ACMS Bayesian(60880) Project: How do predictive coding and free-energy principle work in the brain

Renjun Zhu

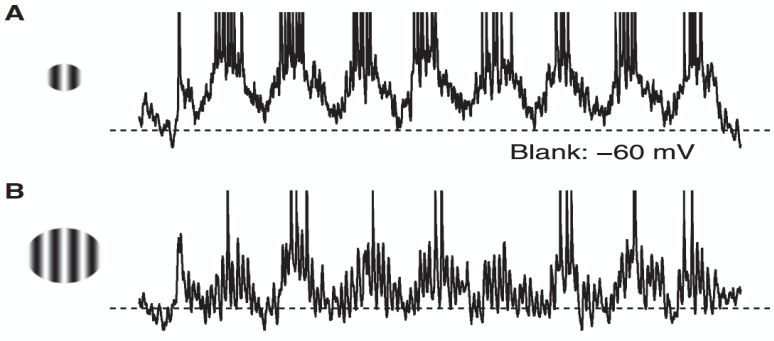
December 18, 2019

Abstract

My work focuses on implementing predictive coding and understanding free-energy principle in the brain, and comparing it with some basic Bayesian Monte Carlo approach. Rao and Balllard first introduced predictive coding model with natural images data to learn features resembling the receptive fields of neurones in the primary visual cortex. The model aims to find the optimal estimates of parameters in the hierarchical neuronal network; whereas free-energy principle proposed by Karl Friston are associated with learning the variance and covariance of the parameters. Although the set up of the two models rely on the knowledge of bayesian prior, both apply gradient ascent or decent method to approximate their estimates.

1 Introduction

Bayesian theory has been well developed in the past few decades to understand uncertainty, and methods such as Monte Carlo(MC) simulations are widely used in many research areas to draw conclusions about model parameters. In the field of neurone science, brains are multi-layer structured and have rich dynamics. Applying bayesian inference in the primary visual cortex can see how the sensory cortex get information from the noisy environment.

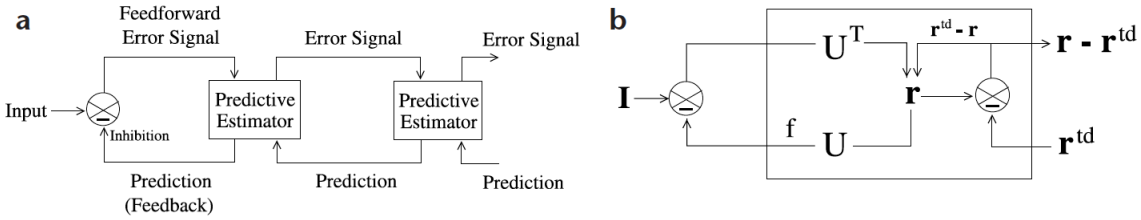


Scientists found from the experiment(see above figure) that when extend stimulus beyond classical receptive field(A, figure), neurones at the centre are suppressed(B, figure) with a decreasing spiking firing rate.

2 Predictive Coding Model

2.1 Model definition

Rao and Ballard used natural images, where the neighbouring pixel intensities are correlated in a level just like the neurones' spiking activities, to capture this end-stopping effect, a reduction of response behaviours(1999). They proposed a model of visual process(see below figure, a): when given an input stimulus, starting from the bottom up, the higher order visual cortical areas(i.e.V2) bring the predictions of the expected neural activities to the lower order areas(i.e. V1) through feedback connections; then the residual errors that are not predicted by V2 are calculated and pass through the feedforward connections in V1.



2.2 Method and Findings

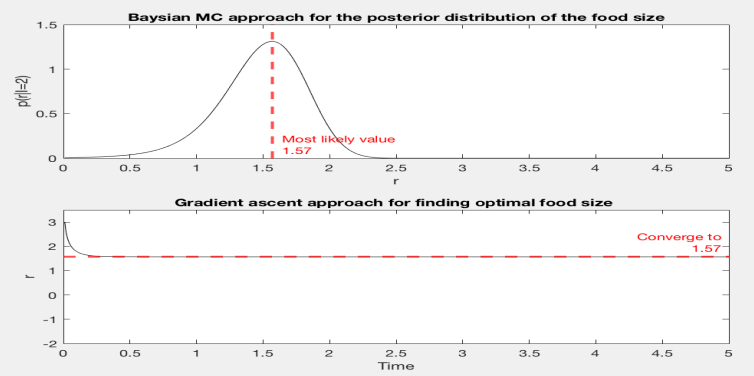
Mathematically, this can be a hierarchical model trying to estimate the hidden causes parameter, $\mathbf{r}$, in the network via synaptic learning, $\mathbf{U}$(see above figure, b). For level 1: $\mathbf{I} = f(\mathbf{U}\mathbf{r}) + \mathbf{n}$, with $f(\mathbf{U}\mathbf{r}) = f(\sum_j U_j r_j)$. r_j correspond to the activities such as the firing rates of neurones or the internal representation of the spatial characteristics of the image $\mathbf{I}$ in Rao and Ballard's original paper(1999). U_j are the columns of $\mathbf{U}$, representing synaptic weights of neurones or basis vectors for generating images. f is the neuronal activation function, which can be linear or nonlinear; typically, it is a sigmoidal function, i.e. $f(x) = tanh(x)$. $\mathbf{n} \sim N(0, \sigma^2)$ is the prediction errors that assume to be Gaussian. Similar set up for the higher lever: $\mathbf{r} = \mathbf{r}^{td} + \mathbf{n}^{td}$, with $\mathbf{r}^{td} = f(\mathbf{U}^h \mathbf{r}^h)$. $\mathbf{r}^{td}$ is the top-down prediction of $\mathbf{r}$, and $\mathbf{n}^{td} \sim N(0, \sigma_{td}^2)$ is the corresponding prediction errors.

Many literature(Rao and Ballard, 1999; Bogacz, 2017) have taken the Gaussian prior on $p(r_j)$ and $p(U_{ij})$, and also for the gaussian likelihood. Instead of directly calculating the posterior distribution, they take the negative logarithm of the posterior, and obtain the optimisation function, $E = E_1 + \alpha \sum_i r_i^2 + \lambda \sum_{i,j} U_{ij}^2$, where $\alpha, \lambda \in \mathbb{R}$. After applying gradient ascent or descent for the optimal estimates of $\mathbf{r}$ and $\mathbf{U}$, the network then has dynamics, $\frac{d\mathbf{r}}{dt} = -\frac{k_1}{2}\frac{\partial E}{\partial \mathbf{r}}$ and synaptic learning, $\frac{d\mathbf{U}}{dt} = -\frac{k_2}{2}\frac{\partial E}{\partial \mathbf{U}}$, where $k_1, k_2 \in \mathbb{R}$.

For illustration, consider a toy example that an animal wants to infer the size of a food item. Apply bayesian MC simulations and comparing it with the gradient ascent method.

Example 2.2.1. Let $\mathbf{I}$ be the light intensity of a food item, say if the observed value is

$\mathbf{I} = 2$, and let $\mathbf{r}$ be the size of a food item that the animal wants to infer with prior knowledge $p(\mathbf{r}) \sim N(\mathbf{r}_p, \Sigma_p^2)$, say the prior mean is $\mathbf{r}_p = 3$ and the prior variance is $\Sigma_p^2 = 1$. If the size, $\mathbf{r}$, is given, as the sensory input is noisy in general, one can treat the light intensity as Gaussian with $p(\mathbf{I}|\mathbf{r}) \sim N(f(\mathbf{r}), \Sigma_\mathbf{I}^2)$,. If we assume all the food have squared shape, then the active function $f(\mathbf{r}) = \mathbf{r}^2$ indicates that on average the light intensity is non-linear related with the size. For simplicity, we can set $\Sigma_\mathbf{I}^2 = 1$. From Baye's theorem, the posterior distribution of the food size is $p(\mathbf{r}|\mathbf{I}) = \frac{p(\mathbf{r})p(\mathbf{I}|\mathbf{r})}{p(\mathbf{I})}$, with $p(\mathbf{I}) = \int p(\mathbf{r})p(\mathbf{I}|\mathbf{r}))d\mathbf{r}$, the normalization term. Below is the comparison from Bayesian MC direct sampling of posterior distribution versus the gradient ascent approach to find optimal food size.



Note that both methods shared some similar results on the estimate of $\mathbf{r}$. The bayesian MC method showed that the most likely value for $\mathbf{r}$ is around 1.57, and the gradient ascent method using Euler updates also converges to $\mathbf{r} = 1.57$.

One remark is that base on the prior knowledge with expectation $\mathbf{r}_p = 3$, it is surprising to see the posterior probability is low at $\mathbf{r} = 3$, this is when $\mathbf{r} = 3$, the expected likelihood, $\mathbb{E}[p(\mathbf{I}|\mathbf{r})]$ takes the active function, and become $f(3) = 3^2 = 9$, which is far different from the observed value $\mathbf{I} = 2$, so $p(\mathbf{I} = 2|\mathbf{r} = 3) \approx 0$

3 Free Energy Principle

3.1 Background

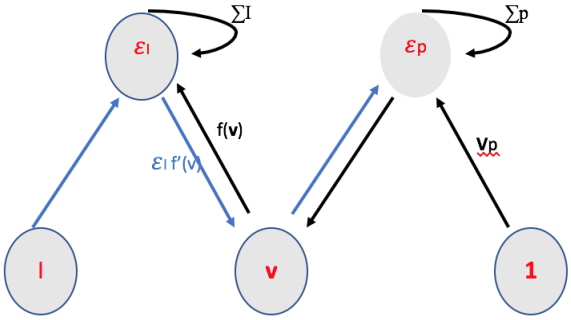
Friston extended the predictive coding model by presenting the sensory uncertainty through learning the variance and covariance of the features parameters via a simple plasticity rule(Friston, 2005). Many of literatures agree with his idea that brain itself is or has an internal model of the environment(Bogacz, 2015). Putting into bayesian framework, one can think the deviation between this internal model and the reality of the external world is the "Surprises". Moreover, the brain does not like surprises, so it aims to minimise the surprise. Then the goal is to obtain posterior distribution, $p(\mathbf{v}|\mathbf{I})$, where $\mathbf{v}$ is another hidden features. If we use q($\mathbf{v}$) to approximate, from KL divergence:

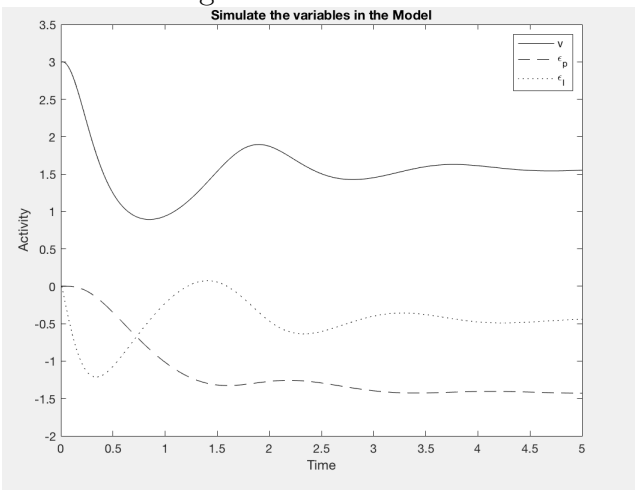
$$
\begin{aligned}
KL[q(\mathbf{v})\|p(\mathbf{v}|\mathbf{I})] &= \int q(\mathbf{v}) log \frac{q(\mathbf{v})}{p(\mathbf{v}|\mathbf{I})} d\mathbf{v} \\
&= \int q(\mathbf{v}) log \frac{q(\mathbf{v})p(\mathbf{I})}{p(\mathbf{v},\mathbf{I})} d\mathbf{v} \\
&= \int q(\mathbf{v}) log \frac{q(\mathbf{v})}{p(\mathbf{v},\mathbf{I})} d\mathbf{v} + \int q(\mathbf{v}) log[p(\mathbf{I})] d\mathbf{v}
\end{aligned}
$$

The first term of the last equation is the free-energy function, denote as $F := \int q(\mathbf{v}) log \frac{q(\mathbf{v})}{p(\mathbf{v},\mathbf{I})} d\mathbf{v}$. There are meanings behind this equation.One might consider the negative of the second term in the last equation, $-\int q(\mathbf{v}) log[p(\mathbf{I})] d\mathbf{v} = -log[p(\mathbf{I})]$ as "Surprises"; then the free-energy function can be expressed as $F = KL[q(\mathbf{v})\|p(\mathbf{v}|\mathbf{I})]$+"Surprise". Since KL divergence is always non-negative, then F is bounded below by the "Surprises". Minimising the "Surprises" is equivalent to minimise F. One can also define "Energy" $= -\int q(\mathbf{v}) log[p(\mathbf{v},\mathbf{I})] d\mathbf{v}$ and "Entropy" $= -\int q(\mathbf{v}) log[q(\mathbf{v})] d\mathbf{v} = \lim_{T\to\infty} \frac{1}{T} \int_0^T -log[q(\mathbf{v})] dt$, so $F =$ "$Energy$"$-$"Entropy". Notice that "Entropy" is essentially the long term average of "Surprises", to get a low "Entropy" meaning to avoid "Surprises", and optimality minimise F.

3.2 Method and Results

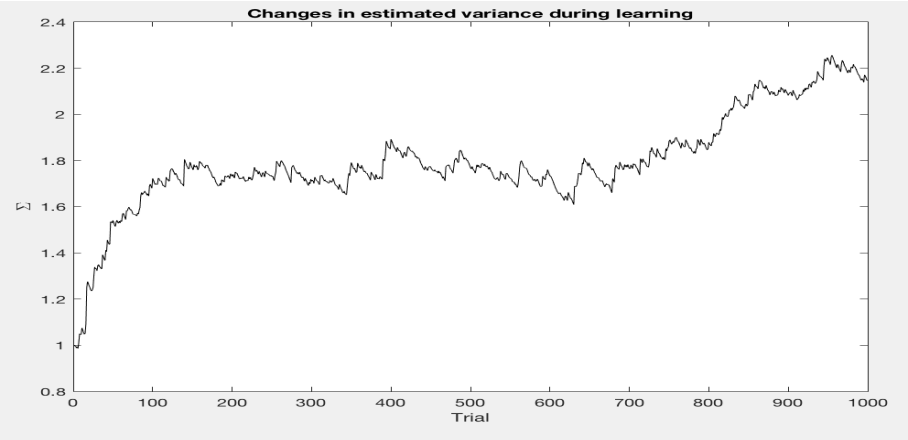
Via free-energy principle, Friston introduced prediction errors as $\epsilon_p = \frac{\mathbf{v}-\mathbf{v}_p}{\Sigma_p}$, to capture how much the food size differs from prior expectation, and $\epsilon_\mathbf{I} = \frac{\mathbf{I}-f(\mathbf{v})}{\Sigma_\mathbf{I}}$, which describing how much the light intensity differs from the expectation if the size of the food is $\mathbf{v}$. Essentially, this is just to standardise the parameters from predictive coding model.





The dynamics(see a diagram on the left) have $\dot{\epsilon_p} = \mathbf{v} - \mathbf{v}_p - \Sigma_p \cdot \epsilon_p$ and $\dot{\epsilon_\mathbf{I}} = \mathbf{I} - f(\mathbf{v}) - \Sigma_\mathbf{I} \cdot \epsilon_\mathbf{I}$. The interpretation is that ϵ_p receives excitatory input(+) from $\mathbf{v}$, inhibitory input(-) from an active neuron via a connection strength $\mathbf{v}_p$, and inhibitory input from itself via a connection strength $\Sigma_\mathbf{p}$. Similar interpretation for $\dot{\epsilon_\mathbf{I}}$. Now applying the gradient ascent method as mentioned before for the new parameters $\epsilon_\mathbf{I}$ and ϵ_p, and comparing with the original parameter, $\mathbf{v}$.

In general, the sensory cortical areas(V1) are hierarchically organised. if assuming the expected neuronal activities in layer v_i depend on the next layer, v_{i+1}, so $\mathbb{E}[\mathbf{v}_i] = f(\mathbf{v}_{i+1})$; then $p(\mathbf{v}_i|\mathbf{v}_{i+1}) \sim N(f(\mathbf{v}_{i+1}), \Sigma_i)$, let $\mathbf{I} = \mathbf{v}_1$. Prediction error on each level converges to $\epsilon_i = \frac{\mathbf{v}_i - f(\mathbf{v}_{i+1})}{\Sigma_i}$, where the variance is $\Sigma_i = \mathbb{E}[(\mathbf{v}_i - f(\mathbf{v}_i))^2]$, and this is the Friston paper interested about. If denote $e_i = \mathbf{v}_i - f(\mathbf{v}_{i+1})$, dynamics of the model becomes: $\dot{\epsilon_i} = \mathbf{v}_i - f(\mathbf{v}_{i+1}) - e_i$, $\dot{e_i} = \Sigma_i \epsilon_i - e_i$, and $\Delta\Sigma_i = \alpha(\epsilon_i e_1 - 1)$ with α as a learning rate. To check the convergence of Σ_i, see below plot with the change of Σ_i in 1000 simulations.



4 Conclusion and Discussion

This course project has broadened my horizons of viewing activities in the brain. Through the study of "bayesian brain hypothese"(Gershman, 2019), I understand how these mathematical models can describe the sensory cortex to "read" our world. However, one may recognise that all the density function here is Gaussian; for the future direction, other reasonable density function need to be discovered, and I am also interested to see other higher dimensional network structure type such as the recurrent network model. Finally, thank you Professor Liu for a wonderful semester and reading my paper this far!

References

[1] R Bogacz. (2015). A tutorial on the free-energy framework for modelling perception and learning *Journal of Mathematical Psychology*. 76(2017) 198-211.

[2] K. Friston. (2009). The free-energy principle: a rough guide to the brain? *Opinion*. 293-300.

[3] S.Gershman. (2019). What does the free energy principle tell us about the brain? *arXiv: 1901.07945vi*. 1-9.

[4] Rao, Rajesh P.N., & Ballard, Dana H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2.* 79-87.