Machine Learning from a Statistical Physics Perspective

with an appendix on

Smart Inference for Covid19 tracing

Marc Mézard
Ecole normale supérieure - PSL University

Prairie’s e-Colloquium
May 6, 2020
Machine Learning, supervised

Input  →  Machine  →  Output
Machine Learning, supervised

Input \rightarrow \text{Machine} \rightarrow \text{Output}

Input \rightarrow \text{Machine} \leftarrow \text{Output}

CAT

DOG
Deep neural networks

Can have 100,000 neurons, 100 layers, more than 1,000,000 parameters

Trained on huge databases, by simple gradient-descent type algorithms

\[ y = f(J_0 + J_1 x_1 + J_2 x_2 + J_3 x_3) \]
Do we understand how it works?

One knows everything (the dream of the neuroscientist)
One understands very little. Accumulated practical knowledge.

No big theoretical progress in the last 25 years
**Machine learning: training**

\[
\xi \rightarrow W \rightarrow y \quad y = f(W, \xi)
\]

Database = \(M\) examples of input-output \((\xi_\mu, y_\mu)\)

**Optimization**

Find \(W\) that minimizes \(\sum \mu [f(W, \xi_\mu) - y_\mu]^2\)
(or other « loss function »)

**Bayesian inference :**

\[
P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0(W) \exp \left( -\beta \sum \mu [f(W, \xi_\mu) - y_\mu]^2 \right)
\]

Unknown Data Prior Inverse temperature
Machine learning: training

Database = $M$ examples of input-output pairs $(\xi_\mu, y_\mu)$

Optimization

Find $W$ that minimizes

$\sum_{\mu} [f(W, \xi_\mu) - y_\mu]^2$

(or other « loss function »)

Bayesian inference:

$$P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0(W) \exp \left( -\beta \sum_{\mu} [f(W, \xi_\mu) - y_\mu]^2 \right)$$

Unknown  Data  Prior  Inverse temperature
Machine learning: generalization

\[ y = f(W, \xi) \]

Database = \( M \) examples of input-output

\[(\xi_\mu, y_\mu)\]
Machine learning: generalization

\[ y = f(W, \xi) \]

Database = \( M \) examples of input-output

**Generalization:** having found the best (a « typical ») set of parameters \( W^* \), compute the performance of the machine on some **new data**

\[ E_g = \sum_{\nu} [y_{\nu} - f(W^*, \xi_{\nu})]^2 \]
Machine learning: training and generalization

Learning: \( P(W \mid \{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0(W) \exp \left( -\beta \sum_\mu \left[ f(W, \xi_\mu) - y_\mu \right]^2 \right) \)

Generalization: \( E_g = \sum_\nu \left[ y_\nu - f(W^*, \xi_\nu) \right]^2 \)

Two main issues:  
- Algorithmic  
- Theoretical
Learning: \( P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0(W) \exp \left( -\beta \sum_\mu [f(W, \xi_\mu) - y_\mu]^2 \right) \)

**Learning problem:** optimization or sampling in a large dimensional space, with a disordered « energy function ». Typical of statistical physics.

**Statistical physics. Models, disorder, ensembles, replicas, message-passing equations...**
Magnets and Ising Model

\[ s_i \in \{\pm 1\} \]

\[ E = -\sum_{ij} J_{ij} s_i s_j \]

Equilibrium: \( P(s_1, \ldots, s_N) = \frac{1}{Z} e^{-E/T} \)

Ferromagnet: \( J_{ij} > 0 \)

At low T: spins align, \( P \) concentrates on 2 ordered states
Magnets and Ising Model

\[ s_i \in \{ \pm 1 \} \]

\[ E = - \sum_{ij} J_{ij} s_i s_j \]

Equilibrium: \[ P(s_1, \ldots, s_N) = \frac{1}{Z} e^{-E/T} \]

Ferromagnet: \( J_{ij} > 0 \)

Phases: \( \langle s_i \rangle = M \)

\[ M = \tanh \left( \sum_j J_{ij} s_j \right) \simeq \tanh (zJM) \]

« Mean Field » (Weiss 1907)
Magnets and Ising Model

\[ s_i \in \{\pm 1\} \]

\[ E = -\sum_{ij} J_{ij} s_i s_j \]

Equilibrium: \( P(s_1, \ldots, s_N) = \frac{1}{Z} e^{-E/T} \)

Ferromagnet: \( J_{ij} > 0 \)

Phases : \( \langle s_i \rangle = M \)

\[ M = \tanh \left( \sum_j J_{ij} s_j \right) \approx \tanh (zJ M) \]

« Mean Field » (Weiss 1907)
Random Magnets: Spin glasses disorder ensemble

\[ s_i \in \{\pm 1\} \]

\[ E_J(s) = -\sum_{ij} J_{ij} s_i s_j \]

\[ P_J(s) = \frac{1}{Z_J} e^{-\beta E_J(s)} \]

CuMn

\[ J_{ij} \sim \mathcal{N}(0, 1/N) \]

Strongly disordered system:
Random Magnets: Spin glasses disorder ensemble

\[ J_{ij} \sim \mathcal{N}(0, 1/N) \]

**Strongly disordered system:**

Spin glass sample described by the whole set of \( J_{ij} \)

\[ s_i \in \{\pm 1\} \]

\[ E_J(s) = -\sum_{ij} J_{ij} s_i s_j \]

\[ P_J(s) = \frac{1}{Z_J} e^{-\beta E_J(s)} \]

\[ J_{ij} \sim \mathcal{N} \left( \frac{J_0}{N}, \frac{1}{N} \right) \]

**Ensemble:**

drawn from a probability distribution. eg iid
Replicas, large deviations

Free energy of sample $J$: $ f_J = -\frac{1}{\beta N} \log Z_J$

Probability of finding a sample with $ f_J = f $: $ e^{N\Phi(f)}$

Almost all samples have $ f_J = f^*$

and the same thermodynamic properties

Reconstruct the large deviation function $ \Phi(f) $ and find $ f^*$

$$\mathcal{E}(Z^n_J) = \int df \ e^{N[-n\beta f + \Phi(f)]} \left[ \int df \ e^{N\Phi(f)} \right]^{-1} \simeq e^{-nN\beta f^*} $$

when $ n \to 0 $

studied in the thermodynamic limit with the Laplace method
Phase Diagram and Message Passing

\[ T = 1/\beta \]

- Par
- Fer
- SG
- F+SG

eg SK model
Phase Diagram and Message Passing

Inhomogeneous Mean Field (cavity) ➝ message passing algorithms BP, AMP, GAMP, VAMP

Built up in the last 40 years, a very fruitful connexion...

\[ h_{i\setminus j} = \frac{1}{\beta} \text{atanh}[\tanh(\beta J_{ki}) \tanh(\beta h_{k\setminus i})] + \frac{1}{\beta} \text{atanh}[\tanh(\beta J_{\ell i}) \tanh(\beta h_{\ell\setminus i})] \]

SG phase = many solutions
1- Glass « phase » : Many pure states, unrelated by symmetry, organized in a hierarchical « ultrametric » structure

2- Many metastable states, unrelated by symmetry

3- « True » ground state : fragile to perturbation!
Machine learning: training

Learning: \( P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0(W) \exp \left(-\beta \sum_\mu [f(W, \xi_\mu) - y_\mu]^2\right) \)

Variables = \( W \).

Disorder in the sample = data base \( \xi_\mu, y_\mu \)

Zero energy = perfect performance of the machine on the training set

Landscape? How does it depend on the problem, the database, the architecture?

Stochastic gradient descent often find a solution. Depends on details, but good…
Machine learning: generalization

Learning: \[ P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0(W) \exp \left( -\beta \sum_\mu [f(W, \xi_\mu) - y_\mu]^2 \right) \]

Generalization: \[ E_g = \sum_\nu [y_\nu - f(W^*, \xi_\nu)]^2 \]

Deep networks work in an « over-parametrized » regime
Makes learning easier
Should degrade generalization: why training with so many parameters does not lead to overfitting?
Machine learning: training and generalization

Usual behavior in statistics

Number of parameters

Generalization

Training
Machine learning: training and generalization

Usual behavior in statistics

With 4 parameters I can fit an elephant (J. von Neumann)
Machine learning: training and generalization

Deep networks

Number of Data vs. Number of parameters
- Generalization
- Training

Graph showing the relationship between the number of data and the number of parameters, illustrating the trade-off between training and generalization.
Machine learning Theory

Some important general statements.

Two layers limited to linearly separable problems
More than two: universal computer

Complexity results. Bounds on the difference between training energy and generalization energy, for different classes of functions

Simple settings. Convex optimization. Linear problems

\[ y = f(W.\xi) \]
Machine learning Theory

What statistical physics can bring:

Tools and concepts for empirical analysis (landscape, learning dynamics)

Precise statements for the asymptotic regime (thermodynamic limit). Phase transitions, learning dynamics.

Requires an ensemble to model the data
Learning:

\[ P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0(W) \exp \left( -\beta \sum_{\mu} [f(W, \xi_\mu) - y_\mu]^2 \right) \]

**Algorithmic studies** typically uses one (or several) databases for \{\xi_\mu, y_\mu\} : data = quenched disorder

\[ \xi_\mu = \text{Examples from the 80's: iid patterns' entries} \]

\[ y_\mu = \text{CAT} \]

**Theoretical analysis** relies on a probabilistic « model of the world ».

Independent patterns drawn from \( P_\xi(\xi) P_y(y|\xi) \)

Quenched disorder

Examples from the 80’s: iid patterns’ entries \( \xi_{i\mu} \sim \mathcal{N}(0, 1) \)
Example from the 90’s:
« teacher-student perceptron »

Teacher: generates parameters $w^*$ from teacher prior $w_i^* = \pm 1$
generates data $\xi_i^\mu = \pm 1$ and target $y^\mu = \text{Sign}(\sum_i w_i^* \xi_i^\mu)$

Student: same architecture, machine-learning for finding its weights $w_i$

\[
y = f(W.\xi)
\]
Example from the 90’s:
« teacher-student perceptron »

**Teacher**: generates parameters $w^*$ from teacher prior $w^*_i = \pm 1$
generates data $\xi_i^\mu = \pm 1$ and target $y^\mu = \text{Sign}(\sum_i w^*_i \xi_i^\mu)$

**Student**: same architecture, machine-learning for finding its weights $w_i$
Example from the 90’s: « teacher-student perceptron »

\[ E_\xi(W) = \sum_{\mu=1}^{P} [\text{Sign}(W . \xi_\mu) - \text{Sign}(W^* . \xi_\mu)]^2 \]

Binary weights. Discrete optimization

Replica analysis= properties of the Gibbs measure

\[ P(W) = \frac{1}{Z} e^{-\beta E_\xi(W)} \]

Gardner 88, Gardner Derrida 89, Gyorgyi 90
Barbier et al. 2019
Example from the 90’s: « teacher-student perceptron »

\[ E_\xi(W) = \sum_{\mu=1}^{P} \left[ \text{Sign}(W.\xi_\mu) - \text{Sign}(W^*.\xi_\mu) \right]^2 \]

Binary weights. Discrete optimization

**Replica analysis** = properties of the Gibbs measure

Gardner 88, Gardner Derrida 89, Gyorgyi 90
Barbier et al. 2019

\[ P(W) = \frac{1}{Z} e^{-\beta E_\xi(W)} \]

Generalization error depends on the typical angle between \( W \)
and \( W^* \)

Phase diagram in the thermodynamic limit

\[
\begin{align*}
N &\to \infty \\
P &\to \infty \quad \alpha = P/N
\end{align*}
\]

Example. Binary perceptron \( w_i \in \{\pm 1\} \)

**Algorithm: TAP-AMP equations**

MM 89, Rangan 2011, Krzakala et al. 2012
Teacher-student binary perceptron

Algorithm based on mean field equations

- Theory
- Gyorgyi 1990

Perfect learning phase transition

Gyorgyi 1990
With discrete weights: Phase transition to perfect generalization when the size of the database reaches the threshold $\alpha_c = 1.245$. Fast message passing algorithm for $\alpha > 1.49$

Nice results, but of little use for understanding realistic networks. Decoupling between theoretical results and practical engineering applications…

Conjectured 30 years ago. Proof: Barbier et al. 2019

With continuous weights $w_i \in \mathbb{R}$:
Generalization error of maximally stable $W$ decreases like $0.5005/\alpha$

Conjectured 30 years ago. Proof: Barbier et al. 2019
Why does it work?

Architecture
Algorithms
Data structure
Why does it work?

Architecture
Algorithms
Data structure

Data structure
- Hidden manifolds and sub manifolds
- Combinatorial structure
- Euclidean correlations

- Analyse data
- Build generative models that can be analyzed fully in some large size limit
- Understand mechanisms
The hidden manifold of data

MNIST

Input space: dimension $28^2 = 784$
The hidden manifold of data

Input space: dimension $28^2 = 784$

Manifold of handwritten digits in MNIST:

Nearest neighbors’ distance: $R_{nn} \sim p^{-1/d}$

$p \sim cR^d$

Grassberger Procaccia 83, Costa Hero 05, Heinz Audibert 05, Ansuini et al. 19, Spigler et al. 19…
The hidden manifold of data

Input space: dimension \(28^2 = 784\)

Manifold of handwritten digits in MNIST:

Nearest neighbors’ distance: \(R_{nn} \sim p^{-1/d}\)

\(p \sim cR^d\)

Grassberger Procaccia 83, Costa Hero 05, Heinz Audibert 05, Ansuini et al. 19, Spigler et al. 19…
The hidden manifold of data

MNIST: \( d = 784 \)

\[ d_{\text{eff}} \approx 15 \]

Spigler et al. 19

Nearest neighbors’ distance:

\[ R_{\text{nn}} \approx p^{-1/d} \]
The hidden manifold of data

MNIST: \( d = 784 \)
\( d_{\text{eff}} \approx 15 \)

Nearest neighbors’ distance:
\[ R_{nn} \approx p^{-1/d} \]

The neural net should answer: this image does not seem to be a handwritten digit

Spigler et al. 19
Structure of the task: perceptual sub-manifolds

MNIST problem: in the 15-dim manifold of handwritten digits, identify the 10 perceptual sub manifolds associated with each digit, of dimensions between 7 and 13...

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7877</td>
<td>6990</td>
<td>7141</td>
<td>6824</td>
<td>6903</td>
</tr>
<tr>
<td>2</td>
<td>8/7/7</td>
<td>13/12/13</td>
<td>14/13/13</td>
<td>13/12/12</td>
<td>12/12/12</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>6876</td>
<td>7293</td>
<td>6825</td>
<td>6958</td>
<td>6903</td>
</tr>
<tr>
<td>5</td>
<td>11/11/11</td>
<td>10/10/10</td>
<td>14/13/13</td>
<td>12/11/11</td>
<td>12/11/11</td>
</tr>
</tbody>
</table>
Structure of the task: perceptual sub-manifolds

$555$

$d_{\text{eff}}(5) \approx 12$

Hein Audibert 05

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7877</td>
<td>6990</td>
<td>7141</td>
<td>6824</td>
<td>6903</td>
</tr>
<tr>
<td>2</td>
<td>8/7/7</td>
<td>13/12/13</td>
<td>14/13/13</td>
<td>13/12/12</td>
<td>12/12/12</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>6876</td>
<td>7293</td>
<td>6825</td>
<td>6958</td>
<td>6903</td>
</tr>
<tr>
<td>5</td>
<td>11/11/11</td>
<td>10/10/10</td>
<td>14/13/13</td>
<td>12/11/11</td>
<td>12/11/11</td>
</tr>
</tbody>
</table>

MNIST problem: in the 15-dim manifold of handwritten digits, identify the 10 perceptual sub manifolds associated with each digit, of dimensions between 7 and 13…

… from an input in 784 dimensions!
An ensemble for the hidden manifold

Pattern $\mu$: $X_{\mu i} = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_{\mu r} \vec{F}_r \right]$  

Data = input patterns built from $R$ features $\vec{F}_r$.

A feature is a $N$ component vector in the input space.

Each pattern is built from a weighted superposition of features (feature $r$ has weight $C_r$):

$$\sum_{r=1}^{R} C_r \vec{F}_r$$

Then apply a nonlinear folding function $f$ to each component.

S. Goldt, F. Krzakala, MM L. Zdeborova

arXiv:1909.11500
An ensemble for the hidden manifold

\[ X_{\mu i} = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_{\mu r} F_{ir} \right] \]

The \( R \)-dimensional data manifold is folded by applying the non-linear function \( f \)
An ensemble for the task

\[ \tilde{X}^\mu = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r^\mu \tilde{F}_r \right] \]

« Latent representation »: \( \{C_r\} \)

Desired output = function of latent representation

Example:
\[ y^\mu = g \left( \sum_{r=1}^{R} \tilde{w}_r C_r^\mu \right) \]

(perceptron in hidden manifold)
An ensemble for the task

\[
\tilde{X}^{\mu} = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r \tilde{F}_r \right]
\]

« Latent representation »: \{C_r\}

Desired output = function of latent representation

Examples:

\[
y^{\mu} = g \left( \sum_{r=1}^{R} \tilde{w}_r C_r^{\mu} \right)
\]

(perceptron in latent space)

\[
y^{\mu} = \sum_{m=1}^{M} \tilde{v}_m \ g \left( \sum_{r=1}^{R} \tilde{w}_{mr} C_r^{\mu} \right)
\]

(2 layers nn in latent space)
Learning from HMM data

Learn using a 2-layer neural net, $K$ hidden units

$$\phi(\vec{X}) = \sum_{k}^{K} v_{k} g \left( \vec{w}_{k} \cdot \vec{X} / \sqrt{N} \right),$$

Training error

$$\frac{1}{2P} \sum_{\mu=1}^{P} \left[ \Phi(\vec{X}^{\mu}) - y^{\mu} \right]^{2}$$

Target label: generated from latent representation $C_{r}^{\mu}$
Learning from HMM data

Learn using a 2-layer neural net, $K$ hidden units

$$\phi(\tilde{X}) = \sum_{k} v_k g \left( \overrightarrow{w_k} \cdot \tilde{X} / \sqrt{N} \right),$$

Training error

$$\frac{1}{2P} \sum_{\mu=1}^{P} \left[ \Phi(\tilde{X}^\mu) - y^\mu \right]^2$$

or other loss function

Target label: generated from latent representation $C^\mu_r$
Learning from HMM data

Learn using a 2-layer neural net, $K$ hidden units

$$\phi(\vec{X}) = \sum_{k} v_k g \left( \vec{w}_k \cdot \vec{X} / \sqrt{N} \right),$$

Training error

$$\frac{1}{2P} \sum_{\mu=1}^{P} \left[ \Phi(\vec{X}^\mu) - y^\mu \right]^2$$

or other loss function

Target label: generated from latent representation $C_r^\mu$

Generalization error: same with $P^*\text{new patterns}$
NB: Hidden manifold and random features

$$\tilde{X}^{\mu} = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r^{\mu} \tilde{F}_r \right]$$

Correlated components \quad iid

Learning a task with a iid database in R dimensions $C_r^{\mu}$

In general not linearly separable. Embed it into a larger dimensional space of features $C_r^{\mu} \rightarrow \tilde{X}_i^{\mu}$

Embedding through quenched (or « lazily learnt ») matrix $F$, and nonlinearity $f$. Special case of HMM

Connexion to Montanari Mei arXiv:1908.05335 and 1911.01544
NB: Hidden manifold and random features

\[ \tilde{X}^\mu = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r^\mu \tilde{F}_r \right] \]

Correlated components \quad iid

In general not linearly separable. Embed it into a larger dimensional space of features

\[ C_r^\mu \rightarrow X_i^\mu \]

Embedding through quenched (or « lazily learnt ») matrix \( F \), and nonlinearity \( f \). Special case of HMM

Connexion to Montanari Mei arXiv:1908.05335 and 1911.01544
Analytic study of the hidden manifold model

\[ \tilde{X} = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r \tilde{F}_r \right] \]

Correlated components iid

Solvable limit = \textbf{thermodynamic limit} with extensive latent dimension \( N \to \infty, R \to \infty, P \to \infty \)

With fixed \( R/N = \gamma, \ P/N = \alpha, \ K \)
Analytic study of the hidden manifold model

\[ \tilde{X} = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r \tilde{F}_r \right] \]

Correlated components \quad iid

balanced:

\[ F_{ri} = O(1) \]

\[ \frac{1}{N} \sum_i F_{ri} F_{si} = O(1/\sqrt{N}) \]

\[ \frac{1}{N} \sum_i F_{ri} F_{ri} = 1 \]
Analytic study of the hidden manifold model

\[ \tilde{X} = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r \tilde{F}_r \right] \]

Correlated components

\[ X_i = f[u_i] \]

\[ u_i = \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r F_{ri} \]

Gaussian, weakly correlated \( O(1/\sqrt{N}) \) when \( F_{ri} \) are balanced and \( O(1) \)

\[ \mathbb{E} (f[u_i] f[u_j]) = \langle f(u) \rangle^2 + \langle uf(u) \rangle^2 \mathbb{E} (u_i u_j) \]

\( u \) Gaussian \( \mathcal{N}(0, 1) \)
Gaussian Equivalence Theorem (GET)

Inputs of hidden units:
\[ u_i = \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r F_{ri} \]

\[ X_i = f[u_i] \quad \text{iid} \]

\[ \lambda^k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^k f[u_i] \]

**GET**: In the thermodynamic limit, the variables \( \lambda^k \) have a Gaussian distribution, with covariance

\[ \mathbb{E}[\tilde{\lambda}^k \tilde{\lambda}^\ell] = (c - a^2 - b^2) W^{k\ell} + b^2 \Sigma^{k\ell} \]

\[ W^{k\ell} = \frac{1}{N} \sum_{i=1}^{N} w_i^k w_i^\ell \quad \Sigma^{k\ell} = \frac{1}{R} \sum_{r=1}^{R} S_r^k S_r^\ell \quad S_r^k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^k F_{ir} \]

\[ c = \langle f(u)^2 \rangle \quad a = \langle f(u) \rangle \quad b = \langle uf(u) \rangle \quad u \text{ Gaussian } \mathcal{N}(0, 1) \]
Learn using a 2-layer neural net, \( K \) hidden units

\[
\Phi(\vec{X}) = \sum_{k=1}^{K} g\left(\vec{w}^k \cdot \vec{X} / \sqrt{N}\right)
\]

\( \vec{X} = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r \vec{F}_r \right] \)

\( \vec{X} \) = inside hidden R-dimensional manifold, folded by function \( f \)

Desired output given constructed from latent representation

\[
\Phi_t(\vec{X}) = \sum_{m=1}^{M} \tilde{g} \left( \sum_{r=1}^{R} \tilde{w}_r^m C_r \right)
\]
Online learning: ODE for SGD

Evolution of the weights during learning

\[(w^k_i)^{\mu+1} - (w^k_i)^{\mu} = -\eta \Delta g'(\lambda^k) f(u_i)\]

\[
\Delta = \sum_{\ell=1}^{K} g(\lambda^\ell) - \sum_{m=1}^{N} \tilde{g}(\nu^m)
\]

New pattern (and therefore new latent representation \(C_r\)) at each time

GET: \(\lambda^k\) and \(\nu^m\) are Gaussian, and the learning dynamics can be analyzed by ordinary differential equations for order parameters like

\[
W^{k\ell} = \frac{1}{N} \sum_{i=1}^{N} w^k_i w^\ell_i
\]

D Saad and S Solla 95, Biehl and Schwarze 95, …
Order parameters

\[ W^{k\ell} = \frac{1}{N} \sum_{i} w_i^k w_i^\ell \]
\[ \Sigma^{k\ell} = \frac{1}{R} \sum_{r=1}^{R} S_r^k S_r^\ell \]
\[ R^{km} = \frac{1}{R} \sum_{r=1}^{R} S_r^k \tilde{w}_r^m \]

where

\[ S_r^k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^k F_{ir} \]

**NB:** mixed order parameter \( R^{km} \) measures the correlation of pre-activation of neuron \( k \) in the student and the weight \( m \) in the latent task. First project student’s weight to latent space (\( S_r^k \)), then measure overlap to teacher.
ODE Theory vs simulations $N=10000$, $D=100$, $M=2$, $K=2$

specializes after $50N$ steps
Larger second layer allows better learning after specialization
Larger second layer allows better learning: experiments on databases (erf)
« Full batch »: perceptron learning and generalized linear regression

\[ y_w (\vec{X}) = g (\vec{w} \cdot \vec{X}) \]
« Full batch »: perceptron learning and generalized linear regression

\[ \bar{X}^\mu = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r^\mu \vec{F}_r \right] \]

Learn from database of \( P \) patterns lying in a hidden manifold

Target output = latent task

\[ y^\mu = \Phi_t(\bar{X}^\mu) = \tilde{g} \left( \sum_{r=1}^{R} \tilde{w}_r \cdot C_r^\mu \right) \]

- Classification \( \tilde{g} (z) = \text{Sign}(z) \)
- Regression \( \tilde{g} (z) = z \)
« Full batch »: perceptron learning and generalized linear regression

\[
\tilde{X}^\mu = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r^\mu F_r \right]
\]

Learn from database of \( P \) patterns lying in a hidden manifold

Target output = latent task

\[
y^\mu = \Phi_t (\tilde{X}^\mu) = \tilde{g} \left( \sum_{r=1}^{R} \tilde{w}_r C_r^\mu \right)
\]

- Classification \( \tilde{g} (z) = \text{Sign}(z) \)

- Regression \( \tilde{g} (z) = z \)

Learning = minimize « loss »

\[
E = \sum_{\mu=1}^{P} \epsilon \left( y^\mu , g(\tilde{w}.\tilde{X}^\mu) \right) + \frac{\lambda}{2} \tilde{w}^2
\]

F. Gerace, B. Loureiro, F. Krzakala, MM, L. Zdeborova
In short

Gardner’s computation: typical volume of weight space compatible with the data \( \{ \tilde{X}_\mu, \Phi_t(\tilde{x}_\mu) \} \). Evaluated with replicas.

The volume can be written in terms of the local input fields to the hidden variables, \( \lambda_\mu^a \).

GET: these are Gaussian variables, independent for different patterns, correlated for one given pattern.
In short

Gardner’s computation: typical volume of weight space compatible with the data $\left\{ \vec{X}_\mu, \Phi_t(\vec{x}_\mu) \right\}$. Evaluated with replicas

The volume can be written in terms of the local input fields to the hidden variables, $\lambda^a_\mu$.

GET: these are Gaussian variables, independent for different patterns, correlated for one given pattern.

Theory vs simulations. Classification, logistic loss, « sign » non-linearities, $R = 200 \quad P = 600 \quad \lambda = 10^{-3}$
"Double descent"

Opper and Kinzel 1995
Spigler et al. 2019
Belkin et al. 2019

\[ R/N = 1/3 \]
\[ \lambda = 10^{-4} \]

Classification task:

\[ \hat{y}^\mu = \text{Sign} \left( \sum_r \tilde{w}_r C_r^\mu \right) \]

Square loss: minimize

\[ \sum_{\mu} (\hat{y}^\mu - \vec{w} \cdot \vec{X}^\mu)^2 \]

zero for \( P < N \)

« capacity » \( \alpha^* = P/N = 1 \)

Logistic loss: « capacity » \( \alpha^* = P/N > 2 \)
NB Phase diagram for learning: Threshold of linear separability

$P/R$

$N/P = 1/\alpha^*$

NB: at large $R$, the data matrix entries are close to iid. Cover’s result $\alpha^* = 2$
Statistical physics for machine learning: Requires better ensembles for data
- Data in submanifolds
- Combinatorial structure

**Hidden Manifold Model**

**Data has « Latent representation »:** \( \{C_r\} \)

**Desired output (task) = function of latent representation**

**Example**

\[
y = g \left( \sum_{r=1}^{R} \tilde{w}_r C_r \right) \quad \tilde{X} = f \left[ \frac{1}{\sqrt{R}} \sum_{r=1}^{R} C_r \tilde{F}_r \right]
\]

- Good learning and generalization phenomenology
- Can be studied analytically: online learning and full batch in the limit where \( R = O(N) \), thanks to a Gaussian Equivalence Theorem
Smart Inference for Covid19 tracing using message passing


COLLABORATING WITH: A. BRAUNSTEIN, LUCA DALL ASTA, ALESSANDRO INGROSSO, INDACO BIAZZO, ANNA PAOLA MUNTONI, (TORINO)

DISCUSSIONS WITH: YOSHUA BENGIO, IRINA RISH, LUCA FERRETTI, IVAN BESTVINA, (MILA)

Information about individuals (age, symptoms,…), known by each individual

Information about contacts (time, duration), known by the two individual in contact

Pb: Infer the probability that each individual be infected

Message-passing provides an efficient solution, without need for a central system, and based on simple exchange of messages that can be encrypted
Smart Inference of People At Risk (SIPAR)

Risk can be estimated more accurately than the list of contacts. Every individual should account for increased risk of his recent contacts and spread the information to other contacts.

Basic tool: Susceptible-Infected-Recovered (SIR) model for individuals

- **Susceptible individuals (S)**: Can be infected
- **Infected individuals (I)**: Can infect others
- **Removed individuals (R)**: Cannot spread or be infected

**Parameters:**
- $\lambda_{ij}(t)$: Infection rate from $j$ to $i$ at time $t$ (contact)
- $\mu_i$: Recovery rate

**Estimate probabilities:**
- $P_S^j(t)$
- $P_I^j(t)$
- $P_R^j(t)$
Smart Inference of People At Risk (SIPAR)

Mean-field equations:

\[
P_i^s(t+1) = P_i^s(t) \left( 1 - \sum_{j \in \partial_i(t)} P_j^i(t) \lambda_{ij}(t) \right)
\]

\[
P_i^r(t+1) = P_i^r(t) + \mu_i P_i^i(t)
\]

\[
P_i^i(t+1) = P_i^i(t) + P_i^s(t) \sum_{j \in \partial_i(t)} P_j^i(t) \lambda_{ij}(t) - \mu_i P_i^i(t)
\]

Can incorporate the knowledge from contacts’ infections (backtrack)

More elaborate = dynamic message passing (cavity equations)

Lokhov, MM, Ohta, Zdeborova PRE 2014, PRE 2015
Comparing tracing and SIPAR: quarantine all symptomatic tested positive and test more according to tracing, or smart inference ranking.

Random geometric contact graph in 2D, scale 1.1, daily on average 7.4 contacts.
Population size= 10 000, \( \lambda=0.02 \), \( \mu=0.03 \). Initially 20 infected \( \tau=5 \), \( \delta=15 \).
Tests 7-21

Contact graph provided by Ferretti and Hinch.
Daily on average 12.7 contacts.
Population size= 50 000, \( \lambda=0.01 \), \( \mu=0.03 \). Initially 10 infected \( \tau=5 \), \( \delta=15 \).
Tests 50-100
The End
Thanks to Florent Krzakala and Lenka Zdeborova, as well as Sebastian Goldt, Federica Gerace, Bruno Loureiro, Galen Reeves, Antoine Baker, Maria Refinetti, Stefano Sarao…

MM, Phys.Rev. E 95 (2017),022117

