The committee machine: Computational to statistical gaps in learning a two-layers neural network

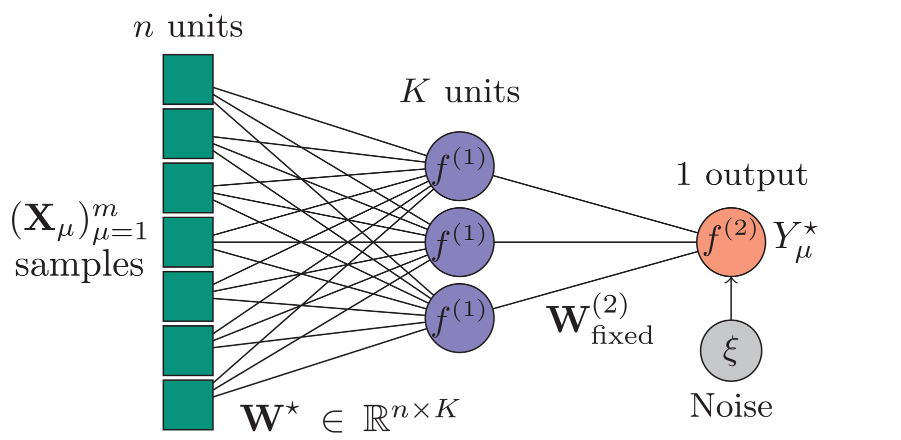
Benjamin Aubin[2], Antoine Maillard[1], Jean Barbier[1,3,4], Florent Krzakala[1], Nicolas Macris[3] & Lenka Zdeborová[2]

[1] *École Normale Supérieure (Paris)* [2] *Institut de Physique Théorique (Saclay)* [3] *École Polytechnique Fédérale de Lausanne* [4] *International Center for Theoretical Physics (Trieste)*

Motivation and outline

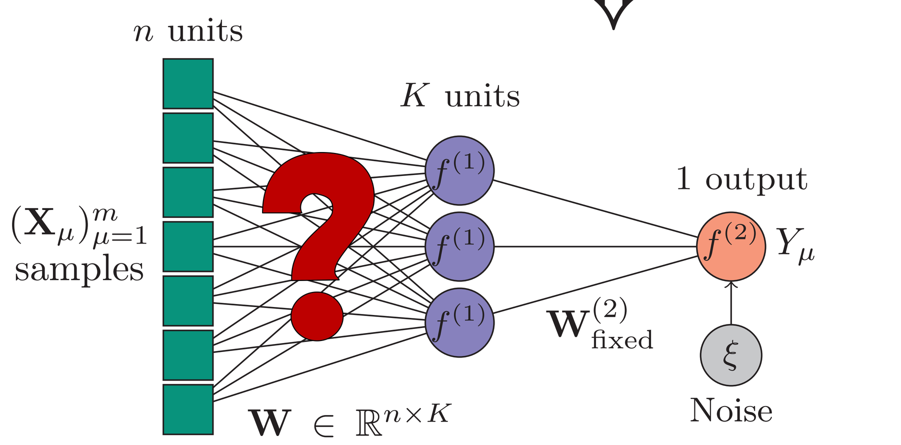
- Understand the *generalization effectiveness* of neural networks.
- **Understand the *typical case* scenario, and locate *phase transitions* in the generalization abilities: when can the network learn efficiently ?**
- Traditional bounds are *worst case* bounds, based on the value of the VC dimension and Rademacher complexity.
- $\Rightarrow$ Complementary approach: we are able to compute the *optimal* generalization error in the typical case, and to compute the corresponding solution in *polynomial* time.
- Drawbacks: we need an i.i.d.-sampled dataset, and the optimal algorithm is slower per iteration than SGD (but often converges in a few iterations only).

A general teacher-student model



TEACHER
Weights : $\mathbf{W}^\star$
Sample : $\mathbf{X}_\mu$
Output : $Y^\star_\mu$

$\Downarrow$



STUDENT
Receives $(\mathbf{X}_\mu, Y^\star_\mu)$
Learns $\mathbf{W}^\star$

- Limit $n\to\infty$ with $\lim_{n\to\infty}\frac{m}{n} = \alpha \in (0,\infty)$ and $K = \Theta_n(1)$.
- Gaussian i.i.d. samples $\mathbf{X}_\mu \sim \mathcal{N}(0, Id)$.
- For all $i \in [|1,n|]$, the weight vector $\mathbf{W}_i = \{W_{il}\}_{l=1}^K$ has a Bayesian prior P_0, with zero mean and well-defined covariance matrix ρ.
- We look at the *typical case*, by averaging over the weights of the teacher and the noise.

At a non-rigorous level, these models were previously investigated by the theoretical physics litterature e.g. [1], [2].

Notations and definitions

- $\mathcal{S}_K^+$: real positive symmetric $K\times K$ matrices.
- $\mathcal{S}_K^+(\rho) = \{q \in \mathcal{S}_K^+ \quad \text{s.t.} \quad q \in \mathcal{S}_K^+ \wedge (\rho - q) \in \mathcal{S}_K^+\}$.
- For $(\mathbf{V},\mathbf{U}) \in \mathbb{R}^K$, $\mathbf{z}_{q,\mathbf{V}}(\mathbf{U}) \equiv q^{1/2}\mathbf{V} + (\rho-q)^{1/2}\mathbf{U}$.
- For $\mathbf{x}\in\mathbb{R}^d$, $\mathcal{D}\mathbf{x} = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}||\mathbf{x}||^2}\mathrm{d}\mathbf{x}$.
- We write $Y_\mu \sim P_{\text{out}}\Big(\cdot \,\Big|\, \frac{1}{\sqrt{n}}\mathbf{X}_\mu\mathbf{W}\Big)$, accounting for $f^{(1)}$, $f^{(2)}$, and the noise.
- Two auxiliary functions for $r\in\mathcal{S}_K^+$ and $q \in \mathcal{S}_K^+(\rho)$:

$$\mathcal{I}_{P_0}(r) \equiv -\int \mathcal{D}\mathbf{Z}\,\mathrm{d}P_0(\mathbf{W}_0) \ln \int \mathrm{d}P_0(\mathbf{W}) e^{-\frac{1}{2}||r^{1/2}(\mathbf{W}_0-\mathbf{W})+\mathbf{Z}||^2}$$

$$\mathcal{I}^{(\rho)}_{\text{out}}(q) \equiv \int \mathrm{d}\hat{\mathbf{Y}}\mathcal{D}\mathbf{V} P_{\text{out}}(\hat{\mathbf{Y}}|\rho^{1/2}\mathbf{V}) \ln P_{\text{out}}(\hat{\mathbf{Y}}|\rho^{1/2}\mathbf{V}) - \int \mathrm{d}\hat{\mathbf{Y}}\mathcal{D}\mathbf{V}\mathcal{D}\mathbf{u}^\star P_{\text{out}}(\hat{\mathbf{Y}}|\mathbf{z}_{q,\mathbf{V}}(\mathbf{u}^\star)) \ln \int \mathcal{D}\mathbf{u} P_{\text{out}}(\hat{\mathbf{Y}}|\mathbf{z}_{q,\mathbf{V}}(\mathbf{u}))$$

Main theoretical result

Theorem: Replica-symmetric formula

(*i*) The normalized *mutual information* $i_n \equiv \frac{1}{n}I(\mathbf{W};\mathbf{Y}|\mathbf{X})$ converges to :

$$i_\infty = \inf_{r\in\mathcal{S}_K^+}\ \sup_{q\in\mathcal{S}_K^+(\rho)} \Big\{\mathcal{I}_{P_0}(r) + \alpha\mathcal{I}^{(\rho)}_{\text{out}}(q) - \frac{1}{2}\mathrm{Tr}[r(\rho-q)]\Big\}$$

Let us call $(r^\star, q^\star)$ the extremizers in this formula.

(*ii*) The *Bayes-optimal* generalization error

$$\epsilon_g^{(n)} \equiv \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\Big[\big(\mathbb{E}_{\mathbf{W}|\mathbf{X}}[Y(\mathbf{XW})] - Y^\star(\mathbf{XW}^\star)\big)^2\Big]$$

converges as $n\to\infty$ to a limit $\epsilon_g(q^\star)$ that only depends on $q^\star$.

$q^\star$ can be interpreted as the *overlap matrix* between the weights of the teacher and the weights of the student : "$q^\star = \frac{1}{n}\mathbf{W}^\intercal\mathbf{W}^\star$".

Sketch of proof

We use an adaptive version of Guerra's interpolation [3], developed in [4]. Extension of the techniques applied in general linear estimation [5].

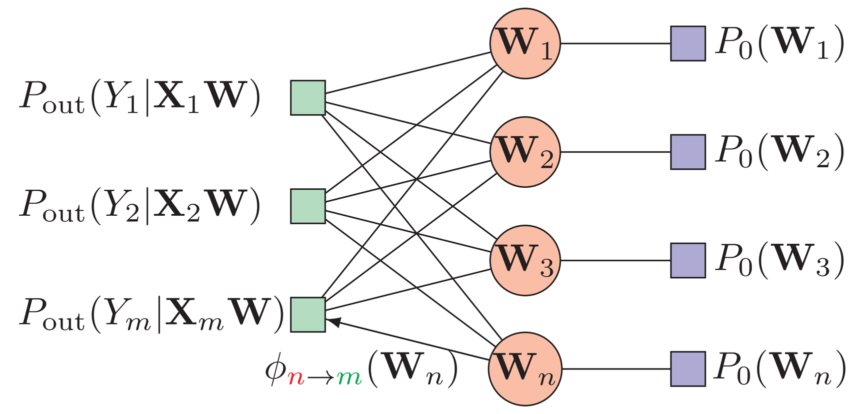
(1) Choose functions $q(t)$, $r(t)$. Two estimations problems:

$$\begin{aligned}
&(a) \quad \mathbf{Y}_0 = r(t)^{1/2}\mathbf{W}_0 + \mathbf{Z} && i_a(t=1) = \mathcal{I}_{P_0}(r(1))\\
&(b) \quad \hat{\mathbf{Y}} \sim P_{\text{out}}(\cdot|\mathbf{z}_{q(t),\mathbf{V}}(\mathbf{u}^\star)) && i_b(t=1) = \mathcal{I}^{(\rho)}_{\text{out}}(q(1))
\end{aligned}$$

(2) Interpolate between the original problem and $(a)+(b)$. The interpolated m.i. verifies $i_n(t=0) = i_n; i_n(t=1) = i_a + i_b$. So $i_n = i_a + i_b - \int_0^1 i_n'(t)\mathrm{d}t$. Choose a *smart path* $\{q(t), r(t)\}$ to conclude.

Approximate message-passing algorithm (1)

Factor graph representation of the interactions:



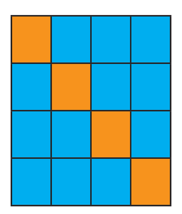
- The "message" $\phi_{i\to\mu}(\mathbf{W}_i)$ is the marginal probability of $\mathbf{W}_i$ in the absence of the node μ. One can write *belief propagation* [6] iterative equations on the set of messages $\{\phi_{i\to\mu}(\mathbf{W}_i)\}_{\mu,i}$.
- Gaussian approximation for messages $\Rightarrow$ AMP algorithm [7] [8].
- Different from SGD : **we do not optimize a cost function !**
- **AMP is conjectured (and often shown) to perform the optimal learning for inference problems of this class.**
- **Rigorous track of AMP via *state evolution* iterations:**

$$q^{t+1}_{\text{AMP}} = \rho - 2\partial_r\mathcal{I}_{P_0}(r^t_{\text{AMP}}), \quad r^{t+1}_{\text{AMP}} = -2\alpha\partial_q\mathcal{I}^{(\rho)}_{\text{out}}(q^t_{\text{AMP}}).$$

It is the variational condition of the replica symmetric-formula !

A special case: the committee machine

$$Y^\star_\mu = \mathrm{sign}\Big[\frac{1}{\sqrt{K}}\sum_{l=1}^K \mathrm{sign}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i}W^\star_{il}\Big)\Big].$$

Symmetry $\Longrightarrow q^\star =$ 

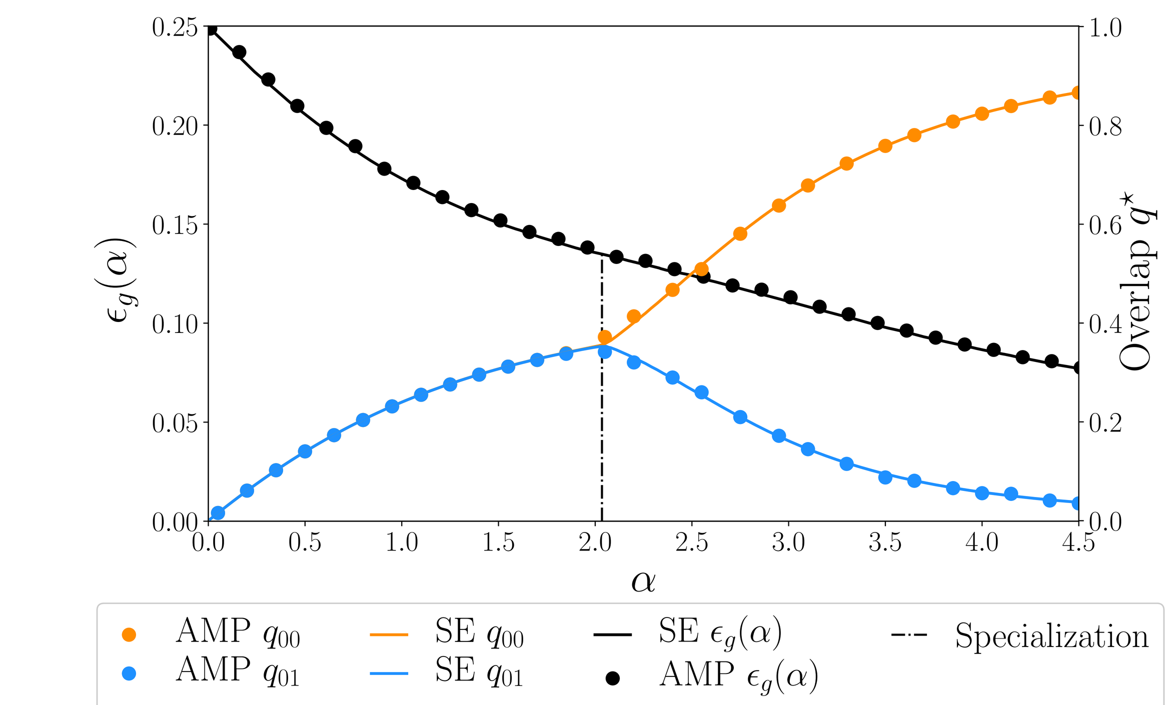
AMP iterative algorithm

- Input: $\mathbf{Y}\in\mathbb{R}^m$, $\mathbf{X}\in\mathbb{R}^{m\times n}$.
- Iteration $t\to t+1$:

$$\begin{aligned}
&\boldsymbol{\omega}^t_\mu = \textstyle\sum_{i=1}^n \Big[X_{\mu i}\hat{\mathbf{W}}^t_i - X^2_{\mu i}(\Sigma^{t-1}_i)^{-1}\hat{C}^t_i\Sigma^{t-1}_i\mathbf{g}^{t-1}_\mu\Big]\\
&V^t_\mu = \textstyle\sum_{i=1}^n X^2_{\mu i}\hat{C}^t_i\\
&\mathbf{g}^t_\mu = G(Y_\mu,\boldsymbol{\omega}^t_\mu,V^t_\mu) \quad ; \quad h^t_\mu = H(Y_\mu,\boldsymbol{\omega}^t_\mu,V^t_\mu)\\
&\mathbf{T}^t_i = \Sigma^t_i\Big(\textstyle\sum_{\mu=1}^m X_{\mu i}\mathbf{g}^t_\mu - X^2_{\mu i}h^t_\mu\hat{\mathbf{W}}^t_i\Big)\\
&\Sigma^t_i = -\Big(\textstyle\sum_{\mu=1}^m X^2_{\mu i}h^t_\mu\Big)^{-1}\\
&\hat{\mathbf{W}}^{t+1}_i = f_W(\mathbf{T}^t_i,\Sigma^t_i) \quad ; \quad \hat{C}^{t+1}_i = f_C(\mathbf{T}^t_i,\Sigma^t_i)
\end{aligned}$$

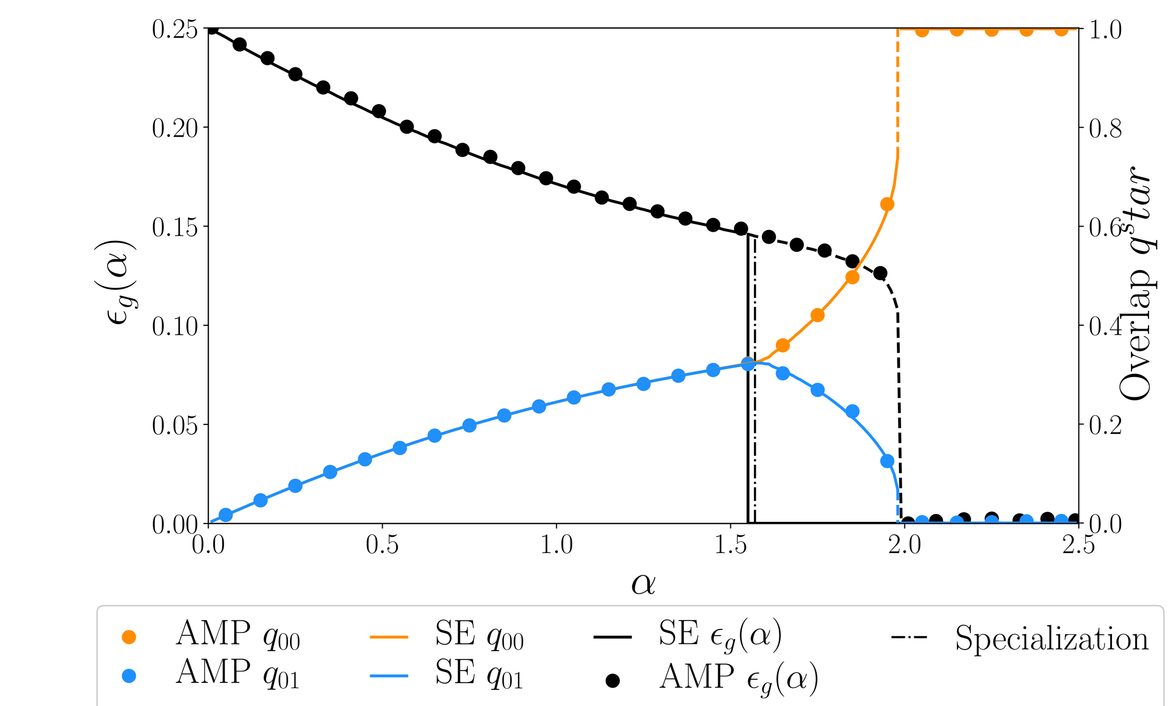
  $[G, H, f_W$ and f_C are simple functions of P_0 and P_{out}.]
- Output: $\{\hat{\mathbf{W}}_i, \hat{C}_i\}$ (mean and variance of the weights).

Small number of hidden neurons (1)



- $K=2$ with *Gaussian prior* on the weights.
- *Specialization* transition at $\alpha_c \simeq 2$: for $\alpha \le \alpha_c$, the student believes the data is linearly separable, i.e. that the teacher has only one hidden unit. At $\alpha \ge \alpha_c$, the student "learns" to differentiate the neurons of the teacher.
- Specialization is here a *second order* phase transition.
- Perfect generalization reached at $\alpha \gg 1$.

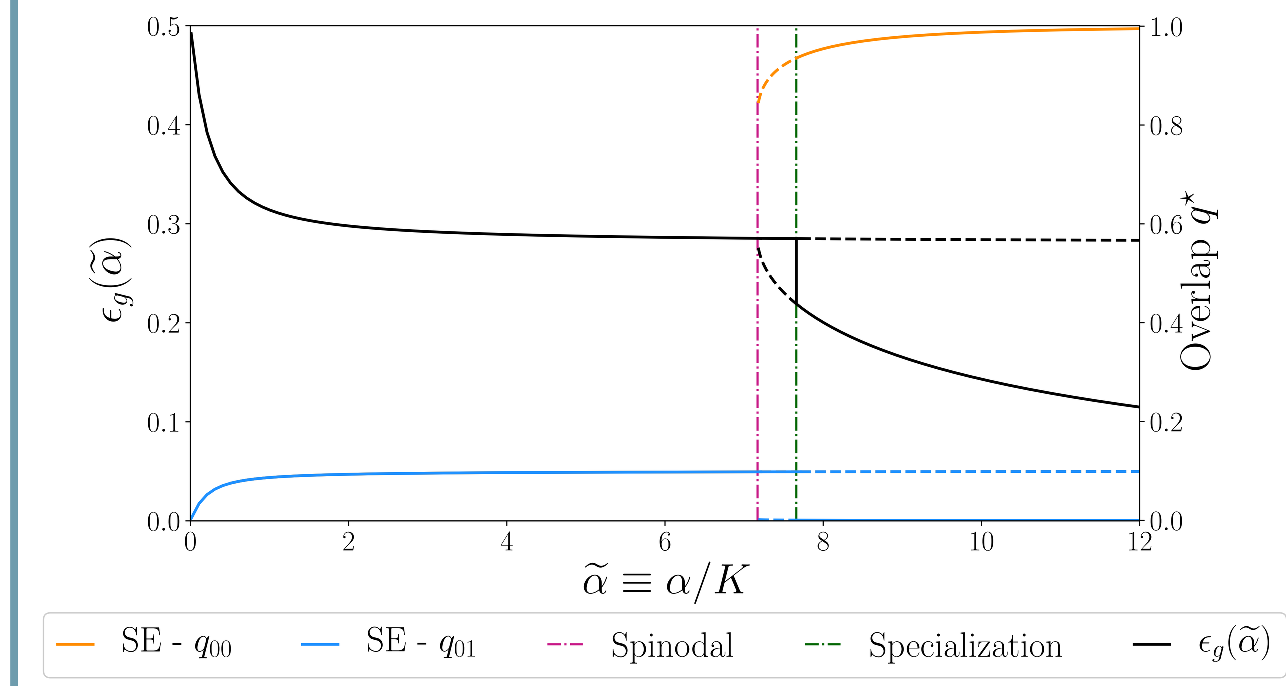
Small number of hidden neurons (2)



- $K=2$ with *Rademacher prior* on the weights.
- Existence of a first order *specialization* transition at $\alpha_c \simeq 1.6$.
- *Perfect generalization* transition: the perfect generalization solution exists for $\alpha \simeq 1.6$, but it is found by AMP for $\alpha \ge 2$.

$\Rightarrow$ Existence of a *computational gap / hard phase*. The bounds of this hard phase are approximately 1.6 and 2.

Large number of hidden neurons



- $K \gg 1$ with *Gaussian prior* on the weights.
- *First order* specialization transition.
- At all $\tilde{\alpha}$, AMP is stuck in the local minimum of the non-specialized solution with high generalization error.
- It is a stable local minima at least up to $\alpha = \Theta(K^2)$!
- *Very large computational gap / hard phase* !

Conclusion and perspectives

- Proof of the heuristic replica formula for a general model of a two-layers neural network with one hidden layer.
- Rigorous computation of the Bayes optimal error and optimal learning to unveil computational gaps.
- Evidence for a specialization phase transition at both small and large K and different priors on the weights.
- Algorithmic evidence for a large 'hard' phase with $K\gg 1$, and for small K with binary weights.
- Algorithmic evidence of a perfect generalization phase transition in large binary networks.
- Side result : these transitions do not appear in linear networks.
- (Some) remaining questions :
  - What happens for K diverging with n? Even at the (non-rigorous) replica level, this is a challenging question.
  - What if the student has a different architecture than the teacher ?
  - What if we learn the second layer ? And a deep network ?

References

[1] T. L. H. Watkin, A. Rau, and M. Biehl. Apr. 1993.
[2] H. S. Seung, H. Sompolinsky, and N. Tishby. Apr. 1992.
[3] F. Guerra. eprint: `arXiv:cond-mat/0205123`.
[4] J. Barbier and N. Macris. eprint: `arXiv:1705.02780`.
[5] J. Barbier et al. eprint: `arXiv:1708.03395`.
[6] M. Mezard and A. Montanari. *Information, physics, and computation.* Oxford University Press, 2009.
[7] D. L. Donoho, A. Maleki, and A. Montanari. eprint: `arXiv:0907.3574`.
[8] S. Rangan. eprint: `arXiv:1010.5141`.