GRAPH REPRESENTATIONS, BACKPROPAGATION AND BIOLOGICAL PLAUSIBILITY

Marco Gori
SAILAB, University of Siena

NeurIPS 2019

OUTLINE

- Learning in structured domains
- Diffusion machines and spatiotemporal locality
- Backpropagation diffusion and biological plausibility

LEARNING IN STRUCTURED DOMAINS

NeurIPS 2019

Graphs as Pattern Models

physicochemical behavior

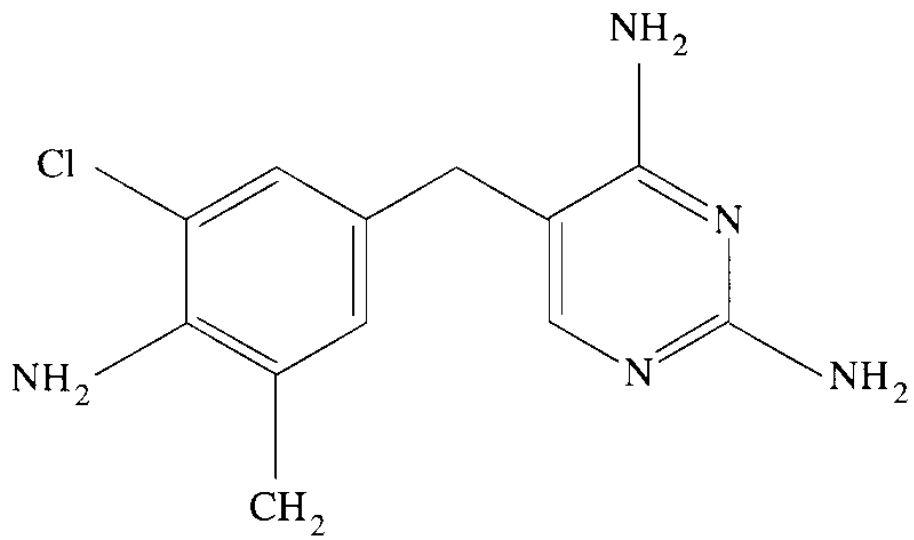
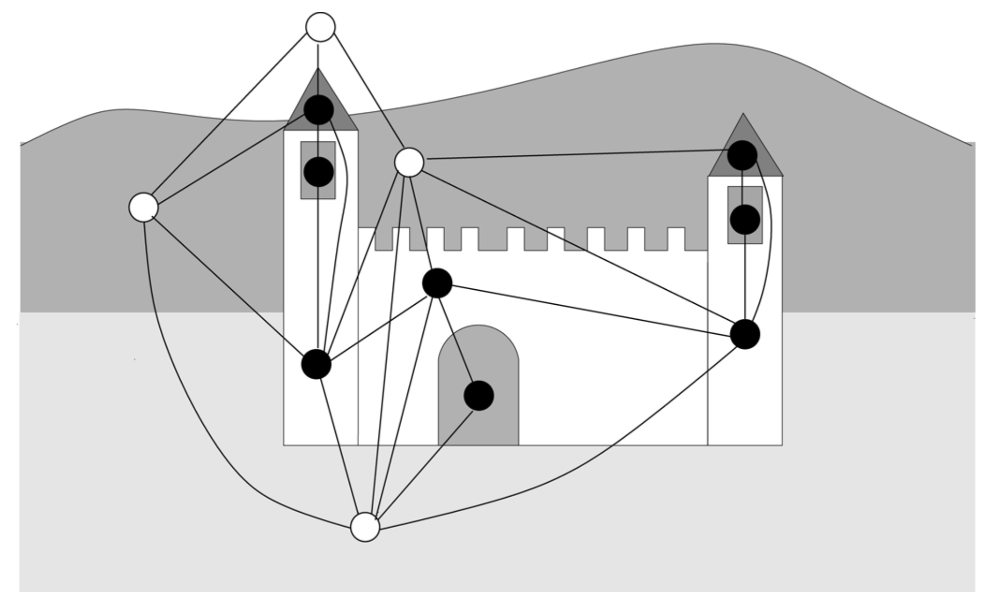


image classification

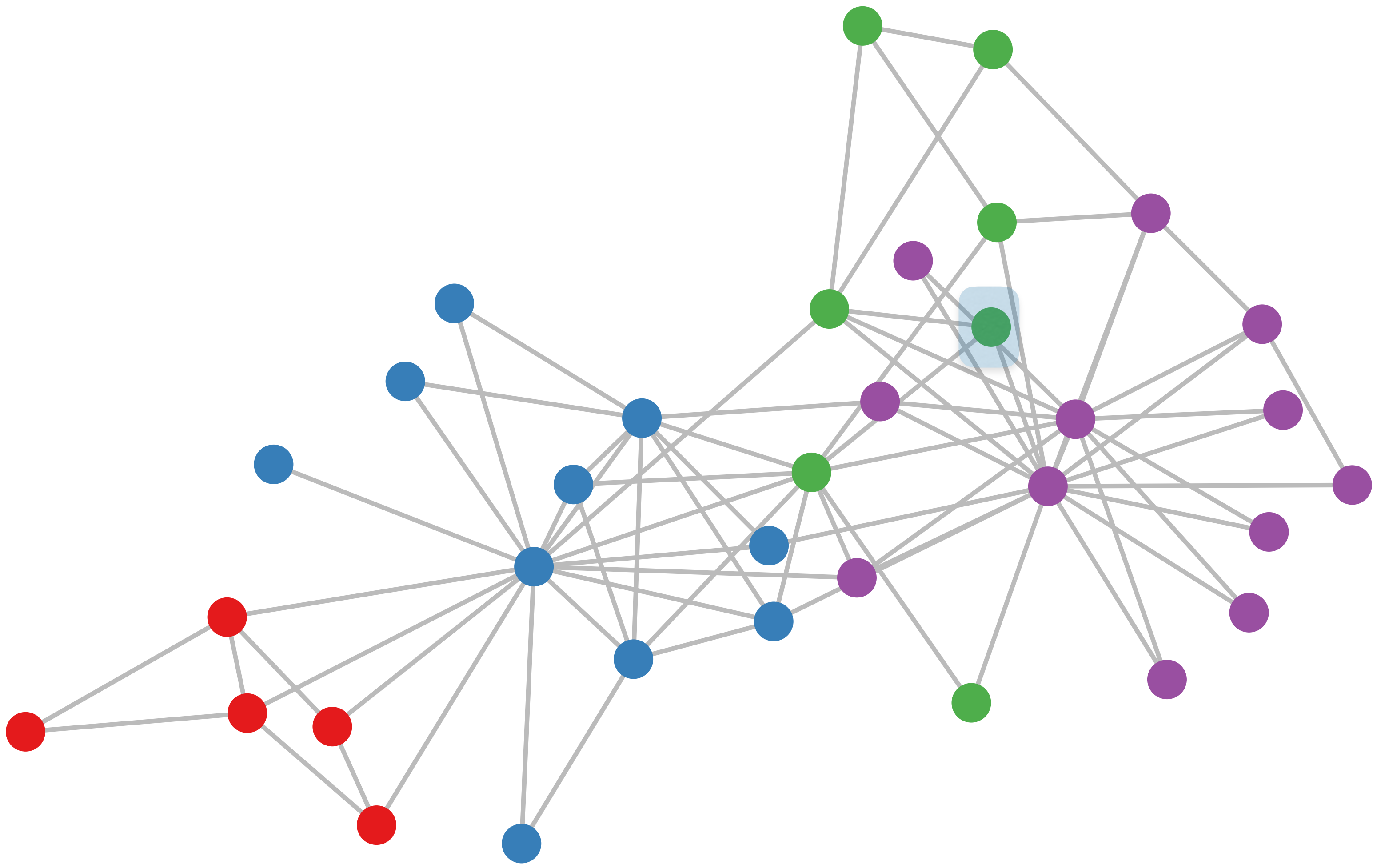


What are the features?

Social nets

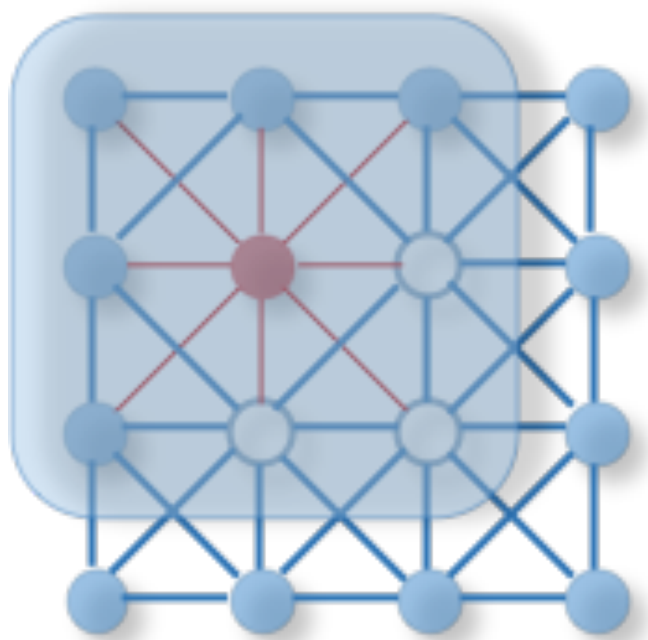
here we need to make prediction at node level!

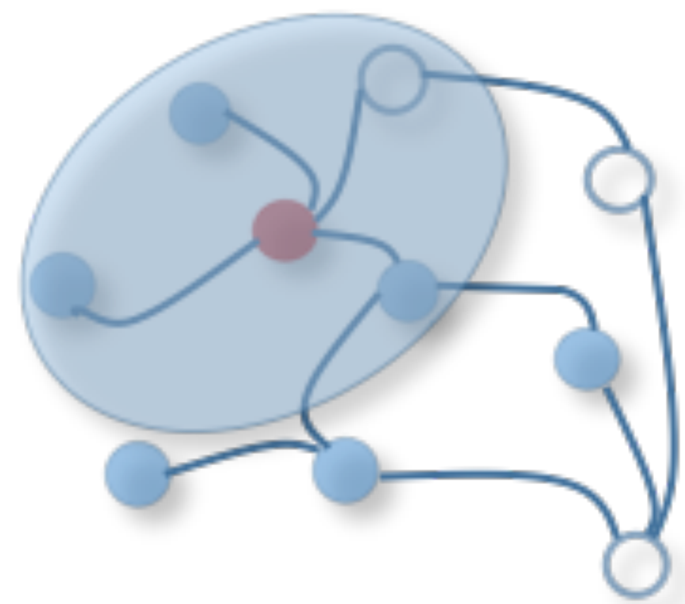
quasi-eqlibrium dynamic models



GRAPH NEURAL NETS

popular and successful mostly thanks to
graph convolutional networks





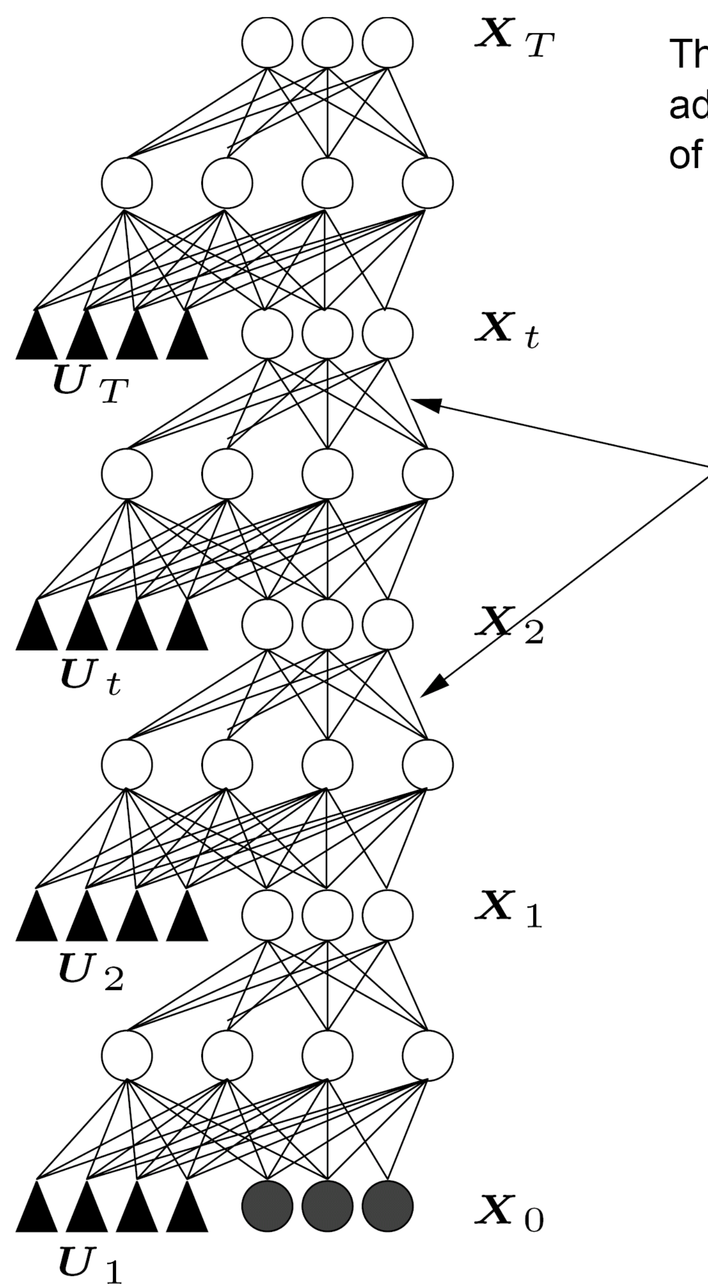
pictures from Z. Wu et al

Non-Euclidean Deep Learning

NeurIPS 2019

HISTORICALLY …
ANOTHER PATH WAS FOLLOWED!

Extension of the idea of time unfolding …



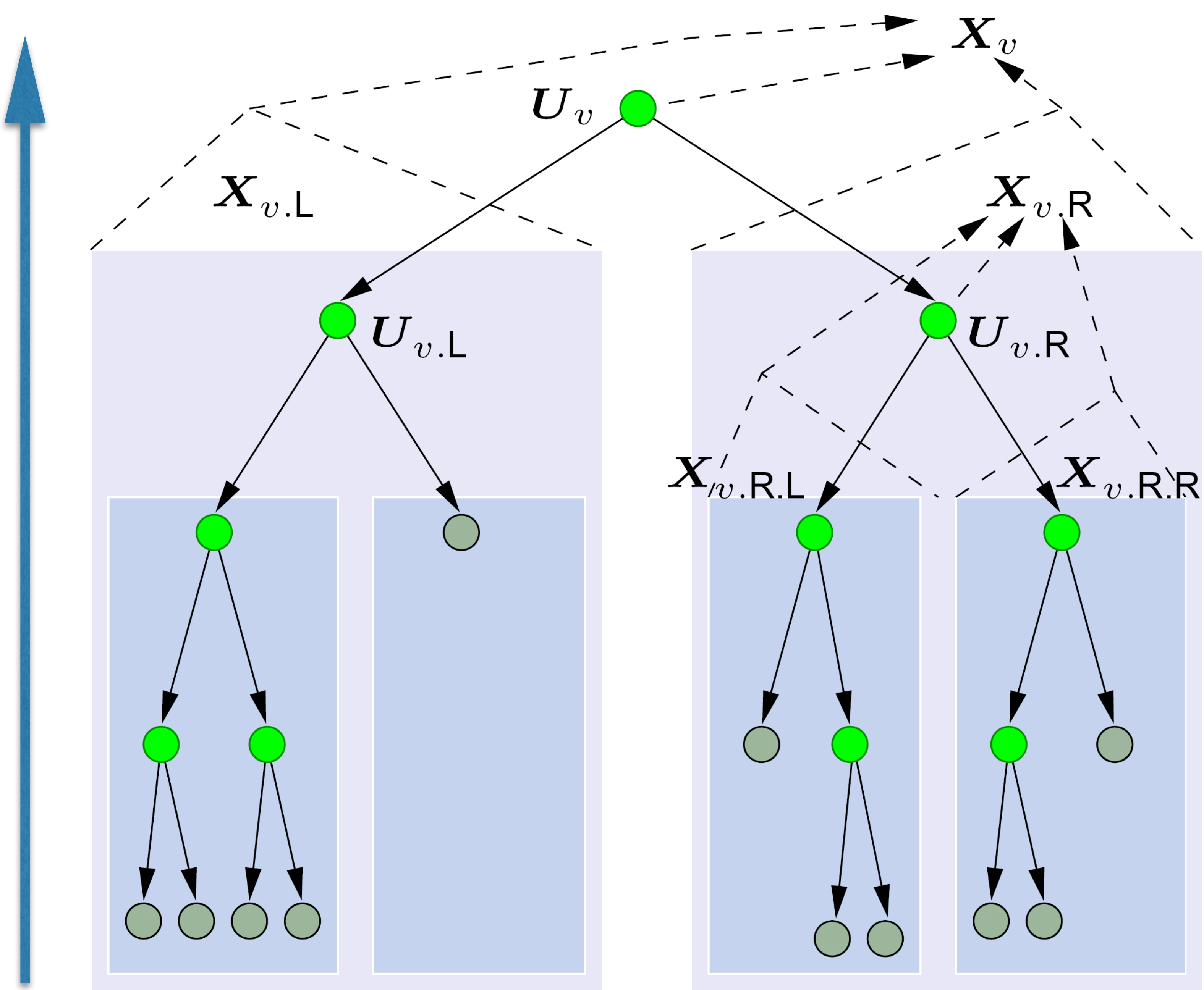
The last state is an adaptive encoding of the whole sequence

Weights are shared (replicated)

The encoding net has a feedforward structure: Gradient can be computed by Backpropagation (through time)

Structure unfolding

The case of binary trees …

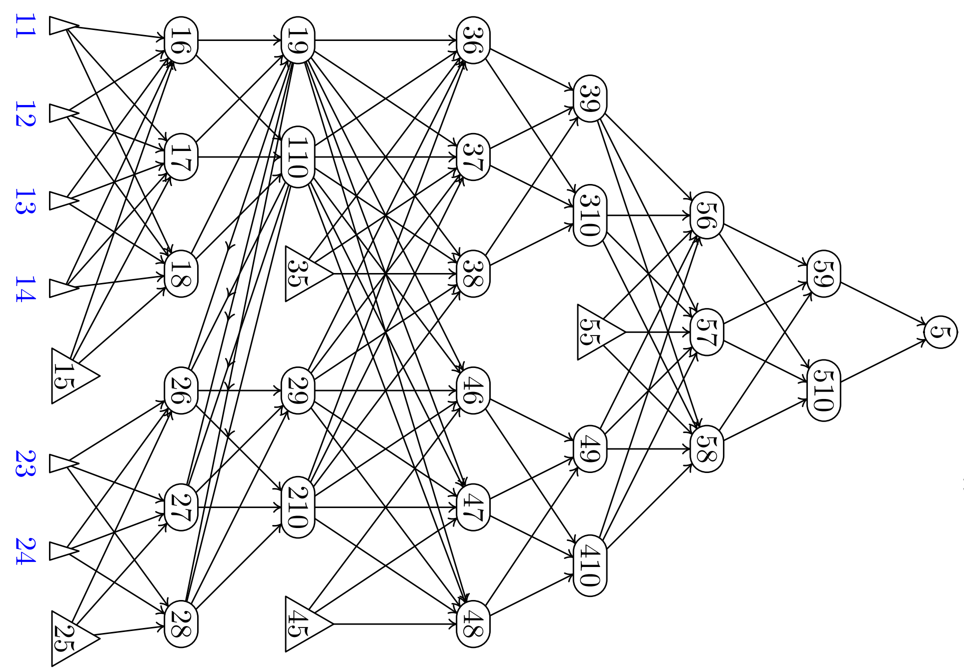


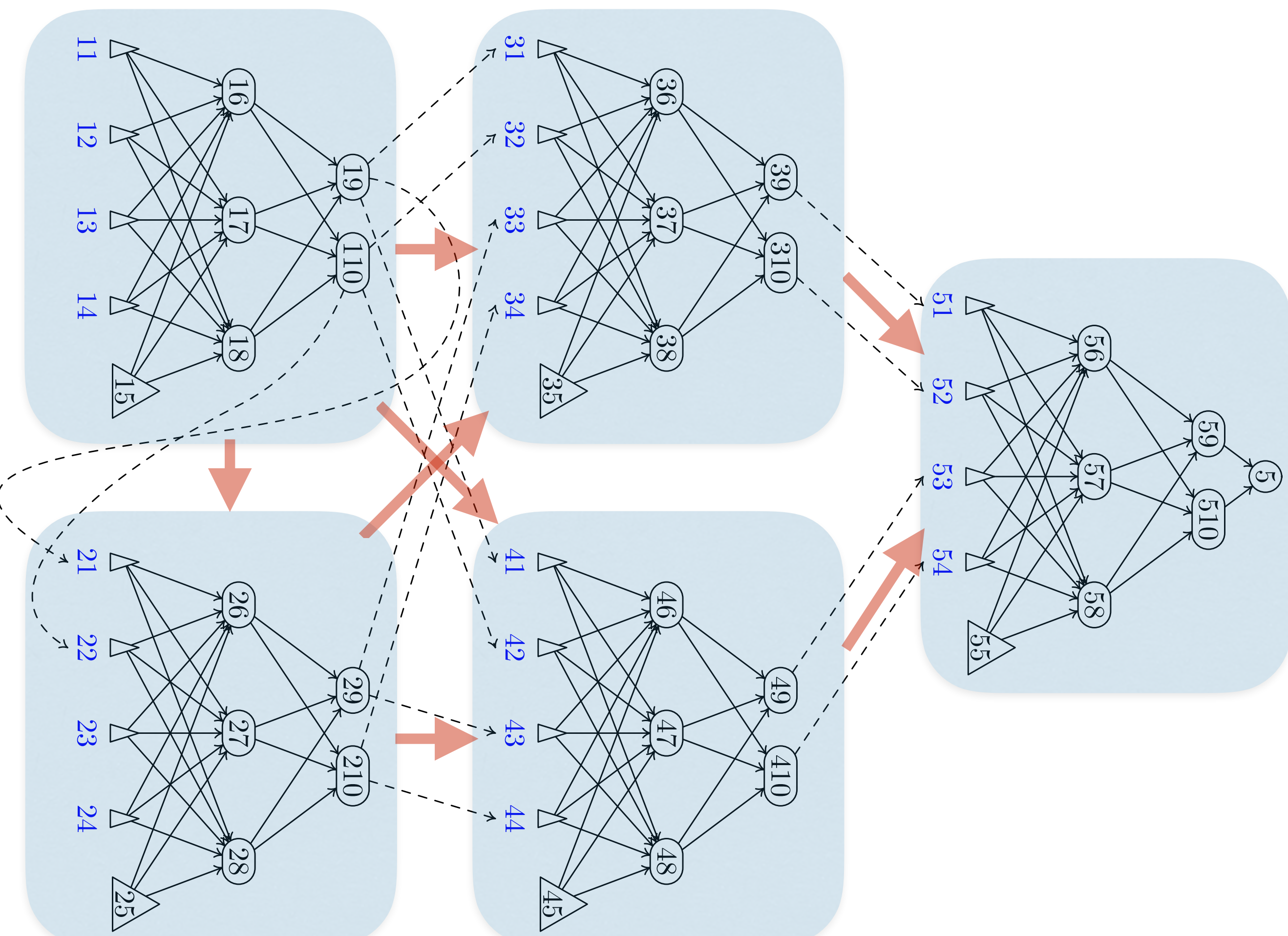
frontier state if v.R is external

$$X_v = f(U_v, X_{v.\mathsf{L}}, X_{v.\mathsf{R}})$$

frontier state if v.L is external

Graph Compiling …





25

A recurrent net arises from cyclic graphs

The Graph Neural Network Model

Gori et al IJCNN 2005, 2009 IEEE-TNN

LEARNING AS A DIFFUSION PROCESS

THE FRAMEWORK OF CONSTRAINED-BASED LEARNING
AND THE ROLE OF TIME COHERENCE

NeurIPS 2019

Natural Laws of Learning
The links with mechanics

Once we believe in ergodicity …
there is no distinction between training and test sets!

laws of learning

regularization term   Loss function of neural net

$$\mathcal{A}_\epsilon = \int_0^T dt\; e^{-t/\epsilon}\left(\frac{1}{2}\epsilon^2\rho\ddot{q}^2 + \frac{1}{2}\epsilon\nu\dot{q}^2 + V(q,t)\right)$$

kinetic energy   potential energy

laws of mechanics

NeurIPS 2019

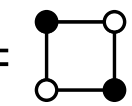
Natural Laws of Cognition: A Pre-Algorithmic Step

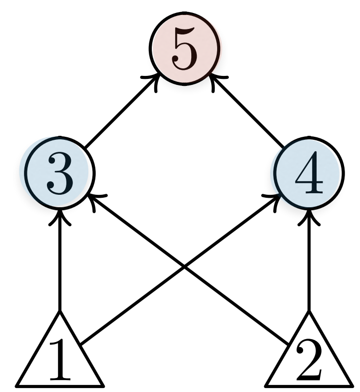
Natural Learning Theory $\leadsto$ Mechanics	**Remarks**
$w_i \leadsto q_i$	Weights are interpreted as generalized coordinates.
$\dot{w}_i \leadsto \dot{q}_i$	Weights variations are interpreted as generalized velocities.
$\upsilon_i \leadsto p_i$	The conjugate momentum to the weights is defined by using the machinery of Legendre transforms.
$A(w) \leadsto S(q)$	The cognitive action is the dual of the action in mechanics.
$F(t, w, \dot{w}) \leadsto L(t, q, \dot{q})$	The Lagrangian F is associated with the classic Lagrangian L in mechanics.
$H(t, w, \upsilon) \leadsto H(t, q, p)$	When using w and υ, we can define the Hamiltonian, just like in mechanics.

weights of neural net as position particles

Constraint Reactions

architectural and environmental constraints

$$\mathcal{L} = \{((0,0),0),((0,1),1),((1,0),1),((1,1),0)\} =$$ 



1 2

$$\mathsf{y}((0,1),(1,0))$$
$$\neg\mathsf{y}((0,0),(1,1)).$$

4 3

"hard" architectural constraints

$$x_{\kappa 3} - \sigma(w_{31}x_{\kappa 1} + w_{32}x_{\kappa 2} + b_3) = 0$$
$$x_{\kappa 4} - \sigma(w_{41}x_{\kappa 1} + w_{42}x_{\kappa 2} + b_4) = 0 \qquad \kappa = 1,2,3,4$$
$$x_{\kappa 5} - \sigma(w_{53}x_{\kappa 3} + w_{54}x_{\kappa 4} + b_4) = 0$$

training set constraints

$$x_{15} = 1,\ x_{25} = 1,\ x_{35} = 0,\ x_{45} = 0$$

Lagrangian Approach

Lagrangian Multipliers

Static Models → holonomic constraints

Dynamic Models → non-holonomic constraints

functional optimization:
variational calculus under subsidiary conditions

Formulation of Learning

holonomic constraints (DAGs)

regularization term   risk function

$$\mathscr{A}(x,W) := \int \frac{1}{2}(m_x|\dot{x}(t)|^2 + m_W|\dot{W}(t)|^2)\,\varpi(t)dt + \mathscr{F}(x,W)$$

$$\mathscr{F}(x,W) := \int F(t,x,\dot{x},\ddot{x},W,\dot{W},\ddot{W})\,dt$$

$$G^j(t,x(t),W(t)) = 0, \qquad 1 \le j \le \nu$$

neural constraints (Einstein's notation)

$$G^j(\tau,\xi,M) := \begin{cases} \xi^j - e^j(\tau), & \text{if } 1 \le j \le \omega; \\ \xi^j - \sigma(m_{jk}\xi^k) & \text{if } \omega < j \le \nu, \end{cases}$$

Proposition 1: Functionally independent for acyclic graphs feedforward nets

Formulation of Learning (con't)

holonomic constraints - any digraph

regularization term — risk function

$$\mathscr{A}(x,W,s) := \underbrace{\int \frac{1}{2}\left(m_x|\dot{x}(t)|^2 + m_W|\dot{W}(t)|^2 + m_s|\dot{s}(t)|^2\right)\varpi(t)dt}_{\text{regularization term}} + \underbrace{\mathscr{F}(x,W,s)}_{\text{risk function}},$$

$$\mathscr{F}(x,W,s) := \int F(t,x,\dot{x},\ddot{x},W,\dot{W},\ddot{W},s)\,dt.$$

neural constraints

slack variables

$$G^j(\tau,\xi,M,\zeta) := \begin{cases} \xi^j - e^j(\tau) + \boxed{\zeta^j}, & \text{if } 1 \le j \le \omega; \\ \xi^j - \sigma(m_{jk}\xi^k) + \boxed{\zeta^j} & \text{if } \omega < j \le \nu. \end{cases}$$

Proposition 2: Functionally independent for any graph

Formulation of Learning (con't)

Non-holonomic constraints (any digraph)

regularization term loss term

$$\mathscr{A}(x, W) = \int \Big(\underbrace{\frac{m_x}{2}|\dot{x}(t)|^2 + \frac{m_W}{2}|\dot{W}(t)|^2}_{\text{regularization term}} + \underbrace{F(t, x, W)}_{\text{loss term}} \Big)\, \varpi(t)\, dt$$

neural constraints

$$\dot{x}^i(t) + cx^i(t) - \sigma(w_{ik}(t)x^k(t)) = 0 \qquad 0 < c < 1$$

Proposition 3: Functionally independent for any graph

Feedforward Networks (DAGs)

$$-m_x\varpi(t)\ddot{x}(t) - m_x\dot{\varpi}(t)\dot{x}(t) - \lambda_j(t)G^j_\xi(x(t),W(t)) + L^x_F(x(t),W(t)) = 0;$$

$$-m_W\varpi(t)\ddot{W}(t) - m_W\dot{\varpi}(t)\dot{W}(t) - \lambda_j(t)G^j_M(x(t),W(t)) + L^W_F(x(t),W(t)) = 0$$

$$\Big(\frac{G^i_{\xi^a}G^j_{\xi^a}}{m_x} + \frac{G^i_{m_{ab}}G^j_{m_{ab}}}{m_W}\Big)\lambda_j = \varpi\big(G^i_{\tau\tau} + 2(G^i_{\tau\xi^a}\dot{x}^a + G^i_{\tau m_{ab}}\dot{w}_{ab} + G^i_{\xi^a m_{bc}}\dot{x}^a\dot{w}_{bc}) + G^i_{\xi^a\xi^b}\dot{x}^a\dot{x}^b + G^i_{m_{ab}m_{cd}}\dot{w}_{ab}\dot{w}_{cd}\big) - \dot{\varpi}(\dot{x}^a G^i_{\xi^a} + \dot{w}_{ab}G^i_{m_{ab}}) + \frac{L^{x^a}_F G^i_{\xi^a}}{m_x} + \frac{L^{w_{ab}}_F G^i_{m_{ab}}}{m_W}$$

instantaneous linear equation

$$L^x_F = F_x - d(F_{\dot{x}})/dt + d^2(F_{\ddot{x}})/dt^2,\ L^W_F = F_W - d(F_{\dot{W}})/dt + d^2(F_{\ddot{W}})/dt^2$$

supervised learning

$$F(t,x,\dot{x},\ddot{x},W,\dot{W},\ddot{W}) = F(t,x) \to L^x_F = \partial_x F, \quad L^w_F = 0$$

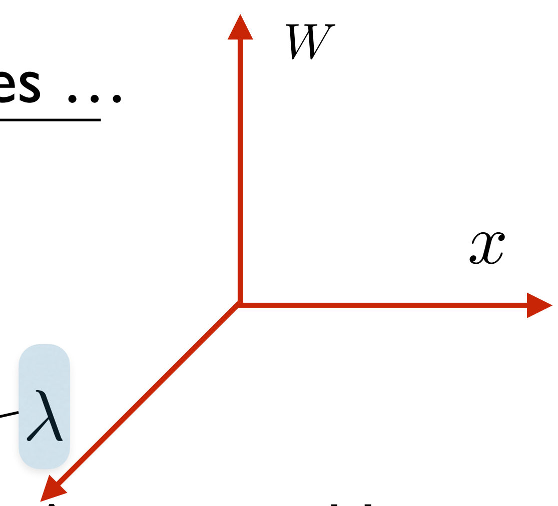
Reduction to Backpropagation

$m_x \to 0$

the chain rule arises ...

$$\dot{W}_{ij} = -\frac{1}{\gamma}\sigma'(w_{ik}x^k)\delta_i x^j;$$

$$G^i_{\xi^a} G^j_{\xi^a} \delta_j = -V_{x^a} G^i_{\xi^a},$$



Augmented Learning Space

$$T\delta = -V_x$$

$$T = \begin{pmatrix} 1 & -\sigma'(w_{21}x^1)w_{21} & 0 \\ 0 & 1 & -\sigma'(w_{32}x^2)w_{32} \\ 0 & 0 & 1 \end{pmatrix}$$

$$\delta_3 = -V_{x^3};$$
$$\delta_2 = \sigma'(w_{32}x^2)w_{32}\delta_3;$$
$$\delta_1 = \sigma'(w_{21}x^1)w_{21}\delta_2.$$

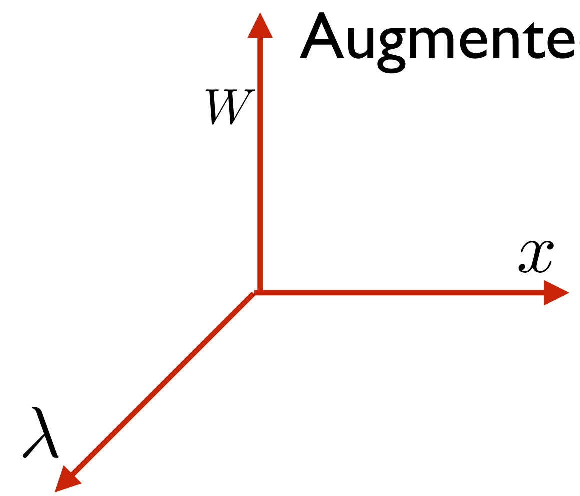
A somewhat surprising kinship with the BP delta-error

Early discovery by Yan Le Cun, 1989

Euler-Lagrange Equations

non-holonomic constraints

intuition: we need to store the multipliers
and provide temporal updating



$$\dot{x}^i(t) + cx^i(t) - \sigma(w_{ik}(t)x^k(t)) = 0;$$

BP-like GNN factorization $\delta_j x^i$

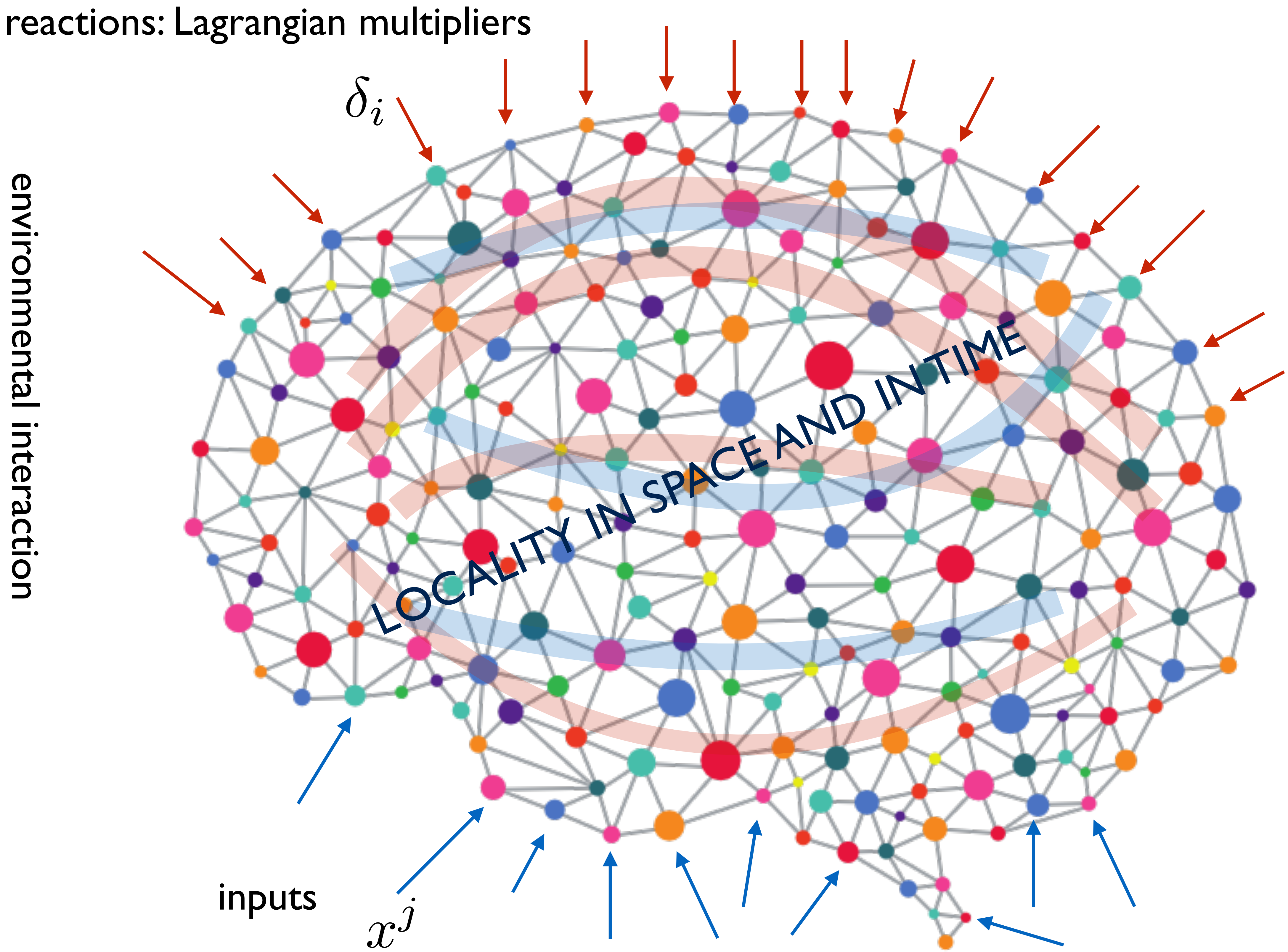
$$\dot{W}(t) = -\frac{1}{\gamma}\delta_j(t)G^j_M(t, x(t), W(t), \dot{x}(t))$$

$$\dot{\delta}(t) = \delta_j(t)G^j_\xi(t, x(t), W(t), \dot{x}(t)) + V_\xi(t, x(t))$$

This makes GNN efficient!

Unlike BPTT and RTRL, learning equations are local in space and time:
connections with Equilibrium Propagation (Y. Bengio et al)

DIFFUSION LEARNING
AND BIOLOGICAL PLAUSIBILITY



Biological Plausibility of Backpropagation

BP diffusion is biologically plausible

Biological concerns should not involve BP,

but the instantaneous map $x^i(t) = \sigma(w_{ik} x^k(t))$

replace with

$\downarrow$

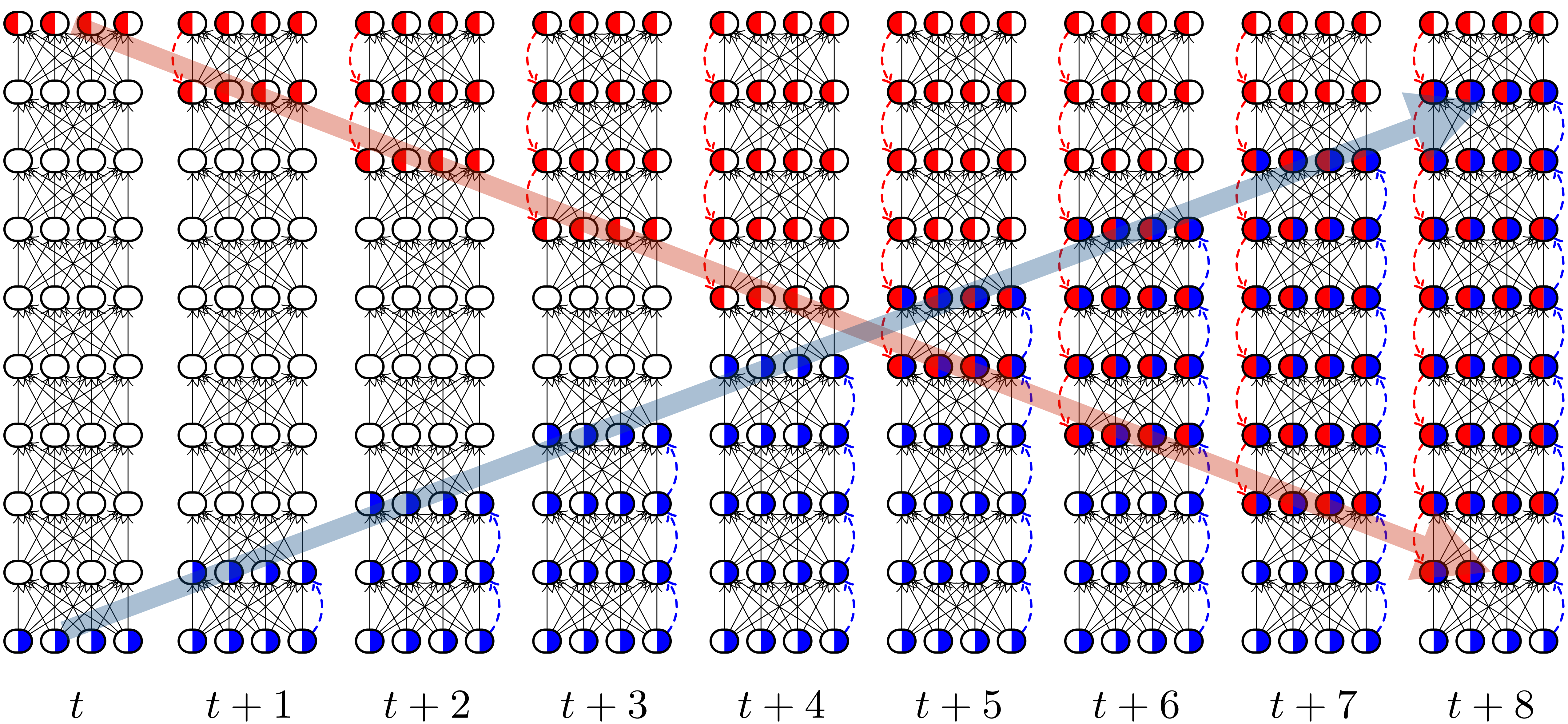
$$x^i(t) = \sigma(w_{ik}(t-1) x^k(t-1))$$
$$\dot{x}^i(t) + c x^i(t) - \sigma(w_{ik}(t) x^k(t)) = 0$$

… clever related comment by Francis Crick, 1989

BP algorithm is NOT biologically plausible

Forward and Backward Waves

BP diffusion is biologically plausible



BP algorithm is NOT biologically plausible

Conclusions

PRELIMINARY EXPERIMENTAL CHECK

- GNN: Success due to convolutional graphs, but the "diffusion path" is still worth exploring
- What happens with deep networks in graph compiling?
- Laws of learning, pre-algorithmic issues, and biological plausibility
- Dynamic models for Lagrangian multipliers (always delta-error): new perspective whenever time-coherence does matter!
- Euler-Lagrangian Learning and SGD

Acknowledgments

Alessandro Betti, SAILAB

Publications

- F. Scarselli et al, "The Graph Neural Network Model," IEEE-TNN, 2009
- A. Betti, M. Gori, and S. Melacci, Cognitive Action Laws: The Case of Visual Features, IEEE-TNNLS 2019
- A. Betti, M. Gori, and S. Melacci, Motion Invariance in Visual Environment, IJCAI 2019
- A. Betti and M. Gori, Backprop Diffusion is Biologically Plausible, arXiv:1912.04635
- A. Betti and M. Gori, Spatiotemporal Local Propagation, arXiv: 1907.05106

Software

Preliminary version

NeurIPS 2019

Machine Learning

A CONSTRAINT-BASED APPROACH

MORGAN KAUFMANN

Marco Gori