CS4540/8803: Algorithmic Theory of Intelligence

Project Guidelines.

Instructions. Form teams of 1-3 students. Each team can choose one of the following options:

- Read a paper and provide an insightful explanation/simpler proof/generalization.
- Prove a conjecture. (This can be a small extension of the results of a paper; for example, arguing that the results hold in a more general setting, or arguing that a stronger result holds in a more specific setting.)
- Formulate a precise conjecture supported by experimental or theoretical evidence.

Possible topics. Below is a list of suggested papers; you may pick the topic from one of these, or suggest your own. Only one group can work on each topic, so you should reach out to reserve the topic if you have one you are particularly interested in. Additionally, please discuss your project idea with the TAs during office hours.

For some of the papers listed, we have included a summary and some sample questions to help you get started. If you choose these topics, you are not limited to these questions (and hopefully new ideas will come up as you read and understand the papers!)

1 Algorithms

1. Learning-augmented algorithms [MV22]. Using predictions from machine learning to the traditional algorithms can sometimes circumvent the worst-case analysis. Can you think of the learning augmented algorithms/ data structures [LLW22] that are provable to be beneficial when the predictions are good?
2. Algorithms for online convex minimization [HAK07]. In the online convex optimization problem, the algorithm is given a sequence of (possibly unrelated) convex cost functions, with the goal of coming up with a strategy that minimizes the total cost. What do these classic algorithms have in common, and how do they exploit the convexity of the cost functions?
3. Optimization under delayed feedback [VDL22]. In classical Bayesian optimization, the algorithm makes a guess and is immediately given information on how well the guess performed. This paper studies the case where the feedback comes after a random amount of time.

2 Neuroscience

1. k-cap as a model of neuron dynamics [RV23]. In the assembly model, the k neurons with the highest synaptic inputs fire at each time step. How does this function behave on different types of connectomes?
2. Biological random projection. The structure of the fly's olfactory system is well-known by neuroscientists. This paper [DSN17] gives evidence that their system implements a nearest-neighbor search algorithm.

3. Random graph models with fixed subgraph densities. Many graphs encountered in applications are sparse, but have dense communities. This paper [PV22] introduces a random graph model which results in many dense subgraphs.

4. Graph sparsification [SS08]. The goal of graph sparsification is to approximate a dense graph G with a sparse graph H, speeding up processing time while preserving key properties.

3 Machine Learning Theory

1. Hardness of agnostic learning. This paper [SSSS11] uses the idea of boosting to give a hardness result for learning kernel-based halfspaces.

2. Learning intersections of halfspaces [Vem10]. Faster algorithms or more general distributions?

3. Learning DNF [KS01].

4 Representation Learning

1. Clustering. Efficient clustering algorithms for high dimensional data. Some previous references: [KVV04, CEM+15].

2. Active learning is allowed to draw random unlabeled examples and ask for the labels of some of them. This paper [BBL06] analyzed the active learning algorithm.

3. Lifelong learning is the setting to learn a sequence of tasks in a stream while we maintain a set of features that help to learn future tasks more efficiently. The paper [CLV22] provides an efficient lifelong learning algorithm for learning halfspaces for logconcave distributions. Can you generalize to other data distribution? Can you generalize to more general function class, such as polynomials?

4. Meta-learning is to use the data from existing tasks for learning algorithms or representations that enable better performance on unseen tasks. [BKT19] gives the result for gradient-based meta-learning though online convex optimization.

5. Reconstruction-based self-supervised learning is the method that learns the representation by solving some reconstruction style pretext tasks, such as predict the masked word. [LLSZ21] gives some results based on conditional independence assumption.

6. Contrastive learning. This is a self-supervised learning method, that learns the representation by pushing similar examples closer while keeping different examples far apart. [HWGM21] provides a way to analyze contrastive learning by augmentation graph on data.

7. Data augmentation. This is a common technique for machine learning. [SMB22] provides some PAC learning results under transformation invariance.

8. Semi-supervised learning. Since labels are expensive, algorithms that use less labels are proposed. [FKKT21] proposes efficient algorithms for learning from coarse labels.

9. Mixup is a way of data augmentation, which uses linear interpolation to get new data. The paper [ZCLG23] uses feature-noise data model and shows that mixup training can effectively learn rare features.

10. Machine unlearning [NRSM21] is the process to let a machine learning model forget about some of its training data points without retraining from sketch. Can you think of efficient algorithms for machine unlearning?

5 Deep Learning Theory

1. When is (S)GD provably efficient for loss minimization? The paper [MBB18] shows the fast convergence of SGD for over-parameterized learning under some assumptions.

2. Implicit bias. When training classifiers with (S)GD, the predictor converges to the direction of max-margin solution, which is referred to as implicit bias. [SHN$^+$18] has proved implicit bias for separable data. Can you prove implicit bias in other settings or show a better convergence rate?

3. Benign overfitting. Seek explanations for generalization despite overfitting. [BLLT20] considers linear regression while [KCCG23] provides risk bounds for learning two-layer ReLU convolutional neural networks.

4. Generalization for interpolating algorithms [BHM18]. In practice learning methods that perfectly fit the training data perform well. When does this heuristic provably work?

5. Double descent [BHMM19, AP20] refers to the phenomenon that when we increase model size, performance first gets worse and then gets better. Can you provide some analysis to explain this phenomenon?

6. Convolution networks. Are convolutional nets the optimal solution for some optimization problem solved by evolution? [LZA20]

7. Self-attention. Transformers can learn functions in-context from examples, without updating their internal weights. Could the self-attention layer be performing a type of gradient descent[VONR$^+$23]?

8. Performative Prediction. In some applications, making predictions in online learning can influence the data distribution. This paper uses game theory to find a stable point in this setting[PZMDH20].

9. Neural Tangent Kernel [JGH18] is a kernel that can be used in the analysis of deep learning. What are its strengths and limitations?

References

[AP20] Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33:11022–11032, 2020.

[BBL06] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 65–72, 2006.

[BHM18] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[BKT19] Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pages 424–433. PMLR, 2019.

[BLLT20] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[CEM+15] Michael B Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 163–172, 2015.

[CLV22] Xinyuan Cao, Weiyang Liu, and Santosh Vempala. Provable lifelong learning of representations. In *International Conference on Artificial Intelligence and Statistics*, pages 6334–6356. PMLR, 2022.

[DSN17] Sanjoy Dasgupta, Charles F Stevens, and Saket Navlakha. A neural algorithm for a fundamental computing problem. *Science*, 358(6364):793–796, 2017.

[FKKT21] Dimitris Fotakis, Alkis Kalavasis, Vasilis Kontonis, and Christos Tzamos. Efficient algorithms for learning from coarse labels. In *Conference on Learning Theory*, pages 2060–2079. PMLR, 2021.

[HAK07] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2007.

[HWGM21] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[KCCG23] Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting for two-layer relu networks. *arXiv preprint arXiv:2303.04145*, 2023.

[KS01] Adam R Klivans and Rocco Servedio. Learning dnf in time. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 258–265, 2001.

[KVV04] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515, 2004.

[LLSZ21] Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

[LLW22] Honghao Lin, Tian Luo, and David Woodruff. Learning augmented binary search trees. In *International Conference on Machine Learning*, pages 13431–13440. PMLR, 2022.

[LZA20] Zhiyuan Li, Yi Zhang, and Sanjeev Arora. Why are convolutional nets more sample-efficient than fully-connected nets? *arXiv preprint arXiv:2010.08515*, 2020.

[MBB18] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

[MV22] Michael Mitzenmacher and Sergei Vassilvitskii. Algorithms with predictions. *Communications of the ACM*, 65(7):33–35, 2022.

[NRSM21] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pages 931–962. PMLR, 2021.

[PV22] Samantha Petti and Santosh S Vempala. Approximating sparse graphs: The random overlapping communities model. *Random Structures & Algorithms*, 61(4):844–908, 2022.

[PZMDH20] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[RV23] Mirabel E Reid and Santosh S Vempala. The k-cap process on geometric random graphs. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3469–3509. PMLR, 2023.

[SHN+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[SMB22] Han Shao, Omar Montasser, and Avrim Blum. A theory of pac learnability under transformation invariances. *Advances in Neural Information Processing Systems*, 35:13989–14001, 2022.

[SS08] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 563–568, 2008.

[SSSS11] Shai Shalev-Shwartz, Ohad Shamir, and Karthik Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM Journal on Computing*, 40(6):1623–1646, 2011.

[VDL22] Arun Verma, Zhongxiang Dai, and Bryan Kian Hsiang Low. Bayesian optimization under stochastic delayed feedback. In *International Conference on Machine Learning*, pages 22145–22167. PMLR, 2022.

[Vem10] Santosh S Vempala. Learning convex concepts from gaussian distributions with pca. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 124–130. IEEE, 2010.

[VONR+23] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[ZCLG23] Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. The benefits of mixup for feature learning. *arXiv preprint arXiv:2303.08433*, 2023.