Comparing Dynamics: Deep Neural Networks versus Glassy systems

M. Baity-Jesi, L.Sagun, M.Geiger, S.Spigler, G. Ben Arous, C. Cammarota, Y. LeCun, M. Wyart, G. Biroli

Smile seminar: COLT/ICML debrief

January 14, 2020

Outline

Motivation and precautions

Motivations

- Nice ideas:
  - **Experimental** paper
  - Lots of **different people** talking together
  - Another point of view : out of equilibrium **statistical physics**

Motivation and precautions

Motivations

- Nice ideas:
  - **Experimental** paper
  - Lots of **different people** talking together
  - Another point of view : out of equilibrium **statistical physics**
- ~~Nice people on a trendy topic~~

Motivation and precautions

Motivations

- Nice ideas:
  - **Experimental** paper
  - Lots of **different people** talking together
  - Another point of view : out of equilibrium **statistical physics**
- ~~Nice people on a trendy topic~~

Precautions

- A different notion of "showing"
- They take precautions in the paper

Comparison with dynamics of glassy systems

Aim of the article: numerical analysis of the training dynamics of Deep Neural Networks (DNN).

1. Comparison with *glassy systems*
2. Infer energy landscape and dynamics of DNN

Comparison with dynamics of glassy systems

Aim of the article: numerical analysis of the training dynamics of Deep Neural Networks (DNN).

1. Comparison with *glassy systems*
2. Infer energy landscape and dynamics of DNN

Model comparison:

DNN		Glassy systems
loss function	$\Longleftrightarrow$	energy
weights	$\Longleftrightarrow$	degrees of freedom
data set	$\Longleftrightarrow$	parameter defining energy
SGD	$\Longleftrightarrow$	quench and Langevin dynamics

Basic facts on glassy dynamics

Model: 3-spin model.

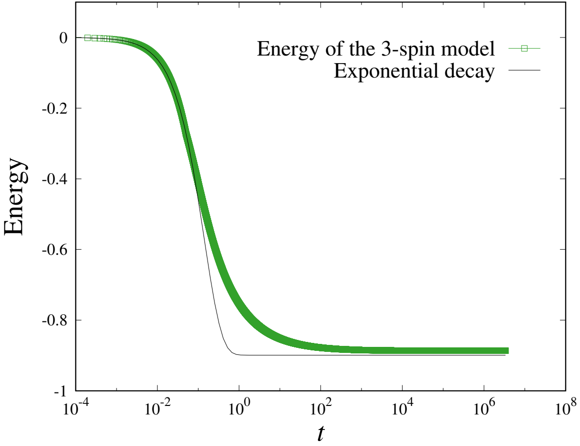
- N spins $(\sigma_i)_{i \leqslant N}$ such that $\sum_i \sigma_i^2 = N$.
- **Interactions**: J i.i.d. centered Gaussian with variance $3/N^2$.
- **Energy** of the system:

$$E = -\sum_{i_1, i_2, i_3} J_{i_1, i_2, i_3} \sigma_{i_1} \sigma_{i_2} \sigma_{i_3}$$

Dynamics: model transition at T^*.

1. Quench from $T_i = \infty$ to $T_f < T^*$
2. Relaxation follows *Langevin dynamics*.

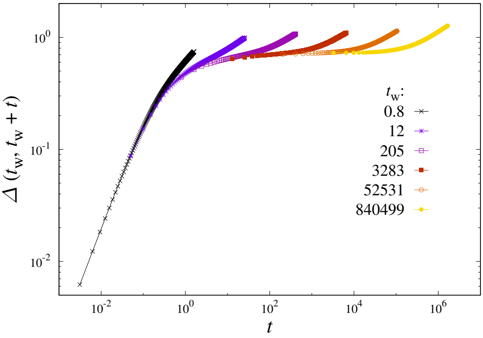
Two main observables: Energy



Two main observables: Mean square displacement

Off-equilibrium: Δ depends on t_w

$$\Delta(t_w, t_w + t) = \frac{1}{N}\sum_{i=1}^{n}(\sigma_i(t_w) - \sigma_i(t_w + t))^2$$

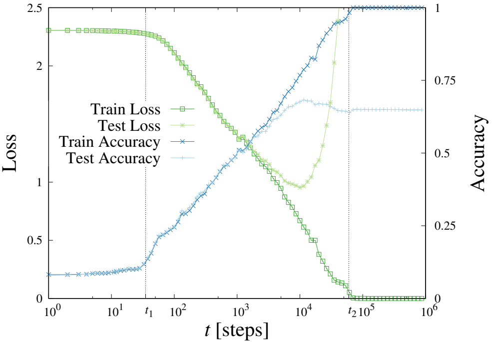


Aging

- The longer t_w, the longer it takes to decorrelate
- Increasingly slow dynamics due to more and more flat directions.
- Dynamic never converges

Comparison with DNN: Energy

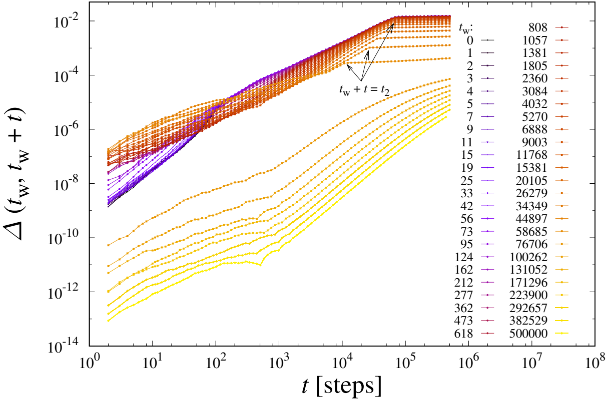
Trained 4 models: from toy model to Resnet18, from MNIST to CIFAR-100.
Show only one: small convolutionnal network on CIFAR-10.



Comparison with glassy dynamics:

- Almost the same
- Optimization: slower decay for $t_1 < t < t_2$
- Reaches bottom of landscape for $t_2 < t$

Comparison with DNN: Mean square displacement

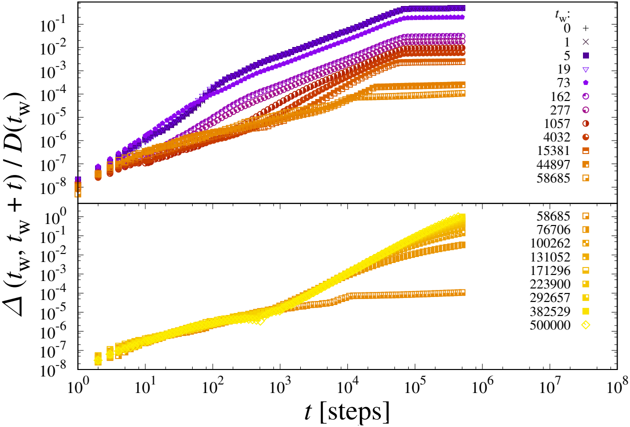


Three regimes

1. $t_w < t_1$: visiting
2. $t_1 < t_w < t_2$: ***aging***: dependence in t_w and plateau
3. $t_w > t_2$: *diffusion at the bottom* (next slide)

Diffusion at the bottom of the landscape

Difference with spin-glasses: no aging after t_2, but diffusion at the bottom of the landscape. Indeed, after rescaling with the right diffusion factor, curbs collapse after t_2:



Discussion

They exhibit **three different regimes**:

1. Initial exploration

Discussion

They exhibit **three different regimes**:

1. Initial exploration
2. Aging dynamics

Discussion

They exhibit **three different regimes**:

1. Initial exploration
2. Aging dynamics
3. Stationnary: diffusive in the bottom of the landscape

$\Longrightarrow$ **Not barrier crossing** but slowing down due to increasing of **flat directions**

A conjecture: existence of a phase transition

Conjecture the existence of a phase transition:

- Over-parametrized models $\longrightarrow$ diffusion in the bottom of the landscape $\longrightarrow$ learn well.
- Under-parametrized models $\longrightarrow$ Glassy dynamics until the end $\longrightarrow$ may take infinite time to learn.