NYU Center for Data Science

For Women in Science FONDATION L'ORÉAL

FLATIRON INSTITUTE

Statistical mechanics of learning: the physicist's view on learning theory

Summer School: Machine Learning in Quantum Physics and Chemistry

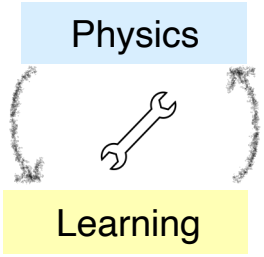
University of Warsaw

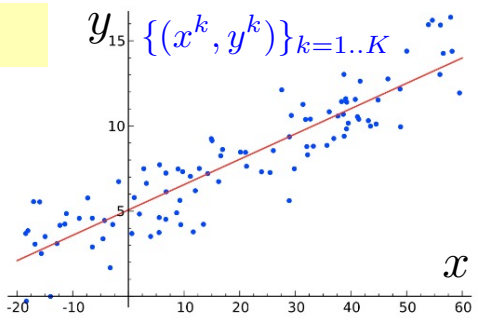
Aug 2021

Marylou Gabrié

(NYU Center for Data Science/Flatiron Institute CCMathematics)

Physics and Machine Learning



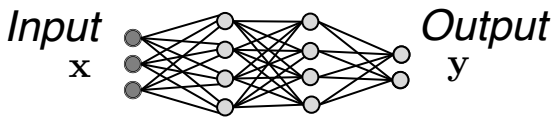


▷ Machine learning for physics

- Since the very beginning …
  - Fitting models to experimental data
  - Data denoising

– e.g. linear regression $\hat{y} = ax + b$

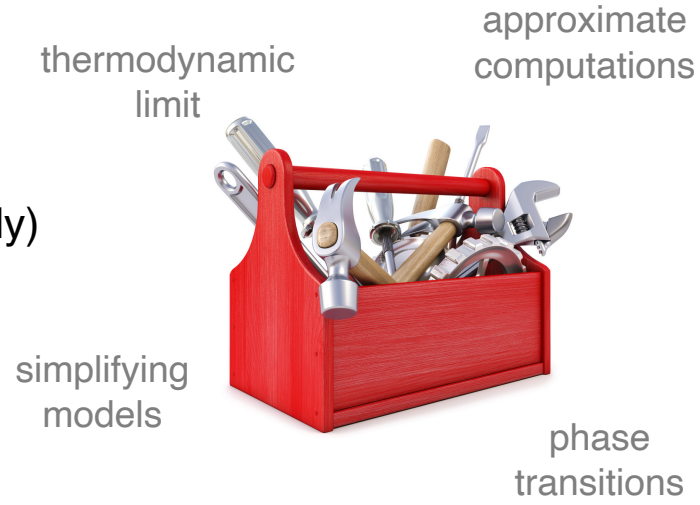
- Recent deep learning “revolution”
  - Many sophisticated application
  (main topic of this school)



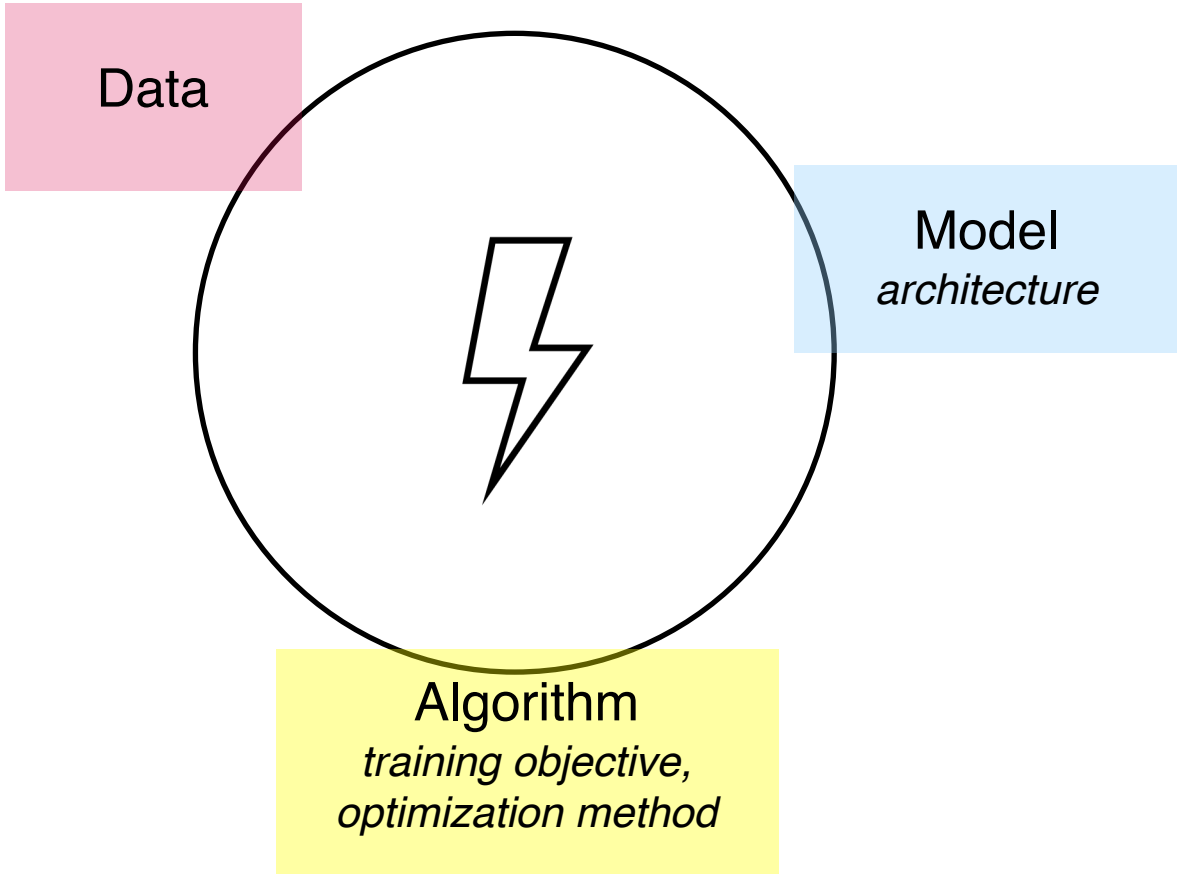
▷ Statistical physics for machine learning

- In the 80s
  - Use the statistical physics toolbox
    - to study toy models of learning (mainly)
    - to design new learning algorithms

- Since mid 2010s: New wave of interest

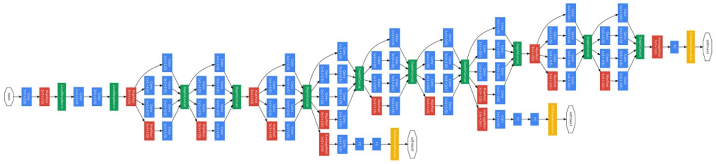


Understanding the successes of learning



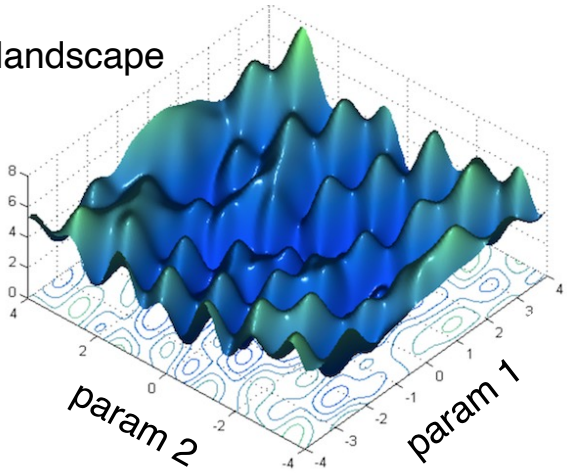
Some puzzles of modern machine learning

Google le Net – inputs 112 × 122 pixels – ~7 millions parameters



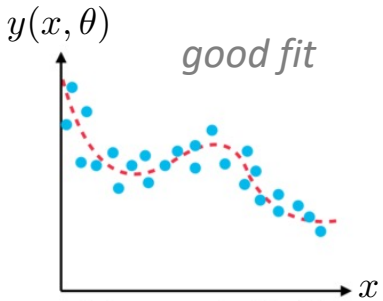
Szegedy et al, CVPR 2015

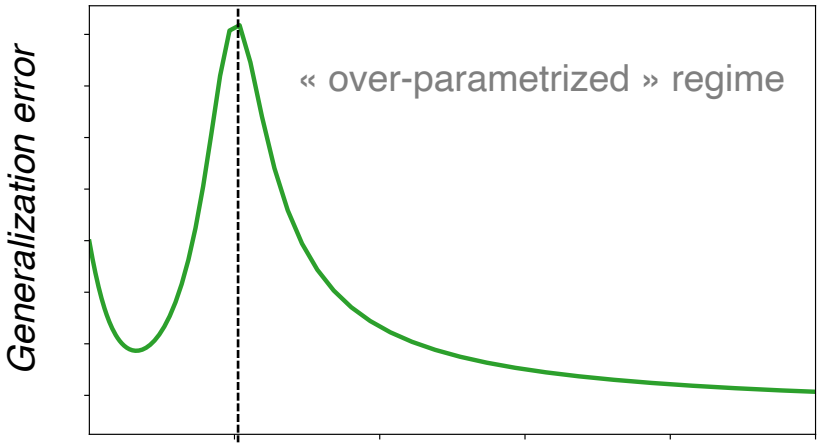
Loss landscape

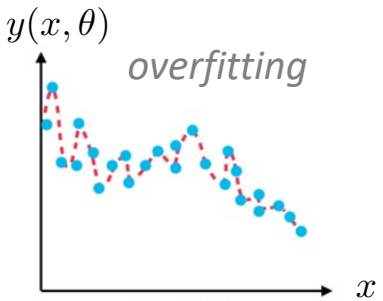


▷ Efficient optimization in non-convex settings

▷ Generalization in the « over-parametrized » regime





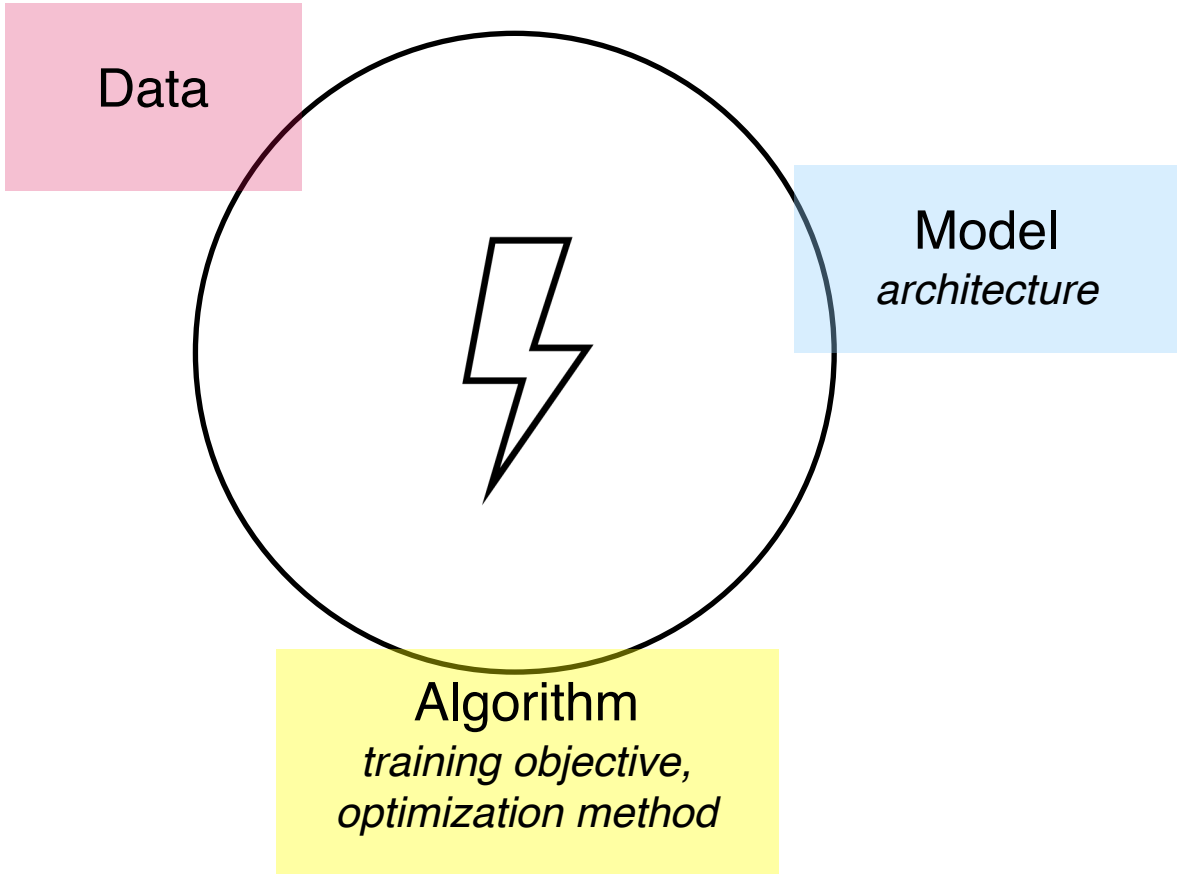


« classical regime »
dilemma biais-variance

 # parameters / model flexibility

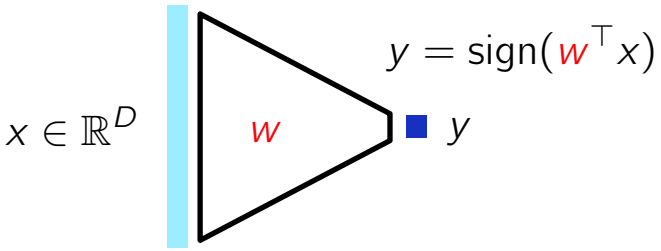
Belkin et al PNAS (2019), Spigler et al J-Phys A (2019)

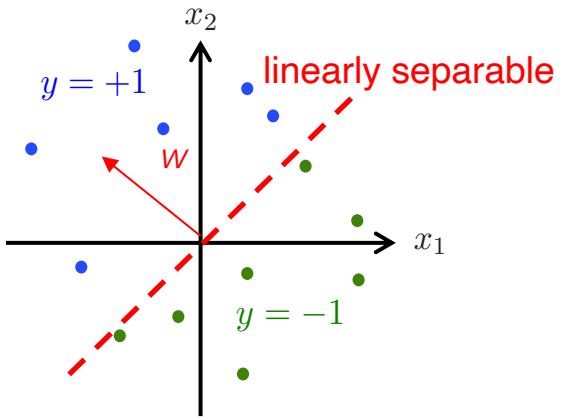
Understanding the successes of learning



Study simple ‘solvable’ models: the startegy of physicists (but not only!)

▷ Example: The perceptron – one neuron model

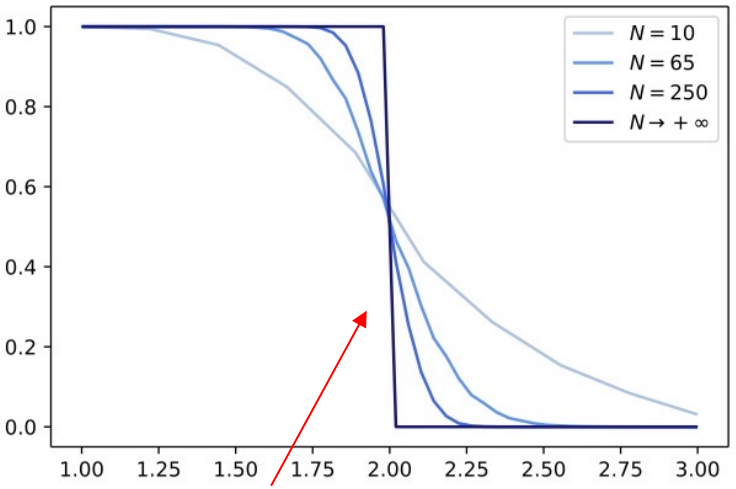




▷ First possible question: What is its capacity? How many data points can it fits?

- Training data: in general position (« linearly independent ») $\{x^{(k)}\}_{k=1}^N$
- Probability random labelling is linearly separable $\{y^{(k)}\}_{k=1}^N$ (Cover 1965)
- As a function of

$$\alpha = \frac{\#\text{ training points}}{\#\text{ parameters}} = \frac{N}{D}$$

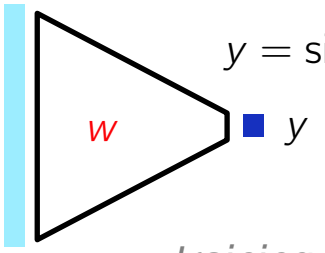


$\alpha = \frac{N}{D}$

Cover, T. M. (1965). *Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern*

Perceptron's capacity reloaded: Statistical physics computation

$x \in \mathbb{R}^D$



training data $\mathcal{D} = \{y^{(k)}, x^{(k)}\}_{k=1}^N$

▷ Volume of weight parameter space satisfying the labelling

$$V_{D,N} = \int_{\Omega(w)} \mathrm{d}w \prod_{k=1}^{N} \delta(y(x^{(k)}; w) - y^{(k)}) \quad > 0$$

each new data point
= constraint/interaction for the weights

at least one solution!

$$V_{D,N} \propto \lim_{\beta \to +\infty} \int_{\Omega(w)} \mathrm{d}w \, e^{-\beta \sum_{k=1}^{N} \left(y(x^{(k)}, w) - y^{(k)}\right)^2}$$

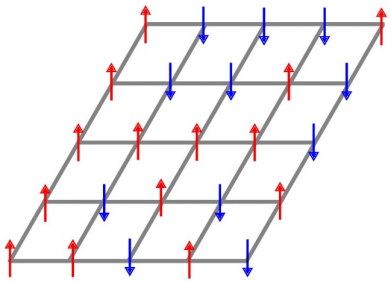
$$V_{D,N} \propto \lim_{\beta \to +\infty} \int_{\Omega(w)} \mathrm{d}w \, e^{-\beta E\left(w, \{x^{(k)}, y^{(k)}\}\right)}$$

effective energy function

▷ Why is this hard? high-dimensional + depends on training set

Gardner, E. (1987). *Maximum Storage Capacity in Neural Networks*
Krauth, W., & Mézard, M. (1989). *Storage capacity of memory networks with binary couplings*

Disordered systems physics and applications to learning



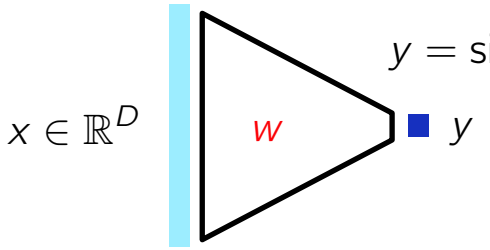
- ▷ Two types of random variables
  - state of the system $s \in \mathbb{R}^d$ « learning parameters »
    spins
  - disordered interactions $J \in \mathbb{R}^N$ « data $\{x_i, y_i\}_{i=1}^N$ = constraint on parameters to learn »
    couplings
- ▷ Boltzmann weight: $\rho(s|J) = e^{-E(s;J)}/\mathcal{Z}_J$ normalisation: $\mathcal{Z}_J = \int_{\mathbb{R}^d} ds\, e^{-E(s;J)}$
  - ex: spin glass $E(s;J) = -\sum_{(i,j)} J_{ij} s_i s_j$ $J_{ij} \sim \mathcal{N}(J_{ij}\,;\,0, 1/\sqrt{N})$ Sherrington - Kirkpatrick (1975)
    disordered couplings
- ▷ Asymptotic computations + concentration on typical states $\lim_{d\to\infty} \mathbb{E}_J[\frac{1}{d}\log \mathcal{Z}_J] \approx \frac{1}{d}\log \mathcal{Z}_J$ $d \gg 1$
  disorder average

Replica computation, variational mean-field methods, high-temperature expansion ...
Physics: Edwards & Anderson (1975), Thouless Anderson Palmer (1977) Mézard, Parisi, Virasoro (1986), etc ...
Mathematics: Talagrand (2006), Panchenko (2014), Subag (2017) etc ...

- ▷ Application to learning since(80s-90s)

Amit, Gutfreund & Sompolinsky (1985), Gardner (1987), Derrida & Nadal (1987), Peterson & Anderson (1987), Krauth & Mézard (1987), Györgyi (1990). Opper, M Haussler, D. (1991) etc ...

Perceptron's capacity reloaded: Statistical physics computation



▷ Volume of weight parameter space satisfying the labelling

$$V_{D,N} \propto \lim_{\beta \to +\infty} \int_{\Omega(w)} \mathrm{d}w \, e^{-\beta E\left(w, \{x^{(k)}, y^{(k)}\}\right)}$$

disorder variables?

▷ Asymptotic computation using the replica method:

- Fix scaling $\alpha = \dfrac{\#\text{ training points}}{\#\text{ parameters}} = \dfrac{N}{D}$ $\qquad V(\alpha) = \lim_{N \to +\infty} \mathbb{E}_{x,y}[V_{D,M}]$
- Fix distribution of data/disorder

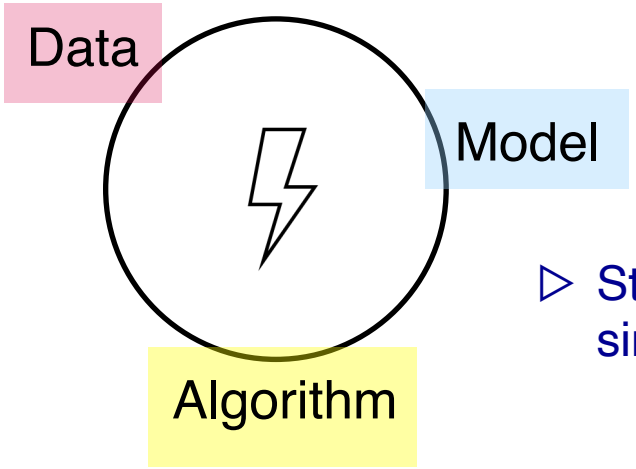
$V(\alpha_c) = 0$

Gaussian inputs $\quad p(x_i^{(k)}) = \mathcal{N}(x_i; 0, 1)$	$\alpha_c = 2$
Binary inputs $p(x_i^{(k)}) \pm 1$, binary weights $\quad w_i \in \{-1, 1\}$	$\alpha_c \approx 0.83$

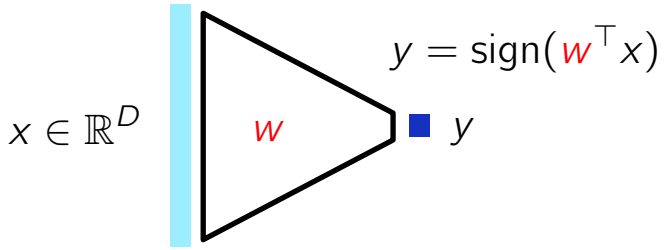
Cover, T. M. (1965). *Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern*
Gardner, E. (1987). *Maximum Storage Capacity in Neural Networks*
Krauth, W., & Mézard, M. (1989). *Storage capacity of memory networks with binary couplings*

Introduction recap

▷ Take into account 3 elements to understand the successes of deep learning

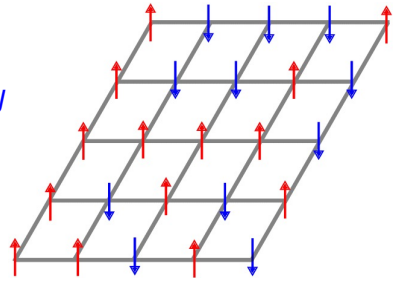


▷ Statistical mechanics of learning focuses on simple solvable models



▷ With the help of methods developped for disordered systems

$$\lim_{d\to\infty} \mathbb{E}_J[\frac{1}{d}\log \mathcal{Z}_J] \approx \frac{1}{d}\log \mathcal{Z}_J$$



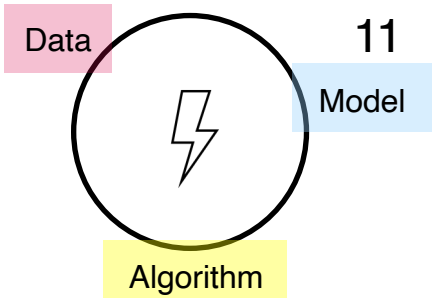
first example:
Compute the capacity of perceptron, how many samples randomly sampled can it classify?

Scope of today's lecture

Discuss classical and recent litterature to give you an idea of what can be studied in machine learning with the statistical mechanics point of view.

- ▷ The teacher-student paradigm
- ▷ Models of data structure
- ▷ Dynamics of learning

Teacher-Student paradigm: A toy model to study generalization

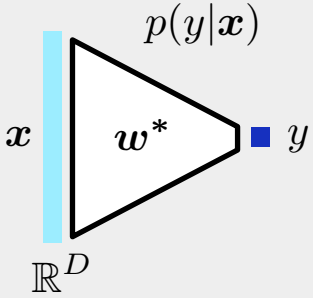


▷ Data-generating model: Teacher

Data

input distribution + *input-output rule*

$p_x(\boldsymbol{x})$



ground truth parameter distribution $p_w(\boldsymbol{w}^*)$

training data

$$\mathcal{D} = \{y^{(k)}, \boldsymbol{x}^{(k)}\}_{k=1}^{N}$$

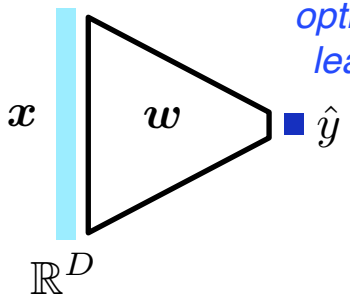
- *Can we fit the data?*
- *Can we recover the teacher rule?*

1. *Choose a model*
2. *Choose an algorithm*

▷ Learning model: Student

Model

ex: matching teacher



optimal model to learn the data!

▷ Learning strategy?

Algorithm

- Loss minimization $\min_{\boldsymbol{w}\in\mathbb{R}^D} \sum_{k=1}^{N} \ell\left(y^{(k)}, \hat{y}(\boldsymbol{x}^{(k)}, \boldsymbol{w})\right)$
- Bayesian posterior

$$p(\boldsymbol{w}|\mathcal{D}) \propto \prod_{k=1}^{N} p(y^{(k)}|\boldsymbol{w}, \boldsymbol{x}^{(k)}) p(\boldsymbol{w})$$

Given the training data and a priori on the student, what information do we have about the student?

Bayes optimal generalization error (perceptron + Gaussian or binary input)

▷ When student model = teacher model, the setting is Bayes optimal

assuming square loss

▷ Bayes optimal generalization error: $\mathcal{E}_g = \mathbb{E}_{w|\mathcal{D}}\mathbb{E}_{x,y}[(y - \hat{y}(x, w))^2]$

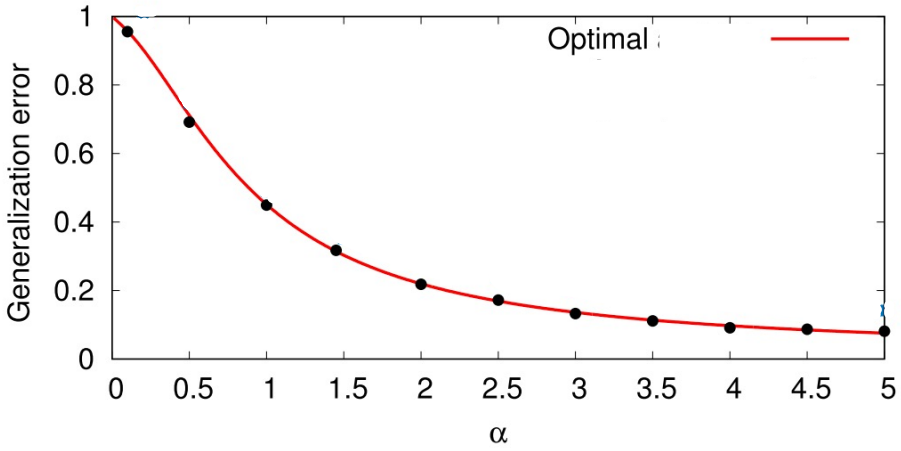
« Mean error over the data if drawing the student parameters from the posterior distribution »

▷ Using the same tool as discussed before

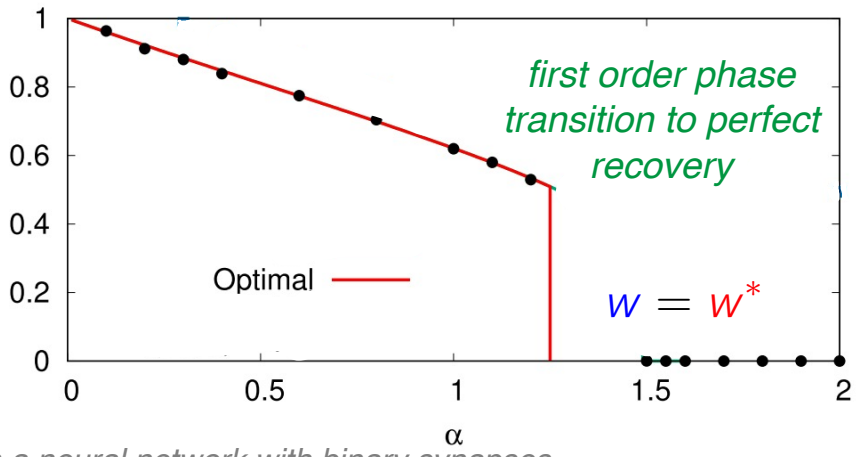
(disorder average, thermodynamic limit, replica computation)

$$\lim_{N\to+\infty} \mathbb{E}_{\mathcal{D}}\Big[\mathcal{E}_g(\mathcal{D})\Big] \quad \underset{N\to\infty}{\to} \quad \mathcal{E}_g(\alpha) \qquad \alpha = \frac{\#\text{ training points}}{\#\text{ parameters}} = \frac{N}{D}$$

$w_i \in \mathbb{R}$ sparse



$w_i \pm 1$

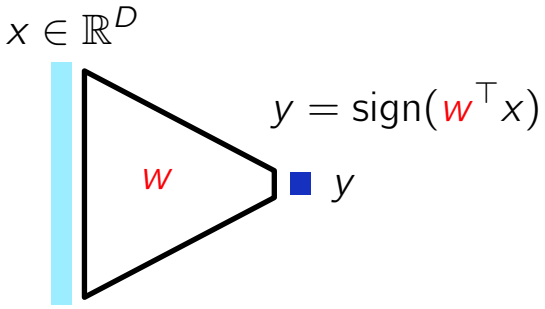


Györgyi, G. (1990). *First-order transition to perfect generalization in a neural network with binary synapses*

Krauth, W., & Mézard, M. (1989). *Storage capacity of memory networks with binary couplings*

Barbier, J. et al. (2018). *Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models*

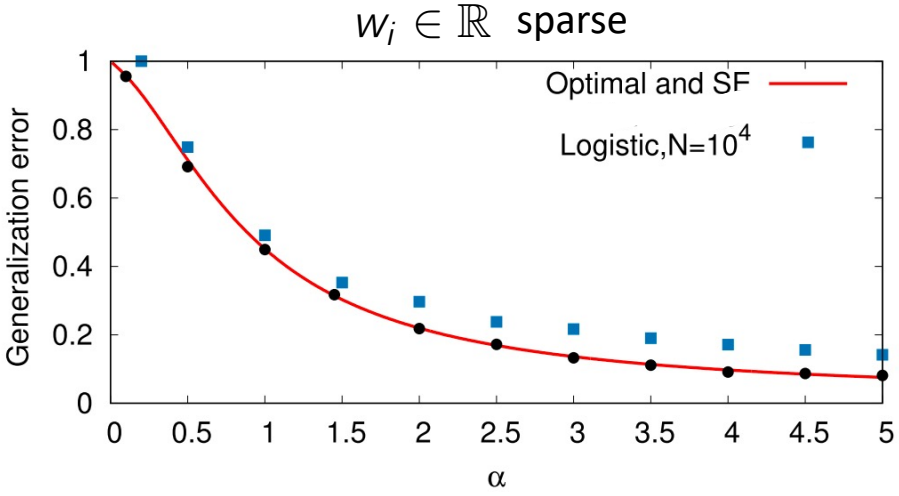
What about other (practical) algorithms? How to train in practice?

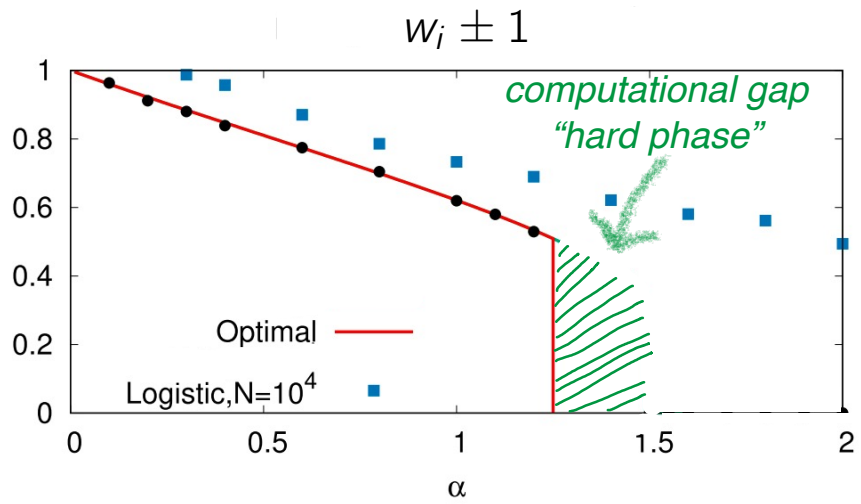


▷ Gradient descent Algorithm

For instance binary classication problem

e.g. Logistic regression



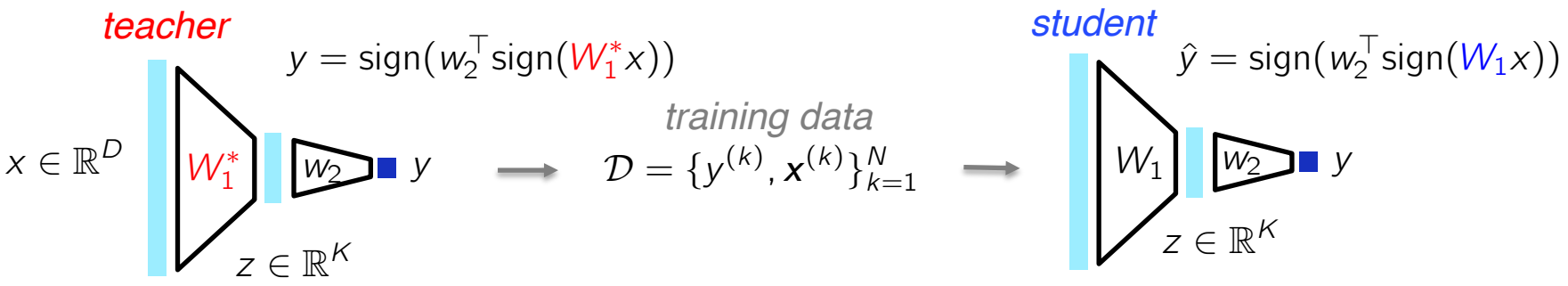


▷ Message passing algorithms (among them belief propagation) Algorithm

- Related to some mean-field computation methods
- Performance predicted with the same tools (think disorder averaged) state evolution (SE)
- Optimal among polynomial time algorithms for preceptron → computational gap

Barbier, J., et al (2018). *Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models.*

A first 2-layer architecture: committee machines



▷ High-dimensional scaling $D \to +\infty$

$$\alpha = \frac{\#\text{ training points}}{\#\text{ parameters}} = \frac{N}{D} = O(1)$$

$$K = \#\text{ hidden layer units} = O(1)$$

▷ Bayes optimal generalization

$$\lim_{N\to+\infty} \mathbb{E}_{\mathcal{D}}\Big[\mathcal{E}_g(\mathcal{D})\Big] \underset{N\to\infty}{\to} \mathcal{E}_g(\alpha)$$

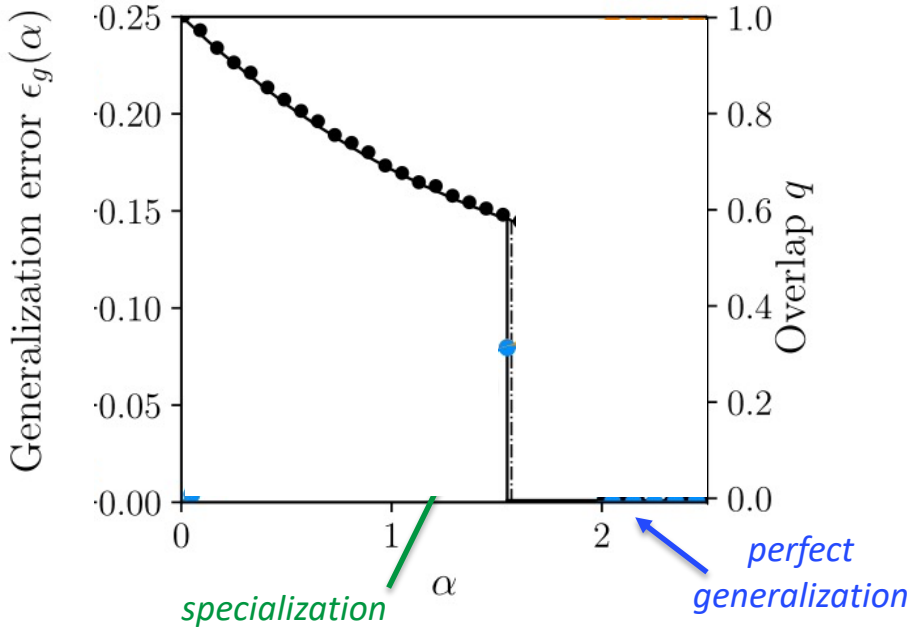
▷ Message passing algorithm

specialization transition!

K = 2, binary weights

$q_{00} =< (W_1^*)_{0,\cdot} ; (W_1)_{0,\cdot} >$

$q_{01} =< (W_1^*)_{0,\cdot} ; (W_1)_{1,\cdot} >$



Schwarze (1993). *Learning a Rule in a Multilayer Neural-Network.*
Schwarze & Hertz (1993). *Generalization in Fully Connected Committee Machines.*
Monasson et ql (2004). *Learning and Generalization Theories of Large Committee-Machines*
Aubin et al (2018). *The committee machine: Computational to statistical gaps in learning a two-layers neural network*

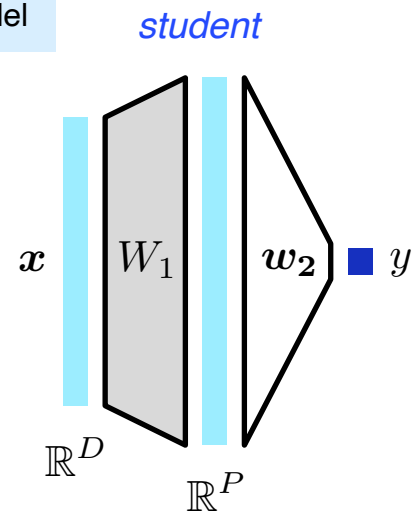
Another multi-layer example: Random features + not matched student and teacher

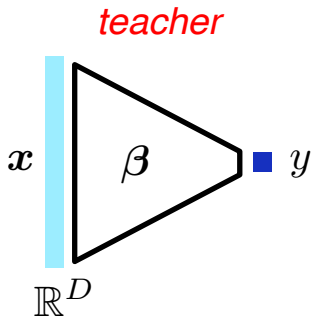
▷ Student: Alike a 2-layer neural network

Model

Rahimi, A. and Recht, B. *NeurIPS* 2018

- Learn a map $\hat{y} = \hat{f}\left(\boldsymbol{w_2}^\top \sigma(W_1 \boldsymbol{x})\right)$
- Weight parameters $\boldsymbol{w_2} \in \mathbb{R}^P, W_1 \in \mathbb{R}^{P \times D}$
- First weight matrix fixed and random (e.g., i.i.d. Gaussian entries)





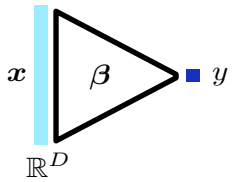
▷ Teacher: Data perceptron $y = f_0(\boldsymbol{\beta}^\top \boldsymbol{x})$

▷ Training objective $\min_{\boldsymbol{w} \in \mathbb{R}^P} \sum_{\mu=1}^{N} \ell\left(y^\mu, \boldsymbol{x}^\mu, \boldsymbol{w}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$

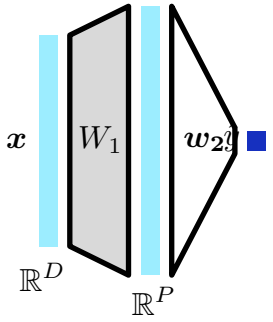
▷ Gibbs posterior formulation $\mu_\beta\left(\boldsymbol{w} \mid \{\boldsymbol{x}^\mu, y^\mu\}\right) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta\left[\sum_{\mu=1}^{N} \ell(y^\mu, \boldsymbol{x}^\mu \cdot \boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\right]}$

$$= \frac{1}{\mathcal{Z}_\beta} \underbrace{\prod_{\mu=1}^{N} e^{-\beta \ell(y^\mu, \boldsymbol{x}^\mu \cdot \boldsymbol{w})}}_{\equiv P_y(\boldsymbol{y} \mid \boldsymbol{w} \cdot \boldsymbol{x}^\mu)} \underbrace{\prod_{i=1}^{P} e^{-\frac{\beta\lambda}{2} w_i^2}}_{\equiv P_w(\boldsymbol{w})}$$

An important intermediary result: Gaussian equivalence principle



$$y = f^0(\boldsymbol{\beta}^\top \boldsymbol{x})$$



$$\hat{y} = \hat{f}\left(\boldsymbol{w_2}^\top \sigma(W_1 \boldsymbol{x})\right)$$

▷ Generalization error in the high-dimensional limit

Gausian input $\boldsymbol{x} \sim \mathcal{N}(0, I_D)$

input dimension $D \to \infty$

parameters $P \to \infty$

training samples $N \to \infty$

$$N/D = \alpha = O(1)$$
$$D/P = \gamma = O(1)$$

$$\lim_{N\to\infty} \mathcal{E}_g = \mathbb{E}_{\lambda,\nu}\left[\left(f^0(\nu) - \hat{f}(\lambda)\right)^2\right]$$

"$\nu = \boldsymbol{\beta}^\top \boldsymbol{x}$"

"$\lambda = \boldsymbol{w_2}^\top \sigma(W_1 \boldsymbol{x})$"

where the scalar fields are jointly Gaussian.

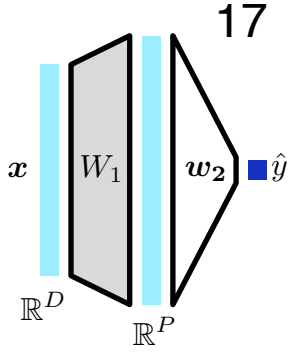
▷ Compute covariance of Gaussian fields?

- random matrix theory tools or replica method from statistical physics!

Karoui, N. El. (2010). *Annals of Statistics*, Mei, S., & Montanari, A. (2019). *ArXiv 1908.05355*. Gerace, F., et al *ICML (2020),* Hu, H., & Lu, Y. M. (2020). *ArXiv 2009.07669*. Goldt, S., et al (2020). *ArXiv, 2006.14709*

Generalization in the random features model

$$\min_{\boldsymbol{w}\in\mathbb{R}^P} \sum_{\mu=1}^{N} \frac{1}{2}\left(y^\mu - \boldsymbol{w}^\top \sigma(W_1 \boldsymbol{x}^\mu)\right)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$



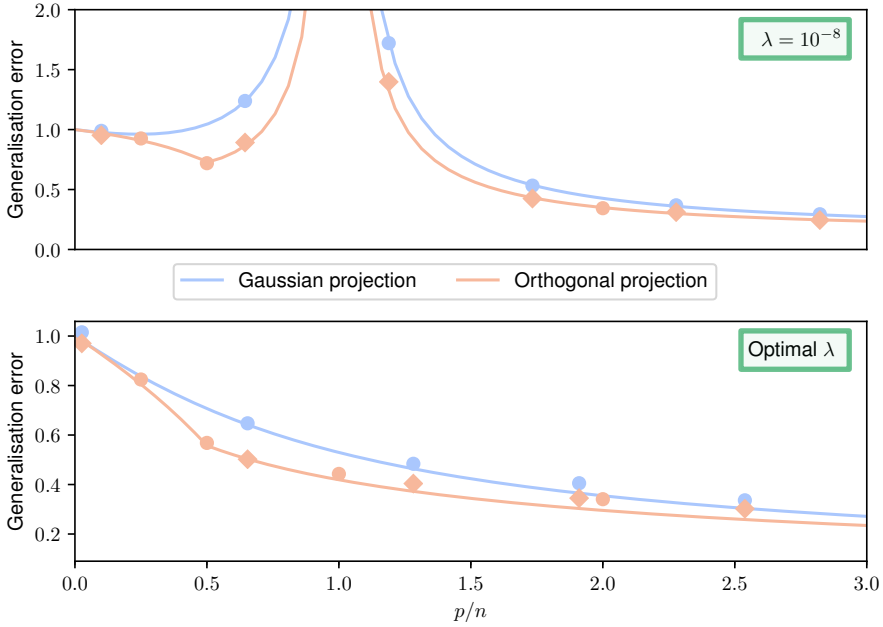
Regression teacher — Data

$$y = \boldsymbol{\beta}^\top \boldsymbol{x} + \epsilon \quad \text{with } \epsilon \sim \mathcal{N}(0\,,\,\Delta)$$

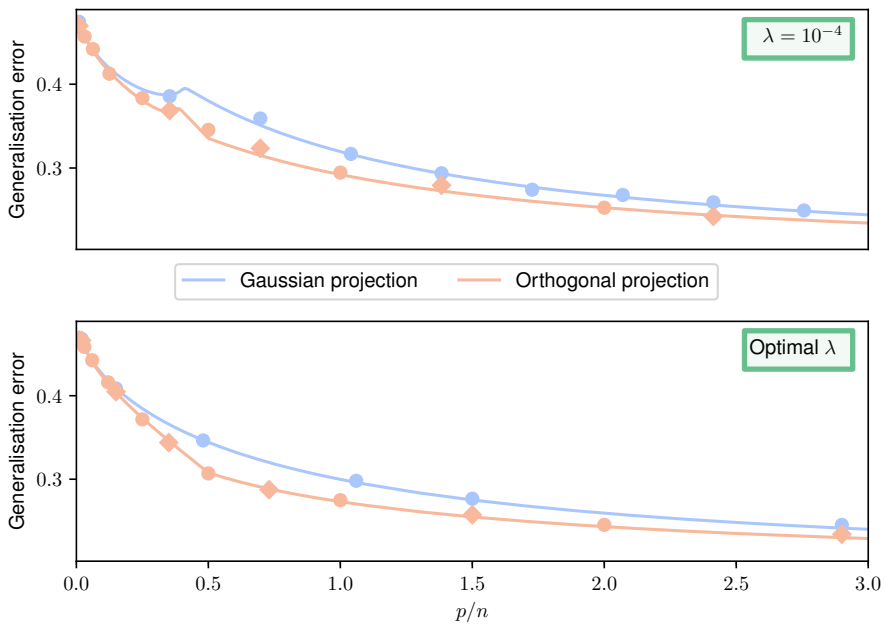
Classification teacher — Data

$$y = \begin{cases} \text{sign}(\boldsymbol{\beta}^\top \boldsymbol{x}) & \text{with prob. } \Delta \\ -\text{sign}(\boldsymbol{\beta}^\top \boldsymbol{x}) & \text{with prob. } 1-\Delta \end{cases}$$

double descent!



slight double descent

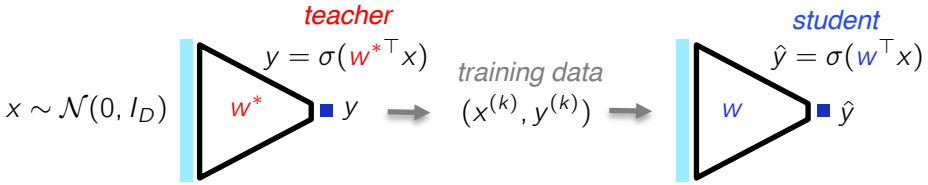


double descent cancelled by appropriate regularization

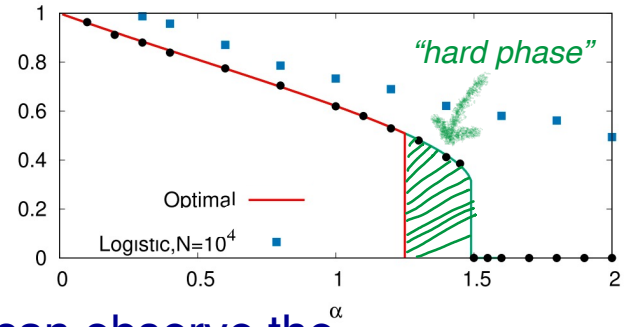
Gerace, F., et al *ICML (2020)*

The teacher-student paradigm: Recap

- The teacher-student scenario consist in creating simple synthetic problem of learning



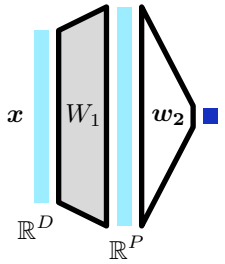
- Matched teacher and student: **Bayes optimal setting**. Information theoretic limit is not always achievable in reasonnable time: **computational gaps**.



examples:
perceptron +
committee machine (2-layer)

- In mis-matched case we can observe the double descent phenonenon in overparametrized model.

example:
perceptron teacher w.
random feature student



QUESTIONS?

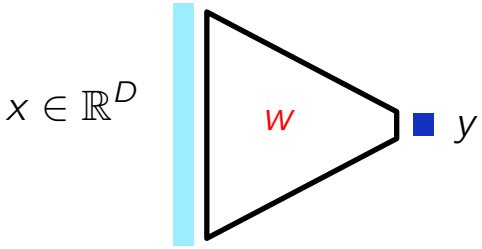
Scope of today's lecture

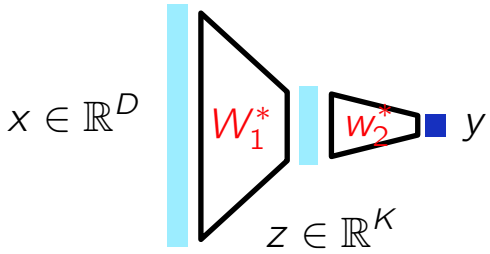
Discuss classical and recent litterature to give you an idea of what can be studied in machine learning with the statistical mechanics point of view.

- ▷ The teacher-student paradigm
- ▷ Models of data structure
- ▷ Dynamics of learning

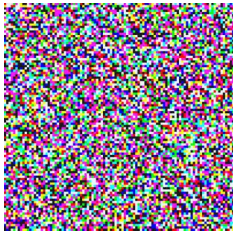
Towards more realistic models of data

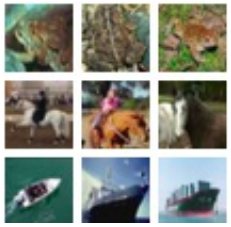
- Sturcture in between the input and outputs
  (e.g. Neural network teacher models)





- Structure in the input



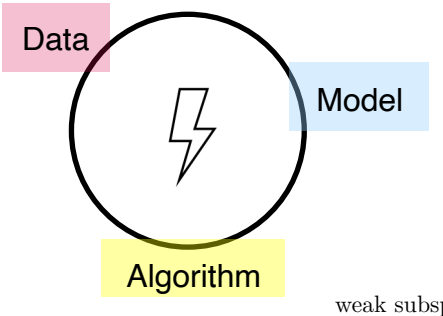


solvability / interpretability

degree of structure / realism

white noise ? *CIFAR images*

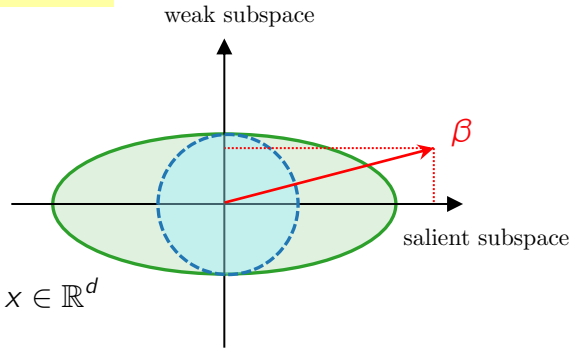
Simple model for data structure: Salient and weak features model



▷ Structure in the inputs Data $x \sim \mathcal{N}(0, \Sigma_x)$

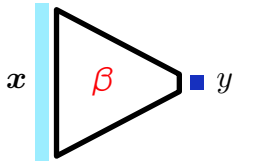
2 cases: - isotropiques $\Sigma_x = I_d$

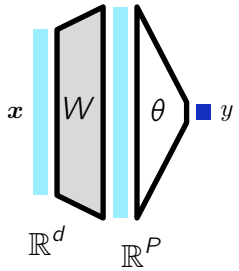
- anisotropique structurée $\Sigma_x = \begin{bmatrix} \sigma_{x,1} I_{\phi_1 d} & 0 \\ 0 & \sigma_{x,2} I_{\phi_2 d} \end{bmatrix}$



▷ Structure input-output (teacher) and learning model (student)

(as we just discussed)

teacher $y = f^0(\beta^\top x)$ 

→ *training data* $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ → student $\hat{y} = \hat{f}\left(\theta^\top \sigma(Wx)\right)$ 

▷ Generalization error (high-dimensional limit)

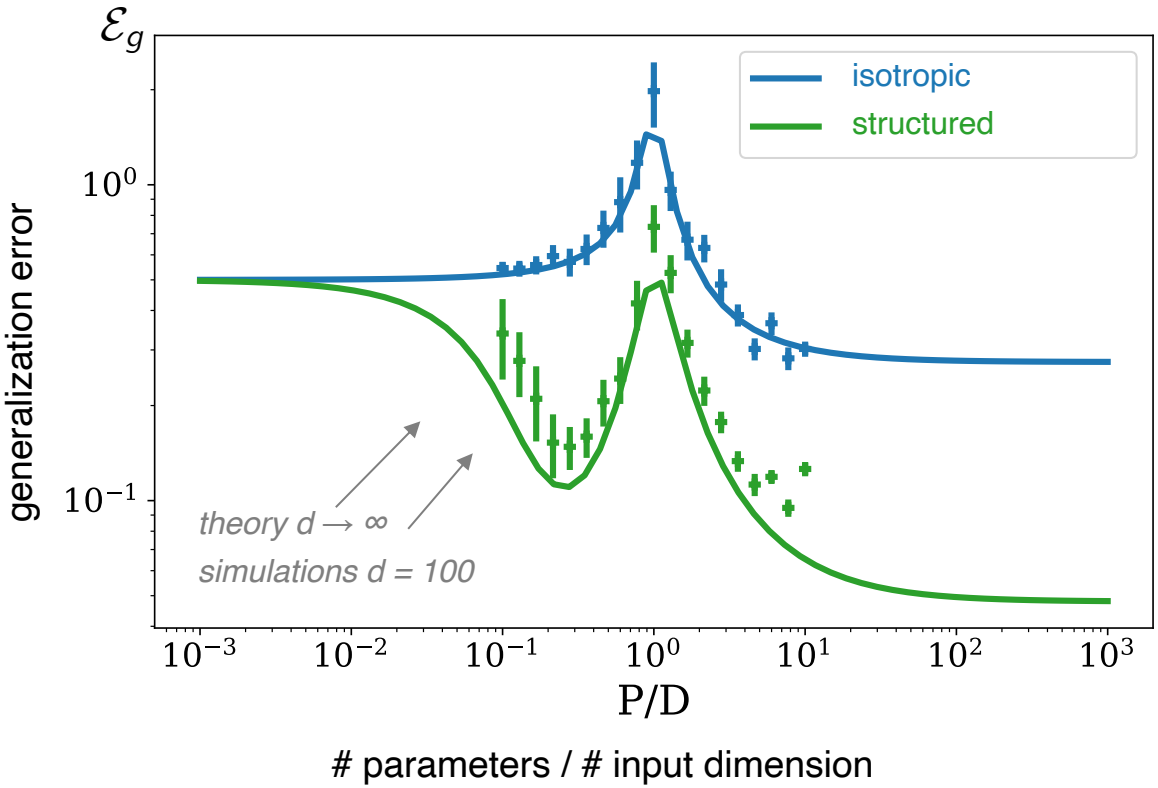
(replica computation)

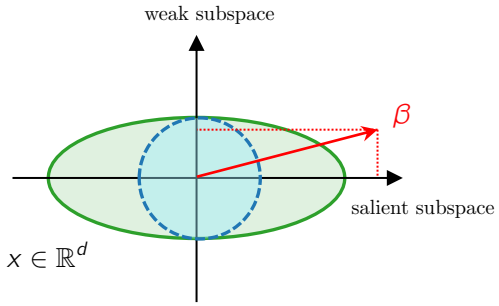
$$\mathcal{E}_g = \mathbb{E}_{\boldsymbol{x}}\left[(y(x) - \hat{y}(x;\theta))^2\right] \approx \lim_{d\to\infty} \mathbb{E}_{\boldsymbol{\beta}}\mathbb{E}_{\mathcal{D}}\left[\mathcal{E}_g\right] = \mathcal{E}_{g,\ell}(\alpha, \gamma, \sigma_{x,12}, \sigma_{\beta,12})$$

$$(d \to \infty, P \to \infty, N \to \infty, \gamma = P/d = O(1), \alpha = N/d = O(1))$$

d'Ascoli, S., Gabrié, M., Sagun, L., & Biroli, G. (2021). More data or more parameters ? Investigating the effect of data structure on generalization. *ArXiv Preprint*, *2103.05524*.

Effect of data structure on double descent

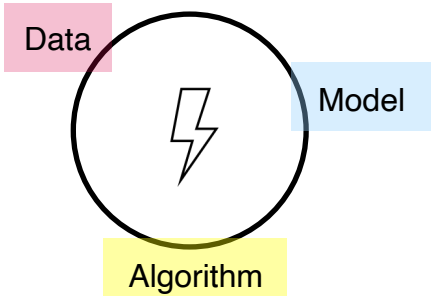




▷ The structure in the data is detected during training before the overfitting peak

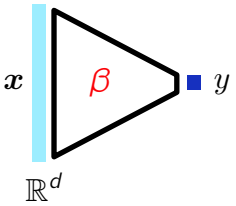
▷ Double descent is exacerbated by the data structure when aligned with the task

Interplay between data structure and loss function in classification 1/2



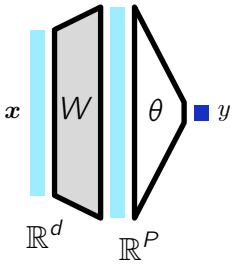
teacher

$$y = \mathrm{sign}(\beta^\top x)$$



student

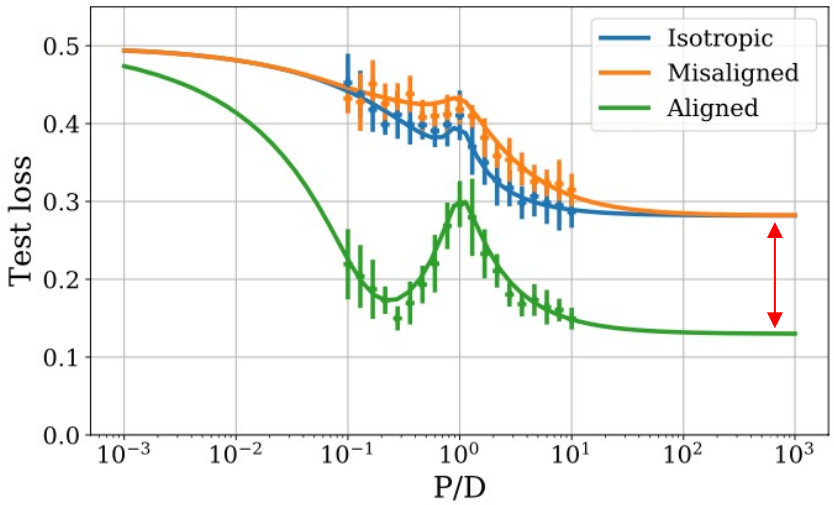
$$\hat{y} = \hat{f}\left(\theta^\top \sigma(Wx)\right)$$



▷ Training objective $\min_{\theta\in\mathbb{R}^P} \sum_{\mu=1}^{N} \ell\left(y^\mu, x^\mu, \theta\right) + \frac{\lambda}{2}\|\theta\|_2^2$

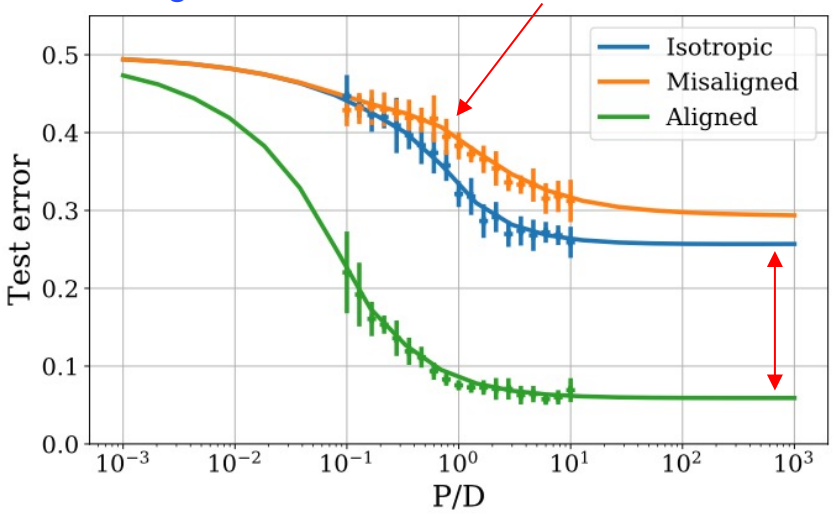
Algorithm

square loss



logistic loss

overfitting peak attenuated

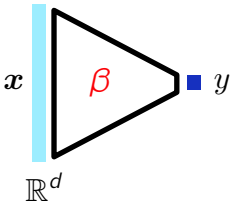


logistic loss takes most advantage of structure

Interplay between data structure and loss function in classification 2/2

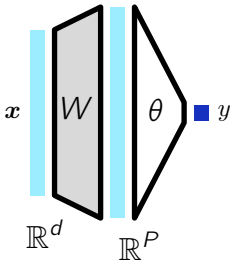
teacher

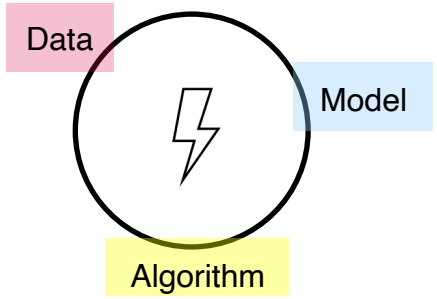
$$y = \mathrm{sign}(\beta^\top x)$$



student

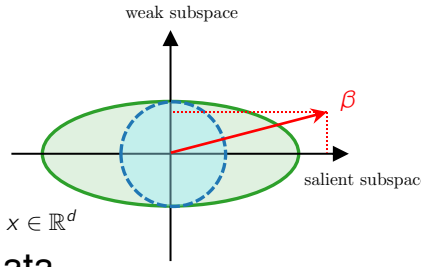
$$\hat{y} = \hat{f}\left(\theta^\top \sigma(Wx)\right)$$



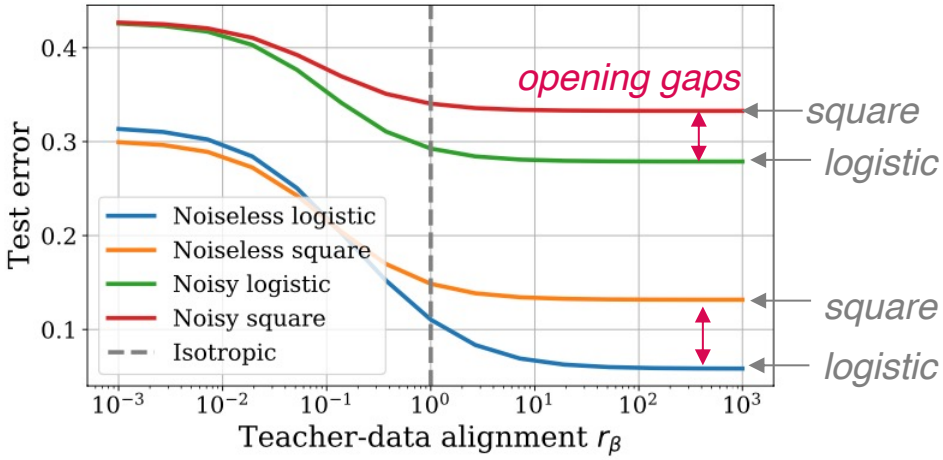


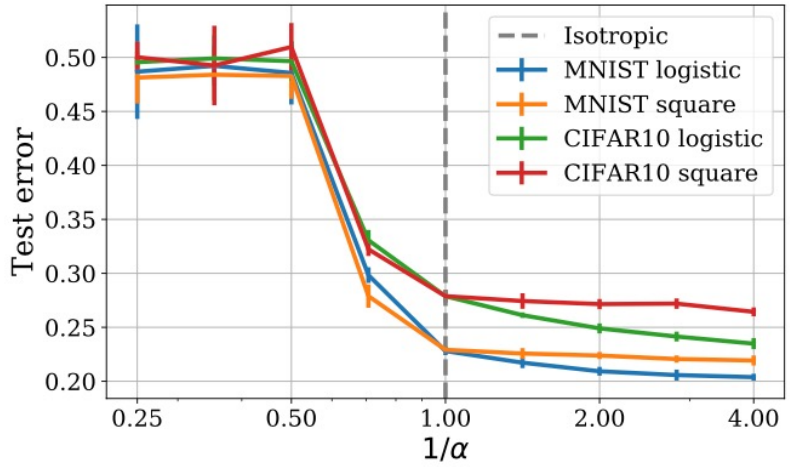
▷ Training objective $\min_{\theta\in\mathbb{R}^P} \sum_{\mu=1}^{N} \ell\left(y^\mu, x^\mu, \theta\right) + \frac{\lambda}{2}\|\theta\|_2^2$

Algorithm



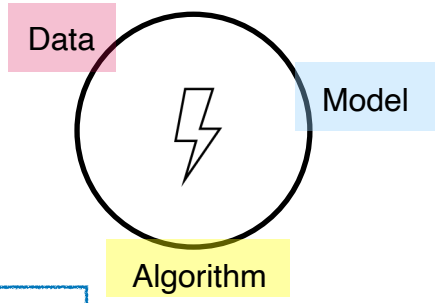
Real data



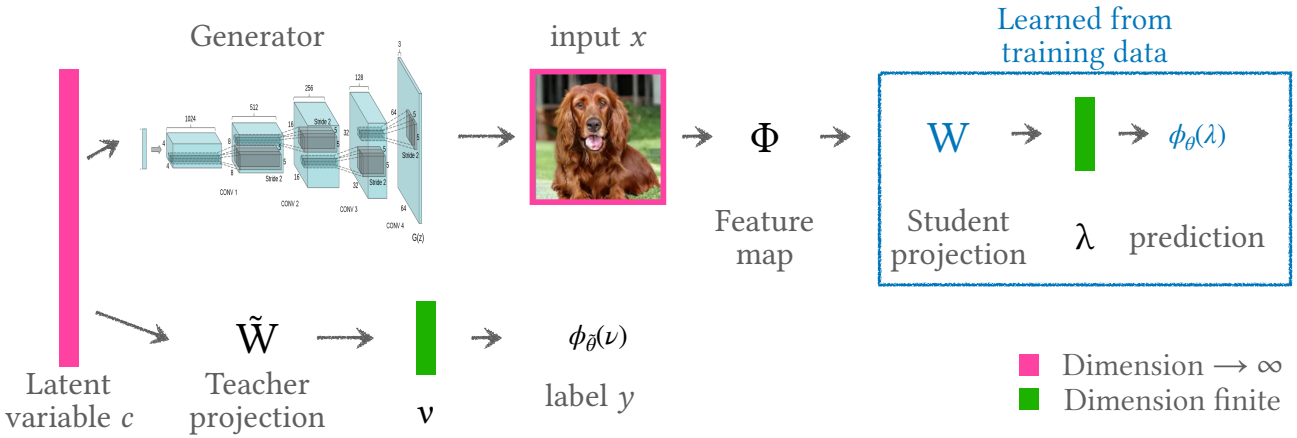


▷ Simple solvable model helps intuition further confirmed by real data experiments

Towards more and more realistic data in the analysis: trained generative models



▷ Using pretrained generators in synthetic data models

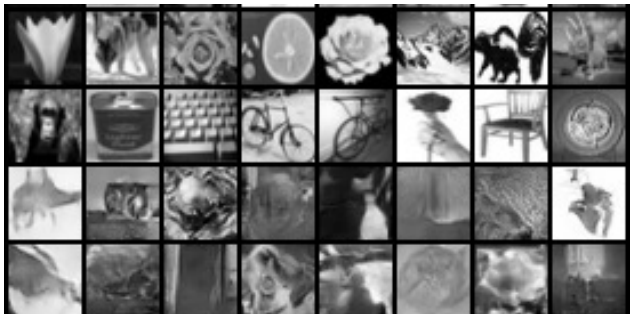


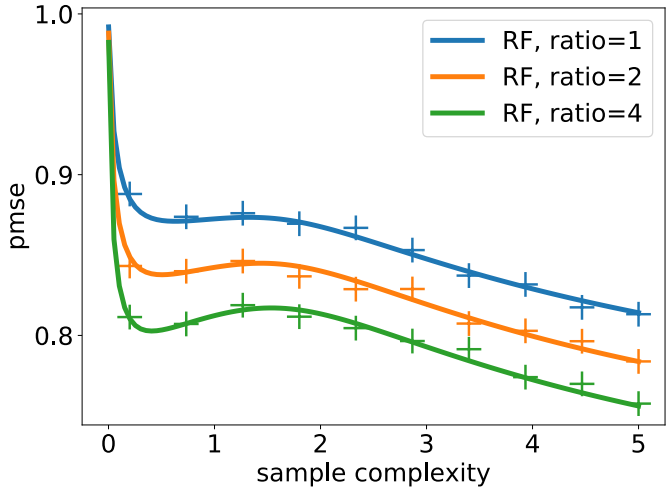
Gaussian equivalence principle here crucial!

▷ Example:
Predicting random features learning on CIFAR

Real

Generated

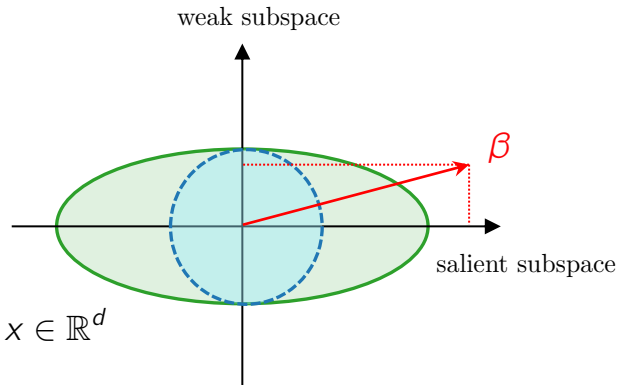




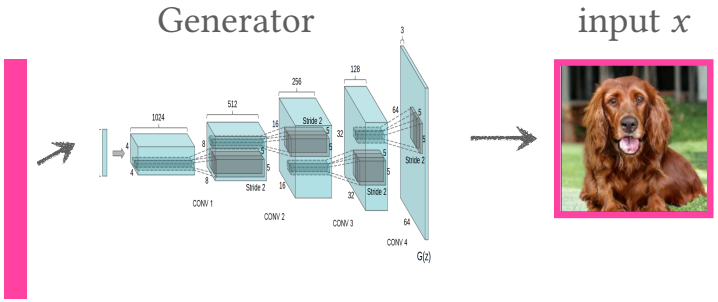
Goldt, S. et al (2020). The Gaussian equivalence of generative models for learning with shallow neural networks. MSML

Models of data structures: Recap

- With simple models of data structures we can start to understand its impact on generalization

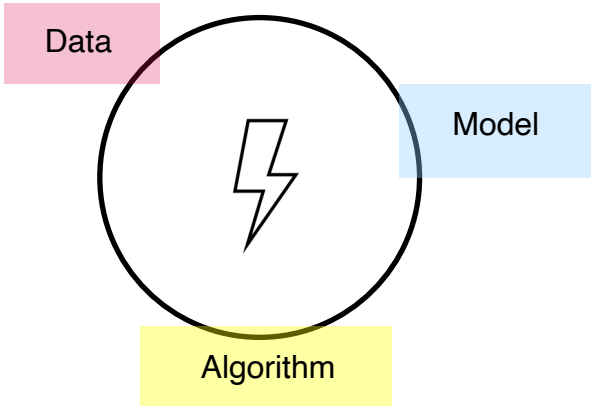


- Synthetic (trained) model can mimic real data and remain amenable to analysis

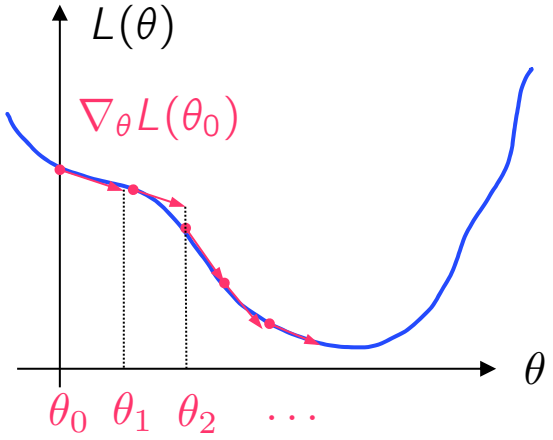


QUESTIONS?

More analysis of algorithms?



- (Stochastic) Gradient descent??
  - Loss $L(\theta) = \frac{1}{N}\sum_{k=1}^{N} \ell[y^{(k)}, \hat{y}(x^{(k)}, \theta)]$
  - Gradient update $\theta^{t+1} \leftarrow \theta^t - \eta \nabla_\theta L(\theta^t)$



Scope of today's lecture

Discuss classical and recent litterature to give you an idea of what can be studied in machine learning with the statistical mechanics point of view.

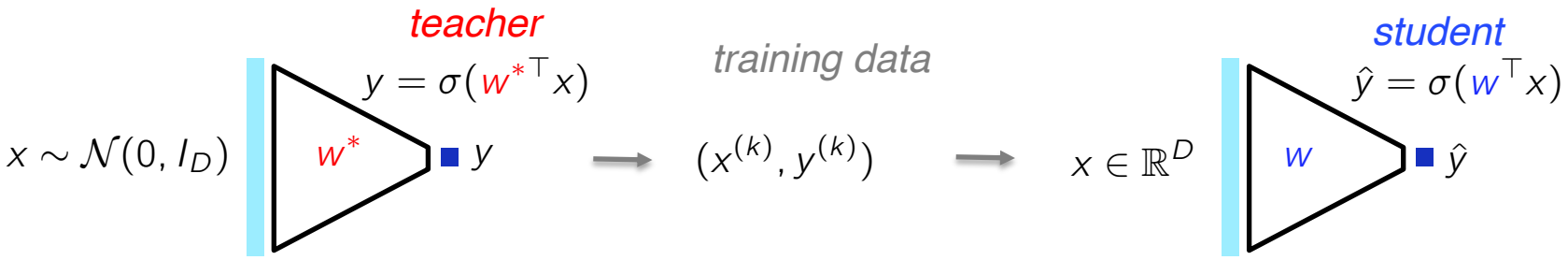
- ▷ The teacher-student paradigm
- ▷ Models of data structure
- ▷ Dynamics of learning

As soon as in the 90s: Analysis of online learning

▷ Online gradient descent

- Stream of data $(x^{(1)}, y^{(1)}),\ (x^{(2)}, y^{(2)}),\ \cdots,\ (x^{(k)}, y^{(k)}),\ \cdots$ *each sample seen once, size of dataset = # iterations*
- Parameter update with each sample $w^{k+1} = w^k - \eta \nabla_w L\left(w^k; (x^{(k)}, y^{(k)})\right)$

▷ Analysis principle: perceptron example



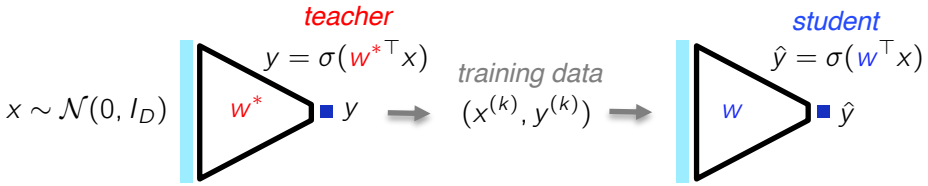
- Square loss for one sample $L(x; w, w^*) = \frac{1}{2}\left(y(x) - \hat{y}(x)\right)^2 = \frac{1}{2}\left(\sigma(w^{*\top} x) - \sigma(w^\top x)\right)^2$
- Parameter update $w^{t+1} - w^t = \eta\left(\sigma(w^{*\top} x) - \sigma(w^\top x)\right)\nabla\sigma(w^\top x)$
- Continuous time limit + high dimensional limit + disorder variable average

made possible by having uncorrelated samples at each update !

Biehl & Riegler. *On-Line Learning with a Perceptron*, 1994, Biehl & Schwarze. *Learning by on-line gradient descent*, 1995. Riegler & Biehl. *On-line backpropagation in two-layered neural networks*, 1995. Saad & Solla. *Exact Solution for On-Line Learning in Multilayer Neural Networks*, 1995. Saad & Solla. *On-line learning in soft committee machines*. 1995.

As soon as in the 90s: Analysis of online learning

▷ Analysis principle: perceptron example



- Follow training in terms of overlaps

$$Q^{(k)} = w^{(k)\top} w^{(k)} \quad \text{« self overlap »}$$
$$R^{(k)} = w^{*\top} w^{(k)} \quad \text{« teacher-student overlap »}$$

- Continuous time limit + high dimensional limit + disorder variable average

$$w^{t+1} - w^t = \eta dt \left(\sigma(w^{*\top} x) - \sigma(w^\top x)\right) \nabla\sigma(w^\top x)$$

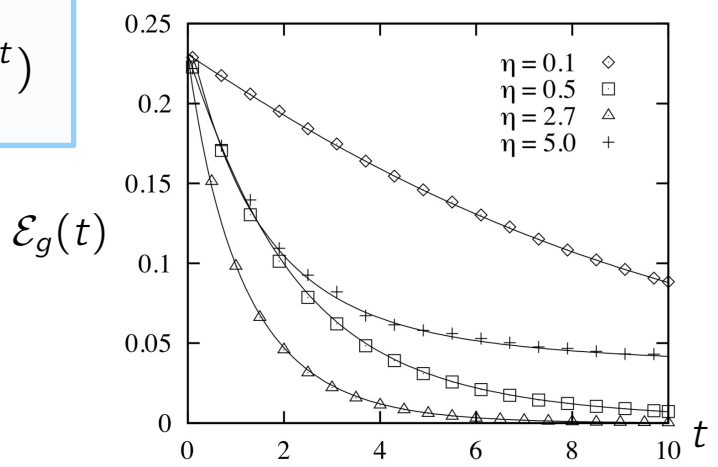
→ closed set of equations on the overlaps

$$\begin{cases} \dfrac{\mathrm{d}R}{\mathrm{d}t} = g_R(R^t, Q^t) \\ \dfrac{\mathrm{d}Q}{\mathrm{d}t} = g_Q(R^t, Q^t) \end{cases}$$

- Generalization error:

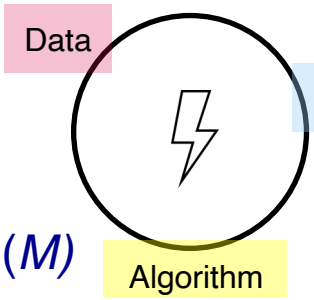
$$\mathcal{E}_g(w^{(k)}, w^*) = \mathbb{E}_x \left[\frac{1}{2}\left(y^*(x) - \hat{y}(x, w^{(k)})\right)^2\right]$$
$$= \mathcal{E}_g(R^{(k)}, Q^{(k)}) \underset{N\to+\infty, dt\to 0}{\longrightarrow} \mathcal{E}_g(t)$$

different convergence rates for different learning rate

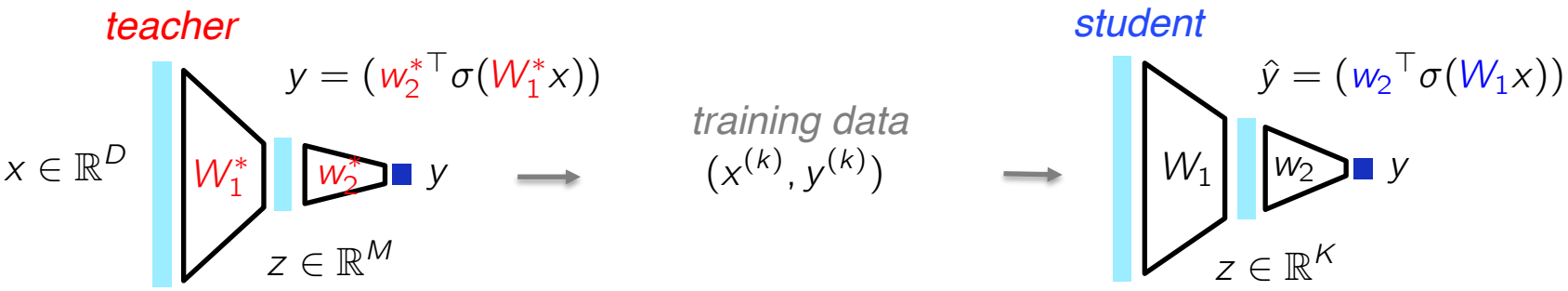


Biehl & Schwarze. (1995) *Learning by on-line gradient descent.*

Committee machines online dynamics



▷ Two layer model with finite number of hidden units
+ possibly more/less hidden units in student (*K*) than teacher (*M*)



▷ Define overlaps again

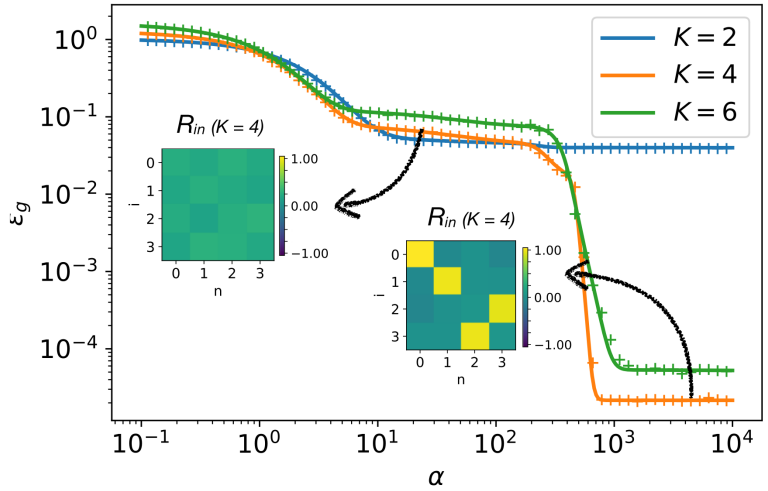
$$R^{(k)} = W_1{}^* W_2{}^{(k)\top} \in \mathbb{R}^{K\times M}$$
$$Q^{(k)} = W^{(k)} W^{(k)\top} \in \mathbb{R}^{K\times K}$$

→ *generalization* $\varepsilon_g(t)$

▷ First only learning first layer + matched models

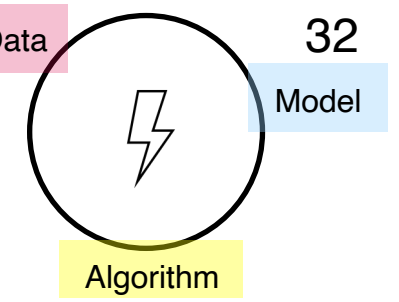
$w_2 = w_1$

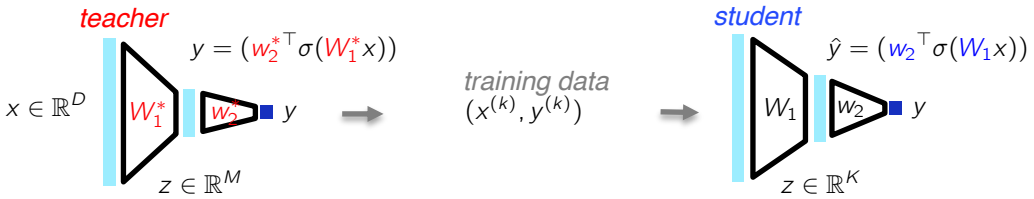
specialization transition:
student neurons match teacher neurones
→ escape plateau in dynamics



Goldt, S. et al (2019). Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Neural Information Processing Systems*, Neurips.*1906.08632*.

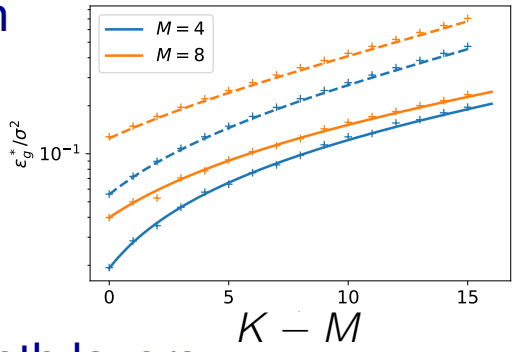
Committee machines online dynamics: overparametrization and feature learning



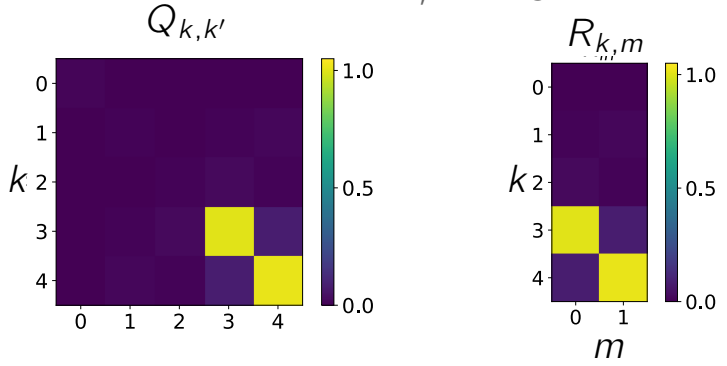


▷ Over parametrization

not helping if the second layer is not learned!



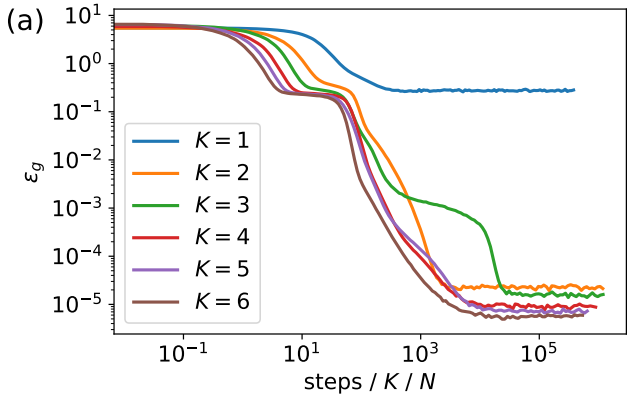
$M = 2, K = 5$

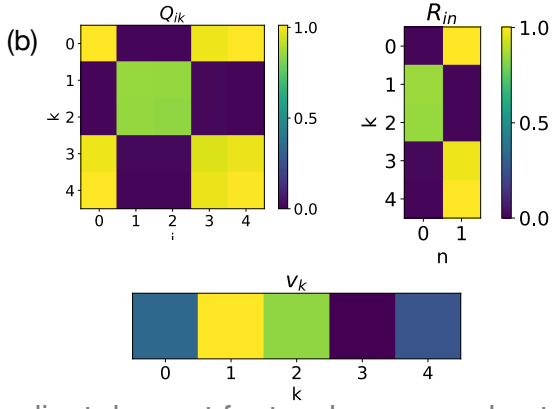


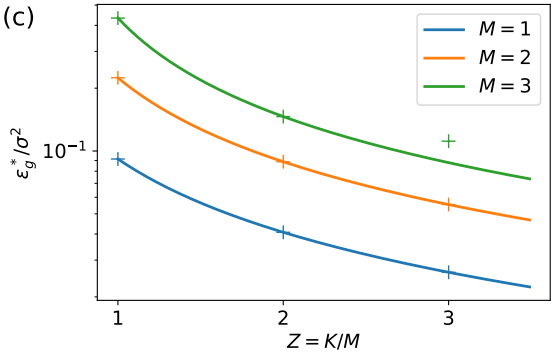
▷ Feature learning = both layers

in the analysis

$w_2^* = (1, 1, ..1) \in \mathbb{R}^K$







overparametrization beneficial by denoising!

Goldt, S. et al (2019). Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. Neurips.

Can we go beyond online gradient descent? Analysis of multi-pass stochastic gradient descent

Involved but possible in simple models

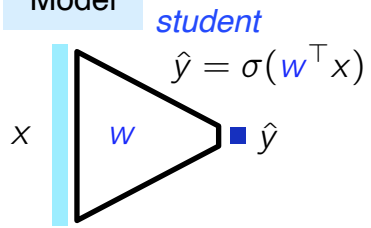
▷ Use of dynamical mean-field theory (DMFT)

Mézar, Parisi, Virasoro (1987), Sompolinsky, Crisanti, Sommers (1988) etc...

▷ Example: Perceptron classification of Gaussian mixtures

Data

Model



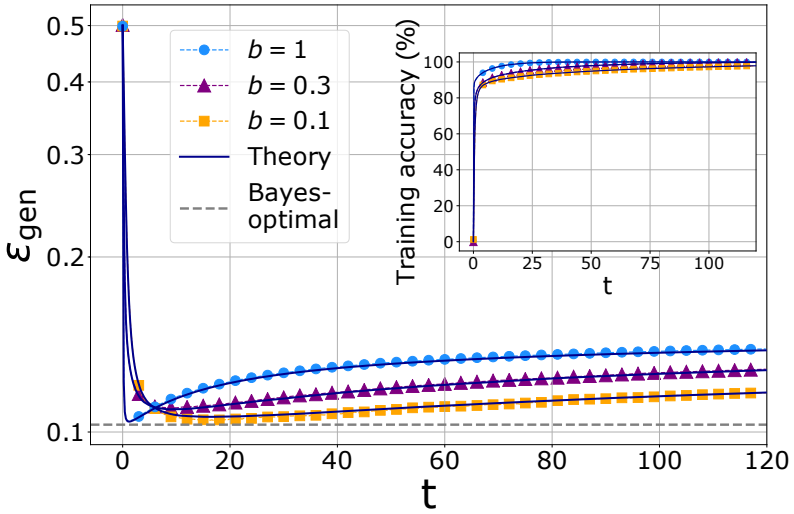
Algorithm

$$\min_w \sum_{i=1}^{N} L(y, \sigma(w^\top x)) + \lambda \|w\|^2$$

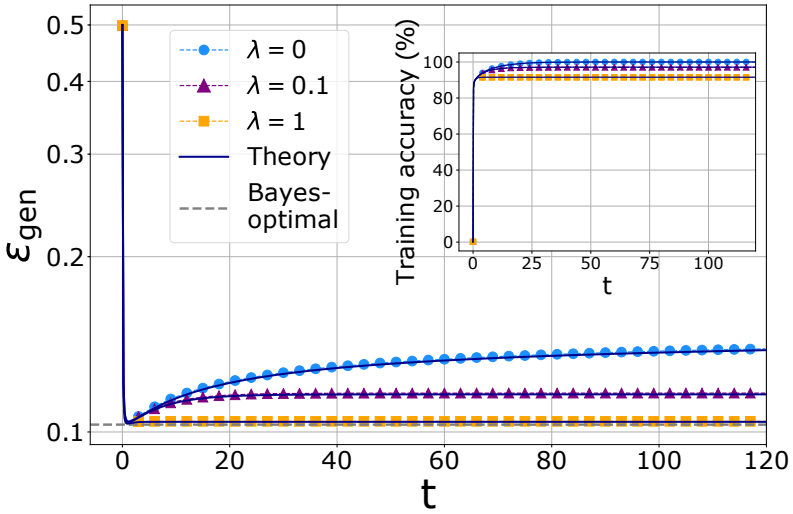
At each time sample in the mini-batch with probability b:**

$$w^{t+1} - w^t = \eta dt \sum_{k=1}^{N} s_k \nabla L\left(y, \sigma(w^\top x^{(k)})\right)$$

Effect of the batchsize



Effect of regularization



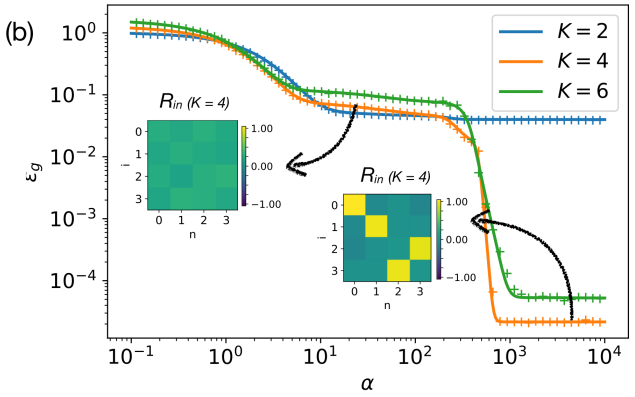
Mignacco, et al (2020). *Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification*. Neurips

Dynamics of learning: Recap

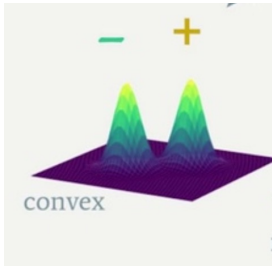
- The dynamics of online gradient descent can be studied in simple models by a set of closed ODEs on the overlaps

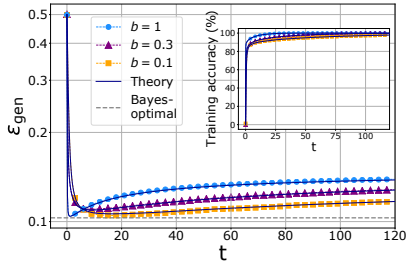
$$\begin{cases} \dfrac{\mathrm{d}R}{\mathrm{d}t} = g_R(R^t, Q^t) \\ \dfrac{\mathrm{d}Q}{\mathrm{d}t} = g_Q(R^t, Q^t) \end{cases} \qquad \mathcal{E}_g(R^{(k)}, Q^{(k)}) \underset{N\to+\infty, dt\to 0}{\longrightarrow} \mathcal{E}_g(t)$$

- In the committee machine training, we can describe the onset of the specialization transition



- Dynamical mean-field theory can help study batch grandient descent and understand the impact of batch-size

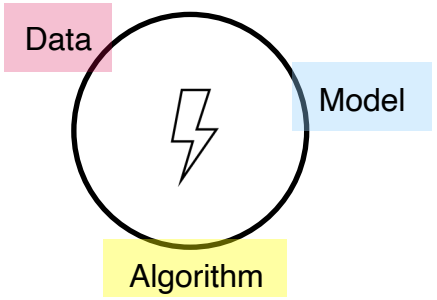


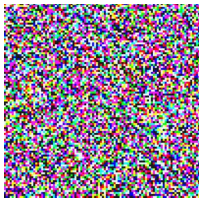


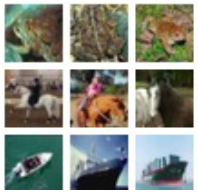
QUESTIONS?

Simple solvable models – what to focus on?

▷ Models of data « *teacher* »







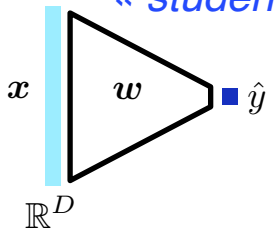
solvability / interpretability

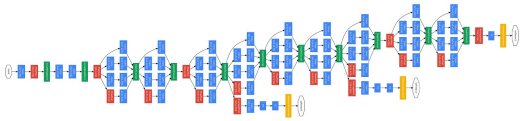
white noise — *?* — *CIFAR images*

degree of structure / realism

▷ Architectures

« *student* »





amenability to analysis

perceptron — *?* — *CIFAR images*

power of fit / realism

▷ Algorithms

amenability to analysis

online gradient descent — *?* — *mini-batch ADAM*

closeness to modern practice

Broadening use of “mean-field approximations” in deep learning literature

▷ “take advantage of concentration effects thanks to randomness in the large size limit”

* Analysis of statistical inference performance

Reviews: - Zdeborová & Krzakala (2016) *Statistical physics of inference: Thresholds and algorithms.*
- **Gabrié** (2020) *Mean field inference methods for neural networks.*

* Signal propagation in deep neural networks

- Trainability of very deep network at init. e.g., Schoenholz et al.(2017*). Deep Information Propagation.*
- Separation of structured data

e.g., Cohen, et al (2020). *Separability and geometry of object manifolds in deep neural networks.*

* Role of over-parametrization in trainability with Gradient Descent methods

- Convergence of SGD for 2-layers neural networks

Chizat & Bach (2018), Mei, Montanari & Nguyen (2018), Rotskoff & Vanden-Eijnden (2018)

- Neural Tangent Kernels, Equivalence to Gaussian processes, “Lazy training”

Jacot et al (2018), Lee et al (2019), review: Bahri et al (2020) *Statistical Mechanics of Deep Learning*

- Online learning e.g., Goldt, et al (2019). Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup

* Landscape, algorithms and interactions

Dauphin et al (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization
Sarao Mannelli & Zdeborová (2020). *Thresholds of descending algorithms in inference problems*.

Selected references

▷ Review papers (among others):

- classical Watkin, T. L. H., Rau, A., & Biehl, M. (1993). The statistical mechanics of learning a rule

- physics & modern M.L. Carleo, G., et. al (2019). *Machine learning and the physical sciences*.
  Gabrié, M. (2019). *Mean-field inference methods for neural networks.*
  Bahri, Y., et al. (2020). Statistical Mechanics of Deep Learning

- relevant statistical physics methods

  Zdeborová, L., & Krzakala, F. (2016). *Statistical physics of inference: Thresholds and algorithms*
  Castellani, T., Cavagna, A., Fisica, D., & Moro, P. A. (2005). *Spin-Glass Theory for Pedestrians*

▷ Books:

- classical stat. mech. of learning
  Engel, A., & Van den Broeck, C. (2001). *Statistical Mechanics of Learning*.
  Opper, M., & Saad, D. (2001). *Advanced mean field methods: Theory and practice*.

- spin glass Mézard, M., Parisi, G., & Virasoro, M. (1986). *Spin Glass Theory and Beyond*

- message passing Mézard, M., & Montanari, A. (2009). *Information, Physics, and Computation*.

THANKS !

Some useful references 1/2

1. Mézard, M., Parisi, G., & Virasoro, M. (1986). *Spin Glass Theory and Beyond* (Vol. 9). WORLD SCIENTIFIC. https://doi.org/10.1142/0271

2. Gardner, E. (1987). Maximum Storage Capacity in Neural Networks. *Europhysics Letters (EPL)*, *4*(4), 481–485. https://doi.org/10.1209/0295-5075/4/4/016

3. Gardner, E., & Derrida, B. (1989). Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, *22*(12), 1983–1994. https://doi.org/10.1088/0305-4470/22/12/004

4. Krauth, W., & Mézard, M. (1989). Storage capacity of memory networks with binary couplings. *Journal de Physique*, *50*(20), 3057–3066. https://doi.org/10.1051/jphys:0198900500200305700

5. Györgyi, G. (1990). First-order transition to perfect generalization in a neural network with binary synapses. *Physical Review A*, *41*(12), 7097–7100. https://doi.org/10.1103/PhysRevA.41.7097

6. Seung, H. S., Sompolinsky, H., & Tishby, N. (1992). Statistical mechanics of learning from examples. *Physical Review A*, *45*(8), 6056–6091. https://doi.org/10.1103/PhysRevA.45.6056

7. Watkin, T. L. H., Rau, A., & Biehl, M. (1993). The statistical mechanics of learning a rule. *Reviews of Modern Physics*, *65*(2), 499–556. https://doi.org/10.1103/RevModPhys.65.499

8. Biehl, M., & Riegler, P. (1994). On-Line Learning with a Perceptron. *Europhysics Letters (EPL)*, *28*(7), 525–530. https://doi.org/10.1209/0295-5075/28/7/012

9. Saad, D., & Solla, S. A. (1995). On-line learning in soft committee machines. *Physical Review E*, *52*(4), 4225–4243.

10. Saad, D., & Solla, S. A. (1995). Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, *74*(21), 4337–4340. https://doi.org/10.1103/PhysRevLett.74.4337

11. Biehl, M., & Schwarze, H. (1995). Learning by on-line gradient descent. *J. Phys. A. Math. Gen.*, *28*(3), 643–656. https://doi.org/10.1088/0305-4470/28/3/018

12. Saad, D. (1999). *On-Line Learning in Neural Networks* (D. Saad (ed.)). Cambridge University Press. https://doi.org/10.1017/CBO9780511569920

13. Opper, M., & Saad, D. (2001). *Advanced mean field methods: Theory and practice*. MIT press.

14. Engel, A., & Van den Broeck, C. (2001). *Statistical Mechanics of Learning*. Cambridge University Press. https://doi.org/10.1017/CBO9781139164542

Some useful references 2/2

15. Castellani, T., Cavagna, A., Fisica, D., & Moro, P. A. (2005). Spin-Glass Theory for Pedestrians. *Journal of Statistical Mechanics: Theory and Experiment*, *5*, 215–266. https://doi.org/10.1088/1742-5468/2005/05/P05012

16. Talagrand, M. (2006). The Parisi formula. *Annals of Mathematics*, *163*(1), 221–263. https://doi.org/10.4007/annals.2006.163.221

17. Mézard, M., & Montanari, A. (2009). *Information, Physics, and Computation*. Oxford University Press.

18. Zdeborová, L., & Krzakala, F. (2016). Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, *65*(5), 453–552. https://doi.org/10.1080/00018732.2016.1211393

19. Barbier, J., Dia, M., Macris, N., & Krzakala, F. (2017). The mutual information in random linear estimation. *54th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2016*, *0*(1), 625–632. https://doi.org/10.1109/ALLERTON.2016.7852290

20. Barbier, J., Krzakala, F., Macris, N., Miolane, L., & Zdeborová, L. (2018). Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models. *Proceedings of the 31st Conference On Learning Theory*, *PMLR 75*, 728–731. http://arxiv.org/abs/1708.03395

21. Gabrié, M., Manoel, A., Luneau, C., Barbier, jean, Macris, N., Krzakala, F., & Zdeborová, L. (2018). Entropy and mutual information in models of deep neural networks. *Advances in Neural Information Processing Systems 31* (Issue Nips, pp. 1826--1836). Curran Associates, Inc. http://papers.nips.cc/paper/7453-entropy-and-mutual-information-in-models-of-deep-neural-networks.pdf

22. Barbier, J., Macris, N., Maillard, A., & Krzakala, F. (2018). The Mutual Information in Random Linear Estimation Beyond i.i.d. Matrices. *IEEE International Symposium on Information Theory - Proceedings*, *2018-June*(3), 1390–1394. https://doi.org/10.1109/ISIT.2018.8437522

23. Aubin, B., Maillard, A., Barbier, J., Krzakala, F., Macris, N., & Zdeborová, L. (2018). The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Neural Information Processing Systems 2018*, *NeurIPS*, 1–44. http://arxiv.org/abs/1806.05451

24. Reeves, G. (2018). Additivity of information in multilayer networks via additive Gaussian noise transforms. *55th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2017*, *2018-Janua*, 1064–1070. https://doi.org/10.1109/ALLERTON.2017.8262855

25. Gabrié, M. (2019). Mean-field inference methods for neural networks. *ArXiv Preprint*, *1911.00890*. http://arxiv.org/abs/1911.00890

26. Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., & Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, *91*(4). https://doi.org/10.1103/revmodphys.91.045002

27. Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S., Sohl-Dickstein, J., & Ganguli, S. (2020). Statistical Mechanics of Deep Learning. *Annual Review of Condensed Matter Physics*, *11*(1), 501–528. https://doi.org/10.1146/annurev-conmatphys-031119-050745

Sketch of a replica computation

▷ Replica trick

$$\mu_\beta\left(\boldsymbol{w} \mid \{\boldsymbol{x}^\mu, y^\mu\}\right) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta\left[\sum_{\mu=1}^N \ell(y^\mu, \boldsymbol{x}^\mu \cdot \boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\right]} \qquad \mathbb{E}_{\mathcal{D}} \log \mathcal{Z}_\beta = \lim_{r \to 0^+} \frac{\partial}{\partial r} \mathbb{E}_{\mathcal{D}} \mathcal{Z}_\beta^r$$

▷ Rewritten in terms of order parameters

diverging dimension

replica potential

$$\mathbb{E}_{\{\boldsymbol{x}^\mu, y^\mu\}} \mathcal{Z}_\beta^r = \int \frac{\mathrm{d}\rho \mathrm{d}\hat{\rho}}{2\pi} \int \prod_{a=1}^r \frac{\mathrm{d}m_s^a \mathrm{d}\hat{m}_s^a}{2\pi} \int \prod_{1 \le a \le b \le r} \frac{\mathrm{d}q_s^{ab} \mathrm{d}\hat{q}_s^{ab}}{2\pi} \frac{\mathrm{d}q_w^{ab} \mathrm{d}\hat{q}_w^{ab}}{2\pi} e^{\boldsymbol{P}\Phi^{(r)}}$$

▷ Saddle point equations

$$\begin{cases} \hat{r}_{s,i} = -2\sigma_{x,i}\frac{\alpha}{\gamma}\partial_{r_{s,i}}\Psi_y(R,Q,M) & r_{s,i} = -\frac{2}{\gamma}\partial_{\hat{r}_{s,i}}\Psi_w\left(\hat{r}_{s,i}, \hat{q}_{s,i}, \hat{m}_{s,i}, \hat{r}_w, \hat{q}_w\right) \\ \hat{q}_{s,i} = -2\sigma_{x,i}\frac{\alpha}{\gamma}\partial_{q_{s,i}}\Psi_y(R,Q,M) & q_{s,i} = -\frac{2}{\gamma}\partial_{\hat{q}_{s,i}}\Psi_w\left(\hat{r}_{s,i}, \hat{q}_{s,i}, \hat{m}_{s,i}, \hat{r}_w, \hat{q}_w\right) \\ \hat{m}_{s,i} = \sigma_{x,i}\frac{\alpha}{\gamma}\partial_{m_{s,i}}\Psi_y(R,Q,M) & m_{s,i} = \frac{1}{\gamma}\partial_{\hat{m}_{s,i}}\Psi_w\left(\hat{r}_{s,i}, \hat{q}_{s,i}, \hat{m}_{s,i}, \hat{r}_w, \hat{q}_w\right) \\ \hat{r}_w = -2\alpha\partial_{r_w}\Psi_y(R,Q,M) & r_w = -2\partial_{\hat{r}_w}\Psi_w\left(\hat{r}_{s,i}, \hat{q}_{s,i}, \hat{m}_{s,i}, \hat{r}_w, \hat{q}_w\right) \\ \hat{q}_w = -2\alpha\partial_{q_w}\Psi_y(R,Q,M) & q_w = -2\partial_{\hat{q}_w}\Psi_w\left(\hat{r}_{s,i}, \hat{q}_{s,i}, \hat{m}_{s,i}, \hat{r}_w, \hat{q}_w\right) \end{cases}$$

▷ Order parameters are precisely what we need

$$m_{s,i} = \lim_{D\to\infty} \frac{1}{D} \mathbb{E}_{\mu_\beta}\left[(\boldsymbol{w}_2 W_1)_i^\top \boldsymbol{\beta}_i\right] \quad q_{s,i} = \lim_{D\to\infty} \frac{1}{D} \mathbb{E}_{\mu_\beta}\left[(\boldsymbol{w}_2 W_1)_i^\top (\boldsymbol{w}_2 W_1)_i\right], \quad q_w = \lim_{P\to\infty} \frac{1}{P} \mathbb{E}_{\mu_\beta}\left[\boldsymbol{w}^\top \boldsymbol{w}\right]$$

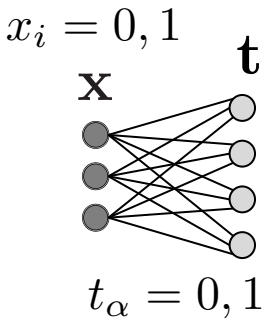
Mezard, M. Parisi, G. & Virasoro M. (1986). *Spin Glass Theory and Beyond;* **Gabrié M**. *J-Phys A* (2020); Gerace F., et al *ICML (2020);* Ascoli S., **Gabrié M.,** Sagun L., & Biroli G. (2021).

On a different note:

A statiscal mechanics inspired learning algorithm

Learning representations with RBMs

▷ Definition Restricted Boltzmann Machine (RBM):



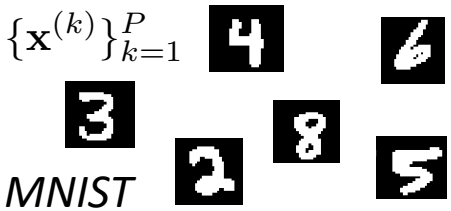
Pairwise *interactions input-hidden units*

$$p(\mathbf{x}, \mathbf{t}) = \frac{1}{\mathcal{Z}} e^{\sum_{i=1}^{N} b_{x,i} x_i + \sum_{\alpha=1}^{M} b_{t,\alpha} t_\alpha + \sum_{\alpha,i} W_{i\alpha} x_i t_\alpha}$$

$$p(\mathbf{x}) = \int \mathrm{d}\mathbf{t}\; p(\mathbf{x}, \mathbf{t}) \longrightarrow$$ *Effective interactions* ***all orders***

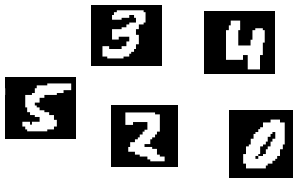
▷ Unsupervised learning:

$\mathcal{D} = \{\mathbf{x}^{(k)}\}_{k=1}^{P}$



MNIST

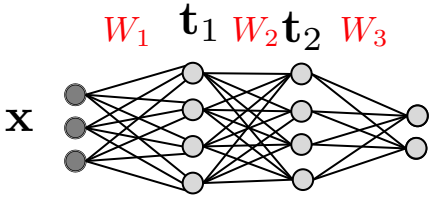
learn $\theta = \{\mathbf{b}_x, \mathbf{b}_t, W\}$ *sample*



$\sim p(\mathbf{x})$

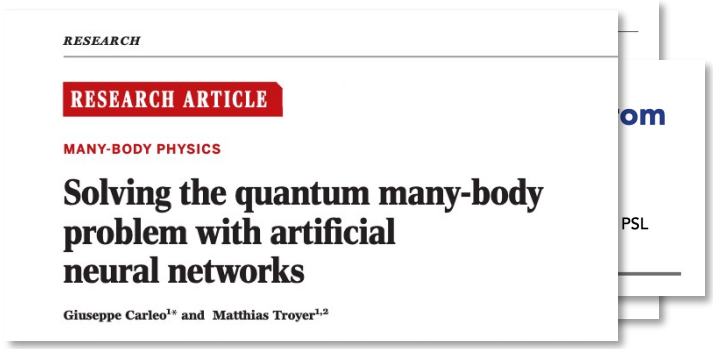
▷ Applications:

Pretraining of deep networks



Biophysics models

Quantum physics



RESEARCH

RESEARCH ARTICLE

MANY-BODY PHYSICS

Solving the quantum many-body problem with artificial neural networks

Giuseppe Carleo[1]* and Matthias Troyer[1,2]

Learning in Restricted Boltzmann Machines (RBMs)

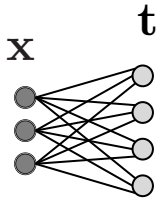
▷ Maximum likelihood learning

$$\ell(W, \mathbf{b}_x, \mathbf{b}_t) = \prod_{k=1}^{P} p(\mathbf{x}^{(k)})$$

Probability of training data according to RBM

+ gradient ascent !

▷ But intractable exact inference



$$\mathcal{Z} = \sum_{\mathbf{x},\mathbf{t}} e^{\sum_{i=1}^{N} b_{x,i} x_i + \sum_{\alpha=1}^{M} b_{t,\alpha} t_\alpha + \sum_{\alpha,i} W_{i\alpha} x_i t_\alpha}$$

⟶ 2^{M+N} *terms !*

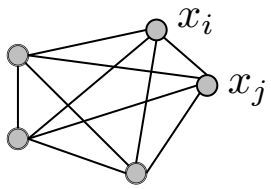
▷ Approximations

- ▹ Approximate Monte Carlo: Contrastive Divergence (Hinton 2002) *state-of-the-art*
- ▹ Mean-field strategies ?
  - Without hidden units: (Kappen, 1999; Cocco & Monasson, 2012; Ricci-Tersenghi 2012 etc ..)
  - With hidden units: (Tieleman et al. 2009; Salakhutdinov & Hinton, 2009) *only naive mean-field*

Can we use other mean-field methods to ease inference and learning ?

From Naive mean-field (NMF) to Thouless-Anderson-Palmer (TAP) approximation

$$E(\mathbf{x}) = -\sum_{i=1}^{N} b_i x_i - \sum_{i,j} W_{ij} x_i x_j \qquad x_i = 0, 1$$



- Free energy functional at **fixed magnetizations** ($\langle x_i \rangle_{\mathbf{m}} = m_i$) **+ without interactions**

≈ ``$\beta \to 0$'' = easy to compute

$$-\beta G(\mathbf{m}) = H_{\mathbf{NMF}}(\mathbf{m}) + \beta \langle \sum_{i=1}^{N} b_i x_i \rangle_{\mathbf{m}} + \beta \sum_{ij} W_{ij} m_i m_j + \frac{\beta^2}{2} {W_{ij}}^2 (m_i - m_i^2)(m_j - m_j^2)$$

entropy — NMF — TAP

- Select **magnetizations of minimum free energy**

$$\frac{\partial G}{\partial m_i} = 0 \Rightarrow m_i = \sigma \left[\beta b_i + \sum_j \beta W_{ij} m_j - \beta^2 {W_{ij}}^2 \left(m_i - \frac{1}{2} \right) (m_j - m_j^2) \right]$$

NMF — TAP

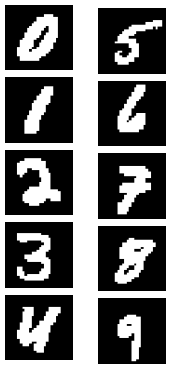
$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

(Thouless, Andersson & Palmer 1977, Plefka 1982, Georges & Yedidia 1999)

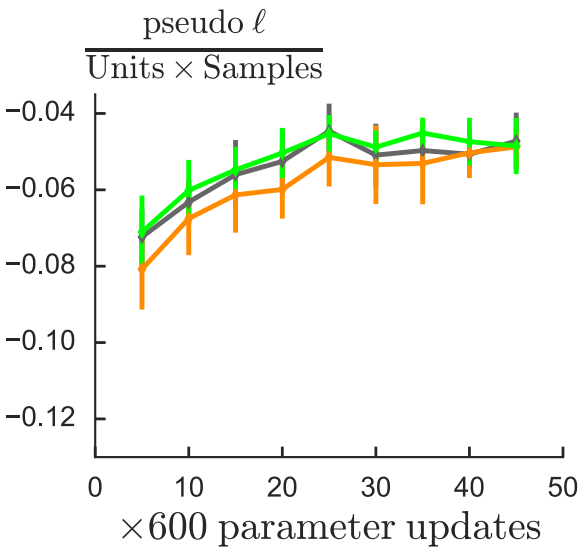
Maximum likelihood training with TAP

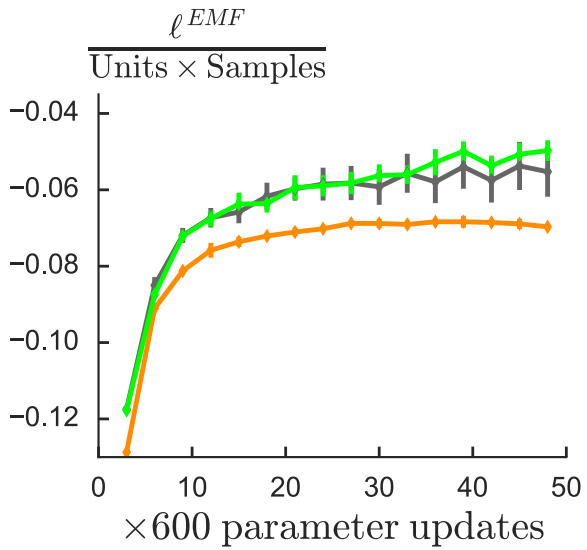
- High-temperature = small weights
- $\ln \mathcal{Z} \approx -G(\mathbf{m})$
- Compare TAP/MCMC (CD) approx of likelihood and gradients

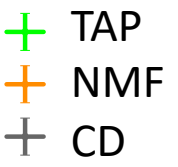
Binarized MNIST



pseudo log-likelihood log-likelihood TAP log-likelihood







comparable running times

Further checks:

- No significant improvements with more orders / AdaTAP
- Extends to real-valued variables

(M.G., E. Tramel & F. Krzakala NIPS 2015)

Following the learned TAP representations 1/2

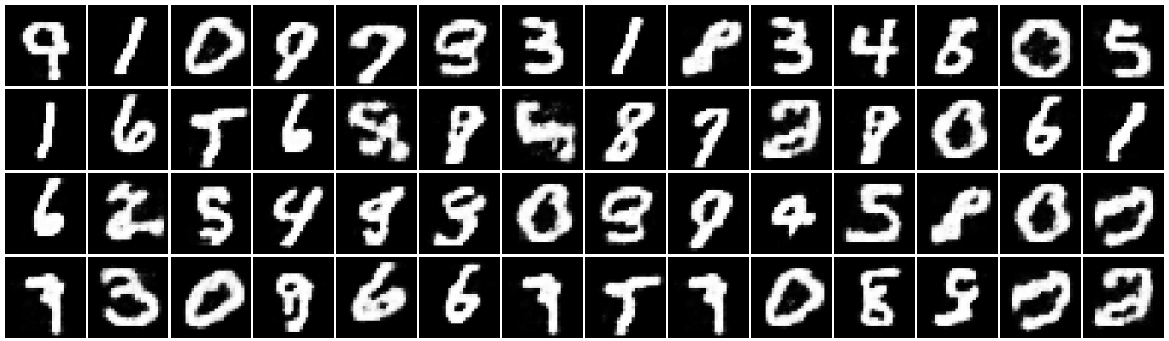
$$m_i^x \leftarrow \sigma\left[\beta b_i + \sum_\alpha \beta W_{i\alpha} m_\alpha^t - \beta^2 {W_{ij}}^2 \left(m_i^x - \frac{1}{2}\right)\left(m_\alpha^t - {m_\alpha^t}^2\right)\right]$$

NMF

TAP

TAP fixed points with trained weights and biases for different initializations

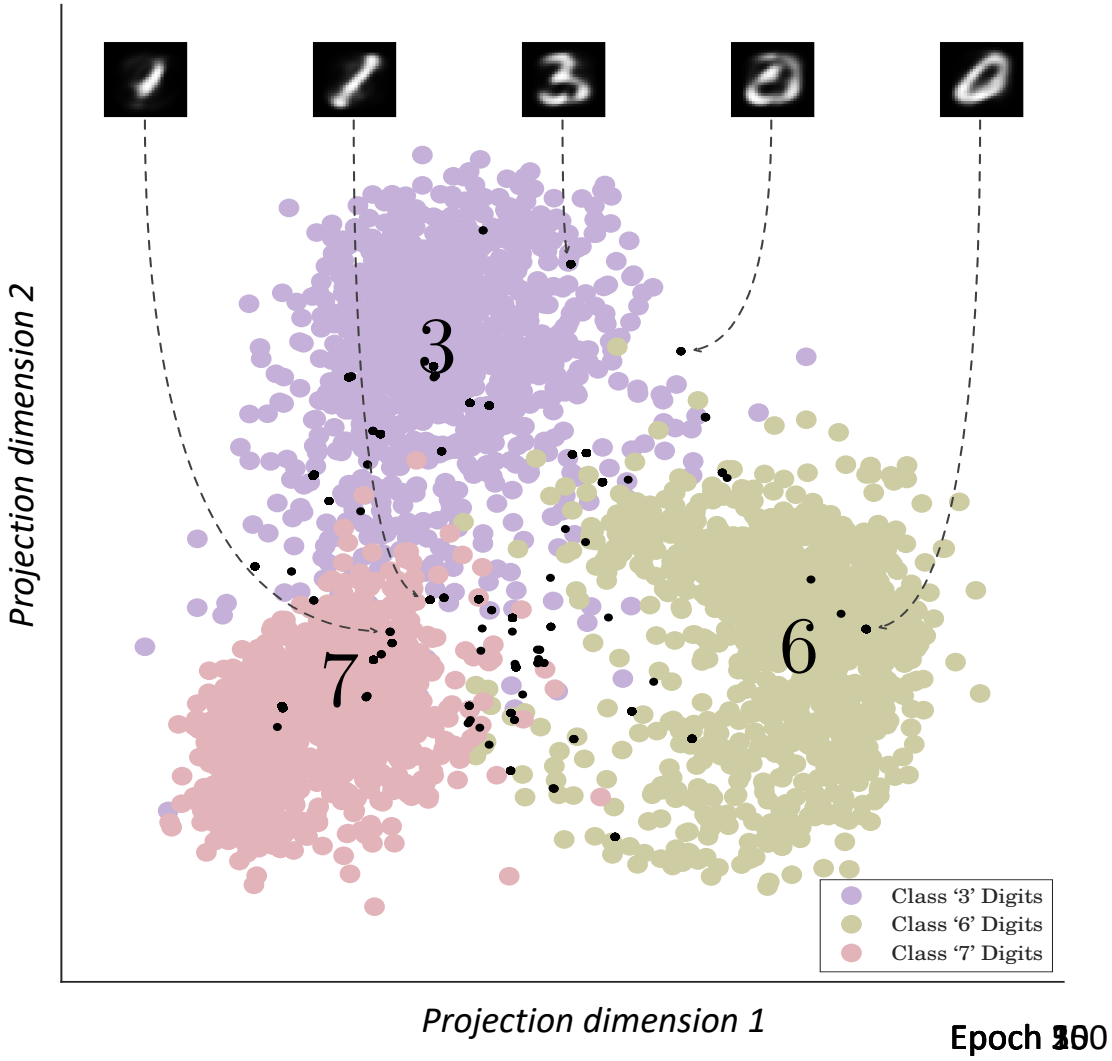
Now for a naive-MF machine



(M.G., E. Tramel & F. Krzakala NIPS 2015)

Following the learned TAP representations 2/2

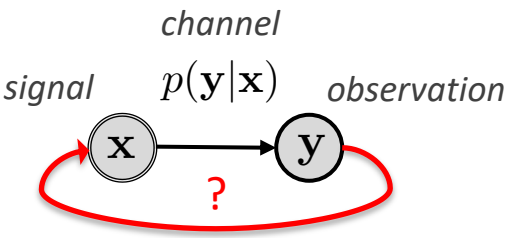
Evolution of magnetizations in configuration space (2d projection)



(E.Tramel, M.G., A. Manoel, F. Caltagirone, F. Krzakala, PRX 2018)

Beyond training: Application for reconstruction

- Reconstruction problems



- **E.g. compressed sensing (CS)**

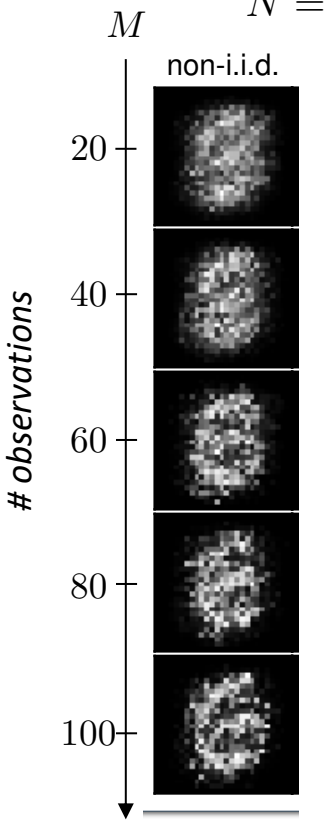
$\mathbb{R}^N$ *sparse signal*

$$\mathbf{y} = F\,\mathbf{x} + \xi \qquad K << N$$

few observations

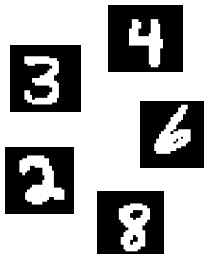
$M << N$

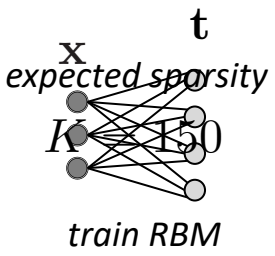
$N = 784$ pixels



- **Exploit prior information**

typical signals





AMP algorithm
≈ TAP for CS
+
RBM TAP equations

RBM learned representations

improve drastically reconstruction

(D.L. Donoho et aL 2009, E. Tramel et al. JSTAT 2016, E. Tramel, A. Manoel, F Caltagirone, M.G., & F. Krzakala ITW 2016)