SYLLABUS FOR MATH 689 SPECIAL TOPICS IN DEEP LEARNING: THEORY AND APPLICATIONS

Course Information

Instructor. Boris Hanin, Blocker-620B, bhanin@math.tamu.edu, 979-845-3261.

Lectures. TR 11:20am-12:35pm (room TBA).

Prerequisites. Working knowledge of linear algebra and probability.

Office Hours. 1-3pm on Wednesdays or by appointment in Blocker-620B.

Grade Composition. The final grade will have three components: final project (50%), paper summary (40%), and in-class participation (10%). The paper summary, due 10/04, is a written summary of article or collection or articles about neural networks. The instructor will suggest many possible articles, athough students are free to choose their own (in consultation with the instructor). The final project is an article that the student plans to submit to ICML 2019. The article should contain original research.

Grading Scale. The final letter grades will be assigned as follows: A $(88\% - 100\%)$, B $(76\% - 87\%)$, C $(64\% - 75\%)$, D $(52\% - 63\%)$, F $(0\% - 51\%)$.

Course Description. This course will give an introduction to both the theory and practice of deep learning. We will cover the practical and theoretical properties of various neural net architectures (fully connected, convolution, recurrent, etc), training neural nets (i.e. optimizers, regularization, backpropagation, learning rate vs. batch size etc), as well a survey of rigorous approaches from probability, theoretical physics, and approximation theory to understanding what neural nets are good for and why they work so well in practice.

The main practical outcome of this course is that every student will write a paper with the goal of submitting it to ICML 2019.

Learning Outcomes. This course will teach you the basic uses of neural networks. You will learn:

(1) the ideas behind and differences between popular neural net architectures: ConvNets, ResNets, RNNs, etc;
(2) some of the practical tricks and considerations for training a neural network: initialization, batch normalization, dropout, early stopping, learning rate decay, etc;
(3) what is theoretically known about the expressive power of neural networks;
(4) what is theoretically known about the loss surfaces of neural networks;
(5) what is theoretically known about neural networks at initialization;

Lecture Schedule. Please find below the lecture and project schedule.

Tues	08/28	Lecture 1	Course overview
Thurs	08/30	Lecture 2	Computational graphs
Tues	09/04	Lecture 3	Representational power of neural nets: [Cyb89], [Bar93]
Thurs	09/06	Lecture 4	Deep vs. shallow: [MPCB14, STR17]
Tues	09/11	Lecture 5	Questions from approximation theory
Thurs	09/13	Lecture 6	Training by backpropagation
Tues	09/18	Lecture 7	SGD practice: momentum, exploding gradients, early stopping
Thurs	09/20	Lecture 8	SGD: saddles [LSJR16], bounded memory [MT17]
Tues	09/25	Lecture 9	Loss surface for linear models: [BH89]
Thurs	09/27	Lecture 10	Loss surface for linear models: [Kaw16]
Tues	10/02	Lecture 11	Loss surface for 1 hidden layer models: [GM17]
Thurs	10/04	Lecture 12	Generalization: [ZBH$^+$16]
Thurs	10/04		**Paper Summary Due**
Tues	19/09	Lecture 13	ConvNets for machine vision
Thurs	10/11	Lecture 14	ResNets: [HZRS16]
Tues	10/16	Lecture 15	Neural nets at initialization: activations [HR18]
Thurs	10/18	Lecture 16	Neural nets at initialization: gradients [Han18]
Tues	10/23	Lecture 17	Neural nets for NLP: word embeddings [LM14, PSM14]
Thurs	10/25	Lecture 18	RNNs: LSTMs [HS97, HBF$^+$01], Seq2Seq [SVL14]
Tues	10/30	Lecture 19	Attention: [VSP$^+$17]
Thurs	11/01	Lecture 20	DL via mean field theory: [PLR$^+$16, RPK$^+$16, SGGSD16]
Tues	11/06	Lecture 21	DL via statistical field theory: [SPSD17]
Thurs	11/08	Lecture 22	Deep reinforcement learning
Tues	11/13	Lecture 23	Deep reinforcement learning
Thurs	11/15	Lecture 24	Deep reinforcement learning
Thurs	11/15		**Final Project Due**
Tues	10/20		No Class: Thanksgiving
Thurs	10/22		No Class: Thanksgiving
Tues	10/27		**Final Presentations**
Thurs	10/29		**Final Presentations**
Tues	11/04		**Final Presentations**

Americans with Disabilities Act (ADA). The Americans with Disabilities Act (ADA) is a federal anti-discrimination statute that provides comprehensive civil rights protection for persons with disabilities. Among other things, this legislation requires that all students with disabilities be guaranteed a learning environment that provides for reasonable accommodation of their disabilities. If you believe you have a disability requiring an accommodation, please contact Disability Services, currently located in the Disability Services building at the Student Services at White Creek complex on west campus or call 979-845-1637. For additional information, visit `http://disability.tamu.edu`.

Academic Integrity. Remember: "An Aggie does not lie, cheat, or steal, or tolerate those who do." For additional information please visit `http://aggiehonor.tamu.edu`.

References

[Bar93] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

[BH89] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[Cyb89] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[GM17] Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. In *Advances in Neural Information Processing Systems*, pages 3656–3666, 2017.

[Han18] Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? *arXiv preprint arXiv:1801.03744*, 2018.

[HBF+01] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[HR18] Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. *arXiv preprint arXiv:1803.01719*, 2018.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Kaw16] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[LM14] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

[LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016.

[MPCB14] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.

[MT17] Michal Moshkovitz and Naftali Tishby. Mixing complexity and its applications to neural networks. *arXiv preprint arXiv:1703.00729*, 2017.

[PLR+16] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[RPK+16] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. *arXiv preprint arXiv:1606.05336*, 2016.

[SGGSD16] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *arXiv preprint arXiv:1611.01232*, 2016.

[SPSD17] Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. A correspondence between random neural networks and statistical field theory. *arXiv preprint arXiv:1710.06570*, 2017.

[STR17] Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. *arXiv preprint arXiv:1711.02114*, 2017.

[SVL14] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

[ZBH+16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016.