Journal of Mathematical Psychology 125 (2025) 102921

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Review

A concise mathematical description of active inference in discrete time

Jesse van Oostrum [a,*], Carlotta Langer [a], Nihat Ay [a,b]

[a] Institute for Data Science Foundations, Hamburg University of Technology, 21073 Hamburg, Germany
[b] Santa Fe Institute, Santa Fe, NM 87501, USA

ARTICLE INFO

Keywords:
Active inference
Free energy principle
Bayesian inference
Tutorial
Mathematical review

ABSTRACT

In this paper we present a concise mathematical description of active inference in discrete time. The main part of the paper serves as a basic introduction to the topic, including a detailed example of the action selection mechanism. The appendix discusses the more subtle mathematical details, targeting readers who have already studied the active inference literature but struggle to make sense of the mathematical details and derivations. Throughout, we emphasize precise and standard mathematical notation, ensuring consistency with existing texts and linking all equations to widely used references on active inference. Additionally, we provide Python code that implements the action selection and learning mechanisms described in this paper and is compatible with `pymdp` environments.

Contents

Introduction

Active inference is a theory that describes the action selection and learning mechanisms of an agent in an environment. We aim to present a concise mathematical description of the theory so that readers interested in the mathematical details can quickly find what they are looking for. We have paid special attention to choosing notation that is more in line with standard mathematical texts and is also descriptive, in the sense that dependencies are made explicit. Hence, the focus of this paper lies on the mathematical details and derivations rather than verbal motivations and justifications.

The paper consists of a main text and an appendix. The main text provides a clear introduction to active inference in discrete time, accessible for people new to the topic. It is divided into two parts: inference, which assumes a given generative model, and learning, which explains how the agent acquires this model. The main text concludes with a worked example of action selection. The appendix delves into finer details and derivations, catering to readers familiar with active inference who seek clarity on the mathematical aspects.

To complement our theoretical exposition, we provide a Python implementation[1] of the action selection and learning mechanisms described in this paper, which is compatible with `pymdp` environments.

* Corresponding author.
E-mail address: jesse.van@tuhh.de (J. van Oostrum).
[1] https://github.com/jessevoostrum/active-inference

https://doi.org/10.1016/j.jmp.2025.102921
Received 25 September 2024; Received in revised form 9 April 2025; Accepted 10 April 2025
Available online 16 May 2025

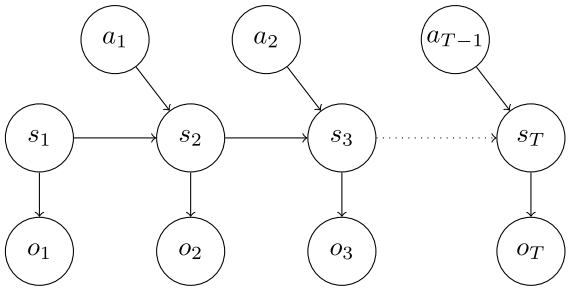

Fig. 1. Graphical representation of the generative model.

This code is more minimalistic, which makes it easier to understand than other implementations such as `SPM` and `pymdp`.

1. Set-up and notation

In this paper we consider an active inference agent acting in a discrete-time setting with a finite-time horizon. This means that we consider a sequence of T time steps and at every time step τ the agent receives an observation o_τ, and performs an action a_τ. We use τ for arbitrary time steps and the letter t to denote the current time step. We use the subscript $_{\tau:\tau'}$ to denote a sequence of variables, e.g. $o_{\tau:\tau'} = (o_\tau, \dots, o_{\tau'})$. A sequence of (future) actions is called a *policy*[2] and is denoted by $\pi_t = a_{t:T}$, with $\pi = \pi_1$. We write $a_{1:t-1}$ for actions that were performed in the past and π_t for future actions that still need to be selected.

The agent models the dynamics of the environment using an internal generative model. This model uses a variable s_τ, called an internal state, to represent the state of the environment[3] at time step τ. The model is given by the following probability distribution:

$$p(o_{1:T}, s_{1:T} | a_{1:T-1}, \theta).$$

This probability distribution factorizes according to the graph in Fig. 1. In the first part of this paper we assume that this generative model is given and need not be learned, and we therefore suppress the dependence on the parameter θ. In the second part we discuss how the model is learned.

Suppose the agent is at time step t. It will have received observations $o_{1:t}$ and performed actions $a_{1:t-1}$. We use $q_t(s_{\tau:\tau'})$ to denote the (approximate) posterior distribution of the generative model, $p(s_{\tau:\tau'} | o_{1:t}, a_{1:t-1})$, and also refer to this as the belief of the agent about the variable $s_{\tau:\tau'}$.

2. Inference

2.1. Action selection according to active inference

Let an agent be at time step t, having received observations $o_{1:t}$ and performed actions $a_{1:t-1}$. According to active inference an agent selects its next action by sampling a policy π_t from the following distribution (equation (10) in Da Costa, Parr, Sajid, Veselic, Neacsu, and Friston (2020)):

$$\sigma\left(-G(\pi_t | o_{1:t}, a_{1:t-1})\right) \tag{1}$$

and selecting the action a_t corresponding to that policy.[4] The function σ denotes the softmax function defined in (48) and G is the expected free energy function given by

$$\begin{aligned} G(\pi_t | o_{1:t}, a_{1:t-1}) \\ = -\Big(\mathbb{E}_{q_t(o_{t+1:T}|\pi_t)} \Big[D_{\mathrm{KL}}\Big(q_t(s_{t+1:T} | o_{t+1:T}, \pi_t) \parallel q_t(s_{t+1:T} | \pi_t) \Big) \Big] \\ + \mathbb{E}_{q_t(o_{t+1:T}|\pi_t)} \Big[\ln p_C(o_{t+1:T}) \Big] \Big) . \end{aligned} \tag{2}$$

Note that according to (1) the agent is more likely to sample policies π_t that have a low expected free energy $G(\pi_t, o_{1:t}, a_{1:t-1})$. In Section 2.2 we discuss equivalent formulations and different interpretations of the expected free energy. The distributions $q_t(o_{t+1:T}|\pi_t)$, $q_t(s_{t+1:T}|o_{t+1:T}, \pi_t)$, $q_t(s_{t+1:T}|\pi_t)$, needed for the calculation of G, are (approximate) posterior distributions of the generative model of the agent after having observed $o_{1:t}$ and performed $a_{1:t-1}$. In Section 2.3 we describe how the agent infers these posterior distributions. The distribution p_C is a preference distribution over observations that we assume is given to the agent. This distribution is distinct from the generative model p.

Remark 1. Note that in certain descriptions of active inference in discrete time also a variational free energy term F appears in the distribution in Eq. (1) (e.g. equation (B.9) in Parr, Pezzulo, and Friston (2022)). This term is only relevant in specific cases that we will discuss in Remark 2 in Appendix A. Furthermore a habit term E is sometimes included that is also considered in Appendix A, but discarded here for simplicity.

2.2. Expected free energy

Recall that Eq. (2) gives the following expression for the expected free energy function:

$$\begin{aligned} G(\pi_t | o_{1:t}, a_{1:t-1}) \\ = -\Big(\mathbb{E}_{q_t(o_{t+1:T}|\pi_t)} \Big[D_{\mathrm{KL}}\Big(q_t(s_{t+1:T} | o_{t+1:T}, \pi_t) \parallel q_t(s_{t+1:T} | \pi_t) \Big) \Big] \\ + \mathbb{E}_{q_t(o_{t+1:T}|\pi_t)} \Big[\ln p_C(o_{t+1:T}) \Big] \Big) . \end{aligned}$$

The first term between the brackets on the RHS is called *epistemic value* or *information gain*. It measures the average change in belief about the future states $s_{t+1:T}$ due to receiving future observations $o_{t+1:T}$. The second term, known as *utility*, quantifies the similarity between the expected future observation distribution and the preferred observation distribution. As previously mentioned, the agent is more likely to sample policies with low expected free energy, which correspond to high information gain and utility.

An equivalent formulation of expected free energy is given by

$$\begin{aligned} G(\pi_t | o_{1:t}, a_{1:t-1}) = \mathbb{E}_{q_t(s_{t+1:T}|\pi_t)} \Big[\mathrm{H}\big[p(o_{t+1:T} | s_{t+1:T}) \big] \Big] \\ + D_{\mathrm{KL}}\Big(q_t(o_{t+1:T} | \pi_t) \parallel p_C(o_{t+1:T}) \Big). \end{aligned} \tag{3}$$

The first term on the RHS is referred to as *ambiguity*. It measures the average uncertainty an agent has about its future observations given knowledge of its future states. The second term is called *expected complexity* or *risk*. It represents the divergence between expected and preferred future observations. The agent favors policies with low ambiguity and risk.

In Appendix C we show that both expressions of the expected free energy are equal.

[2] Note that in a reinforcement learning context the term "policy" has a different meaning.

[3] Note that the number of possible internal states is usually much smaller than the actual number of states the environment can be in.

[4] Note that at every time step a new policy is sampled, only the action a_t for the corresponding time step t is executed, and the rest of the actions in π_t are discarded.

In practice, often the following mean field approximations are made:

$$q_t(s_{t+1:T}) = \prod_{\tau=t+1}^{T} q_t(s_\tau),$$
$$p_C(o_{t+1:T}) = \prod_{\tau=t+1}^{T} p_C(o_\tau).$$

Eqs. (2) and (3) can then be written as follows:

$$G(\pi_t|o_{1:t}, a_{1:t-1}) = \sum_{\tau=t+1}^{T} G_\tau(\pi_t, o_{1:t}, a_{1:t-1}), \tag{4}$$

with

$$\begin{aligned} G_\tau(\pi_t|o_{1:t}, a_{1:t-1}) &= -\left(\mathbb{E}_{q_t(o_\tau|\pi_t)}\left[D_{\mathrm{KL}}\Big(q_t(s_\tau|o_\tau, \pi_t) \parallel q_t(s_\tau|\pi_t) \Big) \right] \right. \\ &\quad \left. + \mathbb{E}_{q_t(o_\tau|\pi_t)}\Big[\ln p_C(o_\tau) \Big] \right) \\ &= \mathbb{E}_{q_t(s_\tau|\pi_t)}\Big[\mathrm{H}\big[p(o_\tau|s_\tau) \big] \Big] + D_{\mathrm{KL}}\Big(q_t(o_\tau|\pi_t) \parallel p_C(o_\tau) \Big). \end{aligned} \tag{5}$$

It is outside the scope of this paper to further derive or motivate the expected free energy. We refer the reader to Appendix B.2.5 in Parr et al. (2022) and Da Costa, Tenka, Zhao, and Sajid (2024), Millidge, Tschantz, and Buckley (2021), Wei (2024) for more details.

2.3. State inference

In this section we describe the simplest form of state inference, which is obtained by applying Bayes' rule. State inference methods as described in e.g. Heins et al. (2022), Parr et al. (2022), Smith, Friston, and Whyte (2022) can be thought of as computationally efficient approximations of what is described here. See Appendix A for more details on these methods.

Above we defined q_t to be the (approximate) posterior of the generative model given $o_{1:t}$, $a_{1:t-1}$. In this section we make the conditioning variables explicit and write $q(\cdot|o_{1:t}, a_{1:t-1})$ instead.

Current and future state inference

We start by studying the generative model that is assumed to be given to the agent. The generative model can be decomposed as follows: (see Fig. 1)

$$\begin{aligned} p(o_{1:T}, s_{1:T}|a_{1:T-1}) &= p(s_{1:T}|a_{1:T-1}) p(o_{1:T}|s_{1:T}) \\ p(s_{1:T}|a_{1:T-1}) &= p(s_1) \prod_{\tau=2}^{T} p(s_\tau|s_{\tau-1}, a_{\tau-1}) \\ p(o_{1:T}|s_{1:T}) &= \prod_{\tau=1}^{T} p(o_\tau|s_\tau). \end{aligned}$$

At every time step the agent updates its belief about the current state it is in. Before having performed any observations, its belief about the current state is equal to the prior belief $p(s_1)$. After receiving o_1 it will update its belief using Bayes' rule:

$$q(s_1|o_1) \propto p(o_1|s_1) p(s_1).$$

Subsequently, it will perform an action a_1 (selected as described in Section 2.1) and receive a next observation o_2. The belief about s_2 is given by

$$\begin{aligned} q(s_2|o_{1:2}, a_1) &\propto p(o_2|s_2, o_1, a_1) p(s_2|o_1, a_1) \\ &= p(o_2|s_2) \sum_{s_1} p(s_2|s_1, o_1, a_1) p(s_1|a_1, o_1) \\ &= p(o_2|s_2) \sum_{s_1} p(s_2|s_1, a_1) q(s_1|o_1). \end{aligned}$$

For a general t the belief about s_t is updated as follows:

$$q(s_t|o_{1:t}, a_{1:t-1}) \propto p(o_t|s_t) \sum_{s_{t-1}} p(s_t|s_{t-1}, a_{t-1}) q(s_{t-1}|o_{1:t-1}, a_{1:t-2}). \tag{6}$$

For future time point $\tau > t$, the belief about the state s_τ is given by

$$q(s_\tau|o_{1:t}, a_{1:\tau-1}) = \sum_{s_{\tau-1}} p(s_\tau|s_{\tau-1}, a_{\tau-1}) q(s_{\tau-1}|o_{1:t}, a_{1:\tau-2}). \tag{7}$$

The belief about a future state given a future observation is calculated as follows:

$$q(s_\tau|o_\tau, o_{1:t}, a_{1:\tau-1}) \propto p(o_\tau|s_\tau) q(s_\tau|o_{1:t}, a_{1:\tau-1}). \tag{8}$$

Future observation inference

In order to compute G in (2), we need a posterior distribution $q(o_\tau|o_{1:t}, a_{1:\tau-1})$ over future observations o_τ. This can be computed as follows:

$$q(o_\tau|o_{1:t}, a_{1:\tau-1}) = \sum_{s_\tau} p(o_\tau|s_\tau) q(s_\tau|o_{1:t}, a_{1:\tau-1}). \tag{9}$$

Past, current, and future state inference

In order to perform inference over states in the past, present and future (which is needed for the learning of the generative model and for the computation of the variational free energy over states), the agent can use the following formula:

$$\begin{aligned} q(s_{1:T}|o_{1:t}, a_{1:t-1}, \pi_t) &\propto p(o_{1:t}|s_{1:T}, a_{1:t-1}, \pi_t) p(s_{1:T}|a_{1:t-1}, \pi_t) \\ &= p(o_{1:t}|s_{1:t}) p(s_{1:T}|a_{1:t-1}, \pi_t) \\ &= \prod_{\tau=1}^{t} p(o_\tau|s_\tau)\ p(s_1) \prod_{\tau=2}^{T} p(s_\tau|s_{\tau-1}, a_{\tau-1}), \end{aligned} \tag{10}$$

where we use $a_{\tau-1}$ in the last term for elements of both $a_{1:t-1}$ and π_t.

3. Learning

In the above section, we have assumed that the agent has access to a generative model $p(o_{1:T}, s_{1:T}|a_{1:T-1}, \theta)$. In this section we discuss how the parameter θ of this model is learned. The generative model consists of three (conditional) categorical distributions that are parametrized by $\theta = (\theta^D, \theta^A, \theta^B)$ in the following way:

$$p(s_1^{(j)}|\theta^D) = \theta^D_j, \tag{11}$$
$$p(o_\tau^{(i)}|s_\tau^{(j)}, \theta^A) = \theta^A_{ij}, \tag{12}$$
$$p(s_\tau^{(j)}|s_{\tau-1}^{(k)}, a_{\tau-1}^{(l)}, \theta^B) = \theta^B_{jkl}, \tag{13}$$

where we have enumerated the elements of the observation, action and state space with the bracketed superscript $^{(\cdot)}$. In order to learn the parameters of the generative model, we adopt a Bayesian belief updating scheme with a Dirichlet prior. (See Appendix B for details on this.) More specifically, the prior over θ is parametrized by $\alpha = (\alpha^D, \alpha^A, \alpha^B)$ and is given by

$$p(\theta|\alpha) = p(\theta^D|\alpha^D) \prod_j p(\theta^A_{\bullet j}|\alpha^A) \prod_{k,l} p(\theta^B_{\bullet kl}|\alpha^B) \tag{14}$$

$$p(\theta^D|\alpha^D) \propto \prod_j \left(\theta^D_j \right)^{\alpha^D_j - 1}, \tag{15}$$

$$p(\theta^A_{\bullet j}|\alpha^A) \propto \prod_i \left(\theta^A_{ij} \right)^{\alpha^A_{ij} - 1}, \tag{16}$$

$$p(\theta^B_{\bullet kl}|\alpha^B) \propto \prod_j \left(\theta^B_{jkl} \right)^{\alpha^B_{jkl} - 1}, \tag{17}$$

where $\theta_{\bullet j}$ denotes the vector $(\theta_{1j}, \ldots, \theta_{nj})$.

Now after performing actions $a_{1:T-1}$ and receiving observations $o_{1:T}$ we want to update our belief about θ according to Bayes' rule. The true posteriors over these parameters are not Dirichlet distributions. (See Appendix B for details). The active inference literature suggests to set the approximate posterior distribution to be a Dirichlet distribution and update the hyperparameter α in the following way (equation (B.12)

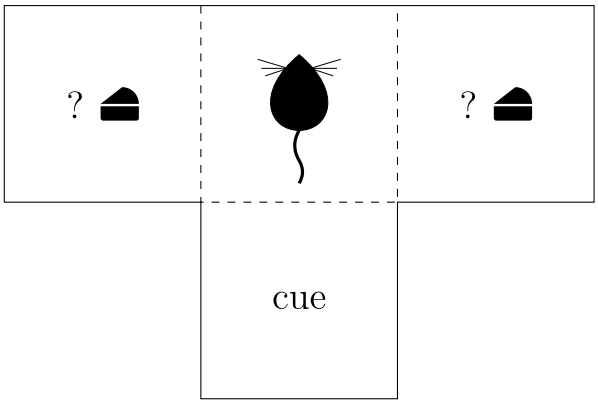

Fig. 2. T-maze environment.

in Parr et al. (2022) and equation (21), (A.6) and (A.7) in Da Costa et al. (2020)):

$$\alpha_j^{D\prime} = \alpha_j^D + q_T\left(s_1^{(j)}\right), \tag{18}$$

$$\alpha_{ij}^{A\prime} = \alpha_{ij}^A + \sum_{\tau=1}^{T} \mathbb{1}_{o^{(i)}}(o_\tau) q_T\left(s_\tau^{(j)}\right), \tag{19}$$

$$\alpha_{jkl}^{B\ \prime} = \alpha_{jkl}^B + \sum_{\tau=2}^{T} q_T\left(s_\tau^{(j)}\right) q_T\left(s_{\tau-1}^{(k)}\right) \mathbb{1}_{a^{(l)}}(a_{\tau-1}). \tag{20}$$

The distributions $q_T\left(s_\tau\right), \tau \in \{1, \ldots, T\}$ are approximate posteriors obtained[5] using the current version of the generative model (before θ has been updated). In Appendix B we elaborate on the origin of this learning rule.

Note the similarity with the standard update rule for Dirichlet priors given in (31). In the standard update rule the element α_{i^*} of the hyperparameter corresponding to the observation $x^{(i^*)}$ is incremented by 1, which makes this observation more likely in the updated distribution. In the updates (18)–(20) the hyperparameters are incremented by the amount of posterior belief in that state or state transition, e.g. α_j^D is incremented by $q\left(s_1^{(j)} | o_{1:T}, a_{1:T-1}\right)$.

In order to go from a Dirichlet distribution $p(\theta|\alpha)$ to an actual value of the parameter that can be used for the generative model, the mean of the distribution can be used, which is given by

$$\begin{aligned}\hat{\theta}_i &= \mathbb{E}_{p(\theta|\alpha)}[\theta_i] \\ &= \frac{\alpha_i}{\sum_j \alpha_j}.\end{aligned} \tag{21}$$

For example, after the learning step, the new distribution over s_1 is given by

$$p(s_1^{(j)} | \hat{\theta}^D) = \frac{\alpha_j^{D\prime}}{\sum_j \alpha_k^{D\prime}}.$$

This concludes the discussion of learning in the context of the active inference framework.

4. Example: T-maze example

In this section we discuss the action selection mechanism of an active inference agent in the T-maze environment depicted in Fig. 2. (See also Section 7.3 of Parr et al. (2022) and Heins (2021).) This illustrates the theory of action selection and state inference that is presented in Section 2.

[5] See Section 2.3.

Description of the (internal) generative model of the agent

State, observation and action spaces

The internal state space of the agent[6] $\mathcal{S}$ has two dimensions,[7] *Location* and *Reward condition*, and can be described as follows:

$$\begin{aligned}\mathcal{S} &= \mathcal{S}^L \times \mathcal{S}^R, \\ \mathcal{S}^L &= \{\texttt{center}, \texttt{right arm}, \texttt{left arm}, \texttt{cue location}\}, \\ \mathcal{S}^R &= \{\texttt{reward on right}, \texttt{reward on left}\}.\end{aligned}$$

A typical element of the state space is written as $s = (s^L, s^R)$.

The observation space $\mathcal{O}$ has three dimensions,[8,9] *Location*, *Reward*, and *Cue*, and can be described as follows:

$$\begin{aligned}\mathcal{O} &= \mathcal{O}^L \times \mathcal{O}^R \times \mathcal{O}^C, \\ \mathcal{O}^L &= \{\texttt{center}, \texttt{right arm}, \texttt{left arm}, \texttt{cue location}\}, \\ \mathcal{O}^R &= \{\texttt{no reward}, \texttt{reward}, \texttt{loss}\}, \\ \mathcal{O}^C &= \{\texttt{cue right}, \texttt{cue left}\}.\end{aligned}$$

A typical element of the observation space is written as $o = (o^L, o^R, o^C)$.

The space of actions $\mathcal{A}$ is described by

$$\begin{aligned}\mathcal{A} = \{&\texttt{move to center}, \texttt{move to right arm}, \\ &\texttt{move to left arm}, \texttt{move to cue location}\}.\end{aligned}$$

Note that these actions are always available, independent of the current location of the agent. A typical element of the action space is written as a.

In the following we use `[dir]` as a placeholder for `left` and `right`, and `[loc]` as a placeholder for the four locations `center`, `right arm`, `left arm` and `cue location`.

Observation kernel $p(o|s)$

We now specify the observation kernel $p(o|s)$ of the generative model of the agent. First note that the observation dimensions are independent, that is

$$p(o|s) = p(o^L|s)p(o^R|s).$$

The beliefs about the location observation given the state are modeled as follows:

$$p(o^L|s) = \begin{cases} 1 & \text{if } o^L = s^L \\ 0 & \text{otherwise,} \end{cases}$$

which implies that the location can be unambiguously inferred from the observation.

The beliefs about the reward observation given the state are modeled as follows:

$$\begin{aligned}p(o^R = \texttt{no reward} | s^L \in \{\texttt{center}, \texttt{cue location}\}, s^R) &= 1, \\ p(o^R = \texttt{no reward} | s^L \notin \{\texttt{center}, \texttt{cue location}\}, s^R) &= 0, \\ p(o^R = \texttt{reward} | s^L = \texttt{[dir] arm}, s^R = \texttt{reward on [dir]}) &= 0.98, \\ p(o^R = \texttt{loss} | s^L = \texttt{right arm}, s^R = \texttt{reward on left}) &= 0.98, \\ p(o^R = \texttt{loss} | s^L = \texttt{left arm}, s^R = \texttt{reward on right}) &= 0.98.\end{aligned}$$

[6] In this example the state of the environment (generative process) is the same as the state space of the internal world model of the agent (generative model). Note that this is in general not the case. In a more realistic setting the state space of the environment will be much more complex than the internal state space.

[7] The dimensions of the state space are sometimes referred to as state factors.

[8] The dimensions of the observation space are sometimes referred to as observation modalities.

[9] Note that we follow here the description from Heins (2021). In Parr et al. (2022) the cue observation is absorbed into the location observation.

This implies that the agent observes no reward when it is in location center or cue location, it observes reward when it is in the same arm as specified by the reward condition with high probability, and it observes loss when it is in the opposite arm of the reward condition with high probability.

The beliefs about the cue observation given the state are modeled as follows:

$$p(o^C = \texttt{cue [dir]} | s^L = \texttt{cue location}, s^R = \texttt{reward on [dir]}) = 1,$$
$$p(o^C = \texttt{cue [dir]} | s^L \in \mathcal{S}^L \setminus \{\texttt{cue location}\}, s^R) = 0.5.$$

This implies that the cue observation is completely informative about the reward condition when the agent is at the cue location, and otherwise independent of the actual reward condition.

Transition dynamics kernel $p(s_{\tau+1}|s_\tau, a_\tau)$

We continue by describing the transition dynamics kernel $p(s_{\tau+1}|s_\tau, a_\tau)$.

$$p(s^L_{\tau+1} = \texttt{[loc]} | s^L_\tau \in \{\texttt{center}, \texttt{cue location}\}, s^R_\tau, a_\tau = \texttt{go to [loc]}) = 1,$$
$$p(s^L_{\tau+1} = \texttt{[dir] arm} | s^L_\tau = \texttt{[dir] arm}, s^R_\tau, a_\tau) = 1.$$

This implies that if the agent is in center or cue location it will be in the location specified by the action in the next time step. If it is in one of the arms however, it will stay there, independent of the choice of action a_τ.

$$p(s^R_{\tau+1} = \texttt{reward on [dir]} | s^R_\tau = \texttt{reward on [dir]}, s^L_\tau, a_\tau) = 1.$$

This implies that the reward condition stays constant throughout the trajectory.

Preference distribution p_C *and prior over states* p_D

The unnormalized preference distribution p_C can be chosen to favor observations with reward and discourage observations with loss as follows:

$$p_C((\texttt{[loc]}, \texttt{no reward}, \texttt{cue [dir]})) = 2,$$
$$p_C((\texttt{[loc]}, \texttt{reward}, \texttt{cue [dir]})) = 3,$$
$$p_C((\texttt{[loc]}, \texttt{loss}, \texttt{cue [dir]})) = 1.$$

Finally we let the prior belief over states p_D be uniform.

Action selection procedure

We will now simulate the trajectory of an agent acting according to active inference. We set the time horizon to $T = 3$, which implies that the policies will have length 2.

Time step 1

The agent starts by receiving an observation $o_1 = (\texttt{center}, \texttt{no reward}, \texttt{cue right})$. It now updates its beliefs about the current state such that

$$q_t(s^L_1 = \texttt{center}, s^R_1 = \texttt{reward on [dir]}) = 0.5.$$

Subsequently it computes its beliefs about future states and observations given a policy using Eqs. (7), (8), (9). For example for $\pi^*_1 = (\texttt{move to cue location}, \texttt{move to left arm})$ we have

$$q_1(s^L_2 = \texttt{cue location}, s^R_2 = \texttt{reward on [dir]} \mid \pi_1 = \pi^*_1) = 0.5,$$
$$q_1(s^L_3 = \texttt{left arm}, s^R_3 = \texttt{reward on [dir]} \mid \pi_1 = \pi^*_1) = 0.5, \quad (22)$$

and

$$q_1(o^L_2 = \texttt{cue location}, o^R_2 = \texttt{no reward}, o^C_2 = \texttt{cue [dir]} \mid \pi_1 = \pi^*_1) = 0.5,$$
$$q_1(o^L_3 = \texttt{left arm}, o^R_3 = \texttt{reward}, o^C_3 = \texttt{cue [dir]} \mid \pi_1 = \pi^*_1) = 0.25,$$
$$q_1(o^L_3 = \texttt{left arm}, o^R_3 = \texttt{loss}, o^C_3 = \texttt{cue [dir]} \mid \pi_1 = \pi^*_1) = 0.25,$$

and for example for $o^*_2 = (\texttt{cue location}, \texttt{no reward}, \texttt{cue left})$ we have

$$q_1(s^L_2 = \texttt{cue location}, s^R_2 = \texttt{reward on left} \mid o_2 = o^*_2, \pi_1 = \pi^*_1) = 1. \quad (23)$$

Note here the reduction of uncertainty about s_2 due to the observation o^*_2, represented by the epistemic value defined in Section 2.2 given by the KL divergence between the distributions (22) and (23).

The agent now computes G and plugs this into Eq. (1) and gets the following posterior distribution over policies:

$a_2 \rightarrow$ / $a_1 \downarrow$	center	right arm	left arm	cue location
center	0.022	0.041	0.041	0.046
right arm	0.041	0.075	0.075	0.083
left arm	0.041	0.075	0.075	0.083
cue location	0.046	0.083	0.083	0.091

and in this scenario it samples a policy $\pi^*_1 = (a^*_1, a^*_2)$ with as first action move to cue location with highest probability.

Time step 2

After having performed action $a^*_1 = \texttt{move to cue location}$, the next observation it receives is $o^*_2 = (\texttt{cue location}, \texttt{no reward}, \texttt{cue right})$. Its belief about the current state is now given by

$$q_2(s^L_2 = \texttt{cue location}, s^R_2 = \texttt{reward on right}) = 1.$$

For instance, when $\pi^*_2 = (\texttt{move to left arm})$, the beliefs about future states are

$$q_2(s^L_3 = \texttt{left arm}, s^R_3 = \texttt{reward on right} \mid \pi_2 = \pi^*_2) = 1,$$

and the observation beliefs

$$q_2(o^L_3 = \texttt{left arm}, o^R_3 = \texttt{reward}, o^C_3 = \texttt{cue [dir]} \mid \pi_2 = \pi^*_2) = 0.01,$$
$$q_2(o^L_3 = \texttt{left arm}, o^R_3 = \texttt{loss}, o^C_3 = \texttt{cue [dir]} \mid \pi_2 = \pi^*_2) = 0.49.$$

Since all uncertainty has already been taken away by the last observation, conditioning on $o^*_3 = (\texttt{left arm}, \texttt{reward}, \texttt{cue left})$ will make no difference to the belief about the state s_3, i.e.

$$q_2(s^L_3 = \texttt{left arm}, s^R_3 = \texttt{reward on right} \mid o_3 = o^*_3, \pi_2 = \pi^*_2) =$$
$$q_2(s^L_3 = \texttt{left arm}, s^R_3 = \texttt{reward on right} \mid \pi_2 = \pi^*_2) = 1,$$

which will cause the epistemic value term in G to be zero.

The agent now calculates G again and obtains the following distribution over policies:

a_2	center	right arm	left arm	cue location
	0.20	0.52	0.08	0.20

and will then sample the policy (move to right arm) with highest probability.

CRediT authorship contribution statement

Jesse van Oostrum: Writing – review & editing, Writing – original draft. **Carlotta Langer:** Writing – review & editing. **Nihat Ay:** Supervision.

Acknowledgments

The authors would like to thank Thomas Parr, Conor Heins, Ryan Smith, Beren Millidge, Pablo Lanillos, Sean Tull, Stephen Mann, Pradeep Kumar Banerjee, Frank Röder and Lance Da Costa for helpful discussions and comments and acknowledge the support of the Deutsche Forschungsgemeinschaft Priority Programme "The Active Self" (SPP 2134).

Appendix A. Variational free energy minimization

Preliminaries

In this section we use x for a general observed variable and z for a general latent variable.

Let $p(x, z) = p(x|z)p(z)$ be a generative model. In order to perform inference over the latent variables after making an observation x, one has to compute the posterior $p(z|x)$. This posterior is often hard to compute directly. One can instead approximate this posterior by finding the distribution $q_x(z)$ in a family of distributions $\mathcal{Q}$ that minimizes the following function:

$$F(\tilde{q}|x) = \sum_z \tilde{q}(z)\left(\ln \tilde{q}(z) - \ln p(x, z)\right),$$

called the variational free energy. Note that we distinguish notationally $\tilde{q}$, which is a generic element of $\mathcal{Q}$ and a variable in F, and q_x, which is the minimizer of F for a fixed x, i.e.

$$q_x = \arg\min_{\tilde{q}\in\mathcal{Q}} F(\tilde{q}|x).$$

Note that when $\mathcal{Q}$ is large enough, for example when z is discrete and $\mathcal{Q}$ is the set of all probability distributions over z, then the minimizer of the free energy is equal to the exact posterior distribution and we have

$$q_x(z) = p(z|x).$$

We can replace $\ln p(x, z)$ in F by a general function f, i.e.

$$F_f(\tilde{q}|x) = \sum_z \tilde{q}(z)\left(\ln \tilde{q}(z) - f(x, z)\right).$$

The minimizer can be found by substituting $g(x, z) = e^{f(x,z)}$ as follows:

$$\begin{aligned} F_f(\tilde{q}, x) &= \sum_z \tilde{q}(z)\left(\ln \tilde{q}(z) - \ln g(x, z)\right) \\ &= \left(\sum_z \tilde{q}(z)\left(\ln \tilde{q}(z) - \ln \frac{g(x, z)}{\sum_{z'} g(x, z')}\right)\right) + \ln \sum_{z'} g(x, z'). \end{aligned}$$

If $\mathcal{Q}$ is again large enough, the minimizer is given by:

$$\begin{aligned} q_x(z) &= \frac{g(x, z)}{\sum_z g(x, z)} \\ &= \frac{e^{f(x,z)}}{\sum_z e^{f(x,z)}} \\ &= \sigma(f(x, z)), \end{aligned} \tag{24}$$

where σ is the softmax function defined in (48) and in the case that $f(x, z) = \ln p(x, z)$ we have

$$\begin{aligned} q_x(z) &= \sigma(\ln p(x, z)) \\ &= p(z|x). \end{aligned}$$

The variational free energy can be written as follows:

$$\begin{aligned} F(\tilde{q}|x) &= D_{\mathrm{KL}}(\tilde{q}(z) \parallel p(z|x)) - \ln p(x) \\ &\geq -\ln p(x), \end{aligned}$$

which shows that the negative of the variational free energy is a lower bound on the evidence (ELBO). By minimizing F w.r.t. $\tilde{q}$ we get the following approximate equality (equation (B.2) in Parr et al. (2022)):

$$F(q_x|x) \approx -\ln p(x). \tag{25}$$

Variational free energy minimization in active inference

Active inference adopts the perspective that perception, action selection and learning can be interpreted as minimizing one single variational free energy function F. In its complete form it can be written as follows:

$$F(\tilde{q}|o_{1:t}, a_{1:t-1}) = \mathbb{E}_{\tilde{q}(s_{1:T},\theta,\pi_t)}\left[\ln \tilde{q}(s_{1:T}, \theta, \pi_t) - \ln p(s_{1:T}, o_{1:t}, \theta, \pi_t|a_{1:t-1})\right].$$

The distributions in the family $\mathcal{Q}$ are assumed to factorize as follows:

$$q(s_{1:T}, \theta, \pi) = q(\theta^D)q(\theta^A)q(\theta^B)q(\pi)\prod_{\tau=1}^{T} q(s_\tau),$$

which is sometimes referred to as the mean-field approximation. Now F can be written as follows:

$$\begin{aligned} F(\tilde{q}|o_{1:t}, a_{1:t-1}) = \mathbb{E}_{\tilde{q}(s_{1:T},\theta,\pi_t)}\Bigg[& \ln \tilde{q}(\pi_t) + \ln \tilde{q}(\theta) + \sum_{\tau=1}^{T} \ln \tilde{q}(s_\tau) - \ln p(\pi_t) \\ & - \ln p(\theta) - \ln p(s_1|\theta^D) - \sum_{\tau=1}^{t} \ln p(o_\tau|s_\tau, \theta^A) \\ & - \sum_{\tau=2}^{T} \ln p(s_\tau|s_{\tau-1}, a_{\tau-1}, \theta^B)\Bigg], \end{aligned} \tag{26}$$

where we use $a_{\tau-1}$ in the last term for elements of both $a_{1:t-1}$ and π_t.

Perception

For studying perception, Eq. (26) can be rewritten as follows:

$$F(\tilde{q}|o_{1:t}, a_{1:t-1}) = \mathbb{E}_{\tilde{q}(\pi_t)}\left[F_{\pi_t}(\tilde{q}|o_{1:t}, a_{1:t-1})\right] + C_{\setminus s}$$

where $C_{\setminus s}$ is independent of $\tilde{q}(s_{1:T})$ and

$$\begin{aligned} & F_{\pi_t}(\tilde{q}|o_{1:t}, a_{1:t-1}) \\ & \quad = \mathbb{E}_{\tilde{q}(s_{1:T})}\left[\ln \tilde{q}(s_{1:T}) - \mathbb{E}_{\tilde{q}(\theta)}[\ln p(s_{1:T}, o_{1:t}|\theta, a_{1:t-1}, \pi_t)]\right]. \end{aligned} \tag{27}$$

Perception according to active inference is minimizing F_{π_t} w.r.t. $\tilde{q}(s_{1:T})$. The minimizer is written $q_t(s_{1:T}|\pi_t) = q(s_{1:T}|\pi_t, o_{1:t}, a_{1:t-1})$.

We can use the factorizing properties to rewrite (27) as follows:

$$\begin{aligned} F_{\pi_t}(\tilde{q}|o_{1:t}, a_{1:t-1}) = & \sum_{\tau=1}^{T} \mathbb{E}_{\tilde{q}(s_\tau)}[\ln \tilde{q}(s_\tau)] - \mathbb{E}_{\tilde{q}(s_1,\theta^D)}[\ln p(s_1|\theta^D)] \\ & - \sum_{\tau=1}^{t} \mathbb{E}_{\tilde{q}(s_\tau,\theta^A)} p(o_\tau|s_\tau, \theta^A) \\ & - \sum_{\tau=2}^{T} \mathbb{E}_{\tilde{q}(s_\tau,s_{\tau-1},\theta^B)} p(s_\tau|s_{\tau-1}, a_\tau, \theta^B), \end{aligned}$$

which is equivalent to equation (6) in Da Costa et al. (2020).[10] One can also get rid of the expectations over θ by replacing them by an estimator $\hat{\theta}$. Then we get

$$\begin{aligned} F_{\pi_t}(\tilde{q}|o_{1:t}, a_{1:t-1}) = & \sum_{\tau=1}^{T} \mathbb{E}_{\tilde{q}(s_\tau)}[\ln \tilde{q}(s_\tau)] - \mathbb{E}_{\tilde{q}(s_1)}[\ln p(s_1|\hat{\theta}^D)] \\ & - \sum_{\tau=1}^{t} \mathbb{E}_{\tilde{q}(s_\tau)} p(o_\tau|s_\tau, \hat{\theta}^A) \\ & - \sum_{\tau=2}^{T} \mathbb{E}_{\tilde{q}(s_\tau,s_{\tau-1})} p(s_\tau|s_{\tau-1}, a_\tau, \hat{\theta}^B), \end{aligned}$$

which is equivalent to (B.4) in Parr et al. (2022).

[10] Note that in Da Costa et al. (2020) only the parameter θ^A is treated as a variable and θ^D and θ^B are considered fixed.

Learning

Updating the θ parameter (learning) happens at the end of an episode ($t = T$). The agent has observed $o_{1:T}$ and performed $a_{1:T-1}$. The variational free energy from Eq. (26) can be written as follows:

$$F(\tilde{q}|o_{1:T},\ a_{1:T-1}) = \mathbb{E}_{\tilde{q}(\theta)}\left[\ln \tilde{q}(\theta) - \mathbb{E}_{q_T(s_{1:T})}[\ln p(o_{1:T}, s_{1:T}, \theta|a_{1:T-1})]\right] + C_{\setminus\theta}, \quad (28)$$

where $C_{\setminus\theta}$ is independent of $\tilde{q}(\theta)$ and we have fixed $q_T(s_{1:T})$ to be the approximate posterior over states, inferred using the current (not-updated) belief $p(\theta)$. If the current belief $p(\theta)$ is a Dirichlet distribution with hyperparameter α, then the minimizer $q_T(\theta)$ of F will also be a Dirichlet distribution with hyperparameter α' as given in (18)–(20). In Appendix B we derive this, and relate it to a well known variational inference algorithm called coordinate ascent variational inference (CAVI).

Action selection

Finally we can also view action selection as the minimization of the variational free energy function (26). We can rewrite this function as follows:

$$\begin{aligned} F(\tilde{q}|o_{1:t}) &= \mathbb{E}_{\tilde{q}(\pi)}\Big[\ln \tilde{q}(\pi) - \ln p(\pi) + \mathbb{E}_{\tilde{q}(s_{1:T})}\left[\ln \tilde{q}(s_{1:T})\right. \\ &\qquad \left. - \ln p(o_{1:t}, s_{1:T}|\pi)\right]\Big] + C_{\setminus\pi} \\ &= \mathbb{E}_{\tilde{q}(\pi)}\left[\ln \tilde{q}(\pi) - \ln p(\pi) + F_\pi(\tilde{q}(s_{1:T})|o_{1:t})\right] + C_{\setminus\pi}, \end{aligned}$$

where F_π is defined in (27), $C_{\setminus\pi}$ is independent of $\tilde{q}(\pi)$, and we replaced the expectation over $\tilde{q}(\theta)$ by an estimator $\hat{\theta}$ and suppress the dependence of the generative model p on $\hat{\theta}$ in the notation. (This is equivalent to the second line in equation (B.7) in Parr et al. (2022).) What is important to note here, is that the agent is trying to infer an action sequence (policy) π of both future and past actions. We have therefore dropped the dependence on $a_{1:t-1}$ in both F and F_π, and instead π is a sequence of action starting at $\tau = 1$ instead of $\tau = t$. In Section 2, we always fixed the past actions to the actions that were actually performed, which is no longer the case here.

We minimize F w.r.t. $\tilde{q}(\pi)$ and $\tilde{q}(s_{1:T})$ and using (24) we get for the minimizers respectively

$$q_t(\pi) = \sigma\left(-\ln p(\pi) + F_\pi(q_t(s_{1:T})|o_{1:t})\right), \quad (29)$$

and $q_t(s_{1:T})$ is the minimizer of F_π. Now we can use

$$p(\pi) = \sigma\left(\ln E(\pi) - G(\pi_t|o_{1:t}, a_{1:t-1})\right),$$

which corresponds to the last line equation (B.7) in Parr et al. (2022).[11] The term $E(\pi)$ is a habit term, signifying what policies the agent is usually exercising. Plugging this back into (29) gives

$$q_t(\pi) = \sigma\left(-\ln E(\pi) + G(\pi_t|o_{1:t}, a_{1:t-1}) + F_\pi(q_t|o_{1:t})\right), \quad (30)$$

which corresponds to equation (B.9) in Parr et al. (2022).

Remark 2. We now try to interpret this derivation conceptually. Note that due to (25) we have the following approximate equality:

$$\begin{aligned} F(\tilde{q}|o_{1:t}) &= \mathbb{E}_{\tilde{q}(\pi)}\left[\ln \tilde{q}(\pi) - \ln p(\pi) + F_\pi(\tilde{q}(s_{1:T}), o_{1:t})\right] + C_{\setminus\pi} \\ &\approx \mathbb{E}_{\tilde{q}(\pi)}\left[\ln \tilde{q}(\pi) - \ln p(\pi) - \ln p(o_{1:t}|\pi)\right] + C_{\setminus\pi} \\ &= \mathbb{E}_{\tilde{q}(\pi)}\left[\ln \tilde{q}(\pi) - \ln p(o_{1:t}, \pi)\right] + C_{\setminus\pi}, \end{aligned}$$

which implies that the minimizer $q_t(\pi)$ is approximately equal to the posterior $p(\pi|o_{1:t})$. In other words, this says that we select the policy that is most probable given the past observations. That is, the agent forgets which past actions it has performed, and tries to infer these based on the past observations. Then it tries to find the most likely sequence of future actions to go with this sequence of past actions. Selecting future actions in this way however only makes sense when certain past action sequences make certain future action sequences more likely. For example, let our action space consist of two actions $\{\texttt{left}, \texttt{right}\}$ and policies consist of sequences of actions of length two. Now suppose that the prior over policies dictates that the agent almost certainly performs the policies $(\texttt{left}, \texttt{left})$ or $(\texttt{right}, \texttt{right})$. This implies that having inferred the first action gives the agent more information about the most likely next action. However, if both next actions are equally likely given a first action, according to the prior, then the likelihood term $p(o_{1:t}|\pi)$ does not have any information about the next action. Note that in the calculation of G the past actions are fixed to the actions that have been performed. Therefore G will not make certain future action sequences more likely based on past possible past action sequences. Therefore, the term F_π in (30) only becomes relevant when the habit term E makes certain future action sequences more likely based on past action sequences.

Appendix B. Derivation of the learning rules

Bayesian belief updating

The learning process of an active inference agent is formulated as Bayesian belief updating over the parameters. In general, Bayesian belief updating can be described as follows. Let θ be the parameter of a model p_θ we want to learn and x the output of this model. We start with a prior belief $p(\theta|\alpha)$ which is parametrized by the hyperparameter α. Now our posterior belief about θ is given by the distribution $p(\theta|x, \alpha)$ which is obtained by Bayes' rule. In some special cases,[12] the posterior distribution belongs to the same parametrized family as the prior, such that $p(\theta|x, \alpha) = p(\theta|\alpha')$. Then the learning can be summarized by the update from α to α'.

Categorical model without latent variables

Now we let the model be a categorical distribution over elements $\{x^{(1)}, \dots, x^{(n)}\}$ parametrized by $\theta = (\theta_1, \dots, \theta_n)$. That is,

$$p(x^{(i)}|\theta) = \theta_i, \quad \forall i \in \{1, \dots, n\}.$$

The prior over the parameter θ is given by the Dirichlet distribution parametrized by the hyperparameter $\alpha = (\alpha_1, \dots, \alpha_n)$. That is,

$$p(\theta|\alpha) \propto \prod_i \theta_i^{\alpha_i - 1}.$$

After observing x^*, the posterior is given by

$$\begin{aligned} p(\theta|x^*, \alpha) &\propto p(x^*|\theta)p(\theta|\alpha) \\ &= \prod_i \theta_i^{\alpha_i - 1 + \mathbb{1}_{x^{(i)}}(x^*)}. \end{aligned}$$

Note that this is again a Dirichlet distribution with hyperparameter α' such that

$$\alpha'_i = \alpha_i + \mathbb{1}_{x^{(i)}}(x^*), \quad \forall i \in \{1, \dots, n\}. \quad (31)$$

That is, the hyperparameter corresponding to the observation x^* is increased by one. This will make this observation more likely in the updated distribution.

[11] It can be argued that calling this a prior is incorrect, since it actually depends on the observations $o_{1:t}$.

[12] For details, see the theory of conjugate priors.

Categorical model with latent variables

Exact posterior

Now we let the model be a joint distribution over the product space of observations x and latent states z, given by $\{x^{(1)}, \dots, x^{(n)}\} \times \{z^{(1)}, \dots, z^{(m)}\}$. The model is parametrized by $\theta = (\theta^D, \theta^A)$ as follows:

$$p(z^{(j)}|\theta^D) = \theta^D_j, \tag{32}$$

$$p(x^{(i)}|z^{(j)}, \theta^A) = \theta^A_{ij}. \tag{33}$$

The prior over the θ is defined as follows:

$$p(\theta|\alpha) = p(\theta^D|\alpha^D) \prod_j p(\theta^A_{\bullet j}|\alpha^A) \tag{34}$$

$$p(\theta^D|\alpha^D) \propto \prod_j \left(\theta^D_j\right)^{\alpha^D_j - 1}, \tag{35}$$

$$p(\theta^A_{\bullet j}|\alpha^A) \propto \prod_i \left(\theta^A_{ij}\right)^{\alpha^A_{ij} - 1}, \tag{36}$$

where $\theta_{\bullet j}$ denotes the vector $(\theta_{1j}, \dots, \theta_{nj})$. After observing $x^* = x^{(i^*)}$, the exact posterior is given by

$$\begin{aligned} p(\theta|x^*, \alpha) &\propto p(x^*|\theta)p(\theta|\alpha) \\ &= \sum_j p(x^*|z^{(j)}, \theta^A)p(z^{(j)}|\theta^D)p(\theta^A|\alpha^A)p(\theta^D|\alpha^D) \\ &\propto \sum_j \theta^A_{i^* j}\theta^D_j \left(\prod_{j'}\prod_{i'} \left(\theta^A_{i'j'}\right)^{\alpha^A_{i'j'}-1}\right)\left(\prod_{j''} \left(\theta^D_{j''}\right)^{\alpha^D_{j''}-1}\right) \\ &= \sum_j \left(\prod_{j'}\prod_{i'} \left(\theta^A_{i'j'}\right)^{\alpha^A_{i'j'}-1+\mathbb{1}_{i^*}(i')\mathbb{1}_j(j')}\right)\left(\prod_{j''} \left(\theta^D_{j''}\right)^{\alpha^D_{j''}-1+\mathbb{1}_j(j'')}\right). \end{aligned}$$

Note that this is no longer a Dirichlet distribution. Below we discuss how the Dirichlet distribution shows up in an algorithm for approximating the true posterior.

Mean-field approximation

We can also approximate the posterior over θ by minimizing the following variational free energy function:

$$F(\tilde{q}|x^*) = \mathbb{E}_{\tilde{q}(z,\theta)}\left[\ln \tilde{q}(z, \theta) - \ln p(x^*, z, \theta|\alpha)\right], \tag{37}$$

where we assume the approximate posterior over both θ and z factorizes as follows:

$$q(z, \theta) = q(z)q(\theta).$$

The coordinate ascent variational inference (CAVI) algorithm (Bishop & Nasrabadi, 2006; Blei, Kucukelbir, & McAuliffe, 2017) updates the distributions $q(z)$ and $q(\theta)$ iteratively, each time holding one distribution fixed while updating the other. More specifically, we can initialize $q(\theta)$ with our prior belief $p(\theta|\alpha)$ and minimize F w.r.t. $\tilde{q}(z)$. The variational free energy now becomes

$$F(\tilde{q}|x^*, q(\theta)) = \mathbb{E}_{\tilde{q}(z)}\left[\ln \tilde{q}(z) - \mathbb{E}_{q(\theta)}[\ln p(z|x^*, \theta)]\right] + C_{\setminus z}, \tag{38}$$

where $C_{\setminus z}$ is independent of $\tilde{q}(z)$. The minimizer $q(z)$ is proportional to

$$q(z) \propto \exp\left(\mathbb{E}_{q(\theta)}[\ln p(z|x^*, \theta)]\right). \tag{39}$$

(See Eq. (24).) We then fix this $q(z)$ and optimize F w.r.t. $\tilde{q}(\theta)$ and get

$$F(\tilde{q}|x^*, q(z)) = \mathbb{E}_{\tilde{q}(\theta)}\left[\ln \tilde{q}(\theta) - \mathbb{E}_{q(z)}[\ln p(x^*, z, \theta|\alpha)]\right] + C_{\setminus \theta}, \tag{40}$$

where $C_{\setminus \theta}$ is independent of $\tilde{q}(\theta)$. The minimizer $q(\theta)$ is proportional to

$$q(\theta) \propto \exp\left(\mathbb{E}_{q(z)}[\ln p(x^*, z, \theta|\alpha)]\right). \tag{41}$$

These two steps are performed iteratively until the beliefs about θ and z have converged.

Note that when $p(x, z, \theta|\alpha)$ is a categorical model with Dirichlet priors, as defined in (32)–(36), we can rewrite Eq. (41) as follows:

$$\begin{aligned} q(\theta) &\propto \exp\left(\mathbb{E}_{q(z)}[\ln p(x^*, z, \theta|\alpha)]\right) \\ &= \exp\left(\mathbb{E}_{q(z)}\left[\ln p(x^*|z, \theta^A) + \ln p(z|\theta^D) + \ln p(\theta|\alpha)\right]\right) \\ &= \exp\left(\sum_j q(z^{(j)})\left[\ln p(x^*|z^{(j)}, \theta^A) + \ln p(z^{(j)}|\theta^D)\right]\right) p(\theta|\alpha) \\ &= \prod_j \left(\theta^A_{i^* j}\right)^{q(z^{(j)})} \left(\theta^D_j\right)^{q(z^{(j)})} \left(\theta^D_j\right)^{\alpha^D_j - 1} \prod_i \left(\theta^A_{ij}\right)^{\alpha^A_{ij}-1} \\ &= \prod_j \left(\theta^D_j\right)^{\alpha^D_j + q(z^{(j)}) - 1} \prod_i \left(\theta^A_{ij}\right)^{\alpha^A_{ij} + \mathbb{1}_{i^*}(i)q(z^{(j)})-1}. \end{aligned}$$

Note that this is again a Dirichlet distribution with updated parameter $\alpha' = (\alpha^{D\prime}, \alpha^{A\prime})$ given by

$$\alpha^{D\prime}_j = \alpha^D_j + q(z^{(j)}), \tag{42}$$

$$\alpha^{A\prime}_{ij} = \alpha^A_{ij} + \mathbb{1}_{i^*}(i)q(z^{(j)}). \tag{43}$$

Update (42) makes latent states with high $q(z)$ more likely in the updated distribution. Update (43) makes sure that latent states are more likely to generate the observation $x^{(i^*)}$, especially those with high $q(z)$. Note the similarity with the update rule (31) for categorical models without latent variables.

POMDP model

Now we let p be the generative model from an active inference agent as described in (11)–(17). Similar to (37) the variational free energy now becomes

$$\begin{aligned} &F(\tilde{q}|o_{1:T}, a_{1:T-1}) \\ &\quad = \mathbb{E}_{\tilde{q}(s_{1:T},\theta)}\left[\ln \tilde{q}(s_{1:T}, \theta) - \ln p(o_{1:T}, s_{1:T}, \theta|a_{1:T-1}, \alpha)\right]. \end{aligned} \tag{44}$$

We use the mean-field approximation $q(s_{1:T}, \theta) = q(s_{1:T})q(\theta)$. Equivalent to (38)–(39) we can start by minimizing F w.r.t. $\tilde{q}(s_{1:T})$ and get a minimizer $q_T(s_{1:T})$ using the current belief about θ. Then, similar to (40)–(41) we can use this minimizer to write F as follows:

$$\begin{aligned} &F(\tilde{q}|o_{1:T}, a_{1:T-1}, q_T(s_{1:T})) \\ &\quad = \mathbb{E}_{\tilde{q}(\theta)}\left[\ln \tilde{q}(\theta) - \mathbb{E}_{q_T(s_{1:T})}[\ln p(o_{1:T}, s_{1:T}, \theta|a_{1:T-1}, \alpha)]\right] + C_{\setminus \theta}, \end{aligned} \tag{45}$$

and work out the minimizer. This gives

$$\begin{aligned} q(\theta) &\propto \exp\left(\mathbb{E}_{q_T(s_{1:T})}[\ln p(o_{1:T}, s_{1:T}, \theta|a_{1:T-1}, \alpha)]\right) \\ &= \exp\left(\mathbb{E}_{q_T(s_{1:T})}\left[\sum_{\tau=1}^T \ln p(o_\tau|s_\tau, \theta^A) + \ln p(s_1|\theta^D) \right.\right. \\ &\qquad \left.\left. + \sum_{\tau=2}^T \ln p(s_\tau|s_{\tau-1}, a_{\tau-1}, \theta^B)\right]\right) p(\theta|\alpha) \\ &= \exp\left(\sum_{s_{1:T}} q_T(s_{1:T})\left[\sum_{\tau=1}^T \ln p(o_\tau|s_\tau, \theta^A) + \ln p(s_1|\theta^D) \right.\right. \\ &\qquad \left.\left. + \sum_{\tau=2}^T \ln p(s_\tau|s_{\tau-1}, a_{\tau-1}, \theta^B)\right]\right) p(\theta|\alpha) \\ &= \prod_j \left(\theta^D_j\right)^{q_T(s_1^{(j)})} \prod_i \left(\theta^A_{ij}\right)^{\sum_{\tau=1}^T \mathbb{1}_{o^{(i)}}(o_\tau) q_T(s_\tau^{(j)})} \\ &\qquad \prod_k \left(\theta^B_{jkl}\right)^{\sum_{\tau=2}^T q_T(s_\tau^{(j)}) q_T(s_{\tau-1}^{(k)}) \mathbb{1}_{a^{(l)}}(a_{\tau-1})} p(\theta|\alpha) \\ &= \prod_j \left(\theta^D_j\right)^{\alpha^D_j + q_T(s_1^{(j)}) - 1} \prod_i \left(\theta^A_{ij}\right)^{\alpha^A_{ij} + \sum_{\tau=1}^T \mathbb{1}_{o^{(i)}}(o_\tau) q_T(s_\tau^{(j)}) - 1} \\ &\qquad \prod_k \left(\theta^B_{jkl}\right)^{\alpha^B_{jkl} + \sum_{\tau=2}^T q_T(s_\tau^{(j)}) q_T(s_{\tau-1}^{(k)}) \mathbb{1}_{a^{(l)}}(a_{\tau-1}) - 1}. \end{aligned}$$

This gives the update rules for α given in (18)–(20).

Remark 3. Note that these update rules are actually just the first iteration of the CAVI algorithm described above. It will therefore in general not minimize the variational free energy in Eq. (44). Instead it minimizes the quantity given in Eqs. (28) and (45) where $q_T(s_{1:T})$ is fixed. Note however that if one would assume $q_T(s_{1:T})$ to be given, the following variational free energy would be the natural choice to minimize:

$$\mathbb{E}_{\tilde{q}(\theta)}\left[\ln \tilde{q}(\theta) - \ln\left(\mathbb{E}_{q_T(s_{1:T})}[p(o_{1:T}, \theta | s_{1:T}, a_{1:T-1}, \alpha)]\right)\right],$$

since this has as minimizer the exact posterior distribution.

Appendix C. Further details

Equivalent formulations of expected free energy

Recall that the expected free energy in Eq. (2) is given by

$$\begin{aligned} G(\pi_t | o_{1:t}, a_{1:t-1}) \\ = -\Big(\mathbb{E}_{q_t(o_{t+1:T}|\pi_t)}\left[D_{\mathrm{KL}}\left(q_t(s_{t+1:T}|o_{t+1:T}, \pi_t) \parallel q_t(s_{t+1:T}|\pi_t)\right)\right] \\ + \mathbb{E}_{q_t(o_{t+1:T}|\pi_t)}\left[\ln p_C(o_{t+1:T})\right]\Big). \end{aligned}$$

We can derive the equivalent formulation from Eq. (3) as follows. We first expand the KL divergence term to get

$$\begin{aligned} G(\pi_t | o_{1:t}, a_{1:t-1}) = \mathbb{E}_{q_t(s_{t+1:T}, o_{t+1:T}|\pi_t)}\Big[\ln q_t(s_{t+1:T}|\pi_t) - \ln q_t(s_{t+1:T}|o_{t+1:T}, \pi_t) \\ - \ln p_C(o_{t+1:T})\Big]. \end{aligned} \tag{46}$$

Using Bayes' rule we can rewrite

$$\begin{aligned} -\ln q_t(s_{t+1:T}|o_{t+1:T}, \pi_t) = -\ln q_t(o_{t+1:T}|s_{t+1:T}, \pi_t) - \ln q_t(s_{t+1:T}|\pi_t) \\ + \ln q_t(o_{t+1:T}|\pi_t). \end{aligned}$$

Plugging this into (46) and using that $q_t(o_{t+1:T}|s_{t+1:T}, \pi_t) = p(o_{t+1:T}|s_{t+1:T})$ we get

$$\begin{aligned} G(\pi_t | o_{1:t}, a_{1:t-1}) &= \mathbb{E}_{q_t(s_{t+1:T}, o_{t+1:T}|\pi_t)}\Big[\ln p(o_{t+1:T}|s_{t+1:T}) + \ln q_t(o_{t+1:T}|\pi_t) \\ &\qquad - \ln p_C(o_{t+1:T})\Big] \\ &= \mathbb{E}_{q_t(s_{t+1:T}|\pi_t)}\Big[\mathrm{H}\big[p(o_{t+1:T}|s_{t+1:T})\big]\Big] \\ &\qquad + D_{\mathrm{KL}}\Big(q_t(o_{t+1:T}|\pi_t) \parallel p_C(o_{t+1:T})\Big), \end{aligned}$$

where the last line is equal to Eq. (3).

Independence between state factors and observation modalities

In order to make the computation of state inference more efficient, the agent can use independencies between different state factors and observation modalities (different dimensions of state and observation space). More specifically, we assume that given a state, the different observation modalities are independent, which translates to:

$$p(o_\tau | s_\tau) = \prod_m p(o_\tau^m | s_\tau).$$

We use superscript m and f to denote a specific observation modalities and state factors respectively. Furthermore, we assume a certain state factor to be independent of all other state factors in the same and previous time step, given the same state factor in the previous time step and the last action, i.e.:

$$p(s_\tau | s_{\tau-1}, a_{\tau-1}) = \prod_f p(s_\tau^f | s_{\tau-1}^f, a_{\tau-1}).$$

For a fixed state factor f Eqs. (6) and (7) now become

$$\begin{aligned} q_t(s_t^f) &\propto p(o_t | s_t^f) \sum_{s_{t-1}^f} p(s_t^f | s_{t-1}^f, a_{t-1}) q_{t-1}(s_{t-1}^f) \\ q_t(s_\tau^f | a_{1:\tau-1}) &= \sum_{s_{\tau-1}^f} p(s_\tau^f | s_{\tau-1}^f, a_{\tau-1}) q_t(s_{\tau-1}^f | a_{1:\tau-2}), \end{aligned} \tag{47}$$

and Eq. (10) becomes

$$q_t(s_{1:T} | \pi_t) \propto \prod_{\tau=1}^{t} \prod_m p(o_\tau^m | s_\tau)\ p(s_1) \prod_{\tau=2}^{T} \prod_f p(s_\tau^f | s_{\tau-1}^f, a_{\tau-1}).$$

Fixed point iteration

Eq. (47) involves the distribution $p(o_t | s_t^f)$. We can however not access this directly. To find an approximate solution, we can use fixed point iteration as follows:

$$q_t^{(i+1)}(s_t^f) \propto \sum_{s_t^{\setminus f}} q_t^{(i)}(s_t^{\setminus f}) p(o_t | s_t) \sum_{s_{t-1}^f} p(s_t^f | s_{t-1}^f, a_{t-1}) q_{t-1}(s_{t-1}^f),$$

where $\setminus f$ denotes the set of all state factors apart from f.

Softmax function

Definition 1. Let $S = \{x^{(1)}, \ldots, x^{(n)}\}$ be a finite set and $\mu : S \to \mathbb{R}$ a function. The *softmax function* σ is given by

$$\sigma(\mu(x^{(i)})) = \frac{e^{\mu(x^{(i)})}}{\sum_j e^{\mu(x^{(j)})}}. \tag{48}$$

Note that $\sigma(\mu(x^{(i)})) > 1$ and $\sum_S \sigma(\mu(x^{(i)})) = 1$. Therefore the softmax function can be used to turn μ into a probability distribution.

References

Bishop, Christopher M., & Nasrabadi, Nasser M. (2006). *Pattern recognition and machine learning, Vol. 4.* (4), Springer.

Blei, David M., Kucukelbir, Alp, & McAuliffe, Jon D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association, 112*(518), 859–877.

Da Costa, Lancelot, Parr, Thomas, Sajid, Noor, Veselic, Sebastijan, Neacsu, Victorita, & Friston, Karl (2020). Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology, 99*, Article 102447. http://dx.doi.org/10.1016/j.jmp.2020.102447.

Da Costa, Lancelot, Tenka, Samuel, Zhao, Dominic, & Sajid, Noor (2024). Active inference as a model of agency. arXiv preprint arXiv:2401.12917.

Heins, Conor (2021). Active inference demo: T-maze environment. https://pymdp-rtd.readthedocs.io/en/latest/notebooks/tmaze_demo.html.

Heins, Conor, Millidge, Beren, Demekas, Daphne, Klein, Brennan, Friston, Karl, Couzin, Iain, et al. (2022). pymdp: A Python library for active inference in discrete state spaces. *The Journal of Open Source Software, 7*(73), 4098.

Millidge, Beren, Tschantz, Alexander, & Buckley, Christopher L. (2021). Whence the expected free energy? *Neural Computation, 33*(2), 447–482.

Parr, Thomas, Pezzulo, Giovanni, & Friston, Karl J. (2022). *Active inference: the free energy principle in mind, brain, and behavior.* MIT Press.

Smith, Ryan, Friston, Karl J., & Whyte, Christopher J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology, 107*, Article 102632.

Wei, Ran (2024). Value of information and reward specification in active inference and POMDPs. arXiv preprint arXiv:2408.06542.