# Summary of *A tutorial on the free-energy framework for modelling perception and learning* by Rafal Bogacz

Marco Casari

University of Turin

Complex system in neuroscience, 12 December 2023



Summary of *A tutorial on the free-energy framework for modelling perception and learning* by Rafal Bogacz

Summary of *A tutorial on the free-energy framework for modelling perception and learning* by Rafal Bogacz

Marco Casari

University of Turin

Complex system in neuroscience, 12 December 2023

- In this presentation the main topics of the paper are presented in order of appearance.
- Source code of proposed exercises is available at https://github.com/mirasac/sistneur/tree/main/code.

## Introduction

- Predictive coding model of Rao and Ballard.



1. Prior predictions are compared to stimuli and the model parameters are updated considering prediction errors, features corresponding to receptive fields in the the primary sensory cortex are learned.

Introduction

- Predictive coding model of Rao and Ballard.

# Introduction

- Predictive coding model of Rao and Ballard.
- Free-energy model of Friston.



1. Prior predictions are compared to stimuli and the model parameters are updated considering prediction errors, features corresponding to receptive fields in the the primary sensory cortex are learned.
2. Weight stimuli by their noise, learn features using their covariance, implement attentional modulation changing the variance of attended features.

Introduction

- Predictive coding model of Rao and Ballard.
- Free-energy model of Friston.

# Introduction

- Predictive coding model of Rao and Ballard.
- Free-energy model of Friston.
- Hebbian plasticity.



1. Prior predictions are compared to stimuli and the model parameters are updated considering prediction errors, features corresponding to receptive fields in the the primary sensory cortex are learned.
2. Weight stimuli by their noise, learn features using their covariance, implement attentional modulation changing the variance of attended features.
3. Synaptic strenght is changed proportionally to activities of pre-synaptic and post-synaptic neurons.

Introduction

- Predictive coding model of Rao and Ballard.
- Free-energy model of Friston.
- Hebbian plasticity.

# Introduction

- Predictive coding model of Rao and Ballard.
- Free-energy model of Friston.
- Hebbian plasticity.
- Free energy minimization.



1. Prior predictions are compared to stimuli and the model parameters are updated considering prediction errors, features corresponding to receptive fields in the the primary sensory cortex are learned.
2. Weight stimuli by their noise, learn features using their covariance, implement attentional modulation changing the variance of attended features.
3. Synaptic strenght is changed proportionally to activities of pre-synaptic and post-synaptic neurons.
4. Minimization of free energy can be seen as the base of many theories of perception.

Introduction

- Predictive coding model of Rao and Ballard.
- Free-energy model of Friston.
- Hebbian plasticity.
- Free energy minimization.

## Working hypotheses

1. Local computation.



1. The state of a neuron is determined only by the synaptic weight and the state of its input neurons.

Working hypotheses

1. Local computation.

## Working hypotheses

1. Local computation.
2. Local plasticity.



1. The state of a neuron is determined only by the synaptic weight and the state of its input neurons.
2. Synaptic plasticity depends only on the activities of pre-synaptic and post-synaptic neurons.



1. Local computation.
2. Local plasticity.

# Working hypotheses

1. Local computation.
2. Local plasticity.
3. Basic neuronal computation.



1. The state of a neuron is determined only by the synaptic weight and the state of its input neurons.
2. Synaptic plasticity depends only on the activities of pre-synaptic and post-synaptic neurons.
3. The state of a neuron is the result of the application of a monotonic function to the linear combination of states and synaptic weights of input neurons.

Working hypotheses

1. Local computation.
2. Local plasticity.
3. Basic neuronal computation.

## Single variable model

- Feature is a scalar variable $v \in \Omega_v$.
- Stimulus is a scalar variable $u \in \Omega_u$.



1. The model describes the inference of a single variable from a single sensory input.

Single variable model

- Feature is a scalar variable $v \in \Omega_v$.
- Stimulus is a scalar variable $u \in \Omega_u$.

# Single variable model

- Feature is a scalar variable $v \in \Omega_v$.
- Stimulus is a scalar variable $u \in \Omega_u$.
- Relation between feature and stimulus is a differentiable function $g : \Omega_v \to \Omega_u$.



Single variable model

- Feature is a scalar variable $v \in \Omega_v$.
- Stimulus is a scalar variable $u \in \Omega_u$.
- Relation between feature and stimulus is a differentiable function $g : \Omega_v \to \Omega_u$.

1. The model describes the inference of a single variable from a single sensory input.
2. In general inferred variable and sensory input are related by some smooth function.

## Single variable model

- Feature is a scalar variable $v \in \Omega_v$.
- Stimulus is a scalar variable $u \in \Omega_u$.
- Relation between feature and stimulus is a differentiable function $g : \Omega_v \to \Omega_u$.
- Sensory input $p(u|v)$ is affected by gaussian noise and it has mean $g(v)$ and variance $\Sigma_u$.



Single variable model

- Feature is a scalar variable $v \in \Omega_v$.
- Stimulus is a scalar variable $u \in \Omega_u$.
- Relation between feature and stimulus is a differentiable function $g : \Omega_v \to \Omega_u$.
- Sensory input $p(u|v)$ is affected by gaussian noise and it has mean $g(v)$ and variance $\Sigma_u$.

1. The model describes the inference of a single variable from a single sensory input.
2. In general inferred variable and sensory input are related by some smooth function.
3. Sensory input and stimulus are drafted from the same space.

## Single variable model

- Feature is a scalar variable $v \in \Omega_v$.
- Stimulus is a scalar variable $u \in \Omega_u$.
- Relation between feature and stimulus is a differentiable function $g : \Omega_v \to \Omega_u$.
- Sensory input $p(u|v)$ is affected by gaussian noise and it has mean $g(v)$ and variance $\Sigma_u$.
- Prior knowledge of the feature $p(v)$ follows a gaussian distribution with mean $v_p$ and variance $\Sigma_p$.



- Feature is a scalar variable $v \in \Omega_v$.
- Stimulus is a scalar variable $u \in \Omega_u$.
- Relation between feature and stimulus is a differentiable function $g : \Omega_v \to \Omega_u$.
- Sensory input $p(u|v)$ is affected by gaussian noise and it has mean $g(v)$ and variance $\Sigma_u$.
- Prior knowledge of the feature $p(v)$ follows a gaussian distribution with mean $v_p$ and variance $\Sigma_p$.

1. The model describes the inference of a single variable from a single sensory input.
2. In general inferred variable and sensory input are related by some smooth function.
3. Sensory input and stimulus are drafted from the same space.
4. Information gained and constantly updated from previous experience.

## Exact solution to the inference problem

- Bayes theorem:

$$p(v|u) = \frac{p(v)p(u|v)}{p(u)} \quad . \tag{1}$$



1. Knowledge of feature depending on a given stimulus is the posterior. Prior knowledge on the feature is the prior, distribution of stimulus is the likelihood.

Exact solution to the inference problem

- Bayes theorem: $p(v|u) = \frac{p(v)p(u|v)}{p(u)}$ . (1)

## Exact solution to the inference problem

- Bayes theorem:

$$p(v|u) = \frac{p(v)p(u|v)}{p(u)} \quad . \tag{1}$$

- Marginal likelihood of stimuli:

$$p(u) = \int_{\Omega_v} p(v)p(u|v)\,\mathrm{d}v \quad . \tag{2}$$



Exact solution to the inference problem

- Bayes theorem:

$$p(v|u) = \frac{p(v)p(u|v)}{p(u)} \quad . \tag{1}$$

- Marginal likelihood of stimuli:

$$p(u) = \int_{\Omega_v} p(v)p(u|v)\,\mathrm{d}v \quad . \tag{2}$$

1. Knowledge of feature depending on a given stimulus is the posterior. Prior knowledge on the feature is the prior, distribution of stimulus is the likelihood.
2. In general, marginal likelihood is difficult to evaluate.

## Exact solution to the inference problem

- Bayes theorem:

$$p(v|u) = \frac{p(v)p(u|v)}{p(u)} \quad . \tag{1}$$

- Marginal likelihood of stimuli:

$$p(u) = \int_{\Omega_v} p(v)p(u|v)\,\mathrm{d}v \quad . \tag{2}$$

- No implementation in simple biological systems.



Exact solution to the inference problem

- Bayes theorem: $p(v|u) = \frac{p(v)p(u|v)}{p(u)}$ . (1)
- Marginal likelihood of stimuli: $p(u) = \int_{\Omega_v} p(v)p(u|v)\,\mathrm{d}v$ . (2)
- No implementation in simple biological systems.

1. Knowledge of feature depending on a given stimulus is the posterior. Prior knowledge on the feature is the prior, distribution of stimulus is the likelihood.
2. In general, marginal likelihood is difficult to evaluate.
3. Complex calculations and infinite nodes are needed to represent each value of the posterior.

## Approximate solution to the inference problem

- Most likely value of the feature is a scalar variable $\phi \in \Omega_v$.



1. Evaluating the mode of the posterior instead of the whole function is more biologically plausible.

Approximate solution to the inference problem

- Most likely value of the feature is a scalar variable $\phi \in \Omega_v$.

## Approximate solution to the inference problem

- Most likely value of the feature is a scalar variable $\phi \in \Omega_v$.
- Equivalent to maximize negative free energy with respect to the feature:

$$F(v,u) = \ln\left(p(v)p(u|v)\right) \quad . \tag{3}$$



1. Evaluating the mode of the posterior instead of the whole function is more biologically plausible.
2. The most likely feature value is the fixed point of the gradient descent method applied to the negative free energy.

Approximate solution to the inference problem

- Most likely value of the feature is a scalar variable $\phi \in \Omega_v$.
- Equivalent to maximize negative free energy with respect to the feature:

$$F(v,u) = \ln\left(p(v)p(u|v)\right) \quad . \tag{3}$$

## Approximate solution to the inference problem

- Most likely value of the feature is a scalar variable $\phi \in \Omega_v$.
- Equivalent to maximize negative free energy with respect to the feature:

$$F(v,u) = \ln\left(p(v)p(u|v)\right) \quad . \tag{3}$$

- Prediction errors:

$$\varepsilon_p = \frac{v - v_p}{\Sigma_p} \quad , \tag{4}$$

$$\varepsilon_u = \frac{u - g(v)}{\Sigma_u} \quad . \tag{5}$$



1. Evaluating the mode of the posterior instead of the whole function is more biologically plausible.
2. The most likely feature value is the fixed point of the gradient descent method applied to the negative free energy.
3. Prediction errors are introduced as new variables to extend the dynamical system and satisfy Hebbian plasticity.

## Neural implementation

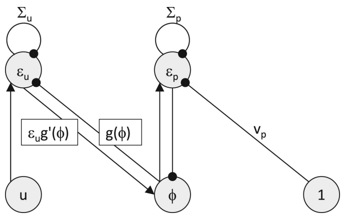



Fig. 3 from article: network implementation of the dynamical system

$$\begin{cases} \dot{\phi} = \varepsilon_u g'(\phi) - \varepsilon_p \\ \dot{\varepsilon}_p = \phi - v_p - \Sigma_p \varepsilon_p \\ \dot{\varepsilon}_u = u - g(\phi) - \Sigma_u \varepsilon_u \end{cases} . \tag{6}$$



- Note that hypotheses are satisfied.

Neural implementation

Fig. 3 from article: network implementation of the dynamical system

## Learning model parameters

- Choose model parameters to maximize $p(u)$.



1. Model parameters are mean and variance of variables.

Learning model parameters

- Choose model parameters to maximize $p(u)$.

## Learning model parameters

- Choose model parameters to maximize $p(u)$.
- Equivalent to maximize negative free energy with respect to parameters:

$$\frac{\partial F}{\partial v_p} = \frac{\phi - v_p}{\Sigma_p} \quad , \tag{7}$$

$$\frac{\partial F}{\partial \Sigma_p} = \frac{1}{2}\left(\frac{(\phi - v_p)^2}{\Sigma_p^2} - \frac{1}{\Sigma_p}\right) \quad , \tag{8}$$

$$\frac{\partial F}{\partial \Sigma_u} = \frac{1}{2}\left(\frac{(u - g(\phi))^2}{\Sigma_u^2} - \frac{1}{\Sigma_u}\right) \quad . \tag{9}$$



1. Model parameters are mean and variance of variables.
2. The fixed point of this dynamical system exists only as sample mean over the occured events of perception, where most likely feature value and stimulus are known.

## Learning model parameters

- Choose model parameters to maximize $p(u)$.
- Equivalent to maximize negative free energy with respect to parameters:

$$\frac{\partial F}{\partial v_p} = \frac{\phi - v_p}{\Sigma_p} \quad , \tag{7}$$

$$\frac{\partial F}{\partial \Sigma_p} = \frac{1}{2}\left(\frac{(\phi - v_p)^2}{\Sigma_p^2} - \frac{1}{\Sigma_p}\right) \quad , \tag{8}$$

$$\frac{\partial F}{\partial \Sigma_u} = \frac{1}{2}\left(\frac{(u - g(\phi))^2}{\Sigma_u^2} - \frac{1}{\Sigma_u}\right) \quad . \tag{9}$$

- Hebbian plasticity is satisfied using prediction errors.



1. Model parameters are mean and variance of variables.
2. The fixed point of this dynamical system exists only as sample mean over the occured events of perception, where most likely feature value and stimulus are known.
3. Without prediction errors, the computation is still local.

Learning model parameters

- Choose model parameters to maximize $p(u)$.
- Equivalent to maximize negative free energy with respect to parameters:

$$\frac{\partial F}{\partial v_p} = \frac{\phi - v_p}{\Sigma_p} \quad , \tag{7}$$

$$\frac{\partial F}{\partial \Sigma_p} = \frac{1}{2}\left(\frac{(\phi - v_p)^2}{\Sigma_p^2} - \frac{1}{\Sigma_p}\right) \quad , \tag{8}$$

$$\frac{\partial F}{\partial \Sigma_u} = \frac{1}{2}\left(\frac{(u - g(\phi))^2}{\Sigma_u^2} - \frac{1}{\Sigma_u}\right) \quad . \tag{9}$$

- Hebbian plasticity is satisfied using prediction errors.

## Learning relation parameter

- Linear relation:

$$g(v, \theta) = \theta v \quad . \tag{10}$$



1. Only one parameter is considered without loss of generality.

Learning relation parameter

- Linear relation:

$$g(v, \theta) = \theta v \quad . \tag{10}$$

## Learning relation parameter

- Linear relation:
$$g(v,\theta) = \theta v \quad . \tag{10}$$
- Nonlinear relation:
$$g(v,\theta) = \theta h(v) \quad . \tag{11}$$



1. Only one parameter is considered without loss of generality.
2. The nonlinearity increases the complexity of the network and partially changes Hebbian plasticity, still keeping it local.

Learning relation parameter

- Linear relation: $g(v,\theta) = \theta v$ . (10)
- Nonlinear relation: $g(v,\theta) = \theta h(v)$ . (11)

## Learning relation parameter

- Linear relation:

$$g(v, \theta) = \theta v \quad . \tag{10}$$

- Nonlinear relation:

$$g(v, \theta) = \theta h(v) \quad . \tag{11}$$

- Gradient of negative free energy for learning:

$$\frac{\partial F}{\partial \theta} = \frac{u - \theta h(\phi)}{\Sigma_u} h(\phi) = \varepsilon_u h(\phi) \quad . \tag{12}$$



Learning relation parameter

- Linear relation: $g(v, \theta) = \theta v$ . (10)
- Nonlinear relation: $g(v, \theta) = \theta h(v)$ . (11)
- Gradient of negative free energy for learning: $\frac{\partial F}{\partial \theta} = \frac{u - \theta h(\phi)}{\Sigma_u} h(\phi) = \varepsilon_u h(\phi)$ . (12)

1. Only one parameter is considered without loss of generality.
2. The nonlinearity increases the complexity of the network and partially changes Hebbian plasticity, still keeping it local.
3. Same consideration of model parameters apply to the relation parameter.

## Free energy framework

- Minimization of Kullback-Leibler divergence:

$$KL(q(v)||p(v|u)) = \int_{\Omega_v} q(v) \ln\left(\frac{q(v)}{p(v|u)}\right) \mathrm{d}v \quad . \tag{13}$$



Free energy framework

- Minimization of Kullback-Leibler divergence:

$$KL(q(v)||p(v|u)) = \int_{\Omega_v} q(v) \ln\left(\frac{q(v)}{p(v|u)}\right) \mathrm{d}v \quad . \tag{13}$$

1. In general, the posterior is approximated by a simpler probability distribution and the divergence between the two is minimized.

## Free energy framework

- Minimization of Kullback-Leibler divergence:

$$KL(q(v)||p(v|u)) = \int_{\Omega_v} q(v) \ln\left(\frac{q(v)}{p(v|u)}\right) \mathrm{d}v \quad . \tag{13}$$

- Definition of negative free energy:

$$F(v,u) = \int_{\Omega_v} q(v) \ln\left(\frac{p(v,u)}{q(v)}\right) \mathrm{d}v \quad . \tag{14}$$



Free energy framework

- Minimization of Kullback-Leibler divergence:

$$KL(q(v)||p(v|u)) = \int_{\Omega_v} q(v) \ln\left(\frac{q(v)}{p(v|u)}\right) \mathrm{d}v \quad . \tag{13}$$

- Definition of negative free energy:

$$F(v,u) = \int_{\Omega_v} q(v) \ln\left(\frac{p(v,u)}{q(v)}\right) \mathrm{d}v \quad . \tag{14}$$

1. In general, the posterior is approximated by a simpler probability distribution and the divergence between the two is minimized.
2. Minimize KL divergence or maximize negative free energy to learn most likely model value, maximize marginal likelihood or maximize negative free energy to learn model parameters.

## Free energy framework

- Minimization of Kullback-Leibler divergence:

$$KL(q(v)||p(v|u)) = \int_{\Omega_v} q(v) \ln\left(\frac{q(v)}{p(v|u)}\right) \mathrm{d}v \quad . \tag{13}$$

- Definition of negative free energy:

$$F(v,u) = \int_{\Omega_v} q(v) \ln\left(\frac{p(v,u)}{q(v)}\right) \mathrm{d}v \quad . \tag{14}$$

- For the models discussed in the paper: $q(v) = \delta(v - \phi)$.



Free energy framework

- Minimization of Kullback-Leibler divergence:

$$KL(q(v)||p(v|u)) = \int_{\Omega_v} q(v) \ln\left(\frac{q(v)}{p(v|u)}\right) \mathrm{d}v \quad . \tag{13}$$

- Definition of negative free energy:

$$F(v,u) = \int_{\Omega_v} q(v) \ln\left(\frac{p(v,u)}{q(v)}\right) \mathrm{d}v \quad . \tag{14}$$

- For the models discussed in the paper: $q(v) = \delta(v - \phi)$.

1. In general, the posterior is approximated by a simpler probability distribution and the divergence between the two is minimized.
2. Minimize KL divergence or maximize negative free energy to learn most likely model value, maximize marginal likelihood or maximize negative free energy to learn model parameters.
3. Equation (3) is recovered using delta function centered in the most likely feature value as probability distribution.

## Multiple variables model

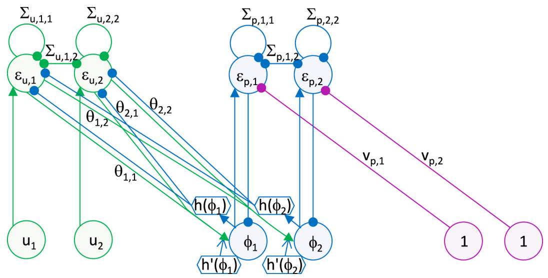
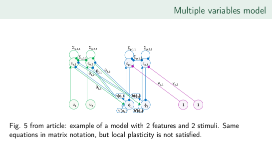



Fig. 5 from article: example of a model with 2 features and 2 stimuli. Same equations in matrix notation, but local plasticity is not satisfied.





- Calculus rules are extended to work elementwise on vectors and matrices, multivariate gaussian distribution and nonlinear relation between variables and stimuli are used.
- The inverse of covariance matrix depends on non-adjacent neurons, Hebbian plasticity is again partially satisfied.

## Hierarchical structure implementation

- Parallel to structure of cortical areas.



1. Information is used and travels in different layers of the cortex.

Hierarchical structure implementation

- Parallel to structure of cortical areas.

## Hierarchical structure implementation

- Parallel to structure of cortical areas.
- Generalized equations for the inference task:

$$\dot{\vec{\phi}}_i = -\vec{\varepsilon}_i + h'(\vec{\phi}_i) \times \mathbf{\Theta}_{i-1}^{\intercal} \vec{\varepsilon}_{i-1} \quad , \tag{15}$$
$$\dot{\vec{\varepsilon}}_i = \vec{\phi}_i - \mathbf{\Theta}_i h(\vec{\phi}_{i+1}) - \mathbf{\Sigma}_i \vec{\varepsilon}_i \quad . \tag{16}$$

- Generalized equations for the learning task:

$$\frac{\partial F}{\partial \mathbf{\Sigma}_i} = \frac{1}{2}(\vec{\varepsilon}_i \vec{\varepsilon}_i^{\intercal} - \mathbf{\Sigma}_i^{-1}) \quad , \tag{17}$$
$$\frac{\partial F}{\partial \mathbf{\Theta}_i} = \vec{\varepsilon}_i h(\vec{\phi}_{i+1})^{\intercal} \quad . \tag{18}$$



1. Information is used and travels in different layers of the cortex.
2. Note the elementwise product and matrices of model and relation parameters.

Hierarchical structure implementation
- Parallel to structure of cortical areas.
- Generalized equations for the inference task:
(15)
(16)
- Generalized equations for the learning task:
(17)
(18)

## Recover local plasticity

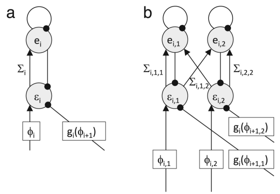
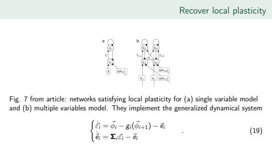



Fig. 7 from article: networks satisfying local plasticity for (a) single variable model and (b) multiple variables model. They implement the generalized dynamical system

$$\begin{cases} \dot{\vec{\varepsilon}}_i = \vec{\phi}_i - g_i(\vec{\phi}_{i+1}) - \vec{e}_i \\ \dot{\vec{e}}_i = \mathbf{\Sigma}_i \vec{\varepsilon}_i - \vec{e}_i \end{cases} . \tag{19}$$





- The update rule for the model parameters is Hebbian and contains the learning rate as hyperparameter of the model.
- Convergence of prediction errors to the sample variances is guaranteed if the most likely feature values change at slower time scales.

# Conclusion

- Stimuli weighted by noise.
- Learn covariance of stimuli.
- Attentional modulation.



Conclusion

- Stimuli weighted by noise.
- Learn covariance of stimuli.
- Attentional modulation.