Deep Learning through Information Geometry

Workshop Proposal to NeurIPS 2020

by

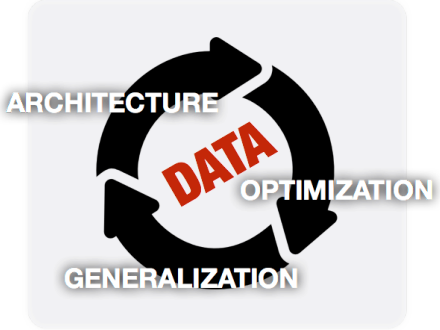
Alexander Alemi (Google LLC), Pratik Chaudhari (Univ. of Pennsylvania), Varun Jog (Univ. of Wisconsin-Madison), Dhagash Mehta (The Vanguard Group), Frank Nielsen (Sony Computer Science Laboratories Inc.), Stefano Soatto (Univ. of California Los Angeles & AWS), and Greg Ver Steeg (Univ. of Southern California)

Abstract

Attempts at understanding deep learning have come from different disciplines, namely physics, statistics, information theory, and machine learning. These lines of investigation have very different modeling assumptions and techniques; it is unclear how their results may be reconciled together. This workshop builds upon the observation that Information Geometry has strong overlaps with these directions and may serve as a means to develop a holistic understanding of deep learning. The workshop program is designed to answer two specific questions. The first question is: how do geometry of the hypothesis class and information-theoretic properties of optimization inform generalization. Good datasets have been a key propeller of the empirical success of deep networks. Our theoretical understanding of data is however poor. The second question the workshop will focus on is: how can we model data and use the understanding of data to improve optimization/generalization in the low-data regime.

1 Description

It has proven difficult to build a theoretical foundation that can sufficiently explain the strong empirical performance of deep networks and push practice to the next level. Three distinct—but interdependent—aspects lie at the heart of this challenge: (i) neural architectures are large and complex, (ii) training these non-convex models is difficult, and (iii) generalization performance of the trained models is difficult to guarantee. Attempts at understanding deep learning have focused on attacking each of these aspects independently, e.g., approximation theory to characterize the hypothesis class, building stochastic optimization theory tailored to deep networks, and adapting statistical learning theory to handle over-parametrized models. While substantial progress has been made on each of the individual aspects, a coherent understanding of deep learning remains elusive.



Our first goal is to bring together individual threads of theoretical work that address the three aspects of deep learning. This involves combining ideas from the diverse fields that have fueled these efforts, namely statistical physics [1, 2], applied mathematics and optimization [3–6] and information theory [7–11] along with statistical learning theory [12–16]. We aim for these different communities to gain appreciation of each other's results, learn each other's language and compare and contrast their results. For instance, generalization is typically studied in a teacher-student setting in statistical physics [17], this is somewhat aligned with the Bayesian

setting [18, 19] used in information theory but these results are quite different from uniform convergence, max-margin, or stability bounds in statistical learning theory [12, 20]. Optimization theory presents similar dissonant discourse [21–24]. Information Geometry [9] has strong overlaps with both these directions; it allows an explicit characterization of both the geometry [25] and information theoretic properties [10, 18] of the hypothesis class and leads to results that qualitatively fit empirical results in the modern literature [13]. It would therefore be fruitful to engage these communities via a unifying platform and seek a direction forward.

Our second observation is that while large real-world datasets have been instrumental in propelling empirical progress, current theories do not sufficiently exploit properties of the data. This is a large intellectual gap, and potentially the key to developing a holistic way to understand deep learning. Theoretical attempts to model the data have been quite successful [26, 27]. A deep, empirical understanding data has also been developed in the vision and NLP communities under the umbrella of transfer learning [28–30]. Bringing forth these insights and cataloging them systematically, vis-a-vis existing theory, will improve our understanding of deep learning, and machine learning in general.

Concretely, we wish to discuss the following questions at this workshop.

1. Statistical learning theory focuses on the complexity of the hypothesis class to bound the generalization gap, but it is clear that this approach won't work for deep networks. Nevertheless, empirically neural networks often exhibit good generalization properties.
   (a) How can we adapt existing theory to exploit the geometry of the hypothesis class?
   (b) The learning algorithm can be viewed as an information processing procedure. What information-theoretic properties of this channel lead to good generalization?
2. How should we build an understanding of data in machine learning? Specifically, how does the dataset (task) in deep learning affect optimization and generalization? How can we adapt learning theory to understand the low-data regime?

2 Similar workshops in the past

Our workshop will discuss topics that lie at the intersection of information theory and geometry but with the goal of making advances in understanding generalization of high-dimensional machine learning models and extrapolating these theories to low labeled-data regimes. The broader topic is aligned with the following two previously organized workshops.

1. A workshop on "Information Theory and Machine Learning" was organized at NeurIPS 2019 (https://sites.google.com/view/itml19/home). This workshop had a broader goal to identify common tools in information theory for representation learning, e.g., information bottleneck, fairness, privacy, generative models and variational inference, error-correcting codes. Our workshop seeks to sharpen the scope and ask (i) how existing ideas in Bayesian statistics and information geometry compare and contrast with results in statistical learning theory, and (ii) understanding the low-data regime (transfer/few-shot learning) through information theory.
2. A workshop on "Geometry in Machine Learning (GiMLi)" was organized at ICML 2018 (http://gimli.cc/2018/). This workshop drew out the connections of geometry in ML, e.g., geometric data, optimization/statistics on manifolds, information geometry and modeling invariances. Our workshop builds upon information geometry from GiMLi while focusing the discussion on generalization and characterizing the information theoretic properties of data.

There have been other workshops such as "Principled Approaches to Deep Learning" (ICML 2017, https://www.padl.ws/), "Theory of Deep Learning" (ICML 2018, https://sites.google.com/site/deeplearningtheory/), "Identifying and

Understanding Deep Learning Phenomena" (ICML 2019, http://deep-phenomena.org), and "Integration of Deep Learning Theories" (NeurIPS 2018, http://nips2018dltheory.rice.edu/) that have tackled topics aligned with our proposed program.

3 Invited speakers

The following speakers have accepted our invitation to deliver a keynote at the workshop.

1. **Shun-ichi Amari (Senior Advisor, RIKEN Brain Science Institute, Japan)**
   Website: https://scholar.google.com/citations?user=cH2eTqAAAAAJ&hl=en
   Prof. Amari is an authority on information geometry and neural networks. He brings a broad perspective on the issues tackled by this workshop, having played a key role across many decades of research on neural networks.

2. **Gintare Karolina Dziugaite (Fundamental Research Scientist, Element AI, Canada)**
   Email: karolina.dziugaite@gmail.com
   Website: https://gkdz.org, https://scholar.google.com/citations?user=5K1QB_8AAAAJ
   Gintare Karolina Dziugaite is a junior researcher (PhD, 2019). Her work on non-vacuous PAC-Bayes generalization bounds for deep networks has been influential.

3. **Marco Gori (Professor, University of Siena, Italy)**
   Email: marco.gori@unisi.it
   Website: https://scholar.google.com/citations?user=wBMRRk0AAAAJ& hl=en& oi=ao
   Prof. Gori is a leading researcher in pattern recognition and game playing. His work develops a unified view of learning and inference using symbolic representations of data as the bridge between the two.

4. **Alexander Rakhlin (Associate Professor, Massachusetts Institute of Technology, USA)**
   Email: rakhlin@mit.edu
   Website: http://www.mit.edu/∼ rakhlin/
   Prof. Rakhlin has done foundational work on statistical learning theory, optimization, and high-dimensional statistics, specific to deep networks. His results are complementary to the understanding of generalization in information geometry. The workshop seeks to develop a refined understanding as to where these two lines of investigation come together/bifurcate.

5. **Ke Sun (Data 61, CSIRO, Sydney, Australia)**
   Email: sunk@ieee.org
   Website: https://scholar.google.com/citations?user=n6AIt34AAAAJ& hl=en
   Ke Sun is a junior researcher (PhD, 2015) who has developed a large body of work on the information geometry of machine learning.

We are considering expanding the program to accommodate one more keynote speaker. This will potentially come at the cost of reducing keynotes to 40' (35' talk, 5' discussion) and contributed talks to 15' (13' talk, 2' discussion). We plan to reach out to Naftali Tishby (https://en.wikipedia.org/wiki/Naftali_Tishby) who has made seminal contributions to using information theory, statistical physics and neural networks.

4 Schedule and logistics

A tentative schedule of the workshop is as follows.

09.20 – 09.30 Opening Remarks

09.30 – 10.45	Keynote 1
10.45 – 11.05	Contributed Talk 1
11.05 – 11.25	Contributed Talk 2
11.25 – 12.10	Keynote 2
Break	
01.30 – 02.15	Keynote 3
02.15 – 03.00	Keynote 4
Break	
03.20 – 03.40	Contributed Talk 3
03.40 – 04.25	Keynote 5
04.25 – 05.00	Panel Discussion and Closing Remarks
05.00 – 06.00	Poster and Breakout Sessions

The entire workshop will be held remotely. The program will run (tentatively) in the PDT time-zone; this is to ensure that attendees across the globe can attend at least parts of the workshop in real-time and ask questions. Zoom will be used for keynotes and contributed talks; the Zoom session will be live-streamed and recorded to YouTube (this will also alleviate the pressure on Zoom for speakers/attendees spread across the world). The poster session will consist of 5' pre-recorded videos by the authors of accepted submissions with a dedicated Slack discussion channel for each poster that will run throughout the day.

Panel Discussion. The panel discussion forms a key part of the program. The panel will consist of all the speakers and will be moderated by the organizers. Questions to the panelists will be invited prior to the workshop in addition to extempore questions. The goal of the panel discussion is to engage the participants and the panelists in a dialogue to chart out concrete ways to make progress on the problems discussed in the workshop. Our goal as organizers has been to bring in keynote speakers that come with different points of views to engage in this dialogue.

Breakout sessions. There will be a breakout session for the attendees after the workshop concludes where they can communicate, with speakers/organizers who may join or amongst themselves, in an informal setting. This session seeks to replicate the coffee conversations that would happen at a regular in-person conference.

Reviewing Process. We expect about 25 accepted (4-page) submissions to the workshop. Each submission will be reviewed by at least 2 reviewers. The reviewing committee will consist of the 7 organizers and 5–10 other researchers. We will advertise the workshop within the affiliating institutions of the organizers and more broadly on social media to encourage a broad spectrum of submissions. Three submissions will be selected for 20' (18' talk, 2' discussion) contributed talks in the main program.

Outcome. The proceedings of the workshop will be consolidated into a report by the organizers. This report will also contain editorials from some of the speakers.

Sponsorship. The workshop will be held remotely and we do not anticipate funding needs beyond infrastructural setup for live-streaming. We are planning to have one best poster award (about $1,000) for which we will seek sponsorship from Amazon Web Services and The Vanguard Group.

5 How do we address diversity?

1. **Diversity of thought**. We have invited keynote speakers spanning computer science, statistics, statistical physics and information theory. These researchers have been working on different aspects of machine learning but often with a shared goal, e.g., information geometry and optimization, PAC-Bayes bounds,

or statistical learning theory all provide different ways of understanding generalization of deep networks. We seek to bring these voices together to make progress. The organizers are equally broad in their interests; their background draws from statistical physics, control and robotics, information theory, and pure mathematics.

2. **Diversity of gender, race, and seniority.** We recognize that keynote speakers at NeurIPS are an inspiration and role-models for the attendees. We have tried to maintain a balance in the speakers across gender and race. In addition to this, we believe that diversity in seniority is equally important because it strongly encourages junior attendees to enter the field; two of our invited speakers are junior researchers within 5 years of beginning their post-doctoral careers.

3. All the **speakers and organizers come from different institutions**, spread across the world, with few ongoing collaborations between them. This will translate to a global audience and potential future collaborations among the speakers/organizers and the attendees. The remote nature of the workshop further facilitates this.

We will link to the NeurIPS 2020 code of conduct from the workshop website. Proceedings will be closely moderated by the organizers to confirm to this code.

6 Timeline

July 31	Workshop acceptance notification
Sept 18	Submission date for workshop contributions, we will consider extending this deadline to be after the ICLR 2020 deadline in order to get fresh-off-the-press submissions
August	Review submissions
Oct 01	Notification of acceptance to the submissions
Oct–Nov	Setup infrastructure for the workshop
Dec 11–12	Workshop

All deadlines will be in AOE time to accommodate submissions from across the globe.

7 Organizers

1. **Alexander Amini (Google LLC)**
   Bio: Alex Alemi is a Senior Research Scientist as Google. Alex recieved his PhD in Physics from Cornell University working under Jim Sethna on theoretical condensed matter. He recieved his Master's in Physics also from Cornell and his Bachelor of Science in Physics from the California Institute of Technology. Alex's research interests lie at the intersection of deep learning and information theory, with a particular focus on trying to not only understand existing objectives and techniques in an information theoretical light but also trying to develop new objectives for representation learning.

   Organizational experience: He was on the Program Committee the Uncertainty in Deep Learning Workshop at ICML 2020.

   Email: alemi@google.com
   Google Scholar: https://scholar.google.com/citations?user=68hTs9wAAAAJ
   Webpage: https://alexalemi.com

2. **Pratik Chaudhari (University of Pennsylvania)**
   Bio: Pratik Chaudhari is an Assistant Professor in Electrical and Systems Engineering and Computer and Information Science at the University of Pennsylvania. Pratik received his PhD (2018) in Computer Science from the University of California Los Angeles; his Master's from the Massachusetts Institute of Technology and his Bachelor's from the Indian Institute of Technology Bombay. From 2018-19, he was a Senior Applied Scientist at Amazon Web Services and a Postdoctoral Scholar in Computing and Mathematical Sciences at the California Institute of Technology. He was a part of nuTonomy Inc. (now Aptiv) from 2014-16. Pratik's research interests lie in developing a theoretical understanding of deep learning. His main contributions in this area have provided an understanding of the energy landscape of deep networks using techniques from statistical physics, information theory and optimization.

   Organizational experience: He was on the program committee of the 2nd Conference on Learning for Dynamics and Control (L4DC) in 2020 (135 submissions, completely virtual) and co-organized the North-East Robotics Colloquium in 2019 (1-day event, 300 attendees, 45 submissions). In addition to these workshops, Pratik has taught two large courses at the University of Pennsylvania: the first one titled "Principles of Deep Learning" had 92 students and 5 TAs, the second one titled "Learning in Robotics" had 86 students and 5 TAs; material (including remote lectures and homework assignments) for both courses was created from scratch.

   Email: pratikac@seas.upenn.edu
   Google Scholar: https://scholar.google.com/citations?hl=en& user=c_z5hWEAAAAJ
   Webpage: https://pratikac.github.io/

3. **Varun Jog (University of Wisconsin-Madison)**
   Bio: Varun Jog is an Assistant Professor at the Electrical and Computer Engineering Department and a fellow at the Grainger Institute for Engineering at the University of Wisconsin-Madison. His research interests include information theory, convex geometry, and machine learning. He is a recipient of the NSF-CAREER Award (2020), the Eli Jury Award from the EECS Department at UC Berkeley (2015), and the Jack Keil Wolf student paper award at ISIT 2015. Varun Jog received his B.Tech. degree in Electrical Engineering from IIT Bombay in 2010, and his Ph.D. in Electrical Engineering and Computer Sciences (EECS) from UC Berkeley in 2015. He was a postdoctoral fellow at the University of Pennsylvania from 2015-2016.

   Organizational experience: He has been a Technical Program Committee member for the IEEE International Symposium on Information Theory 2018, 2019, and 2020.

   Email: vjog@wisc.edu

   Google Scholar: https://scholar.google.com/citations?hl=en& user=4XIn3jAAAAAJ

   Webpage: https://sites.google.com/wisc.edu/vjog

4. **Dhagash Mehta (The Vanguard Group)**

   Bio: Dhagash Mehta is a Principle Research Scientist at The Vanguard Group leading a research group on machine learning and investment strategies. Prior to joining Vanguard, he was a Senior Research Scientist at United Technologies Research Center (now Raytheon Research Center).

   After finishing his Part III Mathematical Tripos at the University of Cambridge, Dhagash pursued his Ph.D. between the University of Adelaide, Australia, and Imperial College London in Applied Math/Theoretical Physics areas. After a few postdoctoral positions at National University of Ireland, Syracuse University and The University of Cambridge, he held research assistant professor positions at North Carolina State University and then at the University of Notre Dame while successfully winning multiple NSF and DARPA research grants. Dhagash has published 55 journal articles and 25+ conference proceedings.

   Research Interests and Expertise: Dhagash's research expertise is in loss landscapes of deep learning, applications of algebraic geometry to machine learning, machine learning and non-convex methods for

portfolio optimization and financial services. He is also on the editorial advisory board at Journal for Financial Data Science.

Organizational Experience: Dhagash has co-organized two medium size conferences (Statistical Topology of Random Manifolds: Theory and Applications; Machine Learning Landscape) at the Abdus Salam International Center for Theoretical Physics Trieste (Italy), multiple mini-symposia at Society for Applied and Industrial Mathematics conferences and multiple sectionals at American Mathematical Society conferences.

Email: dhagashbmehta@gmail.com

Google Scholar: https://scholar.google.com/citations?hl=en& user=J7fyX_sAAAAJ

Webpage: https://www.linkedin.com/in/dhagash-mehta-ph-d-45000111a/

5. **Frank Nielsen (Sony Computer Science Laboratories Inc)**

   Bio: Frank Nielsen (PhD'96, HDR'06) is a senior researcher at Sony Computer Science Laboratories Inc, Tokyo, Japan. His research focuses on the geometric science of information with applications to machine learning, information theory and visual computing. He currently serves as an editor of the "Information Geometry" journal (Springer) and the electronic journal "Entropy" (MDPI). Frank Nielsen taught at Ecole Polytechnique (Palaiseau, France) visual computing (Charles River Media textbook), and high-performance computing for data science (Springer textbook).

   Organizational Experience: Frank co-organized, among others, the biannual conference "Geometric Science of Information" (GSI'13 to GSI'19, Springer LNCS proceedings) and the "Joint Structures and Common Foundation of Statistical Physics, Information Geometry and Inference for Learning", Les Houches (France), 2020. Heis a senior member of the IEEE and the ACM, and a ELLIS fellow.

   Email: Frank.Nielsen@acm.org

   Google Scholar: https://scholar.google.com/citations?hl=en& user=c-cuO9cAAAAJ

   Webpage: https://franknielsen.github.io

6. **Stefano Soatto (University of California Los Angeles and Amazon Web Services)**

   Bio: Dr. Stefano Soatto is Professor of Computer Science and Electrical Engineering, and Director of the UCLA Vision Lab, in the Henry Samueli School of Engineering and Applied Sciences at UCLA. He is also Director of Applied Science at Amazon AI - AWS.

   Dr. Soatto received his Ph.D. in Control and Dynamical Systems from the California Institute of Technology in 1996; he joined UCLA in 2000 after being Assistant and then Associate Professor of Electrical and Biomedical Engineering at Washington University, and Research Associate in Applied Sciences at Harvard University. Between 1995 and 1998 he was also Ricercatore in the Department of Mathematics and Computer Science at the University of Udine - Italy. He received his D.Ing. degree (highest honors) from the University of Padova- Italy in 1992. Dr. Soatto is the recipient of the David Marr Prize for work on Euclidean reconstruction and reprojection up to subgroups. He also received the Siemens Prize with the Outstanding Paper Award from the IEEE Computer Society for his work on optimal structure from motion. He received the National Science Foundation Career Award and the Okawa Foundation Grant. He was Associate Editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) and a Member of the Editorial Board of the International Journal of Computer Vision (IJCV) and Foundations and Trends in Computer Graphics and Vision, Journal of Mathematical Imaging and Vision, SIAM Imaging. He is a Fellow of the IEEE.

   Organizational Experience: Program Co-Chair of CVPR 2005, ICCV 2011, ICCV 2017, SIAM Imaging Conference (2015); Workshop Chair (CVPR 2015), workshops at ICCV 2007, ICCV 2009, tutorials, demos and other events since CDC 1994.

   Email: soatto@cs.ucla.edu

Google Scholar: https://scholar.google.com/citations?user=lH1PdF8AAAAJ

Webpage: http://web.cs.ucla.edu/s̃oatto

7. **Greg Ver Steeg (University of Southern California)**

   Bio: Dr. Ver Steeg is Research Professor of Computer Science at the University of Southern California. Dr. Ver Steeg's research explores practical methods for inferring meaningful structure in complex systems. This work draws on a diverse set of connections between information theory, machine learning, causal inference, and physics. His recent work has applied these tools to solving outstanding problems in diverse domains including biology, natural language, human behavior, and social networks.

   Academic background: Dr. Ver Steeg received his Ph.D. in Theoretical Physics from the California Institute of Technology in 2009, for work in quantum information theory. He started as a postdoc at the USC Information Sciences Institute working on complex networks and is now a Research Associate Professor in the USC CS department, doing work at the intersection of information theory, machine learning, and statistical physics. His work has been recognized with an AFOSR Young Investigator Award, IJCAI Early Career Spotlight, Amazon Research Award, and an Institute Achievement Award.

   Organizational experience: He has served on the program committee for several conferences including UAI, AISTATS, and ICML. He recently co-developed a course on representation learning for CS PhD students that was taught at USC in 2019. He has served as PI or co-PI for multiple teams working on DARPA funded projects.

   Email: gregv@isi.edu

   Google Scholar: https://scholar.google.com/citations?hl=en& user=goLucoIAAAAJ

   Webpage: https://www.isi.edu/people/gregv/about

References

[1] Carlo Baldassi, Federica Gerace, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Learning may need only a few bits of synaptic precision. *Physical Review E*, 93(5):052313, May 2016.

[2] Shun-ichi Amari, Tomoko Ozeki, Ryo Karakida, Yuki Yoshida, and Masato Okada. Dynamics of Learning in MLP: Natural Gradient and Singularity Revisited. *Neural Computation*, 30(1):1–33, January 2018.

[3] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[4] Stanley Osher and Ronald P. Fedkiw. Level Set Methods: An Overview and Some Recent Results. *Journal of Computational Physics*, 169(2):463–502, May 2001.

[5] Lenaic Chizat and Francis Bach. Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss. *arXiv:2002.04486 [cs, math, stat]*, June 2020.

[6] Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillame Carlier. Deep Relaxation: partial differential equations for optimizing deep neural networks. *Journal of Research in the Mathematical Sciences (RMS)*, 2017. [arXiv:1704.04932]; Short version in Proc. of the Workshop on Principled Approaches to Deep Learning, ICML, Aug 6-11, 2017; SIAM Conference on Analysis of Partial Differential Equations, Dec 9-12, 2017; Asilomar Conference on Signals, Systems and Computers, Oct 29-Nov 1, 2017; SIAM Conference of Imaging Sciences, Bologna, June 5-8, 2018.

[7] Naftali Tishby, Fernando C Pereira, and W Bialek. The information bottleneck method. In *Proceedings 37th Allerton Conference on Communication, Control, and Computing*, 1999.

[8] Alessandro Achille and Stefano Soatto. Emergence of Invariance and Disentanglement in Deep Representations. *arXiv:1706.01350 [cs, stat]*, June 2018.

[9] Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer Japan, Tokyo, 2016.

[10] Ke Sun and Frank Nielsen. Lightlike Neuromanifolds, Occam's Razor and Deep Learning. *arXiv:1905.11027 [cs, stat]*, February 2020.

[11] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. *arXiv preprint arXiv:1711.01530*, 2017.

[12] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. In G. Goos, J. Hartmanis, J. van Leeuwen, David Helmbold, and Bob Williamson, editors, *Computational Learning Theory*, volume 2111, pages 224–240. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[13] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. *arXiv:1812.11118 [cs, stat]*, September 2019.

[14] Tengyuan Liang and Alexander Rakhlin. Just Interpolate: Kernel "Ridgeless" Regression Can Generalize. *arXiv:1808.00387 [cs, math, stat]*, February 2019.

[15] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *arXiv:1903.08560 [cs, math, stat]*, November 2019.

[16] Xialiang Dou and Tengyuan Liang. Training Neural Networks as Learning Data-adaptive Kernels: Provable Representation and Approximation Benefits. *arXiv:1901.07114 [cs, math, stat]*, July 2019.

[17] A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, March 2001.

[18] Vijay Balasubramanian. Statistical Inference, Occam's Razor, and Statistical Mechanics on the Space of Probability Distributions. *Neural Computation*, 9(2):349–368, February 1997.

[19] David McAllester. A PAC-Bayesian Tutorial with A Dropout Bound. *arXiv:1307.2118 [cs]*, July 2013.

[20] V Vapnik and A Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory Probab. Appl.*, 16(2):264–280, January 1971.

[21] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8571–8580. Curran Associates, Inc., 2018.

[22] Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. *arXiv:1910.07833 [cs, math, stat]*, May 2020.

[23] Shun-ichi Amari, Hyeyoung Park, and Tomoko Ozeki. Singularities Affect Dynamics of Learning in Neuromanifolds. page 59.

[24] David Saad and Sara A. Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225–4243, October 1995.

[25] Sumio Watanabe. Algebraic Geometry and Statistical Learning Theory. page 296.

[26] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modelling the influence of data structure on learning in neural networks: The hidden manifold model. *arXiv:1909.11500 [cond-mat, stat]*, May 2020.

[27] Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A Group-Theoretic Framework for Data Augmentation. *arXiv:1907.10905 [cs, math, stat]*, February 2020.

[28] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018.

[29] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A Baseline for Few-Shot Image Classification. 2020. Proc. of International Conference of Learning and Representations (ICLR) [arXiv:1909.02729].

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.