A Biologically Plausible 3-factor Learning Rule for Expectation Maximization in Reinforcement Learning and Decision Making

Mohammad Javad Faraji
School of life sciences, Brain mind institute
School of computer and communication sciences
École Polytechnique Fédéral de Lausanne (EPFL)
Lausanne, CH-1015
mohammadjavad.faraji@epfl.ch

Kerstin Preuschoff
Geneva finance research institute
School of economics and management
University of Geneva
Geneva, CH-1211
kerstin.preuschoff@unige.ch

Wulfram Gerstner
School of life sciences, Brain mind institute
School of computer and communication sciences
École Polytechnique Fédéral de Lausanne (EPFL)
Lausanne, CH-1015
wulfram.gerstner@epfl.ch

Abstract

One of the most frequent problems in both decision making and reinforcement learning (RL) is expectation maximization involving functionals such as reward or utility. Generally, these problems consist of computing the optimal solution of a density function. Instead of trying to find this exact solution, a common approach is to approximate it through a learning process.

In this work we propose a functional gradient rule for the maximization of a general form of density-dependent functionals using a stochastic gradient ascent algorithm. If a neural network is used for parametrization of the desired density function, the proposed learning rule can be viewed as a modulated Hebbian rule. Such a learning rule is biologically plausible, because it consists of both local and global factors corresponding to the coactivity of pre/post-synaptic neurons and the effect of neuromodulation, respectively.

We first apply our technique to standard reward maximization in RL. As expected, this yields the standard policy gradient rule in which parameters of the model are updated proportional to the amount of reward. Next, we use variational free energy as a functional and find that the estimated change in parameters is modulated by a measure of surprise signal. Finally, we propose an information theoretical equivalent of existing models in expected utility maximization, as a standard model of decision making, to incorporate both individual preferences and choice variability. We show that our technique can also be applied into such novel framework.

Keywords: Stochastic gradient ascent, reinforcement learning, decision making, expected utility maximization, multi-factor learning rules, free energy, surprise, neuromodulation.

Acknowledgements

This project is funded by the European Research Council (grant agreement no. 268 689).

1 Introduction

Theoretical descriptions of synaptic plasticity have been dominated by Hebb's rule [1] which is based on two major factors: *locality* and *coactivity*. According to Hebb's rule, both pre- and post-synaptic neurons have to be active to make their connection stronger. Empirical studies, however, show the existence of other *global factors* that can influence synaptic plasticity [2]. These global factors correspond to diffusive action of neuromodulators or feedback from the activity state of a whole population. Deficits in activity of the neuromodulatory system (corresponding to global factors) in humans and animals leaves many tasks un-learnable [3]. For instance, Dopamine (DA) as a neuromodulator is used in signaling reward prediction error that takes part in temporal difference (TD) learning algorithms such as Q-learning and SARSA [4]. Acetylcholine (Ach) is another candidate neuromodulator used in signaling alertness [5]. It is thus of interest to expand on Hebbian learning rules and formulate general new synaptic plasticity rules that combine two Hebbian activity factors with one or multiple global factors. The simplest 3-factor learning rule, including two Hebbian terms modulated by a third factor, is an example.

Expectation maximization, on the other hand, is one of the most frequently encountered problems in both decision making [6] and reinforcement learning (RL) [4]. It consists of computing the optimal solution of a density function. It might represent a learning agent's policy in RL, or the likelihood of selecting different choices in a decision making process. We introduce a functional gradient rule for the maximization of a general form of density-dependent functionals, such as reward or utility, using a stochastic gradient ascent algorithm. We obtain a learning rule by which we approximate the optimal solution through a learning process. This learning rule benefits from a biological plausibility if a neural network is used for parametrization of the desired density function. This is consistent with a modulated Hebbian learning rule (i.e., 3-factor learning rule) in which both global and local factors influence the synaptic connections among the neurons. We apply our technique to standard reward maximization in RL and a variational learning problem to show that reward and surprise signals can be interpreted as third factors in this framework. We further propose a more general formalism of expected utility maximization, a standard model of decision making, that can be solved using functional gradient rule. The aim of such a novel approach is to incorporate both individual preferences and choice variability in the decision making process regardless of the specific details of the model used.

2 Methods

We apply a stochastic gradient ascent technique to approximate the optimal density function that maximizes a functional $\mathbb{F}[P] = \langle \mathcal{F}[P] \rangle_P$ where $\langle . \rangle_P$ denotes the average with respect to the probability density $P(x)$ of the random variable X. The term $\mathcal{F}[P]$ might be considered as a general form of reward or utility function which itself depends on the density function P. The general form of the online gradient rule will be derived in the following.

Theorem 1 (functional gradient rule): The stochastic gradient ascent algorithm for maximizing a functional $\mathbb{F}[P] = \langle \mathcal{F}[P] \rangle_P$ over all possible distributions P parametrized by $\theta \in \mathbb{R}^n$ yields the online learning rule,

$$\Delta\theta \propto \tilde{\mathcal{F}} \nabla_\theta \ln P, \tag{1}$$

where the multiplier factor $\tilde{\mathcal{F}}$ is defined as

$$\tilde{\mathcal{F}} = \frac{\partial}{\partial P} \left(P \mathcal{F}[P] \right) = \mathcal{F}[P] + \frac{\partial \mathcal{F}[P]}{\partial \ln P}. \tag{2}$$

Proof: In order to have an online learning rule for $\theta \in \mathbb{R}^n$, we need to find a term $\Delta\theta$ such that $\langle \Delta\theta \rangle_P = \nabla_\theta \mathbb{F}[P]$. Using the nice trick $P \nabla_\theta \ln P = \nabla_\theta P$ we have

$$\begin{aligned}
\nabla_\theta \langle \mathcal{F}[P] \rangle_P &= \int dx\ P \nabla_\theta \mathcal{F}[P] + \mathcal{F}[P] \nabla_\theta P = \int dx\ P \left(\frac{\partial \mathcal{F}[P]}{\partial P} \nabla_\theta P \right) + \mathcal{F}[P] \left(P \nabla_\theta \ln P \right) \\
&= \left\langle \frac{\partial \mathcal{F}[P]}{\partial P} \nabla_\theta P + \mathcal{F}[P] \nabla_\theta \ln P \right\rangle_P = \left\langle \frac{\partial \mathcal{F}[P]}{\partial P} \left(P \nabla_\theta \ln P \right) + \mathcal{F}[P] \nabla_\theta \ln P \right\rangle_P \\
&= \left\langle \left(\frac{\partial \mathcal{F}[P]}{\partial P} P + \mathcal{F}[P] \right) \nabla_\theta \ln P \right\rangle_P = \left\langle \frac{\partial \left(P \mathcal{F}[P] \right)}{\partial P} \nabla_\theta \ln P \right\rangle_P .
\end{aligned} \tag{3}$$

Corollary 1: The multiplier factor $\tilde{\mathcal{F}}$ in the learning rule (1) can be replaced by $\tilde{\mathcal{F}} + c$ where $c \in \mathbb{R}$ is a constant term because

$$\left\langle \left(\tilde{\mathcal{F}} + c \right) \nabla_\theta \ln P \right\rangle_P = \left\langle \tilde{\mathcal{F}} \nabla_\theta \ln P \right\rangle_P + c \left\langle \nabla_\theta \ln P \right\rangle_P , \tag{4}$$

and $\langle \nabla_\theta \ln P \rangle_P = \int dx\ P \nabla_\theta \ln P = \int dx\ \nabla_\theta P = \nabla_\theta \int dx\ P = \nabla_\theta (1) = 0$.

Corollary 2: If $\mathcal{F}[P]$ does not explicitly depend on P or is linear in $\ln P$, then the multiplier factor $\tilde{\mathcal{F}}$ can be replaced by $\mathcal{F}[P]$. The proof is simply done by using **Corollary 1** in (2).

We want to stress that our proposed learning rule (1) can indeed be embedded in the class of biologically plausible 3-factor learning rules, if a neural network is used for parametrization. The term $\tilde{\mathcal{F}}$ represents a globally modulating third factor, depending on the properties of the neuronal ensemble in a nonlocal fashion. The term $\nabla_\theta \ln P$ represents a Hebbian term, which can be shown to depend on both pre- and postsynpatic activity. As an example, we use a population of spiking neural network for learning the density function P. The neuron model that we use here is a generalized linear model (GLM). This model has the form of a Spike Response Model (SRM) with escape noise [7, 8]. The membrane potential $u_i(t)$ of neuron i at time t is given as $u_i(t) = \sum_j w_{ij}(X_j * \phi)(t) + \eta_i(t)$, where w_{ij} is the synaptic efficacy between pre-synaptic neuron j and post-synaptic neuron i, $X_j(t) = \sum_f \delta(t - t_j^f)$ denotes the presynaptic spike train, $\phi(t)$ is the somatic EPSP, and $\eta_i(t) = -\eta_0 \int_0^t ds\ e^{-\frac{t-s}{\tau_a}}\ X_i(s)$ is the adaptation potential (η_0 and τ_a are constants). The spikes are then generated by a stochastic Poisson process with an exponential escape rate $\rho_i(t)$ [8] conditioned on the membrane potentials,

$$\rho_i(t) = \rho_0 \exp(\frac{u_i(t) - \theta}{\Delta U}), \tag{5}$$

where θ and ΔU are physical constants of the neuron. Free parameters $\theta \in \mathbb{R}^n$ by which P is parametrized are synaptic efficacies w_{ij} between neurons. Each sampled observed data x is modeled as a set of spike trains $\{X_i\}$ generated by all the neurons within a neuronal population. $P(x)$ is then modeled as the likelihood of generating each set of spike trains, corresponding to each sampled observed data x, in that population. The likelihood of a particular spike train $x = \{X_i\}$ which is observed in the interval $[0, T]$ can be written as [9, 10]

$$\ln P(x) = \sum_k \int_0^T dt\ [\ln \rho_k(t) X_k(t) - \rho_k(t)], \tag{6}$$

and its gradient with respect to the particular synaptic weight w_{ij} is calculated as (see [10, 11] for details)

$$\nabla_{w_{ij}} \ln P(x) = \frac{1}{\Delta U}\ (X_j * \phi)(t)\ [X_i(t) - \rho_i(t)]. \tag{7}$$

Therefore, we conclude that the learning rule for synaptic weights w_{ij} according to gradient ascent $\Delta w_{ij} \propto \nabla_{w_{ij}} \ln P(x)$ can be calculated locally and is written as a product of two local (Hebbian) factors: $(X_j * \phi)(t)$ which depends on the pre-synaptic neuron j and $[X_i(t) - \rho_i(t)]$ that depends on the state of the post-synaptic neuron i.

3 Results

In this section we describe two examples of using our proposed functional gradient rule. First, reward maximization in the context of RL can be formulated as finding the optimal policy $\pi(a|s)$ that maximizes the expected reward $\langle R(s,a) \rangle_{\pi(a|s)f(s)}$ where $R(s,a)$ denotes the reward for taking action a in state s and $f(s)$ is the density function of state space. The online learning rule (1) for reward maximization in RL is

$$\Delta\theta \propto R \nabla_\theta \ln \pi, \tag{8}$$

where θ is used to parametrize policy π. Note that since reward $R(s,a)$ does not explicitly depend on the policy π, the multiplier factor $\tilde{\mathcal{F}}$ in (1) is the reward $R := R(s,a)$ itself, according to **Corollary 2**. The learning rule (8) is the standard policy gradient rule used in the reward maximization approach known as R-max [12].

Second, variational methods are typically used in complex statistical models which are defined by a joint distribution $p(v,h)$ over a set of observed (visible) v and unobserved (hidden) h variables. The joint distribution p is known as a generative model governed by some adaptive parameters $\theta \in \mathbb{R}^n$. Two main purposes of using variational methods are to analytically approximate the posterior distribution $p(h|v)$ of hidden variables (for statistical inference over them) or to derive a lower bound for a marginal likelihood $p(v) = \sum_h p(v,h)$ of the visible variables (usually for model selection). A computationally tractable lower bound $\mathcal{L}(q; w, \theta)$ for the marginal likelihood $p(v)$ of the visible variables is calculated as

$$\begin{aligned} \ln p(v) &= \ln \sum_h p(v,h) = \ln \sum_h q(h|v)\ \frac{p(v,h)}{q(h|v)} \\ &\geq \sum_h q(h|v)\ \ln\ \frac{p(v,h)}{q(h|v)} := \mathcal{L}(q; w, \theta), \end{aligned} \tag{9}$$

where we have applied Jensen's inequality. Here $w \in \mathbb{R}^m$ denotes adaptive parameters used for expressing $q(h|v)$. It is easy to see that the difference between the true log likelihood $\ln p(v)$ and its approximated lower bound $\mathcal{L}(q; w, \theta)$ is

$$\ln p(v) - \mathcal{L}(q; w, \theta) = \sum_h q(h|v)\ \frac{q(h|v)}{p(h|v)} := D_{KL}(q||p). \tag{10}$$

Therefore, maximizing the lower bound $\mathcal{L}(q;w,\theta)$ is equivalent to minimizing the Kullback-Leibler divergence $D_{KL}(q||p)$ of the true posterior distribution $p(h|v)$ from the approximated one $q(h|v)$. The lower bound $\mathcal{L}(q;w,\theta)$ is known as (negative) variational free energy $\mathbb{F}[q;v]$ in statistical learning [13] that can be expressed as

$$\mathbb{F}[q;v] = -\mathcal{L}(q;w,\theta) = \langle -\ln p(v,h)\rangle_q - H(q). \tag{11}$$

The variational free energy $\mathbb{F}[q;v]$ for each observed variable v can be considered as a measure of its novelty indicating how much the new observed data v is surprising. Here surprise is taken to be the negative log-likelihood $-\ln p(v)$ of observed data v. One can express variational free energy $\mathbb{F}[q;v]$ as $\langle \mathcal{F}[q;v]\rangle_q$ where $\mathcal{F}[q;v] = -\ln p(v,h) + \ln q(h|v)$, denotes the *instantaneous* amount of free energy for observed data v. The online learning rule, suggested by **Theorem 1**, for variational free energy minimization is then given by

$$\Delta w \propto -\mathcal{F}\nabla_w \ln q, \tag{12}$$

where the minus sign is because of the minimization and $\mathcal{F} = \mathcal{F}[q;v]$ is the instantaneous amount of free energy for observed data v. Note that since $\mathcal{F}[q;v]$ is linear in $\ln q$, the multiplier factor $\tilde{\mathcal{F}}$ in (1) would be equal to $\mathcal{F}[q;v]$ according to **Corollary 2**. The significance of learning rule (12) is that it explicitly shows that the amount of estimated change in parameters w for learning q is proportional to the amount of surprise or information contained in the observed data v. In other words, the surprise signal measured by the instantaneous free energy modulates the learning rate such that surprising observed data v yields more change of the parameter w for learning the approximate posterior distribution q.

4 Discussion

A standard model of decision making is *expected utility maximization* [14] in which a decision maker selects a choice $x^* \in \mathcal{X}$ with the highest *subjective expected utility* $U(x^*)$ among all other alternatives $x \in \mathcal{X}$. In a probabilistic framework, it can be interpreted as selecting choice x^* with probability 1 and choosing the rest with probability 0 (i.e. $P(x) = \delta(x - x^*)$ is the corresponding choice selection density function which determines the likelihood of selecting different choices, where $\delta(.)$ denotes the Kronecker delta function). The density function $P(x) = \delta(x - x^*)$ maximizes the expected value of the utility function $\langle U(x)\rangle_P$ among all possible density functions $P(x)$ because

$$\langle U(x)\rangle_P = \sum_x U(x)P(x) \leq \sum_x U(x^*)P(x) = U(x^*) = \langle U(x)\rangle_{\delta(x-x^*)}. \tag{13}$$

In reinforcement learning, however, choosing the action with the highest value function does not allow for sufficient exploration; this requires choice variability, e.g., by adding noise. Furthermore, individual preferences should be incorporated into the decision making processes, such as action selection in RL.

Expected utility theory accounts for individual differences by explicitly modeling different beliefs about the probabilities of different outcomes. Instead of using a stochastic action selection function (such as a sigmoid) we propose an information theoretical equivalent of existing models to incorporate both individual preferences and choice variability. As such we do not need to impose any specific form of constraints existing in different models. In contrast to maximizing just the average utility $\langle U(x)\rangle_P$, maximizing the functional

$$\begin{aligned}\mathbb{F}[P] &= \langle U(x)\rangle_P + \frac{1}{\lambda_1}H(P) - \frac{1}{\lambda_2}H(P,P_0)\\ &= \left\langle U(x) - \frac{1}{\lambda_1}\ln P(x) + \frac{1}{\lambda_2}\ln P_0(x)\right\rangle_P,\end{aligned} \tag{14}$$

yields a choice selection density function $P(x)$ which not only leads to a relatively high average utility, but also allows exploration. Further, it does not allow the solution to be highly different from a reference P_0. While the entropy $H(P) = \langle -\ln P(x)\rangle_P$ of a density function P in (14) models choice variability, the relative entropy $H(P,P_0) = \langle -\ln P_0(x)\rangle_P$ models subjectivity by involving a subjective reference density function P_0. The minus sign in the last term of the first line in (14) penalizes those density functions that are highly different from a subjective reference P_0. The parameters λ_1 and λ_2 control the fuzziness of the solution by changing the weights of the second and third terms, respectively. By taking the derivative of (14) with respect to P and setting it equal to zero, one could find that the functional (14) is maximized by

$$P^*(x) = \arg\max_P \mathbb{F}[P] = \frac{P_0(x)^{\frac{\lambda_1}{\lambda_2}}\, e^{\lambda_1 U(x)}}{Z(\lambda_1,\lambda_2)}, \tag{15}$$

where $Z(\lambda_1,\lambda_2) = \sum_x P_0(x)^{\frac{\lambda_1}{\lambda_2}}\, e^{\lambda_1 U(x)}$ is the normalizing factor. Equation (15) resembles a (modified) Bayes' rule in the sense that the effect of utility U in making the posterior density function P^* is controlled by a free parameter λ_1 and a prior belief P_0 that is affected by the ratio $\frac{\lambda_1}{\lambda_2}$. Although the optimal density P^* that yields the maximal functional value

$\mathbb{F}[P^*] = \frac{1}{\lambda_1}\ln Z(\lambda_1,\lambda_2)$ is explicitly derived in (15), it can also be learned using the functional gradient rule (1). This is because the functional (14) can be expressed as $\langle \mathcal{F}[P] \rangle_P$ where $\mathcal{F}[P] = U(x) - \frac{1}{\lambda_1}\ln P(x) + \frac{1}{\lambda_2}\ln P_0(x)$ is a density-dependent functional.

If the maximizer P^* is approximated by any other density $\tilde{P}$, then its corresponding functional value $\mathbb{F}[\tilde{P}]$ differs from its maximal value $\mathbb{F}[P^*]$ in proportion to the KL divergence $D_{KL}(\tilde{P}||P^*) \geq 0$. This is because,

$$\begin{aligned}
\mathbb{F}[P^*] - \mathbb{F}[\tilde{P}] &= \frac{1}{\lambda_1}\ln Z(\lambda_1,\lambda_2) - \left\langle U(x) - \frac{1}{\lambda_1}\ln \tilde{P}(x) + \frac{1}{\lambda_2}\ln P_0(x) \right\rangle_{\tilde{P}} \\
&= \frac{1}{\lambda_1}\left\langle \ln Z(\lambda_1,\lambda_2) - \ln e^{\lambda_1 U(x)} + \ln \tilde{P}(x) - \frac{\lambda_1}{\lambda_2}\ln P_0(x) \right\rangle_{\tilde{P}} \\
&= \frac{1}{\lambda_1}\left\langle \ln \frac{Z(\lambda_1,\lambda_2)\tilde{P}(x)}{e^{\lambda_1 U(x)} P_0(x)^{\frac{\lambda_1}{\lambda_2}}} \right\rangle_{\tilde{P}} = \frac{1}{\lambda_1}\left\langle \ln \frac{\tilde{P}(x)}{P^*(x)} \right\rangle_{\tilde{P}} = \frac{1}{\lambda_1} D_{KL}\left(\tilde{P}||P^*\right). \qquad (16)
\end{aligned}$$

As an example, we investigate a binary decision making task (such as the two-armed bandit problem) in which a subject has to make a decision between two alternatives $x = 1$ and $x = 0$. The probability $P(x)$ of making decision x is modeled by a Bernoulli distribution parametrized by θ such that $P(x) = \theta^x(1-\theta)^{(1-x)}$. We use $P_0(x) = 0.5$ to incorporate no *a priori* preference in making different decisions. We further assume that $\lambda_1 = \lambda_2 = \lambda$ to make the formula simpler. As such, the optimal probability of making the decision $x = 1$ in our binary example is equal to

$$P^*(x=1) = \frac{e^{\lambda U(x=1)}}{e^{\lambda U(x=1)} + e^{\lambda U(x=0)}} = \frac{1}{1+e^{-\lambda \Delta U}}, \qquad (17)$$

where $\Delta U = U(x=1) - U(x=0)$ is the difference between the decisions' utilities. If $U(x=1) > U(x=0)$, the probability $P^*(x=1)$ of making decision $x=1$ in (17) is greater than 0.5. The parameter λ then determines how big that probability should be for different values of ΔU. Note that the stochastic (sigmoid) action selection function, which is used in the expected utility theorem for modeling choice variability, is explicitly derived in (17) as the optimal solution in the sense that it maximizes the functional (14).

References

[1] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2002.

[2] John NJ Reynolds and Jeffery R Wickens. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15(4):507–521, 2002.

[3] Michael W Decker and James L McGaugh. The role of interactions between the cholinergic system and other neuromodulatory systems in learing and memory. *Synapse*, 7(2):151–168, 1991.

[4] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. Cambridge Univ Press, 1998.

[5] Michael I Posner and Jin Fan. Attention as an organ system. *Topics in integrative neuroscience: From cells to cognition*, pages 31–61, 2004.

[6] Irving L Janis and Leon Mann. *Decision making: A psychological analysis of conflict, choice, and commitment*. Free Press, 1977.

[7] Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

[8] Renaud Jolivet, Alexander Rauch, Hans-Rudolf Lüscher, and Wulfram Gerstner. Predicting spike timing of neocortical pyramidal neurons by simple threshold models. *Journal of computational neuroscience*, 21(1):35–49, 2006.

[9] Jean-Pascal Pfister, Taro Toyoizumi, David Barber, and Wulfram Gerstner. Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning. *Neural computation*, 18(6):1318–1348, 2006.

[10] Danilo J Rezende, Daan Wierstra, and Wulfram Gerstner. Variational learning for recurrent spiking networks. In *NIPS*, pages 136–144, 2011.

[11] Danilo Jimenez Rezende and Wulfram Gerstner. Stochastic variational learning in recurrent spiking networks. *Frontiers in computational neuroscience*, 8, 2014.

[12] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, volume 99, pages 1057–1063, 1999.

[13] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[14] Jack Meyer. Two-moment decision models and expected utility maximization. *The American Economic Review*, pages 421–430, 1987.