Ensembles and Structured Data
in
Physics and Inference

Marc Mézard
Ecole normale supérieure - PSL University

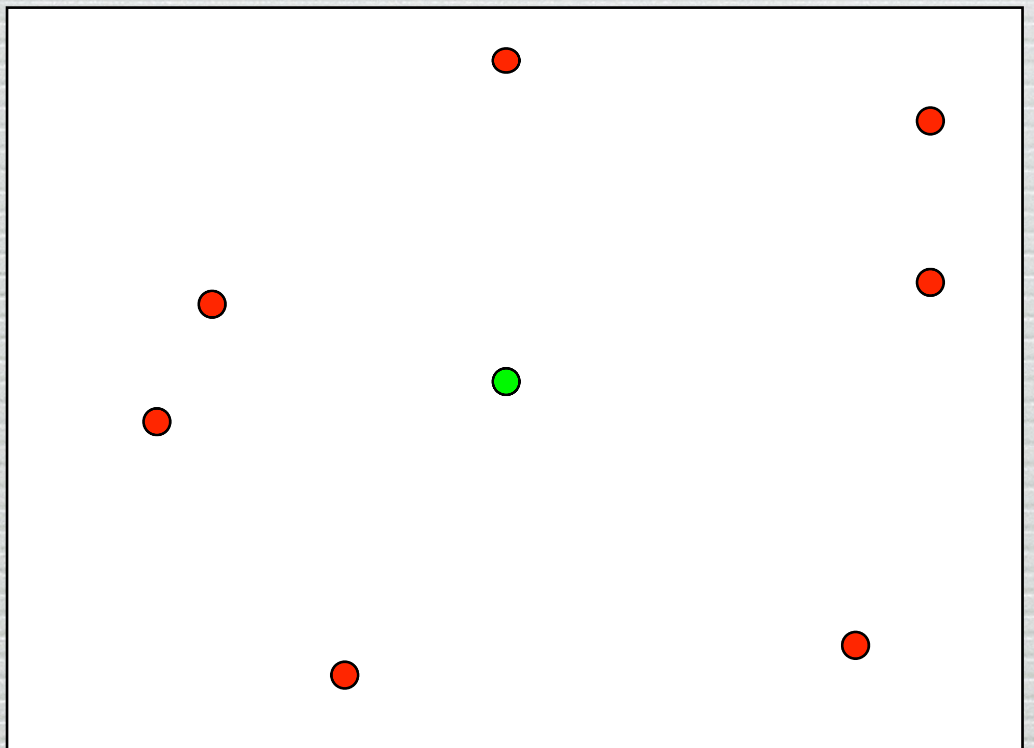
Bangalore, Turing lecture 3
January 8, 2020

Chapter One

Ensembles: 3 examples

Codes: Shannon code ensemble

Unit hypercube
in N dimensions



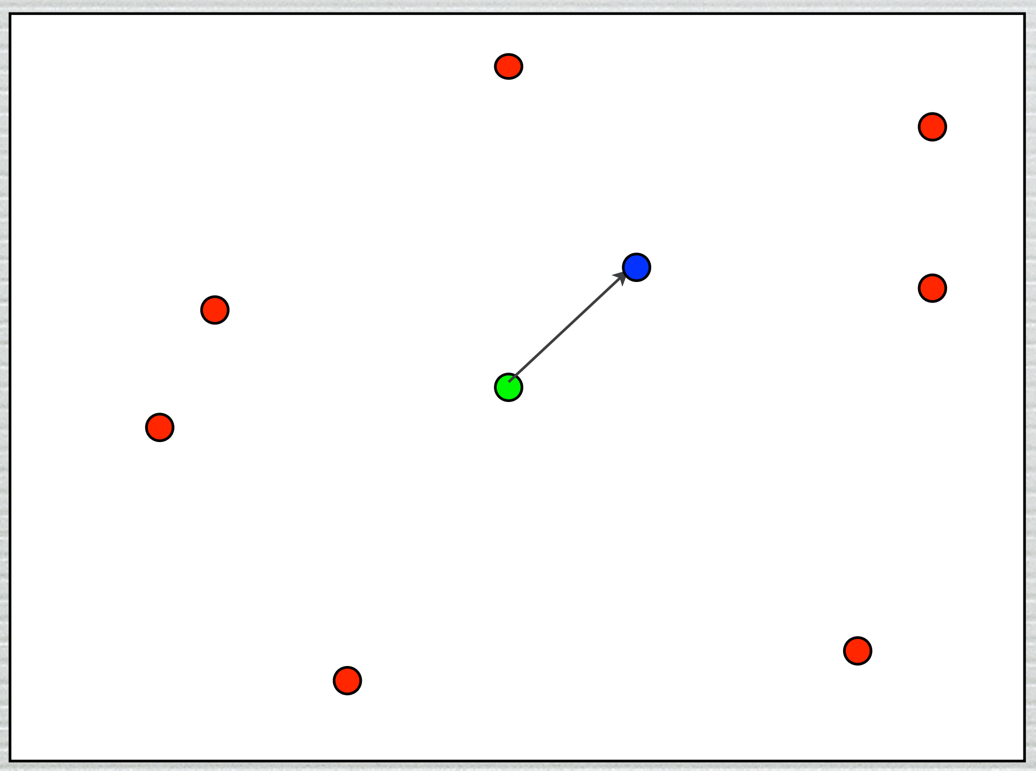
codewords
(random)

sent codeword

2^{RN} iid random points, uniform distribution

Codes: Shannon code ensemble

Unit hypercube
in N dimensions



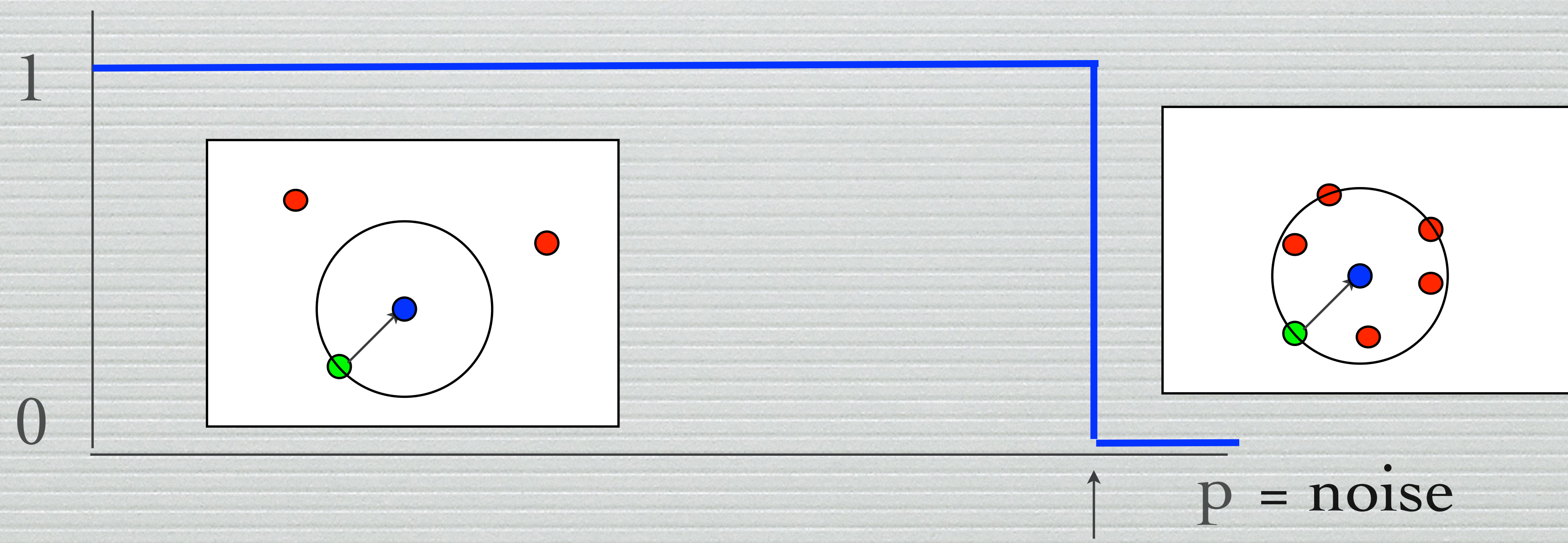
- codewords (random)
- sent codeword
- received word

2^{RN} iid random points, uniform distribution

Phase transitions in decoding

Decoding = find closest codeword

Probability of perfect decoding:



Shannon « bound »

geometric phase transition

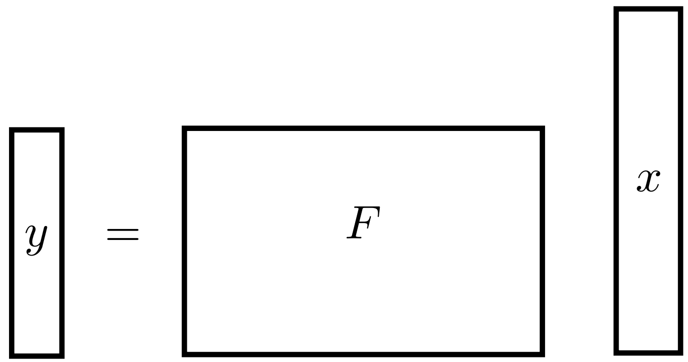
Phase diagram for compressed sensing

N Unknown x

M Linear measurements $y = Fx$

$M < N$

$\alpha = M/N$



x sparse: ρN non-zero components

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

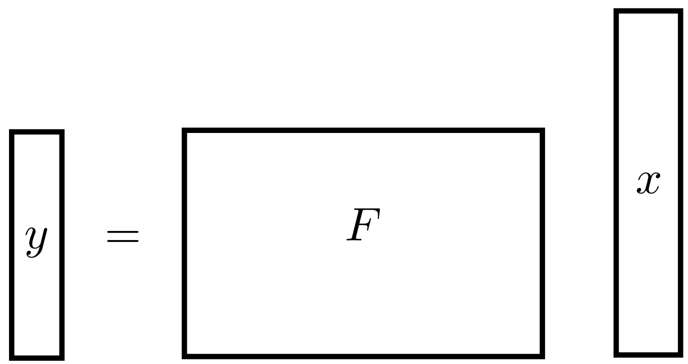
Phase diagram for compressed sensing

N Unknown x

M Linear measurements $y = Fx$

$M < N$

$\alpha = M/N$



x sparse: ρN non-zero components

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 $F \sim \mathcal{N}(0, 1/N)$

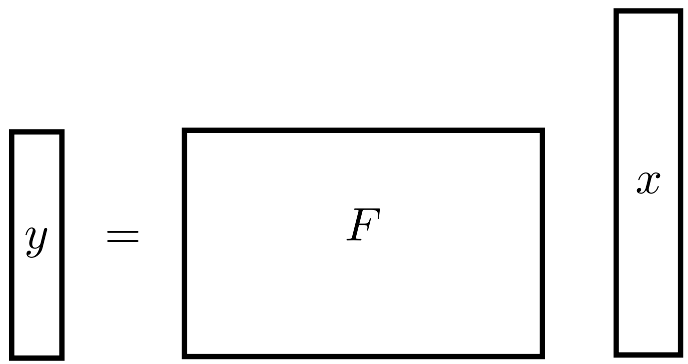
Phase diagram for compressed sensing

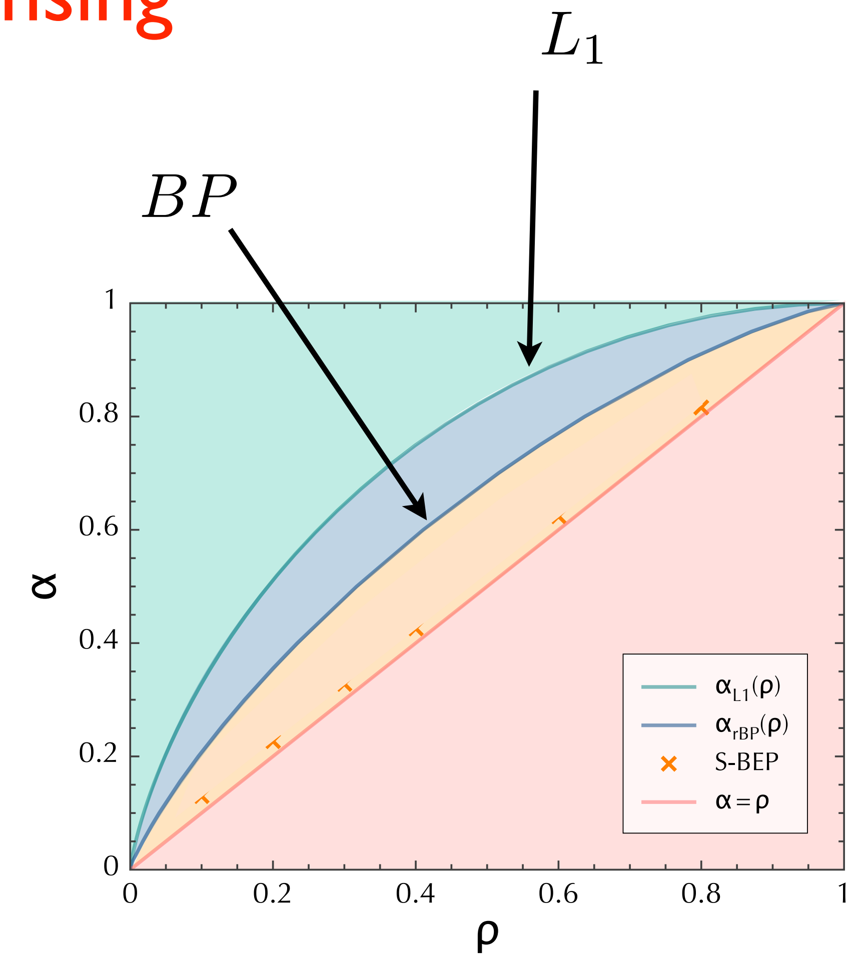
N Unknown x

M Linear measurements $y = Fx$

$M < N$

$\alpha = M/N$





x sparse: ρN non-zero components

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 $F \sim \mathcal{N}(0, 1/N)$

Analytic study: cavity equations, density evolution, replicas, state evolution

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)

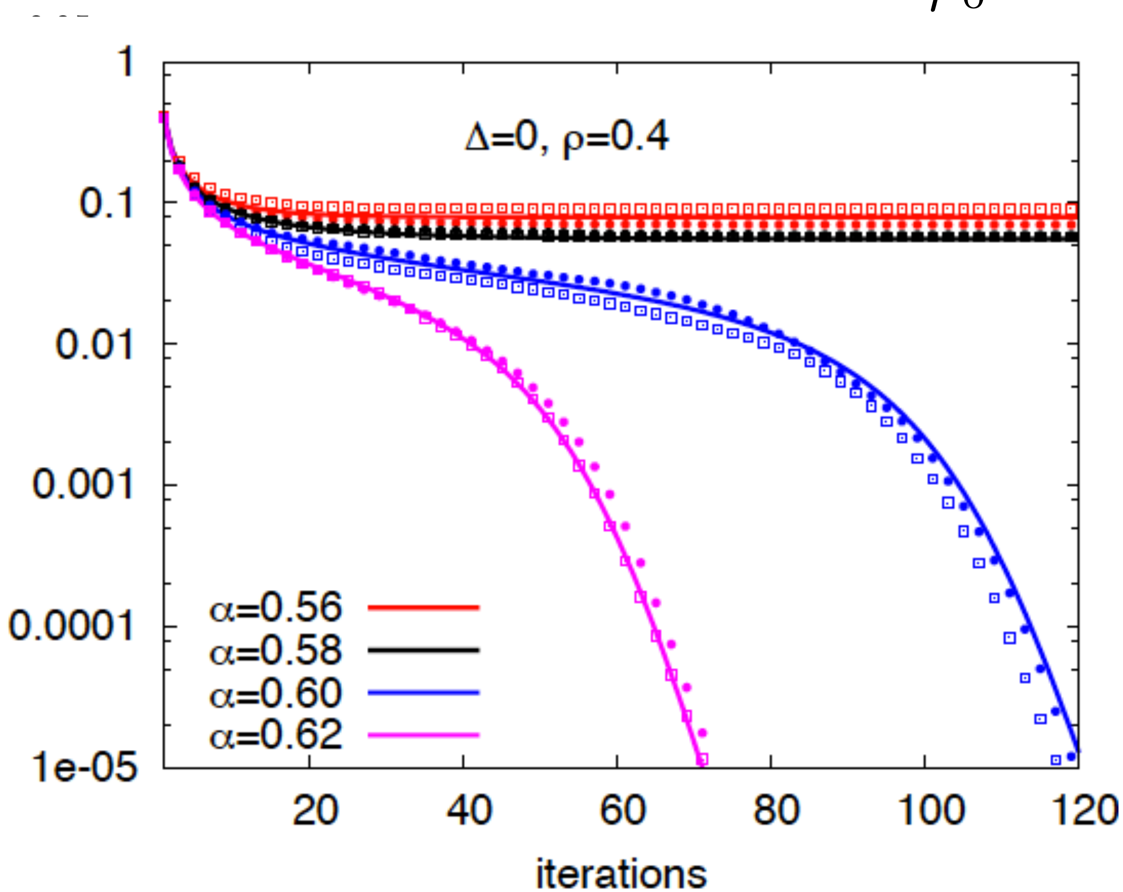
→ analytic control of the BP equations

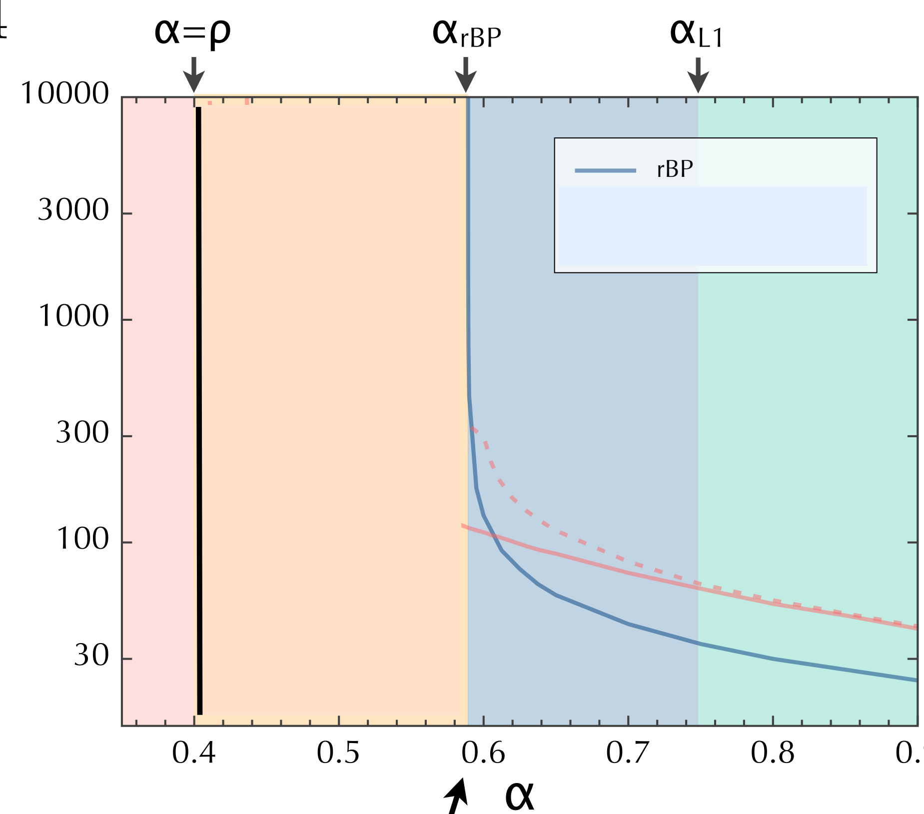
NB rigorous: Bayati Montanari, Lelarge Montanari

Error

BP convergence time

$\rho_0 = .4$

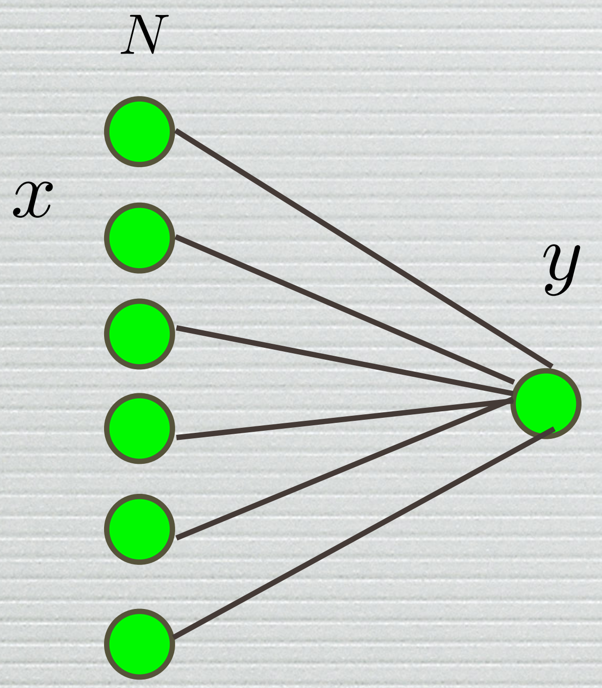




Dynamic glass transition

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

Perceptron learning in the '90s…



Task to be learnt= teacher perceptron

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

Learning= student perceptron

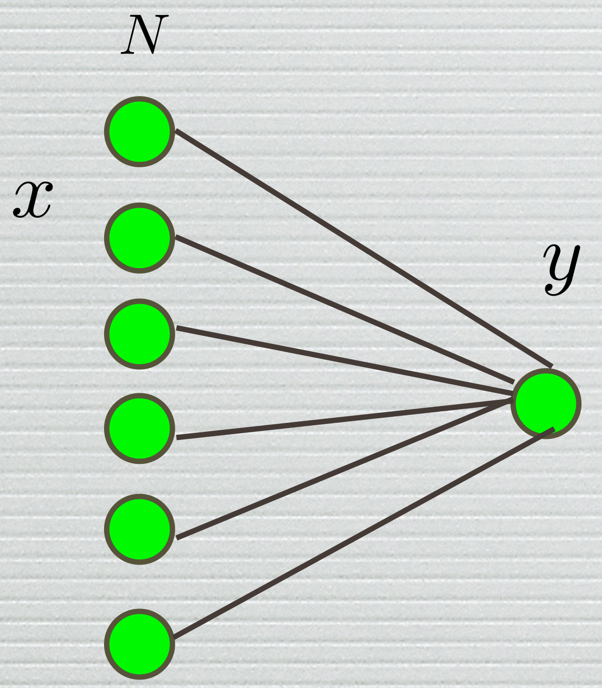
$$y = Sign(K.x) \qquad K_i = \pm 1$$

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

Perceptron learning in the '90s...



Task to be learnt= teacher perceptron

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

Learning= student perceptron

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

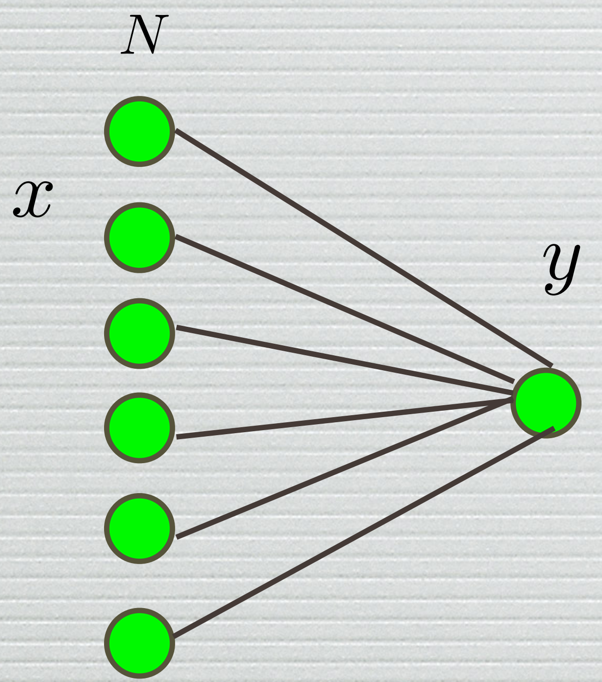
Machine learning: database of P examples x^μ

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 $\sim \mathcal{N}(0, 1/N)$

Perceptron learning in the '90s…



Task to be learnt= teacher perceptron

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

Learning= student perceptron

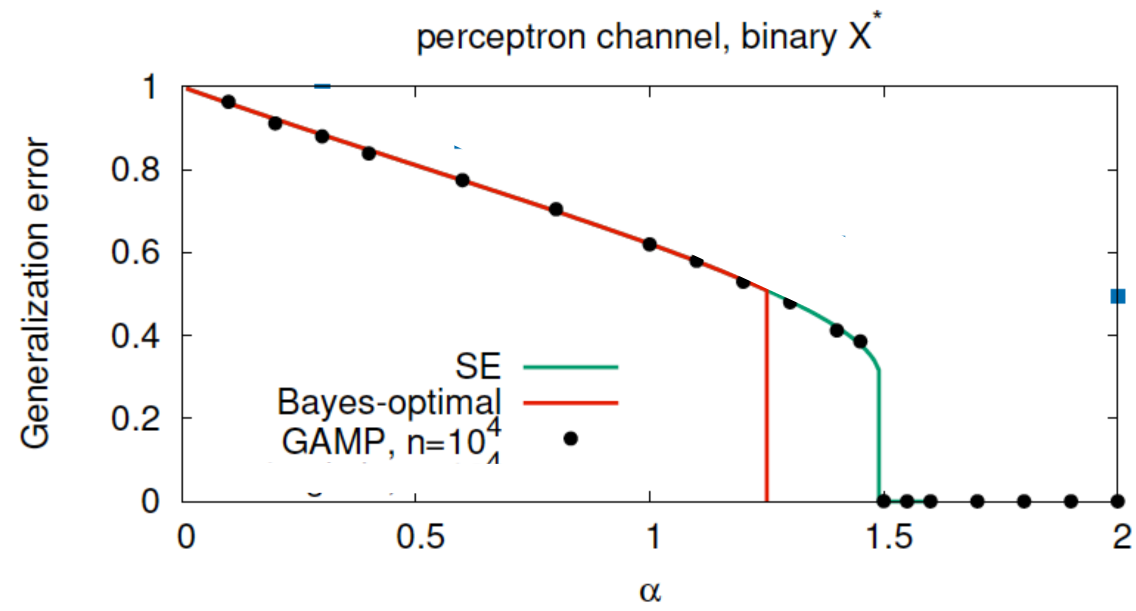
$$y = Sign(K.x) \qquad K_i = \pm 1$$

Thermodynamic limit

$$N, P \to \infty$$
$$\alpha = P/N$$

Transition to perfect generalization

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?

Chapter Two

A first (simple) example:

Hopfield Model
with
Structured Patterns

Hopfield model with structured patterns

Mezard 2017

Neurons = N binary spins: $s_i \in \{\pm 1\}$

Patterns to be memorized

$$\xi_i^\mu = \pm 1,\ i \in \{1, \ldots n\},\ \mu \in \{1, \ldots p\}\ ,$$

$$E = -\frac{1}{2}\sum_{i,j} J_{ij} s_i s_j \qquad J_{ij} = \frac{1}{N}\sum_\mu \xi_i^\mu \xi_j^\mu$$

$$P_J(s) = \frac{1}{Z} e^{(\beta/2)\sum_{i,j} J_{ij} s_i s_j} \qquad Z = \sum_s e^{(\beta/2)\sum_{i,j} J_{ij} s_i s_j}$$

A spin glass, like the SK spin glass, except for the choice of J_{ij}

Naive TAP equations

$$H_i = \sum_k J_{ki} \tanh(\beta H_k) - \beta \tanh(\beta H_i) \sum_k J_{ki}^2 [1 - \tanh^2(\beta H_k)]$$

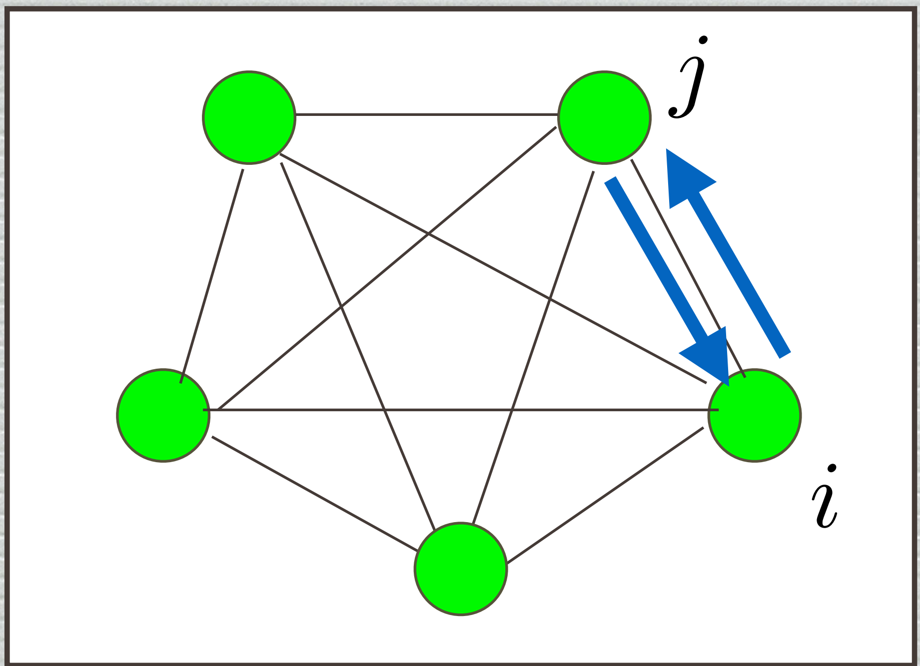
Disordered and infinite range

Naive TAP equations

$$H_i = \sum_k J_{ki} \tanh(\beta H_k) - \beta \tanh(\beta H_i) \sum_k J_{ki}^2 [1 - \tanh^2(\beta H_k)]$$

Disordered and infinite range

WRONG



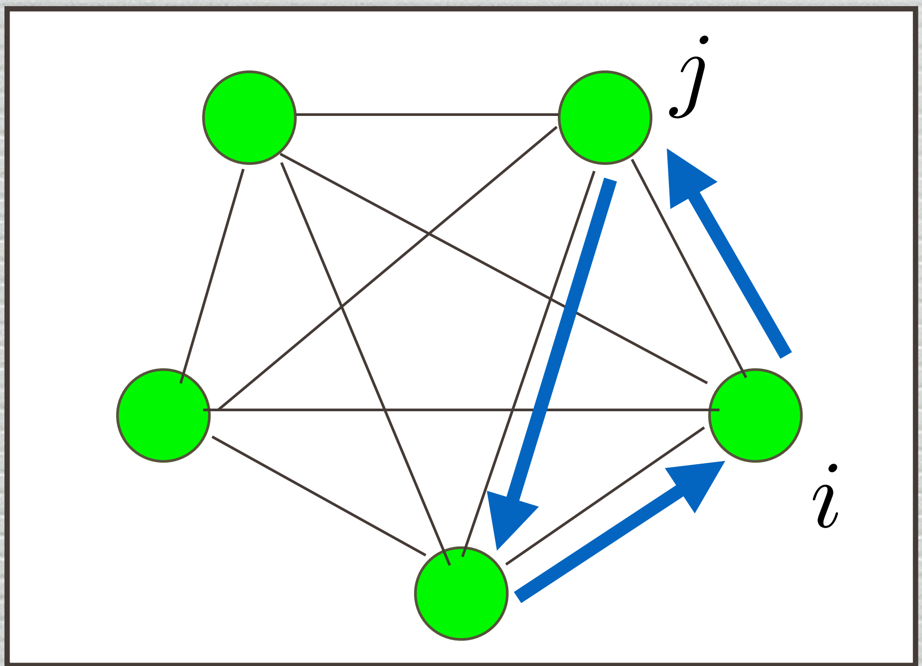
TAP is valid only if indirect interaction from i to j through other sites can be neglected

TAP in the Hopfield model: more subtle!

$$J_{ij} = \frac{1}{N} \sum_{\mu} \xi_i^\mu \xi_j^\mu$$

$$\overline{J_{ij} J_{jk} J_{ki}} \neq 0$$

Indirect interactions matter
« Naive » TAP does not apply



The mean-field equations depend on the quenched disorder ensemble

The Hopfield model as a Restricted Boltzmann Machine

$$Z = \sum_s e^{(\beta/2)\sum_{i,j} J_{ij} s_i s_j} \qquad J_{ij} = \frac{1}{N}\sum_\mu \xi_i^\mu \xi_j^\mu$$

$$Z = \sum_s \exp\left(\frac{\beta}{2N}\sum_\mu \left[\sum_i \xi_i^\mu s_i\right]^2\right)$$

Hubbard Stratonovitch (Gaussian transform) :

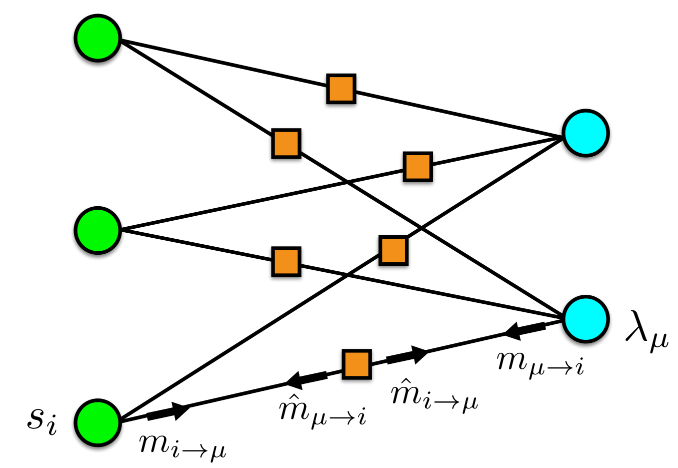
$$Z = \sum_s \int \prod_\mu \frac{d\lambda_\mu}{\sqrt{2\pi\beta}} \; \exp\left[-\frac{\beta}{2}\sum_\mu \lambda_\mu^2 + \beta \sum_{\mu,i} \frac{\xi_i^\mu}{\sqrt{N}} s_i \lambda_\mu\right]$$

$$Z = \sum_s \int \prod_\mu \frac{d\lambda_\mu}{\sqrt{2\pi\beta}} \ \exp\left[-\frac{\beta}{2}\sum_\mu \lambda_\mu^2 + \beta \sum_{\mu,i} \frac{\xi_i^\mu}{\sqrt{N}} s_i \lambda_\mu \right]$$

Spin-variable Pattern-variable Coupling, iid

Hopfield model is a restricted Boltzmann machine, with a specific set of couplings $\frac{\xi_i^\mu}{\sqrt{N}}$ that store P patterns.

iid couplings



TAP-AMP equations in the paramagnetic or SG phase

$$H_i \simeq \sum_\nu \frac{\xi_i^\nu}{\sqrt{N}} A_\nu - \frac{\alpha}{1-\beta(1-q)} \tanh(\beta H_i)$$

$$A_\mu = \frac{1}{\sqrt{N}} \sum_j \xi_j^\mu \tanh(\beta H_j)$$

$$q = \frac{1}{N} \sum_i \tanh^2(\beta H_i)$$

First written in MPV 1987, claimed wrong in Nakanishi-Takayama 1997, Shamir Sompolinsky 2000, actually correct.

Can be used as an algorithm (with right time indices)

More general spin glass ensembles

Neurons = N binary spins: $s_i \in \{\pm 1\}$

$$E = -\frac{1}{2}\sum_{i,j} J_{ij} s_i s_j$$

$$Z = \sum_s e^{(\beta/2)\sum_{i,j} J_{ij} s_i s_j}$$

$$P_J(s) = \frac{1}{Z} e^{(\beta/2)\sum_{i,j} J_{ij} s_i s_j}$$

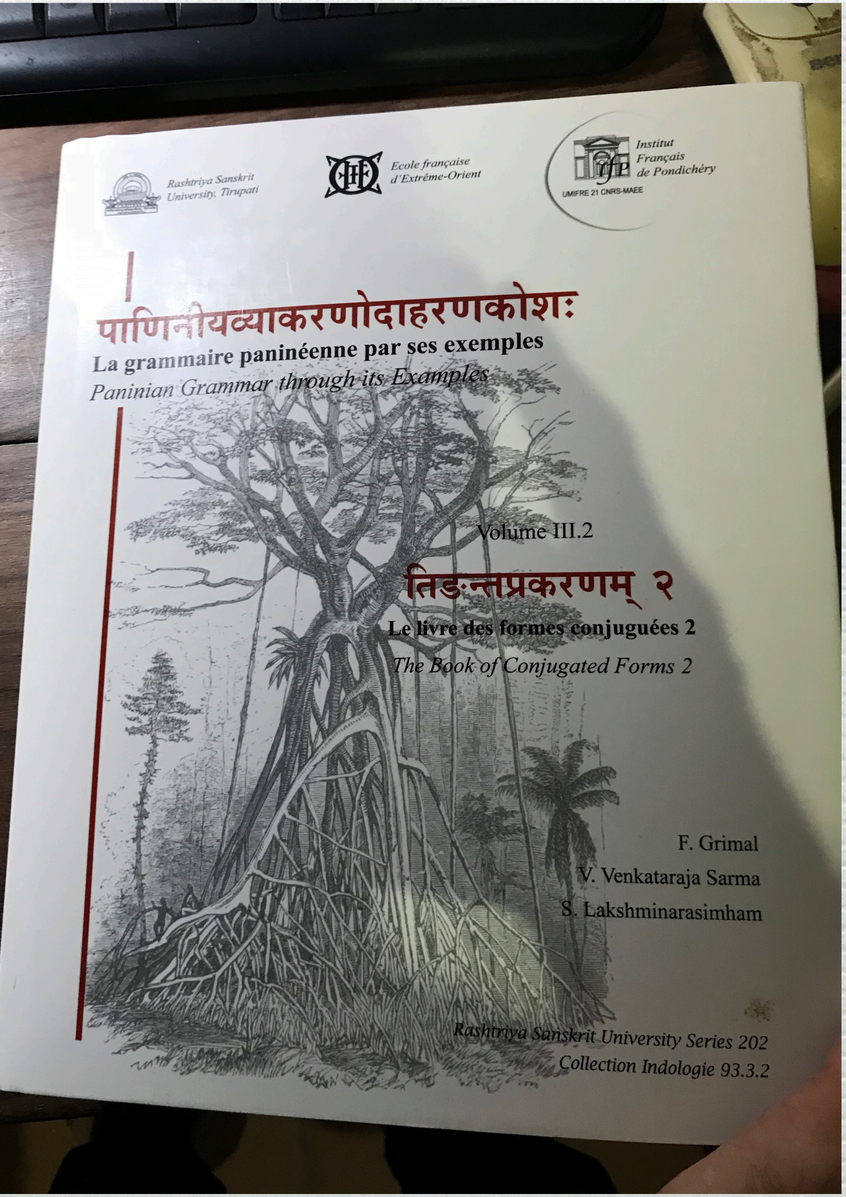
J iid : SK model

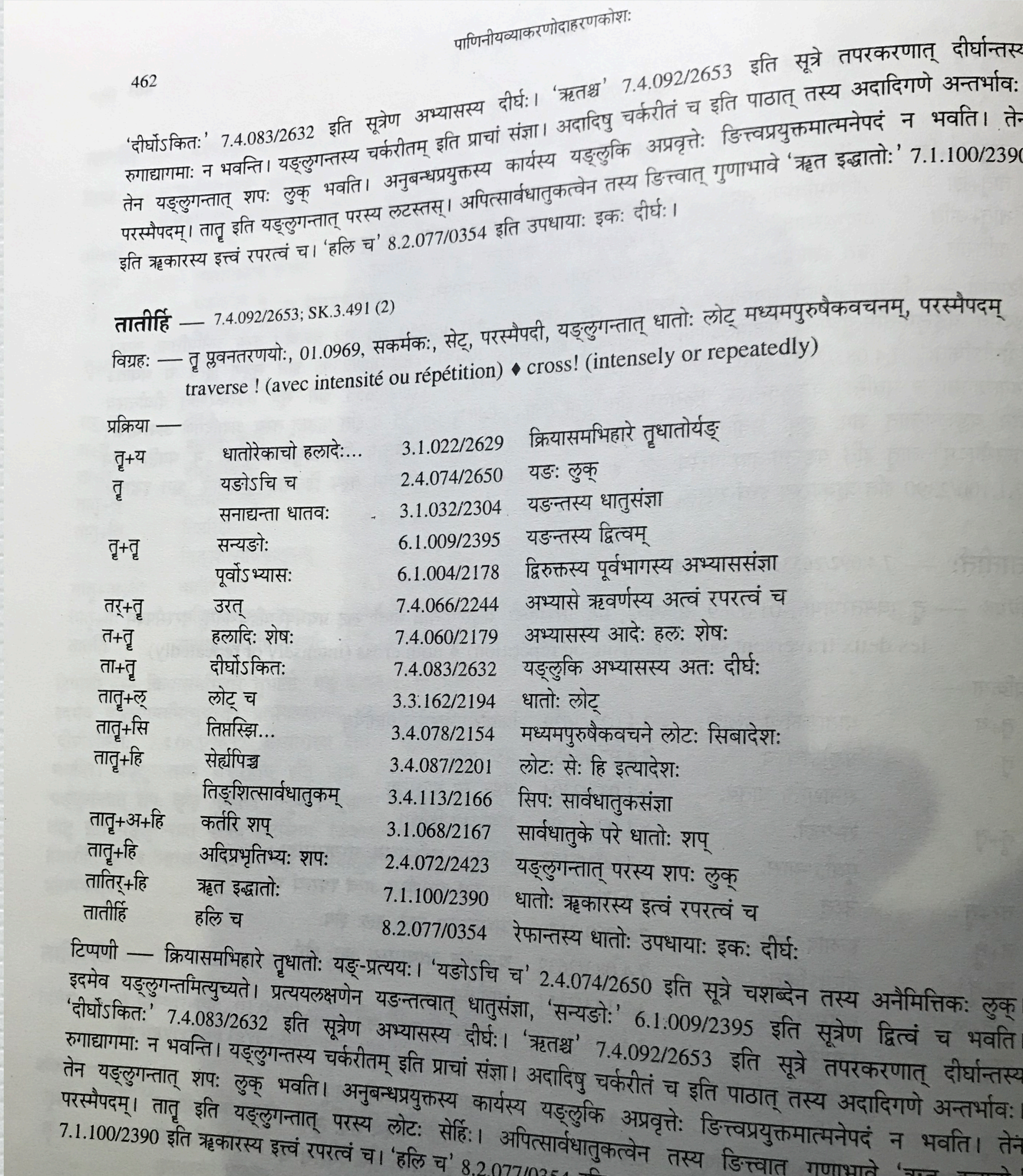
$$J_{ij} = \frac{1}{N}\sum_\mu \xi_i^\mu \xi_j^\mu$$ Hopfield

$$J = ODO^{-1}$$

Random orthogonal (Parisi Potters, Kabashima, ...)

Structured patterns: combinatorial disorder





462 पाणिनीयव्याकरणोदाहरणकोशः

'दीर्घोऽकितः' 7.4.083/2632 इति सूत्रेण अभ्यासस्य दीर्घः। 'ऋतश्च' 7.4.092/2653 इति सूत्रे तपरकरणात् दीर्घान्तस्य रुगाद्यागमाः न भवन्ति। यङ्लुगन्तस्य चर्करीतम् इति प्राचां संज्ञा। अदादिषु चर्करीतं च इति पाठात् तस्य अदादिगणे अन्तर्भावः तेन यङ्लुगन्तात् शपः लुक् भवति। अनुबन्धप्रयुक्तस्य कार्यस्य यङ्लुकि अप्रवृत्तेः ङित्त्वप्रयुक्तमात्मनेपदं न भवति। तेन परस्मैपदम्। तातॄ इति यङ्लुगन्तात् परस्य लटस्तस्। अपित्सार्वधातुकत्वेन तस्य ङित्त्वात् गुणाभावे 'ॠत इद्धातोः' 7.1.100/2390 इति ॠकारस्य इत्त्वं रपरत्वं च। 'हलि च' 8.2.077/0354 इति उपधायाः इकः दीर्घः।

तातीर्हि — 7.4.092/2653; SK.3.491 (2)

यङ्लुगन्तात् धातोः लोट् मध्यमपुरुषैकवचनम्, परस्मैपदम्

विग्रहः — तॄ प्लवनतरणयोः, 01.0969, सकर्मकः, सेट्, परस्मैपदी, traverse ! (avec intensité ou répétition) ♦ cross! (intensely or repeatedly)

प्रक्रिया —

तॄ+य	धातोरेकाचो हलादेः...	3.1.022/2629	क्रियासमभिहारे तॄधातोर्यङ्
तॄ	यङोऽचि च	2.4.074/2650	यङः लुक्
	सनाद्यन्ता धातवः	3.1.032/2304	यङन्तस्य धातुसंज्ञा
तॄ+तॄ	सन्यङोः	6.1.009/2395	यङन्तस्य द्वित्वम्
	पूर्वोऽभ्यासः	6.1.004/2178	द्विरुक्तस्य पूर्वभागस्य अभ्याससंज्ञा
तर्+तॄ	उरत्	7.4.066/2244	अभ्यासे ऋवर्णस्य अत्वं रपरत्वं च
त+तॄ	हलादिः शेषः	7.4.060/2179	अभ्यासस्य आदेः हलः शेषः
ता+तॄ	दीर्घोऽकितः	7.4.083/2632	यङ्लुकि अभ्यासस्य अतः दीर्घः
तातॄ+ल्	लोट् च	3.3.162/2194	धातोः लोट्
तातॄ+सि	तिप्तस्झि...	3.4.078/2154	मध्यमपुरुषैकवचने लोटः सिबादेशः
तातॄ+हि	सेर्ह्यपिच्च	3.4.087/2201	लोटः सेः हि इत्यादेशः
	तिङ्शित्सार्वधातुकम्	3.4.113/2166	सिपः सार्वधातुकसंज्ञा
तातॄ+अ+हि	कर्तरि शप्	3.1.068/2167	सार्वधातुके परे धातोः शप्
तातॄ+हि	अदिप्रभृतिभ्यः शपः	2.4.072/2423	यङ्लुगन्तात् परस्य शपः लुक्
तातिर्+हि	ॠत इद्धातोः	7.1.100/2390	धातोः ॠकारस्य इत्वं रपरत्वं च
तातीर्हि	हलि च	8.2.077/0354	रेफान्तस्य धातोः उपधायाः इकः दीर्घः

टिप्पणी — क्रियासमभिहारे तॄधातोः यङ्-प्रत्ययः। 'यङोऽचि च' 2.4.074/2650 इति सूत्रे चशब्देन तस्य अनैमित्तिकः लुक् इदमेव यङ्लुगन्तमित्युच्यते। प्रत्ययलक्षणेन यङन्तत्वात् धातुसंज्ञा, 'सन्यङोः' 6.1.009/2395 इति सूत्रेण द्वित्वं च भवति 'दीर्घोऽकितः' 7.4.083/2632 इति सूत्रेण अभ्यासस्य दीर्घः। 'ऋतश्च' 7.4.092/2653 इति सूत्रे तपरकरणात् दीर्घान्तस्य रुगाद्यागमाः न भवन्ति। यङ्लुगन्तस्य चर्करीतम् इति प्राचां संज्ञा। अदादिषु चर्करीतं च इति पाठात् तस्य अदादिगणे अन्तर्भावः तेन यङ्लुगन्तात् शपः लुक् भवति। अनुबन्धप्रयुक्तस्य कार्यस्य यङ्लुकि अप्रवृत्तेः ङित्त्वप्रयुक्तमात्मनेपदं न भवति। तेन परस्मैपदम्। तातॄ इति यङ्लुगन्तात् परस्य लोटः सेर्हिः। अपित्सार्वधातुकत्वेन तस्य ङित्त्वात् गुणाभावे '...' 7.1.100/2390 इति ॠकारस्य इत्त्वं रपरत्वं च। 'हलि च' 8.2.077/0354 ...

Structured patterns: combinatorial disorder

Modified Hopfield model: Combinatorial patterns

$$\vec{\xi}^\mu = (\xi_1^\mu, \cdots, \xi_N^\mu)$$

$\vec{\xi}^\mu$ built from superposition of elementary features $\vec{u}^r$

$$\boxed{\vec{\xi}^\mu = \frac{1}{\sqrt{\gamma N}} \sum_r v_r^\mu \, \vec{u}^r}$$

, binary $v_r^\mu \in \{\pm 1\}$

« **Model of the world** »: hierarchy of combinatorial compositions

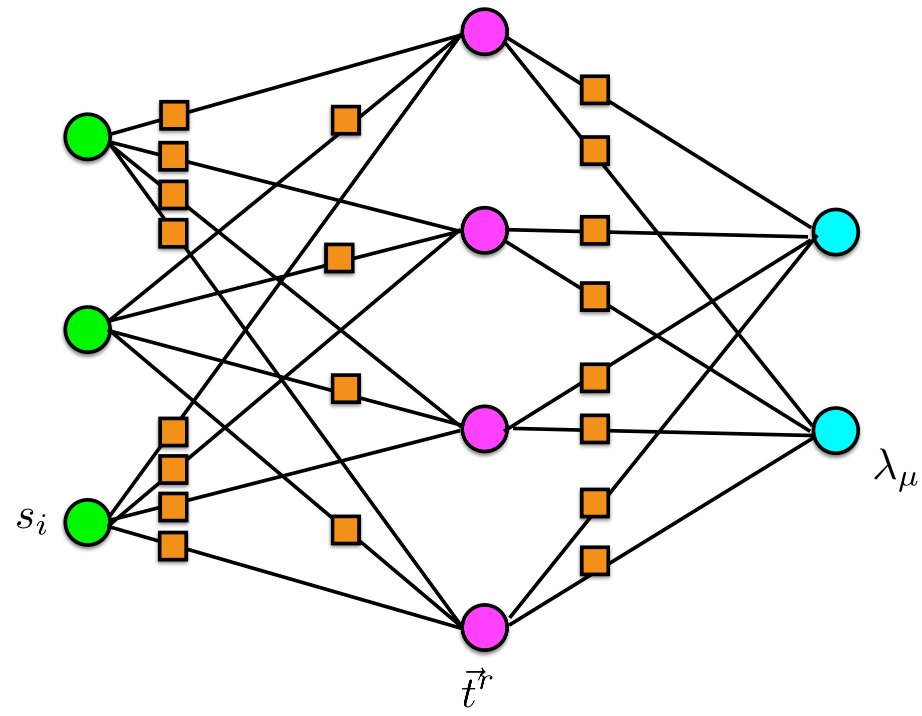
NB: linear manifold. One would like a non-linear one. See later

TAP equations in the Hopfield model with structured patterns

Modified Hopfield model: Combinatorial patterns

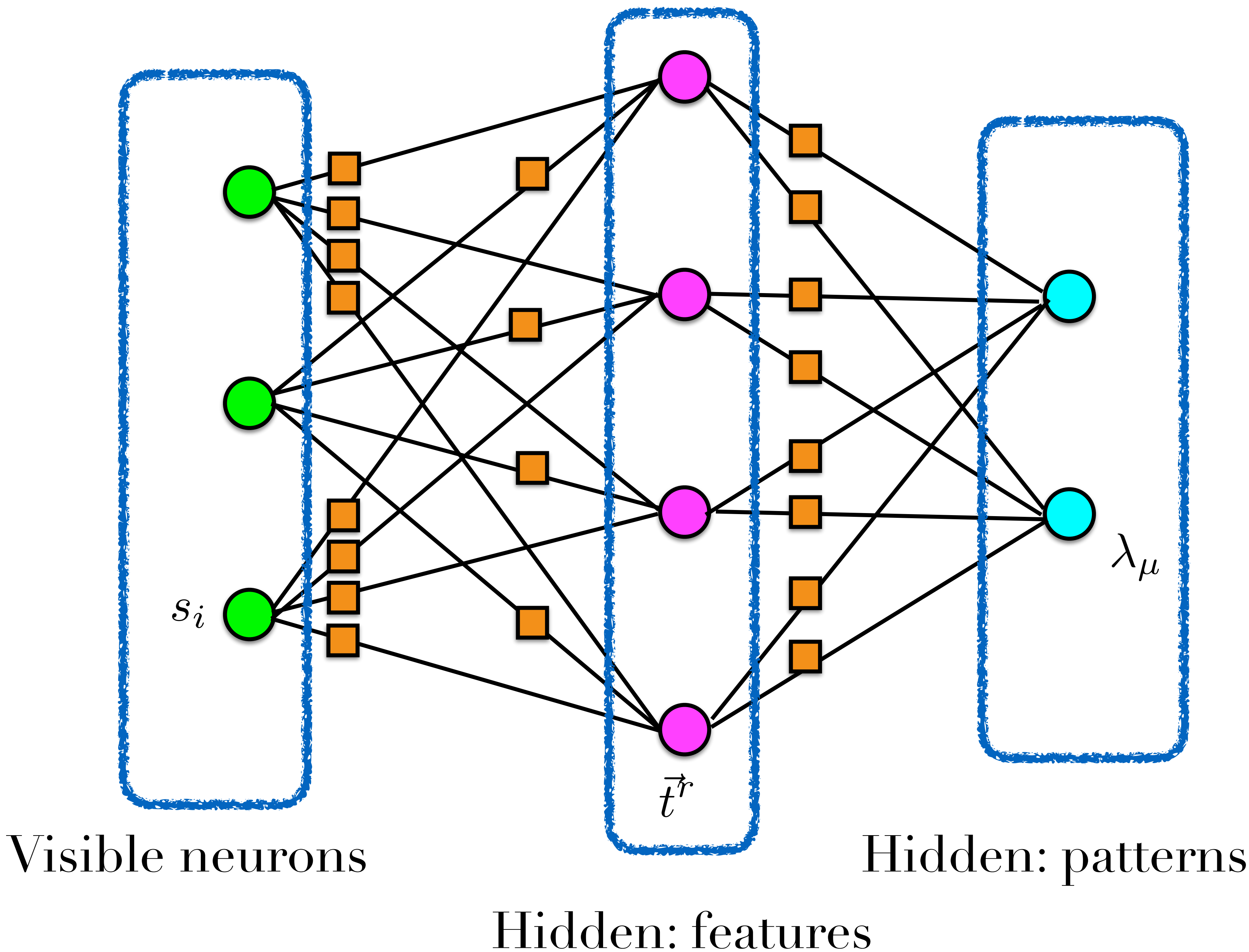
$$Z = \sum_s \int \prod_\mu \frac{d\lambda_\mu e^{-\beta\lambda_\mu^2/2}}{\sqrt{2\pi\beta}} \ \exp\left[\frac{\beta}{\sqrt{\gamma}} \sum_{r=1}^{\gamma N} \left(\frac{1}{\sqrt{N}} \sum_i u_i^r s_i\right) \left(\frac{1}{\sqrt{N}} \sum_\mu v_\mu^r \lambda_\mu\right)\right]$$

Disentangle the last term by another Hubbard Stratonovitch representation

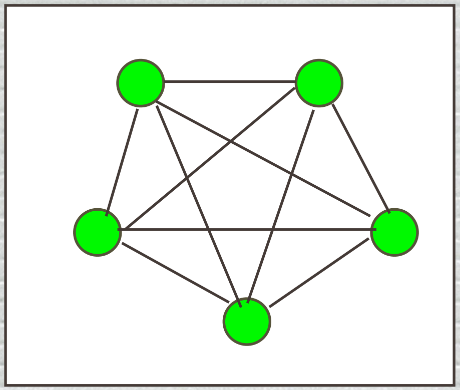


$$Z = \sum_s \int \prod_\mu d\lambda_\mu \int \prod d\vec{t}^r \exp\left[-\frac{\beta}{2}\sum_\mu \lambda_\mu^2 + \beta \sum_{r=1}^{\gamma N}\left(+\frac{1}{\sqrt{\gamma}}U^r V^r - \hat{U}^r U^r - \hat{V}^r V^r\right)\right]$$

$$\exp\left[\frac{\beta}{\sqrt{N}}\sum_{r=1}^{\gamma N}\sum_{i=1}^{N}\hat{U}^r u_i^r s_i + \frac{\beta}{\sqrt{N}}\sum_{r=1}^{\gamma N}\sum_{\mu=1}^{\alpha N}\hat{V}^r v_\mu^r \lambda_\mu + \frac{\beta}{\sqrt{\gamma}}\sum_{r=1}^{\gamma N} U^r V^r\right]$$

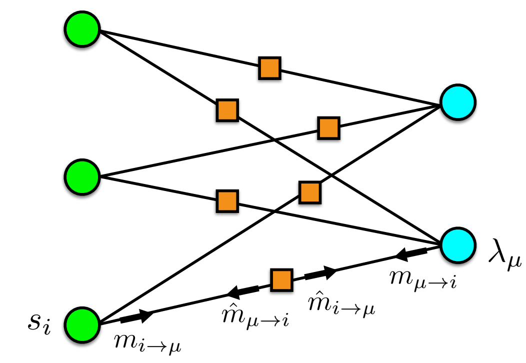

s_i
$\vec{t}^r$
λ_μ
Visible neurons
Hidden: patterns
Hidden: features

TAP equations in the Hopfield model with structured patterns (MM, Phys Rev E 2016)



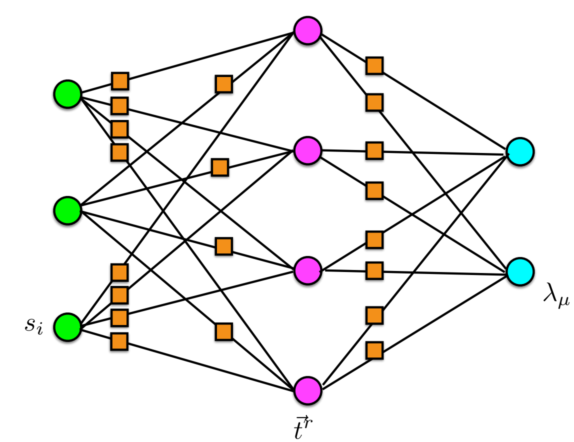
Hopfield model

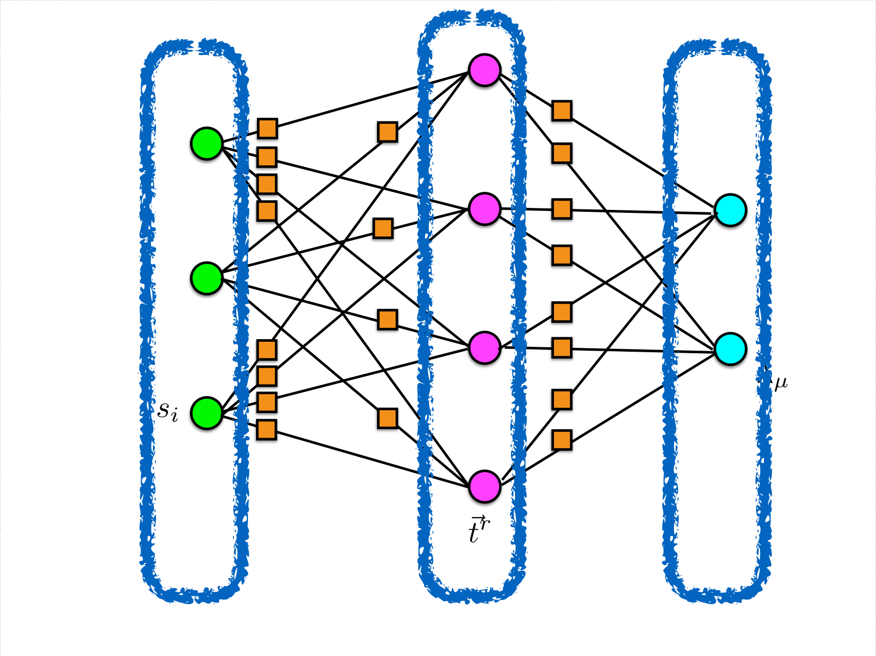
Restricted Boltzmann machine



with combinatorial patterns

Two hidden layers





Hypothesis about the success of deep networks: successive disentanglement of combinatorial correlations?

Visible input → Subfeatures → Features → Patterns

Combinatorial correlations = new type of correlations. Present in images, language, semantics, etc. Crucial (see also Baraniuk, Mossel)

Statistical physics of disordered systems builds upon the study of random ensembles of problems

The choice of ensemble is crucial. eg

$$E = -\frac{1}{2}\sum_{i,j} J_{ij} s_i s_j$$

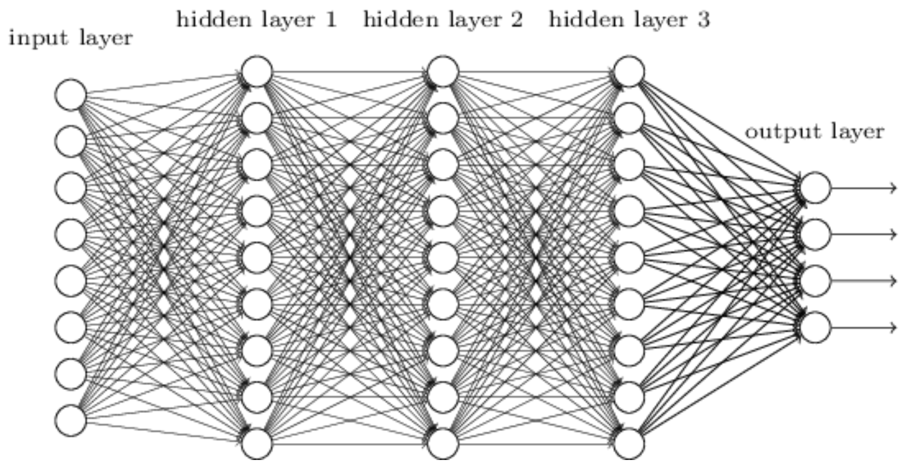
with iid J is very different from J generated with Hebb rule, itself very different from Hebb rule with structured patterns.

Challenge = build better ensembles of the « world », with **combinatorial correlations, hierarchically structured,** and study machine learning with these new ensembles

Chapter Three

The Hidden Manifold Model

Learning in a three-layer feedforward network with structured patterns



Why does it work?

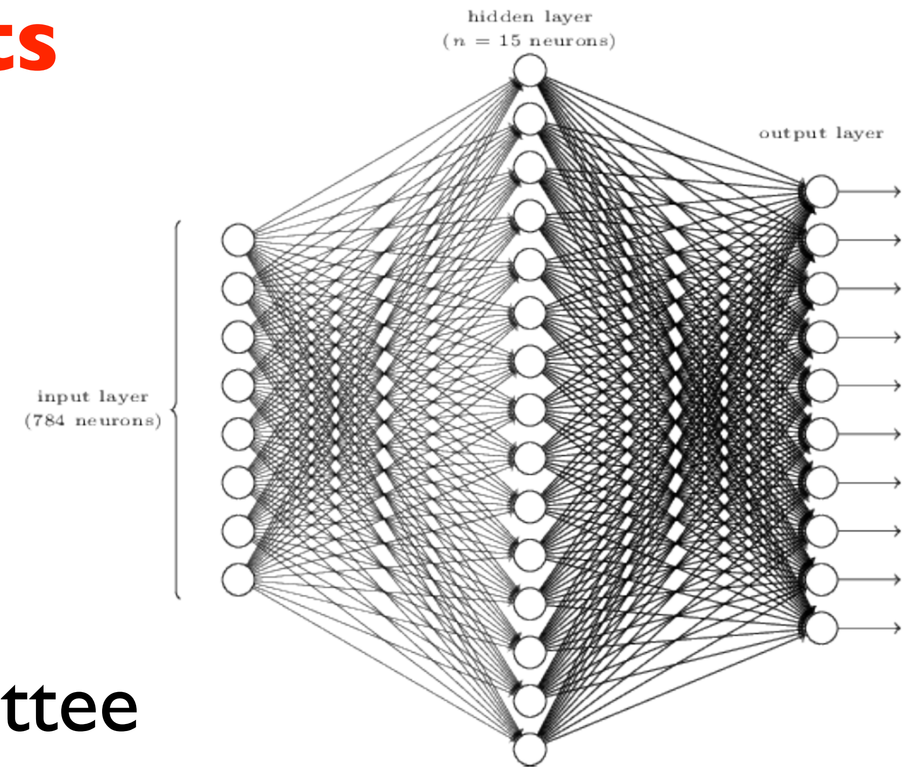
Data structure

- 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

S. Goldt, F. Krzakala, MM, L. Zdeborova
arXiv:1909.11500

Theory: Ensembles of data, ensemble of weights



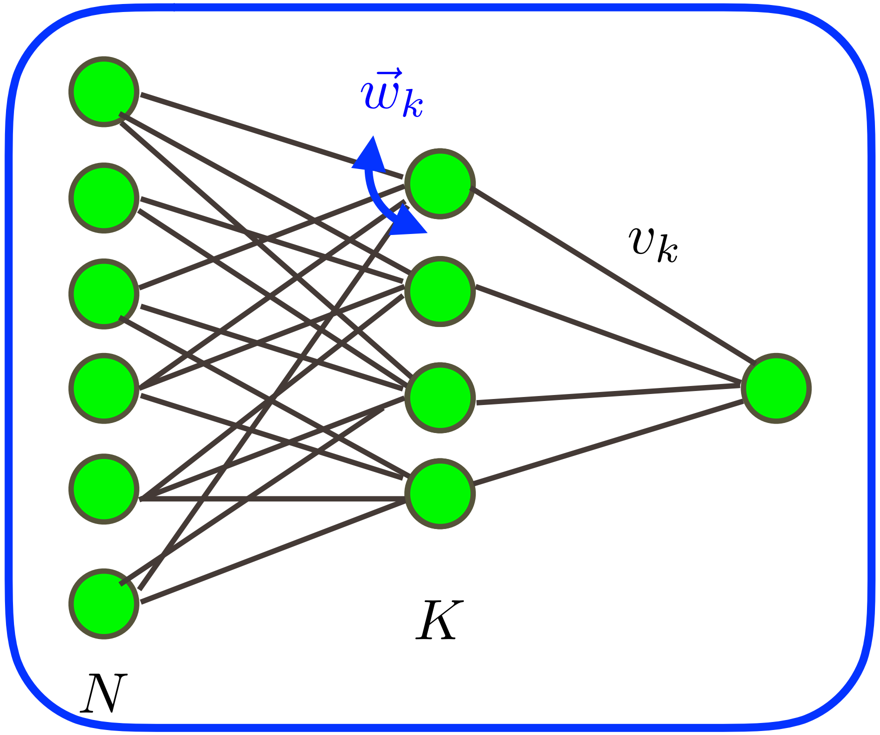
Mostly used so far
Data = input patterns
with **iid entries**

Perceptron learning, committee machine, teacher-student

Pattern μ , input entry i : $X_{\mu i} = \mathcal{N}(0,1)$ $\quad P \times N$ matrix

NB Physicists use P patterns in N dimensions, statisticians use n patterns in p dimensions… **Sorry**

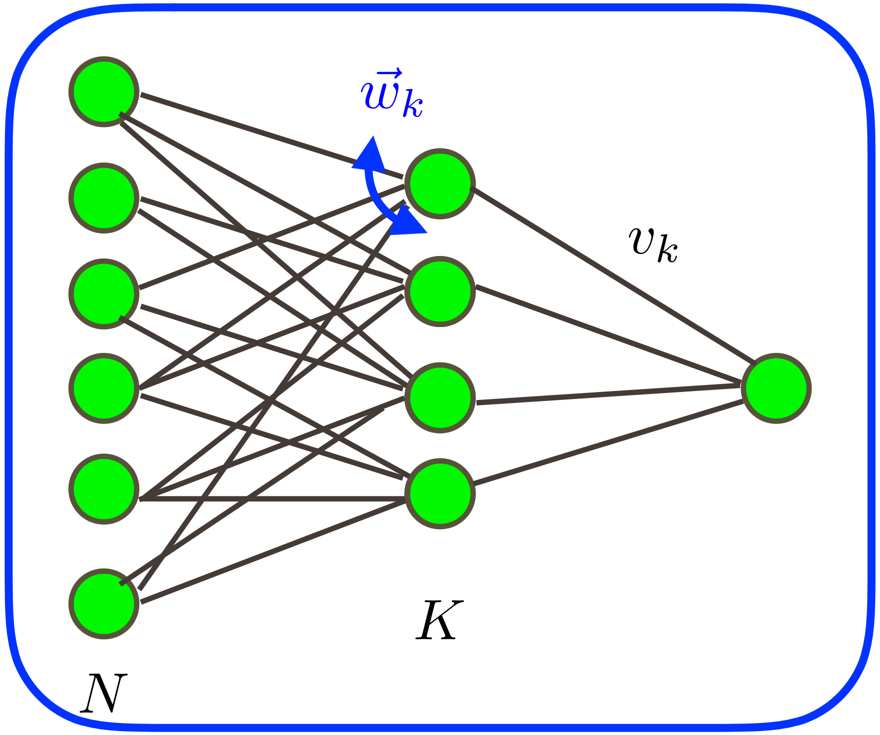
Differences between MNIST and iid data



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

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

Differences between MNIST and iid data



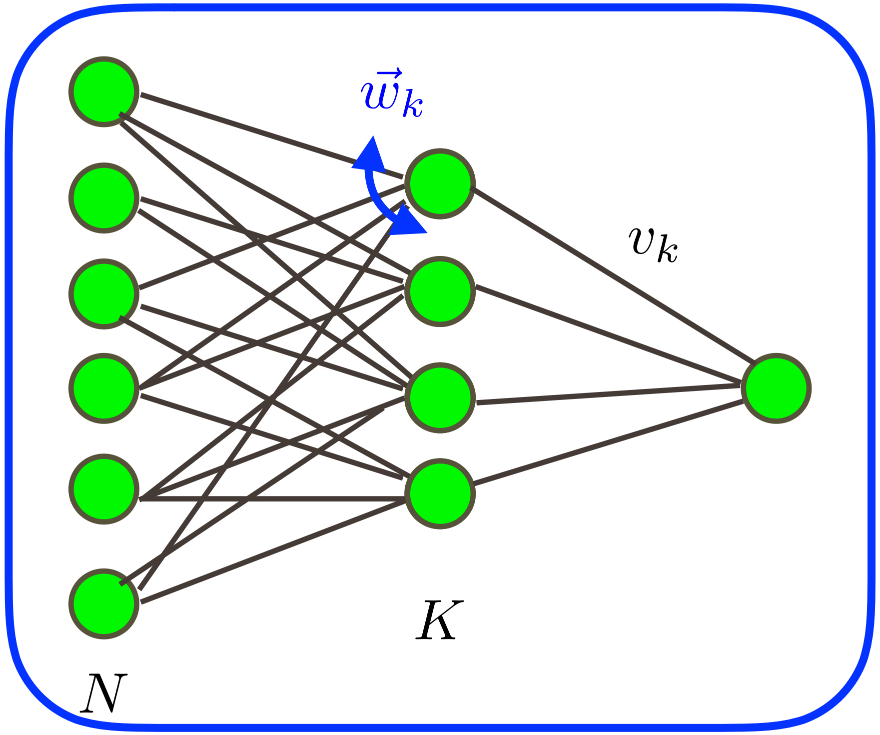
Learn using a 2-layer neural net, K hidden units

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

Task 1: distinguish odd from even numbers in MNIST

$\phi_t(\vec{X}) = 1$ for even digits $\quad \phi_t(\vec{X}) = -1$ for odd digits

Differences between MNIST and iid data



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

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

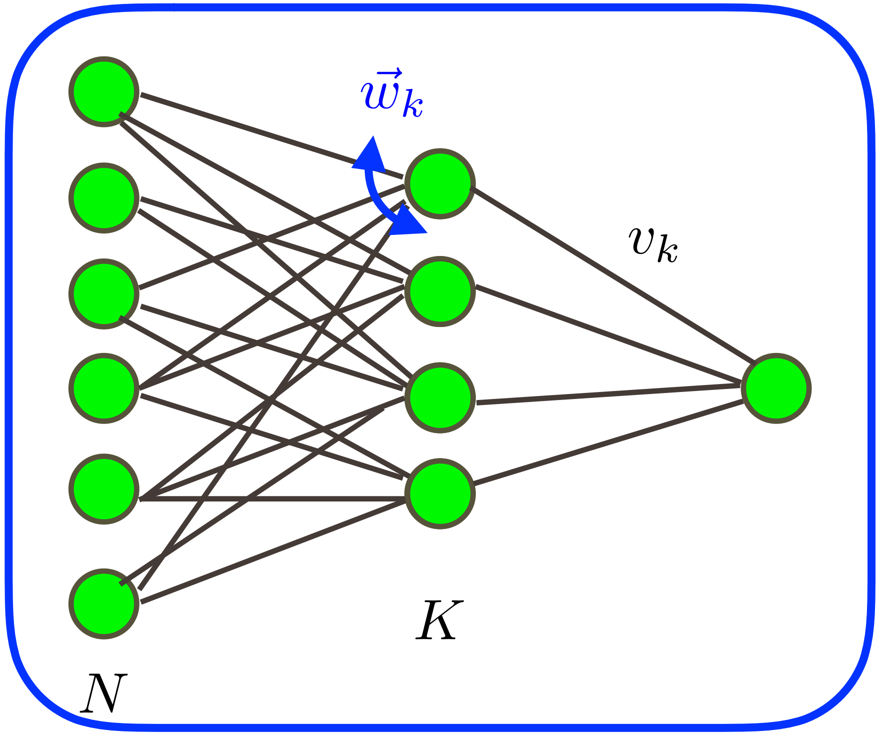
Task 1: distinguish odd from even numbers in MNIST

$\phi_t(\vec{X}) = 1$ for even digits $\quad \phi_t(\vec{X}) = -1$ for odd digits

Task 2: iid input data; desired output given by a 2-layer « teacher network » with M hidden units

$$\phi_t(\vec{X}) = \mathrm{Sign}\left[\sum_{m=1}^{M} \nu_m \, g\left(\vec{\omega}_m.\vec{X}/\sqrt{N}\right)\right]$$

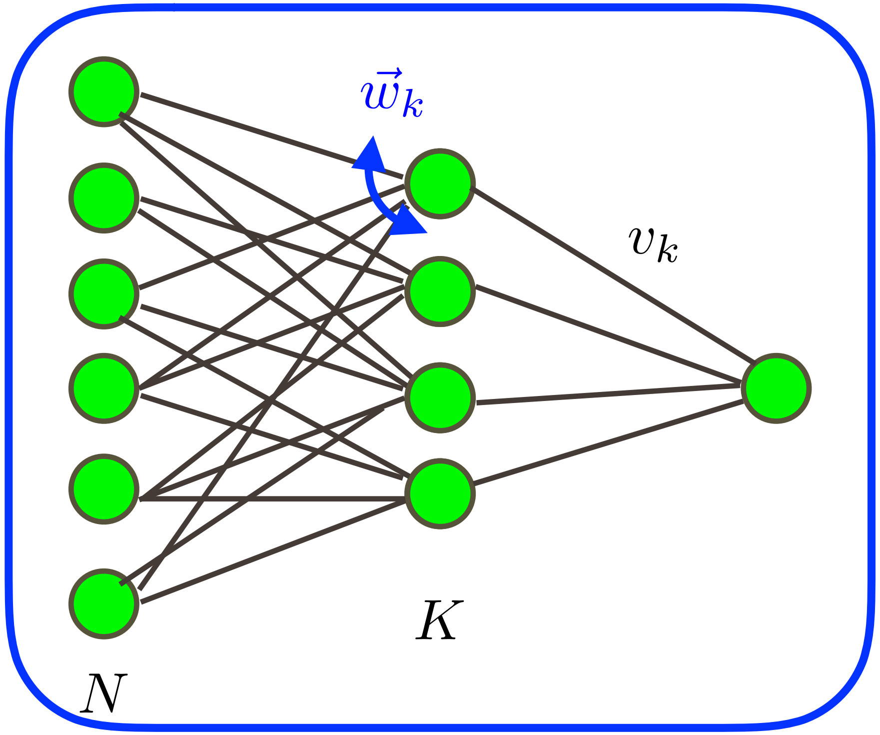
Differences between MNIST and iid data



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

$$\phi(\vec{X}) = \sum_{k}^{K} v_k g\left(\vec{w}_k.\vec{X}/\sqrt{N}\right),$$

Differences between MNIST and iid data



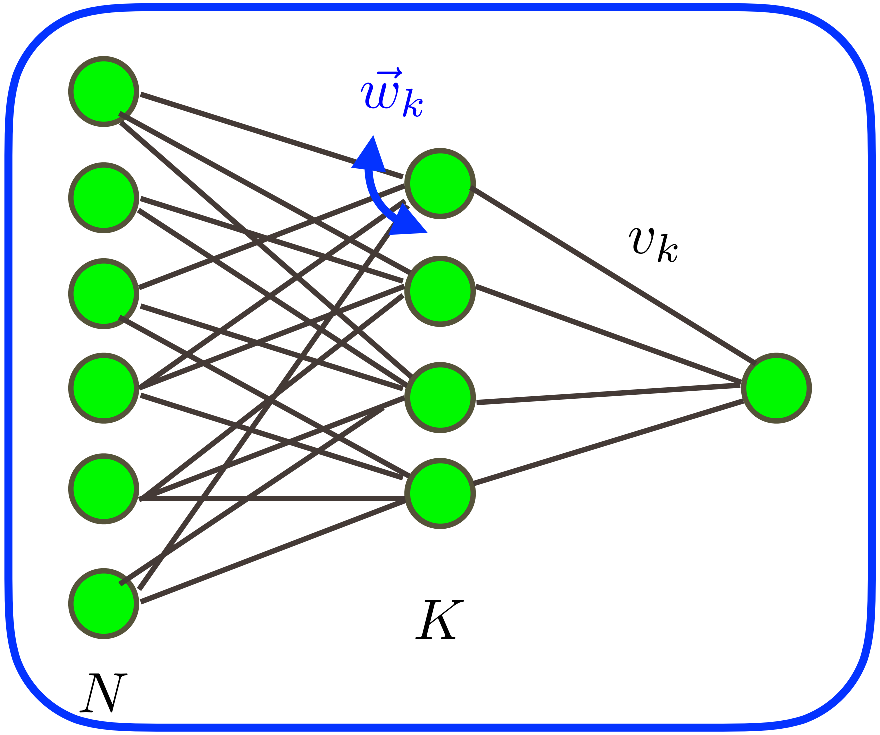
Learn using a 2-layer neural net, K hidden units

$$\phi(\vec{X}) = \sum_{k}^{K} v_k g\left(\vec{w}_k . \vec{X}/\sqrt{N}\right),$$

Training error $$\varepsilon_g = \frac{1}{2P}\sum_{\mu=1}^{P}\theta\left[\phi(\vec{X}_\mu) - \phi_t(\vec{X}_\mu)\right]^2$$

Generalization error: same with P^* new patterns

Differences between MNIST and iid data



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

$$\phi(\vec{X}) = \sum_{k}^{K} v_k g\left(\vec{w}_k.\vec{X}/\sqrt{N}\right),$$

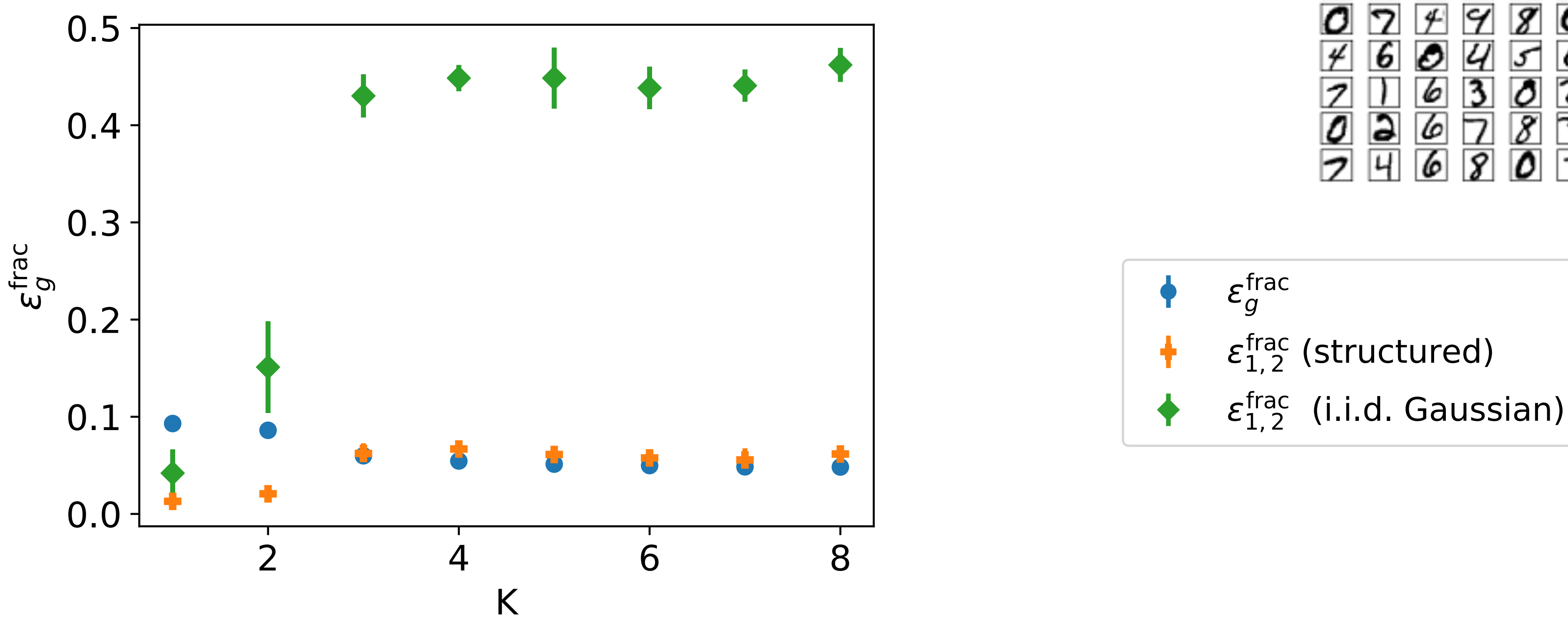
Training error $$\varepsilon_g = \frac{1}{2P}\sum_{\mu=1}^{P}\theta\left[\phi(\vec{X}_\mu) - \phi_t(\vec{X}_\mu)\right]^2$$

Generalization error: same with P^* new patterns

Also monitored: difference between two learning trials with different initial conditions

$$\varepsilon_{12} = \frac{1}{2P}\sum_{\mu=1}^{P}\theta\left[\phi_1(\vec{X}_\mu) - \phi_2(\vec{X}_\mu)\right]^2$$

MNIST data



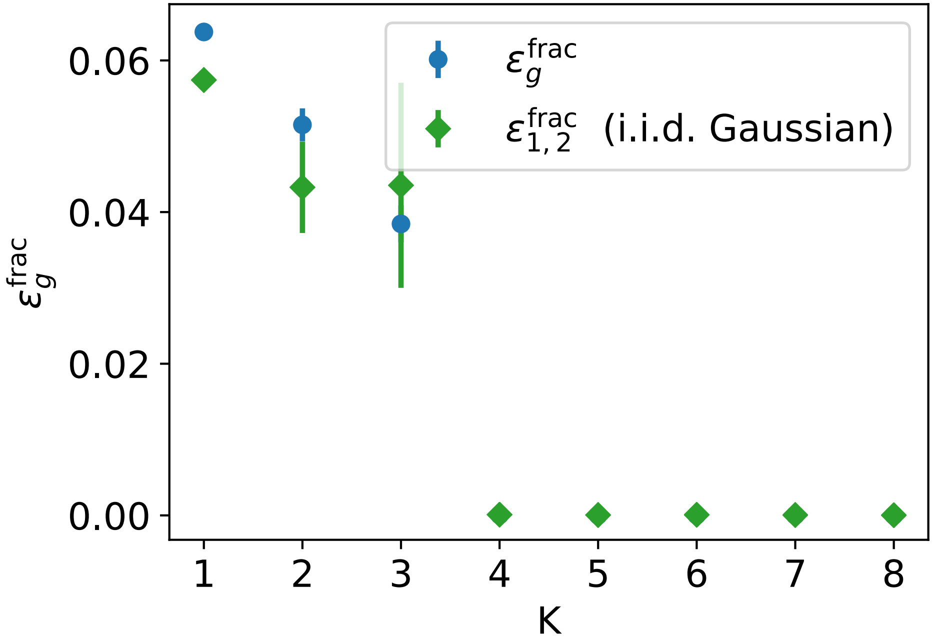
Generalization error decreases with K

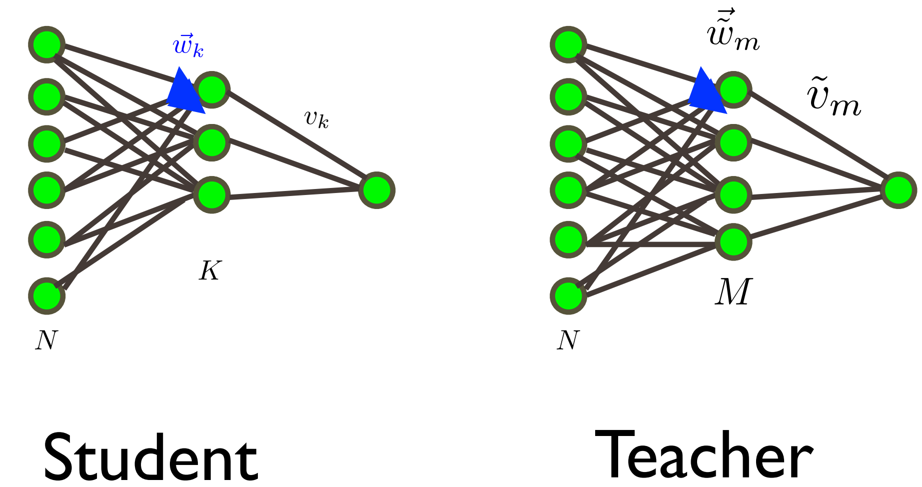
The difference on MNIST between two trials agrees with generalization error

The difference on random images between two trials is large (nearly uncorrelated functions)

iid data and teacher network

$M = 4$

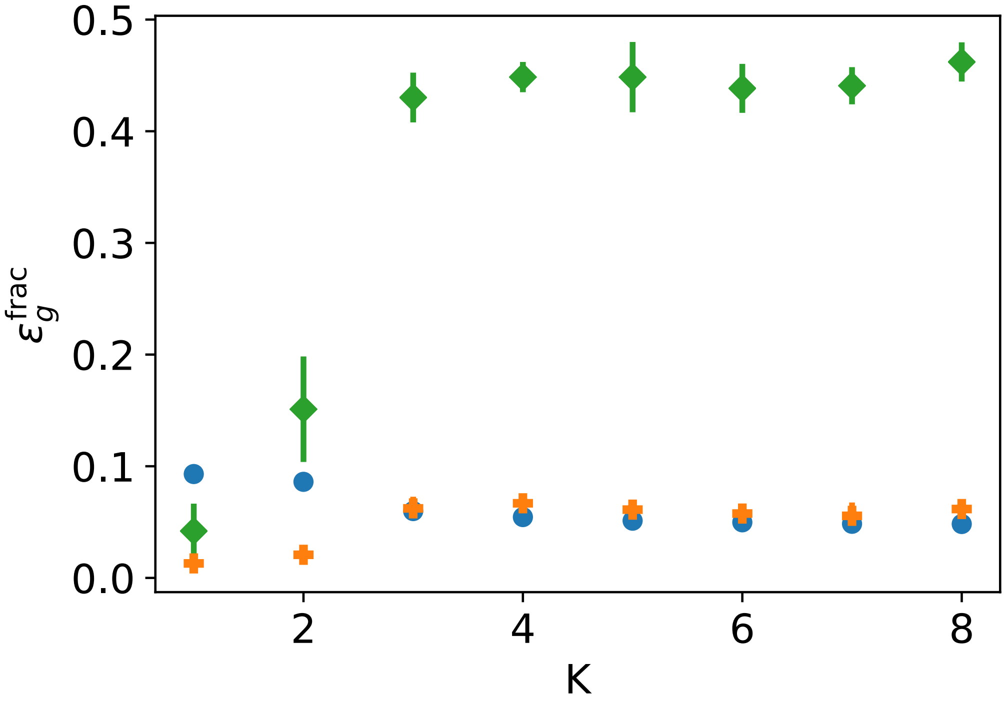


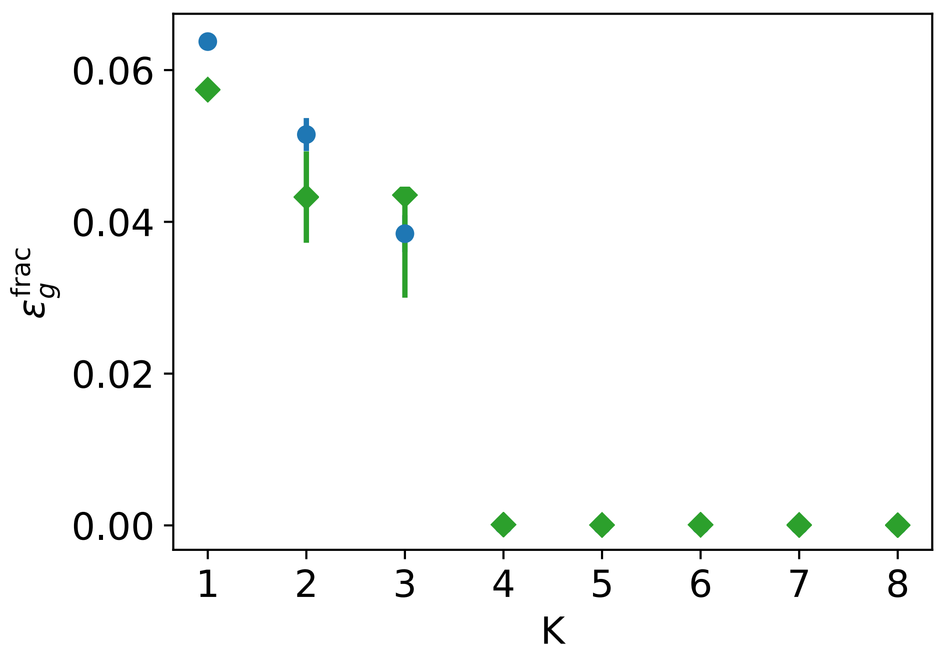


Generalization error decreases with K , vanishes for $K \geq M$

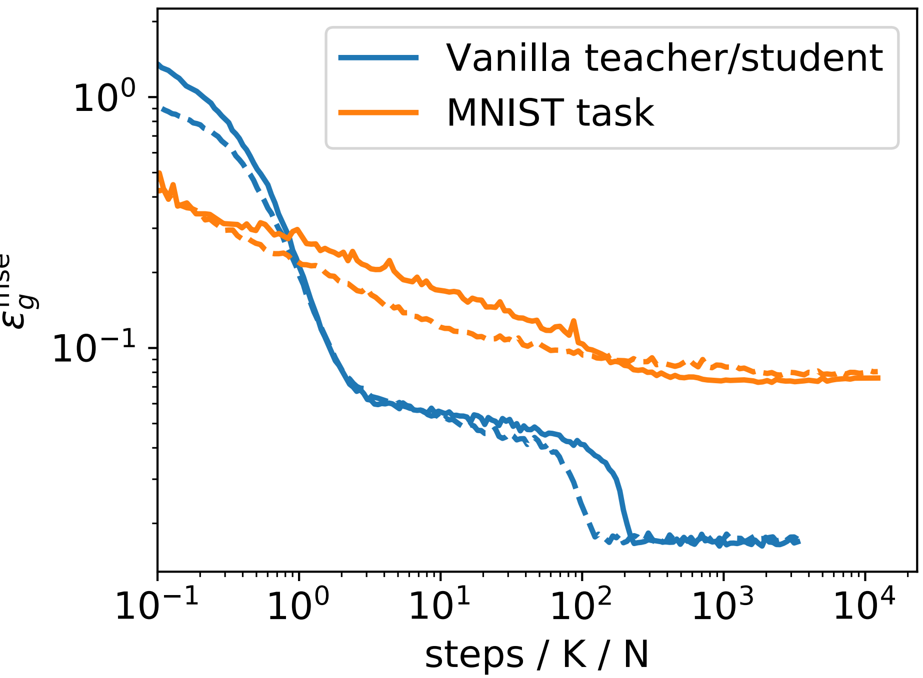
The difference on random images between two trials is equal to ε_g . For $K \geq M$ the two trials learn the same global function

MNIST versus iid data





Learning dynamics

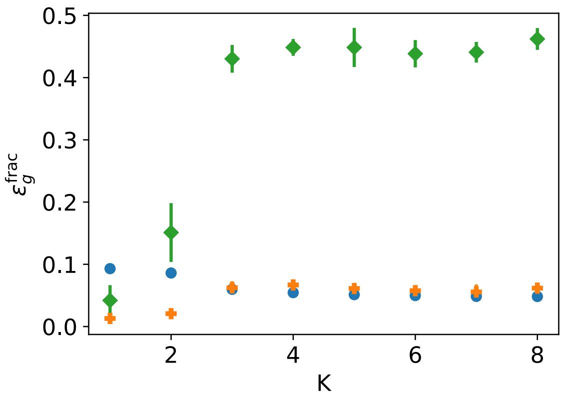


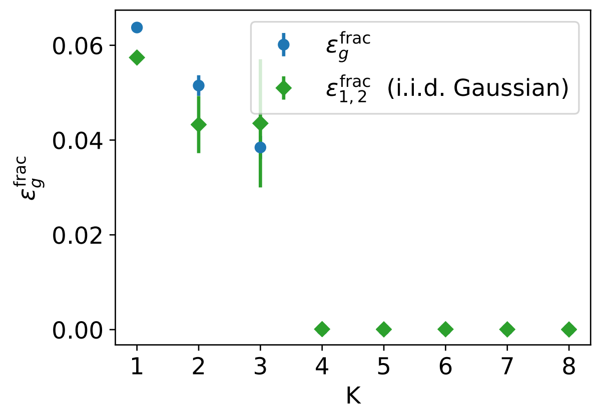
Plateau in the « teacher-student » setup

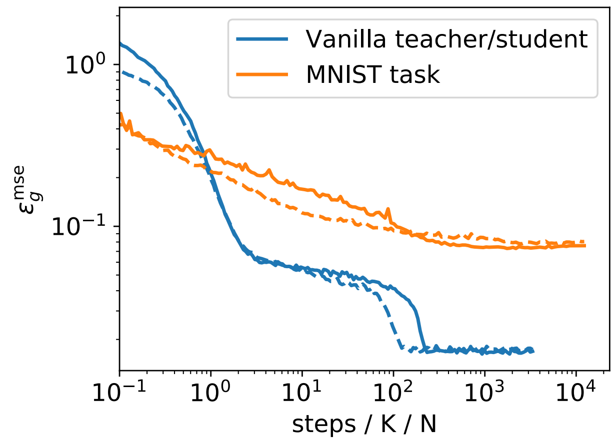
M. Biehl and H Schwarze 95, Saad and Solla 95

After some time the dynamics stabilize in a metastable state where all the hidden units have roughly the same overlap with all the teacher vector. Long plateau before the specialization of hidden units occurs.

Differences between MNIST and iid data







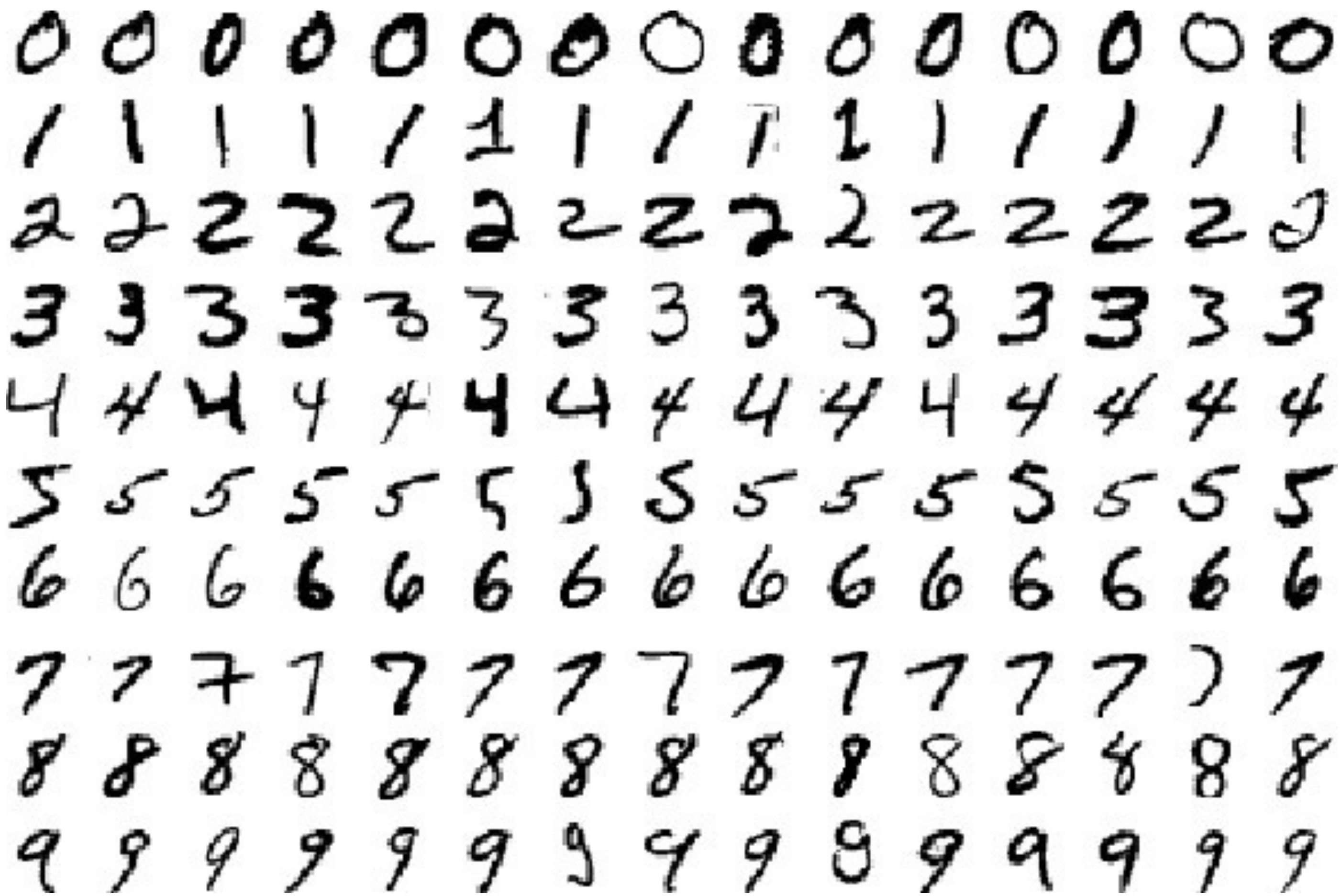
Two different trials learn the same function in iid data teacher-student, completely different functions in MNIST (outside of the hidden manifold)

Plateau in the learning of teacher-student with iid data, not seen in MNIST

What data structure?

The hidden manifold of data

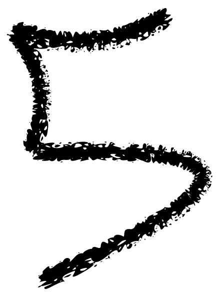
MNIST

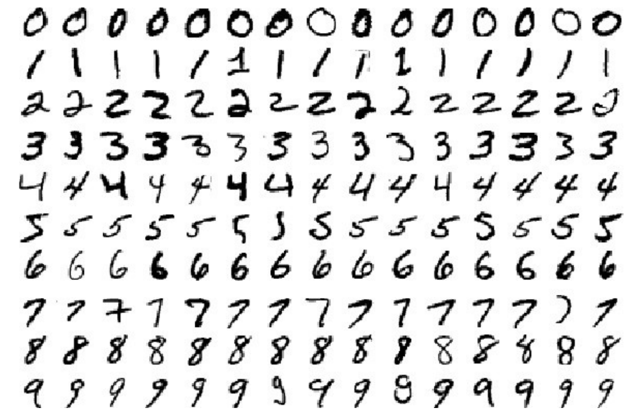


Input space: dimension $28^2 = 784$

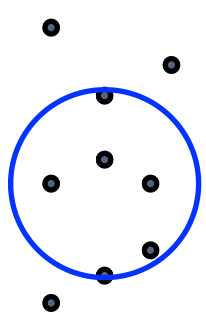
The hidden manifold of data

Input space: dimension $28^2 = 784$





Manifold of handwritten digits in MNIST:

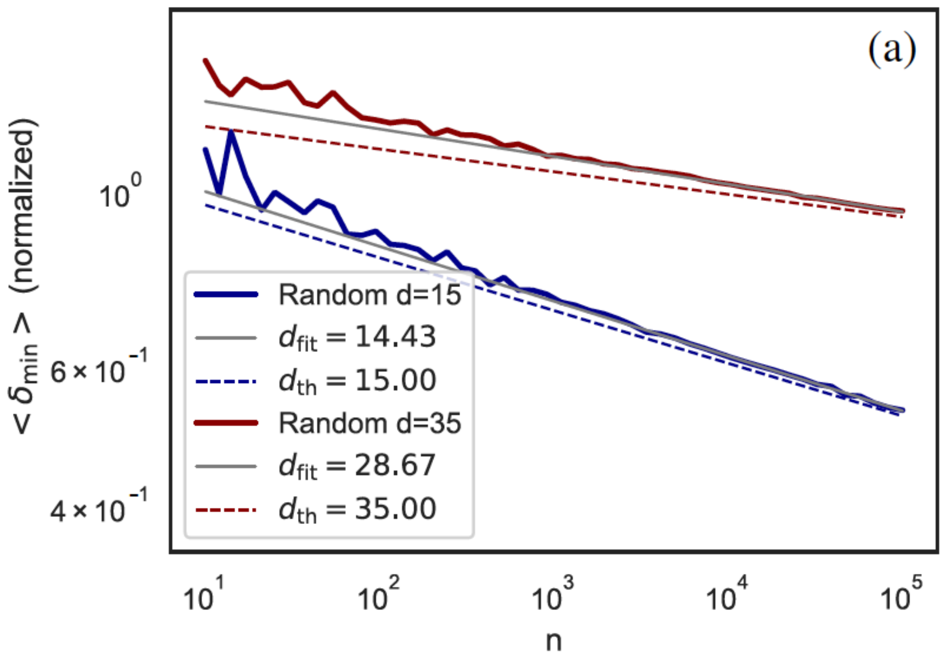


$p \simeq cR^d$

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



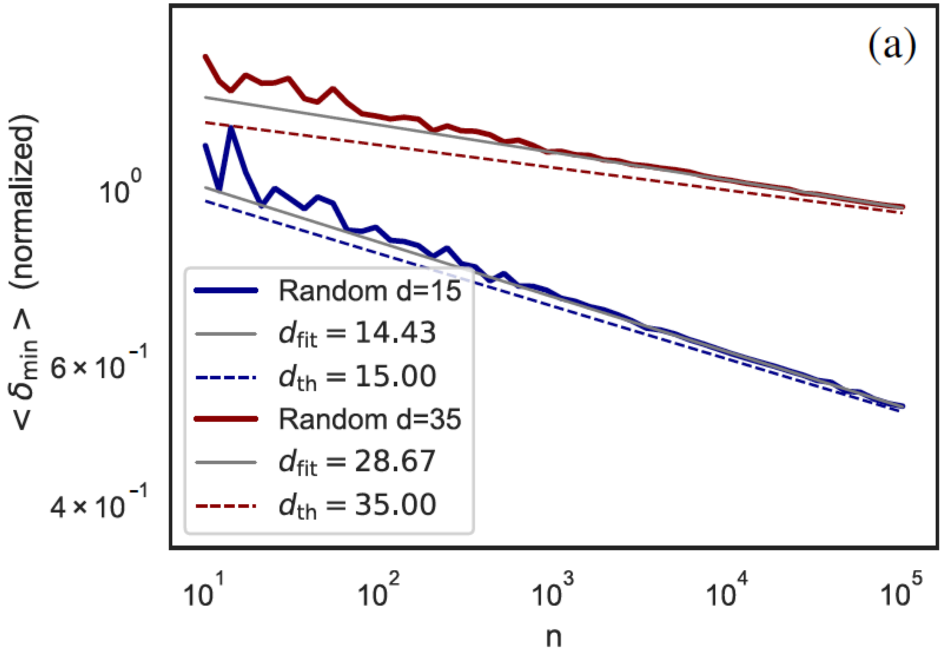
MNIST: $d = 784$

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

Spigler et al. 19

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

The hidden manifold of data

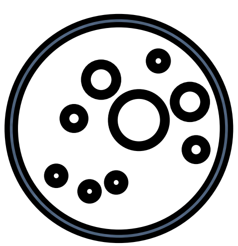


MNIST: $d = 784$

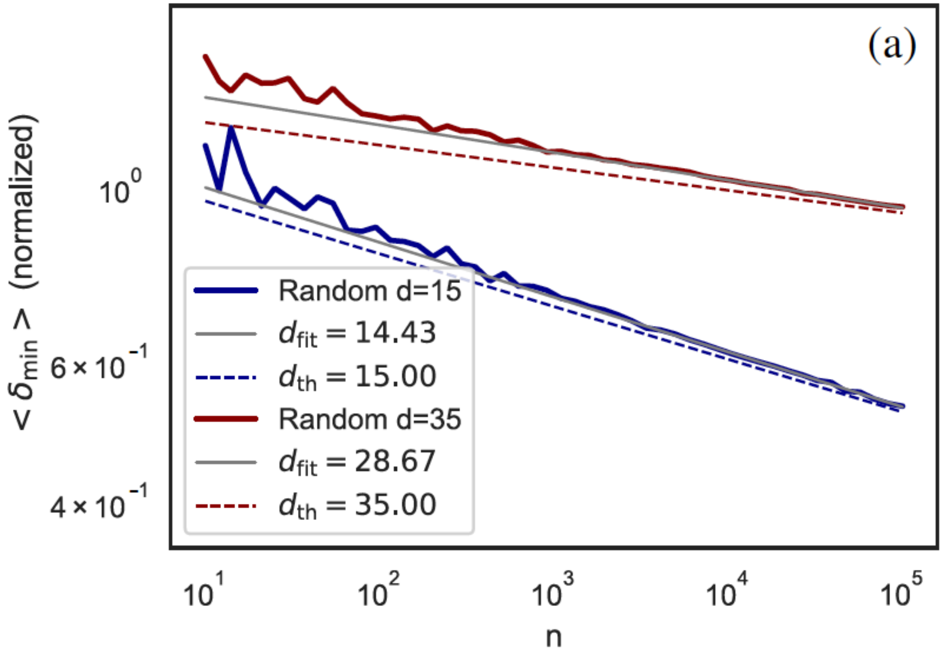
$d_{\mathrm{eff}} \simeq 15$

Spigler et al. 19

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



The hidden manifold of data

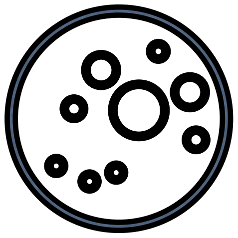


MNIST: $d = 784$

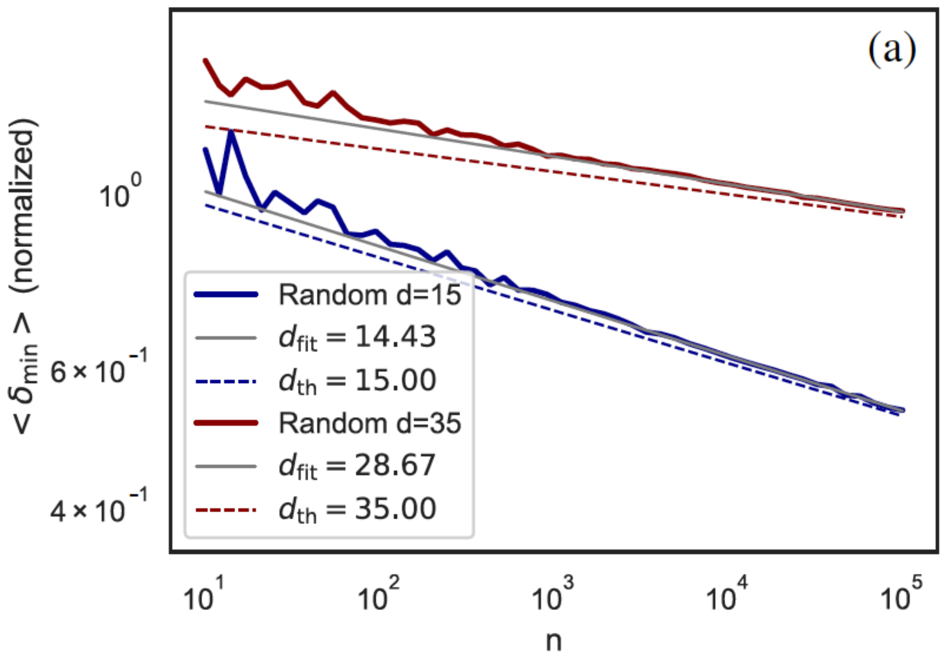
$d_{\mathrm{eff}} \simeq 15$

Spigler et al. 19

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



The hidden manifold of data

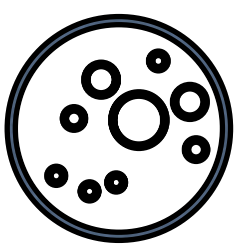


MNIST: $d = 784$

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

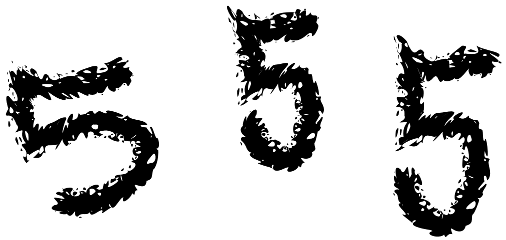
Spigler et al. 19

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



The neural net should answer: this image does not seem to be a handwritten digit

Structure of the task: perceptual sub-manifolds



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

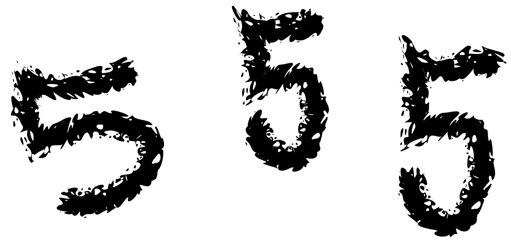
Hein Audibert 05

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**…

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!

A new 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

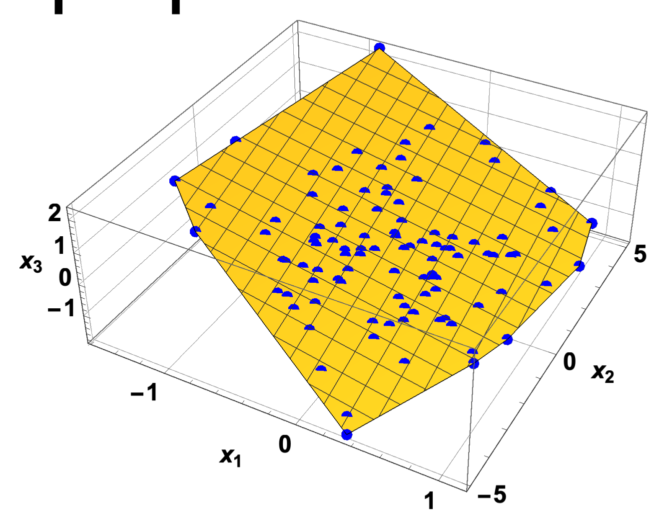
Pattern μ: $$X_{\mu i} = f\left[\frac{1}{\sqrt{R}}\sum_{r=1}^{R} C_{\mu r} F_{ir}\right]$$

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):

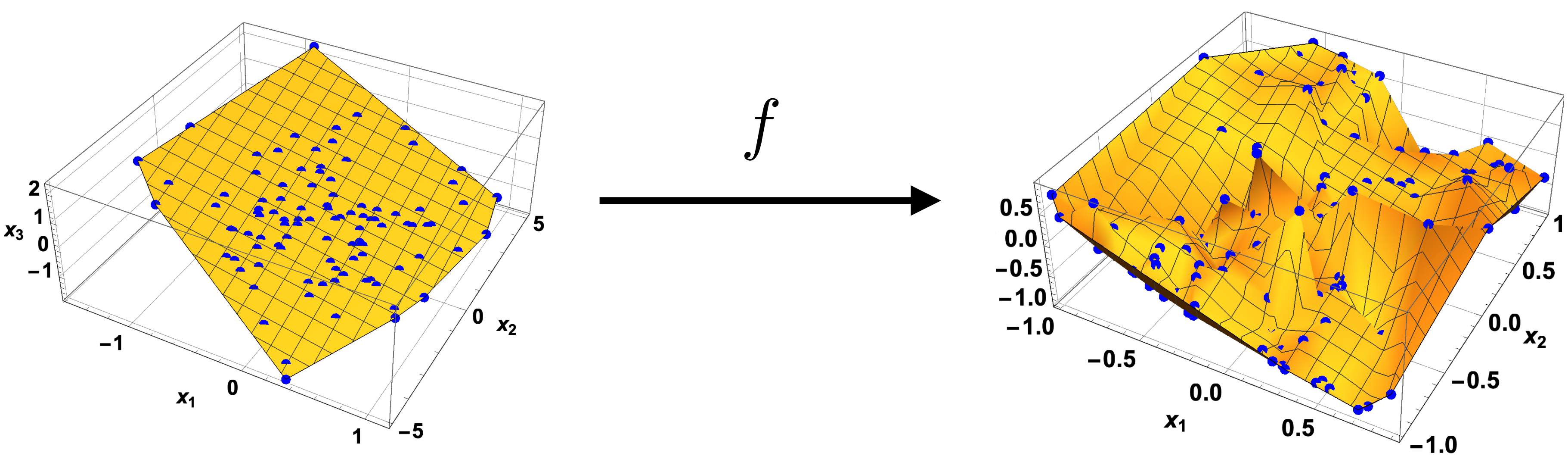
$$\sum_{r=1}^{R} C_r \vec{F}_r$$



An ensemble for the hidden manifold

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

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



An ensemble for the task

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

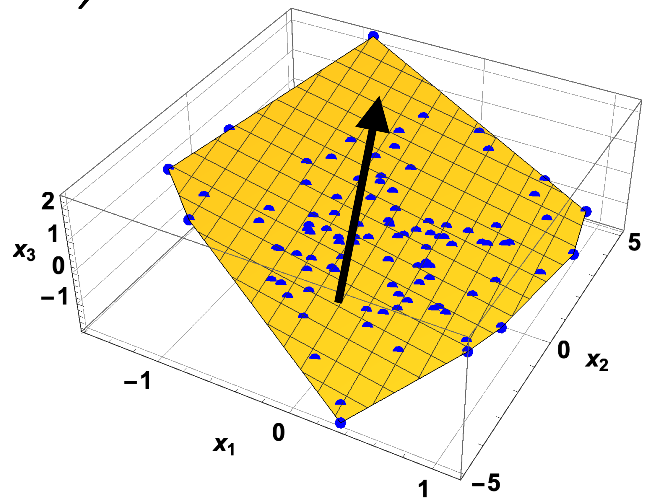
« Latent representation »: $\{C_r\}$

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

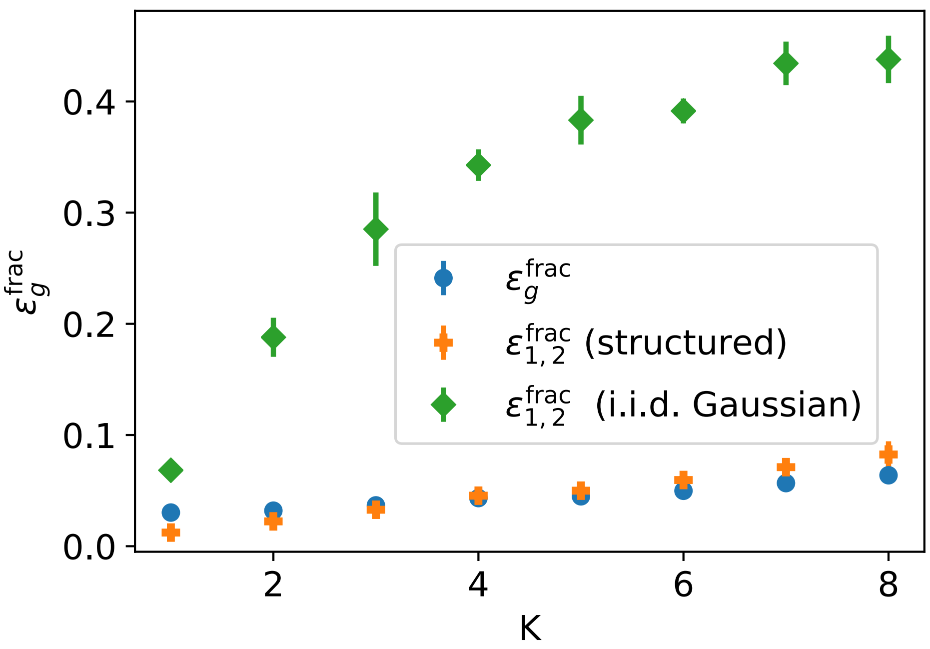
Task $\quad y = \sum_{m=1}^{M} \tilde{v}_m \; g\left(\sum_{r=1}^{R} \tilde{w}_{mr} C_r\right)$

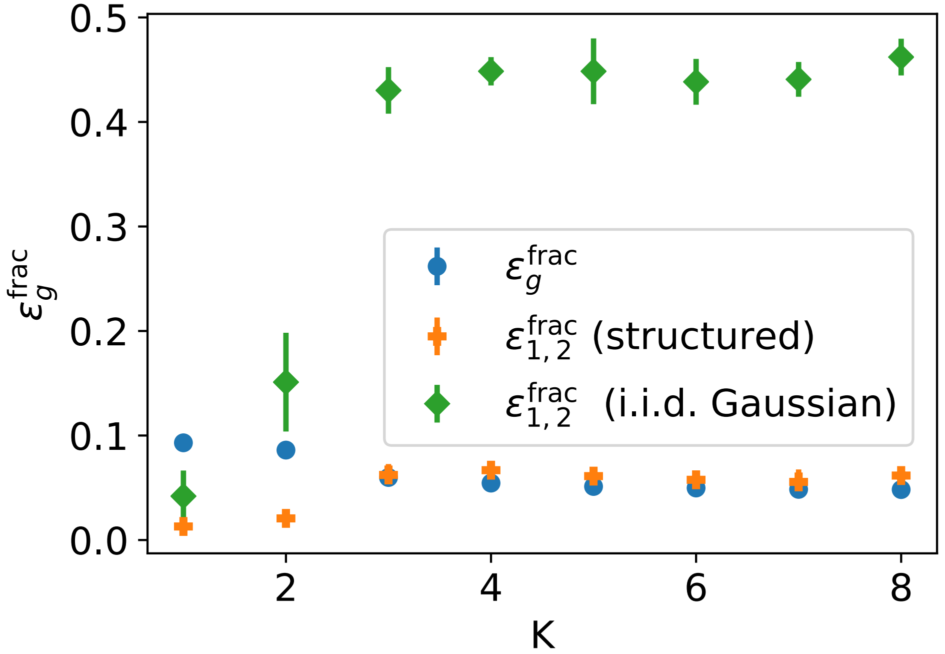
depends on $\{\tilde{w}_m . C\}, \quad m \in \{1, ...M\}$

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

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

Experimenting with the « hidden manifold model »



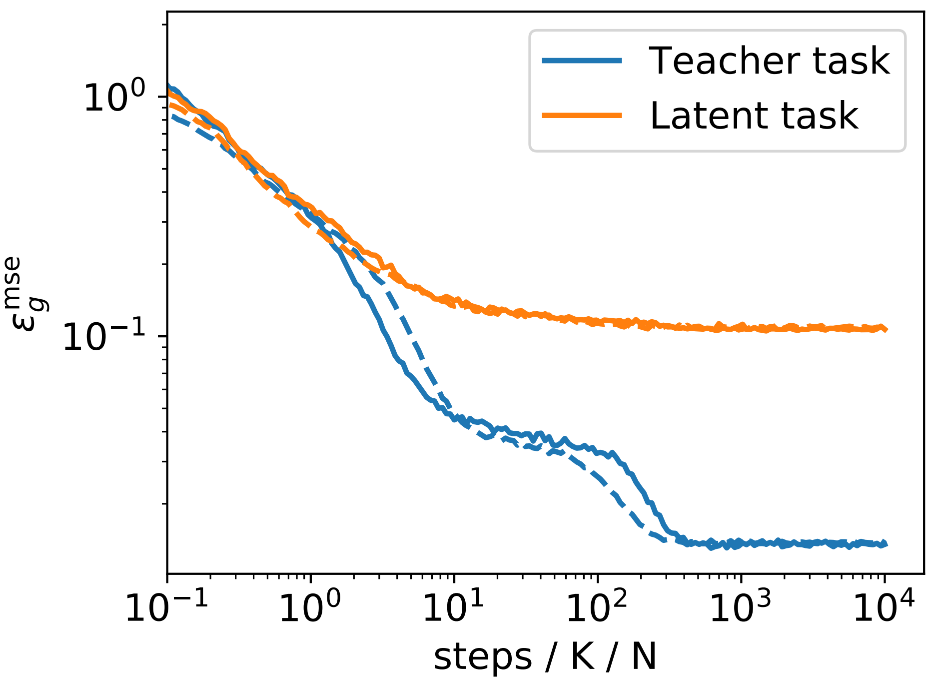


Hidden manifold model

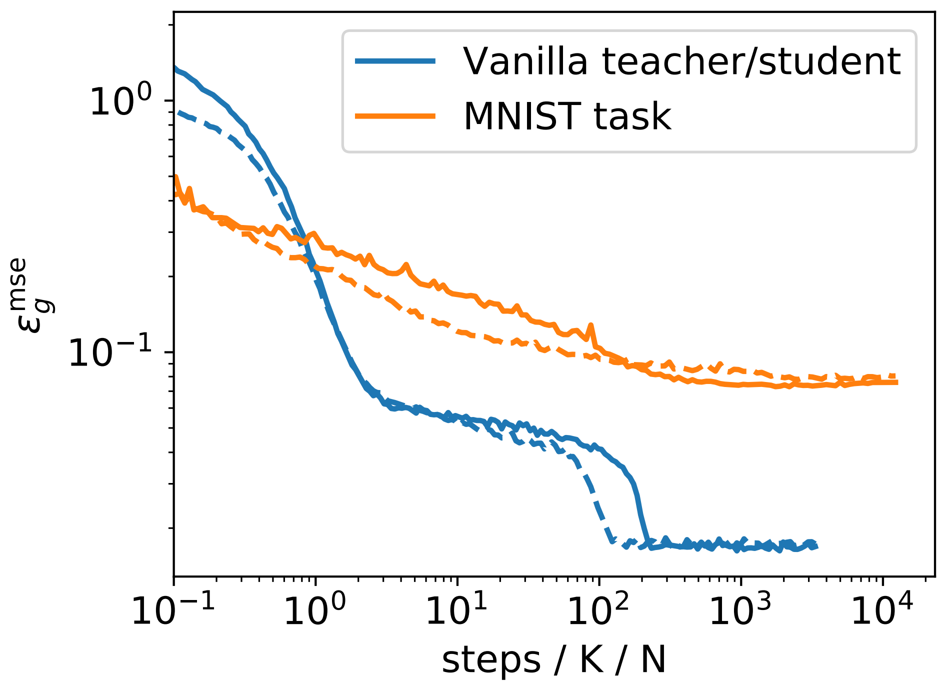
$R = 10$

MNIST

Experimenting with the « hidden manifold model »



Hidden manifold model



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

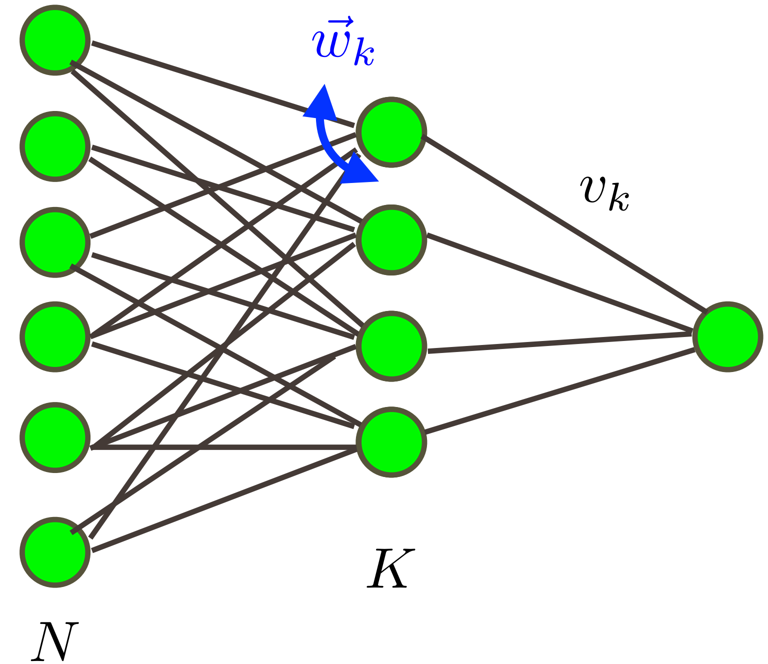
- Does not have the pathologies of teacher-student setup with iid data
- Learning and generalization phenomenology $\sim$ MNIST
- Can be studied analytically

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

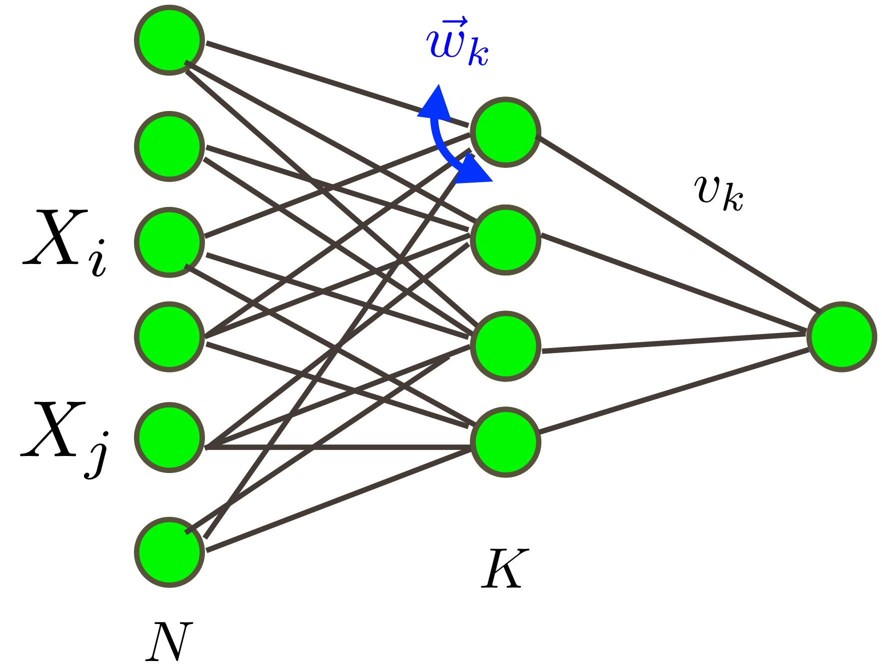
Analytic study of the hidden manifold model

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

Correlated components

iid

balanced:



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

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

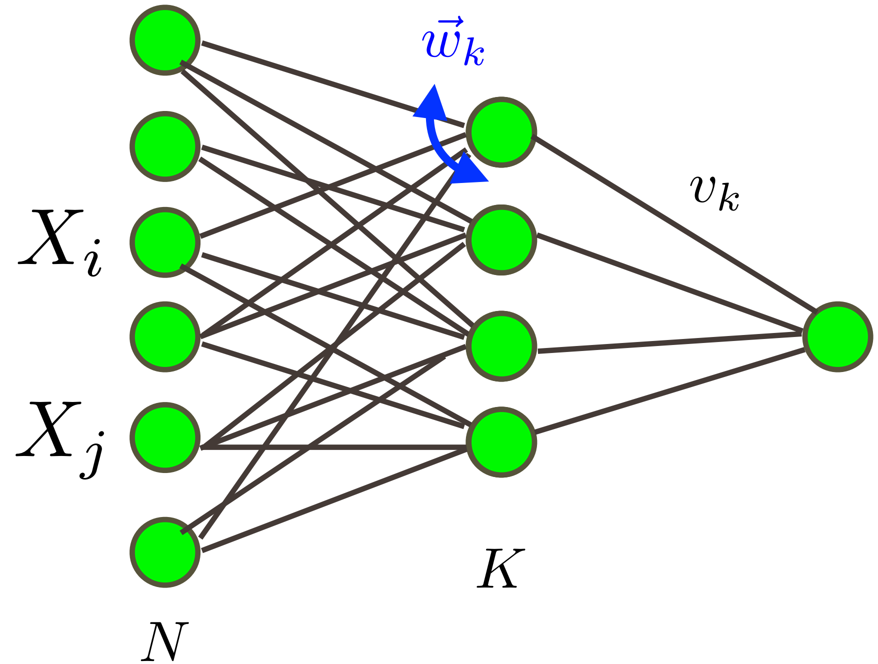
$$\frac{1}{N}\sum_i F_{ri}F_{ri} = 1$$

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 $O(1)$

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

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

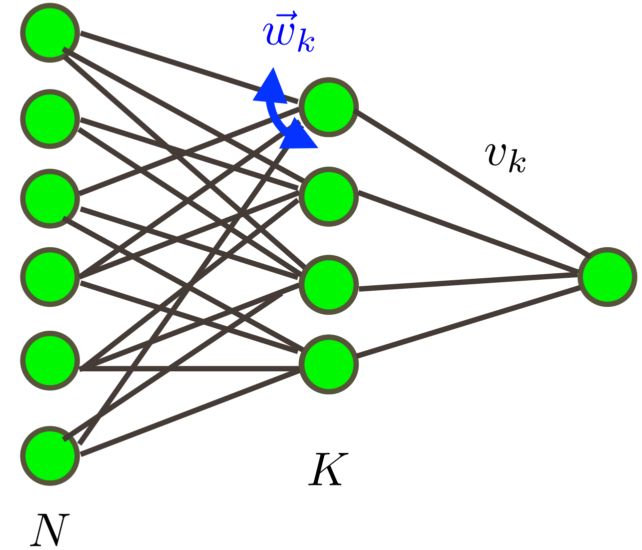
Gaussian Equivalence Theorem (GET)

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

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

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

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

$c = \langle f(u)^2 \rangle \qquad a = \langle f(u) \rangle \qquad b = \langle u f(u) \rangle \qquad u$ 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.

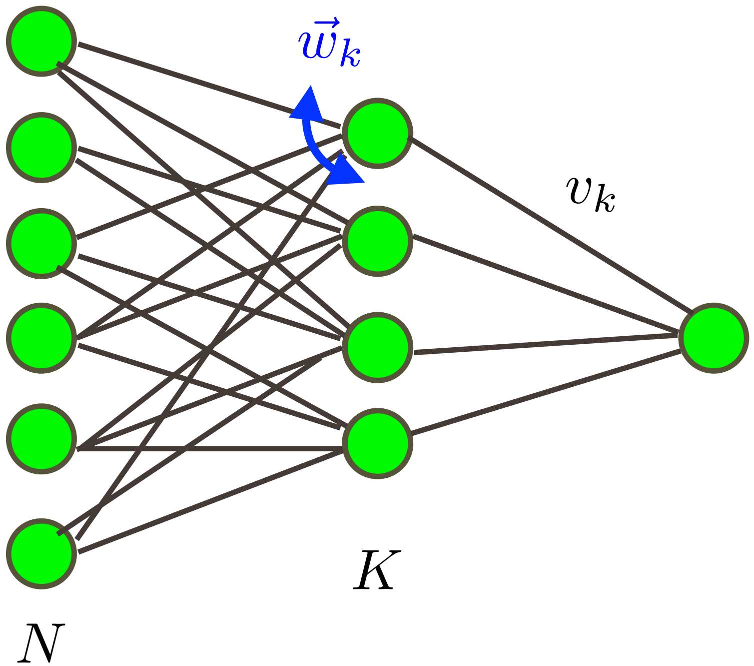
NB: depends on the manifold folding function f only through the three quantities

$c = \langle f(u)^2 \rangle$ $\quad a = \langle f(u) \rangle$ $\quad b = \langle u f(u) \rangle$ $\quad u$ 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, …

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

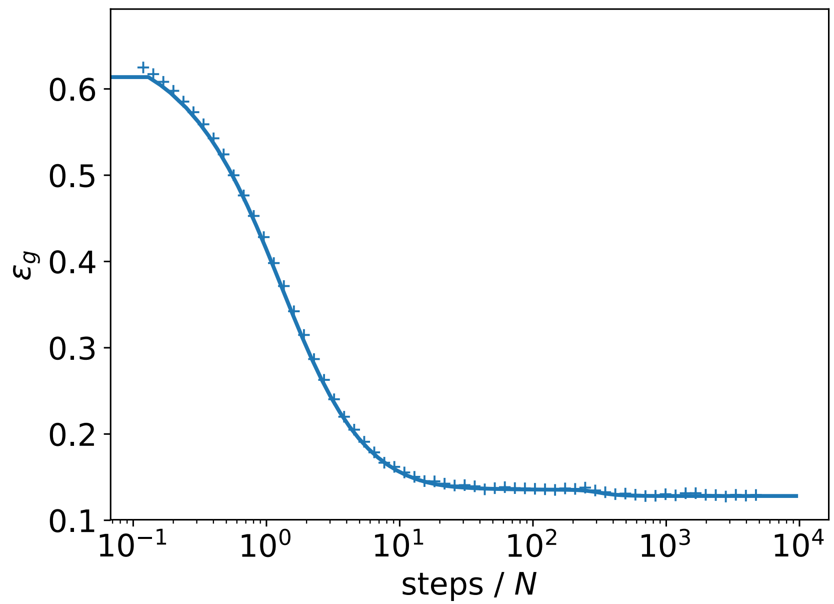
$$\Delta = \sum_{\ell=1}^{K} g(\lambda^\ell) - \sum_{m=1}^{N} \tilde{g}(\nu^m)$$

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

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

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

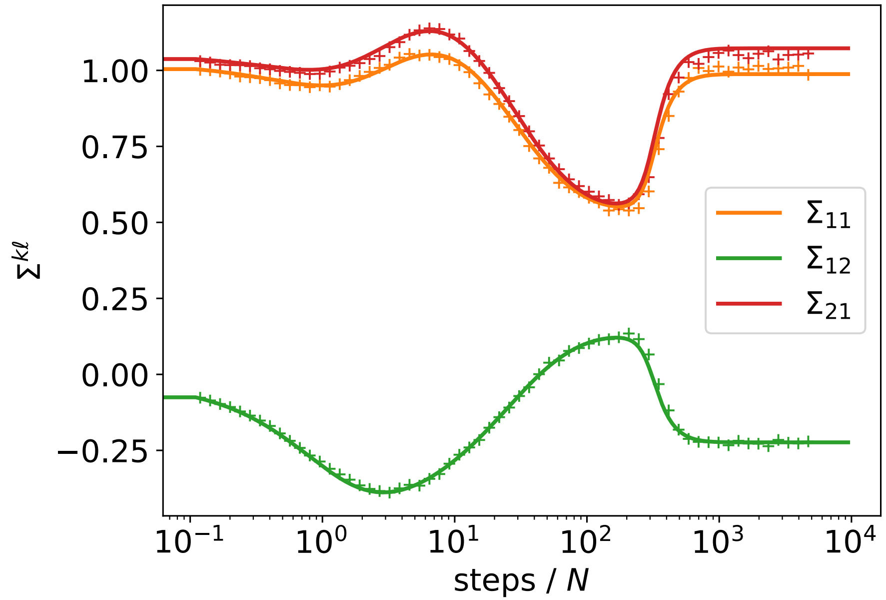
ODE Theory vs simulations N=8000, D=4000, M=2, K=2

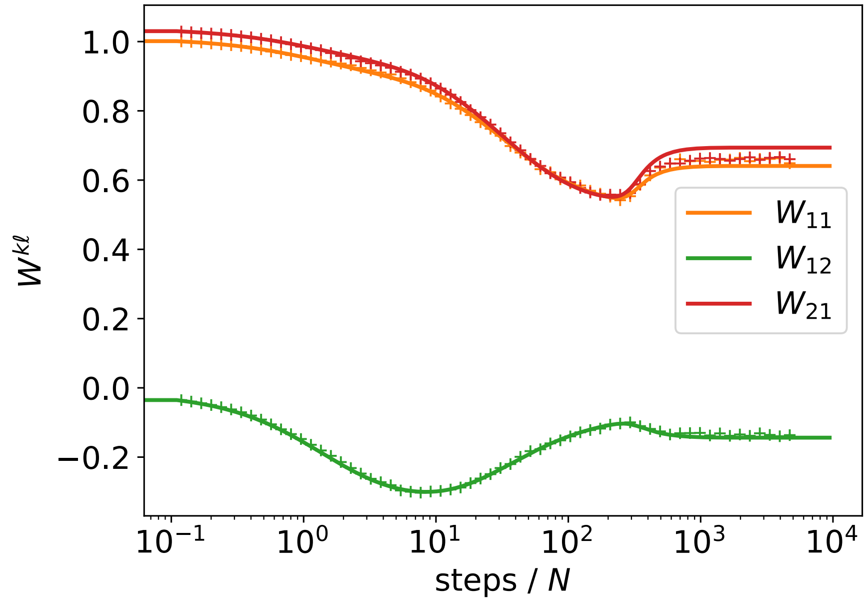


$$S_r^k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^k F_{ir}$$

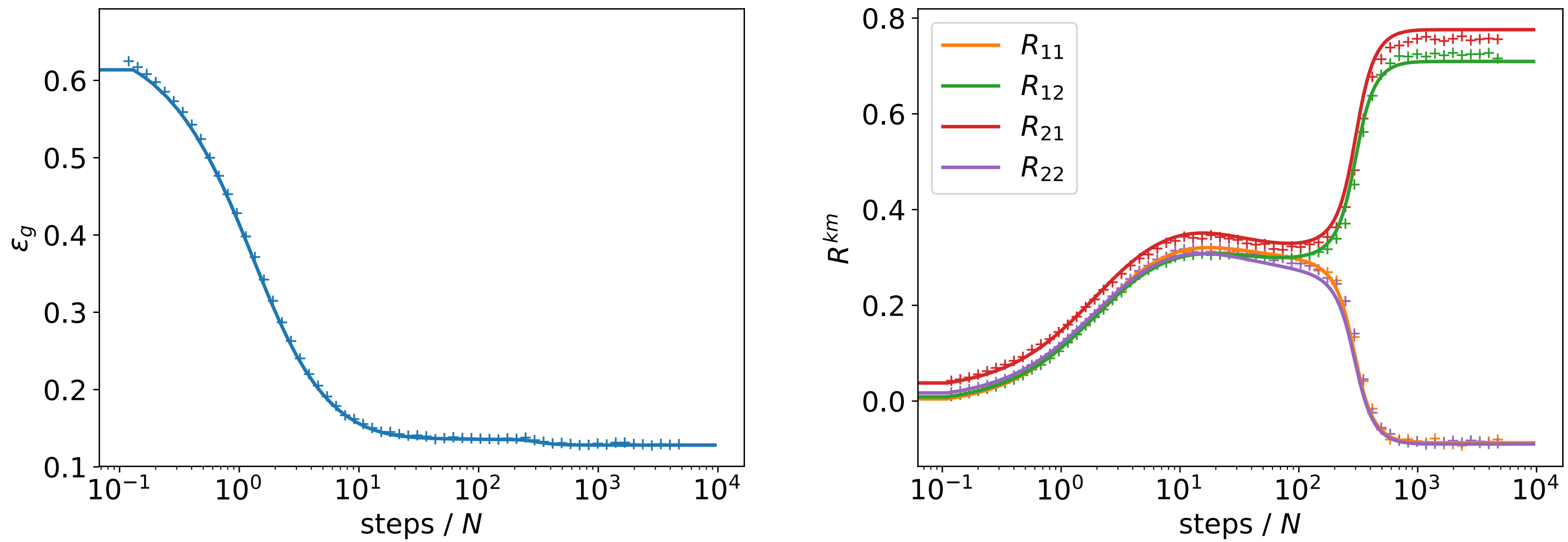
$$W^{k\ell} = \frac{1}{N} \sum_{i} w_i^k w_i^\ell$$

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





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



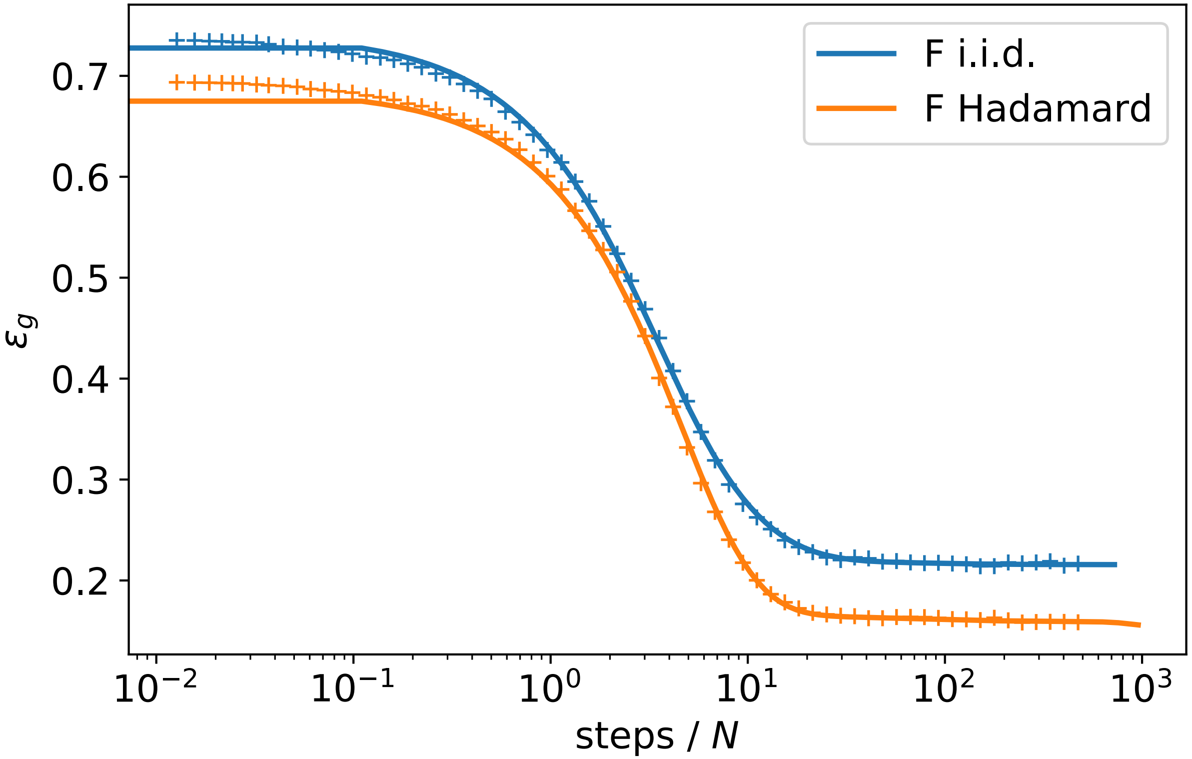
$$S_r^k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^k F_{ir} \qquad R^{km} = \frac{b}{D} \sum_{r=1}^{D} S_r^k \tilde{w}_r^m$$

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

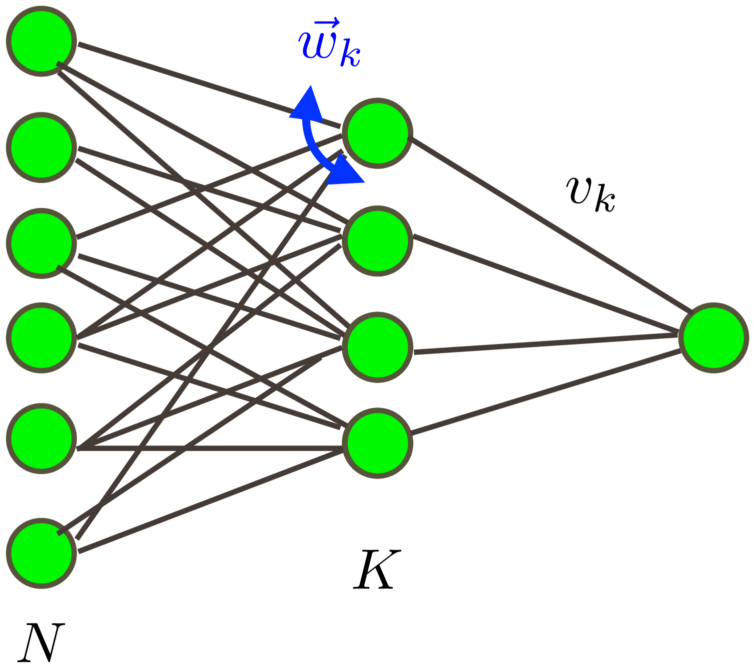
specializes after 100 steps

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

Hadamard F



Phase diagram of Hidden Manifold Model



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

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

Learn from database of P patterns.

Training error

$$E = \sum_{\mu=1}^{P} \epsilon\left[\Phi_t(X_\mu) - \Phi(X_\mu)\right]$$

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

$$Z = \int \prod_{i,k} \left[dw_i^k P_w(w_i^k)\right] e^{-\beta \sum_\mu \epsilon(\Phi_t(\vec{X}_\mu) - \Phi(\vec{X}_\mu))}$$

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

$$Z = \int \prod_{i,k} \left[dw_i^k P_w(w_i^k)\right] e^{-\beta \sum_\mu \epsilon(\Phi_t(\vec{X}_\mu) - \Phi(\vec{X}_\mu))}$$

Compute $\frac{1}{N} \log Z$ averaged over the distribution of latent components $C_{\mu r}$, using replicas $\mathbb{E}_C Z^n \simeq e^{N\Psi(n)}$

$$\mathbb{E}_C \frac{1}{N} \log Z = \Psi'(0)$$

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

$$Z = \int \prod_{i,k} \left[dw_i^k P_w(w_i^k)\right] e^{-\beta \sum_\mu \epsilon(\Phi_t(\vec{X}_\mu) - \Phi(\vec{X}_\mu))}$$

Compute $\frac{1}{N} \log Z$ averaged over the distribution of

latent components $C_{\mu r}$, using replicas $\mathbb{E}_C Z^n \simeq e^{N\Psi(n)}$

$$\mathbb{E}_C \frac{1}{N} \log Z = \Psi'(0)$$

$$Z^n = \int \prod_{ik} \prod_{a=1}^{n} \left[dw_i^{ka} P_w(w_i^{ka})\right] e^{-\beta \sum_{\mu,a} \epsilon(\Phi_t(\vec{X}_\mu) - \Phi^a(\vec{X}_\mu))}$$

Committee with weights w_i^{ka}

$$Z^n = \int \prod_{ik} \prod_{a=1}^{n} \left[dw_i^{ka} P_w(w_i^{ka}) \right] e^{-\beta \sum_{\mu,a} \epsilon(\Phi_t(\vec{X}_\mu) - \Phi^a(\vec{X}_\mu))}$$

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

Natural variables = inputs to hidden neurons

$\lambda_\mu^{ka} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^{ka} f\left[\frac{1}{\sqrt{R}} \sum_{r=1}^{D} F_{ir} C_{\mu r}\right]$	$\nu_\mu^m = \frac{1}{\sqrt{R}} \sum_{r=1}^{R} \widetilde{w}_r^m C_{r\mu}$

GET ⇨ These are joint Gaussian, with known covariance

$$\mathbb{E}_C Z^n = \int \prod_{ika} \left[dw_i^{ka} P_w(w_i^{ka}) \right] \prod_\mu \mathbb{E}_{\lambda;\nu} \exp\left[-\beta \sum_{\mu,a} \epsilon\left(\sum_m \widetilde{g}(\nu_\mu^m) - \sum_k g(\lambda_\mu^{ka}) \right) \right]$$

⇨ The replica computation can be done, for any $\epsilon, g, \tilde{g}\ , K, M$

In short

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.

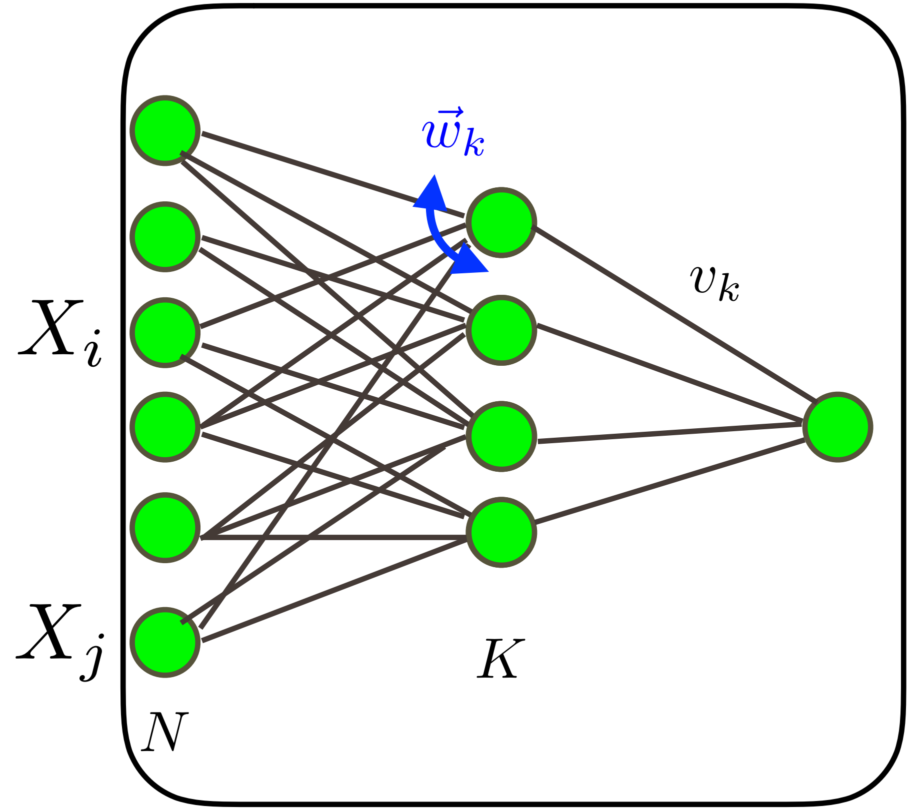
Results… coming soon.

NB: Hidden manifold and random feature

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

Correlated components

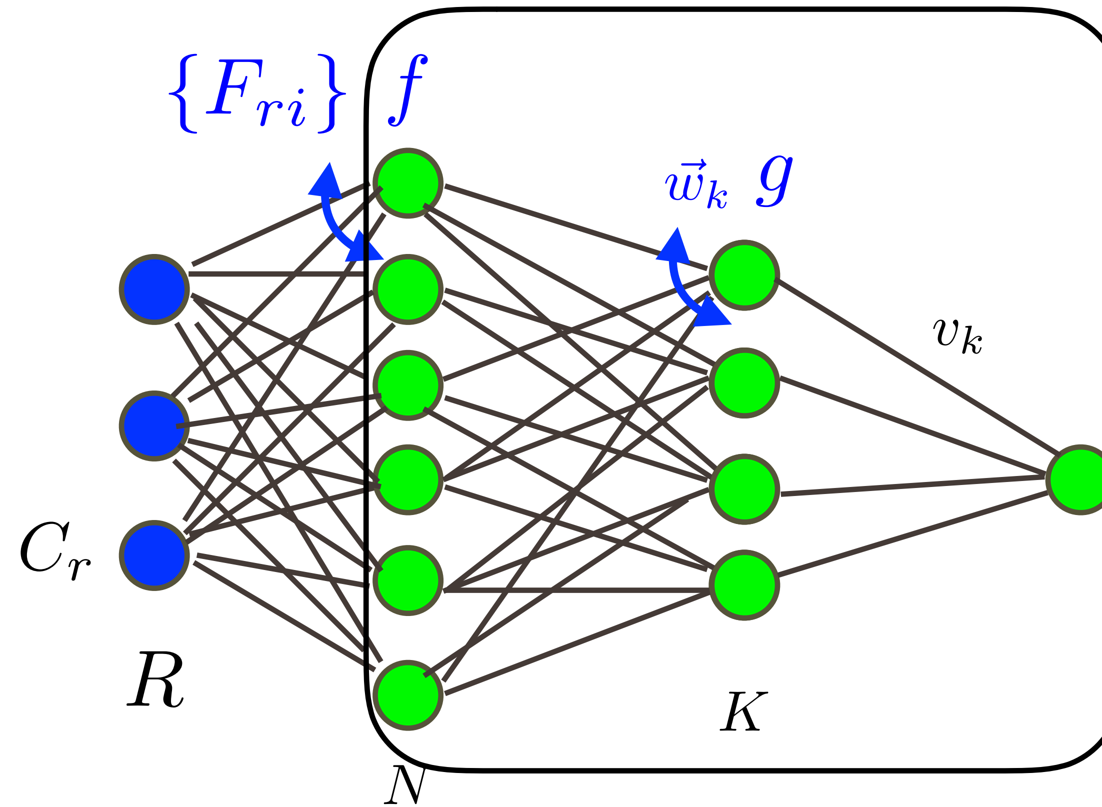
iid



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

Correlated components

iid



Connection between C_r and X_i : F_{ri}

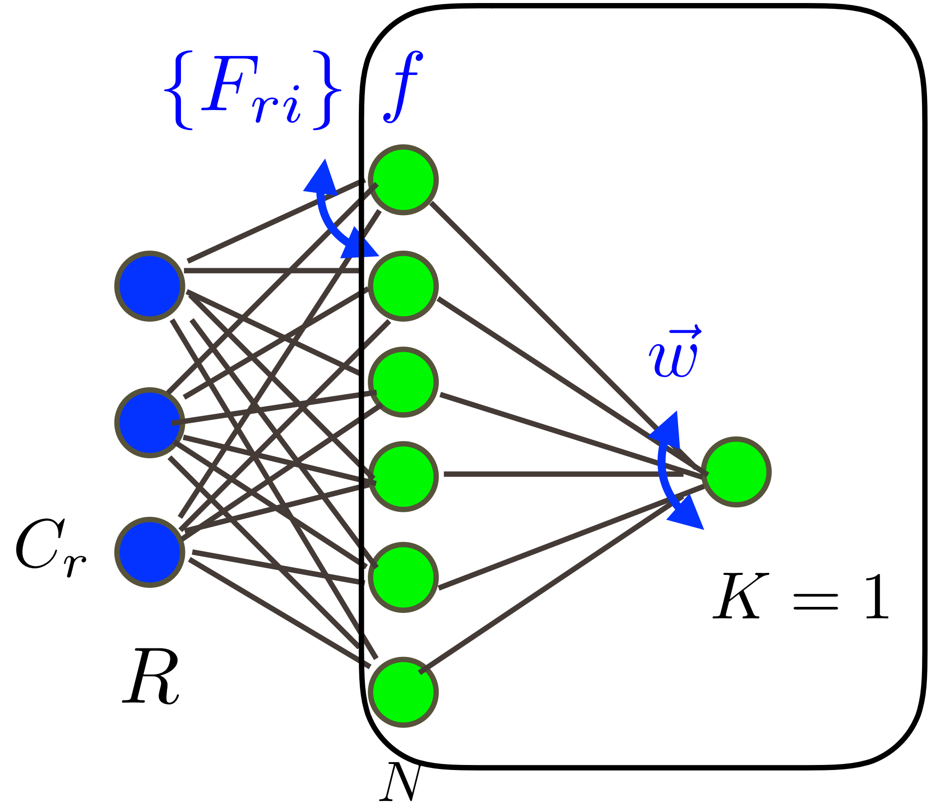
Hidden manifold model = build patterns directly in feature space, from iid coefficients in latent representation

NB: Hidden manifold and random features

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

Correlated components

iid



Connexion to Montanari Mei arXiv:1908.05335

Linear regression

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

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

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

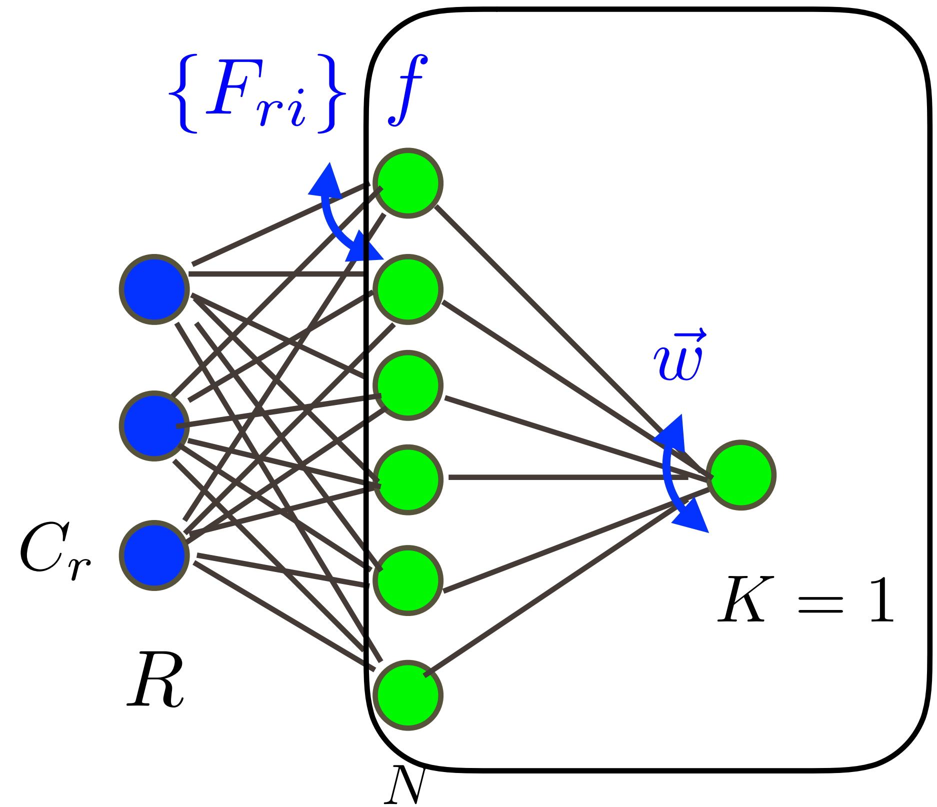
Generalization Montanari et-al 1911.01544

NB: Hidden manifold and random features

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

Correlated components

iid



Connexion to Montanari Mei
arXiv:1908.05335

Task $\quad \Phi_t(\vec{X}) = \sum_{r=1}^{R} \tilde{w}_r^m C_r$

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

with $M = 1$
linear $\tilde{g}$

Generalization Montanari et-al 1911.01544

Linear regression

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

Linear regression of random features is a special case of HMM

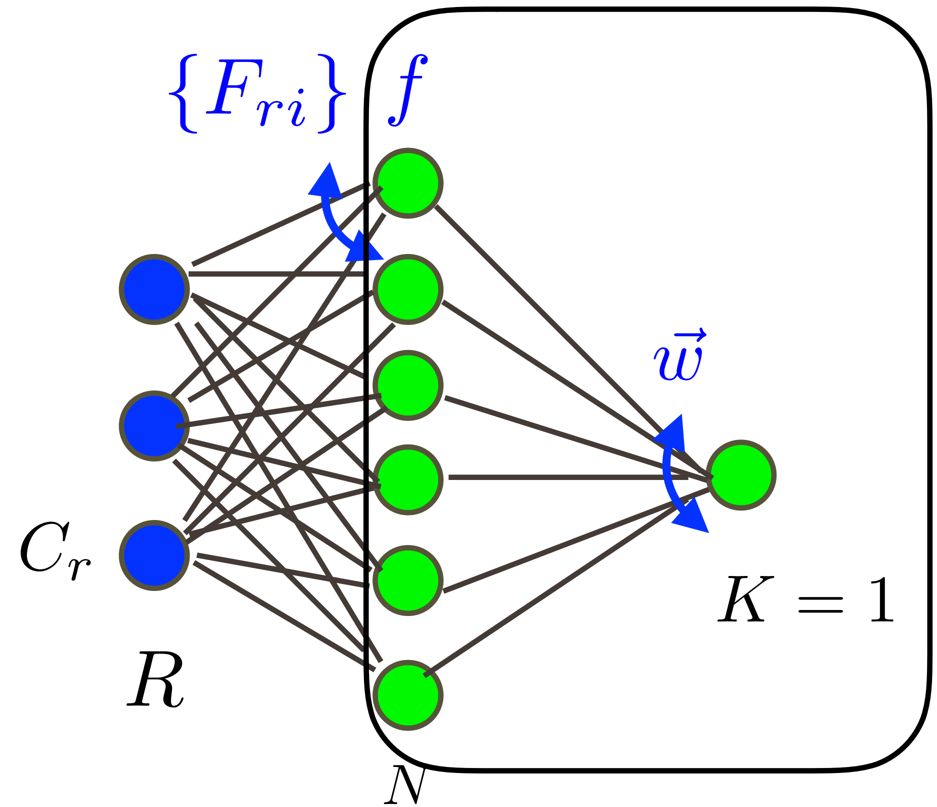
NB: Hidden manifold and random features

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

Correlated components

iid

Connexion to Montanari Mei
arXiv:1908.05335



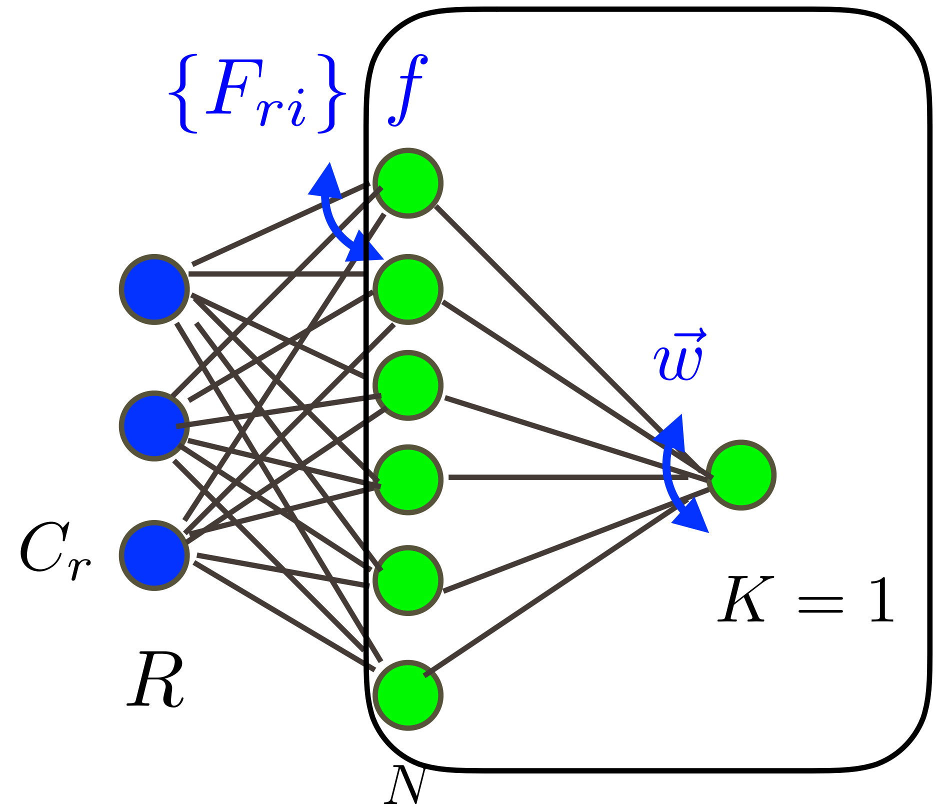
Linear regression

NB: Hidden manifold and random features

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

Correlated components

iid



Connexion to Montanari Mei arXiv:1908.05335

Linear regression

Statistically equivalent to a case where

$$X_{\mu i} = \alpha + \frac{\beta}{\sqrt{R}}\sum_{r=1}^{R} C_{\mu r} F_{ri} + \eta_{\mu i}$$

iid

Consequence of GET and

$$c = \langle f(u)^2 \rangle \qquad a = \langle f(u) \rangle \quad b = \langle u f(u) \rangle$$

NB: applies also to the case where F_{ri} are not random (but they must be « balanced »)

Summary

Data structure is important

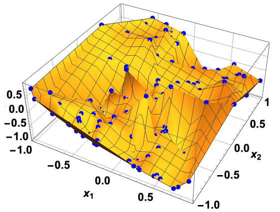
- Hidden manifolds and sub manifolds
- Combinatorial structure

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