Artificial Intelligence : logic, symbolic sequential reasoning

Machine Learning : inductive inference, learning from data, learning from examples

Nowadays, AI comprises both

MACHINE LEARNING

Statistical learning

- strong mathematical foundations
- a lot of theorems
- no biological motivations
- relatively "simple" models

e.g.:
Support Vector Machines

Cybernetics
Connectionism,
Neural Networks,
Deep Learning

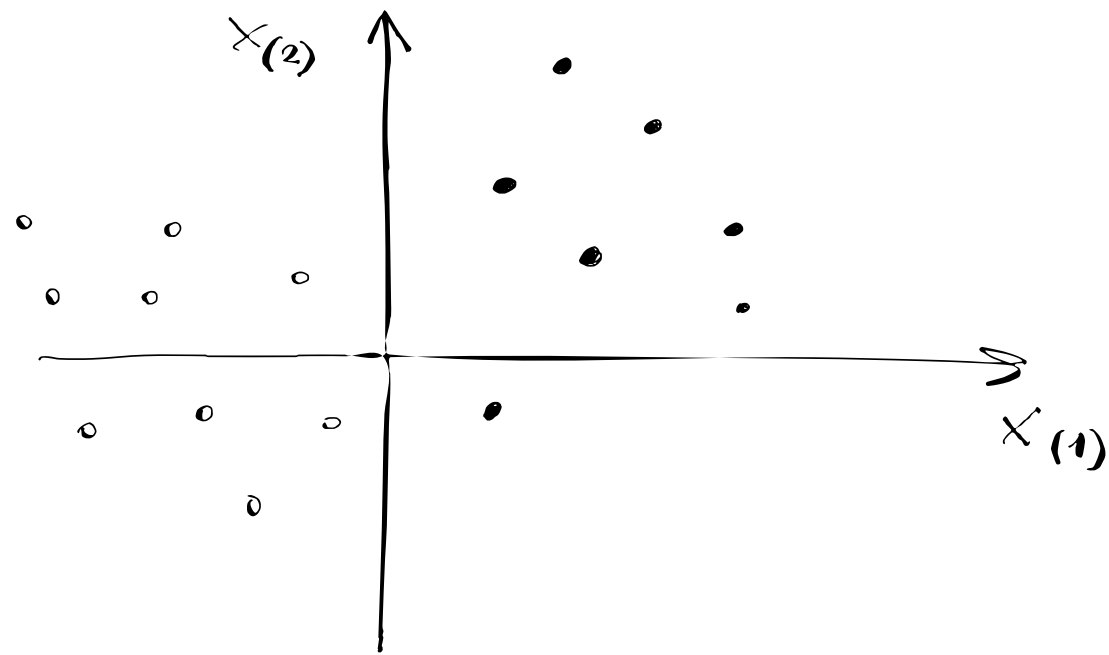
- biologically motivated
- few theorems
- relatively complex models

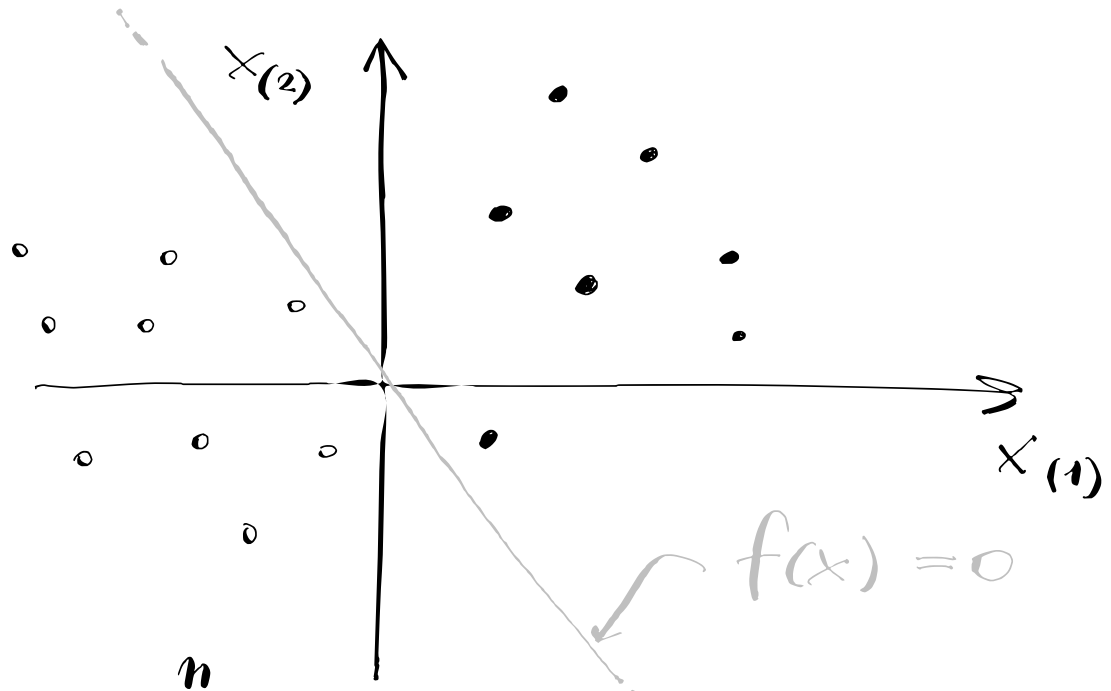
e.g.
Artificial Neural Networks

simplest problem: binary classification

$$T = \{(x_i, y_i) \quad i = 1 \dots L\}$$

$$x_i = \mathbb{R}^n, \quad y_i \in \{-1, +1\}$$





$$f(x) = \sum_{i=1}^{n} w_i x_{(i)} + b = W^T x + b$$

$$y = \text{sign}\left(f(x)\right) = \ominus\left(f(x)\right)$$

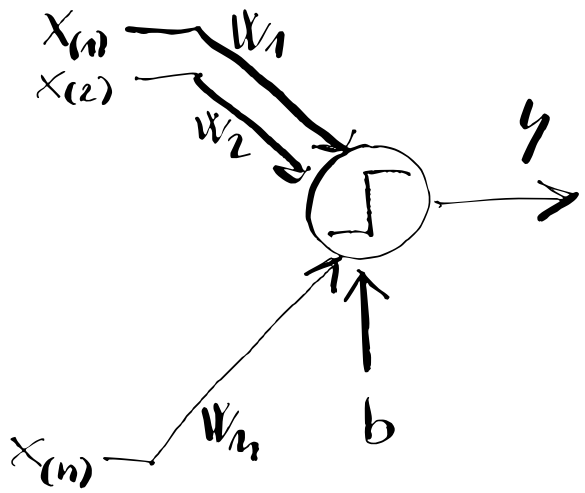
threshold function

STATISTICAL LEARNING

$f(x) = W^T x + b$

choose W and b
OPTIMALLY

CYBERNETICS



"neuron" (McCulloch & Pitts, 1943)

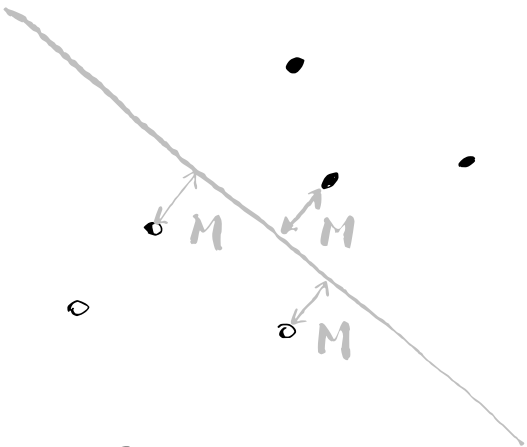
adjust W and b
ITERATIVELY

STATISTICAL LEARNING

optimality criterion:

MAXIMIZE THE MARGIN

(distance to the closest example)



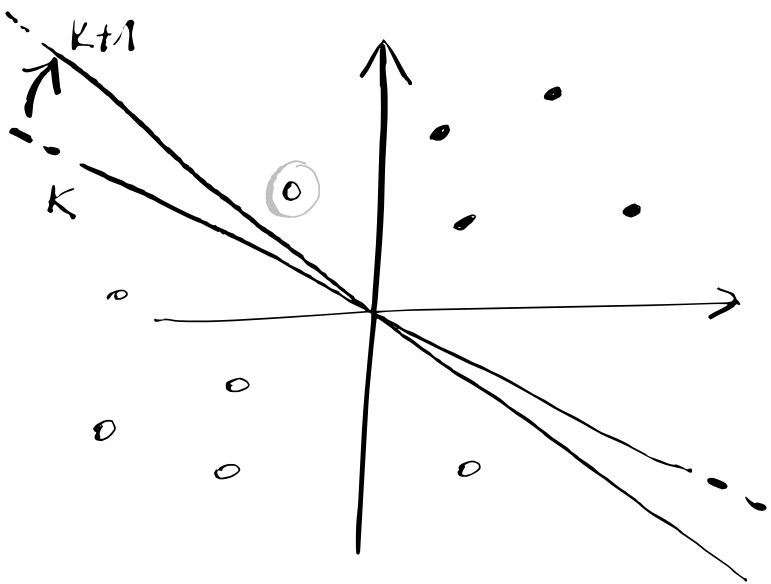
QP problem

Boser & Vapnik, 1992

CYBERNETICS, NN

perceptron's

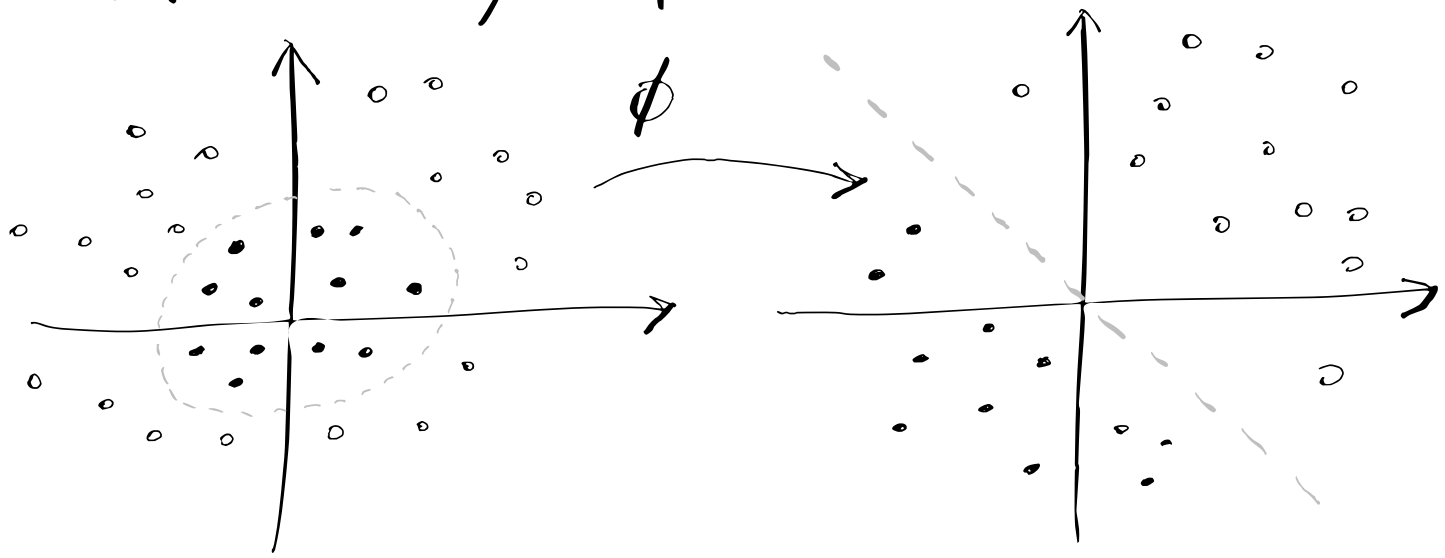
training algorithm



Rosenblatt, 1958

Novikoff, 1962

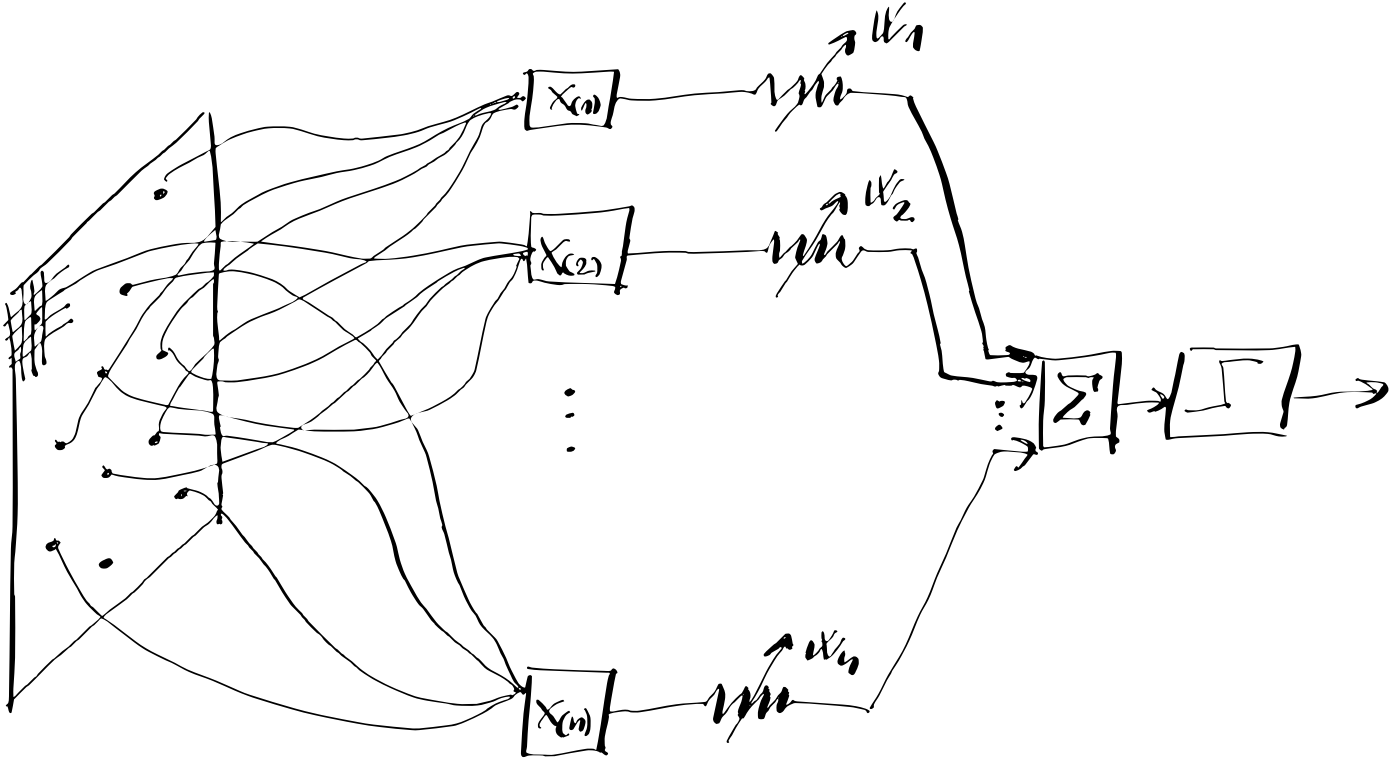
Non linearly - separable data



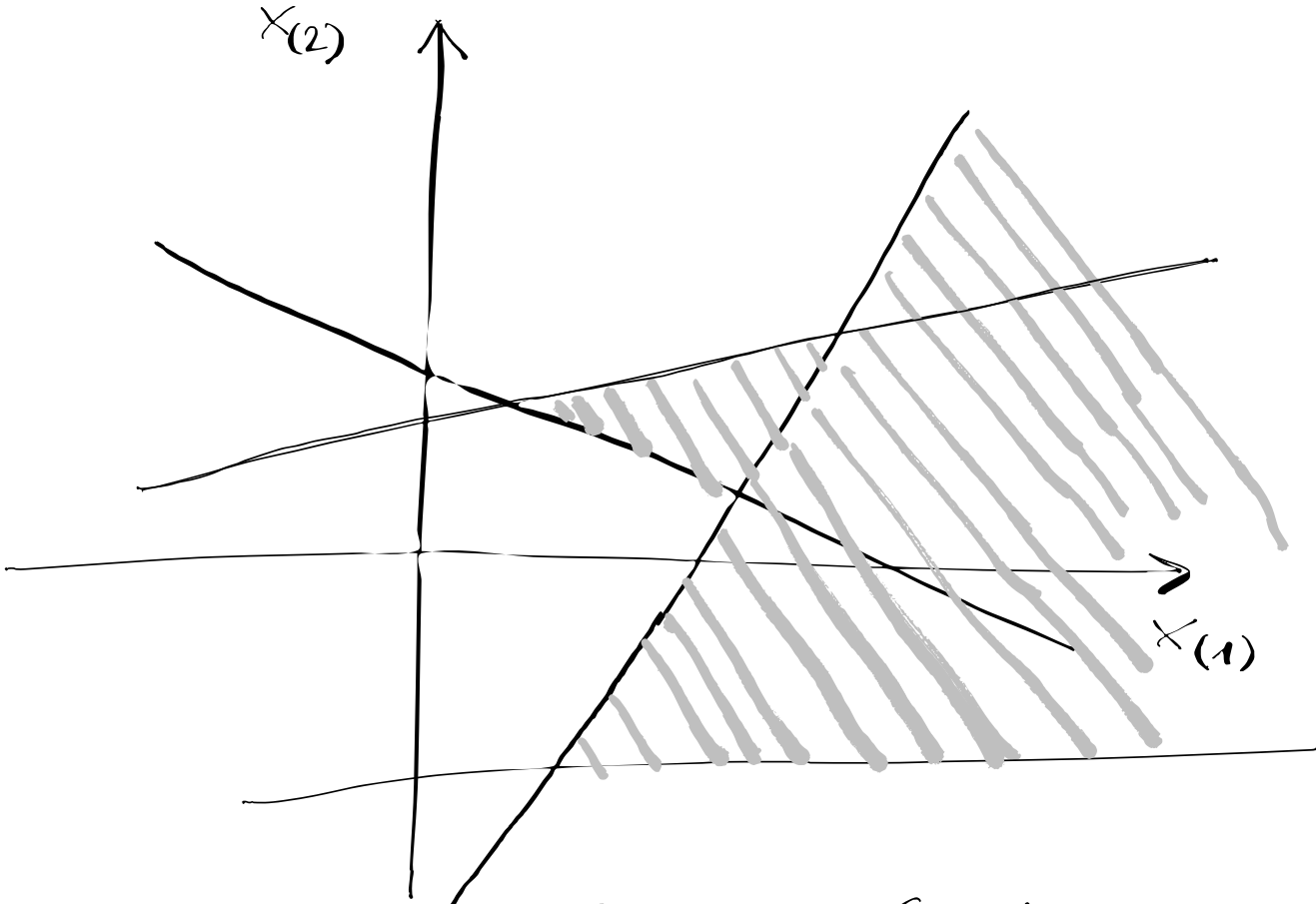
INPUT SPACE

FEATURE SPACE

ϕ: FEATURE MAPPING



RANDOM (!)
CONNECTIONS realize a feature mapping.



non linear decision functions (in the input space) are possible

In 1969, Minsky & Papert point out some limitations of the perceptron (Minsky & Papert, "Perceptrons", 1969).

— logical XOR

Frank Rosenblatt dies in 1971.

NNs are abandoned for a decade

STATISTICAL LEARNING

FIXED feature map
(is a design parameter)

e.g., RBF map,

polynomial map

⇓

EFFICIENT TRAINING IN
∞ - DIMENSIONAL SPACE,
IMPRESSIVE RESULTS IN
MANY TASKS
('90)

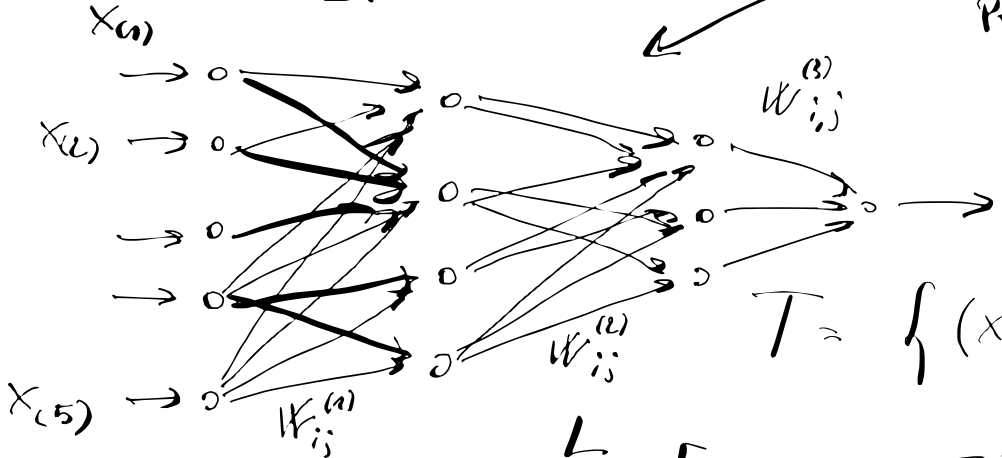
CONNECTIONISM

LEARNED feature map
(Rumelhart et al, 1986)

"ERROR BACKPROPAGATION"

BACKPROP

MULTI LAYER PERCEPTRON



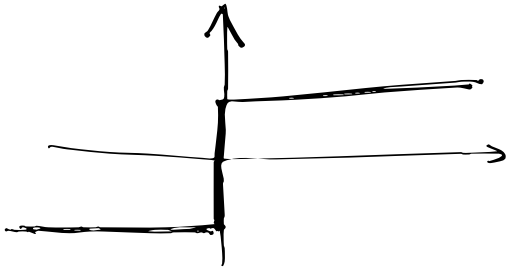
$$T = \{(x_i, y_i) \quad i = 1 \dots L\}$$

empirical error = (6st function)

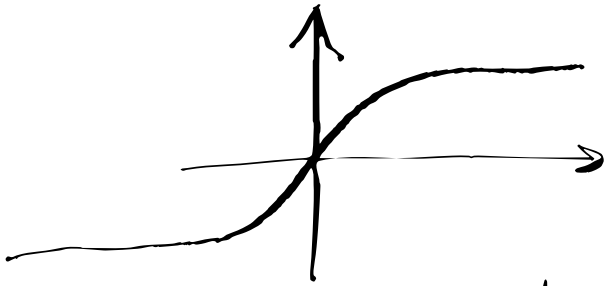
$$\sum_{i=1}^{L} \left[f(x_i) - y_i \right]^2$$

composition of functions

BACKPROP = apply GRADIENT DESCENT to the empirical error; compute the gradient using the CHAIN RULE.

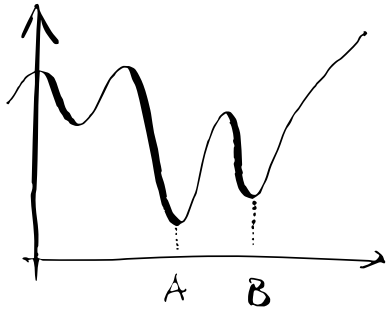


threshold



sigmoid
(differentiable)

problems with backprop: e.g. tanh

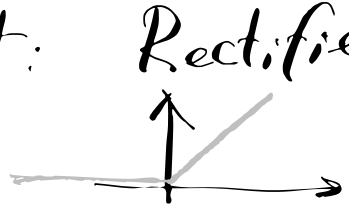


① (multiple bad minima)

② vanishing gradient

③ computationally demanding for large datasets

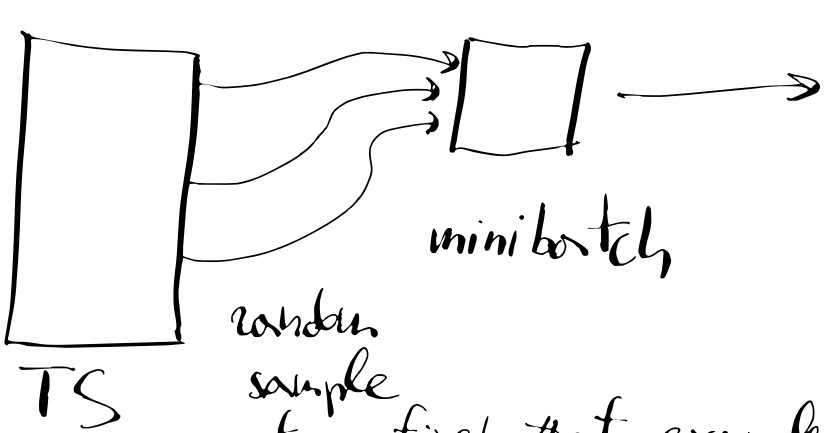
- solution to vanishing gradient: Rectified Linear Unit (ReLU)



[non-differentiability is not a problem because the goal is NOT to arrive to a null gradient point, thus undefined gradient is acceptable]

- solution to computational burden:

Stochastic Gradient Descent (SGD)



Nowadays connectionism is known as

DEEP LEARNING

enabling factors :
- lot of data
- lot of computational power (GPUs)
- some improvements in training (ReLU, dropout, batch normalization)

KEY OBSERVATION :
depth is essential because it allows the extraction of a hierarchy of features

PROCESSING SEQUENCES

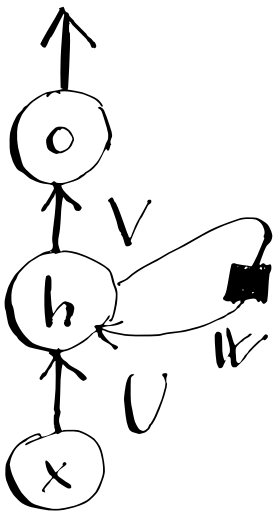
- 1D CNNs

- RECURRENT NNs (backprop through time)
  (extreme form of parameter sharing)

RNN

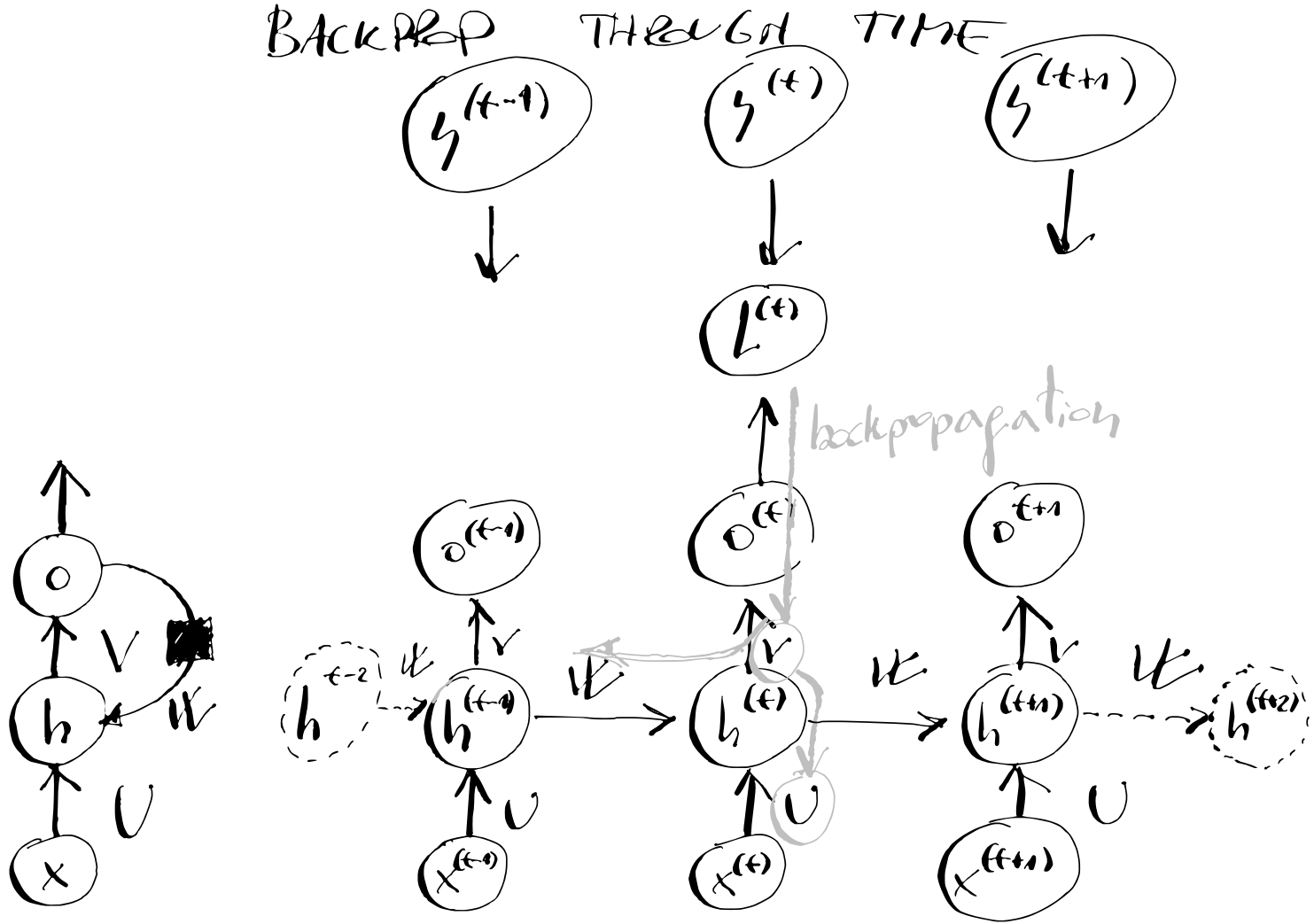
$x^{(1)}, x^{(2)}, \cdots x^{(t)}$ input sequence

$h^{(t)}$ state, $o^{(t)}$ atpt



$$\begin{cases} h^{(t)} = f\left(W h^{(t-1)} + b + U x^{(t)}\right) \\ o^{(t)} = V h^{(t)} \end{cases}$$

Dynamic system


BACKPROP THROUGH TIME
backpropagation

EFFECTIVE VARIANT:

Long Short-Term Memory Network

(LSTM)