Dynamics of a Bayesian hyperparameter

Martin Biehl and Ryota Kanai

Araya Inc., Tokyo, Japan
martin@araya.org

Abstract. The free energy principle which underlies active inference attempts to explain the emergence of Bayesian inference in stochastic processes under the assumption of (non-equilibrium) steady state distributions. We contribute a study of the dynamics of an exact Bayesian inference hyperparameter embedded in a Markov chain that infers the dynamics of an observed process. This system does not have a steady-state but still contains exact Bayesian inference. Our study may contribute to future generalizations of the free energy principle to non-steady state systems.
Our treatment uses well-known constructions in Bayesian inference. The main contribution is that we take a different perspective than that of standard treatments. We are interested in how the dynamics of Bayesian inference look from the outside.

Keywords: Free energy principle · active inference · Markov blankets · Bayesian inference

1 Introduction

One of the most fundamental components of the free energy principle is the approximate Bayesian inference lemma [1]. It claims to provide a sufficient condition for (possibly approximate) Bayesian inference to occur within an ergodic multivariate Markov process. The condition is that there is a partitioning of the variables into internal, active, sensory, and external variables such that the steady-state distribution factorizes in a particular way. If we write μ for internal, a for active, s for sensory, η for external variables and p^* for the steady state density then the required factorization is the conditional independence relation

$$p^*(\mu, \eta|s, a) = p^*(\mu|s, a)p^*(\eta|s, a). \tag{1}$$

This means that (S, A) form a Markov blanket for μ and also for η. However, Bayesian inference can also happen inside processes that don't have steady-state densities. We will illustrate this with two examples below. This explicitly shows that ergodicity and the corresponding Markov blanket condition are only sufficient for Bayesian inference and not necessary.

Often, the dynamics of the hyperparameters[1] of Bayesian inference are relegated to the background and the focus is on how to compute posteriors for a

[1] For example, the pseudo-counts that are accumulated as the parameters of a Dirichlet posterior over the categorical states of a generative process.

given hyperparameter or prior. The embedding of both the observed process as well as the hyperparameter into a Markov chain converts standard results into a setting very similar to that of the free energy principle in [1]. The differences are that we have a discrete countably infininte state space instead of a continuous one, discrete instead of continuous time, and in the current version no actions. We will include actions into our setting in future work. Methods for transitioning to continuous systems are well studied so that we are optimistic that insights from the discrete setting can eventually be carried over to the continuous domain. In general we think that the method of embedding Bayesian inference and possibly also approximate Bayesian inference processes into Markov chains can provide rigorously defined examples of interesting systems whose properties can then be studied from an external point of view. The present work exhibits how this can be done in principle.

We observe that the dynamics of the Bayesian hyperparameter can be specified directly in dependence on the last hyperparameter and the observation. This highlights the fact that the probability distributions representing the belief that is being updated are in some sense unnecessary for the dynamics of the process. They have no effect that isn't captured by the hyperparameter itself. This is similar to the situation of the approximate inference lemma where the most likely internal state only "appears" to engage in approximate Bayesian inference with respect to the external state. If we forgot how we derived the hyperparameter dynamics then all we could say is that they appear to engage in Bayesian inference since there is a belief updating process compatible with their dynamics.

2 IID parameter inference

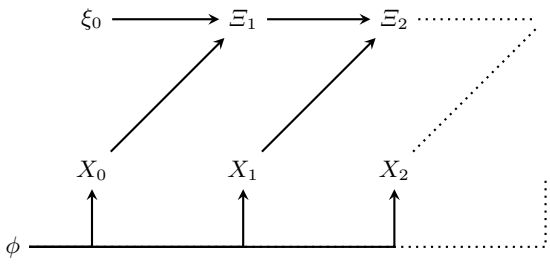

Fig. 1: Bayesian network of the hyperparameter updating process for an IID process with parameter ϕ and initial hyperparameter ξ_0.

Assume as given an identically and independently distributed (IID) random process $(X_t)_{t\in\mathbb{N}}$ with sample space $\mathcal{X}$ specified by a categorical distribution with

parameter $\phi = (\phi_x)_{x\in\mathcal{X}}$ which is a vector of the probabilities of the different outcomes i.e. $\phi_x \in [0,1]$ and

$$\sum_{x\in\mathcal{X}} \phi_x = 1. \tag{2}$$

For each $t \in \mathbb{N}$ we then have

$$p(x_t|\phi) = \prod_x \phi_x^{\delta_{xx_t}} \tag{3}$$

where δ_{xx_t} is the Kronecker delta.

We then assume another process $(\Xi_t)_{t\in\mathbb{N}}$ whose dynamics are those of a Bayesian hyperparameter (specifically a parameter of a Dirichlet distribution over parameters of categorical distributions) that updates to parameterize the posterior after each sample from $(X_t)_{t\in\mathbb{N}}$. More precisely, we imagine that for all $t \in \mathbb{N}$ the outcome ξ_t parameterizes a Dirichlet distribution $q(\hat{\phi}|\xi_t)$ over possible (values of/categorical distribution parameters) ϕ.[2] After observing a new sample x_t the posterior $q(\hat{\phi}|x_t, \xi_t)$ is then well defined. To update ξ_t to ξ_{t+1} we require that ξ_{t+1} is the parameter of the posterior. For this we must assume that there exists $\xi \in \Xi$ such that $q(\hat{\phi}|x_t, \xi_t) = q(\hat{\phi}|\xi)$. More generally, we can require that

$$p(\xi_{t+1}|\xi_t, x_t) := \delta_{f(\xi_t,x_t)}(\xi_{t+1}) \tag{4}$$

with

$$f(\xi_t, x_t) := \arg\min_{\xi} \mathrm{KL}[q(\hat{\Phi}|\xi)||q(\hat{\Phi}|x_t, \xi_t)]. \tag{5}$$

In the case we chose where ξ_t is the parameter of a Dirichlet distribution over categorical parameters the solution to this optimisation is [3]

$$\xi_{t+1} = \xi_t + \delta_{x_t} \tag{6}$$

since ξ_t are vectors with $|\mathcal{X}|$ components we can also write this (maybe more clearly) componentwise, i.e. for each component $x \in \mathcal{X}$:

$$(\xi_{t+1})_x := (\xi_t)_x + (\delta_{x_t})_x \tag{7}$$

Here $(\delta_{x_t})_x := \delta_{x_t x}$ with $\delta_{x_t x}$ the Kronecker delta. In other words, δ_{x_t} is a one-hot encoding of x_t. This defines all the mechanisms/kernels in the Bayesian network Figure 1 which illustrates our setting. We make two observations:

[2] We add a hat to variables that the beliefs encoded by ξ_t range over. This is to highlight that the hatted variables can take different values from the actual ones e.g. when we have a fixed ϕ that defines the IID process then in general the encoded belief $q(\hat{\phi}|\xi_t)$ still ranges over $\hat{\phi} \neq \phi$. A more technical reason is that the hatted variables are in some sense virtual. This should become clearer in the following. A rigorous definition of what "virtual" means is beyond the scope of this paper.

[3] This is the solution because it leads to the KL divergence being zero which means $q(\hat{\Phi}|f(\xi_t, x_t)) = q(\hat{\Phi}|x_t, \xi_t)$. See e.g. [2] for properties of Dirichlet priors for categorical distributions.

- Note that while we defined the dynamics of ξ_t via Bayesian inference/updating, the resulting dynamics are just those of a counter of occurrences. There is no reference anymore to the belief $q(\hat{\Phi}|\xi_t)$.
- The resulting Markov chain is not ergodic. The Markov chain state at time t is defined (ξ_t, x_t). A Markov chain can only be ergodic if all states are recurrent. However, since each component $(\xi_t)_x$ of ξ_t is non-decreasing and one of the components increases at every timestep we can never have $\xi_t = \xi_{t+n}$ for any integer $n > 0$.

2.1 Fully observable Markov chain

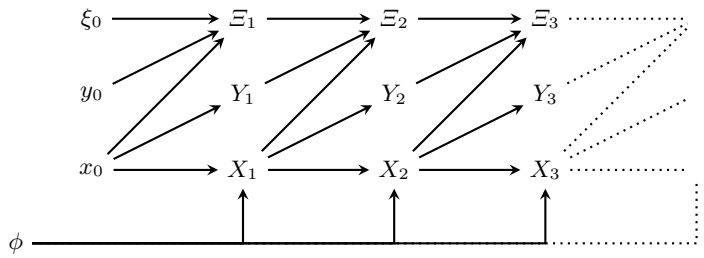

Fig. 2: Bayesian networks of the hyperparameter updating process for a fully observable Markov chain with Markov matrix specified by ϕ, initial Markov chain state x_0, initial stored state y_0, and initial hyperparameter ξ_0. The storage variable Y_t stores the values of X_{t-1} so that Ξ_{t+1} can use the pair $(x_{t-1}, x_t) = (y_t, x_t)$ which indicates the transition that occurred from $t-1$ to t to update.

In order to get an intuition for how to generalise the simple IID case to more interesting cases we look at possibly the next most simple case of inferring the transition probabilities of a time-homogenous Markov chain. Assume as given a time-homogenous finite (discrete-time) Markov chain $(X_t)_{t\in\mathbb{N}}$ with sample space $\mathcal{X}$, initial state $x_0 \in \mathcal{X}$, and Markov matrix (transition probabilities) $p(x_{t+1}|x_t) = \phi_{x_{t+1}x_t}$. Here $(\phi_{x_{t+1}x_t})_{x_{t+1},x_t\in\mathcal{X}}$ is a matrix of probabilities whose columns sum to one i.e. $\phi_{x_{t+1}x_t} \in [0,1]$ and for all $x_t \in \mathcal{X}$

$$\sum_{x_{t+1}\in\mathcal{X}} \phi_{x_{t+1}x_t} = 1. \tag{8}$$

For each $t \in \mathbb{N}$ we then have

$$p(x_{t+1}|x_t, \phi) := \prod_{x'x} \phi_{x'x}^{(\delta_{x_{t+1}\otimes x_t})_{x'x}} \tag{9}$$

where we define for $x, y \in \mathcal{X}$, $\delta_{x\otimes y}$ is a $|\mathcal{X}| \times |\mathcal{X}|$ matrix with

$$(\delta_{x\otimes y})_{ij} := \begin{cases} 1 \text{ if } i = x \text{ and } j = y \\ 0 \text{ else.} \end{cases} \tag{10}$$

We now assume two other processes $(\Xi_t)_{t\in\mathbb{N}}$ and $(Y_t)_{t\in\mathbb{N}}$. Similar to Section 2 dynamics of $(\Xi_t)_{t\in\mathbb{N}}$ are those of a Bayesian hyperparameter (a parameter of a Dirichlet distribution over parameters of $|\mathcal{X}|$ categorical distributions) that updates to parameterize the posterior after each transition from x_{t-1} to x_t. Since x_{t-1} is not available to ξ_{t+1} directly it gets stored in y_t. So the update depends on both x_t and y_t. More precisely, we imagine that for all $t \in \mathbb{N}$ the outcome ξ_t parameterizes a Dirichlet distribution $q(\hat{\phi}|\xi_t)$ over possible (values of/categorical distribution parameters) ϕ. At each timestep ξ_{t+1} is updated in response to the pair (y_t, x_t) where x_t is a new sample from the Markov chain and y_t is the stored previous sample x_{t-1} from the Markov chain. At the same time y_{t+1} is updated by setting it equal to x_t. In this way all values/data necessary for the next update are explicitly present at $t + 1$.

The posterior $q(\hat{\phi}|x_t, y_t, \xi_t)$ which corresponds to $q(\hat{\phi}|x_t, x_{t+1}, \xi_t)$ is then well defined. To update ξ_t to ξ_{t+1} we require that ξ_{t+1} is the parameter of the posterior. For this we must assume that there exists $\xi \in \Xi$ such that $q(\hat{\phi}|x_t, y_t, \xi_t) = q(\hat{\phi}|\xi)$. More generally, we can require that

$$p(\xi_{t+1}|\xi_t, x_t, y_t) := \delta_{f(\xi_t,x_t,y_t)}(\xi_{t+1}) \tag{11}$$

with

$$f(\xi_t, x_t, y_t) := \arg\min_{\xi} \mathrm{KL}[q(\hat{\Phi}|\xi)||q(\hat{\Phi}|x_t, y_t, \xi_t)]. \tag{12}$$

When ξ_t is the parameter of a Dirichlet distribution over categorical parameters this is equivalent to (cmp. [2])

$$f(\xi_t, x_t, y_t) : = \xi_t + \delta_{x_t\otimes y_t}. \tag{13}$$

The update of y_t is just the copying of x_t:

$$p(y_{t+1}|x_t) := \delta_{x_t}(y_{t+1}) \tag{14}$$

With this, all the mechanisms/kernels in the Bayesian network Figure 2 which illustrates our setting are defined.

We make two observations:

- Again the dynamics of ξ_t are just those of a counter of occurrences (of transitions now). There is no reference anymore to a belief.
- The resulting Markov chain is also not ergodic.

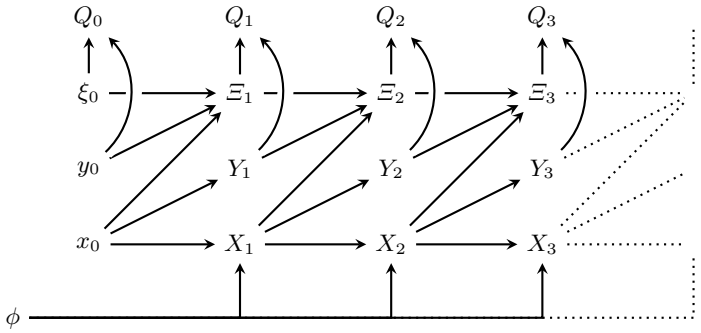

Fig. 3: Bayesian networks of the hyperparameter updating process for a fully observable Markov chain. Here we include the belief distribution as a single random variable Q_t. This random variable takes values in the joint probability distributions over $\hat{\Phi} \times \mathcal{X}^{\mathbb{N}}$ as explained in the text.

Internal belief dynamics We can make the internal belief more explicit by viewing it as a coarse-graining of the Markov chain state (ξ_t, y_t, x_t).

Let $\mathcal{Q}$ be the set of internal belief distributions that each pair (y_t, ξ_t) is mapped to and write Q_t as the random variable that represents the internal belief distribution at time t (see Figure 3). An instance of such a belief distribution is then denote by q_t.

For a particular (external) timestep t the joint distribution q_t is written

$$q_t(\hat{\phi}, \hat{x}_{-1:\infty}) := \prod_{\tau=0}^{\infty} q(\hat{x}_\tau | \hat{x}_{\tau-1}, \hat{\phi}) q_t(\hat{x}_{t-1}) q_t(\hat{\phi}) \tag{15}$$

where the "intitial" distributions $q_t(\hat{x}_{t-1})$ and $q_t(\hat{\phi})$ will be determined from (y_t, ξ_t) via the two functions we discuss below:

$$q_t(\hat{x}_{t-1}) := b_Y(y_t)(\hat{x}_{-1}) \qquad q_t(\hat{\phi}) := b_\Xi(\xi_t)(\hat{\phi}). \tag{16}$$

First define the functions $b_\Xi : \Xi \to \Delta_{\hat{\Phi}}$ and $b_Y : \mathcal{Y} \to \Delta_{\hat{\mathcal{X}}}$ via

$$b_\Xi(\xi)(\hat{\phi}) := \frac{1}{B(\xi)} \prod_{x'x} \hat{\phi}_{x'x}^{(\xi)_{x'x}-1} \quad b_Y(y)(\hat{x}_{-1}) := \delta_y(\hat{x}_{-1}), \tag{17}$$

where $B(\xi)$ is the beta function.

Using these two distributions as building blocks we can define a third function $b_{Y,\Xi} : \mathcal{Y} \times \Xi \to \Delta_{\Theta \times \mathcal{X}^{\mathbb{N}}}$.

$$b_{Y,\Xi}(y, \xi)(\hat{\phi}, \hat{x}_{-1:\infty}) : = \frac{1}{B(\xi)} \prod_{x'x} \hat{\phi}_{x'x}^{(\xi + c(x_{-1:\infty}))_{x'x}-1} \delta_y(x_{-1}). \tag{18}$$

where

$$c(x_{0:t}) := \sum_{\tau=1}^{t-1} \delta_{x_\tau} \otimes \delta_{x_\tau - 1}. \tag{19}$$

With this we can define the dependence of Q_t on the external variables (y_t, ξ_t):

$$p(q_t|y_t, \xi_t) := \delta_{b_{Y,\Xi}(y_t,\xi_t)}(q_t). \tag{20}$$

With this the Bayesian network of Figure 3 is fully defined.

This shows that Q_t for each t is a function of the pair (y_t, ξ_t) and therefore a coarse-graining of the Markov chain state. This highlights the virtual or interpretational nature of the beliefs in this setting. They have no consequence for the next state of the Markov chain.

Acknowledgments

The work by MB and RK on this publication was made possible through the support of a grant from Templeton World Charity Foundation, Inc. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation, Inc.

References

1. Friston, K.: A free energy principle for a particular physics. arXiv:1906.10184 [q-bio] (Jun 2019), http://arxiv.org/abs/1906.10184, arXiv: 1906.10184
2. Minka, T.: Bayesian inference, entropy, and the multinomial distribution. Online tutorial (2003), https://tminka.github.io/papers/minka-multinomial.pdf