Response to Reviewers - Round 2

Title: "BROOD: Bilevel and Robust Optimization and Outlier Detection for Efficient Tuning of High-Energy Physics Event Generators"

Authors: Wenjing Wang, Mohan Krishnamoorthy, Juliane Müller, Stephen Mrenna, Holger Schulz, Xiangyang Ju, Sven Leyffer, Zachary Marshall

We thank Greg Landsberg, Editor in charge, and the three reviewers for their comments. Below please find our responses to each of the remaining issues raised in the review reports. We have numbered all detailed reviewer comments according to our first round of revisions and include only those comments that the reviewers had further comments on. The new changes made in the manuscript are indicated in blue font and the number of the comment corresponding to the changes is indicated where applicable, e.g., "[Reviewer comment iii:] changed text".

Response to the Editor in charge

Dear Authors, The first two referees are happy with the revised version of the manuscript, and the third referee has just submitted a follow-up report, where they request you to address a few minor remaining comments. We ask you to please consider this request, upon which the paper will be published without further delays. Regards, Greg Landsberg Editor in Charge

Response: Dear Editor, in the following we have addressed the reviewer comments.

Response to Reviewer 1

Thank you to the authors for their detailed responses to my comments and for integrating my suggestions into the manuscript. I am satisfied with the responses and I think the paper is now ready for publication.

Response: Thank you for your positive review.

Response to Reviewer 2

Thank you for going through all my comments. Looks great, so let's publish!

Response: We thank the reviewer for your support of our manuscript.

Response to Reviewer 3

I went through the new version of the manuscript and the provided answers to my comments. I am in general very satisfied with the effort made by the authors to improve the paper. I still have a few follow up comments:

Response: Thank you for your positive comments, we appreciate your support and we are addressing your remaining comments below.

iii ***Review round 1 comment:*** *The paper discusses the task of MC tuning and proposes a new procedure to improve over current state of the art, by speeding up the computation and reaching a better agreement on a real-life example. This is done utilising a few new elements, including a better minimisation strategy (a few are proposed) and an outlier removal procedure. I have a few questions regarding the big picture proposed: 1) To which extent the outlier removal is sound? Even in absence of systematic issues, outliers will occur. Removing them might help to established the expectation value of the parameter one is fitting, but any sense of statistical interpretation of the uncertainty range is lost (i.e., coverage is broken). The paper offers no discussion of this point and how crucial this is.*

Review round 1 response: As explained in the response to reviewer comment 25, removing bins or entire observables from the tuning process does not reduce the information required to achieve a good tune. Additionally, the filtering approaches only eliminate parts of the MC model that are highly unlikely to be explained by data. Hence, it is a conservative approach since the range of the function within the domain is usually much larger than the range of the values that could be used to fit the data. The outlier removal is based on the intuition that the models that are highly unlikely to be explained by data could be removed to (a) get a better estimate of the tune, and (b) prevent the algorithms from going into regions of extrapolation. We added this explanation in Section 7.

Review round 2 comment: *The authors insist that removing points is conservative. I don't see it, sorry. Would be nice if they could show what they say. From my experience, removing outliers in a fit (particularly a chisq fit) has an effect on the fit (bias and uncertainty underestimate). I have the impression that what the authors say is true under the assumption that the model you use to fit the data is correct. But they comment at length that this is not the case (maybe because of the underlying assumptions, e.g., fixed-order perturbative calculations, soft-physics modeling, etc). So I am not very convinced of all this and I would like to see a more clear demonstration.*

Review round 2 response: Thank you for your thoroughness on this matter. We would like to clarify our explanation and also change our word choice to better reflect what the outlier removal means here. In our intended meaning, "conservative" was referring to the case of dropping the data for which applying the model would be invalid. We realize that this word choice may be misleading. Our goal is to attain a good model fit only in regions where it is valid. We have no analytic formula for this region, but we infer it from the data itself. We do believe that there are regions where the model is correct, but there are other regions where it is not. If we include data in the fitting from regions where the model is invalid, then our results will most likely be incorrect as well.

We use our surrogate functions to perform the outlier filtering. There are two scenarios of the outcomes of our filtering approach: (1) we remove outliers in regions where the model is invalid, in which case the model cannot describe the data, and thus the approach could be considered "conservative"; (2) we remove data (outliers) from regions where the model is valid, in which case our approach cannot be considered "conservative". In this case, the uncertainties of the parameters will be impacted. However, given our eigentune results, we do not believe that this is of big concern. Please see Sections 3 and 7, blue text, for the update to the manuscript.

e ***Review round 1 comment:*** *Page 15: the offered procedure is effectively a single-hypothesis test. Only Fisher called this kind of test hypothesis test. He disagreed with the Neyman Pearson hypothesis test, which requires two hypotheses to be specified. My understanding is that Fisher's problem is ill-posed, as decades of frequentist literature exposed problems related to this. In particular, no any claim of optimization in this context is typically an overstatement. This point should be discussed.*

Review round 1 response: We thank the reviewer for this comment. Fisher's approach considered the p-value could be interpreted as a continuous measure of evidence against the null hypothesis. On the other hand, Neyman Pearson's approach said one can use the p-value to either reject the null hypothesis, fail to reject the null hypothesis, or end up with type 1/2 errors. Given these definitions, the approach described in Section 3.2 is a Neyman Pearson approach since every subset $\mathcal{B}$ of observable $\mathcal{O}$ is tested to check if the null hypothesis is rejected or otherwise. We realize that the explanation of the optimization formulation may be misleading and we have added a few clarifying sentences below equation (15) to help the reader better understand our approach.

Review round 2 comment: *I disagree. NP requires an alternative hypothesis. Not clear what that is, in your method. A quick google search for a paper that spells this out:* `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4347431/`*. Not a super relevant point. I would just avoid saying hypothesis test and would say null-hypothesis test or some such*

Review round 2 response: We thank the reviewer for the link to the tutorial. Please note that we updated the wording in the text as suggested (see Section 3.2).

l ***Review round 1 comment:*** *Sec 4.8: why not running everything on one machine and make a one-to-one comparison?*

Review round 1 response: Please note that we were not primarily concerned with the run time of the methods, but rather automation, and one goal was to ensure that the code runs on different architectures. See Section 4.9.

Review round 2 comment: *I acknowledge the fact that speed is not the main aspect. But I think the question stands. I find this choice quite odd, also because even if the time is not the main concern of the paper, it is the main concern of this section (see title of 4.9). At least some explanation of the reasoning behind this choice should be given.*

Review round 2 response: We thank the reviewer for their comment. Please note that there was a typo in the manuscript when reporting the specs of the computers we used. Both computers used in the numerical experiments used 1 thread and they both have similar processor performance. Thus, we believe that the reported wall-clock times can serve as a rough guide of the computing effort required by each method. Please see Section 4.9.