Progress and hurdles in the statistical mechanics of deep learning

July 23rd 2020

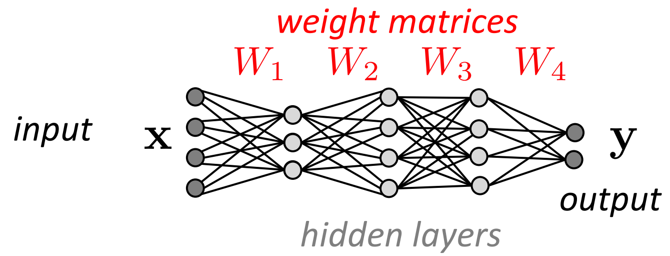
IST seminar series Mathematics, Physics & Machine Learning

Marylou Gabrié (NYU, Flatiron Institute)

Collaborators:

Andre Manoel (Owkin), Jean Barbier (ICTP Trieste), Clément Luneau (EPFL), Nicolas Macris (EPFL), Florent Krzakala (ENS Paris), Lenka Zdeborova (IPHT Saclay)

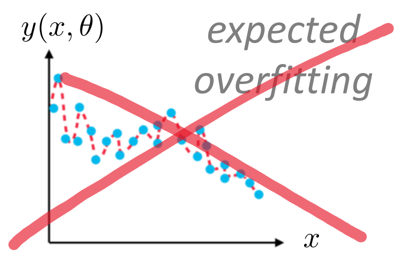
Understanding machine learning with deep neural nets

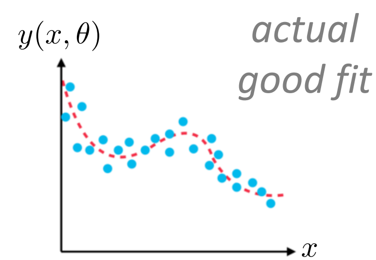


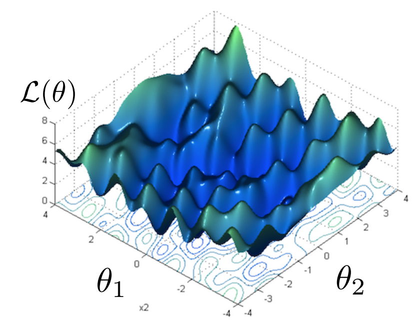
- **Supervised learning with neural networks**
  - training data $\mathcal{D} = \{\mathbf{y}^{(k)}, \mathbf{x}^{(k)}\}_{k=1}^{P}$
  - fit with class of parametrized functions $\mathbf{y} = f(\mathbf{W}_L f(\mathbf{W}_{L-1} \ldots f(\mathbf{W}_1 \mathbf{x})))$

- **Impressive performances (automatic vision, natural language processing etc.)**

- **Interesting properties**
  - universal approximators
  - not prone to overfitting
  - train with local descent algorithm despite of non-convexity





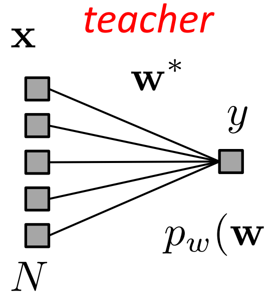


Statistical mechanics of learning, initiated in the 80s

▻ **Focus on simple (solvable) models**

$p_x(\mathbf{x}) =$ *i.i.d*

teacher

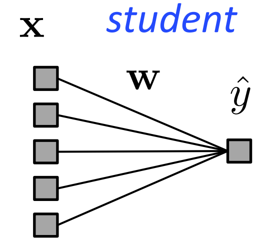


$p_w(\mathbf{w}^*) =$ *i.i.d*

training data

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

student



▻ **Consider the Bayesian posterior statistics** $p(\mathbf{w}|\mathcal{D}) = \dfrac{p(\mathcal{D}|\mathbf{w})\, p_w(\mathbf{w})}{p(\mathcal{D})}$

e.g. Bayes optimal estimator (minimum mean square error)

$$\min_{\hat{\mathbf{w}}} \int \mathrm{d}\mathbf{w}\, (\mathbf{w} - \hat{\mathbf{w}})^2\, p_S(\mathbf{w}\mathcal{D}) \longrightarrow \hat{\mathbf{w}}_{\mathrm{MMSE}} = \int \mathrm{d}\mathbf{w}\, \mathbf{w}\, p_S(\mathbf{w}|\mathcal{D})$$

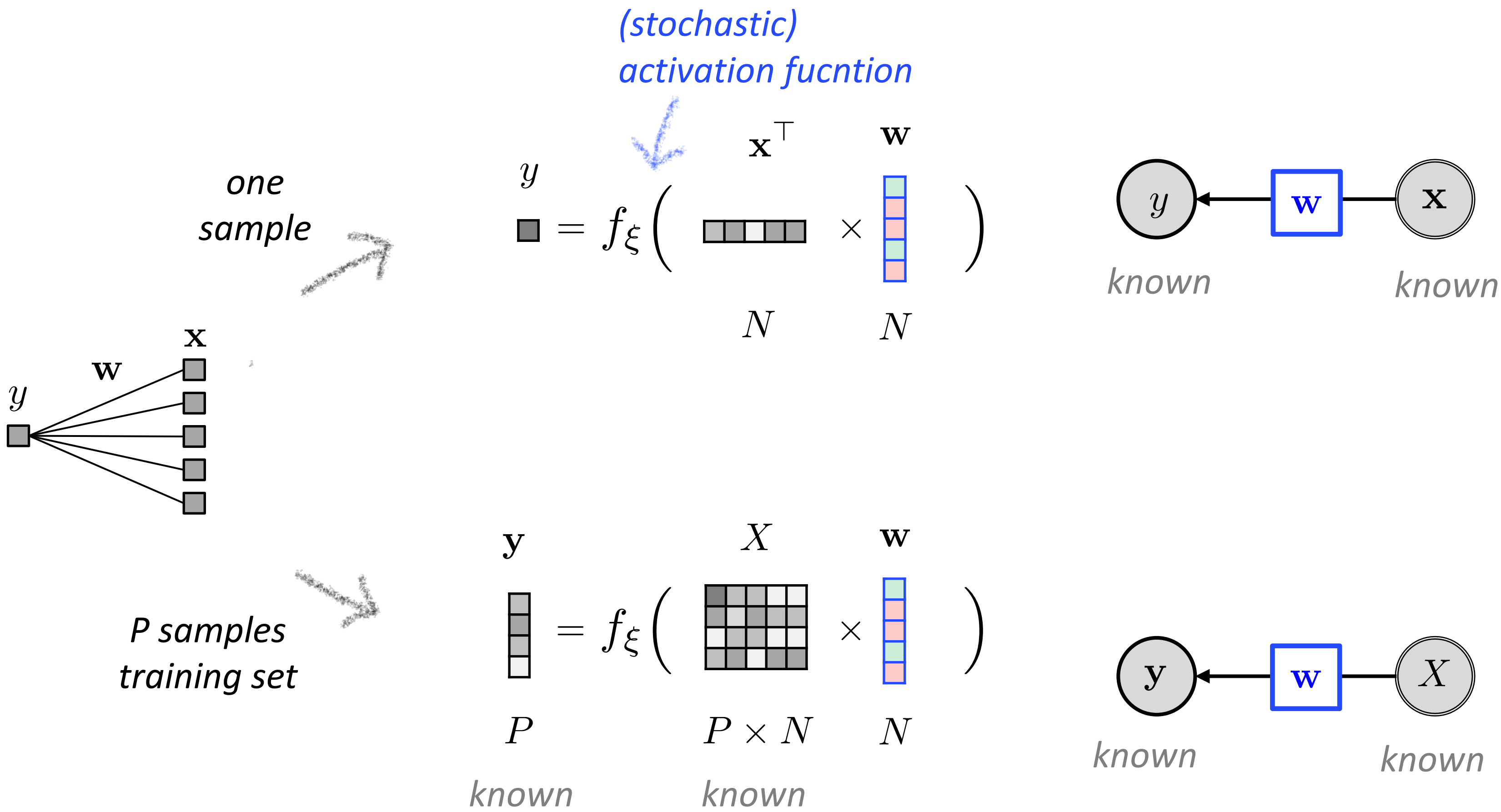
▻ **The thermodynamic limit = infinitely large model**
→ typical cases concentrate at the average $N \to \infty$ $\alpha = P/N$

Mean-field tools from the stat. phys. of disordered systems:

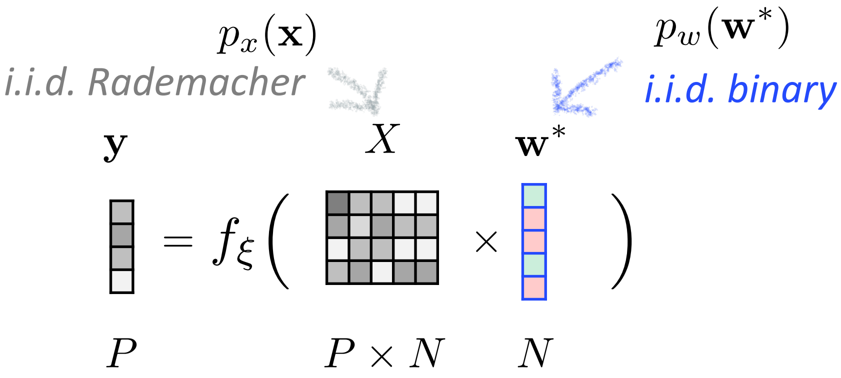
e.g. Bayes optimal square error $$\mathrm{MMSE}(\alpha) = \lim_{N\to\infty} \frac{1}{N} \int \mathrm{d}\mathbf{w}\, (\mathbf{w} - \hat{\mathbf{w}}_{\mathrm{MMSE}})^2\, p_S(\mathbf{w}|\mathcal{D})$$

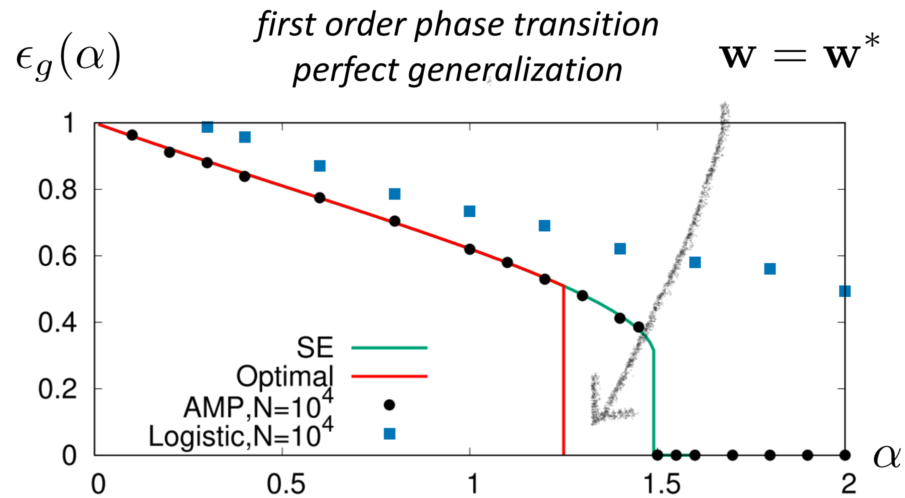
Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985) Amit, D. J., Hanoch, G., & Sompolinsky, H. (1986) Gardner, E. (1987) E. Gardner, B. Derrida, P. Mottishaw (1987), B. Derrida, J.P. Nadal (1987), Peterson, C., & Anderson, J. R. (1987), W. Krauth, M. Mézard (1987), Gardner, E. (1988), Mézard, M. (1989). Györgyi, G. (1990). Opper, M., & Haussler, D. (1991) H.S. Seung, H. Sompolinsky, N. Tishby (1992), T.L.H. Watkin, A. Rau, M. Biehl. (1993), Monasson, R., Zecchina, R. (1995) etc..

Starting point:
Perceptron a.k.a Generalized Linear Model (GLM)



Statistical mechanics of the Perceptron





1. Teacher-student / planted problem
2. Bayesian optimal
3. Mean-field analysis / typical case

generalization error:

teacher *student*

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

$$\epsilon_g(\alpha) = \lim_{N\to\infty} \frac{1}{N}\epsilon_g(N,P) \qquad \alpha = P/N$$

✓ Information theoretic analysis with mean-field replica method [1]

✓ Rigorously proven [2]

✓ (Generalized) Approximate Message Passing (AMP) algorithm [3]

✓ State evolution statistical analysis of algorithm performance [3]

[1] Györgyi (1990). *First-order transition to perfect generalization in a neural network with binary synapses*
[2] Barbier, et al. (2018). *Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models*
[3] Rangan (2011) *Generalized Approximate Message Passing for Estimation with Random Linear Mixing.*

Mean-Field methods for statistical inference analysis
The tools

Information theoretic analysis

Non-rigorous computations of asymptotic posterior statistics

replica method

Mathematical rigorous proofs of the conjecture

Guerra interpolation,
Adaptive interpolation

Algorithms

Message passing algorithms for inference on finite size models

belief propagation (BP),
approximate message passing (AMP),
expectation propagation (EP)

Statistical analysis of asymptotic performance of message passing algorithms

state evolution (SE)

high temperature expansions
(naïve MF, TAP)

"Mean-field approximations" in deep learning literature

- more general than tools above

- neglect correlations thanks to randomness in the thermodynamic (large-size) limit

non-exhaustive!

*** Analysis of statistical of inference** ← Focus of this talk

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.*

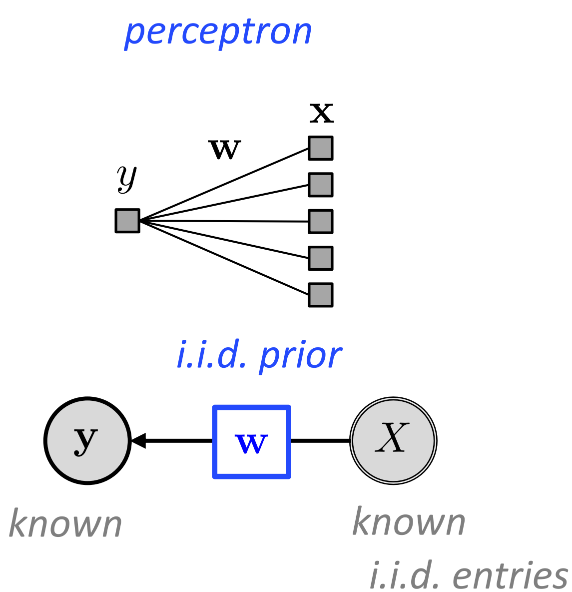
*** Role of over-parametrization in trainability with Gradient Descent methods**

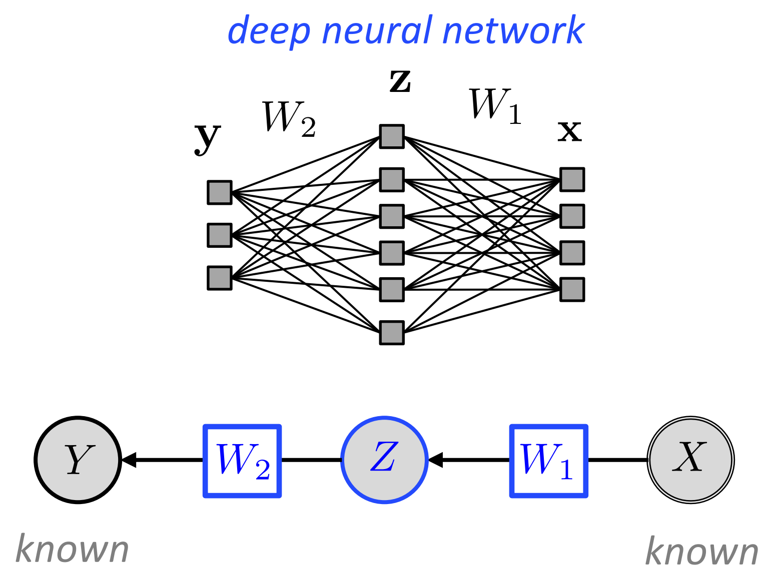
- Convergence of SGD for 2-layers neural networks
  Chizat & Bach (2018), Mei, Montanari & Nguyen (2018), Rotskoff & Vanden-Eijnden (2018)
- Neural Tangent Kernels, Equivalence to Gaussian processes, "Lazy training"
  Jacot et al (2018), Lee et al (2019), review: Bahri et al (2020) *Statistical Mechanics 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

*** Gradient Descent algorithms and landscape interactions**

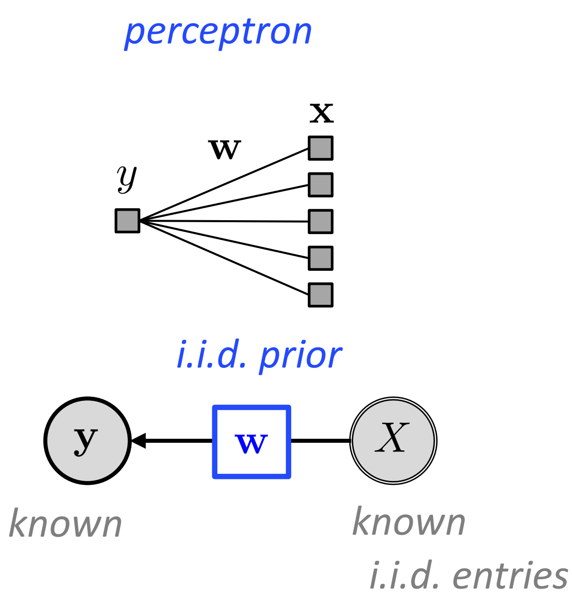
Dauphin et al (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization
Sarao Mannelli & Zdeborova (2020). *Thresholds of descending algorithms in inference problems.*

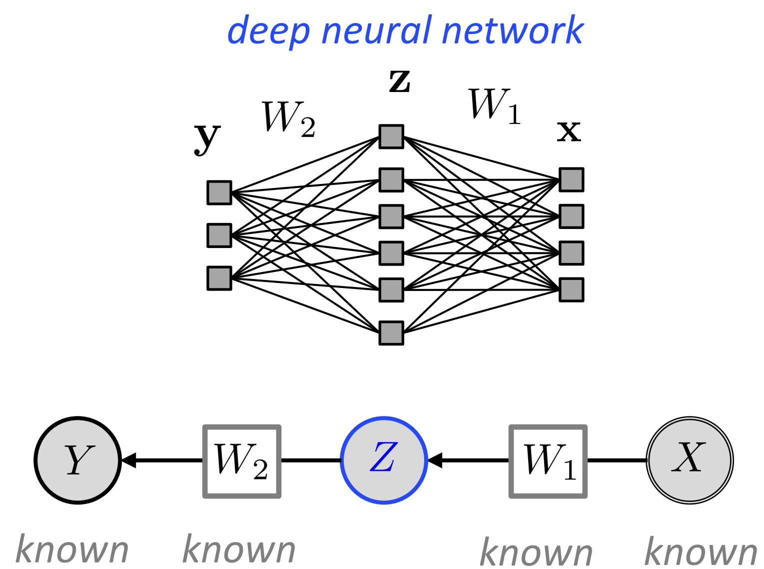
From perceptron/GLM with random i.i.d. matrices to deep neural networks ?





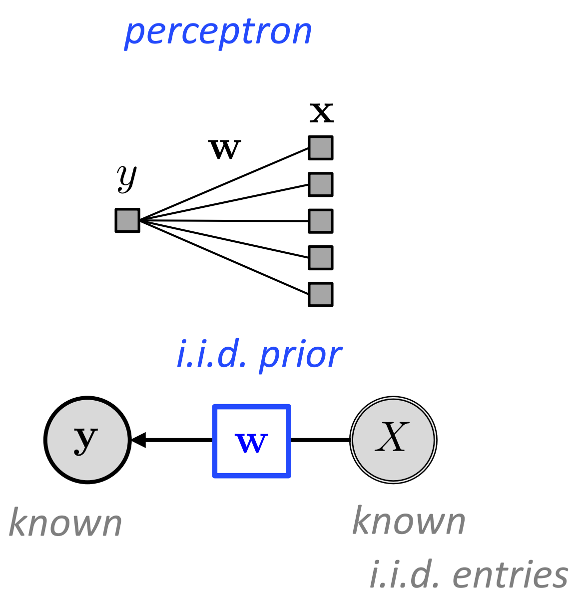
From perceptron/GLM with random i.i.d. matrices to deep neural networks ?

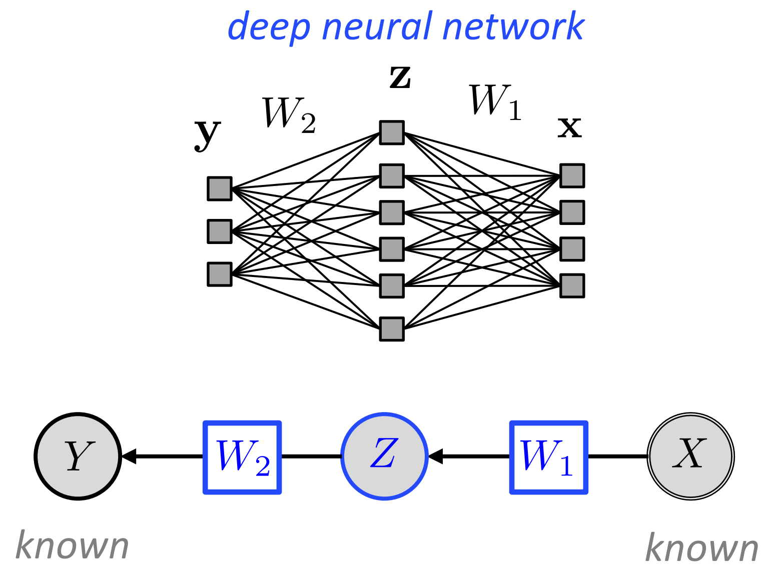




1. **Inference of layers variables in deep networks (with learned weight matrices)**
2. **The challenge of weight inference and structured weights**

From perceptron/GLM with random i.i.d. matrices to deep neural networks ?

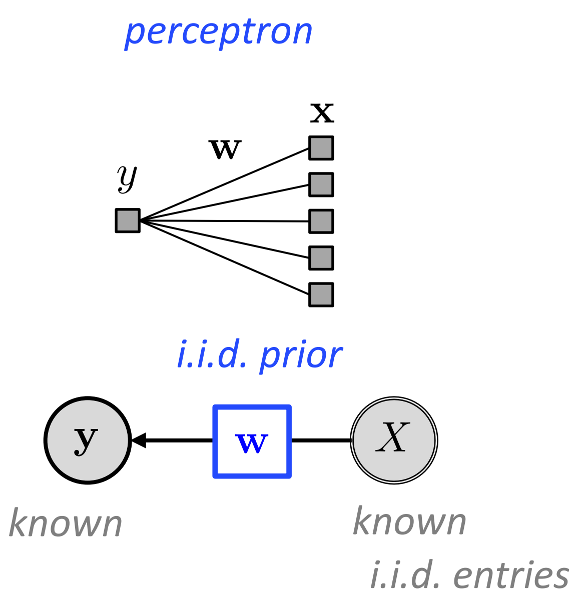


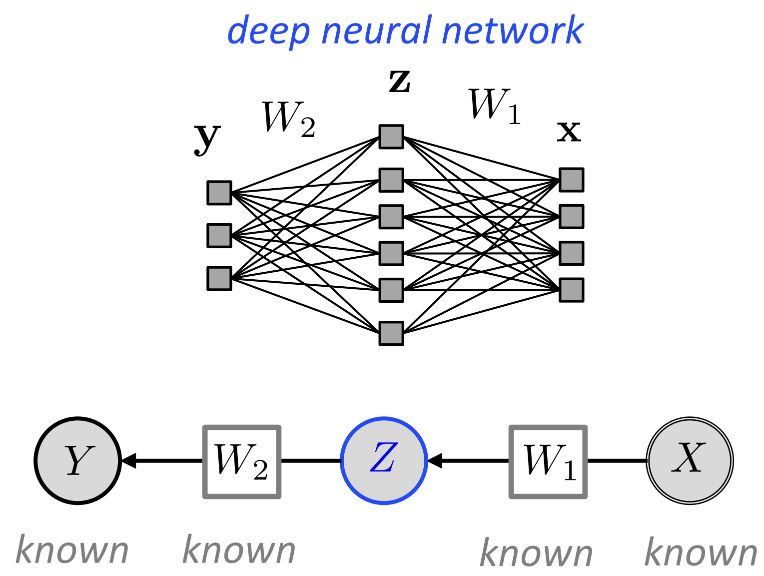


1. **Inference of layers variables in deep networks (with learned weight matrices)**

2. **The challenge of weight inference and structured weights**

From perceptron/GLM with random i.i.d. matrices to deep neural networks ?



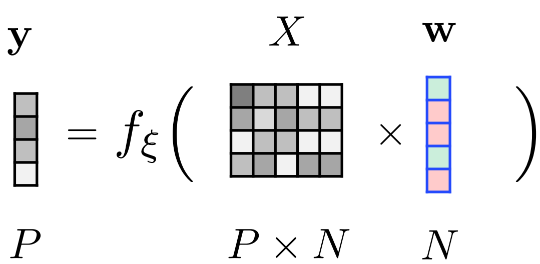


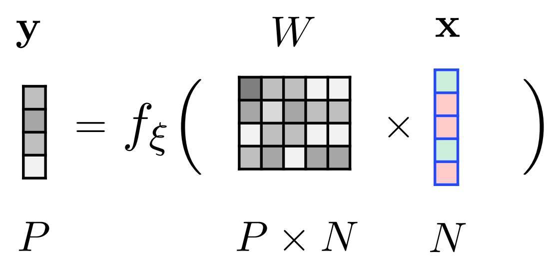
1. **Inference of layers variables in deep networks (with learned weight matrices)**

2. **The challenge of weight inference and structured weights**

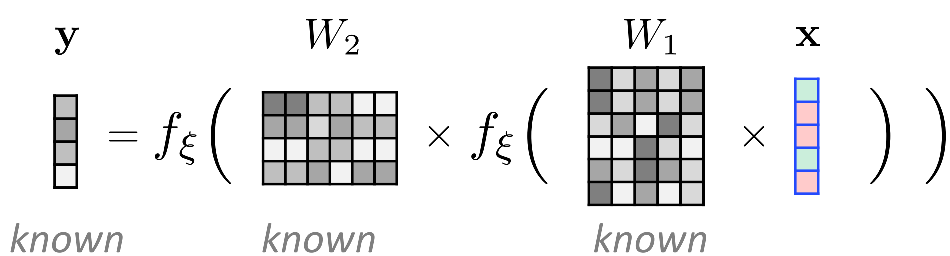
Inferring neural networks layer states from output

Single layer = perceptron / GLM

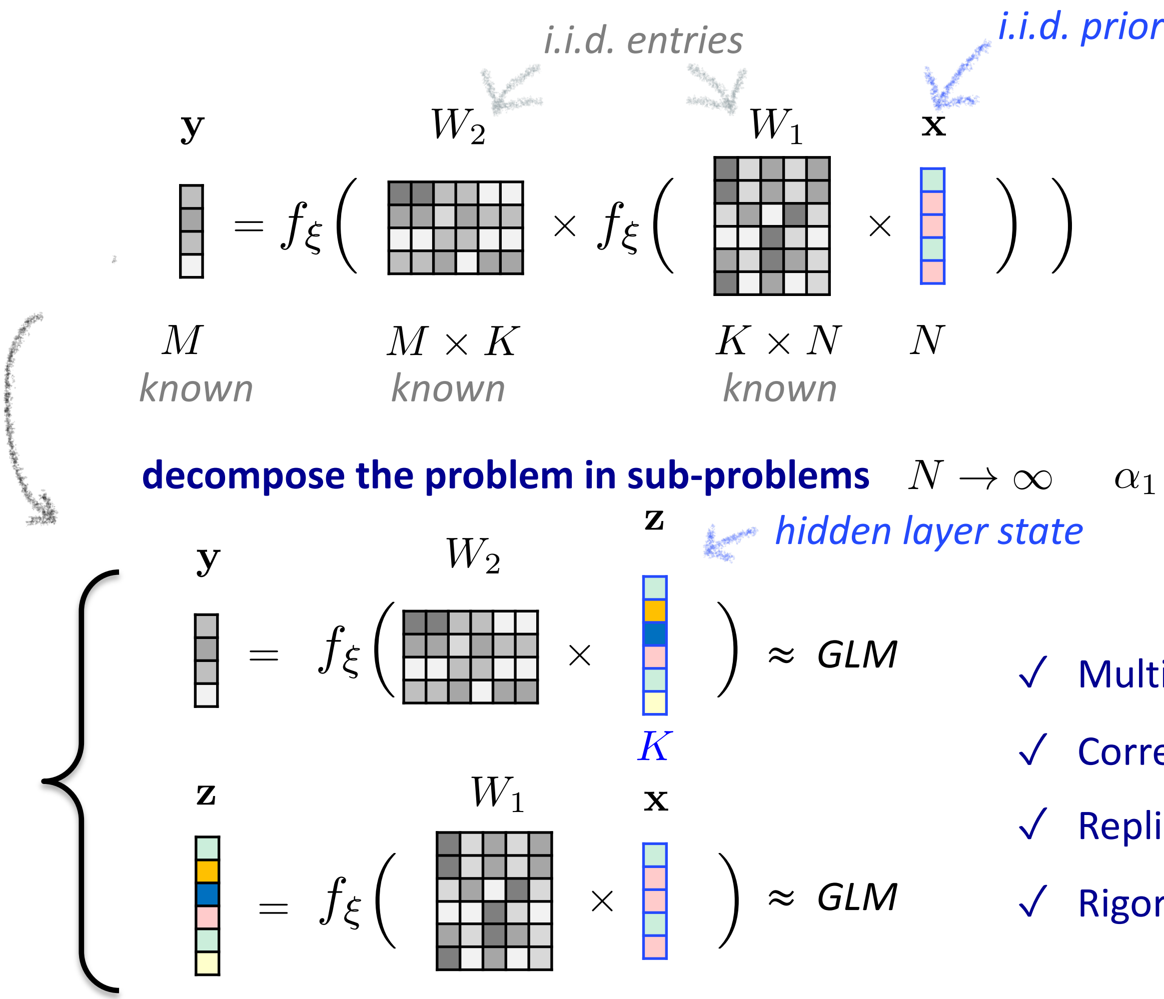




Multi-layer ?



Layers inference in deep neural network with i.i.d weights



N *input dim*

K *hidden dim*

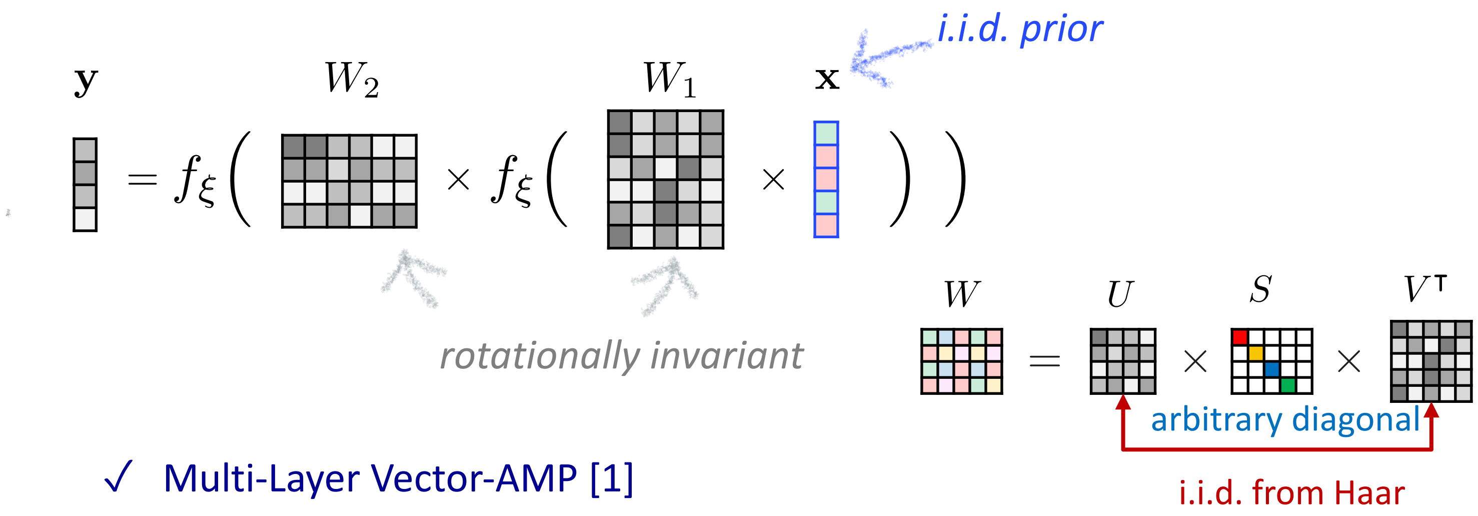
M *output dim*

decompose the problem in sub-problems $N \to \infty \quad \alpha_1 = K/N \quad \alpha_2 = M/N$

- ✓ Multi-Layer AMP (arbitrary depth) [1]
- ✓ Corresponding state evolution (SE) [1]
- ✓ Replica free energy (mmse, entropy) [1]
- ✓ Rigorously proven for 2 layers [2, 3]

[1] Manoel et al (2017) *Multi-layer generalized linear estimation.*
[2] Gabrié et al (2018) *Entropy and mutual information in models of deep neural networks.*
[3] Reeves (2018) *Additivity of Information in Multilayer Networks via Additive Gaussian Noise Transforms.*

Layers inference in deep neural network with weight matrices with correlations



✓ Multi-Layer Vector-AMP [1]

✓ Corresponding state evolution [1]

✓ Replica free energy (mmse, entropy) [2, 3]
(extension of single layer formula by [4])

✗ Proof ?

[1] Fletcher et al (2018*) Inference in deep networks in high dimensions.*

[2] Gabrié et al (2018) *Entropy and mutual information in models of deep neural networks.*

[3] Reeves (2018) *Additivity of Information in Multilayer Networks via Additive Gaussian Noise Transforms*.

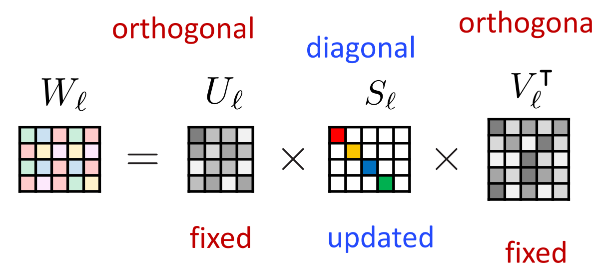
[4] Shinzato & Kabashima (2009) Learning from correlated patterns by simple perceptrons

Explicit weight learning, empirical verification

Learning the weight matrices while remaining rotationally inv.?

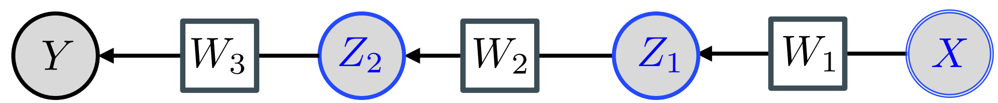
e.g. with gradient descent

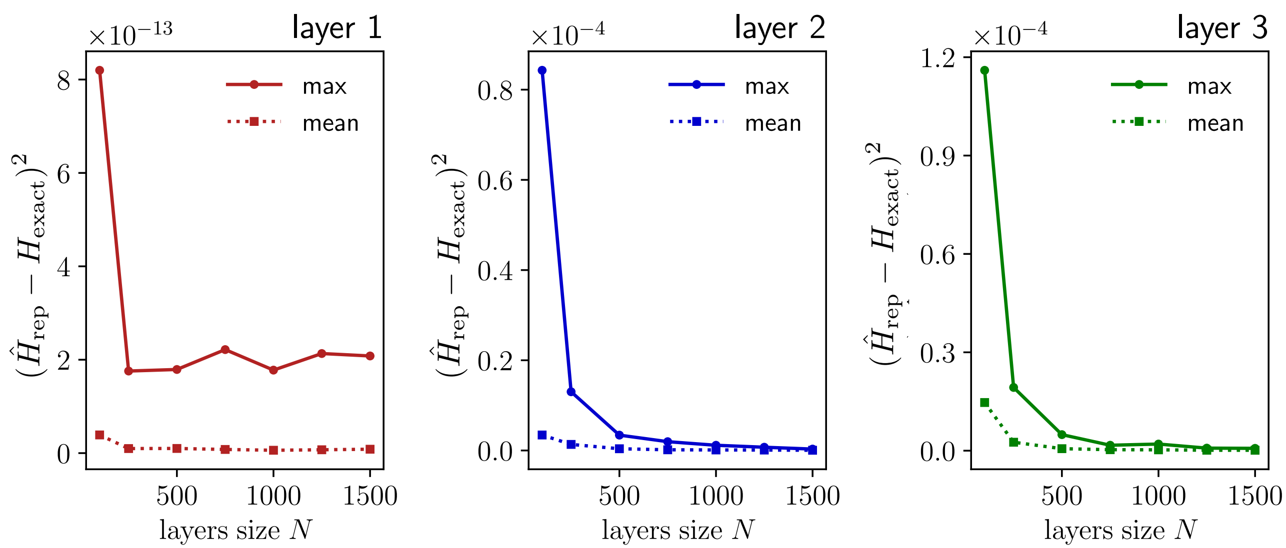
- Initialize Gaussian i.i.d W matrices
- Singular value decomposition
- Only learn spectrum (N degrees of freedom instead of N^2)



Numerical verification?

- Linear networks trained
- Gaussian inputs



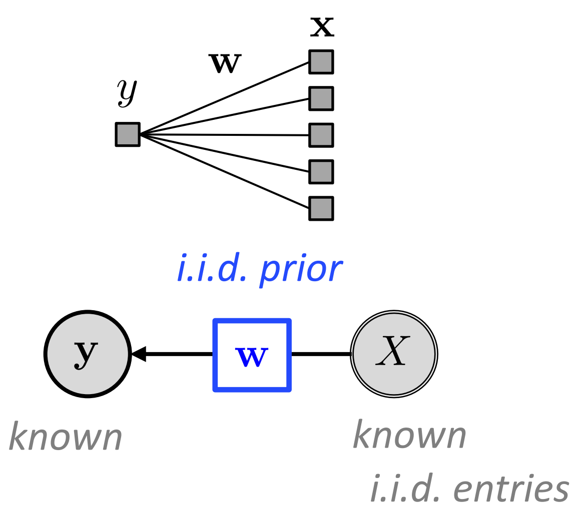


Replica correct with learned matrices

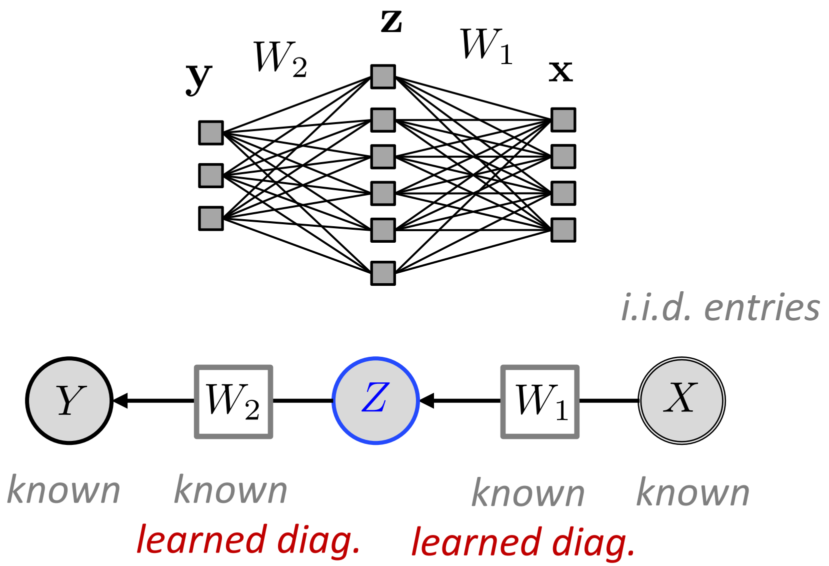
[1] Gabrié et al (2018) *Entropy and mutual information in models of deep neural networks.*

From perceptron/GLM with random i.i.d. matrices to deep neural networks ?

perceptron



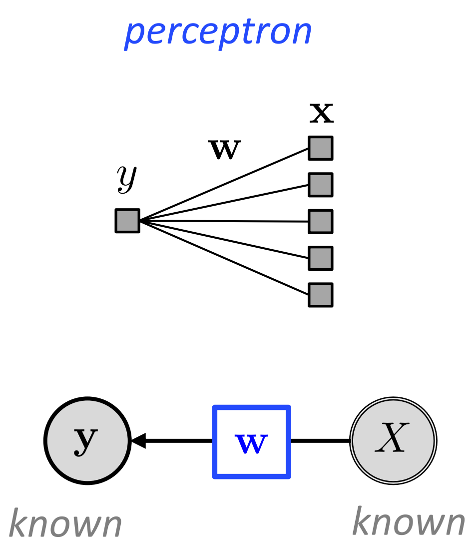
deep neural network

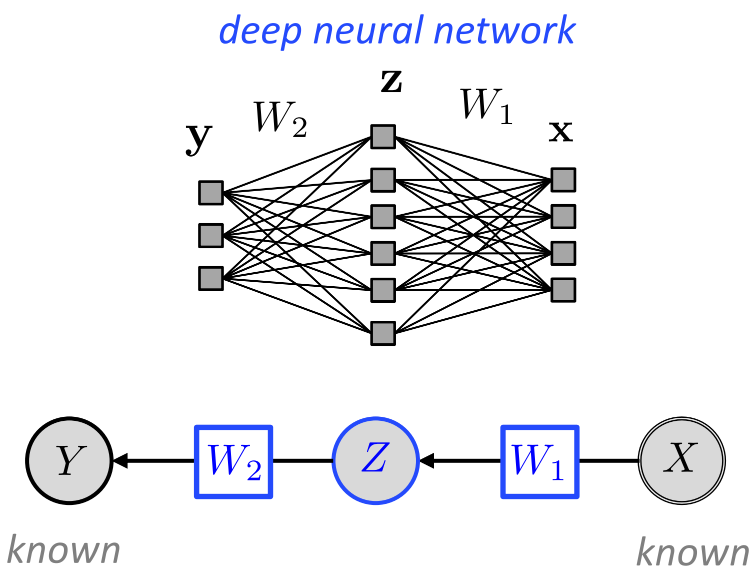


1. **Inference of layers variables in deep networks (with learned weight matrices)**

2. **The challenge of weight inference and structured weights**

From GLM with random i.i.d. matrices to deep neural networks ?





1. **Inference of layers variables in deep networks (with learned weight matrices)**

2. **The challenge of weight inference and structured weights**

Weight inference in deep neural networks decomposed

$$\underset{\substack{M \times P \\ \text{known}}}{Y} = f_\xi\Big(\underset{M \times K}{W_2} \times f_\xi\big(\underset{K \times N}{W_1} \times \underset{\substack{N \times P \\ \text{known}}}{X} \big)\Big)$$

N *input dim*

K *hidden dim*

M *output dim*

P *sample size*

First idea: decompose the inference in sub-problems

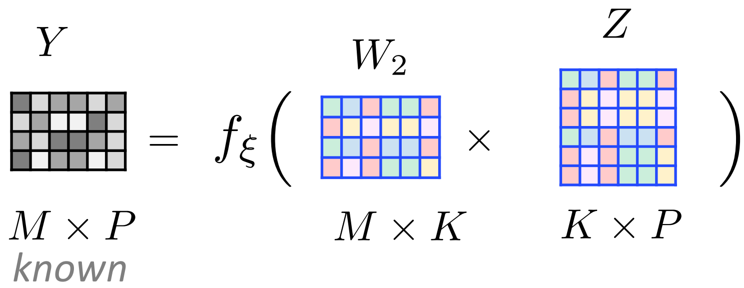
(alike Multi layer - AMP)

hidden layer states over the P samples

▹ $\underset{\text{known}}{Y} = f_\xi\big(W_2 \times \underset{K \times P}{Z} \big)$ ≈ *matrix factorization with rank K*

▹ $Z = f_\xi\big(W_1 \times \underset{\text{known}}{X} \big)$ ≈ *P x GLMs*

Scaling of the size of the hidden layer?



≈ matrix factorization with rank K

$N \to \infty$

▹ $K = O(1)$

- "low-rank matrix factorization": good mean field understanding [1, 2]
- finite number of hidden units, committee machines: great body of work! [3, 4, 5, 6, ..]

[1] Lesieur et al (2016), *MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel*

[2] Lesieur et al (2017), *Constrained Low-rank Matrix Estimation: Phase Transitions, Approximate Message Passing and Applications*

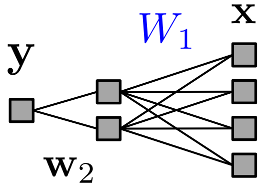
[3]Aubin et al (2018). *The committee machine: Computational to statistical gaps in learning a two-layers neural network*

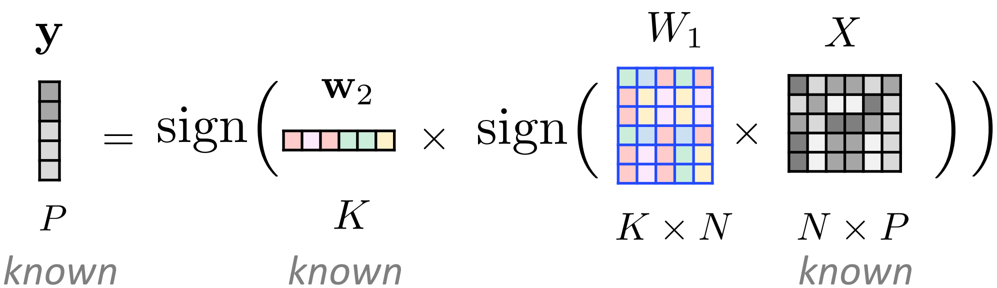
[4] Monasson et ql (2004). *Learning and Generalization Theories of Large Committee-Machines*

[5] Schwarze & Hertz (1993). *Generalization in Fully Connected Committee Machines.*

[6] Schwarze (1993). *Learning a Rule in a Multilayer Neural-Network.*

Phase transitions for committee machines





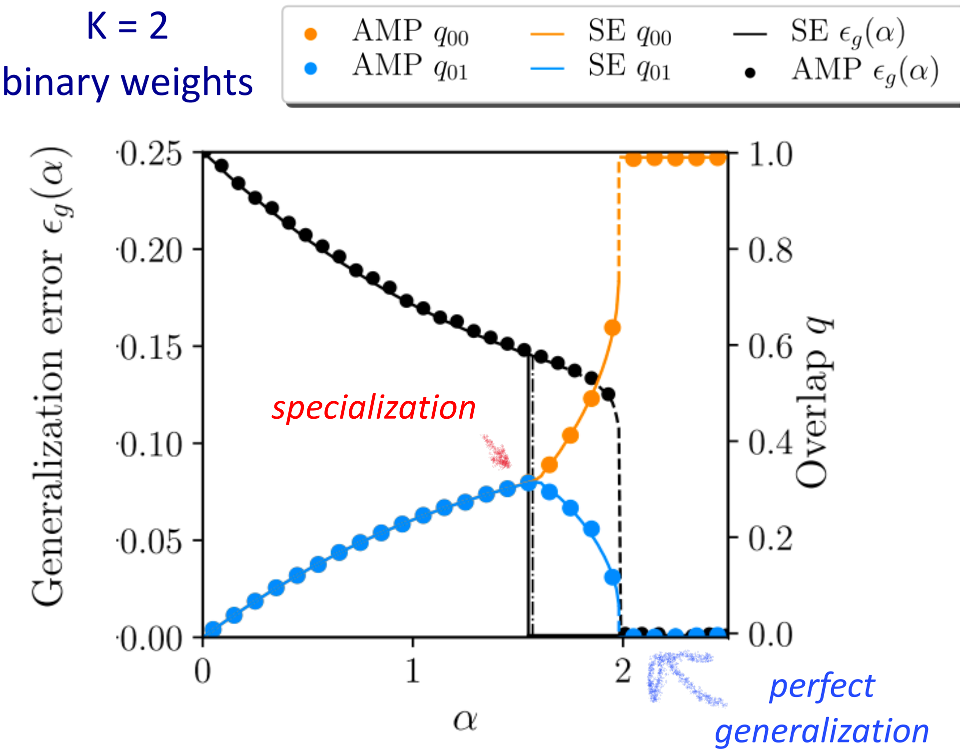
N *input dim*
K *hidden dim*
P *sample size*

K = 2
binary weights

- ✓ Committee-AMP [1]
- ✓ Corresponding state evolution [1]
- ✓ Replica free energy (mmse) [2, 3, 4]
- ✓ Proof [1]

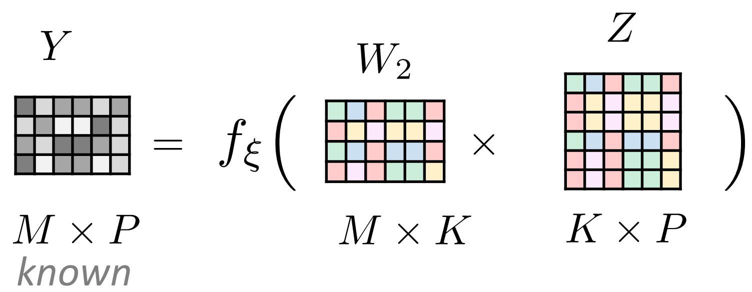
teacher *student*

$$q_{00} = \text{overlap}\{(W_1^*)_{0,.}\,;\,(W_1)_{0,.}\}$$
$$q_{01} = \text{overlap}\{(W_1^*)_{0,.}\,;\,(W_1)_{1,.}\}$$



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

Scaling of the size of the hidden layer?



≈ matrix factorization with rank K

$N \to \infty$

▹ $K = O(1)$

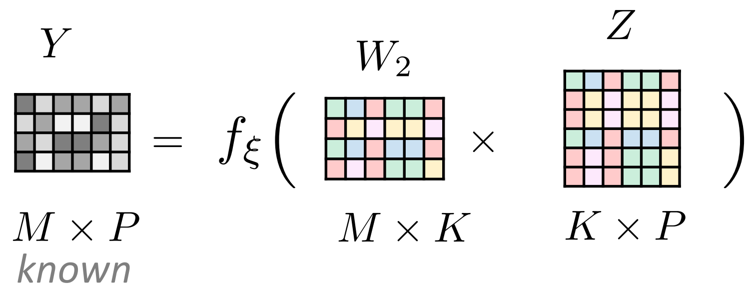
- “low-rank matrix factorization”: good mean field understanding [1, 2]
- finite number of hidden units, committee machines: great body of work! [3, 4, 5, 6, ..]

▹ $K = O(N)$

- “high-rank matrix factorization”: mean-field analysis?
- number of hidden units scaling like the inputs

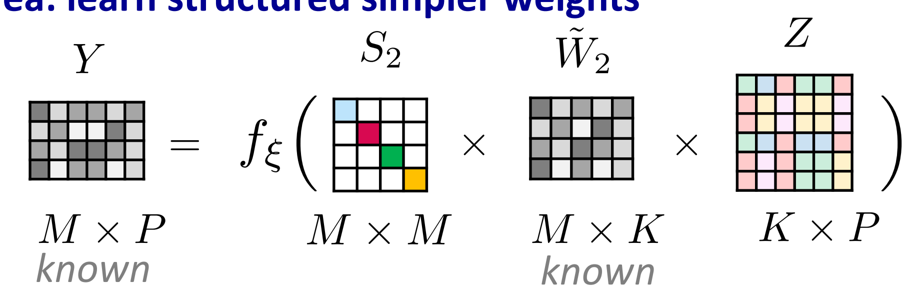
[1] Lesieur et al (2016), *MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel*
[2] Lesieur et al (2017), *Constrained Low-rank Matrix Estimation: Phase Transitions, Approximate Message Passing and Applications*
[3]Aubin et al (2018). *The committee machine: Computational to statistical gaps in learning a two-layers neural network*
[4] Monasson et ql (2004). *Learning and Generalization Theories of Large Committee-Machines*
[5] Schwarze & Hertz (1993). *Generalization in Fully Connected Committee Machines.*
[6] Schwarze (1993). *Learning a Rule in a Multilayer Neural-Network.*

Structured weights inference $K = O(N)$



 # parameters $O(N^2)$

Second idea: learn structured simpler weights

▻ 

 # parameters $O(N)$

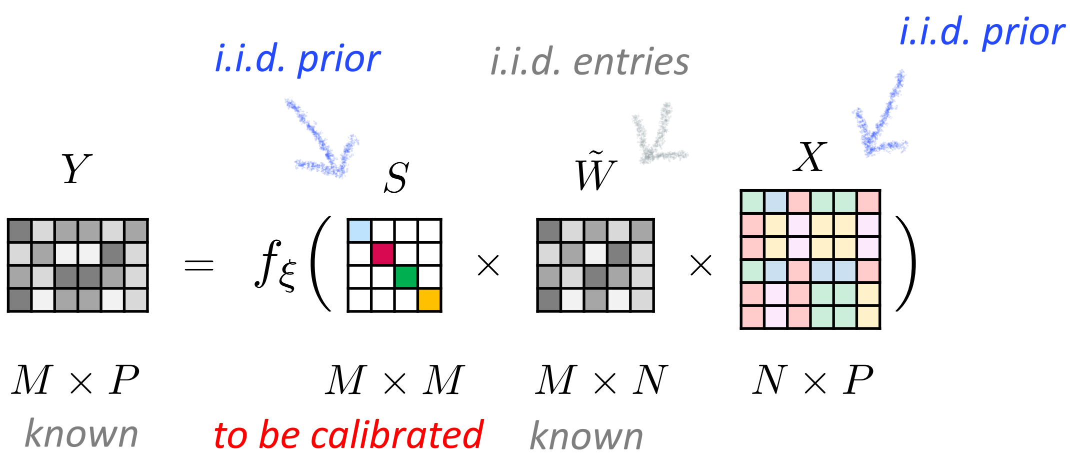
- ▻ Also used in deep learning literature:
  - Speed / memory concerns: e.g. ACDC layers [1], Ensemble learning [2]
  - Theoretical papers: e.g. Porcupine networks [3], Replica entropy [4]

- ▻ Signal processing literature: a.k.a. Blind Calibration

[1] Moczulski et al (2015), *ACDC: A Structured Efficient Linear Layer*
[2] Wen et al (2020), *BatchEnsemble: An Alternative Approach to Efficient Ensemble and Lifelong Learning*
[3] Feizi et al (2016) *Porcupine Neural Networks: (Almost) All Local Optima are Global*
[4] Gabrié et al (2018), *Entropy and mutual information in models of deep neural networks*

Blind calibration mean field analysis

Simultaneous recovery of input signal and "calibration variables"



N *input dim*

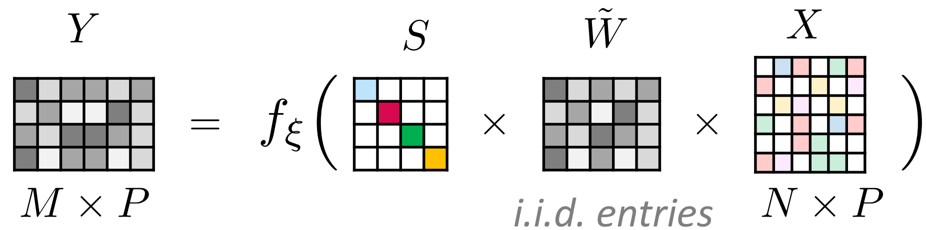
M *output dim*

P *sample size*

- ✓ Calibration - AMP algorithm [1, 2]
- ✓ Corresponding state evolution [3]
- ✓ Replica free energy [3]
- ✗ Rigorous proof

[1] Schulke C. et al (2013), *Blind Calibration in Compressed Sensing using Message Passing Algorithms*
[2] Schulke C. et al (2016), *Blind sensor calibration using approximate message passing*
[3] Gabrié M. et al (2020), *Blind calibration for compressed sensing: State evolution and an online algorithm*

Numerical results for sparse priors



i.i.d. entries

i.i.d. prior

i.i.d. ρ -sparse

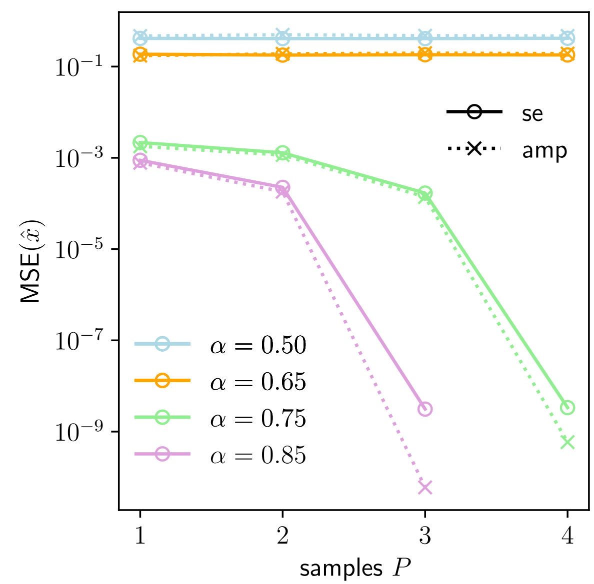
Example sparse signal recovery:

output dim / input dim $\alpha = M/N$

input sparsity ρ

naive count $\alpha_{\min} = \rho \frac{P}{P-1}$

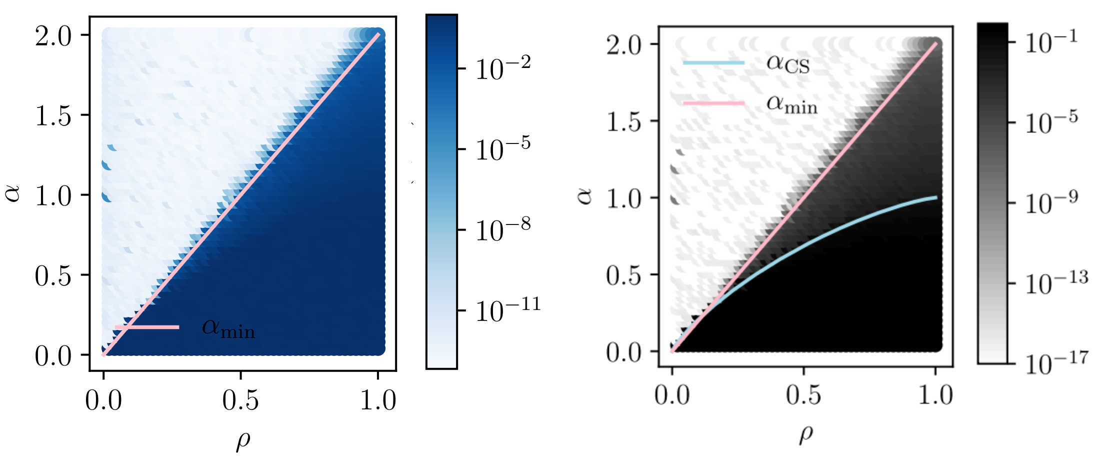
Cal- AMP State evolution



Cal- AMP reconstruction errors ($P = 2$)

S error

X error

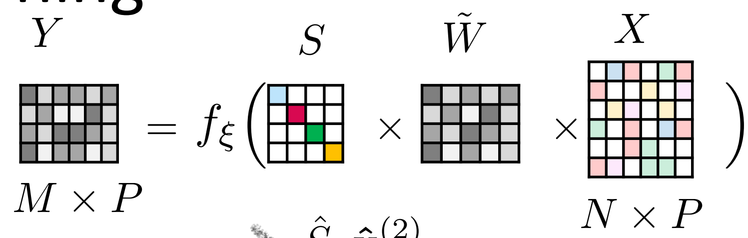


Cal-AMP reconstructs efficiently with a finite number of samples

Good agreement SE and Cal-AMP

[1] Gabrié M. et al (2020), *Blind calibration for compressed sensing: State evolution and an online algorithm*

Statistical mechanics of online learning



observation

$p_S(S)\, p_x(\mathbf{x}) \xrightarrow{\mathbf{y}^{(1)}} \hat{p}(S, \mathbf{x} \,|\, \mathbf{y}^{(1)}) \xrightarrow{\mathbf{y}^{(2)}} \hat{p}(S, \mathbf{x} \,|\, \mathbf{y}^{(2)}) \xrightarrow{\mathbf{y}^{(3)}}$ *etc.*

prior — updated prior — updated prior

approx. posterior $\hat{S}\ \hat{\mathbf{x}}^{(1)}$ — approx. posterior $\hat{S}\ \hat{\mathbf{x}}^{(2)}$

- **Streaming AMP for GLM [1], for blind calibration [2]**

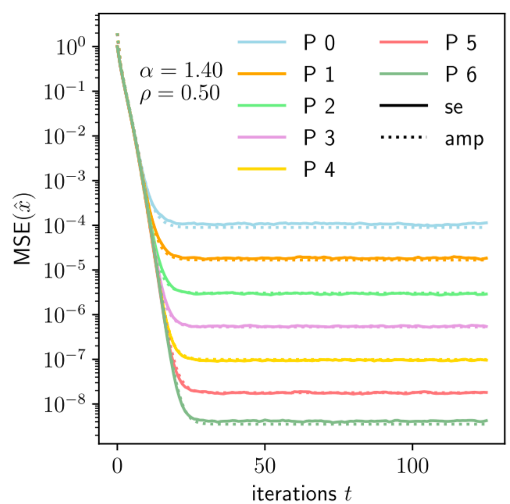
Numerical results:

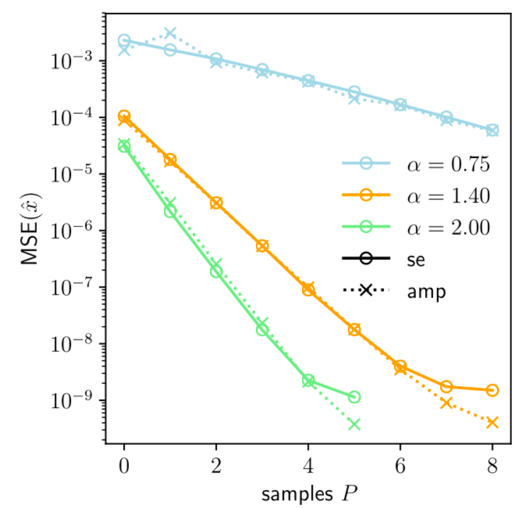
Example of sparse signal recovery

X *i.i.d.* ρ *-sparse*

$\alpha = M/N$

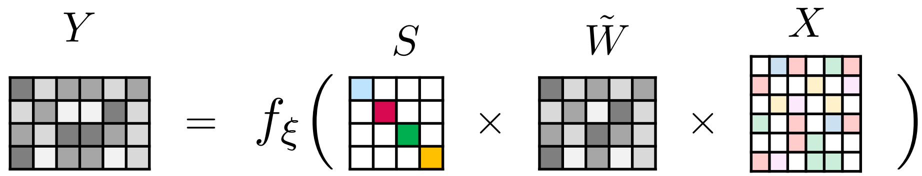
output dim / input dim





[1] Manoel et al. (2018). Streaming Bayesian inference: Theoretical limits and mini-batch approximate message-passing
[2] Gabrié M. et al (2020), *Blind calibration for compressed sensing: State evolution and an online algorithm*

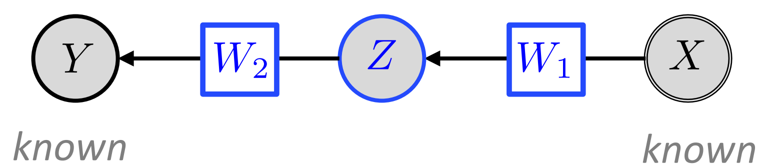
Perspectives for weight inference in deep NNs



Weight inference in hidden layers for the stat mech of deep learning (offline/batch and online/mini-batch)

▻ Perspective: Combine Cal-AMP in layers to infer structured weights in NNs (extensive number of hidden units!)

▻ Challenge: Back to the teacher-student scenario?



Perspectives for mean-field methods for inference and information/computational thresholds

- **More and more complex matrix ensembles (weights, data)**
- **Combining solutions to more complex models**
- **Great open source package for algorithms**

sphinxteam / **tramp**

Tutorial review:

Gabrié (2020), *Mean field inference methods for neural networks* – arXiv/1911.00890

Software:

Baker et al (2020), *Compositional Inference with Tree Approximate Message Passing*

Thank you!