Reviewing Comments regarding the manuscript: "R*: A Robust MCMC Convergence Diagnostic with Uncertainty using Gradient-Boosted Machines"

General Remarks

The authors introduce a novel convergence diagnostic for MCMC by comparing the classification accuracy of gradient-boosted trees, predicting chain ids for drawn samples, to random guessing. I believe that the work is quite relevant but needs some further improvements.

Some high level remarks on the draft.

1. The text contains various small grammatical errors. I suggest the authors make a careful read through the paper to correct those. Specifically, quite a few sentences seem the have a missing "the".
2. The authors sometimes use "ML" in cases in which this does not make much sense to me. An example is the use of "ML accuracy". I'm not sure what "ML accuracy" refers to if not the "accuracy", which is a term from statistical learning theory. If "ML accuracy" refers to something else, I suggest the authors elaborate on the term.
3. The authors sometimes make relatively strong claims without much or any evidence. Examples for this can be found in the "detailed comments" section.
4. I would ask the authors to have a pass through the cited works. It seem to me that some of the citations to not reference the original work. Examples for this can be found in the "detailed comments" section.

Some additional suggestions:

1. It is slightly disappointing that the authors did not evaluate their approach with different ML methods. I would encourage the authors to add a systematic analysis comparing the selected classifier (gradient-boosted trees) against other state of the art classification methods.
2. It is unclear to me how well this approach will work with discrete parameter spaces. I would recommend the authors conduct additional experiments containing draws from a generative process involving discrete variables.
3. I would generally recommend to use (A), (B), respectively, in the figure captions and use more descriptive captions in some cases.

Detailed Comments

Introduction

1. The authors mentioned a few selected probabilistic programming languages that implement a dynamic HMC variant. As this list is incomplete, I suggest the authors extend the listing to provide a more complete picture. Examples for missing PPLs are: Turing.jl [Ge et al. 2018], TensorFlow Probability [Dillon et al. 2017] and Pyro [Bingham et al. 2019].

2. "(Stan in particular (Carpenter et al., 2017) is a great exemplar of this)" I would suggest to remove this suggestive mention of a particular library if not really necessary.

3. "MCMC estimates can have negligible bias with only a relatively small number of draws." -> "MCMC estimates can have A negligible bias with only a relatively small number of draws." ... Maybe elaborate when the bias can be expected to be negligible.

4. "easily become stuck in areas of parameter space" -> "easily become stuck in areas of THE parameter space"

5. I would suggest to rephrase "Unfortunately, anyone who uses MCMC knows that it is full of false dawns: chains can easily become stuck in areas of parameter space, and observation over short intervals mean the sampling distribution appears converged." as this sentence is difficult to understand. Suggestion: "Unfortunately, anyone who uses MCMC algorithms knows that they are full of false dawns. Observing the Markov chain over a short interval may suggest that the sampling distribution is converged, while chains could quickly become stuck in areas of the parameter space."

6. "high posterior density is, time and again, not evidence" ... Consider removing "time and again".

7. There seems to be a confusion with punctuation here (and a missing "the"): "To combat this curse of hindsight, running multiple, independent chains, which have been initialised at diverse areas of parameter space is recommended" consider changing to: "To combat this curse of hindsight, running multiple independent chains, which have been initialised at diverse areas of THE parameter space, is recommended."

8. There are plenty missing "the" before the terms "parameter space". I would suggest the authors make a careful read through the manuscript to correct spelling & grammar errors.

9. "Recently, Stan has adopted more advanced variations on the original" is it necessary to mention Stan here? Maybe consider removing the mention and simply highlight the

advanced variations.

10. "Here, we introduce R∗, a new convergence metric." consider writing "diagnostic" instead to stay consistent with the title. Also, maybe explain what is new about the diagnostic. Currently, this sentence does not convey any information.

11. "This statistic is built on the intuition". I would suggest not to use the term "intuition" in a scientific publication. I suppose the idea behind R* comes, in fact, from the insight that "it should not be possible to discern from a draw's value the chain that generated it" if the chains are mixed.

12. "To maximise predictive accuracy, our chosen ML classifier naturally exploits differences in the full joint distributions between chains, which means it's sensitive to variations across the joint distribution of target model dimensions unlike most existent convergence diagnostics". As far as I can tell, this work does not propose a novel classifier, but only uses existing work. The sentence however, might suggest so. I would, therefore, suggest to rephrase the sentence.

13. "For our ML classifier," again, I belief this could be confusing to the reader as this work does not propose a classifier.

14. I had a look at the citation "Chollet and Allaire, 2018" but couldn't find any justification for the claim that gradient-boosted trees "are known to perform well for the types of tabular data". Could the authors refer me to the respective section in the book?

15. "For the types of problem we test," maybe "we tested," ?

16. "O(seconds)" I fail to understand this use of the O-notation. Could you elaborate? What is the computational complexity of the mentioned Rhat statistics in O-notation?

17. "many iterations" what is many?

18. "the time taken is longer.". I'm not sure I understand? Is the mentioned the computational complexity or not? I fail to understand this sentence.

19. "we describe in detail the method for calculating R∗ and its uncertainty". Rephrase.

20. "and elsewhere". Where? Please cite the respective works.

Method

1. "sampling distribution for any dimension, θ, in the target distribution" -> The introduction of the notation is a bit unclear. It seem theta refers to the value rather than the dimension. So maybe use use "any dimension (j) in the target..." instead to stay consistent with the use in the remaining draft, see Sec. 3.4 for example.

2. Figure 1 (caption): Maybe use (A) and (B) instead. This plots also don't show a prediction result, rather the comparison of the marginals for mixed and unmixed chains. I would also recommend to use bolder lines and explain the visualisation a bit more in detail.
3. Figure 2 (caption): Again, I don't think it is accurate to say "prediction" here, as this is simply a visualisation of the joint rather than the marginals. Also please use (A) and (B) or similar instead. I would also ask to add more descriptive text to the caption as this visualisation effectively illustrates the key insight of the paper.
4. "we train a supervised machine learning (ML) model" -> ML has already been introduced in the introduction. Please remove.
5. "whether classification accuracy" -> "whether THE classification accuracy"
6. "above the “null” case," -> This is a bit cryptic. I think what the authors mean is that the accuracy should be above random guessing. Could you rephrase or elaborate.
7. "of ML accuracy" -> There is no such thing as ML accuracy. I suppose the authors mean "of the accuracy".
8. "null accuracy," -> Again, unclear what that means.
9. "gives a recipe for calculating" -> I'm not sure what a recipe for calculating something should be. Maybe rephrase to "provides a pseudo-code implementation of...".
10. "The ML classifier" -> I'm not sure it is necessary to always refer to the classifier as a ML classifier. Maybe consider dropping the ML in the remainder of the draft.
11. "which experience has dictated to be a highly predictive framework for use in tabular data" Again I could not find any details about this claim in the book. Could you please refer me to the respective page?
12. "each chain has dimensions: $X \in RS \times RK$, where S is the number of ..." Is X really the dimension? What is X? This needs some more careful introduction of the notation.
13. "70% of draws for training and 30% for testing" -> Why was 70/30 chosen?
14. "taking O(seconds) on a desktop computer" I fail to understand what the O-notation of seconds is. Why not denote the O-notation of gradient-boosted trees? I refer the authors to [Ke et al. 2017] for details. Or maybe I misunderstand what the authors refer to.
15. "This shows that as the number of samples increased, variation in R_* declined." This is to be expected as the generalisation error of the classifier will decrease.
16. "more training data leads to better ML models." I don't understand this sentence. Maybe rephrase to "more training data can lead to a lower generalisation error of the classifier." Also note that increased training data does not necessarily imply that the performance of the classifier will increase.
17. "randomness in R_* calculation" -> "randomness in the calculation of R_*"
18. "about convergence being drawn" I don't understand this. Please rephrase.
19. "The computational cost of doing this may, of course, be unreasonable." Explain what unreasonable means and why it is "unreasonable" or remove this sentence.
20. "variation in R_* comes from sampling from the probability simplex" Are those class probabilities callibrated?

21. "shown in Figure 3B was generated by repeatedly recomputing R_* using Algorithm 1 " -> Maybe consider using the same scale and x-axis for both plots so that they are easily comparable.
22. "fixed number of chains: four." -> Remove "four".
23. "classification becomes a harder problem when there are more categories," Is this true? Could you please elaborate and provide a reference discussion on the complexity of classification problems with large number of classes.
24. "where all chains bar one have" -> "where all chains BUT one have"
25. "because it is harder to classify chains when there are more of them." Again, I'm not convinced this is actually true. In fact there is theoretical evidence that the opposite might be the case in some situations. See for example: [Abramovich & Pensky 2019].
26. "The performance of GBM, like all ML methods, depends on its hyperparameters." This is not entirely true as there are hyperparameter free methods.
27. "We use Stan's NUTS algorithm" Maybe consider adding a reference here.
28. "he latter thinned by a factor of 5" Why was this necessary?
29. "remains stubbornly shifted" What does "stubbornly shifted" mean? I suggest you drop the "stubbornly" here.
30. "Rather than run the MCMC" -> "Rather than RUNNING the MCMC"
31. General remark to section 3.2.2. It is difficult to assess the experiment without details about the configuration of the NUTS sampler used. I would suggest the authors provide a more complete description of the experimental setup.
32. "and this allows us" -> "which allows us"
33. "eq." -> Uppercase
34. "Fig. 8C variable importance" -> "Fig. 8C THE variable importance"
35. "high values mean a variable is more important" Please elaborate how feature importance is measured in gradient-boosted trees.
36. "this illustrates that variable importance provides information complementary to R and ESS" Why does it imply this?
37. "Since the R_* distribution indicated non-convergence for both parameterisations, we ran each model for sixty-times as long, although thinned by a factor of 3, resulting in 10,000 post-warm-up iterations across each of 4 chains." Please rephrase this sentence.
38. "remains stubbornly " What does this mean. How is stubborn defined? Please consider removing this.
39. "One measure of distributional "closeness" is the KL-divergence, which, in this case, could be used to measure the divergence from target to sampling distribution: if the target distribution is known, fitting a kernel density estimator (KDE) to samples allows an approximate (typically univariate) measure of KL-divergence to be calculated for each dimension. " This is a very long and complicated sentence. Consider rephrasing it.
40. "fitting a kernel density estimator (KDE) to samples allows an approximate (typically univariate) measure of KL-divergence" Why is this the case if the target distribution is not tractable? If it would be tractable and exists in close-form, why not use it directly?
41. I appreciate the excursion in 3.3.1 describing why certain measures will fail for heavy-

tailed distributions.

42. "after c.10,000 iterations" What does the c. stand for?
43. "After c.550 iterations," Again, this notation has not been introduced. What is c.550?
44. "The model can be parameterised in two ways, as described in Vehtari et al. (2020)" I believe those parameterisation have been introduced much earlier. Please cite the original work introducing a reparameterization of the 8-schools model instead.
45. "for the NUTS algorithm (Betancourt, 2017)," Please use the correct citation for NUTS here [Hoffman & Gelman 2014].
46. "To probe the predictive power of the ML classifier, we investigated how predictive accuracy varies across parameter space." -> "To probe the predictive power of the classifier, we investigated how THE predictive accuracy varies across THE parameter space."
47. "we group MCMC draws into deciles and draw from the R_* distribution for each decile." Are you using the whole data set here or only the test set? This is unclear. If it is the test set, how does this relate to the predictive capacities of the classifier?
48. "eduction in ML classification accuracy" Again, the is not such thing as a "ML classification accuracy". I suppose the authors refer to "classification accuracy", which is a term from statistical learning theory, if so please rephrase.
49. "this was to test that ML classification" Again, there is not such thing as "ML classification". Please rephrase.
50. "didn't become prone to overfitting in this limit." Maybe instead: "the classifier did not overfitting in this scenario."
51. "sampling was done using Stan's NUTS algorithm." Please describe the hyper-parameters used.

Discussion

1. "we used supervised machine learning (ML) classifiers to quantify" Only one method was used, consider rephrasing.
2. "a null model (which predicts a chain's identity uniformly at random)" This is the first time the "null" model is properly described. Maybe consider moving this description into the methods section.
3. "extracting ML-predicted chain" There is no such things as a "ML-predicted chain". Please rephrase.
4. "includes this information in building a model" I believe the authors refer to "train a model" or "learn a model" not "build a model". Consider rephrasing.
5. "because the classification boundaries for this task are unlikely to be very complex compared to (say) machine vision tasks. " Why would this be the case?
6. "target (RUser4512, 2018)." Please replace this citation with a proper citation of a scientific work, such as [Abramovich & Pensky].
7. "If so, this suggests that larger models" Please rephrase.

8. "larger models (usually needing more MCMC iterations) may currently be beyond the reach of R_*." I suggest the authors briefly discuss alternative approaches, which would allow the application of R_* in such scenarios. Examples for this could the the use of deep learning techniques or data reduction techniques.
9. "our ML calculation method" What is a ML calculation method?

References

- Ge, Xu & Ghahramani "Turing: a language for flexible probabilistic inference.", in proceedings of AISTATS, 2018.
- Dillon et al. "TensorFlow Distributions", in Arxiv, 2017
- Bingham et al. "Pyro: Deep universal probabilistic programming", in Journal of Machine Learning Research, 2019.
- Ke et at. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", in proceedings of NeurIPS 2017.
- Abramovich & Pensky "Classification with many classes: Challenges and pluses", in Journal of Multivariate Analysis, 2019.
- Hoffman & Gelman "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo", in Journal of Machine Learning Research, 2014.