Active Inference in continuous time notes

Federico Maggiore

September 29, 2020

Summary notation

- $\mathbf{x} = \{x_i\}_{i=1}^{D}$ environmental variables of the D-dimensional space constituting latent or hidden states;
- $\mathbf{s} = \{s_i\}_{i=1}^{S}$ body sensors input;
- $P(\mathbf{x}, \mathbf{s})$ *G-density*;
- $P(\mathbf{x}|\mathbf{s})$ *Posterior*;
- $P(\mathbf{s}|\mathbf{x})$ *Likelihood*;
- $P(\mathbf{x})$ *Prior*;
- $P(\mathbf{s}) = \int P(\mathbf{s}|\mathbf{x})P(\mathbf{x})d\mathbf{x}$ *marginal likelihood*
- $Q(\mathbf{x})$ *R-density*
- $F \equiv \int Q(\mathbf{x}) \ln \frac{Q(\mathbf{x})}{P(\mathbf{x},\mathbf{s})} d\mathbf{x}$ *Variational Free Energy*
- $L(\mu, s) \equiv -\ln P(\mu, s)$ *Laplace-encoded energy*

References: Manuel and Christopher 2019 Christopher L. et al. 2017

Free Energy Principle (FEP)

The goal of an agent is to determine the probability of the hidden states given some sensory inputs:

$$P(\mathbf{x}|\mathbf{s}) = \frac{P(\mathbf{x}, \mathbf{s})}{P(\mathbf{s})} = \frac{P(\mathbf{s}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{s})} \tag{1}$$

with

- $P(\mathbf{x}, \mathbf{s})$ *G-density*, beliefs about the states assumed to be encoded by the agent;
- $P(\mathbf{x}|\mathbf{s})$ *Posterior*, i.e. probability of hidden causes x given observed sensory data;

- $P(\mathbf{s}|\mathbf{x})$ *Likelihood*, i.e. organism's assumptions about sensory input $\mathbf{s}$ given the hidden causes $\mathbf{x}$;
- $P(\mathbf{x})$ *Prior*, i.e. agent's beliefs about hidden causes **before** that $\mathbf{s}$ are received;
- $P(\mathbf{s}) = \int P(\mathbf{s}|\mathbf{x})P(\mathbf{x})d\mathbf{x}$ *marginal likelihood*, i.e. normalization factor.

For the agent it's not necessary to compute the complete posterior distribution, it has only to find the hidden state -or at least a good approximation- that maximize the posterior, i.e. $\arg\max_{\mathbf{x}} P(\mathbf{x}|\mathbf{s})$. The problem with the exact Bayesian scheme, is that $P(\mathbf{s})$ is often impossible to calculate, and moreover $P(\mathbf{x}|\mathbf{s})$ may not take a standard shape and could not have a summary statistics.

A biologically plausible technique consist in using an auxiliary distribution $Q(\mathbf{x})$ called *recognition density* (*R-density*) that has to be optimized to became a good approximation of the posterior.

In order to do this the Kullback-Leibler divergence is minimized:

$$\begin{aligned} D_{KL}(\, Q(\mathbf{x}) \,||\, P(\mathbf{x}|\mathbf{s}) \,) &= \int Q(\mathbf{x}) \ln \frac{Q(\mathbf{x})}{P(\mathbf{x}|\mathbf{s})} d\mathbf{x} \\ &= \int Q(\mathbf{x}) \ln \frac{Q(\mathbf{x})P(\mathbf{s})}{P(\mathbf{x},\mathbf{s})} d\mathbf{x} \\ &= \int Q(\mathbf{x}) \ln \frac{Q(\mathbf{x})}{P(\mathbf{x},\mathbf{s})} d\mathbf{x} + \ln P(\mathbf{s}) \int Q(\mathbf{x}) d\mathbf{x} \\ &= F + \ln P(\mathbf{s}) \end{aligned} \tag{2}$$

where

- $F \equiv \int Q(\mathbf{x}) \ln \frac{Q(\mathbf{x})}{P(\mathbf{x},\mathbf{s})} d\mathbf{x} = -\left\langle \ln P(\mathbf{x},\mathbf{s}) \right\rangle_Q - \left\langle \ln Q(\mathbf{x}) \right\rangle_Q$ is the *Variational Free Energy* (VFE), a quantity that depends on the R-density and the knowledge about the environment i.e. the G-density $P(\mathbf{s},\mathbf{x}) = P(\mathbf{s}|\mathbf{x})P(\mathbf{x})$ that we are assuming the agent has.
- $\ln P(\mathbf{s})$ is a term independent of the recognition density $Q(\mathbf{x})$ ($\Rightarrow$ minimizing F with respect to $Q(\mathbf{x})$ will minimize the D_{KL})

Laplace approximation

Often optimizing F for arbitrary $Q(\mathbf{x})$ is particularly complex. Moreover, it is assumed that neural activity parametrise sufficient statistic. For these reasons, a common approximation is to assume that the R-density take a Gaussian form.

One dimensional case

Let's assume that the R-density $Q(x)$ has a peak at point μ. The Taylor-expansion of the logarithm around this peak is

$$\ln Q(x) \simeq \ln Q(\mu) - \frac{1}{2} \frac{(x-\mu)^2}{\Sigma} \tag{3}$$

with

$$\frac{1}{\Sigma} = -\frac{\partial^2}{\partial x^2} \ln Q(x) \bigg|_{x=\mu} \tag{4}$$

Now it is possible to approximate the probability distribution $Q(x)$ with the distribution

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{\sqrt{2\pi\Sigma}}\, e^{\frac{(x-\mu)^2}{2\Sigma}} \tag{5}$$

i.e. a Gaussian distribution that has been normalized using the factor $Q(\mu)\sqrt{2\pi\Sigma}$.
Now the VFE can be written as follow

$$\begin{aligned} F &\approx \int \mathcal{N}(x;\mu,\Sigma)(-\frac{1}{2}\ln(2\pi\Sigma) - \frac{(x-\mu)^2}{2\Sigma})dx - \int \mathcal{N}(x;\mu,\Sigma)\ln P(x,s)dx \\ &= -\frac{1}{2}\ln(2\pi\Sigma) - \frac{1}{2\Sigma}\int \mathcal{N}(x;\mu,\Sigma)(x-\mu)^2 dx - \int \mathcal{N}(x;\mu,\Sigma)\ln P(x,s)dx \\ &= -\frac{1}{2}\ln(2\pi\Sigma) - \frac{1}{2} - \int \mathcal{N}(x;\mu,\Sigma)\ln P(x,s)dx \end{aligned} \tag{6}$$

To end up with an analityc model of the FEP, further simplifications and assumptions are needed to evaluate the last term[1].
Let's first assume that the R-density is sharply peaked at its mean value and that $P(x,s)$ is a smooth function of x: under these assumptions is possible to consider the integrated function appreciably non-zero only near the peak, and is possible to use a second order Taylor expansion of the $L(x,s) \equiv -\ln P(x,s)$ around $x = \mu$.

$$L(x,s) \approx L(\mu,s) + \left[\frac{dL(x,s)}{dx}\right]_{x=\mu}(x-\mu) + \frac{1}{2}\left[\frac{d^2L(x,s)}{dx^2}\right]_{x=\mu}(x-\mu)^2 \tag{7}$$

$$\Downarrow$$

$$\begin{aligned} -\int \mathcal{N}(x;\mu,\Sigma)\ln P(x,s)dx \approx \int \mathcal{N}(x;\mu,\Sigma)\Big\{ & L(\mu,s) + \left[\frac{dL(x,s)}{dx}\right]_{x=\mu}(x-\mu) \\ & + \frac{1}{2}\left[\frac{d^2L(x,s)}{dx^2}\right]_{x=\mu}(x-\mu)^2\Big\} \\ = L(\mu,s) & + \left[\frac{dL(x,s)}{dx}\right]_{x=\mu}\left(\int \mathcal{N}(x;\mu,\Sigma)\mu dx - \mu\right) \\ & + \frac{1}{2}\left[\frac{d^2L(x,s)}{dx^2}\right]_{x=\mu}\int \mathcal{N}(x;\mu,\Sigma)(x-\mu)^2 \\ = L(\mu,s) & + \frac{1}{2}\left[\frac{d^2L(x,s)}{dx^2}\right]_{x=\mu}\Sigma \end{aligned} \tag{8}$$

Now is possible to rewrite the variational free energy as

$$F(\mu,\Sigma,s) \approx L(\mu,s) + \frac{1}{2}\left(\left[\frac{d^2L(x,s)}{dx^2}\right]_{x=\mu}\Sigma - \ln(2\pi\Sigma) - 1\right) \tag{9}$$

with $L(\mu,s)$ said *Laplace-encoded energy*, and the variational free energy written as a function and not anymore as a functional.

[1] At the end we expand the implications for the interpretation of brain functions due to this issue.

Since the goal is to minimize the Kullback-Leibler divergence trough the minimization of the VFE, is possible to simplify further removing the Σ dependency taking the derivative with respect this and imposing $\frac{dF}{d\Sigma} = 0$

$$\frac{dF}{d\Sigma} = \frac{1}{2}\left(\left[\frac{d^2L(x,s)}{dx^2}\right]_{x=\mu} - \frac{1}{\Sigma}\right) = 0 \tag{10}$$

$$\Rightarrow \Sigma = \left[\frac{d^2L(x,s)}{dx^2}\right]^{-1}_{x=\mu} \equiv \Sigma^* \tag{11}$$

The final form of the VFE is then

$$F \approx L(\mu, s) - \frac{1}{2}\ln(2\pi\Sigma^*) \tag{12}$$

that can be also written getting rid of the constant variance term

$$F \approx L(\mu, s) \tag{13}$$

Multivariate case

Generalizing for a density $Q(\mathbf{x})$ over a D-dimensional space $\mathbf{x}$ with peak at $\boldsymbol{\mu}$, let's go trough the same procedure:

$$\ln Q(\mathbf{x}) \simeq \ln Q(\boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{14}$$

with

$$\left[\boldsymbol{\Sigma}^{-1}\right]_{i,j} = -\frac{\partial^2}{\partial x_i \partial x_j}\ln Q(\mathbf{x})\bigg|_{\mathbf{x}=\boldsymbol{\mu}} \tag{15}$$

Now let's approximate $Q(\mathbf{x})$ with the multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^K \det\boldsymbol{\Sigma}}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{16}$$

and rewrite the VFE as follow:

$$\begin{aligned} F \approx & \int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\left[-\frac{1}{2}\ln\left((2\pi)^D\det\boldsymbol{\Sigma}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]d\mathbf{x} \\ & - \int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\ln P(\mathbf{x},\mathbf{s})d\mathbf{x} \\ = & -\frac{1}{2}\ln\left((2\pi)^D\det\boldsymbol{\Sigma}\right) - \int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]d\mathbf{x} \\ & - \int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\ln P(\mathbf{x},\mathbf{s})d\mathbf{x} \end{aligned} \tag{17}$$

Let's focus on the second term making as first thing the change of variables with unitary Jacobian, $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$.

$$\int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]d\mathbf{x} = \int \mathcal{N}(\mathbf{y};\mathbf{0},\boldsymbol{\Sigma})\left[\frac{1}{2}\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}\right]d\mathbf{y} \tag{18}$$

After that, since $\mathbf{\Sigma^{-1}}$ is a symmetric and real matrix, the spectral theorem guarantees the existence of an orthonormal matrix $\mathbf{U}$ such that $\mathbf{U}^T\mathbf{\Sigma}^{-1}\mathbf{U} = \mathbf{\Lambda}$, with $\mathbf{\Lambda}$ diagonal matrix containing the eigenvalues $\{\lambda_i\}_{i=1}^D$ of $\mathbf{\Sigma}^{-1}$ and $\mathbf{U}$ containing as columns the eigenvectors of $\mathbf{\Sigma}^{-1}$, so adding the identity matrix $\mathbf{I} = \mathbf{U}\mathbf{U}^T$ we get

$$\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} = \mathbf{y}^T\mathbf{U}\mathbf{U}^T\mathbf{\Sigma}^{-1}\mathbf{U}\mathbf{U}^T\mathbf{y} = \mathbf{z}^T\mathbf{\Lambda}\mathbf{z} = \sum_{i=1}^{D}\lambda_i z_i^2\,, \tag{19}$$

where $\mathbf{z} = \mathbf{U}^T\mathbf{y}$ is the $\mathbf{y}$ representation in the orthonormal basis given by the eigenvectors of $\mathbf{\Sigma}^{-1}$. Therefore moving to the variable $\mathbf{z}$ (the Jacobian of this change of variables is unitary too), we obtain

$$\frac{1}{\sqrt{(2\pi)^D \det\mathbf{\Sigma}}}\int \frac{1}{2}\left[\sum_{i=1}^{D}\lambda_i z_i^2\right] e^{-\frac{1}{2}\sum_{i=1}^{D}\lambda_i z_i^2} d\mathbf{z} = \frac{1}{2}\sqrt{\frac{\prod_{i=1}^{D}\lambda_i}{(2\pi)^D}}\sum_{i=1}^{D}\int \lambda_i z_i^2\, e^{-\frac{1}{2}\sum_{i=1}^{D}\lambda_i z_i^2} d\mathbf{z} \tag{20}$$

indicating with $\mathbf{z}_{\not i} \equiv \{z_j\}_{j\neq i}$

$$\frac{1}{2}\sqrt{\frac{\prod_{i=1}^{D}\lambda_i}{(2\pi)^D}}\sum_{i=1}^{D}\int \lambda_i z_i^2\, e^{-\frac{1}{2}\lambda_i z_i^2} dz_i \int e^{\frac{1}{2}\sum_{j\neq i}\lambda_i z_j^2} d\mathbf{z}_{\not i} = \frac{1}{2}\sqrt{\frac{\prod_{i=1}^{D}\lambda_i}{(2\pi)^D}}\sum_{i=1}^{D}\lambda_i\sqrt{\frac{2\pi}{\lambda_i}}\frac{1}{\lambda_i}\sqrt{\frac{(2\pi)^{D-1}}{\prod_{\not i}\lambda_{\not i}}} \tag{21}$$

$$\begin{aligned}\int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\mathbf{\Sigma})\left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] d\mathbf{x} &= \int \mathcal{N}(\mathbf{y};\mathbf{0},\mathbf{\Sigma})\left[\frac{1}{2}(\mathbf{y})^T\mathbf{\Sigma}^{-1}(\mathbf{y})\right] d\mathbf{y} \\ &= \frac{1}{\sqrt{(2\pi)^K \det\mathbf{\Sigma}}}\int e^{-\frac{1}{2}(\mathbf{y})^T\mathbf{\Sigma}^{-1}(\mathbf{y})}\left[\frac{1}{2}(\mathbf{y})^T\mathbf{\Sigma}^{-1}(\mathbf{y})\right] d\mathbf{y}\end{aligned} \tag{22}$$

Let's now evaluate the third term of Eq.(17)

$$\vdots$$

$$F \approx L(\boldsymbol{\mu}, \mathbf{s}) \tag{23}$$

Laplace encoded free energy

In this case indeed this approximation is used in the following manner (in the following only the univariate case is presented in detail since it captures all the relevant assumptions).

First of all the R-densities are assumed Gaussian distributions

$$Q(x) \equiv \mathcal{N}(x; \mu, \Sigma) = \mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{2\pi\Sigma}} e^{\frac{(x-\mu)^2}{2\Sigma}}$$

This approximation is particularly useful to approximate integrals.

Laplace encoded free energy

Let's first rewrite F as follow:

$$F = \langle \ln P(\mathbf{x}, \mathbf{s}) \rangle_Q - \langle \ln Q(\mathbf{x}) \rangle_Q$$

with L called *Laplace encoded energy* and the second term is referred as entropy. Often optimizing F for arbitrary $Q(x)$ is particularly complex, so a common approximation

References

Christopher L., Buckley et al. (Dec. 2017). "The free energy principle for action and perception: A mathematical review". In: *Journal of Mathematical Psychology* 81, pp. 55–79. DOI: 10.1016/j.jmp.2017.09.004. URL: https://doi.org/10.1016/j.jmp.2017.09.004.

Manuel, Baltieri and Buckley Christopher (Mar. 2019). "PID Control as a Process of Active Inference with Linear Generative Models". In: *Entropy* 21.3, p. 257. DOI: 10.3390/e21030257. URL: https://doi.org/10.3390/e21030257.