Rough landscapes
from machine learning to glasses and back

Chiara Cammarota

King's College London

Marco Baity-Jesi, Levent Sagun, Mario Geiger, Stefano Spigler, Gerard Ben Arous, Chiara Cammarota, Yann LeCun, Matthieu Wyart, Giulio Biroli

PMLR 80:314-323, 2018

At the Crossroad of Physics and Machine Learning

11.02.2019 KITP

Supervised learning

$\{\mathbf{X}^\alpha\}$

Images

$\{Y^\alpha\}$

Labels

Input data (Train set) ~ 10^3 x 10^4

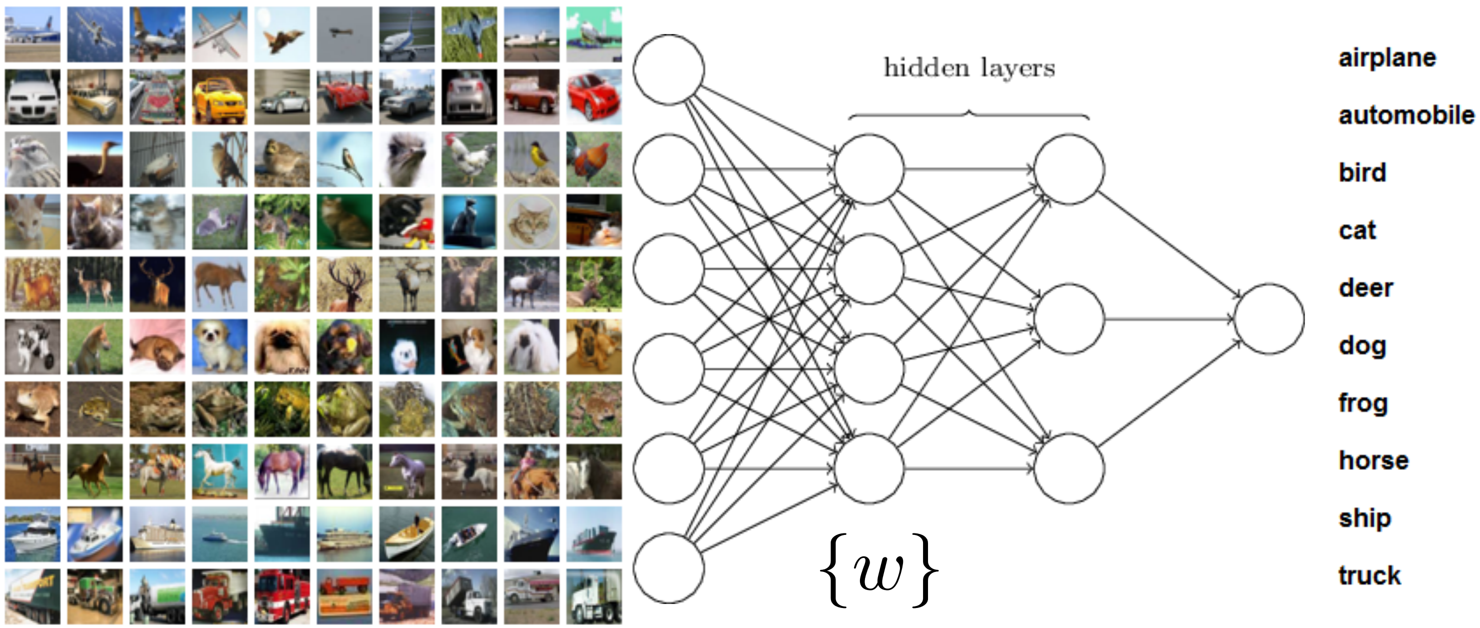
Parameters ~ 10^8

Goal: find the right weights as to classify unseen images (Test set)

Supervised learning

$\{\mathbf{X}^\alpha\}$

Images



$\{Y^\alpha\}$

Labels

Input data (Train set) ~ 10^3 x 10^4

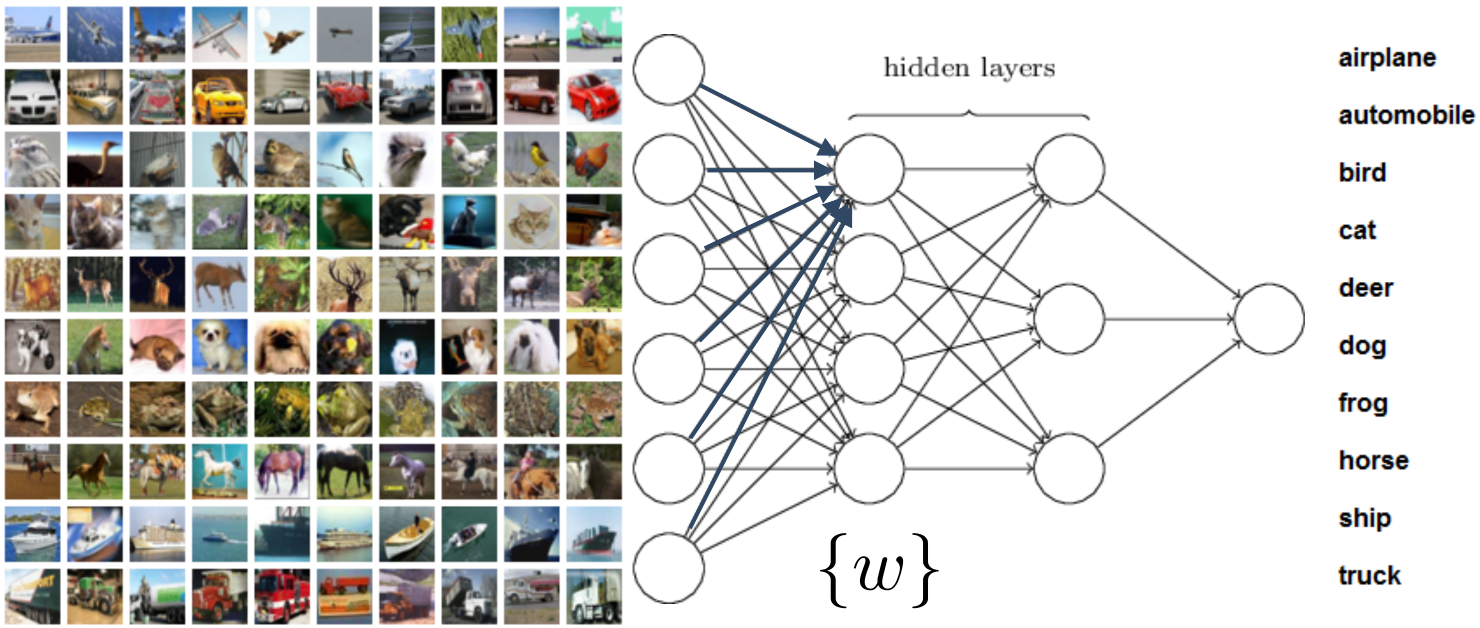
Parameters ~ 10^8

Goal: find the right weights as to classify unseen images (Test set)

Supervised learning

$\{\mathbf{X}^\alpha\}$

Images



$\{Y^\alpha\}$

Labels

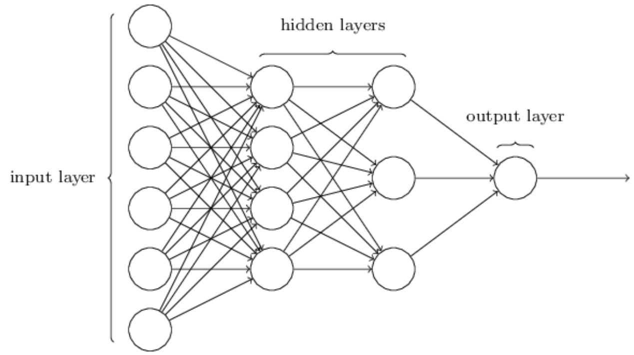
Input data (Train set) ~ 10^3 x 10^4

Parameters ~ 10^8

Goal: find the right weights as to classify unseen images (Test set)

Training Dynamics

Let's start by correctly classifying the Train set



distance between output and correct answer, i.e.

$$l(\{w\};\{\mathbf{X}^\alpha\},\{Y^\alpha\}) = (Y^\alpha - f(\{w\};\{\mathbf{X}^\alpha\}))^2$$

Loss function
$$\mathcal{L}\{w\} = \frac{1}{M}\sum_{\alpha}^{M} \ell(\{w\};\{\mathbf{X}^\alpha\},\{Y^\alpha\})$$

Learning (training): minimise the Loss function from random initial condition

Gradient Descent
$$\mathbf{w}(t+\Delta t) = \mathbf{w}(t) - \eta\nabla_w\mathcal{L}\{w\}$$

Stochastic Gradient Descent
$$\mathbf{w}(t+\Delta t) = \mathbf{w}(t) - \eta\sum_{\alpha}^{B}\nabla_w\ell(\{w\};\{\mathbf{X}^\alpha\},\{Y^\alpha\})$$

How good is Training?

Learning is strikingly good! …but why?

Many answers are hidden in the rough landscape of Loss function, but very complex!

What is the shape of the Loss landscape?

Why and how does training succeed?

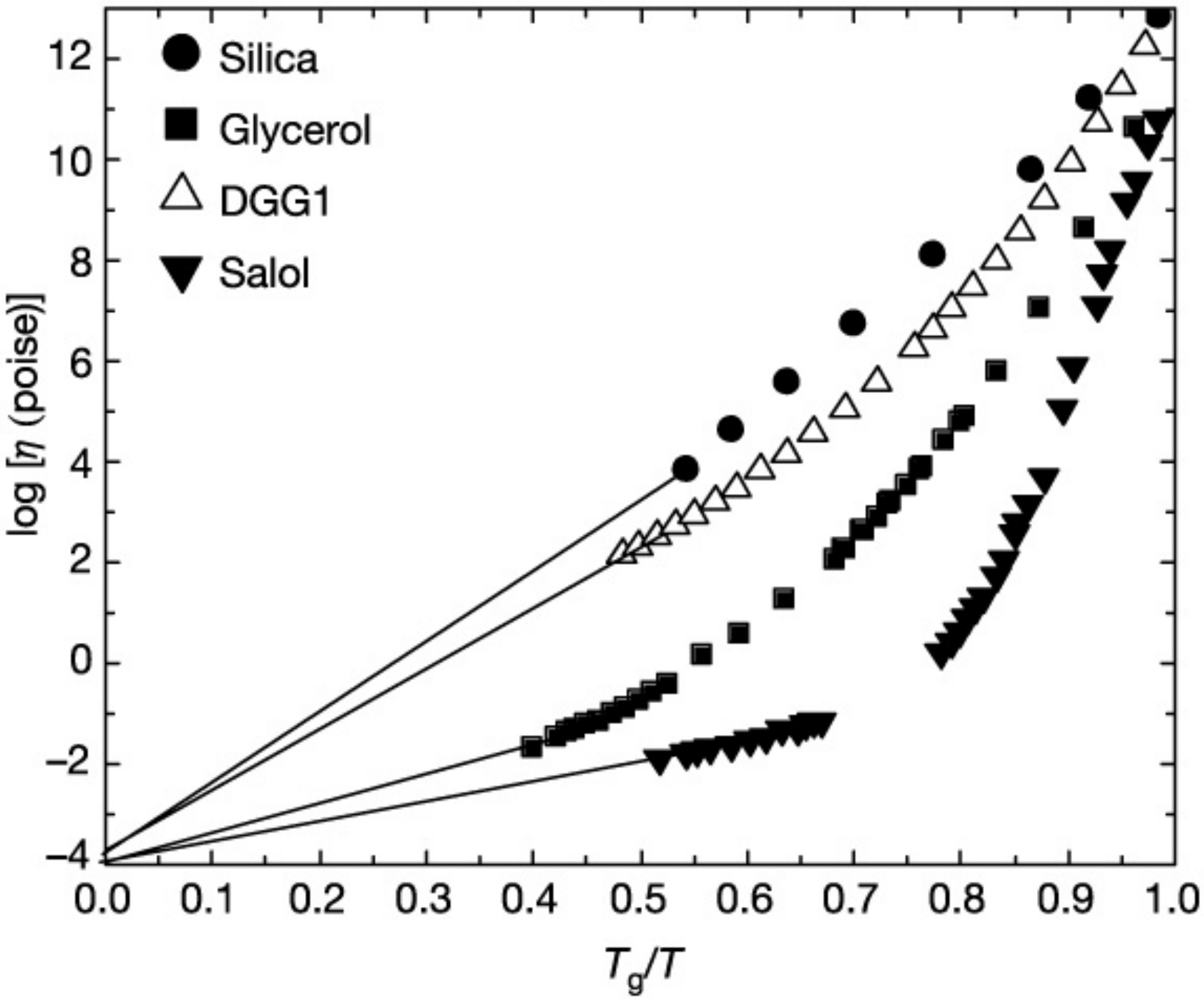
Choromanska et al. 2015, Baldassi et al. 2016, Soudry Charmon 2016, Freeman Bruna 2017, Soudry Hoffer 2018

Dauphin et al. 2014, Sagun et al. 2014, Lee et al. 2016, Jastrzebski et al. 2017

good minima,
fat minima,
rare minima,
are there minima?
role of saddles

The glass-formers paradigm

Same challenge met in physics of glass-formers



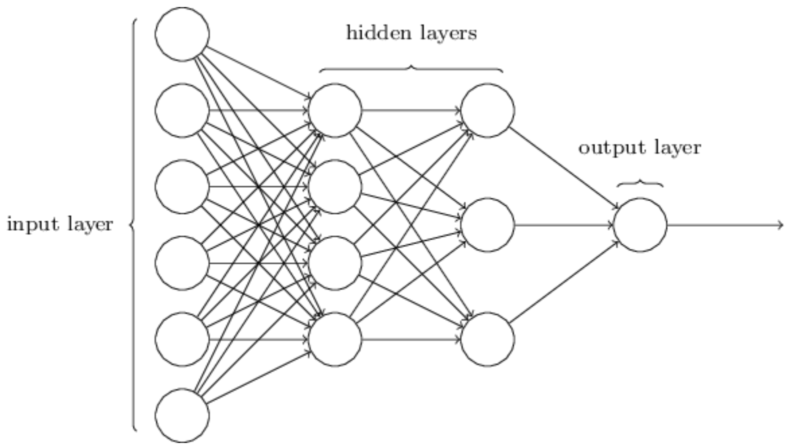
~ 50 years of research

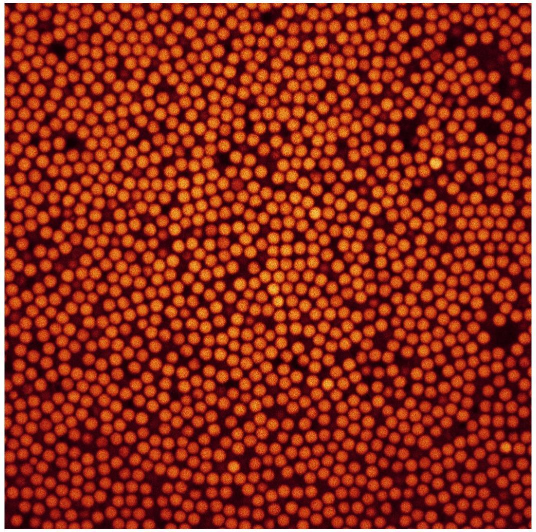
Goldstein 1968

$$\tau(T) \sim \exp(\Delta(T)/T)$$

$$\Delta(T) \uparrow \quad T \downarrow$$

From ML to models of glasses





$$\mathcal{L}\{w\} = \frac{1}{M}\sum_{\alpha}^{M} \ell(\{w\}; \{\mathbf{X}^{\alpha}\}, \{Y^{\alpha}\})$$

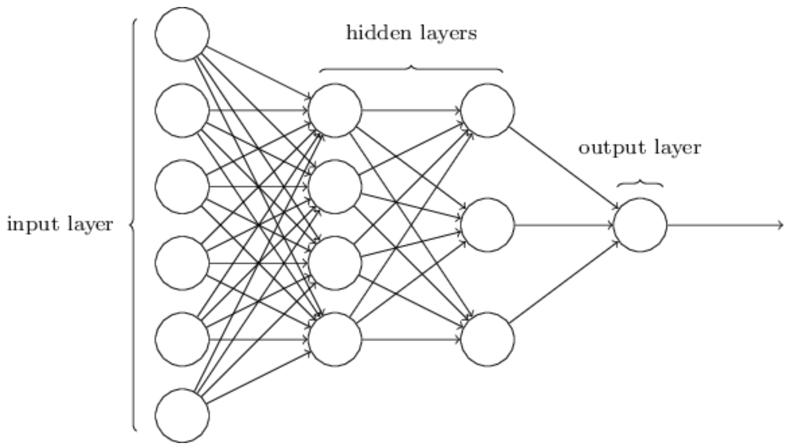
$$H = \sum_{(i,j)} V(|\mathbf{r}_i - \mathbf{r}_j|)$$

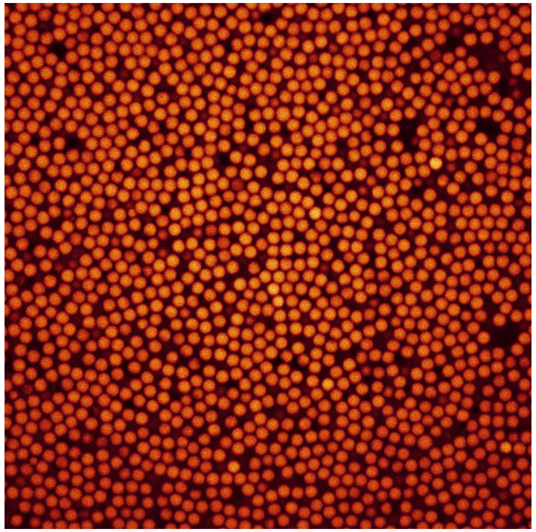
Parameters ~ 10^8

Real systems 10^23 particles

Numerical simulations 10^3 particles

From training to glass dynamics





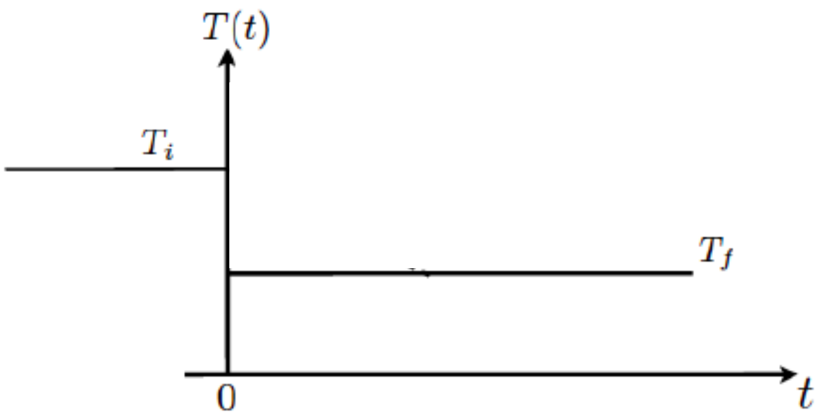
Learning (training) : minimise the Loss function from random initial condition

$$\dot{r}_{\alpha,i}(t) = -\nabla_{\alpha,i} H$$

$$\mathbf{w}(t+\Delta t) = \mathbf{w}(t) - \eta \nabla_w \mathcal{L}\{w\}$$

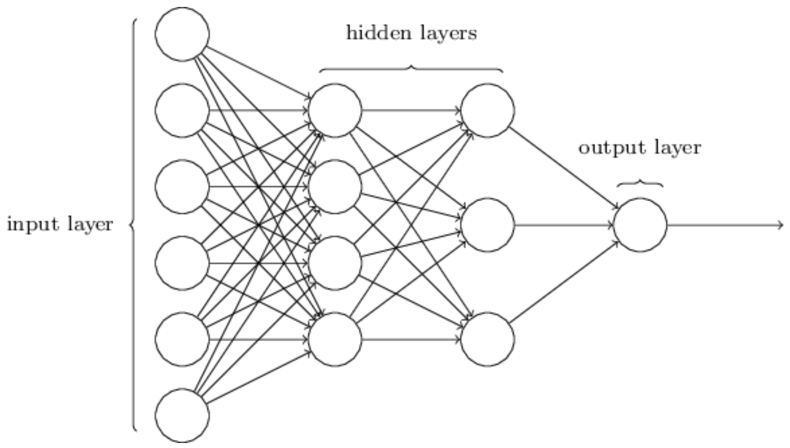
Every particle moves to minimise the Energy

$$\mathbf{w}(t+\Delta t) = \mathbf{w}(t) - \eta \sum_{\alpha}^{B} \nabla_w \ell(\{w\}; \{\mathbf{X}^\alpha\}, \{Y^\alpha\})$$



Quenches : rapid coolings from high temperature, i.e. almost random initial configuration

From training to glass dynamics

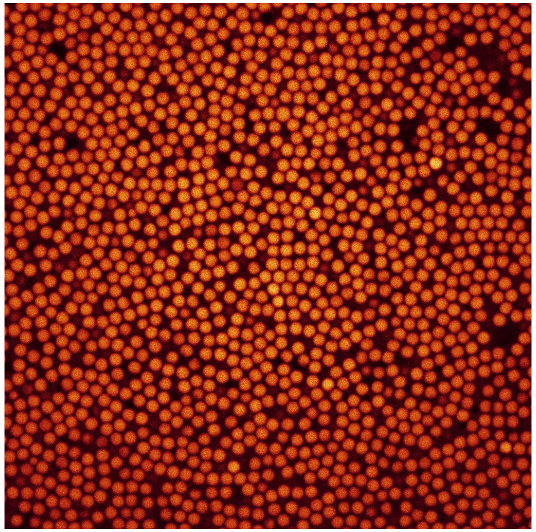


Learning (training) : minimise the Loss function from random initial condition

$$\mathbf{w}(t+\Delta t) = \mathbf{w}(t) - \eta \nabla_w \mathcal{L}\{w\}$$

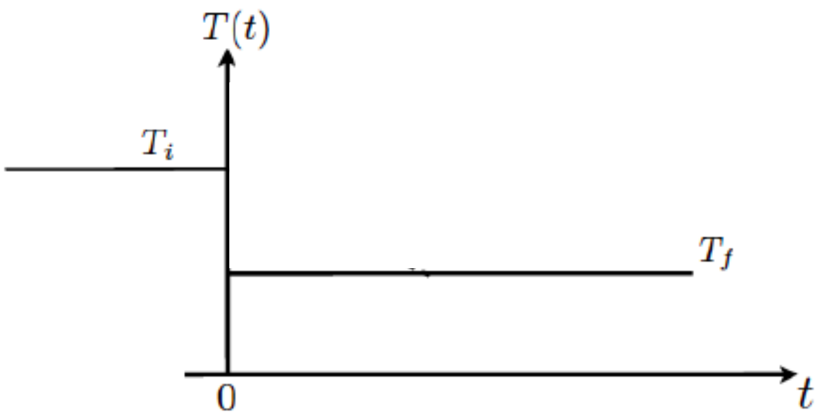
$$\mathbf{w}(t+\Delta t) = \mathbf{w}(t) - \eta \sum_{\alpha}^{B} \nabla_w \ell(\{w\}; \{\mathbf{X}^\alpha\}, \{Y^\alpha\})$$

Quenches : rapid coolings from high temperature, i.e. almost random initial configuration

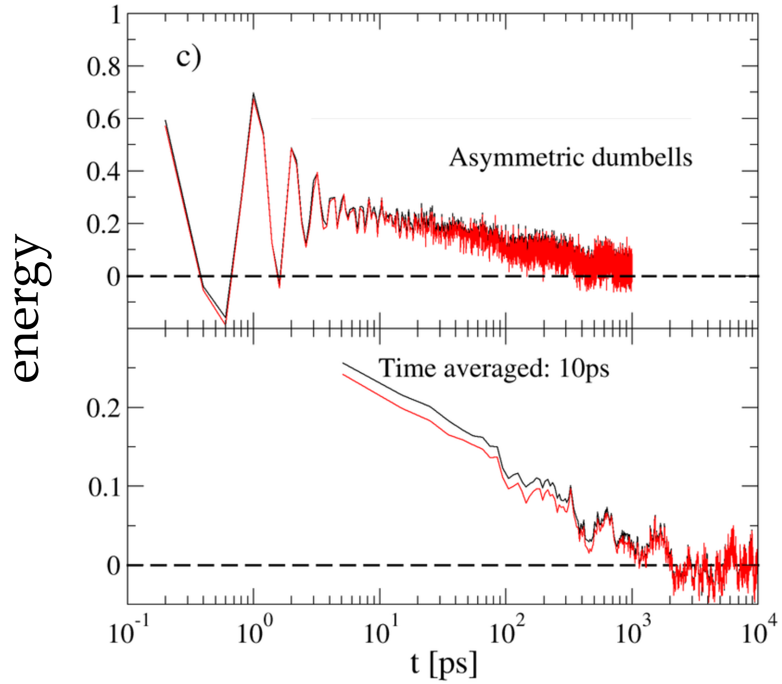


$$\dot{r}_{\alpha,i}(t) = -\nabla_{\alpha,i} H + \eta_{\alpha,i}(t)$$

Every particle moves to minimise the Energy + thermal noise



Aging dynamics



Sciortino 2005

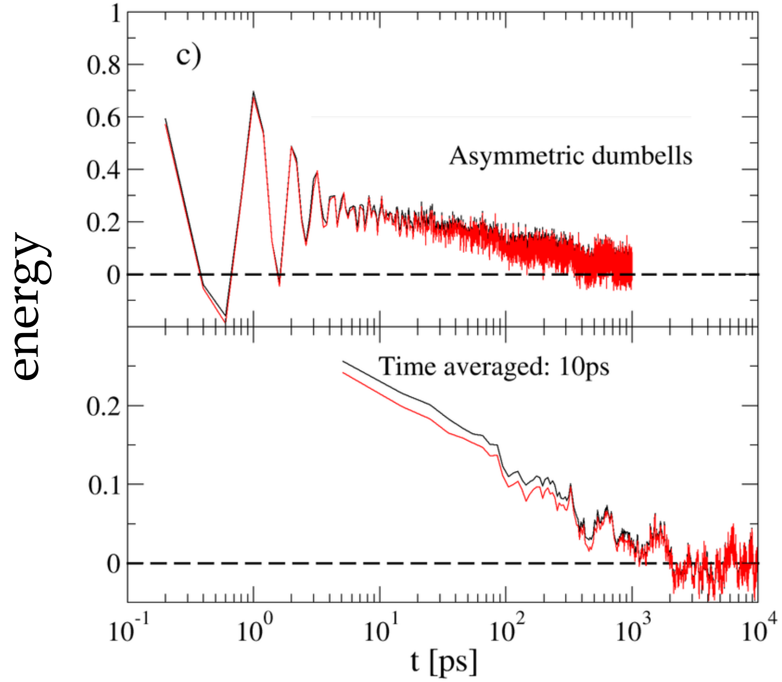
logarithmic decay (many timescales -> AGING!)

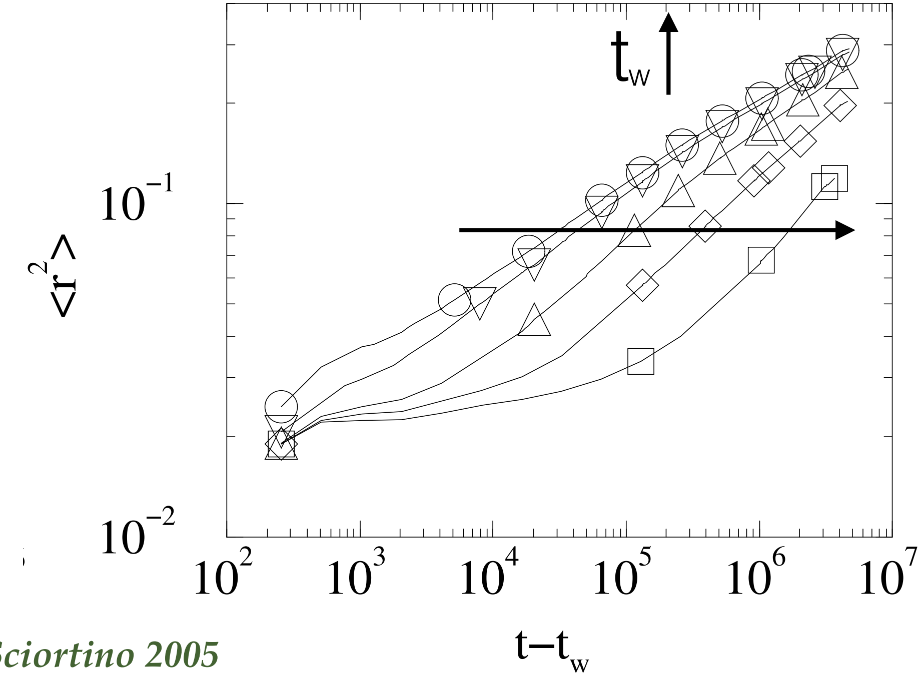
mean squared displacements

$$\langle r^2 \rangle = \frac{1}{N} \sum_i (\mathbf{r}_i(t_w) - \mathbf{r}_i(t))^2$$

the older the system, the longer it takes to move

Aging dynamics





Sciortino 2005

logarithmic decay (many timescales -> AGING!)

mean squared displacements

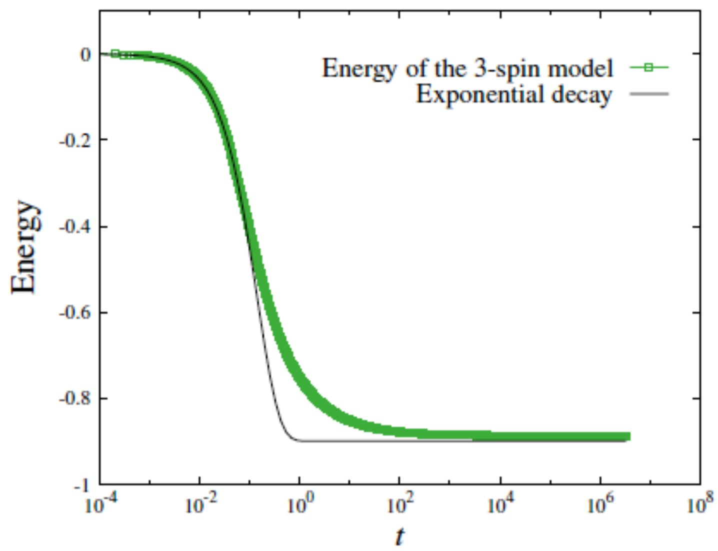
$$\langle r^2 \rangle = \frac{1}{N} \sum_i (\mathbf{r}_i(t_w) - \mathbf{r}_i(t))^2$$

the older the system, the longer it takes to move

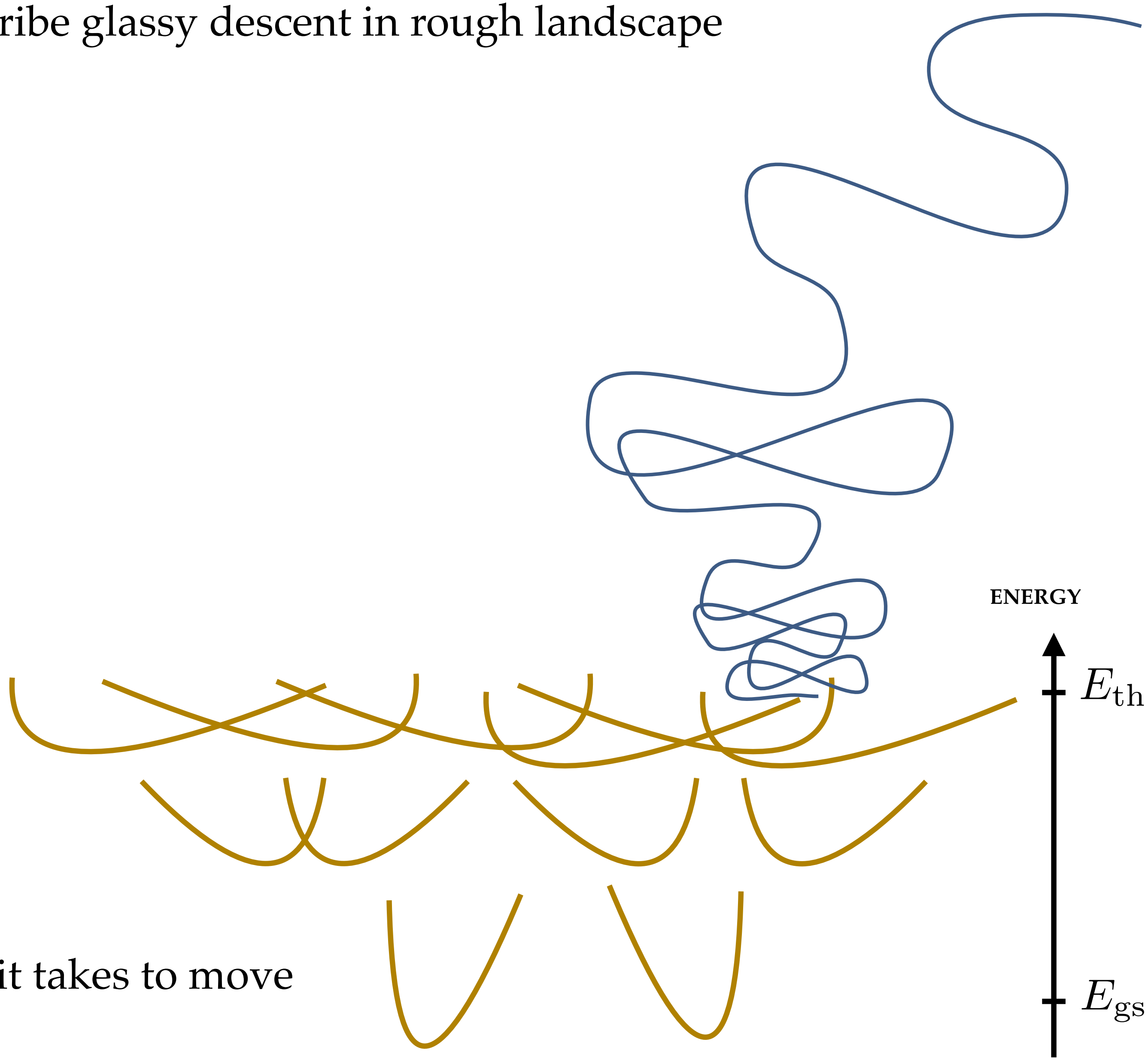
Aging in Mean Field models

Mean Field spin models to describe glassy descent in rough landscape

Slow decrease & trapping above the GS



Cugliandolo, Kurchan 1993



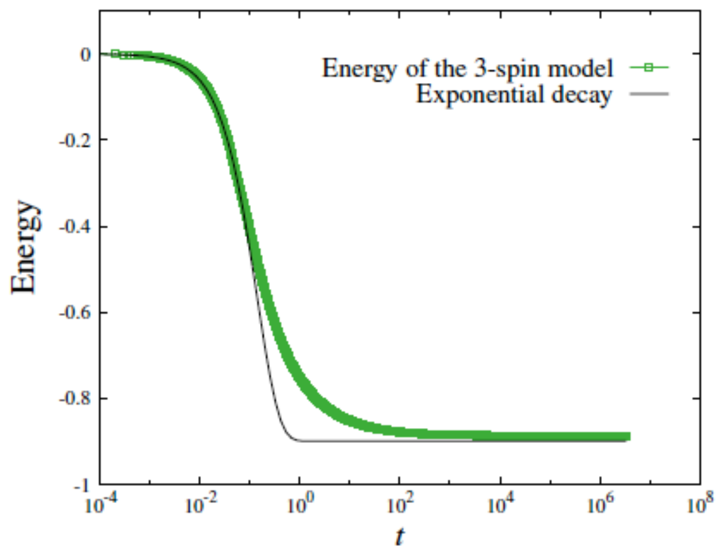
slower than exponential

the older the system, the longer it takes to move
no need of minima for that!

Aging in Mean Field models

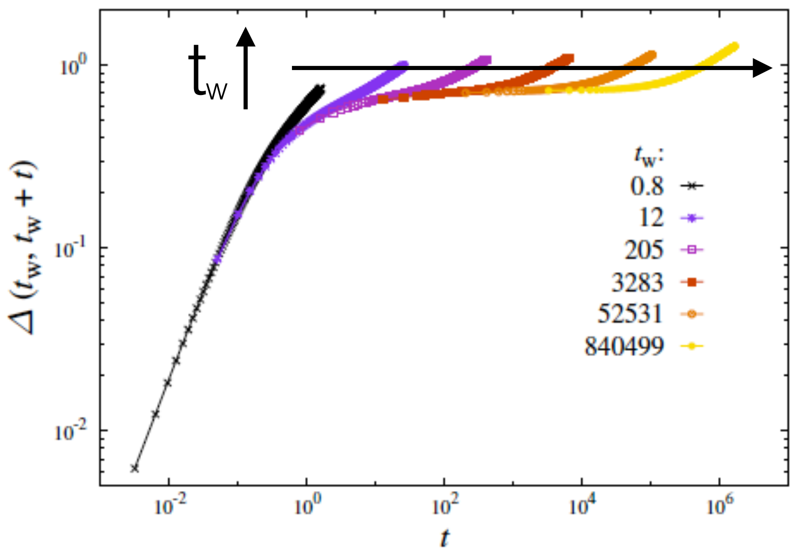
Mean Field spin models to describe glassy descent in rough landscape

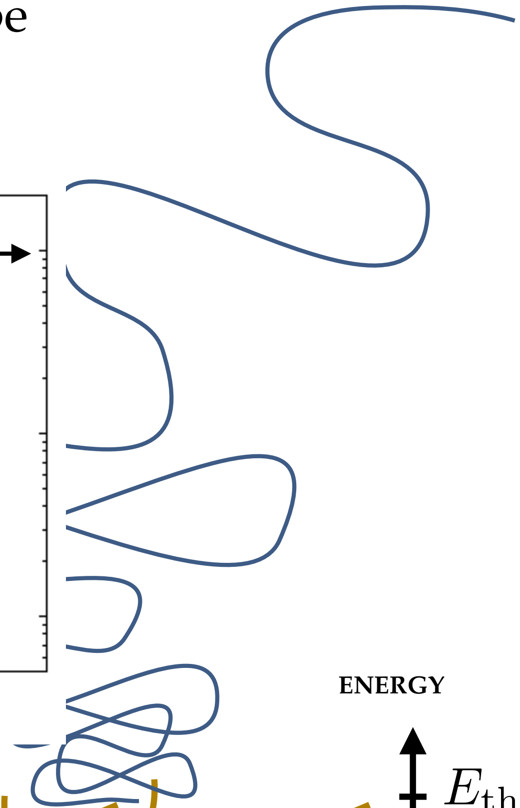
Slow decrease & trapping above the GS

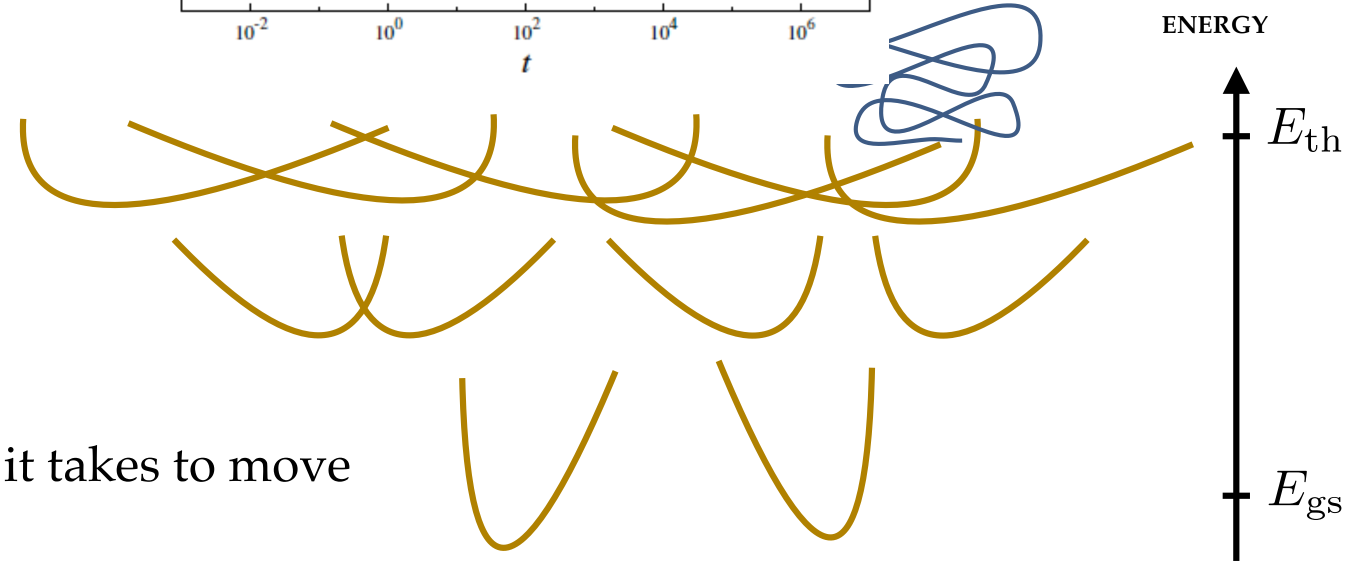


Cugliandolo, Kurchan 1993

Aging







slower than exponential

the older the system, the longer it takes to move

no need of minima for that!

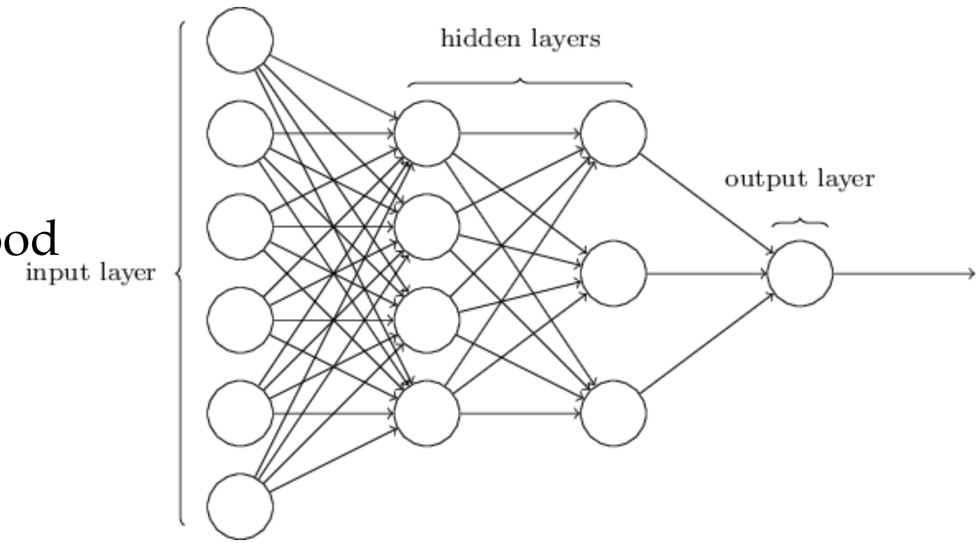
Comparing Dynamics

MBJ, LS, MG, SS, GBA, CC, YLC, MW, GB (2018)

Toy model: 1 hidden layer, ReLU, sigmoid in output, MSE as a loss

Fully connected: 3 hidden layers, ReLU, log likelihood

Small Net: 2 hidden convolutional layers,
2 fully connected ReLU, log likelihood



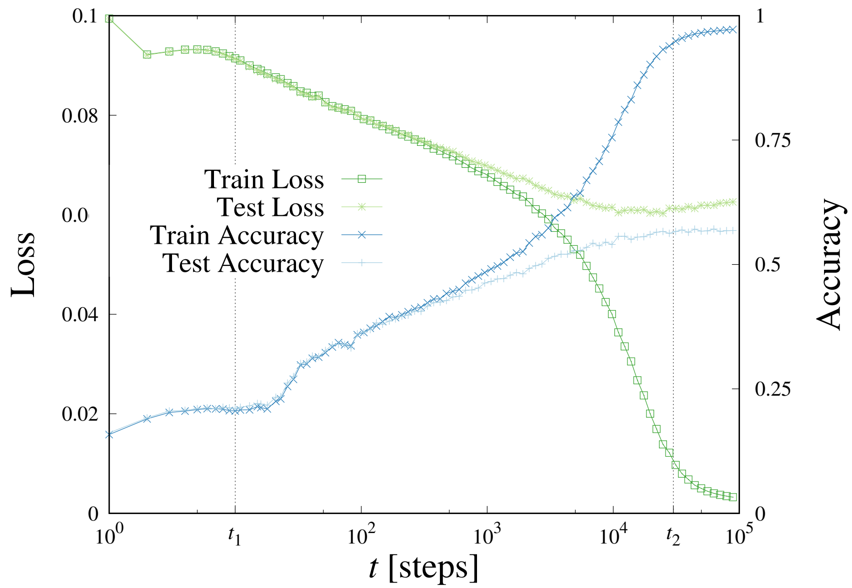
ResNet18: 18 hidden convolutional layers

MNIST, CFAR-10, CFAR-100

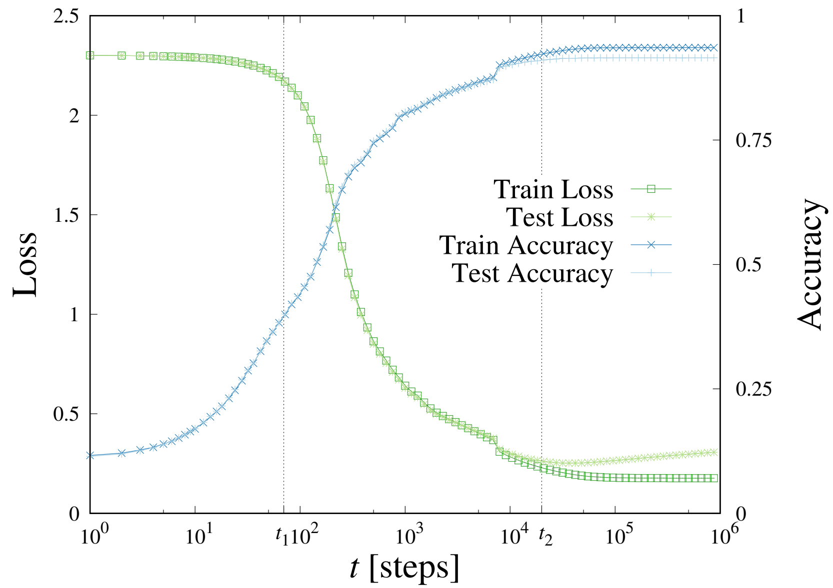
$$\mathcal{L}\{w\} = \frac{1}{M}\sum_{\alpha}^{M} \ell(\{w\}; \{\mathbf{X}^\alpha\}, \{Y^\alpha\}) \qquad \Delta(t_w, t_w + t) = \frac{1}{N}\sum_i (w_i(t_w) - w_i(t_w + t))^2$$

Aging during Learning

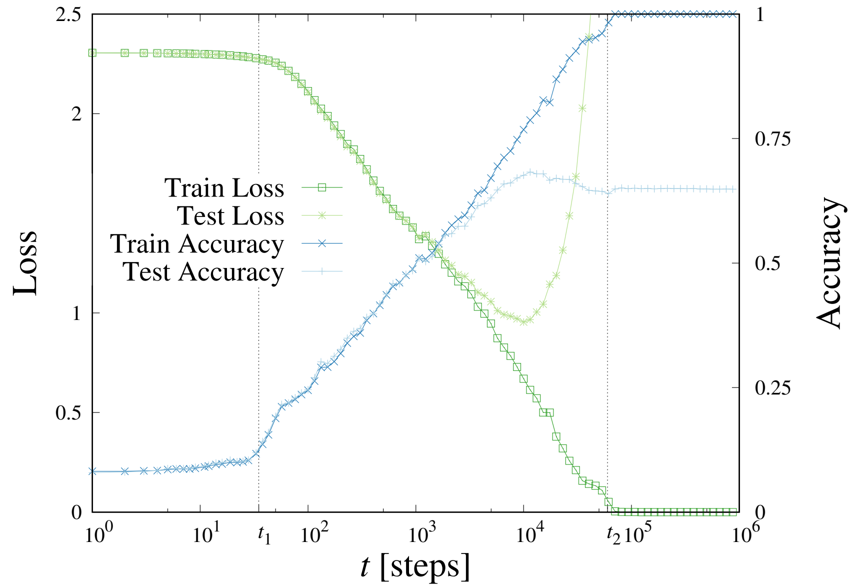
MBJ, LS, MG, SS, GBA, CC, YLC, MW, GB (2018)



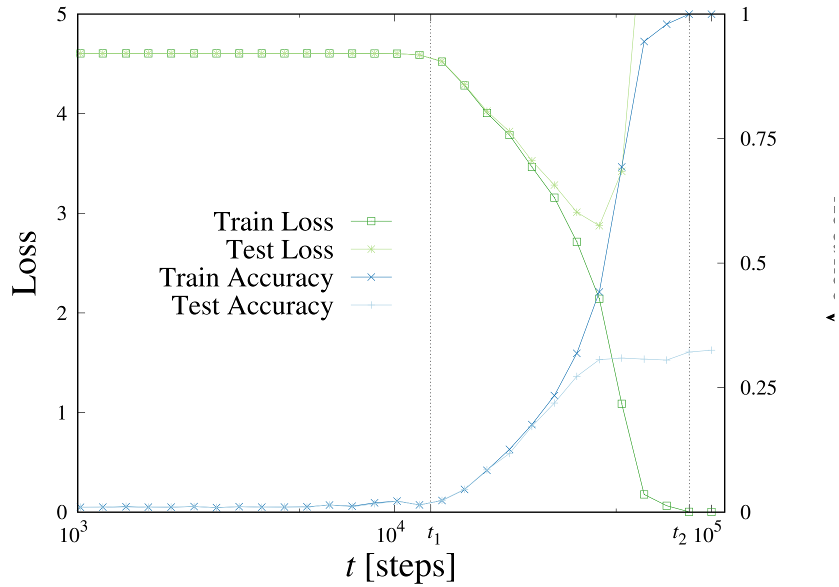
(a) Toy Model on CIFAR-10 $m = 10^4$, $B = 100$, $\alpha = 0.1$.



(b) Fully Connected on MNIST, $B = 128$, $\alpha = 0.01$.



(c) Small Net on CIFAR-10, $B = 100$, $\alpha = 0.01$.

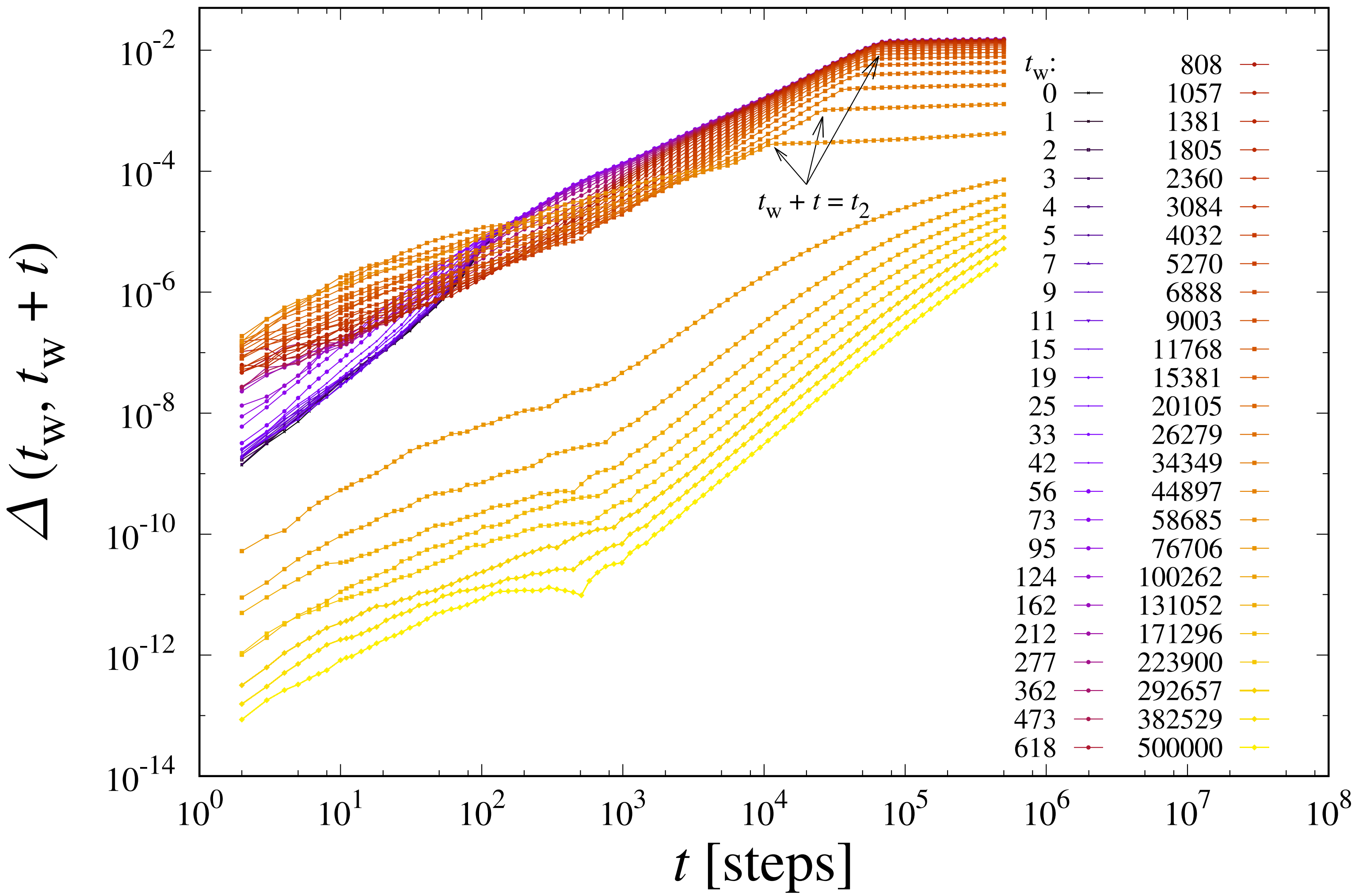


(d) ResNet-18 on CIFAR-100, $B = 64$, $\alpha = 0.01$.

Learning as Interrupted Aging and Diffusion

MBJ, LS, MG, SS, GBA, CC, YLC, MW, GB (2018)

$$\Delta(t_w, t_w + t) = \frac{1}{N}\sum_i (w_i(t_w) - w_i(t_w + t))^2$$

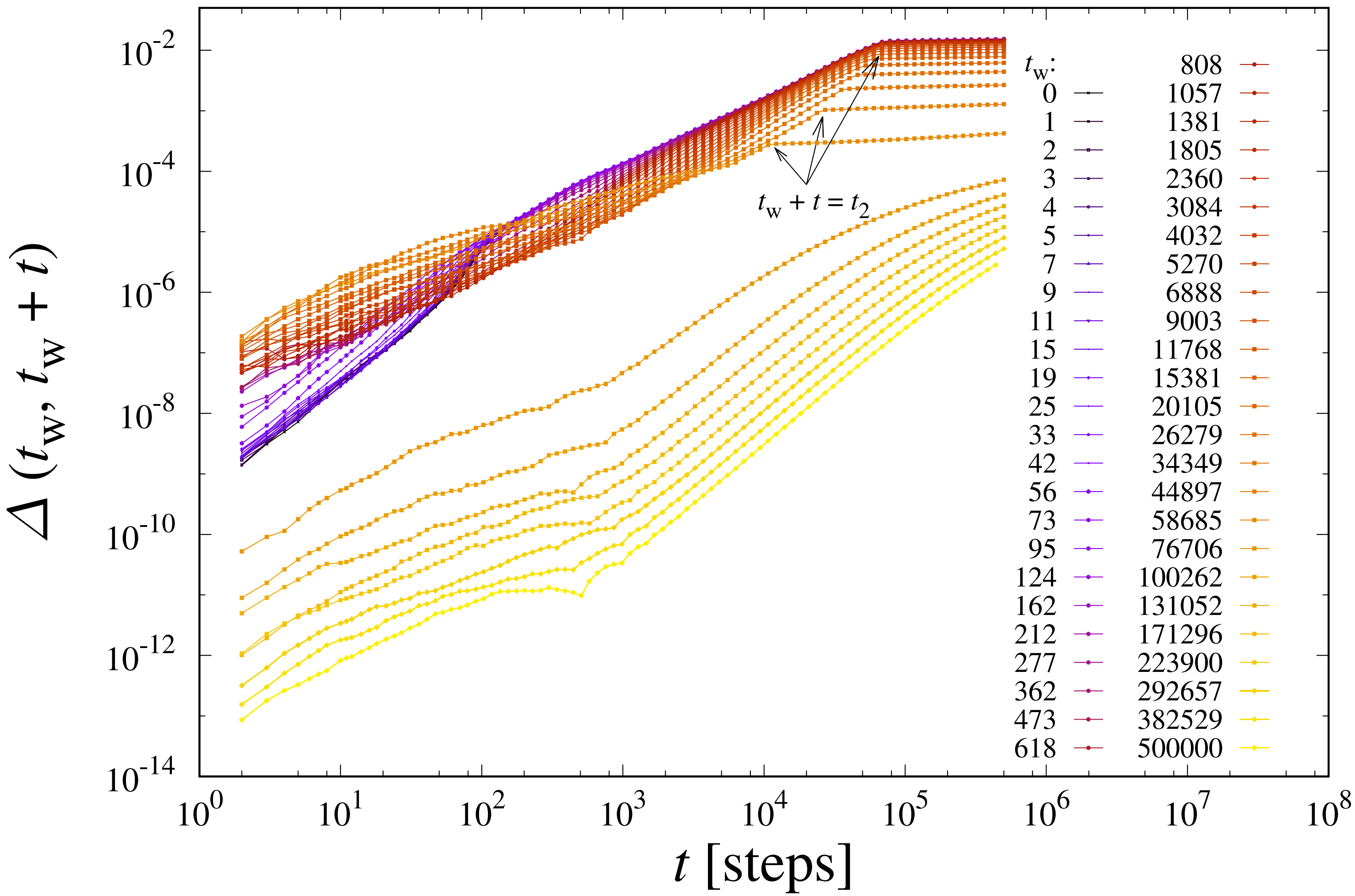


Flat bottom of the Loss landscape

Learning as Interrupted Aging and Diffusion

MBJ, LS, MG, SS, GBA, CC, YLC, MW, GB (2018)

$$\Delta(t_w, t_w + t) = \frac{1}{N}\sum_i (w_i(t_w) - w_i(t_w + t))^2 \qquad D = \frac{1}{|\text{train set}|}\sum_{\alpha \in train\ \ set} |\nabla \mathcal{L}_\alpha - \nabla \mathcal{L}_{\text{emp}}|^2$$



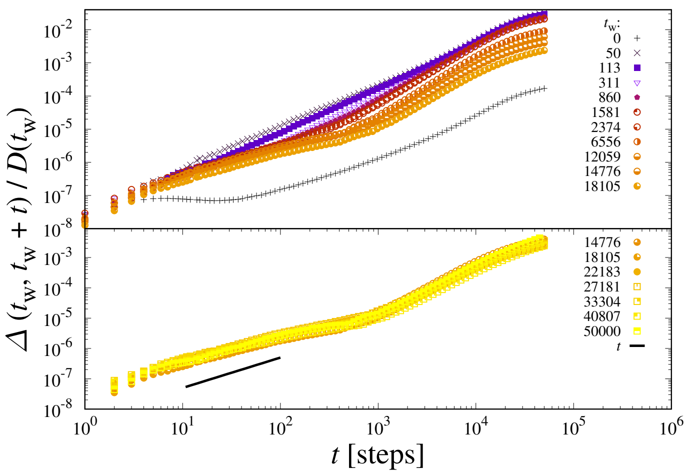
Flat bottom of the Loss landscape

Learning as Interrupted Aging and Diffusion

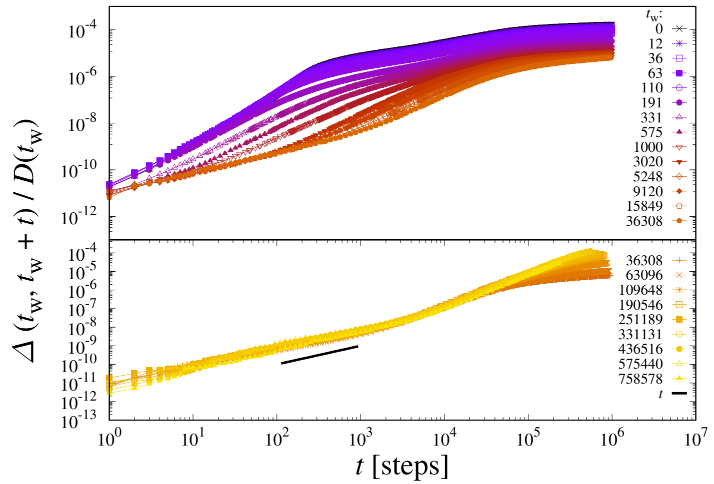
MBJ, LS, MG, SS, GBA, CC, YLC, MW, GB (2018)

$$\Delta(t_w, t_w + t) = \frac{1}{N}\sum_i (w_i(t_w) - w_i(t_w + t))^2$$

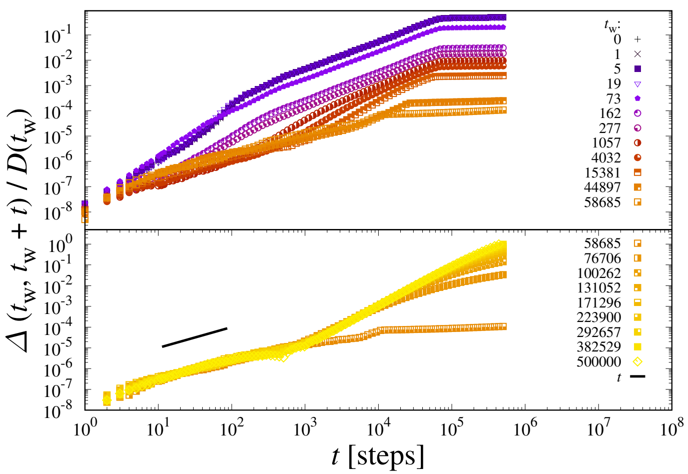
$$D = \frac{1}{|\text{train set}|}\sum_{\alpha \in train\ \ set} |\nabla \mathcal{L}_\alpha - \nabla \mathcal{L}_{\text{emp}}|^2$$



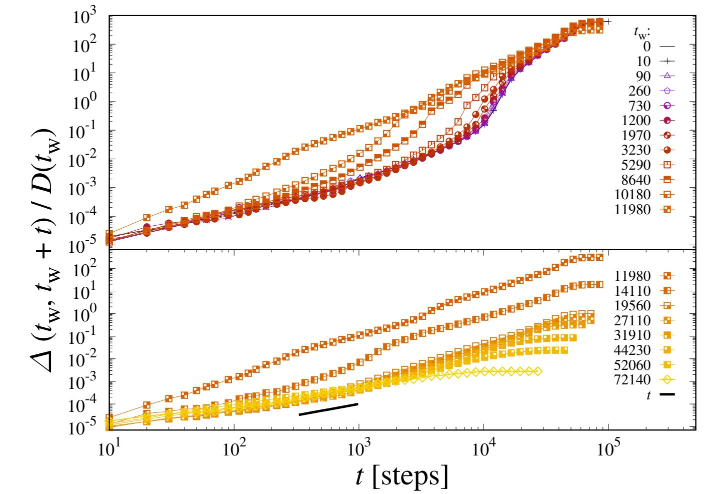
(a) Toy Model on CIFAR-10, $B = 100$, $\alpha = 0.1$.



(b) Fully Connected on MNIST, $B = 128$, $\alpha = 0.01$.



(c) Small Net on CIFAR-10, $B = 100$, $\alpha = 0.01$.



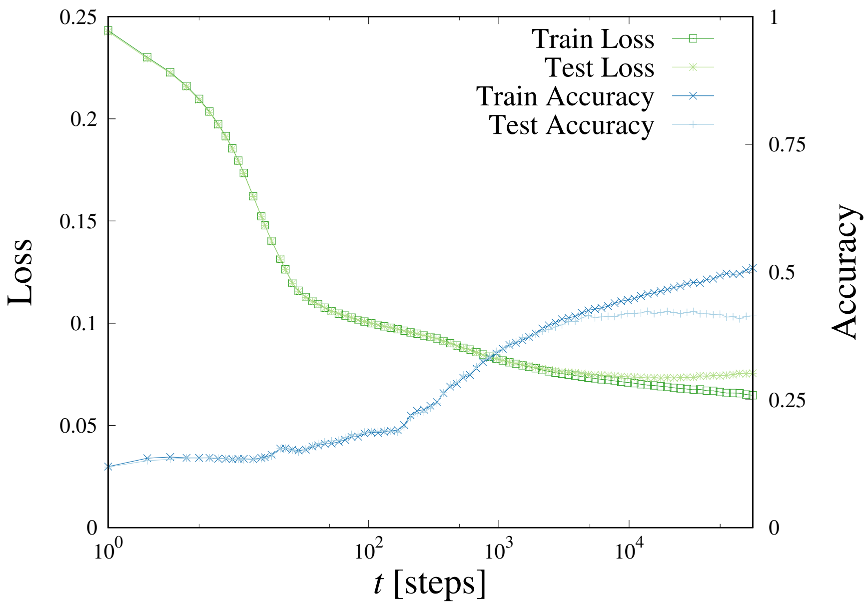
(d) ResNet-18 on CIFAR-100. $B = 64$, $\alpha = 0.01$.

Flat bottom of the Loss landscape

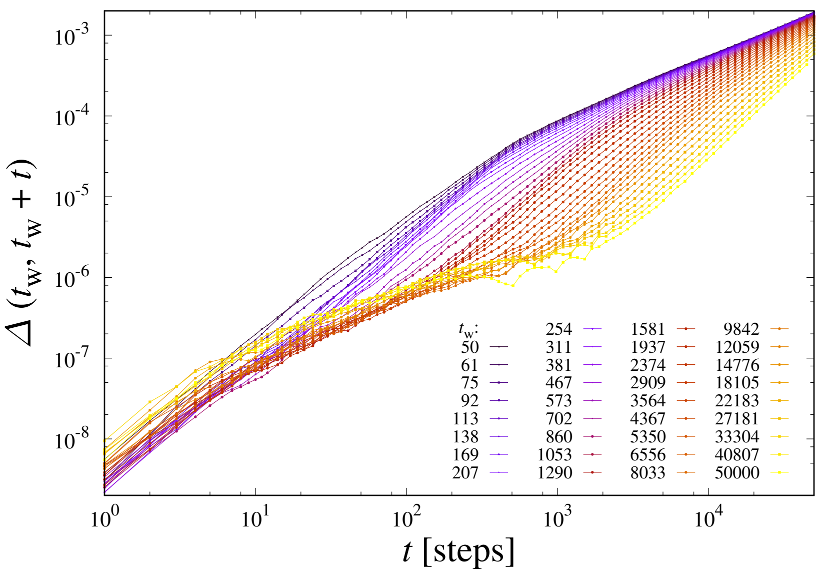
Uninterrupted Aging in under-parametrised NN

MBJ, LS, MG, SS, GBA, CC, YLC, MW, GB (2018)

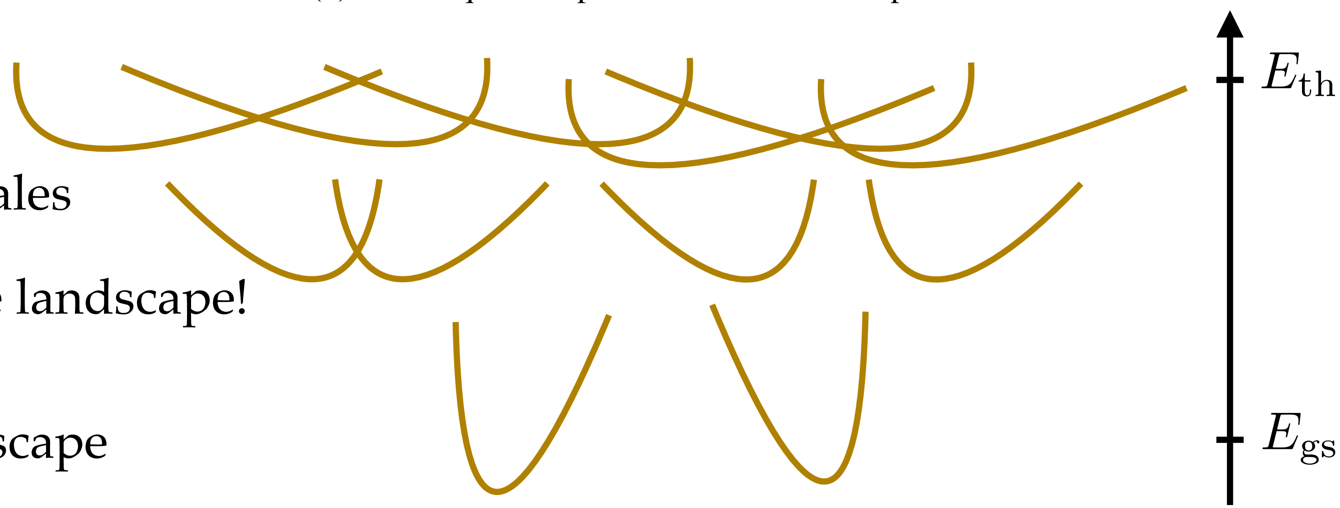
Toy model: 1 hidden layer (MUCH SMALLER), ReLU, sigmoid in output, MSE as a loss



(a) Loss of the under-parametrized model.



(b) Mean square displacement of the under-parametrized model.



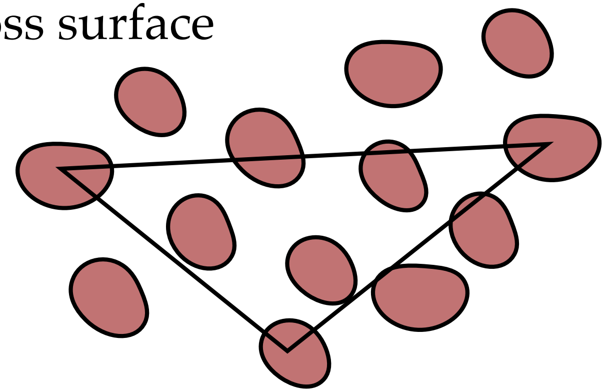
Aging on infinitely long timescales

Not getting to the bottom of the landscape!

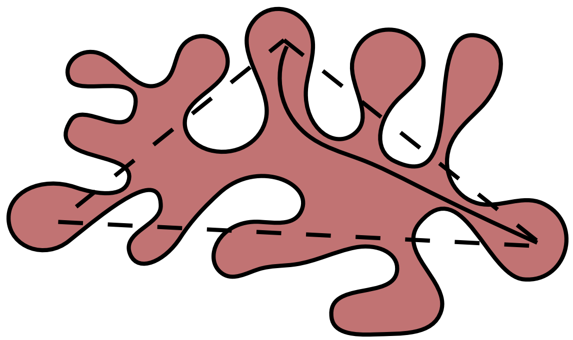
Rough bottom of the Loss landscape

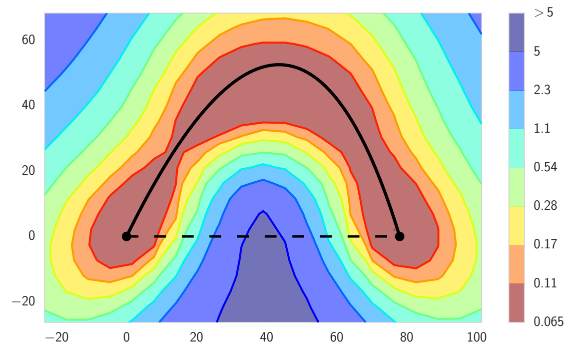
Conclusion, open questions

Aging, complex loss surface



Interrupted Aging, flat loss bottom *Draxler et al. 2018, Garipov et al. 2018, Sagun et al. 2018*





Study of the overparametrised / underparametrised transition

Spigler, Geiger et al. 2018; Geiger et al. 2019

Is there a way to speed up learning in the under-parametrised regime?

Are there bad minima left even in over-parametrised? Why do we avoid them?

Is underlying structure of data important?

Thank you!