23/07/2017

SURYA GANGULI

WEAVING TOGETHER STATISTICAL MECHANICS, MACHINE LEARNING AND NEUROSCIENCE

INTRODUCTION

High dimensional statistics → assumption about simplicity of the data ↔ associated algorithms

Theoretical neuroscience ↔ machine learning successes, for which we know all of the ingredients
↳ could the method we use to understand these (still to be invented) paved the way of understanding the brain.

OUTLINE: 1. Oldies but goodies.
2. High dimensional statistics: theory and experiments.
3. Deep learning: theory and practice.

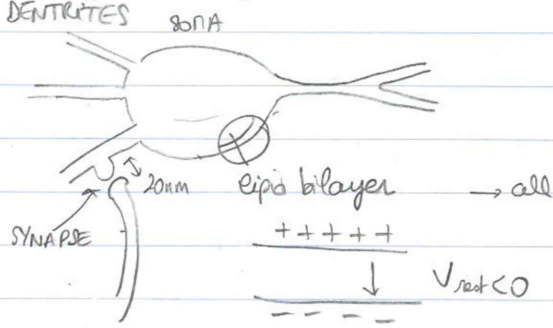
REF: STATISTICAL MECHANICS OF

LECTURE I OLDIES BUT GOODIES

Hodgin Huxley → quantitative description of action potential.

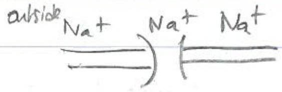
Hopfield → an idea of memory

Hodgin Huxley:



→ all information transported by this voltage.

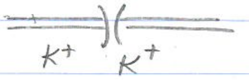
① SODIUM CHANNEL



if for any reason $V_{rest}\nearrow$
the channels open
Na$^+$ rushes in

FAST POSITIVE FEEDBACK

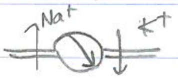
② POTASSIUM CHANNEL



if $V_{rest}\nearrow$, K$^+$ channels open slowly
K$^+$ rushes out

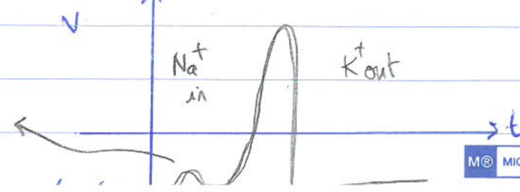
SLOW NEGATIVE FEEBACK.

③ energy cost to being excitable to maintain the non equilibrium steady state



pumping constantly!

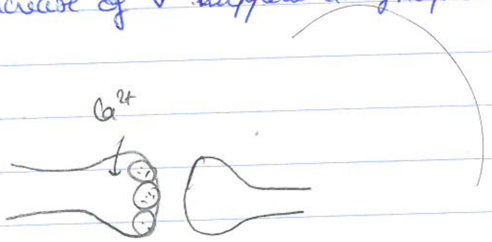
→ we have the minimal ingredients of an excitable system in biology:



increase under threshold of excitation

→ The spike happening somewhere on the membrane will then propagate through a wave along the membrane.

→ Increase of V happens at synapse:



SPIKE FROM PRESYNAPTIC NEURONS
↳ Ca^{2+} rushes in,
↳ vesicules with neurotransmitter open and release
↳ neurotransmitter diffuses
↳ post synaptic neuron open neurotransmitter gated ion channels.

↳ complicated transmission, that actually contains many degrees of freedom that allow a very sensible fine tuning of the spikes → store memories for a lifetime ?.

→ Then to study the brain, we need to coarse grain... REF: Model Neuron: From Hogkin Huxley to Hopfield (Abbot and Kepler 1990).

\+ time scale separation: spikes ~ 2ms.
cognition ~ 100 ms.

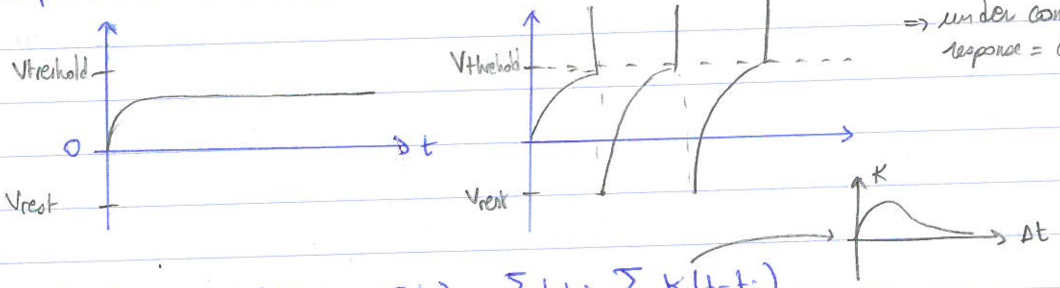
↳ INTEGRATE AND FIRE: "leaky integrator of membrane voltage".

$$\boxed{\tau \frac{dV}{dt} = -V + I(t)}$$

→ if V > Vthreshold emit a spike, then go back to Vrest.

input current I(t) < Vthreshold.

input I(t) > Vthreshold



⇒ under constant input current response = constant rate of spikes.

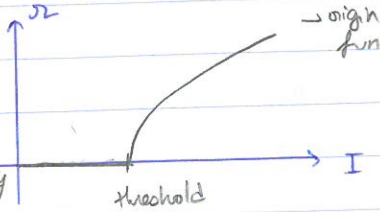
⇒ collective modelling: $I_i(t) = \sum_j W_{ij} \sum_{t_j} K(t-t_j)$

SYNAPTIC STRENGTH   time of spike of j

Rk: Can formulate mean field theory of integrate and fire networks → a bit complicated.

↳ SPIKE RATES: one more step of coarsening.

r = rate = spikes per unit time:



→ origin of non linear activation function of artificial NN

Ex: Compute analytically

$$\boxed{\tau \frac{dr_i}{dt} = -r_i + \sum_j W_{ij} \varphi(r_j^{ext}) + I_i^{ext}}$$

Rk: fastest neuron 200 Hz
spontaneous random neuron 3 Hz

↳ <u>SPIN NEURON</u> $\underline{s_i(n+1) = \mathrm{sgn}\Big(\sum_j J_{ij} s_i(n) + h_i\Big)}$ → zerotemperature Glauber dyn of spin system –

$\overset{\text{time}}{\downarrow}$ (over n)

<u>Rk</u>: Other separation of time scales with the learning, for which we have on much longer time scales plasticity = changes on the J_{ij} / W_{ij}.

HOPEFIELD MODEL

What is a memory? → <u>pattern completion</u>: associate | name with picture –
| stimulus with action –

↳ Store on N neurons, P memories.

memories: ξ_i^μ $\begin{cases} i = 1 \dots N \\ \mu = 1 \dots P \end{cases}$

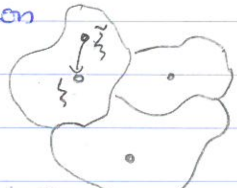
$\mathcal{N}(0,1)$

RANDOM iid MEMORIES

→ we wish that if initializing in $\tilde{\xi}_i^\mu$, system will converge to ξ_i^μ cleaning up the corrupted memory.

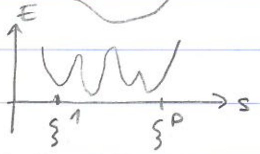
→ image of basins of attraction

→ how to implement that?



↳ Hopefield realized that one needs a dynamics with decreasing energy to rule out the oscillations → ok Glauber

↳ How to have energy minimums $E = -\sum_{ij} J_{ij} s_i s_j$ close to ξ_i^μ ?



→ idea that neurons that fire together, wire together. (Hebb).

$\Delta J_{ij} = \xi_i^\mu \xi_j^\mu$ each time see a pattern

↳ after seeing all the patterns: $\boxed{J_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu}$ → $E = -\frac{1}{N} \sum_{ij} \sum_{\mu} \xi_i^\mu \xi_j^\mu s_i s_j$

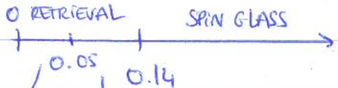
$$= -\frac{1}{N} \sum_{\mu=1}^{P} \Big(\sum_i \xi_i^\mu s_i\Big)^2 = -N \sum_{\mu=1}^{P} (m^\mu)^2$$

with $\boxed{\begin{cases} E = -N \sum_{\mu=1}^{P} (m^\mu)^2 \\ m^\mu = \frac{1}{N} \sum_i \xi_i^\mu s_i \end{cases}}$ magnetisation of μ^{th} memory = overlap

↳ looking good

↳ what about interference between memories? Is $\overset{\text{ORDER PARAMETER}}{m} = (m^1, \dots, m^P) = (\sim 1, 0, 0, 0 \dots)$ stable?

↳ replica calculation :



$\alpha = P/N$ "loading"!

lowest free energy

also superposition of memories

↳ yet not too problematic, given an initialization close to a pattern –

Rk: Correlated patterns? Pseudo inverse rule?

→ Is ξ_i^μ a fixed point of T=0 dynamics? sufficient constraint to be in retrieval phase!

first pattern: $\xi_i^1 = \mathrm{sgn}\left(\frac{1}{N} \sum_j \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \xi_j^1 \right)$

$$= \mathrm{sgn}\left(\underbrace{\frac{1}{N} \sum_j \xi_i^1}_{\xi_i^1} + \underbrace{\frac{1}{N} \sum_{\mu=2}^{P} \sum_j \xi_i^\mu \xi_j^\mu \xi_j^1}_{} \right)$$

interference of competing memories.

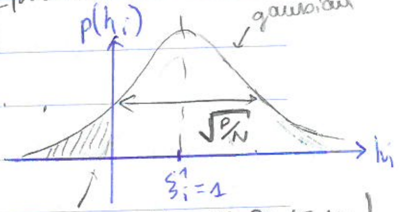
↳ CLT $\sim \mathcal{N}(0, \alpha)$

$\frac{1}{N^2} \times \underbrace{P \times N}_{\text{nb terms}} \times O(1)$

prefactor

prior variance

$\Rightarrow \xi_i^1 = \mathrm{sgn}(\underbrace{\xi_i^1 + \eta_i^1}_{h_i^1})$ with $\eta_i \sim \mathcal{N}(0, \frac{P}{N})$



could be a problem if P/N too big!

/07/2017

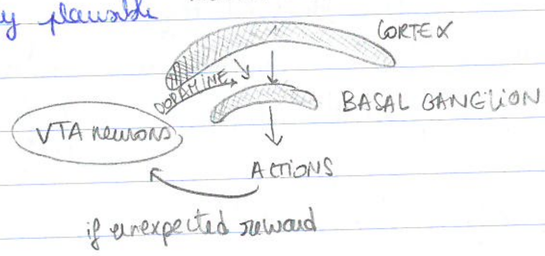
ABOUT LEARNING : (UNSUPERVISED / SUPERVISED)

eg: STATISTICAL STRUCTURE OF NATURAL IMAGES → NETWORK CONNECTIVITY → NETWORK ACTIVITY } what are the rules of these feedback.

↳ q°: fixed points?

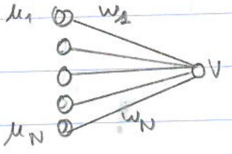
Different learning tasks:
- unsupervised learning
- supervised learning $\vec{x} \to \vec{y} \overset{?}{=} \vec{y}^*$
- reinforcement learning not told $\vec{y}^*$, but tell you whether good/bad direction
  - ↳ more biologically plausible

THE MEANING OF LIFE:



if unexpected reward

UNSUPERVISE LEARNING : What is Hebbian learning doing?



a single neuron $v = \vec{w} \cdot \vec{u}$ linear

Imagine streaming input pattern $\vec{u}^\alpha$ $\alpha = 1..n$

And having weights changing according to hebb rule: $\Delta \vec{w} = \lambda \vec{u}^\alpha v^\alpha$

biological assumption to the behavior of a neuron.

small lambda (slow plasticity)

↳ after n inputs: $\vec{w}(n) = \vec{w}(0) + \sum_{\alpha=1}^{n} \lambda \vec{u}^\alpha v^\alpha \to \boxed{\tau \frac{d\vec{w}}{dt} = \langle \vec{u} v \rangle_{P(\vec{u})}}$

↳ plug-in $v = \vec{w}^T u \to \tau \frac{dw_i}{dt} = \langle \sum_k w_k u_k u_i \rangle = \sum_k w_k \underbrace{\langle u_k u_i \rangle}_{Q_{ki}} \Rightarrow \boxed{\tau \frac{d\vec{w}}{dt} = Q\vec{w}}$

learning driven by second order correlation for linear neurons!
↳ of $P(u)$