Statistical Mechanics Methods for Discovering Knowledge from Modern Production Quality Neural Networks

Charles H. Martin
Calculation Consulting
San Francisco, CA 94122
charles@CalculationConsulting.com

Michael W. Mahoney
ICSI and Department of Statistics
University of California at Berkeley
Berkeley, CA 94720
mmahoney@stat.berkeley.edu

ABSTRACT

There have long been connections between statistical mechanics and neural networks, but in recent decades these connections have withered. However, in light of recent failings of statistical learning theory and stochastic optimization theory to describe, even qualitatively, many properties of production-quality neural network models, researchers have revisited ideas from the statistical mechanics of neural networks. This tutorial will provide an overview of the area; it will go into detail on how connections with random matrix theory and heavy-tailed random matrix theory can lead to a practical phenomenological theory for large-scale deep neural networks; and it will describe future directions.

KEYWORDS

statistical mechanics, neural networks, random matrix theory, heavy-tailed random matrix theory

ACM Reference Format:
Charles H. Martin and Michael W. Mahoney. 2019. Statistical Mechanics Methods for Discovering Knowledge from Modern Production Quality Neural Networks. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332294

DESCRIPTION AND OUTLINE

While techniques from statistical mechanics have long had connections with neural networks [10, 18, 32], work in statistical learning theory in the last few decades has largely ignored it. Recent work has demonstrated, however, that traditional statistical learning theory and stochastic optimization theory often do not provide even a qualitative guide to the performance of practical deep neural networks. Motivated by this, very recent work has begun to revisit statistical mechanics approaches to learning.

The tutorial will provide an overview of these recent developments, aimed toward a typical conference attendee. In particular, it will bring to the members of the community an awareness of this alternate approach to learning and generalization, including its long history from the 1980s and before; highlight and summarize recent failings of statistical learning theory and stochastic optimization theory at explaining even qualitative properties of deep neural networks in computer vision and natural language processing; describe how one can use recent results in heavy tailed random matrix theory to construct a phenomenological theory of learning; use this theory to operationalize certain aspects of the statistical mechanics approach to learning and generalization, in order to make predictions for production-scale models; and highlight connections with other related works that in recent years have used techniques from statistical mechanics.

Historical Overview. We will set the context by providing a brief overview of early work on the connections between statistical mechanics and neural networks [10, 18, 32].

Review The Foundational Material. We will review the foundational material in the statistical mechanics approach to neural networks. For this, we will draw from [12, 14, 30, 34].

Differences with Traditional Learning Theory. We will describe differences between the traditional approaches to statistical learning theory and the statistical mechanics approach, including problem formulation, e.g., worst case uniform bounds versus typical average case predictions.

Failings of Traditional Approaches. We will describe recent work that has pointed to fundamental limitations of traditional statistical learning theory and traditional stochastic optimization theory, including the effect of adding noise to labels [36], changing batch size and step size [16], etc.

A Simple Statistical Mechanics Model. In light of the failings of popular machine learning approaches, we will describe a very simple model from statistical mechanics that qualitatively captures these phenomena [23]. This model is perhaps the simplest model that captures the observed phenomena, and it suggests that revisiting statistical mechanics approaches will be useful more generally.

Random Matrix Theory and Heavy-tailed Random Matrix Theory. We will describe developments in random matrix theory and heavy-tailed random matrix theory. This will include both from the perspective of physics and strongly-correlated system theory [9, 33] as well as mathematics and statistics [5, 6, 15]. We will describe how this approach via strongly-correlated systems is complementary to and yet a substantial departure from the original statistical mechanics approaches to neural networks. Importantly, while the empirically-observed heavy-tailed phenomena render certain aspects of the theory more difficult (e.g., since the concentration properties of the random variables are less good), they help with other aspects of the theory (e.g., many of the pathologies of Gaussian-based spin glasses can be avoided).

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332294

Theory of Implicit Heavy Tailed Self Regularization. We will outline the recent theory of implicit heavy tailed self regularization [24, 26], including the motivations, the large body of empirical results, and the known and suspected limitations. This theory uses ideas from the statistical mechanics of learning, but it does so in light of empirical results that are seen in modern state-of-the-art networks. An important part of this is using these methods to develop a theory that can make useful predictions for the practitioner. Thus, a focus will be on explaining how the theory can be used in practice for extracting insight from production-quality pre-trained models.

Extensions and Applications of the Theory. We will describe recent work on extending and applying the theory. This will start with describing novel statistical mechanics based capacity control metrics to predict trends in generalization accuracy for models such as those in the VGG and ResNet series [25], comparing and contrasting this with recent work using capacity control metrics such as the log Frobenius norm [17, 29]. This will also include a discussion of connections with recent work on large batch size [13, 31], energy landscape methods [7, 8, 11, 27, 35], phase transitions in related phenomenon [4, 28], etc.

GOALS AND TARGET

The goals of the tutorial are the following:

- to bring to the members of the community an awareness of this alternate approach to learning, generalization, and extracting insight from data;
- to provide an overview to members of the community of the basic ideas of the theory and how that theory is quite different than recently-popular techniques;
- to describe recent empirical results using that theory to make strong predictions about the generalization properties of production-scale models; and
- to describe fruitful future directions in which the theory can be used in practical large-scale data analysis settings.

There have been several recent related seminars [19–22], which have been based on our recent results on this topic [23–26]. The tutorial will cover more generally the basic motivation, theoretical and empirical results, and future directions at a level understandable by students and researchers attending the conference. It will also cover recent software repositories to reproduce published results (e.g., [1], based on [24, 26]; and [3], based on [25]) as well as the `WeightWatcher` package [2].

REFERENCES

[1] 2018. ImplicitSelfRegularization. https://github.com/CalculatedContent/ImplicitSelfRegularization.

[2] 2018. WeightWatcher. https://pypi.org/project/WeightWatcher/.

[3] 2019. PredictingTestAccuracies. https://github.com/CalculatedContent/PredictingTestAccuracies.

[4] M. Advani and S. Ganguli. 2016. *Statistical Mechanics of High-Dimensional Inference.* Technical Report Preprint: arXiv:1601.04650.

[5] G. Ben Arous and A. Guionnet. 2008. The spectrum of heavy tailed random matrices. *Communications in Mathematical Physics* 278, 3 (2008), 715–751.

[6] A. Auffinger and S. Tang. 2016. Extreme eigenvalues of sparse, heavy tailed random matrices. *Stochastic Processes and their Applications* 126, 11 (2016), 3310–3330.

[7] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina. 2016. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proc. Natl. Acad. Sci. USA* 113, 48 (2016), E7655–E7662.

[8] A. J. Ballard, R. Das, S. Martiniani, D. Mehta, L. Sagun, J. D. Stevenson, and D. J. Wales. 2017. Energy landscapes for machine learning. *Physical Chemistry Chemical Physics* 19 (2017), 12585–12603. Issue 20.

[9] J. P. Bouchaud and M. Potters. 2003. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management.* Cambridge University Press.

[10] J. D. Cowan. 1967. *Statistical Mechanics of Neural Networks.* Ft. Belvoir: Defense Technical Information Center.

[11] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. 2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Annual Advances in Neural Information Processing Systems 27: Proceedings of the 2014 Conference.* 2933–2941.

[12] A. Engel and C. P. L. Van den Broeck. 2001. *Statistical mechanics of learning.* Cambridge University Press, New York, NY, USA.

[13] N. Golmant, N. Vemuri, Z. Yao, V. Feinberg, A. Gholami, K. Rothauge, M. W. Mahoney, and J. Gonzalez. 2018. *On the Computational Inefficiency of Large Batch Sizes for Stochastic Gradient Descent.* Technical Report. Preprint: arXiv:1811.12941.

[14] D. Haussler, M. Kearns, H. S. Seung, and N. Tishby. 1996. Rigorous Learning Curve Bounds from Statistical Mechanics. *Machine Learning* 25, 2 (1996), 195–236.

[15] I. M. Johnstone. 2001. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* (2001), 295–327.

[16] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. 2016. *On large-batch training for deep learning: generalization gap and sharp minima.* Technical Report Preprint: arXiv:1609.04836.

[17] Q. Liao, B. Miranda, A. Banburski, J. Hidary, and T. Poggio. 2018. *A surprising linear relationship predicts test performance in deep networks.* Technical Report Preprint: arXiv:1807.09659.

[18] W. A. Little. 1974. The existence of persistent states in the brain. *Math. Biosci.* 19 (1974), 101–120.

[19] M. W. Mahoney. February 2019. Seminar at ACM SF-SIG. https://www.youtube.com/watch?v=2qF8TezRwS0.

[20] M. W. Mahoney. September 2018. Seminar at Simons Institute. https://simons.berkeley.edu/talks/9-24-mahoney-deep-learning.

[21] C. H. Martin. December 2018. Seminar at ICSI. https://www.youtube.com/watch?v=6Zgul4oygMc.

[22] C. H. Martin. June 2018. Seminar at LBNL. https://www.youtube.com/watch?v=_Ni5UDrVwYU.

[23] C. H. Martin and M. W. Mahoney. 2017. *Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior.* Technical Report Preprint: arXiv:1710.09553.

[24] C. H. Martin and M. W. Mahoney. 2018. *Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning.* Technical Report Preprint: arXiv:1810.01075.

[25] C. H. Martin and M. W. Mahoney. 2019. *Heavy-Tailed Universality Predicts Trends in Test Accuracies for Very Large Pre-Trained Deep Neural Networks.* Technical Report Preprint: arXiv:1901.08278.

[26] C. H. Martin and M. W. Mahoney. 2019. Traditional and Heavy-Tailed Self Regularization in Neural Network Models. In *Proceedings of the 36st International Conference on Machine Learning.*

[27] J. Pennington and Y. Bahri. 2017. Geometry of Neural Network Loss Surfaces via Random Matrix Theory. In *Proceedings of the 34th International Conference on Machine Learning.* 2798–2806.

[28] J. Pennington, S. S. Schoenholz, and S. Ganguli. 2018. *The Emergence of Spectral Universality in Deep Networks.* Technical Report Preprint: arXiv:1802.09979.

[29] T. Poggio, Q. Liao, B. Miranda, A. Banburski, X. Boix, and J. Hidary. 2018. *Theory IIIb: Generalization in Deep Networks.* Technical Report Preprint: arXiv:1806.11379.

[30] H. S. Seung, H. Sompolinsky, and N. Tishby. 1992. Statistical mechanics of learning from examples. *Physical Review A* 45, 8 (1992), 6056–6091.

[31] C. J. Shallue, J. Lee, J. Antognini, J. Sohl-Dickstein, R. Frostig, and G. E. Dahl. 2018. *Measuring the Effects of Data Parallelism on Neural Network Training.* Technical Report. Preprint: arXiv:1811.03600.

[32] H. Sompolinsky. 1988. Statistical Mechanics of Neural Networks. *Physics Today* 41, 12 (1988), 70–80.

[33] D. Sornette. 2006. *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools.* Springer-Verlag, Berlin.

[34] T. L. H. Watkin, A. Rau, and M. Biehl. 1993. The statistical mechanics of learning a rule. *Rev. Mod. Phys.* 65, 2 (1993), 499–556.

[35] Z. Yao, A. Gholami, Q. Lei, K. Keutzer, and M. W. Mahoney. 2018. *Hessian-based Analysis of Large Batch Training and Robustness to Adversaries.* Technical Report. Preprint: arXiv:1802.08241.

[36] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. 2016. *Understanding deep learning requires rethinking generalization.* Technical Report Preprint: arXiv:1611.03530.