Coding and Learning in Networks

Matilde Marcolli

Caltech, Winter 2026
Ma191b: Geometry of Neuroscience

Some references:

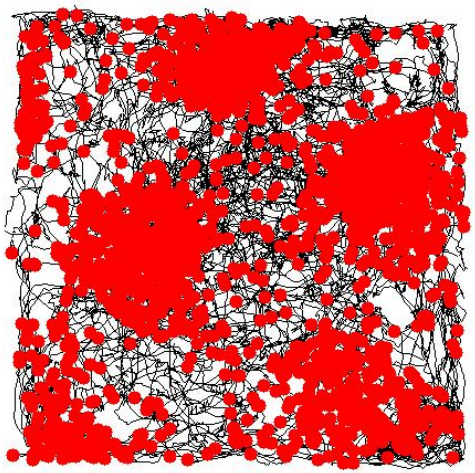
- R.Desimone, T.D.Albright, C.G.Gross, C.Bruce, *Stimulus-selective properties of inferior temporal neurons in the macaque*, J Neurosci. 4 (1984) N.8, 2051–2062.
- J.A.Cardin, M.Carlén, K.Meletis, U.Knoblich, F.Zhang, K.Deisseroth, L.H.Tsai, C.Moore, *Driving fast-spiking cells induces gamma rhythm and controls sensory responses*, Nature, 459 (2009) 7247, 663–667.
- G.G.Gregoriou, S.J.Gotts, H.Zhou, R.Desimone, *Long-range neural coupling through synchronization with attention*, Progress in brain research 176 (2009) 35–45.
- G.G.Gregoriou, S.J.Gotts, H.Zhou, R.Desimone, *High-frequency, long-range coupling between prefrontal and visual cortex during attention*, Science 324 (2009) N. 5931, 1207–1210.
- A.Pasupathy, C.E.Connor, *Shape representation in area V4: position-specific tuning for boundary conformation*, J Neurophysiol. 86 (2001) N.5, 2505–2519.
- Le Chang, Doris Y. Tsao, *The Code for Facial Identity in the Primate Brain*, Cell 169 (2017), 1013–1028

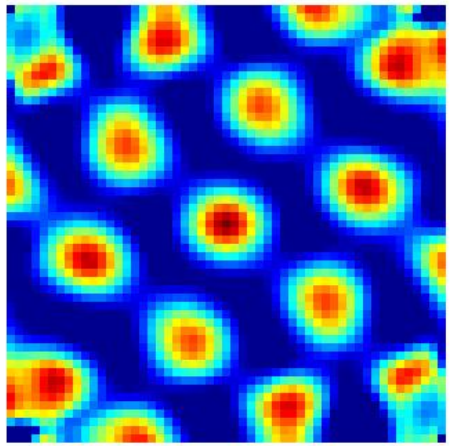
• individual neurons response shows greater variability (for repeated same stimulus) than what expected based on the fact that a neuron integrates large number of synaptic inputs (central limit theorem, expect small fluctuations)

• to explain high variability in response: average synaptic input is sub-threshold for spiking (balancing of excitatory and inhibitory inputs) and activity generated by above-threshold fluctuations: extremely sensitive to small fluctuations (especially from *correlated inputs*)

• some neurons in the inferior temporal (IT) cortex respond selectively to highly specific complex objects (though most IT neurons do not appear to be "detectors" for complex objects): experiments by Desimone et al. (1984)

• Place and grid cells in the rodent hippocampus: moving to new environment, spatial activity patterns of hippocampal place cells remap (grid cells different firing pattern), or only partial "rate remapping" where grid cells unaltered firing patterns





Trajectory of a rat through a square environment (black) and locations of firing of a particular grid cell; spatial autocorrelation of the neuronal activity of the grid cell

• Phase precession of grid and place cell firing: consistent with phase coding of displacement along the current direction of motion

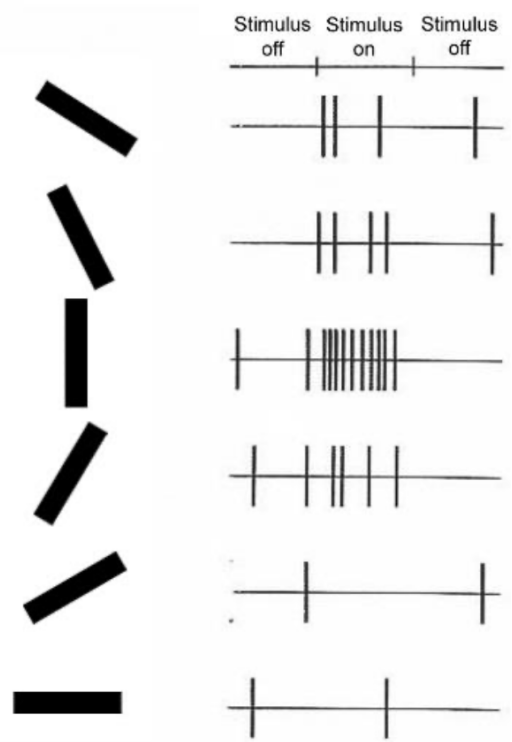
• Experiments: paired-recordings in frontal eye field and area V4 show stimulus in their joint receptive field leads to enhanced oscillatory coupling between the two areas (Gregoriu et al. 2009)

• overlapping RFs (receptive fields) for V4 and FEF (frontal eye field) sites; measured normalized firing rates averaged across the population of cells in FEF and V4

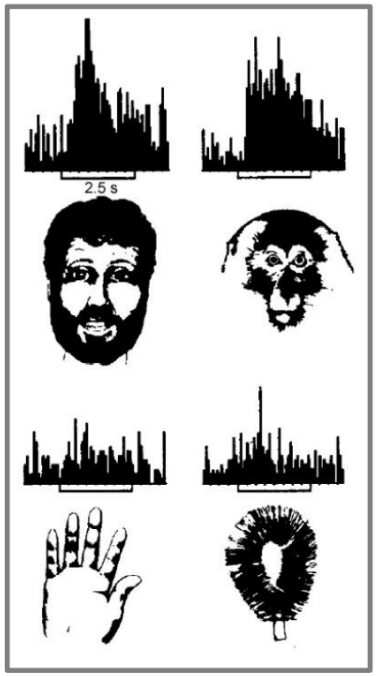
• Oscillatory synchronization of neural assemblies in response to stimuli detected in the olfactory and visual systems of several vertebrates and invertebrates (Stopfer at al. 1997): oscillatory synchronization of neuronal assemblies is essential for fine sensory discrimination

• Cortical gamma oscillations produce neural ensemble synchrony: timing of sensory input relative to a gamma cycle determined amplitude and precision of responses; experiments on responses of a 4 RS cell to whisker stimulus in mice at different temporal phases relative to the induced gamma oscillation (Cardin et al. 2009)

Single neuron firing rate



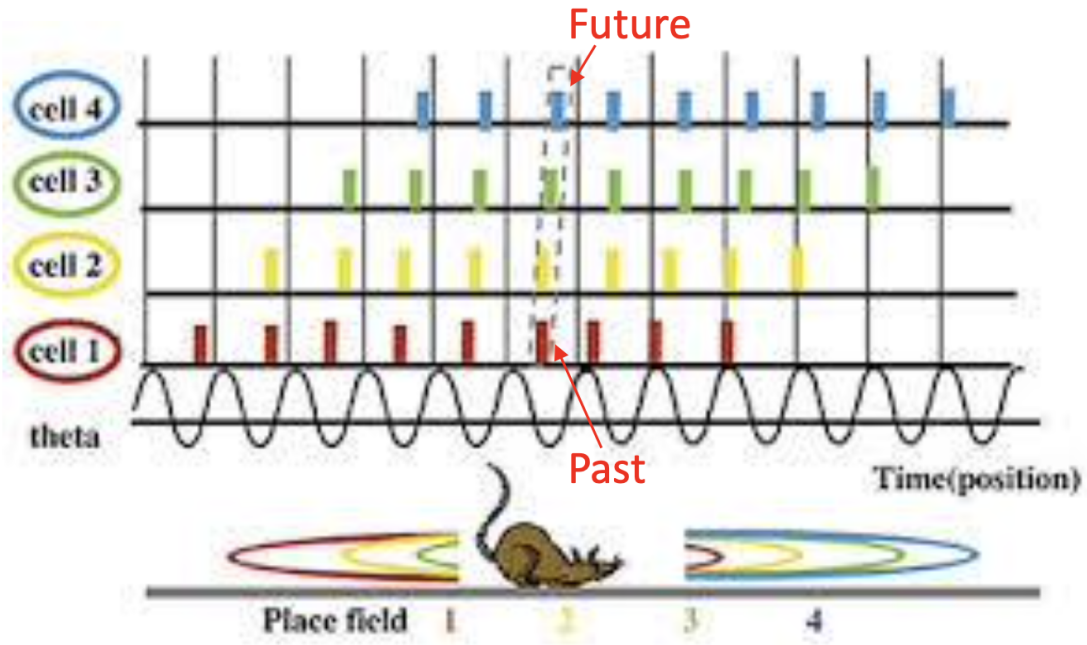
Hubel and Wiesel, J. Physiol., 1959



Desimone et al., J Neurosci., 1984

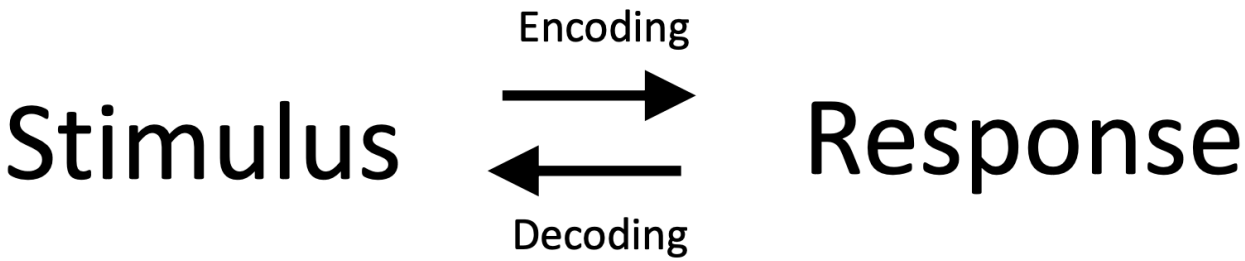
Single neuron spike phase

A place cell fires in one place in a square box



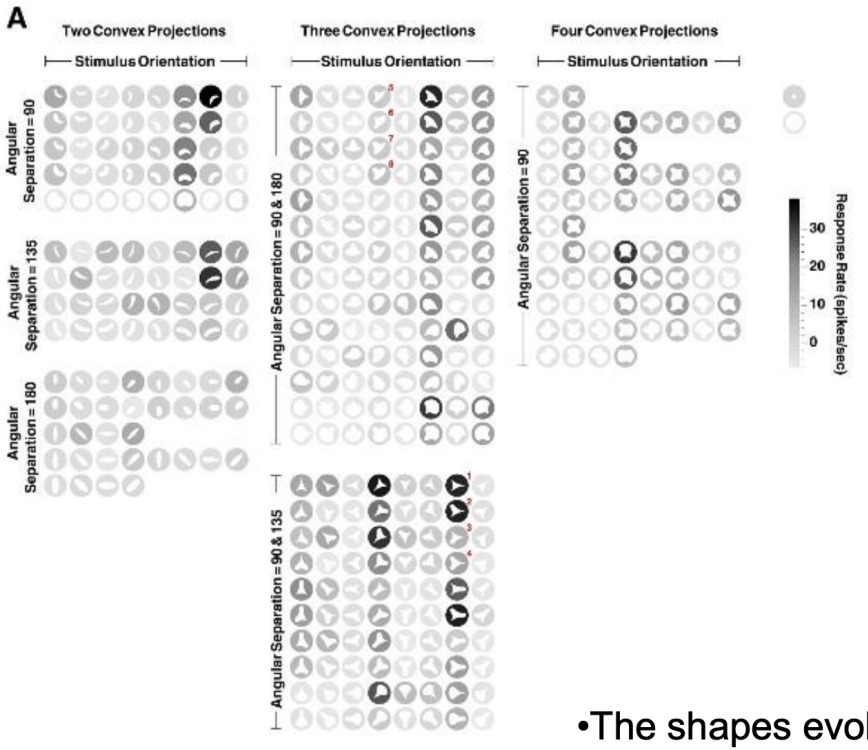
Phase precession

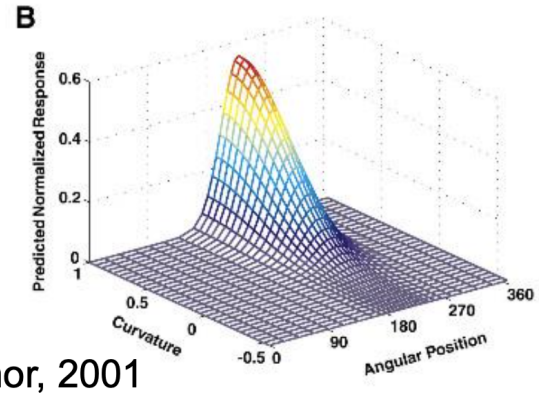
Neural Codes: Decoding/Encoding



- Encoding: model, fit parameters based on responses to a training set
- Decoding: invert the model, or use Bayesian inference to relate P(s|r) to P(r|s)

Example: reconstructing shapes from V4 activity: set of stimuli combining convex/concave boundary elements in closed shapes



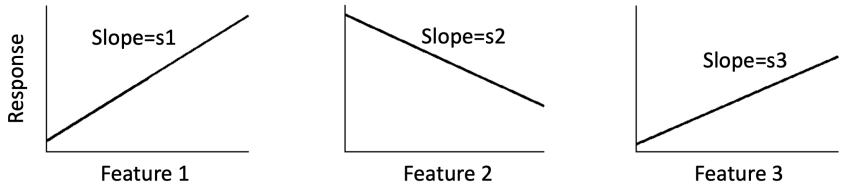


Pasupathy & Connor, 2001

•The shapes evoking strongest responses were characterized by a consistent type of boundary conformation at a specific position within the stimulus.

Example: reconstructing a face from face patch activity

Ramp-shaped tuning implies linear relationship between features and responses



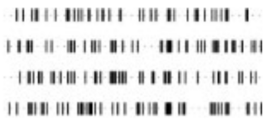
$$Response = s1 \cdot feature1 + s2 \cdot feature2 + \cdots s50 \cdot feature50 + c$$

In short, $\vec{R} = S \cdot \vec{F} + \vec{C}$

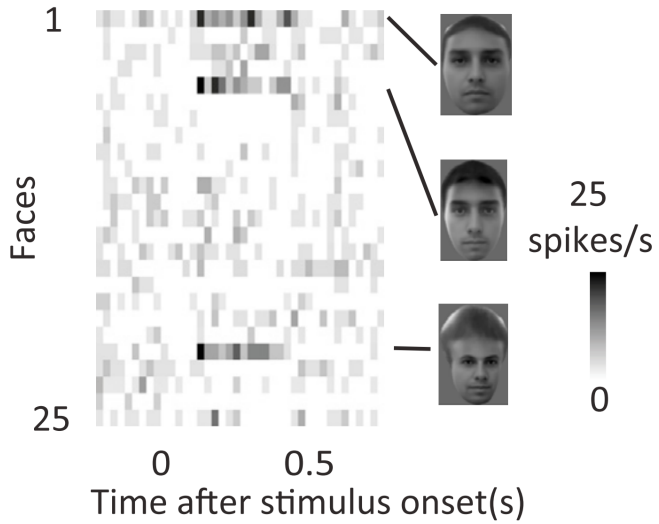
↓ Invert transformation

$\vec{F} = W \cdot \vec{R} + \vec{C}'$

Decoding face identity



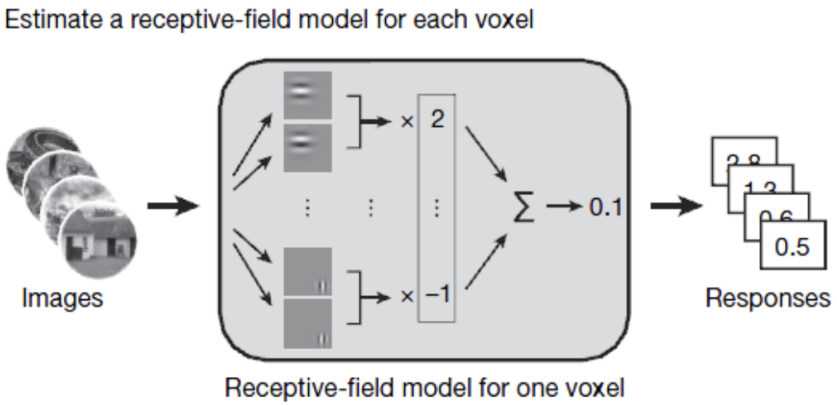
$$\underbrace{\vec{F}}_{\text{50-d Face feature vector}} = W \cdot \underbrace{\vec{R}}_{\text{Cell responses}} + \vec{C}$$



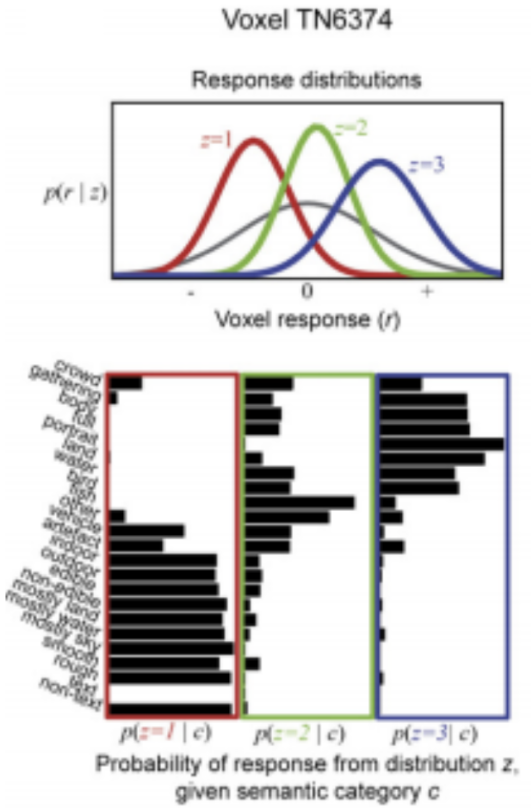
L. Chang, D.Tsao 2017

Example:

Reconstructing natural scenes from fMRI activity



Structural encoding model



Semantic encoding model

Philosophical problem

- "V1 neurons represent orientation"
- "V4 neurons represent curvature"
- "Face neurons represent facial shape and appearance"
- "Olfactory neurons represent smells"
- "Decision neurons represent decisions"

How does brain know what a particular neuron's firing represents?

Modeling of encoding/decoding, representation, and learning in networks; mathematical theory of learning; problem of 'qualia' and conscious experience, how to model?

Artificial networks that learn: a short history and some key ideas

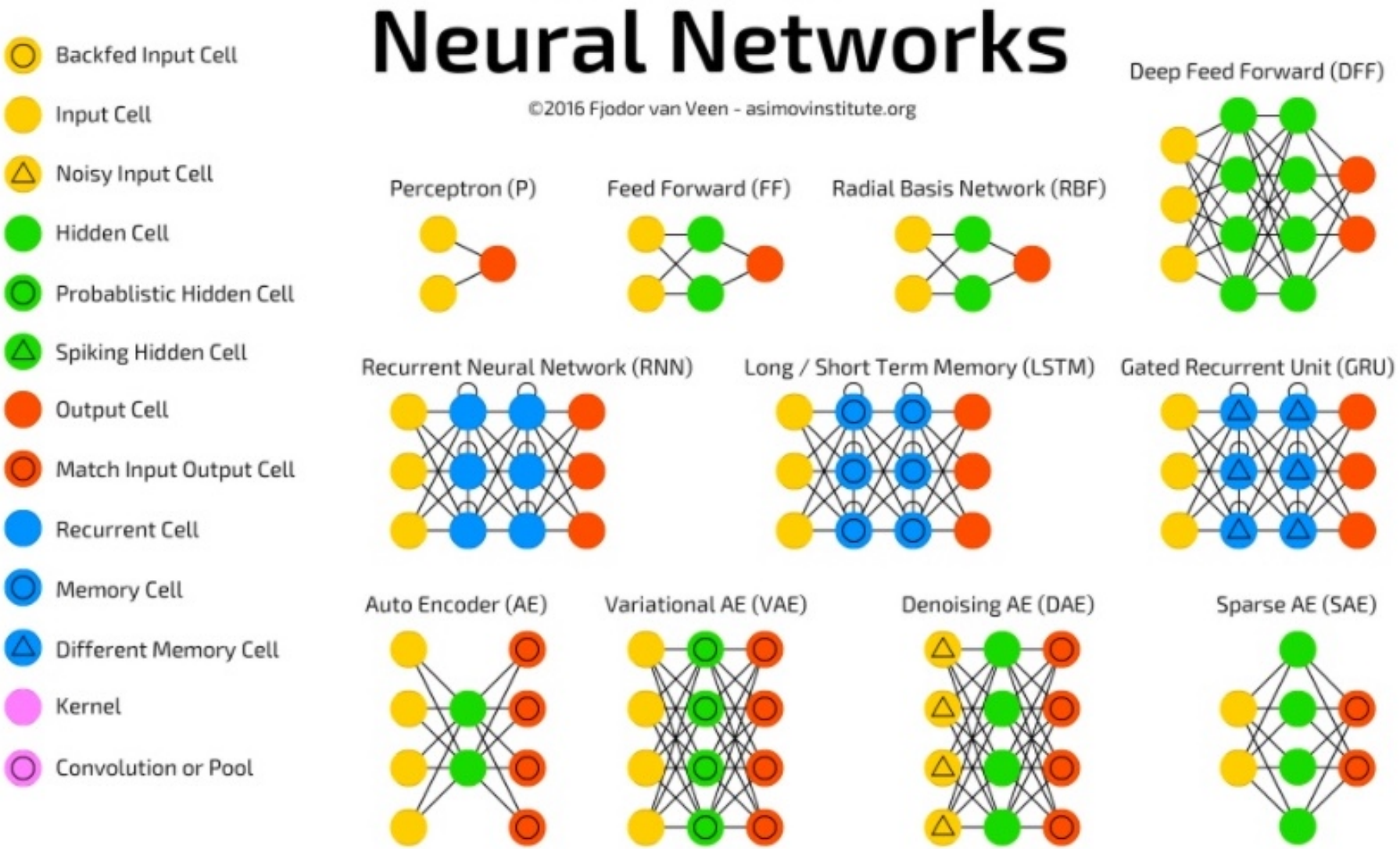


Image: The Asimov Institute

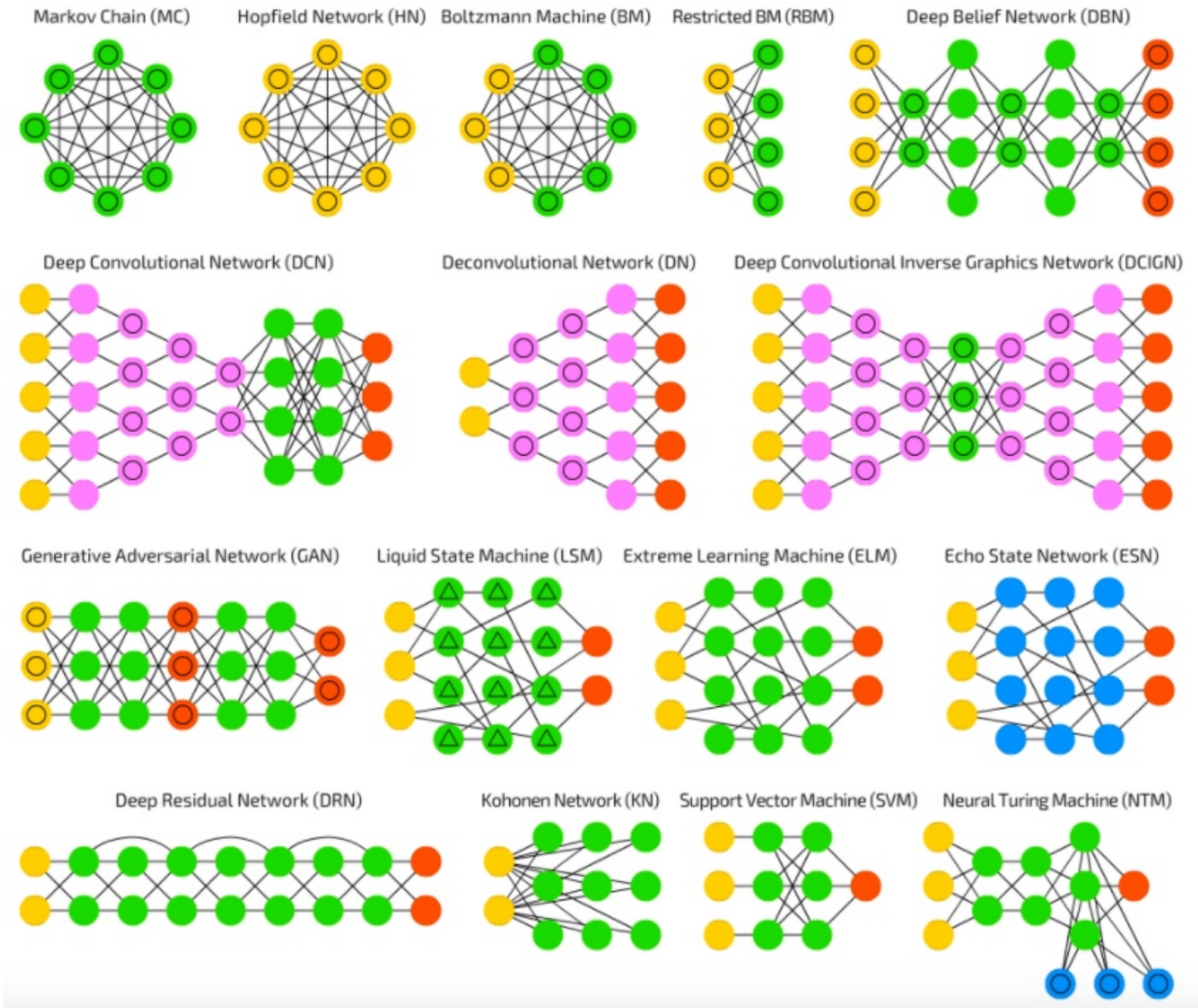


Image: The Asimov Institute

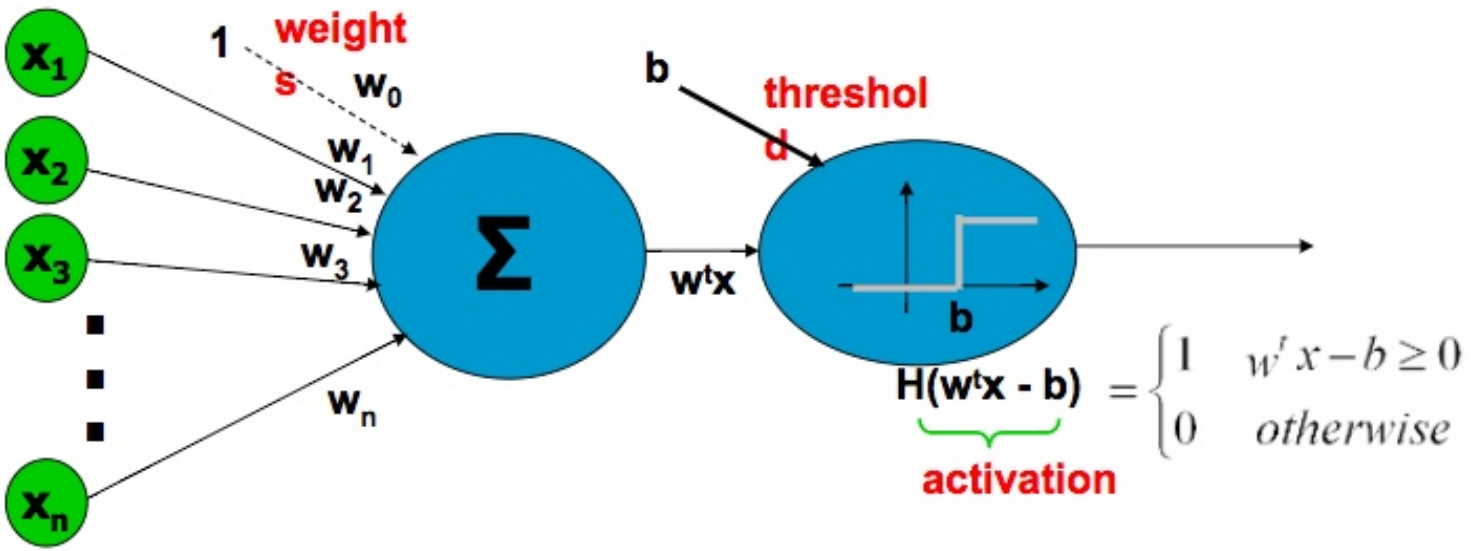
References:

- Ian Goodfellow, Yoshua Bengio, Aaron Courville, *Deep Learning*, MIT Press, 2017.

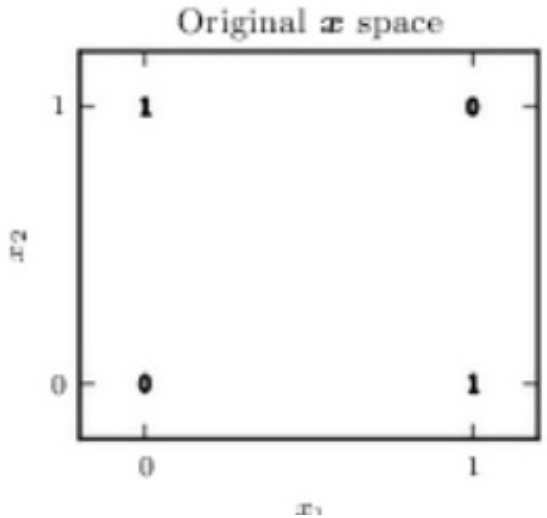
Historical Origins

- Cybernetics: "artificial neurons"

Warren McCulloch & Walter Pitts (1943)



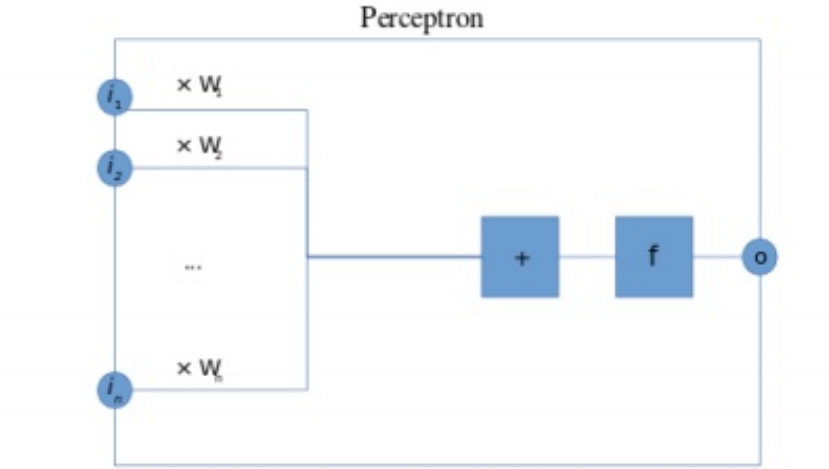
- Marvin Minsky, Seymour Papert, *Perceptron*, MIT Press, 1969
- Perceptron: an algorithm for supervised learning of binary classifiers (decides if a given vector belongs to a certain class or not), linear classifier (separation by hyperplanes)
- problem: shown impossible for these class of network to learn XOR function



- but multi-layer perceptrons can produce XOR function)

Perceptron

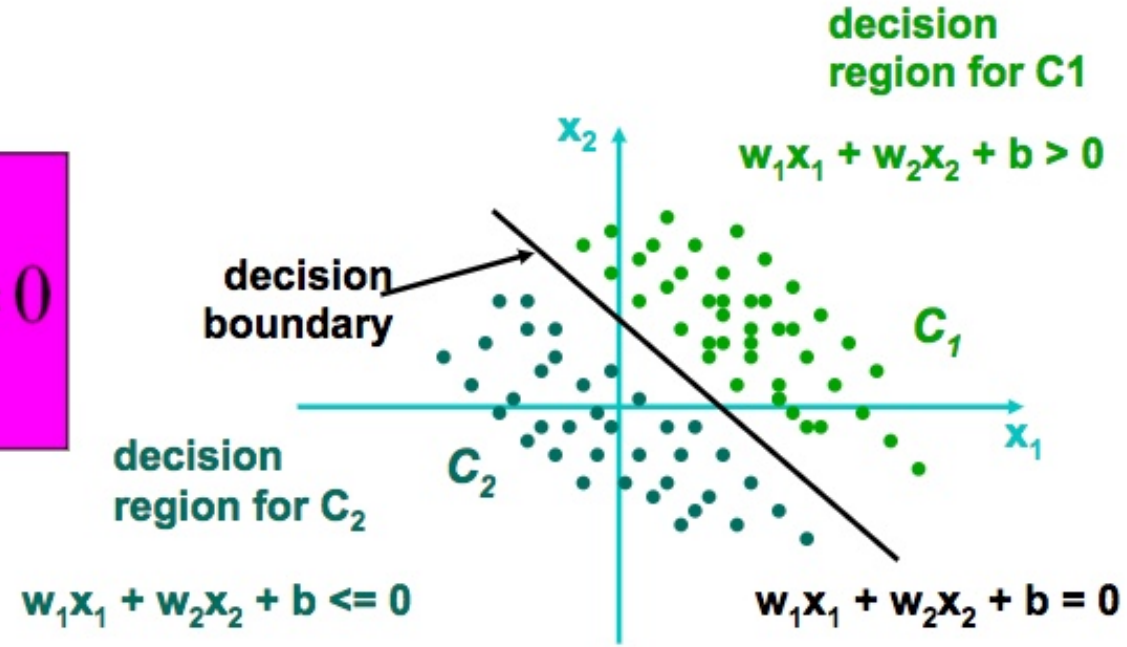
$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases}$$



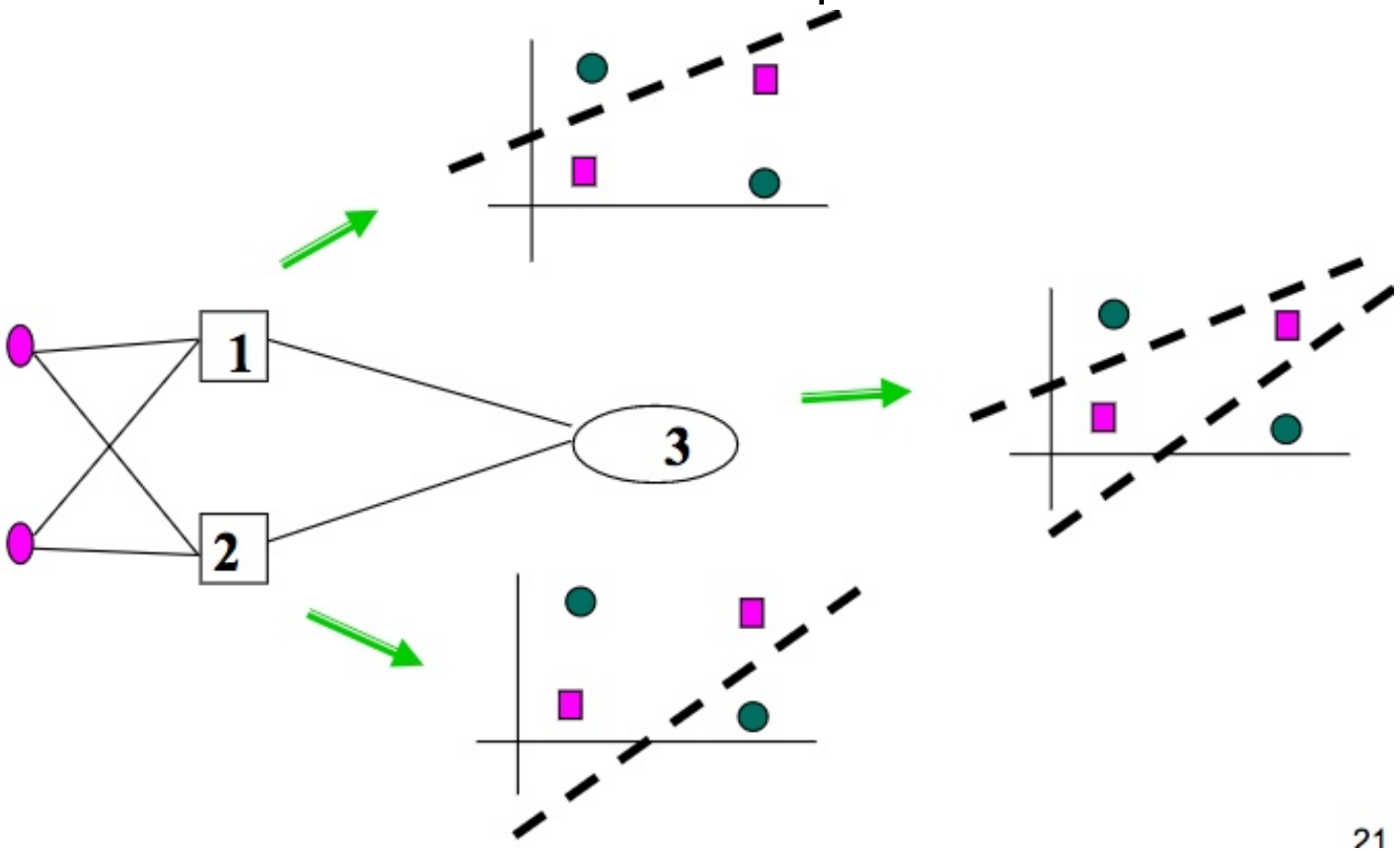
$$o = f\left(\sum_{k=1}^{n} i_k \cdot W_k\right)$$

Linear classification:

$$\sum_{i=1}^{m} w_i x_i + b = 0$$

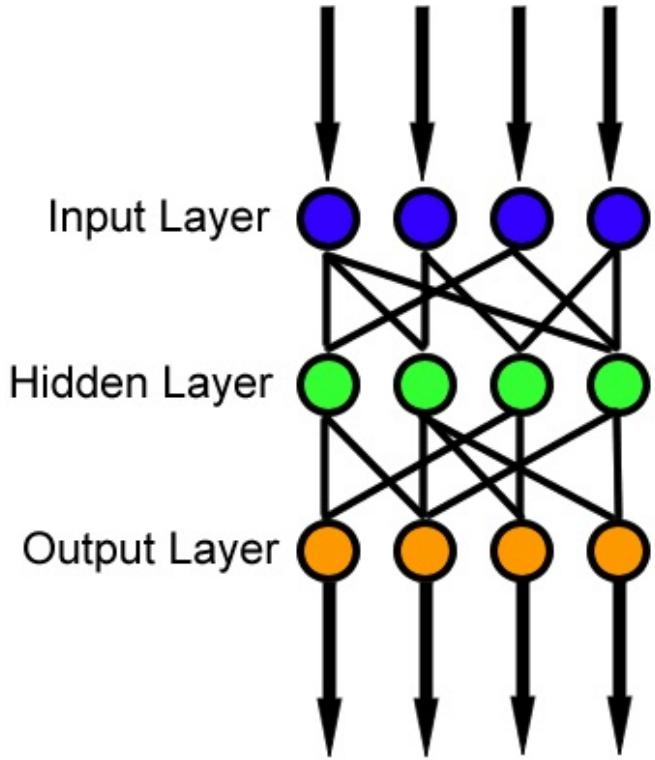


Solution to the XOR problem:



21

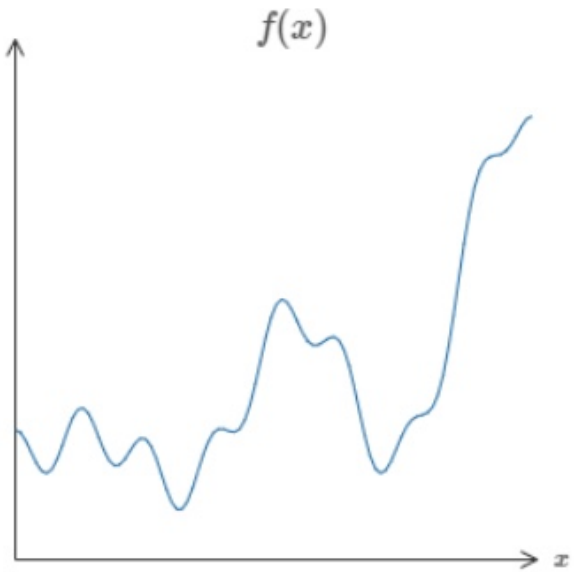
Feedforward Neural Networks

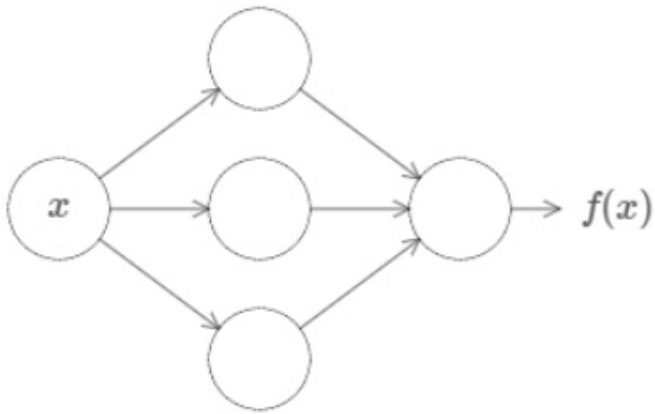


- acyclic: connections between the units do not form a cycle
- information flows always in one direction

Universal Approximation Theorem

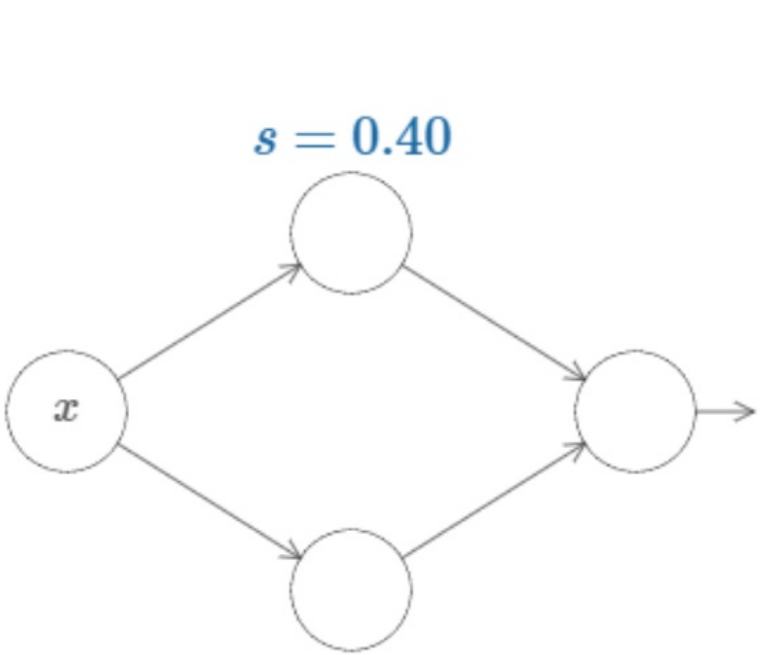
- Question: given a continuous function $f : \mathbb{R} \to \mathbb{R}$ can construct a neural network that at input x outputs a very good approximation of $f(x)$?

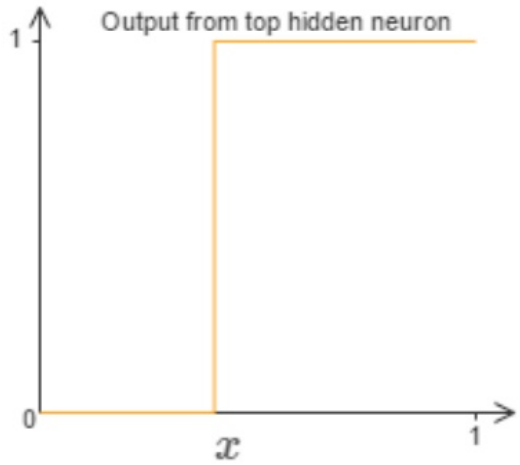




- Answer: yes this is always possible

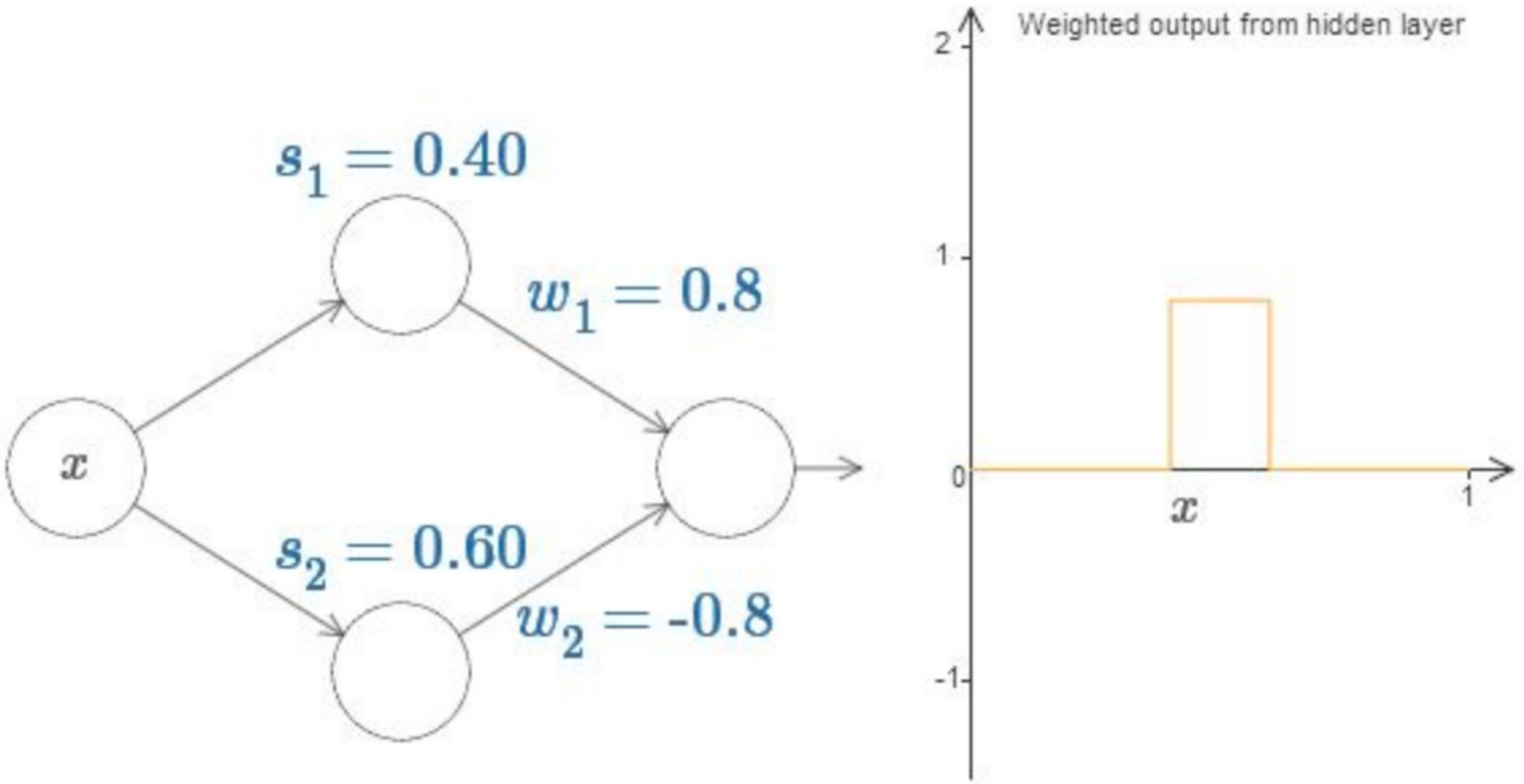
• Key idea: combinations of thresholds compute locally constant functions



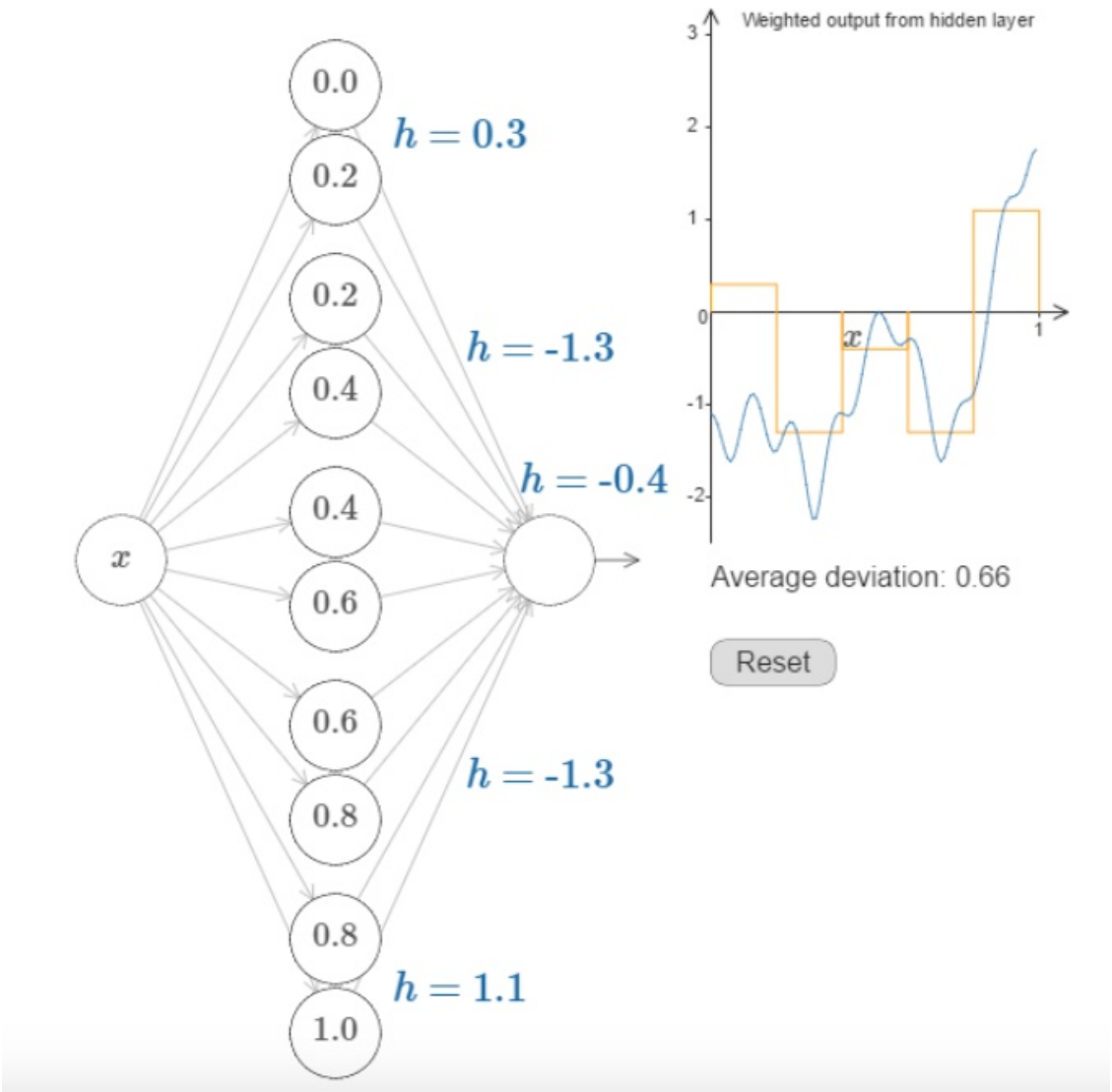


s=-b/w, big w and b results in step

single neuron computes $\sigma(wx + b)$ with $\sigma(x) = (1 + e^{-x})^{-1}$ resulting in a step (approximation)



combined effect of different nodes produce characteristic functions of intervals



characteristic functions of intervals approximate continuous functions

Computational experiment demonstrating increase in efficiency with depth

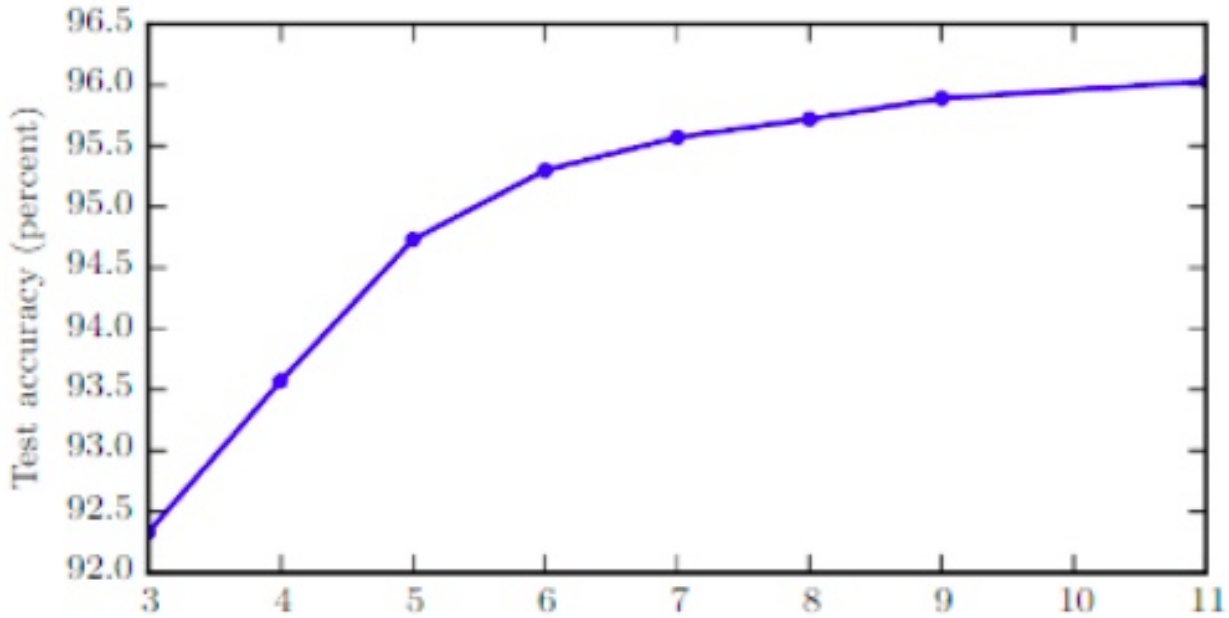


Figure 6.6: Empirical results showing that deeper networks generalize better when used to transcribe multi-digit numbers from photographs of addresses. Data from Goodfellow *et al.* (2014d). The test set accuracy consistently increases with increasing depth. See figure 6.7 for a control experiment demonstrating that other increases to the model size do not yield the same effect.

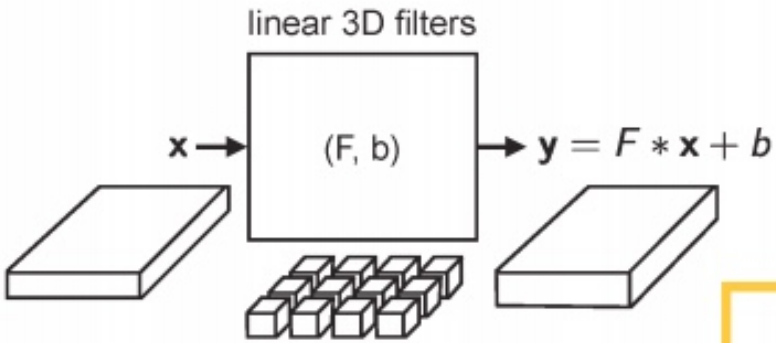
Convolutional Neural Networks

(special type of multi-layered percepton)

Linear convolution

57

A bank of "3D" linear filters



$$y_{ijq} = b_q + \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} \sum_{k=1}^{K} x_{u+i,v+j,k} f_{u,v,k,q}$$

Linear, translation invariant, local:

- Input **x** = H × W × K array
- Filter bank F = H' × W' × K × Q array
- Output **y** = (H - H' + 1) × (W - W' + 1) × Q array

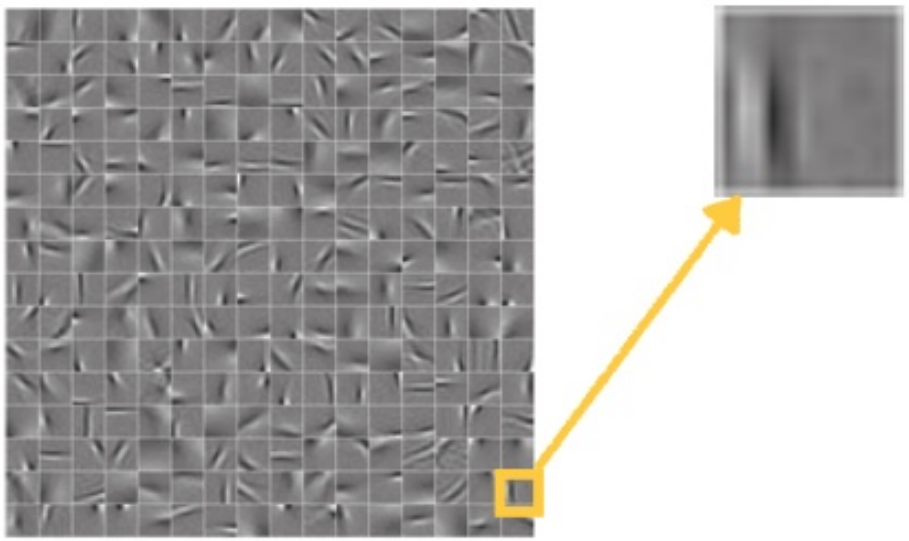
Linear convolution

Filter bank example

A bank of 256 filters (learned from data)

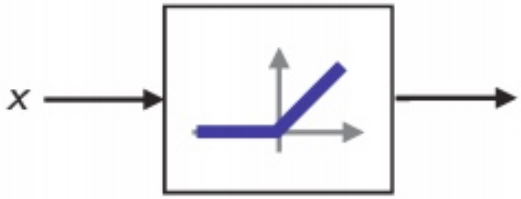
Each filter is 1D (it applies to a grayscale image)

Each filter is 16×16 pixels



Activation functions

Scalar non-linearity



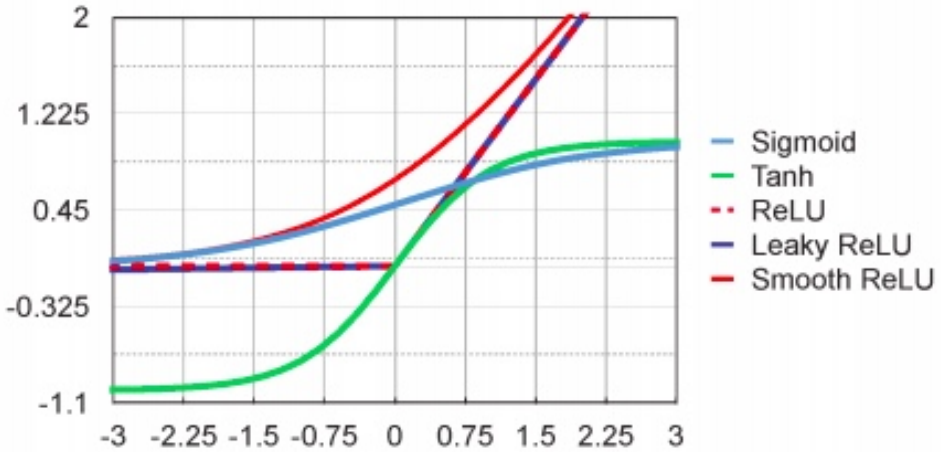
$$y = \frac{1}{1+e^{-x}} \quad \text{sigmoid}$$

$$y = \tanh(x) \quad \text{hyperb. tan}$$

$$y = \max\{0, x\} \quad \text{ReLU}$$

$$y = \log(1+e^{x}) \quad \text{Soft ReLU}$$

$$y = \epsilon x + (1-\epsilon)\max\{0, x\} \quad \text{Leaky ReLU}$$

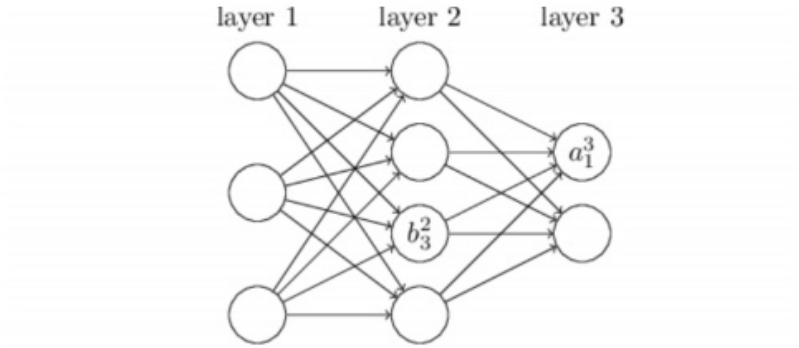


multiple layers: linear convolution, non-linear gating, linear convolution, ...

Back-propagation (learning technique)

- output value compared to (known) correct answer: compute error-function
- computed error fed back through the network
- algorithm that adjusts weights to reduce value of error function
- to adjust weights: optimization by gradient descent by computing derivatives of the error function with respect to weight parameters, and upgrading weights so that error decreases (Note: requires use of differentiable activation function)
- criticism of back-propagation: (Geoffrey Hinton) need an objective function, for which need a measure of distance between predicted value and labeled training data... problematic for unsupervised learning

Stochastic Gradient Descent via Back propagation



$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right) \qquad C = \frac{1}{2n}\sum_x \|y(x) - a^L(x)\|^2$$

Summary: the equations of backpropagation

$$\delta^L = \nabla_a C \odot \sigma'(z^L) \quad \text{(BP1)}$$

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l) \quad \text{(BP2)}$$

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l \quad \text{(BP3)}$$

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l \quad \text{(BP4)}$$

$a^L(x)$ vector of activations output of network with input x; $y(x)$ desired output; $\odot$ Hadamard product (componentwise product of two vectors); δ^ℓ errors in level ℓ; δ^L error in output layer

Stochastic Gradient Descent via Back propagation

1. **Input a set of training examples**

2. **For each training example x:** Set the corresponding input activation $a^{x,1}$, and perform the following steps:

   - **Feedforward:** For each $l = 2, 3, \ldots, L$ compute $z^{x,l} = w^l a^{x,l-1} + b^l$ and $a^{x,l} = \sigma(z^{x,l})$.

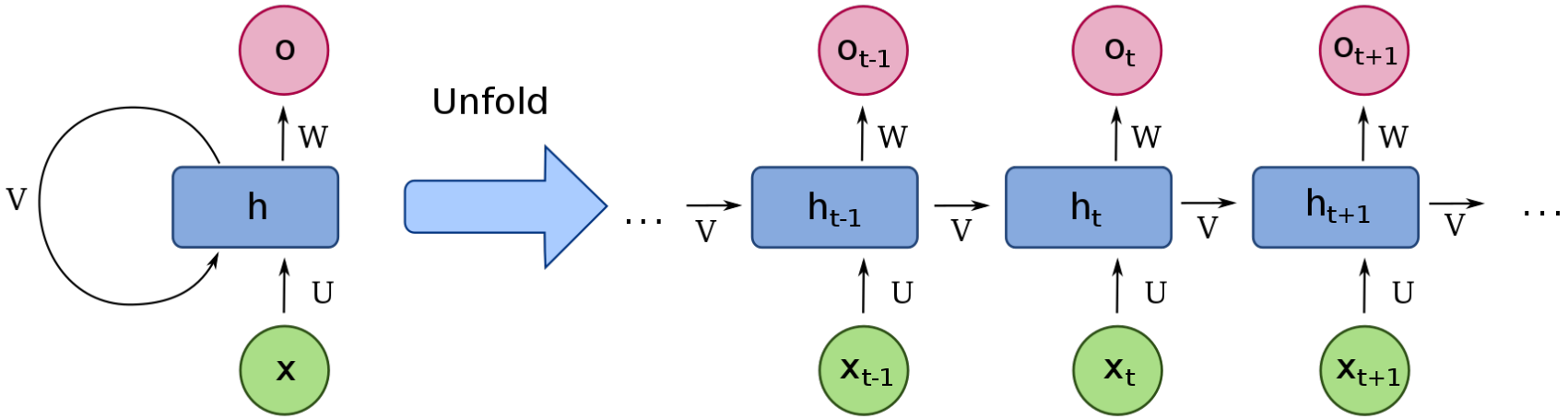
   - **Output error $\delta^{x,L}$:** Compute the vector $\delta^{x,L} = \nabla_a C_x \odot \sigma'(z^{x,L})$.

   - **Backpropagate the error:** For each $l = L-1, L-2, \ldots, 2$ compute $\delta^{x,l} = ((w^{l+1})^T \delta^{x,l+1}) \odot \sigma'(z^{x,l})$.

3. **Gradient descent:** For each $l = L, L-1, \ldots, 2$ update the weights according to the rule $w^l \to w^l - \frac{\eta}{m} \sum_x \delta^{x,l} (a^{x,l-1})^T$, and the biases according to the rule $b^l \to b^l - \frac{\eta}{m} \sum_x \delta^{x,l}$.

- Examples
  - Hopfield networks
  - Boltzmann machines

• directed graphs that allow cycles, storage internal states (memory), time delays, feedback loops, controlled states (gated states)

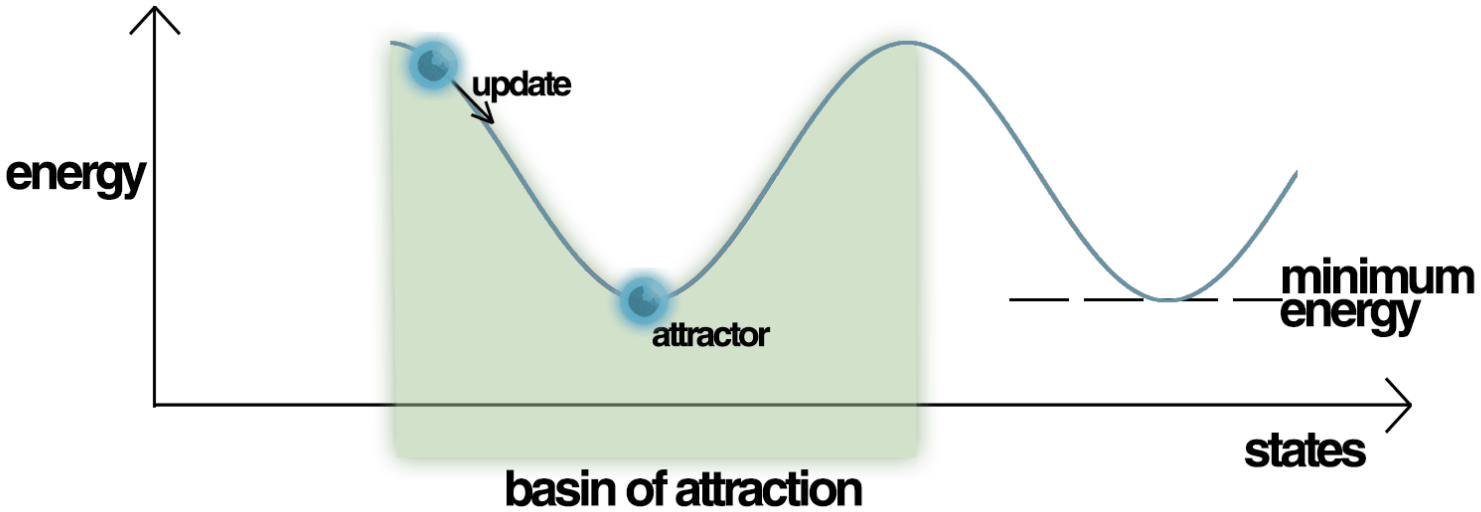


Hopfield Network

- historical connection between statistical physics of spin glass models and neural networks
- nodes variables $s_i = \pm 1$, update

$$s_i = \begin{cases} +1 & \sum_j w_{ij} s_j \geq \theta_i \\ -1 & \text{otherwise} \end{cases}$$

$$E = -\frac{1}{2} \sum_{i,j} w_{ij} s_i s_j - \sum_i \theta_i s_i$$



Energy landscape of the Hopfield network

Mehta-Schwab approach:

• Main idea: Kadanoff's "variational renormalization group scheme" for spin systems can be mapped exactly to a Deep Neural Network built of stacked layers of Restricted Boltzmann Machines (RBM), with a variational procedure based on minimizing the Kullback–Leibler divergence

Kadanoff's variational renormalization

- Leo P. Kadanoff, Anthony Houghton, Mehmet C. Yalabik, *Variational Approximations for Renormalization Group Transformations*, Journal of Statistical Physics, Vol. 14, (1976) No. 2, 171–203

• statistical mechanics problem: calculation of free energy in terms of a sum over states

• approximate recursion relations give upper and lower bounds on free energy

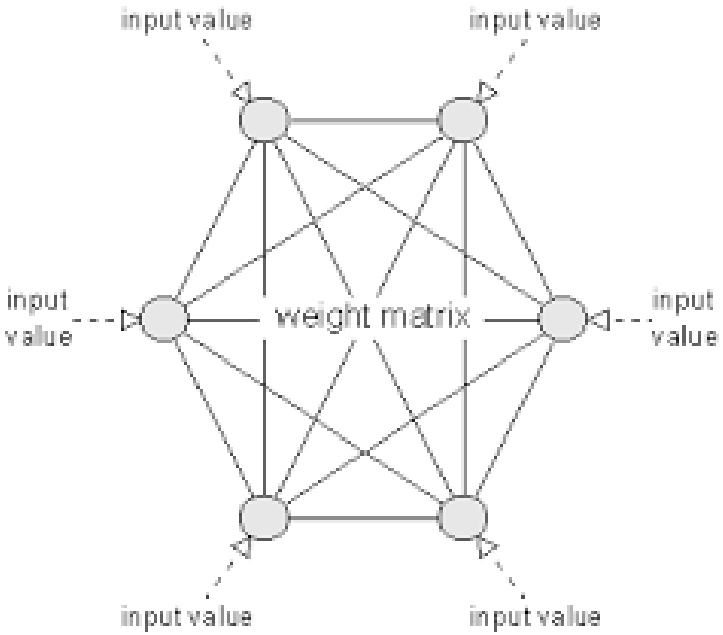
• optimized by treating parameters within the renormalization equations variationally

Ising Model on a graph G

- Hamiltonian

$$\mathcal{H} = -\sum_{v} B_v x_v - \sum_{e} J_e x_{s(e)} x_{t(e)}$$

- all nodes are "visible nodes": this type of spin glass model in statistical physics same as the Hopfield Associative Memory in neural network theory



• for other statistical physics systems more general Hamiltonians with many-body terms

$$\mathcal{H}(x) = -\sum_i \kappa_i x_i - \sum_{i,j} \kappa_{ij} x_i x_j - \sum_{ijk} \kappa_{ijk} x_i x_j x_k \cdots$$

• partition function: sum over configurations

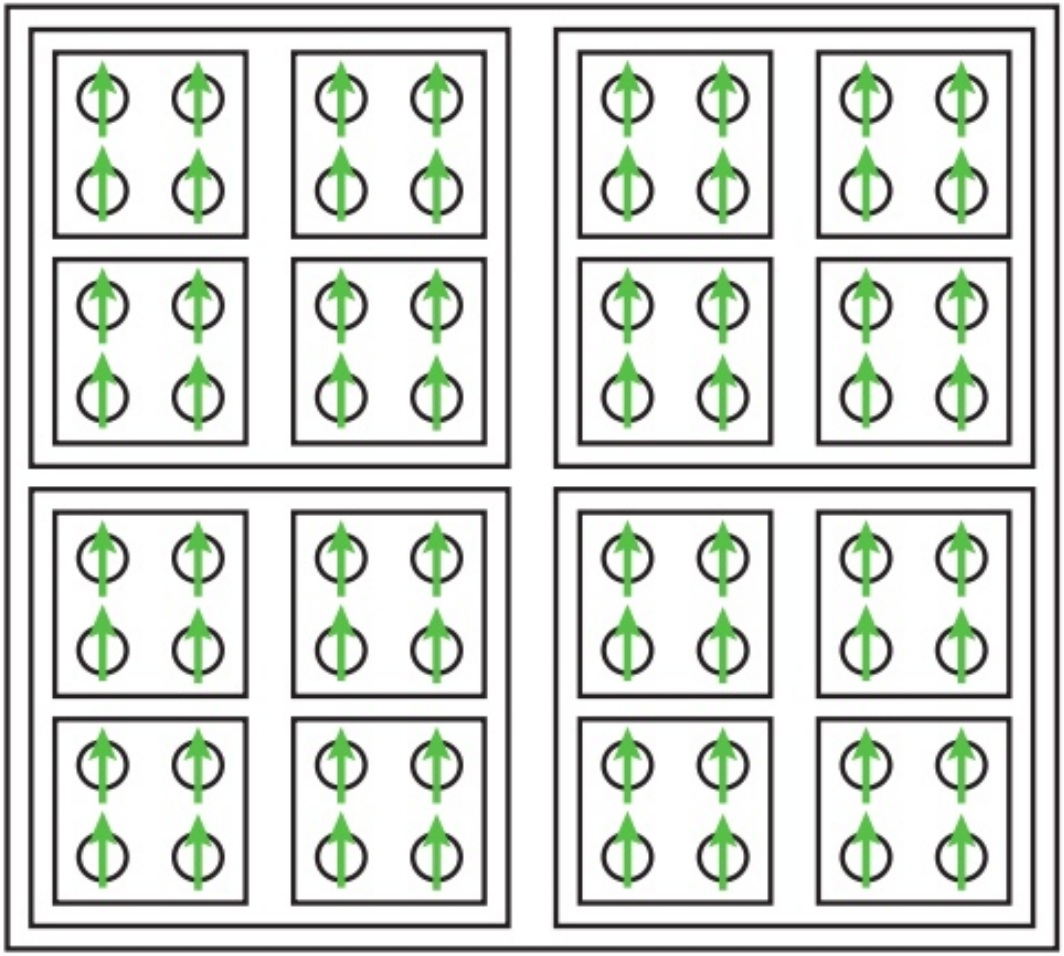
$$\mathcal{Z} = \sum_x e^{-\mathcal{H}(x)}$$

thermodynamic parameter $\beta = 1/T$ inverse temperature

• free energy

$$\mathcal{F} = -\log \mathcal{Z} = -\log \sum_x e^{-\beta \mathcal{H}(x)}$$

• physical system is coarse grained by introducing "block" variables that average spins in a block, effective behavior, sequence of successive coarse graining



free energy (after a rescaling) is preserved

- Hamiltonian $\mathcal{H}(x)$ describes system at *fine grained scales*
- construct a correlation function $V(x, h)$ that couples it to the next level of *coarse graining*

$$\sum_h e^{-V(x,h)} = 1$$

for all x so partition function remains unchanged (and free energy)

- joint Hamiltonian

$$\mathcal{H}(x, h) = \mathcal{H}(x) + V(x, h)$$

$$\mathcal{Z} = \sum_x e^{-\beta\mathcal{H}(x)} = \sum_h \sum_x e^{-\beta\mathcal{H}(x,h)}$$

- renormalized effective Hamiltonian acting on hidden nodes

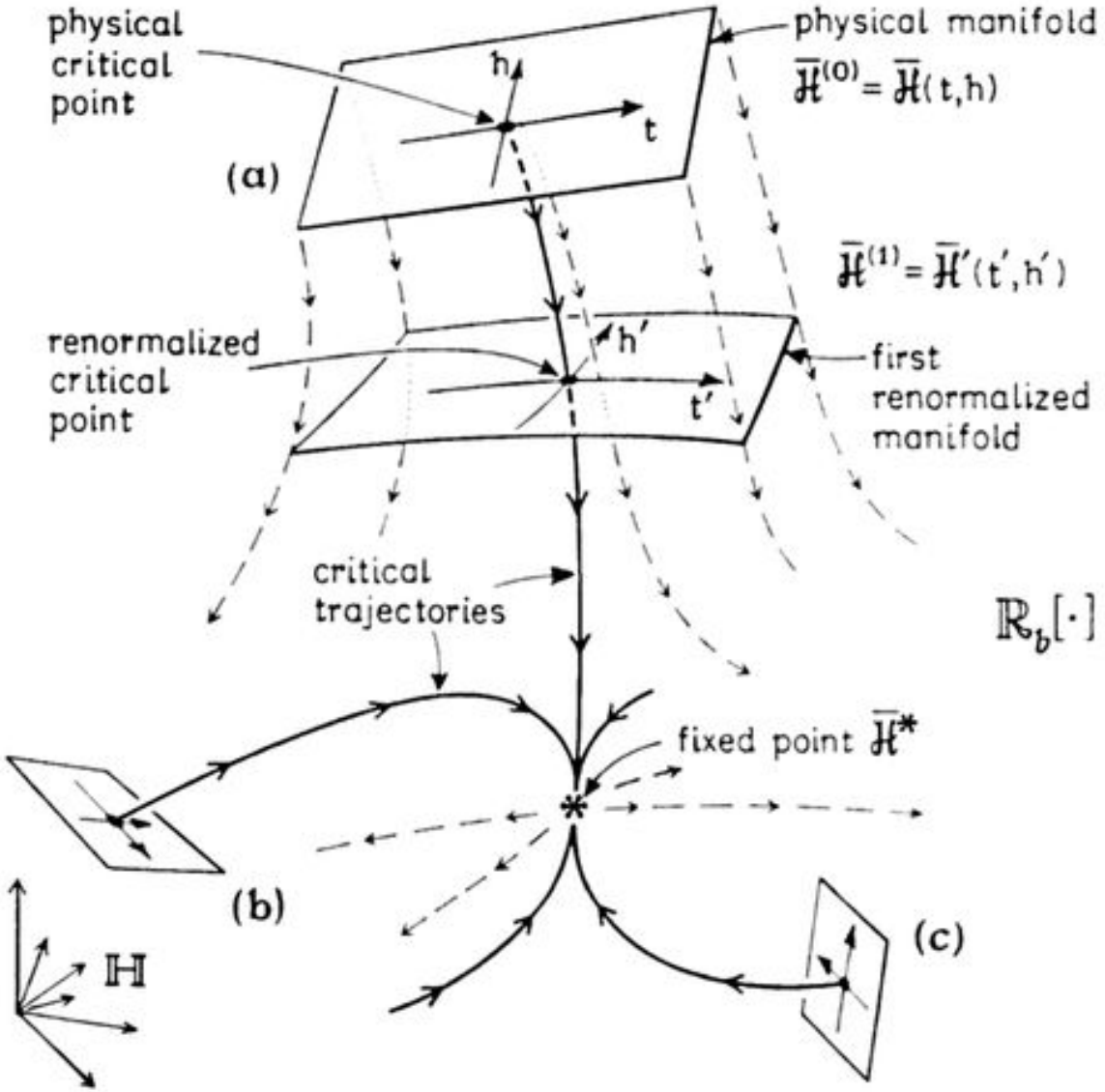
$$\bar{\mathcal{H}}(h) = \log \sum_x e^{-\beta V(x,h)} e^{-\beta \mathcal{H}(x)}$$

$$\bar{\mathcal{Z}} = \sum_h e^{-\beta \bar{\mathcal{H}}(h)}$$

- want $V(x, h)$ that minimizes the free energy difference between

$$\bar{\mathcal{F}} = -\log \bar{\mathcal{Z}} \quad \text{and} \quad \mathcal{F} = -\log \mathcal{Z}$$

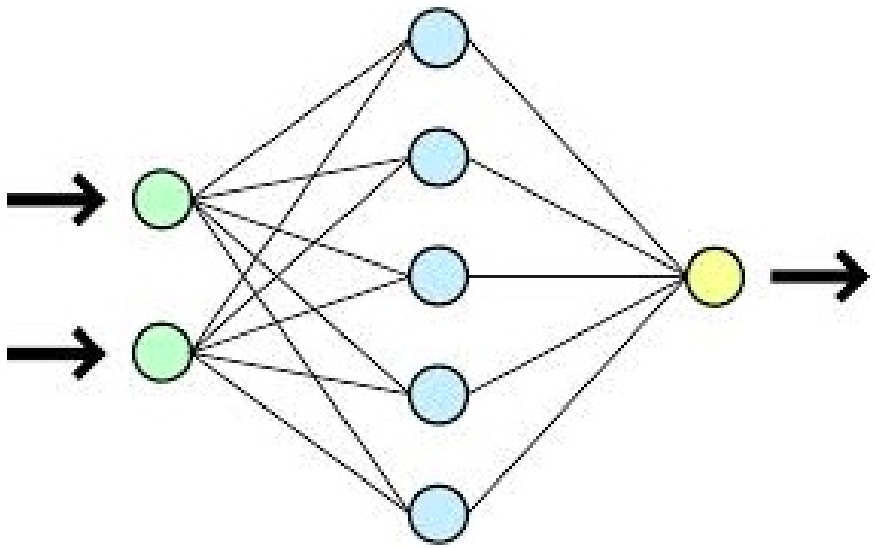
- variational problem: Kadanoff–Houghton–Yalabik computed explicit lower bounds for the minimizer $V(x, h)$ for given systems (eg Ising model on a graph)


physical
critical
point
h
t
(a)
physical manifold
$\overline{\mathcal{H}}^{(0)} = \overline{\mathcal{H}}(t,h)$
$\overline{\mathcal{H}}^{(1)} = \overline{\mathcal{H}}'(t',h')$
renormalized
critical
point
h'
t'
first
renormalized
manifold
critical
trajectories
$\mathbb{R}_b[\cdot]$
fixed point $\overline{\mathcal{H}}^*$
(b)
(c)
$\mathbb{H}$

Variational RG and neural networks

• idea: introducing layers of nodes with hidden variables in a neural network is a form of Renormalization related to *scale* change

• modify the Hopfield network architecture to introduce hidden nodes: Hinton's restricted Boltzmann machines (RBM)



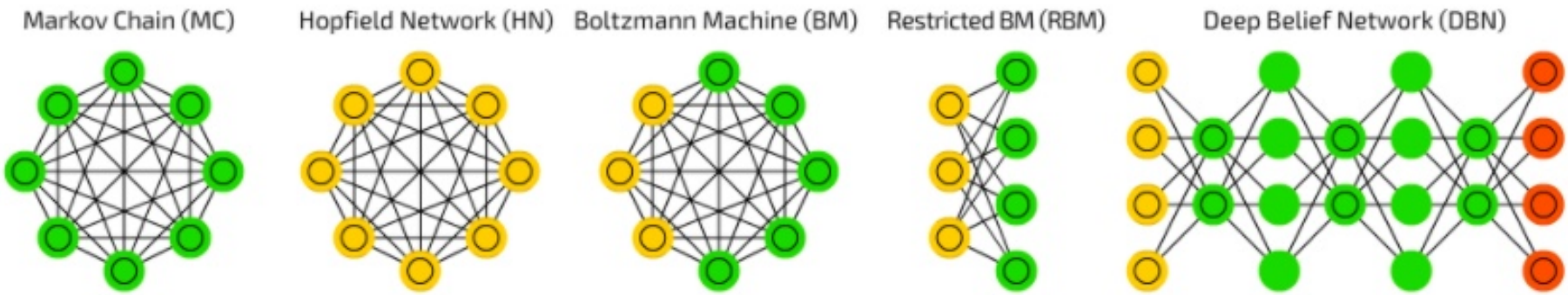
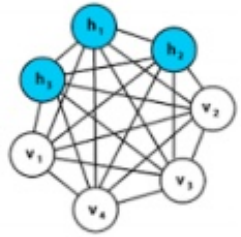


Image: The Asimov Institute

(yellow=backfed input node, green=probabilistic hidden node, red=match input output node)

Boltzmann machine



A Boltzmann machine, like a Hopfield network, is a network of units with an "energy" defined for the network. It also has binary units, but unlike Hopfield nets, Boltzmann machine units a stochastic. The global energy, E, in a Boltzmann machine is identical in form to that of a Hopfield network:

$$E = -\left(\sum_{i<j} w_{ij}\, s_i\, s_j + \sum_i \theta_i\, s_i\right)$$

Where:

- w_{ij} is the connection strength between unit j and unit i.
- s_i is the state, $s_i \in \{0, 1\}$, of unit i.
- θ_i is the bias of unit i in the global energy function. ($-\theta_i$ is the activation threshold for the unit.)

Often the weights are represented in matrix form with a symmetric matrix W, with zeros along the diagonal.

The difference in the global energy that results from a single unit i being 0 (off) versus 1 (on), written ΔE_i, assuming a symmetric matrix of weights, is given by:

$$\Delta E_i = \sum_{j>i} w_{ij}\, s_j + \sum_{j<i} w_{ji}\, s_j + \theta_i$$

We can now finally solve for $p_{\mathrm{i=on}}$, the probability that the i-th unit is on.

$$p_{\mathrm{i=on}} = \frac{1}{1 + \exp(-\frac{\Delta E_i}{T})}$$

There are two phases to Boltzmann machine training, and we switch iteratively between them. One is the "positive" phase where the visible units' states are clamped to a particular bin (according to P^+). The other is the "negative" phase where the network is allowed to run freely, i.e. no units have their state determined by external data. Surprisingly enough, the gr given by the very simple equation (proved in Ackley et al.[3]):

$$\frac{\partial G}{\partial w_{ij}} = -\frac{1}{R}[p_{ij}^{+} - p_{ij}^{-}]$$

where:

- p_{ij}^{+} is the probability of units *i* and *j* both being on when the machine is at equilibrium on the positive phase.
- p_{ij}^{-} is the probability of units *i* and *j* both being on when the machine is at equilibrium on the negative phase.
- R denotes the learning rate

- Energy functional for RBM:

$$E(x,h) = x^t B + x^t W h + C^t h$$

- probability distribution

$$\mathbb{P}(x,h) = \frac{e^{-\beta E(x,h)}}{\mathcal{Z}}, \quad \mathcal{Z} = \sum_{x,h} e^{-\beta E(x,h)}$$

- distributions for visible and hidden nodes: marginals

$$\mathbb{P}(x) = \sum_h \mathbb{P}(x,h) = \sum_h \frac{e^{-\beta E(x,h)}}{\mathcal{Z}}$$

$$\mathbb{P}(h) = \sum_x \mathbb{P}(x,h) = \sum_x \frac{e^{-\beta E(x,h)}}{\mathcal{Z}}$$

- Hamiltonians for visible and hidden nodes

$$\mathcal{H}(x) = -\log \sum_{h} e^{-E(x,h)}$$

$$\mathcal{H}(h) = -\log \sum_{x} e^{-E(x,h)}$$

- training of RBMs: comparing free energies on training data and validation data and minimize difference in free energy

- the parameters in the RBM are chosen to minimize the Kullback–Leibler divergence between the true distribution of the data and the variational distribution

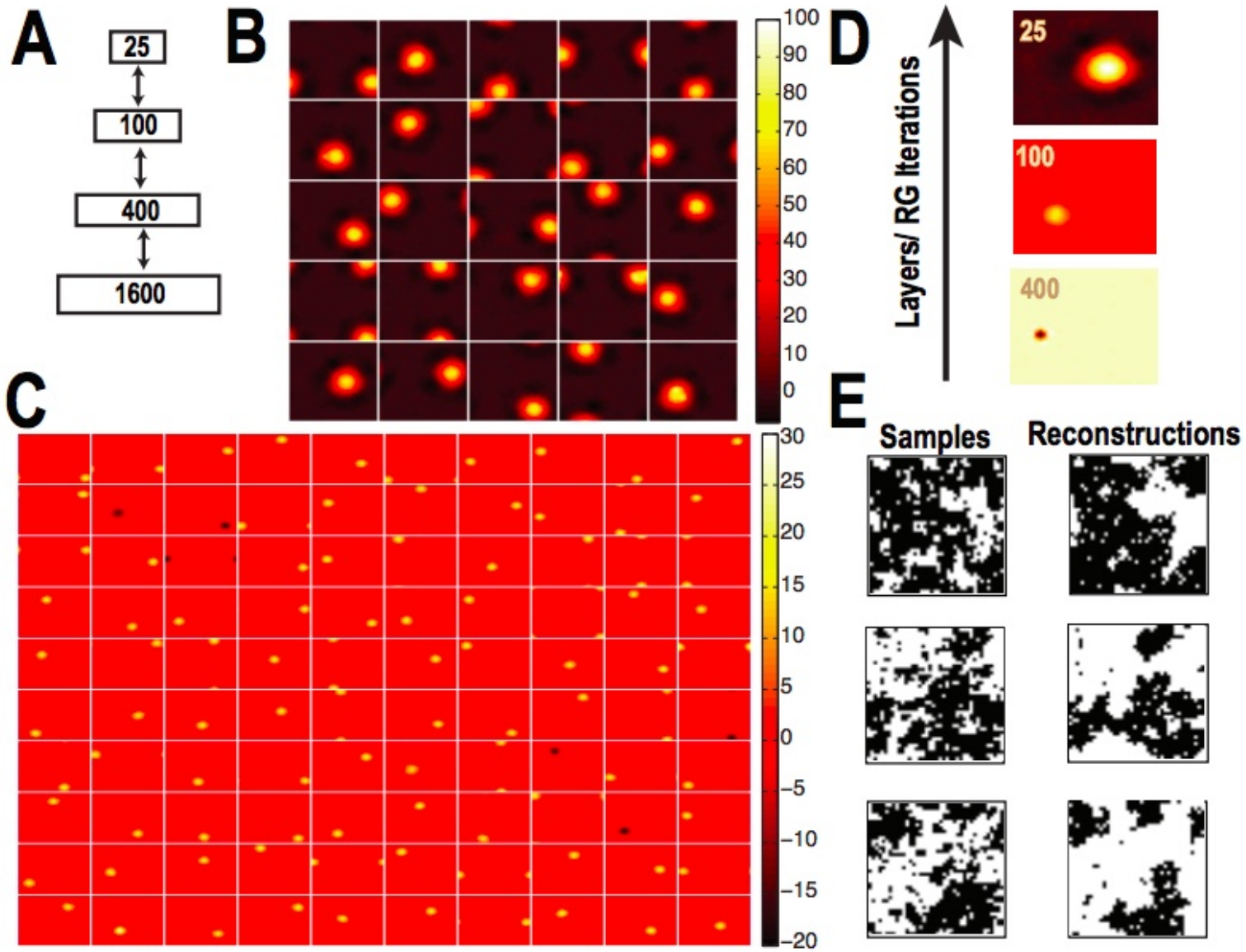
• the *variational distribution* is the one obtained as marginal

$$\mathbb{P}(x) = \sum_h \mathbb{P}(x, h) = \sum_h \frac{e^{-\beta E(x,h)}}{\mathcal{Z}}$$

• so minimizing the difference in free energy between training data and validation data can be done by minimizing the Kullback–Leibler divergence

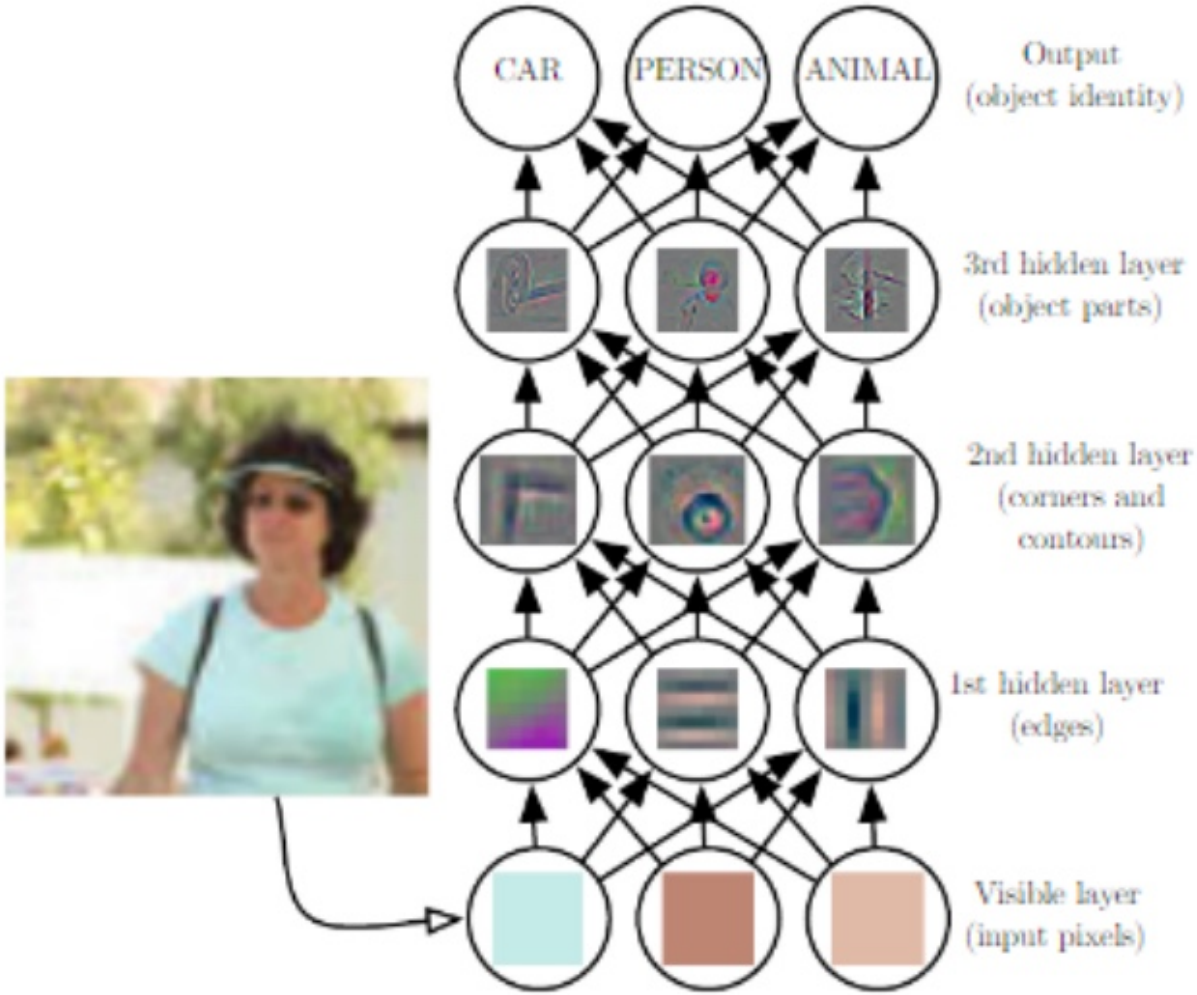
$$\mathrm{KL}(P|\mathbb{P}) = \sum_x P(x) \log \frac{P(x)}{\mathbb{P}(x)}$$

• so can map each step of the Kadanoff–Houghton–Yalabik variational RG method to a RBM: resulting architecture is a stacked layers of Restricted Boltzmann Machines (RBM), Deep Belief Network (DBN)



from Mehta–Schwab: 2D Ising Model simulated by a DNN of RBMs

- what it shows: Kadanoff variational RG algorithm can be implemented on a network given by a stack of RBMs
- what it claims: Deep Learning is a form of Renormalization



Mathematical theory of learning:

an approach to studying and estimating limits of learning and learnability in neural networks

- F. Cucker, S. Smale, *On the mathematical foundations of learning*, Bulletin of the American Math. Society 39 (2001) N.1, 1–49.

• General problem: when two sets of random variables x, y are probabilistically related

- relation described by probability distribution $P(x, y)$
- some square loss problem (minimization problem)

$$E(f) = \int (y - f(x))^2 \, P(x, y) \, dx \, dy$$

- distribution itself unknown, but minimize empirical error

$$E_N(f) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2$$

over a set of random sampled data points $\{(x_i, y_i)\}_{i=1,\ldots,N}$

- if f_N minimizes empirical error, want that the probability

$$\mathbb{P}(\|E(f_N) - E_N(f_N)\| > \epsilon)$$

is sufficiently small

- Problem depends on the function space where f_N lives

General setting

• F. Cucker, S. Smale, *On the mathematical foundations of learning*, Bulletin of the American Math. Society 39 (2001) N.1, 1–49.

- X compact manifold, $Y = \mathbb{R}^k$ (for simplicity $k = 1$), $Z = X \times Y$ with Borel measure ρ
- ξ random variable (real valued) on probability space (Z, ρ)
- expectation value and variance

$$\mathbb{E}(\xi) = \int_Z \xi \, d\rho, \qquad \sigma^2(\xi) = \mathbb{E}((\xi - \mathbb{E}(\xi))^2) = \mathbb{E}(\xi^2) - \mathbb{E}(\xi)^2$$

- function $f : X \to Y$, *least squares error* of f

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 \, d\rho$$

  measures average error incurred in using $f(x)$ as a model of the dependence between y and x
- Problem: how to minimize the error?

- conditional probability $\rho(y|x)$ (probability measure on Y)
- marginal probability $\rho_X(S) = \rho(\pi^{-1}(S))$ on X, with projection $\pi : Z = X \times Y \to X$
- relation between these measures

$$\int_Z \phi(x,y)\, d\rho = \int_X \left(\int_Y \phi(x,y)\, d\rho(y|x) \right) d\rho_X$$

- breaking of $\rho(x,y)$ into $\rho(y|x)$ and $\rho_X(S)$ is breaking of Z into input X and output Y

- regression function $f_\rho : X \to Y$

$$f_\rho(x) = \int_Y y \, d\rho(y|x)$$

- *assumption*: f_ρ is bounded
- for fixed $x \in X$ map Y to $\mathbb{R}$ via

$$y \mapsto y - f_\rho(x)$$

- expectation value is zero so variance

$$\sigma^2(x) = \int_Y (y - f_\rho(x))^2 \, d\rho(y|x)$$

- averaged variance

$$\sigma_\rho^2 = \int_X \sigma^2(x) \, d\rho_X = \mathcal{E}(f_\rho)$$

  measures how "well conditioned" ρ is
- Note: in general ρ and f_ρ not known but ρ_X known

- error, regression, and variance:

$$\mathcal{E}(f) = \int_X ((f(x) - f_\rho(x))^2 + \sigma_\rho^2)\, d\rho_X$$

- What this says: σ_ρ^2 is a lower bound for the error $\mathcal{E}(f)$ for all f, and $f = f_\rho$ has the smallest possible error (which depends only on ρ)
- why identity holds:

$$\begin{aligned}
\mathcal{E}(f) &= \int_Z (f(x) - f_\rho(x) + f_\rho(x) - y)^2 \\
&= \int_X (f(x) - f_\rho(x))^2 + \int_X \int_Y (f_\rho(x) - y)^2 \\
&\qquad\qquad +2 \int_X \int_Y (f(x) - f_\rho(x))(f_\rho(x) - y) \\
&= \int_X (f(x) - f_\rho(x))^2 + \sigma_\rho^2.
\end{aligned}$$

Goal: "learn" (= find a good approximation for) f_ρ given random samples of Z

- $Z^N \ni z = ((x_1, y_1), \ldots, (x_N, y_N))$ sample set of points (x_i, y_i) independently drawn with probability ρ
- empirical error

$$\mathcal{E}_z(f) = \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2$$

- for random variable ξ *empirical mean*

$$\mathbb{E}_z(\xi) = \frac{1}{N} \sum_{i=1}^{N} \xi(z_i, y_i)$$

- given $f : X \to Y$ take $f_Y : Z \to Y$ to be $f_Y : (x, y) \mapsto f(x) - y$

$$\mathcal{E}(f) = \mathbb{E}(f_Y^2), \qquad \mathcal{E}_z(f) = \mathbb{E}_z(f_Y^2)$$

Facts of Probability Theory

(quantitative versions of law of large numbers)

- ξ random variable on probability space Z with mean $\mathbb{E}(\xi) = \mu$ and variance $\sigma^2(\xi) - \sigma^2$
- Chebyshev: for all $\epsilon > 0$

$$\mathbb{P}\left\{ z \in Z^m : \left| \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mu \right| \geq \epsilon \right\} \leq \frac{\sigma^2}{m\epsilon^2}$$

- Bernstein: if $|\xi(z) - \mathbb{E}(\xi)| \leq M$ for almost all $z \in Z$ then $\forall \epsilon > 0$

$$\mathbb{P}\left\{ z \in Z^m : \left| \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mu \right| \geq \epsilon \right\} \leq 2 \exp\left(- \frac{m\epsilon^2}{2(\sigma^2 + \frac{1}{3} M\epsilon)} \right)$$

- Hoeffding:

$$\mathbb{P}\left\{ z \in Z^m : \left| \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mu \right| \geq \epsilon \right\} \leq 2 \exp\left(- \frac{m\epsilon^2}{2M^2} \right)$$

Defect Function of $f : X \to Y$

$$L_z(f) := \mathcal{E}(f) - \mathcal{E}_z(f)$$

discrepancy between error and empirical error (only $\mathcal{E}_z(f)$ measured directly)

- estimate of defect if $|f(x) - y| \leq M$ almost everywhere, then $\forall \epsilon > 0$, with σ^2 variance of f_Y^2

$$\mathbb{P}\{z \in Z^m \,:\, |L_z(f)| \leq \epsilon\} \geq 1 - 2\epsilon \exp\left(-\frac{m\epsilon^2}{2(\sigma^2 + \frac{1}{3}M^2\epsilon)}\right)$$

- from previous Bernstein estimate taking $\xi = f_Y^2$
- when is $|f(x) - y| \leq M$ a.e. satisfied? e.g. for $M = M_\rho + P$

$$M_\rho = \inf\{\bar{M} \,:\, \{(x,y) \in Z \,:\, |y - f_\rho(x)| \geq \bar{M}\} \text{ measure zero } \}$$

$$P \geq \sup_{x \in X} |f(x) - f_\rho(x)|$$

Hypothesis Space

- a learning process requires a datum of a class of functions (hypothesis space) within which the best approximation for f_ρ
- $C(X)$ algebra of continuous functions on topological space X
- $\mathcal{H} \subset C(X)$ compact subset (not necessarily subalgebra)
- look for minimizer (not necessarily unique)

$$f_{\mathcal{H}} = \operatorname{argmin}_{f \in \mathcal{H}} \int_Z (f(x) - y)^2$$

because $\mathcal{E}(f) = \int_X (f - f_\rho)^2 + \sigma_\rho^2$ also minimizer

$$f_{\mathcal{H}} = \operatorname{argmin}_{f \in \mathcal{H}} \int_X (f - f_\rho)^2$$

- continuity: if for $f \in \mathcal{H}$ have $|f(x) - y| \leq M$ a.e., bounds

$$|\mathcal{E}(f_1) - \mathcal{E}(f_2)| \leq 2M \|f_1 - f_2\|_\infty$$

and for $\mathcal{E}_z$ also, so $\mathcal{E}$ and $\mathcal{E}_z$ continuous

- compactness of $\mathcal{H}$ ensures existence of minimizer but not uniqueness (a uniqueness result when $\mathcal{H}$ convex)

Empirical target function $f_{\mathcal{H},z}$

- minimizer (non unique in general)

$$f_{\mathcal{H},z} = \text{argmin}_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

Normalized Error

$$\mathcal{E}_{\mathcal{H}}(f) = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}})$$

$\mathcal{E}_{\mathcal{H}}(f) \geq 0$ vanishing at $f_{\mathcal{H}}$

Sample Error $\mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},z})$

$$\mathcal{E}(f_{\mathcal{H},z}) = \mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},z}) + \mathcal{E}(f_{\mathcal{H}}) = \int_X (f_{\mathcal{H},z} - f_\rho)^2 + \sigma_\rho^2$$

estimating $\mathcal{E}(f_{\mathcal{H},z})$ by estimating sample and approximation errors, $\mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},z})$ and $\mathcal{E}(f_{\mathcal{H}})$ one on $\mathcal{H}$ the other independent of sample z

bias-variance trade-off

- bias = approximation error; variance = sample error
  - fix $\mathcal{H}$: sample error $\mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},z})$ decreases by increasing number m of samples
  - fix m: approximation error $\mathcal{E}(f_{\mathcal{H}})$ decreases when enlarging $\mathcal{H}$
- procedure:
  1. estimate how close $f_{\mathcal{H},z}$ and $f_{\mathcal{H}}$ depending on m
  2. how to choose $\dim \mathcal{H}$ when m is fixed
- first problem: how many examples need to draw to say with confidence $\geq 1-\delta$ that $\int_X (f_{\mathcal{H},z} - f_{\mathcal{H}})^2 \leq \epsilon$?

Uniformity Estimate (Vapnik's Statistical Learning Theory)

• covering number: S metric space, $s > 0$, number $\mathcal{N}(S, s)$ minimal $\ell \in \mathbb{N}$ so that $\exists$ disks in S radii s covering S; for S compact $\mathcal{N}(S, s)$ finite

• uniform estimate: $\mathcal{H} \subset C(X)$ compact, if for all $f \in \mathcal{H}$ have $|f(x) - y| \leq M$ a.e., then $\forall \epsilon > 0$

$$\mathbb{P}\{z \in Z^m : \sup_{f \in \mathcal{H}} |L_z(f)| \leq \epsilon\} \geq 1 - \mathcal{N}(\mathcal{H}, \frac{\epsilon}{8M}) 2 \exp\left(-\frac{m\epsilon^2}{4(2\sigma^2 + \frac{1}{3}M^2\epsilon)}\right)$$

with $\sigma^2 = \sup_{f \in \mathcal{H}} \sigma^2(f_Y^2)$

• main idea: like previous "estimate of defect" but passing from a single function to a family of functions, using a uniformity based on "covering number"

Estimate of Sample Error

- $\mathcal{H} \subset C(X)$ compact, with $|f(x) - y| \leq M$ a.e. for all $f \in \mathcal{H}$, and $\sigma^2 = \sup_{f \in \mathcal{H}} \sigma^2(f_Y^2)$, then $\forall \epsilon > 0$

$$\mathbb{P}\{z \in Z^m : \mathcal{E}_{\mathcal{H}}(f_z) \leq \epsilon\} \geq 1 - \mathcal{N}(\mathcal{H}, \frac{\epsilon}{16M}) 2 \exp\left(-\frac{m\epsilon^2}{8(4\sigma^4 + \frac{1}{3}M^2\epsilon)}\right)$$

- obtained from previous estimate using $L_z(f) = \mathcal{E}(f) - \mathcal{E}_z(f)$
- so answer to first question: to ensure probability above $\geq 1 - \delta$ need to take at least

$$m \geq \frac{8(4\sigma^4 + \frac{1}{3}M^2\epsilon)}{\epsilon^2}\left(\log(2\mathcal{N}(\mathcal{H}, \frac{\epsilon}{16M})) + \log(\frac{1}{\delta})\right)$$

obtained by setting

$$\delta = \mathcal{N}(\mathcal{H}, \frac{\epsilon}{16M}) 2 \exp\left(-\frac{m\epsilon^2}{8(4\sigma^4 + \frac{1}{3}M^2\epsilon)}\right)$$

- need various techniques for estimating covering numbers $\mathcal{N}(\mathcal{H}, s)$ depending on the choice of the compact set $\mathcal{H}$

Second Question: Estimating the Approximation Error

$$\mathcal{E}(f_{\mathcal{H},z}) = \mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},z}) + \mathcal{E}(f_{\mathcal{H}})$$

focus on $\mathcal{E}(f_{\mathcal{H}})$, which depends on $\mathcal{H}$ and ρ

$$\int_X (f_{\mathcal{H}} - f_\rho)^2 + \sigma_\rho^2$$

second term independent of $\mathcal{H}$ so focus on first; f_ρ bounded, but not in $\mathcal{H}$ nor necessarily in $C(X)$

- Main idea: use finite dimensional hypothesis space $\mathcal{H}$; estimate in terms of growth of eigenvalues of an operator
- Main technique: Fourier analysis; Hilbert spaces

Fourier Series: start with case of $X = T^n = (S^1)^n$ torus

• Hilbert space $L^2(X)$ Lebesgue measure with complete orthonormal system

$$\phi_\alpha(x) = (2\pi)^{-n/2} \exp(i\alpha \cdot x), \quad \alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{Z}^n$$

Fourier series expansion

$$f = \sum_{\alpha \in \mathbb{Z}^n} c_\alpha \, \phi_\alpha$$

• finite dimensional subspaces $\mathcal{H}_N \subset L^2(X)$ spanned by ϕ_α with $\|\alpha\| \leq B$, dimension $N(B)$ number of lattice points in ball radius B in $\mathbb{R}^n$

$$N(B) \leq (2B)^{n/2}$$

• $\mathcal{H}$ hypothesis space: ball $\mathcal{H}_{N,R}$ of radius R in $\| \cdot \|_\infty$ norm in $\mathcal{H}_N$

Laplacian

- on torus $X = T^n$ Laplacian $\Delta : \mathcal{C}^\infty(X) \to \mathcal{C}^\infty(X)$

$$\Delta(f) = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}$$

Fourier series basis ϕ_α are eigenfunctions of $-\Delta$ with eigenvalue $\|\alpha\|^2$

- more general X: bounded domain $X \subset \mathbb{R}^n$ with smooth boundary ∂X and a complete orthonormal system ϕ_k of $L^2(X)$ (Lebesgue measure) of eigenfunctions of Laplacian with

$$-\Delta(\phi_k) = \zeta_k \, \phi_k, \quad \phi_k|_{\partial X} \equiv 0, \quad \forall k \geq 1$$

$$0 < \zeta_1 \leq \zeta_2 \leq \cdots \leq \zeta_k \leq \cdots$$

- subspace $\mathcal{H}_N$ of $L^2(X)$ generated by $\{\phi_1, \ldots, \phi_N\}$
- hypothesis space $\mathcal{H} = \mathcal{H}_{N,R}$ ball of radius R for $\|\cdot\|_\infty$ in $\mathcal{H}_N$

Construction of $f_{\mathcal{H}}$

- Lebesgue measure μ on X and measure ρ (marginal probability ρ_X induced by ρ on $Z = X \times Y$)
- consider regression function

$$f_\rho(x) = \int_Y y \, d\rho(y|x)$$

- assumption f_ρ bounded on X so in $L^2_\rho(X)$ and in $L^2_\mu(X)$
- choice of R: assume also that $R \geq \|f_\rho\|_\infty$, which implies $R \geq \|f_\rho\|_\rho$
- then $f_{\mathcal{H}}$ is orthogonal projection of f_ρ onto $\mathcal{H}_N$ using inner product in $L^2_\rho(X)$
- goal: estimate approximation error $\mathcal{E}(f_{\mathcal{H}})$ for this $f_{\mathcal{H}}$

Distorsion factor:

- identity function on bounded functions extends to

$$J : L^2_\mu(X) \to L^2_\rho(X)$$

- distorsion of ρ with respect to μ

$$D_{\rho\mu} = \|J\|$$

operator norm: how much ρ distorts the ambient measure μ

- reasonable assumption: distorsion is finite
- in general ρ not known, but ρ_X is known, so $D_{\rho\mu}$ can be computed

Weyl Law

• Weyl law on rate of growth of eigenvalues of the Laplacian (acting on functions vanishing on boundary of domain $X \subset \mathbb{R}^n$)

$$\lim_{\lambda \to \infty} \frac{N(\lambda)}{\lambda^{n/2}} = (2\pi)^{-n} B_n Vol(X)$$

B_n volume of unit ball in $\mathbb{R}^n$; $N(\lambda)$ number of eigenvalues (with multiplicity) up to λ

• Weyl law: Li–Yau version

$$\zeta_k \geq \frac{n}{n+2} 4\pi^2 \left(\frac{k}{B_n \, Vol(X)} \right)^{2/n}$$

P. Li and S.-T. Yau, *On the parabolic kernel of the Schrödinger operator*, Acta Math. 156 (1986), 153–201

• from this get a weaker estimate, using explicit volume B_n

$$\zeta_k \geq \left(\frac{k}{Vol(X)} \right)^{2/n}$$

Approximation Error and Weyl Law

- norm $\|\cdot\|_K$: for $f = \sum_{k=1}^{\infty} c_k \phi_k$ with ϕ_k eigenfunctions of $-\Delta$

$$\|f\|_K := \left(\sum_{k=1}^{\infty} c_k^2 \zeta_k \right)^{1/2}$$

like L^2-norm but weighted by eigenvalues of Laplacian in ℓ^2 measure of $c = (c_k)$

- Approximation Error Estimate: for $\mathcal{H}$ and $f_{\mathcal{H}}$ as above

$$\mathcal{E}(f_{\mathcal{H}}) \leq D_{\rho\mu}^2 \left(\frac{k}{Vol(X)} \right)^{2/n} \|f_\rho\|_K^2 + \sigma_\rho^2$$

- proved using Weyl law and estimates

$$\|f_\rho - f_{\mathcal{H}}\|_\rho = d_\rho(f_\rho, \mathcal{H}_N) \leq \|J\| \, d_\mu(f_\rho, \mathcal{H}_N)$$

$$d_\mu(f_\rho, \mathcal{H}_N)^2 = \| \sum_{k=N+1}^{\infty} c_k \phi_k \|_\mu^2 = \sum_{k=N+1}^{\infty} c_k^2 = \sum_{k=N+1}^{\infty} c_k^2 \zeta_k \frac{1}{\zeta_k} \leq \frac{1}{\zeta_{N+1}} \|f_\rho\|_K^2$$

where $f_\rho = \sum_k c_k \phi_k$

Solution of the bias-variance problem

- mimimize $\mathcal{E}(f_{\mathcal{H},z})$ by minimizing both sample error and approximation error
- minimization as a function of $N \in \mathbb{N}$ (for the choice of hypothesis space $\mathcal{H} = \mathcal{H}_{N,R}$)
- select integer $N \in \mathbb{N}$ that minimizes $\mathcal{A}(N) + \epsilon(N)$ where $\epsilon = \epsilon(N)$ as in previous estimate of sample error and

$$\mathcal{A}(N) = D_{\rho\mu}^2 \left(\frac{k}{Vol(X)} \right)^{2/n} \|f_\rho\|_K^2 + \sigma_\rho^2$$

- from previous relation between m, $R = \|f_\rho\|_\infty$, δ and ϵ obtain

$$\epsilon - \frac{288M^2}{m} \left(N \log(\frac{96RM}{\epsilon}) + 1 + \log(\frac{1}{\delta}) \right) \geq 0$$

find N that minimizes ϵ with this constraint

- no explicit closed form solution for N minimizing $\mathcal{A}(N) + \epsilon(N)$ but can be estimated numerically in specific cases

Applications of Cucker–Smale theory of learning in neuroscience and in machine learning

- Tomaso A. Poggio and Fabio Anselmi, *Visual Cortex and Deep Networks*, MIT Press, 2016
- V. Maiorov, *Approximation by neural networks and learning theory*, Journal of Complexity 22 (2006) 102–117
- Tomaso A. Poggio and Steve Smale, *The Mathematics of Learning: Dealing with Data*, Notices AMS, May 2023, 537–544

What is the nature of the representations learned by deep networks?

"An essential ingredient of ergo-learning strategy is a search for symmetry--repetitive patterns--in flows of signals. Even more signicantly, an ergo system creates/identies such patterns by reducing/compressing "information" and by structuralizing "redundancies" in these flows." --Gromov

A connection between category theory & machine learning?