Learning by Local Entropy Maximization: the effective landscape of neural networks learning algorithms

Riccardo Zecchina

Bocconi University, Milan
Institute for Data Science

Carlo Baldassi (Bocconi, Milan)

Federica Gerace
Carlo Lucibello
Luca Saglietti

Politecnico di Torino
Human Genetics Foundation

Alessandro Ingrosso
Columbia University

Christian Borgs
Jennifer Chayes
Leon Bottou
Yann Lecun
Pratik Chaudhari
Stefano Soatto
Bert Kappen

Microsoft Research New England
FB research
UCLA

Radboud University Nijmegen
Plan of the talk

- Geometrical structure of minima in non-convex random optimization and learning problems
  - Clustering and symmetry breaking
  - The Local Entropy Measure reveals the existence of subdominant high local density regions in weight space.
  - Accessibility and Local Bayesian predictions

- Algorithms from a local entropy measure
  - The Robust Ensemble: an “out-of-equilibrium” measure
  - Real replicas algorithms: MCMC, SGD and Belief Propagation
  - Connections with DNNs
  - ...
What makes a constraint satisfaction problem or a learning problem extracted from a natural distribution hard to solve?

Basic example: Random K-SAT

- Let $C_K(N)$ be the set of all $2^K \binom{N}{K}$ possible K-clauses on $x_1, x_2, \ldots, x_N$
- Select uniformly, independently and with replacements $M = \alpha N$ clauses from $C_K(N)$ to generate a K-cnf formula $F_N(K, \alpha)$

$$F = (x_1 \lor x_27 \lor \bar{x}_3) \land (\bar{x}_{11} \lor x_3 \lor x_2) \land \ldots \land (x_9 \lor \bar{x}_8 \lor \bar{x}_{30})$$

**Question:** does $F_N(K, \alpha)$ have a truth assignment?
\textbf{Factor Graphs for CSPs}

- $N$ discrete variables $\{x_i\}$, e.g., Boolean, spins, colors
- Constraints $E_a$, $a = 1, \ldots, M$ involving vars $\{x_{i(a)}\}$

\[ E_a = \begin{cases} 
0 & \text{if } \{x_{i(a)}\} \text{ satisfy constraint} \\
1 & \text{otherwise}
\end{cases} \]

Cost/Energy function:

\[ E = \sum_{a=1}^{\alpha N} E_a [\{x\{i(a)\}\}] \]

\[ E_a = (x_{i_1} \lor \bar{x}_{i_2} \lor x_{i_3}) \]

$M = \alpha N$
The SAT threshold conjecture.

For each $k \geq 2$, random $k$-SAT has a sharp satisfiability threshold $\tilde{\delta}$.

Increasing $P_{SAT}$ with $n \to \infty$ as $n \to \infty$—that is, a single critical value $\tilde{\delta}$ separates SAT from UNSAT (with high probability in the limit $n \to \infty$; fixed $k$).

$E_0 = 0$

$E_0 > 0$

$\alpha = \frac{M}{N}$

$P(SAT) \to 1$ as $n \to \infty$

$P(UNSAT) \to 1$ as $n \to \infty$

median computational complexity

number of clauses per variable $\alpha$
Finding isolated solutions is hard. In the last 15 years many physicists, mathematicians and CS have contributed to various aspects of these results … the scenario is by now rigorously established.
Gibbs measure decomposition

\[ P(w) = \frac{1}{Z} \prod_{a=1}^{M} \Psi_a(\{w_{(i,j)} \in a\}) \]

**RS:** \( P_1 = 1 \)

**1RSB-d:** \( \mathcal{N} = e^{\Sigma N} \)

**1RSB-s:** \( \mathcal{N} = \text{sub-exp} \)

\[ P_\ell = \sum_{\{w \in A_\ell\}} P(w) \]

\( P_1 > P_2 > P_3 > ... \)

---

![Graphical representation](image)
Learning as a CSP problem

Constraints: one for each pattern

\[ f(W; \sigma^\mu, \xi^\mu) = \delta(\sigma^\mu; \sigma(W, \xi^\mu)) \]

\[ \sum_j W_{j,k} \xi^\mu_j - \gamma_k = 0 \]

\[ \sigma^\mu = \theta(\sum_j W_{k} \tau_k - \gamma) \]

Fully connected factor graph
“Old” (90s) statistical physics results: a relatively similar scenario

Geometry of space of solution and internal representations in MLP learning random patterns with continuous weights (zero errors landscape)

Fractional volume of weights storing the patterns

\[ V = \frac{\int d\mathbf{w} \delta(\mathbf{w}^2 - 1) \prod_{\mu} \delta(\sigma^{\mu}; \sigma(\mathbf{w}, \xi^{\mu}))}{\int d\mathbf{w} \delta(\mathbf{w}^2 - 1)} \]
How does learning take place in large scale DNNs?

Learning algorithms: Variants of gradient back-propagation

However, successful algorithms never “simply” minimize the loss.

Why?
The simplest non-convex neural device: perceptron with discrete weights

Analytical results generalise to arbitrary number of levels and multiple layers.
Non-convex minimum “energy” problem

Given a set of i.i.d. random examples (p=1/2):

$$\{(\xi_i^\mu = \pm 1, \sigma^\mu = \pm 1)\} \quad i = 1, \ldots, N \quad \mu = 1, \ldots, \alpha N$$

Find \( \mathbf{W} \) such that

$$\sigma^\mu = \sigma(\mathbf{W}, \xi^\mu) \quad \forall \mu$$

$$\iff \alpha N \quad \text{constraints on} \quad \{W_i\}$$

Cost-energy function

$$E(\mathbf{W}) = \sum_{\mu} \Theta \left( -\sigma^\mu \text{sgn}(\mathbf{W} \cdot \xi^\mu) \right) = \# \text{ number of errors}$$

$$\Theta(x) = \begin{cases} 
1 & x \geq 0 \\
0 & x < 0 
\end{cases}$$
Phase diagram (~1990)

- At $E=0$, minima are very narrow and isolated

\[ S = \log (\# \text{optimal } W \text{ assignments}) \]

\[ \alpha_c = 0.83 \]

\[ \alpha = P/N \]

For decades, heuristic local search algorithms were believed to fail in finding solutions for any extensive number of patterns.

some classical papers:


Geometry of the space of solutions in the binary perceptron:

Franz-Parisi potential: entropy at distance $d$, sampling from typical solution $J$

$$F(x) = \left\langle \frac{1}{Z(T')} \sum_J \Theta \left( \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i \xi_i^\mu \right) \ln \sum_w \Theta \left( \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \xi_i^\mu \right) e^{x J \cdot w} \right\rangle_{\xi}$$

$\min_{\alpha} (\alpha) \approx O(N)$

H. Huang, Y. Kabashima (2014) ($q=1$ known since the 80's)

$$\alpha_c = \frac{P_{max}}{N} \simeq 0.83$$

learning impossible for random patterns ?
Golf course for any $\alpha$? Efficient learning impossible?

Message-passing algorithms (2006, A. Braunstein RZ) and its simplifications work well.

success probability

$n \simeq 10^6$

something unclear …
Weight enumerator functions computed with BP, relative to a solution found by an algorithm and to a typical solution (planted)

Evidence for Random Walks analysis:

N=1001
samples=50

$W \sim O(N)$
Learning in rare regions? Large deviations analysis

\[ \tilde{\xi}^\mu \rightarrow \begin{array}{c}
\xi \in \{-1, 1\}^N \\
W_i \in \{\pm 1\}, \ i = 1, \ldots, N \\
\tau (W, \xi) = \text{sign} (W \cdot \xi) 
\end{array} \]

Characteristic function:

\[ X_\xi (W) = \prod_{\mu=1}^{\alpha N} \Theta (\sigma^\mu \tau (W, \xi^\mu)) = 1 \text{ iff all patterns are correctly classified} \]

**Number of solutions** within Hamming distance \(d\) from a given weight vector \(\tilde{W}\):

\[
\mathcal{N} (\tilde{W}, d) = \sum_{\{W\}} X_\xi (W) \delta \left( W \cdot \tilde{W}, N (1 - 2d) \right)
\]

\[ A \cdot B = \sum_{j=1}^{N} A_j B_j \]
Local entropy measure

number of solutions within a distance $d$

$$\mathcal{N}(\tilde{W}, d) = \sum_{\{W\}} x_\xi(W) \delta(W \cdot \tilde{W}, N (1 - 2d))$$

“energy” = local entropy

$$\mathcal{E}_d(\tilde{W}) \equiv - \log \mathcal{N}(\tilde{W}, d)$$

Local Entropy Measure

maximally dense for $y \to \infty$

$$\mathcal{P}(\tilde{W}) \propto e^{-y \mathcal{E}_d(\tilde{W})}$$

normalisation

$$Z(d) = \sum_{\{\tilde{W}\}} x_\xi(\tilde{W}) e^{-y \mathcal{E}_d(\tilde{W})}$$

By the replica/cavity method we can compute the expectation of the local entropy in the large $N$ limit

$$\mathcal{I}_I (d, y) = - \left\langle \mathcal{E} \left( \tilde{W} \right) \right\rangle_{\xi, \tilde{W}} = \frac{1}{N} \left\langle \log \mathcal{N} \left( \tilde{W}, d \right) \right\rangle_{\xi, \tilde{W}}$$

$$\mathcal{I}_I (d, y) = \partial_y \left( y \mathcal{F} (d, y) \right) \quad \text{internal entropy}$$

$$\mathcal{I}_E (d, y) = -y \left( \mathcal{F} (d, y) + \mathcal{I}_I (d, y) \right) \quad \text{external entropy}$$
Large deviation analysis

in the 90s we were not aware of this fundamental structural property
close to capacity the dense clusters breaks up

the shape is not spherical …
Probability to give the same answer of the teacher on a new input

Teacher

Student

Making predictions

Optimal Bayesian prediction:

\[ P(\sigma|\xi^{new}, \{\xi^\mu, \sigma_\mu\}) = \int dW \, P(\sigma|W, \xi^{new}, \{\xi^\mu, \sigma_\mu\}) \, P(W, \{\xi^\mu, \sigma_\mu\}) \]
Dense states have the propensity to generalise

- contribution to the Bayesian integral from the dense cluster
- the Teacher is an isolated weight vector
Prediction error \( \sim 1.2 \% \) with binary weights

no overfitting with size
Generalisation to multiple state variables

Few levels almost saturate performance

Principled algorithm: 
Local Entropy driven Simulated Annealing

Objective Function: 
search for configurations which maximize the local entropy (minimize the “energy”)

\[ \mathcal{E} (\tilde{W}) = - \log \mathcal{N} (\tilde{W}, d) \]

1. SA moves

2. Belief Propagation method to estimate the local entropy
We performed extensive simulations and studied the scaling properties of EdMC in conjunction with at least squares fit (Fig. 1). A typical trajectory of standard Monte Carlo is uncapable of reaching a zero energy configuration, and gets systematically trapped in low energy states (our MC is terminated after a fixed number of iterations). In Figure 4 we show the scaling of MC and EdMC, also comparing the number of EdMC necessary to reach the average energy, also comparing the number of EdMC necessary to reach the average energy.

Contrast to simulated annealing. Figure 2 is a log-log plot of the number of iterations per temperature change, and gets systematically trapped in low energy states (our MC is terminated after a fixed number of iterations).

The power scaling of simulated annealing has to be confronted with the almost linear behaviour of zero temperature EdMC, which is reported in Figure 3. The situation is similar to that of Perceptron Learning Problem.

Local should not be interpreted as *infinitesimal*: the local entropy is the log of the number of optimal configuration within a hyper-sphere of radius $O(N)$ or fractional volume $O(1)$. 
What we have learned from non-convex 1-2 layer NN learning random random patterns?

✓ The loss function presents an exponential proliferation of metastable states which trap SA or full batch Langevin dynamics

✓ HOWEVER, there exist rare dense regions (small but still of extensive size) which are accessible to simple non-detailed-balance stochastic algorithms. These regions have good generalisation capabilities.

✓ Accessibility and generalization are not in conflict

✓ The Local Entropy Measure amplifies the weight of these regions (from exponentially small to dominant!)

✓ shape of dense regions depends on the data, difficult to study analytically even for random patterns
Successful algorithms never “simply” minimize the loss.

Why?

Because the stationary measure of the stochastic learning process should not be the equilibrium Gibbs measure of the loss function. Many (simple!) out-of-equilibrium processes are attracted by the rare dense states (wide minima).
From Local entropy measure to the Robust Ensemble

$$
\mathcal{P}(\tilde{W}) \propto e^{-y\mathcal{E}_d(\tilde{W})} \quad \mathcal{E}_d(\tilde{W}) \equiv -\log \mathcal{N}(\tilde{W}, d)
$$

We may write

$$
\mathcal{P}(\tilde{W}) \propto \lim_{\beta \to \infty} \left( \sum_{\{W\}} e^{-\beta E(W) + \gamma W \cdot \tilde{W}} \right)^y
$$

where $\gamma$ is a Lagrange multiplier controlling the distance

\[ y \text{ integer } \Rightarrow \text{ multiple real replicas} \]

\[
\mathcal{P}(\tilde{W}) \propto \lim_{\beta \to \infty} \left( \sum_{\{W\}} e^{-\beta E(W) + \gamma W \cdot \tilde{W}} \right)^y = \lim_{\beta \to \infty} \prod_{a=1}^{y} \sum_{\{W^a\}} e^{-\beta E(W^a) + \gamma W^a \cdot \tilde{W}} = \]

\[
= \lim_{\beta \to \infty} \sum_{\{W^1, W^2, \ldots, W^y\}} e^{-\beta \sum_{a=1}^{y} E(W^a) + \gamma \sum_{a=1}^{y} W^a \cdot \tilde{W}}
\]

\[
= \lim_{\beta \to \infty} \sum_{\{W^1, W^2, \ldots, W^y\}} e^{-\beta \sum_{a=1}^{y} E(W^a) + \gamma \sum_{a=1}^{y} \sum_{j=1}^{N} W_j^a \tilde{W}_j}
\]
Robust Ensemble

\[ \mathcal{P}_{RE}(\tilde{W}, \{W^a\}) \propto e^{-\beta \sum_{a=1}^{y} E(W^a)} + \gamma \sum_{a=1}^{y} \sum_{j=1}^{N} W_j^a \tilde{W}_j \]

Marginalizing the center

\[ \hat{\mathcal{P}}_{RE}(\{W^a\}) \propto e^{-\beta(\sum_{a=1}^{y} E(W^a) - \frac{1}{\beta} \sum_{j=1}^{N} \log(2 \cosh(\gamma \sum_{a=1}^{y} W_j^a)))} \]

Expectations of observables:

\[ E[f(\tilde{W})] = \sum_{\tilde{W}} \sum_{\{W^a\}} f(\tilde{W}) \mathcal{P}_{RE}(\tilde{W}, \{W^a\}) \]
Replicated MC

\[ E(W) = \sum_{a=1}^{y} E(W^a) - \frac{1}{\beta} \sum_j \log(2 \cosh(\gamma \sum_{a=1}^{y} W^a_j)) \]

1) \( \Delta E = E(W') - E(W) \) can be computed efficiently when \( W' \) and \( W \) differ in one weight
2) efficient MC sampling for rejection rate reduction (non trivial)
3) most probable of the centroid value: \( \tilde{W}_j = \text{sign} \sum_{a=1}^{y} W^a_j \) (typically \( E(\tilde{W}) \leq \langle E(W^a) \rangle_a \))

\[ \alpha = 0.3 \]
\[ y = 3 \]

Notice: landscape of local minima could be different from the MC using BP
Replicated Stochastic Gradient Descent

\[ H \left( \{W^a\} \right) = \sum_{a=1}^{y} E \left( W^a \right) + \frac{1}{\beta} \sum_{j=1}^{N} \log \left( e^{-\frac{\gamma}{2} \sum_{a=1}^{y} (W_j^a - 1)^2} + e^{-\frac{\gamma}{2} \sum_{a=1}^{y} (W_j^a + 1)^2} \right) \]

\[ \frac{\partial H}{\partial W_i^a} \left( \{W^b\} \right) = \frac{\partial E}{\partial W_i} \left( W \right) \bigg|_{W=W^a} + \frac{\gamma}{\beta} \left( \tanh \left( \gamma \sum_{b=1}^{y} W_i^b \right) - W_i^a \right) \]

\[ (W_i^a)^{t+1} = (W_i^a)^t - \eta \frac{1}{m(t)} \sum_{\mu \in m(t)} \frac{\partial E^\mu}{\partial W_i} \bigg|_{W=(W^a)^t} + \eta' \left( \tanh \left( \gamma \sum_{b=1}^{y} (W_i^b)^t \right) - (W_i^a)^t \right) \quad \eta' = \frac{\gamma}{\beta \eta} \]

\[ N = 1605 \text{ weights} \quad K = 5 \]

![Graph showing error rate and epochs to solution as functions of patterns per synapse α for non-interacting (γ = 0) and interacting (γ > 0) cases. The error rate increases with α, and the epochs to solution also increase, especially for interacting cases.]
Replicated Belief Propagation: focusing BP (fBP) ~ BP with reinforcement

\[ m_{* \rightarrow j}^{t+1} = \tanh \left( (y - 1) \tanh^{-1} \left( m_{j \rightarrow *}^t \tanh \gamma \right) \right) \tanh \gamma \]

fBP becomes a solver looking for high density regions of solution. Interesting convergence properties (to be further studied).
Replicated BP is also an analytical tool: phase diagram on NN with one hidden layer

committee machine with $N = 1605, K = 5, y = 7$, increasing $y$ from 0 to 2.5, averages on 10 samples. Top: local entropy versus distance to the reference $W^*$ for various $\alpha$ (error bars not shown for clarity). The topmost grey curve ($\alpha = 0$) is an upper bound, representing the case where all configurations within some distance are solutions.

The case of Deep Networks

Modified sampling measure:

\[ P(x') \propto e^{y \Phi(x')} \]

Local free entropy:

\[
\Phi(x') = \log \int_x e^{-\beta f(x)} e^{-\lambda \|x - x'\|^2} dx
\]

where: \( x \) are the continuous weights, and \( f(x) \) is the loss/energy function

Langevin dynamics:

```
Algorithm 1: Entropy-SGD algorithm

Input: current weights \( x \), Langevin iterations \( L \)
Hyper-parameters: scope \( \gamma \), learning rate \( \eta \), SGLD step size \( \eta' \)

// SGLD iterations:
1 \( x', \mu \leftarrow x; \)
2 for \( \ell \leq L \) do
3 \( \Xi^\ell \leftarrow \) sample mini-batch;
4 \( dx' \leftarrow \frac{1}{m} \sum_{i=1}^{m} \nabla_x f(x'; \xi_i) - \gamma (x - x'); \)
5 \( x' \leftarrow x' - \eta' dx' + \sqrt{\eta'} \epsilon \sim \mathcal{N}(0, I); \)
6 \( \mu \leftarrow (1 - \alpha) \mu + \alpha x'; \)

// Update weights:
7 \( x \leftarrow x - \eta \gamma (x - \mu) \)
```

**Entropy-SGD: Biasing Gradient Descent into Wide Valleys**

Local Entropy & Robust Ensemble

\[ \beta \rightarrow \infty \quad \text{sampling from} \]

\[ P_{RE}(\tilde{x}) \propto \int dx^1 \ldots dx^y e^{-\beta \phi(\tilde{x}, \{x^a\})} \]

\[ \phi(\tilde{x}, \{x^a\}) = \sum_{a=1}^{y} E(x^a) + \frac{\lambda}{\beta} \sum_{a=1}^{y} \|x^a - \tilde{x}\|^2 \]

Elastic Averaging SGD with momentum

\[ \min_{x^1, \ldots, x^p, \tilde{x}} \frac{1}{p} \sum_{i=1}^{p} \left( E[f(x^i, \xi^i)] + \frac{\rho}{2} \|x^i - \tilde{x}\|^2 \right) \]

Work in progress with L. Bottou, L. Sagun, J. Chayes, C. Borgs, C. Baldassi, …

Deep learning with Elastic Averaging SGD

Sixin Zhang  
Courant Institute, NYU  
zsx@cims.nyu.edu

Anna Choromanska  
Courant Institute, NYU  
achoroma@cims.nyu.edu

Yann LeCun  
Center for Data Science, NYU & Facebook AI Research  
yann@cims.nyu.edu

Loss function:
\[
\min_{x^1, \ldots, x^p, \tilde{x}} \sum_{i=1}^{p} \mathbb{E}[f(x^i, \xi^i)] + \frac{\rho}{2} \|x^i - \tilde{x}\|^2,
\]

with Momentum SGD.

CIFAR dataset with the 7-layer convolutional neural network.

The idea of Flat Minima is not new:
ON LARGE-BATCH TRAINING FOR DEEP LEARNING:
GENERALIZATION GAP AND SHARP MINIMA

Nitish Shirish Keskar
Northwestern University
Evaston, IL 60208
keskar.nitish@u.northwestern.edu

Dheevatsa Mudigere
Intel Corporation
Bangalore, India
dheevatsa.mudigere@intel.com

Jorge Nocedal
Northwestern University
Evaston, IL 60208
j-nocedal@northwestern.edu

Ping Tak Peter Tang
Intel Corporation
Santa Clara, CA 95054
peter.tang@intel.com

ITAYHUBARA

Mikhail Smelyanskiy
Intel Corporation
Santa Clara, CA 95054
mikhail.smelyanskiy@intel.com

Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to $+1$ or $-1$

Matthieu Courbariaux$^{1}$
Itay Hubara$^{2}$
Daniel Soudry$^{3}$
Ran El-Yaniv$^{2}$
Yoshua Bengio$^{1,4}$

$^1$Université de Montréal
$^2$Technion - Israel Institute of Technology
$^3$Columbia University
$^4$CIFAR Senior Fellow

“Although BNNs are slower to train, they are nearly as accurate as 32-bit float DNNs.”
Conclusion and what next

Theoretical framework:
out-of-equilibrium statistical physics and large deviations studies are a key framework for understanding learning phenomena

Next algorithmic developments:

• Accessible dense states in DNN, connections with regularization techniques (dropout), temporal version of local entropy
• An opportunity for acceleration?
• Simple forms of stochastic learning process
• Learning with low precision weights: can we design new neural hardware?
• Generalization to unsupervised learning
• …