Supplementary Materials

Explicit parameterization of the finite horizon POMDP

Active inference rests on the tuple $(O, U, S, T, \Pi, R, P, Q)$:

- A finite set of outcomes, $O$
- A finite set of control states or actions, $U$
- A finite set of hidden states, $S$
- $T = 0, ..., T$ a finite set which determines the temporal horizon
- A finite set of time-sensitive policies, $\Pi$
- A generative process $R(\tilde{o}, \tilde{s}, \tilde{u})$ that generates a probabilistic outcome $o \in O$ from (hidden) state $s \in S$ and action $u \in U$
- A generative model $P(\tilde{o}, \tilde{s}, \pi, z)$ with outcome $o \in O$, (hidden) state $s \in S$, policies $\pi \in \Pi$ and model parameters $z$.
- An approximate posterior $Q(\tilde{s}, \pi, z) = Q(s_o|\pi)Q(s_{\tau}|\pi)Q(\pi)Q(z)$ over states $s \in S$, policies $\pi \in \Pi$ and model parameters.

The generative process describes transitions between hidden (unobserved) states in the world that generate (observed) outcomes. Their transitions depend on action, which depends on posterior beliefs about the next state. Subsequently, these beliefs are formed using a generative model of how outcomes are generated. The generative model (based on partially observable MDP) describes what the agent believes about the world, where beliefs about hidden states and policies are encoded by expectations. Here actions are part of the generative process in the world and policies are part of the generative model of the agent.
Pseudo-code for active inference: belief updating and action selection

Initialize the following:
Probability of seeing outcomes, given states, likelihood: $A$
Probability of transitioning between states, given an action: $B$
Log probability of agent’s preferences about outcomes: $C$
Probability of state the agent believes it is at the beginning of each trial: $D$

for $\tau = 1 : T$ do

Sample state, $s$ based on generative process

Sample outcome $o$ based on likelihood matrix $A$

Variational updates of expected states, $s$ under sequential policies
(gradient descent on $F$)

Evaluate expected free energy $G$ of policies $\pi$

Bayesian model averaging of expected states $s$ over policies $\pi$

Select action with the lowest expected free energy

end

Accumulation of (concentration) parameters for learning update based on learning rate
Pseudo-code for Q-Learning

Initialize the following:
Q-value function; $Q(s,a)$
Initialize parameter for exploration; $\epsilon$
Specify learning rate, $\alpha$ and discount factor, $\gamma$

for $\tau = 1 : T$ do

    Sample exploration rate threshold from a random uniform distribution, $U(0,1)$

    Choose action based on $\max_a(Q(s,:))$ if exploration rate threshold is greater than $\epsilon$, else choose random action

    Execute $a^*$ and receive $r, s'$

    Update $Q(s,a)$: using $(1 - \alpha) * Q(s,a) + \alpha * (r + \gamma * \max_a(Q(s',:)))$ 

    $s = s'$

    Update exploration parameter $\epsilon$: $\epsilon$-- decay rate

end
Pseudo-code for Bayesian Model-Based Reinforcement Learning using Thompson Sampling

Initialize the following:
\( \Theta_t, \Theta_r \) as uniform
Probability of transitioning between states, given an action, transition model; \( \Theta_t \)
Probability of receiving reward, given a state, reward function; \( \Theta_r \)

Repeat:

Sample \( \Theta_{t,1}, \ldots, \Theta_{t,k} \sim Pr(\Theta_t) \forall a \)
Sample \( \Theta_{r,1}, \ldots, \Theta_{r,k} \sim Pr(\Theta_r) \forall a \)

\( Q^*_\hat{\theta}_{t,i,\theta_{r,i}} \leftarrow \text{solve } MDP_{\theta_{t,i},\theta_{r,i}} \)

\( \hat{Q}(s, a) \leftarrow \frac{1}{k} \sum_{i=1}^{k} Q^*_\theta_{t,i,\theta_{r,i}}(s, a) \)

\( a^b \leftarrow \max_a \hat{Q}(s, a) \)

Execute \( a^* \) and receive \( r, s' \)

\( b(\Theta_t) \leftarrow b(\Theta_t)Pr(s'|s, a, \Theta_t) \)
\( b(\Theta_r) \leftarrow b(\Theta_r)Pr(r|s, a, s', \Theta_r) \)

\( s \leftarrow s' \)

end