NYU Center for Data Science

FLATIRON INSTITUTE

For Women in Science
FONDATION L'ORÉAL

Statistical physics for machine learning

February 11th 2020

Machine Learning in Physics
VDSP-ESI Winter School 2020

Marylou Gabrié (NYU, Flatiron Institute)

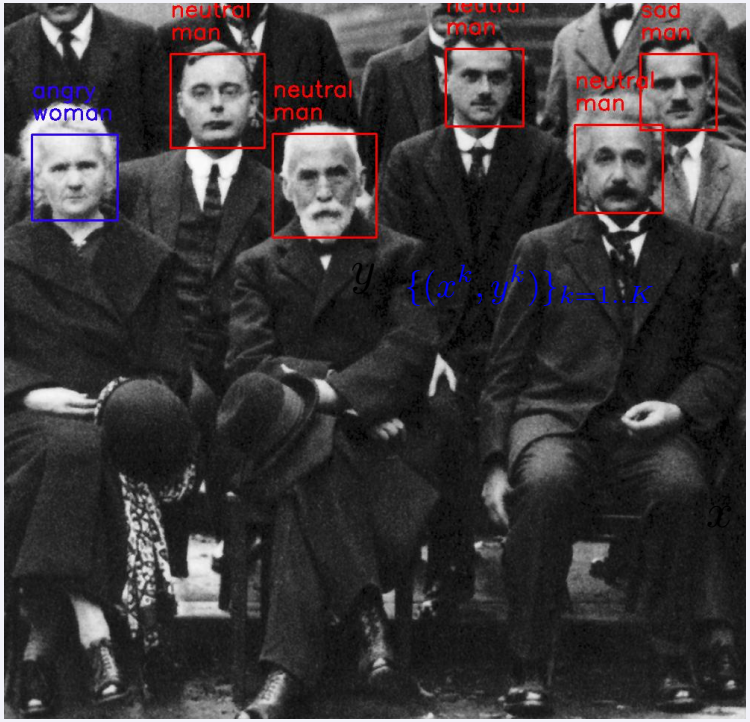
Alia Abbara (LPENS)

Artificial Neural Networks behind "AI" breakthroughs

"Artificial intelligence": solve highly complex tasks with a computer

examples: play go, identify faces etc.

AlphaGo, Silver et al. 2016, 2017)



(Arriaga et al. 2017)

Artificial Neural Networks behind "AI" breakthroughs

"Artificial intelligence": solve highly complex tasks with a computer

examples: play go, identify faces etc.

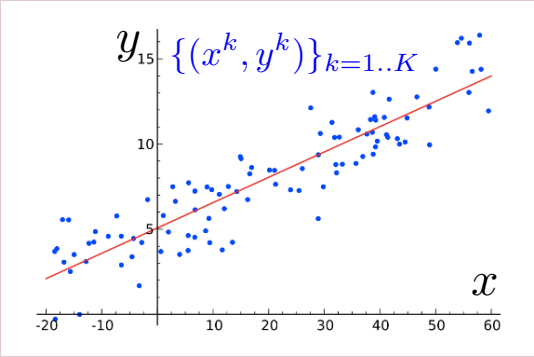
Machine learning: algorithm able to learn from examples

simple example : linear regression

- task: predict y from x from set of training points

- method: fit parametric model

$$\hat{y} = ax + b$$



Artificial Neural Networks behind "AI" breakthroughs

"Artificial intelligence": solve highly complex tasks with a computer

examples: play go, identify faces etc.

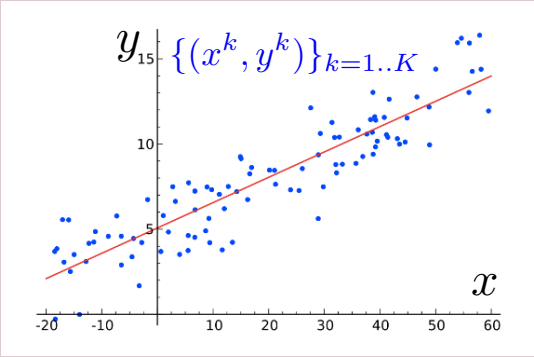
Machine learning: algorithm able to learn from examples

simple example : linear regression

- task: predict y from x from set of training points

- method: fit parametric model

$$\hat{y} = ax + b$$



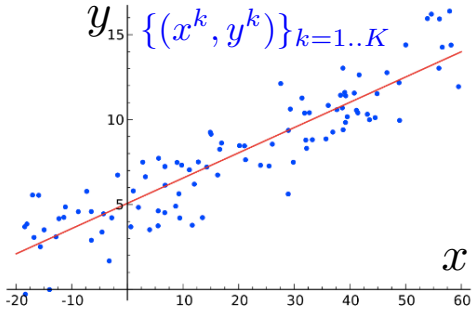
Deep learning: use deep neural networks as parametric model

harder example : automatic vision (self driving cars), machine translation

Physics and machine learning

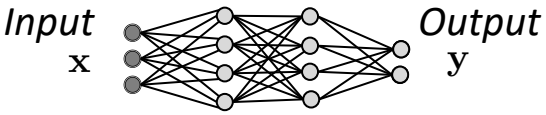
Machine learning for physics

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



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

- Recent deep learning “revolution”



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 (cf tutorial)

- New wave of interest

thermodynamic limit

approximate computations

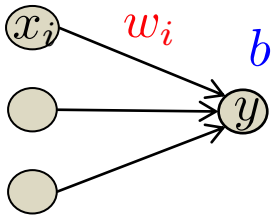
simplifying models

phase transitions

QUICK INTRODUCTION TO NEURAL NETWORKS

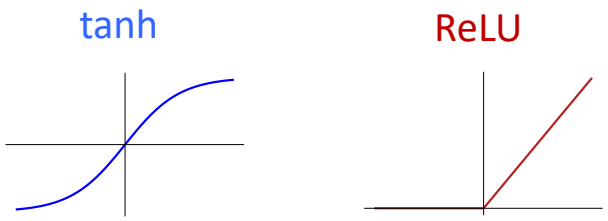
Deep neural networks are parametric models …

An artificial neuron

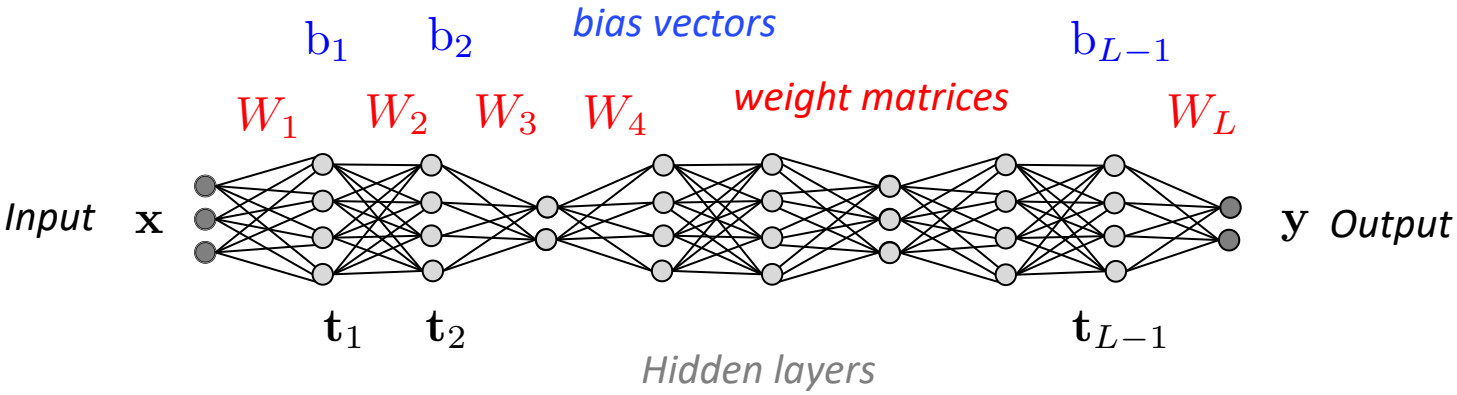


$$y = f(\sum_i w_i x_i + b) = f(\mathbf{w}^T \mathbf{x} + b)$$

(non-linear) activation function



A (deep) neural networks is a combination of artificial neural neurons

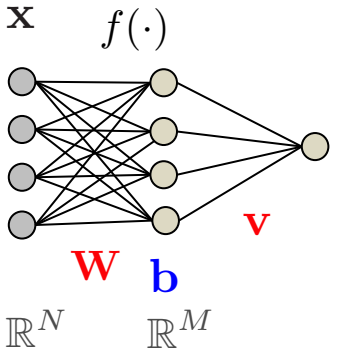


$$\mathbf{y} = f(\mathbf{W}_L f(\mathbf{W}_{L-1} \ldots f(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) .. + \mathbf{b}_{L-1}) + \mathbf{b}_L)$$

… with a great representative power

Theory: Universal approximation theorem (Cybenko 1989, Hornik 1991)

> "Any **continuous function on** $\mathbb{R}^N$ can be well approximated by a **2-layer neural network**."



Practice: Deep neural network supervised learning

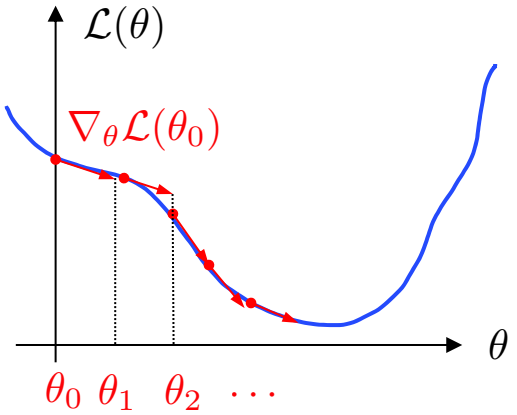
$$\hat{\mathbf{y}}(\mathbf{x};\theta) = f(\mathbf{W}_L f(\mathbf{W}_{L-1} \ldots f(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) .. + \mathbf{b}_{L-1}) + \mathbf{b}_L)$$

- Large collection of parameters $\theta = \{\mathbf{W}_1, \mathbf{W}_2.., \mathbf{W}_L, \mathbf{b}_1, \mathbf{b}_2.., \mathbf{b}_L\}$
- Training data and training-loss

$$\mathcal{D} = \{(\mathbf{x}^k, \mathbf{y}^k)\}_{k=1}^P \qquad \mathcal{L}(\theta) = \frac{1}{K}\sum_k \underbrace{\ell[y^k, \hat{y}(x^k,\theta)]}_{= \|y^k - \hat{y}(x^k,\theta)\|^2}$$

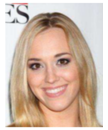
- Optimization by gradient descent

$$\min_\theta \mathcal{L}(\theta) \qquad \theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$$

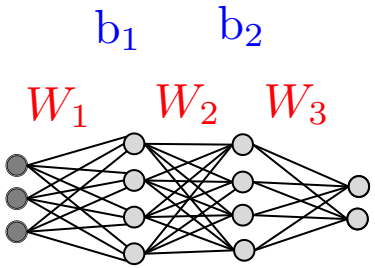


Supervised learning example: image classification

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

$\mathbf{x}$	$\mathbf{y}$
	*Carrie*
	*Steve*
	*Ben*
	*Carrie*
	*Steve*
	*Ben*
	*Carrie*
	*Steve*
	*Ben*
	*Carrie*
	*Steve*
	*Ben*

new picture from celebrity of the data set

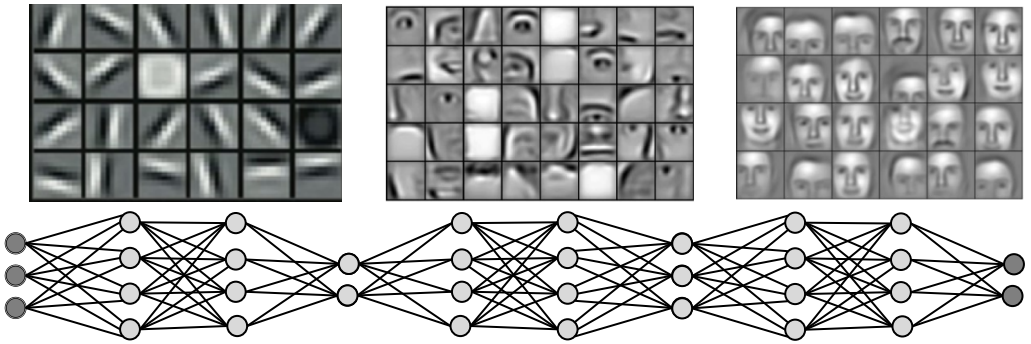


Prof Dellago !

Why does deep learning work so well ?

Intuition:

- Hierarchical and invariant representations

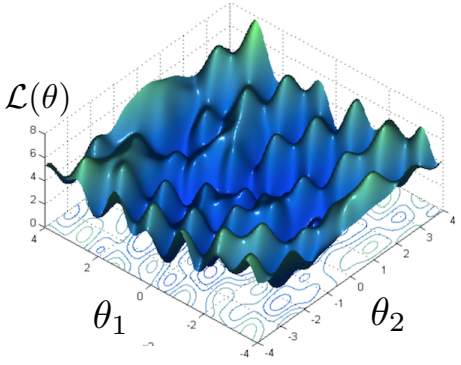


(Lee et al. 2016)

Prof Dellago !

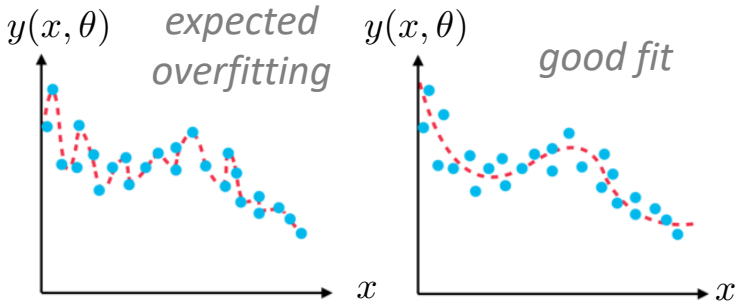
Puzzles:

- Optimization:



- Generalization:

$$\#\ \text{parameters} \gg \#\ \text{points}$$



Examples of questions for the theory of learning

- **How many training examples are needed to generalize ?**
- **Is gradient descent the optimal algorithm ?**
- **Can a given parametric model (e.g. neural network) fit perfectly a dataset ?**

Plan of attack

1. **The curse of dimensionality, disordered systems physics and models**
   A. Storage of the perceptron
   B. Spin glasses, thermodynamic limit, and disorder averages
   C. Some models and results on the perceptron

2. **Teacher-students and Bayesian inference**
   A. Modelling learning 2.0
   B. Bayesian inference
   C. Optimality and information theoretic limits

3. **Algorithmic performance**
   A. Algorithm complexity
   B. Computational hard phase

4. **Dynamics of learning**
   A. Online learning in the perceptron
   B. Dynamics of overlaps and generalization error

Plan of attack

1. **The curse of dimensionality, disordered systems physics and models**
   A. Storage of the perceptron
   B. Spin glasses, thermodynamic limit, and disorder averages
   C. Some models and results on the perceptron

2. **Teacher-students and Bayesian inference**
   A. Modelling learning 2.0
   B. Bayesian inference
   C. Optimality and information theoretic limits

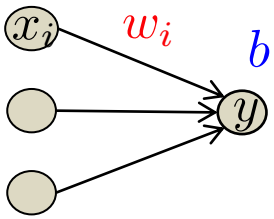
3. **Algorithmic performance**
   A. Algorithm complexity
   B. Computational hard phase

4. **Dynamics of learning**
   A. Online learning in the perceptron
   B. Dynamics of overlaps and generalization error

Perceptron storage problem

The simplest neural network

$\mathbf{x} \in \mathbb{R}^N \qquad y \in \{-1, 1\}$



$\mathbb{R}^N \qquad \mathbb{R}$

$$y(\mathbf{x}; \mathbf{w}, b) = \text{sign}(\sum_i w_i x_i + b) = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$$

training set:

$$\mathbf{x}^{(k)}, y^{(k)} \sim p(\mathbf{x}, y) \qquad k = 1 \cdots P$$

Volume of parameters fitting the data:

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

at least one solution!

curse of dimensionality !

Very hard question because

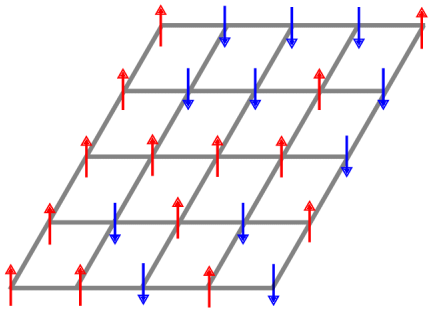
- High dimensional integral
- Even if distribution fixed depends on the precise training set $\{\mathbf{x}^{(k)}, y^{(k)}\}_{k=1}^{P}$

Gardner, E. (1987). *Maximum Storage Capacity in Neural Networks.*

Physics of disordered systems – the toy spin glass

Spin glass energy (N spins)

$$E(\mathbf{s}; J) = -\sum_{(i,j)} J_{ij} s_i s_j$$



Couplings distributions

- Ising ferromagnet $J_{ij} = J_0 > 0 \quad \forall (i,j)$
- Spin glass, disorder interactions $J_{ij} \sim P_J(J_{ij}) \quad \text{i.i.d.}$

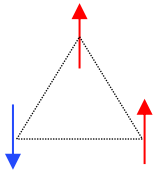
different environment for each spin!!

$$J_{ij} \sim \mathcal{N}(J_{ij}\,;\, 0, 1/\sqrt{N})$$

Sherrington - Kirkpatrick (1975)

Energy landscape?

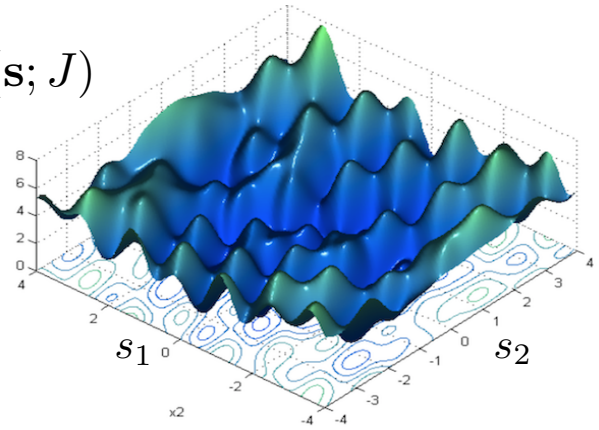
- Ground state configuration?



frustration may occur !

- Many local minima, metastable states

$E(\mathbf{s}; J)$



slow dynamics of relaxation: theoretical model for glass

Castellani, T., Cavagna, A., Fisica, D., & Moro, P. A. (2005). *Spin-Glass Theory for Pedestrians.*
Mézard, M., Parisi, G., & Virasoro, M. (1986). *Spin Glass Theory and Beyond*

Physics of disordered systems – thermodynamic limit and self averaging

Free energy density $f_N(J) = -\frac{\beta}{N} \ln \mathcal{Z}_N(J) = -\frac{\beta}{N} \ln \int \mathrm{d}\mathbf{s}\, e^{-\beta E(\mathbf{s};J)}$

For self averaging disorder

$$f_N(J) \xrightarrow[N \to \infty]{} \lim_{N\to\infty} \int \mathrm{d}J\, P_J(J) \big(f_N(J) \big) = f \approx f_N(J)$$

thermodynamic limit

disorder average

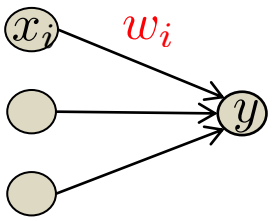
quenched free energy:
independent of disorder realization
typical free energy!

Large literature of "mean-field" methods to compute/approximate quenched averages

- Variational methods (c.f. tutorial naive mean-field)
- Replica method
- Temperature expansions

Back to -
Perceptron storage problem - as a disordered system

$\mathbf{x} \in \mathbb{R}^N$ $\quad y \in \{0,1\}$



$\mathbb{R}^N$

$$y(\mathbf{x};\mathbf{w}) = \mathrm{sign}(\sum_i w_i x_i) = \mathrm{sign}(\mathbf{w}^T\mathbf{x})$$

training set:

$$\mathbf{x}^{(k)}, y^{(k)} \sim p(\mathbf{x}, y) \qquad k = 1 \cdots P$$

each new data point = constraint/interaction for the weights

Volume $$V_{P,N} = \int_{\Omega(\mathbf{w})} \mathrm{d}\mathbf{w} \prod_{k=1}^{P} \delta(y(\mathbf{x}^{(k)};\mathbf{w}) - y^{(k)})$$

disorder variables $\mathbf{x}^{(k)}, y^{(k)} \sim p(\mathbf{x}, y)$

$$V_{P,N} \propto \int_{\Omega(\mathbf{w})} \mathrm{d}\mathbf{w}\, e^{-\beta \sum_{k=1}^{P} \left(y(\mathbf{x}^{(k)};\mathbf{w}) - y^{(k)}\right)^2}$$

$\beta \to +\infty$

$$V_{P,N} \propto \int_{\Omega(\mathbf{w})} \mathrm{d}\mathbf{w}\, e^{-\beta E(\mathbf{w};\{\mathbf{x}^{(k)};y^{(k)}\})}$$

Gardner, E. (1987). *Maximum Storage Capacity in Neural Networks.*

Back to -
Perceptron storage problem - thermodynamic limit

Volume $V_{P,N} \propto \int_{\Omega(\mathbf{w})} \mathrm{d}\mathbf{w}\, e^{-\beta E(\mathbf{w};\{\mathbf{x}^{(k)};y^{(k)}\})}$

disorder variables $\mathbf{x}^{(k)}, y^{(k)} \sim p(\mathbf{x}, y)$

Thermodynamic limit: $\begin{cases} N \to \infty \\ P \to \infty \end{cases}$ $\alpha = P/N = \dfrac{\#\,\text{points}}{\#\,\text{parameters}}$ fixed

controls difficulty

Results for different models using the mean-field replica method $V(\alpha_c) = 0$

Model 1:

- **data:** random Gaussian inputs, random binary outputs
- **weights:** normalized $p_0(\mathbf{w}) \propto \delta\Big(\sum_{i,j} w_{ij}^2/N - 1\Big)$

$\alpha_c = 2$

Model 2:

- **data:** random binary inputs, random binary outputs
- **weights:** binary $p_0(\mathbf{w}) \propto \prod_{i,j} \delta(w_{ij} - 1) + \delta(w_{ij} + 1)$

$\alpha_c \simeq 0.83$

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

Small recap 1

1. **The curse of dimensionality, disordered systems physics and models**
   A. Storage of the perceptron
   B. Spin glasses, thermodynamic limit, and disorder averages
   C. Some models and results on the perceptron

Important concepts
- Physics of disordered systems
- Modelling assumptions

Illustrations
- on the perceptron – classification of random data

Perspective
- Yet up to now random data, can we do better?

Plan of attack

1. **The curse of dimensionality, disordered systems physics and models**
   A. Storage of the perceptron
   B. Spin glasses, thermodynamic limit, and disorder averages
   C. Some models and results on the perceptron

2. **Teacher-students and Bayesian inference**
   A. Modelling learning 2.0
   B. Bayesian inference
   C. Optimality and information theoretic limits

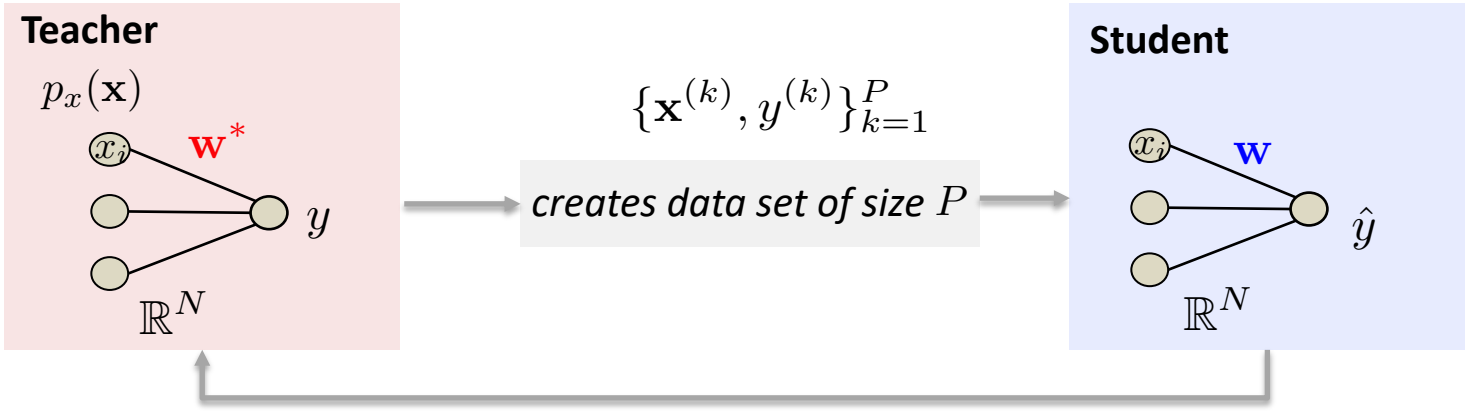
3. **Algorithmic performance**
   A. Algorithm complexity
   B. Computational hard phase

4. **Dynamics of learning**
   A. Online learning in the perceptron
   B. Dynamics of overlaps and generalization error

Learning a rule – teacher-student scenario

In storage problem first solved by physicists $\mathbf{x}^{(k)}, y^{(k)} \sim p_x(\mathbf{x})p_y(y)$

Use a model of relationship between inputs and outputs $\mathbf{x}^{(k)}, y^{(k)} \sim p_x(\mathbf{x})p_T(y|\mathbf{x})$



- Training error: how do we fit the training set? ← similar to storage problem
- Generalization error: can we predict output of new input $\mathbf{x}$? ← monitors overfitting!

Gardner, E., & Derrida, B. (1989*). Three unfinished works on the optimal storage capacity of networks*

Generalization error

Definition

$$\mathcal{E}(\mathbf{w}, \mathbf{w}^*) = \mathbb{E}_{\mathbf{x}} \left[\ell\big(y^*(\mathbf{x}), y(\mathbf{x})\big) \right]$$

$$= \int_{\mathbb{R}^N} \mathrm{d}\mathbf{x}\, p_x(\mathbf{x}) \left[1 - \theta\big(\mathrm{sign}(\mathbf{w}^{*\top}\mathbf{x})\mathrm{sign}(\mathbf{w}^{\top}\mathbf{x})\big) \right]$$

← classical sign perceptron

$$= \int_{\mathbb{R}^N} \mathrm{d}\mathbf{x}\, p_x(\mathbf{x}) \left[\ell\big(f(\mathbf{x}; \mathbf{w}^*), f(\mathbf{x}; \mathbf{w})\big) \right]$$

← more generally

↑ parametrized class of models

↑ model parameters

↑ population average
high (!!) dimensional integral

Evaluation

- "Monte-Carlo sampling", testing set $\mathcal{E}_N(\mathbf{w}, \mathbf{w}^*) \approx \frac{1}{P} \sum_{\ell=1}^{P} \left[\left(y^*(\mathbf{x}^{(\ell)}) - y(\mathbf{x}^{(\ell)}) \right)^2 \right]$

Physicists typical question - theoretical prediction?

Before:

Optimal storage?

- given distributions of inputs
- given distribution of outputs
- given weight domains

$$y(\mathbf{x};\mathbf{w}) = \mathrm{sign}(\mathbf{w}^T\mathbf{x})$$

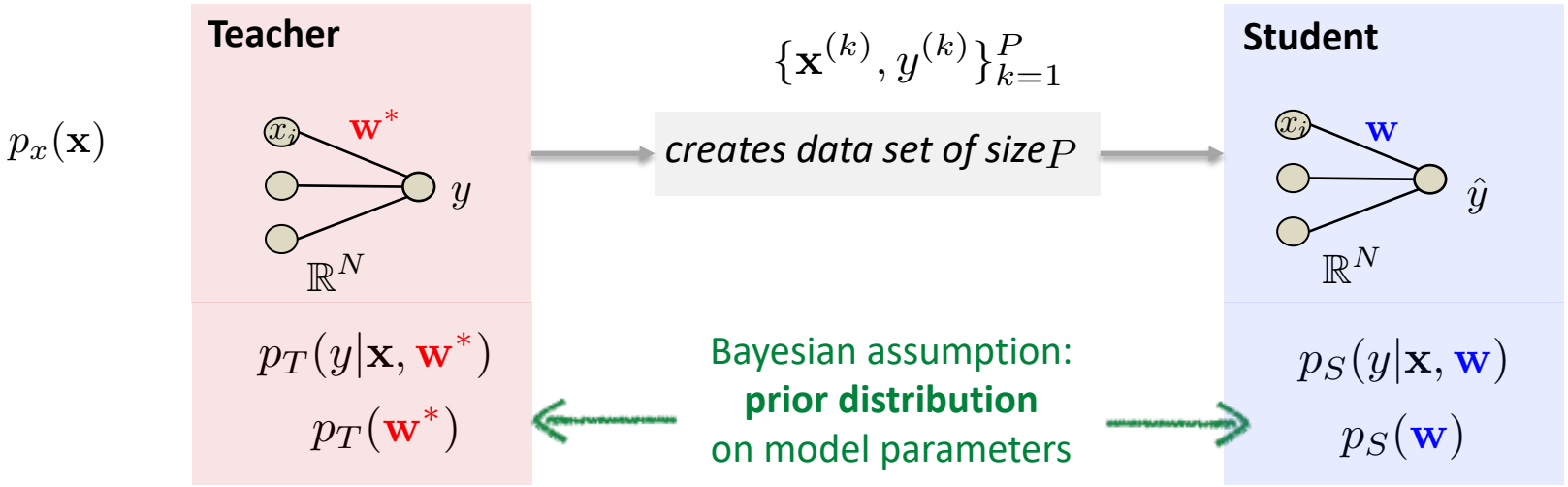
Now:

Optimal generalization error?

$$\mathcal{E}(\mathbf{w},\mathbf{w}^*)$$

Bayesian inference

Full teacher-student scenario modelling



Bayesian posterior distribution

for one point

$$p_S(\mathbf{w}|y^{(k)}, \mathbf{x}^{(k)}) = \frac{p_S(y^{(k)}|\mathbf{x}^{(k)}, \mathbf{w})\, p_S(\mathbf{w})}{p_S(y^{(k)}|\mathbf{x}^{(k)})}$$

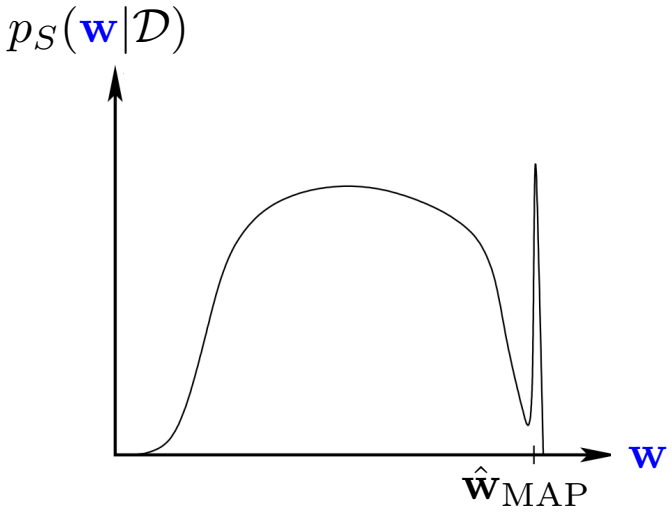
for full training set

$$p_S(\mathbf{w}|\{y^{(k)}, \mathbf{x}^{(k)}\}_{k=1}^{P}) = \frac{\prod_{k=1}^{P} p_S(y^{(k)}|\mathbf{x}^{(k)}, \mathbf{w})\, p_S(\mathbf{w})}{\prod_{k=1}^{P} p_S(y^{(k)}|\mathbf{x}^{(k)})}$$

Bayesian estimators

Maximum a posteriori (MAP)

$$\hat{\mathbf{w}}_{\mathrm{MAP}} = \arg\max_{\mathbf{w}} p_S(\mathbf{w}|\{y^{(k)}, \mathbf{x}^{(k)}\}_{k=1}^P)$$



Posterior mean - minimum mean-square error (MMSE)

$$\hat{\mathbf{w}}_{\mathrm{MMSE}} = \int \mathrm{d}\mathbf{w}\, \mathbf{w}\, p_S(\mathbf{w}|\{y^{(k)}, \mathbf{x}^{(k)}\}_{k=1}^P)$$

minimize squared error expectation
with respect to the posterior

$$\min_{\hat{\mathbf{w}}} \int \mathrm{d}\mathbf{w}\, (\mathbf{w} - \hat{\mathbf{w}})^2\, p_S(\mathbf{w}|\{y^{(k)}, \mathbf{x}^{(k)}\}_{k=1}^P)$$

Bayes optimality

Matched teacher and student

$$\frac{p_T(y|\mathbf{x}, \mathbf{w}^*)}{p_T(\mathbf{w}^*)} = \frac{p_S(y|\mathbf{x}, \mathbf{w})}{p_S(\mathbf{w})}$$

student knows exactly the statistical model used to generate the data

- Student posterior captures true posterior for teacher parameter

$$p_S(\mathbf{w}|y, \mathbf{x}) = \frac{p_S(y|\mathbf{x}, \mathbf{w})\, p_S(\mathbf{w})}{p_S(y|\mathbf{x})} = \frac{p_T(y|\mathbf{x}, \mathbf{w})\, p_T(\mathbf{w})}{p_T(y|\mathbf{x})} = p_T(\mathbf{w}|y, \mathbf{x})$$

- Definition of an optimal generalization error (MMSE)

$$\mathcal{E}(\mathbf{w}^*, \mathcal{D}) = \mathbb{E}_{\mathbf{x}}\left[\left(y^*(\mathbf{x}) - \mathbb{E}_{\mathbf{w}|\mathcal{D}}[y(\mathbf{x})]\right)^2\right]$$

optimal algorithm given the rule and the data

Optimal generalization w.r.t. sample complexity

$$\mathcal{E}(\mathbf{w}^*, \mathcal{D}) = \mathbb{E}_{\mathbf{x}} \left[\left(y^*(\mathbf{x}) - \mathbb{E}_{\mathbf{w}|\mathcal{D}}[y(\mathbf{x})] \right)^2 \right]$$

Average over the disorder + thermodynamic limit + mean-field replica method:

$$\int \mathrm{d}\mathbf{w}^* \, p_T(\mathbf{w}^*) \, \mathbb{E}_{\mathcal{D}} \left[\mathcal{E}(\mathbf{w}^*, \mathcal{D}) \right] \underset{N \to \infty}{\to} \mathcal{E}(\alpha)$$

with:

$$\alpha = P/N = \frac{\#\,\text{training points}}{\#\,\text{weights}}$$

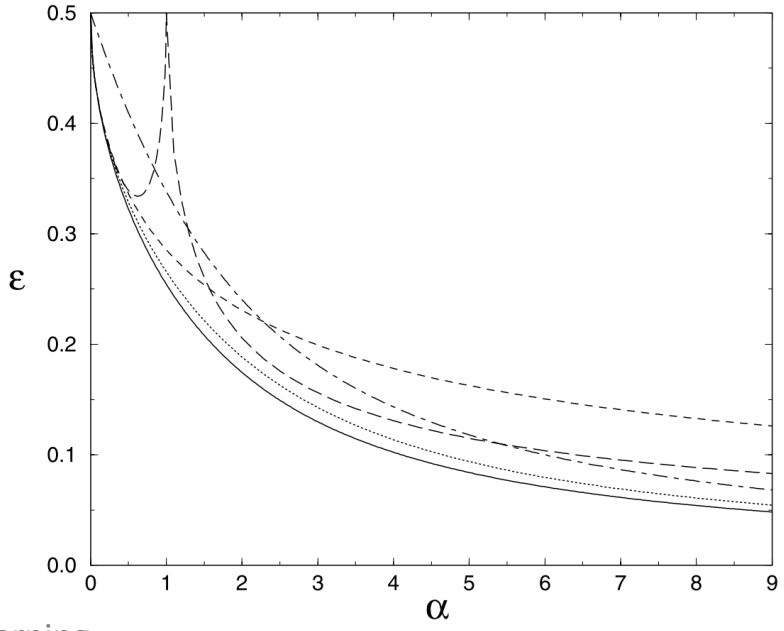
average with prior of teacher rule

expectation over datasets of size N

example I:

- binary input patterns $x_i^{(k)} = \pm 1$
- teacher prior: uniform on the sphere

$$p_T(\mathbf{w}^*) = (2\pi e)^{-N/2} \delta\left(\sum_{i=1}^{N} w_i^{*2} - N \right)$$



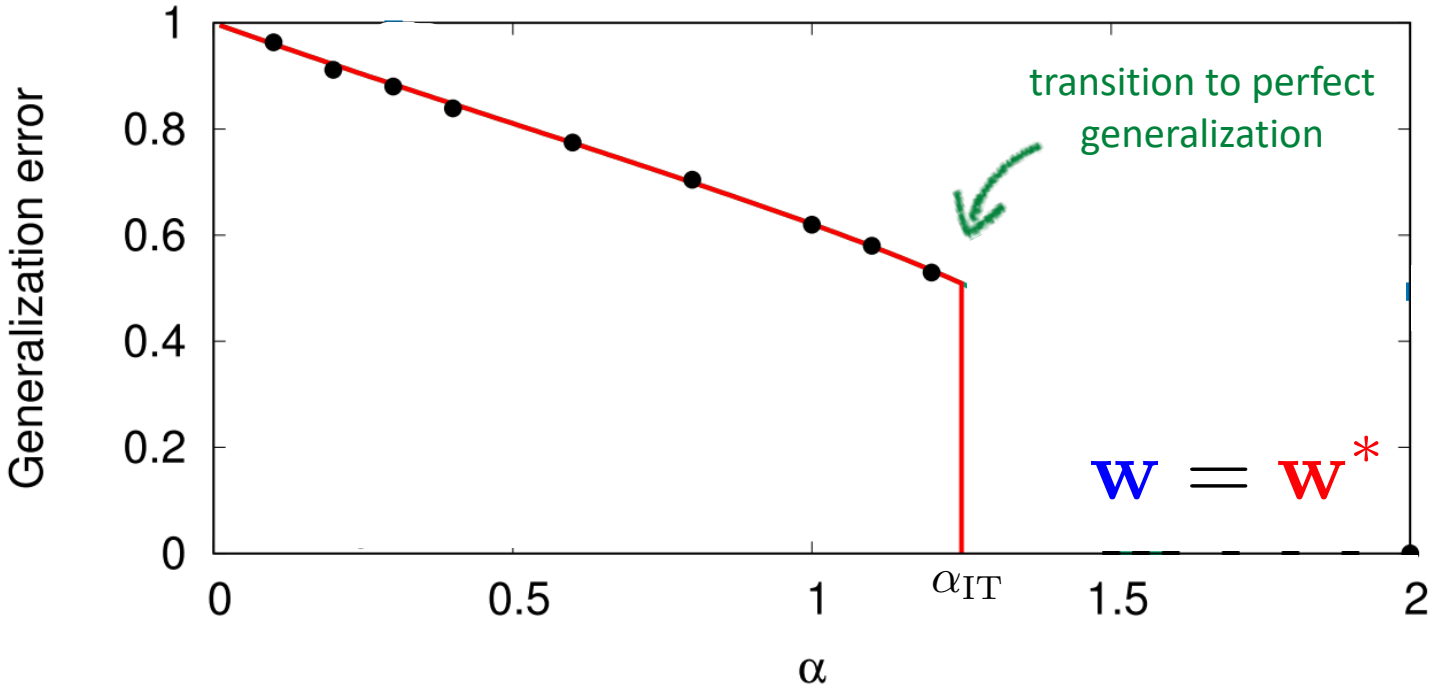
Engel, A., & Van den Broeck, C. (2001). Statistical Mechanics of Learning

Optimal generalization w.r.t. sample complexity

example II:

- binary input patterns $x_i^{(k)} = \pm 1$
- binary weights $w_i = \pm 1$

$$\alpha = P/N = \frac{\#\,\text{training points}}{\#\,\text{weights}}$$



- sharp change of behavior as a function of the control parameter
- discontinuous jump → analog of first order phase transition

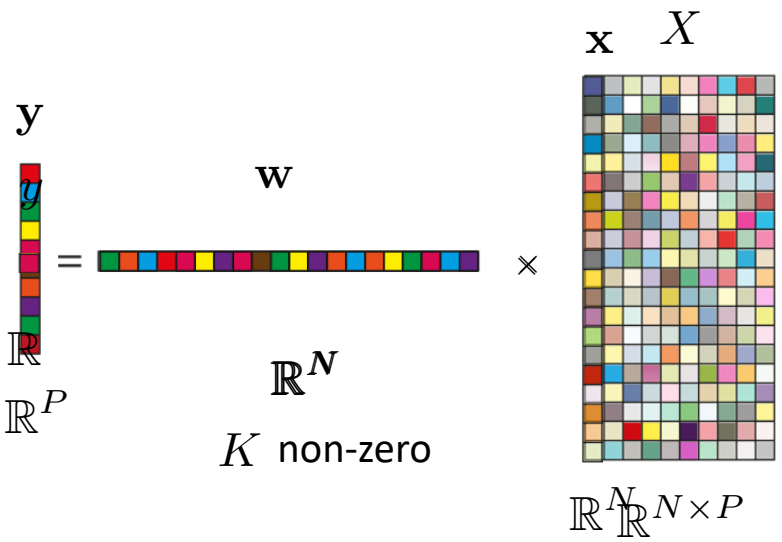
Györgyi, G. (1990). *First-order transition to perfect generalization in a neural network with binary synapses*
Krauth, W., & Mézard, M. (1989). *Storage capacity of memory networks with binary couplings*
Barbier, J. et al. (2018). *Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models*

More intuitions on the information theoretic limit

- **Linear activation** $y(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x}$ **?** $\alpha_{\mathrm{IT}} = 1$

- **Absolute value activation** $y(\mathbf{x}; \mathbf{w}) = |\mathbf{w}^T \mathbf{x}|$ **?** $\alpha_{\mathrm{IT}} = 1$

- **Linear activations with sparse weights ?** $\alpha_{\mathrm{IT}} = K/N = \rho$



Small recap 2

2. Teacher-students and Bayesian inference

A. Modelling learning 2.0
B. Bayesian inference
C. Optimality and information theoretic limits

Important concepts
- Bayesian inference
- Bayes optimal setting
- Information theoretic threshold

Illustrations
- Teacher-student perceptron
- Different activations and variables distributions

Perspective
- Is learning doable up to the information theoretic transition?

Plan of attack

1. **The curse of dimensionality, disordered systems physics and models**
   A. Storage of the perceptron
   B. Spin glasses, thermodynamic limit, and disorder averages
   C. Some models and results on the perceptron

2. **Teacher-students and Bayesian inference**
   A. Modelling learning 2.0
   B. Bayesian inference
   C. Optimality and information theoretic limits

3. **Algorithmic performance**
   A. Algorithm complexity
   B. Computational hard phase

4. **Dynamics of learning**
   A. Online learning in the perceptron
   B. Dynamics of overlaps and generalization error

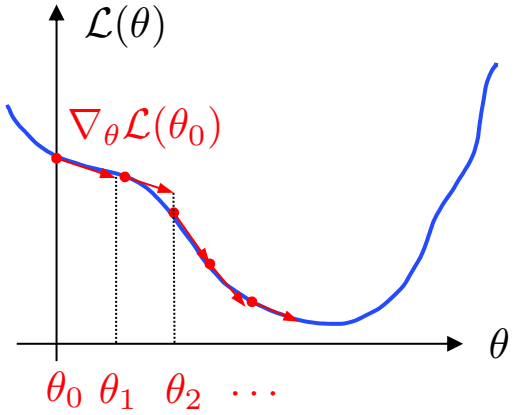
Algorithmic complexity

Important notion of learning theory:

algorithmic complexity = number of operations to execute the algorithm as a function of the size of the problem

e.g. gradient descent on N parameters for K steps:

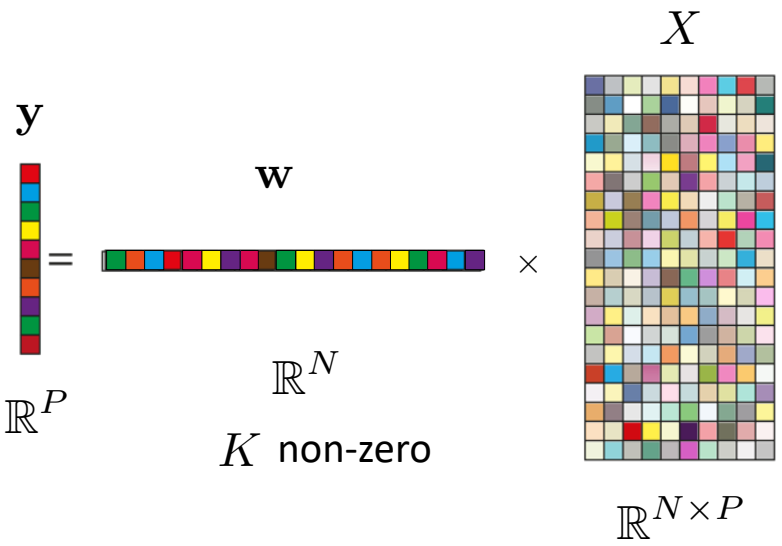
- N operations per step $\theta_i \leftarrow \theta - \eta \partial_{\theta_i} \mathcal{L}(\theta)$
- K steps
- algorithm complexity $O(N \times K)$



- **Polynomial time: good ! (Logarithmic: even better)**
- **Exponential time: not realistic for large problems !**

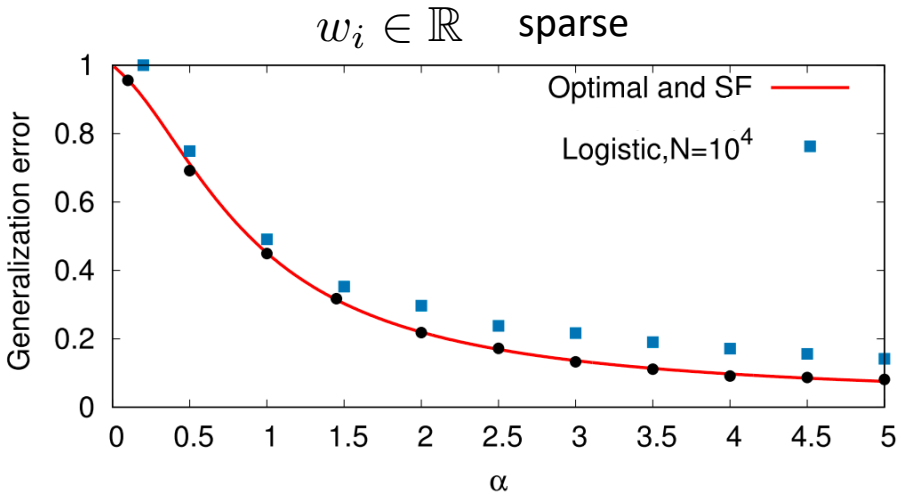
More intuitions on the information theoretic limit

- **Linear activation** $y(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x}$ **?** $\alpha_{\mathrm{IT}} = 1$
  - matrix inversion → **polynomial** time
- **Absolute value activation** $y(\mathbf{x}; \mathbf{w}) = |\mathbf{w}^T \mathbf{x}|$ **?** $\alpha_{\mathrm{IT}} = 1$
  - test all possible combination of signs for the rows → **exponential** time
- **Linear activations with sparse weights ?** $\alpha_{\mathrm{IT}} = K/N = \rho$
  - test all possible locations → **exponential** time

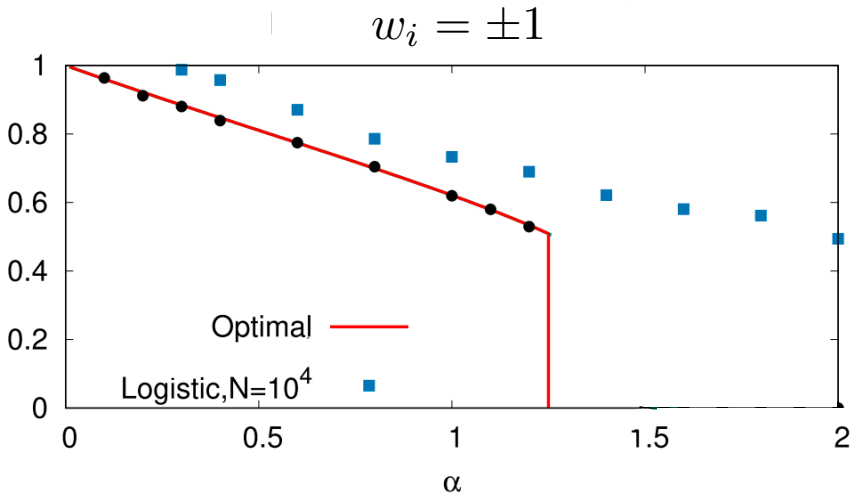


Perceptron learning with logistic regression

- **Logistic regression =**
  popular learning algorithm for binary classification based on maximum likelihood



- Only slightly worse than Bayes optimal



- Worse than Bayes optimal
- Missing transitions to optimal generalization

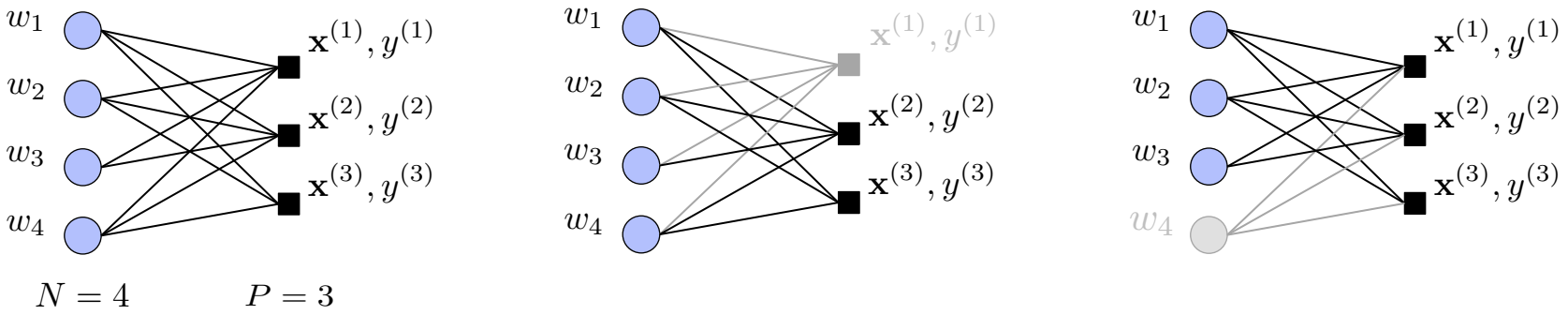
Barbier, J., et al (2018).
Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models.

A special class of algorithms: Message passing

Efficient algorithms to approximate joint distributions for certain models

- e.g. perceptron posterior $p_S(\mathbf{w}|\{y^{(k)}, \mathbf{x}^{(k)}\}_{k=1}^P) \propto \prod_{k=1}^P p_S(y^{(k)}|\mathbf{x}^{(k)}, \mathbf{w})\, p_S(\mathbf{w})$

representation of posterior (factor graph)



- output of algorithm: approximation of marginal $p_S(w_i|\{y^{(k)}, \mathbf{x}^{(k)}\}_{k=1}^P)$

Message passing, origin and optimality

- **Related to mean-field approximation in physics used to compute partition functions (a.k.a. high dimensional integrals)**

- **Rediscovered several times in different fields:**
  - Hans Bethe. *Statistical Theory of Superlattices*. **1935**
  - Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. **1988**

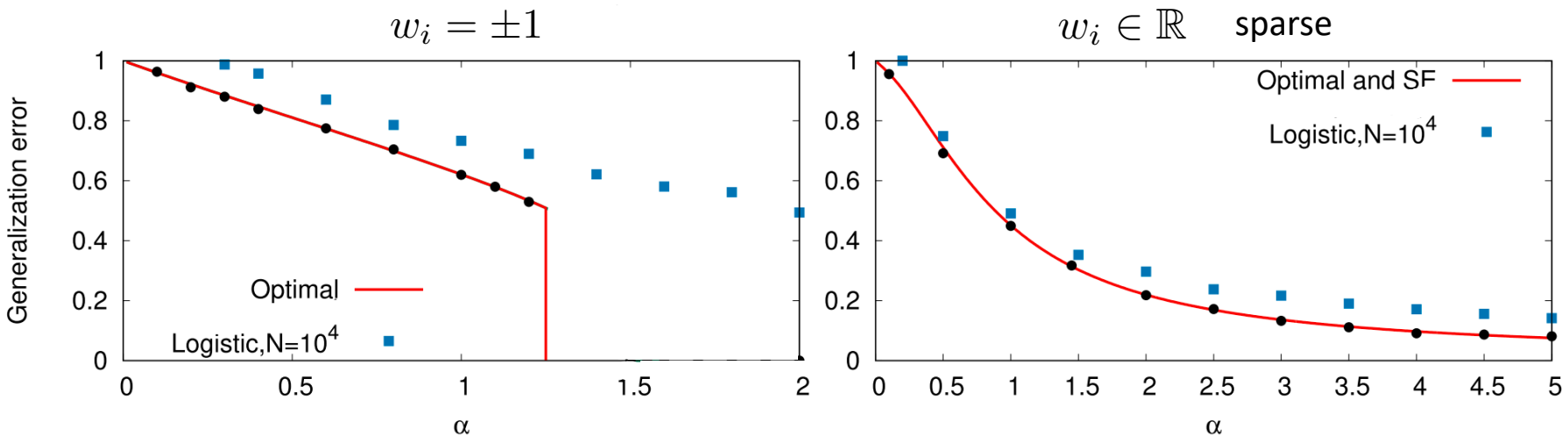
- **Many variants:**
  - Belief Propagation (BP)
  - Approximate Message Passing (AMP)
  - Expectation Propagation (EP)

- **Believed to be optimal among polynomial algorithms in certain cases:**
  - Either get to the Bayes optimal solution
  - Or achieve the lowest possible error among polynomial time algorithms

Very important fact !

Perceptron learning with logistic regression

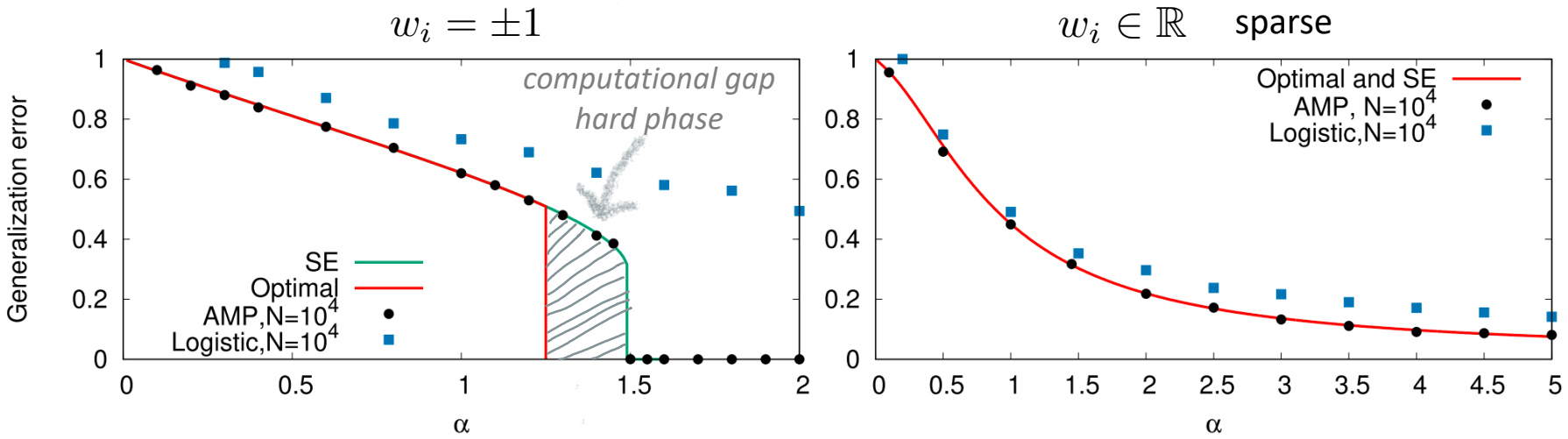
- **Logistic regression =**
  popular learning algorithm for binary classification based on maximum likelihood



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

Perceptron learning with logistic regression and AMP

- **Logistic regression =**
  popular learning algorithm for binary classification based on maximum likelihood
- **AMP = message passing algorithm believed to be optimal in this model**



- For small sample complexity follows Bayes opt
- Missing transitions to optimal generalization
- Eventually reaching zero

- Similar to Bayes optimal

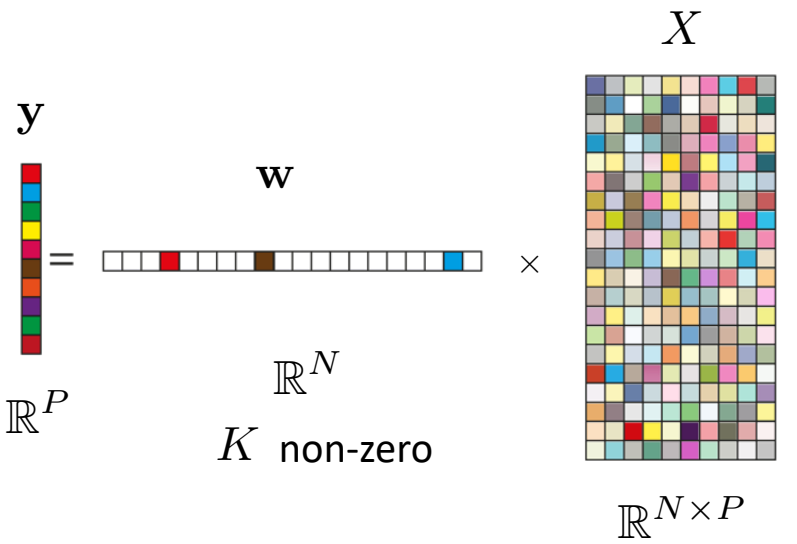
- **Discovery of a hard phase and a computational gap in certain cases**

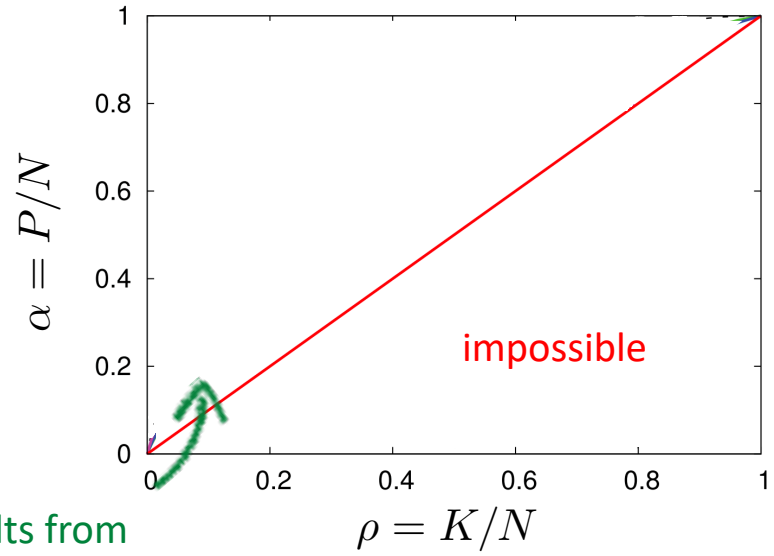
Barbier, J., et al (2018).
Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models.

More intuitions on the computational gap & hard phase

- **Linear activations with sparse weights ?** $y(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x}$

  - test all possible locations → exponential time $\alpha_{\rm IT} = K/N = \rho$





results from
message passing algorithms

Zdeborová, L., & Krzakala, F. (2016). Statistical physics of inference: Thresholds and algorithms

Small recap 3

3. Algorithmic performance

A. Algorithm complexity

B. Computational hard phase

Important concepts

- Algorithmic complexity
- Computational gaps
- Message passing algorithms

Illustrations

- Teacher-student perceptron
- With different activations

Perspective

- Can we study the dynamics of learning algorithms?

Plan of attack

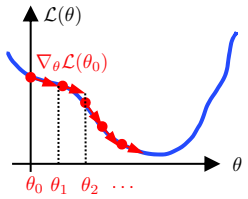
1. **The curse of dimensionality, disordered systems physics and models**
   A. Storage of the perceptron
   B. Spin glasses, thermodynamic limit, and disorder averages
   C. Some models and results on the perceptron

2. **Teacher-students and Bayesian inference**
   A. Modelling learning 2.0
   B. Bayesian inference
   C. Optimality and information theoretic limits

3. **Algorithmic performance**
   A. Algorithm complexity
   B. Computational hard phase

4. **Dynamics of learning**
   A. Online learning in the perceptron
   B. Dynamics of overlaps and generalization error

Online stochastic gradient descent



- **data stream** $(\mathbf{x}^{(1)}, y^{(1)}),\ (\mathbf{x}^{(2)}, y^{(2)}),\ \cdots,\ (\mathbf{x}^{(k)}, y^{(k)}),\ \cdots$ **i.i.d samples**
- **update for each data point** $\mathbf{w}^{k+1} = \mathbf{w}^k - \eta \nabla_{\mathbf{w}} \ell\big(\mathbf{w}^k; (\mathbf{x}^{(k)}, y^{(k)})\big)$
  - ▹ time index = sample index: k
- **use each data point only once !**

Online = stochastic gradient descent with mini-batch of size one
+ possibly infinite data set

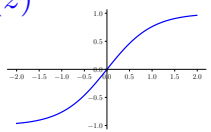
<u>Uncorrelated training data between updates</u> → eased the analysis !

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.

Online SGD for the student perceptron

Analysis for

$f(z) = \mathrm{erf}(z)$



- perceptron with continuous output $y(\mathbf{x}) = f(\mathbf{w}^\top \mathbf{x})$
- **matched** teacher and student $y^*(\mathbf{x}) = f(\mathbf{w}^{*\top}\mathbf{x})$
- gaussian centered inputs $p_x(\mathbf{x}) = \mathcal{N}(\mathbf{x}; 0, I_N)$
- square loss $\ell(\mathbf{x}; \mathbf{w}, \mathbf{w}^*) = \frac{1}{2}\left(y^*(\mathbf{x}) - y(\mathbf{x})\right)^2 = \frac{1}{2}\left(f(\mathbf{w}^{*\top}\mathbf{x}) - f(\mathbf{w}^\top \mathbf{x})\right)^2$

- **Weight vector update** $\mathbf{w}^{k+1} = \mathbf{w}^k - \eta \nabla_{\mathbf{w}} \ell\left(\mathbf{w}^k; (\mathbf{x}^{(k)}, y^{(k)})\right)$
- **Overlap variables** $Q^k = \mathbf{w}^\top \mathbf{w}^k$

  $R^k = \mathbf{w}^{*\top} \mathbf{w}^k$

alignment of student vector with teacher vector

- **Generalization error** $\mathcal{E}_g(\mathbf{w}^k, \mathbf{w}^*) = \mathbb{E}_x\left[\frac{1}{2}\left(y^*(\mathbf{x}) - y_{\mathbf{w}^k}(\mathbf{x})\right)^2\right] = \mathcal{E}_g(R^k, Q^k)$

Biehl & Riegler. (1994) *On-Line Learning with a Perceptron*

Thermodynamic limit = continuous time limit + disorder average

Online update rule $\mathbf{w}^{t+1} - \mathbf{w}^t = \frac{\eta}{N}\left(f(\mathbf{w}^\top \mathbf{x}) - f(\mathbf{w}^{*\top}\mathbf{x})\right)\nabla_{\mathbf{w}} f(\mathbf{w}^\top \mathbf{x})$

Thermodynamic limit $N \to \infty$

Disorder average $\begin{cases} h = \mathbf{w}^T \mathbf{x} \\ h^* = \mathbf{w}^{*T}\mathbf{x} \end{cases}$ ← correlated Gaussian variables $\begin{cases} \text{mean} & \mathbb{E}_x[h] = \mathbb{E}_x[h^*] = 0 \\ \text{variance} & \mathrm{Var}_x[h] = Q \\ \text{covariance} & \mathbb{E}_x[h^* h] = R \end{cases}$

Continuous time limit $\mathbf{w}^{t+1} - \mathbf{w}^t \propto \frac{\eta}{N} \quad \to \quad \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} \qquad t = k/N$

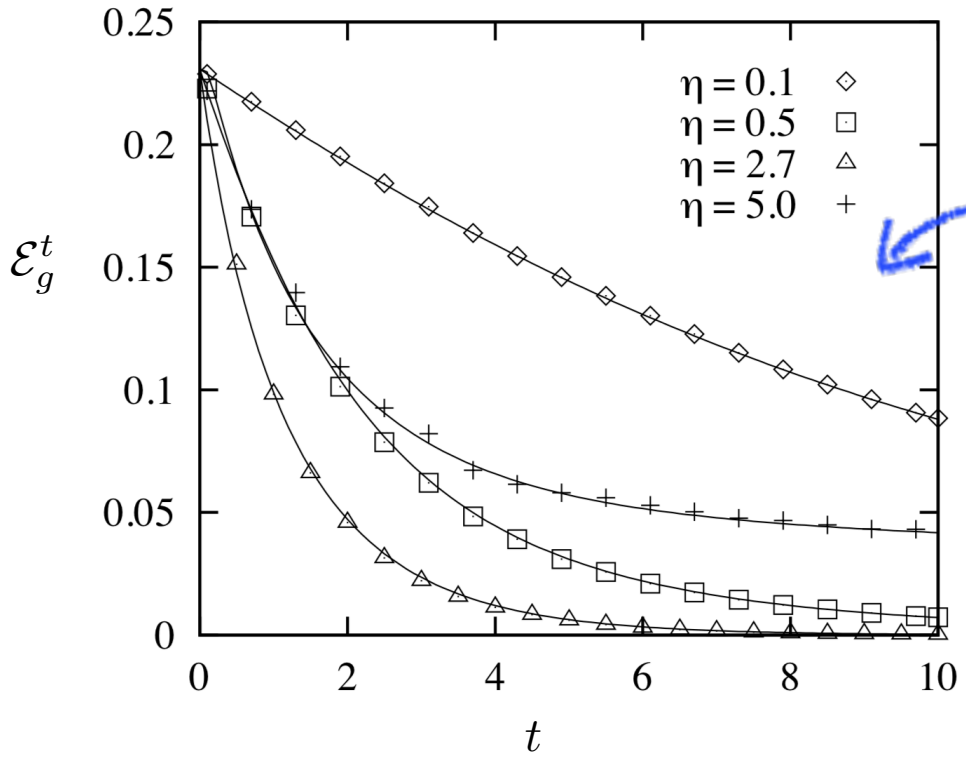
$$R^{k+1} - R^k \to \frac{\mathrm{d}R}{\mathrm{d}t} \qquad Q^{k+1} - Q^k \to \frac{\mathrm{d}Q}{\mathrm{d}t}$$

Closed set of time evolution equations

Finally

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

→ generalization error $\mathcal{E}_g^t = \mathcal{E}_g(R^t, Q^t)$



different convergence rates
for different learning rate

not decaying to zero
for too large learning rate

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

Small recap 4

4. Dynamics of learning

A. Online learning in the perceptron

B. Dynamics of overlaps and generalization error

Important concepts

- Fully online learning
- Overlap variables
- Closed set of equations for asymptotic dynamics

Illustrations

- Teacher-student perceptron
- with erf activations and Gaussian inputs

Plan of attack

1. **The curse of dimensionality, disordered systems physics and models**
   A. Storage of the perceptron
   B. Spin glasses, thermodynamic limit, and disorder averages
   C. Some models and results on the perceptron

2. **Teacher-students and Bayesian inference**
   A. Modelling learning 2.0
   B. Bayesian inference
   C. Optimality and information theoretic limits

3. **Algorithmic performance**
   A. Algorithm complexity
   B. Computational hard phase

4. **Dynamics of learning**
   A. Online learning in the perceptron
   B. Dynamics of overlaps and generalization error

What else can done in this direction of research?

Analyze other types of models?

- Support vector machines
- A type of two layers architectures (committee machines)
- Deep neural networks at initialization
- Matrix factorization
- Community detection on graphs
- etc …

Parallel track of research make the predictions of mean-field methods rigorous!

- Talagrand, M. (2006). The Parisi formula.
- Reeves, G. (2018). *Additivity of information in multilayer networks via additive Gaussian noise transforms*
- Gabrié, M., et al. (2018). *Entropy and mutual information in models of deep neural networks.*
- Barbier, J., et al. (2018). *The Mutual Information in Random Linear Estimation Beyond i.i.d. Matrices.*
- Barbier, J., et al. (2017). *The mutual information in random linear estimation.*
- Aubin, B., et al. (2018). *The committee machine: Computational to statistical gaps in learning a two-layers neural network.*
- *etc …*

A few applications to design practical learning algorithms

- One example in tutorial !

References selected

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*.

References all 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

References all 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