Statistical physics and statistical inference

Marc Mézard

Ecole normale supérieure
PSL University

Salam lecture 3
ICTP - Trieste
January 29, 2020

What is inference?

Statistics

Infer a hidden rule, or hidden variables, from data.

Restricted sense : find parameters of a probability distribution

Urn with 10.000 balls. Draw 100, find 70 white balls and 30 black

Best guess for the composition of the urn? How reliable? Probability that it has 6000 white- 4000 black?

If only black and white balls , with fraction x of white,

probability to pick-up 70 white balls is $\binom{100}{70} x^{70}(1-x)^{30}$

Log likelihood of x : $L(x) = 70 \log x + 30 \log(1-x)$

Maximum at $x^* = .7$ Probability of .6 : $e^{L(.6)-L(.7)}$

Bayesian inference

Unknown parameters	x	Prior	$P(x)$
Measurements	y	Likelihood	$P(y\|x)$

Posterior $$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Bayesian inference

Unknown parameters	x	Prior	$P(x)$
Measurements	y	Likelihood	$P(y\|x)$

Posterior $$P(\boxed{x}|y) = \frac{P(y|\boxed{x})P(\boxed{x})}{P(y)}$$

Bayesian inference

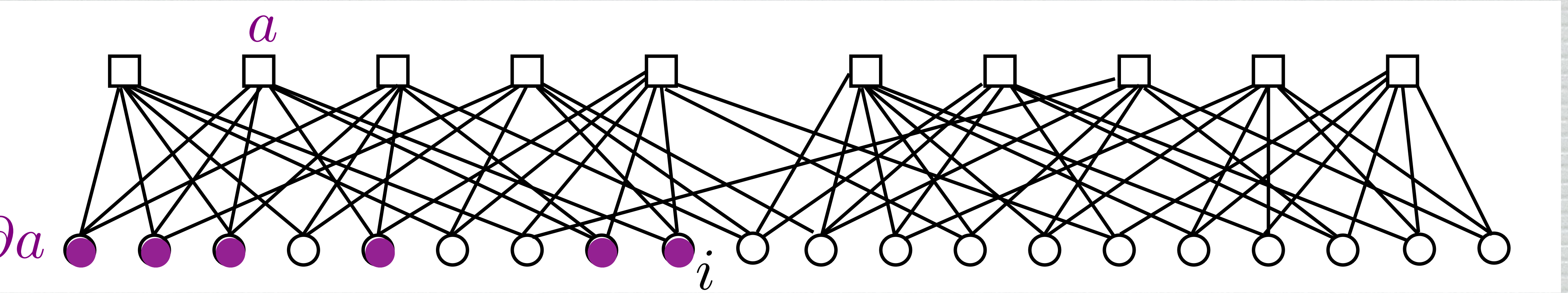
Unknown parameters	x	Prior	$P(x)$
Measurements	y	Likelihood	$P(y\|x)$

Posterior $$P(\boxed{x}|y) = \frac{P(y|\boxed{x})P(\boxed{x})}{P(y)}$$

E.g. error correcting codes

x	reconstructed message
y	received message
$P(y\|x)$	transmission channel
$P(x)$	codebook

Decoding as an inference problem



$$P(x_1,\ldots,x_N|\underbrace{\boxed{y_1,\ldots,y_N}}_{\text{received}}) = \frac{1}{Z}\boxed{\prod_i \psi_i(x_i|y_i)}\boxed{\prod_a \mathbb{I}\left(\sum_{i\in\partial a} x_i = 0\ (mod\ 2)\right)}$$

A priori knowledge of the channel

$P(y|x)$

Parity check constraints

$P(x)$

Statistical inference: general scheme

Challenge = rules with **many hidden parameters.** eg : machine learning with large machine and big data, decoding in communication,…

$x = (x_1, \dots, x_N) \qquad N \gg 1$

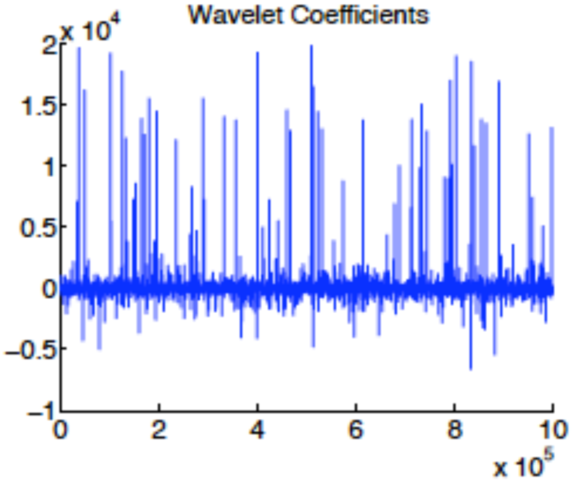
Many measurements $\qquad y = (y_1, \dots, y_M) \qquad M \gg 1$

Measure of the amount of data $\qquad \alpha = M/N$

→ **Algorithms**

→ **Prediction on the quality of inference**, on the performance of the algorithms, on the type of situations where they can be applied

First example : Compressed sensing



From 65.536 wavelet coefficients, keep 25.000

(From Candes-Wakin)

How to acquire the image directly in the compressed form?

Applications in MRI, tomography, etc.

The simplest compressed-sensing problem: reconstruct a signal from linear measurements

Consider a system of linear measurements

$$\boxed{y = Fx}$$

Measurements

$$y = \begin{pmatrix} y^1 \\ . \\ . \\ y^M \end{pmatrix}$$

(e.g. wavelet components)

Signal $x = \begin{pmatrix} x^1 \\ . \\ . \\ . \\ . \\ x^N \end{pmatrix}$

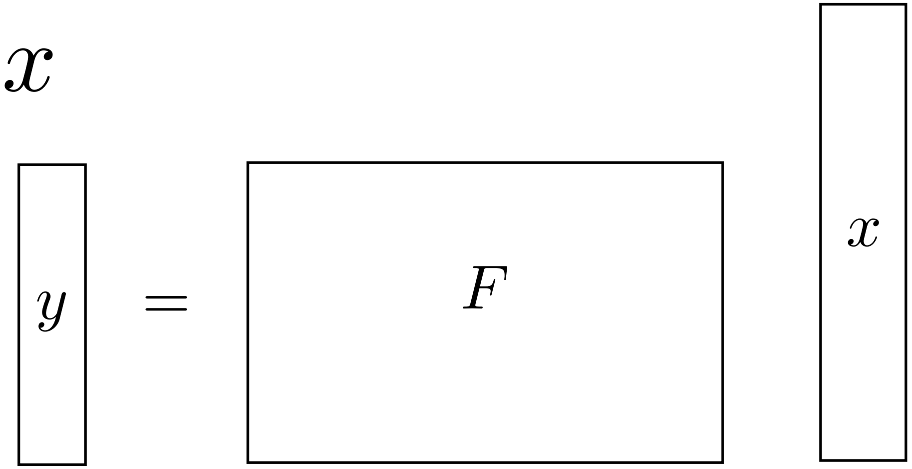
$F = M \times N$ matrix

Pb: Find x when $M < N$ and x is sparse

The problem: $y = Fs$ and s is sparse, i.e. it has R components $\neq 0$

$R < M < N$ y is observed, F is known. Find s

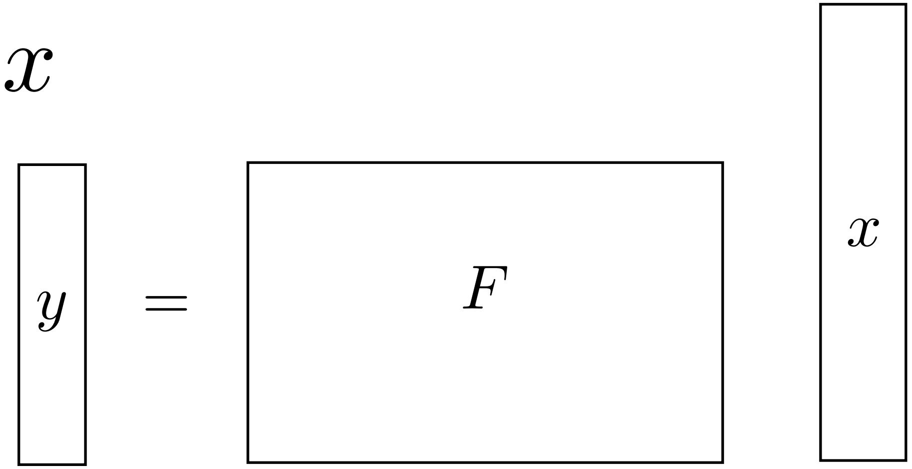
Study the linear system $y = Fx$



The problem: $y = Fs$ and s is sparse, i.e. it has R components $\neq 0$

$R < M < N$ y is observed, F is known. Find s

Study the linear system $y = Fx$



Exploit the sparsity of the original s

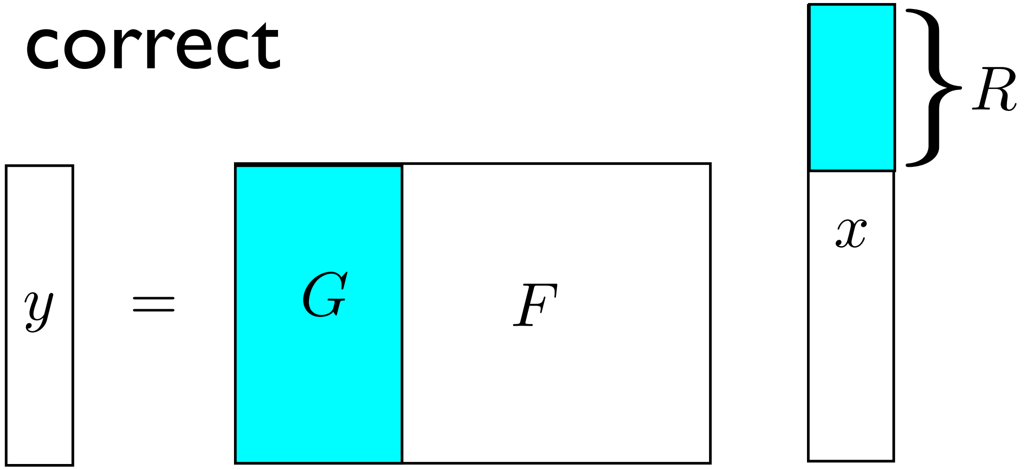
The problem: $y = Fs$ and s is sparse

R components $\neq 0$

➡ Study the linear system $y = Fx$

A ‘simple’ solution: guess the positions where $x_i \neq 0$ and check if it is correct

e.g. $x_1, \ldots, x_R \neq 0$



$G = \{\ R$ first columns of $F\ \}$

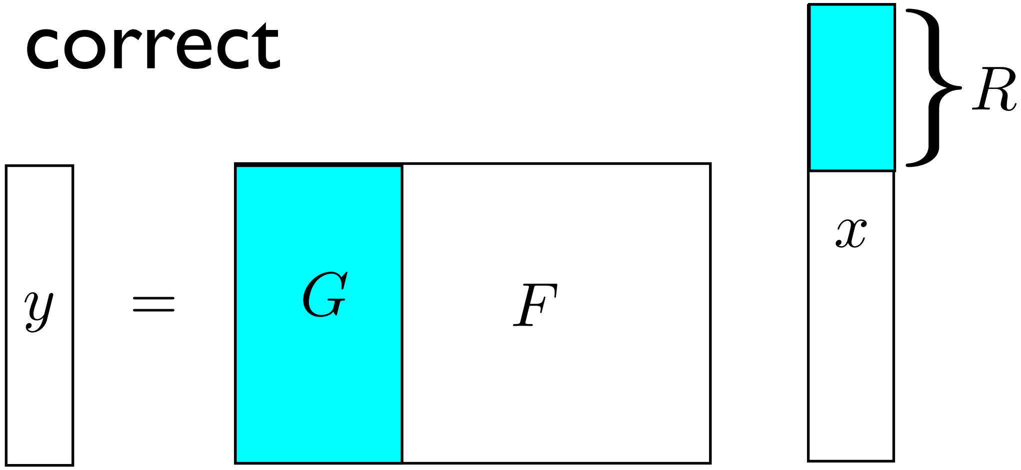
Solve : $y^\mu = \sum_{i=1}^{R} G^{\mu i} x_i \qquad \mu = 1, \ldots, M$

The problem: $y = Fs$ and s is sparse

R components $\neq 0$

➡ Study the linear system $y = Fx$

A 'simple' solution: guess the positions where $x_i \neq 0$ and check if it is correct



e.g. $x_1, \ldots, x_R \neq 0$

$G = \{\, R$ first columns of $F \,\}$

Solve : $y^\mu = \sum_{i=1}^{R} G^{\mu i} x_i \qquad \mu = 1, \ldots, M$

$R < M$ ➡ too many equations

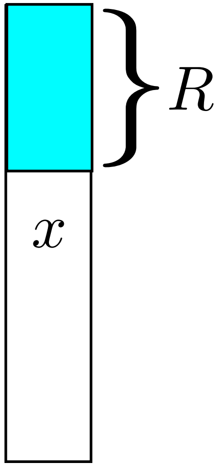
➡ generically inconsistent (no solution), except if the guess of locations of $x_i \neq 0$ was correct

The problem: $y = Fs$ and s is sparse

R components $\neq 0$

→ Study the linear system $y = Fx$

A 'simple' solution: guess the positions where $x_i \neq 0$ and check if it is correct



e.g.

$G = \{$

Solv

$R < M$

$$\binom{N}{R}$$ possible guesses

→ generically inconsistent (no solution), except if the guess of locations of $x_i \neq 0$ was correct

Phase diagram

«Thermodynamic limit»

$N \gg 1$ variables

$R = \rho N$ non-zero variables

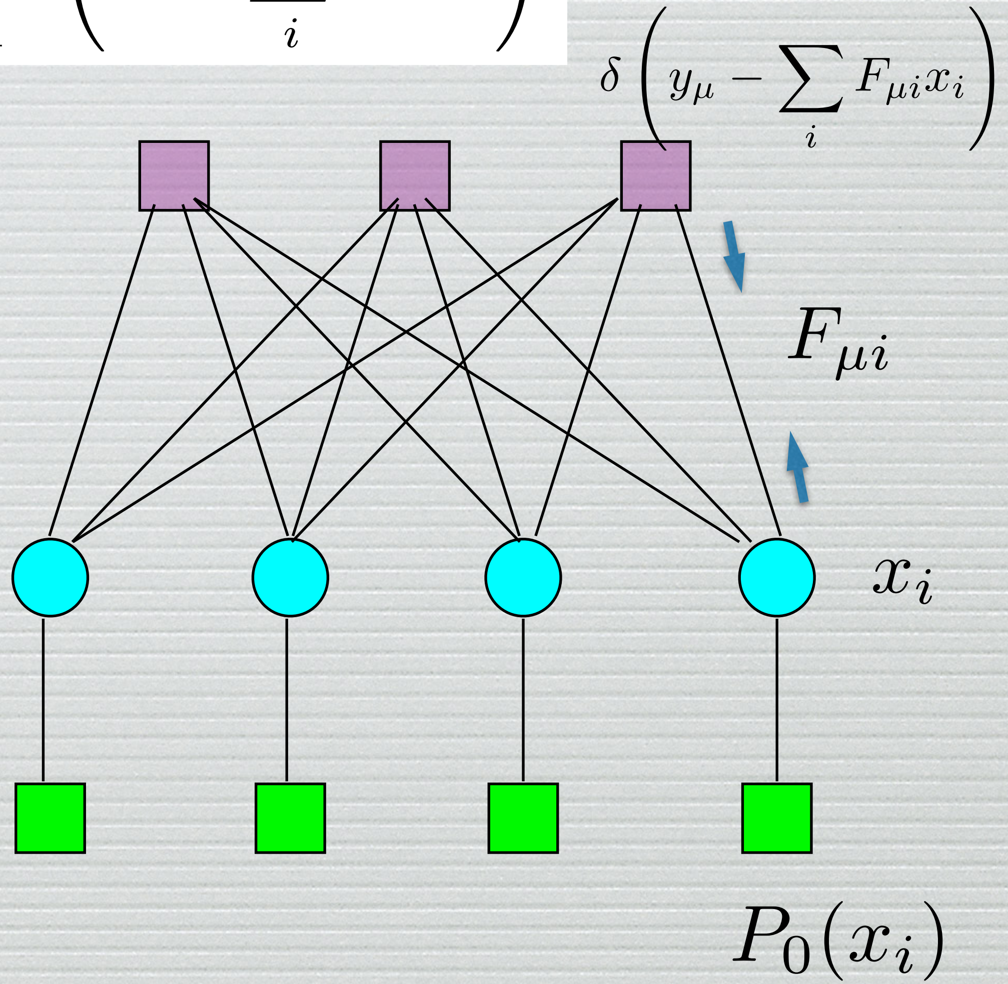
$M = \alpha N$ equations

- Solvable by enumeration when $\alpha > \rho$ but $O(e^N)$
- ℓ_1 norm approach

  Find a N- component vector x such that the M equations $y = Fx$ are satisfied and $||x||_1$ is minimal
- AMP = Bayesian approach + cavity mean-field equations

$$P(\mathbf{x}) = \prod_{i=1}^{N} [(1-\rho)\delta(x_i) + \rho\phi(x_i)] \quad \prod_{\mu=1}^{P} \delta\left(y_\mu - \sum_i F_{\mu i} x_i\right)$$

$$P(\mathbf{x}) = \prod_{i=1}^{N} [(1-\rho)\delta(x_i) + \rho\phi(x_i)] \quad \prod_{\mu=1}^{P} \delta\left(y_\mu - \sum_i F_{\mu i} x_i\right)$$



$F_{\mu i}$: iid, known

Spin glass with multispin interactions, infinite range: write mean field equations.

Messages: $m_{i\to\mu}(x_i)$
$m_{\mu\to i}(x_i)$

Becomes Gaussian in the thermodynamic limit

Mézard 1989, Opper Winther 96, Kabashima 2003, 2008 , Donoho Maleki Montanari 2009, Rangan+ 2011, Krzakala+ 2012, …

BP equations

$$a_{i\to\mu} = \int \mathrm{d}x_i \, x_i \, m_{i\to\mu}(x_i)$$

$$v_{i\to\mu} = \int \mathrm{d}x_i \, x_i^2 \, m_{i\to\mu}(x_i) - a_{i\to\mu}^2$$

$$m_{\mu\to i}(x_i) = \frac{1}{\tilde{Z}^{\mu\to i}} e^{-\frac{x_i^2}{2} A_{\mu\to i} + B_{\mu\to i} x_i}$$

$$m_{i\to\mu}(x_i) = \frac{1}{\tilde{Z}^{i\to\mu}} \left[(1-\rho)\delta(x_i) + \rho\phi(x_i)\right] e^{-\frac{x_i^2}{2} \sum_{\gamma\neq\mu} A_{\gamma\to i} + x_i \sum_{\gamma\neq\mu} B_{\gamma\to i}}$$

BP equations

$$\boxed{a_{i\to\mu}} = \int \mathrm{d}x_i \, x_i \, m_{i\to\mu}(x_i)$$

$$\boxed{v_{i\to\mu}} = \int \mathrm{d}x_i \, x_i^2 \, m_{i\to\mu}(x_i) - a_{i\to\mu}^2$$

$$m_{\mu\to i}(x_i) = \frac{1}{\tilde{Z}^{\mu\to i}} e^{-\frac{x_i^2}{2}\boxed{A_{\mu\to i}}+\boxed{B_{\mu\to i}}x_i}$$

$$m_{i\to\mu}(x_i) = \frac{1}{\tilde{Z}^{i\to\mu}} \left[(1-\rho)\delta(x_i) + \rho\phi(x_i)\right] e^{-\frac{x_i^2}{2}\sum_{\gamma\neq\mu} A_{\gamma\to i} + x_i \sum_{\gamma\neq\mu} B_{\gamma\to i}}$$

Four «messages» sent along each edge $i - \mu$

($4NM$ numbers) can be simplified to $O(N)$parameters

From « cavity messages »

TAP equations

$$a_{i\to\mu} = \int \mathrm{d}x_i \, x_i \, m_{i\to\mu}(x_i)$$

$$v_{i\to\mu} = \int \mathrm{d}x_i \, x_i^2 \, m_{i\to\mu}(x_i) - a_{i\to\mu}^2$$

To full local distribution

$$a_i = \int dx_i \; m_i(x_i) x_i = \langle x_i \rangle$$

$$v_i = \int dx_i \; m_i(x_i) x_i^2 - a_i^2 = \langle x_i^2 \rangle - a_i^2$$

TAP = coupled equations between the 2N variables $a_i, \; v_i$

Iteration ⟶ algorithm : GAMP

Statistical study ⟶ phase diagram and control of the algorithm

Analytic study

$$P(\mathbf{x}) = \prod_{i=1}^{N} [(1-\rho)\delta(x_i) + \rho\phi(x_i)] \quad \prod_{\mu=1}^{P} \delta\left(y_\mu - \sum_i F_{\mu i} x_i\right)$$

Replica method allows to compute the «free entropy»

$$\boxed{\Phi(D) = \lim_{N\to\infty} \frac{1}{N} \log P(D)}$$

where $P(D)$ is the probability that reconstructed x is at distance D from original signal s.

$$\boxed{D = \frac{1}{N} \sum_i (x_i - s_i)^2}$$

Cavity method shows that the order parameters of the BP iteration flow according to the gradient of the replica free entropy Φ («density evolution» eqns)

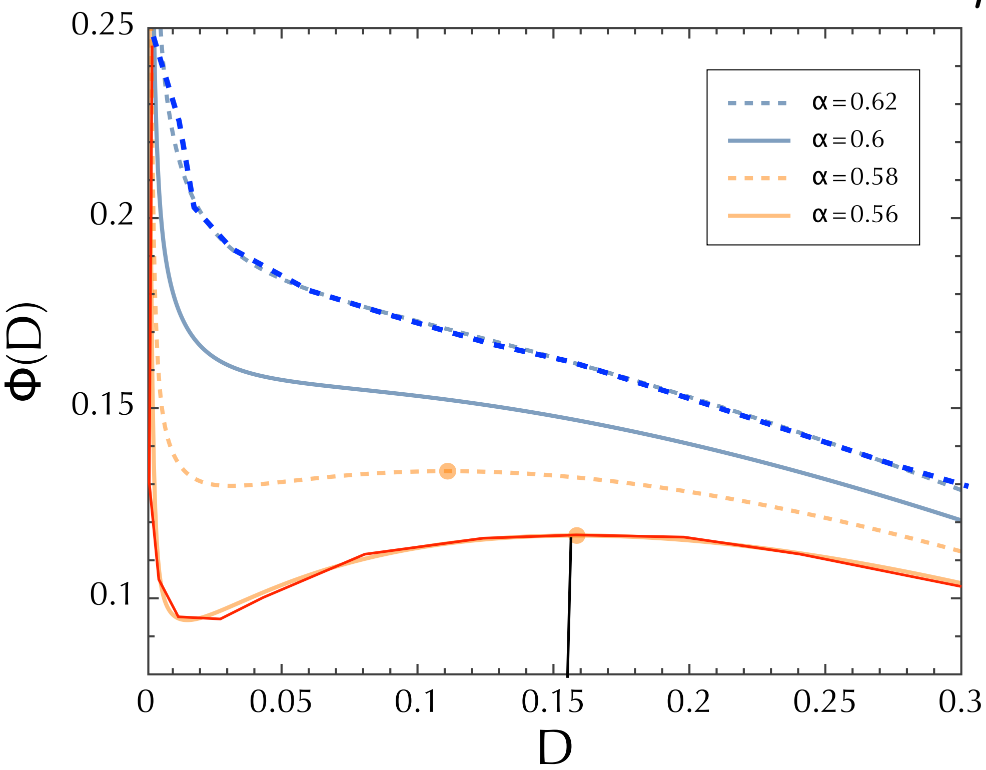
⇒ analytic control of the BP equations

NB rigorous: Bayati Montanari, Lelarge Montanari

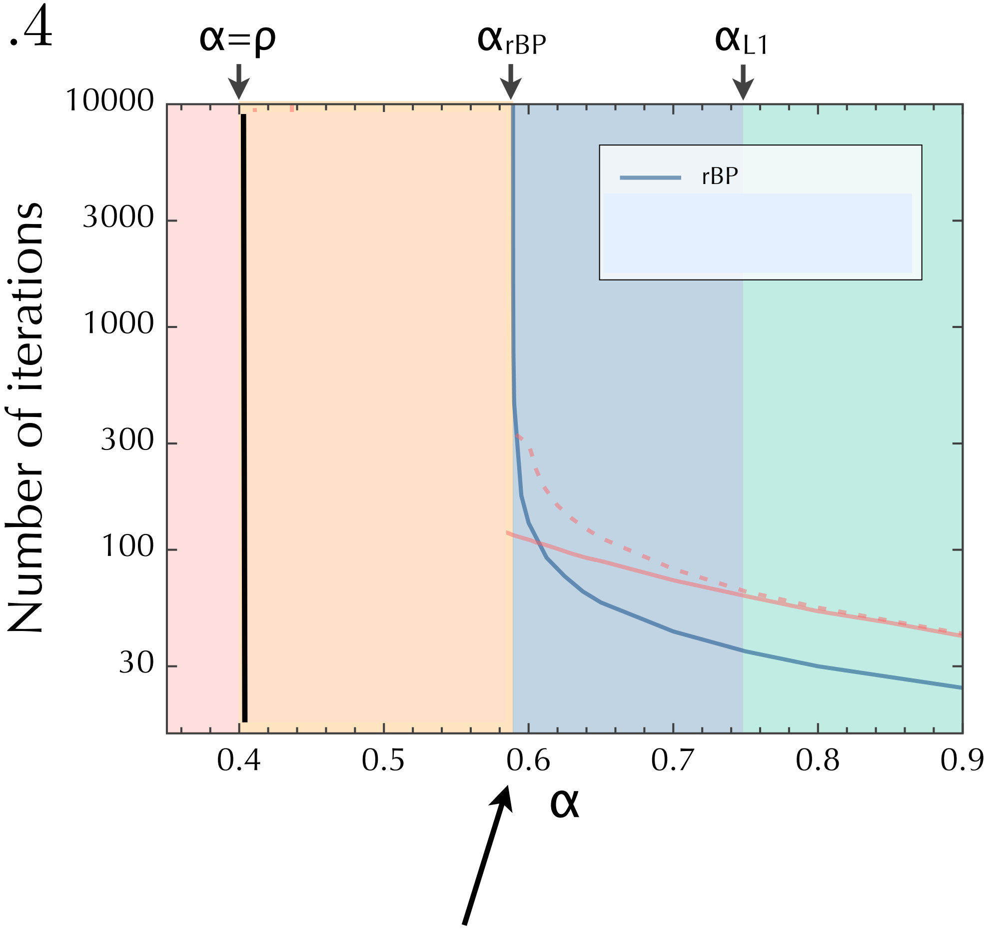
Free entropy $\sim \log P(D)$

BP convergence time

$\rho_0 = .4$



distance to native state



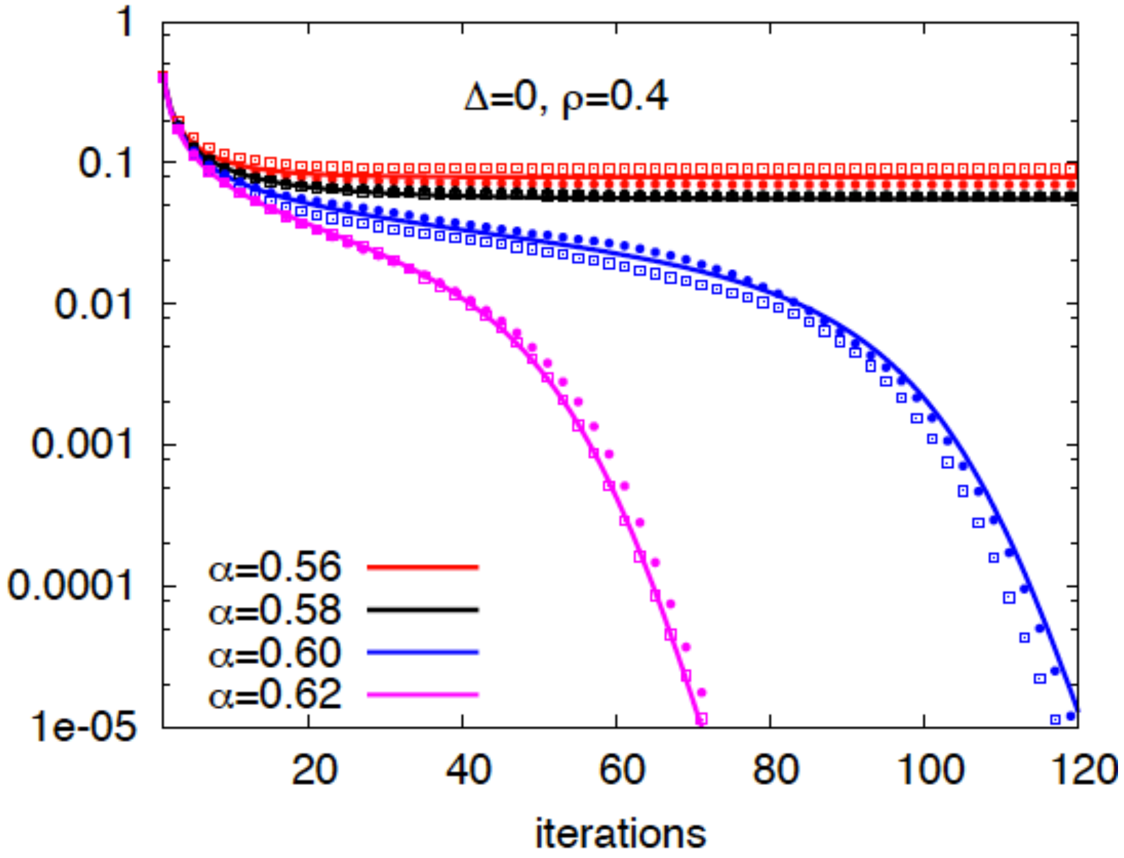
Dynamic glass transition

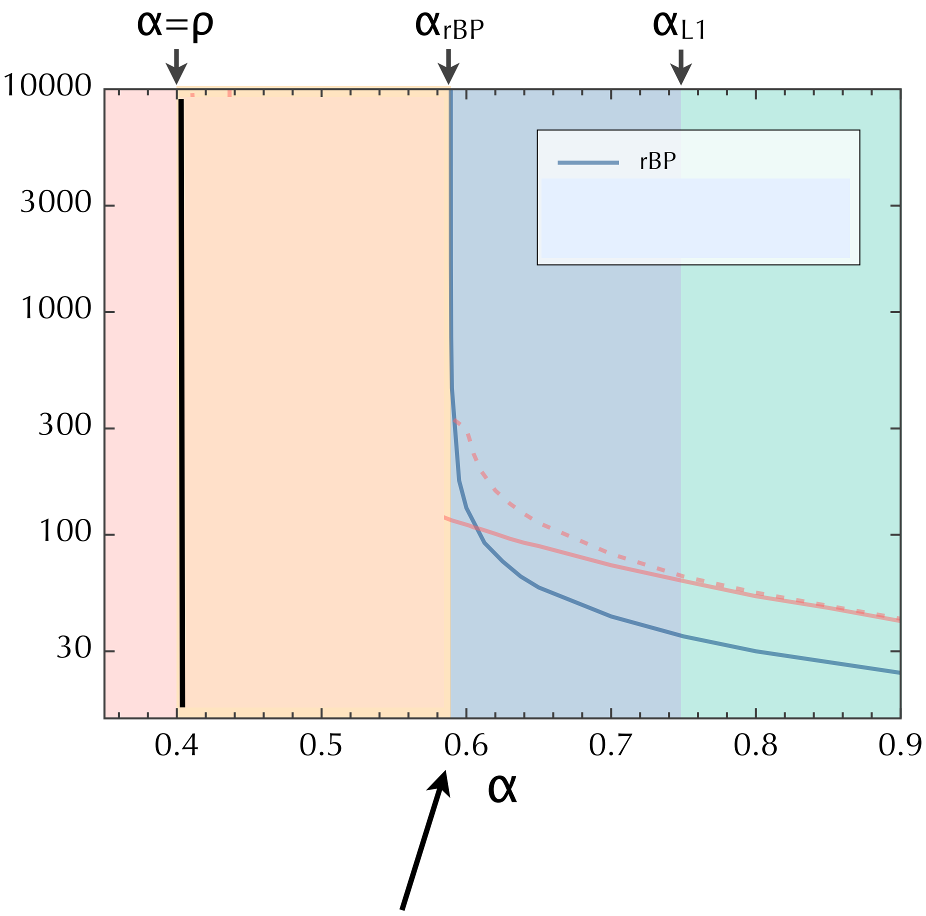
When α is too small, BP is trapped in a **glass phase**

Error

BP convergence time

$\rho_0 = .4$

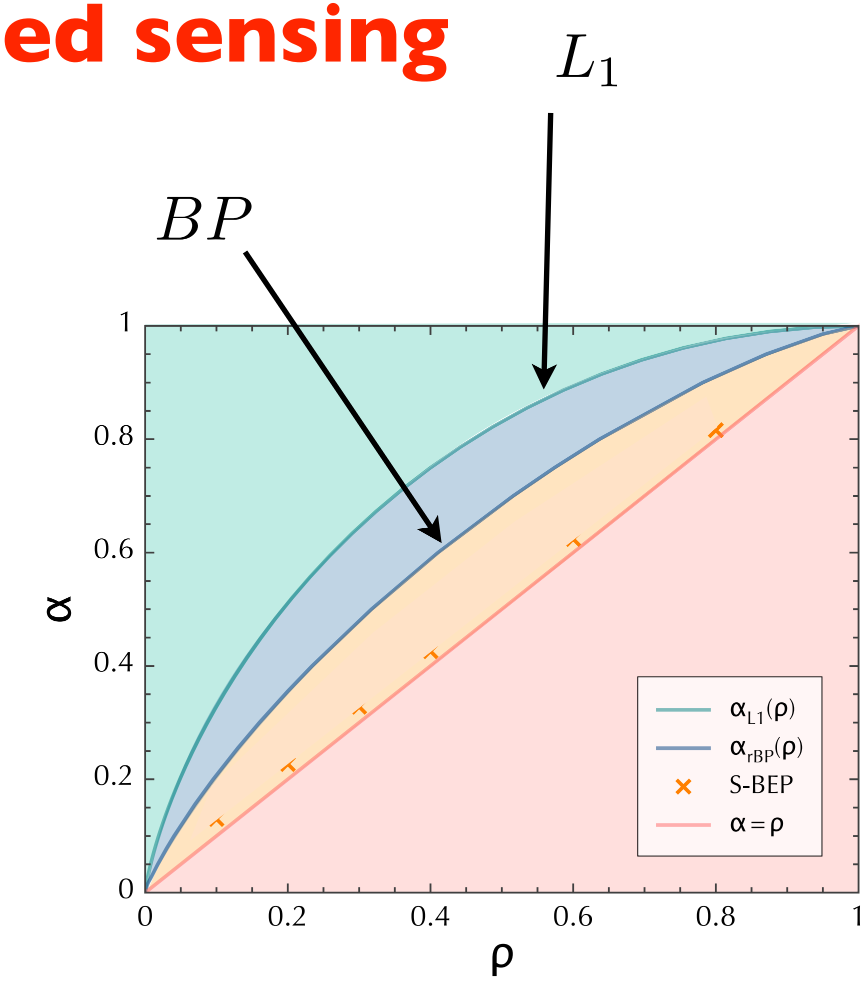




Dynamic glass transition

NB comparison of theory (replica, cavity, density evolution) and numerical experiment

Phase diagram for compressed sensing



L_1 Find a N- component vector x such that the M equations are satisfied and $||x||_1$ is minimal

BP Bayesian approach: max of $P(x|y)$ studied by BP

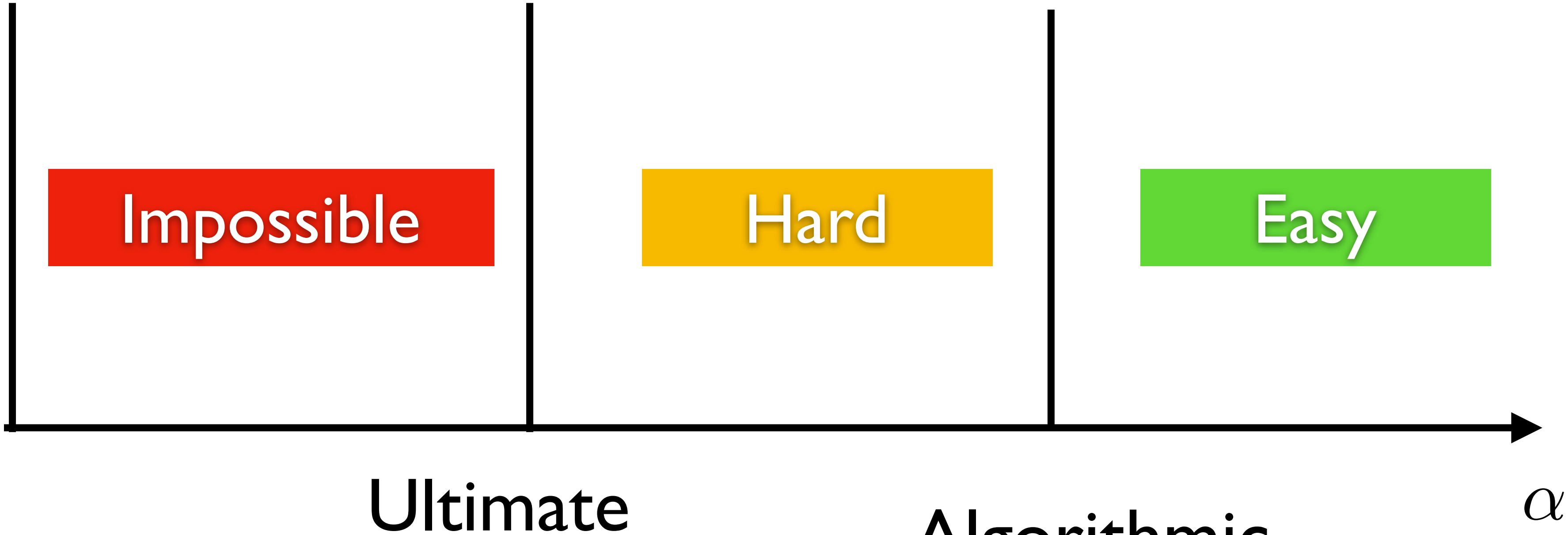
Ensemble: iid elements of $F \sim \mathcal{N}(0, 1/N)$

Analysis of random instances : phase transitions

N (real) variables, M measurements (linear functions)

Analysis of random instances : phase transitions

Reconstruction of signal using BP. Fixed ρ, increase α



Impossible

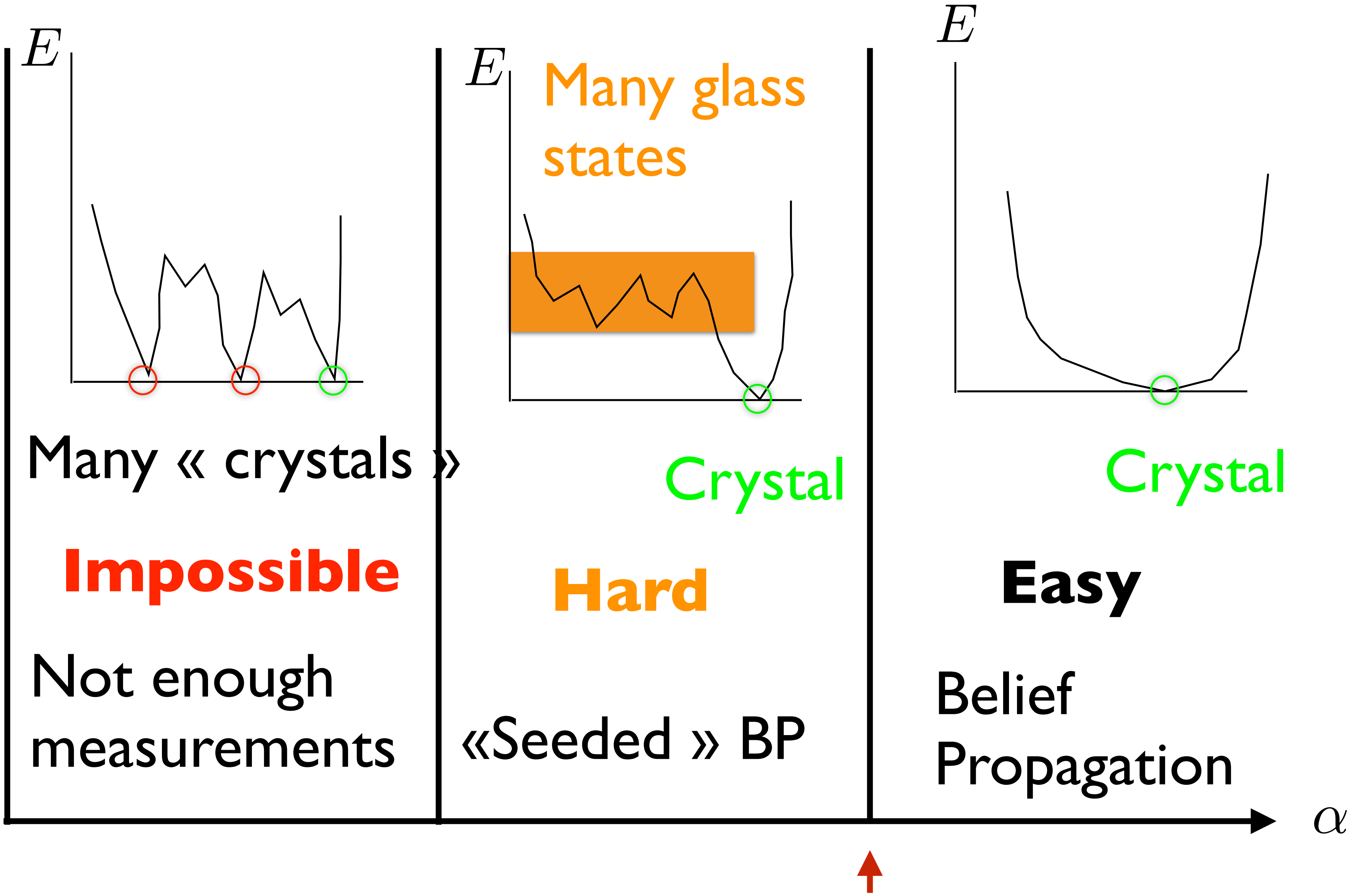
Hard

Easy

α

Ultimate (information theoretic) threshold

Algorithmic threshold



Many « crystals »

Impossible

Not enough measurements

Many glass states

Crystal

Hard

«Seeded » BP

Crystal

Easy

Belief Propagation

Dynamical phase transition. Ubiquitous in statistical inference. Conjecture « All local algorithms freeze »… How universal?

Getting around the glass trap

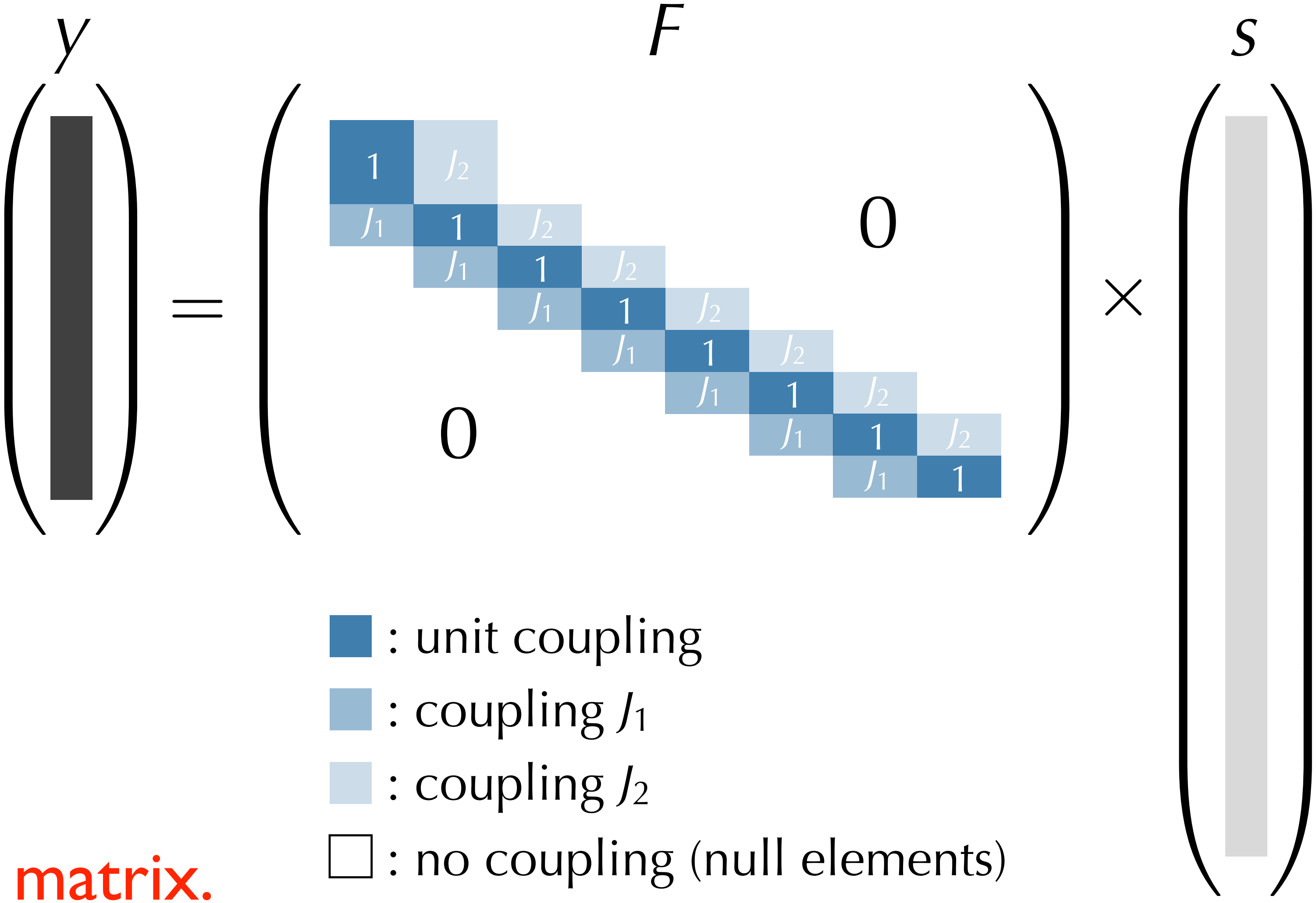
Design the matrix F so that one nucleates the naive state (crystal nucleation idea,
...borrowed from error correcting codes : « spatial coupling »)

Felström-Zigangirov,
Kudekar Richardson Urbanke,
Hassani Macris Urbanke,
...

«Seeded BP»

Nucleation and seeding

Nucleation and seeding



🟦 : unit coupling

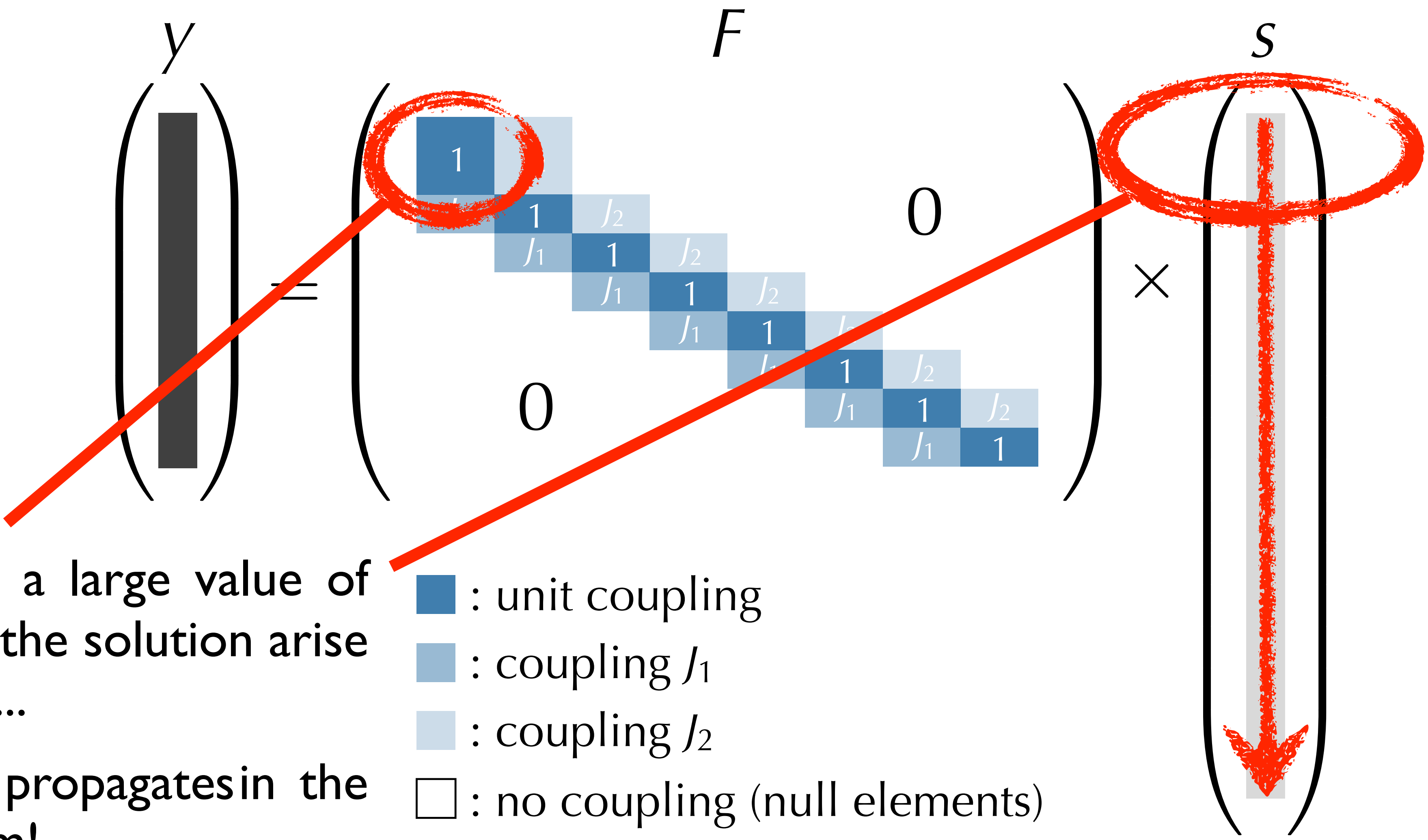
▨ : coupling J_1

░ : coupling J_2

□ : no coupling (null elements)

Structured measurement matrix.
Variances of the matrix elements

$F_{\mu i} =$ independent random Gaussian variables, zero mean and variance $J_{b(\mu)b(i)}/N$



Block 1 has a large value of M such that the solution arise in this block...

... and then propagatesin the whole system!

■ : unit coupling

▨ : coupling J_1

░ : coupling J_2

□ : no coupling (null elements)

$$L = 8$$

$$N_i = N/L$$

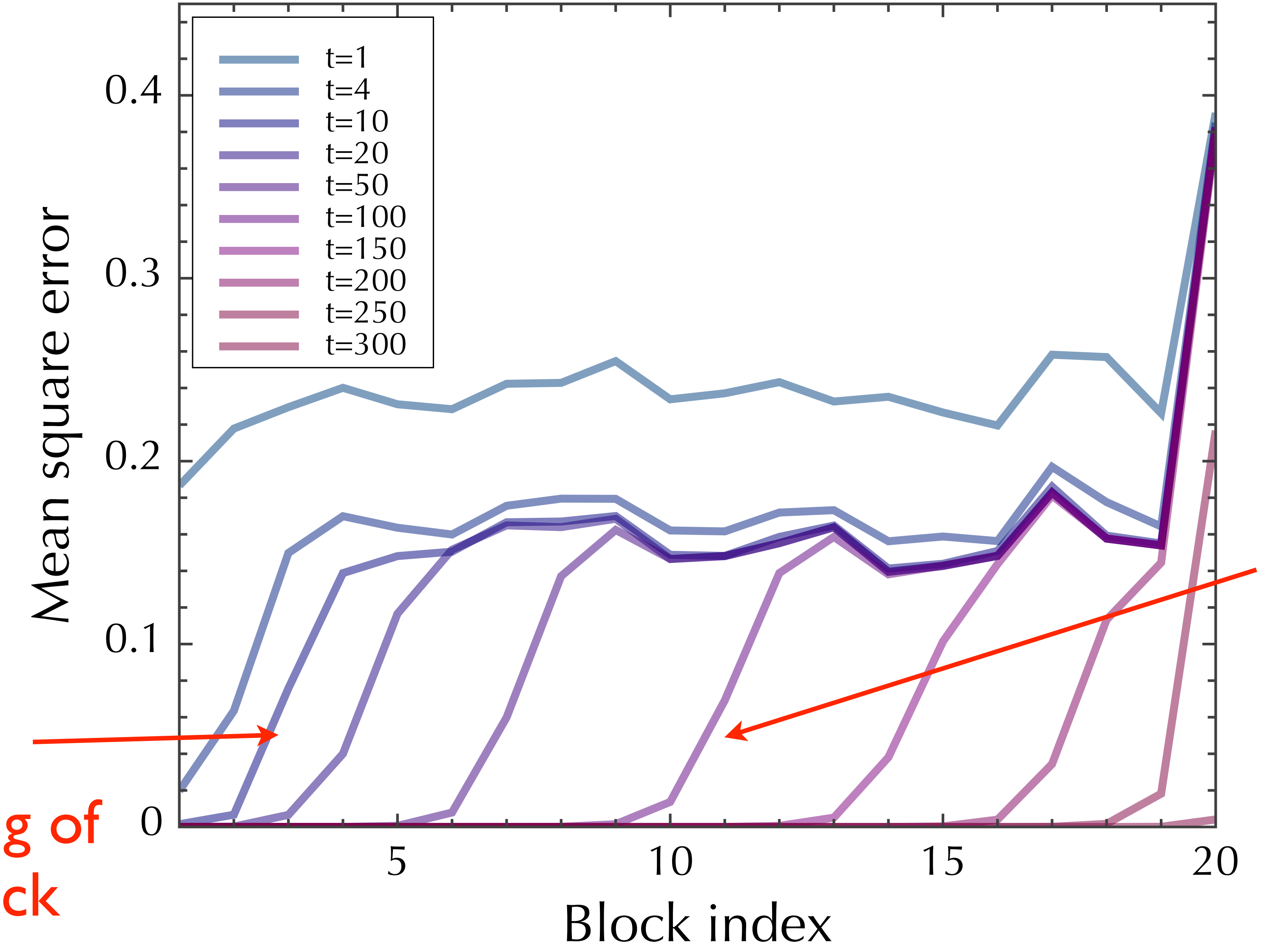
$$M_i = \alpha_i N/L$$

$$\alpha_1 > \alpha_{BP}$$

$$\alpha_j = \alpha' < \alpha_{BP} \qquad j \geq 2$$

$$\alpha = \frac{1}{L}\left(\alpha_1 + (L-1)\alpha'\right)$$

Numerical study



$t = 100$

decoding of blocks 1 to 9

$t = 10$

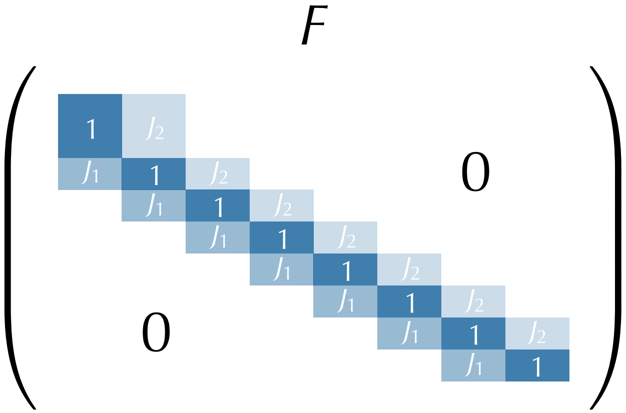
decoding of first block

$L = 20$ $N = 50000$ $\rho = .4$ $J_1 = 20$ $\alpha_1 = 1$

$J_2 = .2$ $\alpha = .5$

Performance of the probabilistic approach + message passing + parameter learning+ seeding matrix

$$Z = \int \prod_{j=1}^{N} \mathrm{d}x_j \prod_{i=1}^{N} [(1-\rho)\delta(x_i) + \rho\phi(x_i)] \prod_{\mu=1}^{M} \delta\left(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i\right)$$



- Simulations
- Analytic approaches (replicas and cavity)

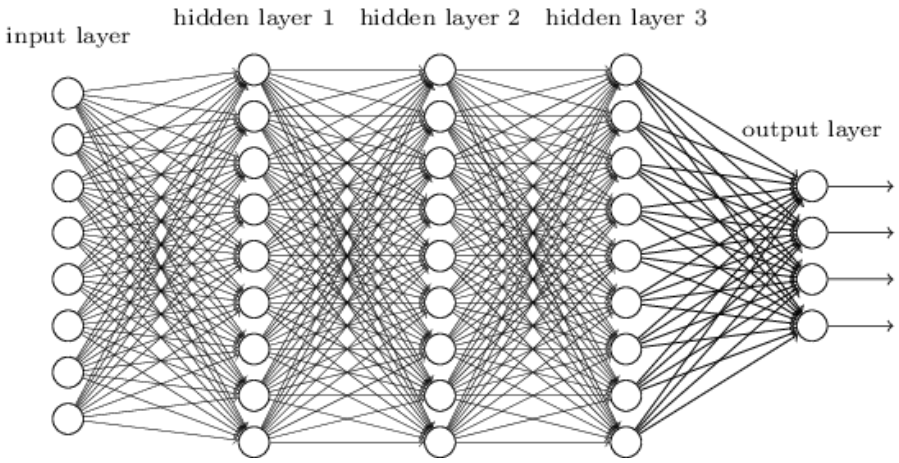
$$\boxed{\rightarrow \alpha_c = \rho_0}$$

Reaches the ultimate information-theoretic threshold

Proof: Donoho Javanmard Montanari

Part Two

Back to Machine learning: the importance of data structure



Why does it work?

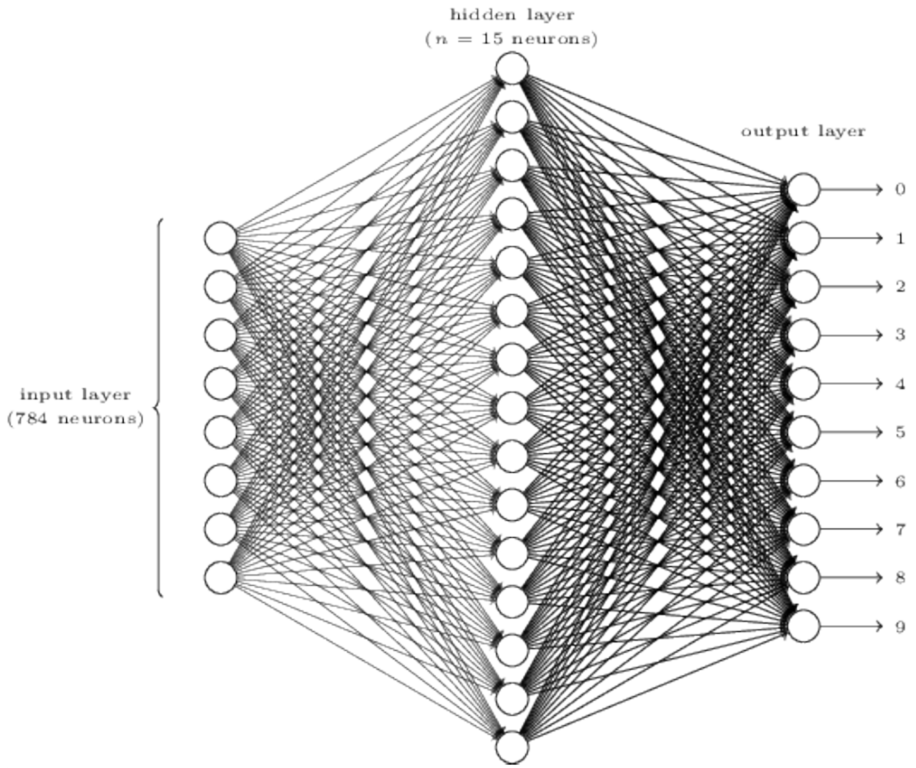
Data structure

- Hidden manifolds and sub manifolds
- Combinatorial structure
- Euclidean correlations

- Analyse data
- Build generative models that can be analyzed fully in some large size limit
- Understand mechanisms

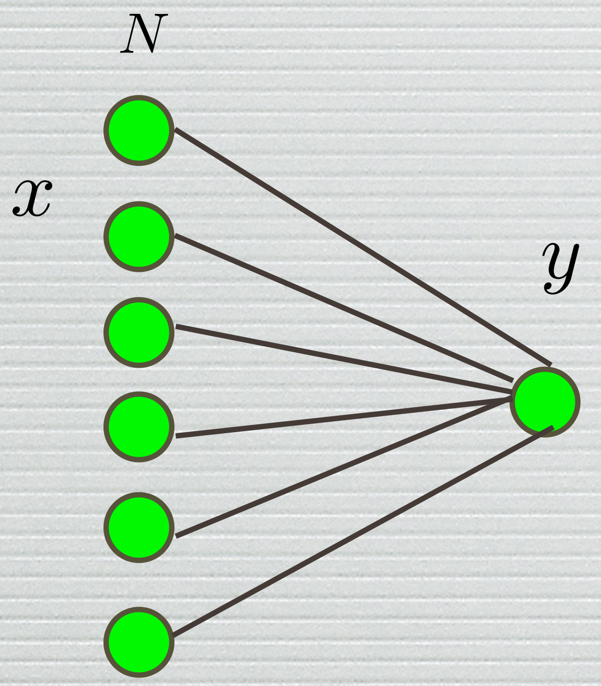
Theory: Ensembles of data, prior on weights

Mostly used so far
Data = input patterns
with **iid entries**



Perceptron learning, committee machine, teacher-student
Many results in the 90's

Analytic study of perceptron learning



Task to be learnt= teacher perceptron

$$y = Sign(J.x) \qquad J_i = \pm 1$$

Learning= student perceptron

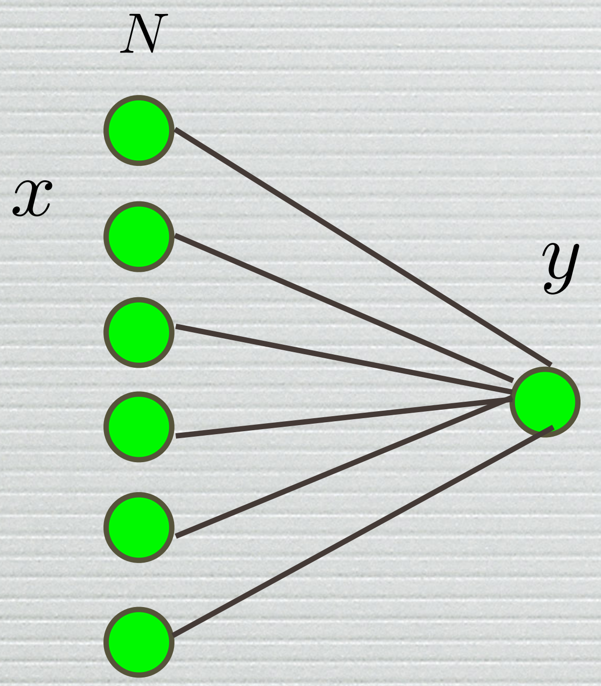
$$y = Sign(K.x) \qquad K_i = \pm 1$$

Machine learning: database of P examples x^μ

and the desired labels $y^\mu = Sign(J.x^\mu)$

Learn the components of K. Compute the generalization error

Analytic study of perceptron learning



Task to be learnt= teacher perceptron

$$y = Sign(J.x) \qquad J_i = \pm 1$$

Learning= student perceptron

$$y = Sign(K.x) \qquad K_i = \pm 1$$

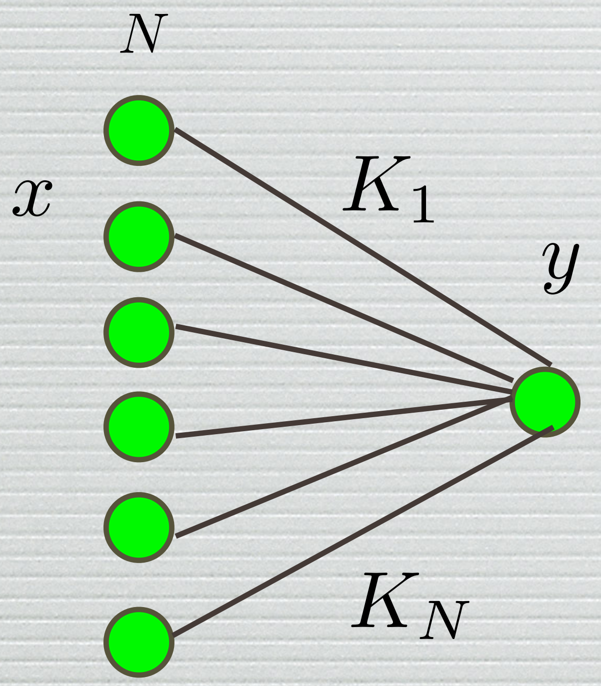
Machine learning: database of P examples x^μ

and the desired labels $y^\mu = Sign(J.x^\mu)$

Learn the components of K. Compute the generalization error

Ensemble: iid x_i^μ eg $\sim \mathcal{N}(0, 1/N)$

Analytic study of perceptron learning



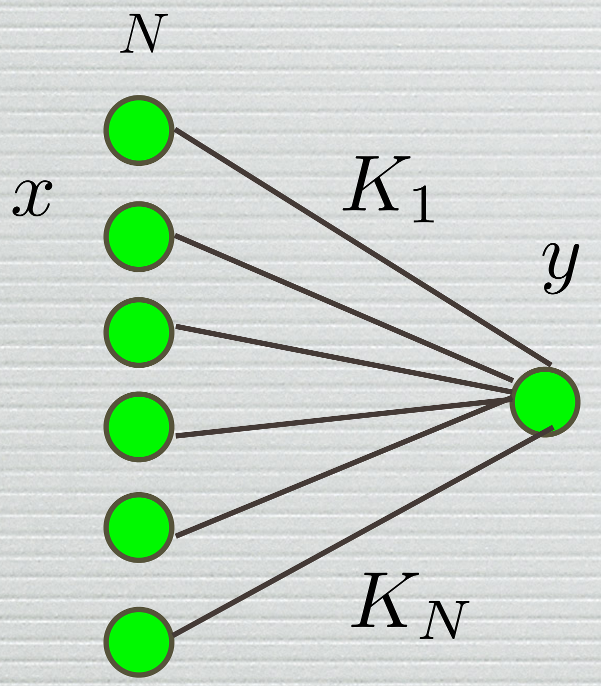
Task to be learnt= teacher perceptron

$$y^\mu = Sign(J.x^\mu)$$

Learning= student perceptron

$$y = Sign(K.x) \qquad K_i = \pm 1$$

Analytic study of perceptron learning



Task to be learnt= teacher perceptron

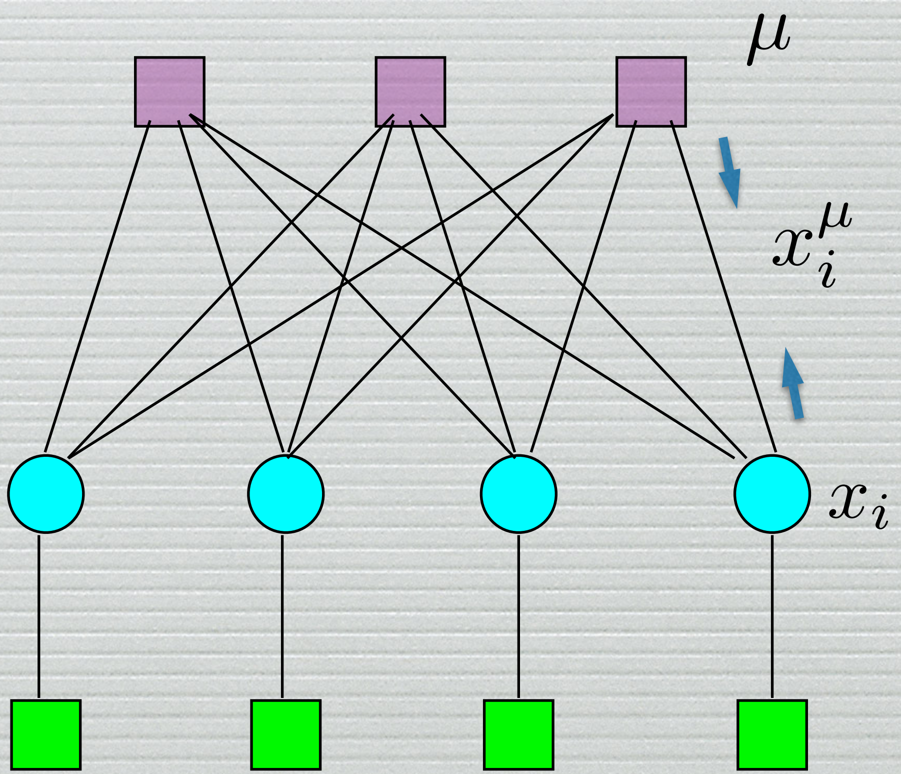
$$y^\mu = Sign(J.x^\mu)$$

Learning= student perceptron

$$y = Sign(K.x) \qquad K_i = \pm 1$$

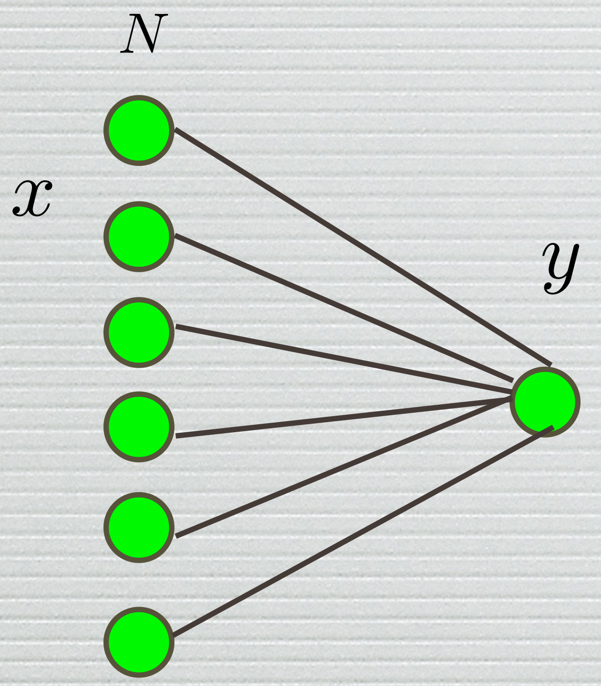
Statistical physics of learning:

$$P(K) = \frac{1}{Z} \prod_{\mu=1}^{P} \delta\left(y^\mu, Sign(K.x^\mu)\right)$$



Similar to compressed sensing !

Analytic study of perceptron learning



Task to be learnt= teacher perceptron

$$y = Sign(J.x) \qquad J_i = \pm 1$$

Learning= student perceptron

$$y = Sign(K.x) \qquad K_i = \pm 1$$

Thermodynamic limit

$$N, P \to \infty$$

$$\alpha = P/N$$

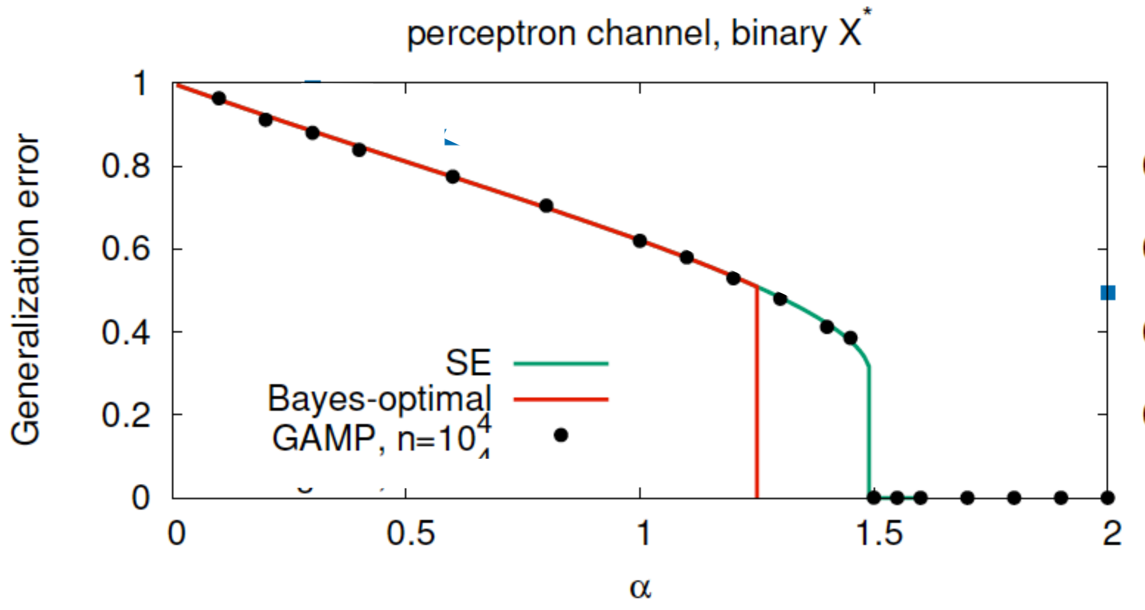
Replicas

Algorithm (BP-cavity)

MM 1989, Opper Haussler 1991

Braunstein Zecchina 2006

Györgyi 1990; Barbier et al 2018



Statistical-physics and probabilistic tools

Precise statements in the thermodynamic limit both on the phase diagram, and on the behavior of some classes of algorithms.
But limited to an ensemble of disorder (in compressed sensing: choice of F)

Complementary to other theoretical approaches that apply to a large class of problems (eg L_1 norm applies to all F with RIP properties), or to the worst case

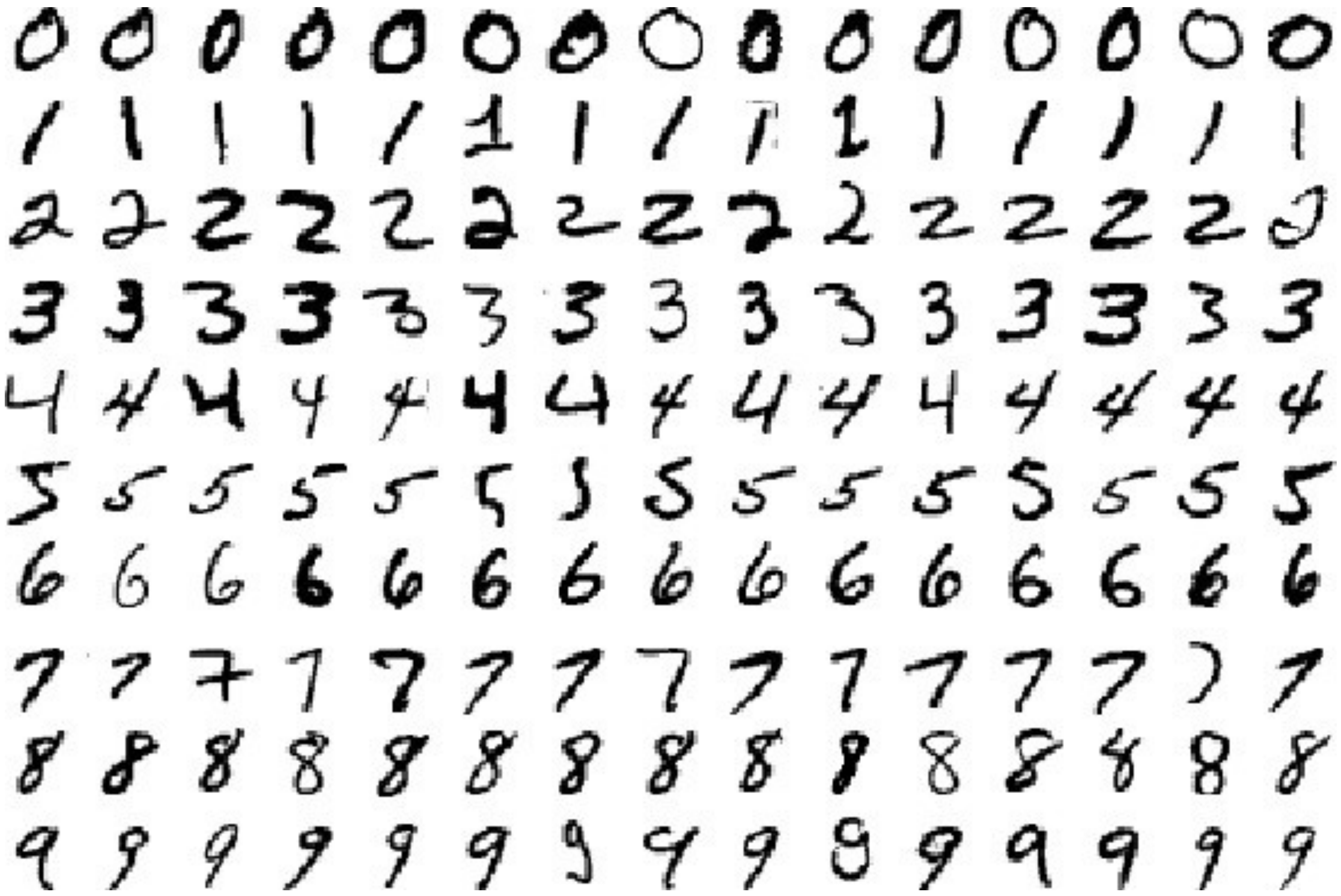
What ensembles can be studied?

What ensembles have been studied?

Does it matter?

The hidden manifold of data

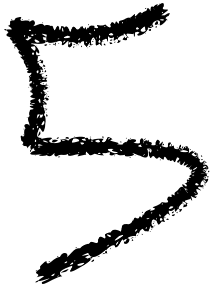
MNIST

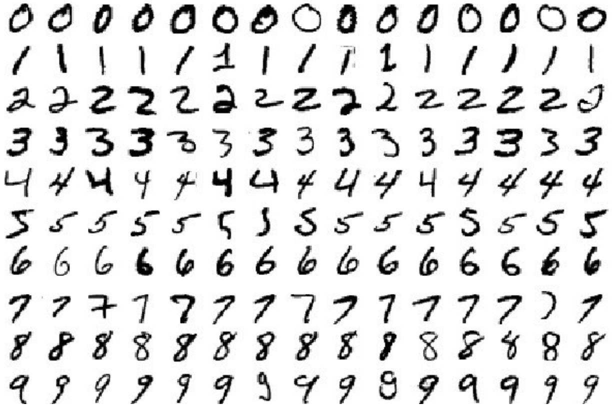


Input space: dimension $28^2 = 784$

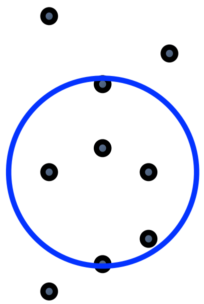
The hidden manifold of data

Input space: dimension $28^2 = 784$





Manifold of handwritten digits in MNIST:

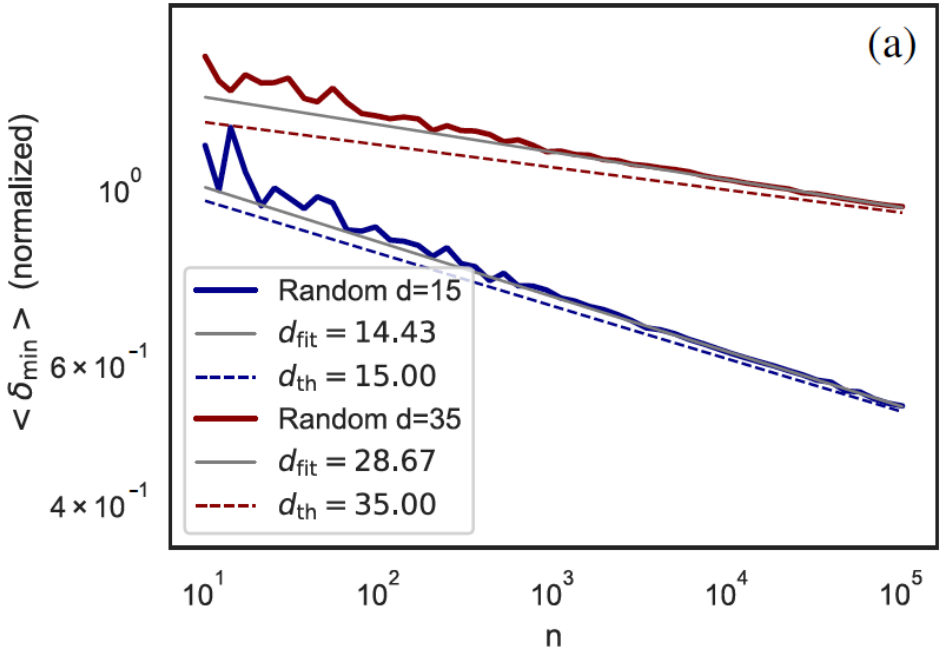


$p \simeq cR^d$

Nearest neighbors' distance: $R_{nn} \simeq p^{-1/d}$

Grassberger Procaccia 83, Costa Hero 05, Heinz Audibert 05, Ansuini et al. 19, Spigler et al. 19…

The hidden manifold of data



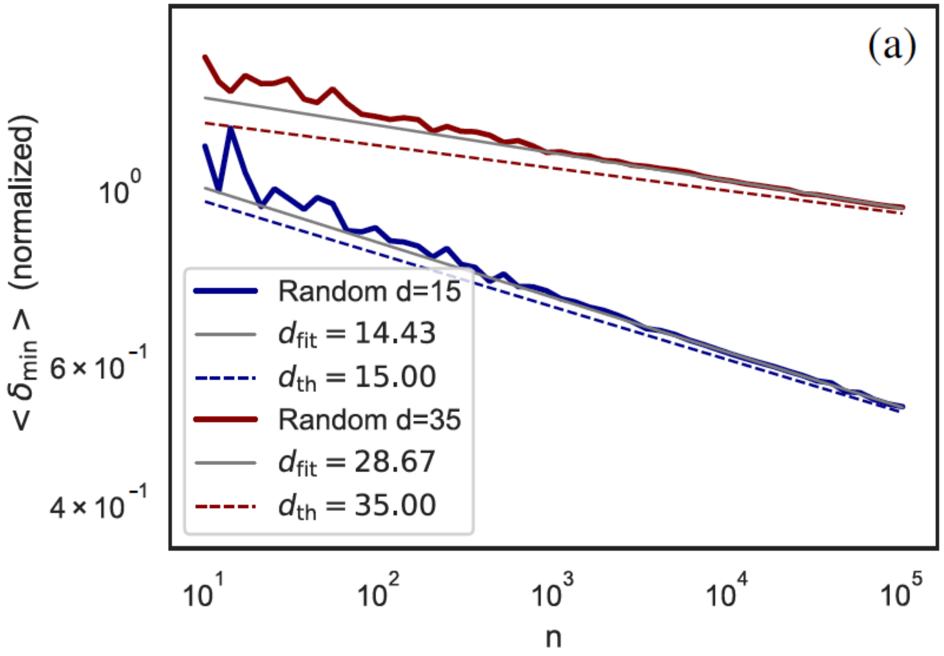
MNIST: $d = 784$

$d_{\mathrm{eff}} \simeq 15$

Spigler et al. 19

Nearest neighbors' distance : $R_{nn} \simeq p^{-1/d}$

The hidden manifold of data

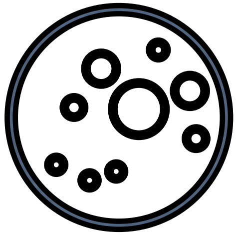


MNIST: $d = 784$

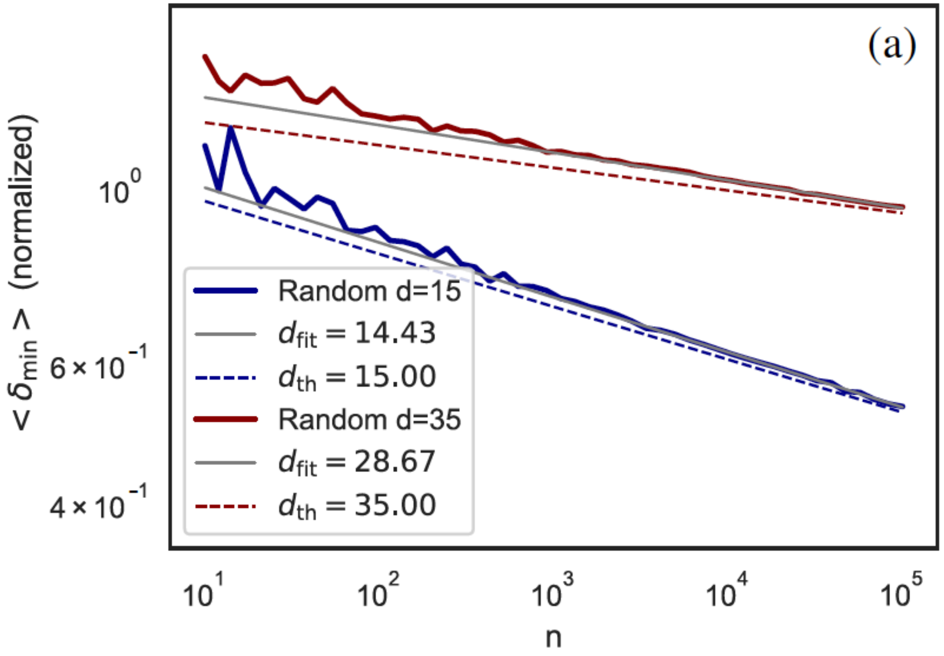
$d_{\text{eff}} \simeq 15$

Spigler et al. 19

Nearest neighbors' distance : $R_{nn} \simeq p^{-1/d}$



The hidden manifold of data

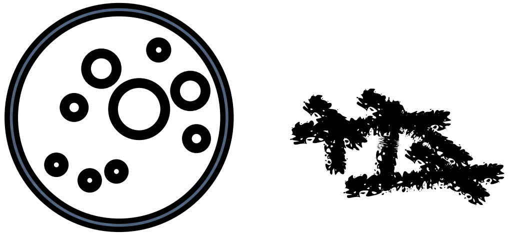


MNIST: $d = 784$

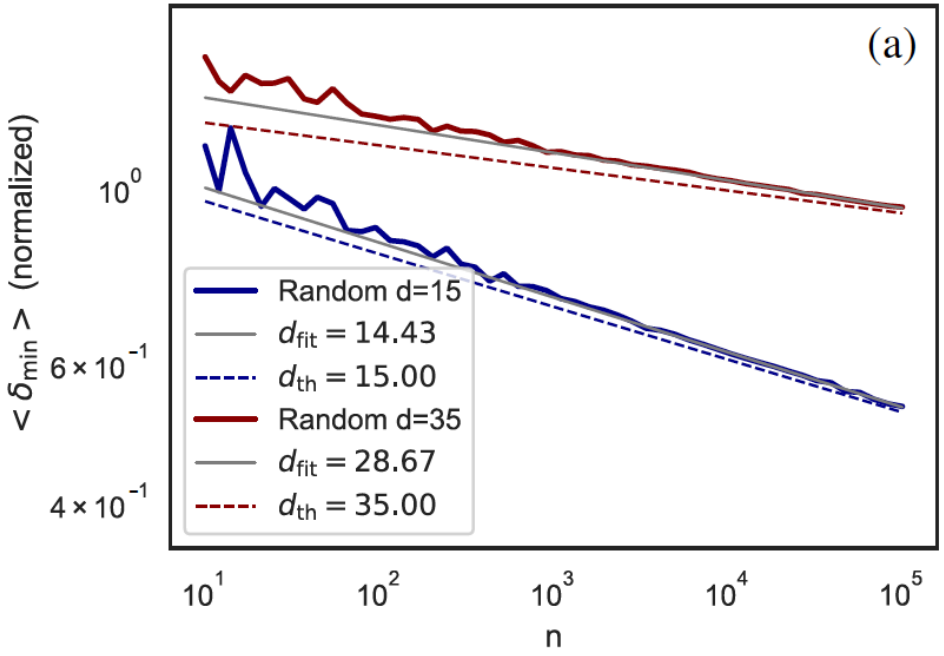
$d_{\text{eff}} \simeq 15$

Spigler et al. 19

Nearest neighbors' distance : $R_{nn} \simeq p^{-1/d}$



The hidden manifold of data

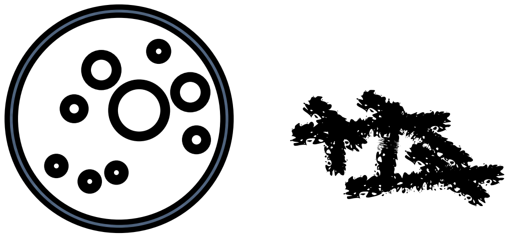


MNIST: $d = 784$

$d_{\text{eff}} \simeq 15$

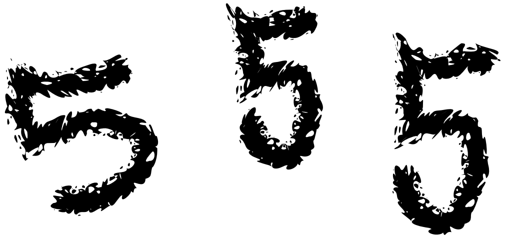
Spigler et al. 19

Nearest neighbors' distance : $R_{nn} \simeq p^{-1/d}$



The neural net should answer: this image does not belong to the category of handwritten digits on which I have been traines

Structure of the task: perceptual sub-manifolds



$d_{\text{eff}}(5) \simeq 12$

Hein Audibert 05

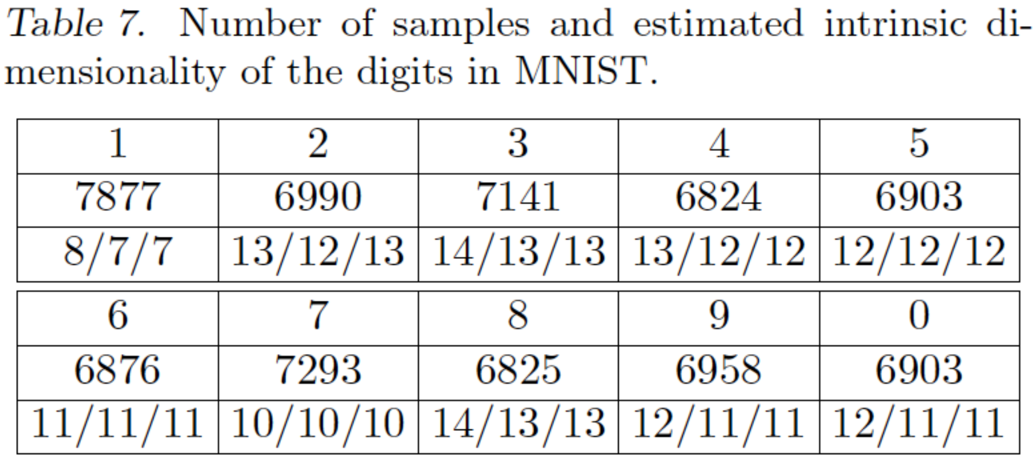


Table 7. Number of samples and estimated intrinsic dimensionality of the digits in MNIST.

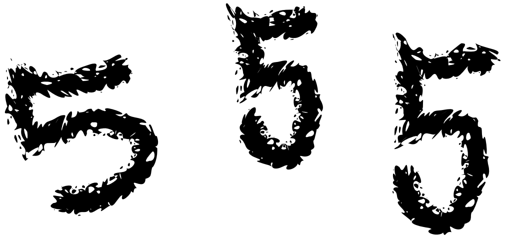
1	2	3	4	5
7877	6990	7141	6824	6903
8/7/7	13/12/13	14/13/13	13/12/12	12/12/12

6	7	8	9	0
6876	7293	6825	6958	6903
11/11/11	10/10/10	14/13/13	12/11/11	12/11/11

MNIST problem: in the **15-dim manifold** of handwritten digits, identify the **10 perceptual sub manifolds** associated with each digit, of **dimensions between 7 and 13**…

… from an input in 784 dimensions!

Structure of the task: perceptual sub-manifolds



$$d_{\mathrm{eff}}(5) \simeq 12$$

Hein Audibert 05

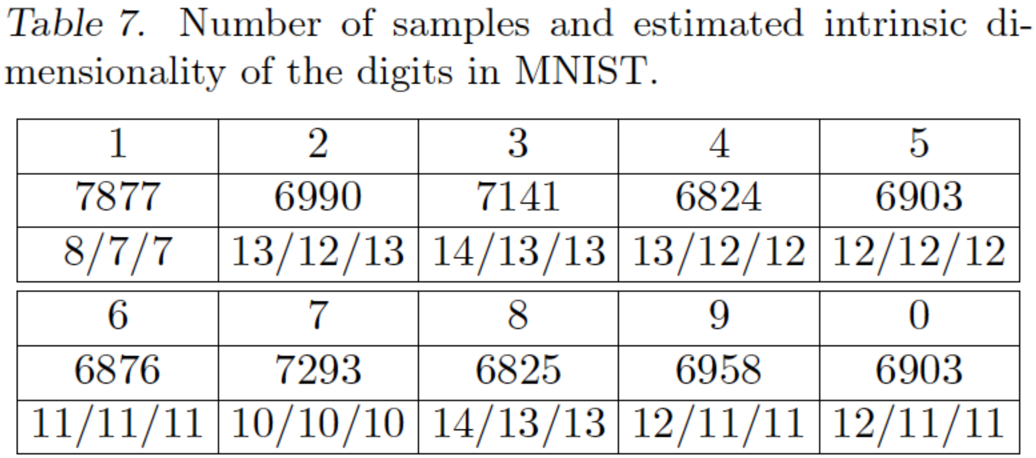


Table 7. Number of samples and estimated intrinsic dimensionality of the digits in MNIST.

1	2	3	4	5	6	7	8	9	0
7877	6990	7141	6824	6903	6876	7293	6825	6958	6903
8/7/7	13/12/13	14/13/13	13/12/12	12/12/12	11/11/11	10/10/10	14/13/13	12/11/11	12/11/11

MNIST problem: in the **15-dim manifold** of handwritten digits, identify the **10 perceptual sub manifolds** associated with each digit, of **dimensions between 7 and 13**…

… from an input in 784 dimensions!

Very different from iid inputs !

An ensemble
for the hidden manifold
and for the task to be achieved

S. Goldt, F. Krzakala MM L. Zdeborova

arXiv:1909.11500

An ensemble for the hidden manifold

Pattern μ :

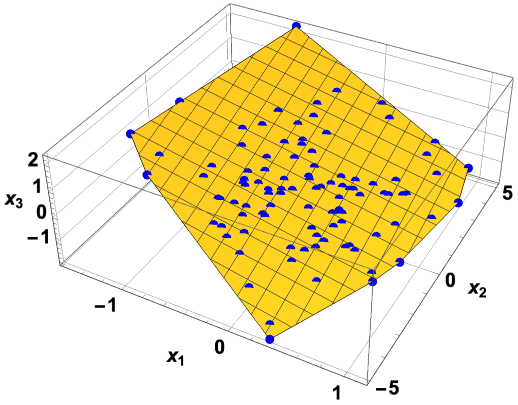
$$X_{\mu i} = f\left[\frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_{\mu r}F_{ir}\right]$$

Data = input patterns built from R features $\vec{F}_r$

A feature is a N component vector in the input space

Each pattern is built from a weighted superposition of features (feature r has weight C_r):

$$\sum_{r=1}^{R} C_r \vec{F}_r$$



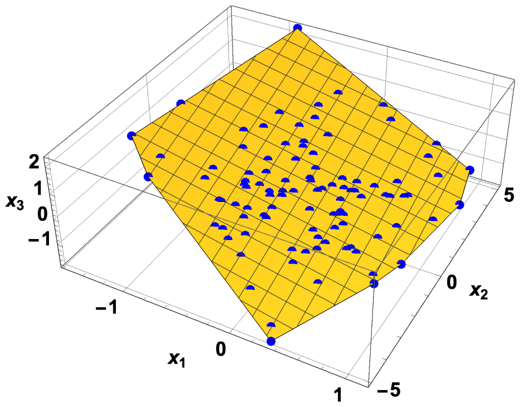
An ensemble for the hidden manifold

Pattern μ:
$$X_{\mu i} = f\left[\frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_{\mu r} F_{ir}\right]$$

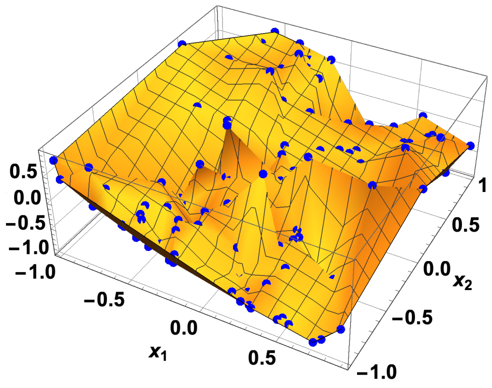
Data = input patterns built from R features $\vec{F}_r$

A feature is a N component vector in the input space

Each pattern is built from a weighted superposition of features (feature r has weight C_r):
$$\sum_{r=1}^{R} C_r \vec{F}_r$$



The R-dimensional data manifold is folded by applying the non-linear function f



An ensemble for the task

$$\vec{X} = f\left[\frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_r \vec{F}_r\right]$$

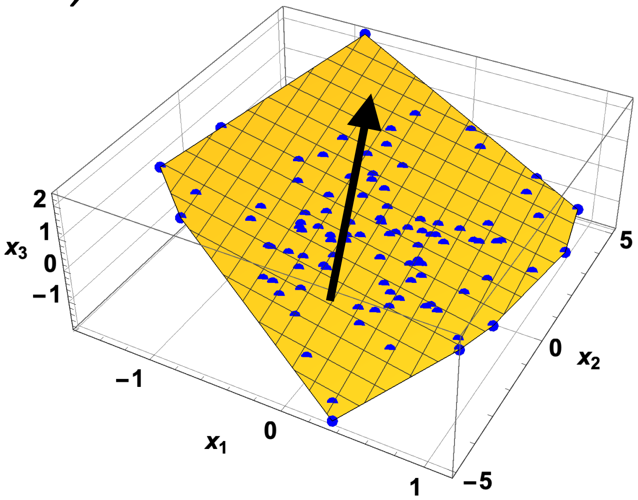
« Latent representation »: $\{C_r\}$

iid

Desired output = **function of latent representation**

Examples: $y = g\left(\sum_{r=1}^{R} \tilde{w}_r C_r\right)$

(perceptron in hidden manifold)



An ensemble for the task

$$\vec{X} = f\left[\frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_r \vec{F}_r\right]$$

« Latent representation »: $\{C_r\}$

Desired output (task) = function of latent representation

Examples: $y = g\left(\sum_{r=1}^{R} \tilde{w}_r C_r\right)$ (perceptron in latent space)

$$y = \sum_{m=1}^{M} \tilde{v}_m \; g\left(\sum_{r=1}^{R} \tilde{w}_{mr} C_r\right)$$

(2 layers nn in latent space)

Manifold of data and sub manifolds of the task

$$\vec{X} = f\left[\frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_r \vec{F}_r\right]$$

« Latent representation »: $\{C_r\}$

Hidden manifold of data: folded R-dimensional manifold

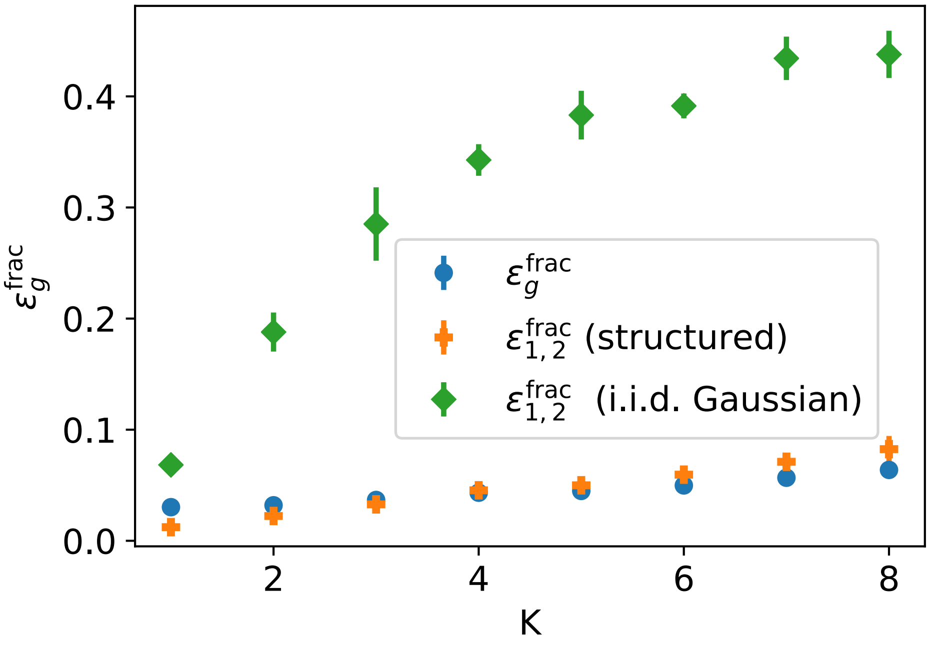
Task $y = \sum_{m=1}^{M} \tilde{v}_m \, g\left(\sum_{r=1}^{R} \tilde{w}_{mr} C_r\right)$

depends on $\{\tilde{w}_m.C\}, \quad m \in \{1, ...M\}$

where $\{\tilde{w}_m\}$ and C live in a R-dim space

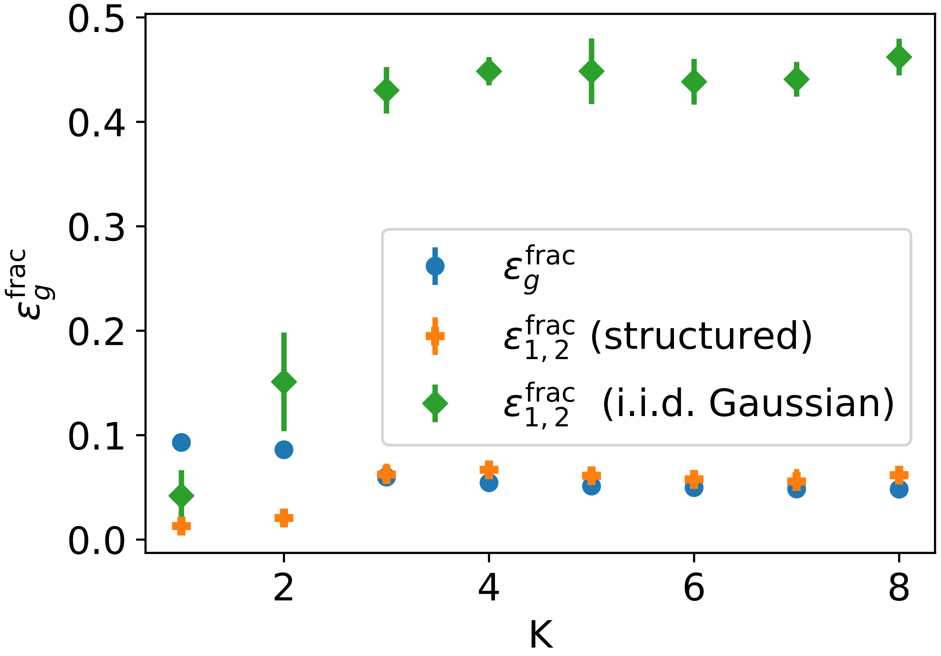
For $M < R$ perceptual sub manifold = moving in directions orthogonal to the $\{\tilde{w}_m\}$, in latent space

Experimenting with the « hidden manifold model »



Hidden manifold model

$R = 10$



MNIST

Hidden manifold model

$$\vec{X} = f\left[\frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_r \vec{F}_r\right]$$

Data. « Latent representation »: $\{C_r\}$

Desired output (task) = function of latent representation

Example $$y = g\left(\sum_{r=1}^{R} \tilde{w}_r C_r\right)$$

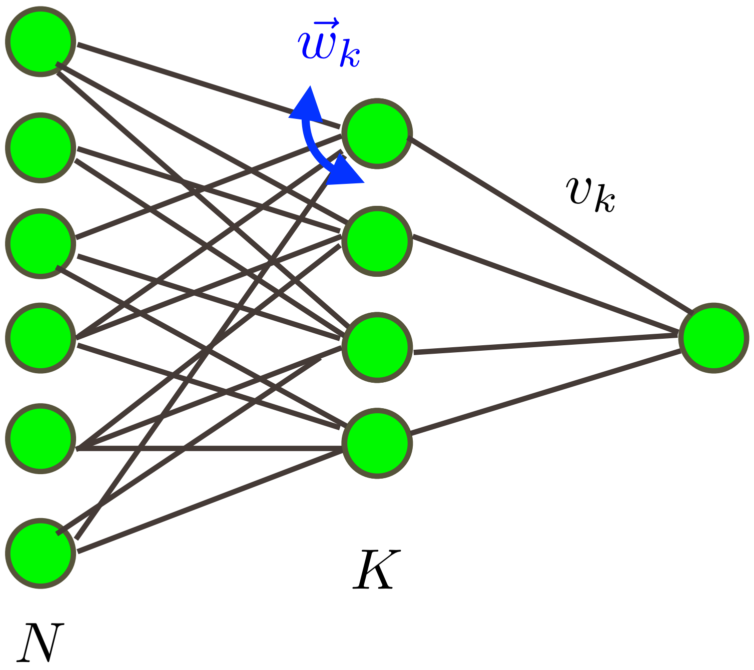
- Does not have the pathologies of teacher-student setup with iid data
- Learning and generalization phenomenology $\sim$ MNIST
- Can be studied analytically: online learning and phase diagram

Analytic study of the hidden manifold model

$$\vec{X} = f\left[\frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_r \vec{F}_r\right]$$

Correlated components

iid



Solvable limit = thermodynamic limit with extensive latent dimension $N \to \infty,\ R \to \infty,\ P \to \infty$

With fixed $R/N = \gamma\,,\quad P/N = \alpha\,,\quad K$

Analytic study of the hidden manifold model

$$\vec{X} = f\left[\frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_r \vec{F}_r\right]$$

Correlated components

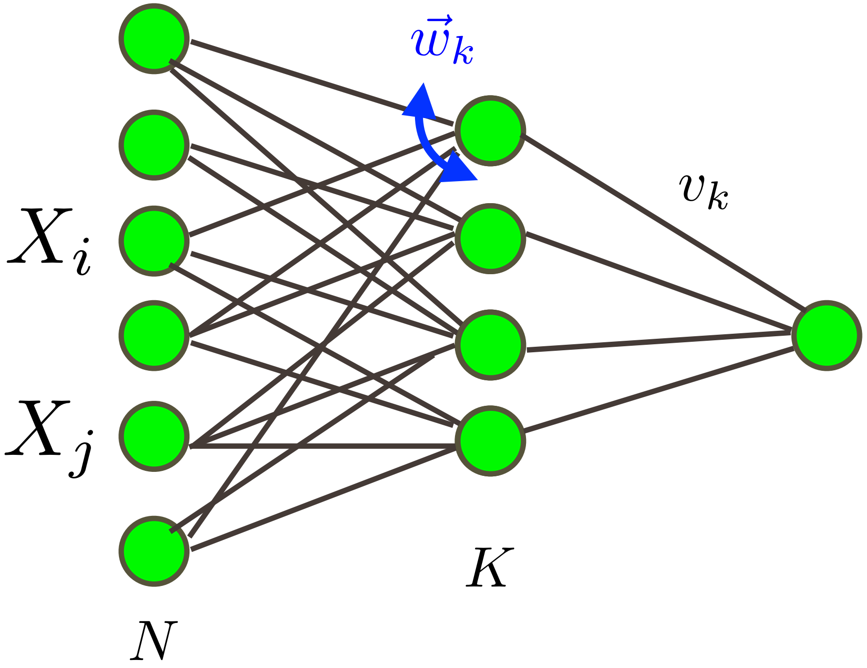
iid

balanced:

$$F_{ri} = O(1)$$

$$\frac{1}{N}\sum_i F_{ri}F_{si} = O(1/\sqrt{N})$$

$$\frac{1}{N}\sum_i F_{ri}F_{ri} = 1$$

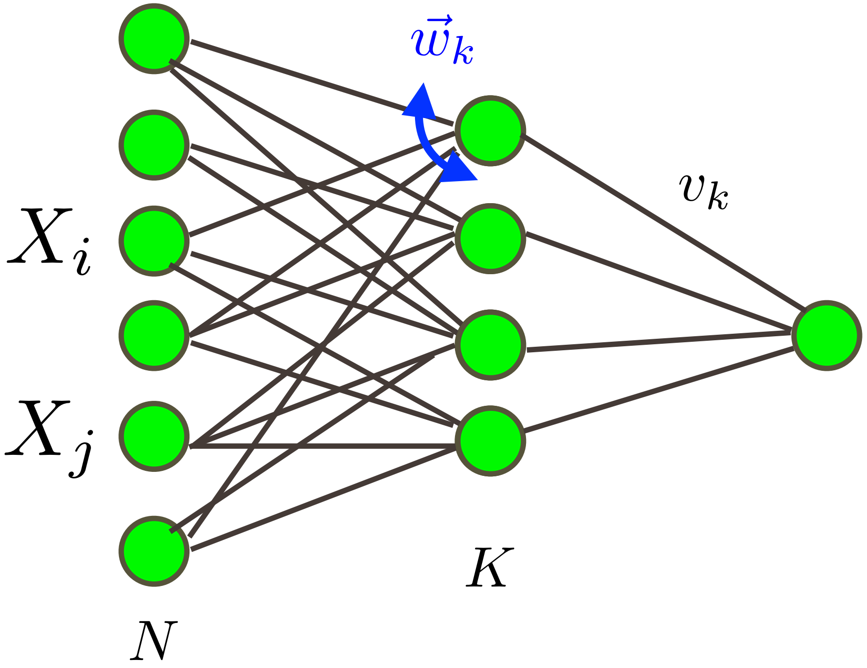


Analytic study of the hidden manifold model

$$\vec{X} = f\left[\frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_r \vec{F}_r\right]$$

Correlated components

iid



$$X_i = f[u_i]$$

$$u_i = \frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_r F_{ri}$$

Gaussian, weakly correlated $O(1/\sqrt{N})$ when F_{ri} are balanced and $O(1)$

$$\mathbb{E}\left(f[u_i]f[u_j]\right) = \langle f(u)\rangle^2 + \langle uf(u)\rangle^2 \mathbb{E}\left(u_i u_j\right)$$

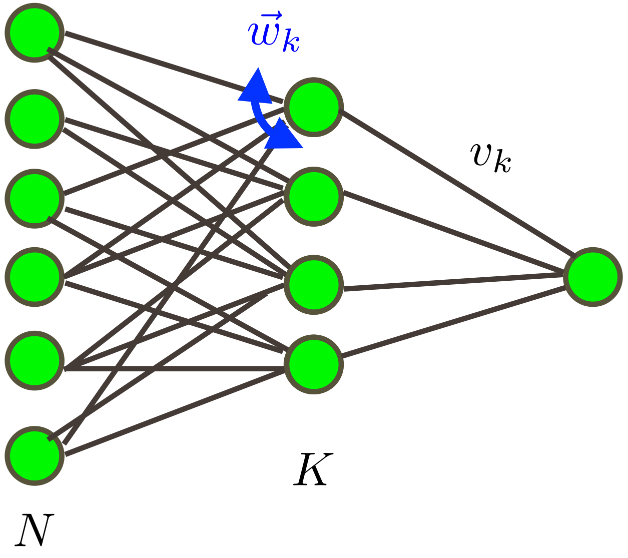
u Gaussian $\mathcal{N}(0,1)$

Gaussian Equivalence Theorem (GET)

$$u_i = \frac{1}{\sqrt{R}} \sum_{r=1}^{R} \boxed{C_r} F_{ri}$$

iid

$$X_i = f[u_i]$$



Inputs of hidden units: $\lambda^k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^k f[u_i]$

GET: In the thermodynamic limit, the variables λ^k have a Gaussian distribution, with covariance

$$\mathbb{E}[\tilde{\lambda}^k \tilde{\lambda}^\ell] = (c - a^2 - b^2) W^{k\ell} + b^2 \Sigma^{k\ell}$$

$$W^{k\ell} \equiv \frac{1}{N} \sum_{i=1}^{N} w_i^k w_i^\ell \qquad \Sigma^{k\ell} \equiv \frac{1}{R} \sum_{r=1}^{R} S_r^k S_r^\ell \qquad S_r^k \equiv \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^k F_{ir}$$

$c = \langle f(u)^2 \rangle$ $\quad a = \langle f(u) \rangle$ $\quad b = \langle u f(u) \rangle$ $\quad u$ Gaussian $\mathcal{N}(0,1)$

Gaussian Equivalence Theorem (GET)

$$u_i = \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r F_{ri}$$

$$X_i = f[u_i]$$

Inputs of hidden units:

$$\lambda^k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^k f[u_i]$$

GET in a nutshell: in the thermodynamic limit (with extensive latent dimension of the hidden manifold, $R = \gamma N$), the inputs of hidden units have Gaussian distribution. Then the model is solvable.

NB: F_{ri} and w_i^k are not necessarily random, but balanced

$$S^{k_1 k_2 \ldots k_p}_{r_1 r_2 \ldots r_q} = \frac{1}{\sqrt{N}} \sum_i w_i^{k_1} w_i^{k_2} \ldots w_i^{k_p} F_{ir_1} F_{ir_2} \ldots F_{ir_q} = O(1)$$

Gaussian Equivalence Theorem (GET)

$$u_i = \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r F_{ri}$$

$$X_i = f[u_i]$$

Inputs of hidden units:

$$\lambda^k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^k f[u_i]$$

GET in a nutshell: in the thermodynamic limit (with extensive latent dimension of the hidden manifold, $R = \gamma N$), the inputs of hidden units have Gaussian distribution. Then the model is solvable.

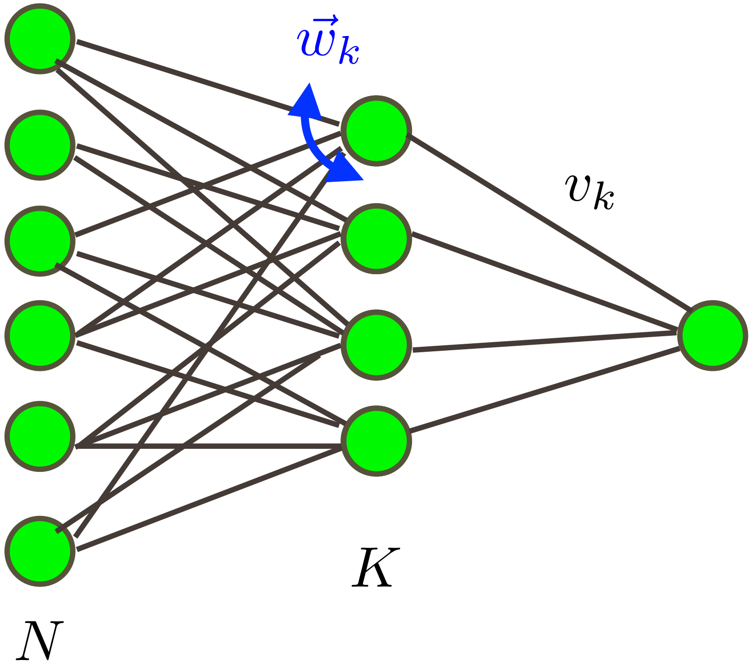
NB: depends on the manifold folding function f only through the three quantities

$c = \langle f(u)^2 \rangle$ $a = \langle f(u) \rangle$ $b = \langle u f(u) \rangle$ u Gaussian $\mathcal{N}(0,1)$

Any folding function f is statistically equivalent to a quadratic one

$$f(u) = \alpha + \beta u + \gamma u^2$$

Online learning of Hidden Manifold Model



Learn using a 2-layer neural net, K hidden units

$$\Phi\left(\vec{X}\right) = \sum_{k=1}^{K} g\left(\vec{w}^k.\vec{X}/\sqrt{N}\right)$$

$$\vec{X} = f\left[\frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_r\vec{F}_r\right]$$

$\vec{X}$ = inside hidden R-dimensional manifold, folded by function f

Desired output given constructed from latent representation

$$\Phi_t(\vec{X}) = \sum_{m=1}^{M} \tilde{g}\left(\sum_{r=1}^{R} \tilde{w}_r^m C_r\right)$$

Online learning: ODE for SGD

Evolution of the weights during learning

D Saad and S Solla 95, Biehl and Schwarze 95, …

$$\left(w_i^k\right)^{\mu+1} - \left(w_i^k\right)^{\mu} = -\frac{\eta}{\sqrt{N}} \Delta g'(\lambda^k) f(u_i)$$

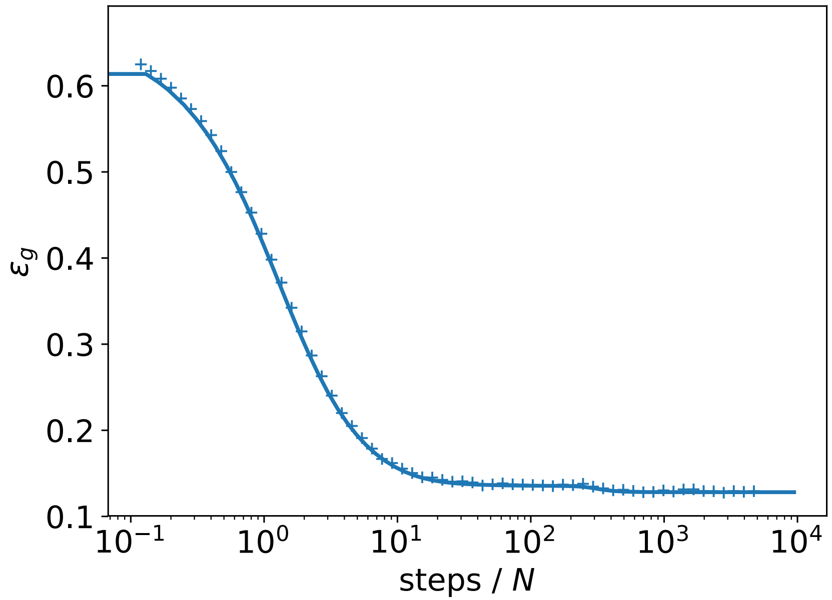
$$\Delta = \sum_{\ell=1}^{K} g(\lambda^\ell) - \sum_{m=1}^{N} \tilde{g}(\nu^m)$$

New pattern (and therefore new latent representation C_r) at each time

GET: $\lambda^k$and ν^m are Gaussian, and the learning dynamics can be analyzed by ordinary differential equations for order parameters like

$$W^{k\ell} \equiv \frac{1}{N} \sum_{i=1}^{N} w_i^k w_i^\ell$$

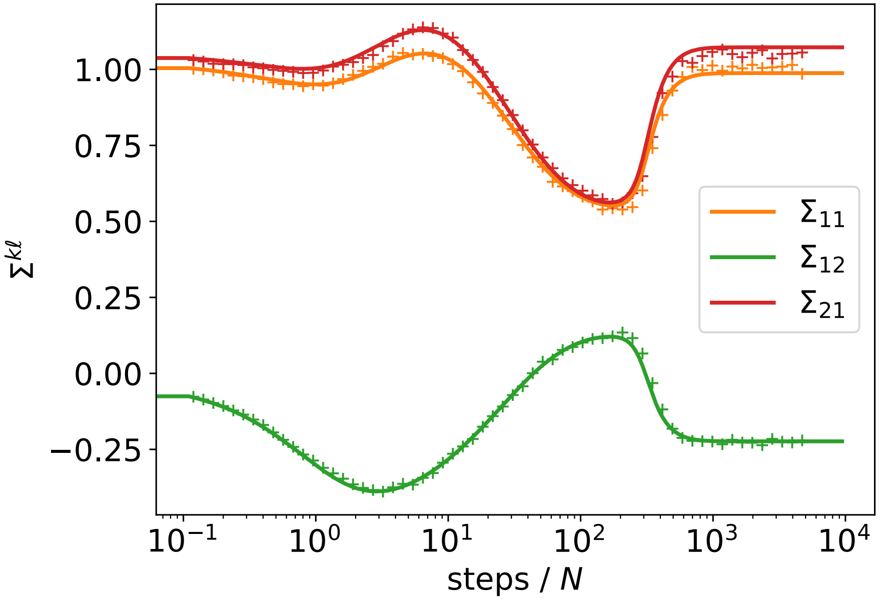
ODE Theory vs simulations N=8000, D=4000, M=2, K=2

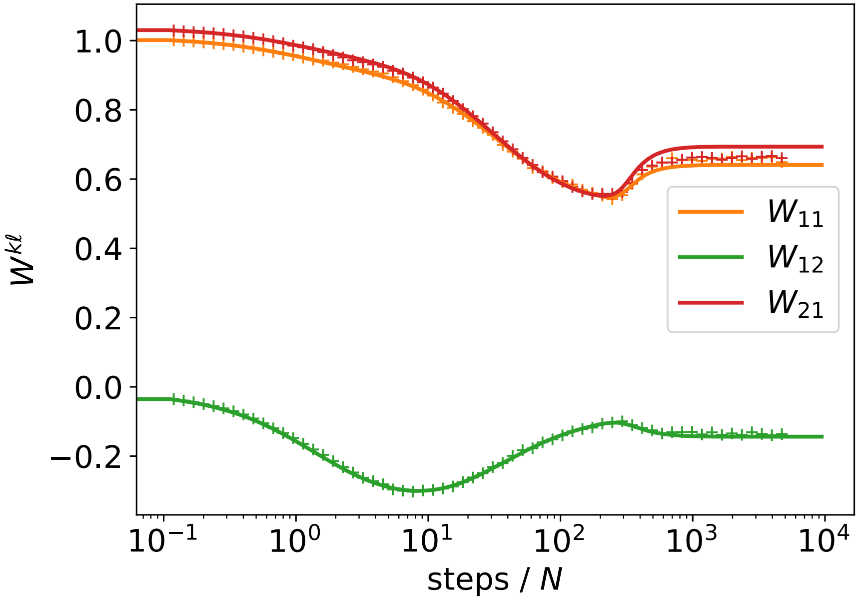


$$S_r^k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^k F_{ir}$$

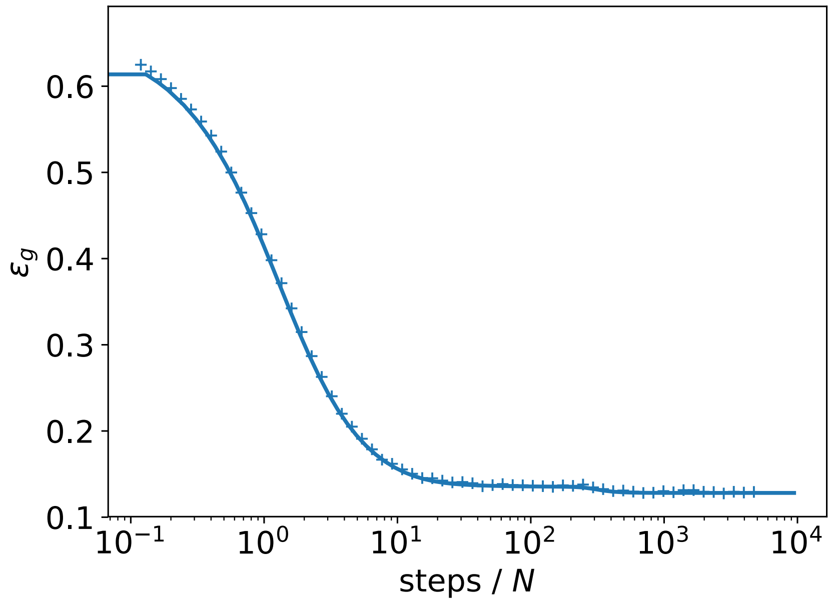
$$W^{k\ell} = \frac{1}{N} \sum_i w_i^k w_i^\ell$$

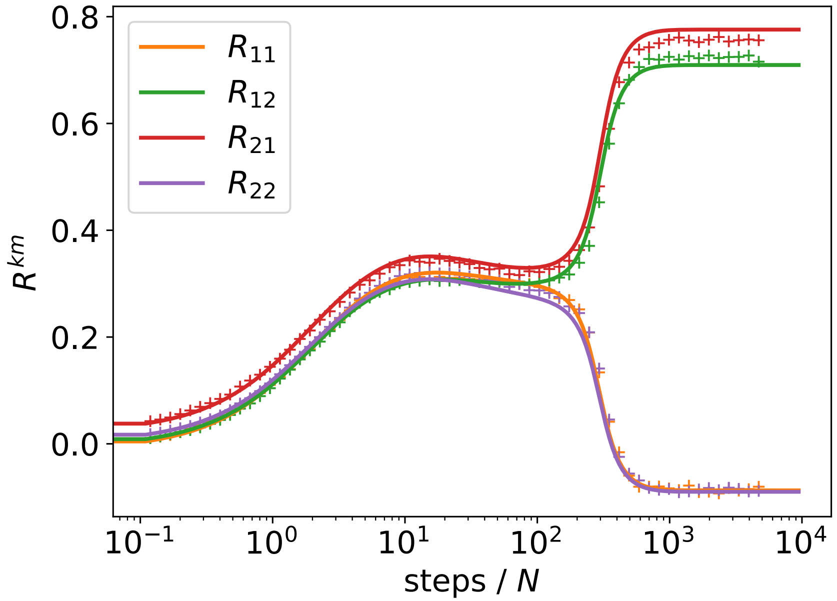
$$\Sigma^{k\ell} = \frac{1}{D} \sum_{r=1}^{D} S_r^k S_r^\ell$$





ODE Theory vs simulations N=8000, D=4000, M=2, K=2





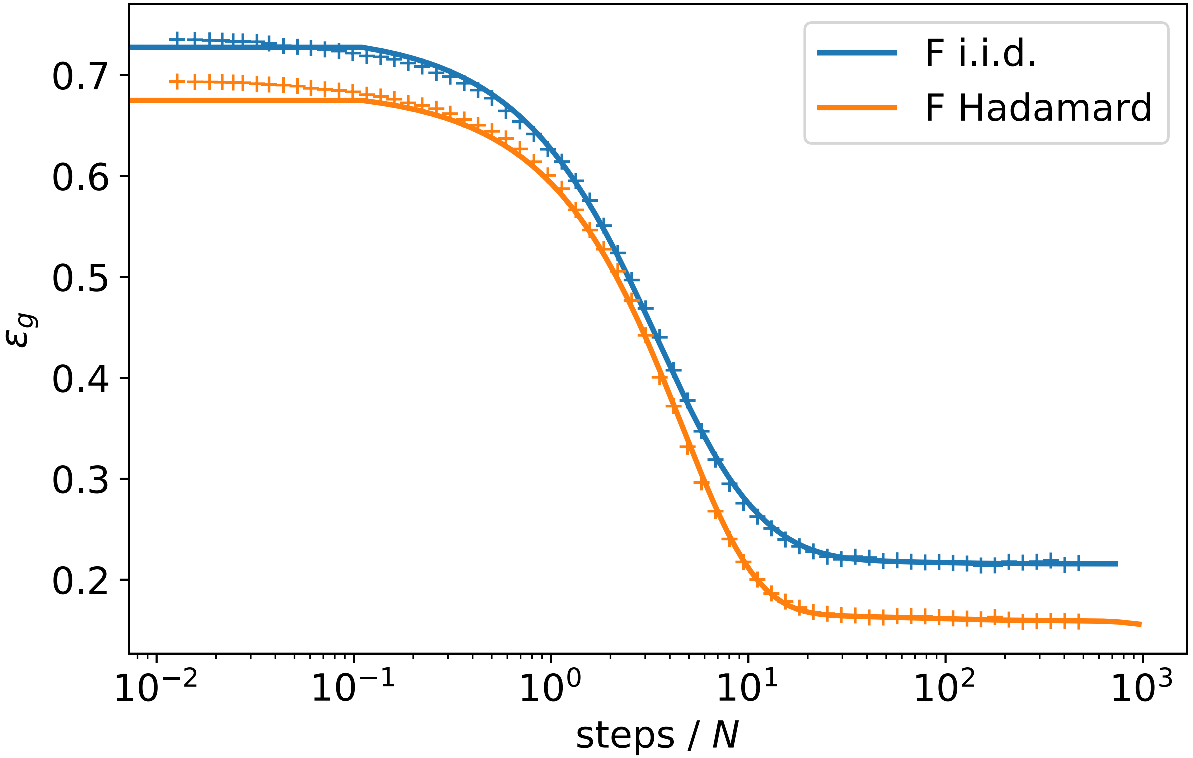
$$S_r^k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^k F_{ir} \qquad R^{km} = \frac{b}{D} \sum_{r=1}^{D} S_r^k \tilde{w}_r^m$$

correlation of pre-activation of neuron k in the student and the weight m in the latent task

specializes after 100 steps

ODE Theory vs simulations N=1023, D=1023, M=2, K=2

Hadamard F



Phase diagram of Hidden Manifold Model

Gardner's computation: volume of space in w_i^k compatible with the data $\left\{\vec{X}_\mu, \Phi_t(\vec{x}_\mu)\right\}$

Evaluated with replicas

The volume can be written in terms of the local input fields to the hidden variables, λ_μ^{ka} .

The GET shows that these are Gaussian variables, independent for different patterns, correlated for one given pattern. Finite number of correlations between nk variables, so the computation can be done.

Results… coming soon (Federica Gerace, Bruno Loureiro, Florent Krzakala, Lenka Zdeborova , MM, in preparation).

Summary

Data structure is important

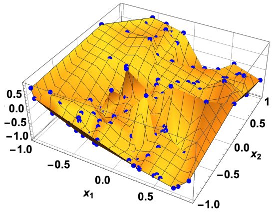
- Hidden manifolds and sub manifolds
- Combinatorial structure

Hidden Manifold Model

Data has « Latent representation »: $\{C_r\}$

Desired output (task) = function of latent representation

Example $y = g\left(\sum_{r=1}^{R} \tilde{w}_r C_r\right)$ $\vec{X} = f\left[\frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r \vec{F}_r\right]$



- Does not have the pathologies of teacher-student setup with iid data
- Learning and generalization phenomenology $\sim$ MNIST
- Can be studied analytically : online learning and full batch in the limit where $R = O(N)$, thanks to a Gaussian Equivalence property

Statistical inference and statistical physics

Infer a hidden rule, or hidden variables, from data. Many variables, big data chosen from an *ensemble* → stat. physics

Physics approach:

- mean-field cavity equations → efficient algorithm
- replica method → phase diagram, and control of algorithm
- frequent pattern of phase diagram:

Impossible

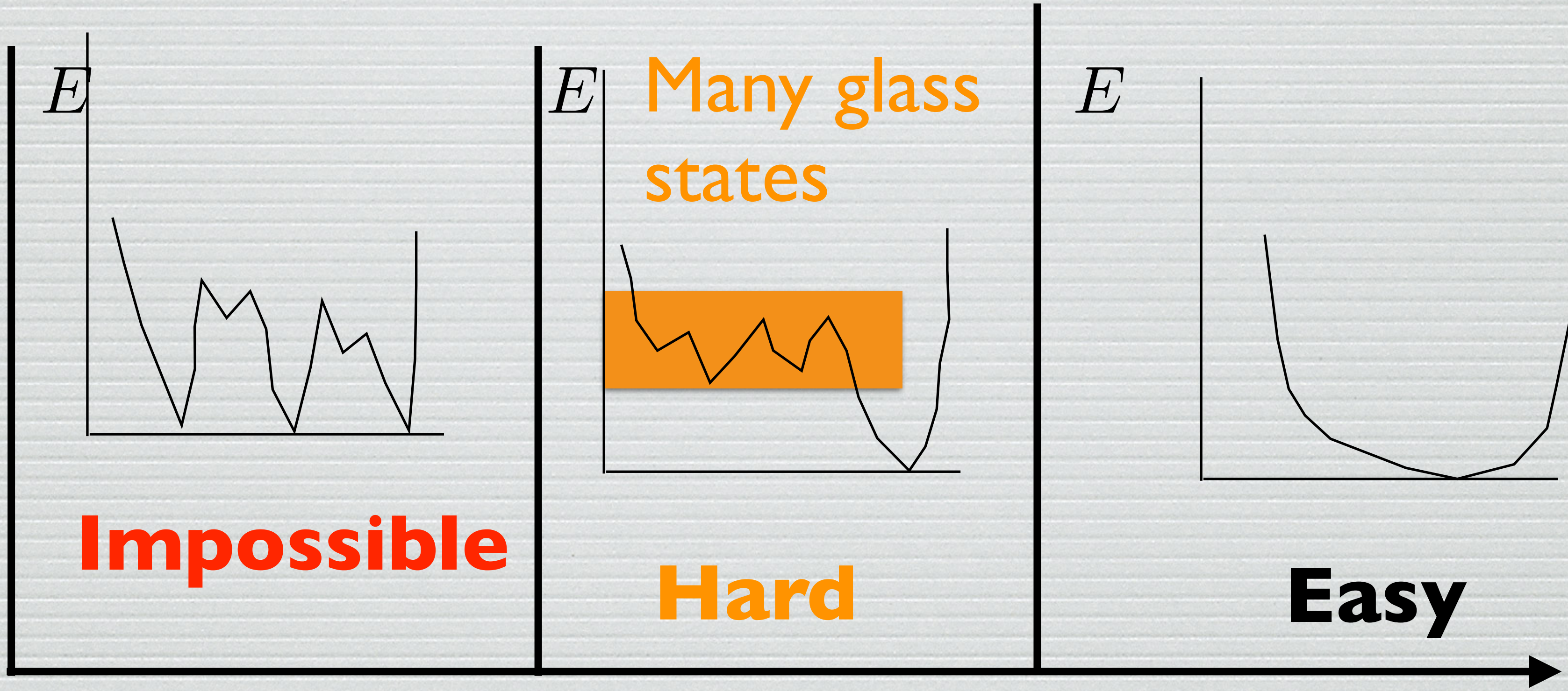
Hard

Easy

Info. theory threshold

Algorithmic threshold

Amount of data



Many glass states

Impossible

Hard

Easy

Impossible

Hard

Easy

Info. theory threshold

Algorithmic threshold

Amount of data

Statistical inference and statistical physics

Infer a hidden rule, or hidden variables, from data. Many variables, big data chosen from an *ensemble* → stat. physics

Physics approach:

- mean-field cavity equations → efficient algorithm
- replica method → phase diagram, and control of algorithm
- frequent pattern of phase diagram

Relevance for machine learning: data and task structure is probably crucial. Define *new ensembles*, like eg the Hidden Manifold Model