Additionally, the derivative of the energy term for $j = i$ is $\bar{q}^{\,i}_{x_i,x_i}(t) \equiv \mathbf{E}_{\mu\backslash i(t)}\left[q_{x_i,x_i|U_i}\right]$. Next, the derivative of the entropy term is $\gamma^i_{x_i,x_i}(t)/\mu^i_{x_i}(t)$. Finally, the derivative of f with respect to $\frac{d}{dt}\mu^i_{xi}(t)$ is $-\lambda^i_{x_i}(t)$. Plugging in these derivatives into Equation (23) we obtain

$$\bar{q}^{\,i}_{x_i,x_i}(t) + \psi^i_{x_i}(t) - \frac{\gamma^i_{x_i,x_i}}{\mu^i_{x_i}(t)} + \frac{d}{dt}\lambda^i_{x_i}(t) = 0 \quad . \tag{24}$$

Next, the derivative w.r.t. $\gamma^i_{x_i,y_i}(t)$ gives us

$$\ln\mu^i_{x_i}(t) + \ln\tilde{q}^i_{x_i,y_i}(t) - \ln\gamma^i_{x_i,y_i}(t) + \lambda^i_{y_i}(t) - \lambda^i_{x_i}(t) = 0 \quad . \tag{25}$$

Denoting $\rho^i_{x_i}(t) = \exp\{\lambda^i_{x_i}(t)\}$, Equation (25) becomes

$$\gamma^i_{x_i,y_i}(t) = \mu^i_{x_i}(t)\tilde{q}^i_{x_i,y_i}(t)\frac{\rho^i_{y_i}(t)}{\rho^i_{x_i}(t)} \quad ,$$

which is the algebraic equation of γ. Using this result and the definition of $\gamma^i_{x_i,x_i}$ we have

$$\gamma^i_{x_i,x_i}(t) = -\sum_{y_i\neq x_i}\gamma^i_{x_i,y_i}(t) = -\mu^i_{x_i}(t)\sum_{x_i,y_i}\tilde{q}^i_{x_i,y_i}(t)\frac{\rho^i_{y_i}(t)}{\rho^i_{x_i}(t)}.$$

Plugging this equality into (24) and using the identity $\frac{d}{dt}\rho^i_{x_i}(t) = \frac{d}{dt}\lambda^i_{x_i}(t)\rho^i_{x_i}(t)$ gives

$$\frac{d}{dt}\rho^i_{x_i}(t) = -\rho^i_{x_i}(t)\left(\bar{q}^{\,i}_{x_i,x_i}(t) + \psi^i_{x_i}(t)\right) - \sum_{y_i\neq x_i}\tilde{q}^i_{x_i,y_i}\rho^i_{y_i}(t) \quad .$$

Thus the stationary point of the Lagrangian matches the updates of (16–17). ■

Appendix G. Proof of Proposition 7

Proof We denote the time at the vertex $t_0 = (\mathsf{b}_1, T)$, the time just before as $t_1 = (\mathsf{b}_1, T-h)$ and the times just after it on each branch $t_2 = (\mathsf{b}_2, h)$ and $t_3 = (\mathsf{b}_3, h)$, as in Figure 13.

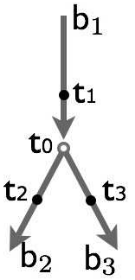


Figure 13: Branching process with discretization points of Lemma 7.

The marginals $\mu^i_{x_i}(\mathsf{b}_1, t)$ are continuous, as they are derived from the forward differential equation. To derive the propagation formula for the $\rho^i_{x_i}(t)$ requires additional care. The $\rho^i_{x_i}(t)$ have been derived from the constraints on the time-derivative of $\mu^i_{x_i}(t)$. In a vertex this constraint is threefold, as we now have the constraints on b_1

$$\frac{\mu^i_{x_i}(t_0) - \mu^i_{x_i}(t_1)}{h} = \sum_{y_i} \gamma^i_{x_i,y_i}(t_1)$$

and those on the other branches b_k for $k = 2,3$

$$\frac{\mu^i_{x_i}(t_k) - \mu^i_{x_i}(t_0)}{h} = \sum_{y_i} \gamma^i_{x_i,y_i}(t_0) \quad .$$

The integrand of the Lagrangian corresponding to point t_0 is

$$\begin{aligned} \mathcal{L}_{|t_0} &= \tilde{\mathcal{F}}(\eta;\mathbb{Q})_{|t_0} + \lambda^0(t_1)\left(\frac{\mu^i_{x_i}(t_0) - \mu^i_{x_i}(t_1)}{h} - \sum_{y_i}\gamma^i_{x_i,y_i}(t_1)\right) \\ &- \sum_{k=2,3}\lambda^k(t_0)\left(\frac{\mu^i_{x_i}(t_k) - \mu^i_{x_i}(t_0)}{h} - \sum_{y_i}\gamma^i_{x_i,y_i}(t_0)\right) \quad , \end{aligned}$$

as this is the only integrand which involves $\mu_{x_i}(t_0)$, the derivative of the Lagrangian collapses into

$$\begin{aligned} \frac{\partial}{\partial \mu^i_{x_i}(t_0)}\mathcal{L} &= \frac{\partial}{\partial \mu^i_{x_i}(t_0)}\mathcal{L}_{|t_0} \\ &= \frac{\lambda^0(t_1)}{h} - \left(\frac{\lambda^2(t_0)}{h} + \frac{\lambda^3(t_0)}{h}\right) + \frac{\partial}{\partial \mu^i_{x_i}(t_0)}\tilde{\mathcal{F}}(\eta;\mathbb{Q})_{|t_0} = 0 \quad . \end{aligned}$$

Rearranging the previous equation and multiplying by h, we get

$$\lambda^0(t_1) = \lambda^2(t_0) + \lambda^3(t_0) + \frac{\partial}{\partial \mu^i_{x_i}(t_0)}\tilde{\mathcal{F}}(\eta;\mathbb{Q})_{|t_0} h \quad .$$

Looking at (24) we can see that as t_0 is not a leaf, and thus $\mu^i_{x_i}(t_0) > 0$ and the derivative of the functional cannot diverge. Therefore, as $h \to 0$ this term vanishes and we are left with

$$\lambda^0(t_1) = \lambda^2(t_0) + \lambda^3(t_0)$$

which after taking exponents gives (21). ■

References

C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor. Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems 20*. MIT Press, 2007.

K. L. Chung. *Markov Chains with Stationary Transition Probabilities*. Springer Verlag, Berlin, 1960.

T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Comput. Intell.*, 5(3):142–150, 1989.

M. Dewar, V. Kadirkamanathan, M. Opper, and G. Sanguinetti. Parameter estimation and inference for stochastic reaction-diffusion systems: application to morphogenesis in d. melanogaster. *BMC Systems Biology*, 4(1):21, 2010.

T. El-Hay, N. Friedman, D. Koller, and R. Kupferman. Continuous time markov networks. In *Proc. Twenty-second Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

T. El-Hay, N. Friedman, and R. Kupferman. Gibbs sampling in factorized continuous-time markov processes. In *Proc. Twenty-fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

T. El-Hay, I. Cohn, N. Friedman, and R. Kupferman. Continuous-time belief propagation. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

G. Elidan, I. Mcgraw, and D. Koller. Residual belief propagation: informed scheduling for asynchronous message passing. In *Proc. Twenty-second Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

Y. Fan and C. R. Shelton. Sampling for approximate inference in continuous time Bayesian networks. In *Tenth International Symposium on Artificial Intelligence and Mathematics*, 2008.

Y. Fan and C. R. Shelton. Learning continuous-time social network dynamics. In *Proc. Twenty-fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

J. Felsenstein. *Inferring Phylogenies*. Sinauer, 2004.

C. W. Gardiner. *Handbook of Stochastic Methods*. Springer-Verlag, New-York, third edition, 2004.

I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Prentice-Hall, 1963.

K. Gopalratnam, H. Kautz, and D. S. Weld. Extending continuous time bayesian networks. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI)*, pages 981–986. AAAI Press, 2005.

M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational approximations methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge MA, 1999.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

A. Lipshtat, H. B. Perets, N. Q. Balaban, and O. Biham. Modeling of negative autoregulated genetic networks in single cells. *Gene*, 347:265, 2005.

K. P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002.

B. Ng, A. Pfeffer, and R. Dearden. Continuous time particle filtering. In *Proc. of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.

U. Nodelman, C. R. Shelton, and D. Koller. Continuous time Bayesian networks. In *Proc. Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 378–387, 2002.

U. Nodelman, C. R. Shelton, and D. Koller. Learning continuous time Bayesian networks. In *Proc. Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 451–458, 2003.

U. Nodelman, C. R. Shelton, and D. Koller. Expectation maximization and complex duration distributions for continuous time Bayesian networks. In *Proc. Twenty-first Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 421–430, 2005a.

U. Nodelman, C. R. Shelton, and D. Koller. Expectation propagation for continuous time Bayesian networks. In *Proc. Twenty-first Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 431–440, 2005b.

M. Opper and G. Sanguinetti. Variational inference for Markov jump processes. In *Advances in Neural Information Processing Systems 20*. MIT Press, 2007.

M. Opper and G. Sanguinetti. Learning combinatorial transcriptional dynamics from gene expression data. *Bioinformatics*, 26(13):1623–1629, 2010.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition, 2007.

S. Rajaram, T. Graepel, and R. Herbrich. Poisson-networks: A model for structured point processes. In *Proc. Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, January 2005.

A. Ruttor and M. Opper. Approximate parameter inference in a stochastic reaction-diffusion model. In *Proc. Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9, pages 669–676, 2010.

G. Sanguinetti, A. Ruttor, M. Opper, and C. Archambeau. Switching regulatory models of cellular stress response. *Bioinformatics*, 25(10):1280–1286, 2009.

S. Saria, U. Nodelman, and D. Koller. Reasoning at the right time granularity. In *Proc. Twenty-third Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

A. Simma, M. Goldszmidt, J. MacCormick, P. Barham, R. Black, R. Isaacs, and R. Mortier. Ct-nor: Representing and reasoning about events in continuous time. In *Proc. Twenty-fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1:1–305, 2008.

J. Yu and J. L. Thorne. Dependence among sites in RNA evolution. *Mol. Biol. Evol.*, 23:1525–37, 2006.