Details of habit model inspired from active inference (Bogacz, 2020)

An agent determines action a based on reward R. R is sum of the reward currently presented and expected reward in the near future.

$$R = r + v \quad \cdots (1)$$

The current reward r is determined physically (presented or not). The v is defined by valuation system, which is described later. The a is assumed as action intensity (e.g., the response rates to a lever).

Bogacz (2020) presumes that the action is emitted in cooperation of goal-directed and habit systems. The goal directed system is defined as expectation of R under certain action a and the current state s, $P(R|a,s)$. The habit system is defined as the history of actions performed under state s, $P(a|s)$. Thus, unlike previous habit models, Bogacz (2020)'s model postulate goal-directed and habit systems encode probabilities of different quantities.

If we look habit system as a prior and goal-directed system as a likelihood distribution. We can define a posterior $P(a|R,s)$ as follows, using Bayes rule;

$$P(a|R,s) = \frac{p(R|a,s)p(a|s)}{p(R|s)} \quad \cdots (2)$$

The agents' task is to choose the action a, which maximizes the posterior $P(a,|R,s)$.

One need to select a specific type of distribution, of which shape is characterized by sufficient statistics. Bogacz (2020) uses Gaussian distribution for both goal-directed $P(R|a,s)$ and habit $P(a|s)$. The Gaussian distribution is expressed as,

$$P(x) = f(x;\mu,\Sigma) = \frac{1}{\sqrt{2\pi\Sigma}} \exp\left(-\frac{(x-\mu)^2}{2\Sigma} \right) \quad \cdots (3)$$

where μ and Σ are mean and variance parameters.

Bogacz models goal-directed and habitual behaviors using free parameters q and h:

$$P(R|a,s) = f(R;aqs,\Sigma_g) \quad \cdots (4)$$

$$P(a|s) = f(a;hs,\Sigma_h) \quad \cdots (5)$$

The state s encodes stimulus intensity, thus continuous values in this context. The mean of goal-directed and habit system are aqs and hs, respectively. The Σ_g and Σ_h are variance parameters of each system. The task of agent is choose optimal a in ongoing trial and learn parameters q, h, Σ_g, and Σ_h once it obtains a reward.

Intuitively, the goal-directed system estimate reward R in proportion to action intensity a, representing R-O association. The habit system only relies on state s, but is independent from R, representing S-R association.

Because both likelihood and prior are Gaussian distribution, the posterior has analytic solution. However, Bogacz (2020) assume that the actual organism is unlikely to directly evaluate (2). Thus, approximate inference is utilized.

The free energy is defined as follows;

$$F = \ln[P(R \mid s, a)P(a \mid s)] \cdots (6)$$

This is a logarithm of numerator in (2). This is a special case of standard definition of variational free energy found in other literatures (e.g., Bogacz, 2017, Tschantz et al., 2020). The relationship is described in the Appendix.

Bogacz (2020) assumes whole behavior can be decomposed into “planning” and “learning”. During planning, the reward is not presented yet, and thus $r = 0$ resulting in $R = v$ according to (1). The action is performed in a way to maximize the derivative of F with respect to a.

$$\dot{a} = \frac{\partial F}{\partial a} \cdots (7)$$

Let us calculate (7), by assuming (4) and (5).

$$\begin{aligned} F &= \ln[P(R \mid a, s)P(a \mid s)] \\ &= \ln P(R \mid a, s) + \ln P(a \mid s) \end{aligned}$$

The logarithm of Gaussian distribution can be computed following,

$$\ln f(x; \mu, \Sigma) = \frac{1}{2}\left(- \ln \Sigma + -\frac{(x - \mu)^2}{\Sigma} \right) - \frac{1}{2} \ln 2\pi$$

We used $\ln \exp(a) = a$, and $\ln \frac{1}{\sqrt{\Sigma}} = \ln \Sigma^{-1/2} = -\frac{1}{2}\Sigma$, and $\ln ab = \ln a + \ln b$ in this transformation. Thus, free energy here can be expressed:

$$F = \frac{1}{2}\left(-\ln \Sigma_g + -\frac{(R - aqs)^2}{\Sigma_g} - \ln \Sigma_h + -\frac{(a - hs)^2}{\Sigma_h} \right) + C \cdot \cdot \cdot (8)$$

where C is a constant term.

We would like to choose action, which maximize $\frac{\partial F}{\partial a}$.

$$\dot{a} = \frac{\partial F}{\partial a} = \frac{(R - aqs)qs}{\Sigma_g} - \frac{hs - a}{\Sigma_h} \cdot \cdot \cdot (9)$$

For notational simplicity, we introduce the prediction error terms

$$\delta_g = \frac{R - aqs}{\Sigma_g} \cdot \cdot \cdot (10)$$

$$\delta_h = a - hs \cdot \cdot \cdot (11)^1$$

The equation (10) can be interpreted as reward prediction error, whereas (11) represents different prediction error that quantify what extent the current action was differed from previous.

$$\frac{\partial F}{\partial a} = \delta_g qs + \frac{hs - a}{\Sigma_h} \cdot \cdot \cdot (12)$$

Bogacz (2020) used Euler method to update action intensity in his simulation.

Once reward is obtained ($R = r$), the agent involves “learning”, at which it updates parameters. Specifically, the parameter updates follows the derivative of F with respect to each parameter.

$$\Delta q \sim \frac{\partial F}{\partial q} = \delta_g as \cdot \cdot \cdot (13)$$

---

[1] This is not typo. Bogacz (2020) defined (11) without scaling by Σ_h.

$$\Delta h \sim \frac{\partial F}{\partial h} = \delta_h s \cdots (14)$$

$$\Delta \Sigma_g \sim \frac{\partial F}{\partial \Sigma_g} \sim (\delta_g \Sigma_g)^2 - \Sigma_g \cdots (15)$$

$$\Delta \Sigma_h \sim \frac{\partial F}{\partial \Sigma_h} \sim \delta_h^2 - \Sigma_h \cdots (16)$$

The equations (13) and (14) can be derived
Note that the derivatives of F with respect to variances are approximation, but not exact solutions. The exact derivative $\frac{\partial F}{\partial \Sigma_h}$, for example, is $\frac{\partial F}{\partial \Sigma_h} = \frac{(a-h)^2 - \Sigma_h}{2\Sigma_h^2}$. Thus, equation (16) ignores scaling $2\Sigma_h^2$. This approximation is useful when Σ is much lower than 1. It is intuitive to depict how updates differ between exact and approximated derivatives.

Actual update equations are defined using learning parameters are expressed using learning parameters;

$$\dot{q} = \alpha_g \delta_g a s \cdots (17)$$

$$\dot{h} = \alpha_h \delta_h s \cdots (18)$$

$$\dot{\Sigma}_g = \alpha_{\Sigma_g} \left((\delta_g \Sigma_g)^2 - \Sigma_g \right) \cdots (19)$$

$$\dot{\Sigma}_h = \alpha_{\Sigma_h} (\delta_h^2 - \Sigma_h) \cdots (20)$$

where α_g, α_h, α_{Σ_g}, and α_{Σ_h} are learning parameters (fixed values in the simulation).
Moreover, the dynamics of prediction errors can be defined:

$$\tau_\delta \dot{\delta}_g = r + v - a q s - \Sigma_g \delta_g \cdots (21)$$

$$\tau_\delta \dot{\delta}_h = a - h s - \delta_h \cdots (22)$$

One can easily show these dynamics result in prediction error terms defined at (10) and (11) when setting equal to 0.

$$\tau_\delta \dot{\delta}_g = 0$$

$$r + v - a q s - \Sigma_g \delta_g = 0$$

$$\Sigma_g \delta_g = r + v - a q s$$

$$\delta_g = \frac{r + v - a q s}{\Sigma_g}$$

The right side of final equality is the definition of prediction error (10).

Up to now, we used $R = r + v$, but have not defined v yet. Bogacz (2020) assumes that v is computed by 'valuation' system. He utilized temporal difference learning framework in reinforcement learning, which mimics the activities of dopaminergic neurons (Montague et al., 1996).

Time is assumed to be divided into interval I, and the state of the environment is represented by a column vector $\bar{s}_v$.

$\tau \dot{v} = \bar{w}\bar{s}_v - v$ ・・・(23)

$\tau_\delta \dot{\delta}_v = r + v - v_{t-I} - \delta_v$ ・・・(24)

$\dot{\bar{w}} = \alpha_v \delta_v \bar{e}$ ・・・(25)

$\tau \dot{\bar{e}} = \lambda \bar{e}_{t-I-3\tau} + \bar{s}_v^T - \bar{e}$ ・・・(26)

$\tau \dot{r} = r_0 - r$ ・・・(27)

where $\bar{w}$ is a vector of parameters describing how much reward can be expected after stimulus appearing in a particular interval, and τs are time constants. The $\bar{e}$ is a parameter called 'eligibility trace' associated with weights $\bar{w}$.

The equation (23) determines the changes of values v converges to $v = \bar{w}\bar{s}$.

The equation (24) represents 'temporal difference prediction error', which utilize not only $R = r + v$, but also previous valuation v_{t-I}. For example, if we set $I = 0.2$ seconds, the equation (24) means valuation system refers 200ms past values to compute prediction error. The prediction errors converges $R - v_{t-I}$ in accordance with the standard temporal difference learning (Sutton and Barto, 1998)
When $t \leq I$, v_{t-I} is set to 0.

The equation (25) describes how the weights are updated according to the prediction errors. The weights are updated based on not only prediction error, but also eligibility trace $\bar{e}$, which is defined in equation (26). Intuitively, eligibility trace determines when the weights are updated. The parameter λ is fixed to 0.9 and τ is fixed 0.02 across simulation.

The equation (27) is a bit conceptually tricky. The r is reward signal, and r_0 is the actual value to the reward. The reward signal r is determined based on the pay-off of action intensity:

$r = 5\tanh(3a/5) - a$ ・・・(28)

The function appears inverse U-like shape. However, during extinction trials, the reward signal is always 0, and the agent must pay only cost. Bogacz (2020) simply assumes the cost as negative action intensity:

$r = -a$ ・・・(29)

The r_0 is either 0 or 1 according to the actual reward presentation.

In simulations involving selection of action intensity, the time represented by the valuation system was divided into intervals of $I = 0.2$. The stimulus was presented at time $t = 1$, while the reward was given at time $t = 2$, thus the valuation system represented the value of 5 time intervals (i.e. vectors $\bar{w}$, $\bar{s}_v$ and $\bar{e}$ had 5 elements each).

For instead, in the actual simulation, if a reward is presented $t = 2$, and $I = 0.2$, then total trial length is $T = 2.2$ seconds. Assuming $\Delta t = 0.001$, the number of states is $T/\Delta t = 2200$. The number of 'micro state' is defined $(reward\ time - CS\ time)/I = (2-1)/0.2 = 5$. Thus, the number of elements $\bar{w}$ and $\bar{s}_v$ are five.

Supplement (1) : in the simulation, state values s_v are encoded as matrix 5×2200, and in each time step t, $\bar{w}'\bar{s}_v$ is calculated using $\bar{s}_v = s_v[:, t]$. Then, v, which appears in (1), is updated as a scalar value, using (23).

Supplement (2) : A index $t - I - 3\tau$ in equation (26) refers the previous e. Since $I = 0.2$ and $\tau = 0.02$, resulting in $-(I + 3\tau) = -0.26$, the trace is assumed as 260ms.

Appendix : Bogacz (2020)'s free energy is negative free energy of others

According to definition (6), $F = \ln[P(R|s,a)P(a|s)]$ in Bogacz (2020)'s paper. We can re-write $P(R|s,a)P(a|s) = P(R,a|s)$, which corresponds to generative model in others' formulation.

A conventional variational free energy in the FEP literatures is define as

$$F_v = \int Q(x) \ln \frac{Q(x)}{P(x,o)} dx$$

where x represents hidden variables that the agent are evaluating, o observation (e.g., sensory inputs) and $Q(x)$ is approximated distribution to infer the posterior $P(x \mid o)$, which is encoded by the agent so-called 'recognition density'. The $P(x, o)$ is called generative model, which the agents have. Given that $P(x, o) = P(x \mid o)P(o)$, the variational free energy F_v is re-written:

$$
\begin{aligned}
F_v &= \int Q(x) \ln \frac{Q(x)}{P(x \mid o)P(o)} dx \\
&= \int Q(x) \left(\ln \frac{Q(x)}{P(x \mid o)} - \ln P(o) \right) dx \\
&= \int Q(x) \ln \frac{Q(x)}{P(x \mid o)} dx - \int Q(x) \ln P(o) dx \\
&= D_{KL}[Q(x) \| p(x \mid o)] - \ln P(o)
\end{aligned}
$$

where D_{KL} denotes Kullback-Leibler divergence. Since Kullback-Leibler divergence is non-negative, one can show variational free energy F_v is upper bound of surprise. Therefore, agents must recognize, learn, and perform action in order to reduce F_v in ordinary FEP.

$$
\begin{aligned}
&F_v = D_{KL}[Q(x) \| p(x \mid o)] - \ln P(o) \\
&D_{KL}[Q(x) \| p(x \mid o)] = F_v + \ln P(o) \geq 0 \\
&F_v \geq - \ln P(o)
\end{aligned}
$$

These brief formulation of F_v appears in standard FEP literatures, and basic presumption of these is that the agent tries to approximate the posterior belief $P(x \mid o)$ using recognition density $Q(x)$.

If one assumes $Q(x)$ as delta distribution, we can show that F defined in (6) corresponds to negative variational free energy $-F_v$.

$$
\delta(x) = \begin{cases} \infty \ (x = 0) \\ 0 \ (x \neq 0) \end{cases},
$$

and practically, one can define recognition density,

$$
Q(x) = \delta(x - \phi)
$$

where ϕ is the current inference of agent to x (for example, “the current temperature x is ϕ degrees”). This assumption of delta distribution as recognition density means that the agent guesses the single scalar value of x, but not its uncertainty. The delta distribution has a property:

$$\int \delta(x) dx = 1$$

$$\int \delta(x) f(x) dx = f(0)$$

These are definite integrations but the intervals were omitted. Then, substituting recognition density by delta distribution,

$$\begin{aligned} F_v &= \int Q(x) \ln \frac{Q(x)}{P(x,o)} dx \\ &= \int \delta(x-\phi) \ln \frac{\delta(x-\phi)}{P(x,o)} dx \\ &= \int \delta(x-\phi) \Big(\ln \delta(x-\phi) - \ln P(x,o) \Big) dx \\ &= -\ln P(x,o) \\ &= -\ln P(o \mid x) P(x) \end{aligned}$$

Since $\ln 1 = 0$, integral of $\ln \delta(x)$ is result in 0. Considering $o = R$ (what the agent observes), and $x = a$ (what the agents must infer), we obtain

$$F_v = -\ln[P(R \mid a) P(a)]$$

Finally, one can condition state s above equation, and notice the relationship F_v and Bogacz (2020)’s F.

$$F = -F_v$$

Because the F here is the negative variational free energy, the agents must perform action a in a way to maximize F.