MECHANICAL MODELS OF FEATURE LEARNING

By Ivan Dokmanić

August 8, 2025, Cargèse

SADA signals and data group

Universität Basel

erc

PRACTICAL DEEP LEARNING IS

DEEP

NON-LINEAR

FEATURE LEARNING

PRACTICAL DEEP LEARNING IS

DEEP

NON-LINEAR

FEATURE LEARNING

A GREAT DEAL OF THEORY ADDRESSES

SHALLOW LEARNING

LINEAR ARCHITECTURES

NON-FEATURE LEARNING

AN ACCIDENT: THE MICROSCOPIC PICTURE IS 100% KNOWN

AN ACCIDENT: THE MICROSCOPIC PICTURE IS 100% KNOWN

$$\frac{pV}{T} = \text{const}\,.$$

BOYLE 1662
CHARLES 1787,
GAY-LUSSAC 1802,
. . .

AN ACCIDENT: THE MICROSCOPIC PICTURE IS 100% KNOWN

$$\frac{pV}{T} = \text{const}\,.$$

BOYLE 1662
CHARLES 1787,
GAY-LUSSAC 1802,
. . .

$$p = \frac{1}{3}\frac{N}{V}\overline{mv^2}$$

$$f(v) \propto v^2 e^{-\frac{mv^2}{2k_B T}}$$

MAXWELL & BOLTZMANN 1860-1872

THE BLACK BOX SYNDROME

THE BLACK BOX SYNDROME

THE BLACK BOX SYNDROME

I AM NOT A BLACK BOX.

SCIENTIFIC THEORIES

Appetizer 3 First principles of phenomenological?

PHENOMENOLOGICAL MODELS

First Course 13 Springs, blocks, and folding rulers

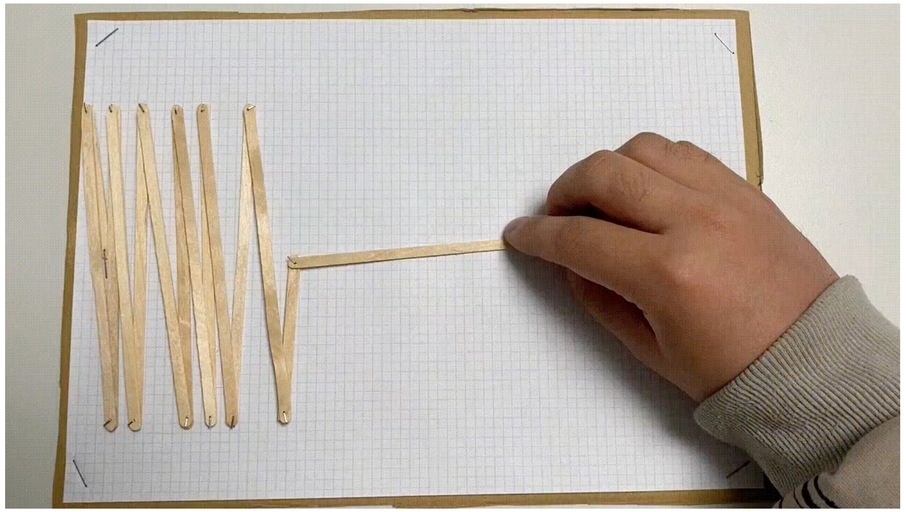
FIRST-PRINCIPLE ATTEMPTS

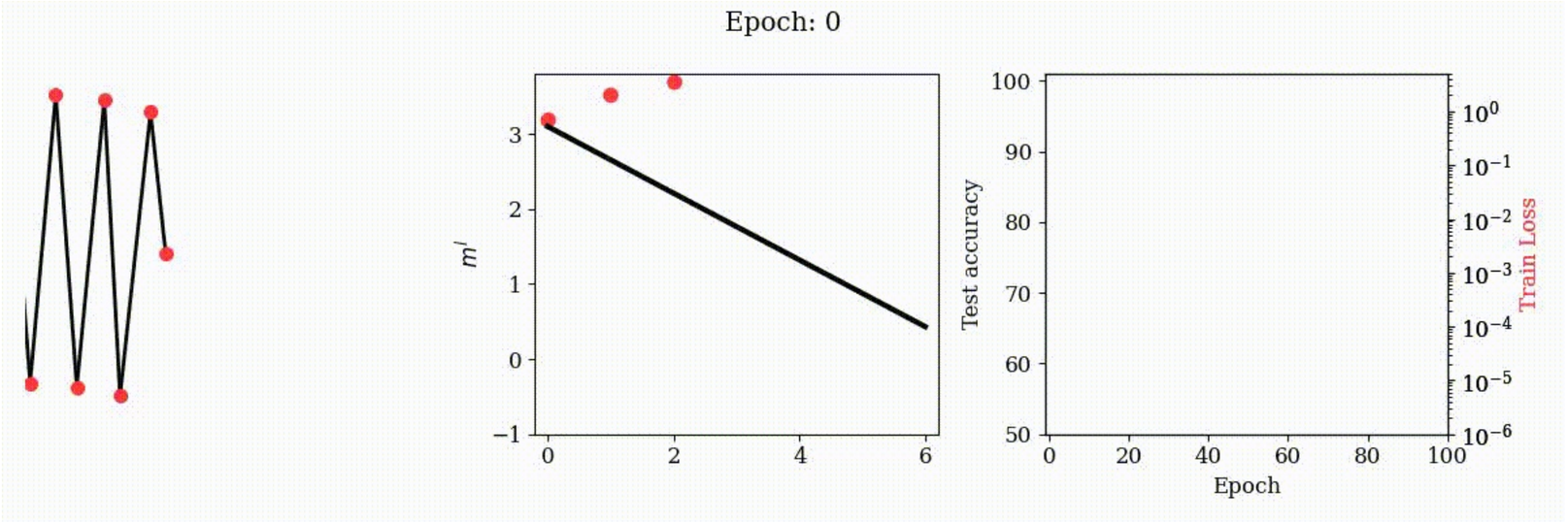
Second Course 26 Deep mean field theory and all that

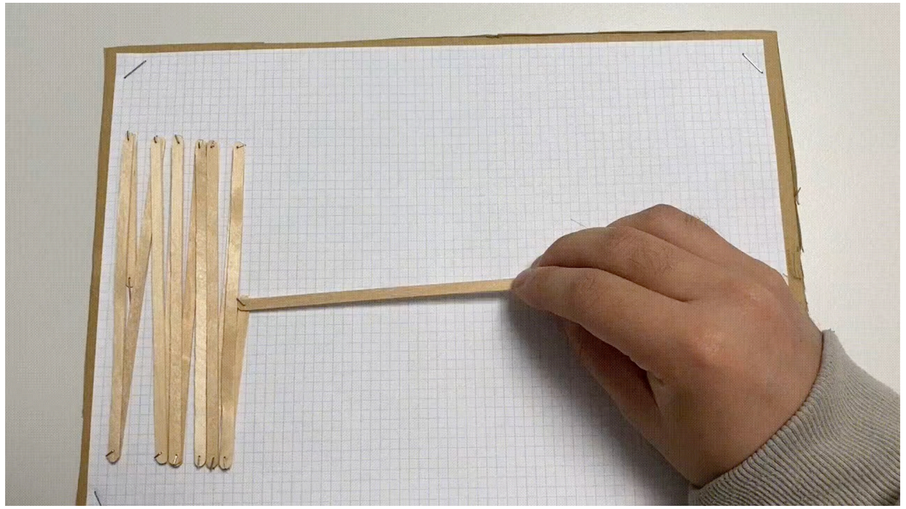
Dessert 34 ENDNOTES

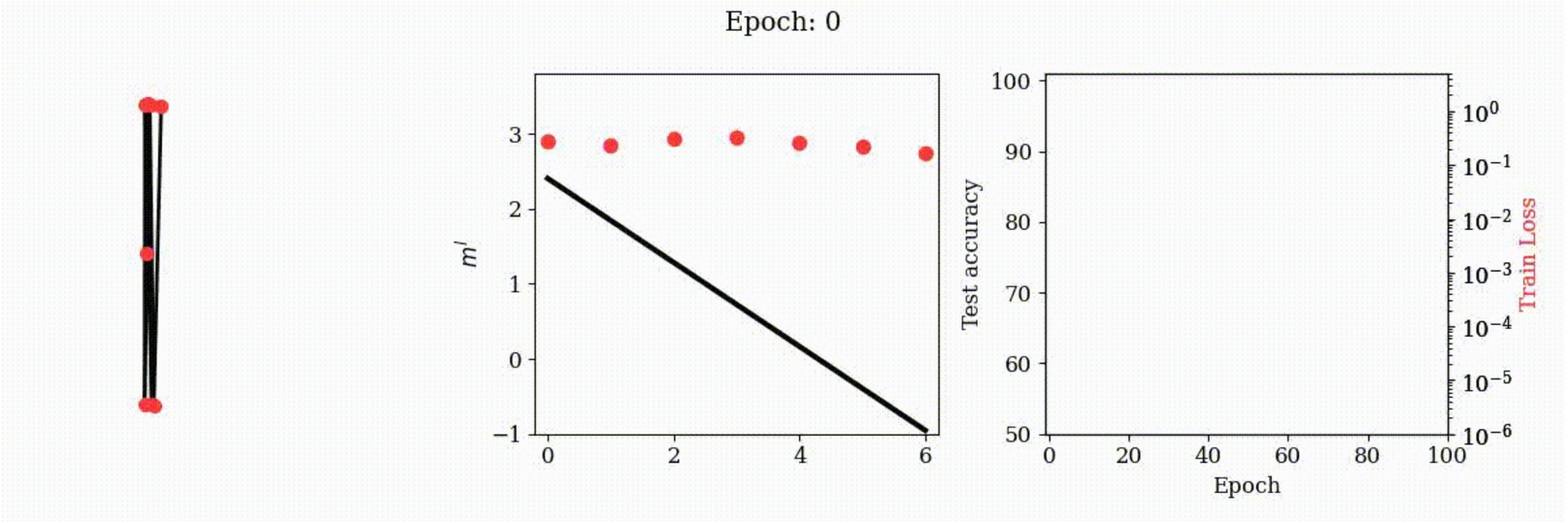
Where I quote Boltzmann

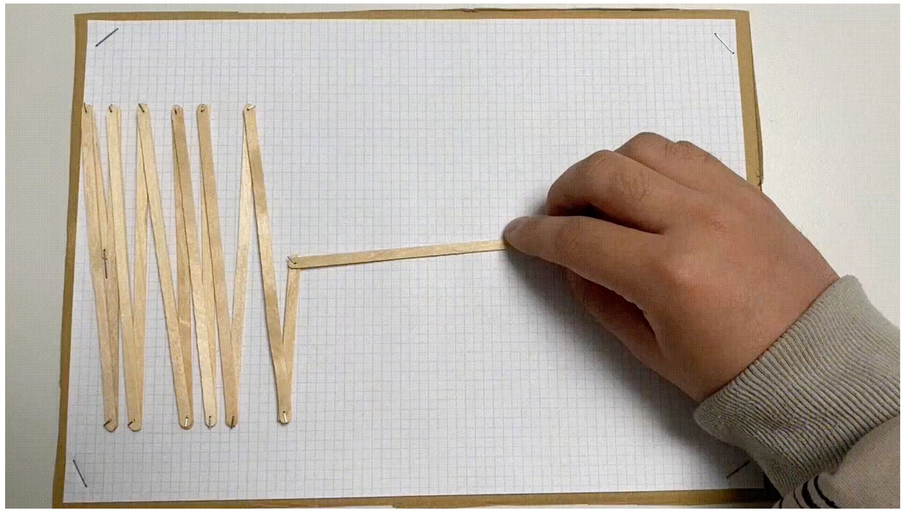
TRUE & USEFUL MACROSCOPIC STATEMENTS ABOUT DEEP NETS?

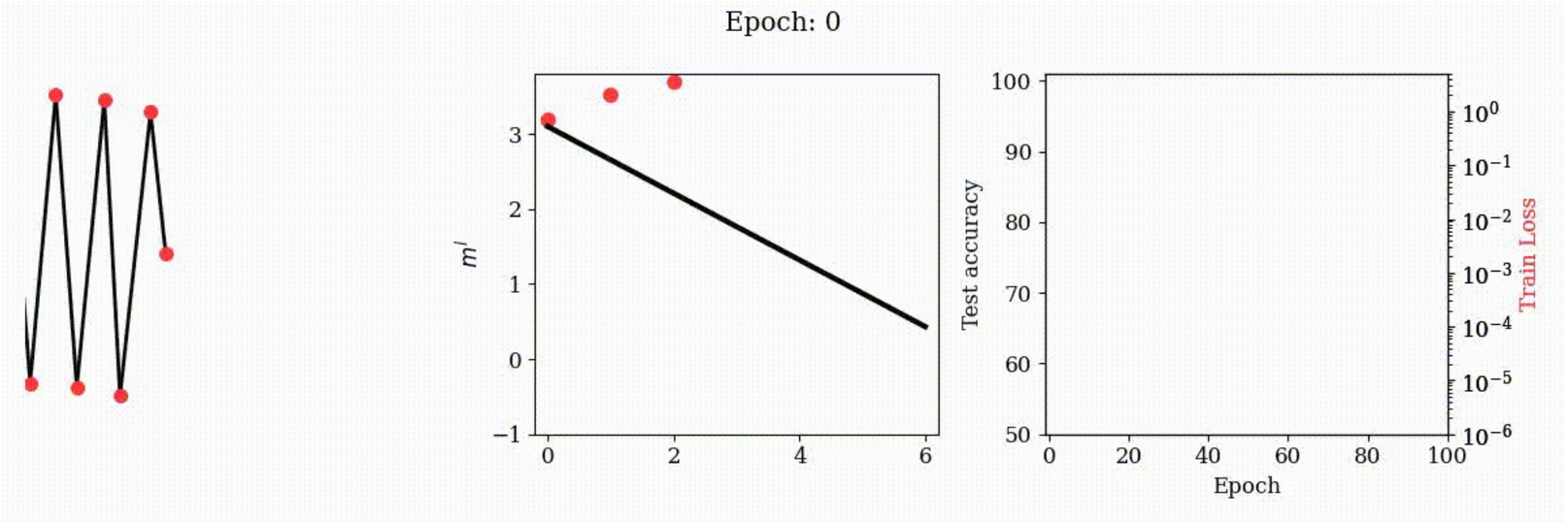


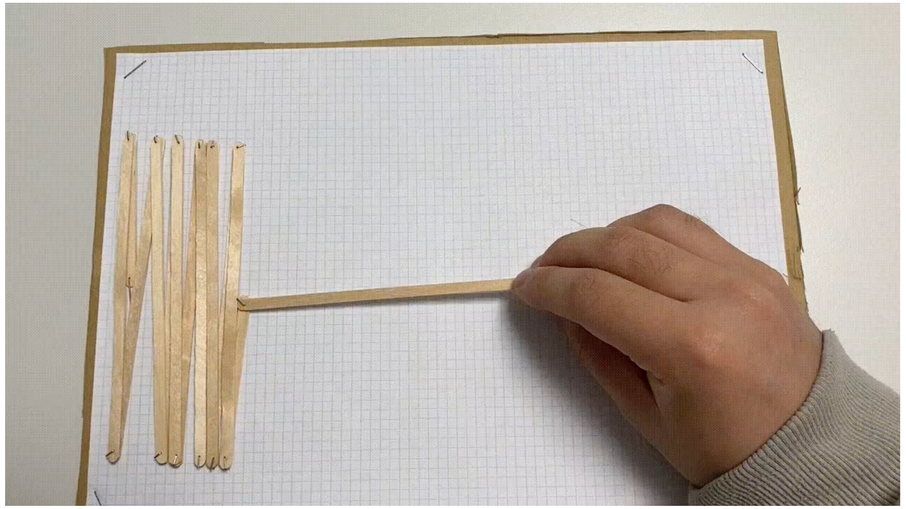

Epoch: 0
m^l
Test accuracy
Train Loss
Epoch

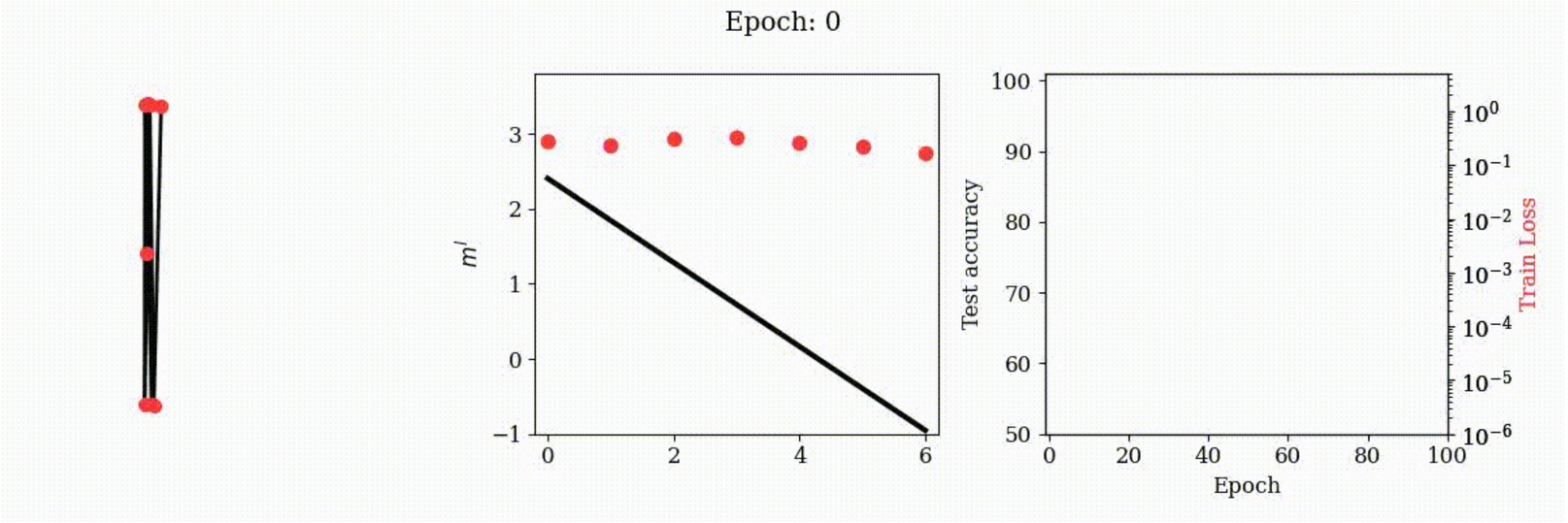


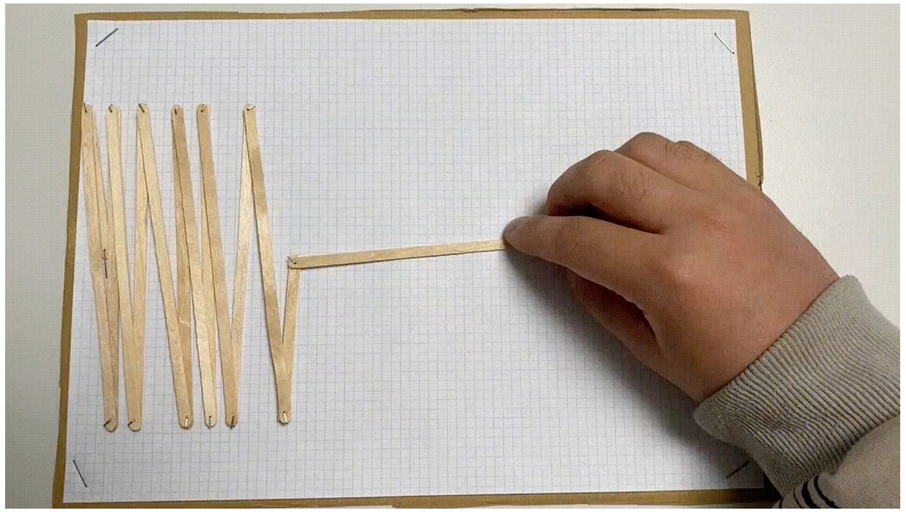

Epoch: 0
m^l
Test accuracy
Train Loss
Epoch

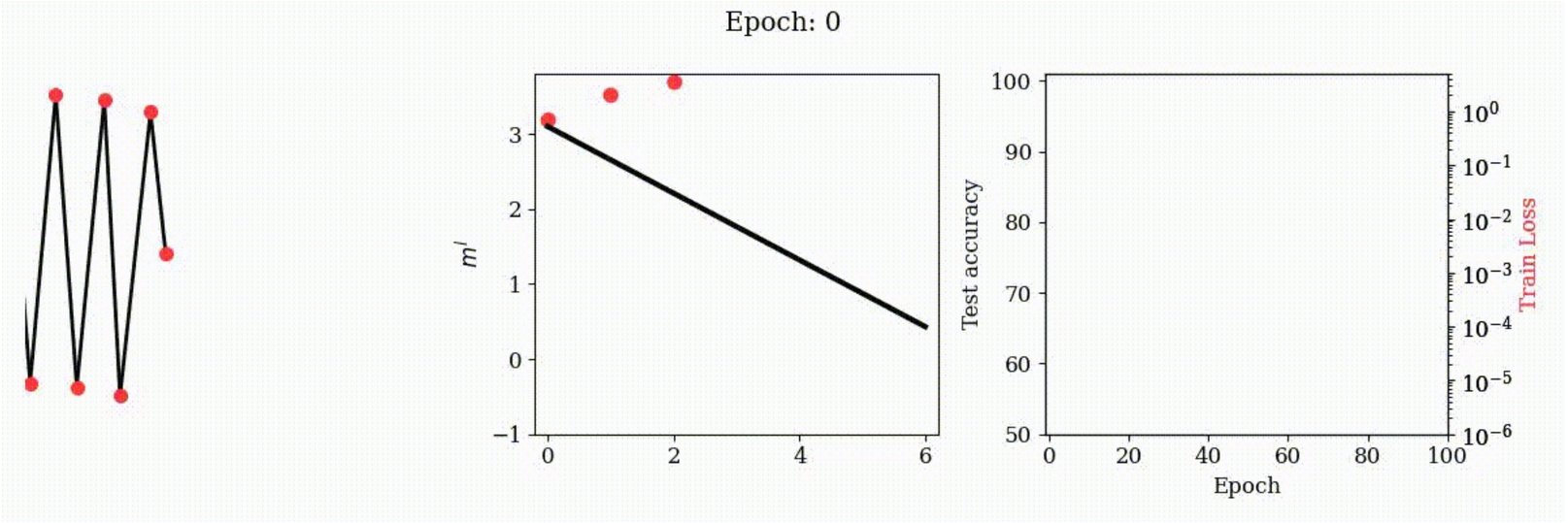


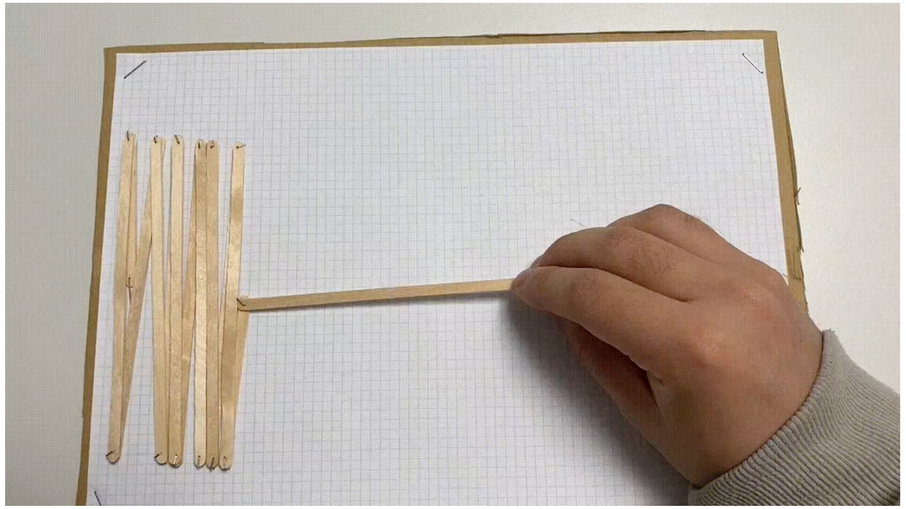

Epoch: 0
m^l
Test accuracy
Train Loss
Epoch

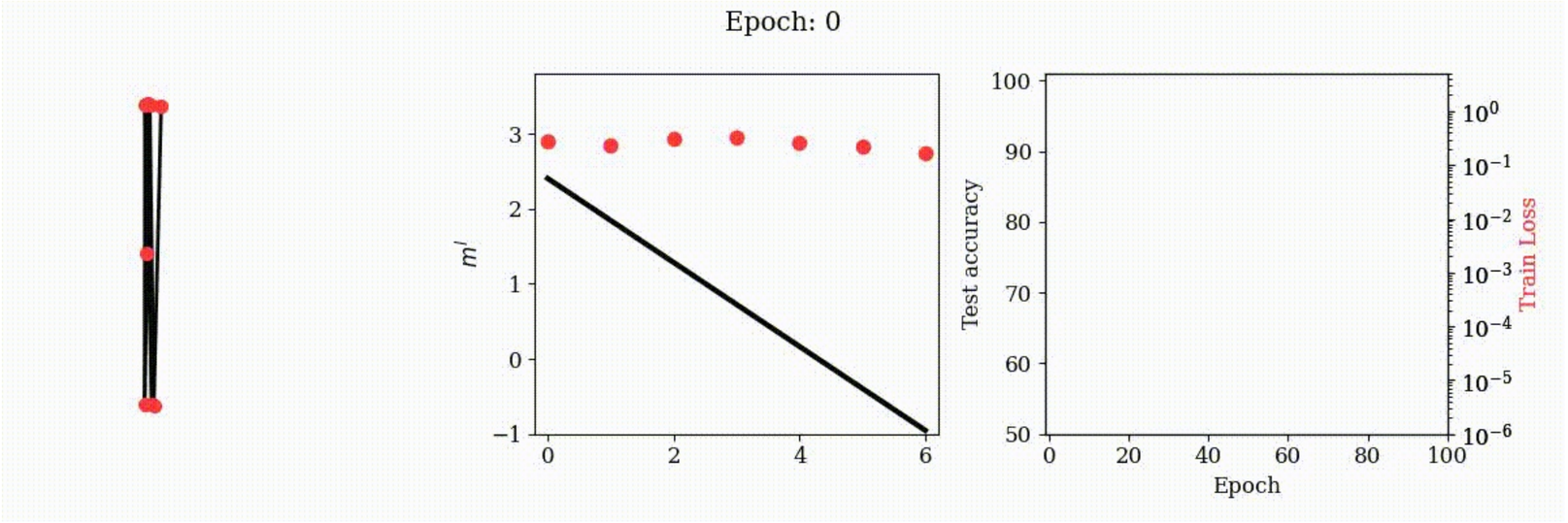


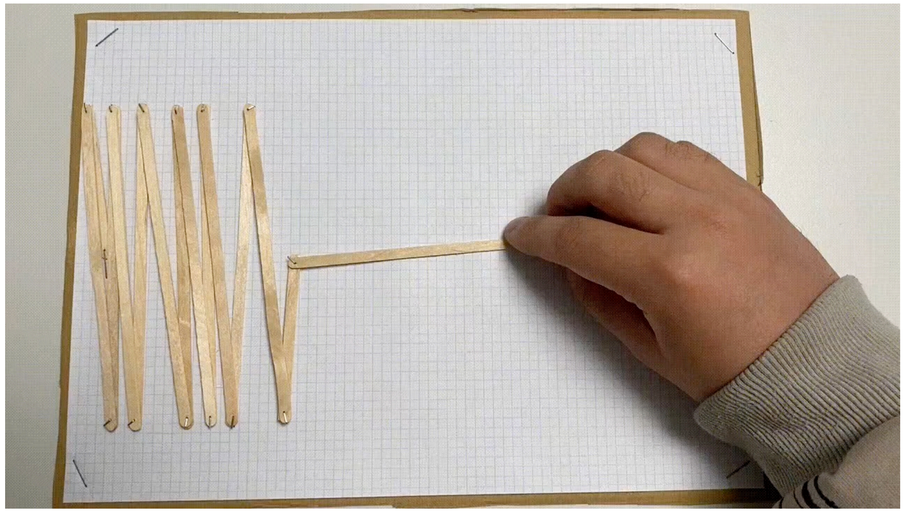

Epoch: 0
m^l
Test accuracy
Train Loss
Epoch

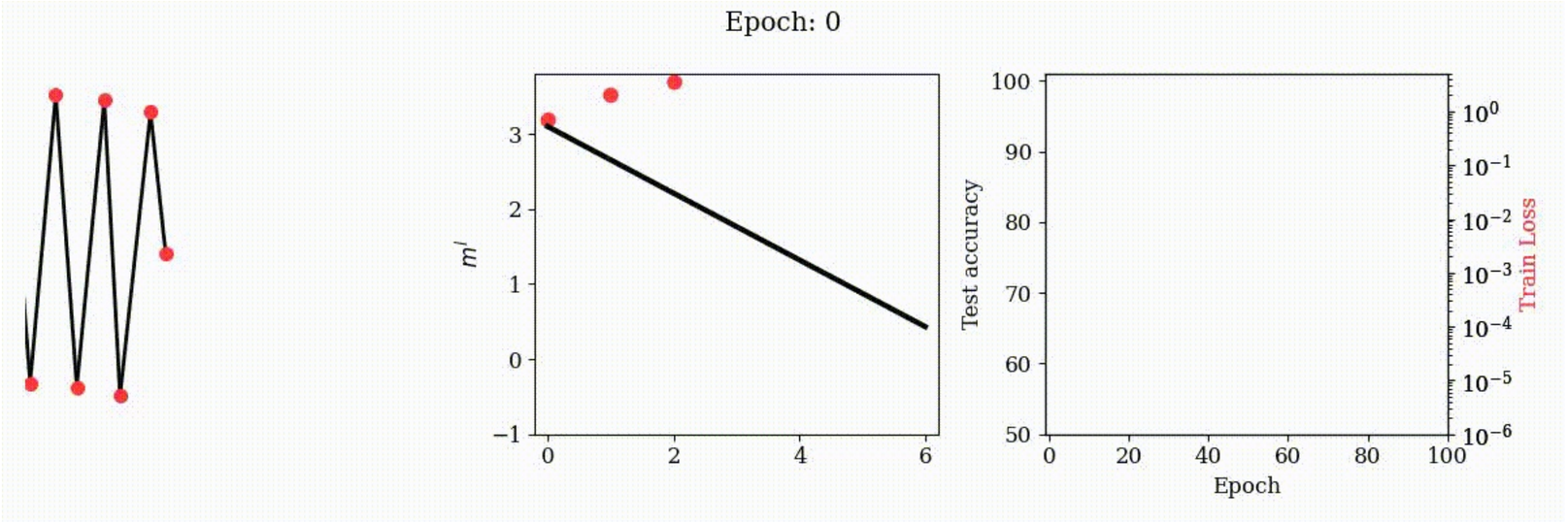


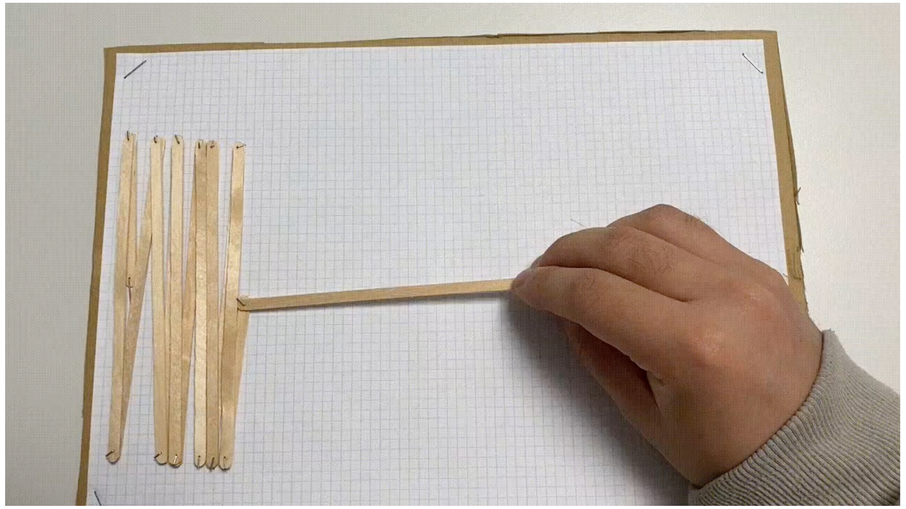

Epoch: 0
m^l
Test accuracy
Train Loss
Epoch

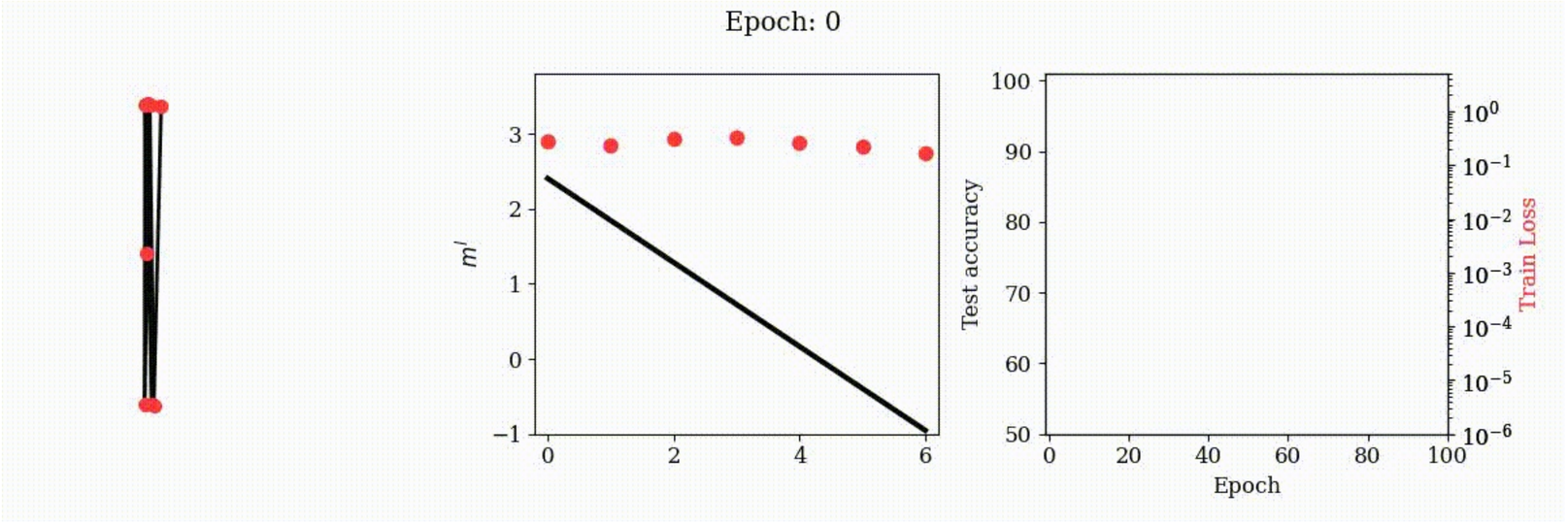


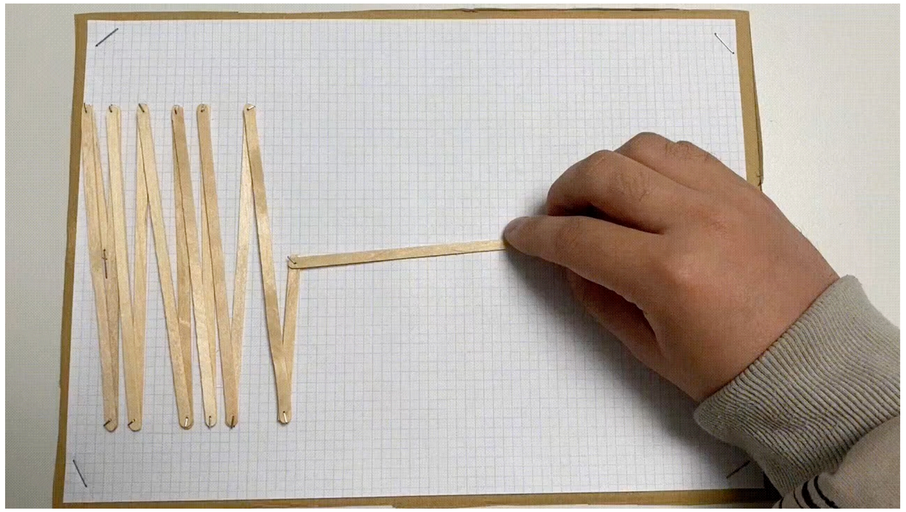

Epoch: 0
m^l
Test accuracy
Train Loss
Epoch

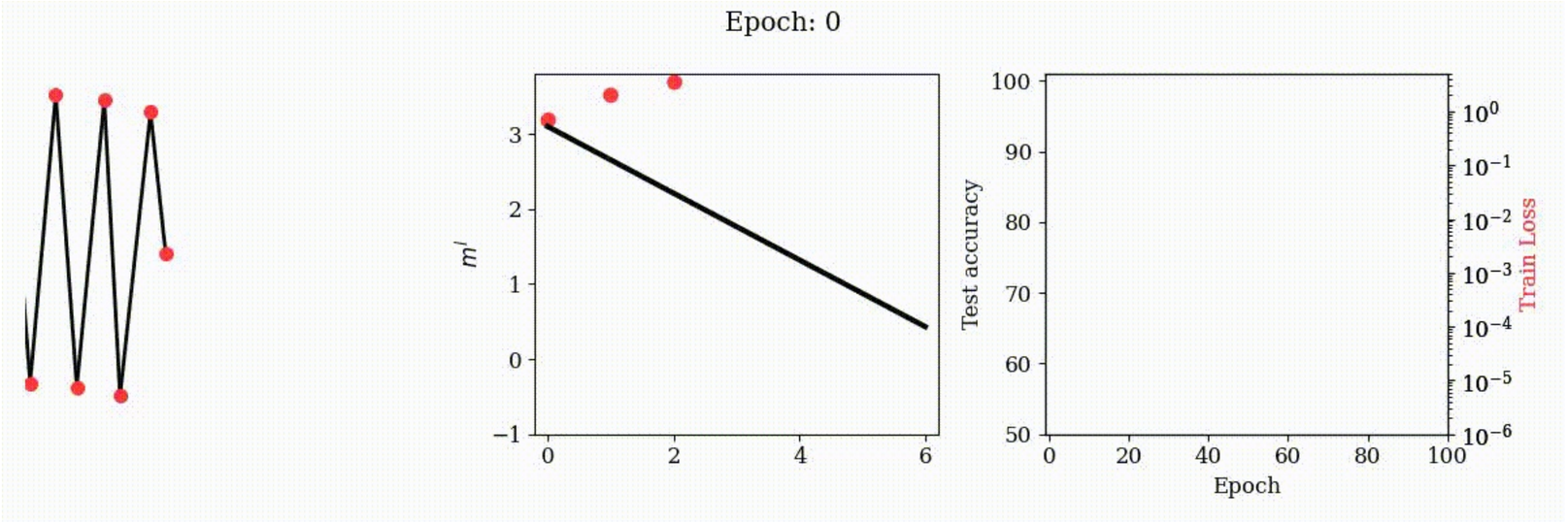


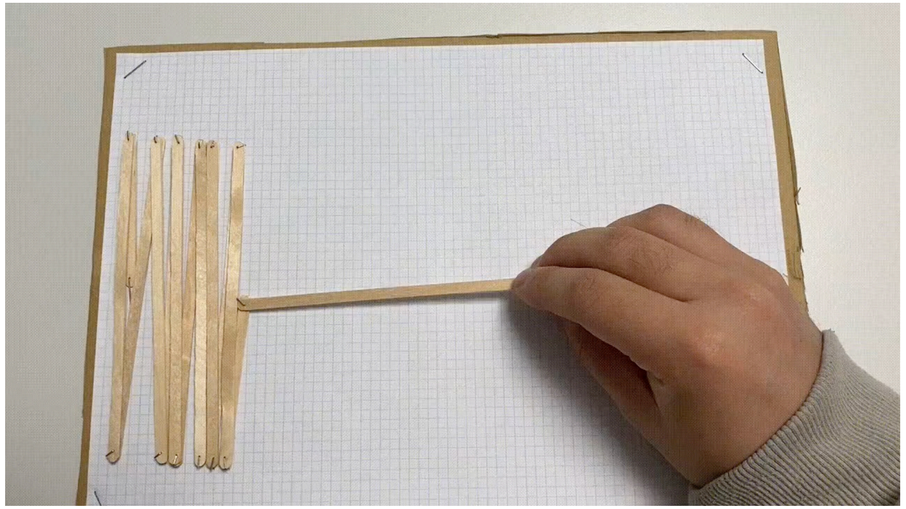

Epoch: 0
m^l
Test accuracy
Train Loss
Epoch

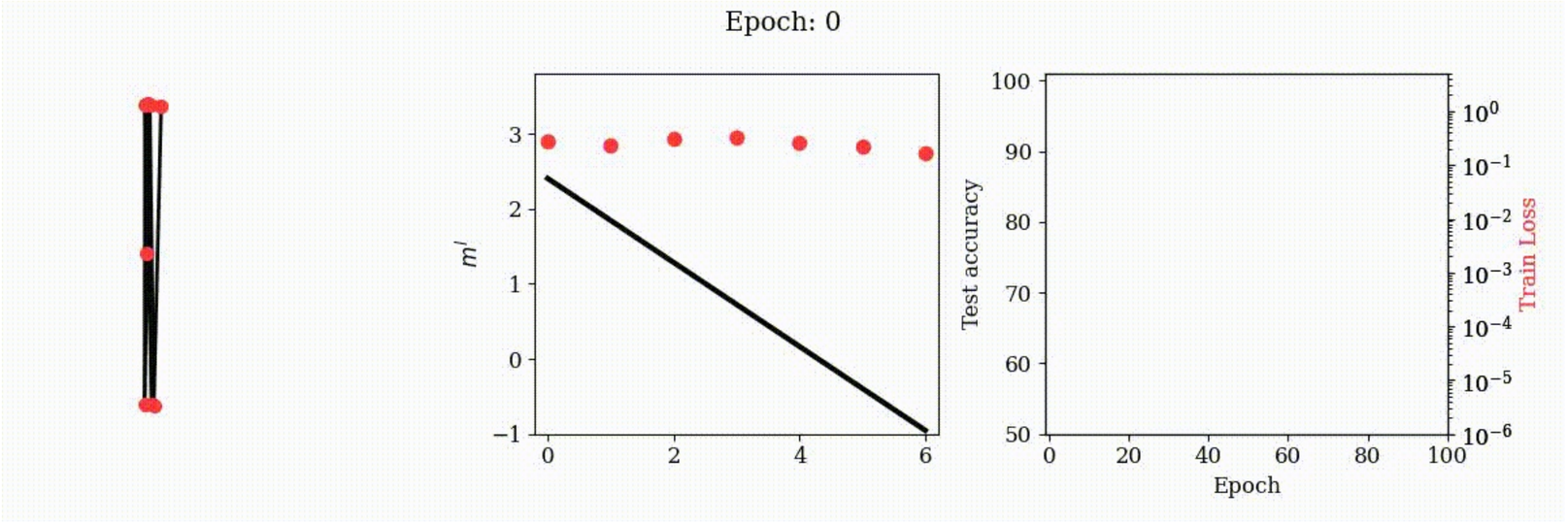



Epoch: 0
m^l
Test accuracy
Train Loss
Epoch

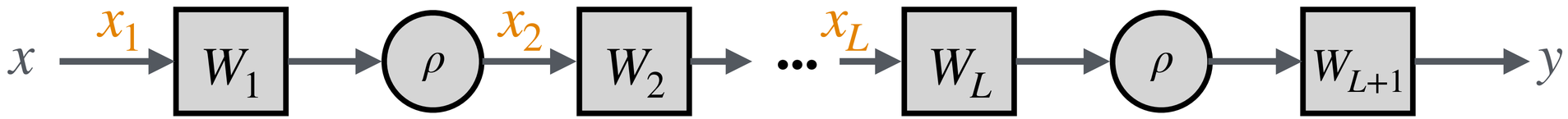



Epoch: 0
m^l
Test accuracy
Train Loss
Epoch

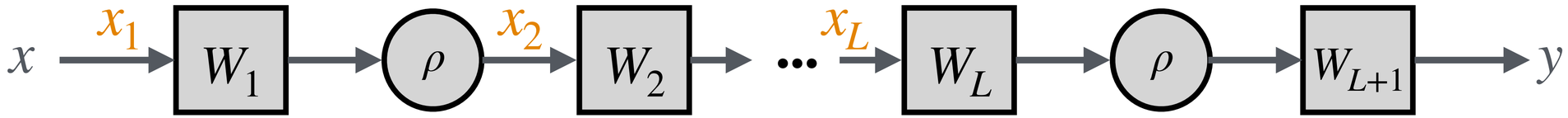



Epoch: 0
m^l
Test accuracy
Train Loss
Epoch

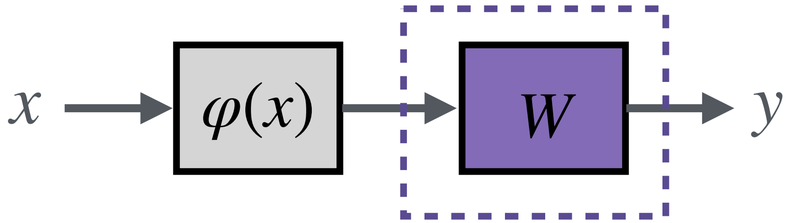
FEATURE LEARNING



FEATURE LEARNING



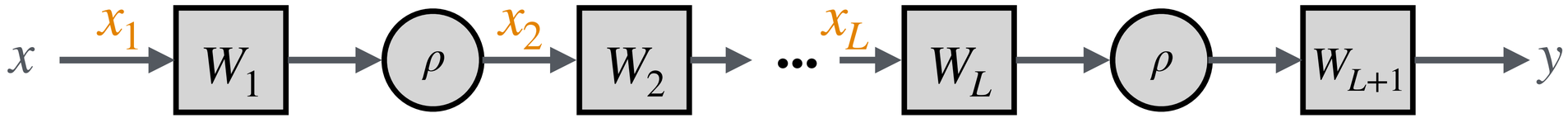
Non-feature learning (old)



only W is trained

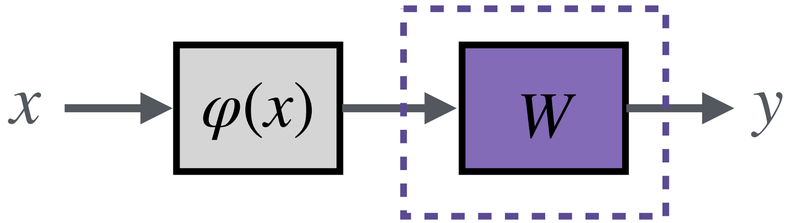
$$f_W(x) = W \circ \varphi(x)$$

FEATURE LEARNING

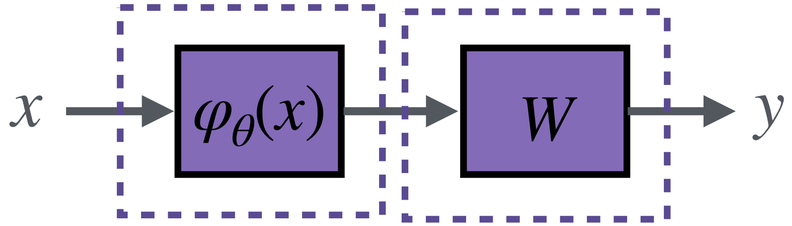


Non-feature learning (old)

Feature learning (new)



only W is trained

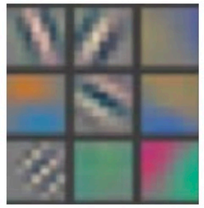


everything learns

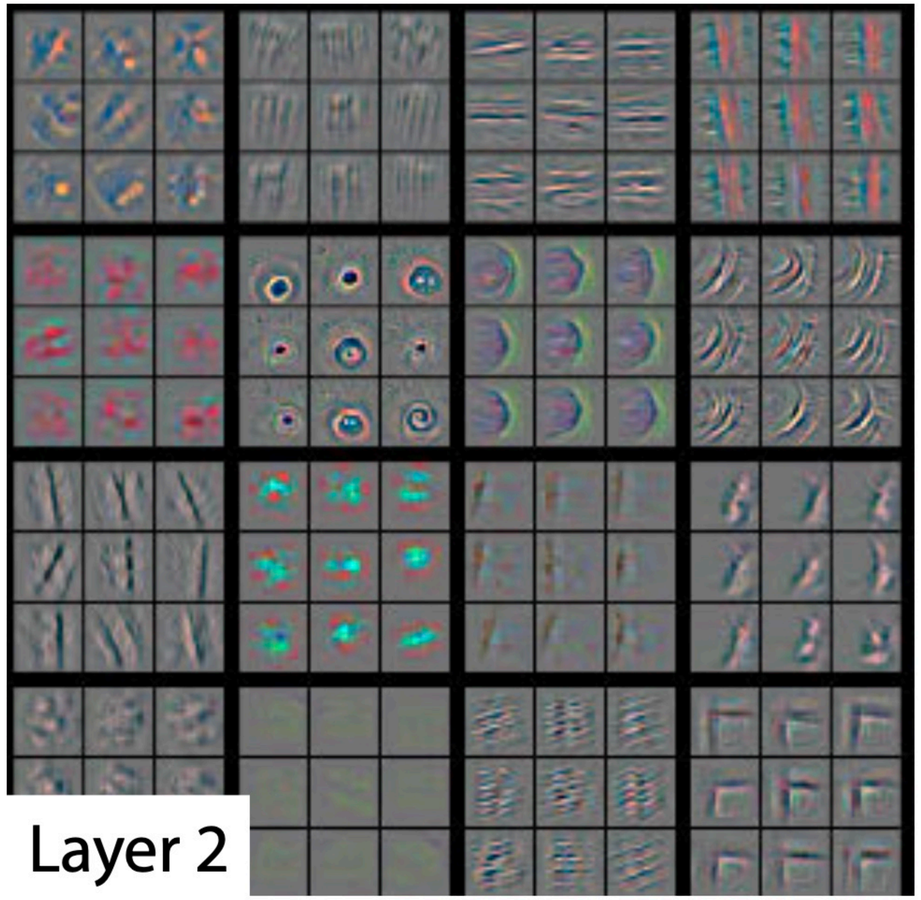
$$f_W(x) = W \circ \varphi(x)$$

$$f_{W,\theta}(x) = W \circ \varphi_\theta(x)$$

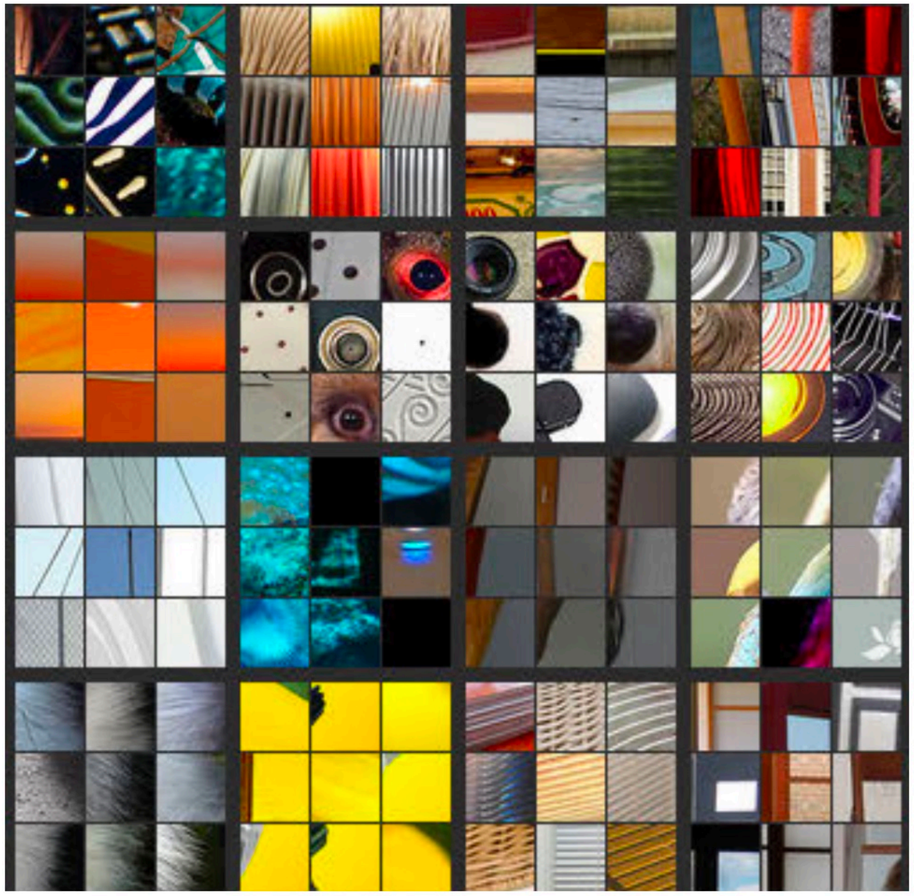
ZEILER & FERGUS, 2013

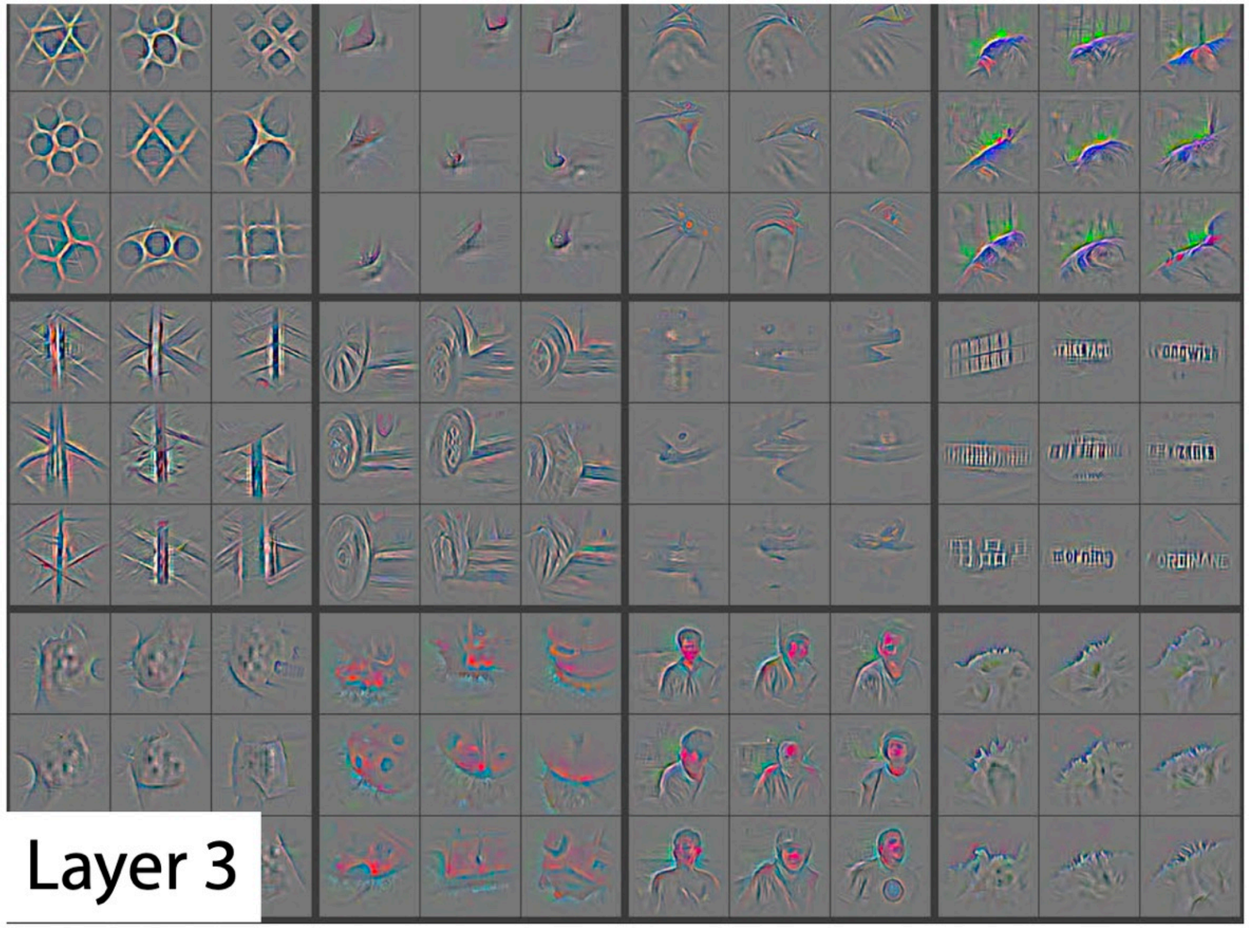


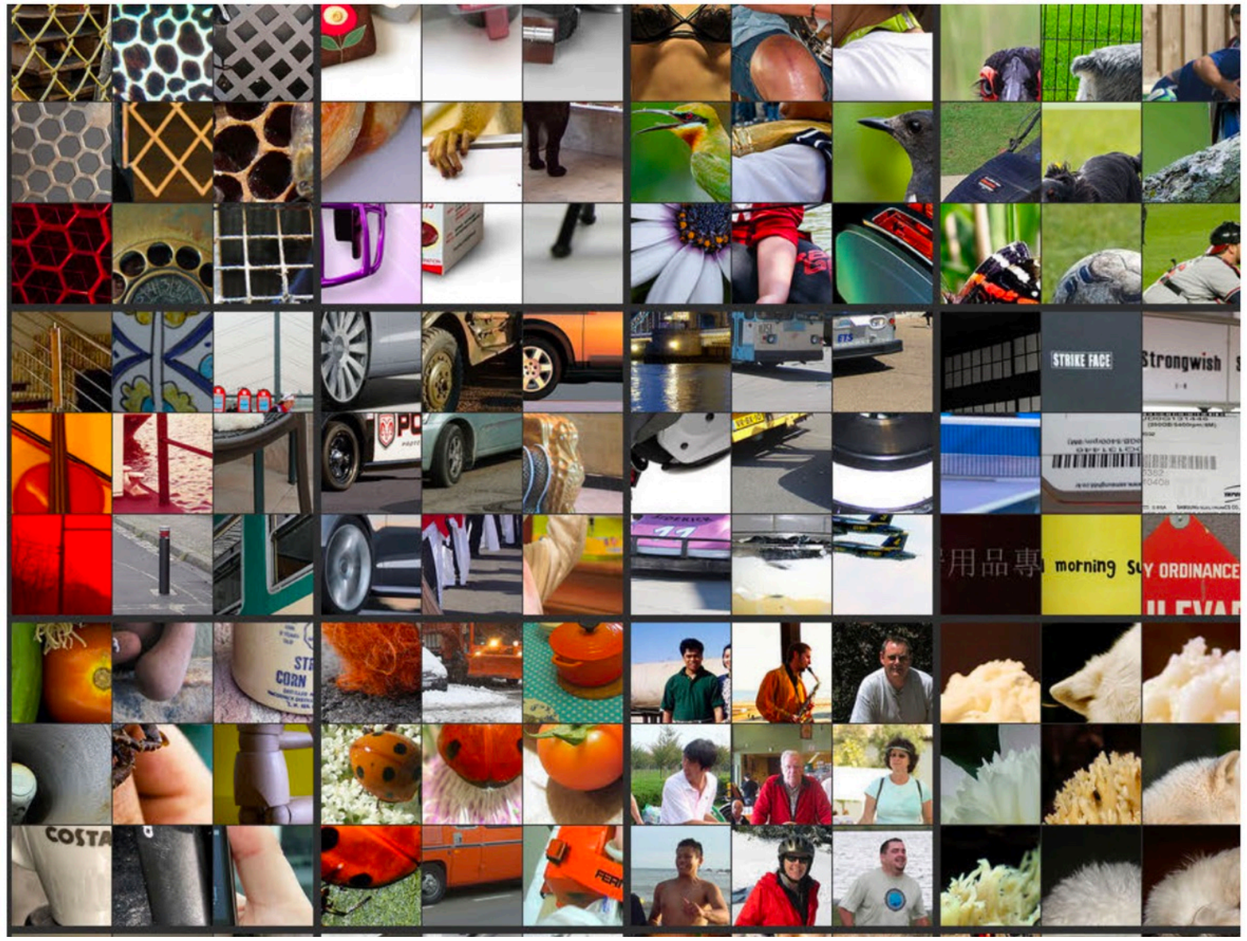
Layer 1



Layer 2




Layer 3


STRIKE FACE
Strongwish
morning
ORDINANCE
COSTA

2013

Feature Learning in Deep Neural Networks – Studies on Speech Recognition Tasks

Dong Yu, Michael L. Seltzer, Jinyu Li[1], Jui-Ting Huang[1], Frank Seide[2]
Microsoft Research, Redmond, WA 98052
[1]Microsoft Corporation, Redmond, WA 98052
[2]Microsoft Research Asia, Beijing, P.R.C.
{dongyu,mseltzer,jinyli,jthuang,fseide}@microsoft.com

Chen et al. 2022

Nichani et al. 2023

Cui et al. 2024

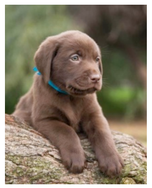
A series of works from Misha Belkin, Cengiz, and many others...

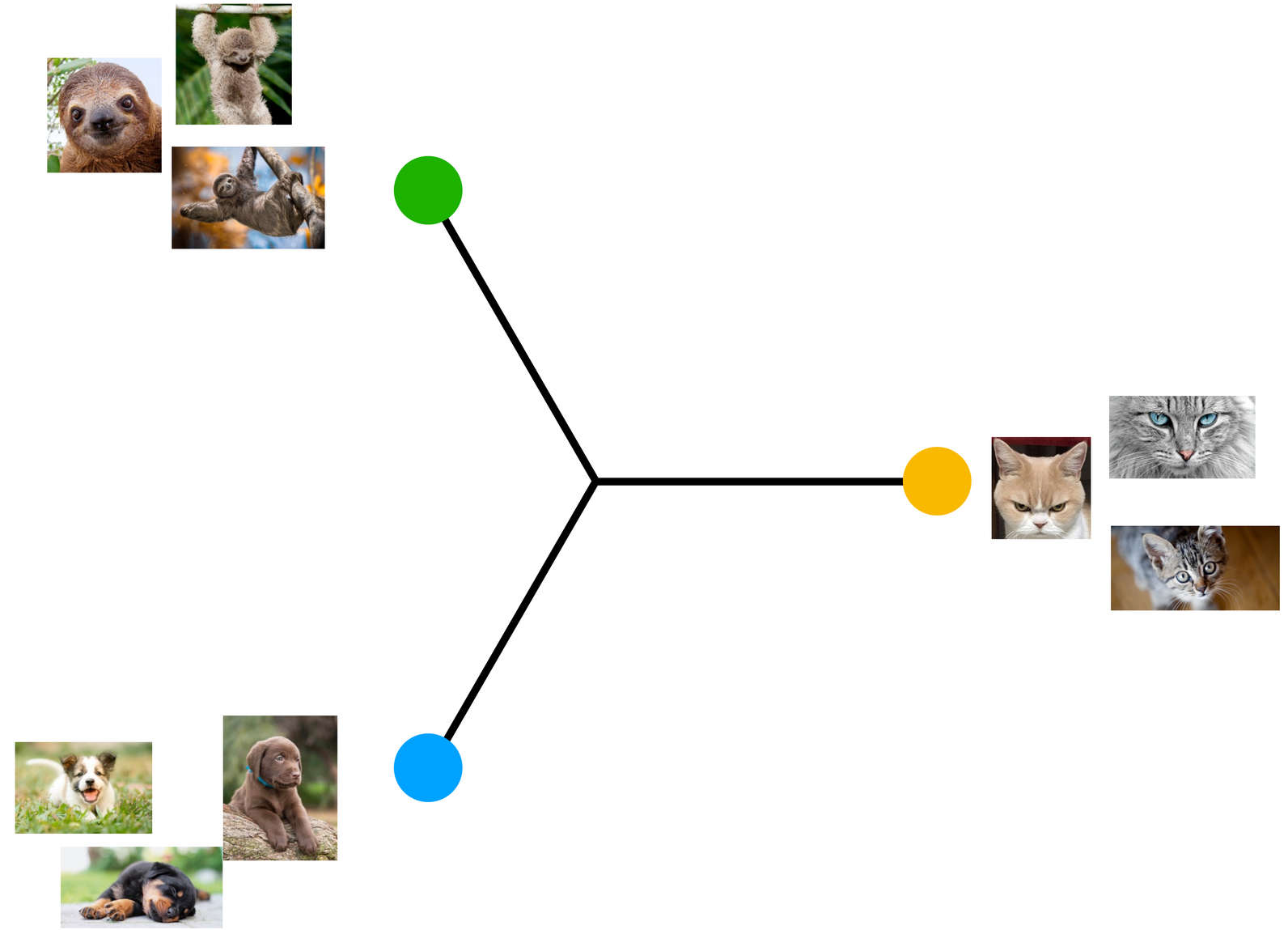
...





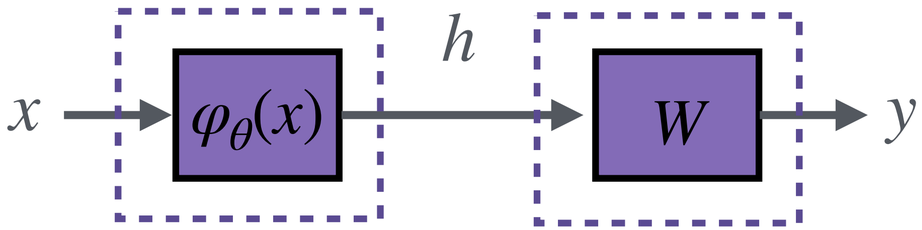






Papyan, Han, Donoho, “Prevalence of neural collapse during the terminal phase of deep learning training,” PNAS 2020

NEURAL COLLAPSE

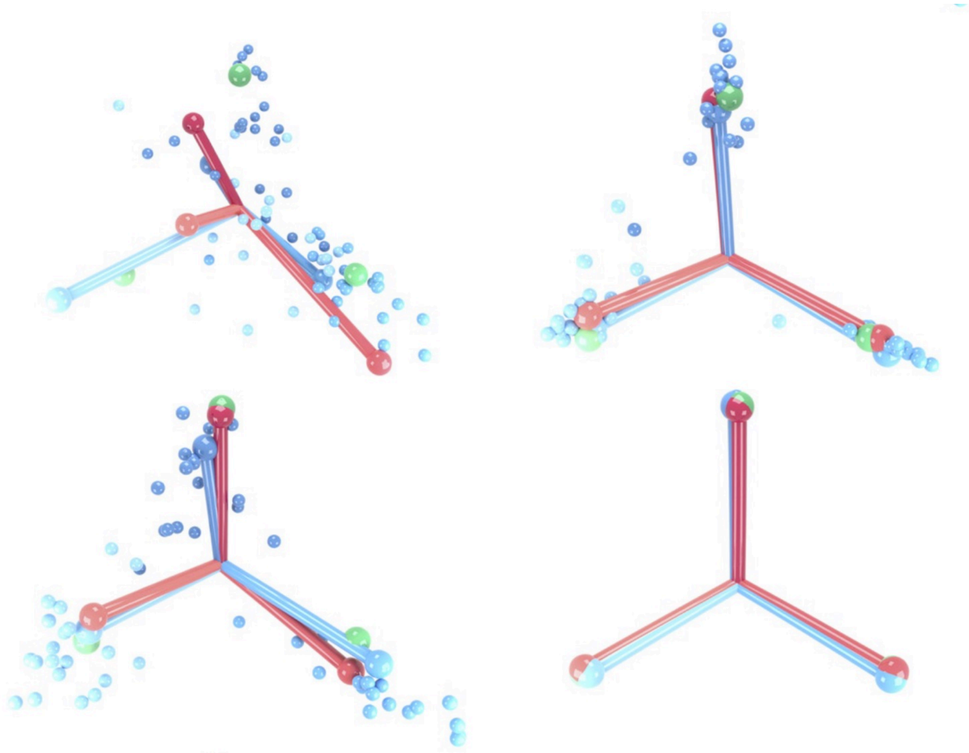


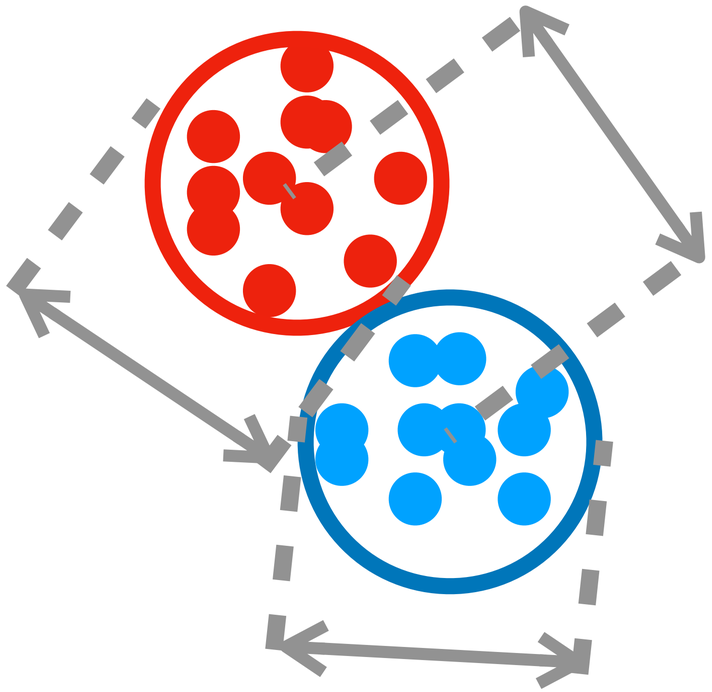
K classes, each with N_k examples

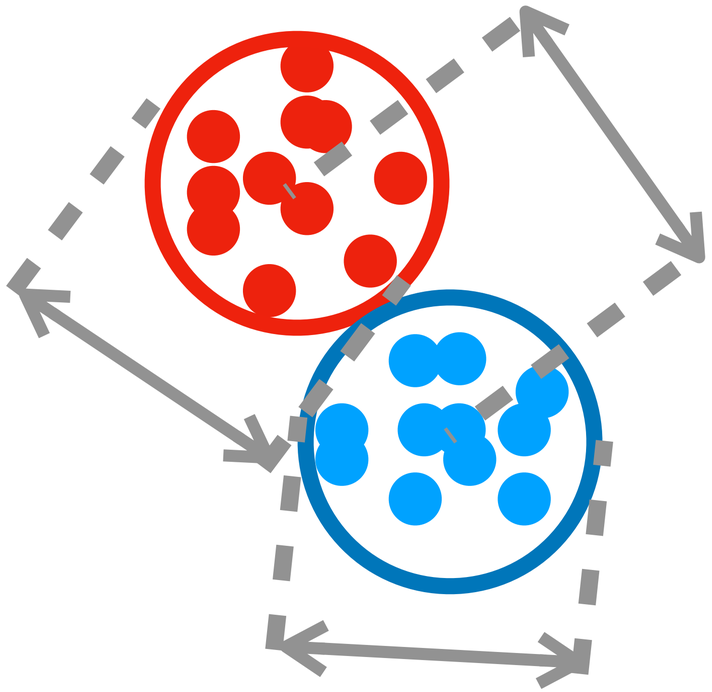
$$\Sigma_k = \mathrm{Cov}(h_{kn})_{n=1}^{N_k} \qquad \mu_k = \mathrm{Ave}_n(h_{kn})$$

$$\Sigma_w = \mathrm{Ave}_k(\Sigma_k) \qquad \Sigma_b = \mathrm{Cov}(\mu_k)$$

$$\mathrm{Tr}(\Sigma_w \Sigma_b^+) \to 0$$

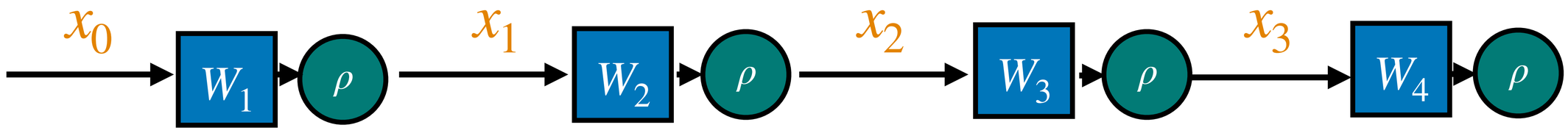


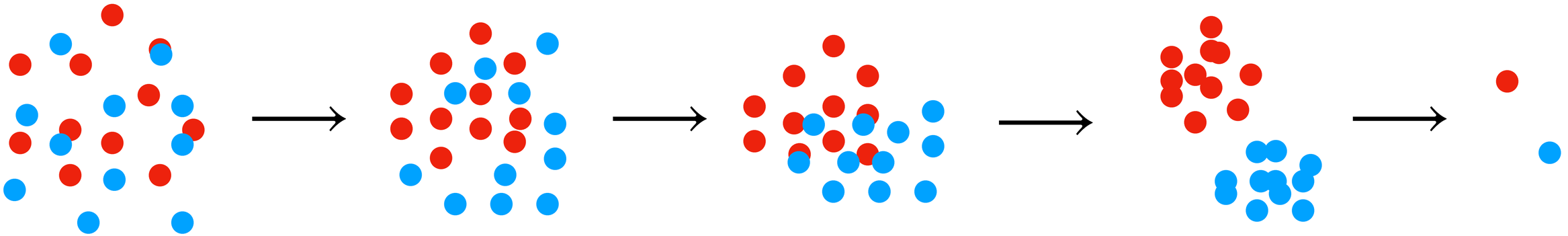




$$D = \log \frac{\text{within-class variance}}{\text{between-class variance}}$$

DEEP NEURAL COLLAPSE





[Papyan, Han, Donoho, (PNAS 2022)]

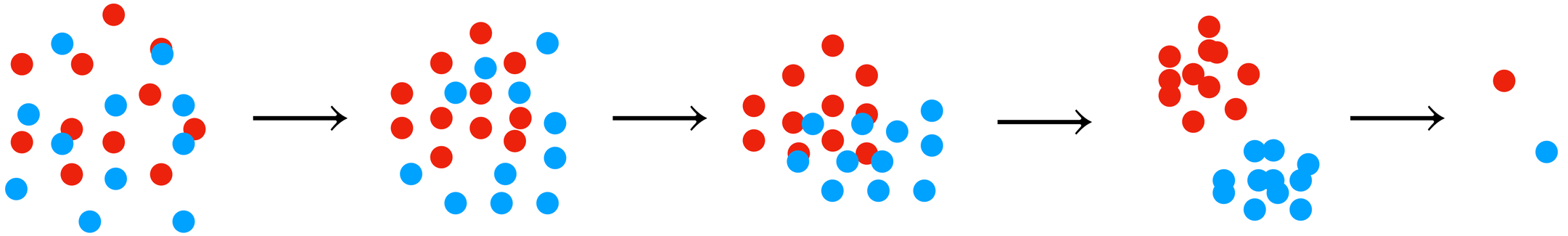
[Súkeník, Mondelli, Lampert, Nuemr (NeurIPS 2023)]

[Rangamani et al. (ICML 2023)]

...

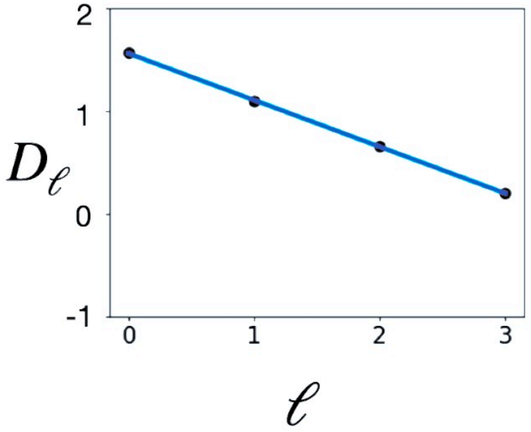
A LAW OF DATA SEPARATION

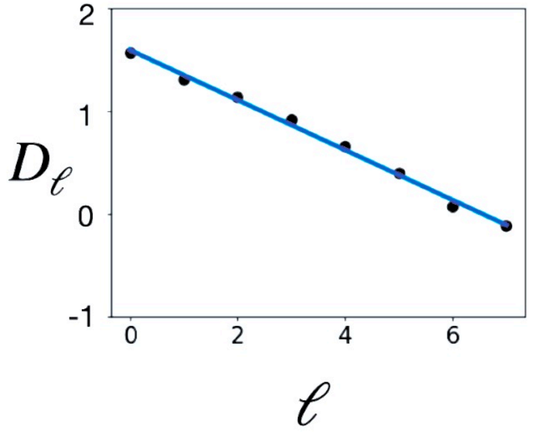
HE, SU; PNAS 2023

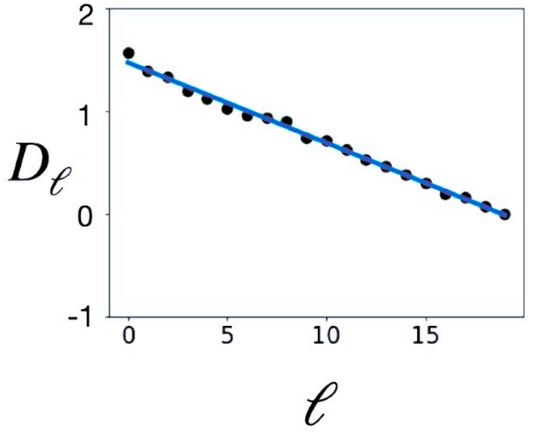


SGD / Adam

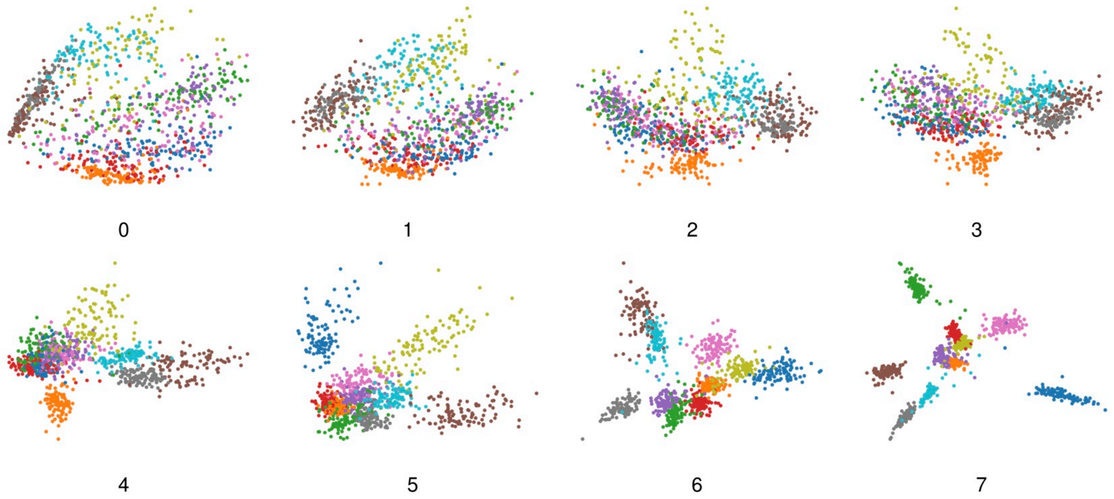
MNIST, CIFAR10

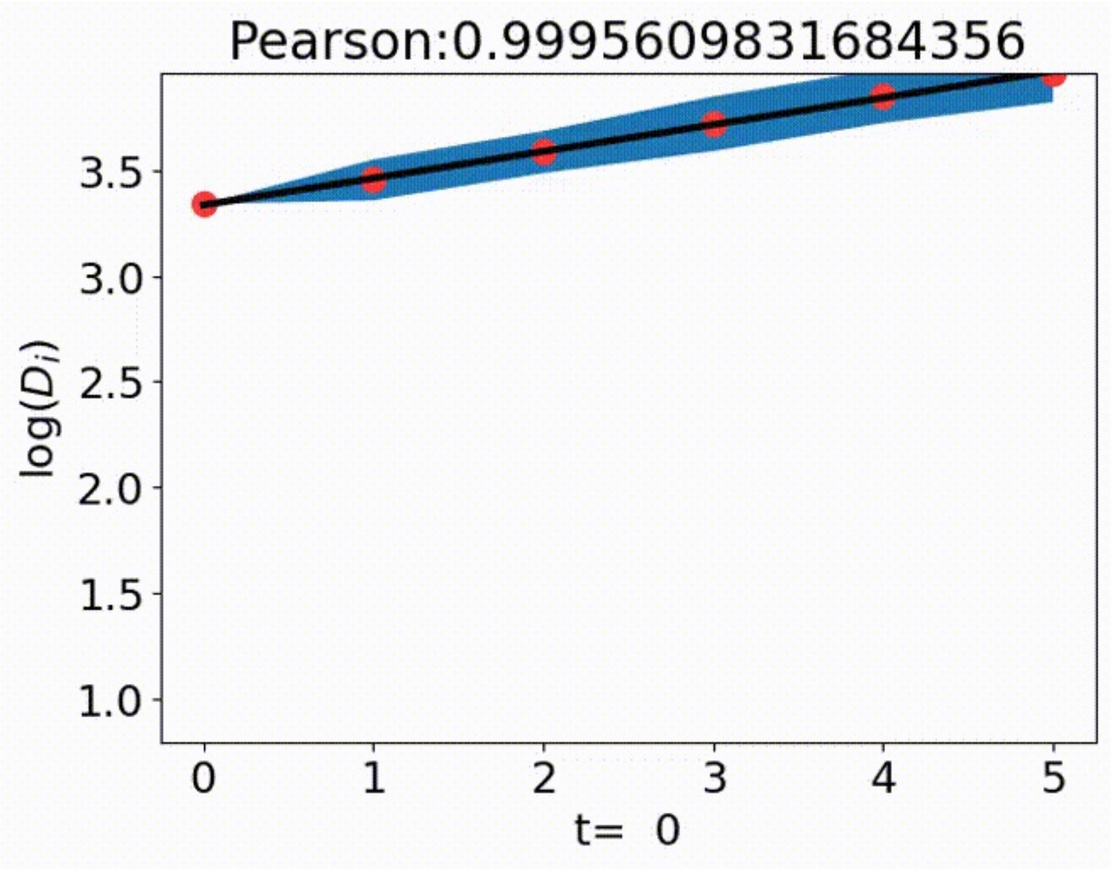




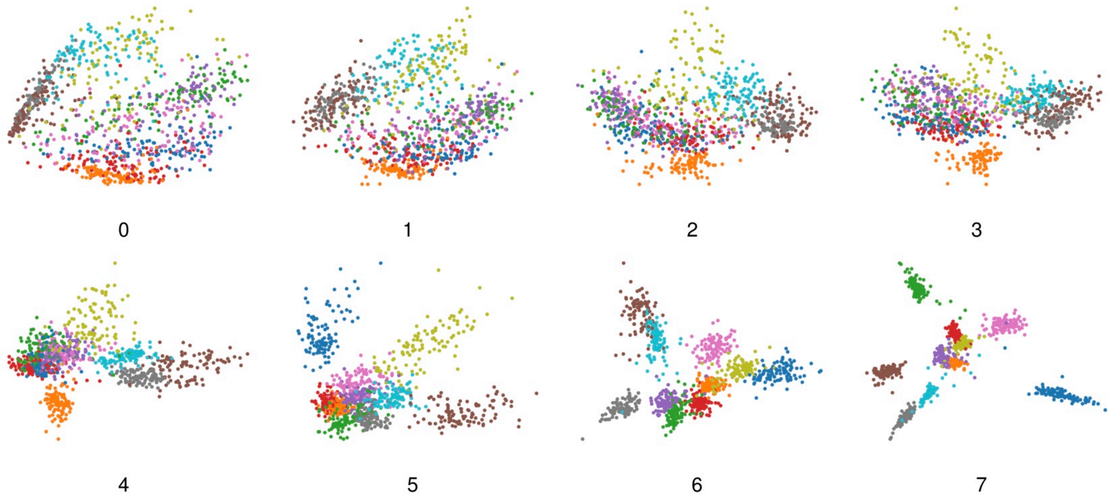


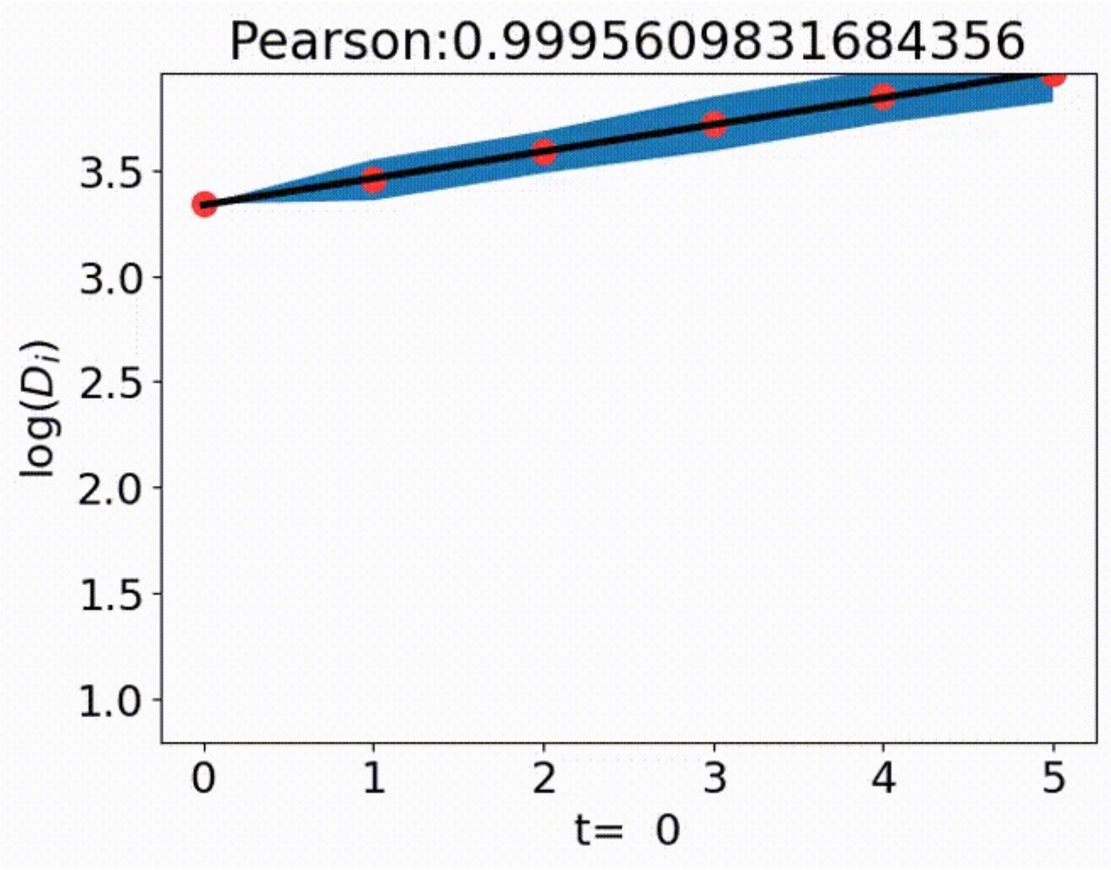
$D_\ell = \text{trace } \Sigma_w \Sigma_b^+$ or $D_\ell = \text{trace } \Sigma_w / \text{trace} \Sigma_b$





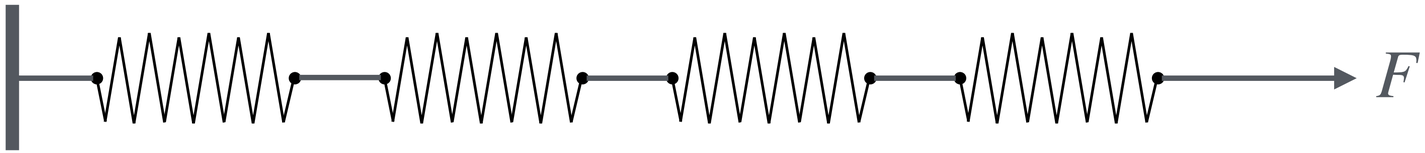
$D_\ell = \text{trace } \Sigma_w \Sigma_b^+$ or $D_\ell = \text{trace } \Sigma_w / \text{trace} \Sigma_b$





A 1ST MECHANICAL ANALOGY

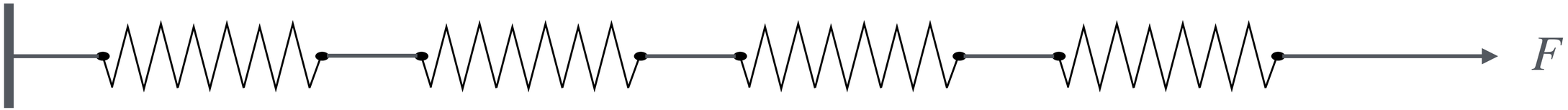
EQUISEPARATION IS SOMEHOW A LINEAR PHENOMENON



$$h_L = \frac{LF}{k} \qquad w_\ell = \frac{F}{k}$$

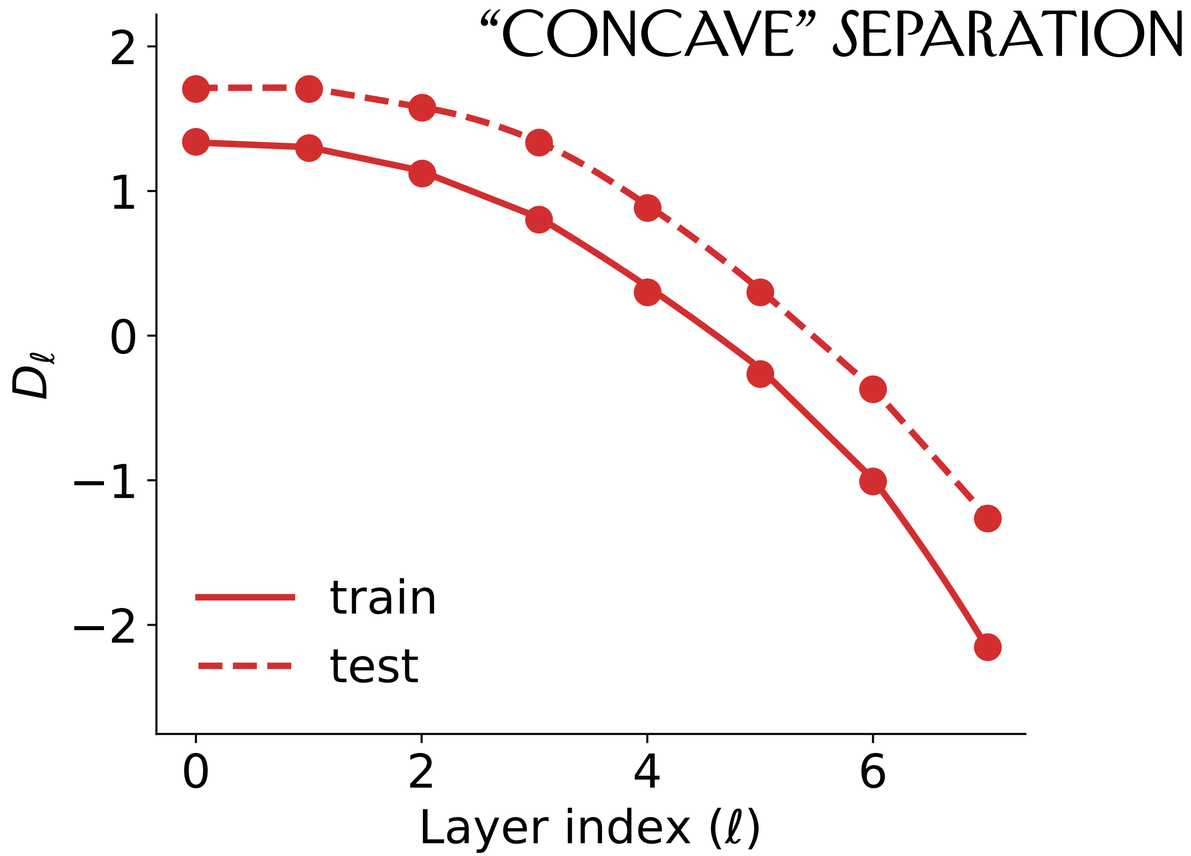
A MECHANICAL ANALOGY

EQUISEPARATION IS *SOMEHOW* A LINEAR PHENOMENON



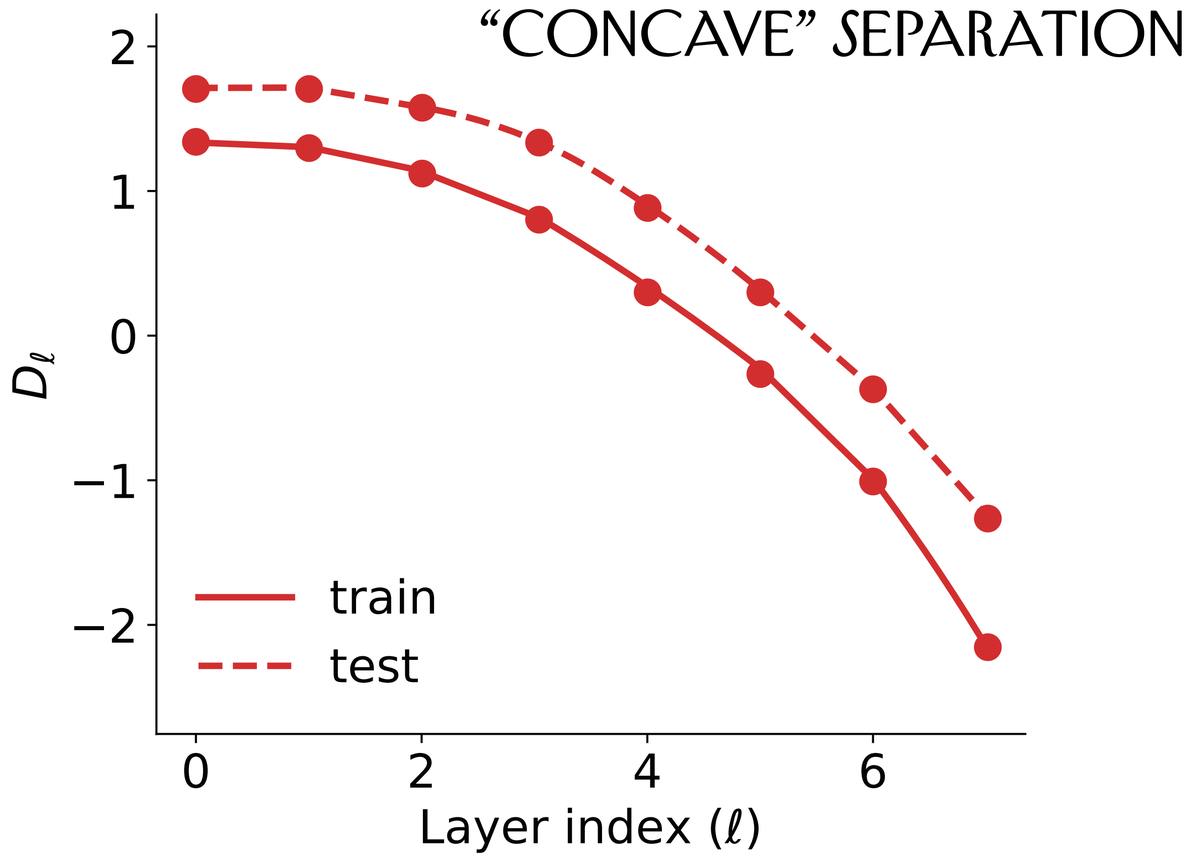
$$h_L = \frac{LF}{k} \qquad w_\ell = \frac{F}{k}$$

A LAW OF DATA SEPARATION?



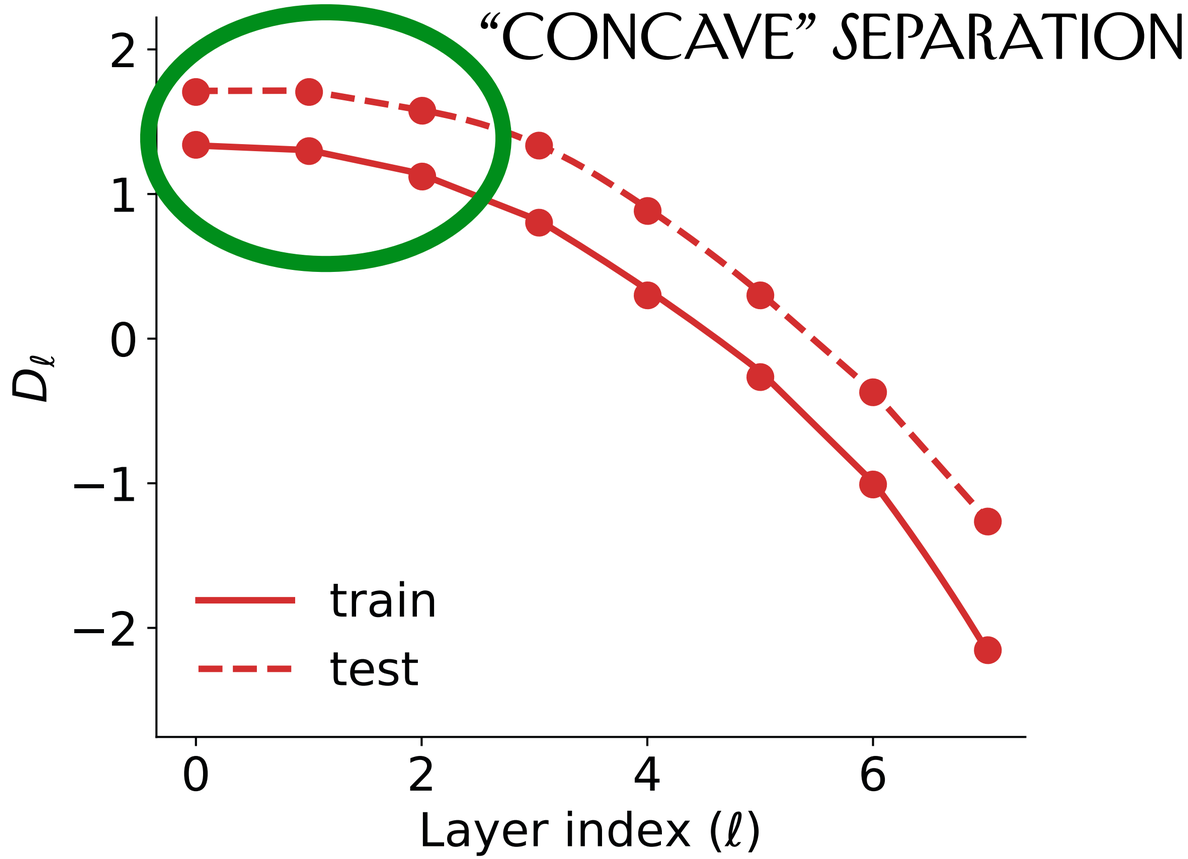
8-layer ReLU ResNet | MNIST | 10000 examples | full batch | SGD | lr = 0.0001, acc = 82.49%

A LAW OF DATA SEPARATION?



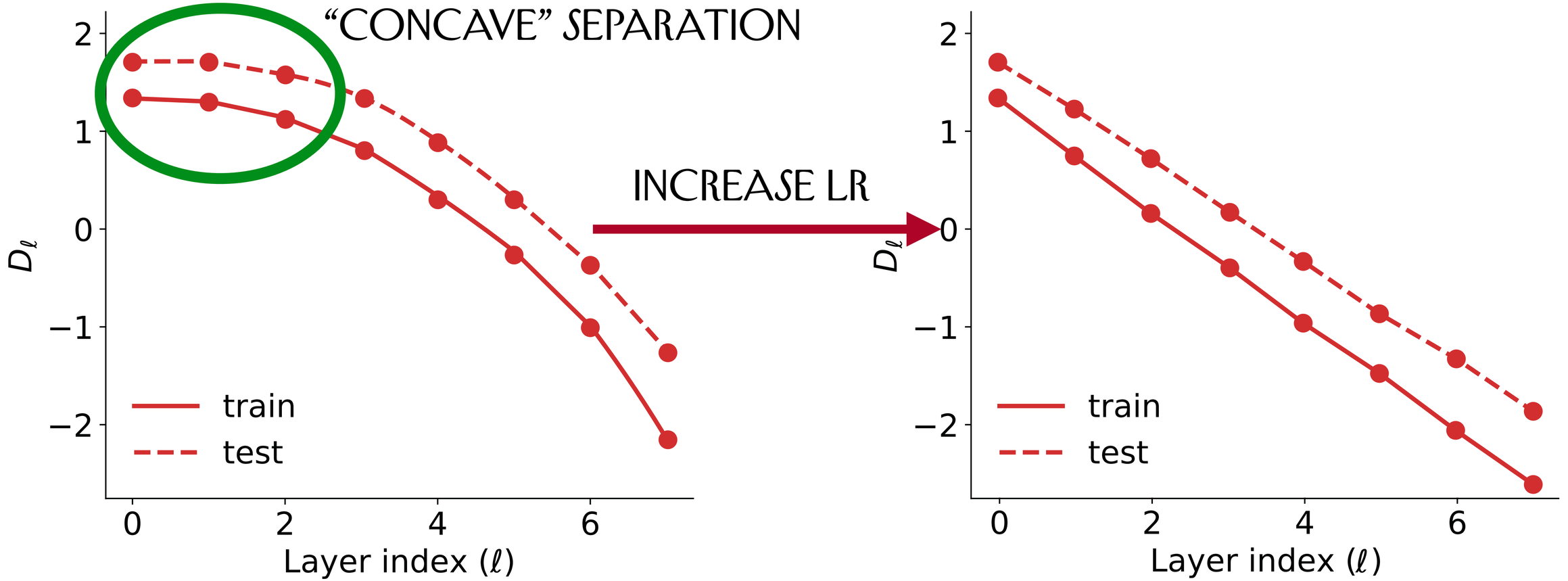
8-layer ReLU ResNet | MNIST | 10000 examples | full batch | SGD | lr = 0.0001, acc = 82.49%

A LAW OF DATA SEPARATION?



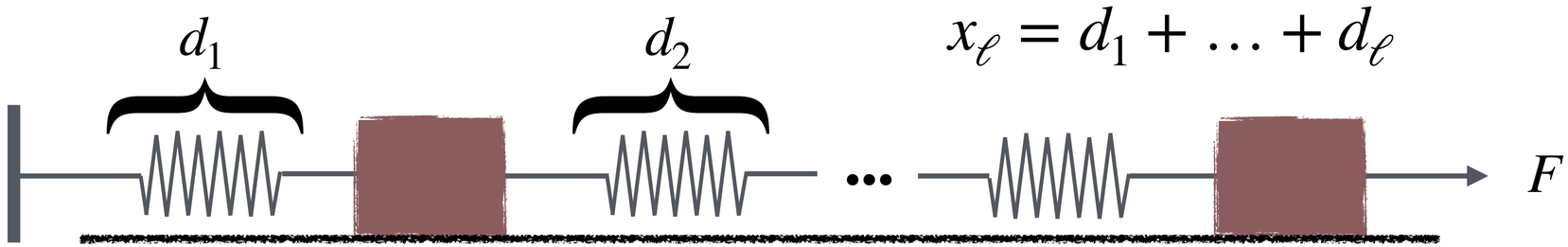
8-layer ReLU ResNet | MNIST | 10000 examples | full batch | SGD | lr = 0.0001, acc = 82.49%

A LAW OF DATA SEPARATION?

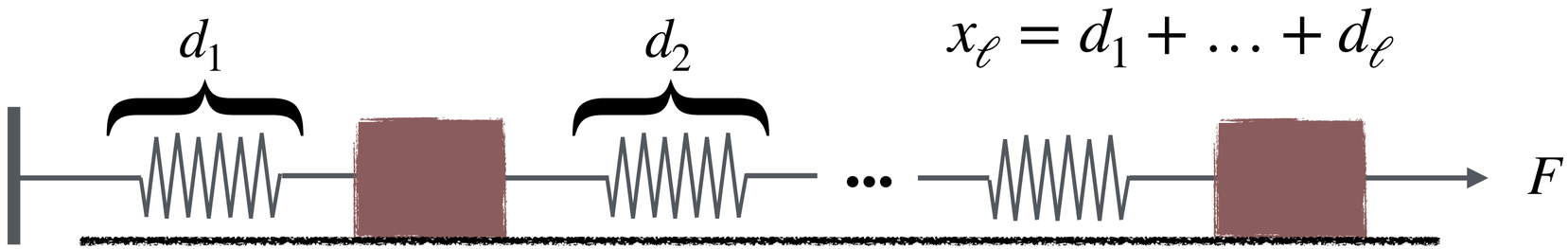


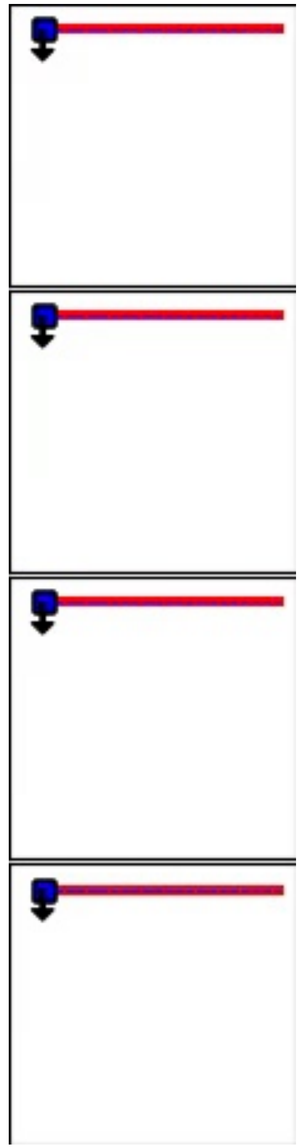
8-layer ReLU ResNet | MNIST | 10000 examples | full batch | SGD | lr = 0.0001, acc = 82.49%

$$\ddot{x}_\ell = k(d_{\ell+1} - d_\ell) - \gamma \dot{x}_\ell - f_\ell$$

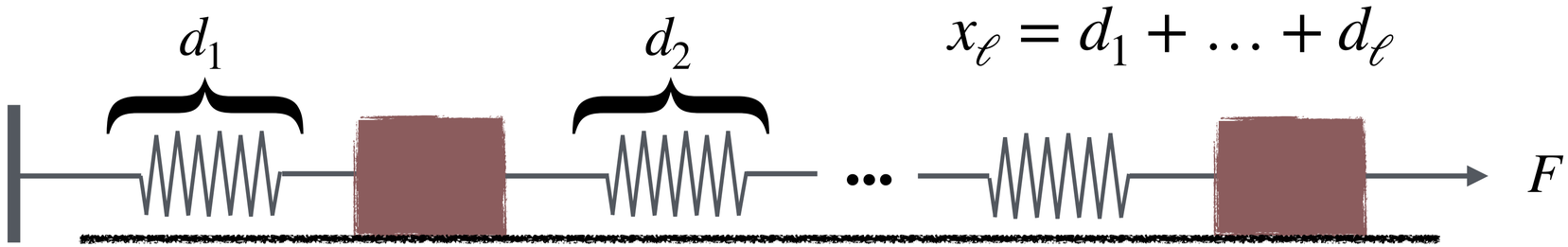


$$\ddot{x}_\ell = k(d_{\ell+1} - d_\ell) - \gamma \dot{x}_\ell - f_\ell$$

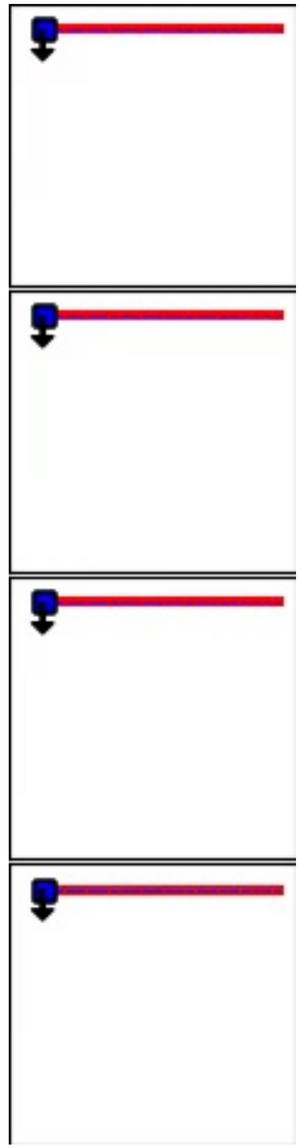




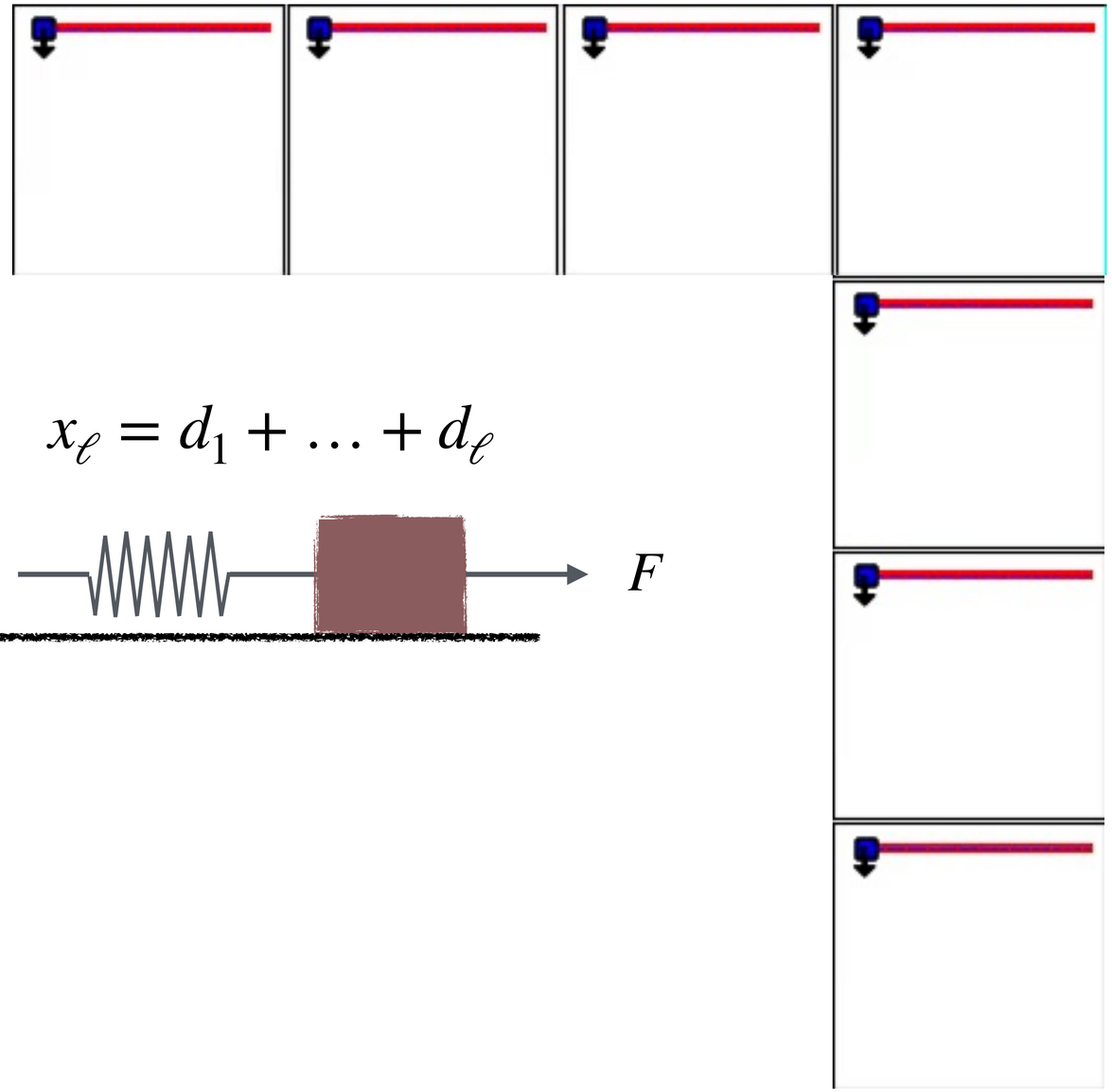
$$\ddot{x}_\ell = k(d_{\ell+1} - d_\ell) - \gamma \dot{x}_\ell - f_\ell$$

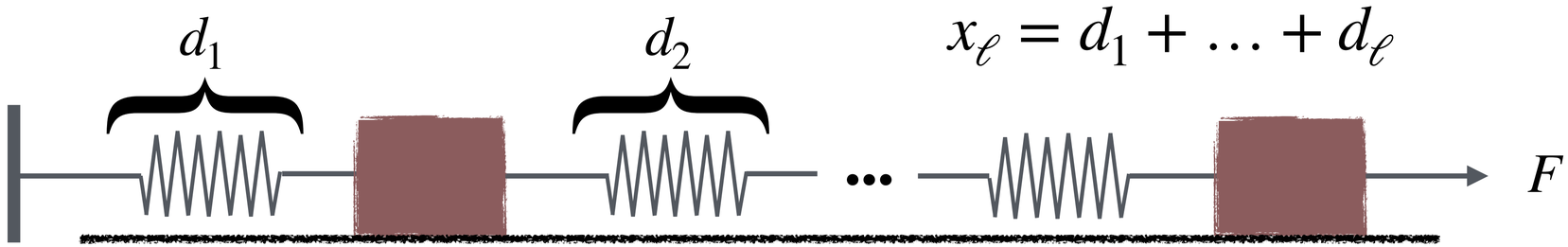


$$\ddot{x}_\ell = k(d_{\ell+1} - d_\ell) - \gamma \dot{x}_\ell - f_\ell + \sigma \xi_\ell$$



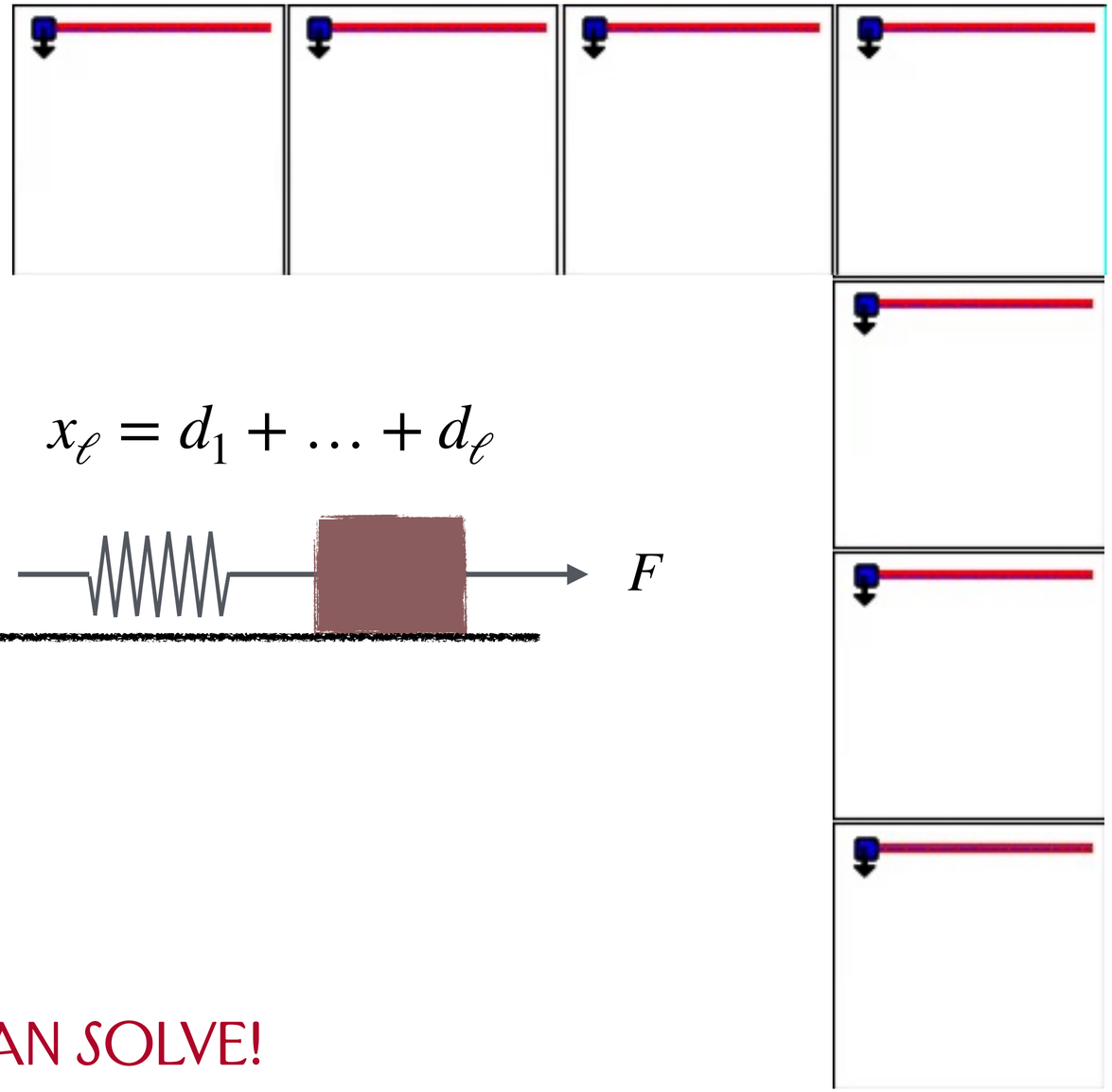
$$\ddot{x}_\ell = k(d_{\ell+1} - d_\ell) - \gamma \dot{x}_\ell - f_\ell$$

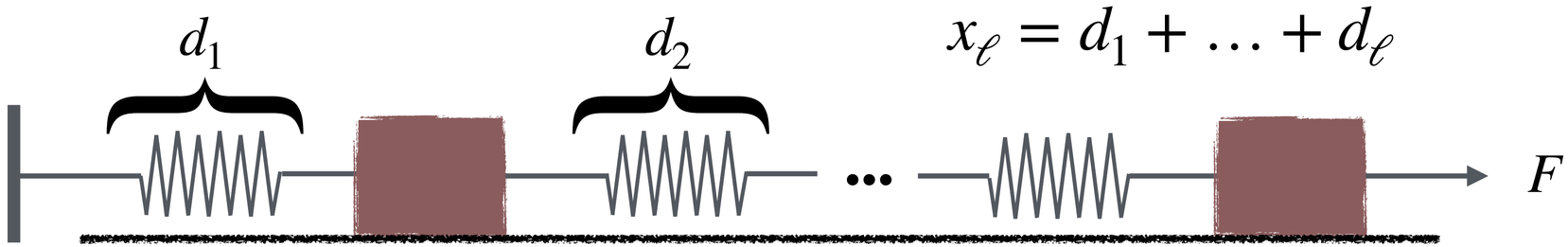




$$\ddot{x}_\ell = k(d_{\ell+1} - d_\ell) - \gamma \dot{x}_\ell - f_\ell + \sigma \xi_\ell$$

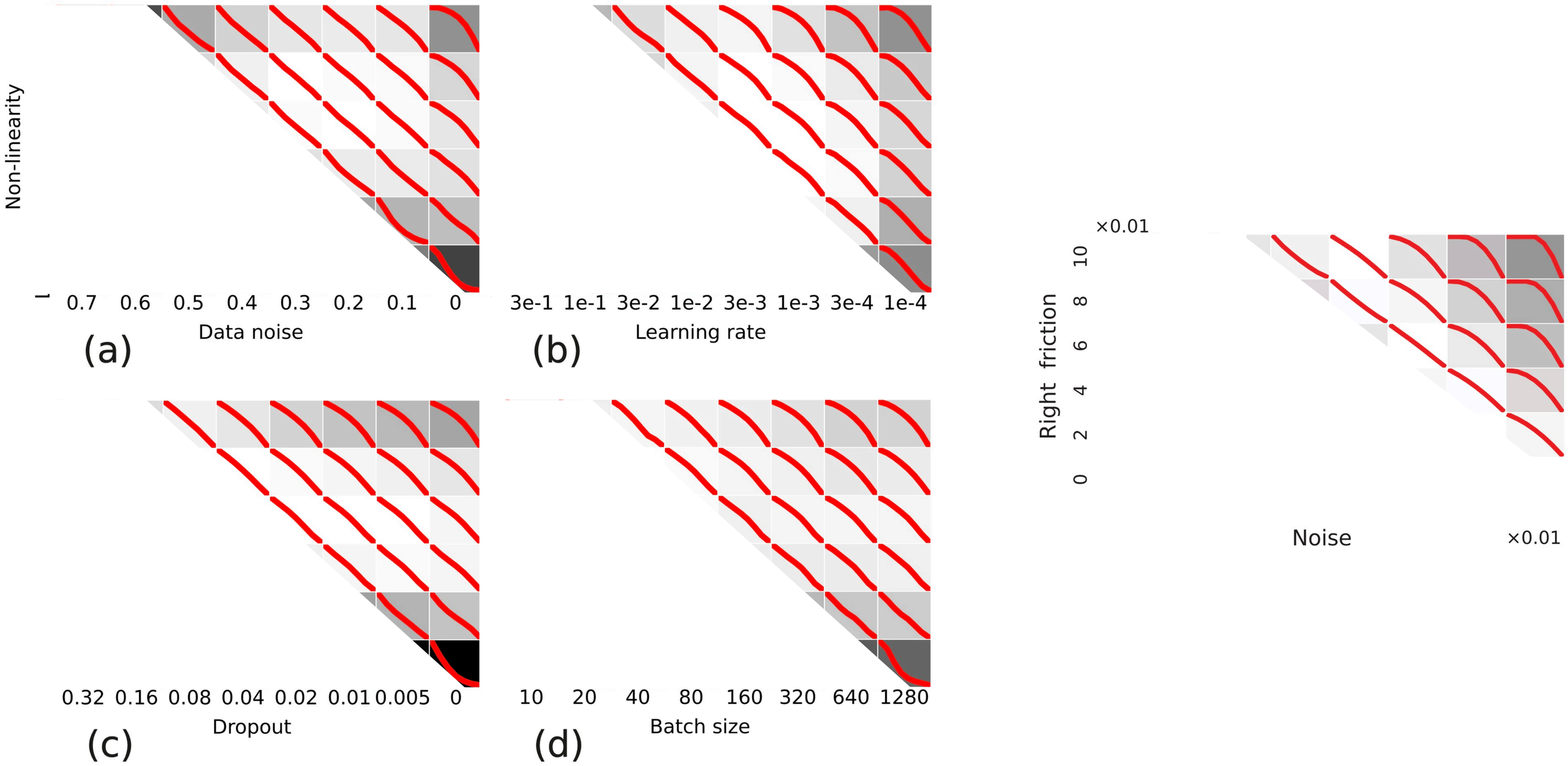
$$\ddot{x}_\ell = k(d_{\ell+1} - d_\ell) - \gamma \dot{x}_\ell - f_\ell$$

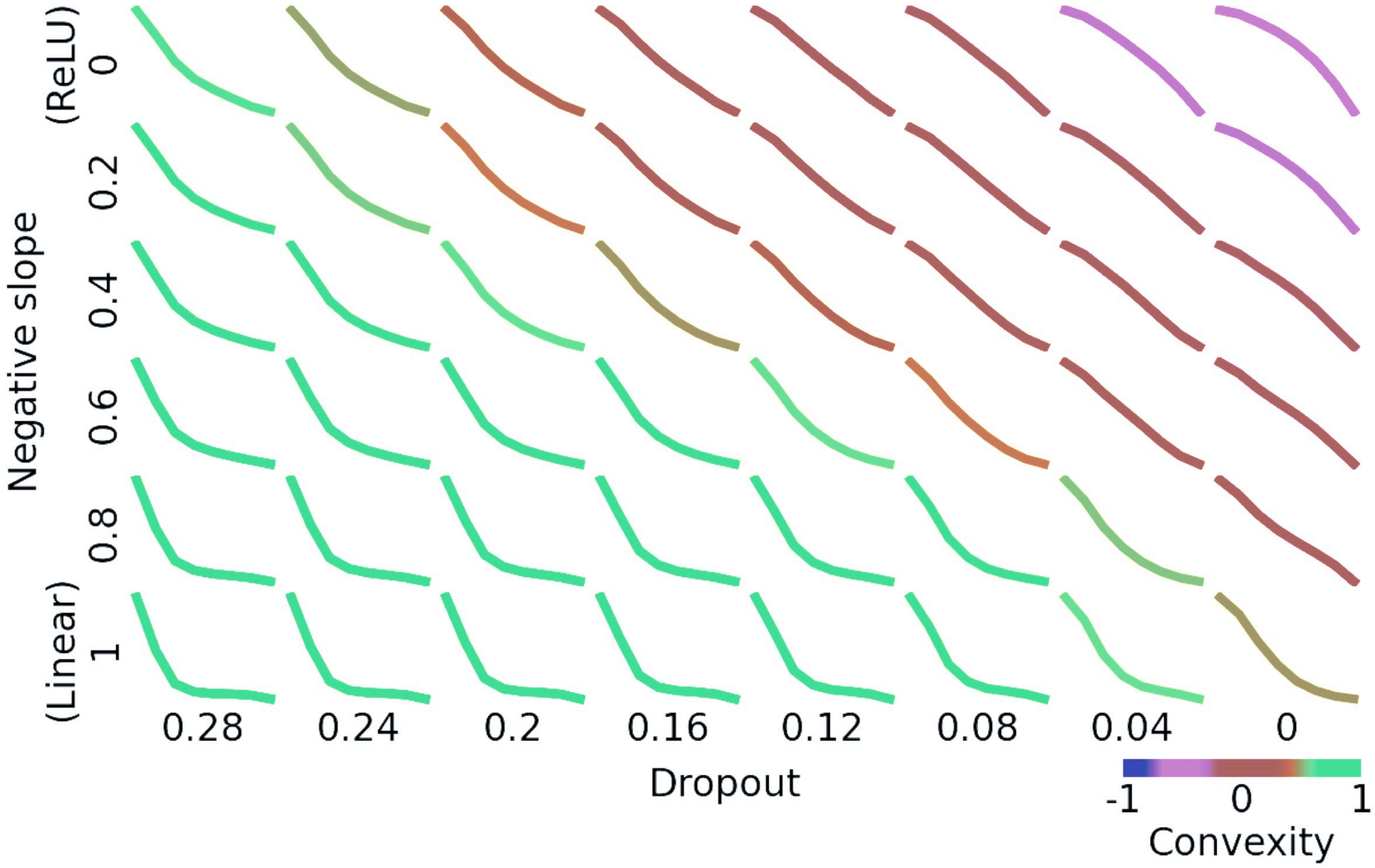




$$\ddot{x}_\ell = k(d_{\ell+1} - d_\ell) - \gamma \dot{x}_\ell - f_\ell + \sigma \xi_\ell$$

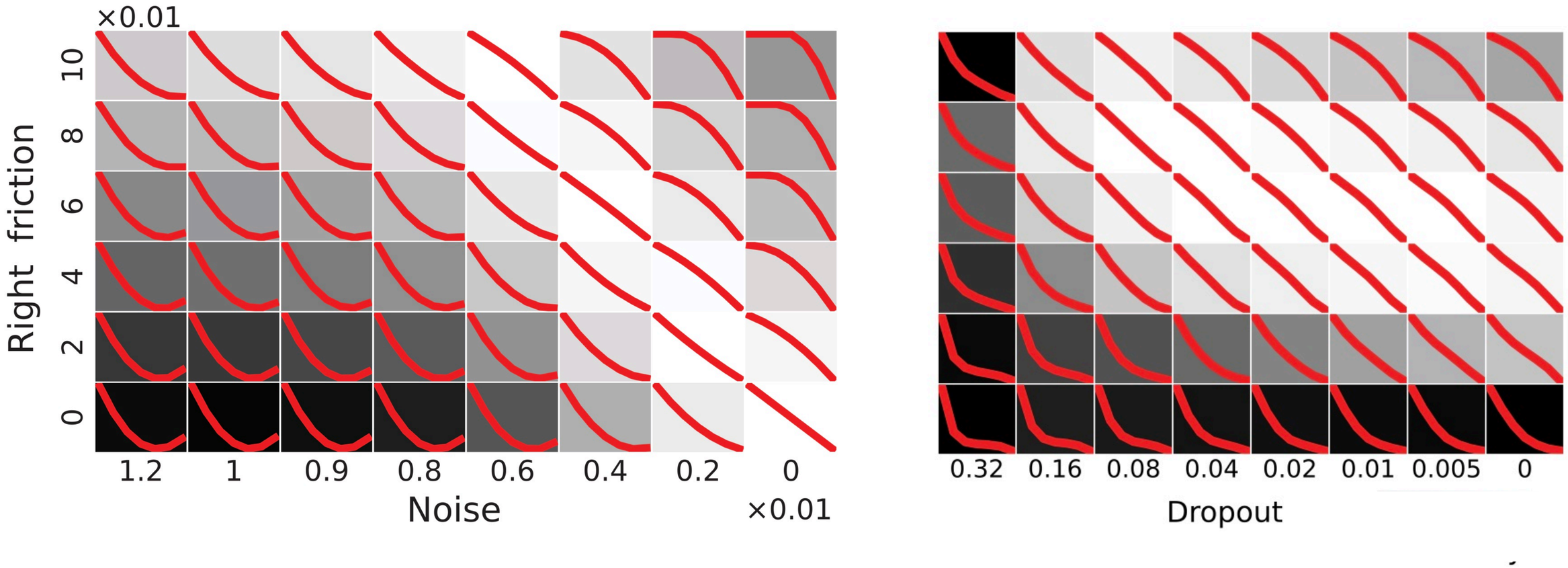
A LOW DIMENSIONAL MODEL THAT WE CAN SOLVE!


Non-linearity
1
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
Data noise
(a)
3e-1 1e-1 3e-2 1e-2 3e-3 1e-3 3e-4 1e-4
Learning rate
(b)
0.32 0.16 0.08 0.04 0.02 0.01 0.005 0
Dropout
(c)
10 20 40 80 160 320 640 1280
Batch size
(d)
Right friction
×0.01
0 2 4 6 8 10
Noise
×0.01

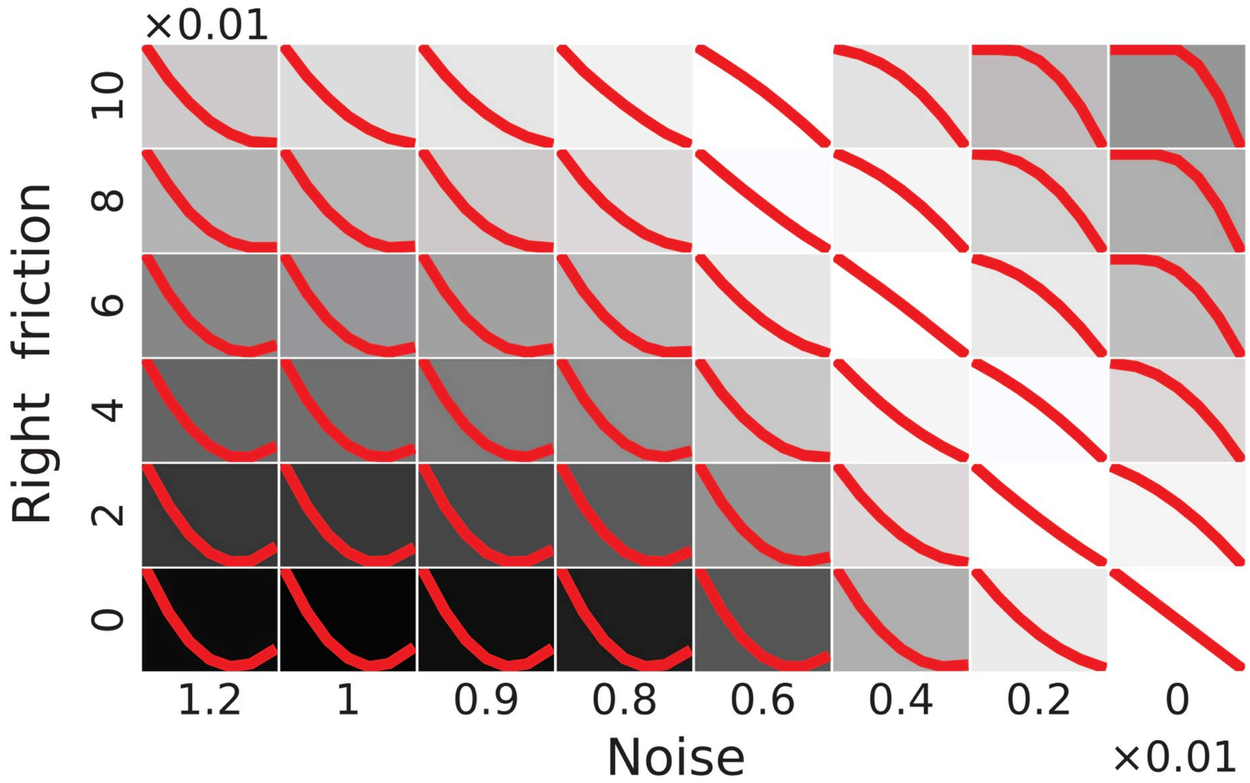

Negative slope
(ReLU)
0
0.2
0.4
0.6
0.8
1
(Linear)
0.28
0.24
0.2
0.16
0.12
0.08
0.04
0
Dropout
-1
0
1
Convexity

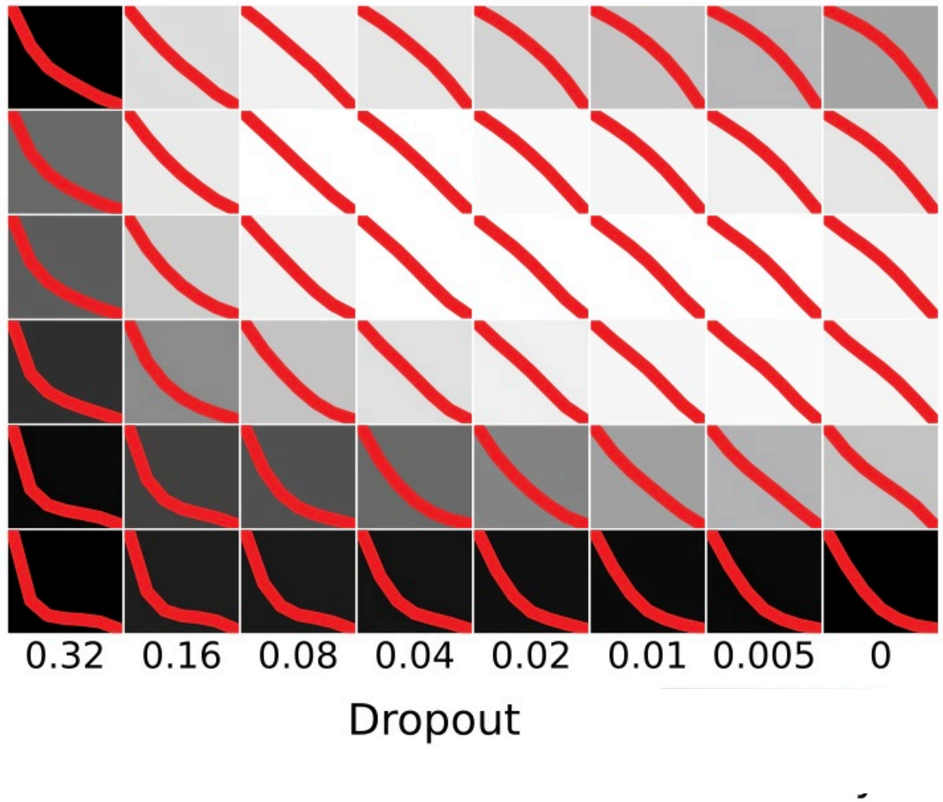
FORWARD VS BACKWARD PASS

$$F_\ell = \begin{cases} k(w_{\ell+1} - w_\ell) & \text{if } \dot{h}_\ell = 0 \text{ and } -\mu_r \le k(w_{\ell+1} - w_\ell) \le \mu_\ell \\ -\mu_\ell & \text{if } \dot{h}_\ell > 0 \text{ or } (\dot{h}_\ell = 0 \text{ and } k(w_{\ell+1} - w_\ell) > \mu_\ell) \\ \mu_r & \text{if } \dot{h}_\ell < 0 \text{ or } (\dot{h}_\ell = 0 \text{ and } k(w_{\ell+1} - w_\ell) < -\mu_r) \end{cases}$$

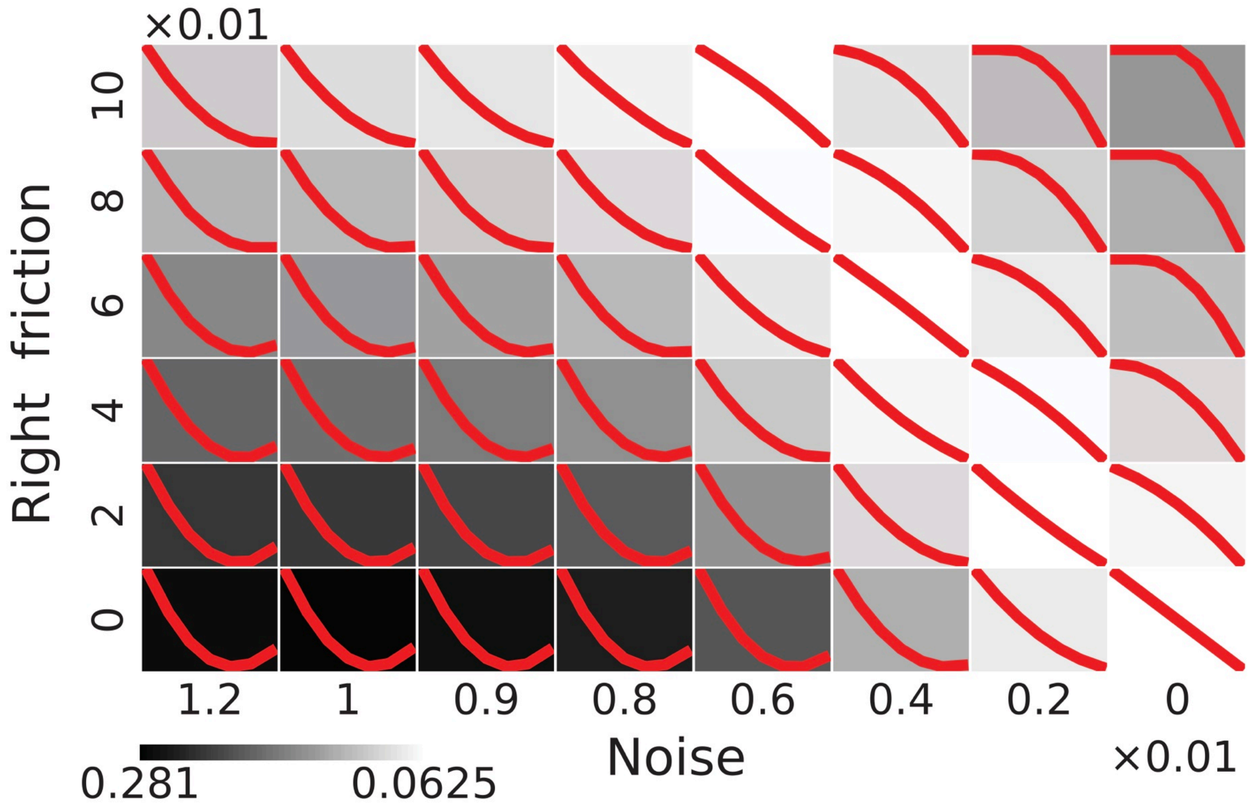

×0.01
Right friction
10
8
6
4
2
0
1.2
1
0.9
0.8
0.6
0.4
0.2
0
Noise
×0.01
0.32
0.16
0.08
0.04
0.02
0.01
0.005
0
Dropout

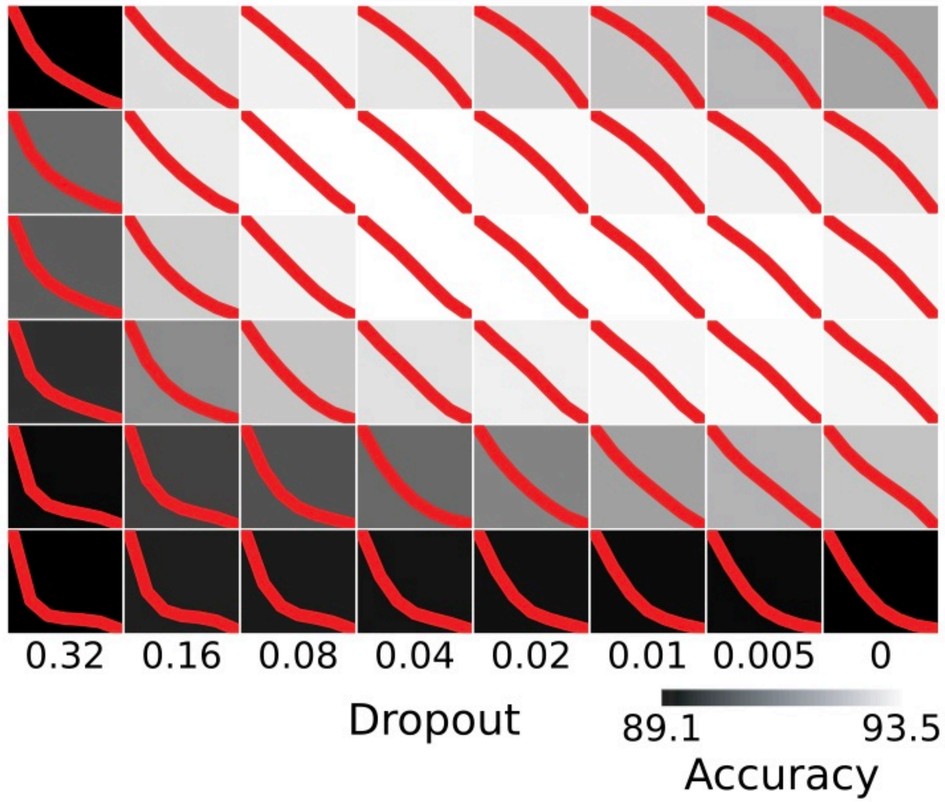
OK SO WHAT?

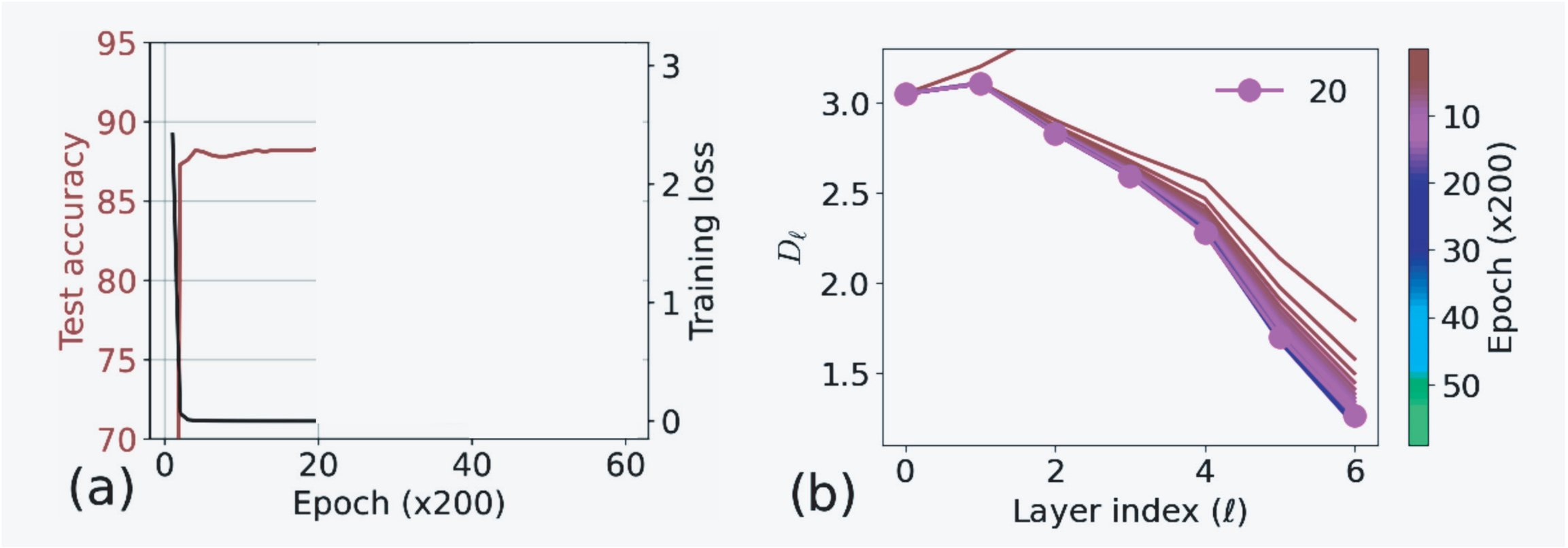

×0.01
Right friction
10
8
6
4
2
0
1.2
1
0.9
0.8
0.6
0.4
0.2
0
Noise
×0.01


0.32
0.16
0.08
0.04
0.02
0.01
0.005
0
Dropout

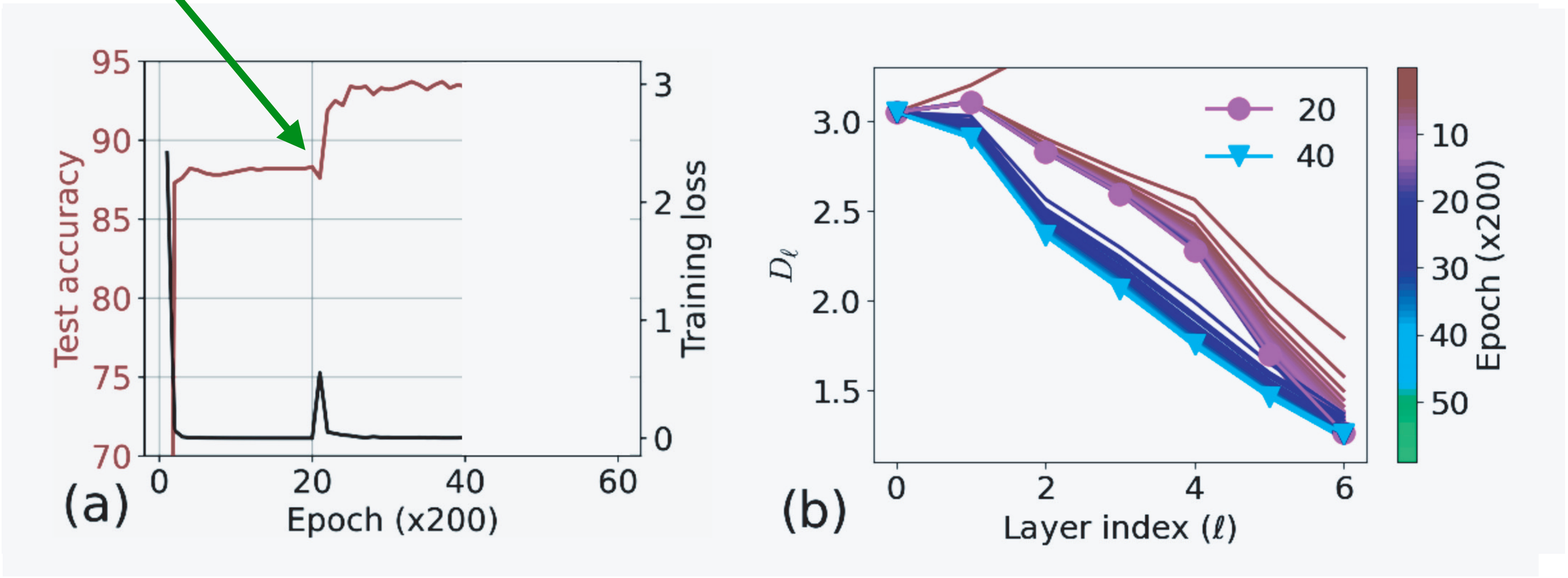
OK SO WHAT?


×0.01
Right friction
10
8
6
4
2
0
1.2
1
0.9
0.8
0.6
0.4
0.2
0
Noise
×0.01
0.281
0.0625


0.32
0.16
0.08
0.04
0.02
0.01
0.005
0
Dropout
89.1
93.5
Accuracy

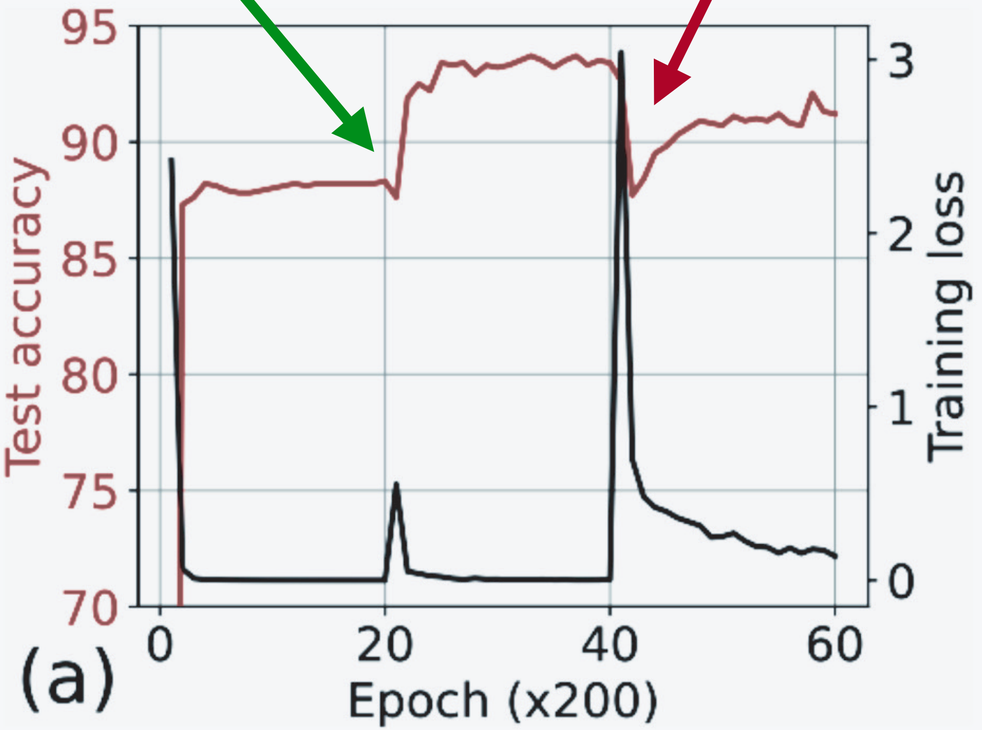

95
90
85
80
75
70
Test accuracy
3
2
1
0
Training loss
0
20
40
60
Epoch (x200)
(a)
D_ℓ
3.0
2.5
2.0
1.5
20
0
2
4
6
Layer index (ℓ)
(b)
10
20
30
40
50
Epoch (x200)

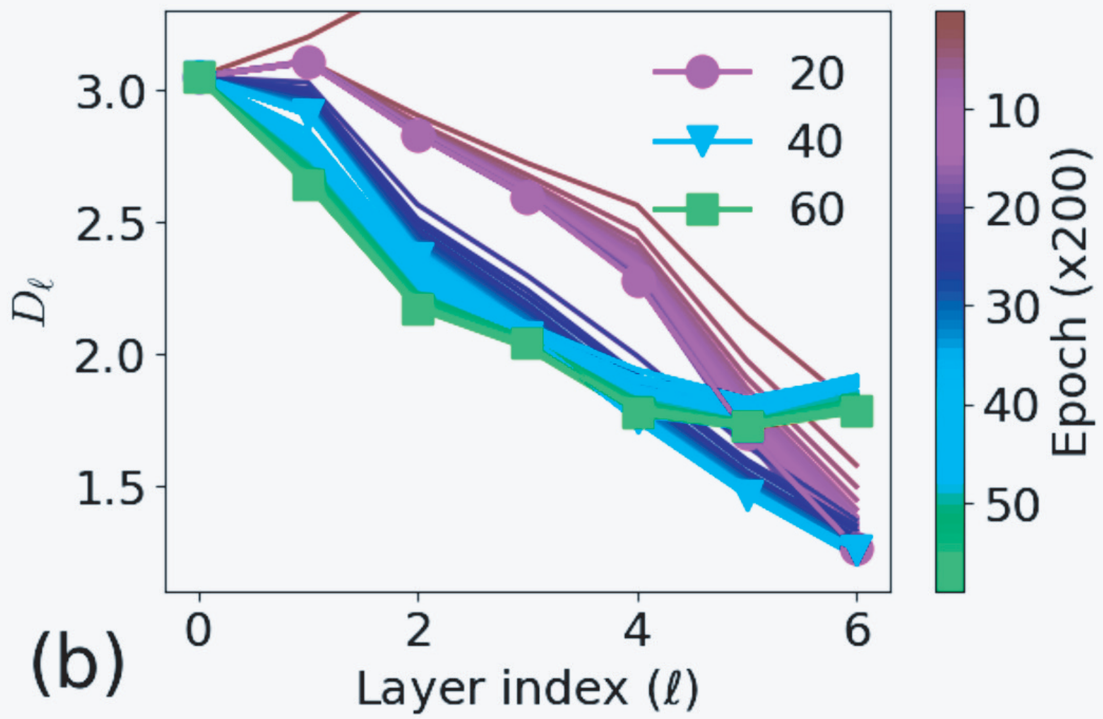
ADD 5% DROPOUT



ADD 5% DROPOUT

ADD 30% DROPOUT

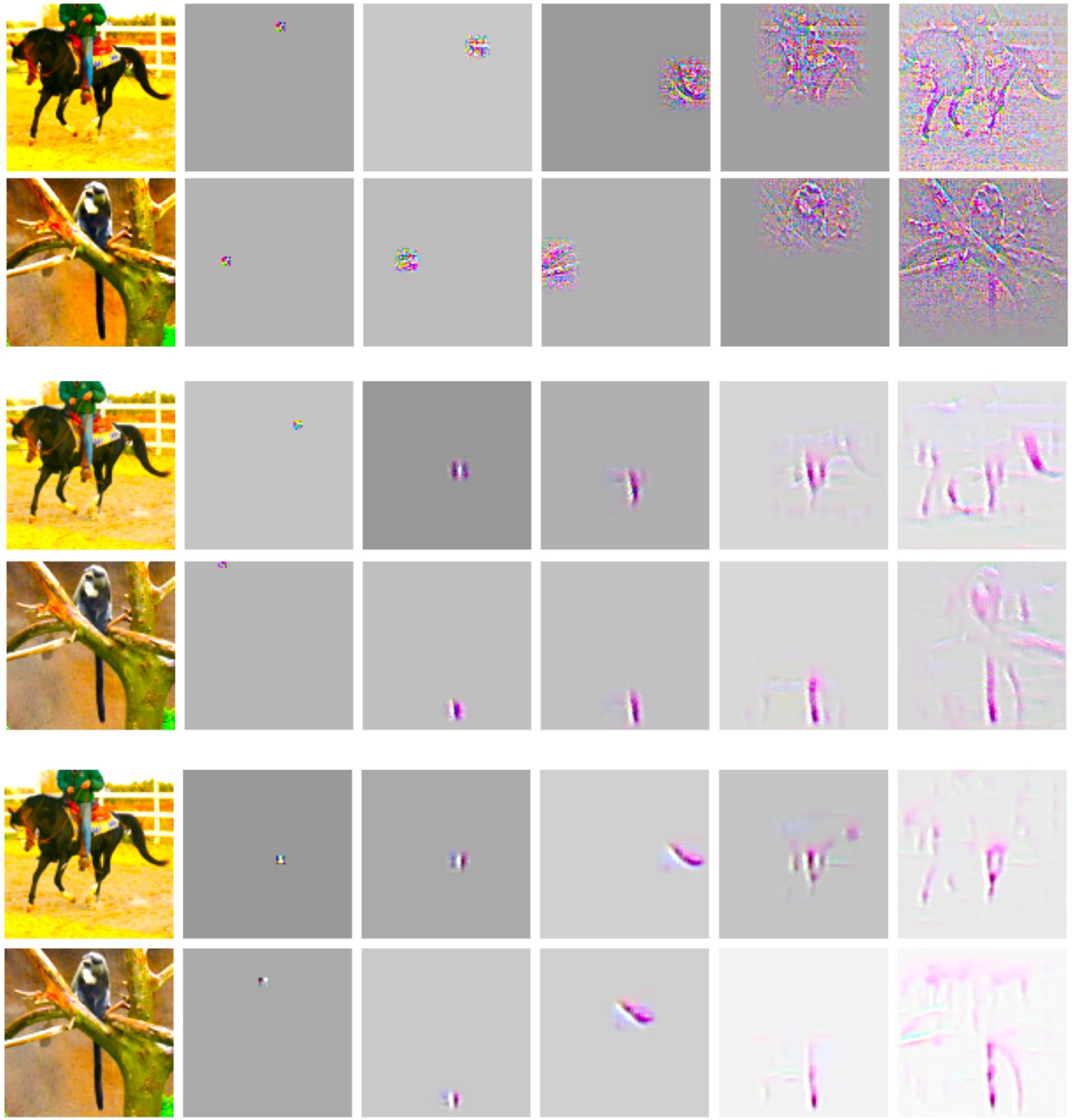




CONCAVE

LINEAR

CONVEX



ONGOING ATTEMPTS AT A FIRST-PRINCIPLES THEORY

DMFT

$$\boldsymbol{h}_m^1 = \frac{1}{\sqrt{D}}\boldsymbol{W}^0\boldsymbol{x}_m, \qquad \boldsymbol{h}_m^{\ell+1} = \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell \phi\left(\boldsymbol{h}_m^\ell\right)$$

$$\boldsymbol{h}_m^{L+1} = \frac{1}{\sqrt{N}}\boldsymbol{w}_L \cdot \phi\left(\boldsymbol{h}_m^L\right), \qquad f_m = \frac{1}{\gamma_0\sqrt{N}}\boldsymbol{h}_m^{L+1}$$

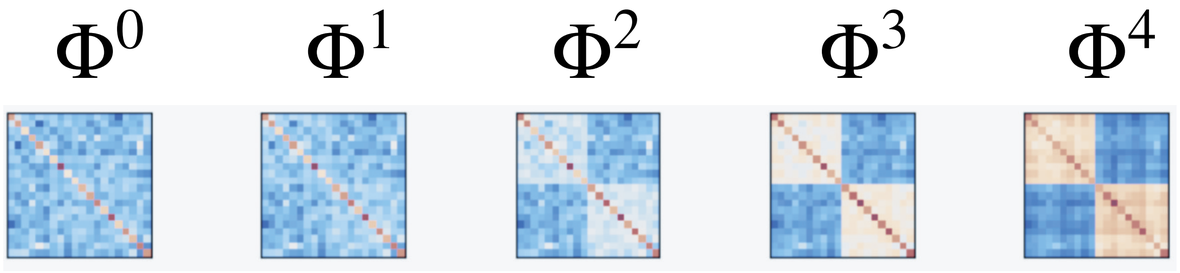
FEATURE KERNEL

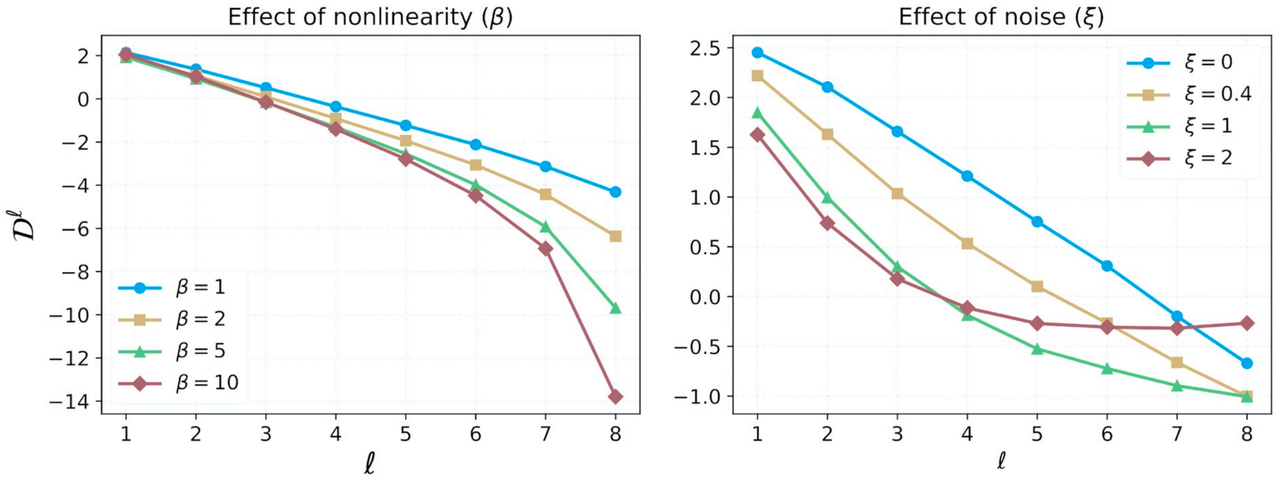
$$\Phi_{mp}^\ell(t,s) = \frac{1}{N}\left\langle \phi\left(\boldsymbol{h}_m^\ell(t)\right), \phi\left(\boldsymbol{h}_p^\ell(s)\right)\right\rangle$$

GRADIENT KERNEL

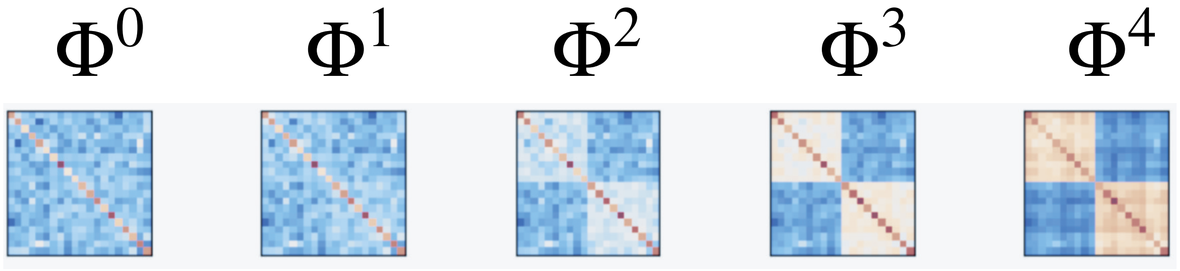
$$G_{mp}^\ell(t,s) = \frac{1}{N}\left\langle \boldsymbol{g}_m^\ell(t), \boldsymbol{g}_p^\ell(s)\right\rangle$$

Bordelon, Pehlevan; 2022

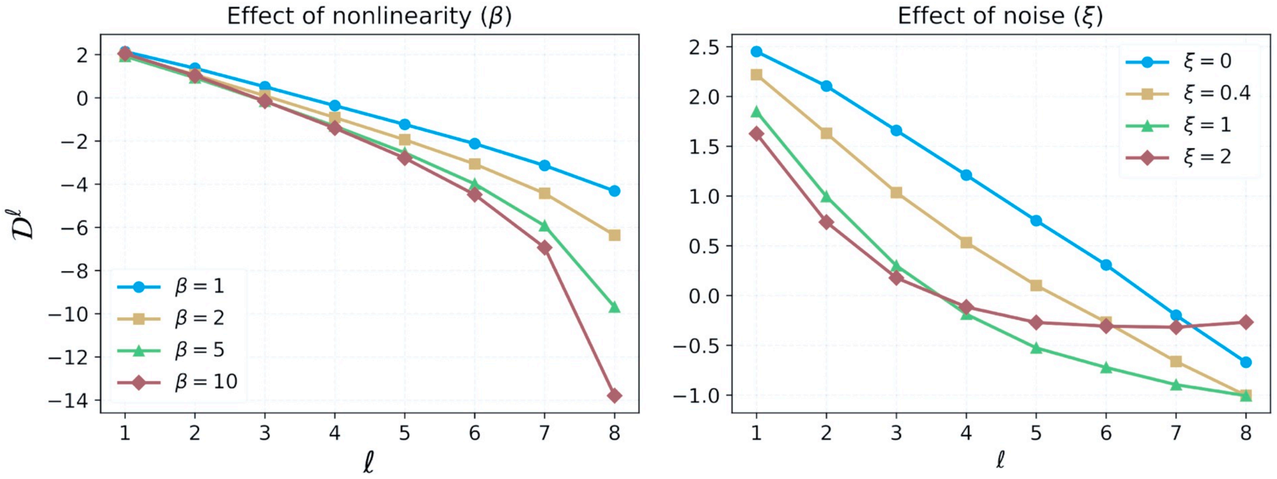

Φ^0
Φ^1
Φ^2
Φ^3
Φ^4


Effect of nonlinearity (β)
$\mathcal{D}^\ell$
ℓ
$\beta = 1$
$\beta = 2$
$\beta = 5$
$\beta = 10$
Effect of noise (ξ)
ℓ
$\xi = 0$
$\xi = 0.4$
$\xi = 1$
$\xi = 2$

WHITE DATA, BINARY CLASSIFICATION, LINEAR NET, VERY SMALL INIT



$$\mathscr{D}_\ell = \log\left(\frac{M-2}{H_y^\ell}\right) \qquad H_y^\ell := \boldsymbol{y}^\top \boldsymbol{\Phi}^\ell \boldsymbol{y}/M$$



$$H_y^\ell = (a(t))^\ell$$

$$\frac{\mathrm{d}a}{\mathrm{d}t} = \gamma_0 \Delta a^L,$$

$$\frac{\mathrm{d}\Delta}{\mathrm{d}t} = -(L+1)\Delta a^{2L}$$

REPARAMETERIZATION

JACOT ET AL. 2022; "FEATURE LEARNING IN L^2 REGULARIZED DNNS"

$$\hat{Z}_0 := X \qquad Z_\ell = W_\ell \hat{Z}_{\ell-1} \qquad \hat{Z}_\ell = \sigma(Z_\ell)$$

REPARAMETERIZATION

JACOT ET AL. 2022; "FEATURE LEARNING IN L^2 REGULARIZED DNNS"

$$\hat{Z}_0 := X \qquad Z_\ell = W_\ell \hat{Z}_{\ell-1} \qquad \hat{Z}_\ell = \sigma(Z_\ell)$$

$$W_\ell = Z_\ell \hat{Z}^+_{\ell-1} + W^\perp_\ell \qquad W^\perp_\ell \hat{Z}_{\ell-1} = 0$$

REPARAMETERIZATION

JACOT ET AL. 2022; "FEATURE LEARNING IN L^2 REGULARIZED DNNS"

$$\hat{Z}_0 := X \qquad Z_\ell = W_\ell \hat{Z}_{\ell-1} \qquad \hat{Z}_\ell = \sigma(Z_\ell)$$

$$W_\ell = Z_\ell \hat{Z}_{\ell-1}^{+} + W_\ell^{\perp} \qquad W_\ell^{\perp} \hat{Z}_{\ell-1} = 0$$

$$\|W_\ell\|_F^2 = \|Z_\ell \hat{Z}_{\ell-1}^{+} + W_\ell^{\perp}\|_F^2 + \|W_\ell^{\perp}\|^2$$

REPARAMETERIZATION

JACOT ET AL. 2022; "FEATURE LEARNING IN L^2 REGULARIZED DNNS"

$$\hat{Z}_0 := X \qquad Z_\ell = W_\ell \hat{Z}_{\ell-1} \qquad \hat{Z}_\ell = \sigma(Z_\ell)$$

$$W_\ell = Z_\ell \hat{Z}_{\ell-1}^{+} + W_\ell^{\perp} \qquad W_\ell^{\perp} \hat{Z}_{\ell-1} = 0$$

$$\|W_\ell\|_F^2 = \|Z_\ell \hat{Z}_{\ell-1}^{+} + W_\ell^{\perp}\|_F^2 + \|W_\ell^{\perp}\|^2 \quad \Rightarrow \quad W_\ell^{\perp} = 0$$

$$C(Z_L) = \|Z_L - Y\|_F^2$$

ORIGINAL LOSS

$$\mathscr{L}(W_1, \ldots, W_L) = C(Z_L) + \frac{\lambda}{2} \sum_{\ell=1}^{L} \|W_\ell\|_F^2$$

$$C(Z_L) = \|Z_L - Y\|_F^2$$

ORIGINAL LOSS

$$\mathscr{L}(W_1, \ldots, W_L) = C(Z_L) + \frac{\lambda}{2} \sum_{\ell=1}^{L} \|W_\ell\|_F^2$$

REPARAM 1

$$\mathscr{L}(Z_1, \ldots, Z_L) = C(Z_L) + \frac{\lambda}{2} \sum_{\ell=1}^{L} \|Z_\ell \hat{Z}_{\ell-1}^{+}\|_F^2$$

$$C(Z_L) = \|Z_L - Y\|_F^2$$

$$K_\ell = Z_\ell^T Z_\ell$$

$$\widehat{K}_\ell = \hat{Z}_\ell^T \hat{Z}_\ell$$

ORIGINAL LOSS

$$\mathscr{L}(W_1, \ldots, W_L) = C(Z_L) + \frac{\lambda}{2} \sum_{\ell=1}^{L} \|W_\ell\|_F^2$$

REPARAM 1

$$\mathscr{L}(Z_1, \ldots, Z_L) = C(Z_L) + \frac{\lambda}{2} \sum_{\ell=1}^{L} \|Z_\ell \hat{Z}_{\ell-1}^+\|_F^2$$

$$C(Z_L) = \|Z_L - Y\|_F^2$$

$$K_\ell = Z_\ell^T Z_\ell$$

$$\widehat{K}_\ell = \hat{Z}_\ell^T \hat{Z}_\ell$$

ORIGINAL LOSS

$$\mathscr{L}(W_1, \ldots, W_L) = C(Z_L) + \frac{\lambda}{2} \sum_{\ell=1}^{L} \|W_\ell\|_F^2$$

REPARAM 1

$$\mathscr{L}(Z_1, \ldots, Z_L) = C(Z_L) + \frac{\lambda}{2} \sum_{\ell=1}^{L} \|Z_\ell \hat{Z}_{\ell-1}^+\|_F^2$$

REPARAM 2

$$\mathscr{L}(\mathscr{K}, Z_L) = C(Z_L) + \frac{\lambda}{2} \sum_{\ell=1}^{L} \|K_\ell \, \widehat{K}_{\ell-1}^+\|_F^2$$

MACROSCOPIC DYNAMICS

WORK IN PROGRESS! POSSIBLE MISTAKES!

$$\varphi_\ell = \mathrm{Tr}\left[K_\ell\right]$$

MACROSCOPIC DYNAMICS

WORK IN PROGRESS! POSSIBLE MISTAKES!

$$\varphi_\ell = \mathrm{Tr}\left[K_\ell\right] \qquad \hat{\varphi}^+_\ell = \mathrm{Tr}\left[\widehat{K}^+_\ell\right]$$

MACROSCOPIC DYNAMICS

WORK IN PROGRESS! POSSIBLE MISTAKES!

$$\varphi_\ell = \mathrm{Tr}\left[K_\ell\right] \qquad \hat{\varphi}^+_\ell = \mathrm{Tr}\left[\widehat{K}^+_\ell\right] \qquad \hat{\psi}_\ell = \mathrm{Tr}\left[K_\ell P_{\ell-1}\right]$$

MACROSCOPIC DYNAMICS

WORK IN PROGRESS! POSSIBLE MISTAKES!

$$\hat{Z}^{+}_{\ell-1}\hat{Z}_{\ell-1}$$

$$\varphi_\ell = \mathrm{Tr}\left[K_\ell\right] \qquad \hat{\varphi}^{+}_\ell = \mathrm{Tr}\left[\widehat{K}^{+}_\ell\right] \qquad \hat{\psi}_\ell = \mathrm{Tr}\left[K_\ell P_{\ell-1}\right]$$

MACROSCOPIC DYNAMICS

WORK IN PROGRESS! POSSIBLE MISTAKES!

$$\hat{Z}^{+}_{\ell-1}\hat{Z}_{\ell-1}$$

$$\varphi_\ell = \mathrm{Tr}\left[K_\ell\right] \qquad \hat{\varphi}^{+}_{\ell} = \mathrm{Tr}\left[\widehat{K}^{+}_{\ell}\right] \qquad \hat{\psi}_\ell = \mathrm{Tr}\left[K_\ell P_{\ell-1}\right]$$

$$\omega = \frac{1}{N}\langle Z_L, Y\rangle$$

MACROSCOPIC DYNAMICS

WORK IN PROGRESS! POSSIBLE MISTAKES!

$\hat{Z}^{+}_{\ell-1}\hat{Z}_{\ell-1}$

$$\varphi_\ell = \mathrm{Tr}\left[K_\ell\right] \qquad \hat{\varphi}^{+}_\ell = \mathrm{Tr}\left[\widehat{K}^{+}_\ell\right] \qquad \hat{\psi}_\ell = \mathrm{Tr}\left[K_\ell P_{\ell-1}\right]$$

$$\omega = \frac{1}{N}\langle Z_L, Y\rangle$$

$$\lambda = \frac{1}{2}$$

$$\frac{d\varphi_\ell}{dt} = -\psi_\ell + \psi_{\ell+1}$$

$$\frac{d\psi_\ell}{dt} = -a_\ell\psi_\ell\hat{\varphi}^{+}_{\ell-1} + b_\ell\psi_{\ell+1}\hat{\varphi}^{+}_{\ell-1} + c_\ell\psi_{\ell+1}\hat{\varphi}^{+}_\ell - d_\ell\frac{\psi^2_\ell\hat{\varphi}^{+}_\ell}{\varphi_{\ell-1}}$$

$$\frac{d\hat{\varphi}^{+}_\ell}{dt} = \frac{1}{2}e_\ell\hat{\varphi}^{+}_\ell\hat{\varphi}^{+}_{\ell-1} - \frac{1}{2}f_\ell\frac{\hat{\varphi}^{+}_\ell\psi_{\ell+1}}{\varphi_\ell}$$

LINEAR NETWORK

$$\varphi_\ell = \psi_\ell = \hat{\varphi}_\ell^+$$

LINEAR NETWORK

$$\varphi_\ell = \psi_\ell = \hat{\varphi}^+_\ell$$

$$\dot{\varphi}_\ell = \frac{\varphi_{\ell+1}}{\varphi_\ell} - \frac{\varphi_\ell}{\varphi_{\ell-1}}$$

$$\dot{\varphi}_L = 2\omega - 2\varphi_L - \frac{\varphi_L}{\varphi_{L-1}}$$

$$\dot{\omega} = 1 - \omega - \frac{\omega}{\varphi_{L-1}}$$

LINEAR NETWORK

$$\varphi_\ell = \psi_\ell = \hat{\varphi}^+_\ell$$

$$\dot{\varphi}_\ell = \frac{\varphi_{\ell+1}}{\varphi_\ell} - \frac{\varphi_\ell}{\varphi_{\ell-1}}$$

$$\dot{\varphi}_L = 2\omega - 2\varphi_L - \frac{\varphi_L}{\varphi_{L-1}}$$

$$\dot{\omega} = 1 - \omega - \frac{\omega}{\varphi_{L-1}}$$

$\Longrightarrow$ A STATIONARY POINT

$$\log \varphi_{\ell+1} - \log \varphi_\ell = \log \varphi_\ell - \log \varphi_{\ell-1}$$

LINEAR NETWORK

$$\varphi_\ell = \psi_\ell = \hat{\varphi}^+_\ell$$

$$\dot{\varphi}_\ell = \frac{\varphi_{\ell+1}}{\varphi_\ell} - \frac{\varphi_\ell}{\varphi_{\ell-1}}$$

$$\dot{\varphi}_L = 2\omega - 2\varphi_L - \frac{\varphi_L}{\varphi_{L-1}}$$

$$\dot{\omega} = 1 - \omega - \frac{\omega}{\varphi_{L-1}}$$

$\Longrightarrow$ A STATIONARY POINT

$$\log \varphi_{\ell+1} - \log \varphi_\ell = \log \varphi_\ell - \log \varphi_{\ell-1}$$

NONLINEAR NETWORK (NO NOISE)

$$\log \varphi_{\ell+1} - \log \hat{\varphi}^+_\ell = \log \varphi_\ell - \log \hat{\varphi}^+_{\ell-1}$$

ENDNOTES

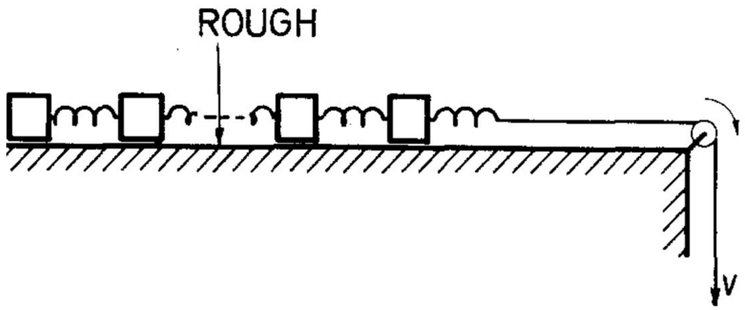
SPRINGS & BLOCKS

Bulletin of the Seismological Society of America. Vol. 57, No. 3, pp. 341–371. June, 1967

MODEL AND THEORETICAL SEISMICITY

By R. Burridge and L. Knopoff

ABSTRACT



ɜ. 3. Schematic diagram of the laboratory model.

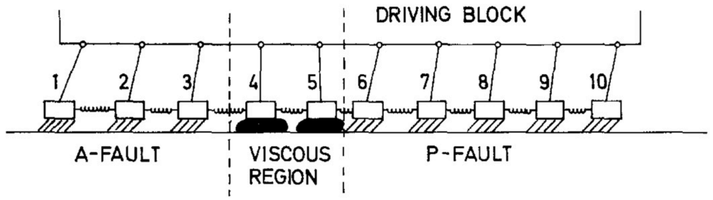
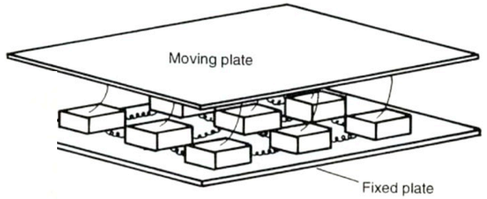


Fig. 15. Schematic diagram of the numerical model.

NEURONS, DYNAMICS AND COMPUTATION

Brains have long been regarded as biological computers. But how do these collections of neurons perform computations?

John J. Hopfield



Spring–block model of earthquakes is somewhat analogous to models of sheets of neurons that account for individual neuron spikes. Springs connecting the blocks correspond to synaptic connections between neurons, while the slipping of a block corresponds to the firing of a neuron. Consequently the types of complex behavior seen in the earthquake models can be expected to recur in neurobiology. (Adapted from ref. 17.) **Figure 7**

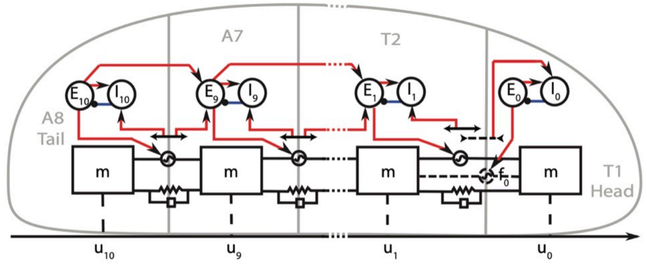
Integrative neuromechanics of crawling in *D. melanogaster* larvae

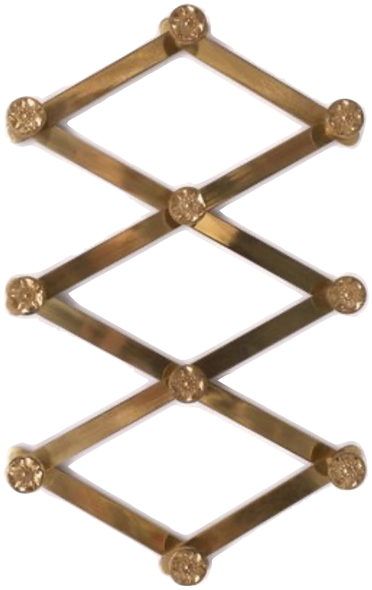
Cengiz Pehlevan[1,2], Paolo Paoletti[3], L Mahadevan[4,5,6,7,8]*

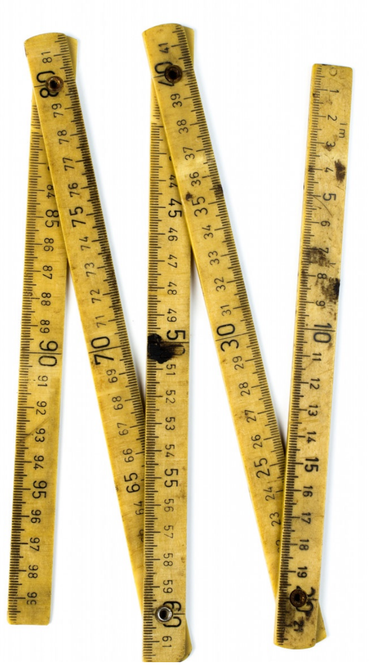
VNC

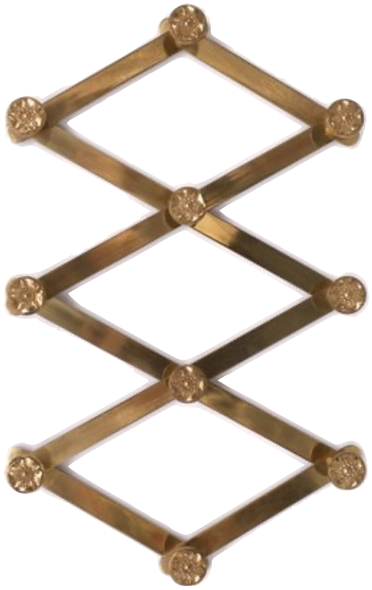
body & muscle mechanics

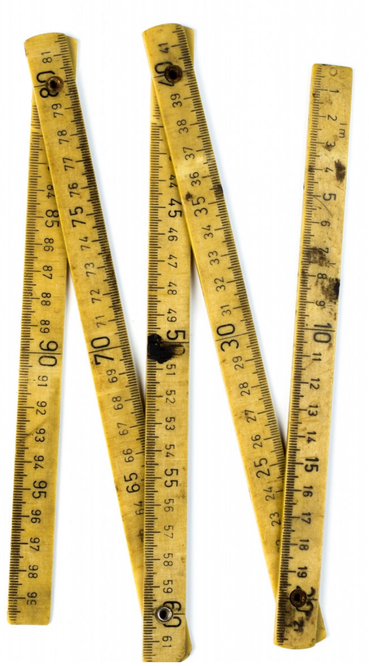
environment (solid friction)











CONNECTIONS WITH THE WORK OF GERARD & AUKOSH

autonomous dynamics? (somehow)

we insist on understanding what happens over layers

CONNECTIONS WITH THE WORK OF GERARD & AUKOSH

autonomous dynamics? (somehow)

we insist on understanding what happens over layers

NOISE IN DYNAMICAL SYSTEMS

linearization / stabilization understood in signal processing and control theory

high dimensional setting is exciting, with focus on cascades (signal flow)

CONNECTIONS WITH THE WORK OF GERARD & AUKOSH

autonomous dynamics? (somehow)

we insist on understanding what happens over layers

NOISE IN DYNAMICAL SYSTEMS

linearization / stabilization understood in signal processing and control theory

high dimensional setting is exciting, with focus on cascades (signal flow)

BILDTHEORIE

you can give a sense of training dynamics to the uninitiated!

HERTZ, BOLTZMANN, . . .

It is the ubiquitous task of science to explain the more complex in terms of the simpler; or, if preferred, to represent [anschaulich darstellen] the complex by means of clear pictures [bilder] borrowed from the sphere of simpler phenomena.

HERTZ, BOLTZMANN, . . .

It is the ubiquitous task of science to explain the more complex in terms of the simpler; or, if preferred, to represent [anschaulich darstellen] the complex by means of clear pictures [bilder] borrowed from the sphere of simpler phenomena.

PHYSICAL REVIEW LETTERS **134,** 257301 (2025)

Editors' Suggestion

Spring-Block Theory of Feature Learning in Deep Neural Networks

Cheng Shi,[1] Liming Pan,[2] and Ivan Dokmanić[1,3,*]

[1]*Departement Mathematik und Informatik, University of Basel, Spiegelgasse 1, 4051 Basel, Switzerland*
[2]*School of Cyber Science and Technology, University of Science and Technology of China, 230026 Hefei, China*
[3]*Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 306 North Wright Street, Urbana, Illinois 61801, USA*