Active Inference in continuous time notes

Federico Maggiore

December 1, 2020

Contents

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;
- $\boldsymbol{\mu} = \{\mu_i\}_{i=1}^{D}$ inner brain state representing the hidden environmental variables $\mathbf{x}$
- $P(\mathbf{x}, \mathbf{s})$ *Joint 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*;
- ε *Prediction error*;
- Σ^{-1} *Precision*

References: Baltieri and C. Buckley 2019, C. L. Buckley et al. 2017, Bogacz 2017,

1 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

- $P(\mathbf{x}, \mathbf{s})$ *joint 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 joint 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 with respect to the recognition density $Q(\mathbf{x})$ ($\Rightarrow$ minimizing F with respect to $Q(\mathbf{x})$ will minimize the D_{KL})

2 One dimensional case

For the sake of simplicity, let's build the framework first in the one dimensional case, to repeat later on all the steps for the multivariate case.

2.1 Laplace approximation

Often optimizing F for arbitrary $Q(x)$ 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.

Let us 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

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

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 us 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 $x = \mu$.

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

$$\Downarrow$$

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

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

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

that can be further simplified if the L function has a second order polynomial form[2], that implies that the second-order derivative of L with respect to x results in a constant factor that is useless and can be ignored[3], leading to

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

About Laplace Approximation Carrying on a theory based on an approximation that leads to Eq.(13), means to assume that the brain represents only the most likely environmental cause of sensory data and nothing else about the distribution. However, as we're going to see, the uncertainties are encoded directly in the form of the joint density.

2.2 Building the generative model

Thanks to the Laplace approximation, we've been able to write the VFE in terms of the Laplace-encoded energy $L(\mu, s)$, that in turn depends on the joint density $P(x, s)$. In this function are encoded the brain beliefs about the environmental causes of the sensory input and the beliefs *a priori* about environmental states.

Therefore, what needs to be built is a *generative model*, that is a model in which is encoded how the brain believe the world works and where all the hypothesis about the agent's behaviour are formalized.

2.2.1 Static Model

Let us consider a simple case of an agent that believes in an environment with hidden state x that stimulates a sensory channel s. As we've seen in Sec.(2.1), the brain will represent the environment only through the inner state μ, and what remains to do is to explicit the mapping between brain states and sensory data that will allows to make explicit the joint density.

Let us assume that the agent believes its sensory input are generated by

$$s = g(x) + \mathcal{N}(s; 0, \Sigma_s) , \tag{14}$$

[2] as we will see in Sec.(2.2)

[3] let us remind that the final goal is to minimize F with respect to x

with g generic function that expresses the relation between states and sensory input, to which is summed a noise represented by the normal distribution with zero mean and variance Σ_s. This assumption means that we can write

$$P(s|x) = \mathcal{N}(s; g(x), \Sigma_s) = \frac{1}{\sqrt{2\pi\Sigma_s}} e^{\frac{(s-g(x))^2}{2\Sigma_s}} . \tag{15}$$

Moreover let us also assume that the agent also has a prior knowledge regarding the environmental state given by $\bar{\mu}$ that is linked with the inner state through

$$x = \bar{\mu} + \mathcal{N}(x; 0, \Sigma_x) \tag{16}$$

$$\Rightarrow P(x) = \mathcal{N}(x; \bar{\mu}, \Sigma_x) = \frac{1}{\sqrt{2\pi\Sigma_x}} e^{\frac{(x-\bar{\mu})^2}{2\Sigma_x}} . \tag{17}$$

Now that we have specified a likelihood and a prior, is possible to determine the joint density

$$P(x, s) = P(s|x)P(x) \tag{18}$$

and consequently the Laplace-encoded energy

$$\begin{aligned} L(\mu, s) &= -\ln P(s|\mu) - \ln P(\mu) \\ &= \frac{1}{2}\ln(2\pi\Sigma_s) + \frac{(s-\mu)^2}{2\Sigma_s} + \frac{1}{2}\ln(2\pi\Sigma_\mu) + \frac{(\mu-\bar{\mu})^2}{2\Sigma_\mu} \\ &= \frac{\varepsilon_s^2}{2\Sigma_s} + \frac{\varepsilon_\mu^2}{2\Sigma_\mu} + \frac{1}{2}\ln(\Sigma_s\Sigma_\mu) + \ln(2\pi) , \end{aligned} \tag{19}$$

where the ε terms are said *prediction errors* and measure the discrepancy respectively between the actual sensory data s and the outcome of its prediction $g(\mu)$ and between μ itself and its prior expectation $\bar{\mu}$. Therefore the former ε_s describes sensory prediction errors, the latter ε_μ model prediction errors (i.e. how brain states deviate from their expectation) and each one is weighted with the the corresponding inverse of the variance Σ_s^{-1} and Σ_μ^{-1} (which are often said *precisions*). As said at the end of Sec.(2.1), since the L function has a quadratic form is possible to ignore all the terms apart from the following

$$F \approx \frac{1}{2}\left[\frac{\varepsilon_s^2}{\Sigma_s} + \frac{\varepsilon_\mu^2}{\Sigma_\mu} + \ln(\Sigma_s\Sigma_\mu)\right] . \tag{20}$$

2.2.2 Dynamic Model

Let's formulate a possible implementation of inference in a dynamically changing environment.

As the static case, is assumed that the agent believes its sensory input are generated in a similar manner with respect to Eq.(14), in particular

$$s(x, t) = g(x) + z_s(t) . \tag{21}$$

Additionally, is assumed that the agent model of environmental dynamic follows the Langevin-type equation

$$\frac{dx(t)}{dt} = f(x) + z_x(t) . \tag{22}$$

Both z and w are terms representing noise and that we will specify later.

Using a *generalised state-space model*, in which the state of a dynamical system is represented in terms of increasingly higher order derivative of its state variables, in combination with local linearity approximation on higher orders of motion[4] suppressing non-linear terms in the partial derivatives, is possible to obtain

$$\begin{aligned} s &= g(x) + z_s(t) & x' &= f(x) + z_x(t) \\ s' &= \frac{\partial g}{\partial x} x' + z'_s & x'' &= \frac{\partial f}{\partial x} x' + z'_x \\ s'' &\simeq \frac{\partial g}{\partial x} x'' + z''_s & x''' &\simeq \frac{\partial f}{\partial x} x'' + z''_x \\ &\vdots & &\vdots \end{aligned} \tag{23}$$

where we have used the notation

$$s' = \frac{ds}{dt}\,, \qquad x' = \frac{dx}{dt}\,, \qquad s'' = \frac{d^2 s}{dt^2}\,, \qquad x'' = \frac{d^2 x}{dt^2}\,, \quad \dots \tag{24}$$

and where $z_s, z'_s, z''_s, \dots, z_x, z'_x, z''_x, \dots$ are the noises source at each dynamic order. Considering the previous linear approximation as equalities, Eq.(23) can be expressed in the more compact form

$$\tilde{s} = g(\tilde{x}) + \tilde{z_s} \quad , \quad \tilde{x}' = f(\tilde{x}) + \tilde{z_x} \tag{25}$$

using the notation

$$\tilde{s} = (s, s', s'', \dots) \equiv (s_{[0]}, s_{[1]}, s_{[2]}, \dots) \quad , \quad \tilde{x} = (x, x', x'', \dots) \equiv (x_{[0]}, x_{[1]}, x_{[2]}, \dots)\,, \tag{26}$$

where

$$s_{[n]} \equiv \frac{d^n}{dt^n} s = s'_{[n-1]} \quad , \quad x_{[n]} \equiv \frac{d^n}{dt^n} x = x'_{[n-1]}\,, \tag{27}$$

$$\tilde{x}' \equiv D\tilde{x} = \frac{d}{dt}(x, x', x'', \dots) = (x', x'', x''', \dots) \equiv (x_{[1]}, x_{[2]}, x_{[3]}, \dots) \tag{28}$$

and

$$\tilde{g} \equiv (g_{[0]}, g_{[1]}, g_{[2]}, \dots) \quad , \quad \tilde{f} \equiv (f_{[0]}, f_{[1]}, f_{[2]}, \dots) \tag{29}$$

with

$$\begin{aligned} g_{[0]} &\equiv g(x) & f_{[0]} &\equiv f(x) \\ g_{[1]} &\equiv \frac{\partial g}{\partial x} x_{[0]} & f_{[1]} &\equiv \frac{\partial f}{\partial x} x \\ &\vdots & &\vdots \\ g_{[n]} &\equiv \frac{\partial g}{\partial x} x_{[n]} & f_{[n]} &\equiv \frac{\partial f}{\partial x} x_{[n]} \\ &\vdots & &\vdots \end{aligned} \tag{30}$$

[4]Without this approximation the model would scale-up very quickly becoming complicated and unwieldy fairly quickly. This approximation becomes exact when f and g are linear.

Considerations about noise sources the stochastic terms $z_s(t)$ and $z_x(t)$ in Eq.(21) and Eq.(22) are analytic and form stochastic equations based on Stratonovich calculus, with well defined covariances of $\tilde{z_s}(t) = (z_{s[0]}, z_{s[1]}, z_{s[2]}, \dots)$ and $\tilde{z_x}(t) = (z_{x[0]}, z_{x[1]}, z_{x[2]}, \dots)$. This property is important to define a non-Markovian process, because in Ito's formulation, based on Wiener noise, the autocorrelation functions can be seen as strictly equal to delta functions representing perfect white noise not existing in real world. **Da approfondire**
A common approximation that brings to a really simple form of the joint density $P(x, s)$ is the one in which the covariances between dynamical orders are assumed equal to zero, i.e. independent noise sources, that leads to

$$\begin{aligned} P(\tilde{s}|\tilde{x}) &= P(s_{[0]}, s_{[1]}, s_{[2]}, \dots | x_{[0]}, x_{[1]}, x_{[2]}, \dots) = \prod_{n=0}^{\infty} P(s_{[n]}|x_{[n]}) \\ P(\tilde{x}) &= P(x_{[0]}, x_{[1]}, x_{[2]}, \dots) = P(x_{[0]}) \prod_{n=0}^{\infty} P(x_{[n+1]}|x_{[n]}) \end{aligned} \tag{31}$$

Assuming in addition that at all dynamics orders $z_{s[n]}$ and $z_{x[n]}$ have a Gaussian form we can write

$$P(s_{[n]}|x_{[n]}) = \mathcal{N}(s_{[n]}; g_{[n]}, \Sigma_{s[n]}) = \frac{1}{\sqrt{2\pi\Sigma_{s[n]}}} e^{-\frac{(s_{[n]} - g_{[n]})^2}{2\Sigma_{s[n]}}} \tag{32}$$

and

$$P(x_{[n+1]}|x_{[n]}) = \mathcal{N}(x_{[n+1]}; f_{[n]}, \Sigma_{x[n]}) = \frac{1}{\sqrt{2\pi\Sigma_{x[n]}}} e^{-\frac{(x_{[n+1]} - f_{[n]})^2}{2\Sigma_{x[n]}}} . \tag{33}$$

In dynamical systems the $x_{[0]}$ prior at each time step is usually omitted because the agent does not explicitly desire any $x_{[0]}$, giving it a flat prior (i.e. infinite variance) that eliminates the corresponding term from the VFE.

In conclusion the joint density can be write as

$$P(\tilde{x}, \tilde{s}) = \prod_{n=0}^{\infty} P(s_{[n]}|x_{[n]}) P(x_{[n+1]}|x_{[n]}) \tag{34}$$

that, inserted in the Laplace-encoded energy lead to

$$\begin{aligned} L(\tilde{\mu}, \tilde{s}) &= \sum_{n=0}^{\infty} \left[\frac{(s_{[n]} - g_{[n]})^2}{2\Sigma_{s_{[n]}}} + \frac{1}{2}\ln(2\pi\Sigma_{s_{[n]}}) \right] + \left[\frac{(\mu_{[n+1]} - f_{[n]})^2}{2\Sigma_{\mu_{[n]}}} + \frac{1}{2}\ln(2\pi\Sigma_{\mu_{[n]}}) \right] \\ &= \sum_{n=0}^{\infty} \frac{\varepsilon_{s_{[n]}}^2}{2\Sigma_{s_{[n]}}} + \frac{\varepsilon_{\mu_{[n]}}^2}{2\Sigma_{\mu_{[n]}}} + \frac{1}{2}\ln(\Sigma_{s_{[n]}}\Sigma_{\mu_{[n]}}) + \ln(2\pi) \end{aligned} \tag{35}$$

with $\varepsilon_{s_{[n]}} \equiv s_{[n]} - g_{[n]}$ and $\varepsilon_{\mu_{[n]}} \equiv \mu_{[n+1]} - f_{[n]}$ nth component of $\tilde{\varepsilon_s}$ and $\tilde{\varepsilon_\mu}$ that, as in the static case, are said *prediction errors*, which encode respectively the discrepancy between sensory data $\tilde{s}$ and its prediction $\tilde{g}$ and the difference between the expected higher-order output $\tilde{\mu}'$ and its generation $\tilde{f}$.

Usually only dynamics up to a finite order n_{max} are considered, and this is done by setting

$$\mu_{[n_{max}+1]} = z_{\mu[n_{max}]} \tag{36}$$

with $\Sigma_{\mu_{[n_{max}]}}$ large, so that the corresponding error term will be close to zero and can be ignored in the Laplace-encoded energy, meaning that the order below is unconstrained and free to change in a

way that best fits the incoming sensory data.
Finally, since also in this case the L function has a quadratic form, when writing down the VFE is possible to ignore all the terms apart from

$$F \approx \frac{1}{2} \sum_{n=0}^{n_{max}} \left[\frac{\varepsilon^2_{s_{[n]}}}{\Sigma_{s_{[n]}}} + \frac{\varepsilon^2_{\mu_{[n]}}}{\Sigma_{\mu_{[n]}}} + \ln(\Sigma_{s_{[n]}} \Sigma_{\mu_{[n]}}) \right] \tag{37}$$

2.3 VFE minimisation

In Sec.(2.1) we have expressed, in an approximated form, the Variational Free Energy in terms of the Laplace-encoded Energy, which depends on joint density $P(\mathbf{x}, \mathbf{s})$, that we have seen how to build starting from the generative model in Sec.(2.2). Now we are going to see a possible implementation of a biologically plausible mechanism to minimise VFE.

In the Free Energy Principle framework, it is proposed that the innate dynamics of the neural activity evolves in such a way that it implement a gradient descent scheme on the VFE.
In particular, in the static model case, a brain state μ is updated between two (internal) sequential steps t and $t + dt$ as

$$\mu^{t+dt} = \mu^t - k \cdot \nabla_\mu F = \mu^t - k \cdot \nabla_\mu L(\mu, s)\,, \tag{38}$$

with k learning rate parameter that has to be tuned and $\nabla_\mu L(\mu, s)$ that goes to zero when a minimum of the L function is reached.
In the dynamic case instead,

3 Multivariate case

Now that we have seen how to organize an active inference framework in the one dimensional case, let's see how to scale up the dimensions of our system remembering that $\mathbf{x} = \{x_i\}_{i=1}^D$, $\mathbf{s} = \{s_i\}_{i=1}^S$ and $\boldsymbol{\mu} = \{\mu_i\}_{i=1}^D$.

Let's recall that the goal is to minimize the Kullback-Leibler divergence

$$D_{KL}(\, Q(\mathbf{x}) \,||\, P(\mathbf{x}|\mathbf{s})\,) = F + \ln P(\mathbf{s}) = -\left\langle \ln P(\mathbf{x},\mathbf{s})\right\rangle_Q + \left\langle \ln Q(\mathbf{x})\right\rangle_Q \tag{39}$$

that, since $P(\mathbf{s})$ is independent with respect to the recognition density $Q(\mathbf{x})$, is equivalent to minimize the Variational Free Energy

$$F = \int Q(\mathbf{x}) \ln Q(\mathbf{x}) d\mathbf{x} - \int Q(\mathbf{x}) \ln P(\mathbf{x},\mathbf{s}) d\mathbf{x} \tag{40}$$

3.1 Laplace approximation

As done before, let's start assuming that neural activity parametrise sufficient statistic, in particular let's approximate the R-density $Q(\mathbf{x})$ with a multivariate Gaussian over the D-dimensional space $\mathbf{x}$, with peak at $\boldsymbol{\mu}$, and let's do the same procedure done in Sec.(2.1).

$$\ln Q(\mathbf{x}) \simeq \ln Q(\boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \tag{41}$$

with

$$\left[\hat{\boldsymbol{\Sigma}}^{-1}\right]_{i,j} = -\frac{\partial^2}{\partial x_i \partial x_j} \ln Q(\mathbf{x})\bigg|_{\mathbf{x}=\boldsymbol{\mu}} \tag{42}$$

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

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\boldsymbol{\Sigma}}) = \frac{1}{\sqrt{(2\pi)^D \det \hat{\boldsymbol{\Sigma}}}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{43}$$

and rewrite the VFE as follow:

$$\begin{aligned} F \approx & \int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\boldsymbol{\Sigma}}) \left[-\frac{1}{2} \ln\left((2\pi)^D \det \hat{\boldsymbol{\Sigma}}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right] d\mathbf{x} + \\ & - \int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\boldsymbol{\Sigma}}) \ln P(\mathbf{x},\mathbf{s}) d\mathbf{x} \\ = & -\frac{1}{2} \ln\left((2\pi)^D \det \hat{\boldsymbol{\Sigma}}\right) - \int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\boldsymbol{\Sigma}}) \left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right] d\mathbf{x} + \\ & - \int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\boldsymbol{\Sigma}}) \ln P(\mathbf{x},\mathbf{s}) d\mathbf{x} \end{aligned} \tag{44}$$

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

$$\int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\boldsymbol{\Sigma}}) \left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right] d\mathbf{x} = \int \mathcal{N}(\mathbf{y};\mathbf{0},\hat{\boldsymbol{\Sigma}}) \left[\frac{1}{2}\mathbf{y}^T \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y} \right] d\mathbf{y} \tag{45}$$

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

$$\mathbf{y}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{y} = \mathbf{y}^T\mathbf{U}\mathbf{U}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{U}\mathbf{U}^T\mathbf{y} = \mathbf{z}^T\mathbf{\Lambda}\mathbf{z} = \sum_{i=1}^{D}\lambda_i z_i^2\,, \tag{46}$$

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

$$\frac{1}{\sqrt{(2\pi)^D \det\hat{\mathbf{\Sigma}}}}\int \frac{1}{2}\left[\sum_{i=1}^{D}\lambda_i z_i^2\right] e^{-\frac{1}{2}\sum_{i=1}^{D}\lambda_i z_i^2} d\mathbf{z} = \frac{1}{2}\sqrt{\frac{\prod_{i=1}^{D}\lambda_i}{(2\pi)^D}}\sum_{i=1}^{D}\int \lambda_i z_i^2\, e^{-\frac{1}{2}\sum_{i=1}^{D}\lambda_i z_i^2} dz_i \tag{47}$$

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

$$\begin{aligned}
&\frac{1}{2}\sqrt{\frac{\prod_{i=1}^{D}\lambda_i}{(2\pi)^D}}\sum_{i=1}^{D}\int \lambda_i z_i^2\, e^{-\frac{1}{2}\lambda_i z_i^2} dz_i \int e^{\frac{1}{2}\sum_{j\neq i}\lambda_i z_j^2} d\mathbf{z}_{\not i} = \\
&\qquad = \frac{1}{2}\sqrt{\frac{\prod_{i=1}^{D}\lambda_i}{(2\pi)^D}}\sum_{i=1}^{D}\lambda_i\sqrt{\frac{2\pi}{\lambda_i}}\frac{1}{\lambda_i}\sqrt{\frac{(2\pi)^{D-1}}{\prod_{\not i}\lambda_{\not i}}} = \frac{1}{2}\sqrt{\frac{\prod_{i=1}^{D}\lambda_i}{(2\pi)^D}}\sum_{i=1}^{D}\sqrt{\frac{(2\pi)^D}{\prod_{i=1}^{D}\lambda_i}} = \frac{D}{2}
\end{aligned} \tag{48}$$

In conclusion then

$$\int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\mathbf{\Sigma}})\left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] d\mathbf{x} = \frac{D}{2} \tag{49}$$

$$\Rightarrow F \approx -\frac{1}{2}\ln\left((2\pi)^D \det\hat{\mathbf{\Sigma}}\right) - \frac{D}{2} - \int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\mathbf{\Sigma}})\ln P(\mathbf{x},\mathbf{s})d\mathbf{x} \tag{50}$$

Let us now evaluate the last term. As done before, let us assume that the R-density is sharply peaked at its mean value (i.e. both variance of and covariances between variables are small) and that $P(\mathbf{x},\mathbf{s})$ is a smooth function of $\mathbf{x}$: 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 $L(\mathbf{x},\mathbf{s}) \equiv -\ln P(\mathbf{x},\mathbf{s})$ around $\mathbf{x} = \boldsymbol{\mu}$.

$$L(\mathbf{x},\mathbf{s}) \approx L(\boldsymbol{\mu},\mathbf{s}) + (\mathbf{x}-\boldsymbol{\mu})^T\left[\nabla L(\mathbf{x},\mathbf{s})\right]_{\mathbf{x}=\boldsymbol{\mu}} + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\left[\nabla^2 L(\mathbf{x},\mathbf{s})\right]_{x=\mu}(\mathbf{x}-\boldsymbol{\mu}) \tag{51}$$

$$\Downarrow$$

$$\begin{aligned}
-\int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\mathbf{\Sigma}})\ln P(\mathbf{x},\mathbf{s})d\mathbf{x} &\approx \int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\mathbf{\Sigma}})\Big\{L(\boldsymbol{\mu},\mathbf{s}) + (\mathbf{x}-\boldsymbol{\mu})^T\left[\nabla L(\mathbf{x},\mathbf{s})\right]_{\mathbf{x}=\boldsymbol{\mu}} + \\
&\quad + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\left[\nabla^2 L(\mathbf{x},\mathbf{s})\right]_{x=\mu}(\mathbf{x}-\boldsymbol{\mu})\Big\}d\mathbf{x} \\
&= L(\boldsymbol{\mu},\mathbf{s}) + \left(\int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\mathbf{\Sigma}})\mathbf{x}^T\, d\mathbf{x} - \boldsymbol{\mu}^T\right)\left[\nabla L(\mathbf{x},\mathbf{s})\right]_{\mathbf{x}=\boldsymbol{\mu}} \\
&\quad + \frac{1}{2}\int \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\hat{\mathbf{\Sigma}})(\mathbf{x}-\boldsymbol{\mu})^T\left[\nabla^2 L(\mathbf{x},\mathbf{s})\right]_{x=\mu}(\mathbf{x}-\boldsymbol{\mu})\, d\mathbf{x} \\
&\overset{?}{=} L(\boldsymbol{\mu},\mathbf{s}) + \frac{D}{2}\left[\nabla^2 L(\mathbf{x},\mathbf{s})\right]_{x=\mu}\hat{\mathbf{\Sigma}}
\end{aligned} \tag{52}$$

$$\vdots$$

$$F \approx L(\boldsymbol{\mu}, \mathbf{s}) \tag{53}$$

3.2 Building the generative model

Now that we've expressed the VFE in terms of the Laplace-encoded Energy $L(\boldsymbol{\mu}, \mathbf{s})$, let's develop the generative model through the building of the joint density $P(\mathbf{x}, \mathbf{s})$.

3.2.1 Static Model

Here the agent believes in an environment with hidden state $\mathbf{x}$ that stimulates a sensory channel $\mathbf{s}$ and that it builds the joint density using a likelihood and a prior.

The prior is built from the expectations

$$\mathbf{x} = \bar{\boldsymbol{\mu}} + \mathbf{z}_{\bar{\boldsymbol{\mu}}}\,, \tag{54}$$

where $\mathbf{z}_{\bar{\boldsymbol{\mu}}}$ is a D-dimensional vector describing correlated noise with zero mean and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\mu}}}$ that allows us to write

$$P(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \bar{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\mu}}}) = \frac{1}{\sqrt{(2\pi)^D \det \hat{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\mu}}}}}\, e^{-\frac{1}{2}(\mathbf{x}-\bar{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\mu}}}^{-1}(\mathbf{x}-\bar{\boldsymbol{\mu}})}\,. \tag{55}$$

Similarly the sensory inputs are assumed to be generated by

$$\mathbf{s} = \boldsymbol{g}(\mathbf{x}) + \mathbf{z}_{\mathbf{s}}\,, \tag{56}$$

with again $\mathbf{z}_{\mathbf{s}}$ S-dimensional vector describing correlated noise with zero mean and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{s}}$ that brings to the likelihood

$$P(\mathbf{s}|\mathbf{x}) = \mathcal{N}(\mathbf{s}; \boldsymbol{g}(\mathbf{x}), \hat{\boldsymbol{\Sigma}}_{\mathbf{s}}) = \frac{1}{\sqrt{(2\pi)^S \det \hat{\boldsymbol{\Sigma}}_{\mathbf{s}}}}\, e^{-\frac{1}{2}(\mathbf{s}-\boldsymbol{g}(\mathbf{x}))^T \hat{\boldsymbol{\Sigma}}_{\mathbf{s}}^{-1}(\mathbf{s}-\boldsymbol{g}(\mathbf{x}))}\,. \tag{57}$$

Now is possible to write the Laplace-encoded energy

$$\begin{aligned} L(\boldsymbol{\mu}, \mathbf{s}) &= \frac{1}{2}\ln\left((2\pi)^S \det \hat{\boldsymbol{\Sigma}}_s\right) + \frac{1}{2}\left(\mathbf{s} - \boldsymbol{g}(\boldsymbol{\mu})\right)^T \hat{\boldsymbol{\Sigma}}_{\mathbf{s}}^{-1}\left(\mathbf{s} - \boldsymbol{g}(\boldsymbol{\mu})\right) \\ &\quad + \frac{1}{2}\ln\left((2\pi)^D \det \hat{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\mu}}}\right) + \frac{1}{2}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\mu}}}^{-1}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \\ &= \frac{1}{2}\boldsymbol{\varepsilon}_{\mathbf{s}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{s}}^{-1}\boldsymbol{\varepsilon}_{\mathbf{s}} + \frac{1}{2}\boldsymbol{\varepsilon}_{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\mu}}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\mu}} \\ &\quad + \frac{1}{2}\ln\left(\det \hat{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\mu}}} \det \hat{\boldsymbol{\Sigma}}_s\right) + \frac{1}{2}\ln\left((2\pi)^D (2\pi)^S\right)\,, \end{aligned} \tag{58}$$

where again we have introduced the prediction errors $\boldsymbol{\varepsilon}_{\mathbf{s}} \equiv \mathbf{s} - \boldsymbol{g}(\boldsymbol{\mu})$ and $\boldsymbol{\varepsilon}_{\boldsymbol{\mu}} \equiv \boldsymbol{\mu} - \bar{\boldsymbol{\mu}}$ and we're able to write the VFE approximated form getting rid of all constants and the term depending on $\hat{\boldsymbol{\Sigma}}^{\star}$:

$$F \approx \frac{1}{2}\boldsymbol{\varepsilon}_{\mathbf{s}}^{T}\hat{\boldsymbol{\Sigma}}_{\mathbf{s}}^{-1}\boldsymbol{\varepsilon}_{\mathbf{s}} + \frac{1}{2}\boldsymbol{\varepsilon}_{\boldsymbol{\mu}}^{T}\hat{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\mu}}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\mu}} + \frac{1}{2}\ln\left(\det\hat{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\mu}}}\det\hat{\boldsymbol{\Sigma}}_{s}\right) . \tag{59}$$

Often, the VFE is further simplified making another assumption, that is the *mean field approximation*, which implies statistical independence between environmental variables and between sensory inputs[5], leading to a likelihood and a prior factorised respectively

$$P(\mathbf{x}) = \prod_{j=1}^{D}\mathcal{N}(x_j; \bar{\mu}_j, \Sigma_{\bar{\boldsymbol{\mu}}}^{(j,j)}) \tag{60}$$

$$P(\mathbf{s}|\mathbf{x}) = \prod_{i=1}^{S}\mathcal{N}(s_i; g_i(\mathbf{x}), \Sigma_{\mathbf{s}}^{(i,i)}) \tag{61}$$

and a VFE with form

$$\begin{aligned} F &\approx \sum_{i=1}^{S}\left[\frac{(\varepsilon_{\mathbf{s}}^{(i)})^2}{2\Sigma_{\mathbf{s}}^{(i,i)}} + \frac{1}{2}\ln\Sigma_{\mathbf{s}}^{(i,i)}\right] + \sum_{j=1}^{D}\left[\frac{(\varepsilon_{\boldsymbol{\mu}}^{(j)})^2}{2\Sigma_{\bar{\boldsymbol{\mu}}}^{(j,j)}} + \frac{1}{2}\ln\Sigma_{\bar{\boldsymbol{\mu}}}^{(j,j)}\right] \\ &= \sum_{i=1}^{S}\left[\frac{(\varepsilon_{\mathbf{s}}^{(i)})^2}{2\Sigma_{\mathbf{s}}^{(i,i)}}\right] + \sum_{j=1}^{D}\left[\frac{(\varepsilon_{\boldsymbol{\mu}}^{(j)})^2}{2\Sigma_{\boldsymbol{\mu}}^{(j,j)}}\right] + \frac{1}{2}\ln\left(\det\hat{\boldsymbol{\Sigma}}_{\bar{\boldsymbol{\mu}}}\det\hat{\boldsymbol{\Sigma}}_{s}\right) . \end{aligned} \tag{62}$$

[5] $\Rightarrow \forall\ i \neq j\ \Sigma_{\bar{\boldsymbol{\mu}}}^{(i,j)} = 0\,, \Sigma_{\mathbf{s}}^{(i,j)} = 0$

References

Baltieri, Manuel and Christopher Buckley (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.

Bogacz, Rafal (2017). "A tutorial on the free-energy framework for modelling perception and learning". In: *Journal of Mathematical Psychology* 76. Model-based Cognitive Neuroscience, pp. 198–211. ISSN: 0022-2496. DOI: https://doi.org/10.1016/j.jmp.2015.11.003. URL: http://www.sciencedirect.com/science/article/pii/S0022249615000759.

Buckley, Christopher L. 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.