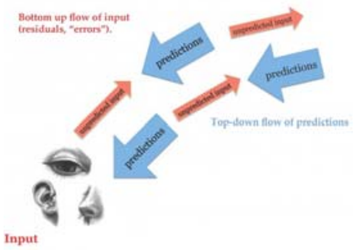


Precision weighting of prediction errors

Predictive Processing Research Seminar

Predictive Processing

Brain as **prediction machine**

- *The brain continuously makes predictions about future sensory evidence based on its current best model of the causes of such evidence*

Bayesian Brain

- *The brain combines prior knowledge with sensory evidence (from various sources) in a Bayesian way*

Hierarchical Brain

- *The brain is organized in a hierarchical way, where "high level" information influences "low level" information and vice versa*

Key sub-processes

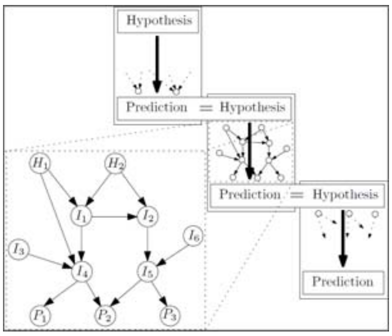
- Making **predictions** of expected input based on the generative models
- Comparing predicted inputs with actual inputs and **computing prediction error**
- **Explaining away** prediction errors (minimizing prediction error)
- **Learning** and adapting generative models

From conceptual idea to formal model

- Predictive processing is assumed to **explain and unify all of cognition**, including higher cognition
- To model, e.g., complex social interactions, Theory of Mind, intention recognition, and problem solving, we need rich enough knowledge structures to model dependences (see discussion of this in Griffiths et al. (2010))
- Our formalization: use discrete **causal Bayesian networks** as knowledge structures to describe predictive processing

Griffiths, Chater, Kemp, Perfors, & Tenenbaum, (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357-364.

Computational model – prediction generation



Hypothesis variables
Hyp = $\{H_1, H_2\}$

Prediction variables
Pred = $\{P_1, P_2, P_3\}$

Intermediate variables
Int = $\{I_1, \ldots I_6\}$

Computational model – error estimation

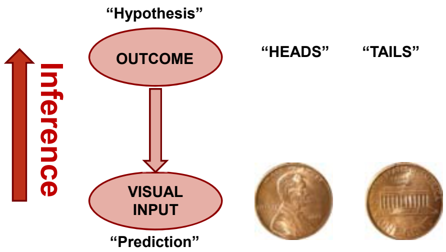
- Prediction and Observation are **probability distributions** over the prediction variables *Pred*
- Prediction is defined as computing the **posterior distribution** $\Pr_{(Pred)}$ given the parameters in the network
- Prediction error is set difference $\Pr_{(obs)}$ - $\Pr_{(Pred)}$
- Estimating the size of this error is defined as computing a KL- divergence or **relative entropy** between predicted distribution and observed distribution

$$D_{\mathrm{KL}}(\Pr_{(\mathrm{Obs})} \| \Pr_{(\mathrm{Pred})}) = \sum_{\mathbf{p} \in \Omega(\mathrm{Pred})} \Pr_{\mathrm{Obs}}(\mathbf{p}) \log \left(\frac{\Pr_{\mathrm{Obs}}(\mathbf{p})}{\Pr_{\mathrm{Pred}}(\mathbf{p})} \right)$$

Precision of prediction & prediction error

- **Precision** is a property of both a **prediction** and of a **prediction error**
- **Precision of a prediction** (expected precision) describes how much uncertainty there is in a prediction (and consequently, how informative the actual observation of what was predicted will be)
- **Precision of a prediction error** describes what proportion of this uncertainty can be attributed to inherent stochastic nature of the process that caused the outcome of the prediction → **precision-weighted prediction errors**

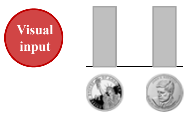
Example: tossing coins

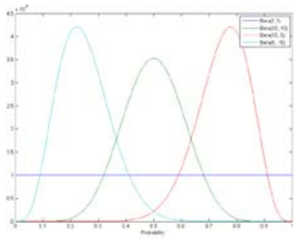


Hyperprior

- Hyperpriors are **priors over priors**
- Define a distribution P(Outcome) over Heads and Tails
- A hyperprior now describes a **distribution over x**, such that P(Outcome = Heads) = x [and P(Outcome = Tails) = 1 – x]
- What does it mean and what does it look like?

Beta distribution





$$P(p;\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

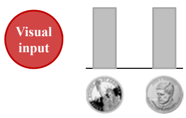
$$\mathrm{B}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

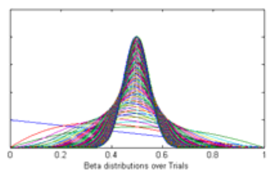
$$\mathrm{P}(X=x) = \frac{\alpha}{\alpha+\beta}$$

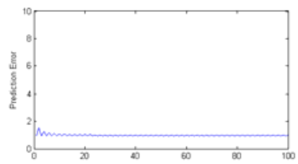
Beta function as hyperprior

- Hyperpriors are **conjugate priors** over the corresponding likelihood function
- Conjugate here means: if the likelihood distribution is of the family X, choosing a conjugate prior ensures that the **posterior distribution** is also of the family X
- In particular, a beta distribution is a conjugate prior over a binomial distribution (in this case: outcomes of coin tosses)
- Dirichlet distributions are conjugate priors over categorical distributions, Gaussians are conjugate priors over themselves, etc.

Bayesian Updating







s = heads, n-s = tails

$$P(p;\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

Precision-weighted prediction error

- Precision-weighted prediction error describes the **size of the effect** of a prediction error on the updating of the model
- The higher this *pwpe*, the **bigger the effect** on the generative model a prediction error is
- The higher this *pwpe*, the **more reducible uncertainty** there is in the environment
- We define this *pwpe* as the **KL divergence** between the hyperprior 'before' and 'after' updating with the new data
- Note that this is a **normative** measure, not a descriptive!

Precision-weighted prediction error (2)

OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Dopamine, Affordance and Active Inference

Karl J. Friston[1*], Tamara Shiner[1], Thomas FitzGerald[1], Joseph M. Galea[2], Rick Adams[1], Harriet Brown[1], Raymond J. Dolan[1], Rosalyn Moran[1], Klaas Enno Stephan[1], Sven Bestmann[2]

1 The Wellcome Trust Centre for Neuroimaging, University College London, Queen Square, London, United Kingdom, 2 Sobell Department of Motor Neuroscience and Movement Disorders, University College London Institute of Neurology, Queen Square, London, United Kingdom

Abstract

The role of dopamine in behaviour and decision-making is often cast in terms of reinforcement learning and optimal decision theory. Here, we present an alternative view that frames the physiology of dopamine in terms of Bayes-optimal behaviour. In this account, dopamine controls the precision or salience of (external or internal) cues that engender action. In other words, dopamine balances bottom-up sensory information and top-down prior beliefs when making hierarchical inferences (predictions) about cues that have affordance. In this paper, we focus on the consequences of changing tonic levels of dopamine firing using simulations of cued sequential movements. Crucially, the predictions driving movements are based upon a hierarchical generative model that infers the context in which movements are made. This means that we can confuse agents by changing the context (order) in which cues are presented. These simulations provide a (Bayes-optimal) model of contextual uncertainty and set switching that can be quantified in terms of behavioural and electrophysiological responses. Furthermore, one can simulate dopaminergic lesions (by changing the precision of prediction errors) to produce pathological behaviours that are reminiscent of those seen in neurological disorders such as Parkinson's disease. We use these simulations to demonstrate how a single functional role for dopamine at the synaptic level can manifest in different ways at the behavioural level.

- This is a **descriptive** measure!

Precision-weighted prediction error (2)

"The Bayesian perspective suggests that there are only two sorts of things that need to be inferred about the world; namely, the **state of the world** and **uncertainty about that state**.

We have suggested that predicted states of the world are encoded in terms of synaptic activity, while uncertainty is encoded by **synaptic gain that encodes the precision** (inverse amplitude or variance) of random fluctuations about predicted states.

If true, this means that modulators of synaptic gain (like **dopamine**) do not report perceptual content but the context in which percepts are formed. In other words, **dopamine reports the precision** or salience of sensorimotor constructs (representations) encoded by the activity of the synapses they modulate."

The many faces of precision

frontiers in PSYCHOLOGY REVIEW ARTICLE

The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science")

- Off-line simulation / planning of actions without execution
- Eliminating (expected) noise at lower levels
- Updating generative models ('learning' – in terms of Clark: "Sculpting effective connectivity")

Reconciling both views

- **Normative view**: the precision weighted prediction error is the net effect of an observation on the generative model
- PWPE = $D_{KL}(f(x;\ \alpha + 1,\ \beta)\ ||\ f(x;\ \alpha,\ \beta))$
- **Descriptive view**: the precision weighted prediction error is a weight (signal gain) on the prediction error
- PWPE = W × $D_{KL}(P_{OBS}\ ||\ P_{PRED})$
- Can we reconcile these two views? YES! We can derive (analytically) W in terms of the hyperparameters (α and β)

Derivation

Prediction: $\Pr(X = x) = \frac{\alpha}{\alpha+\beta}$ and $\Pr(X = \bar{x}) = \frac{\beta}{\alpha+\beta}$

Observation: $\Pr(X = x) = 1$

Prediction error:

$$D_{KL}(\Pr_{Obs}||\Pr_{Pred}) = \sum_{p\in\Omega(Obs)} \Pr_{Obs}(p) \ln\left(\frac{\Pr_{Obs}(p)}{\Pr_{Pred}(p)}\right)$$

$$= 1 \times \ln\left(1/\tfrac{\alpha}{\alpha+\beta}\right) + 0 \times \ln\left(0/\tfrac{\beta}{\alpha+\beta}\right) = \ln\left(\frac{\alpha+\beta}{\alpha}\right)$$

Precision-weighted Prediction error:

$$D_{KL}(f'(x)||f(x)) = \left\langle \ln\left(\frac{f'(x)}{f(x)}\right)\right\rangle_{f'(x)}$$

For $f'(x; \alpha + 1, \beta)$ and $f(x; \alpha, \beta)$

http://bariskurt.com/kullback-leibler-divergence-between-two-dirichlet-and-beta-distributions/

Derivation (2)

$$f(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad \mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$f'(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} x^{\alpha}(1-x)^{\beta-1}$$

$$\mathrm{D_{KL}}(\mathrm{f'(x)} \| \mathrm{f(x)}) = \left\langle \ln\left(\frac{f'(x)}{f(x)}\right) \right\rangle_{f'(x)} = \langle \ln f'(x) - \ln f(x) \rangle_{f'(x)}$$

Derivation (2)

$$\begin{aligned}
&\langle \ln f'(x) - \ln f(x) \rangle_{f'(x)} && = \\
&\langle \ln\Gamma(\alpha+\beta+1) - \ln\Gamma(\alpha+1) - \ln\Gamma(\beta) + (\alpha+\beta)\ln x \\
&\quad - \ln\Gamma(\alpha+\beta) + \ln\Gamma(\alpha) + \ln\Gamma(\beta) - (\alpha+\beta-1)\ln x \rangle_{f'(x)} && = \quad (1) \\
&\ln\Gamma(\alpha+\beta+1) - \ln\Gamma(\alpha+\beta) + \ln\Gamma(\alpha) - \ln\Gamma(\alpha+1) + \langle \ln x \rangle_{f'(x)} && = \quad (2)
\end{aligned}$$

Use that $\langle \ln x \rangle_{f'(x)} = \psi(\alpha+1) - \psi(\alpha+\beta+1)$ and $\Gamma(x) = (x-1)!$

(Result by Bariç Kurt) (by definition)

$$\begin{aligned}
&(\ln(\alpha+\beta)! - \ln(\alpha+\beta-1)!) - (\ln(\alpha)! - \ln(\alpha-1)!) \\
&\quad + \psi(\alpha+1) - \psi(\alpha+\beta+1) && = \quad (3) \\
&\ln(\alpha+\beta) - \ln\alpha + \psi(\alpha+1) - \psi(\alpha+\beta+1) && = \quad (4)
\end{aligned}$$

Derivation (3)

Harmonic number: $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} = \sum_{k=1}^{n} \frac{1}{k}$

Digamma function: $\psi(n) = H_{n-1} - \gamma$

where $\gamma = \lim_{n\to\infty}\left(-\ln n + \sum_{k=1}^{n}\frac{1}{k}\right)$

is the Euler-Mascheroni constant, approximating 0.57721

Important property: $\lim_{x\to\infty} H_x = \ln x$

Derivation (4)

$$\begin{aligned}
&\ln(\alpha+\beta) - \ln\alpha + \psi(\alpha+1) - \psi(\alpha+\beta+1) && = \quad (4) \\
&\ln(\alpha+\beta) - \ln\alpha + (H_\alpha - \gamma) - (H_{\alpha+\beta} - \gamma) && = \quad (5) \\
&\ln(\alpha+\beta) - \ln\alpha + H_\alpha - H_{\alpha+\beta} && \quad\;\; (6)
\end{aligned}$$

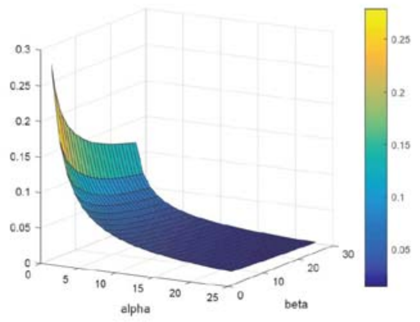
$$W \times \mathrm{D_{KL}}(\mathrm{Pr_{Obs}} \| \mathrm{Pr_{Pred}}) = W \times \ln\left(\frac{\alpha+\beta}{\alpha}\right) = \ln\left(\frac{\alpha+\beta}{\alpha}\right) + H_\alpha - H_{\alpha+\beta}$$

$$W = 1 + \frac{H_\alpha - H_{\alpha+\beta}}{\ln\left(\frac{\alpha+\beta}{\alpha}\right)}$$

Recall: $\lim_{x\to\infty} H_x = \ln x$

So in the long run, when the model becomes stable, W becomes zero!

Weights over time



Conclusion

- Precision-weighted prediction errors for model updating: The more stable the model, the **less impact** a prediction error has on the model
- When models are stable and represent the inherent **stochastic** nature of the environment, the prediction error is 'just' the amount of **irreducible uncertainty** or information in the signal (think of coin toss!)
- Formal definition of `weight' in terms of hyper-parameters of the generative model