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

Benjamin Aubin, Antoine Maillard, Jean Barbier
Nicolas Macris, Florent Krzakala & Lenka Zdeborová
< Can we efficiently learn a teacher network from a limited number of samples? >
« Can we efficiently learn a teacher network from a limited number of samples? »

Teacher:

\((X_i)_{i=1}^n\) samples

\[ W^* \in \mathbb{R}^{p \times K} \]

\( p \) features

\( K \) hidden units

output
« Can we efficiently learn a teacher network from a limited number of samples? »

Teacher:

✓ Committee machine: second layer fixed

[Schwarze’93]
Can we efficiently learn a teacher network from a limited number of samples?

**Teacher:**

- Committee machine: second layer fixed
  - [Schwarze'93]
- i.i.d samples

$$\mathbf{W}^* \in \mathbb{R}^{p \times K}$$
« Can we efficiently learn a teacher network from a limited number of samples? »

**Teacher:**

✓ Committee machine: second layer fixed

[Schwarze'93]

✓ i.i.d samples

**Student:**

![Diagram of a committee machine and a student network with fixed and variable weights.](image)
« Can we efficiently learn a teacher network from a limited number of samples? »

**Teacher:**

✓ Committee machine: second layer fixed  
  [Schwarze’93]

✓ i.i.d samples

**Student:**

!? Committee machine: second layer fixed

(W)^n_i=1 samples

(W* ∈ ℝ^p×K)

Output

(W*(2)(2) fixed

K hidden units

(Xi)^(n)i=1 samples

(f(1))

(f(2))

(f(2))

(f(1))

(Yi*)

(f(1))

(f(2))

(f(2))

(f(1))

(Yi)

(W)

(Yi*)

(W*(2) fixed

K hidden units

(W)^n_i=1 samples

(W* ∈ ℝ^p×K)
Can we efficiently learn a teacher network from a limited number of samples?

Teacher:
✓ Committee machine: second layer fixed
   [Schwarze'93]
✓ i.i.d samples

Student:
✓ Learning task possible?
« Can we efficiently learn a teacher network from a limited number of samples? »

Teacher:
- Committee machine: second layer fixed
  [Schwarze'93]
- i.i.d samples

Student:
- Learning task possible?
- Computational complexity?
Motivation

- Traditional approach
  - Worst case scenario/PAC bounds: VC-dim & Rademacher complexity
  - Numerical experiments
Motivation

➡ Traditional approach

- Worst case scenario/PAC bounds: VC-dim & Rademacher complexity
- Numerical experiments

➡ Complementary approach

✓ Revisit the statistical physics typical case scenario [Sompolinsky’92, Mezard’87]: i.i.d data coming from a probabilistic model
Motivation

➡ Traditional approach

- Worst case scenario/PAC bounds: VC-dim & Rademacher complexity
- Numerical experiments

➡ Complementary approach

✓ Revisit the statistical physics typical case scenario [Sompolinsky’92, Mezard’87]: i.i.d data coming from a probabilistic model
✓ Theoretical understanding of the generalization performance

✓ Regime: $p \to \infty$, $\frac{n}{p} = \Theta(1)$
Main result (1) - Generalization error

- Information theoretically optimal generalization error (Bayes optimal case)

\[ \epsilon_g^{(p)} \equiv \frac{1}{2} \mathbb{E}_{x, \mathbf{w}*} \left[ \left( \mathbb{E}_{\mathbf{w}|x}[Y(X\mathbf{w})] - Y^*(X\mathbf{w}^*) \right)^2 \right] \xrightarrow{p \to \infty} \epsilon_g(q^*) \]
Main result (1) - Generalization error

- Information theoretically optimal generalization error
  (Bayes optimal case)

\[ \epsilon_g^{(p)} = \frac{1}{2} \mathbb{E}_{x,w^*} \left[ (\mathbb{E}_{w|x} [Y(xw)] - Y^*(xw^*))^2 \right] \xrightarrow{p \to \infty} \epsilon_g(q^*) \]
Main result (1) - Generalization error

- Information theoretically optimal generalization error (Bayes optimal case)

\[
\epsilon_g(p) \equiv \frac{1}{2} \mathbb{E}_{X,W^*} \left[ \left( \mathbb{E}_{W|X} [Y(WX)] - Y^*(WX^*) \right)^2 \right] \xrightarrow{p \to \infty} \epsilon_g(q^*)
\]
Main result (1) - Generalization error

- Information theoretically optimal generalization error
  (Bayes optimal case)

\[
\epsilon_g^{(p)} \equiv \frac{1}{2} \mathbb{E}_{X,W^*} \left[ (\mathbb{E}_{W|X}[Y(XW)] - Y^*(XW^*))^2 \right] \xrightarrow{p \to \infty} \epsilon_g(q^*)
\]

- \(q^*\): extremizing the variational formulation of this mutual information:

\[
\lim_{p \to \infty} \frac{1}{p} I(W; Y|X) = \sup_{r \in S_K^+} \inf_{q \in S_K^+} \left\{ \psi_{P_0}(r) + \alpha \Psi_{\text{out}}(q) - \frac{1}{2} \text{Tr}(rq) \right\} + \text{cst}
\]
Main result (1) - Generalization error

- Information theoretically optimal generalization error
  (Bayes optimal case)

\[ \epsilon_g(p) \equiv \frac{1}{2} \mathbb{E}_{X, W^*} \left[ \left( \mathbb{E}_W | X [Y(XW)] - Y^*(XW^*) \right)^2 \right] \xrightarrow{p \to \infty} \epsilon_g(q^*) \]

- \( q^* \): extremizing the variational formulation of this mutual information:

\[
\lim_{p \to \infty} \frac{1}{p} I(W; Y | X) = - \sup_{r \in S_K^+} \inf_{q \in S_K^+} \left\{ \psi_{P_0}(r) + \alpha \Psi_{\text{out}}(q) - \frac{1}{2} \text{Tr}(rq) \right\} + \text{cst}
\]

Heuristic replica mutual information well known in statistical physics since 80’s
Main result (1) - Generalization error

- Information theoretically optimal generalization error
  (Bayes optimal case)

\[ \epsilon_g^{(p)} \equiv \frac{1}{2} \mathbb{E}_{X,W^*} \left[ \left( \mathbb{E}_{W|X} Y(XW) - Y^*(XW^*) \right)^2 \right] \xrightarrow{p \to \infty} \epsilon_g(q^*) \]

- \( q^* \): extremizing the variational formulation of this \textit{mutual information}:

\[ \lim_{p \to \infty} \frac{1}{p} I(W; Y|X) = - \sup_{r \in \mathcal{S}_K^+} \inf_{q \in \mathcal{S}_K^+} \left\{ \psi_{P_0}(r) + \alpha \Psi_{\text{out}}(q) - \frac{1}{2} \text{Tr}(rq) \right\} + \text{cst} \]

Heuristic replica mutual information well known in statistical physics since 80’s

✓ Main contribution: rigorous proof by adaptive (Guerra) interpolation
Main result (2) - Message Passing Algorithm
Main result (2) - Message Passing Algorithm

Traditional approach:

- Minimize a loss function. Not optimal for limited number of samples.
Main result (2) - Message Passing Algorithm

Traditional approach:
- Minimize a loss function. Not optimal for limited number of samples.

Approximate Message Passing (AMP) algorithm:
- Expansion of BP equations on a factor graph. Closed set of iterative equations. Estimates marginal probabilities \( m_j(w_j) \)

![Factor graph representation of the committee machine](image)
Main result (2) - Message Passing Algorithm

Traditional approach:
- Minimize a loss function. Not optimal for limited number of samples.

Approximate Message Passing (AMP) algorithm:
- Expansion of BP equations on a factor graph. Closed set of iterative equations. Estimates marginal probabilities $m_j(w_j)$
- Conjectured to be \textit{optimal} among polynomial algorithms
- Can be \textit{tracked rigorously} (state evolution given by critical points of the replica mutual information) [Montanari-Bayati '10]

\begin{align*}
P_{\text{out}}(Y_i | X_i W) &= w_j P_0(w_j) \\
&= m_{j \rightarrow i}(w_j)
\end{align*}

Factor graph representation of the committee machine
Gaussian weights - sign activation

Large number of hidden units $K = \Theta_p(1)$
Gaussian weights - sign activation

Large number of hidden units $K = \Theta_p(1)$

Generalization error $\epsilon_g(\tilde{\alpha})$

- Bayes optimal $\epsilon_g(\tilde{\alpha})$
- AMP $\epsilon_g(\tilde{\alpha})$
- Discontinuous specialization

$\tilde{\alpha} = \frac{\text{(# of samples)}}{\text{(#hidden units} \times \text{input size)}}$
Gaussian weights - sign activation

Large number of hidden units \( K = \Theta_p(1) \)

Generalization error \( \epsilon_g(\tilde{\alpha}) \)

- Bayes optimal \( \epsilon_g(\tilde{\alpha}) \)
- AMP \( \epsilon_g(\tilde{\alpha}) \)
- Discontinuous specialization

\( \tilde{\alpha} = (\# \text{ of samples})/(\# \text{hidden units} \times \text{input size}) \)
Gaussian weights - sign activation

Large number of hidden units $K = \Theta_p(1)$

- Bayes optimal $\epsilon_g(\tilde{\alpha})$
- AMP $\epsilon_g(\tilde{\alpha})$
- Discontinuous specialization

Generalization error $\epsilon_g(\tilde{\alpha})$

Non-specialized hidden units

$\tilde{\alpha} = (\text{# of samples})/(\text{#hidden units} \times \text{input size})$
Large number of hidden units $K = \Theta_p(1)$

Generalization error $\epsilon_g(\tilde{\alpha})$
- Bayes optimal $\epsilon_g(\tilde{\alpha})$
- AMP $\epsilon_g(\tilde{\alpha})$
- Discontinuous specialization

$\tilde{\alpha} = (\# \text{of samples})/\text{(#hidden units } \times \text{input size)}$
Large number of hidden units $K = \Theta_p(1)$

- Bayes optimal $\epsilon_g(\tilde{\alpha})$
- AMP $\epsilon_g(\tilde{\alpha})$
- Discontinuous specialization

Generalization error $\epsilon_g(\tilde{\alpha})$

Non-specialized hidden units

$\tilde{\alpha} = (\#\text{ of samples})/(\#\text{hidden units} \times \text{input size})$
Gaussian weights - sign activation

Large number of hidden units $K = \Theta_p(1)$

- Bayes optimal $\epsilon_g(\tilde{\alpha})$
- AMP $\epsilon_g(\tilde{\alpha})$
- Discontinuous specialization

Generalization error $\epsilon_g(\tilde{\alpha})$

Non-specialized hidden units

Specialized hidden units

$\tilde{\alpha} = (\# \text{ of samples})/(\#\text{hidden units} \times \text{input size})$
Large number of hidden units $K = \Theta_p(1)$

Generalization error $\epsilon_g(\tilde{\alpha})$

- Bayes optimal $\epsilon_g(\tilde{\alpha})$
- AMP $\epsilon_g(\tilde{\alpha})$
- Discontinuous specialization

Non-specialized hidden units

Specialized hidden units

$\tilde{\alpha} = (#\text{ of samples})/(\#\text{hidden units} \times \text{input size})$
Large number of hidden units $K = \Theta_p(1)$

- Bayes optimal $\epsilon_g(\tilde{\alpha})$
- AMP $\epsilon_g(\tilde{\alpha})$
- Discontinuous specialization

Generalization error $\epsilon_g(\tilde{\alpha})$

$\tilde{\alpha} = \frac{\text{(# of samples)}}{\text{(##hidden units \times input size)}}$

Non-specialized hidden units

Specialized hidden units

Computational gap

Gaussian weights - sign activation
Gaussian weights - sign activation

Large number of hidden units $K = \Theta_p(1)$

- Bayes optimal $\epsilon_g(\tilde{\alpha})$
- AMP $\epsilon_g(\tilde{\alpha})$
- Discontinuous specialization

Generalization error $\epsilon_g(\tilde{\alpha})$

$\tilde{\alpha} = (\text{# of samples})/(\text{#hidden units} \times \text{input size})$

Computational gap

Non-specialized hidden units

Specialized hidden units

TO KNOW MORE:

https://github.com/benjaminaubin/TheCommitteeMachine

Poster #111