Paths in the Loss Landscapes of Neural Networks along Flat Regions and Symmetries

POLITECNICO
MILANO 1863

DIPARTIMENTO DI ELETTRONICA
INFORMAZIONE E BIOINGEGNERIA

Fabrizio Pittorino

Statistical physics & machine learning back together again

Cargèse, July 2023

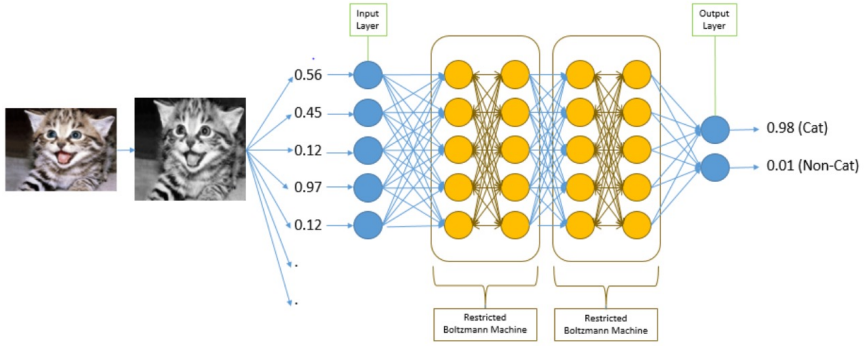
Artificial Intelligence Lab
Bocconi University, Milan

B. Annesi, C. Baldassi, C. Feinauer, C. Lauditi, C. Lucibello, E. Malatesta,

M. Mézard, R. Pacelli, G. Perugini, L. Saglietti, R. Zecchina…

Neural Networks Optimization

Feed forward neural network



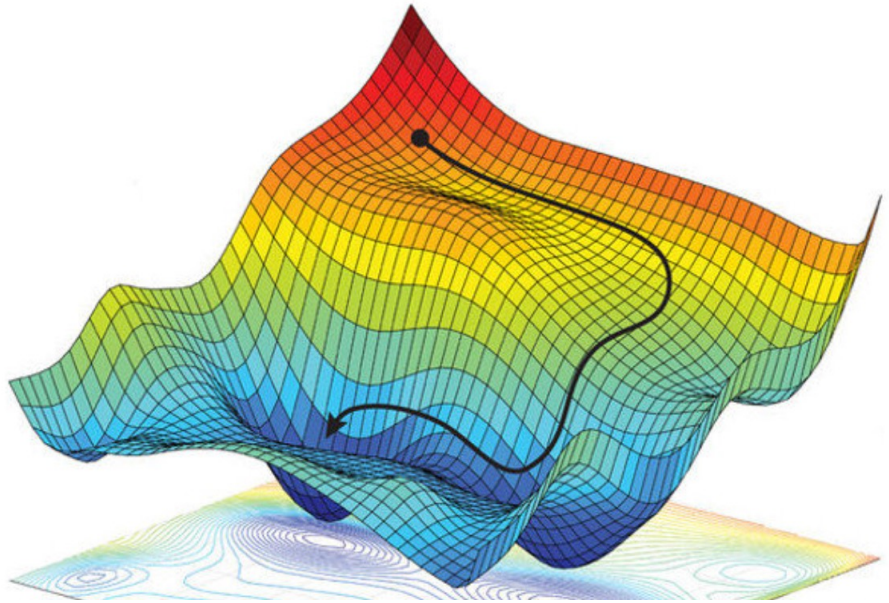
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

where N is the number of data points, f_i the value returned by the model and y_i the actual value for data point i.

SGD dynamics:

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \boldsymbol{\nabla} f^{\mathcal{B}}\left[\boldsymbol{\theta}(t)\right]$$

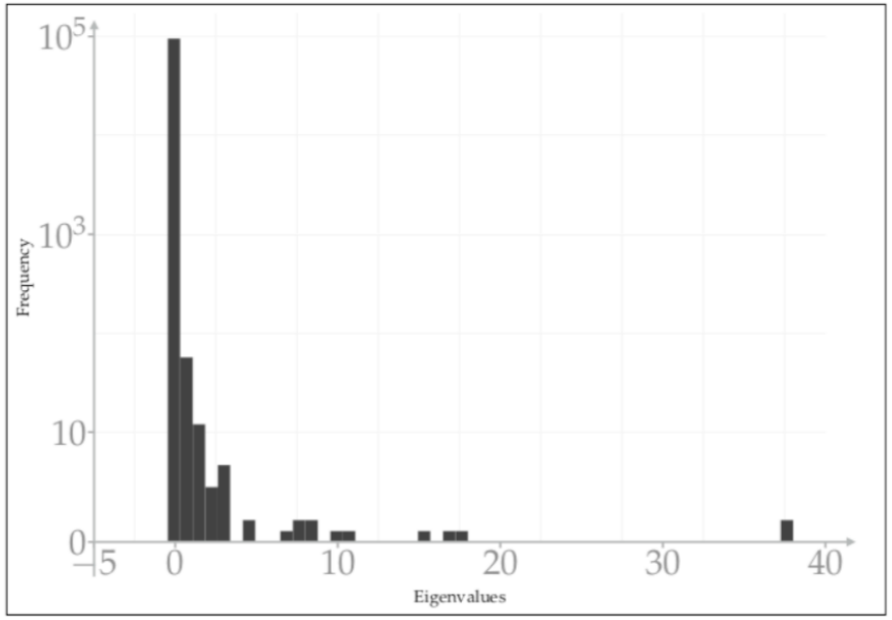
$$f^{\mathcal{B}}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{B}|}\sum_{\alpha \in \mathcal{B}} f_\alpha(\boldsymbol{\theta})$$



Robustly achieve low error on the training set (do not get stuck in bad local minima) and good generalization properties.

Complex Loss Landscapes in Neural Networks

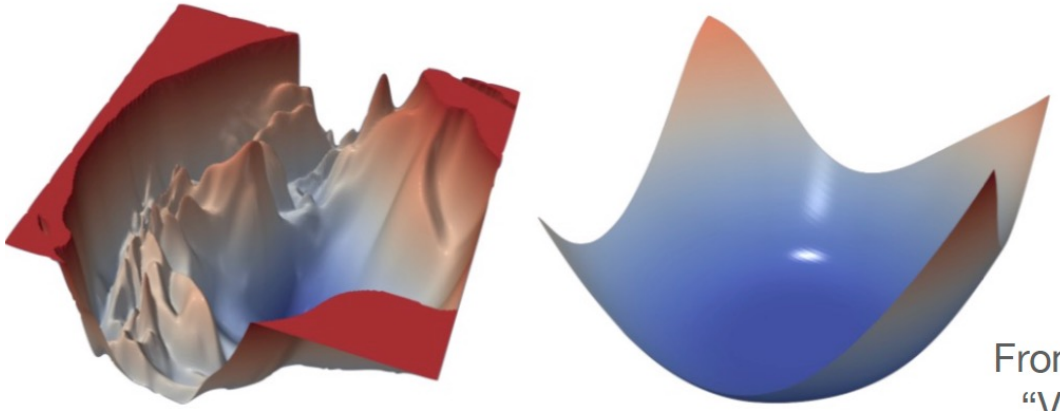
- The spectra of the Hessian in a quasi-minimum. Many flat directions.



From Chaudhary et al '17 "EntropySGD..."

- Architectural choices (e.g. loss, activations, batch-norm, skip-connections) influence the roughness and the large-scale structure of the landscape

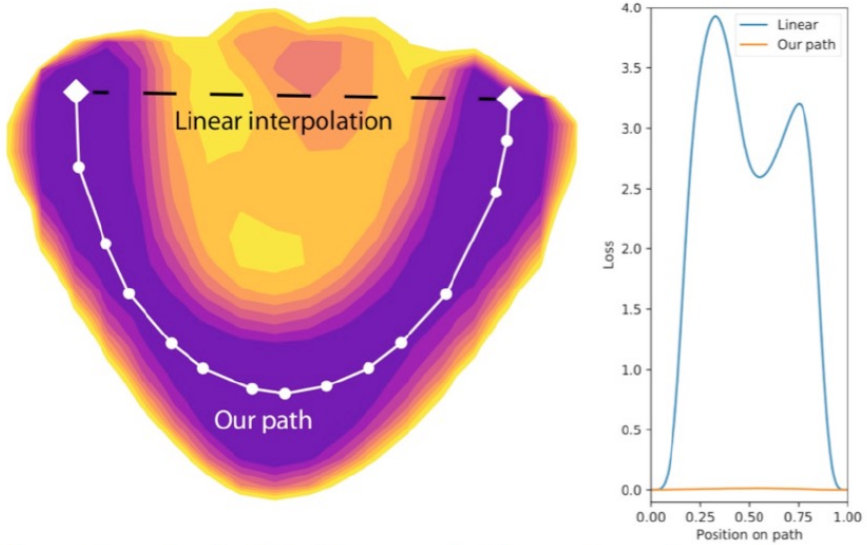
- SGD batch size anti-correlates with minima width and with generalization



(a) ResNet-110, no skip connections (b) DenseNet, 121 layers

From Li et al. '17, "Visualizing..."

- There seems to be a flat non-convex "bottom" connecting "accessible" minimizers.



From Draxler et al. '18, "Essentially no barriers..."

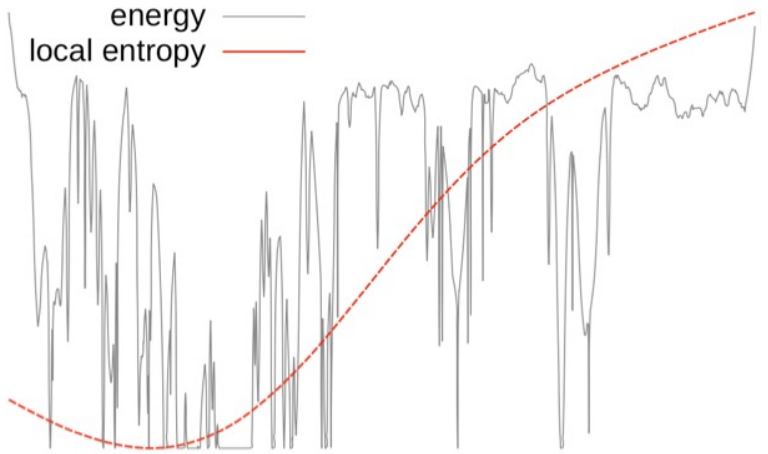
Detecting flat minima: local entropy and local energy

- Main idea: large-deviation analysis, **bias the statistical measure towards dense (wide, flat) regions**
- In practice: define a "**local entropy**" (basically the free volume of a region of the configuration space)

$$\Phi\left(\tilde{W};\beta,\gamma\right)=\log\sum\nolimits_{\{W\}}\exp\left(-\beta\mathcal{L}_{\mathrm{NE}}\left(W\right)-\gamma d\left(W,\tilde{W}\right)\right)$$

"count all configurations below a certain loss and within a certain distance from a reference"

- Instead of minimizing the loss (the energy), **minimize the local entropy**



A proxy of the local entropy:

- The **local energy** is a simple flatness measure:

$$\delta E_{\mathrm{train}}(w,\sigma)=\mathbb{E}_z\,E_{\mathrm{train}}(w+\sigma z\odot w)-E_{\mathrm{train}}(w)$$

Where the noise is: $z\sim\mathcal{N}(0,I_N)$

C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PRL, 2015
C. Baldassi, **F. Pittorino**, R. Zecchina, PNAS, 2020

Entropic Algorithms: Replicated-SGD

A class of entropic algorithms can be derived starting from

$$p(w) \propto e^{-\beta y \mathcal{L}_{\rm LE}(w)}$$

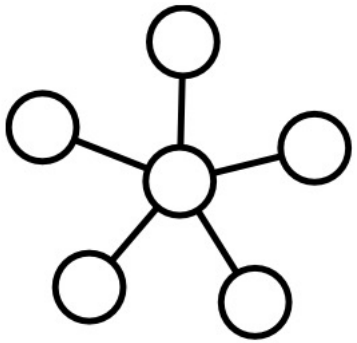
$$\mathcal{L}_{\rm LE}(w) = -\frac{1}{\beta}\log \int dw'\ e^{-\beta \mathcal{L}(w') - \frac{1}{2}\beta\gamma \|w-w'\|^2}$$

Algorithm 2: Replicated-SGD (rSGD)

Input : $\{w^a\}$
Hyper-parameters : y, η, γ, K

for $t = 1, 2, \ldots$ **do**
  $\bar{w} \leftarrow \frac{1}{y}\sum_{a=1}^{y} w^a$
  for $a = 1, \ldots, y$ **do**
    $\Xi \leftarrow$ sample minibatch
    $dw^a \leftarrow \nabla \mathcal{L}\left(w^a; \Xi\right)$
    if $t = 0 \bmod K$ **then**
      $dw^a \leftarrow dw^a + K\gamma\left(w^a - \bar{w}\right)$
    $w^a \leftarrow w^a - \eta\, dw^a$

For **y integer** we have the statistical measure of a system with **y+1 replicas.** We integrate out the original replica and obtain

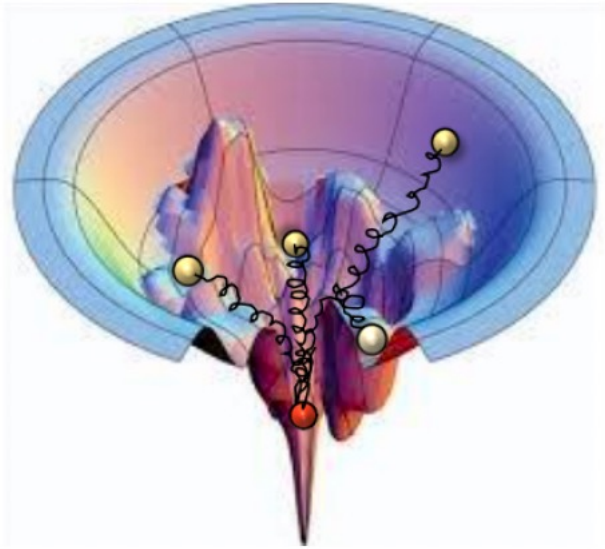


$$p(\{w^a\}_{a=1}^{y}) \propto e^{-\beta \mathcal{L}_{\rm R}(\{w^a\})}$$

where

$$\mathcal{L}_R(\{w^a\}_a) = \sum_{a=1}^{y}\mathcal{L}(w^a) + \frac{1}{2}\gamma\sum_{a=1}^{y}\|w^a - \bar{w}\|^2,$$

with $\bar{w} = \frac{1}{y}\sum_a w^a$. Now one can perform SGD on the replicated loss.

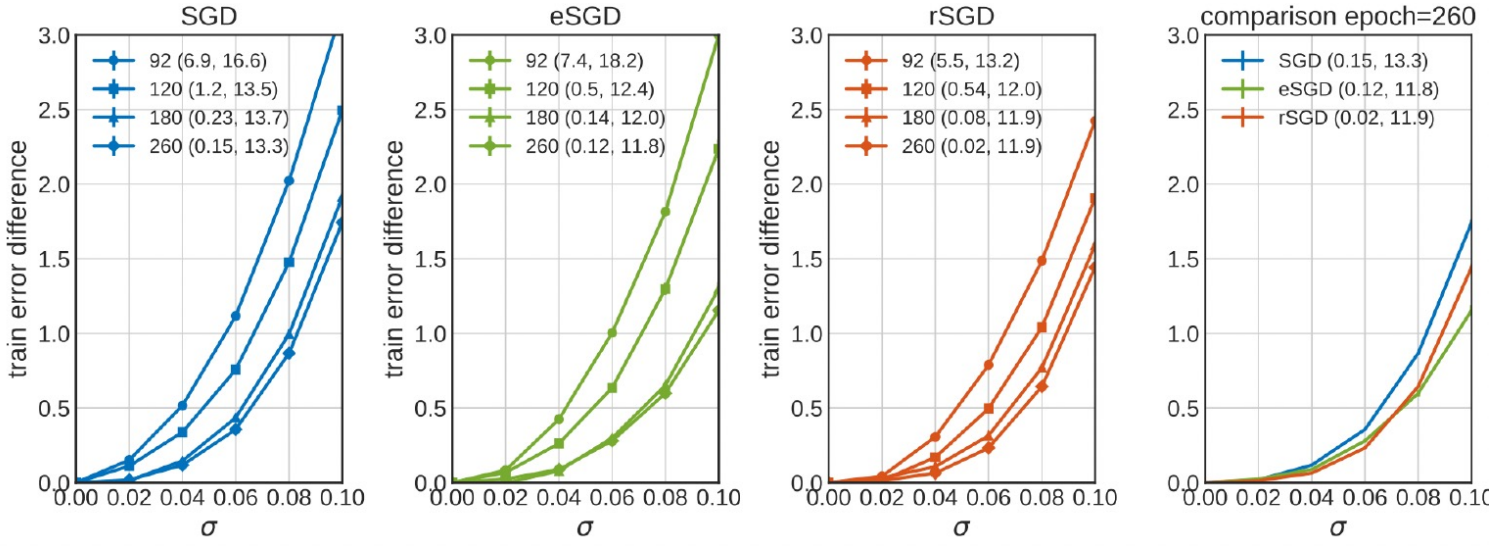


[1] Baldassi, Borgs, Chayes, Ingrosso, Lucibello, Saglietti, Zecchina, PNAS ‘16
[2] **Pittorino**, Lucibello, Feinauer, Perugini, Baldassi, Demyanenko, Zecchina, ICLR 2021

Deep Networks: flatness and generalisation

- State-of-the-art models on CIFAR-10 and CIFAR-100 (standard ML benchmark datasets)

Dataset	Model	Baseline	rSGD	eSGD	rSGD×y
CIFAR-10	SmallConvNet	16.5 ± 0.2	15.6 ± 0.3	14.7 ± 0.3	14.9 ± 0.2
	ResNet-18	13.1 ± 0.3	12.4 ± 0.3	12.1 ± 0.3	11.8 ± 0.1
	ResNet-110	6.4 ± 0.1	6.2 ± 0.2	6.2 ± 0.1	5.3 ± 0.1
	PyramidNet+ShakeDrop	2.1 ± 0.2	2.2 ± 0.1		1.8
CIFAR-100	PyramidNet+ShakeDrop	13.8 ± 0.1	13.5 ± 0.1		12.7
	EfficientNet-B0	20.5	20.6	20.1 ± 0.2	19.5
Tiny ImageNet	ResNet-50	45.2 ± 1.2	41.5 ± 0.3	41.7 ± 1	39.2 ± 0.3
	DenseNet-121	41.4 ± 0.3	39.8 ± 0.2	38.6 ± 0.4	38.9 ± 0.3



Local Entropy is expensive to compute, we compute the cheap Local Energy:

$$\delta\epsilon(w) = \mathbb{E}_z\, \epsilon(w(1+\sigma z)) - \epsilon(w)$$

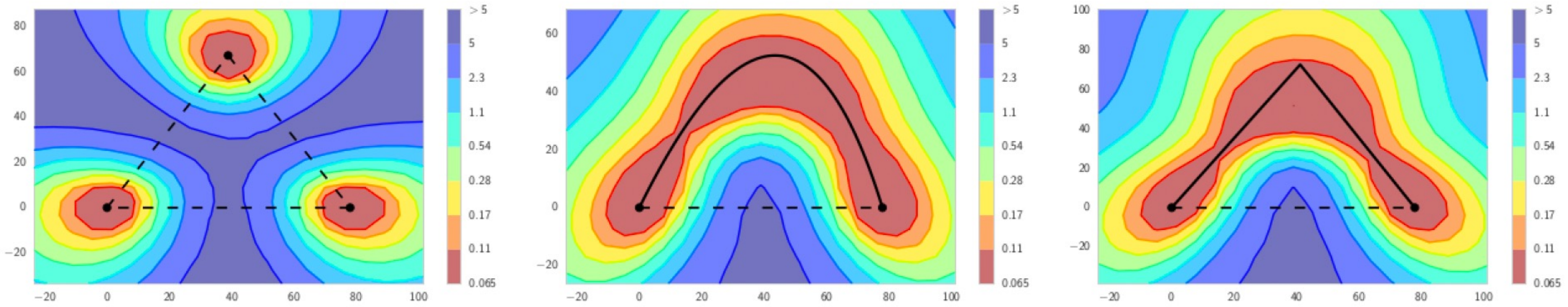
Confirming that entropic algorithm find flatter minima and generalize better

Pittorino, Lucibello, Feinauer, Perugini, Baldassi, Demyanenko, Zecchina, ICLR 2021

Paths in the neural network landscapes

- can be explored in terms of the *loss* or *error* space
- determine the dynamics of gradient descent
- exhibit non-trivial symmetries
- points at the same height can have drastically different generalization properties, connected to **flatness**
- minima found independently can often be connected with relatively simple paths

Loss Landscapes exhibit non-trivial Symmetries



Plot from *Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., & Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of dnns. Advances in neural information processing systems.*

Our work

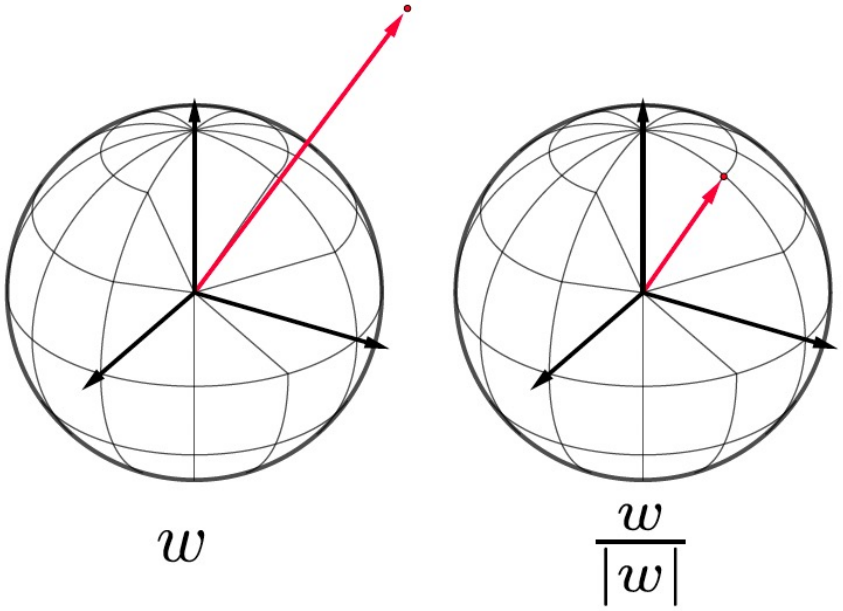
- studies the error landscape
- … after removing symmetries
- … around solutions found with different algorithms (entropic algorithms and standard algorithms)
- … in networks with continuous and binary weights

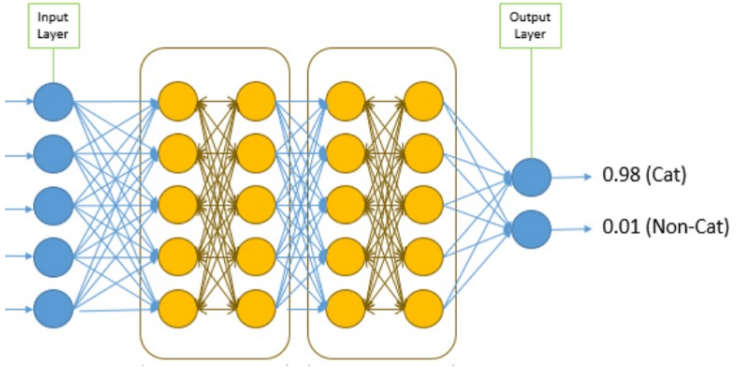
Symmetries: Weight-Norm (rescaling)

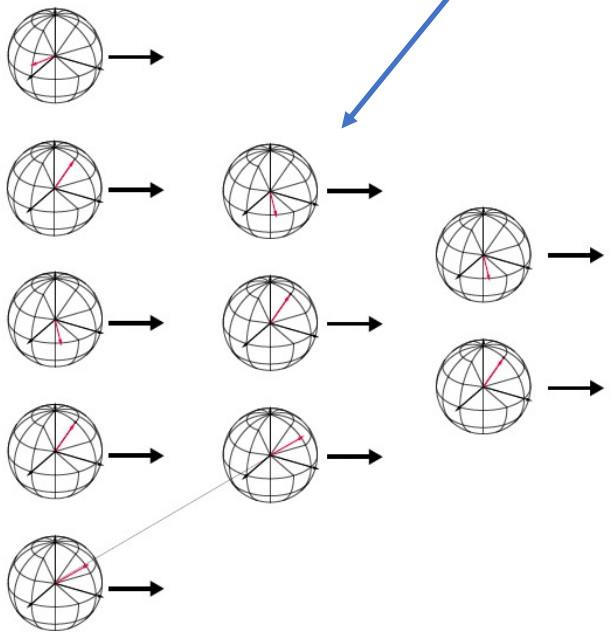
For the *ReLU* activation function we have $f(\alpha x) = \alpha f(x)$ for $\alpha \geq 0$ and therefore

$$f(\vec{w} \cdot \vec{x}) = |\vec{w}| f\left(\frac{\vec{w}}{|\vec{w}|}\vec{x}\right),$$

which means we can *push* the weight norms to the next layer.







Every neuron becomes a hypersphere with norm=1

The complete network is a product of hyperspheres

[The last layer precedes a `argmax` and since we are only interested in the error we can normalize it *globally* with a positive factor]

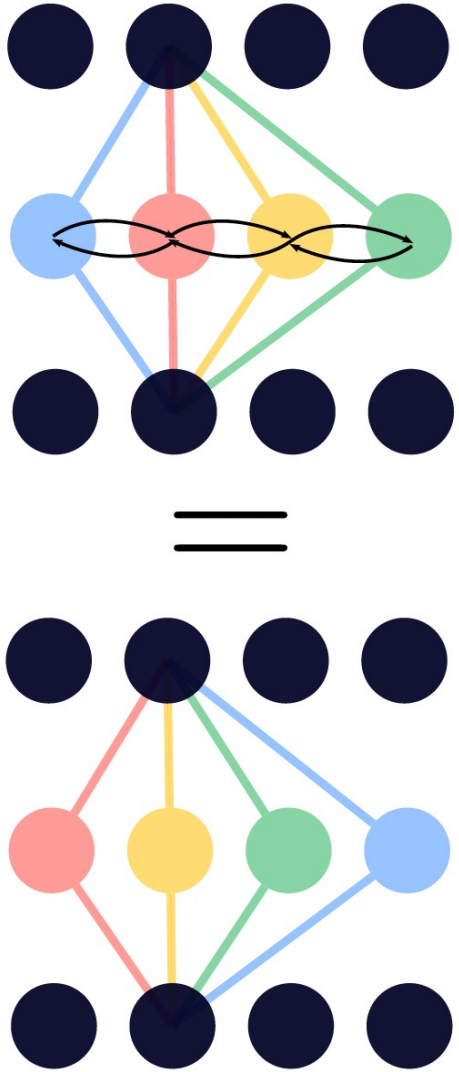
Symmetries: Permutation

- Neurons within a layer can be exchanged as long as incoming and outgoing connections are taken care of
- The same goes for kernels in convolutional layers
- Networks need to be *aligned* before comparing them

Simple Algorithm for Matching

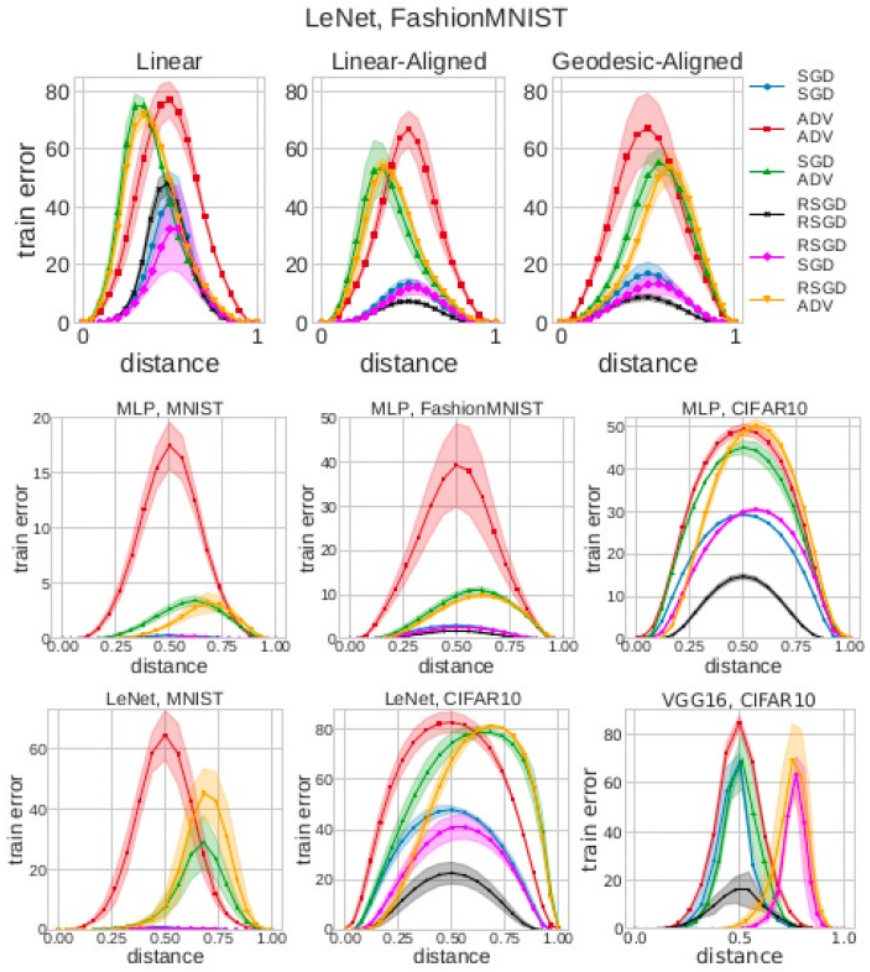
```
Input: Two normalized NNs with parameters A[1..L], B[1..L] and L layers
for l = 1 to L - 1 do
    π = Match(A[l] , B[l])
    PermutePrev(B[l], π)
    PermuteNext(B[l+1] , π ` )
end for
```

`Match` takes two sets of parameter vectors and solves a weighted bipartpite graph matching problem using cosine similarities between vectors as weights.



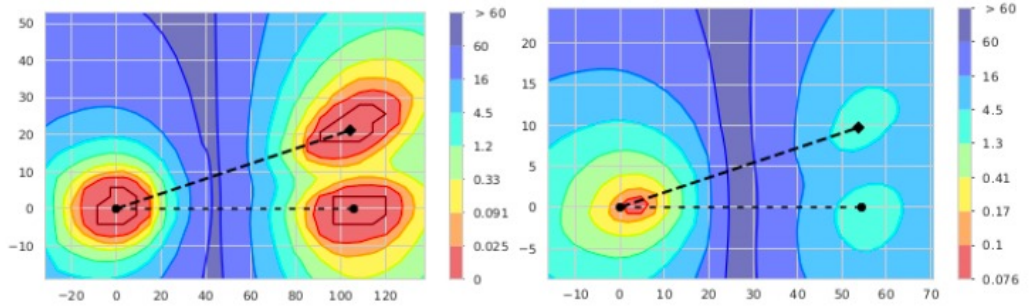
Pittorino, Ferraro, Perugini, Feinauer, Baldassi, Zecchina,
Deep Networks on Toroids: Removing Symmetries Reveals the Structure of Flat Regions in the Landscape Geometry

Mode Connectivity



- SGD: Stochastic Gradient Descent
- RSGD: *Replicated SGD* (finds flatter minima)
- ADV: Adversarial Initialization
- Linear: Straight path between minima
- Geodesic: Path on normalized sub-manifold

VGG16 on Cifar10



- Left Panel: Unnormalized
- Right Panel: Normalized
- Left Points: RSGD (finds flatter minima)
- Right Points: unaligned/aligned SGD with adversarial initialization

Difference is only visible *after* symmetry removal

Pittorino, Ferraro, Perugini, Feinauer, Baldassi, Zecchina, Deep Networks on Toroids: Removing Symmetries Reveals the Structure of Flat Regions in the Landscape Geometry

Simplest non-convex continuous NN model

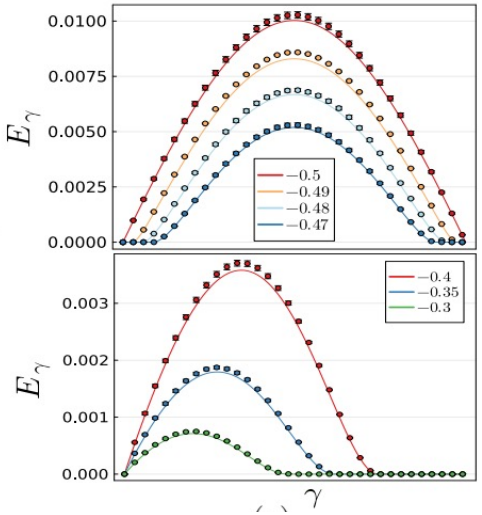
Negative perceptron $\kappa_E < 0$

$$\boldsymbol{W} \cdot \boldsymbol{\xi}^\mu > \kappa_E \sqrt{N}\,, \quad \mu \in [P]$$

Clarissa Lauditi's Poster

Uniform probability density over solutions

$$p_{\boldsymbol{\xi},\kappa_E}(\boldsymbol{W}) = \frac{1}{Z_{\boldsymbol{\xi},\kappa_E}} \delta(\|\boldsymbol{W}\|^2 - N) \prod_{\mu=1}^{P} \Theta(\boldsymbol{W} \cdot \boldsymbol{\xi}^\mu - \kappa_E \sqrt{N})$$



y=2

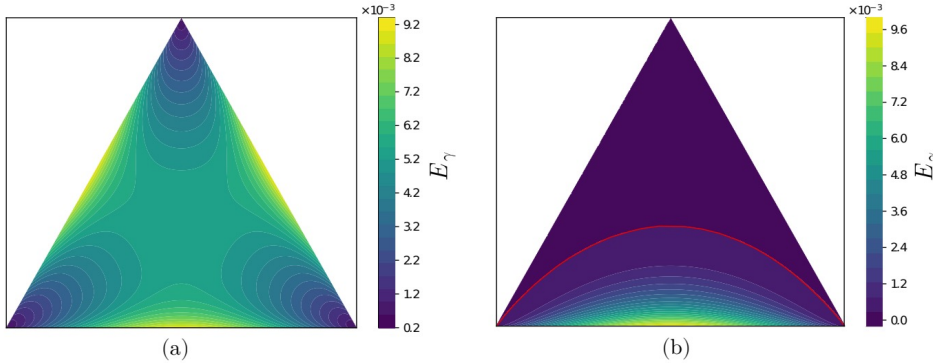
Main result

Typical energy landscape between groups of y solutions

$$E_{\boldsymbol{\gamma}} = \lim_{N \to +\infty} \mathbb{E}_{\boldsymbol{\xi}} \left\langle \Theta\big(- \boldsymbol{W}_{\boldsymbol{\gamma}} \cdot \boldsymbol{\xi}^\mu + \kappa_E \sqrt{N} \big) \right\rangle_{k_1,\ldots,k_y}$$

Where the interpolating solutions are the geodesic projection on the N-sphere of the y-symplex

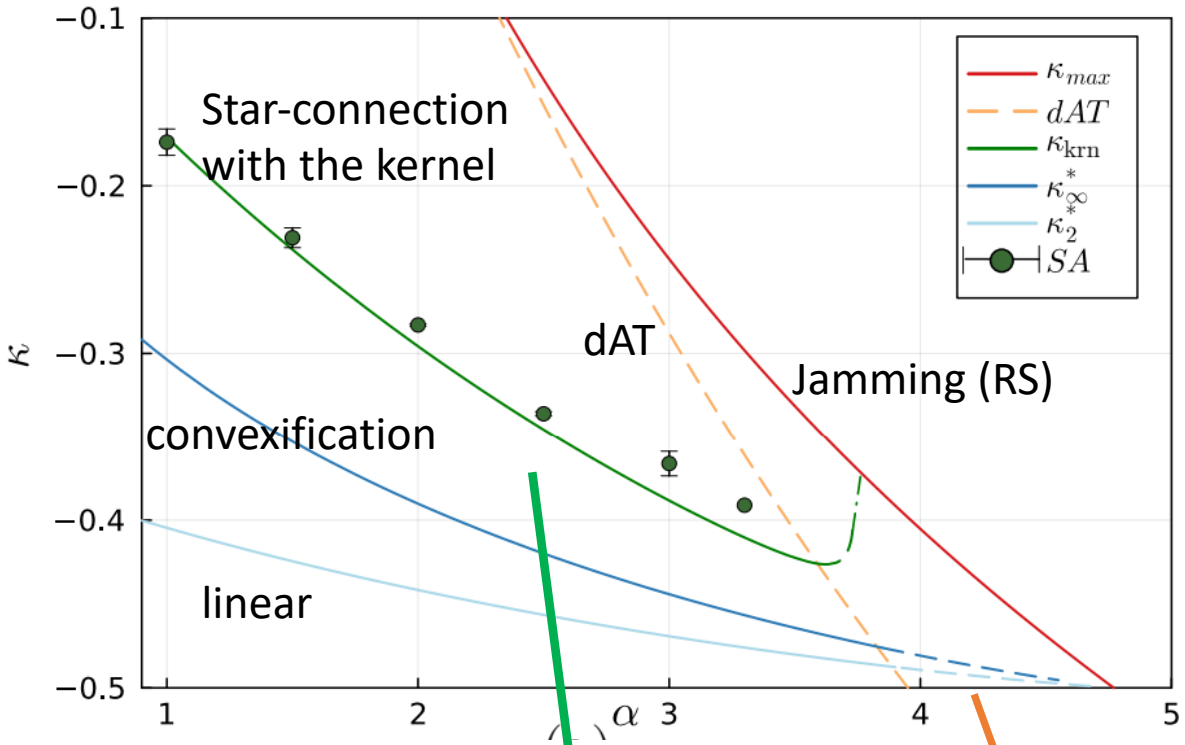
$$\boldsymbol{W}_{\boldsymbol{\gamma}} = \frac{\sqrt{N} \sum_{r=1}^{y} \gamma_r \boldsymbol{W}^r}{\| \sum_{r=1}^{y} \gamma_r \boldsymbol{W}^r \|}$$



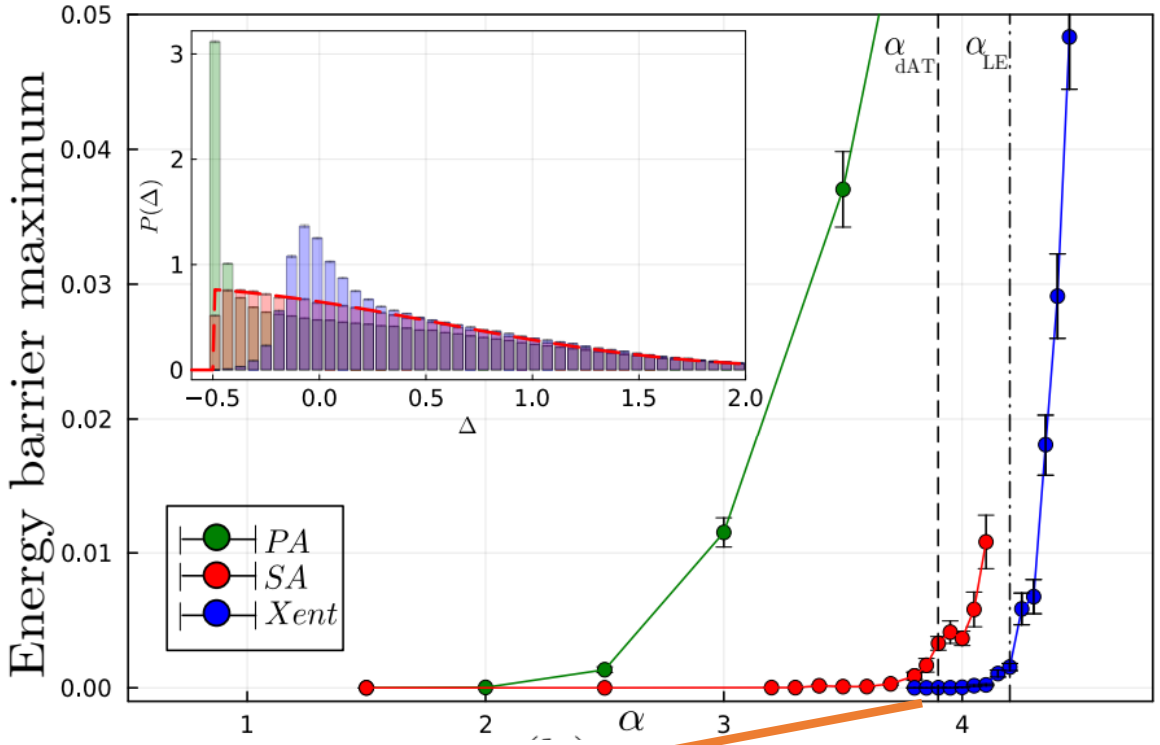
y=3

The star-shaped space of solutions of the spherical negative perceptron, Annesi, Lauditi, Lucibello, Malatesta, Perugini, **Pittorino**, Saglietti, https://arxiv.org/abs/2305.10623

Coalescence thresholds and disconnection transitions

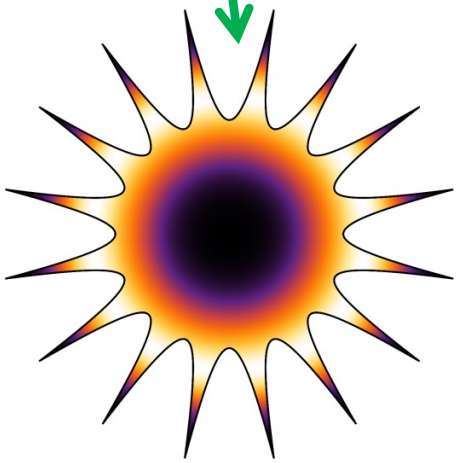


Coalescence thresholds



Numerically estimated disconnection transitions

Star-shaped set of solutions



And then augmenting the constraint density α other disconnection transitions,
the "local entropy transition" (Typical and atypical solutions in non-convex neural networks with discrete and continuous weights, Baldassi, Malatesta, Perugini, Zecchina, https://arxiv.org/abs/2304.13871)…

Only one basin?

The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks

Rahim Entezari[1], Hanie Sedghi[2], Olga Saukh[1], and Behnam Neyshabur[3]

[1]TU Graz / CSH Vienna
[2]Google Research, Brain Team
[3]Google Research, Blueshift Team

simpler view of SGD solutions. We first state our conjecture informally:

Most SGD solutions belong to a set $\mathcal{S}$ whose elements can be permuted in such a way that there is no barrier on the linear interpolation between any two permuted elements in $\mathcal{S}$.

The above conjecture suggests that most SGD solutions end up in the same basin in the loss landscape after proper permutation (see Figure 1 left panel). We acknowledge that the above conjecture is bold.

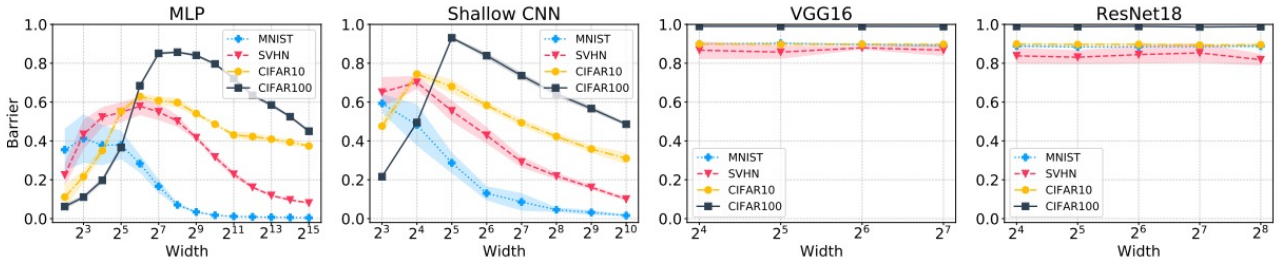


Figure 2: **Effect of width on barrier size.** From **left to right**: one-layer MLP, two-layer Shallow CNN, VGG-16 and ResNet-18 architectures on MNIST, CIFAR-10, SVHN, CIFAR-100 datasets. For large width sizes the barrier becomes small. This effect starts at lower width for simpler datasets such as MNIST and SVHN compared to CIFAR datasets. A closer look reveals a similar trend to that of double-descent phenomena. MLP architectures hit their peak at a lower width compared to CNNs and a decreasing trend starts earlier. For VGG and ResNet, the barrier size is saturated at a high value and does not change due to the effect of depth as discussed in Figure 3.

Git Re-Basin: Merging Models modulo Permutation Symmetries

Samuel K. Ainsworth *skainswo@cs.washington.edu*

Jonathan Hayase *jhayase@cs.washington.edu*

Siddhartha Srinivasa *siddh@cs.washington.edu*
School of Computer Science and Engineering
University of Washington

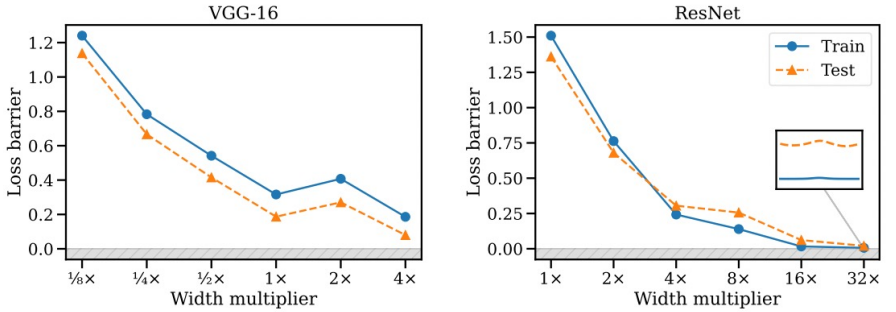
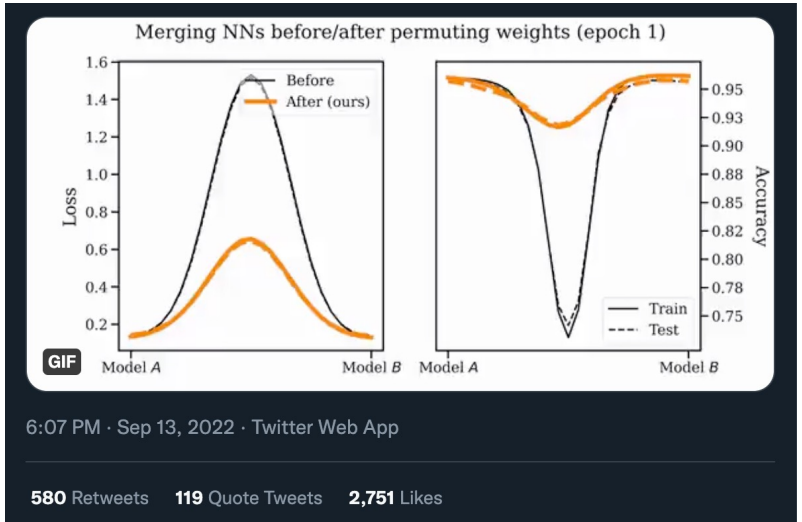
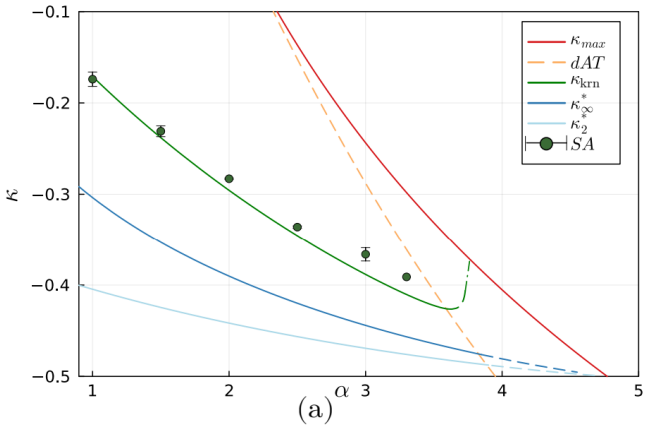


Figure 4: **Wider models exhibit better linear mode connectivity.** Training convolutional and ResNet architectures on CIFAR-10, we ablate their width and visualize their loss barriers after weight matching. Notably, we achieve zero-barrier LMC between ResNet models, the first such demonstration to the best of the authors' knowledge.



Perspectives: extension to realistic models & RS study of disconnection transitions?

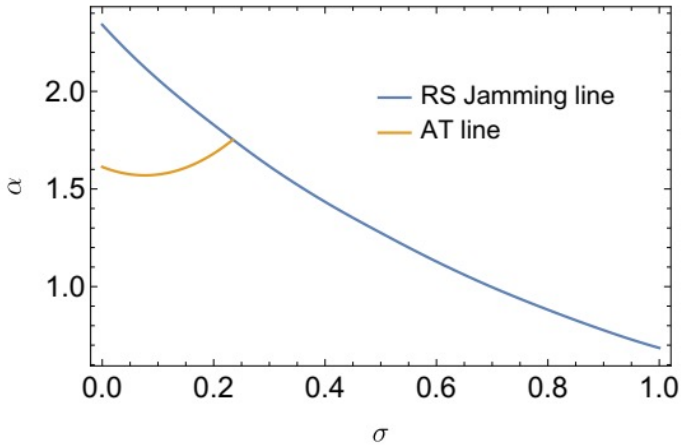
a) The disconnection transitions take place in the fullRSB region



b) The jamming line in the fullRSB region gives rise to critical isostatic regime

Multilayer neural networks have been numerically shown to be in the hypostatic universality class

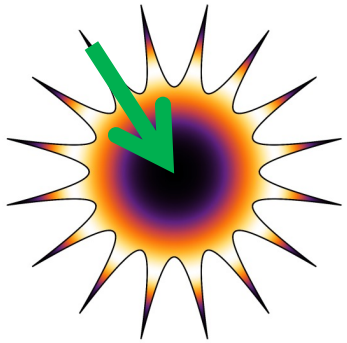
The jamming transition as a paradigm to understand the loss landscape of deep neural networks, Geiger et al.

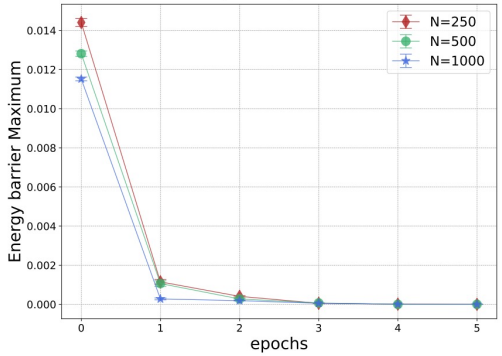


Soft-committee (tree committee machine with Erf activation functions, K=3) possible candidate to analitically study disconnection transitions (RS-solvable) and also more realistic (hypostatic)

Jamming in multilayer supervised learning models, Franz, Hwang, Urbani

The convex core is attractive for SGD.





Maximum barrier height along SGD dynamics initializing in typical solutions

<u>What is the role of the star-shape in SGD optimization and generalization?</u>

Conclusions

- The learning loss landscape of neural networks is characterized by flat minima that have good generalization capabilities
- Starting from this observation, algorithms enhancing flatness and generalization can be designed
- When comparing two or more networks (as in the case of paths in the loss landscape) it is necessary to take into account neural networks symmetries (rescaling and permutation symmetries)
- Debate on the "true" structure of the loss landscape (just one *connected* and *convex* minimum?)
- Take into account permutation symmetry in replica calculations
- Does star-shapedness (->RS non-convex regime) hold in realistic NN models?
- Is it beneficial for SGD optimization and generalization?
- Is it possible to study disconnection transitions in RS (solvable) regime for continuous models?