Statistical Physics, Neural Networks and Combinatorial Optimization

By Michael R. Douglas, Affiliate Faculty, Simons Center for Geometry and Physics

Photo courtesy Michael R. Douglas

Introduction

The idea of humanoid intelligent automata goes back to the ancient Greeks, and intensive study of artificial intelligence began soon after the construction of the first electronic computers in the 1940's. But the 2010s will be remembered as the decade when it came into its own, begining with the 2011 success of IBM's Watson at the game show Jeopardy, through the 2019 development of GPT-2, a program that among other things can write real-sounding newspaper stories, and shows no sign of slowing down. Above all, the 2016 triumph of DeepMind's AlphaGo program over Lee Sedol, the second ranked Go grandmaster in the world, convinced many people (including myself) that AI had reached a new level.

These advances did not spring out of nowhere, and in fact many of the underlying concepts were developed decades ago, with important contributions by mathematicians and physicists. One theme, which we will focus on here, was to invent idealized models of neural systems, which could be analyzed using tools from statistical physics. This work was largely theoretical, and for many years computers were not powerful enough to use these models for practical applications. This started to change in the 90s, and now of course we are in the opposite situation where new practical results far outrun our theoretical understanding.

As an applied field, machine learning (ML) has closer connections to statistics and applied mathematics than to pure mathematics and physics. Still, the early connection with statistical physics led to concepts and methods which remain quite important. In this article we will discuss the relation between disordered systems and combinatorial optimization, which grew out of the work on spin glasses of Giorgio Parisi and the ENS school.

Before we begin, let me briefly survey a few more areas at the interface between AI/ML, mathematics and physics, which I discuss in more detail in my talk [14]. As we have all heard, scientific computation and data analysis is being transformed by methods developed for machine learning. Collider experiments and astronomical observatories produce huge datasets, whose analysis is increasingly challenging. Rather than develop custom methods for each problem, general concepts in statistics and ML can be adapted for a wide variety of problems [10]. Even for much studied problems in computational physics and applied mathematics, such as protein folding, *ab initio* quantum computations and fluid mechanics, new methods are producing dramatic advances, such as the recent success of DeepMind's AlphaFold [25]. Many canonical structures in mathematics can be studied using ML techniques, and a particularly well developed example is Kähler-Einstein metrics [3, 8, 13, 15]. Many of these interests intersect in the mathematical study of feed-forward networks for representing functions, in particular as solutions to partial differential equations as in [16, 17, 23].

What about questions that directly address the central problems of AI? These include the ability to understand and use language, and to make and execute plans. Many researchers in AI have argued that the cleanest domain in which to study these matters is pure mathematics, and have posed questions such as: Can a computer find a mathematical proof, or discover a mathematical concept? In fact there is a very direct analogy between finding a proof and playing a game of solitaire, which enables

the use of the same techniques behind AlphaGo for mathematical theorem proving. These connections are actively studied by researchers in AI and theorem proving [1].

Statistical Physics and the Hopfield Model

Consider a piece of copper (nonmagnetic) doped with a small amount of manganese (magnetic). This is a prototypical example of a spin glass, modeled by the Hamiltonian

$$H = \sum_{i,j=1}^{N} J_{ij} S_i S_j. \quad (1)$$

The random positions of the impurity atoms are idealized by taking each of the couplings Jij (for a given i and j) as an independent random variable (say, Gaussian with zero mean).[1] Spin glasses display all sorts of strange behaviors such as an anomalously slow decay of induced magnetization. One signal of this is frustration, groups of pairwise interactions with conflicting signs and no common ground state. This leads to a very large number of local minima of the energy separated by potential barriers, leading to this slow decay.

During the 70s a theoretical understanding of spin glasses was developed, with major insights coming from the development of the replica method by Parisi and collaborators [20]. Without going into details, the starting point is the observation that to analyze a spin glass one takes sums over spin configurations to get a free energy, which must then be averaged over random couplings. This second average is not weighed by the spin partition function (it is"quenched") and to do it, one employs a mathematical "trick" of considering N_{rep} replicas (copies) of the spin system and taking $N_{rep} \to 0$. While fictitious, the replica structure actually reflects many qualitative properties of the system.

One of these properties is the "ultrametric structure" of the landscape and its basins of attraction. To describe this, let us consider the set of approximate ground states, spin configurations with $H(S) < (1-\epsilon) H_{min}$, where H_{min} is the ground state (minimum possible) energy and $\epsilon > 0$ is some small parameter. As we increase ϵ, the set becomes larger—and the way this happens is that different basins merge, in a way that determines a hierarchical or "tree" structure. Using an assumption of "complete replica symmetry breaking," one can compute the overlaps between different basins at different ϵ, making the picture quantitative.

The replica method leads to insights into disorder, which have been applied to many fields: just to consider biology, there is work on protein folding [9], the immune system [26], and many works in neuroscience. The first of these was Hopfield's model of memory [2, 18, 19]. This describes the process by which a group of neurons, given a stimulus, can classify it into one of a few groups (call these "memories") and produce an appropriate response for each group. Hopfield's motivating example was the sense of taste in the snail, which can identify hundreds of distinct substances and learn which ones are good to eat. But the model has interest far beyond neuroscience, as the classification problem is very central to machine learning.

To define the Hopfield model, we start by writing Eq. (1). The variables S_i are interpreted as "activations" of neurons, for example we might set $S_i = +1$ if neuron i is firing and $S_i = -1$ if not. The coupling J_{ij} is the strength of a synapse connecting neurons i and j. A value $J < 0$ favors correlated firing, so corresponds to an excitatory synapse. One can also have $J < 0$ and inhibition. All these analogies to neuroscience are suggestive but very imprecise, for example Eq. (1) requires the couplings to be symmetric, which is almost never the case in real life. Still, as physicists we can look for properties independent of such details.

Associative memory is idealized as follows. We have a list of configurations $S_i^{(1)}$, $S_i^{(2)}$ and so on, each a set of activations which represents one of the memories. We then start the system in an arbitrary configuration $S_i(t = 0)$, and we want to show there is a choice of couplings J_{ij} and some dynamics such that the system will evolve towards whichever of the $S_i^{(a)}$ is most similar to the original stimulus—say, has the largest dot product $\sum_i S_i(t=0) S_i^{(a)}$. The dynamics thus defines a basin of attraction (a region around $S_i^{(a)}$ which evolves to it), in other words a set of stimuli that the model will associate with memory "a".

[1] The original Edwards–Anderson spin glass Hamiltonian multiplies J_{ij} by a position dependent $f(|\vec{x}_i - \vec{x}_j|)$ term to suppress couplings between distant atoms, but this is not essential for our discussion. Setting $f = 1$ one obtains the Sherrington-Kirkpatrick model, with similar properties.

Assuming that the dynamics favors configurations with low energy, it is not hard to see that the following choice of couplings could work:

$$J_{ij} = -\frac{1}{n}\sum_{a=1}^{n} S_i^{(a)} S_j^{(a)} \qquad (2)$$

One starts by checking that each $S_i^{(b)}$ is a local minimum. Of course the term with $a = b$ will be large and negative, $H = -N/n$ where N is the number of neurons. Then, if the memories are randomly distributed in some sense (say, each component is independently chosen to be ±1 with equal probability), each of the other contributions will be of order $\sqrt{N/n}$. The sum of these competing effects could then be estimated as $\sqrt{N/n}$. So, for $n << N$, they should not disturb the local minimum at $S_i^{(b)}$. Thus, each of the n memories will be a local minimum.

This argument is borne out by simulation—one finds that Eq. (2) has the desired minima as long as $n \lesssim 0.14N$, while for larger n the competing effects spoil this. As N becomes large this threshold becomes sharp and in this sense there is a phase transition in the model. This observation was very influential as we explain shortly. The threshold ratio $\alpha \sim 0.14$ can be computed using the replica method, and eventually rigorous arguments appeared [27].

Although drastically simplified compared to real neural systems, Hopfield's hope was that his model might capture some real world phenomena. An example was presented by Michail Tsodyks at the SCGP January 2020 workshop "Physics of Neural Circuits and Network Dynamics" [24]. A classic experiment in human memory is free recall, in which one gives a subject a list of words to learn and later repeat, in any order, with no prompts or clues. This requires not just learning the words but also generating them spontaneously, in a way that could be modeled by the attractor dynamics of the Hopfield model. The model can explain observable quantities such as the number of items that can be recalled as a function of time, and the distribution of recall times by item.

Combinatorial Optimization

From the point of view of computer science, the problem of finding minima of Eq. (1) is part of the field of combinatorial optimization, a broad term for problems that require extremizing a function of a large number of discrete variables. The traveling salesman problem is a well known example, in which the variables are the choice of successive cities for the salesman to visit and the function is the total distance traveled. Like many such problems, this one is in the complexity class NP, meaning that while it is easy to check that a proposed route is short, it is hard to find short routes (see [4,12] for precise explanations). The problem of finding low energy configurations of Eq. (1) for generic J_{ij} (taking $S_i = \pm 1$) is also in NP (it is a special case of the "quadratic assignment problem").

These problems are also NP-hard, meaning that if we could solve them efficiently, we could solve any other problem in NP. This type of universality is central to computational complexity theory and suggests that ideas relevant for Eq. (1) have broader applicability. Indeed a straightforward generalization of Eq. (1) leads to a very central problem in computer science, the SAT (satisfaction) problem. Still, taking the variables $S_i = \pm 1$, we add higher order interactions, of the form

$$H_{SAT} = \sum_{i,j,k=1}^{n} F_{ijk}(S_i, S_j, S_k). \qquad (3)$$

Each F_{ijk} is chosen from one of the eight functions

$$F^{\pm\pm\pm}(a,b,c) = 1 - \frac{1}{8}(1 \pm a)(1 \pm b)(1 \pm c). \qquad (4)$$

The idea is that each choice $S_i = +1$ or -1 represents a Boolean variable x_i which can be true or false. The functions Eq. (4) represent logical clauses which are the AND of three terms, either x_i for $1 + S_i$ or NOT x_i for $1 - S_i$, and take the value 0 if the clause is true (satisfied) or 1 if false. Thus the value of Eq. (3) is the number of clauses which are not satisfied, and a configuration with $H_{SAT} = 0$ satisfies all of the clauses. More generally one could take clauses with k variables and thus k'th order interactions, to get the k-SAT problem. As physics intuition would suggest, $k = 2$ is easier, while the computational complexity class is the same for all $k \geq 3$.

One can show that any formula in propositional logic

can be rewritten in terms of such clauses, so this is a general formulation of the problem of solving systems of equations in Boolean logic. In fact, any algorithm can be reformulated in terms of such a system, so an efficient algorithm for solving it would solve a wide class of problems: 3-SAT is the prototypical NP-complete problem. It is generally believed that no such efficient algorithm exists, and the conjecture that P ≠ NP is one of the Clay Millennium Prize problems.

Of course the statement that problems in NP are difficult is a worst-case statement, and many instances are easy to solve. For example, if no NOT appears in the equations then we can simply take all the variables to be true. More generally, one expects these problems to be easy if there are very few clauses (most assignments work), easy if there are very many clauses (which will contradict each other), and harder in an intermediate case.

To make this precise, we define the random k-SAT problem in which the number of clauses and variables is fixed, but the specific set of clauses (or Hamiltonian Eq. (3)) is chosen by uniformly sampling among the possibilities. We can then ask for the probability that an instance is satisfiable, or the expected number of satisfying assignments.

Define $\alpha = n/N$ to be the number of clauses divided by the number of variables. In figure 10.4 we plot the fraction of satisfying assignments as a function of α for $k = 2, 3$, for various choices of N. As one increases α, one sees a steep drop for both k, and the N dependence on the $k = 3$ graph suggests that this might become sharp as $N \to \infty$. This intuition is supported by the second graph which plots the computational effort required to find a solution as a function of α—this peaks at a slightly larger value of α.

This is suggestive of a phase transition, and this idea was confirmed using the replica method by Parisi and collaborators [21]. The same approach can be applied to other random problems in combinatorial optimization and statistics, and such phase transitions are very common. A general idea which has emerged is that such problems often have two thresholds, one information theoretic (whether a solution exists) and one computational (whether a solution can be efficiently found), with both transitions becoming sharp in a suitable limit of large problem size. For a comprehensive overview see the recent survey [28].

Let us finish by coming full circle, and asking: What do these ideas have to say about the original problem Eq. (1)? Considered as a problem in combinatorial optimization, we are given a matrix of couplings J_{ij} , and our task is to find the configuration of spins $S_i \in \{\pm 1\}$ with the lowest possible energy, call it E_{min}. Not only is this hard, it is even hard to find approximate ground states with $H < (1 - \epsilon)E_{min}$[7].

But, this is for the worst case. Suppose we ask for an algorithm which works for most choices of J_{ij} , i.e. with high probability in the Gaussian distribution. Such an algorithm, called incremental approximate message passing (IAMP), was recently proposed by Montanari and collaborators [5, 22], based directly on the ultrametric structure of pure states. Without going into details (given in [6]), the algorithm descends the tree of approximate ground states in a step-wise fashion. They even proved it works, in time linear in the number of spins (for a given $\epsilon > 0$), under the hypothesis of complete replica symmetry breaking.

Such optimization problems are the foundations of ML, and such results give hope that the (so far) very empirical study of ML will soon be complemented by an equally rich and powerful theoretical framework. ♦

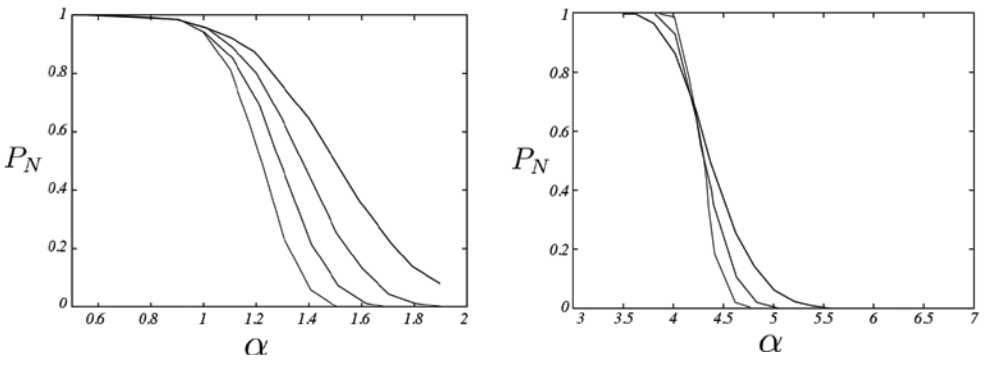

Figure 10.4
The probability that a formula generated from the random K-SAT ensemble is satisfied, plotted versus the clause density α. Left: $K = 2$; Right: $K = 3$.
Mezard and Montanari, Information, Physics, and Computation, Oxford University Press, 2009.

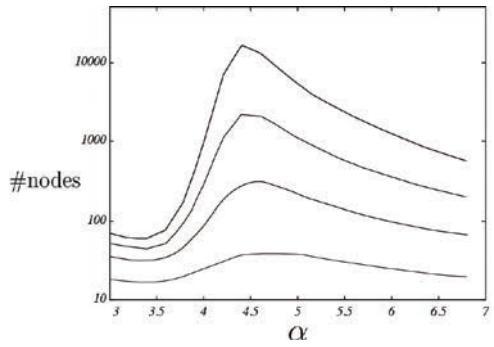

Fig 10.5 Computational effort of our DPLL algorithm applied to random 3-SAT formulae. The logarithm of the number of branching nodes was averaged over 10^4 instances. From bottom to top: N = 50, 100, 150, 200.
Mezard and Montanari, Information, Physics, and Computation, Oxford University Press, 2009

References

[1] Conference on Artificial Intelligence and Theorem Proving, http://aitp-conference.org/2021/
[2] Amit, Daniel J., Hanoch Gutfreund, and Haim Sompolinsky. "Spin-glass models of neural networks." Physical Review A 32, no. 2 (1985): 1007.
[3] Anderson, L.B., Gerdes, M., Gray, J., Krippendorf, S., Raghuram, N. and Ruehle, F., 2020. Moduli-dependent Calabi-Yau and SU (3)-structure metrics from Machine Learning. arXiv preprint arXiv:2012.04656.
[4] Arora, Sanjeev, and Boaz Barak. *Computational Complexity: A Modern Approach.* Cambridge University Press, 2009.
[5] Alaoui, A.E., Montanari, A. and Sellke, M., 2020. Optimization of mean-field spin glasses. arXiv:2001.00904 .
[6] Alaoui, A.E. and Montanari, A., 2020. Algorithmic Thresholds in Mean Field Spin Glasses. arXiv preprint arXiv:2009.11481.
[7] Arora, S., Berger, E., Elad, H., Kindler, G. and Safra, M., 2005, October. On non- approximability for quadratic programs. In 46th Annual IEEE Symposium on Founda- tions of Computer Science (FOCS'05) (pp. 206-215). IEEE.
[8] Ashmore, A., He, Y.H. and Ovrut, B.A., 2020. Machine Learning Calabi–Yau Metrics. Fortschritte der Physik, 68(9), p.2000068.
[9] Bryngelson, Joseph D., and Wolynes,Peter G. "Spin glasses and the statistical me- chanics of protein folding." Proceedings of the National Academy of sciences 84, no. 21 (1987): 7524-7528.
[10] Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L. and Zdeborová, L., 2019. Machine learning and the physical sciences. Reviews of Modern Physics, 91(4), p.045002.
[11] Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4), pp.303-314.
[12] Denef, Frederik, and Michael R. Douglas. "Computational complexity of the landscape: Part I." Annals of physics 322, no. 5 (2007): 1096-1142.
[13] Donaldson, S.K., 2005. Some numerical results in complex differential geometry. arXiv preprint math/0512625.
[14] https://cbmm.mit.edu/news-events/events/brains-minds-machines-seminar-series-how-will-we-do-mathematics-2030
[15] Douglas, M. R., Lakshminarasimhan, S. and Qi, Y. "Numerical Calabi-Yau metrics from holomorphic networks," to appear in the proceedings of MSML 2021, arXiv:2012.04797 .
[16]Weinan, E., 2020. Machine Learning and Computational Mathematics. arXiv:2009.14596.
[17] Han, J., Jentzen, A. and Weinan, E., 2018. Solving high-dimensional partial differen- tial equations using deep learning. Proceedings of the National Academy of Sciences, 115(34), pp.8505-8510.
[18] Hertz, John A. *Introduction to the Theory of Neural Computation.* CRC Press, 2018.
[19] Hopfield, John J. "Neural networks and physical systems with emergent collective computational abilities." Proceedings of the national academy of sciences 79, no. 8 (1982): 2554-2558.
[20] Mézard, Marc, Giorgio Parisi, and Miguel Angel Virasoro. *Spin Glass Theory and Beyond: An Introduction to the Replica Method and its Applications*. Vol. 9. World Scientific Publishing Company, 1987.
[21] Mézard, M., Parisi, G. and Zecchina, R., 2002. Analytic and algorithmic solution of random satisfiability problems. Science, 297(5582), pp.812-815.
[22] Montanari, Andrea. "Optimization of the Sherrington–Kirkpatrick Hamiltonian." SIAM Journal on Computing 0 (2021): FOCS19-1. arXiv:1812.10897 .
[23] Raissi, M. and Karniadakis, G.E., 2018. Hidden physics models: Machine learning of nonlinear partial differential equations. Journal of Computational Physics, 357, pp.125- 141.
[24] Recanatesi, S., Katkov, M., Romani, S. and Tsodyks, M., 2015. Neural network model of memory retrieval. Frontiers in computational neuroscience, 9, p.149.
[25] Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W., Bridgland, A. and Penedones, H., 2020. Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), pp.706-710.
[26] Stein, Daniel L., and Charles M. Newman. *Spin glasses and Complexity.* Vol. 4. Princeton University Press, 2013.
[27] Talagrand, Michel. "Rigorous results for the Hopfield model with many patterns." Probability theory and related fields 110, no. 2 (1998): 177-275.
[28] Zdeborová, L. and Krzakala, F., 2016. Statistical physics of inference: Thresholds and algorithms. Advances in Physics, 65(5), pp.453-552. arXiv:1511.02476