RECURRENT BACK PROPAGATION

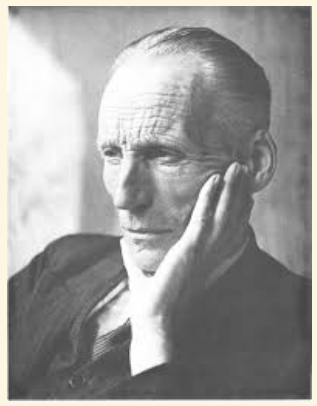
ITERATE THROUGH FIXED POINT

OUTLINE OF THIS LECTURE

- Fixed point theorems
- Fixed point iteration for root finding
- BPTT in 5 minutes or less
- Implicit function differentiation
- Almeida-Pineda algorithm (RBP)

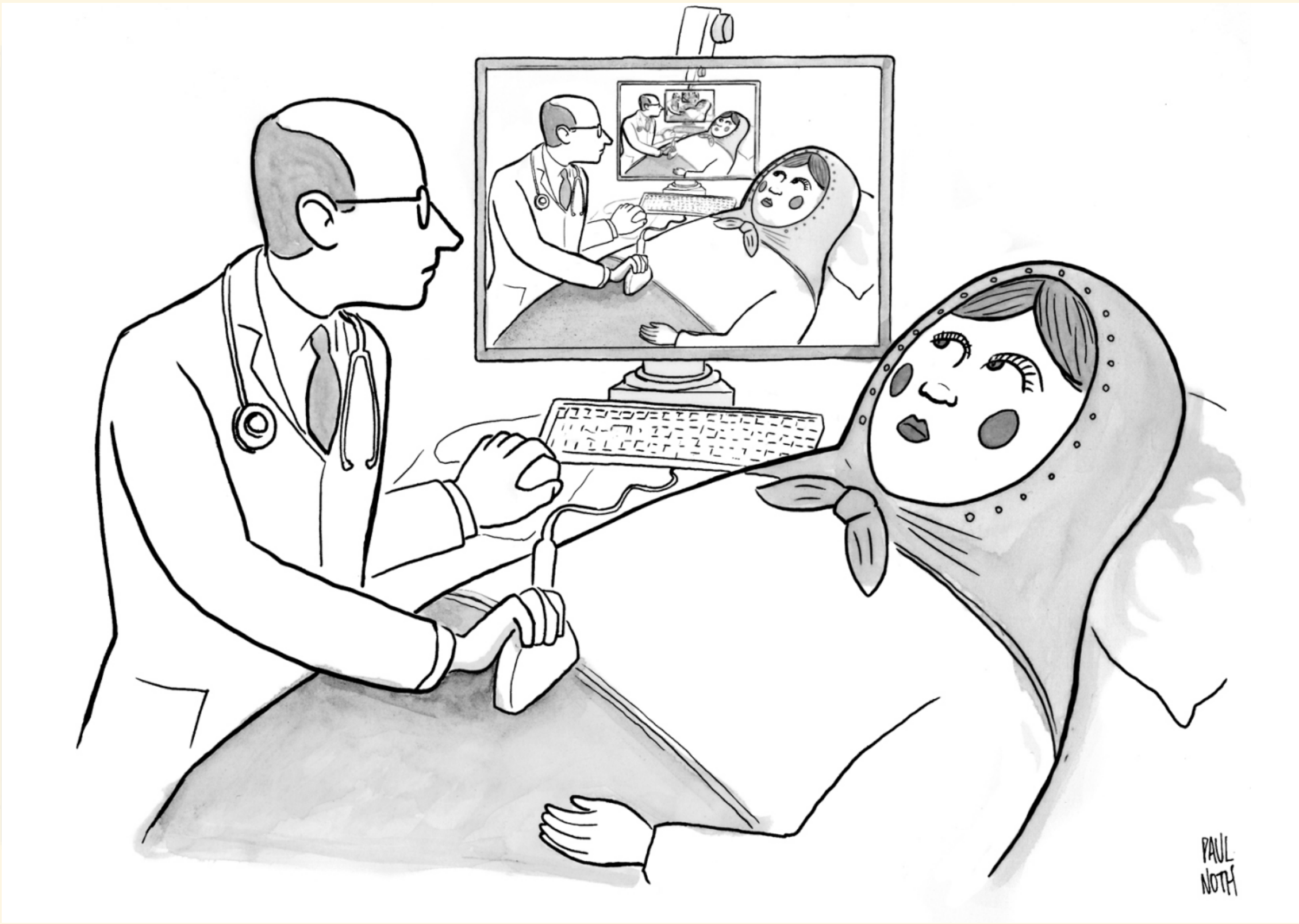
FIXED POINT THEOREMS

BROUWER'S FIXED-POINT THEOREM

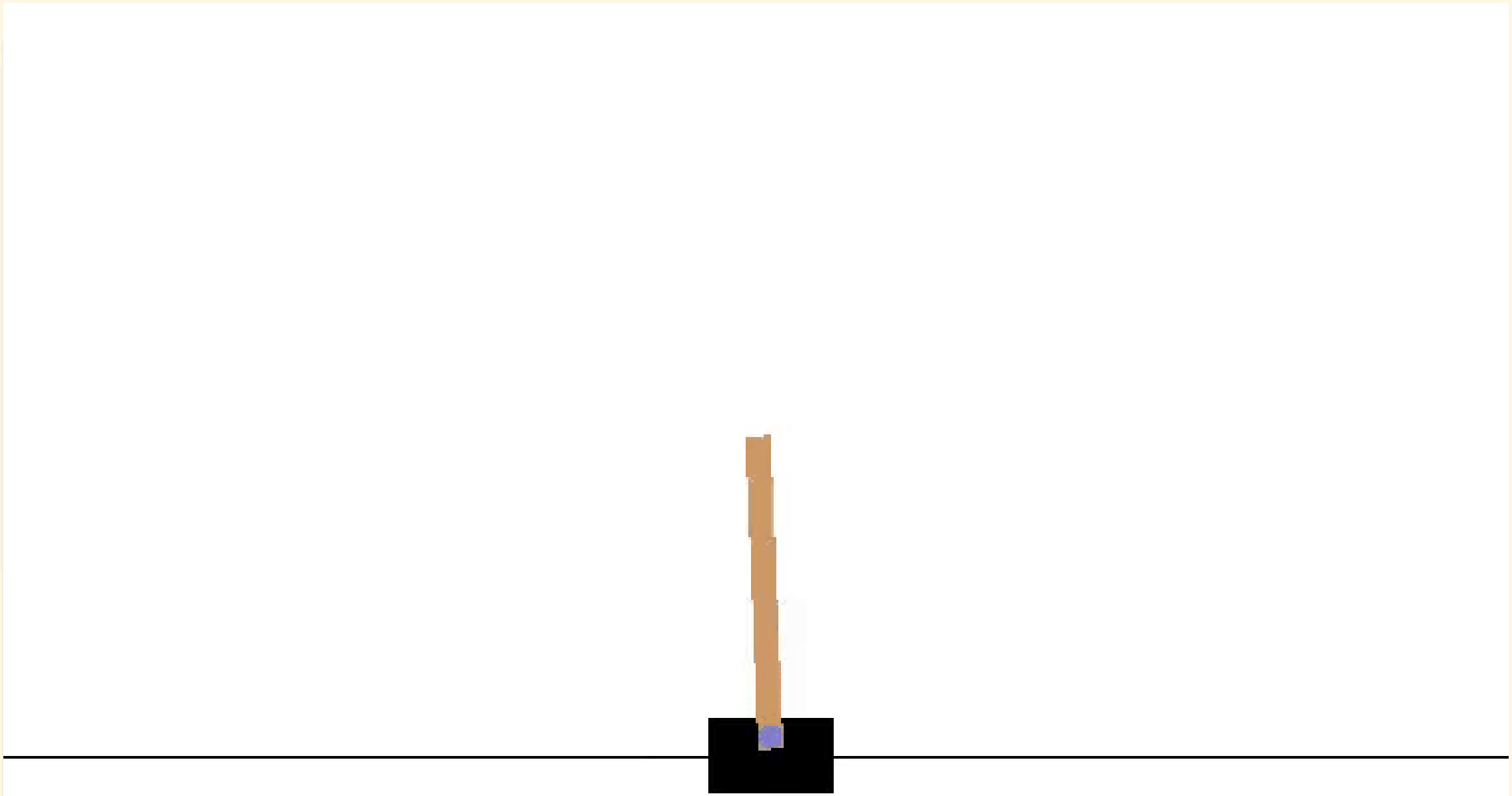


L.E.J. Brouwer

Every continuous function from a convex compact subset K of a Euclidean space to K itself has a fixed point.


PAUL
NOTH


Crumpled City
VENICE
B
C
LIDO
D
Sacca di S. Girolamo
Sacca di S. Alvise
Isola di S. Elena
5
6



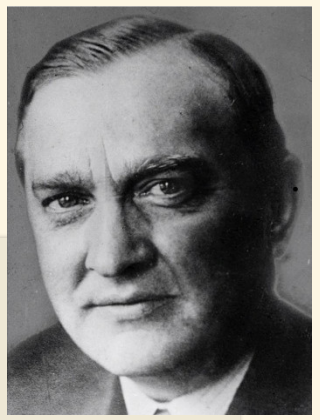
FIXED POINT ITERATION FOR ROOT FINDING

$$x^2 - x - 1 = 0$$

$x^2 = x + 1$
$x = 1 + \frac{1}{x}$
$x_{n+1} = 1 + \frac{1}{x_n}$
Pick initial $x_0 = 2$
$x_1 = 1 + \frac{1}{2} = 1.5$
$x_2 = 1 + \frac{1}{1.5} = 1.6666$
$x_3 = 1 + \frac{1}{1.6666} = 1.6$
$x_4 = 1 + \frac{1}{1.6} = 1.625$
$x_5 = 1 + \frac{1}{1.625} = 1.612538462$
Converges to 1.618

$x^2 - x = 1$
$x(x-1) = 1$
$x = \frac{1}{x-1}$
$x_{n+1} = \frac{1}{x_n - 1}$
Pick initial $x_0 = 1.6$
$x_1 = \frac{1}{1.6-1} = 1.6666$
$x_2 = \frac{1}{1.66666-1} = 1.5$
$x_3 = \frac{1}{1.5-1} = 2$
$x_4 = \frac{1}{2-1} = 1$

BANACH'S FIXED-POINT THEOREM


Stefan Banach

A metric space M is called ***complete*** *if every Cauchy sequence in M converges in M*

Let (X, d) be a ***complete metric space****. Then a map $T: X \to X$ is called a* ***contraction mapping*** *on X if there exists $q \in [0, 1)$ such that*

$$d(T(x), T(y)) \leq qd(x, y), \forall x, y \in X$$

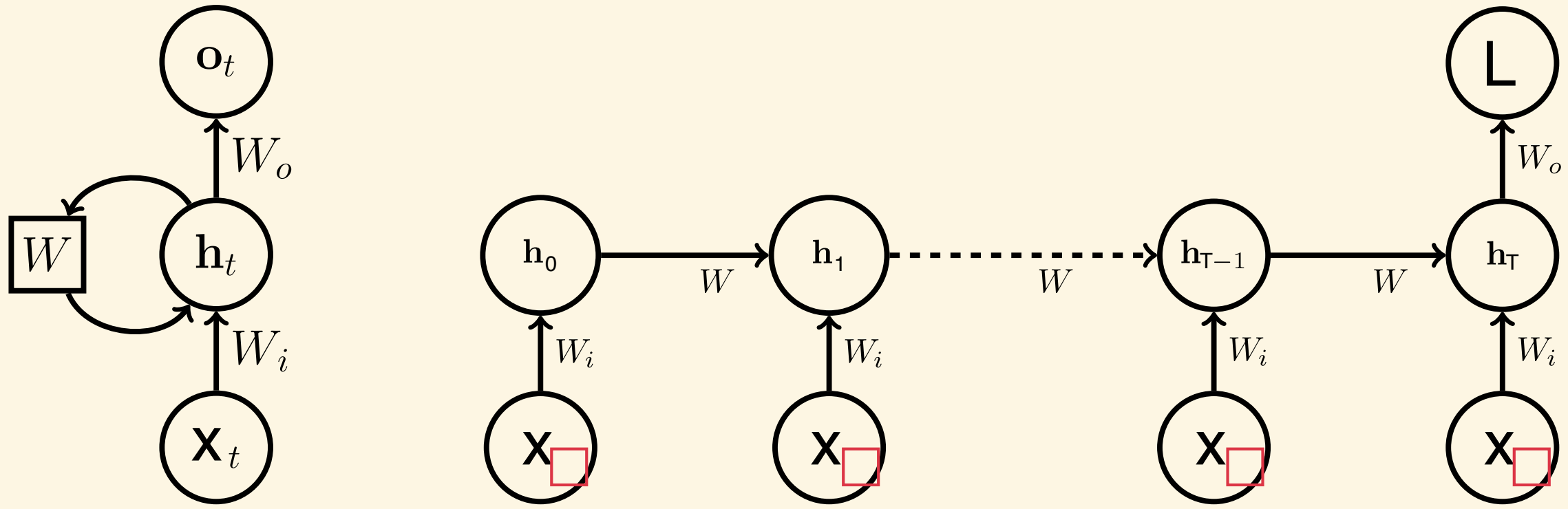
Theorem. *Let (X, d) be a non-empty complete metric space with a contraction mapping $T: X \to X$. Then T admits a unique fixed-point x^* in X (i.e. $T(x^*) = x^*$).*

Furthermore, x^* can be found as follows: start with an arbitrary element x_0 in X and define a sequence $\{x_n\}$ by $x_n = T(x_{n-1})$ for $n \geq 1$. Then $x_n \to x^*$.

Wikipedia

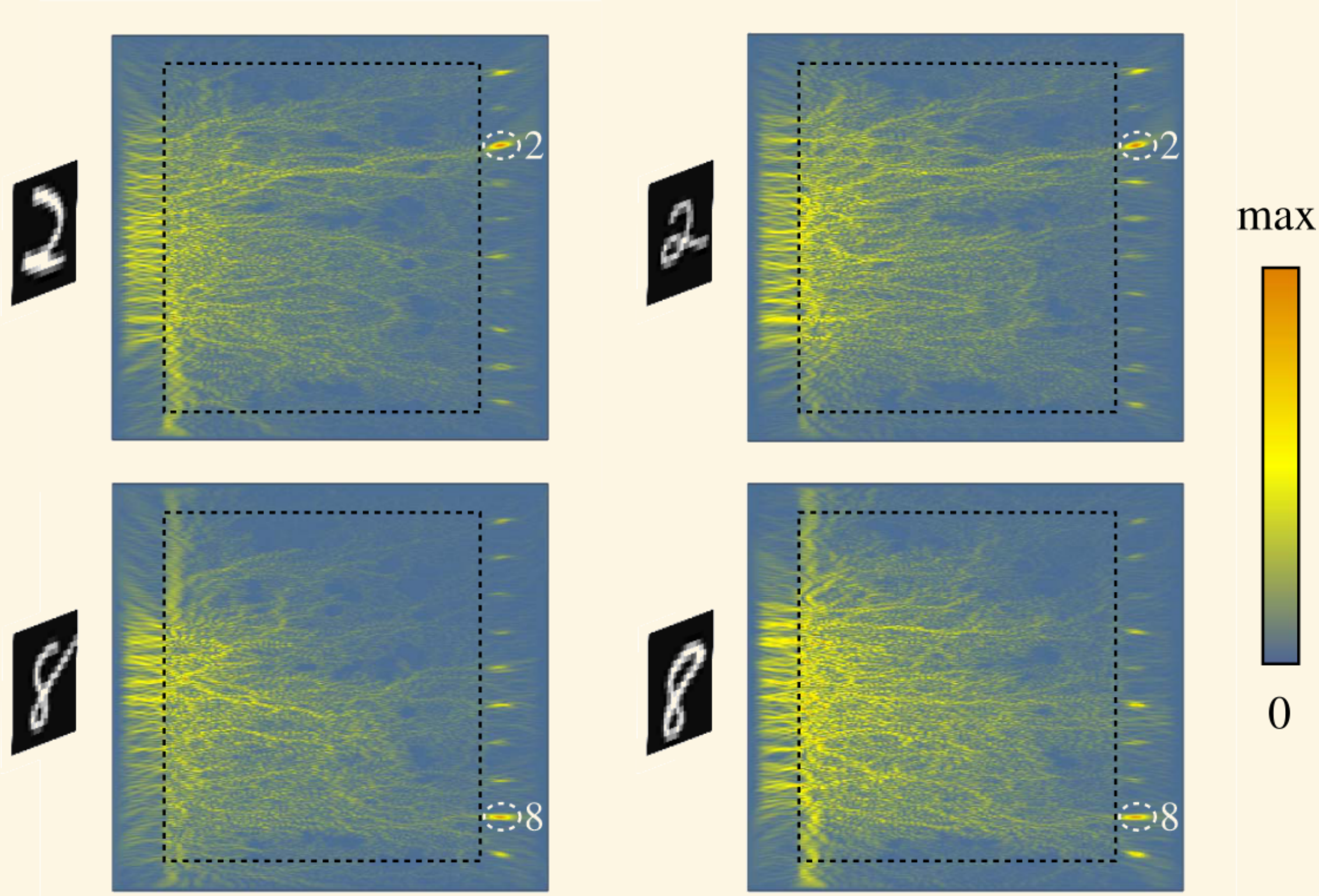
RECURRENT BACK PROPAGATION

RNN WITH THE SAME INPUT AT EVERY t

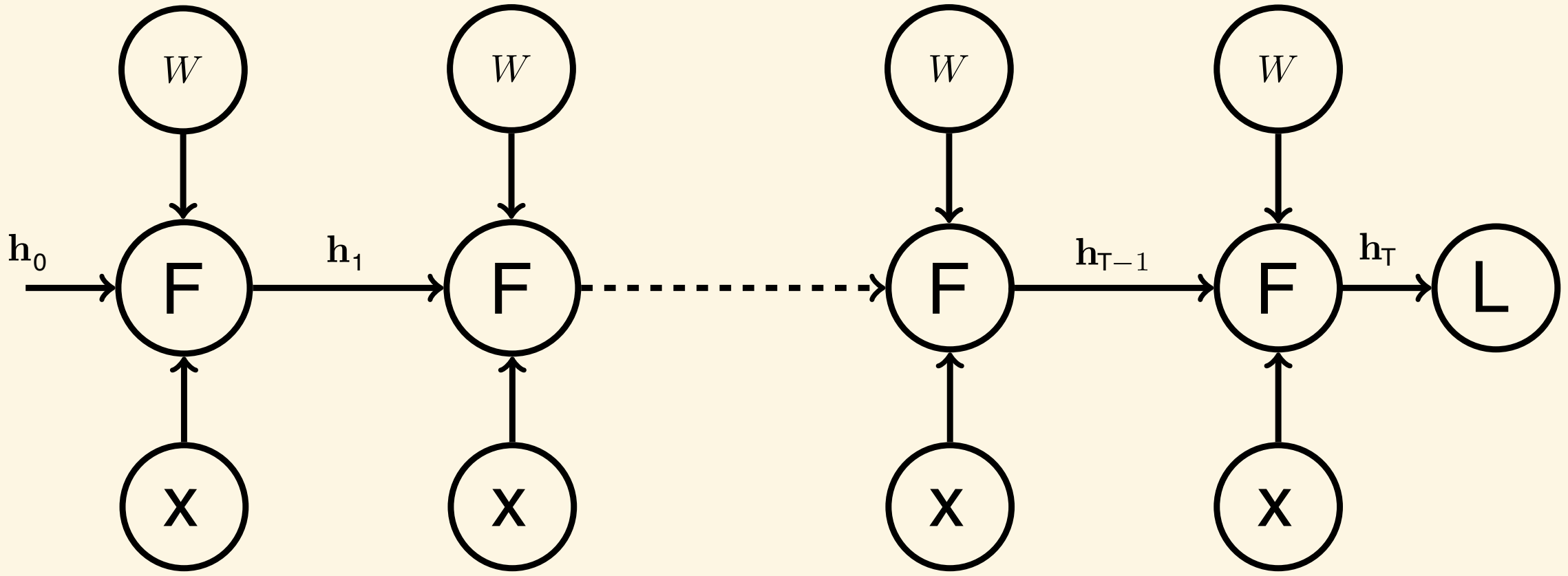


DEEP LEARNING IS BALLISTIC

DEEP LEARNING IS BALLISTIC



RNN AS A MAP



$$\mathbf{h}_{t+1} = F(\mathbf{x}, \mathbf{w}, \mathbf{h}_t)$$
$$\mathbf{h}^* = F(\mathbf{x}, \mathbf{w}, \mathbf{h}^*)$$

BACK PROPAGATION THROUGH TIME

$$\frac{\partial L}{\partial \mathbf{w}^T} = \frac{\partial L}{\partial \mathbf{h}^T} \frac{\partial \mathbf{h}^T}{\partial \mathbf{w}^T}$$

$$\frac{\partial L}{\partial \mathbf{w}^{T-1}} = \frac{\partial L}{\partial \mathbf{h}^T} \frac{\partial \mathbf{h}^T}{\partial \mathbf{h}^{T-1}} \frac{\partial \mathbf{h}^{T-1}}{\partial \mathbf{w}^{T-1}}$$

$$\frac{\partial L}{\partial \mathbf{w}} = \frac{\partial L}{\partial \mathbf{h}^T} \left(\frac{\partial \mathbf{h}^T}{\partial \mathbf{w}} + \frac{\partial \mathbf{h}^T}{\partial \mathbf{h}^{T-1}} \frac{\partial \mathbf{h}^{T-1}}{\partial \mathbf{w}} \dots \right)$$

$$\frac{\partial L}{\partial \mathbf{w}} = \frac{\partial L}{\partial \mathbf{h}^T} \sum_{k=1}^{T} \left(\prod_{i=T-k+1}^{T-1} J_{F,\mathbf{h}} \left(\mathbf{x}, \mathbf{w}, \mathbf{h}^i \right) \right) \frac{\partial F}{\partial \mathbf{w}} \left(\mathbf{x}, \mathbf{w}, \mathbf{h}^{T-k} \right)$$

"THINKING" AND CURRYING


Haskell Curry

$$F(x) = \sqrt{x}$$

$$F(x, u) = \frac{1}{2}\left(u + \frac{x}{u}\right)$$

$$T(u) = F(x, u)$$

$$u_{n+1} = T(u_n)$$

IMPLICIT RNN

$$\Psi(\mathbf{w}, \mathbf{h}) = \mathbf{h} - \mathrm{F}(\mathbf{x}, \mathbf{w}, \mathbf{h})$$

$$\frac{\partial \Psi}{\partial \mathbf{w}}(\mathbf{w}, \mathbf{h}^*) = \frac{\partial \mathbf{h}}{\partial \mathbf{w}}(\mathbf{h}^*) - \frac{dF}{d\mathbf{w}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*)$$

$$J_{F,\mathbf{h}}\left(\mathbf{h}^*\right) = \frac{\partial F}{\partial \mathbf{h}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*)$$

$$\frac{dF}{d\mathbf{w}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*) = \frac{\partial F}{\partial \mathbf{w}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*) + \frac{\partial F}{\partial \mathbf{h}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*)\, \frac{\partial h}{\partial \mathbf{w}}(\mathbf{h}^*)$$

$$\frac{dF}{d\mathbf{w}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*) = \frac{\partial F}{\partial \mathbf{w}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*) + J_{F,\mathbf{h}}\left(\mathbf{h}^*\right)\, \frac{\partial h}{\partial \mathbf{w}}(\mathbf{h}^*)$$

IMPLICIT DIFFERENTIATION

$$\frac{\partial \Psi}{\partial \mathbf{w}}(\mathbf{w}, \mathbf{h}^*) = \frac{\partial \mathbf{h}}{\partial \mathbf{w}}(\mathbf{h}^*) - \frac{dF}{d\mathbf{w}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*)$$

$$\frac{dF}{d\mathbf{w}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*) = \frac{\partial F}{\partial \mathbf{w}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*) + J_{F,\mathbf{h}}\left(\mathbf{h}^*\right) \frac{\partial h}{\partial \mathbf{w}}(\mathbf{h}^*)$$

$$\frac{\partial \Psi}{\partial \mathbf{w}}(\mathbf{w}, \mathbf{h}^*) = \left(I - J_{F,\mathbf{h}}\left(\mathbf{h}^*\right)\right) \frac{\partial \mathbf{h}}{\partial \mathbf{w}}(\mathbf{h}^*) - \frac{\partial F}{\partial \mathbf{w}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*) = \mathbf{0}$$

$$\frac{\partial \mathbf{h}}{\partial \mathbf{w}}(\mathbf{h}^*) = \left(I - J_{F,\mathbf{h}}\left(\mathbf{h}^*\right)\right)^{-1} \frac{\partial F}{\partial \mathbf{w}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*)$$

IMPLICIT DIFFERENTIATION

$$\frac{\partial L}{\partial \mathbf{w}} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial \mathbf{h}}(\mathbf{h}^*)\left(I - J_{F,\mathbf{h}}(\mathbf{h}^*)\right)^{-1}\frac{\partial F}{\partial \mathbf{w}}(\mathbf{x}, \mathbf{w}, \mathbf{h}^*)$$

Let's introduce an auxiliary variable

$$\mathbf{z} = \left(I - J_{F,\mathbf{h}}^T(\mathbf{h}^*)\right)^{-1}\left(\frac{\partial L}{\partial y}\frac{\partial y}{\partial \mathbf{h}}(\mathbf{h}^*)\right)^T$$

Multiply both sides by $\left(I - J_{F,\mathbf{h}}^T(\mathbf{h}^*)\right)$

$$\mathbf{z} - J_{F,\mathbf{h}}^T(\mathbf{h}^*)\,\mathbf{z} = \left(\frac{\partial L}{\partial y}\frac{\partial y}{\partial \mathbf{h}}(\mathbf{h}^*)\right)^T$$

Iterate to convergence

$$\mathbf{z} = J_{F,\mathbf{h}}^T(\mathbf{h}^*)\,\mathbf{z} + \left(\frac{\partial L}{\partial y}\frac{\partial y}{\partial \mathbf{h}}(\mathbf{h}^*)\right)^T$$

RNN AS A MAP

1: **Initialization:** initial guess z_0, *e.g.*, draw uniformly from $[0, 1]$, $i = 0$, threshold ϵ
2: **repeat**
3: $\quad i = i + 1$
4: $\quad z_i = J_{F,h^*}^{\top} z_{i-1} + \left(\frac{\partial L}{\partial y} \frac{\partial y}{\partial h^*} \right)^{\top}$
5: **until** $\|z_i - z_{i-1}\| < \epsilon$
6: $\frac{\partial L}{\partial w_F} = z_i^{\top} \frac{\partial F(x, w_F, h^*)}{\partial w_F}$
7: Return $\frac{\partial L}{\partial w_F}$

$$D\left(\mathbf{z}, \mathbf{x}, \mathbf{w}, \mathbf{h}^*\right) = \mathbf{z}^T F\left(\mathbf{x}, \mathbf{w}, \mathbf{h}^*\right)$$

Use autograd on both terms

$$\mathbf{z} = \frac{\partial D}{\partial \mathbf{h}}(\mathbf{z}, \mathbf{x}, \mathbf{w}, \mathbf{h}^*) + \left(\frac{\partial L}{\partial y} \frac{\partial y}{\partial \mathbf{h}}(\mathbf{h}^*) \right)^T$$

IMPLICIT FUNCTION DIFFERENTIATION

IMPLICIT FUNCTION

$$F(x, y) = x^2 + y^2 - 1$$

TOTAL DERIVATIVE

Differentiation of implicit functions. Suppose we wish to find the derivative of y, where y is a function of x defined implicitly by the relation

$$F(x, y) = 0 \tag{38}$$

between these variables. If x and y satisfy the relation (38) and we give x the increment Δx, then y will receive an increment Δy such that $x + \Delta x$ and $y + \Delta y$ again satisfy (38). Consequently*

$$F(x + \Delta x, y + \Delta y) - F(x, y) = \frac{\partial F}{\partial x}\Delta x + \frac{\partial F}{\partial y}\Delta y + \alpha\sqrt{\Delta x^2 + \Delta y^2} = 0.$$

Thus, provided $\partial F/\partial y \neq 0$, it follows that

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = y'_x = -\frac{\dfrac{\partial F}{\partial x}}{\dfrac{\partial F}{\partial y}}.$$

In this way we have obtained a method for finding the derivative of an implicit function y without first solving the equation (38) for y.

CONCLUDING REMARKS

REFERENCES

Original

- Almeida, L. B. A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. In IEEE International Conference on Neural Networks, pp. 609–618, 1987
- Pineda, F. J. Generalization of back-propagation to recurrentneural networks.Physical review letters, 59(19):2229,1987
- Feynman, R. P. Forces in molecules.Physical Review, 56(4):340, 1939

Recent "Rediscoveries"

- Liao, R., Xiong, Y., Fetaya, E., Zhang, L., Yoon, K., Pitkow, X., Urtasun, R. & Zemel, R.. (2018). Reviving and Improving Recurrent Back-Propagation. Proceedings of the 35th International Conference on Machine Learning, in PMLR 80:3082-3091
- Bai, S., Kolter, J.Z. and Koltun, V., 2019. Deep equilibrium models. In Advances in Neural Information Processing Systems (pp. 688-699).
- Rajeswaran, A., Finn, C., Kakade, S.M. and Levine, S., 2019. Meta-learning with implicit gradients. In Advances in Neural Information Processing Systems (pp. 113-124).
- Jeon Y., Lee M., Young J. ChoiDifferentiable Fixed-Point Iteration Layer 2020