WHAT HAS STATISTICAL PHYSICS TO SAY ABOUT MACHINE LEARNING?

Lenka Zdeborová
(IPhT, CEA Saclay, France)

PAISS, 3.-5. 10. 2019, Paris
ML mostly developed by **engineering design process**: Define an objective (e.g. to reach the best accuracy on ImageNet). Create a tool that reaches the objective.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Method</th>
<th>Top 1 Accuracy</th>
<th>Top 5 Accuracy</th>
<th>Number of params</th>
<th>Extra Training Data</th>
<th>Paper Title</th>
<th>Year</th>
<th>Paper</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FixResNeXt-101 32x48d</td>
<td>86.4%</td>
<td>98.0%</td>
<td>829M</td>
<td>✓</td>
<td>Fixing the train-test resolution discrepancy</td>
<td>2019</td>
<td>📄</td>
<td>⚪</td>
</tr>
</tbody>
</table>

➡ **Deep learning is a revolutionary engineering progress.**

**Science** aims to understand behaviour of existing world. Do we understand why FixResNeXt-101 works?

➡ **Science/understanding of deep learning is in its infancy.**
Do we need understanding? Isn’t engineering enough, simply because “it works”?

Some open questions:

- Why don’t heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn’t backpropagation head for a poor local minima?


Still not answered!
TOWARDS THEORY OF DEEP LEARNING?

Inter-play of three ingredients

- Information theory
- Statistics
- Signal processing
- Data structure
- Algorithm
- Architecture
- Approximation theory
- Learning theory
- Computer science
- Optimisation theory

See also: E. Mossel, Deep Learning Boot Camp in Simons Institute, Berkeley (June 2019).
LONG-LASTING FRIENDSHIP BETWEEN MACHINE LEARNING AND STATISTICAL PHYSICS
Yann LeCun is with Levent Sagun and 3 others. August 30

Stéphane Mallat's tutorial at the "Statistical Physics and Machine Learning back together" summer school in Cargese, Corsica.

There is a long history of theoretical physicists (particularly condensed matter physicists) bringing ideas and mathematical methods to machine learning, neural networks, probabilistic inference, SAT problems, etc.

In fact, the wave of interest in neural networks in the 1980s and early 1990s was in part caused by the connection between spin glasses and recurrent nets popularized by John Hopfield. While this caused some physicists to morph into neuroscientists and machine learners, most of them left the field when interest in neural networks wanted in the late 1990s.

With the prevalence of deep learning and all the theoretical questions that surround it, physicists are coming back!

Many young physicists (and mathematicians) are now working on trying to explain why deep learning works so well. This summer school is for them.

We need to find ways to connect this emerging community with the ML/AI community. It’s not easy because (1) papers submitted by physicists to ML conferences rarely make it because of a lack of qualified reviewers; (2) conference papers don’t count in a physicist’s CV.

http://cargese.krzakala.org
MODELS

- **In data science, models are** used to fit the data (e.g. linear regression: Best straight line that captures the dependence of $y$ on $x$?). In physics we could call those an “ansatz”.

- **In physics, models are** the main tool for understanding.

$$P(\{S_i\}_{i=1,\ldots,N}) = \frac{e^{-\beta \mathcal{H}}}{Z}$$

$$\mathcal{H} = -J \sum_{(ij) \in \mathcal{E}} S_i S_j$$

magnetism of materials
WHAT TO MODEL IN DEEP LEARNING?

We aim to reproduce the salient behaviours of the real system.
Iterative process of improving the model.
WHEN CAN A NEURAL NETWORK LEARN A TEACHER-NEURAL NETWORK?

**Teacher-network**

- Generates data $X$, $n$ samples of $p$ dimensional data, e.g. random input vectors.
- Generates weights $w^*$, e.g. iid random.
- Generates labels $y$.

**Student-network**

- Observes $X$, $y$, the architecture of the network.
- How does the best achievable generalisation error depend on the number of samples $n$?
Yoshua Bengio at France in AI’18: On challenges of deep learning towards AI.
Imagine yourself approaching another planet and observing the bits of information exchanged by aliens communicating with each other.

Unlike on Earth, their communication channel is noisy, but like on Earth, bandwidth is expensive. The best way to communicate is to maximally compress the messages, which leads to sequences of random bits being actually exchanged.

If we only observe the compressed messages, there is no way we can ever understand the alien language.
Overparametrization may help optimization: folklore experiment \( \text{e.g.}[\text{Livni et al.'14}] \)

Generate labeled data by feeding random input vectors into depth 2 net with hidden layer of size \( n \)

Still no theorem explaining this...

Much easier to train a new net with bigger hidden layer!
Take random iid Gaussian $X_{\mu i}$ and random iid $w_i^*$ from $P_w$.

Create $y_\mu = \varphi \left( \sum_{i=1}^{p} X_{\mu i} w_i^* \right)$.

High-dimensional regime: $n \to \infty$, $p \to \infty$, $\alpha \equiv n/p = \Omega(1)$.
First-order transition to perfect generalization in a neural network with binary synapses

Géza Györgyi*
School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332-0430
(Received 9 February 1990)

Learning from examples by a perceptron with binary synaptic parameters is studied. The examples are given by a reference (teacher) perceptron. It is shown that as the number of examples increases, the network undergoes a first-order transition, where it freezes into the state of the reference perceptron. When the transition point is approached from below, the generalization error reaches a minimal positive value, while above that point the error is constantly zero. The transition is found to occur at $a_{GD} = 1.245$ examples per coupling.

- **Binary teacher-weights:** $w^* \in \{-1,1\}^p$
- **Phase transition** in the generalization error’s dependence on sample complexity.

\[
\alpha_{GD} = 1.245 \quad \alpha_{AT} = 1.493
\]

\[
\alpha = n/p
\]
• Solution for any activation function, general class of priors on weights.

• Regions of optimality of approximate message passing (=TAP) algorithm.

• Rigorous proof that the replica solution for the teacher-student model is correct.

Barbier, Krzakala, Macris, Miolane, LZ, arXiv:1708.03395, COLT’18, PNAS’19
Def. “quenched” free energy: \[ f = \lim_{p \to \infty} \frac{1}{p} \mathbb{E}_{y, X} \log Z(y, X) \quad \alpha = \frac{p}{n} \]

Theorem 1:

\[ f = \sup_m \inf_{\hat{m}} f_{RS}(m, \hat{m}) \]

\[ f_{RS}(m, \hat{m}) = \Phi_{P_X}(\hat{m}) + \alpha \Phi_{P_{out}}(m; \rho) - \frac{m\hat{m}}{2} \]

where

\[ \Phi_{P_X}(\hat{m}) \equiv \mathbb{E}_{z, x_0} \left[ \ln \mathbb{E}_x \left[ e^{\hat{m}x_0 + \sqrt{\hat{m}}xz - \hat{m}x^2/2} \right] \right] \]

\[ \Phi_{P_{out}}(m; \rho) \equiv \mathbb{E}_{v, z} \left[ \int \text{d}y \ P_{out}(y|\sqrt{m}v + \sqrt{\rho - m}z) \ln \mathbb{E}_w \left[ P_{out}(y|\sqrt{m}v + \sqrt{\rho - m}w) \right] \right] \]

\[ x, x_0 \sim P_w \quad z, v, w \sim \mathcal{N}(0, 1) \quad \rho = \mathbb{E}_{P_w}(w^2) \]
Def. “quenched” free energy: \( f = \lim_{p \to \infty} \frac{1}{p} \mathbb{E}_{y, X} \log Z(y, X) \quad \alpha = \frac{p}{n} \)

**Theorem 1:**

\[
\begin{align*}
  f &= \sup_{m} \inf_{\hat{m}} f_{RS}(m, \hat{m}) \\
  f_{RS}(m, \hat{m}) &= \Phi_{P_X}(\hat{m}) + \alpha \Phi_{P_{out}}(m; \rho) - \frac{m \hat{m}}{2}
\end{align*}
\]

**Theorem 2:** Optimal generalisation error

\[
\mathcal{E}_{gen} = \mathbb{E}_{v, \xi} \left[ f_{\xi}(\sqrt{\rho} v)^2 \right] - \mathbb{E}_{v} \mathbb{E}_{w, \xi} \left[ f_{\xi}(\sqrt{m^*} v + \sqrt{\rho - m^*} w) \right]^2
\]

where \( m^* \) is the extremizer of \( f_{RS} \).
Algorithm 2 Generalized Approximate Message Passing (G-AMP)

**Input:** y
**Initialize:** $a^0_0, v^0_0, g_{out,\mu}^0, t = 1$

**repeat**

AMP Update of $\omega_\mu, V_\mu$

$$
V_\mu^t \leftarrow \sum_i F_{\mu i}^2 v_{i, t-1}^t
$$

$$
\omega_\mu^t \leftarrow \sum_i F_{\mu i} a_{i, t-1}^{t-1} - V_\mu^t g_{out,\mu}^t
$$

AMP Update of $\Sigma_i, R_i, g_{out,\mu}$

$$
g_{out,\mu}^t \leftarrow g_{out}(\omega_\mu^t, y_\mu, V_\mu^t)
$$

$$
\Sigma_i^t \leftarrow \left[ - \sum_\mu F_{\mu i}^2 \partial_\omega g_{out}(\omega_\mu^t, y_\mu, V_\mu^t) \right]^{-1}
$$

$$
R_i^t \leftarrow a_i^{t-1} + \Sigma_i^t \sum_\mu F_{\mu i} g_{out,\mu}^t
$$

AMP Update of the estimated marginals $a_i, v_i$

$$
a_i^t \leftarrow f_a(\Sigma_i^t, R_i^t)
$$

$$
v_i^t \leftarrow f_v(\Sigma_i^t, R_i^t)
$$

**t \leftarrow t + 1**

**until** Convergence on $a, v$

**output:** $a, v$.

Simple to implement, only matrix multiplications, $O(N^2)$

\[
f_a(\Sigma, R) = \int dx \frac{x P_X(x) e^{-\frac{(x-R)^2}{2\Sigma}}}{\int dx P_X(x) e^{-\frac{(x-R)^2}{2\Sigma}}}, \quad f_v(\Sigma, R) = \Sigma \partial R f_a(\Sigma, R).
\]

\[
g_{out}(\omega, y, V) \equiv \frac{\int dz P_{out}(y|z)(z - \omega) e^{-\frac{(z-\omega)^2}{2V}}}{V \int dz P_{out}(y|z)e^{-\frac{(z-\omega)^2}{2V}}}.
\]
SPHERICAL PERCEPTRON

\[ \phi(z) = \text{sign}(z) \quad P_w = \mathcal{N}(0,1) \quad n \to \infty \quad p \to \infty \quad n/p = \Omega(1) \]

- Generalisation error
- Optimal AMP algorithm
- Logistic regression

Graph showing generalisation error vs. \# of samples per dimension.
$y_\mu = \text{sign} \left( \sum_{i=1}^{p} X_{\mu i} w_i \right)$

$w_i \in \{-1, +1\}$

$n \to \infty$

$p \to \infty$

$n/p = \Omega(1)$

$y_\mu = \text{sign} \left( \sum_{i=1}^{p} X_{\mu i} w_i \right)$

$w_i \in \{-1, +1\}$
$y_\mu = \text{sign} \left( \sum_{i=1}^{p} X_{\mu i} w_i \right)$ \quad w_i \in \{-1, +1\}$

$n \to \infty$ \quad $p \to \infty$ \quad $n/p = \Omega(1)$
SYMMETRIC-DOOR PERCEPTRON

\[ y_\mu = \text{sign} \left( \left| \sum_{i=1}^{p} X_{\mu i} w_i \right| - K \right) \]

\[ w_i \in \{-1, +1\} \]

\[ n \rightarrow \infty \quad p \rightarrow \infty \quad \frac{n}{p} = \Omega(1) \]

![Graph showing symmetric door channel, binary \( X^* \)]

Generalisation error

# of samples per dimension \( n/p \)
Is this bringing us towards the theory of deep learning?
color-code:
described so far
needed

- structured data
- data structure
- iid inputs, teacher outputs
- multiple wide layers
- no hidden units
- architecture
- algorithm
- message passing
- gradient-descent-based
Committee machine

Model from Schwarze’92.


Limit: \( n \to \infty \)  \( p \to \infty \)  \( \alpha = n/p = \Omega(1) \)  \( K = O(1) \)
Specialization phase transition = hidden units specialise to correlate with specific features.

Consequence: Sharp threshold for number of samples below which linear regression is the best thing to do.
$y_\mu = \text{sign} \left[ \sum_{a=1}^{K} \text{sign} \left( \sum_{i} X_{\mu,i} w_{i,a} \right) \right]$

hidden units $K \gg 1$

- Large algorithmic gap:
  - IT threshold: $n > 7.65Kp$
  - Algorithmic threshold $n > \text{const} \cdot K^2p$

![Computational Gap Diagram](image-url)
color-code:
described so far
needed

algorithm
message passing
gradient-descent-based

structured data
data structure
multiple wide layers
only few hidden units
iid inputs, teacher outputs
architecture
How do the gradient-descent-based algorithms compare to the performance of approximate message passing?

Deep learning is fuelled by gradient descent. Understanding is needed!

Progress recently: Linear networks (trivial fixed point). Lazy training (NTK) networks close to initialization. Infinitely wide single layer networks.
GRADIENT-BASED ALGORITHMS

\[ \dot{x}_i(t) = -\mu(t)x_i(t) - \frac{\partial \mathcal{H}}{\partial x_i} + \eta_i(t) \]

- **T=1 Langevin algorithm**: At large time (exponentially) samples the posterior measure.
- **T=0 Gradient flow**.

Where do they go in large constant time?
• High-dimensional. Non-convex loss. Random perceptron?

• Notion of a “good” configuration (~ generalisation error) beyond lowest-loss configuration. Teacher-student perceptron? Hard to analyze (Agoritsas, Biroli, Urbani, Zamponi’18)

• Solvability: Error of gradient flow and Langevin algorithm follows a closed-form tractable equation. Spiked tensor model?

• Have (hopefully) behaviour that has a large universality class.
Signal $x^*$ on a sphere, observe a matrix $Y$ and tensor $T$ as:

$$Y_{ij} = \frac{1}{\sqrt{N}} x^*_i x^*_j + \xi_{ij} \quad \xi_{ij} \sim \mathcal{N}(0, \Delta_2)$$

$$T_{i_1 \ldots i_p} = \frac{\sqrt{(p-1)!}}{N(p-1)/2} x^*_i \ldots x^*_p + \xi_{i_1 \ldots i_p} \quad \xi_{i_1, \ldots, i_p} \sim \mathcal{N}(0, \Delta_p)$$

Corresponding Hamiltonian (loss function, log-likelihood)

$$\mathcal{H}(x) = -\frac{1}{\Delta_2 \sqrt{N}} \sum_{i<j} Y_{ij} x_i x_j - \frac{\sqrt{(p-1)!}}{\Delta_p N(p-1)/2} \sum_{i_1<\ldots<i_p} T_{i_1 \ldots i_p} x_{i_1} \ldots x_{i_p}$$

spherical constraint: $\sum_{i=1}^{N} x_i^2 = N$

Planted version of the mixed 2+p spherical spin glass model.
Bayes-optimal inference = computation of marginals/local magnetization of the Boltzmann measure at $T=1$.

➡ Langevin algorithm.

Maximum likelihood inference = computing the ground state.

➡ Gradient flow.
DYNAMICAL MEAN FIELD THEORY

The same model without spike: mixed spherical p-spin glass

Mean field theory of glassy dynamics:

Analytical Solution of the Off-Equilibrium Dynamics of a Long-Range Spin-Glass Model

L. F. Cugliandolo and J. Kurchan
Dipartimento di Fisica, Università di Roma, La Sapienza, I-00185 Roma, Italy
and Istituto Nazionale di Fisica Nucleare, Sezione di Roma I, Roma, Italy
(Received 8 March 1993)

We study the nonequilibrium relaxation of the spherical spin-glass model with p-spin interactions in the $N \to \infty$ limit. We analytically solve the asymptotics of the magnetization and the correlation and response functions for long but finite times. Even in the thermodynamic limit the system exhibits "weak" (as well as "true") ergodicity breaking and aging effects. We determine a functional Parisi-like order parameter $P_d(q)$ which plays a similar role for the dynamics to that played by the usual function for the statics.

PACS numbers: 75.10.Nr, 02.50.-r, 05.40.+j, 64.60.Cn

Proof of this without spike: BenArous, Dembo, Guionnet’06.
\[ C_N(t, t') = \frac{1}{N} \sum_{i=1}^{N} x_i(t)x_i(t'), \]

\[ \overline{C}_N(t) = \frac{1}{N} \sum_{i=1}^{N} x_i(t)x_i^*, \]

\[ R_N(t, t') = \frac{1}{N} \sum_{i=1}^{N} \partial x_i(t)/\partial h_i(t')|_{h_i=0}, \]

\[ Q(x) = x^2/(2\Delta_2) + x^p/(p\Delta_p). \]

\[ N \to \infty \]

\[
\frac{\partial}{\partial t} C(t, t') = 2R(t', t) - \mu(t)C(t, t') + Q'(\overline{C}(t))\overline{C}(t') + \int_0^t dt'' R(t, t'')Q''(C(t, t''))C(t', t'') + \int_0^t dt'' R(t', t'')Q'(C(t, t'')),
\]

\[
\frac{\partial}{\partial t} R(t, t') = \delta(t - t') - \mu(t)R(t, t') + \int_0^t dt'' R(t, t'')Q''(C(t, t''))R(t'', t'),
\]

\[
\frac{\partial}{\partial t} \overline{C}(t) = -\mu(t)\overline{C}(t) + Q'(\overline{C}(t)) + \int_0^t dt'' R(t, t'')\overline{C}(t'')Q''(C(t, t'')),
\]

Langevin algorithm (T=1)

\[
\frac{\partial}{\partial t} C(t, t') = -\tilde{\mu}(t)C(t, t') + Q'(\overline{C}(t))\overline{C}(t') + \int_0^t dt'' R(t, t'')Q''(C(t, t''))C(t', t'') + \int_0^t dt'' R(t', t'')Q'(C(t, t'')),
\]

\[
\frac{\partial}{\partial t} R(t, t') = -\tilde{\mu}(t)R(t, t') + \int_{t'}^t dt'' R(t, t'')Q''(C(t, t''))R(t'', t'),
\]

\[
\frac{\partial}{\partial t} \overline{C}(t) = -\tilde{\mu}(t)\overline{C}(t) + Q'(\overline{C}(t)) + \int_0^t dt'' R(t, t'')\overline{C}(t'')Q''(C(t, t'')),
\]

Gradient flow (T=0)
LANGEVIN STATE EVOLUTION
(NUMERICAL SOLUTION)

\[ \Delta_2 = 0.7 \]

correlation with ground truth

AMP

igin.com/sphinxteam/spiked_matrix-tensor
GRADIENT-FLOW PHASE DIAGRAM

\[ p = 3 \]

Gradient flow

Langevin hard

H.

Impossible

\[ \frac{1}{L_2} \]

\[ \Delta_p \]
POPULAR “EXPLANATION”

Increasing the SNR

Trivialisation
Annealed entropy of local minima (at m=0 also quenched):

\[ \tilde{\Sigma}_{\Delta_2, \Delta_P}(m, \epsilon_2, \epsilon_p) = \frac{1}{2} \log \left( \frac{p-1}{\Delta_P} + \frac{1}{\Delta_2} \right) + \frac{1}{2} \log(1 - m^2) \]

\[-\frac{1}{2} \left( \frac{m^{p-1}}{\Delta_P} + \frac{m}{\Delta_2} \right)^2 (1 - m^2) - \frac{p \Delta_P}{2} \left( \epsilon_p + \frac{m^p}{p \Delta_P} \right)^2 \]

\[-\Delta_2 \left( \epsilon_2 + \frac{m^2}{2 \Delta_2} \right)^2 + \Phi(t) - L(\theta, t), \]

where:

\[ \Phi(t) = \frac{t^2}{4} + 1_{|t| > 2} \left[ \log \left( \sqrt{\frac{t^2}{4} - 1} + \frac{|t|}{2} \right) - \frac{|t|}{4} \sqrt{t^2 - 4} \right] \]

\[ L(\theta, t) = \begin{cases} 
\frac{1}{4} \int_{\theta + \frac{1}{\theta}}^t \sqrt{y^2 - 4} dy - \frac{\theta}{2} \left( t - \left( \theta + \frac{1}{\theta} \right) \right) & \text{for } \theta > 1, 2 \leq t < \frac{\theta^2 + 1}{\theta} \\
\frac{t^2 - (\theta + \frac{1}{\theta})^2}{8} & \text{for } \theta > 1, 2 \leq t < \frac{\theta^2 + 1}{\theta} \\
\infty & \text{for } t < 2 \\
0 & \text{otherwise.} 
\end{cases} \]

Similar to Ben Arous, Mei, Song, Montanari, Nica’17; Ros, Ben Arous, Biroli, Cammarota’18 for spiked tensor model.
SPURIOUS MINIMA DO NOT NECESSARILY CAUSE GF TO FAIL

Gradient flow
Landscape trivialisation

H. Impossible

$\Delta_p$

$p=3$
WHAT IS GOING ON?

$p = 3, \Delta_p = 1.0$

$\varepsilon$

$t$

$1/\Delta_2 = 1.5$

$1/\Delta_2 = 1.9$

$1/\Delta_2 = 2.3$

$1/\Delta_2 = 2.7$

$N = 65535, \quad \Delta_p = 1.0$

$\varepsilon$

$t$
WHAT IS GOING ON?

Threshold energy in the non-planted model (m=0)

$p = 3, \Delta p = 1.0$

\(\varepsilon\)

\(\frac{1}{\Delta_2} = 1.5\)
\(\frac{1}{\Delta_2} = 1.9\)
\(\frac{1}{\Delta_2} = 2.3\)
\(\frac{1}{\Delta_2} = 2.7\)
Dynamics first goes to the threshold states (replicon condition):

\[
\frac{T^2}{(1 - q^{th})^2} = (p - 1) \frac{(q^{th})^{p-2}}{\Delta_p} + \frac{1}{\Delta_2}
\]

AMP state evolution at fixed \( q \), determines stability of \( T=0 \):

\[
m^{t+1} = \frac{1 - q}{T} \left( \frac{m^t}{\Delta_2} + \frac{(m^t)^{p-2}}{\Delta_p} \right)
\]

Leads to the Langevin/gradient-flow transition (conjecture):

\[
\frac{1}{\Delta_2^2} = (p - 1) \frac{(1 - T\Delta_2)^{p-2}}{\Delta_p} + \frac{1}{\Delta_2}
\]
\[ \Delta_{2}^{GF} = \frac{-\Delta_{p} + \sqrt{\Delta_{p}^{2} + 4(p-1)\Delta_{p}}}{2(p-1)} \]

GRADIENT-FLOW PHASE DIAGRAM
\[ \Delta_{2}^{\text{Lang}} = \sqrt{\frac{\Delta_{3}}{2}} \]

Langevin Phase Diagram

- $p=3$

- Hard
- Easy
- Impossible
LANDSCAPE ANALYSIS

Sarao, Biroli, Cammarota, Krzakala, LZ’19

Increasing the SNR

Former minima develop a negative slope in the direction of the spike!

Trivialisation
CONCLUSION ON GRADIENT-ALGORITHMS

- **Gradient flow** worse than **Langevin algorithm**, both worse than **AMP**. How can GF & LA be improved to reach the AMP threshold?

- Gradient flow (sometimes) works even when spurious local minima are present. Quantified with the Kac-Rice approach.

- First time we have a closed-form conjecture for the threshold of gradient-based algorithms including constants. **Applicable beyond the present model?**
TOWARDS THEORY OF DEEP LEARNING?

color-code:
described so far
needed

structured data, teacher outputs
data structure

multiple wide layers
only few hidden units
architecture

algorithm
message passing
gradient-descent-based
Teacher/student:
Plateau in learning dynamics, due to specialisation (Saad, Solla’95). MNIST (even vs odd classification): No plateau ...

Goldt, Krzakala, Mézard, LZ; arXiv:1909.11500
MNIST VS TEACHER/STUDENT

MNIST (odd vs even):
Two independent students do not learn the same function!

Teacher/student:
Two independent students learn the same function!
HIDDEN MANIFOLD MODEL

Input data: \[ X \in \mathbb{R}^{n \times p} \quad C \in \mathbb{R}^{n \times d} \quad F \in \mathbb{R}^{d \times p} \]

\( n \) samples, \( p \) input & \( d \) latent dimension.

Input on low-dimensional manifold.

\[ X = f(CF) \]

\( C, F \) iid matrices.

Vanilla teacher/student

True labels:

Depend on the latent coordinates \( C \).

\[ \tilde{y}_\mu = \sum_{m=1}^{M} \tilde{v}_m g \left( \langle \tilde{w}_m, C_\mu \rangle \right) \]

\[ y_\mu = \sum_{m=1}^{M} v_m g \left( \langle w_m, X_\mu \rangle \right) \]

\( X \) is iid matrix
MNIST (odd vs even): Two independent students do not learn the same function!

Hidden manifold (d=10): Two independent students do not learn the same function!
Teacher acting on X: Plateau in learning dynamics
MNIST & hidden manifold: No plateau ...
CONCLUSION ON HIDDEN MANIFOLD

• The hidden manifold model reproduces/captures behaviour of learning-dynamics on MNIST.

• Both (i) low-dimensional structure of input, and (ii) labels depending on the latent representation are needed.

• TODO: Solve analytically.

• TODO: Generalize to be able to demonstrate the advantage of depth.
TOWARDS THEORY OF DEEP LEARNING?

color-code:
described so far
needed

hidden manifold data
data structure
algorithm
message passing
gradient-descent-based

iid inputs, teacher outputs
multiple wide layers
only few hidden units
architecture
Physics has many useful tools applicable in high-dimensional inference and learning.
REFERENCES FOR THIS TALK

- Barbier, Krzakala, Macris, Miolane, LZ; *Optimal errors and phase transitions in high-dimensional generalized linear models*; COLT’18, PNAS’19, arXiv:1708.03395.


