View Reviews

Paper ID
6027

Paper Title
Average-case Complexity of Teaching Convex Polytopes via Halfspace Queries

Reviewer #1

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
In this paper the authors study the teaching complexity of learning polytopes in the average case. So in learning theory, teaching complexity is the number of labelled examples provided by a helpful teacher to learner to help the learning process. In this paper, they look at what is the minimum number of such helpful labelled examples are sufficient in order to decide if a target point (which is the input to a learner) is within a polytope or outside. Moreover, interestingly, they also consider the *average* case complexity of this problem, i.e., suppose a point is picked at random then how many teaching queries suffice. Their main result is in the average case, to learn a intersectionn polytyope in R^d, it takes Theta(d) queries, whereas in the worst case it'd have taken Omega(n) queries. They also present other results in this paper when the arrangement of the halfspaces for the polytope are different but I think the above is their main contribution.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
Strengths of this paper: I think in recent times understanding teaching complexity is becoming relevant given that in machine learning it is natural to consider not-necessarily-adverserial examples like in PAC learning and so on. From a theoretical point of view, I think learning the intersection of halfspaces/polytopes is interesting, but I'm also not overly convinced (as I mentioned below) of its ML relevance for NeurIPS. The final result is also surprising and neat, and even the proofs have nice technical contributions.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
Weaknesses of this work: 1) I feel there should be more motivation as to why learning polytopes is important for NeurIPS. I understand there have been a few works in the past on PAC or agnostic learning intersection of halfspaces/polytopes and so on, but they were more elegant mathematically and more TCS motivated. Whereas this paper they say it is ML motivated but I feel I am not convinced if there is a strong ML motivation.
2) I feel the paper is written in a overly complicated way. I mention a few remarks below. The notation can be considerably improved to aid a reader instead of make it look unnecessarily technical.
3) In general, it is not surprising that this is the correct notion of average case analysis. In learning, I guess smoothed analysis is the well-established notion of average case, so it'd be nice if the authors comment about this.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
As far as I checked, the math in the proof seemed right (but I didn't check all the details in the supplementary material)

5. Clarity: Is the paper well written?
It is written in a slightly dense fashion (but I think that's just the abstract). The main paper (which might be the supplementary material) is written better. If this paper gets accepted, I encourage the authors to make significant effort in improving the abstract and making it more accessible to a reader.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?

Unfortunately, there is not much prior work discussed in this paper. I feel there should be a much better discussion of prior works in the introduction.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:
1) Firstly, there have been works done by O'Donnell, Servedio, Tan (STOC'19) which says we can learn polytopes efficiently in PAC/agnostic model, how does that compare to your work? There have also been many prior works to theirs.

2) In line 54, can you motivate your average case better? I feel the right notion of average case should be in some form of smoothed analysis setting where there are perturbations to the examples instead of looking at uniformly random examples. Although your model is interesting, I'd like more justification

3) Line 73, do you want eta . z <= b? instead of strict equality? I'm not sure if its a convex set if you have equality.

4) Lines 73-89, I feel the notation is overly complicated. Any polytope is defined by a m by n matrix W and n-dimensional b. Think of each row as specifying a constraint and the polytope lives in R^n, so we want Wx <=b. And f or your labelling function f is 1 if x satisfies Wx <=b. Am i correct? The subscripts n and d can be implicit in the dimensions of W and need not be explicitly written everywhere

5) Can you comment on the general teaching dimension using your results? I guess recursive teaching dimension is a minimization problem and should be a lower bound to your "average* case analysis. It'd be nice to comment on that

9. Please provide an "overall score" for this submission.
7: A good submission; accept.

10. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes

Reviewer #2

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
Update following the response:
I agree with the response.

The paper presents average case bounds on the teaching complexity of a linear classifier: assume that the universe contains n points in R^d, with unknown labels, satisfying linear separability, and the goal is to learn a linear classifier that classifies *all* the points correctly. This papers considers the teaching complexity, namely, the number of specially selected points that a teacher has to send the learner in order to identify the classifier *exactly*. Namely, the teacher has access to the “true” classifier, and the “teaching set” that she sends to the learner has to uniquely identify the classifications on all n points: there is only one classifier consistent with the

teaching set.
For a worst-case classifier, Omega(n) points are necessary. However, the paper shows that if the classifier is selected uniformly at random (among all linear classifiers), then on average a teaching set of size O(d) suffices (assuming that the points lie in general position).
There are other settings discussed in this paper:
(1) learning polytopes via halfspace queries: here there are n hyperplanes in R^d that split the space into multiple connected components, each being a polytope. Given an unknown polytope, the goal of the learner is to learn it. Each query is a halfspace, specified by one of the n hyperplanes. This setting is dual to learning classifiers, and the same results apply.
(2) Learning \phi-dichotomies: this is the setting where each point has a representation in some R^{d_\phi}, and the goal is to learn a classifier in R^{d_\phi}. This is clearly equivalent to the setting of learning a linear separator.
(3) Active learning: In this setting, O(d\log n) queries are sufficient on average, while Omega(n) in the worst case.
(4) The authors suggest the notion of “relaxed general position”, which seems to be more restrictive than general position. In this setting, there are improved bounds.
** Notice that all upper and lower bounds are on teaching by the worst-case ERM. Namely, the training set has to uniquely identify the classifier. If we were to allow any reconstruction, then a teaching dimension of d would suffice: indeed, the teacher can provide the d support vectors to the learner, and the reconstruction is by a maximal margin classifier.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
The claims are sound and I believe them to be correct, and the statements seem to be novel. In particular, the paper presents a comprehensive study with tight upper and lower bounds. Since the task of learning *exactly* by a worst-case ERM is less studied, this paper poses a significant contribution. The proofs are neat and elegant.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
It is not clear to me how relevant is the theoretical framework, as explained below:
(1) No motivation was given to teaching complexity, and it is not clear to me how it can be applied.
(2) The goal in this paper is to learn by a worst-case ERM. However, if the learner learns by a maximal margin classifier, it suffices for the teacher to send only d samples (i.e. the support vectors).
(3) No motivation was given for the notion of “relaxed general position” that is highlighted in the paper. In particular, it does not seem likely for machine learning data to satisfy this condition.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
Yes

5. Clarity: Is the paper well written?
The paper is written formally and clearly. However, it involves multiple technical definitions of convex geometry, and may not be easy to read for some of the machine learning audience. It is worth changing the terminology to be more aligned with the ML community, and perhaps move the focus towards linear classification (or, the equivalent \phi-dichotomy) as it is more fundamental than learning a polygon.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
Some references to relevant prior work are missing. In particular, a comparison with classical learning models where the goal is to produce a good classifier based on the data (say, a classifier with zero loss), rather than show that any classifier consistent with the data works well. Under these settings, there exists a teaching set of size d (as I discussed above), and there are also related works on active learning with sample size poly(d,log(n)) (it would be nice to compare those to your results).

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:
Relation to prior work:
Some references to relevant prior work are missing. In particular, a comparison with classical learning models

where the goal is to produce a good classifier based on the data (say, a classifier with zero loss), rather than show that any classifier consistent with the data works well. Under these settings, there exists a teaching set of size d (as I discussed above), and there are also related works on active learning with sample size poly(d,log(n)) (it would be nice to compare those to your results).

9. Please provide an "overall score" for this submission.
6: Marginally above the acceptance threshold.

10. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes

Reviewer #3

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
The paper considers the class of regions induced by hyperplanes in R^d. One way to think of these is as the class where the domain corresponds to n hyperplanes in d dimensions and the hypothesis class corresponds to the connected components of the component of the planes with the labelling given by whether the region is on the positive or negative side of the hyperplane.

The paper first studies the teaching complexity of the class i.e. number of labelled data points needed to uniquely identify a function in the class. In the worst case, this can grow with n. The authors define a notion called d'-relaxed general position, which is a generalization of the usual notion of general position, and show that the average case complexity in this setting is O(d'). The proof follows by counting the number for (d-1)-faces and regions induced by the hyperplane arrangement. The authors also contrast this with the learning complexity where a learner queries points for the label and show that learning complexity is O(d log n) in the average case.

The authors also use techniques from the hyperplane arrangement setting to the phi-separable dichotomy settings to get bounds on the teaching complexity.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
The problem considered by the authors is interesting. The average case complexity of various tasks in machine learning as opposed to the worst case and showing gaps between the two is an interesting direction. The authors relate the teaching complexity to the combinatorics of hyperplane arrangements, and use results from the area to bound the average case complexity. The authors also use the results from the hyperplane arrangements to more general setting of dichotomies.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
Many of the results in the paper follow very easily from previous work. Let us consider counting the number of facets and regions in what the authors call d'-relaxed general position. One proof is considering the essentialization [standard notion as defined in An Introduction to Hyperplane Arrangements by R. Stanley] of the arrangement and noting that the essentialization is d' dimensional, and is in general position. Now the counting formula follows from Zaslavski's formula used for general position. The authors' proof is actually similar in spirit to the one sketched

above and can be seen as writing the argument above from first principles. Another simple proof is by noting that the inclusion lattice for an arrangement in d'-general position is the d' truncated Boolean lattice and same arguments used to show counting formula for the general position now give the formulas that the authors claim. Thus, I would be surprised if the theorems claimed by the authors is not known to experts in the field in some form.

Some of the presentation of the results in comparison to the previous work could be inproved. For example from the abstract, "As our main result, we show that the average-case teaching complexity is Θ(d), which is in sharp contrast to the worst-case teaching complexity of Θ(n)". As the author too note later, the O(d) bound for general hyperplane arrangements already follows from the work of Fukuda et al. So the new statement here is the lower bound, which should be explicitly clarified. Lower bound also seems to be only for arrangements in d' general position, which as noted earlier seems to follow easily from previous work.

Minor correction: Should the RHS of the equation A.1 not have a factor of 2 since each face gets counts for two regions to which it is adjacent?

The upper and lower bounds in the iid case also need to be explained more clearly.

Another concern is the presentation. At many points the authors cite lemmas from previous work without explicitly writing said lemmas. It can sometimes be difficult for the reader. Some of the statements of some lemmas and definition (for example definition 1 and Lemma 9) are slightly unclear and could cause confusion to the reader. It would greatly help the reader if the presentation of the paper is reworked.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
Up to my reading all the claims made in the paper are correct. But also see the comments in the weakness section for a more in depth account of the issues.

5. Clarity: Is the paper well written?
The presentation of the paper needs improvement. Some of the definitions and lemmas are unclearly stated. I also think the authors should more clearly state their contributions. Again, see weakness section for a more in depth account of the issues.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
Some of the main claims in the paper seem to follow from previous work and techniques. It would be helpful if the authors had a more thorough comparison to previous work. Again, see weakness section for a more in depth account of the issues.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

9. Please provide an "overall score" for this submission.
4: An okay submission, but not good enough; a reject.

10. Please provide a "confidence score" for your assessment of this submission.
4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes

Reviewer #4

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
This paper discusses the problem of teaching convex polytopes, i.e., regions induced by intersecting arbitrary n-halfspaces in R^d, assuming the presence of a perfect teaching oracle. It shows the average teaching complexity of this problem is \Theta(d) instead of \Theta(n) in the worst case analysis. Notably, the average teaching complexity has no dependence on n. When learning, the learning complexity is \Theta(n) for iid queries and \Theta(d log n) when the learner chooses the points to query. All the results hold under a relaxed assumption that the hyperplanes are in general position in R^d.

The hypothesis class H_{n,d} is a set of n-hyperplanes and the learner has access to a subset of labelled instances, Q\subseteq H_{n,d}\times {+1,-1} called halfspace queries. The goal is to teach a region r* with the smallest Q (a teaching set of minimum size). The average teaching complexity were defined in the paper as the expected size of Q. The paper shows exact forms (Theorem 1 and Proposition 2) for the number of regions and faces respectively induced by intersections of halfspaces in the d'-relaxed general position setting and used the results to compute the teachability of the region.

AFTER AUTHOR RESPONSES:
I think the current content of the paper failed at making useful connections to existing results. I am changing my scores to reflect this.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
The results are theoretical with practical implications. To my best understanding, the analysis is comprehensive. The results certainly add to the current knowledge on teaching and learning-by-query complex hypothesis classes. I believe providing tight bounds on the average teaching complexity of polytopes is of general interest to the wide community of NeurIPS, machine learning, and learning theory.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
The paper focuses solely on the average teaching complexity in the classical teaching paradigm e.g., finding a teaching set of minimum size upfront for every target concept. Given the complexity of the tackled classes, it is certainly understandable not to discuss the results in the light of other related teaching complexity parameters e.g., [8] or Gao, Z., Ries, C., Simon, H. U., & Zilles, S. (2017). Preference-based teaching. The Journal of Machine Learning Research, 18(1), 1012-1043.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
I am not expert in the area but I can tell that the methodology is elegant and the results are detailed.

5. Clarity: Is the paper well written?
The paper is well-written with vey nice figures and examples to follow.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
The authors did connect their results to the related work in the literature. However, I believe there could be more direct connection to the broad field of learning theory that is missing in the discussion (see the comments below). I would also suggest adding direct related results to Table 1 to emphasize on the improved or new results of this paper.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:
I have four comments:
1) When I first read the paper, I thought "Halfspace Query" is a new query type. I believe they are simply membership queries as defined in Angluin's seminal work [A1988]. I am not sure though. Can you comment on the relation between learning from membership queries alone, probably the most well-studied setting in learning-by-queries, and the halfspace queries learning setting as defined in lines 90-111? If they are related it would be nice to

connect them explicitly in the paper.
2) Can you comment on the relation between the average teaching complexity in the paper vs the classic definition of average teaching dimension for a concept class C? 1/|C| \sum_{c\in C} TD(c,C) see, for example, [LSW2007] where TD(c,C) is the classical teaching dimension for a concept c. Also, can we relate this reduction in the average teaching complexity to some “hard-to-teach” regions? What those look like?

3) I would appreciate if the authors sketch Algorithm 2 within the paper, if possible. The algorithm wasn’t obvious to me from the teaching complexity discussion. Regarding Algorithm 2, from my understanding the teaching complexity lower bound active learning complexity, therefore, can we comment on the optimality of Algorithm 2. Can we claim that Algorithm 2 is optimal (or within a factor of log n).

4) Given \Theta(d log(n)) learning complexity, can we say the algorithm is also attribute-efficient as the complexity is sublinear in n? Especially when d is very small compared to n. See for example [KS2006].

[A1988] Angluin, D (1988) Queries and concept learning. Machine Learning, 2(4) 319-342.
[LSW2007] Lee, H. K., Servedio, R. A., & Wan, A. (2007). DNF are teachable in the average case. Machine learning, 69(2-3), 79-96.
[KS2006] Klivans, A. R., & Servedio, R. A. (2006). Toward attribute efficient learning of decision lists and parities. Journal of Machine Learning Research, 7(Apr), 587-602.

9. Please provide an "overall score" for this submission.
6: Marginally above the acceptance threshold.

10. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes