Hierarchical Gaussian Filtering of Sufficient Statistic Time Series for Active Inference

Christoph Mathys[1,2,3(✉)] and Lilian Weber[3,4]

[1] Interacting Minds Centre, Aarhus University, Aarhus, Denmark
chmathys@cas.au.dk
[2] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy
[3] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland
[4] University of Oxford, Oxford, UK

Abstract. Active inference relies on state-space models to describe the environments that agents sample with their actions. These actions lead to state changes intended to minimize future surprise. We show that surprise minimization relying on Bayesian inference can be achieved by filtering of the sufficient statistic time series of exponential family input distributions, and we propose the hierarchical Gaussian filter (HGF) as an appropriate, efficient, and scalable tool for active inference agents to achieve this.

Keywords: Active inference · Exponential families · Message passing · Precision-weighted prediction errors · Hierarchical Gaussian filter

1 Introduction

Active inference [3] is a framework for modelling and programming the behaviour of agents negotiating their continued existence in a given environment. Under active inference, an agent chooses its actions such that they minimize the free energy of its model of the environment. In order to do this, the agent needs to perform inference on the state of the environment and its own internal control states which generate actions.

The agent performing active inference and the researcher modelling such an agent have a converging interest in a simple, modular, and automated algorithm that allows them to perform free energy minimization with complex hierarchical models. Accordingly, there have recently been advances in developing an automated algorithmic framework for free energy minimization in active inference [1,7].

In this paper, we are concerned with the filtering of environmental input which reaches the agent through its Markov blanket. We show that exponential-family input distributions can be inferred by tracking the mean of the sufficient statistics of the inputs by passing simple update messages which amount

T. Verbelen et al. (Eds.): IWAI 2020, CCIS 1326, pp. 52–58, 2020.
https://doi.org/10.1007/978-3-030-64919-7_7

to precision-weighted prediction errors. For stationary input distributions, this implements exact Bayesian inference. In the more common case of non-stationary input distributions, we propose to apply hierarchical Gaussian filtering [4,5] to the sufficient statistic time series, resulting in approximate Bayesian inference with a dynamic learning rate.

2 Bayesian Inference Reduced to Mean-Tracking

2.1 Mean Tracking and Exponential Weighting

As a preliminary, we note that the arithmetic mean $\bar{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i$ of a time series $\{x_1, x_2, \ldots, x_n\}$ can be updated sequentially from $\bar{x}_n$ to $\bar{x}_{n+1}$ when a new observation x_{n+1} occurs.

$$\bar{x}_{n+1} = \bar{x}_n + \frac{1}{n+1}\left(x_{n+1} - \bar{x}_n\right) \tag{1}$$

If we take the previous mean $\bar{x}_n$ to be a prediction for the new observation x_{n+1}, then the difference $x_{n+1} - \bar{x}_n$ is a *prediction error*. The update to $\bar{x}_n$ then amounts to adding the prediction error weighted by $1/(n+1)$. As n grows, the weight of prediction errors approaches zero, which ensures the equal weighting of all observations in the mean.

As a further preliminary, we note that if we replace the weight $1/(n+1)$ of the prediction error with a constant *learning rate* $\alpha \in [0, 1]$, we no longer get the mean $\bar{x}_n$ of the time series but the exponentially weighted average q_n.

$$q_{n+1} = q_n + \alpha\left(x_{n+1} - q_n\right) \tag{2}$$

With $q_0 := 0$ and $\gamma := 1 - \alpha$, this can be written in closed form,

$$q_n = (1-\gamma)\sum_{i=0}^{n-1} \gamma^i x_{n-i}, \tag{3}$$

which makes apparent the exponential downweighting of observations x_i as they lie further in the past.

2.2 A Conjugate Prior Which Reduces Bayesian Inference to Mean Tracking for Exponential Families

Exponential families of probability distributions are those which can be written in the form

$$p\left(\boldsymbol{x}|\boldsymbol{\vartheta}\right) = f_{\boldsymbol{x}}(\boldsymbol{\vartheta}) := h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{t}(\boldsymbol{x}) - b(\boldsymbol{\vartheta})\right), \tag{4}$$

where $\boldsymbol{x}$ is a (possibly) vector-valued observation, $\boldsymbol{\vartheta}$ is a parameter vector, $h(\boldsymbol{x})$ is a normalization constant, $\boldsymbol{\eta}(\boldsymbol{\vartheta})$ is the so-called 'natural' parameter vector,

$\boldsymbol{t}(\boldsymbol{x})$ is the sufficient statistic vector, and $b(\boldsymbol{\vartheta})$ is a scalar function. If we choose as our prior

$$p\left(\boldsymbol{\vartheta}|\boldsymbol{\xi},\nu\right) = g_{\boldsymbol{\xi},\nu}(\boldsymbol{\vartheta}) := z\left(\boldsymbol{\xi},\nu\right)\exp\left(\nu\left(\boldsymbol{\eta}\left(\boldsymbol{\vartheta}\right)\cdot\boldsymbol{\xi} - b(\boldsymbol{\vartheta})\right)\right), \tag{5}$$

where $\boldsymbol{\xi}$ is a hyperparameter vector, $\nu > 0$ a scalar hyperparameter, and $z(\boldsymbol{\xi},\nu)$ the normalization constant

$$z(\boldsymbol{\xi},\nu) := \left(\int \exp\left(\nu\left(\boldsymbol{\eta}\left(\boldsymbol{\vartheta}\right)\cdot\boldsymbol{\xi} - b(\boldsymbol{\vartheta})\right)\right)\mathrm{d}\boldsymbol{\vartheta}\right)^{-1}, \tag{6}$$

then the posterior has the same form as the prior (i.e., it is *conjugate*) with updated hyperparameters

$$\nu \leftarrow \nu + 1 \tag{7}$$

$$\boldsymbol{\xi} \leftarrow \boldsymbol{\xi} + \frac{1}{\nu+1}\left(\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{\xi}\right). \tag{8}$$

A proof of this is in the Appendix.

In other words, with the prior introduced in Eq. 5, Bayesian inference with exponential family models reduces to tracking the mean of the sufficient statistic $\boldsymbol{t}(\boldsymbol{x}_i)$ of the observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$. For a single observation $\boldsymbol{x}$, inference amounts to updating the hyperparameter $\boldsymbol{\xi}$ with the sufficient statistic $\boldsymbol{t}(\boldsymbol{x})$ under the assumption that there have been ν previous observations with sufficient statistic $\boldsymbol{\xi}$.

3 Predictive Distributions

Agents performing active inference minimize the free energy of their model of the environment by minimizing prediction errors regarding their observations (in the long run; in the short run, it is necessary to risk surprises that won't kill us in order to gain the information needed to avoid being dead in the long run). Therefore, the decisive goal and outcome of model-based inference is the *predictive distribution* $\hat{f}$ of inputs $\boldsymbol{x}$. In the present framework, this is

$$\hat{f}_{\boldsymbol{\xi},\nu}(\boldsymbol{x}) := \int f_{\boldsymbol{x}}(\boldsymbol{\vartheta}) g_{\boldsymbol{\xi},\nu}(\boldsymbol{\vartheta})\mathrm{d}\boldsymbol{\vartheta}. \tag{9}$$

For the univariate Gaussian with unknown mean and precision, we will call this the *Gaussian-predictive* distribution $\mathcal{NP}$:

$$\begin{aligned}\hat{f}_{\boldsymbol{\xi},\nu}(\boldsymbol{x}) &= \mathcal{NP}\left(x;\boldsymbol{\xi},\nu\right)\\ &:= \sqrt{\frac{1}{\pi(\nu+1)\left(\xi_{x^2} - \xi_x^2\right)}}\frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)}\\ &\quad\left(1 + \frac{(x-\xi_x)^2}{(\nu+1)\left(\xi_{x^2} - \xi_x^2\right)}\right)^{-\frac{\nu+2}{2}}\end{aligned} \tag{10}$$

For $\xi_x = 0$ and $\xi_{x^2} = 1$, this becomes a Student's-t distribution with $\nu+1$ degrees of freedom. Figure 1 shows how the Gaussian-predictive distribution $\mathcal{NP}$ works in practice, i.e., how it adapts as ν, ξ_x, and ξ_{x^2} are updated sequentially according to Eqs. 7 and 8.

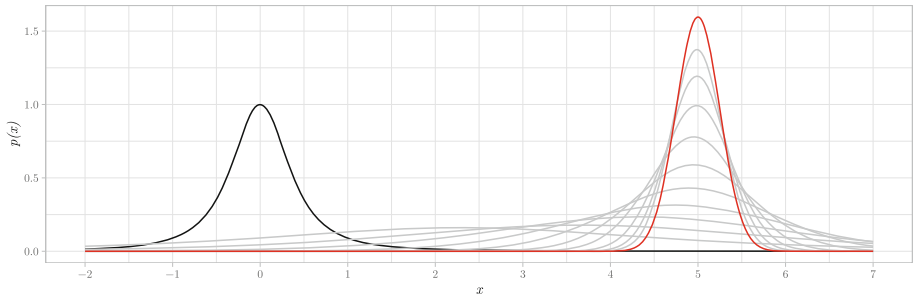

Fig. 1. Sequential updates to the Gaussian-predictive distribution $\mathcal{NP}$ in response to 1024 samples drawn from a Gaussian with mean 5 and standard deviation 1/4 (red). Initial hyperparameters were $\xi_x = 0$, $\xi_{x^2} = 1/8$, and $\nu = 1$, corresponding to the initial $\mathcal{NP}$ in black. Updated predictive distributions after $2, 4, 8, \ldots, 1024$ samples are shown in grey. (Color figure online)

4 Filtering of Sufficient Statistics for Non-stationary Input Distributions

Active inference agents find themselves in environments where the distributions underlying their observations are non-stationary. In such a setting, older observations have less value for inference about the present than newer ones. Using the hyperparameter update scheme introduced above is then inappropriate because it leads to predictive distributions which rely on outdated information and are overconfident because they overestimate the amount of good information they have. However, since our update scheme relies on tracking the mean of the sufficient statistics of the observations, that is, simply on filtering the sufficient statistic time series, we can apply any known filtering method to this time series and use its output to construct predictive distributions. For example, instead of applying Eq. 1, we could use Eq. 2, which amounts to an exponential down-weighting of observations into the past. Using a constant learning rate in this way corresponds to holding ν constant in Eq. 8. As is evident from Eq. 10, this means that the predictive distribution retains its fat tails, meaning that an agent will experience much less surprise at observations far from the predictive mean. However, keeping ν constant raises the question what value to choose for it, and when to change it.

A solution to this is the application of a hierarchical Gaussian filter (HGF) [4,5] to the sufficient statistic time series. The HGF, which contains the Kalman filter as a special case, allows for filtering with an adaptive learning rate

which is adjusted according to a continually updated prediction about the volatility of the environment. Updates in the HGF are precision-weighted prediction errors derived from a hierarchical volatility model by variational approximation. For example, in the case of a Gaussian input distribution as in Fig. 1, input x would be filtered by an HGF, allowing for a posterior predictive distribution that dynamically adapts to a volatile input distribution. Figure 2 shows an example of how this procedure yields an adaptive ν, which falls in response to changes in the input distribution and so ensures that the predictive distribution remains fat-tailed at all times.

Fig. 2. Example of a time series (input x, top panel, fine blue line) filtered with an HGF (posterior mean ξ_x, top panel, red line), which infers the ground truth μ (top panel, yellow line) well in a volatile environment. Comparison of the HGF updates with Eq. 8 yields implied ν (bottom panel). This never rises above 12, ensuring a fat-tailed predictive distribution. In stable phases, implied ν rises; in volatile phases, it falls. (Color figure online)

5 Discussion

We have shown a way to do exact Bayesian inference with exponential-family models simply by tracking the mean of the sufficient statistic function as observations occur. For this to work, the prior introduced in Eq. 5 is crucial, but its

significance has not been recognized before. The approach introduced here is novel. While our prior appears in [2] and seems to have been forgotten since, the resulting updates are there written in a form that obscures their meaning as (precision-)weighted prediction errors and makes it obvious that the relation to mean-tracking was not seen. However, once this is apparent, it supports a *filtering perspective on hyperparameter updates,* which opens up new possibilities such as the HGF filtering proposed in Sect. 4. Additionally, our prior has the benefit of a ready interpretation: ν virtual previous observations with sufficient statistic $\boldsymbol{\xi}$.

For active inference agents, it is critical to predict observations in a way that allows for non-stationary generative processes in the environment. In the framework we propose, this can be achieved by filtering the sufficient statistics of the input distribution using an HGF. This allows predictive distributions to keep a shape (precise but fat-tailed and able to adapt quickly in response to prediction errors) that optimally serves the purpose of minimizing surprise in the long run.

This perspective can be expanded to include networks of HGF nodes where the input distribution and its associated filter are the window into the deeper layers of the network. These deeper layers encode the agent's model of its environment, and it is the free energy of this model that the agent endeavours to minimize by active inference. The present work is therefore a natural complement to recent work on an automated algorithmic framework for free energy minimization in active inference [1,6,7]. The simple message-passing nature of the hyperparameter updates we are proposing fits naturally into message passing schemes in deep networks.

Appendix: Proof of Eqs. 7 and 8

By Bayes' theorem we have

$$\begin{aligned} p\left(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{\xi},\nu\right) &\propto p\left(\boldsymbol{x}|\boldsymbol{\vartheta}\right) p\left(\boldsymbol{\vartheta}|\boldsymbol{\xi},\nu\right) \\ &= f_{\boldsymbol{x}}(\boldsymbol{\vartheta}) g_{\boldsymbol{\xi},\nu}(\boldsymbol{\vartheta}) \\ &= h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{t}(\boldsymbol{x}) - b(\boldsymbol{\vartheta})\right) \\ &\qquad\qquad z\left(\boldsymbol{\xi},\nu\right) \exp\left(\nu\left(\boldsymbol{\eta}\left(\boldsymbol{\vartheta}\right) \cdot \boldsymbol{\xi} - b(\boldsymbol{\vartheta})\right)\right) \\ &\propto \exp\left(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \left(\boldsymbol{t}(\boldsymbol{x}) + \nu\boldsymbol{\xi}\right) - (\nu+1) b(\boldsymbol{\vartheta})\right) \end{aligned}$$

We only need to prove that the argument of the exponential function has the required form. Normalization takes care of the rest. Rearranging the argument gives us

$$\begin{aligned} &\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \left(\boldsymbol{t}(\boldsymbol{x}) + \nu\boldsymbol{\xi}\right) - (\nu+1) b(\boldsymbol{\vartheta}) \\ &\qquad = (\nu+1)\left(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \frac{1}{\nu+1}\left(\boldsymbol{t}(\boldsymbol{x}) + \nu\boldsymbol{\xi}\right) - b(\boldsymbol{\vartheta})\right) \\ &\qquad\qquad = (\nu+1)\left(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \left(\boldsymbol{\xi} + \frac{1}{\nu+1}\left(\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{\xi}\right)\right) - b(\boldsymbol{\vartheta})\right). \end{aligned}$$

From this, it follows that

$$p\left(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{\xi},\nu\right)=g_{\boldsymbol{\xi}',\nu'}(\boldsymbol{\vartheta})$$

with

$$\begin{aligned}\nu' &= \nu+1\\ \boldsymbol{\xi}' &= \boldsymbol{\xi}+\frac{1}{\nu+1}\left(\boldsymbol{t}(\boldsymbol{x})-\boldsymbol{\xi}\right)\end{aligned}$$

References

1. de Vries, B., Friston, K.J.: A factor graph description of deep temporal active inference. Front. Comput. Neurosci. **11** (2017). https://doi.org/10.3389/fncom.2017.00095
2. Diaconis, P., Ylvisaker, D.: Conjugate priors for exponential families. Ann. Stat. **7**(2), 269–281 (1979)
3. Friston, K.J., Daunizeau, J., Kiebel, S.J.: Reinforcement learning or active inference? PLoS ONE **4**(7), e6421 (2009). https://doi.org/10.1371/journal.pone.0006421
4. Mathys, C., Daunizeau, J., Friston, K.J., Stephan, K.E.: A Bayesian foundation for individual learning under uncertainty. Front. Hum. Neurosci. **5**, 39 (2011). https://doi.org/10.3389/fnhum.2011.00039
5. Mathys, C., et al.: Uncertainty in perception and the Hierarchical Gaussian Filter. Front. Hum. Neurosci. **8**, 825 (2014). https://doi.org/10.3389/fnhum.2014.00825
6. Şenöz, İ., de Vries, B.: Online variational message passing in the hierarchical Gaussian filter. In: 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6 (Sep 2018). https://doi.org/10.1109/MLSP.2018.8517019
7. van de Laar, T.W., de Vries, B.: Simulating active inference processes by message passing. Front. Robot. AI **6** (2019). https://doi.org/10.3389/frobt.2019.00020