9 Reviews for the "Implicit Maximum Likelihood Estimation" Paper

The paper on "Implicit Maximum Likelihood Estimation" was submitted to and rejected from NIPS 2018. Attached are the reviews, the rebuttal and the meta-review, in the interest of promoting discussion. Comments are most welcome; please contact the authors at `ke.li@eecs.berkeley.edu` and `malik@eecs.berkeley.edu`.

View Reviews

Paper ID
2604

Paper Title
Implicit Maximum Likelihood Estimation

Reviewer #1

Questions

1. Please provide an "overall score" for this submission.
7: A good submission; an accept. I vote for accepting this submission, although I would not be upset if it were rejected.

2. 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.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals. You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.
++ The paper addresses the three fundamental problems that implicit models usually suffer from -- mode collapse, vanishing gradient, and training instability. The nearest-neighbor based algorithm proposed here truly mitigates all of these issues and hence could be a strong alternative for training any implicit model used in practice.
++ None of the assumptions in Theorem 1 seem too restrictive, so this is a strong paper that I would vote for an accept. The flexibility of training a model with the arguments in the KL divergence swapped is yet another strength of the paper.
++ I noticed a few problems with the writing. Please fix the following:
-- ".. if the goal is generate high-quality samples .."
-- "has less capacity that what's necessary to fit .."

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?
3: Very confident

Reviewer #2

Questions

1. Please provide an "overall score" for this submission.
4: An okay submission, but not good enough; a reject. I vote for rejecting this submission, although I would

not be upset if it were accepted.

2. 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.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals. You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.
Summary: This paper proposes a nearest-neighbor search-based approach for learning implicit models.

The authors claimed their approach has the advantage in avoiding: mode collapse, vanishing gradients, and training instability. My major concerns are about the experiments, which I think is a little weak to support the claims:
1) The only quantitative metric used in this paper is Parzen window estimates, which should be generally be avoided as suggested in “Theis et al. 2016. ICLR. A NOTE ON THE EVALUATION OF GENERATIVE MODELS.”

2) All experiments are only conducted on fair simple datasets, like MNIST and CIFAR10. Sample quality is not persuasive (e.g. Figure 1).

3) I am concerned that the running time and the generative performance of the proposed model critically depend on the performance (hyper-parameters `k`?) of the nearest neighbor search algorithm. It would help if the authors could demonstrate how the generative performance is related to the accuracy of the nearest-neighbor search module.

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?
2: Somewhat confident

Reviewer #3

Questions

1. Please provide an "overall score" for this submission.
10: Top 5% of accepted NIPS papers. Truly groundbreaking work. I will consider not reviewing for NIPS again if this submission is rejected.

2. Please provide a "confidence score" for your assessment of this submission.
5: You are absolutely certain about your assessment. You are very familiar with the related work.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals. You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.

I've really enjoyed this paper, which I consider one of the best papers I've recently read. In a nutshell, this work constitutes the first likelihood-free generative model the training process of which can be shown to be equivalent to maximizing
likelihood under some conditions. Specifically, these conditions are quite moderate, thus realistic to achieve, since the entail that the model is finite and the number of data examples is finite. This is in stark contrast, e.g. to GANs, which necessitate infinite examples and samples.

The technical development of the method, which relies on the idea of finding the nearest sample to each data example and optimizing the model parameters to pull the sample towards it, is both novel and correct. The provided theorems are also correct, and were ask absolutely needed for the paper to be complete.

The experiments are based on standard benchmarks, and provide satisfactory comparisons. However, some indicative figures concerning the computational costs of the method are also needed.

A question I have is how the vanilla setting of Euclidean distance affects algorithm performance. Of course, the authors have discussed this selection, and tried to reassure that this is not much of an issue. However, it would be good if they had provided some empirical evidence supporting this claim. At least, they must discuss how they intend to explore this aspect in future work.
'

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?
3: Very confident

We thank the reviewers for their feedback. R1 & R3 characterized the paper in the following terms:

- R3: "I've really enjoyed this paper, which I consider one of the best papers I've recently read. In a nutshell, this work constitutes the first likelihood-free generative model the training process of which can be shown to be equivalent to maximizing likelihood under some conditions. Specifically, these conditions are quite moderate, thus realistic to achieve, since the entail that the model is finite and the number of data examples is finite. This is in stark contrast, e.g. to GANs, which necessitate infinite examples and samples. The technical development of the method, which relies on the idea of finding the nearest sample to each data example and optimizing the model parameters to pull the sample towards it, is both novel and correct. The provided theorems are also correct, and were ask absolutely needed for the paper to be complete."
- R1: "The paper addresses the three fundamental problems that implicit models usually suffer from – mode collapse, vanishing gradient, and training instability. The nearest-neighbor based algorithm proposed here truly mitigates all of these issues and hence could be a strong alternative for training any implicit model used in practice. None of the assumptions in Theorem 1 seem too restrictive, so this is a strong paper that I would vote for an accept. The flexibility of training a model with the arguments in the KL divergence swapped is yet another strength of the paper."

We thank R1 & R3 for their suggestions and will fix the typos and include results on computational costs and impact of distance metrics in the camera-ready. R2's concerns are about experiments, to which we'll now respond.

The contributions of our paper are theoretical rather than empirical in nature. The point of the paper is to introduce a new method that overcomes mode collapse, vanishing gradients and training instability, not necessarily to demonstrate state-of-the-art image synthesis results. The value of the paper stems from the foundation it lays for a new research direction, upon which subsequent empirical work can be built. The claims about overcoming the three issues can be validated by the theoretical analysis alone; experiments merely supplement the analysis.

Regarding R2's comments on particular choices of evaluation metrics:

1) It is important to measure performance of generative models in terms of *both* their abilities to cover all modes of the data distribution and to generate plausible samples. While Parzen window estimates certainly have limitations (which we also noted in our paper), we point out that there is currently no better alternative quantitative metric for measuring coverage. Notably, other quantitative metrics like Inception scores measure sample quality, rather than coverage, and are therefore not replacements for Parzen window estimates. We are not claiming that Parzen window estimates are reliable measures of sample quality or estimates of the true log-likelihood, but are rather using them to demonstrate empirically the lack of mode collapse/dropping. Please see the discussion on lines 178-201 for details.

2) Sample quality is not necessarily indicative of a model's ability to estimate the density of the underlying data distribution accurately. In the same paper that R2 refers to (Theis et al., 2016), it was pointed out that "qualitative as well as quantitative analyses based on model samples can be misleading about a model's density estimation performance, as well as the probabilistic model's performance in applications other than image synthesis." When a model has the freedom to drop modes, it can effectively choose which modes it wants to model and can therefore trivially achieve good sample quality by dropping all but a few modes. So, sample quality would correlate with density estimation performance *only* when a model is guaranteed to cover all modes. As the methods that achieve state-of-the-art sample quality have the freedom to drop modes, the fact that their samples look more visually appealing than our samples does *not* necessarily mean they they are able to learn the underlying data distribution more accurately than our method.

Even if image synthesis were the end goal, we note that our method compares favourably to other methods at similar stages of development. For example, the samples for CIFAR-10 shown in Figure 1 are noticeably better than the samples shown in the initial GAN and PixelRNN papers. Later iterations of these methods incorporate additional supervision in the form of pretrained weights and/or make task-specific modifications to the architecture and training procedure, which were critical to achieving state-of-the-art sample quality. Because this is the initial paper on a new approach, we avoided these practically motivated enhancements because they are less grounded in theory, would obfuscate the key idea/insight and may give the impression that they are crucial in practice, but will explore them in future work.

3) Actually, only a value of $k = 1$ makes sense in the context of our method; as the theoretical analysis shows, a value of $k > 1$ would not maximize the likelihood of individual data examples. Because the nearest neighbour search algorithm is fairly fast, the running time of the overall algorithm is not very sensitive to its hyperparameters. For example, for our CIFAR-10 experiments, at the beginning of each outer iteration, we performed nearest neighbour search for 8,000 queries over 200,000 samples, each of which is 3072-dimensional. Constructing the data structure took 8.01 seconds, and querying took 1.31 seconds on a 4-year-old six-core CPU. This is relatively insignificant compared to the amount of time taken by backpropagation, which takes 181.85 seconds for 100 iterations of SGD on a 1080 Ti GPU. We'll include a discussion of this in the camera-ready.

View Meta-Reviews

Paper ID
2604

Paper Title
Implicit Maximum Likelihood Estimation

META-REVIEWER #1

META-REVIEW QUESTIONS

1. Please recommend a decision for this submission.
Reject

2. Please provide a meta-review for this submission. Your meta-review should explain your decision to the authors. Your comments should augment the reviews, and explain how the reviews, author response, and discussion were used to arrive at your decision. Dismissing or ignoring a review is not acceptable unless you have a good reason for doing so. If you want to make a decision that is not clearly supported by the reviews, perhaps because the reviewers did not come to a consensus, please justify your decision appropriately, including, but not limited to, reading the submission in depth and writing a detailed meta-review that explains your decision.
This has been a highly-discussed submission. I have carefully reviewed the paper, the author response, and the reviewer discussion and am weighing in with Reviewer 2. The points that Reviewer 2 raise are worth carefully considering in a revised version of the manuscript; I did not find that author feedback sufficiently addressed these points, nor were the other Reviewers able to specifically defend them.

I encourage the authors to take these comments to heart and improve the presentation, especially with regard to related work, and also to flesh out the theoretical connections that the reader would expect to see. If the theoretical presentation is to remain as it is, then the reviewers capture a fair expectation that the NIPS community would expect to see a serious empirical evaluation in its stead.