Threshold of Descending Algorithms in Inference Problems

Stefano **Sarao Mannelli**
Institut de Physique Théorique, CEA-Saclay

Gradient Based Algorithms

$$\frac{d\theta(t)}{dt} = -\nabla \mathcal{L}[\theta(t)] + \eta(t)$$

Parameter/ Estimator

Loss function/ Hamiltonian

(thermal) noise

- **Langevin algorithm:** η white Gaussian with variance 2T
- **Gradient Flow GF:** no η term

Gradient Based Algorithms

$$\frac{d\theta(t)}{dt} = -\nabla\mathcal{L}[\theta(t)] + \eta(t)$$

Parameter/ Estimator

Loss function/ Hamiltonian

(thermal) noise

- **Langevin algorithm:** η white Gaussian with variance 2T
- **Gradient Flow GF:** no η term

variations :

- **Momentum:** adding inertial term
- **Stochastic GF:** ∇ acts on batches of the training set

Gradient Based Algorithms

$$\frac{d\theta(t)}{dt} = -\nabla \mathcal{L}[\theta(t)] + \eta(t)$$

Parameter/ Estimator

Loss function/ Hamiltonian

(thermal) noise

- **Langevin algorithm:** η white Gaussian with variance 2T
- **Gradient Flow GF:** no η term

variations :

- **Momentum:** adding inertial term
- **Stochastic GF:** ∇ acts on batches of the training set

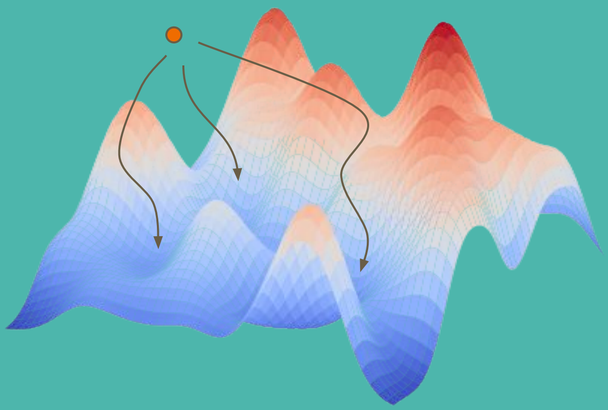
Gradient Based Algorithms

$$\frac{d\theta(t)}{dt} = -\nabla\mathcal{L}[\theta(t)] + \eta(t)$$

Parameter/ Estimator

Loss function/ Hamiltonian

(thermal) noise



- **Q:** in average, does it find the best solution?
- **Q:** what is the role of local minima?

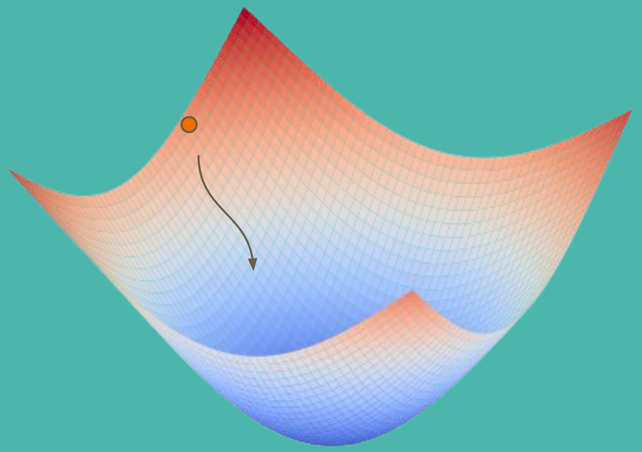
Gradient Based Algorithms

$$\frac{d\theta(t)}{dt} = -\nabla\mathcal{L}[\theta(t)] + \eta(t)$$

Parameter/ Estimator

Loss function/ Hamiltonian

(thermal) noise



- **Q:** in average, does it find the best solution?
- **Q:** what is the role of local minima?

Trivialization is not necessary to find the optimal solution.

In the analysed model, we show that only some local minima are relevant for the algorithmic performance.

We can characterize the algorithmic threshold.

characterize the dynamics

- Linear Neural Networks [Bős, Opper ’97; Saxe, McClelland, Ganguli ‘13]
- One-pass SGD [Saad, Solla ’95 ; Saad ’09; Goldt, Advani, Saxe, Krzakala, Zdeborová '19; Goldt, Mézard, Krzakala, Zdeborová ‘19]
- SGD in 2-layer networks with diverging hidden layer size [Rotskoff, Vanden-Eijnden ’18; Mei, Montanari, Nguyen ’18; Chizat, Bach '18]
- Dynamical Mean Field Theory [Mézard, Parisi, Virasoro ’87; Sompolinsky, Crisanti, Sommers ‘88; Georges, Kotliar, Krauth, Rozenberg ’96; Agoritsas, Biroli, Urbani, Zamponi ‘18; Mignacco, Krzakala, Urbani, Zdeborová ‘20; Krishnamurthy, Can, Schwab ‘20]

characterize the dynamics

- Dynamical Mean Field Theory [Mézard, Parisi, Virasoro '87; Sompolinsky, Crisanti, Sommers '88; Georges, Kotliar, Krauth, Rozenberg '96; Agoritsas, Biroli, Urbani, Zamponi '18; Mignacco, Krzakala, Urbani, Zdeborová '20; Krishnamurthy, Can, Schwab '20]
  - ➢ disordered systems, recurrent neural networks, inference and optimization problems
  - ➢ GD, SGD, Langevin dynamics
  - ➢ it maps the dynamical equation into an effective dynamical equation with coloured noise (whose stochastic process depends on the dynamics itself!)

Spiked Matrix-Tensor Model

$$\mathcal{L}(x) = ||xx^T - Y||_2^2$$
$$+||x^{\otimes p} - T||_2^2$$

With:

$$x, x^* \in \mathbb{S}^{N-1}, \; \xi \sim \mathcal{N}$$

$$Y_{ij} = x_i^* x_j^* + \sqrt{\Delta_2}\, \xi_{ij}$$

$$T_{i_1 \ldots i_p} = x_{i_1}^* \ldots x_{i_p}^* + \sqrt{\Delta_p}\, \xi_{i_1 \ldots i_p}$$

Spiked Matrix-Tensor Model

$$\mathcal{L}(x) = ||xx^T - Y||_2^2$$
$$+||x^{\otimes p} - T||_2^2$$

- Closed expression for DMFT
- Coexistence of many phases for $\Delta_2, \Delta_p = O(1)$
- In general, many techniques can be applied

Spiked Matrix-Tensor Model

$$\mathcal{L}(x) = ||xx^T - Y||_2^2 + ||x^{\otimes p} - T||_2^2$$

$$C(t,t') = \lim_{N\to\infty} x(t)\cdot x(t')$$

$$R(t,t') = \lim_{N\to\infty} \sum_{i=1}^{N} \frac{\delta x_i(t)}{\delta \eta_i(t')}$$

$$m(t) = \lim_{N\to\infty} x(t)\cdot x^*$$

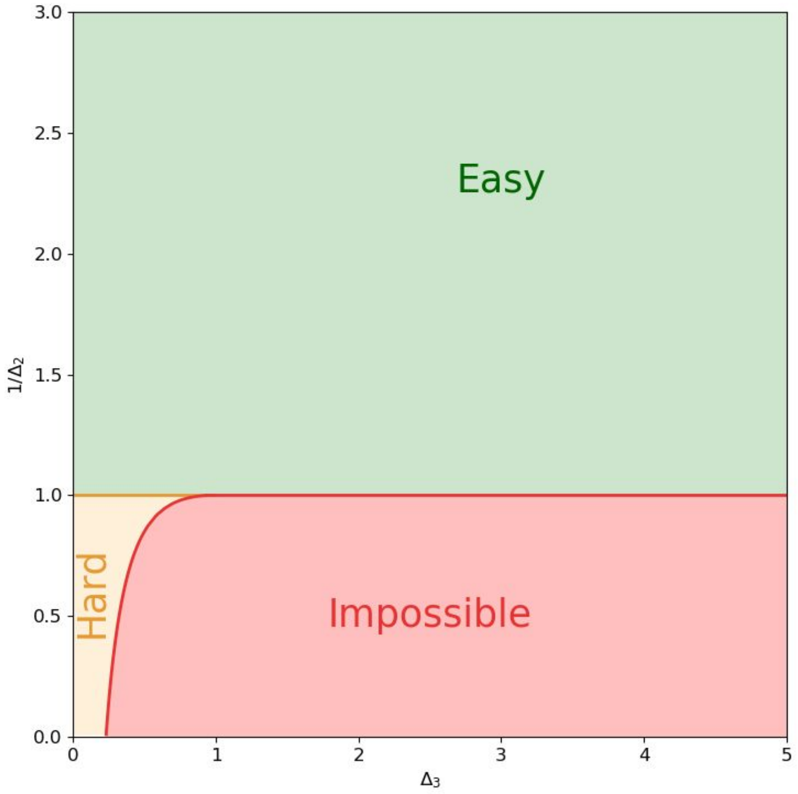
Call $Q(x) = x^2/2\Delta_2 + x^p/p\Delta_p$:

$$\frac{\partial}{\partial t}C(t,t') = 2T\,R(t',t) - \mu(t)C(t,t') + Q'(m(t))m(t') + \int_0^t dt''\,R(t,t'')Q''(C(t,t''))C(t',t'') + \int_0^{t'} dt''\,R(t',t'')Q'(C(t,t'')),$$

$$\frac{\partial}{\partial t}R(t,t') = \delta(t-t') - \mu(t)R(t,t') + \int_{t'}^{t} dt''\,R(t,t'')Q''(C(t,t''))R(t'',t'),$$

$$\frac{\partial}{\partial t}m(t) = -\mu(t)m(t) + Q'(m(t)) + \int_0^t dt''\,R(t,t'')m(t'')Q''(C(t,t'')),$$

$$\mu(t) = T + Q'(m(t))m(t) + \int_0^t dt''\,R(t,t'')\left[Q''(C(t,t''))C(t,t'') + Q'(C(t,t''))\right].$$



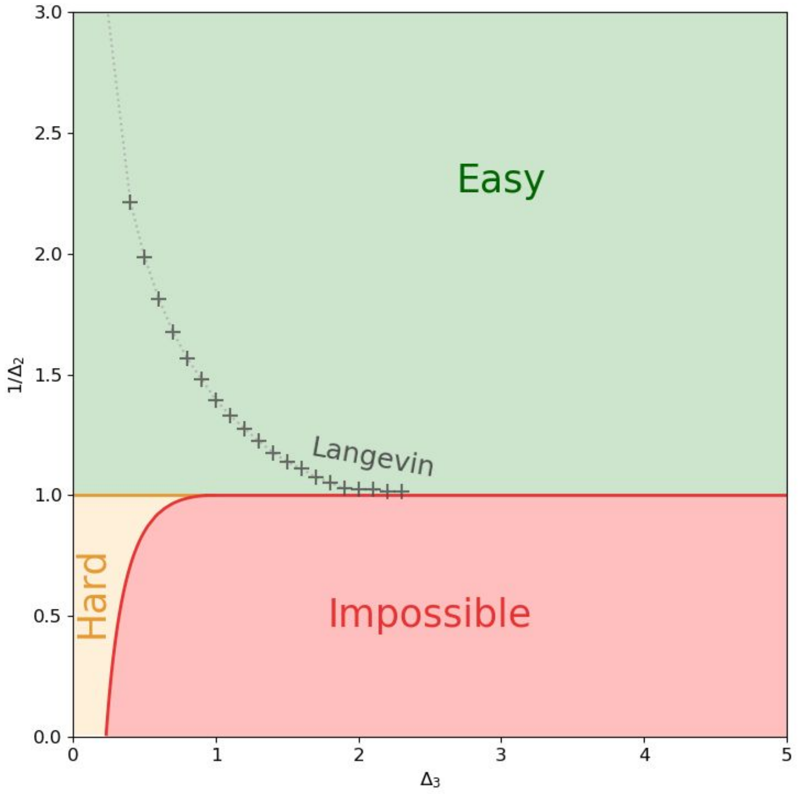
Phase diagram AMP

The 3 phases of the Approximate Message Passing AMP phase diagram:

Easy : AMP from random initialization finds the optimal solution

Hard : the optimal solution is better than random guessing but AMP cannot find it if initialized at random

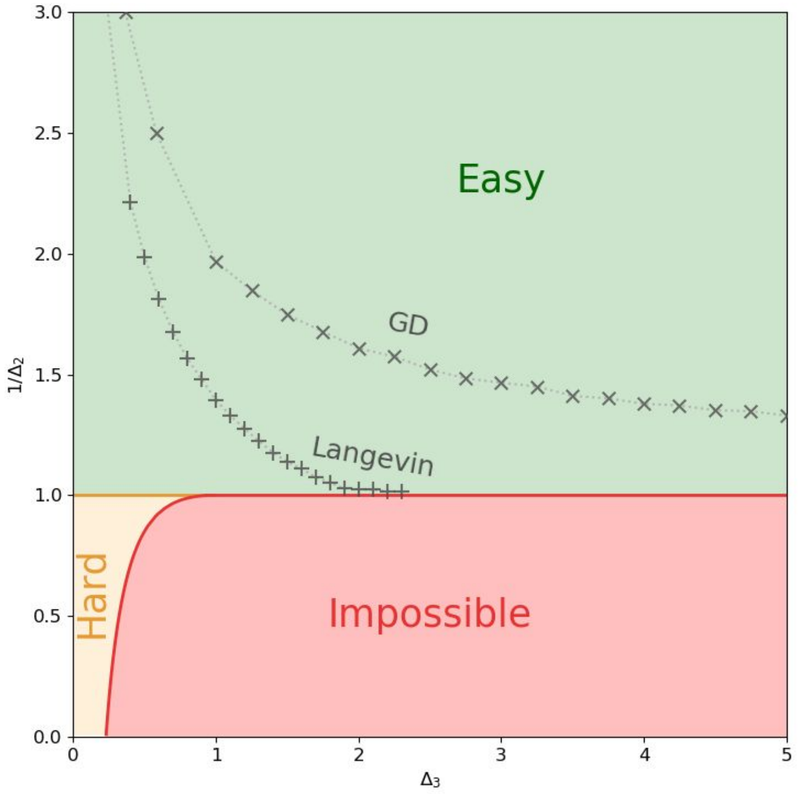
Impossible : the problem is information theoretically impossible.



Phase diagram Langevin algorithm

Extrapolate numerically the threshold from DMFT equations.

Langevin algorithm with `T=1` in the long time limit samples the posterior distribution. Bayes optimal.



Extrapolate numerically the threshold from DMFT equations.

Gradient flow has a worse algorithmic threshold then Langevin. As expected.

Phase diagram gradient flow

What does the landscape of this model look like ?

Kac-Rice to characterize the distribution of minima [Ben Arous, Mei, Montanari, Nica '17; Ros, Ben Arous, Biroli, Cammarota '18; SM, Krzakala, Urbani, Zdeborová '19]

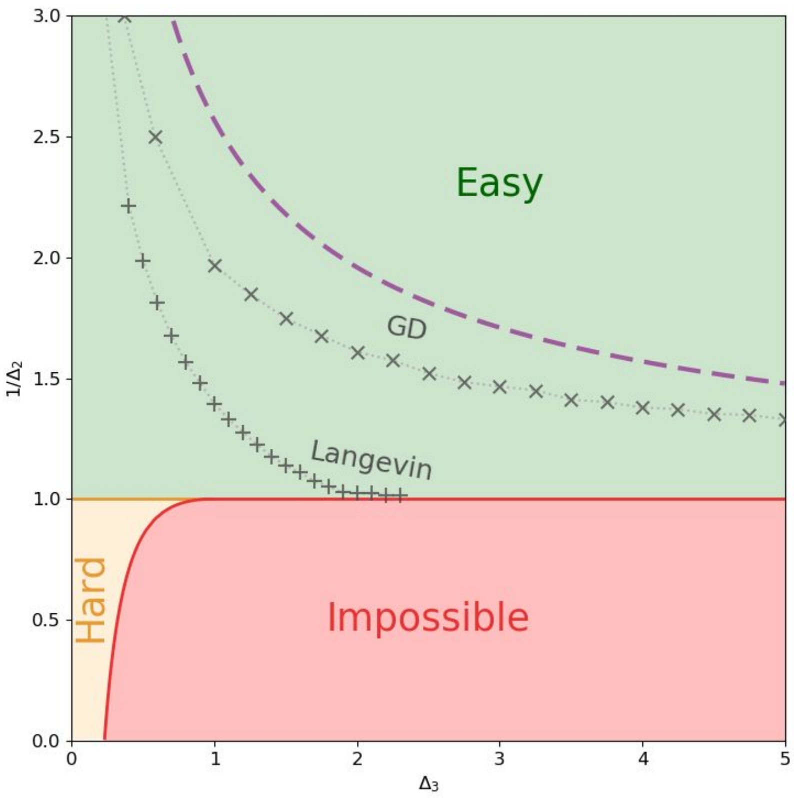
Complexity: $\Sigma = \log[\text{avg \# minima}]/N$

Trivialization transition

What does the landscape of this model look like ?

Kac-Rice to characterize the distribution of minima [Ben Arous, Mei, Montanari, Nica ‘17; Ros, Ben Arous, Biroli, Cammarota ‘18; SM, Krzakala, Urbani, Zdeborová ‘19]

Complexity: `Σ=log[avg # minima]/N`

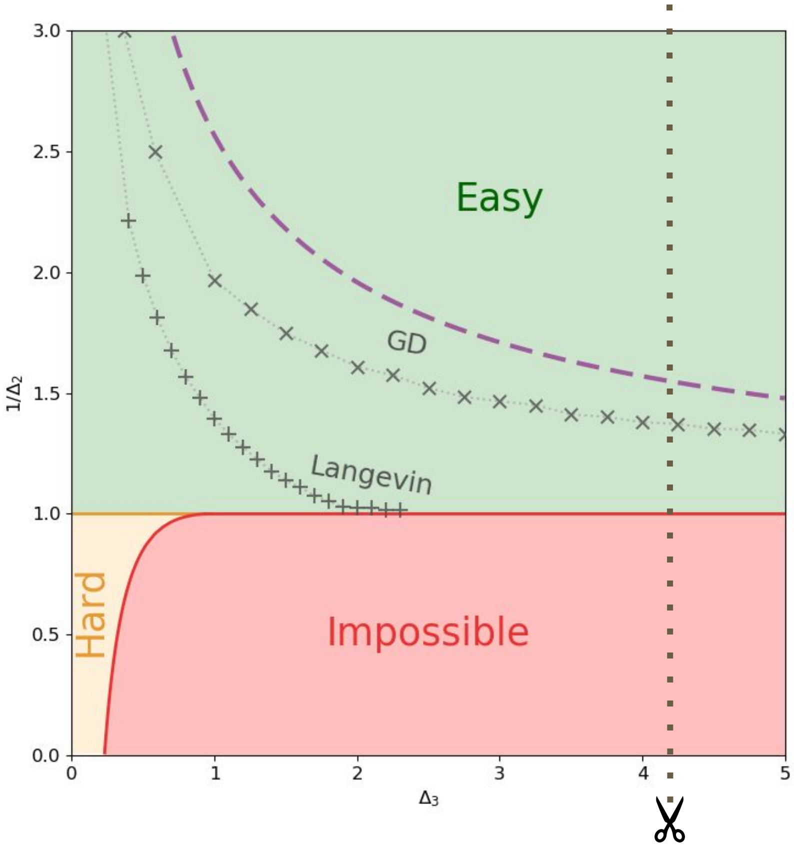


Trivialization transition

What does the landscape of this model look like ?

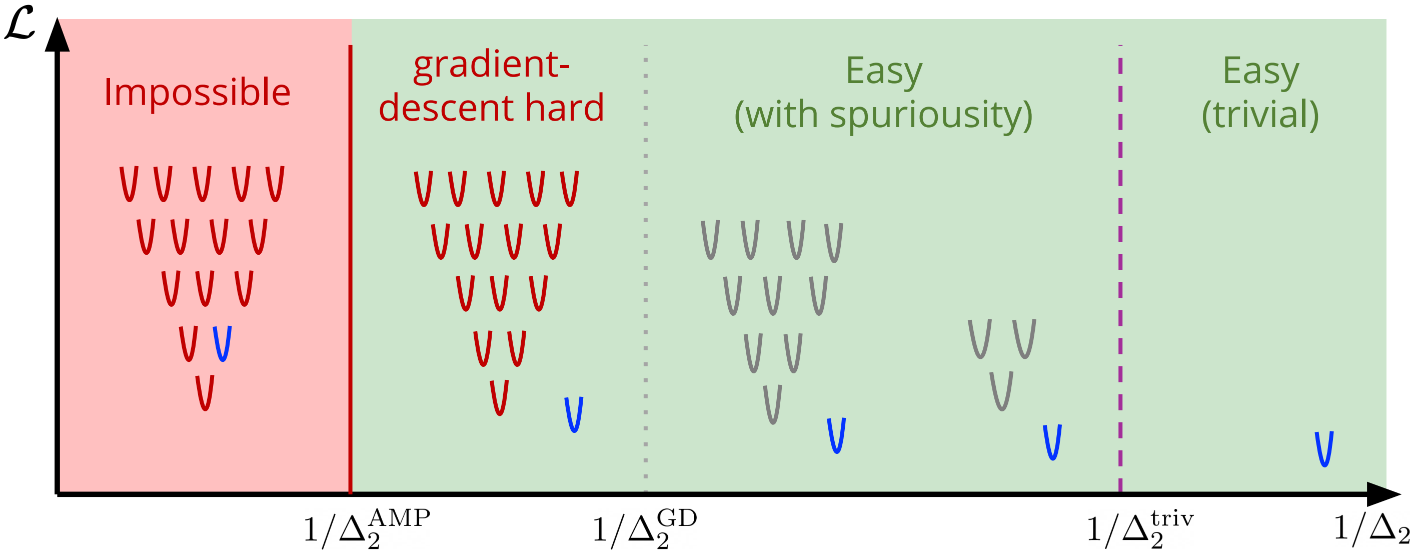
Kac-Rice to characterize the distribution of minima [Ben Arous, Mei, Montanari, Nica ‘17; Ros, Ben Arous, Biroli, Cammarota ‘18; SM, Krzakala, Urbani, Zdeborová ‘19]

Complexity: `Σ=log[avg # minima]/N`

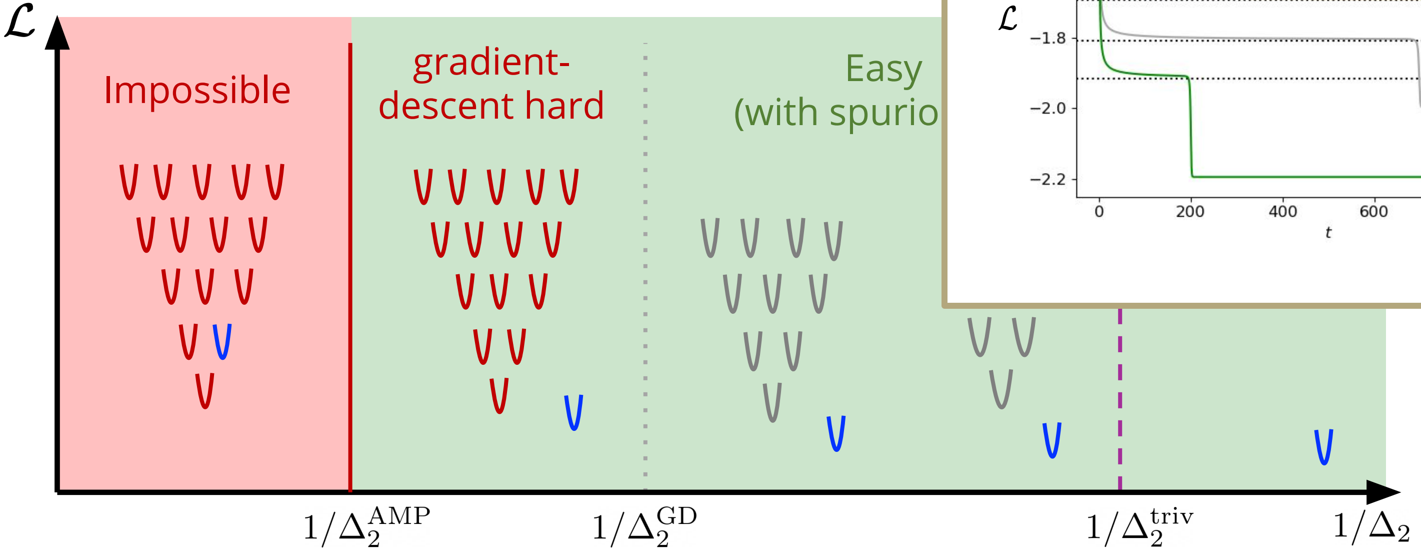


Trivialization transition

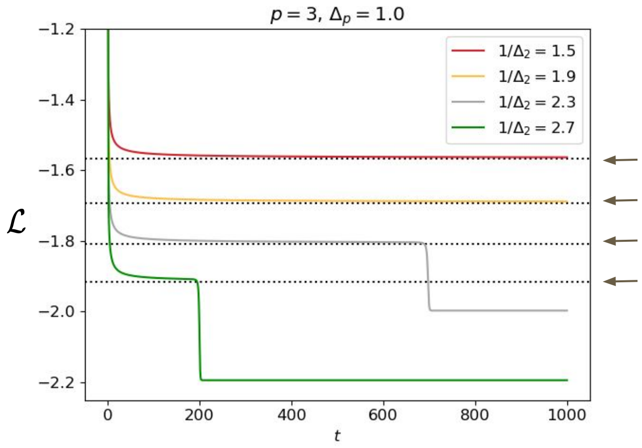
When does GF converge ?



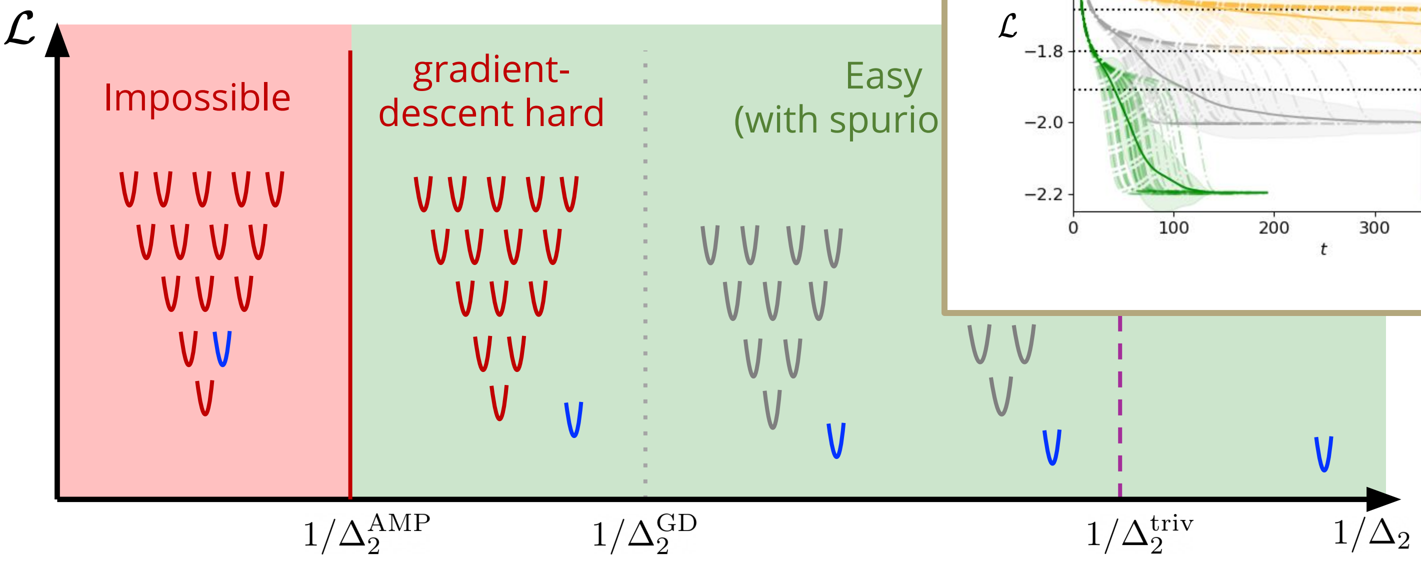
When does GF conv



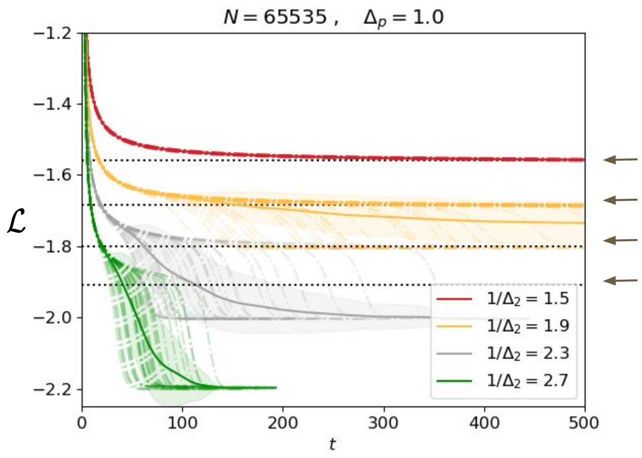
[Cugliandolo, Kurchan '93]



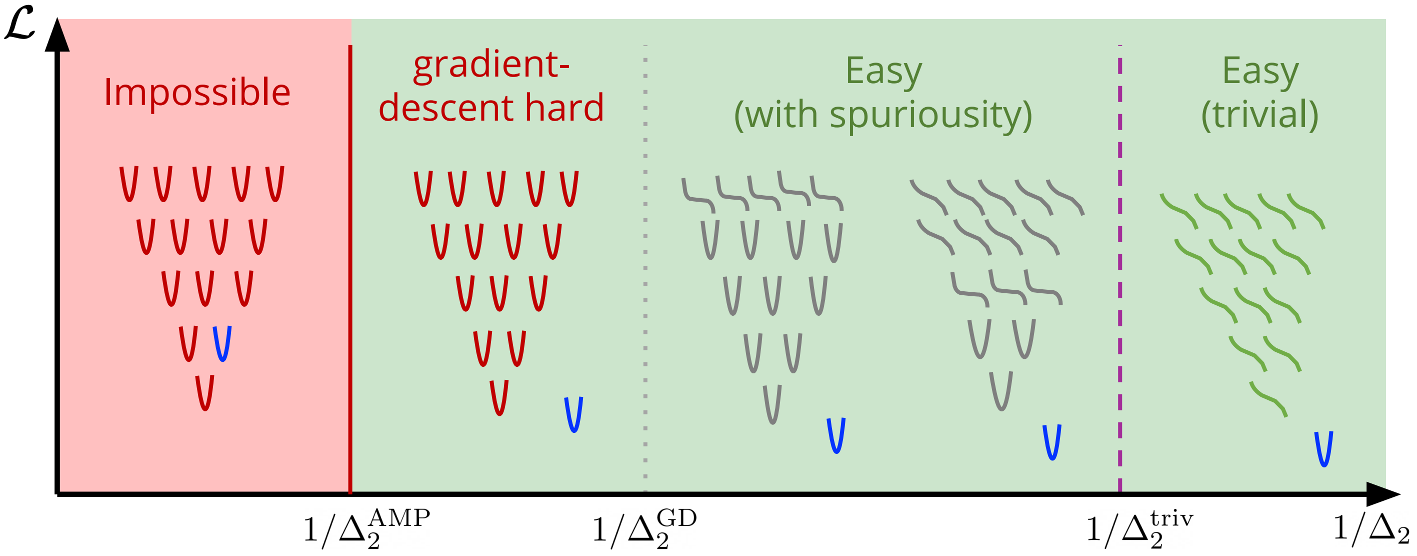
When does GF conv

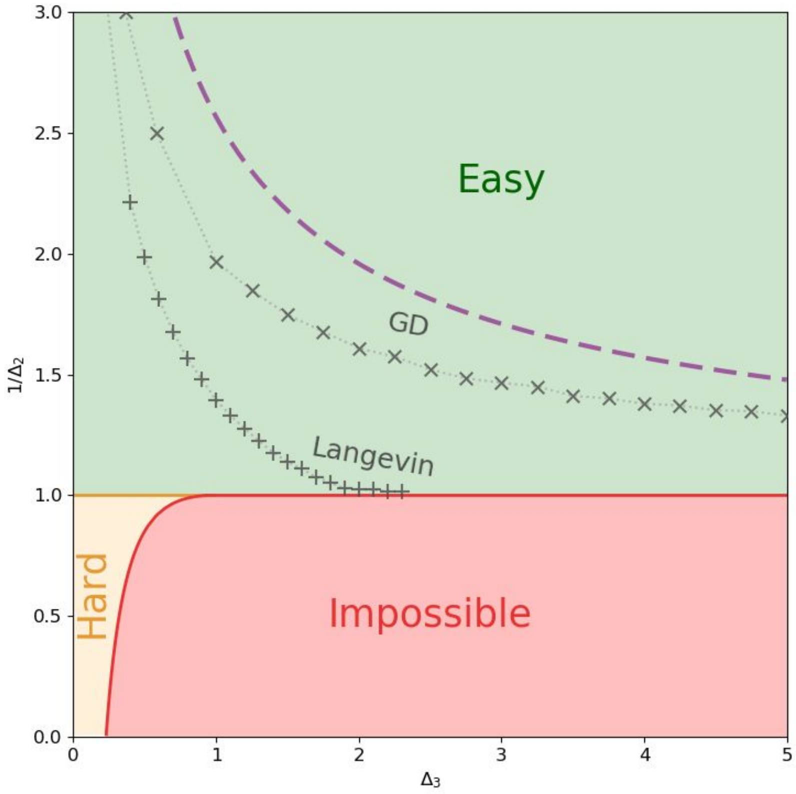


[Cugliandolo, Kurchan '93]



When does GF converge ?



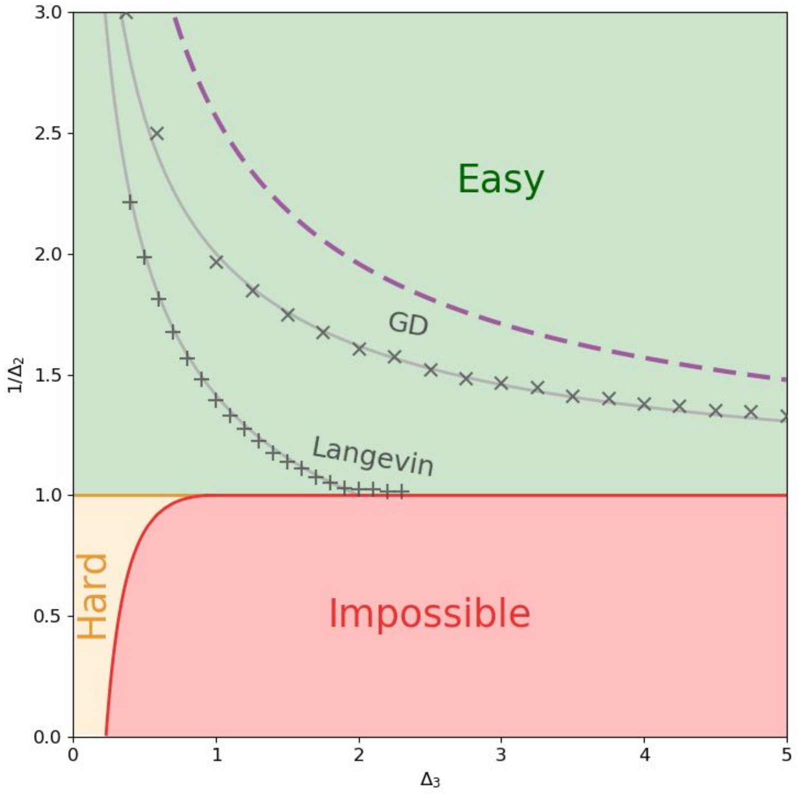


Phase diagram (so far)

Stability of threshold states

- Threshold states : $\frac{T^2}{(1-q)^2} = (p-1)\frac{q^{p-2}}{\Delta_p} + \frac{1}{\Delta_2}$
- Stability : $T\Delta_2 = 1 - q$

$$\Delta_p = \frac{\Delta_2^2(p-1)(1-T\Delta_2)^{p-2}}{1-\Delta_2}$$



Phase diagram (final)

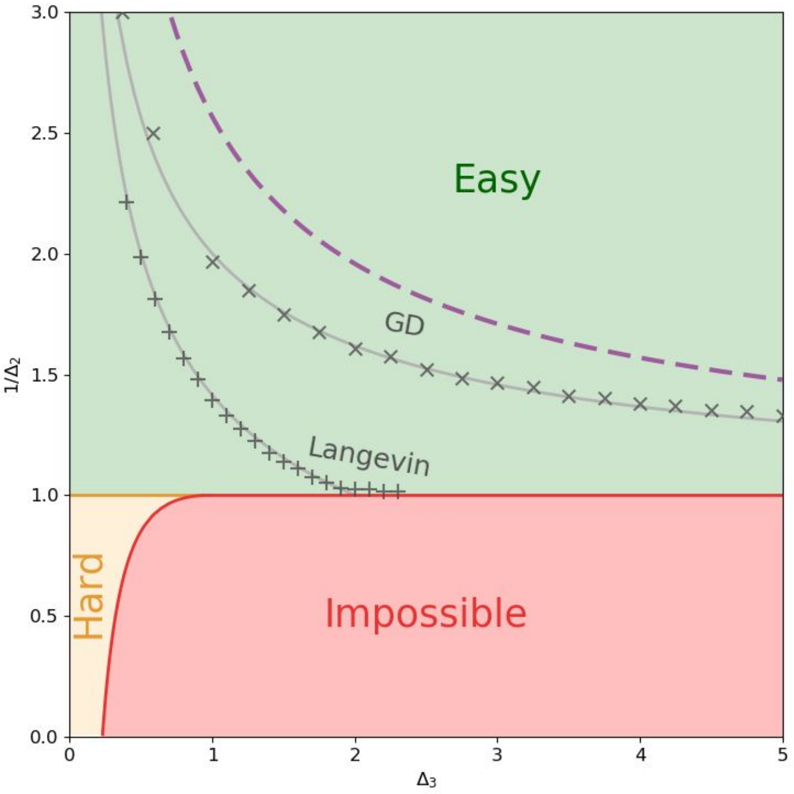
Stability of threshold states

- Threshold states :

$$\frac{T^2}{(1-q)^2} = (p-1)\frac{q^{p-2}}{\Delta_p} + \frac{1}{\Delta_2}$$

- Stability : $T\Delta_2 = 1 - q$

$$\Delta_p = \frac{\Delta_2^2(p-1)(1-T\Delta_2)^{p-2}}{1-\Delta_2}$$



Phase diagram (final)

Conclusions

- GD can escape positive complexity regions,
- role of the stability of the threshold states.

New results

- GD in phase retrieval [2006.06997]:
  from α=#samples/dimension
  critical O(log N) to O(1)

Thank you.

Refs. for this talk

- Marvels and pitfalls of the Langevin algorithm in noisy high-dimensional inference. **SSM**, Biroli, Cammarota, Krzakala, Urbani, Zdeborova. PRX 10, 011057;
- Thresholds of descending algorithms in inference problems. **SSM**, Zdeborova. J.Stat.Mech., 2020(3):034004;
- Who is afraid of big bad minima? analysis of gradient-flow in spiked matrix-tensor models. **SSM**, Biroli, Cammarota, Krzakala, Urbani, Zdeborova. NeurIPS'19;
- Passed&Spurious: Descent algorithms and local minima in spiked matrix-tensor models. **SSM**, Krzakala, Urbani, Zdeborova. ICML'19.