Research Statement

Kevin Zhou

My research primarily concerns the interactions between mathematical logic, in particular model theory, and aspects of computer science related to machine learning and artificial intelligence. My Ph.D. thesis (available at https://kevinzhou96.github.io/thesis.pdf) consists of two projects: new results in query learning of automata, and weighted first-order model counting.

1. Query Learning of Automata

This project appears in my thesis and will be published in the proceedings of the 22nd International Symposium on Automated Technology for Verification and Analysis (ATVA 2024). My results center around using a new method for proving upper bounds on the number of equivalence and membership queries needed to learn in the setting of query learning in order to find new upper bounds for the query complexity of two generalization of deterministic finite automata: advice DFAs and nominal DFAs.

Query learning, sometimes known as *active learning*, is a setting of machine learning in which the learner attempts to learn an unknown target function by interactively posing queries to an oracle. Two common queries considered are equivalence queries and membership queries. In an equivalence query, the learner submits a function as a hypothesis, succeeding if the hypothesis is equal to the target, and to which the oracle responds with an input for which the hypothesis and target disagree otherwise. In a membership query, the learner submits an element of the domain and the oracle responds with the value of the target on that domain element.

Learning various forms of automata using queries is a long-studied field with many applications, including in automatic verification and model checking. It was initiated by Angluin in 1987 with the introduction of the L^* algorithm which learns regular languages using equivalence and membership queries, with the number of queries needed being upper bounded by a polynomial in the size of the minimal DFA recognizing the language as well as the length of the longest counterexample given to an equivalence query [1]. Angluin's algorithm has been adapated to various other settings, such as tree automata [12], nondeterministic finite automata [5], and ω-automata [2]. Recently, Chase and Freitag applied ideas from model theory to develop a characterization of query learnability in terms of finite Littlestone and consistency dimensions, as well as an polynomial upper bound on query complexity based on those quantities [7]. This reflects similar equivalences such as that between PAC-learnabilty and VC-dimension and between online learnability and finite Littlestone dimension. In recent years, these equivalences have given rise to many fruitful interactions between computational learning theory and model theory, starting with Laskowski's observation in 1992 that finite VC-dimension corresponds to the model-theoretic notion of NIP [9] and more recently with Chase and Freitag's observation that finite Littlestone dimension corresponds to model-theoretic stability [6].

My work involves applying Chase and Freitag's bounds to advice DFAs and nominal DFAs. Advice DFAs [8] generalize classical DFAs with the addition of a fixed infinite-length advice string that the automaton reads in parallel with the input string, and are useful in modeling situations where the behavior of the automaton changes in a fixed way over time. By proving upper bounds on the Littlestone dimension and consistency dimension of the class of languages recognized by advice

DFAs, I gave the first known query learning bound for advice DFAs. A full statement of the result is as follows: let $\mathcal{L}_k^{\text{adv}}(n, m)$ be the set of languages over an alphabet of size k recognized by an advice DFA on at most n states, restricted to strings of length at most m.

Theorem 1. *The (EQ+MQ)-query complexity of $\mathcal{L}_k^{adv}(n, m)$ with queries from $\mathcal{L}_k^{adv}(2n, m)$ is $O(n^3 mk \log n)$.*

Nominal DFAs, first introduced by Bojańczyk, Klin, and Lasota [4], generalize classical DFAs to infinite alphabets and state sets. This is useful in settings where there are infinitely many options for data values, such as in XML documents (where arbitrary strings can appear as attribute values) or in software verification (in order to deal with pointers or arbitrary function parameters). However, this generalization is highly nontrivial, since simply relaxing the "finite" restrictions in DFAs to "infinite" results in cardinality issues which make computation intractable. To remedy this, nominal DFAs impose constraints based on invariance under symmetries that reflect real-world operations such as comparing data values for equality or under some linear order. As with nominal DFAs, I prove an upper bound on the query complexity for nominal DFAs via Littlestone and consistency dimensions. A full statement of the result requires several technical definitions, but is summarized as follows: given a G-alphabet A, let $\mathcal{L}_A^{\text{nom}}(n, k)$ denote the set of G-languages recognized by a nominal DFA whose state set has at most n orbits and has dimension at most k.

Theorem 2. *For a fixed G-alphabet A, the (EQ+MQ)-query complexity of $\mathcal{L}_A^{nom}(n, k)$ with queries from $\mathcal{L}_A^{nom}(n, k)$ is at most $\frac{n^{O(k)}}{k^k}$.*

Previous bounds on query complexity of nominal DFAs were obtained by Moerman et al. [11], who developed a nominal version of Angluin's L^* algorithm. My result improves on this previous bound with a better asymptotic dependence on n as well as removing dependence on the length of the longest counterexample given to an equivalence query.

2. Weighted First-Order Model Counting

This project appears in my thesis, and studies interactions between model-theoretic combinatorics and topics related to statistical relational learning.

The classic boolean SAT problem asks to determine whether or not there is an assignment of Boolean variables that satisfies a given Boolean formula. It is a canonical NP-complete problem with wide-ranging applications. The counting version of this problem, #SAT, asks how many distinct assignments satisfy the formula, and is the starting point for various model counting problems. It can be generalized by changing the underlying logic to first-order logic, as well as by giving a method for assigning a weight to each structure and computing the weighted sum. These generalizations form the basis for the weighted first-order model counting problem. Most existing work in this area has focused on determining the computational complexity of computing the weighted first-order model count for sentences falling into various fragments of first-order logic.

Weighted first-order model counting is closely related to problems in *statistical relational learning*, in which one aims to model probabilistic relationships between objects which have a rich interconnected relational structure. Such problems occur in real-world knowledge bases which contain millions or billions of rows of relational data, for which conducting probabilistic inference is an intensive computational task. Weighted model counting provides a flexible framework for encoding probabilistic queries and inference in such settings.

Independently of work on the weighted setting, the unweighted version of the first-order model

counting problem has been extensively studied in the combinatorics literature. In particular, much work has focused on understanding the asymptotic growth rates of *hereditary properties* of various combinatorial objects, which are classes of objects that are defined by a universal theory. A typical result proves a "jump" in the possible asymptotic growth rates, that is, showing that the number of objects of size n must either be at most $f(n)$ or at least $g(n)$, where $f(n)$ has strictly slower asymptotic growth rate than $g(n)$. The most general version of such a result was given by Laskowski and Terry, who prove a complete classification of the jumps in growth rates of hereditary properties of $\mathcal{L}$-structures for any finite relational language $\mathcal{L}$ [10].

My work primarily involves unifying these two perspectives. Since the results on hereditary properties often involve strong structural characterizations of classes falling into the various growth rates, one may hope that this can shed some light on the computational aspects of the weighted model counting problem. In the other direction, restricting to the fragments of first-order logic studied in weighted model counting may yield stronger classification results for unweighted model counting.

As an example of a result in the first direction, I showed that there is an explicit formula for the weighted model count of sentences whose unweighted model count falls into the slowest possible growth rate. While the exact statement of the theorem is fairly technical; however, it can be summarized as follows:

Theorem 3 (informal)**.** *Let ϕ be a first-order sentence in a finite relational language whose unweighted model count is bounded by an exponential in the size of the domain. Then there are parameters $t, K, c_1, \ldots, c_t \in \mathbb{N}$ depending oonly on ϕ such that there is an exact formula for the weighted model count of ϕ depending only on the parameters and the size of the domain.*

In the opposite direction, one heavily-studied fragment of first-order logic is $\mathbf{FO}^2$, in which formulas are only allowed to use at most two logical variables. The weighted model counting problem for $\mathbf{FO}^2$ has been shown to be computable in polynomial time [13, 14, 3]. By analyzing their proof, I derived a sharper dichotomy for the unweighted model count of $\mathbf{FO}^2$ sentences, summarized as follows:

Theorem 4. *Let ϕ be a universal $\mathbf{FO}^2$ sentence. Then the unweighted model count of ϕ is either upper bounded by an exponential function or lower bounded by a function of the form 2^{Cn^2}, where n is the size of the domain and C is a constant.*

This stands in contrast to Laskowski and Terry's classification result, which has four possible growth rates, the aforementioned two being the slowest and fastest possible.

3. Future Work

There is still much work to be done in both of these areas. In query learning of automata, it is worth seeing if Chase and Freitag's method of computing bounds for Littlestone and consistency dimension can be applied to other settings of automata and how those results compare to other approaches such generalizations of the L^* algorithm. In weighted model counting, there is still opportunity to understand how other parts of the unweighted growth classification can yield computational results for the weighted model counting problem. Additionally, there are extensions of $\mathbf{FO}^2$, such as those allowing for counting quantifiers or a linear order axiom, whose weighted model count problem have also been shown to be computable in polynomial time. It would be of interest to see how these fragments also relate to the classification of growth rates for the unweighted model counting problem.

References

[1] Dana Angluin. Learning regular sets from queries and counterexamples. *Inf. Comput.*, 75:87–106, 1987.

[2] Dana Angluin and Dana Fisman. Learning regular omega languages. *Theoretical Computer Science*, 650:57–72, 2016. Algorithmic Learning Theory.

[3] Paul Beame, Guy Van den Broeck, Eric Gribkoff, and Dan Suciu. Symmetric weighted first-order model counting. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '15, page 313–328, New York, NY, USA, 2015. Association for Computing Machinery.

[4] Mikołaj Bojańczyk, Bartek Klin, and Sławomir Lasota. Automata theory in nominal sets. *Logical Methods in Computer Science*, Volume 10, Issue 3, August 2014.

[5] Benedikt Bollig, Peter Habermehl, Carsten Kern, and Martin Leucker. Angluin-style learning of nfa. In *IJCAI*, pages 1004–1009, 07 2009.

[6] Hunter Chase and James Freitag. Model theory and machine learning. *The Bulletin of Symbolic Logic*, 25(3):319–332, 2019.

[7] Hunter Chase and James Freitag. Bounds in query learning. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1142–1160. PMLR, 2020.

[8] Alex Kruckman, Sasha Rubin, John Sheridan, and Ben Zax. A myhill-nerode theorem for automata with advice. *Electronic Proceedings in Theoretical Computer Science*, 96, 10 2012.

[9] Michael C. Laskowski. Vapnik-chervonenkis classes of definable sets. *Journal of The London Mathematical Society-second Series*, 45:377–384, 1992.

[10] Michael C. Laskowski and Caroline Terry. Jumps in speeds of hereditary properties in finite relational languages. *Journal of Combinatorial Theory, Series B*, 2022.

[11] Joshua Moerman, Matteo Sammartino, Alexandra Silva, Bartek Klin, and Michał Szynwelski. Learning nominal automata. In *Proceedings of the 44th ACM SIGPLAN Symposium on Principles of Programming Languages*, POPL 2017, page 613–625, New York, NY, USA, 2017. Association for Computing Machinery.

[12] Yasubumi Sakakibara. Learning context-free grammars from structural data in polynomial time. In *Proceedings of the First Annual Workshop on Computational Learning Theory*, COLT '88, page 330–344, San Francisco, CA, USA, 1988. Morgan Kaufmann Publishers Inc.

[13] Guy Van den Broeck. On the completeness of first-order knowledge compilation for lifted probabilistic inference. In *Neural Information Processing Systems*, 2011.

[14] Guy Van den Broeck, Wannes Meert, and Adnan Darwiche. Skolemization for weighted first-order model counting. In *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'14, page 111–120. AAAI Press, 2014.