Nested Free-Energy Loops with an Epistemic Bonus

Philosopher M. Sunny and OpenAI `o3`

April 2025

1 Preliminaries

The agent interacts with hidden states $\{s_t, o_t, a_t\}_{t=1}^{\infty}$. Its *generative model* p_θ contains both dynamics $p_\theta(s_{t+1} \mid s_t, a_t)$ and emissions $p_\theta(o_t \mid s_t)$. A variational distribution q_φ approximates the intractable true posterior.

Symbol	Meaning
s_t	latent world state at step t
o_t	sensory input at t
a_t	action emitted at t
p_θ	dynamics + emissions ("world model")
q_φ	agent's belief over states
$\mathcal{H}[p]$	Shannon entropy $-\sum_x p(x) \ln p(x)$

Table 1: Notation

Interpretation of $\mathrm{D}_{\mathrm{KL}}\big(q_\varphi(s_t) \,\|\, p_\theta(s_t \mid o_t)\big)$. It is the *Kullback–Leibler divergence* $\mathrm{D}_{\mathrm{KL}}\big(q_\varphi(s_t) \,\|\, p_\theta(s_t \mid o_t)\big)$ between

1. the agent's current belief $q_\varphi(s_t)$, and
2. the Bayesian ideal $p_\theta(s_t \mid o_t)$ after a perfect update.

Hence it measures residual inference error (a.k.a. surprise).

Inner free energy

$$F_t(\varphi, \theta) = \mathrm{D}_{\mathrm{KL}}\big(q_\varphi(s_t) \,\|\, p_\theta(s_t \mid o_t)\big), \tag{1}$$

with updates

$$\varphi \leftarrow \varphi - \eta_\varphi \nabla_\varphi F_t, \qquad \theta \leftarrow \theta - \eta_\theta \nabla_\theta F_t. \tag{2}$$

2 Outer (planning) loop

2.1 Counterfactual roll-outs

Before acting the agent *imagines* a sequence $\mathbf{a} = (a_t, \dots, a_{t+H-1})$ and rolls it out inside p_θ. These **counterfactual simulations** probe consequences without incurring physical cost or risk.

2.2 Expected free energy with epistemic bonus

$$
\begin{aligned}
G_\beta(\mathbf{a}) &= \mathbb{E}\big[\mathrm{D}_{\mathrm{KL}}\big(q_\varphi(s_{t+H}) \,\|\, p_{\mathrm{goal}}\big)\big] && \text{(risk / goal mismatch)} \\
&\quad + \mathbb{E}[\mathcal{H}[p_\theta(o_{t+1:t+H} \mid s_{t:t+H})]] && \text{(ambiguity)} \\
&\quad - \beta\,\mathbb{E}\big[\mathrm{D}_{\mathrm{KL}}\big(q_\varphi(s_{t+H}) \,\|\, q_\varphi^{\mathrm{prior}}\big)\big] && \text{(information gain).} \qquad (3)
\end{aligned}
$$

- $\beta = 0$ – pure exploitation (no curiosity term).
- $\beta > 1$ – strong drive for information gain.

2.3 Action selection

$$a_t^\star \;=\; \arg\min_{\mathbf{a}} \, G_\beta(\mathbf{a}). \tag{4}$$

3 Energetic rationale

Planning is worthwhile only if computation plus execution costs less energy than repeated blind trials:

$$E_{\mathrm{sim}} + E_{\mathrm{execution}} \;<\; E_{\text{blind-trial}}.$$

Equivalently,

$$\mathbb{E}[F_{t+1:t+H} \mid a_t^\star] \;\le\; \mathbb{E}[F_{t+1:t+H} \mid \text{uninformed actions}]\,.$$

4 Toy 1-D world (7 cells)

Cells $0 \dots 6$ with cell 0 the abyss, 6 the goal, and 3 a potential hazard.

Hazard prior	β	Qualitative policy	Typical path
0.45	0	exploit	$1 \to 6$
	> 0	probe then exploit	$1 \to 2 \to 3 \to 6$
0.90	large	avoid	$1 \to 2 \to 3$ (halt)

Table 2: Regimes in the toy world.

Script `toy_fep.py` reproduces these behaviours.

How is D_{KL} evaluated?

For the *discrete* state space of our toy world ($s \in \{0, \dots, 6\} \times \{\text{hazard\_false}, \text{hazard\_true}\}$) the divergence reduces to a finite sum

$$\mathrm{D}_{\mathrm{KL}}\big(q(s) \,\|\, p(s)\big) \;=\; \sum_s q(s)\, \ln \frac{q(s)}{p(s)} \quad \text{(nats).} \tag{5}$$

Numerical details.

- Entries where $q(s) = 0$ contribute 0 (since $\lim_{x \to 0^+} x \ln x = 0$).
- If $p(s) = 0$ while $q(s) > 0$ the divergence is $+\infty$, signalling an impossible event under the model. In code we *clip* $p(s)$ to a tiny floor (10^{-12} in `toy_fep.py`) to keep the result finite and differentiable.
- The continuous analogue switches the sum for an integral $\int q(x) \ln \frac{q(x)}{p(x)} \, dx$, but all conceptual statements in the note carry over unchanged.

Equation (5) is what the `entropy(...)` helper in the Python script computes.

5 Broader links

- Eq. (1) is the inner (Friston) free energy.
- Eq. (3) shows the outer epistemic loop.
- In language models, sampling width/depth acts like tuning β.