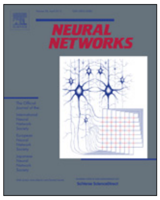
Neural Networks 46 (2013) 62–74

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



# Neuron as a reward-modulated combinatorial switch and a model of learning behavior

Marat M. Rvachev [*]

*108-29 67th drive, Forest Hills, NY 11375, USA*

ARTICLE INFO

*Article history:*
Received 27 February 2012
Revised and accepted 26 April 2013

*Keywords:*
Learning and memory
Neural computation
Synaptic plasticity
Neuron two-layer neural network model
Reinforcement learning

ABSTRACT

This paper proposes a neuronal circuitry layout and synaptic plasticity principles that allow the (pyramidal) neuron to act as a "combinatorial switch". Namely, the neuron learns to be more prone to generate spikes given those combinations of firing input neurons for which a previous spiking of the neuron had been followed by a positive global reward signal. The reward signal may be mediated by certain modulatory hormones or neurotransmitters, e.g., the dopamine. More generally, a trial-and-error learning paradigm is suggested in which a global reward signal triggers long-term enhancement or weakening of a neuron's spiking response to the preceding neuronal input firing pattern. Thus, rewards provide a feedback pathway that informs neurons whether their spiking was beneficial or detrimental for a particular input combination. The neuron's ability to discern specific combinations of firing input neurons is achieved through a random or predetermined spatial distribution of input synapses on dendrites that creates synaptic clusters that represent various permutations of input neurons. The corresponding dendritic segments, or the enclosed individual spines, are capable of being particularly excited, due to local sigmoidal thresholding involving voltage-gated channel conductances, if the segment's excitatory and absence of inhibitory inputs are temporally coincident. Such nonlinear excitation corresponds to a particular firing combination of input neurons, and it is posited that the excitation strength encodes the combinatorial memory and is regulated by long-term plasticity mechanisms. It is also suggested that the spine calcium influx that may result from the spatiotemporal synaptic input coincidence may cause the spine head actin filaments to undergo mechanical (muscle-like) contraction, with the ensuing cytoskeletal deformation transmitted to the axon initial segment where it may modulate the global neuron firing threshold. The tasks of pattern classification and generalization are discussed within the presented framework.

## 1. Introduction

The field of reinforcement learning (RL) solves the problem of sequential decision making by an agent receiving delayed numerical rewards (Sutton & Barto, 1998). The field can be viewed as originating from two major threads: the idea of learning by trial and error that started in the psychology of animal learning (e.g., Thorndike, 1911), and the problem of optimal control and its solution using value functions and dynamic programming (Bellman, 1957). An important branch of the RL theory is the temporal difference (TD) class models for the phasic activity of midbrain dopamine neurons (Montague, Dayan, Person, & Sejnowski, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). The dopamine activity is believed to encode a reward prediction error (RPE) signal that guides learning in the frontal cortex and the basal ganglia (Bush & Mosteller, 1951a, 1951b; Schultz, 1998, 2006). Most scholars active in dopamine studies believe that the dopamine signal adjusts synaptic strengths in a quantitative manner until the subject's estimate of the value of current and future events is accurately encoded in the frontal cortex and basal ganglia (Glimcher, 2011).

This paper considers the problem of instantaneous decision making by an agent receiving immediate rewards within an RL-type framework. A trial-and-error learning paradigm is suggested in which the reward signal modulates memory in (cortical) neurons that act as combinatorial switches. The reward signal may come from an "elementary" reward generator such as that reflecting pain or satisfaction of hunger; it may also involve an RPE-type or "critic"-type (Sutton & Barto, 1998) signal mediated by dopamine and/or other agents that could convey positive as well as negative reward components as was first suggested in Daw, Kakade, and Dayan (2002).

The first contributing thread to the presented model, as in the classical RL theory, is the idea of learning by trial and error and reinforcement of favorable outcomes. The idea, as expressed in Edward Thorndike's "Law of Effect" (Thorndike, 1911), is: "Of several responses made to the same situation those which are accompanied

[*] Tel.: +1 3477294664.
*E-mail address:* rvachev@alum.mit.edu.

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.04.010

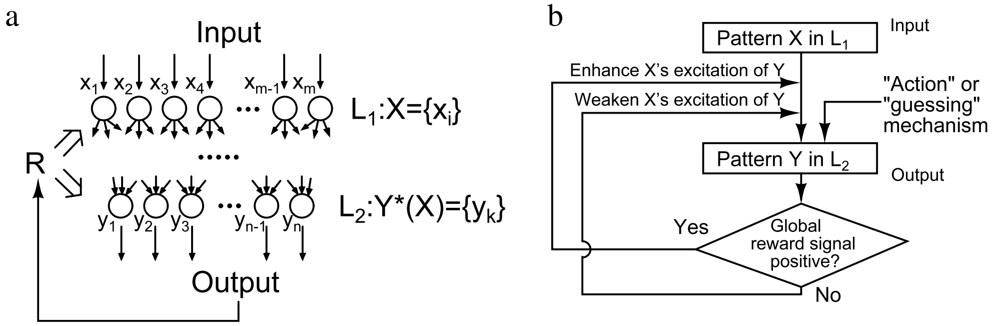


**Fig. 1.** The organism-level learning problem and an outline of the suggested solution. (a) Formulation of the problem. Neurons $x_i$, $i = 1, \dots, m$ in layer $L_1$ connect to neurons $y_k$, $k = 1, \dots, n$ in layer $L_2$. A pattern of excitations $X = \{x_i\}$, if responded to by a pattern of excitations $Y = \{y_k\}$, elicits a positive or negative reward $R$ resulting from the interaction of the generated motor behavior with the environment. The problem is: given an arbitrary $X$, excite $Y^*(X)$ that would lead to positive $R$. (b) Outline of the suggested solution. Learning proceeds by trial and error. Excitation of pattern $X$ excites a pattern $Y(X)$, possibly with the help of an "action" mechanism (e.g., depolarization to all $L_2$ neurons until a certain level of the aggregate $L_2$ output activity is achieved, as discussed in Section 2.3). A "guessing" mechanism introduces variations in the excited patterns $Y$. $X$'s excitation of those $Y$ that lead to positive (negative) $R$ is enhanced (weakened).

or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections to the situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond". This idea is widely regarded as a basic principle underlying much behavior (Campbell, 1960; Cziko, 1995; Dennett, 1981; Hilgard & Bower, 1975).

The second contributing thread is a novel idea that, given proper neuronal circuitry layout, pyramidal neurons can process information by switching the neuron output based on active input neuron combinations. This idea builds on the Two-Layer Neural Network (TLNN) model for the pyramidal neuron (Poirazi, Brannon, & Mel, 2003). Additional computational advantages that could make the idea possible may be provided by mechanical force generated at the dendritic spines and stretch-activation of $\mathrm{Na}^+$ channels at the axon initial segment. An interesting feature of the presented framework is its ability to distil reusable abstract concepts about the environment, making learning with the low-dimensional feedback signal, the reward, efficient.

### 1.1. Problem formulation

The following organism-level learning problem is posed. For simplicity, the neuronal activity states are considered to be binary: "firing" or "not firing". Given an arbitrary combination $X$ of firing neurons in a (perhaps sensory) input layer $L_1$, activate a corresponding "optimal" combination $Y^*(X)$ of firing neurons in a (perhaps motor) output layer $L_2$ (Fig. 1(a)). The optimal combination $Y^*(X)$ is defined as one that produces the motor behavior that results in a positive global reward signal $R$ in the organism. As such, $Y^*(X)$ can be an arbitrary combination of $L_2$ neurons from a combinatorics perspective. The reward signal $R$, in biological terms, may be mediated by certain modulatory neurotransmitters or hormones that are diffusely delivered to generally trainable neurons. It is assumed that in biological systems $R$ can be activated by evolutionarily hardwired circuits, such as when hunger is satisfied, as well as by higher mental processes, e.g., due to the organisms' subjective evaluation of the motor behavior as being satisfactory given the sensory inputs.

It is suggested that the learning process proceeds in a trial-and-error fashion. Given a firing combination $X$ variations are introduced in the firing combination $Y$ with the $X$'s excitation of those $Y$ that lead to positive $R$ being enhanced while $X$'s excitation of those $Y$ that lead to negative $R$ being weakened (Fig. 1(b)). Details of this suggested process are discussed in more detail in Section 5. First, a more elementary learning task is considered: given an arbitrary firing combination $X$ long-term strengthen excitation of an $L_2$ neuron $y_k$, specifically by $X$, if the subsequent reward $R$ is positive. Conversely, long-term weaken excitation of $y_k$, specifically by $X$, if $R$ is negative (Fig. 2).

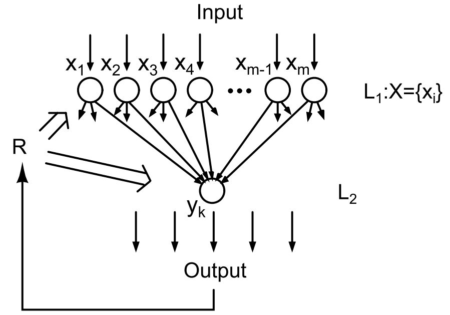


**Fig. 2.** The single-neuron learning problem. $L_1$ neurons $x_i$ connect to an $L_2$ neuron $y_k$. Long-term enhance (weaken) $y_k$ excitation by those combinations $X$ for which the following $y_k$ excitation resulted in a positive (negative) $R$. The enhancement and weakening of excitation may involve long-term potentiation (LTP) and long-term depression (LTD) processes that are influenced by both the combinatorics of the problem and the reward $R$, as suggested in Section 2.2.1.

## 2. Solution to the single-neuron combinatorial switching problem

### 2.1. Local dendritic integration as the basis for combinatorial memory

The following mechanism is posited as the solution and is illustrated in Figs. 3 and 4. $L_1$ neurons connect to the $y_k$ dendrites at random or predetermined locations, forming spatially localized (and possibly overlapping) "synapse neighborhoods" $N_j$ that contain various permutations of input neurons. Sufficient depolarization of the dendritic and/or spine interior within the $j$th neighborhood, caused by the temporal coincidence of the neighborhood's excitatory and absence of inhibitory inputs, causes $N_j$ excitation. The $N_j$ excitation drives local input–output function $F_j$ that has a "combinatorial memory" input–output component $C_j$ that possesses the following properties: (1) $C_j$ expression is long-term enhanced (weakened) if the neighborhood $N_j$ is excited, this is closely followed by a back-propagating action potential (BPAP) at $N_j$, and the immediately following $R$ is positive (negative), and (2) compared to other drivers of neuron stimulation, $C_j$ can substantially contribute to the $y_k$ excitation. Note that the input–output