A Statistical Physics Approach to the Exploitation-Exploration Dilemma in Human Adaptive Behavior

DEC DÉPARTEMENT D'ÉTUDES COGNITIVES ENS LNC²

Constantin Vaillant-Tenzer [1 3 4] Etienne Koechlin [1 2]

[1]Ecole Normale Supérieure - PSL [2]INSERM [3]Université Paris Cité [3]Sorbonne Université

Introduction

Our aim is to give a principled approach to understand the **adaptability** of human and mammals behavior.

Previous approaches

There were **Utility maximization** (Ferrari-Toniolo et al. 2021); and the elaboration of **the free energy principle**, inducing minimization of surprise (Friston 2009). But the transfer function between exploration parameter (entropy) and exploitation (rewards) is arbitrary. **Descriptive heuristic approaches** (Gigerenzer and Gaissmaier 2011, Newell 2005) are not predictive in general, being individual and situational dependent.

Postulates

Behavior is driven by learning world models but constrained by resources. There is an evolutionary sense to be satisfied to continue learning world models.

The postulates in equations

Maximizing acquired information

$$\arg \max_{A \in \mathrm{Var}((\Omega,\mathcal{T},\mathbb{P}),(\Omega,\mathcal{T}))} \left(\sum_{a \in A(\Omega)} I\left(\theta, \varphi(a)|a\right) \mathbb{P}(A=a) \right)$$

Constraint on resources

$$U - \sum_{a \in A(\Omega)} \varsigma(a)\mathbb{P}(a) \geq 0 \qquad (1)$$

With $\varsigma(a) = E_\theta R_\theta(\varphi(a)) - \omega(a)\mathbb{P}(a) - c_I I\left(\theta, \varphi(a)|a\right)$

Constraint on entropy

Regularization constraint on selection effort :

$$-\sum_{a \in A(\Omega)} \ln(\mathbb{P}(a))\mathbb{P}(a) - S \geq 0 \qquad (2)$$

General form of the solution

The **unique** distribution behavior function is, with $\lambda > 0$ and $\beta > 0$:

$$\forall a \in A(\Omega),\ \mathbb{P}(a) = \frac{1}{z}\exp\left(\beta\left(I(\theta,\varphi(a)|a) - \lambda\varsigma(a)\right)\right)$$

With $z = \sum_{a \in A(\Omega)} \exp\left(\beta\left(I(\theta,\varphi(a)|a) - \lambda_1\varsigma(a)\right)\right)$ is the partition function.

An approximation

Computing the Lagrangian λ and β in computationally complex. We hypotheses that brain does a most accurate computationally simple approximation, assuming β big (or S small) and with the numerical saturation of both constraint.

$$\mathbb{P}(a) \approx \frac{1}{z}\exp\left(S_N \left(\underbrace{\frac{I(a)}{I(b)}}_{\text{Exploration}} - \underbrace{N\varsigma(a)\frac{3U - \varsigma(b)}{(U-\varsigma(b))^2}}_{\textit{Exploitation}} \right)\right) \qquad (3)$$

With $S_N := \left(\ln \frac{N-1}{S} - \ln\ln \frac{N-1}{S}\right)$, $I_m := \frac{1}{N}\sum I(a)$ and b the most probable choice. The approximation on $\lambda \approx N\frac{(I(b))(3U-\varsigma(b))}{(U-\varsigma(b))^2} = O\left(3\frac{I(b)-I_m}{U}\right)$ is in $O\left(e^{-\beta}\right)$ and the one on $\beta \approx \frac{S_N}{I(b)}$ is in $O\left(\frac{\ln\beta}{\beta}\right)$.

Application to bandits

This model can be experimentally applied and tested on bandits : a gambler arrives in a casino and has several slot machines, he/she does not know anything about. What choices would he/she make depending on his resources? Since the optimal Bayesian way of learning is just counting, we can assume that subjects make information inferences using Dirichlet processes (Domenech, Rheims, and Koechlin 2020).

All the parameters of our model can be **explicitly computed** in this situation.

probaDyn10_rep_approx_xp.png

Figure 1. Average on 100 subjects of the predicted probability of action on a 10 armed bandits (fixed probabilities of 0.07, 0.14,...,0.7). The light blue curve corresponds to the best arm and so on. Our model classifies the arms.

probaDyn10_rep_approx_100_xp.png

Figure 2. Zoom on the 200 first trials. We can observe oscillations.

From simulations we were able to make the following predictions:

1. Games are played in order of probability and the frequency of each arm's choice is close to probability matching;
2. Compared to probability matching, subjects over-play low payoff arms and under-play high payoff arms;
3. There is a principle of long-term exploration that persists: subjects continue to select sub-optimal bandits over time;
4. At the beginning, subjects explore the different bandits until then they run out of resources and exploit;
5. There are periodic oscillations that continue over time. The main frequency is an inverse function of the arm number and is independent of the initial amount of

Tiredness

To take into account experimental reality, we can add a physiological cost that linearly increases every trial. The coefficient is randomly shared across simulated subjects through a normal distribution.

probaDyn10_rep_approx_xpsim.png

Figure 4. Simulated average on 100 subjects of the predicted probability of action on a 10 armed bandits (fixed probabilities of 0.07, 0.14,...,0.7), taking into account a physiological cost:

$$\text{trial-number} \times \mathcal{N}(0.00148, 0.00146).$$

The light blue curve corresponds to the best arm and so on. Our model classifies the arms.

Generalization for continuous environments

This model may also represent motor actions or in general continuous action spaces. One wishes to find the probability measure m that maximizes :

$$\int_\Omega I\left(\theta, \varphi(a)|a\right) m(a) d\mu(a)$$

The constraint on a resource takes a continuous form and the constraint on entropy can be written trough Kullback-Leibler divergence.

Next steps

There are still much things to do! This includes **comparing** with the other theories mentioned in introduction ; develop more **general version** of the model (continuous spaces, irregular time frames, risk aversion, etc.) and test them and perform **model based fMRI** to asses the neuroscience pertinence of our model. Deep learning will also be part of our journey, both to be able to use effectively and compute parameters in very general and natural case and also to simulate neural circuits of executive functions within the brain.

References

Domenech, Philippe, Sylvain Rheims, and Etienne Koechlin (2020). "Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex". In: *Science* 369.6507, eabb0184.

Ferrari-Toniolo, Simone et al. (2021). "Nonhuman primates satisfy utility maximization in compliance with the continuity axiom of Expected Utility Theory". In: *Journal of Neuroscience* 41.13, pp. 2964–2979.

Friston, Karl (2009). "The free-energy principle: a rough guide to the brain?" In: *Trends in cognitive sciences* 13.7, pp. 293–301.

Gigerenzer, Gerd and Wolfgang Gaissmaier (2011). "Heuristic decision making". In: *Annual review of psychology* 62, pp. 451–482.

Newell, Ben R (2005). "Re-visions of rationality?" In: