EUROPHYSICS LETTERS 15 December 2004

Europhys. Lett., **68** (6), pp. 874–880 (2004)

DOI: 10.1209/epl/i2004-10297-y

Artificial ion channels and spike computation in modulation-doped semiconductors

A. Nogaret[1], N. J. Lambert[1], S. J. Bending[1] and J. Austin[2]

[1] *Department of Physics, University of Bath - Bath BA2 7AY, UK*

[2] *Department of Computer Science, University of York - York YO10 5DD, UK*

received 29 June 2004; accepted in final form 25 October 2004
published online 30 November 2004

PACS. 73.63.-b – Electronic transport in nanoscale materials and structures.
PACS. 87.18.Sn – Neural networks.

Abstract. – Semiconductor pn microwires are shown to replicate the physical characteristics of biological nerve fibres that condition the propagation and summation of electrical spikes. The spatio-temporal response of the nanowire to sequences of synaptic impulses is modelled using finite elements. We explain how networks of pn wires could be interconnected to construct fully asynchronous neural networks. These ideas are applied to the elementary XOR problem.

Introduction. – Semiconductor pn junctions are used in electronic devices as a means of controlling slowly varying currents [1]. The well-known exponential conductance of the pn junction describes what is essentially a one-dimensional transport property since translational invariance is always assumed in the plane. Modulated semiconductors may however be perceived as richer and subtler physical systems capable of modelling the dynamic behaviour of ionic solutions in the solid state. As already noted by Shockley, a close analogy exists between electrons, holes, donors, acceptors in semiconductors and acids, bases, cations, anions in water solutions [2]. One information processing device that functions by regulating the passage of ions across a semipermeable membrane is the nerve cell of living organisms. The signalling of electrical impulses along its nerve fibres was investigated by Hodgkin and Huxley in the 1950s [3]. They found that electrical signals would propagate and interfere according to a universal diffusion equation parametrised by the conductance of the membrane to each ion specie and the geometry of the fibre. This diffusion equation continuously integrates all synaptic inputs over space and time and as such is responsible for the parallel processing of information in the nerve cell. Electronic sum and threshold neurons have been realised at the integrated circuit level to add weighted synaptic inputs, although in a synchronous manner [4]. On the other hand, the temporal response of a nerve fibre to an excitatory current has successfully been replicated by modelling the membrane conductance [5] or by asynchronous circuits using cryo-cooled diodes [6, 7]. Recent clinical experiments [8, 9] and stochastic models of neural activity [10–14] have shown that traditional semiconductor devices are not ideally suited to computing spikes. A new physical medium is therefore needed for propagating depolarisation waves analogue to signalling in biological nerve fibres.

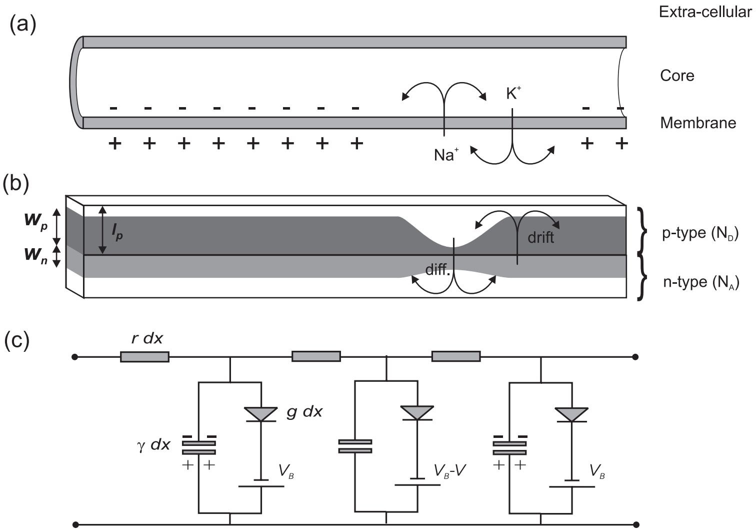


Fig. 1 – (a) Nerve capillary; (b) pn wire; (c) equivalent electrical circuit.

In this paper, we outline the electrical similarity between the semiconductor pn junction and the membrane of biological nerve cells. We show that the propagation and summation of synaptic inputs in the plane of pn wires is determined by a non-linear diffusion equation. A finite-element solver allows us to study the effect of the wire parameters, the amplitude and the duration of synaptic inputs on the speed and diffusion length of spikes. We found that the non-linearity of the diffusion equation manifests through a memory effect of the pn wire to the passage of earlier spikes that has the effect of enhancing spike interferences. We explain how networks of pn wires could be interconnected to perform spatio-temporal neural computations and illustrate this by solving the XOR problem. Our approach paves the way to massively parallel computation on ultra-small scales in simpler, analogue, asynchronous networks and is a step closer to replicating actual neural behaviour. The maturity of the technology on pn junctions together with the advances of modern nanofabrication suggest that artificial networks of spiking neurons could be constructed to study the emergence of synchrony and learning in parallel with experiments on living organisms.

Artificial ion channels. – The examination of nerve fibres and pn wires in fig. 1a) and b) reveals a number of elements that perform a similar electric function. At rest the fibre membrane maintains a gradient of ion concentration that segregates potassium ions inside the membrane and sodium ions outside. Each action potential drives the diffusion of its ion specie across the membrane. At rest, the membrane is more permeable to K^+ than to Na^+. This predominantly results in the accumulation of positive charge outside the fibre until the induced electric field prevents further K^+ outflow. At equilibrium, drift and diffusion currents compensate each other and a potential barrier is established across the membrane. The barrier height (V_m) is given by Nernst equation in table I. Now considering the pn junction in fig. 1b), the chemical imbalance of majority and minority electrons across the depletion region leads to a potential barrier (V_b) that is defined by the same equation —see table I. This barrier is considerably higher in semiconductors because the energy gap is very effective in segregating majority and minority carriers.

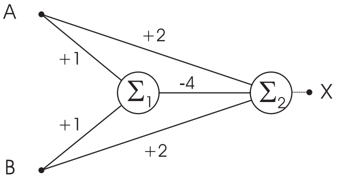
Table I – *The physical properties of nerve membranes and pn junctions.* n_i, N_A, N_D *are the concentrations of intrinsic carriers, acceptor and donor impurities.* $[\mathrm{K}^+]_{\mathrm{in/out}}$, $[\mathrm{Na}^+]_{\mathrm{in/out}}$ *are the potassium and sodium concentrations, respectively, inside and outside the nerve fibre.*

	pn junction	Nerve capillary
Charge carriers	electrons, holes	K^+, Na^+
Membrane	Depletion region 100–1000 nm	Lipid-protein membrane 8–1000 nm
Membrane resting potential	$V_\mathrm{b} = \frac{k_\mathrm{B}T}{e}\ln(\frac{n_\mathrm{i}^2}{N_\mathrm{A}N_\mathrm{D}}) \approx -1.0\,\mathrm{V}$	$V_\mathrm{m} = \frac{k_\mathrm{B}T}{e}\ln(\frac{[\mathrm{K}^+][\mathrm{Na}^+]_{\mathrm{out}}}{[\mathrm{K}^+][\mathrm{Na}^+]_{\mathrm{in}}}) \approx -0.07\,\mathrm{V}$
High-res. channel	p + GaAs electrode, $r \approx 10^{10}\,\Omega/\mathrm{m}$	Intra-cellular, $r \approx 10^{12}$–$10^{14}\,\Omega/\mathrm{m}$
Low-res. channel	n + GaAs electrode	Extra cellular
Capacitance	Depleted region $C \approx 0.1\,\mu\mathrm{F/cm}^2$	Cell capacitance $C \approx 1$–$1000\,\mu\mathrm{F/cm}^2$
Non-linearity	Exponential conductance	Voltage gated Na^+ channels

Electrical signaling along the membrane depends on both the radial conductivity through the membrane and the axial resistivity along the core of the nerve capillary [3]. Because of steric constraints, the resistivity of the capillary is larger than outside. In fig. 1b), the p-type electrode substitutes to the core of the capillary whereas the n-type electrode models the low-resistivity ionic solution outside. In a semiconductor like GaAs, the p-type mobility is 20 times lower than the n-type. To a good approximation, the n-type resistivity may therefore be neglected. The p-type doping level is a useful parameter that will simulate the effect of varying the capillary diameter without the need to adopt a tubular geometry. As shown in table I, p-doped GaAs wire ($N_\mathrm{A} = 1 \times 10^{21}\,\mathrm{m}^{-3}$) will have a lower resistivity per unit length than a capillary of similar cross-section.

The radial conductivity through the nerve membrane exhibits strong non-linearity. This is because Na^+ channels are voltage controlled and only open when the membrane depolarisation crosses a threshold voltage. In fig. 1a) the depolarisation front opens Na^+ channels which triggers a *diffusive* influx of Na^+ ions. The accumulation of positive charge inside the capillary reverses the K^+ *drift*, and produces an outflow of K^+ in an attempt to restore the membrane resting potential. For the pn wire the physical picture looks *a priori* simpler as there is only one exponential conductivity for both electrons and holes which causes the equivalent non-linearity. We argue, however, that in the dynamical regime, Shockley's assumption of charge neutrality may not hold [15]. Instead the appropriate boundary condition requires that the current be zero at the edges of the pn wire. It follows that electrons that have diffused across the depletion region will then spread along the p-type electrode until they are sucked back by the electric field of the depletion region further down the wire. The cross-section of the p-type electrode in fig. 1b) being smaller than the electron diffusion length, few electrons are expected to be lost be recombination. In real pn junctions, the drift current increases as the square root of the depletion width [1], therefore electrons will return to the n-type electrode through a different path shown in fig. 1b). Because of the boundary condition both drift and diffusion current will cancel. To summarize, the opposite dependence of the drift and diffusion current upon the width of the depletion region induces a spatial separation of the diffusion and drift currents similar to that of Na^+ and K^+ currents in the nerve membrane.

The third electrical parameter is the capacitance of the nerve membrane that determines its charging/discharging rate and thus the speed of signals. In our pn wire, this is simply the depletion layer capacitance. The diffusion capacitance [1] is completely negligible in GaAs due to the exceedingly low concentration of minority carriers. Table I compares the capacitance per unit length of a nerve capillary and a pn junction. The difference is explained by the three



XOR

A	B	Σ_1	Σ_2	X
0	0	0	0	0
1	0	1	2	1
0	1	1	2	1
1	1	2	0	0

Fig. 2 – A binary network of two sum and threshold neurons Σ_1 and Σ_2 computing the XOR truth table. The synaptic weights are shown above each connection and the threshold for a neuron to fire is 1.5.

orders of magnitude higher dielectric constant of living tissues and emphasizes the significant advantages to be gained in terms of speed and miniaturisation in working with semiconductors.

Diffusion model. – The equivalent electrical circuit of the pn wire is shown in fig. 1c). $g(V)$ and $\gamma(V)$ are the pn conductance and capacitance per unit length, $r(V)$ is the axial resistance of the p-type electrode per unit length and the resistance of the n-electrode is neglected. Simple circuit analysis applied to a portion dx of the wire gives

$$\frac{\partial V}{\partial t} = \frac{1}{\gamma(V)}\frac{\partial}{\partial x}\left(\frac{1}{r(V)}\frac{\partial V}{\partial x}\right) - \frac{g(V)}{\gamma(V)}V. \tag{1}$$

Equation (1) means that a transient voltage applied via a synaptic input at one point of the wire will diffuse along it. The voltage dependence of r and γ arises from local variations in the p-type depletion layer ($w_{\mathrm{p}}(V)$). In particular, $r(V)$ is close to divergence because the p-doped region (l_{p}) is only marginally thicker than $w_{\mathrm{p}}(V = 0)$. We model $g(V)$ with the conductance of the non-ideal pn junction which is appropriate for GaAs [1]. A small-signal analysis of eq. (1) (for $V < k_{\mathrm{B}}T/e$) is useful to have an analytical estimation of the diffusion parameters. If the synaptic voltage is held constant, the bias will decay exponentially along the wire with diffusion length $\lambda = 1/\sqrt{gr}$. The impedance of the wire, as viewed by a synaptic connection, is $Z_{\mathrm{in}} = 1/(g\lambda) \sim 100\,\mathrm{M}\Omega$, the diffusion speed along the wire is $s = \gamma^{-1}\sqrt{g/r}$.

If two synaptic inputs are applied at different points of the pn wire, the diffusion equation will integrate these two signals and compute the sum at a remote point. The sum will be compared to a voltage threshold so that the signal be eventually regenerated. For our pn wires to have any use in neural computations, they must incorporate synaptic weights, either positive or negative so that voltage spikes interfere constructively or destructively. A simple device that integrates positive and negative weights is the XOR network shown in fig. 2 [16]. If one assumes neurons Σ_1 and Σ_2 to fire above a threshold of 1.5, the input combinations $(A, B) = (0, 1)$, $(1, 0)$ give a weighted sum of 2 at the input of Σ_2 that will then fire. When $(A, B) = (1, 1)$, Σ_1 fires thus inhibiting the two other inputs to Σ_2. Σ_2 does not fire thus fulfilling the XOR truth table.

The XOR function. – The equivalent XOR circuit implemented with pn wires is shown in fig. 3a). The essential difference between the two networks is that in fig. 3a), the spatio-temporal summation is a mathematical property of the diffusion equation instead of being the product of binary computation. The weights may be set in two ways, either by giving each wire a width that allow them to carry different power. This presents the advantage of not perturbing the signal propagation since eq. (1) is independent of the wire width. Weights may also be set via the amplitude of the synaptic voltage. This is a more attractive method with respect to learning as the weights may be changed dynamically. This, however, has a small effect on the signal speed. Systematic modelling of eq. (1) in a straight pn wire showed

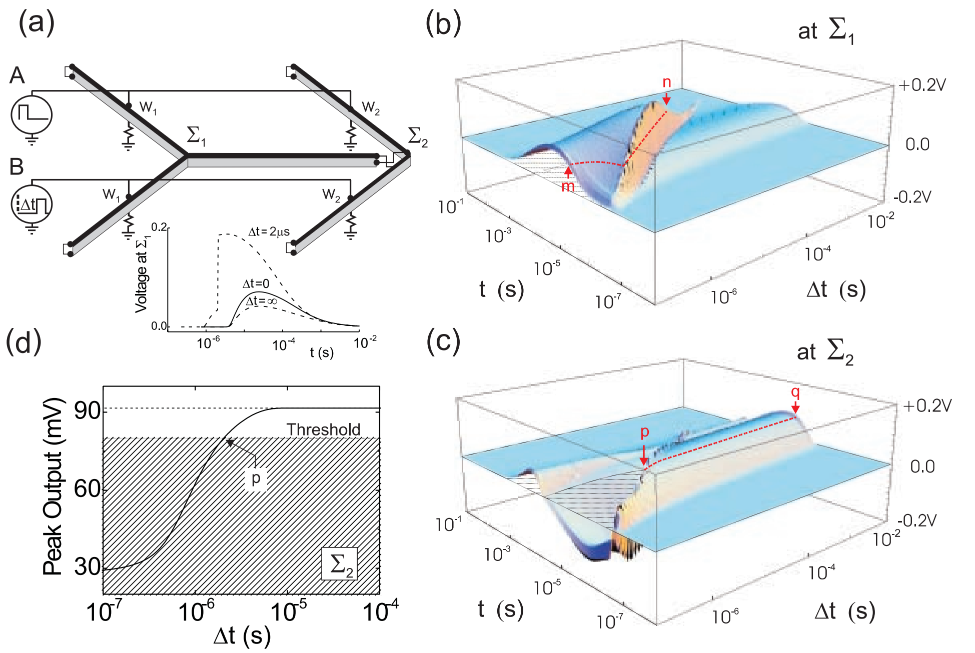


Fig. 3 – (a) Analogue XOR network performing parallel computation with pn wires. Inputs A and B supply 0.1 ns square pulses delayed by Δt. These pulses reach pn wires through four synapses whose weights w_1 and w_2 are set by choosing the appropriate resistances in series with Z_{in}. Signals interfere at $\Sigma_{1,2}$, where they are eventually regenerated by a positive-feedback device grown atop the junction. (b) Summation of spikes A and B at Σ_1. $t = 0$ corresponds to the start of pulse A. The dotted contour (m-n) shows the 80 mV regeneration threshold. (c) Signal waveform at Σ_2 showing signal inhibition. (d) Dependence of the peak voltage in (c) upon Δt. Parameters used in the model are: $N_{\text{A}} = N_{\text{D}} = 10^{21}\,\text{m}^{-3}$, $n_{\text{i}} = 2 \times 10^{12}\,\text{m}^{-3}$, $\mu_{\text{n}} = 0.8\,\text{m}^2\text{V}^{-1}\text{s}^{-1}$, $\mu_{\text{p}} = 0.04\,\text{m}^2\text{V}^{-1}\text{s}^{-1}$, $l_{\text{p}} = 900\,\text{nm}$, $w_{\text{p}}(V = 0) = 850\,\text{nm}$, $\tau = 10^{-8}\,\text{s}$ (GaAs recombination time), $N_{\text{T}} = 5 \times 10^{18}\,\text{m}^{-3}$ (trap density), for practical purposes the wire width is 50–200 nm. (b), (c) in colour on-line.

that the diffusion speed is divided by two when the synaptic voltage drops from +0.8 V to +0.2 V; a consequence of the voltage-dependent coefficients of eq. (1). Negative weights may be realised, as shown in fig. 3a), by connecting together n and p ends of two wires so that positive spikes annihilate at the junction.

The circuit in fig. 3a) was modelled by solving eq. (1) with a finite-element code for each wire section and by matching the appropriate boundary conditions at the wire ends. Open ended wire terminations were set at $V = 0$ to avoid spike reflections. Synaptic inputs were connected at the centre of each 1.0 μm wire section by a Ohmic contact. The weights, w_1 and w_2, were set by the resistances in series with the pn wire. These were chosen so that inputs A and B would give a pulse of amplitude 0.3 V (0.5 V) across the wire at locus w_1 (w_2, respectively). Inputs A and B were assumed to generate 0.1 ns wide square pulses of constant height. Pulse A starts at $t = 0$ whereas pulse B is delayed by time Δt. The signal is regenerated through the positive feedback of a negative differential device fabricated in series with the wire at points Σ_1 and Σ_2. Under an appropriate external bias, this monostable changes state above a 80 mV threshold and a square 0.5 V pulse is applied to the wire for a 0.1 ns duration. Figure 3b) shows the time evolution of the voltage at Σ_1 as the delay between A and B increases. If the two synapses w_1 fire simultaneously ($\Delta t = 0$), the spikes they induce meet at Σ_1 $t = 20\,\mu$s later when the voltage peaks at 70 mV. In contrast, if the synaptic shots are too far apart to interfere ($\Delta t = 10^{-2}$ s), see fig. 3b), the spike due to A or B has a peak voltage of only 41 mV. The difference in spike amplitude when a single synapse or two synapses fire simultaneously, already demonstrates the analogue summation performed by eq. (1). An interesting and rather unexpected situation occurs between points m and n. Here the spike peaks almost immediately after B fires. The interference of the two spikes is maximum at $\Delta t = 20\,\mu$s when it reaches 190 mV. This is well above the 80 mV threshold represented by the dashed line. The wire behaves as if the passage of the first spike facilitates the diffusion of the second spike. The reason for this memory effect lies in the non-linear conductance of the pn junction. As mentioned above, the signal speed varies as the square root of the conductance and is therefore an exponential function of the voltage. Consequently, a residual elevation of the wire bias on the path of the second spike, will have the effect of considerably increasing its speed, hence the observed effect. We note that Σ_1 will only fire when two synapses fire at a relatively close interval of $< 200\,\mu$s which corresponds to segment $m - n$.

Figure 3c) shows the voltage waveform at the input of Σ_2. At $\Delta t = 10^{-7}$ s, the pulses incoming from the w_2 synapses are annihilated by the pulse from Σ_1. Since the peak is less than the 80 mV threshold, Σ_2 does not fire. The short positive spike preceding the negative dip is because the inhibitory action is delayed by the firing of Σ_1. As Δt increases, destructive interference becomes incomplete. At point p, $\Delta t \approx 2\,\mu$s, the peak amplitude crosses the 80 mV threshold and Σ_2 starts firing. If Δt increases further up to point n, $\Delta t \approx 0.5$ ms, Σ_1 stops firing all together and inhibitory action stops. When $\Delta t \to \infty$, the network behaves as if only input A is firing. The resulting spike has a maximum voltage of 92 mV which exceeds the 80 mV threshold (dashed line). Σ_2 will thus fire giving the required XOR output for $(V_A, V_B) = (0.5\,\text{V}, 0)$, $(0, 0.5\,\text{V})$. If both synapses fire at sufficiently short interval, *i.e.* before point p, destructive interference prevent Σ_2 from firing. Figure 3d) plots the voltage peak value at Σ_2 as a function of Δt: when $\Delta t < 2\,\mu$s the peak voltage is below the 80 mV threshold thus the network outputs zero for $(V_A, V_B) = (0.5\,\text{V}, 0.5\,\text{V})$. For A and B firing within this time interval, the network fulfills the XOR truth table.

Discussion. – The principles of parallel analogue computation implemented in the above example may be generalised to more complex networks. Weights will be set by synaptic voltages without these affecting too much spike propagation. We have seen, for example, that it takes $20\,\mu$s for the spike to reach the regeneration point whether the synaptic pulse is 0.3 V or 0.5 V. Among the salient features of pn networks is that the width of the p-type electrode $(l_\text{p} - w_\text{p}(V))$ must not be allowed to cancel, otherwise the axial resistance will diverge and charge will accumulate at an inhibitory junction. The annihilation of two spikes requires that charges spread along reverse biassed junctions, therefore the width of the p-type electrode ought to be finite. Another important aspect is the time constant describing the decay of the voltage spike at $t > 20\,\mu$s in fig. 3. One may establish a parallel between this time constant and the inactivation time of biological membranes [17]. Our decay time depends on the rate at which the spike current sinks at the extremities of the wire and, to a lesser extent, on the electron-hole recombination time. The wire lengths of the XOR example give a decay time of $500\,\mu$s to be compared with a few milliseconds for the inactivation time of biological membranes. If the pn wire is re-excited within the decay time, it will have conserved the memory of previous spikes.

Analogue neural computation otherwise presents a number of advantages. The propagation and summation is an intrinsic property of the diffusion equation rather than the result of a digital operation [4, 5]. Diffusion length and time scales may be precisely engineered by the semiconductor growth conditions, doping and structural parameters. Since there is no stringent material requirement, one might envision using MBE re-growth for integrating dense 3D networks of pn wires. There is also no theoretical limit to miniaturisation, since eq. (1) is independent of the wire width. As compared with biological networks, the use of semiconductors means that the charging rate $(r\gamma)^{-1}$ is considerably faster than in nerve membranes. As compared to other schemes like quantum computation, parallelism is implemented with established technology and at room temperature. Our pn networks, however, present areas of improvement. One is the rather large inactivation time. Another is the computational effort required to calculate weights which will grow with the complexity of the network.

In summary, we have demonstrated that networks of pn wires are capable of performing analogue parallel computation. Artificial networks with no equivalent in nature could be constructed to test spike computation theories over a wider range of conditions than realised in biological networks in order to understand them.

REFERENCES

[1] Sze S. M., *Physics of Semiconductor Devices* (John Wiley, New York) 1981.

[2] Shockley W., *Transistor technology evokes new physics*, Nobel Lecture, December 11, 1956, in *Nobel Lectures, Physics 1942-1962* (Elsevier Publishing, Amsterdam) 1964, p. 344.

[3] Hodgkin A. L. and Huxley A. F., *J. Physiol.*, **117** (1952) 500.

[4] Mead C., *Analog VLSI and Neural Systems* (Addison-Wesley) 1989.

[5] Mahowald M. and Douglas R., *Nature*, **354** (1991) 515.

[6] Coon D. D. and Perera A., *Neural Networks*, **2** (1989) 143.

[7] Coon D. D. and Perera A., *Int. J. Electron.*, **63** (1987) 61.

[8] Mori T. and Shoichi K., *Phys. Rev. Lett.*, **88** (2002) 218101.

[9] Glass L., *Nature*, **410** (2001) 277.

[10] Bohte S. M., Kok J. N. and La Poutre H., *Neurocomputing*, **48** (2002) 17.

[11] Dhamala M., Jirsa V. K. and Ding M., *Phys. Rev. Lett.*, **92** (2004) 074104.

[12] Denker M. *et al.*, *Phys. Rev. Lett.*, **92** (2004) 074103.

[13] Wang X., Zhan M., Choy-Heng L. and Lai Y.-C., *Phys. Rev. Lett.*, **92** (2004) 074102.

[14] Timme M., Wolf F. and Geisel T., *Phys. Rev. Lett.*, **92** (2004) 074101.

[15] The standard screening time ≈ 20 ps assumes the p-type electrode to be larger than the diffusion length which is not realised here.

[16] Minsky M. L. and Papert S. A., *Perceptrons* (MIT Press) 1987.

[17] Kandel E., Schwartz J. and Jessel T., *Principles of Neural Science* (McGraw-Hill) 2000.