View Reviews

Paper ID
3534

Paper Title
Distributional Generalization: Characterizing Classifiers Beyond Test Error

Reviewer #8

Questions

1. [Summary] Please summarize the main claims/contributions of the paper in your own words (1-2 sentences or paragraphs).
The paper makes a novel attempt at conjecturing and defining a notion of generalization for models in the overparametrized regime.

2. [Detailed comments] Describe the strengths and weaknesses of the work, with respect to the following criteria: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, relation with prior work, clarity of writing, and relevance to the ICML community.
Strength:
1. The author has made clear a concept that is very hard to clarify, that is, the generalization error of an overparameterized model needs to be understood conditionally
2. The experiment is convincing; the conjectures and definitions are well-motivated
3. Relation to classical notion of generalization is discussed
Weaknesses:
1. A key object in the theory is the distinguished L, but it is not clear how this may be found or estimated; please discuss this

Detailed comments:
1. typo: line 80, right panel, "we experimentally stress test"
2. Figure 2A: font invisible; please enlarge figure or font size
3. Figure 4: font invisible; please enlarge figure or font size

==========
After rebuttal:

I think the merit of the paper is to bring to a clearer form a well-documented phenomenon that no previous work seems to have studied specifically. The theoretical value of the work is limited at this stage, but I can foresee its relevance for deep-learning-based algorithm designs -- especially for the situations where mislabeling exists, and natural mislabelings tend to be correlated with some classes but not others -- which is the setting of the present work. Actually, to my knowledge, the conjecture and the phenomena stated in this paper seem to be the subconscious, or "hidden", assumption for many label-noise related algorithms; maybe one good way for this paper to improve is to discuss in more detail how the discoveries could serve as the "hidden" basis for the relevant algorithms, and this will make the practical contribution of this paper clearer. I will therefore keep my original score.

3. [Relevance and Significance] (Is the subject matter important? Does the problem it tries to address have broad interests to the ICML audience or has impact in a certain special area? Is the proposed technique important, and will this work influence future development?)
Solid contribution to relevant problem

4. [Novelty] (Is relation to prior work well-explained, does it present a new concept or idea, does it improve the existing methods, or extend the applications of existing practice?)
One idea that surprised me by its originality, solid contributions otherwise

5. [Technical quality] (Is the approach technically sound. The claims and conclusions are supported by flawless arguments. Proofs are correct, formulas are correct, there are no hidden assumptions.)
Technically strong, highly general results, advanced techniques

6. [Experimental evaluation] (Are the experiments well designed, sufficient, clearly described? The experiments should demonstrate that the method works under the assumed conditions, probe a variety of aspects of the novel methods or ideas, not just the output performance, present comparisons with prior work, test the limits and check the robustness of the novel methods or ideas, and demonstrate their practical relevance.)
Solid, informative evaluation w.r.t all 5 criteria

7. [Clarity] (Is the paper well-organized and clearly written, should there be additional explanations or illustrations?)
Very clear, only minor flaws.

8. [Reproducibility] (are there enough details to reproduce the major results of this work?)
Yes

9. [Questions for authors] Please provide questions for authors to address during the author feedback period. (Optional, to help authors focus their response to your review.)
Please answer the weakness I raised.

10. Please provide an "overall score" for this submission.
Weak Accept: Borderline, tending to accept

11. [Confidence] Please provide your confidence in your assessment of this submission.
I am knowledgeable and willing to defend my evaluation, but there's a chance I missed something.

16. Please acknowledge that you have read the author rebuttal. If your opinion has changed, please summarize the main reasons in the Detailed comments sections.
Agreement accepted

Reviewer #9

Questions

1. [Summary] Please summarize the main claims/contributions of the paper in your own words (1-2 sentences or paragraphs).
Consider a binary classification task (for simplicity) on a real world dataset of size n sampled iid from distribution D^n. Assume that there are K different "natural/meaningful" categories that can also be defined over the input domain of this distribution, that are different from the categories in the binary classification task. The claim of the paper is that if we train an interpolating classifier on the binary task (which has no knowledge of these K categories), the correlation between the classifier prediction and any one of these K categories over the training set will be close to the correlation between the true task label and that category over the entire input distribution.

In simple words, a classifier implicitly preserves the correlation between a target label and an "implicit" label between training and test set, even though these implicit labels are never seen by the classifier during training. In experiment 1 of the paper, the target labels would be animals/objects and implicit label would be cats (which happens to be a sub-category of animals).

2. [Detailed comments] Describe the strengths and weaknesses of the work, with respect to the following criteria: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of

the contribution, relation with prior work, clarity of writing, and relevance to the ICML community.
Strengths:

1. The observation presented in this paper are intriguing and fundamentally important towards understanding generalization in machine learning models.

2. The framework presented in this paper for understanding these observations are original and though provoking.

3. The presented framework reveals more structural aspects of generalization than the classical generalization theory which only talks about the expected test error. The presented framework talks about the prediction probability of individual classes and how it generalizes from training set to test set.

4. There is a sufficient variety in experiments to convince the reader of the claimed phenomenon.

5. Definition 1 is very intuitive and can perhaps serve as a useful tool in future analysis. It would be great to see more discussion on it; for instance, perhaps sample complexity would play a role in determining which features are distinguishable vs not. Another point of discussion would be that the existence of a distinguishable feature L is highly entangled with the function class of the model being used in the procedure A. For instance consider that the domain of the data distribution D is 1 dimensional. In this case, if the function class is a sinusoid function, then one of the partitions in L could be {…-4pi,-2pi,0,2pi,4pi…}, while if we use the sign function as our model class, one of the partitions could be the set of positive real numbers. However, this definition clearly seems to be intended more toward natural data setting.

Weakness:

While the ideas presented in the paper are very interesting, there are certain concerns:

1. The main concern is around conjecture 1. The assumption and implications of conjecture 1 are not clear.

a. Assumption: What are natural distributions? Section 3.4 vaguely mentions that the conjecture should hold in “natural settings”. This is not a precise enough definition to make the conjecture useful. For instance, noise is an inherent part of natural data. So is noise a part of natural distributions? If so, which types of noise are? Another question is: do the labels y of a natural distribution D mentioned in the conjecture need to be distinguishable features?

b. Implications: There is no actual discussion on what the statement of the conjecture means (specifically Eq 3,4). The paragraph below the conjecture simply say the LHS and RHS of Eq 4 have a small distance between them and that it holds for all distinguishable features L. But what does this closeness of LHS and RHS mean? This is an important point that is central to the conjecture which is not discussed properly. This needs to be especially clarified in the context of “natural distribution”, perhaps with examples where (x,y) and L are specified.

Here is an example of an ambiguity that arises at the moment due to the lack of this proper discussion. In the binary classification problem between objects and animals in experiment 1, the labels of cats (animals) is flipped with 30% probability to objects. However, this flipping of labels results in a dataset that is not “natural” anymore. So why does the conjecture still apply? Essentially, due to lack of concreteness, it is not clear under what scenarios the conjecture applies.

2. “Constant partition” experiment (line 296): Since the “constant partition” experiment focuses on showing that the marginal probability of each class remains the same between training and test set under class imbalance, it is somewhat important to study the impact of bootstrapping the minority class during training. Bootstrapping does

not actually change the training data distribution, but simply changes how we sample from it during training. Specifically, I believe the claim in conjecture 1 is agnostic to bootstrapping samples during training as it gets subsumed by the procedure A in conjecture 1. However, it might change the joint distribution on the LHS while not changing RHS in conjecture 1 (because RHS does not depend on A but the data distribution has class imbalance by design). Hence if using bootstrapping during training pushes LHS and RHS apart, this could be a drawback of the conjecture.

On a minor note, the terms "interpolating classifier", "interpolating method", or "training to interpolating" are mentioned throughout the paper but never defined, which made reading a bit difficult.

Overall I think the empirical observations in the paper are very interesting. The general idea of the proposed framework is also novel and interesting. However, there are a few moving parts in conjecture 1 (mentioned above).

3. [Relevance and Significance] (Is the subject matter important? Does the problem it tries to address have broad interests to the ICML audience or has impact in a certain special area? Is the proposed technique important, and will this work influence future development?)
Solid contribution to relevant problem

4. [Novelty] (Is relation to prior work well-explained, does it present a new concept or idea, does it improve the existing methods, or extend the applications of existing practice?)
Several novel and surprising contributions

5. [Technical quality] (Is the approach technically sound. The claims and conclusions are supported by flawless arguments. Proofs are correct, formulas are correct, there are no hidden assumptions.)
A paper that may be strong in other respects, but not technically

6. [Experimental evaluation] (Are the experiments well designed, sufficient, clearly described? The experiments should demonstrate that the method works under the assumed conditions, probe a variety of aspects of the novel methods or ideas, not just the output performance, present comparisons with prior work, test the limits and check the robustness of the novel methods or ideas, and demonstrate their practical relevance.)
Sufficient evaluation w.r.t. most criteria

7. [Clarity] (Is the paper well-organized and clearly written, should there be additional explanations or illustrations?)
Mostly clear, but improvements needed, as recommended in the detailed comments.

8. [Reproducibility] (are there enough details to reproduce the major results of this work?)
Yes

10. Please provide an "overall score" for this submission.
Weak Reject: Borderline, tending to reject

11. [Confidence] Please provide your confidence in your assessment of this submission.
I am knowledgeable and willing to defend my evaluation, but there's a chance I missed something.

Reviewer #11

Questions

1. [Summary] Please summarize the main claims/contributions of the paper in your own words (1-2 sentences or paragraphs).
Proposed a conjecture that the model output of an interpolating classifier will match the label distribution conditioned on certain "distinguishable features". The conjecture provides a more refined understanding on

generalization beyond the averaged error, and is supported by various neural network experiments.

2. [Detailed comments] Describe the strengths and weaknesses of the work, with respect to the following criteria: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, relation with prior work, clarity of writing, and relevance to the ICML community.
Strength:

The understanding of overparameterized models is an important problem for both theory and applied researchers. The empirical phenomenon highlighted in this paper is to my knowledge novel, and I think it is an interesting and important direction to characterize the prediction of interpolating classifiers in finer resolutions (beyond the averaged test error).
Additionally, the experiments and figures in the main text are well presented.

Weakness:

My major concern is that the submission does not
(i) adequately address why the studied phenomenon is important and what kind of insights it provides;
(ii) outline a plausible conjecture on the underlying mechanism (which may be empirically validated).
Specifically,

1. There should be more discussion on the implication (either theoretical or practical) of the findings.
For instance, does this property provide new generalization guarantee (as in the case for local elasticity; see below) that potentially remedies the failure of traditional generalization bounds in explaining interpolating models?
Or how do these results inform a practitioner in terms of selecting an architecture or training procedure?
Without these additional demonstrations, it is unclear how much a deep learning researcher should care about this particular property of interpolating predictors.

2. Since there's not really an attempt to explain the underlying mechanism of this phenomenon, it is also unclear of how robust it is across different losses or optimizers. For example, it is known that the cross-entropy loss has rather specific implicit bias towards the end of training [Papyan et al. 2020], and different optimizers may give predictors with different local properties [Amari et al. 2020].
It would be nice if these individual factors can be isolated.

Papyan et al. 2020. Prevalence of neural collapse during the terminal phase of deep learning training.
Amari et al. 2020. When does preconditioning help or hurt generalization?

3. Related to the previous point, I feel that additional experiments should be conducted to fully characterize this phenomenon.
A student-teacher setup would be ideal, as it allows for more fine-grained comparison of the test distribution beyond the hard decisions (which may not reveal certain structures in the network output / logits).
In addition, if the authors believe that local properties of the trained classifier is key to the observed feature calibration (as in nearest neighbors), then experiments should be performed on models with varying degree of "locality"; this could be kernels with different bandwidth, or possibly neural networks trained with different optimizers or regularizers.

4. I might be missing something, but I don't quite understand why overparameterization and interpolation is required for the proposed phenomenon to be present (or is this not the case?). If so, what is it that prevents underparametrized or regularized (such as weight decay or early stopping) models to have similar property?

5 (minor). Regarding the comment on page 3, it's not surprising that an overparameterized model does not

approach the Bayes optimal classifier -- such optimality is usually achieved by carefully designed interpolators. I'm not sure if this can be seen as a unique observation of the current submission.

I have read the author's reply, which addressed some but not all of my concerns.
I agree that in a paper that presents an empirical observation, a possible mechanism of the phenomenon is not necessarily needed. However, in the absence of such explanation, the paper needs to be stronger in other aspects, and at this point I am not fully convinced that (i) the presented phenomenon has sufficient theoretical and practical importance, (ii) the phenomenon is universal in the vaguely defined "natural distributions" across different objectives and training procedures (regression and SVM on kernel models would not be conclusive).
Note that this is in contrast to the implicit regularization of matrix factorization mentioned in the rebuttal, which has a fairly precise conjecture and important theoretical implications.

Hence I am inclined to keep my current score.

3. [Relevance and Significance] (Is the subject matter important? Does the problem it tries to address have broad interests to the ICML audience or has impact in a certain special area? Is the proposed technique important, and will this work influence future development?)
Solid contribution to relevant problem

4. [Novelty] (Is relation to prior work well-explained, does it present a new concept or idea, does it improve the existing methods, or extend the applications of existing practice?)
One idea that surprised me by its originality, solid contributions otherwise

5. [Technical quality] (Is the approach technically sound. The claims and conclusions are supported by flawless arguments. Proofs are correct, formulas are correct, there are no hidden assumptions.)
A paper that may be strong in other respects, but not technically

6. [Experimental evaluation] (Are the experiments well designed, sufficient, clearly described? The experiments should demonstrate that the method works under the assumed conditions, probe a variety of aspects of the novel methods or ideas, not just the output performance, present comparisons with prior work, test the limits and check the robustness of the novel methods or ideas, and demonstrate their practical relevance.)
Sufficient evaluation w.r.t. most criteria

7. [Clarity] (Is the paper well-organized and clearly written, should there be additional explanations or illustrations?)
Mostly clear, but improvements needed, as recommended in the detailed comments.

8. [Reproducibility] (are there enough details to reproduce the major results of this work?)
Yes

9. [Questions for authors] Please provide questions for authors to address during the author feedback period. (Optional, to help authors focus their response to your review.)
Most of my concerns are outlined in the comments above. A few more questions:

1. How does the finding in this submission relate to the local elasticity outlined in [He and Su 2020]?

He and Su 2020. The local elasticity of neural networks.

2. The authors mentioned that the feature calibration property may not hold beyond "natural" distributions and certain class of algorithms, but the details are not specified. What would be a counterexample to this property?

3. To reiterate the points in the previous comment, why should a theoretician or practitioner pay attention to this

conjectured distributional generalization? How does it facilitate our theoretical understanding of interpolating models or practical algorithmic choices?

10. Please provide an "overall score" for this submission.
Weak Reject: Borderline, tending to reject

11. [Confidence] Please provide your confidence in your assessment of this submission.
I am quite sure about my evaluation. It's unlikely, although possible that I missed something that should affect my ratings.

16. Please acknowledge that you have read the author rebuttal. If your opinion has changed, please summarize the main reasons in the Detailed comments sections.
Agreement accepted

Reviewer #13

Questions

1. [Summary] Please summarize the main claims/contributions of the paper in your own words (1-2 sentences or paragraphs).
The paper considers a finer measure of generalization performance than just the test error. The authors show that for datasets in which one of the training classes is mislabeled 30% of the time, similar percent of misclassifications happens in the test dataset for that specific class.

2. [Detailed comments] Describe the strengths and weaknesses of the work, with respect to the following criteria: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, relation with prior work, clarity of writing, and relevance to the ICML community.
The experimental observation that partially mislabeling a class in training leads to similar drop test performance on that class is very interesting and does indeed suggest that we should start considering finer measures of generalization than just test error.

The main theoretical contribution of the paper is the Feature Calibration conjecture the authors propose. They do note that it is not clear which classifiers or distributions satisfy this conjecture, but they expect "natural distributions" to work - it is not at all clear however why that would be the case.

In providing the notion of Distributional Generalization as depending on a set of tests, the authors in principle extend the usual notion of generalization, but it would have been interesting to propose more tests than just those of the feature calibration.

What would have been a good baseline sanity check would be to make sure that the drop in test performance for the mislabeled class is not due to an effectively smaller number of samples in that class.

The figures in the paper are not entirely clear. For example, in Figure 1, it is clear that one of the classes is mislabeled 30% of the time, and looking at the test we see this leads to 20% misclassification for that class. However, it is not at all clear whether for this network this is the only type of misclassifications the network makes on the test set - what is the actual test performance here?

The important question is whether this actually can be applied in any automated way to more complicated problems than the very interpretable case of "similar" classes like different breeds of dogs? How would this be useful in something like medical scan images or particle accelerator data?

---------- After Author Response ----------

While the authors responded to some of my questions, the points raised by other reviewers, particularly on the lack of clarity in defining what makes a "natural distribution" lead me to keep my original score. While this is an interesting empirical observation, I think the paper could be strengthened with some more experiments on various architectures and a cleaner definition used in the main conjecture.

3. [Relevance and Significance] (Is the subject matter important? Does the problem it tries to address have broad interests to the ICML audience or has impact in a certain special area? Is the proposed technique important, and will this work influence future development?)
Reasonable contribution to a minor problem

4. [Novelty] (Is relation to prior work well-explained, does it present a new concept or idea, does it improve the existing methods, or extend the applications of existing practice?)
One idea that surprised me by its originality, solid contributions otherwise

5. [Technical quality] (Is the approach technically sound. The claims and conclusions are supported by flawless arguments. Proofs are correct, formulas are correct, there are no hidden assumptions.)
A paper that may be strong in other respects, but not technically

6. [Experimental evaluation] (Are the experiments well designed, sufficient, clearly described? The experiments should demonstrate that the method works under the assumed conditions, probe a variety of aspects of the novel methods or ideas, not just the output performance, present comparisons with prior work, test the limits and check the robustness of the novel methods or ideas, and demonstrate their practical relevance.)
Insufficient or lacking evaluation in 1-2 criteria, but sufficient w.r.t. the other criteria

7. [Clarity] (Is the paper well-organized and clearly written, should there be additional explanations or illustrations?)
Mostly clear, but improvements needed, as recommended in the detailed comments.

8. [Reproducibility] (are there enough details to reproduce the major results of this work?)
Yes

9. [Questions for authors] Please provide questions for authors to address during the author feedback period. (Optional, to help authors focus their response to your review.)
- In the related work section, you comment on line 135 that interpolation can significantly affect the decision boundary and is not benign - but how does this match the experimental observation that training beyond data separation is actually beneficial and leads to higher test performance?

- Similarly, on lines 162-164 you claim that overparametrization does not give a Bayes-optimal classifier. How is it then that experimentally we see often that the larger the overparametrization, the better we do on the test set?

- In the Conjecture 1, do you have any intuition what the existence of adversarial images implies for distinguishable features? How does this impact Thm. 1 and any potential extensions to deep nets?

10. Please provide an "overall score" for this submission.
Weak Reject: Borderline, tending to reject

11. [Confidence] Please provide your confidence in your assessment of this submission.
I am quite sure about my evaluation. It's unlikely, although possible that I missed something that should affect my ratings.

16. Please acknowledge that you have read the author rebuttal. If your opinion has changed, please summarize the main reasons in the Detailed comments sections.
Agreement accepted