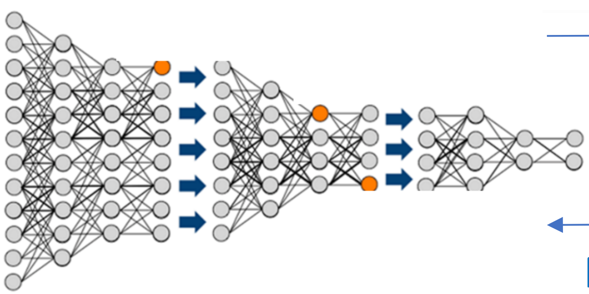
Loss landscape in deep learning: Role of a "Jamming" transition

Matthieu Wyart, PCSL
Physics Institute, EPFL

Mario Geiger, Stefano Spigler, Levent Sagun, Stephane d'Ascoli, Giulio Biroli



Tkatchenko's talk

Montanari's talk

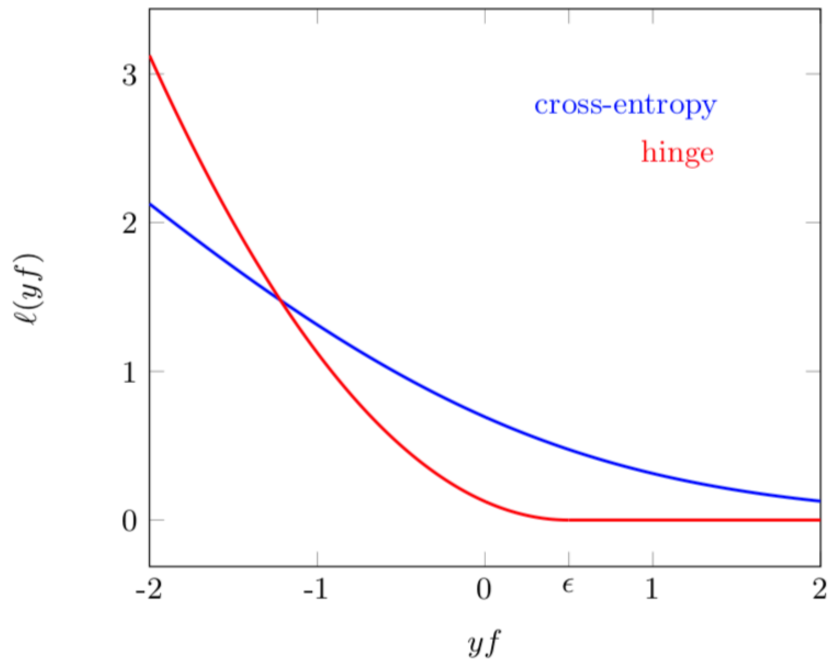
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Set-up

- binary classification task, P training data $\{\mathbf{x}_i, y_i = \pm 1\}$
- Deep net $f_{\mathbf{W}}(\mathbf{x}_i)$ with N parameters.
- Seek W^* Such that $\text{sign}(f_{\mathbf{W}^*}(\mathbf{x}_i)) = y_i$
- Learning: descent in loss function

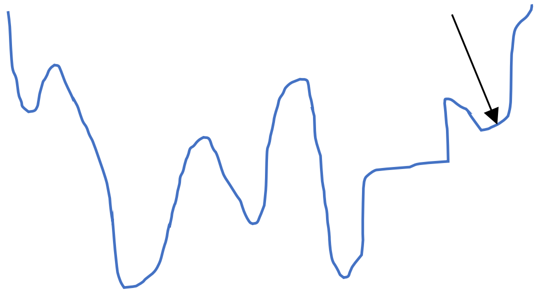
$$\mathcal{L} = \frac{1}{P} \sum_i l(y_i f_{\mathbf{W}}(\mathbf{x}_i))$$

- Typically: cross-entropy



Motivation 1 to describe landscape

- Mezard's talk: why not stuck in bad local minima?
- Choromanska et al. 15: p-spin landscape



- More recent literature: Landscape has plenty of flat directions

Spectrum of Hessian (Sagun's talk)

- Cause: Over-parametrization, N large

Theory: flat directions must be present then *Soudry, Hoffer 17' Cooper 18'* Dynamics: *Baity-Jesy et al. 18'* (Sagun's talk)

Here: Evolution of landscape with N? Sharp transition?

Motivation 2: role of depth?

- Montanari's talk: Universal approximation theorem, one hidden layer *Cybenko 89*

- Why is depth useful then? Expressive power greater Ganguli's talk

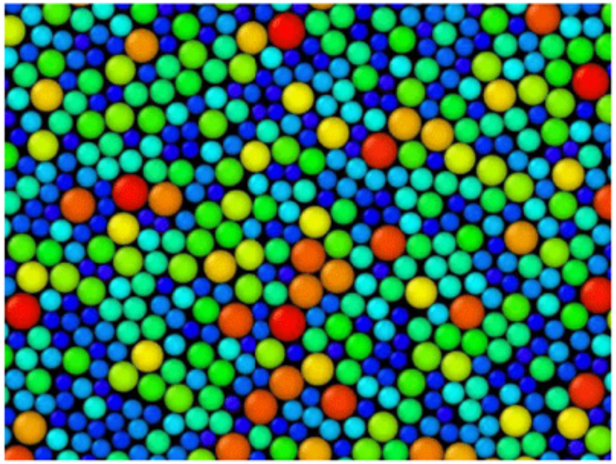
One idea: deep nets more expressive, can fit real data with less parameters, and so generalize better

- Zhang et al. 17: nets can be handcrafted that can fit any data with

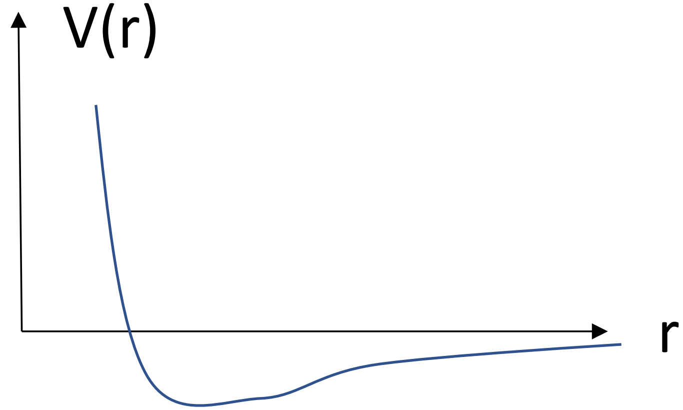
Rather small number of parameters N ~ P.

Can similar solutions be learnt? Effects of depth on the landscape??

Energy landscape in structural glass



Potential energy: $U(\{\mathbf{R}_\alpha\}) = \sum_{\alpha,\beta} V(r_{\alpha,\beta})$

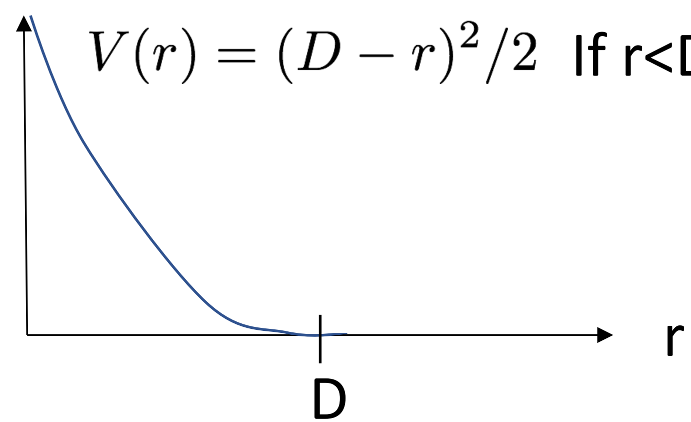


Questions:

-local aspect of landscape: properties of the Hessian of U (vibrational, elastic properties)?

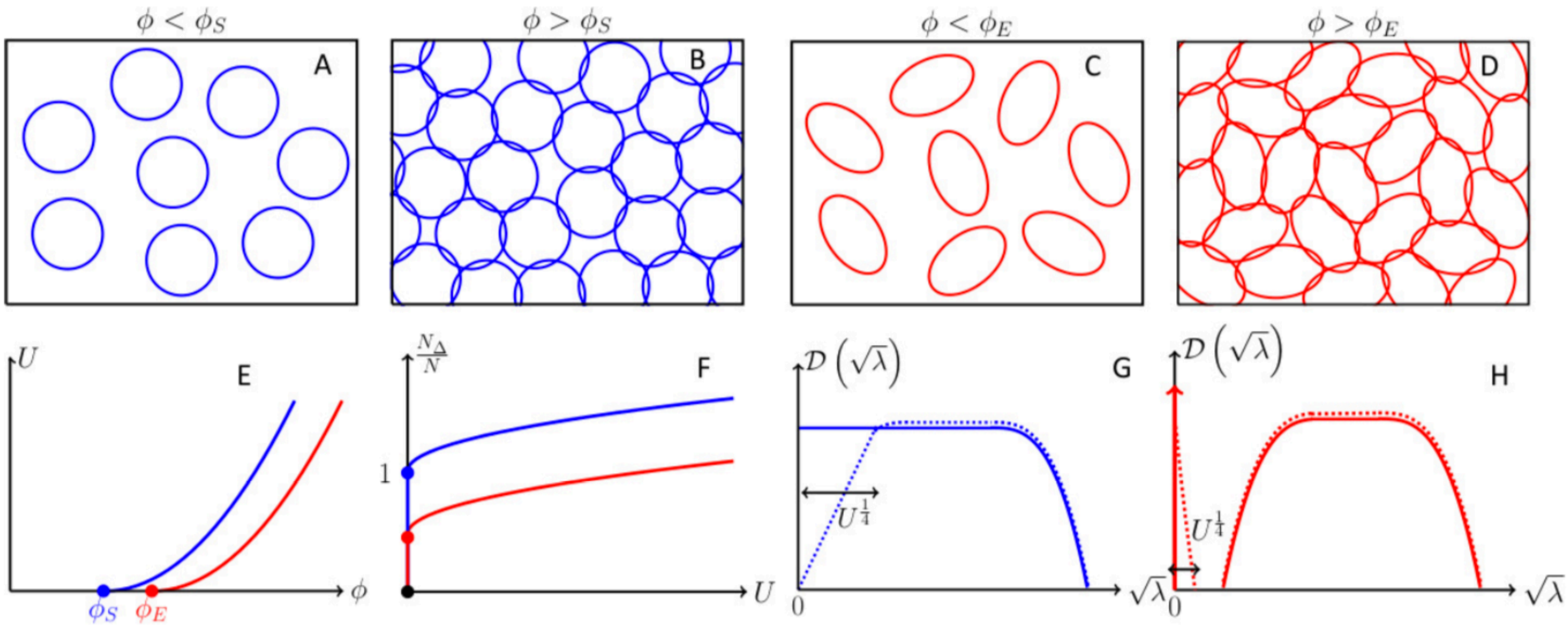
-Non-local aspects of landscape: non-linear response, Flow, etc…

$V(r) = (D-r)^2/2$ If r<D



One approach: finite range potential (granular materials)

Jamming transition: SAT-UNSAT transition with continuous degrees of freedom



<u>Sphere:</u>

-Isostatic $N_\Delta = N$

-Flat spectrum at threshold

<u>Ellipses:</u>

-hypostatic $N_\Delta < N$

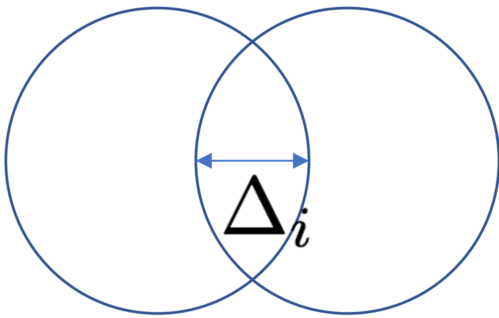
-gapped spectrum

<u>Theory spectrum:</u> Eff. Medium: *M.W. 10', During et al. 13', DeGiuli et al. 14',*
Variational*: Yan et al. 16'* Infinite dimensions: *Franz et al. 15 , Ikeda et al. 18'*

Argument $N_\Delta \leq N$ as jamming approached from above. *Tkachenko, Witten 99*

m : sets of pairs of particles in contact (constraints not satisfied)

Δ_i : overlap between two particles $\Delta_i = D - r_i$



Near jamming:

$$\Delta_i(\{\mathbf{R}_\alpha\}) = 0 \quad \forall\ i \in m$$

Intersection N_Δ manifolds of dimension N-1

Solutions can exist only if more degrees of freedom than constraints

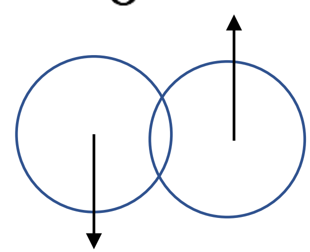
$$\boxed{N_\Delta \leq N}$$

Argument $N = N_\Delta$ for spheres *MW, Silbert, Nagel, Witten 05*

Hessian $N \times N$ matrix

$$\boxed{H = \underbrace{\sum_{i \in m} \nabla\Delta_i \otimes \nabla\Delta_i}_{H_0} + \underbrace{\sum_{i \in m} \Delta_i H_{\Delta i}}_{H_p}}$$

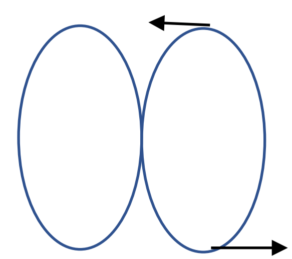
Rank(H_0) $\leq N_\Delta$



<u>Spheres:</u> H_p negative definite

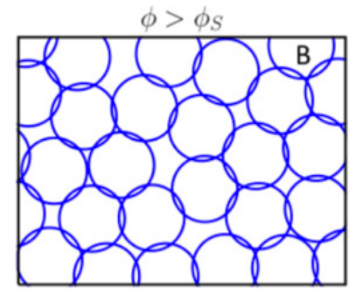
Modes in kernel of H_0 will be unstable:

stability implies H_0 is full rank $N_\Delta \geq N \Rightarrow N = N_\Delta$

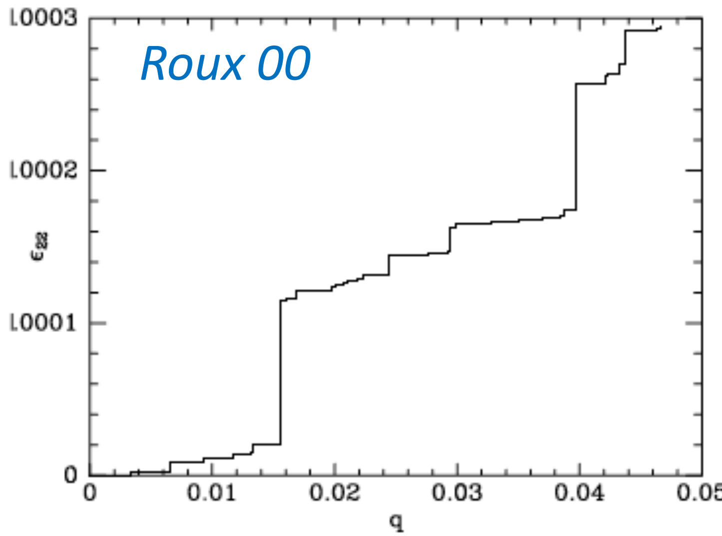


<u>Ellipses</u>: H_p not negative definite,
can stabilize kernel H_0: $N < N_\Delta$

Non-local landscape property



- Distribution forces $\boxed{P_+(\Delta) \sim \Delta^\theta}$
- Distribution gaps $\boxed{P_-(\Delta) \sim (-\Delta)^{-\gamma}}$



- marginal stability implies

$\boxed{\gamma = (1-\theta)/2}$ and crackling noise.

MW 12, Lerner et al. 13, Muller MW 15

- Infinite dimension calculations can compute these exponents

Charbonneau et. al 14. and crackling *Franz Spigler 17*

Continuous symmetry breaking $\gamma = 0.41 \quad \theta = 0.42$

Deep learning?

- Direct analogy between perceptron (in some regime) and jamming of hard spheres *Franz and Parisi 16, Franz et al. 17.* (see Zamponi's talk)

- Deep learning? <u>key idea:</u> finite range

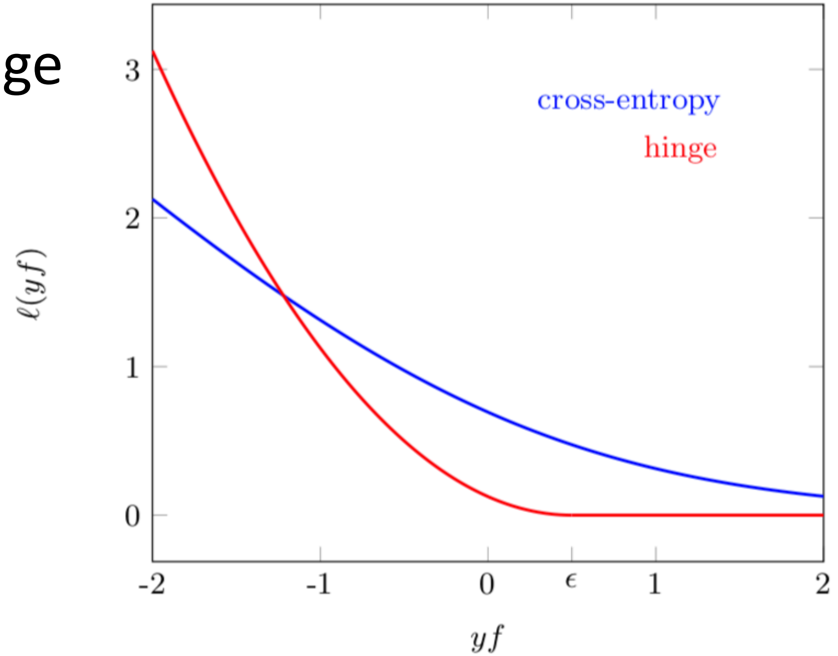
loss function

$$\mathcal{L} = \frac{1}{P}\sum_i l(y_i f_{\mathbf{w}}(\mathbf{x}_i))$$

- Hinge loss

$$\boxed{\Delta_i = \epsilon - f_{\mathbf{w}}(\mathbf{x}_i) y_i}$$

$$\boxed{\mathcal{L} = \frac{1}{P}\sum_{i \in m} \Delta_i^2} \quad i \in m \text{ if } \Delta_i < 0$$



Analogy: particles in contact= misclassified datum

Predictions

- sharp transition at N^* from under-parametrized with $\mathcal{L} > 0$ to over-parametrized ($\mathcal{L} = 0$) as N increases

- Hessian decomposition still holds

$$H = \sum_{i \in m} \nabla \Delta_i \otimes \nabla \Delta_i + \sum_{i \in m} \Delta_i H_{\Delta i} = H_0 + H_p$$

H_p *not negative definite*

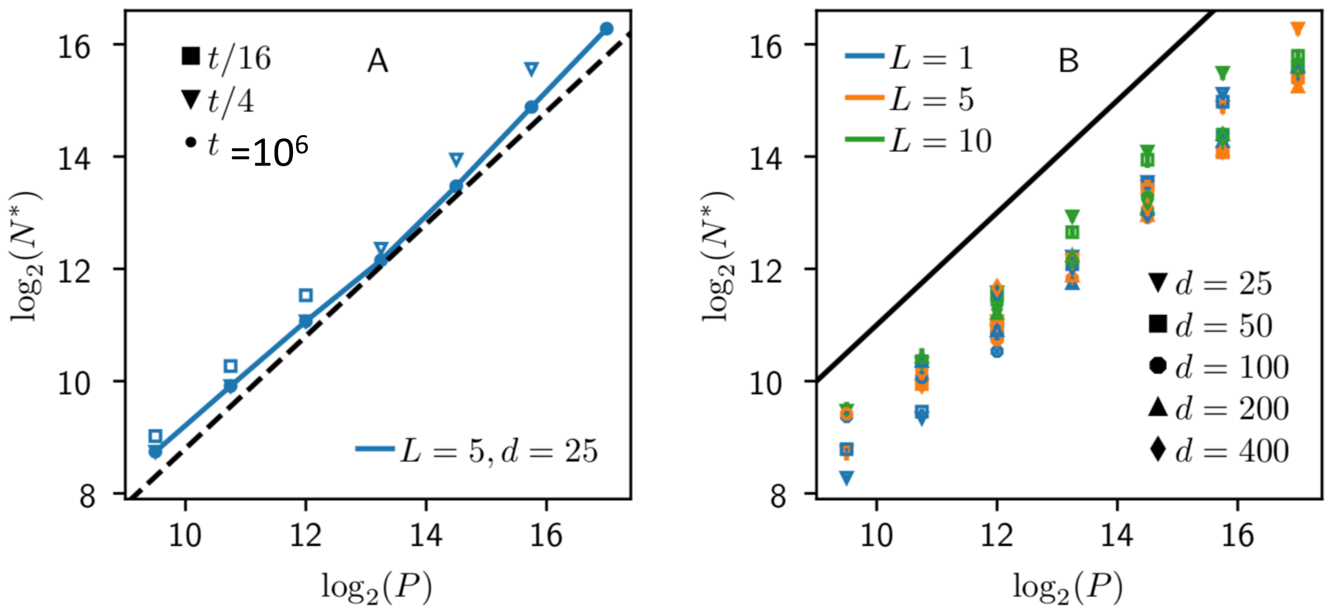
$\Rightarrow$ Hypostatic scenario $\boxed{N_\Delta / N^* < 1}$

- Stability implies $N_\Delta > N_-$ for continuous $f_{\mathbf{w}}(\mathbf{x}_i)$

If $N_- \sim N$ then $N^* < C_0 P$

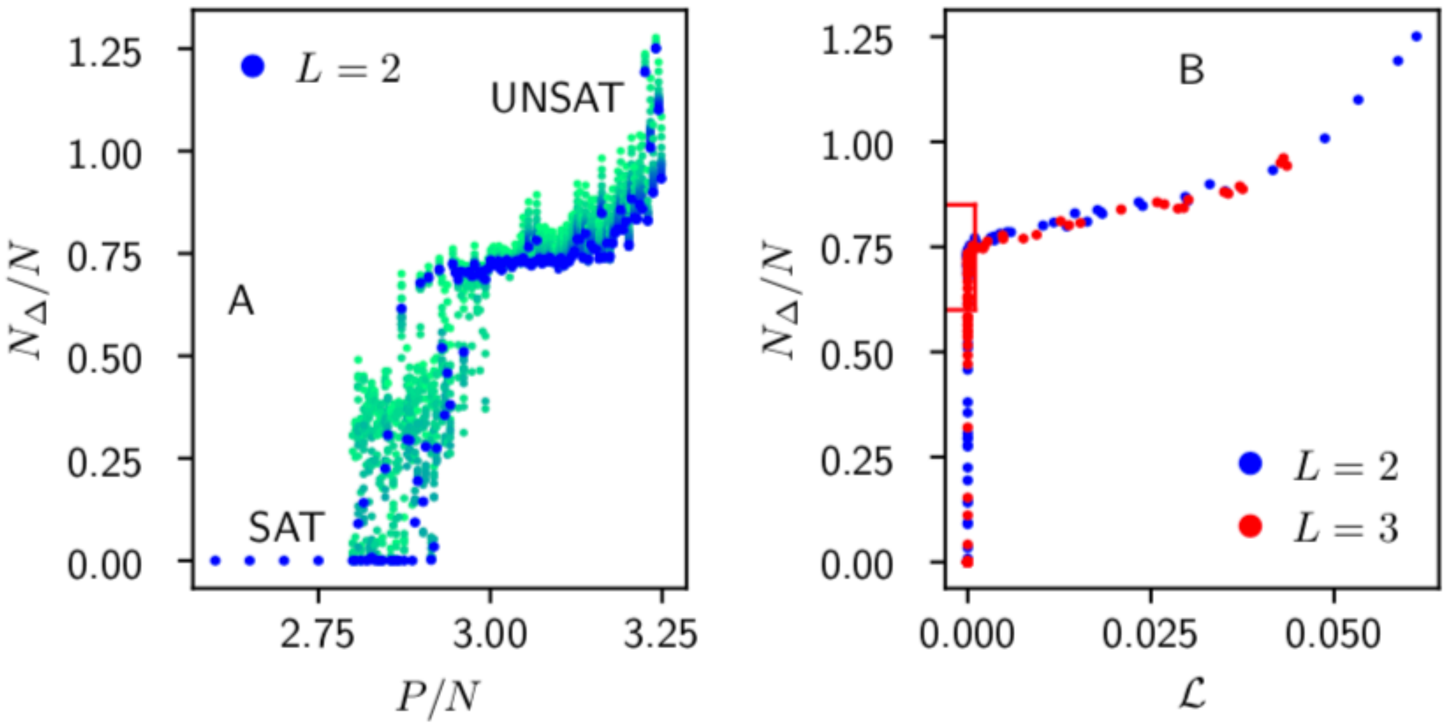
Empirical tests

- Fully connected Network, Depth L, width h, Relu neurons, Gradient descent
- Random data $\mathbf{x}_i \in S^d$ random label $y_i = \pm 1$



- long times: apparent convergence to $\dfrac{N^*}{P} \to r_c$
- No systematic dependence on depth L

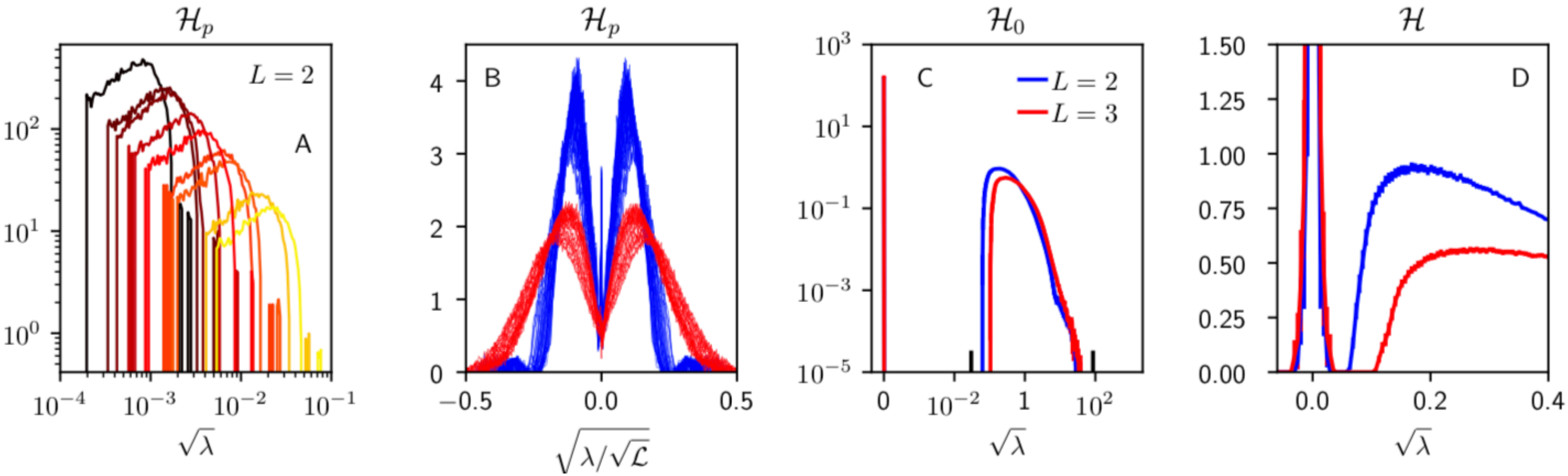
Discontinuous density of constraints N_Δ/N



green t=3*10^5 blue t=10^7

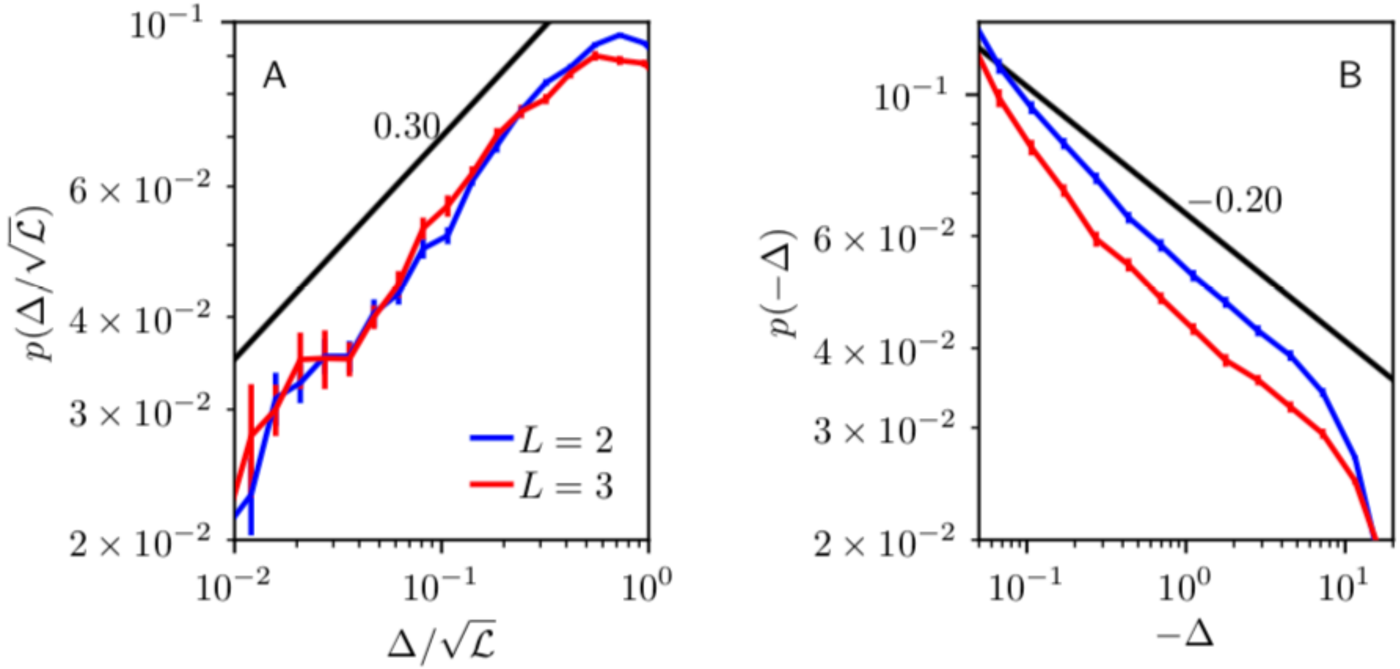
- Discontinuous jump $N_\Delta/N \approx 0.75$: hypostatic

Spectrum of the Hessian



- spectrum of H_p scale as $\Delta \sim \mathcal{L}^{1/2}$
- appears symmetric
- full hessian gapped: flat and stiff valleys
- minima loss presumably not reached

Transition characterized by new exponents



$$P_+(\Delta) \sim \Delta^{\theta} \qquad P_-(\Delta) \sim \Delta^{-\gamma}$$

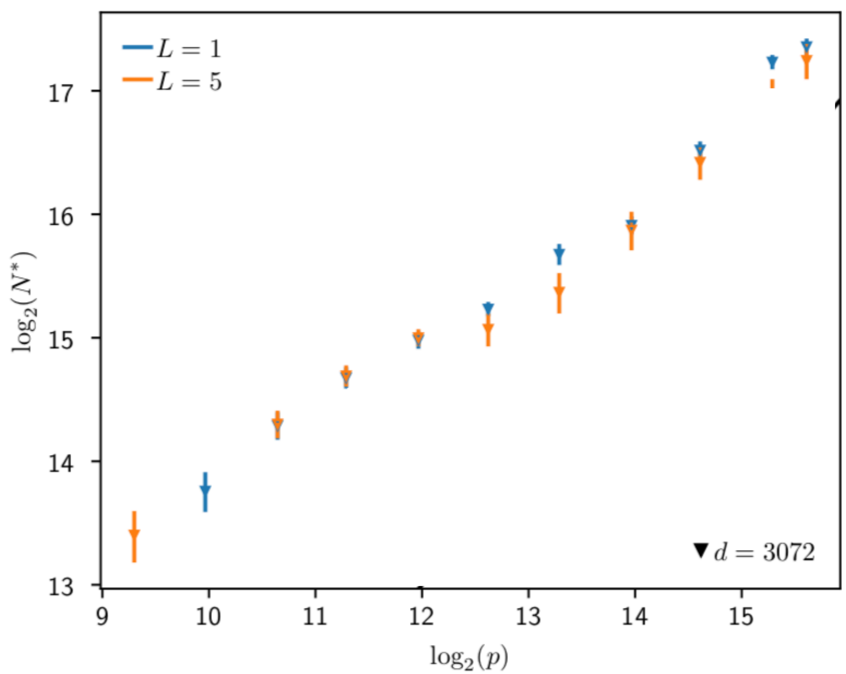
$\boxed{\theta = 0.3}$ $\boxed{\gamma = 0.2}$ appear independent of dimensions

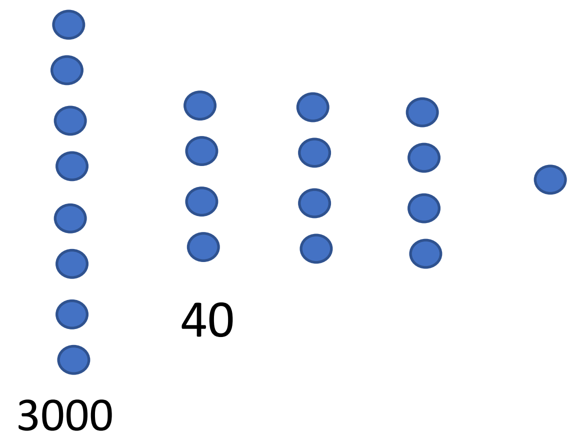
- suggests marginal stability. Learning = crackling?

Real data: Cifar 10

P= $6*10^4$ images
d=32^2*3
10 classes

cross entropy, N* corresponds to 98% train accuracy





transition independent of depth

Ability of fully-connected rectangular nets to fit data depends on N, *not on depth L* both for CIFAR10 and random data

Conclusion

- phase transition over to under-parametrized in deep nets similar to

jamming ellipses. controlled by N, new exponents

- *Why not stuck in poor minima of the loss?*

-> Minima require sufficient constraints to exist, not achievable below transition

- *Role of depth associated to enhanced expressivity?*

-> In 2 simple examples, deep fully connected nets needed same number of parameters than shallow ones to fit data.

- reference point where landscape properties change a lot. Study

-learning (avalanches of change of constraints?)

- generalization