SBI Blueprint workshop notes

https://indico.cern.ch/event/1600677/timetable/

Thursday

Introduction

Simulation-Based Inference — An Introduction (Gilles)

Questions:

- Ricardo: Slide 17: In the local likelihood model, does it make sense to calculate $Z(\theta)$?
    - The normalization is a function of the parameters, \theta, and so the way that you calculate it is to histogram the outputs and then normalize to the histogram.
- Giordon: Slide 20: can we use SBI to maximize the exclusion potential instead of the discovery potential ?
    - Lukas: If you look at a SUSY discovery plot there are two axes (mass, exclusion) and there we optimize on a fixed summary statistic and then different parameter points have different shapes. If we parameterize the summary stat on the mass point then you get better exclusion. Similarly, with SBI you get better exclusions.
- Alex: Is there something optimal for CLs?
- Lukas: We have the extended likelihoods and in SBI we generally look at the event likelihood, but if the total number of events is parameter dependent then we would need to also learn information about the cardinality of the dataset.
    - c.f. Ricardo's talk
- Lukas: Slide 20: Power of SBI comes readjusting per parameter point. The point is *not* that things are binned (that doesn't matter) it is that it is a suboptimal human engineered feature.
- Lukas: The joint score (SALLY) seems to be more data efficient. Is it clear why? Should we change our simulators?
    - Gilles: Unclear. The same improvements haven't been seen in astrophysics applications.
    - (Ricardo) Gilles also mentioned that for methods that use a lot of joint information in the loss, the randomness coming from these sources could actually hurt training performance, and using simple classifiers may actually be better.
    - Ricardo (used SALLY, attempting an answer): SALLY uses the joint score around the reference point. For e.g. ROLR where the two points are far from each other, the joint likelihood ratio could have large per-event variations and would require more training samples to converge properly. The trade-off is that ROLR gives you actually the likelihood ratio which you can use directly, while SALLY only gives

the score, which then one has to histogram. Additionally SALLY is only optimal around the reference point, where linear effects in the POI dominate.

Simulation-Based Inference at the LHC

Simulation-Based Inference in ATLAS (Tae Hyoun)

- Ricardo: in your systematics model, for |alpha| < 1 you do not take into account the event kinematics, while you do for |alpha| > 1.
    - Jay: yes, this interpolation choice is based on HistFactory (InterpCode 4), but even for |alpha|<1, the kinematic dependence is still taken into account since this weight is multiplied by the ratios derived with CARL on the nominal sample
- ?
    - Jay: You can optimize so that it is likelihood independent.
- Davide: How do you choose a reference hypothesis p_ref? What does this choice introduces in the fit? For example, support problems, systematics acceptance effects,
    - Jay: in ATLAS we defined the reference hypothesis using inclusive signal in a preselection region. After the preselections, the signal cannot go to zero, and this prevents the problem.
    - Jay: the choice of the fixed p_ref cancels out in the profiled likelihood ratio, so it is not affected by systematics

Simulation-Based Inference in CMS (Sergio Sanchez)

Questions:

- Lukas: you mentioned you're using the SALLY method, which requires some gradient information at the matrix element level?
- Tae: In situations in which everything floats it seems that you obtain better intervals. Is that true?
    - Sergio: Not sure (didn't work on this analysis), but we should follow up.
    - It is possible to have BSSM interference and this needs to be considered.

Statistical Modeling (Systematic Uncertainties)

HistFactory Introduction (Alex)

- Ricardo: Slide 12: Is there any heuristic to choose between the four options (code 1-4)?
    - In practice, we use code 4 which is a 6th order polynomial. But in practice you need to make a decision. These are all assumptions about the modeling, and if this assumption starts to matter, then you have other more important problems.

HistFactory v2: Systematics Modeling with Gaussian Processes (Jay)

- Ricardo: What is the computational cost for using GPs?
    - Jay: GPs are very fast to evaluate.
    - Jonas: You still need to keep all of your simulated samples in memory to do this though?
        - Jay: No, you're still using the histograms per point.
- How are you going to use the covariance?
    - C.f. link on Slide 16. You can introduce a Gaussian constraint term that depends on the covariance
- How do you ensure that the uncertainties from the GP are not underestimated?
    - Jay: They are Bayesian methods and so you are reliant on your prior which is propagated to the posterior

Unbinned measurements with machine-learned systematic uncertainties (Ricardo)

Questions:

- Giordon: Slide 18: Can you clarify what coverage means here?
    - The classical Frequentist concept of coverage.
- Giordon: Slide 16: ordering of impacts changes between binned and unbinned?
    - We believe in the unbinned case some inputs to the estimator can better constrain certain nuisance parameters, changing the order
- Some systematics are designed to be an eigendecomposition of the nuisance space, does that help/affect this setup?
    - Could have independent systematics but there is no guarantee that their impact on the final likelihood is also factorized — there might be correlations.
- On ensembling vs. calibration (slide 20): why should we prefer one over the other?
    - Maybe the effect of the nuisance is not actually linear, but the model is restricted to linear.
- Nick: Slide 13: Does this calibrate the response one dimension of the parameter space at at time?
    - This is only for balancing the process mixture at nominal point
    - Basically, because nuisances affect processes in a different way, we must be sure we are doing a very good job of predicting the process fraction in all phase space
- Nick: a comment on 16, the constraints on TES are quite extreme, in a more realistic setting do we have a way to understand when we are constraining nuisances better than object performance groups and when our model is artificially constraining these parameters?
- Stefan: What samples was the multiclass network trained on and how were the systematics included into the differential model?:
    - The multiclass network was trained on just the nominal samples, while the nuisance parameters can be conveniently factorised out in the differentiable model

- Stefan: How was the multiclass isotonic regression implemented? What is the calibration sequence that you take in your case?
    - Signal vs all backgrounds, and then iteratively calibrating the largest remaining backgrounds against the rest of the backgrounds.
- Open question!
- On the isotonic regression
    - Why ensembling before isotonic regression?
    - (Ricardo) I think the question here is if we can choose one or the other in some cases.
- (Ricardo) clarification on isotonic regression (from arXiv:2505.05544). The team did one vs. other isotonic calibration for each process separately. The calibrated output for the signal process is retained, then for the other processes the output from the one vs-all isotonic regression is rescaled such that the sum of all the outputs sums to 1 (conservation of probability).

$$\hat{g}_p(\boldsymbol{x}) = \begin{cases} \mathrm{IREG}(g^*_{\mathrm{H}}(\boldsymbol{x})), & \text{if } p = \mathrm{H} \\ \frac{\mathrm{IREG}(g^*_p(\boldsymbol{x}))(1-\mathrm{IREG}(g^*_{\mathrm{H}}(\boldsymbol{x})))}{\sum_{q=\mathrm{Z},\mathrm{t\bar{t}},\mathrm{VV}} \mathrm{IREG}(g^*_q(\boldsymbol{x}))}, & \text{otherwise.} \end{cases} \quad (36)$$

Neural Networks For Likelihood (Ratio) Estimation

Direct Likelihood Learning and Uncertainty Profiling with Normalizing Flows (Davide)

Questions

- Ricardo: Have you compared this to the unbinned unfolded scenario?
    - This is going to be looked at.
- Kyle: If the goal is to do the unfolding part, there is a fair amount of literature from the non-physics community that might not be recognizable at first (Gilles has a paper on this).
- Levi: Do you have an idea of how this approach scales with simulation budget?
    - Probably doesn't require much more for up to 5-dimensions, though this might change with 20 dimensions.

Neural (Quasi-)Probabilistic Likelihood Ratio Estimation (Matthew, Stephen)

Questions

- Ricardo (to not have to be the first to ask a question again): is this preferred to first refining the sample to remove negative weights like done in https://arxiv.org/abs/2505.03724 ?
    - This should be done as first approach as negative weights still encode physical information which may be lost when resampling.

    - Kyle: need some experience to see how things work in practice. If one can remove a step (e.g. pre-processing) that would be ok. If one knows that the distributions are physical and positive everywhere, that's fine. If one is in a situation where one has negative densities, then one nice thing about this approach is that one can find those regions and then decide later what to do with them.
    - Stephen: https://en.wikipedia.org/wiki/Data_processing_inequality in principle, a data processing step will always lose information. If one can skip a processing step then one does not risk losing information.
- Giordon: Slide 6 (General Setup): All the training is done with MC simulation?
    - Depends on the application. Some of the applications mentioned previously were systematics, and you could use simulation to learn the differences between different systematics.

Friday

Toolkits for Simulation-Based Inference

Toolkit for Simulation-Based Inference (IRIS-HEP) (Jay)

- Slide 14: New (beyond MadMiner) is the HistFactory-style assumptions for estimating the \Deltas using NNs
- Can use the same assumptions from HistFactory v2 in a binned space. GPs are an excellent non-parameteric replacement. Can perform better than a NN in the situation in which you don't have enough simulation in some parameter areas.

Questions:

- Ragansu: What machine learning models are allowed in the setup? What forms of datasets are allowed as well?
    - The existing API currently supports just simple MLP, but this could change in the future.
- (Ricardo) Is there a way to somehow automate validation such that it can be integrated in the process ? What about parametric POI implementations using NNs ~ e.g. CARL for EFT ? What about interoperability (e.g. BYONN)
    -
- RD: For the offline analysis in ATLAS we had to combine everything with RooFit. How does combination of non-ROOT tools with RooFit work?
    - There is a serialized workspace format that could be used for combinations, though there aren't necessarily tools to do this yet.

    - Alex: In the statistical ecosystem workshop that happened over the last 2 days and there is active work happening here between ROOT and IRIS-HEP. There seem to be paths forward that could make things easier than expected.
    - Mo: It seems like it could be possible to propagate JAX likelihoods into RooFit.

NEEDLE: Workflow Orchestration for Large-Scale NSBI Deployment (Kylian, Levi)

- [Alex] I'll have to miss the discussion session but happy to chat more about dask / dask-awkward / uproot.dask, we have lots of experience from IRIS-HEP projects. Going with dask-awkward may not be the best longer term choice (https://github.com/dask-contrib/dask-awkward/issues/598, can talk more at this workshop).
- Giordon: Slide 18: (Alex's question above is identical)
- Giordon: You mentioned the choice of dask, but have you explore other things like TaskVine,or rather how easy is it to swap out?
    - …
- Giordon: From the perspective of analysis facilities, is there anything special that is needed from the hardware perspective? You mentioned that if the file sizes are too large to keep in memory, then you need to locally access the data-per-file which maybe requires larger local disk for the nodes that process, or perhaps just more robust network infrastructure to enable efficient streaming of the data?
    - …
-

SBI Toolkit Modification and robust Uncertainty Quantification (Jingjing)

Questions:

- Ricardo: The assumption requires a criteria which is opposite to where MC stat matters?
    - For extremely small MC Stat this wouldn't apply. Would need an empirical test in a real analysis.

Broader Discussion

- How do you store something like a CARL model in the format?

SBI with semi-parameterized Density Ratios

Building summaries with event2vec (Nick, Prasanth)

Questions:

- For the one-hot summary vector which axis drives the (couldn't hear the rest of the question)

    - The sample space of the analysis. You can keep increasing the number of bins, but that will have diminishing returns. At some point the statistical noise will go up and increasing the number of bins will be detrimental.

Parametrized Optimal Observable Approach to NSBI (Matthew)

Questions:

- Jay: Slide 5: Have a method to import a serailized model into RooFit. How does this work?
    - Don't have a schema specification defined but can make one.
- Alex: In the case that you have a parameterized observable, are there things that happen that break asymptotics? At some point are you becoming discontinuous? Do you need a minimum bin count? (c.f. Slide 9) Throughout minimization there is a certain amount of data events in a bin, at the *next* step there might be a different amount because events shifted bins. How does this conceptually work? Does it practically matter?
    - Matthew: See jumps in the NLL come from poor quality optimizations.
- Alex: How do you do the optimization in terms of infrastructure? Do you end up with expressions for the model prediction and data yield in every bin as a function of all the parameters (POI and nuisance parameters)? How to validate this?
    - Yes, fully parameterized in every bin. In practice dependence on the parameterized observable was small. Can validate by sampling parameter points and comparing to the prediction.

Tooling issues for binned nSBI analyses with pyhf and how to overcome them (Malin)

Questions:

- Giordon: Slide 4: Can you only do this if you have some sort of morphing model?
    - Malin: Yes. If you can't build your signal in an analytical form then you would have a problem.
- Giordon: In the context of overfitting we don't try to train on all signal points in SUSY. When you're building these models, are you using a subset?
    - Malin: It comes down to the physics. In DiHiggs there aren't MC simulations for all of the signal samples and so are already doing morphing.
- Alex: Curious on the point of how you actually implement the morphing. Do you take a single template and then have an expression for how this template changes? Can see applications outside of SBI where something like the top mass would require templates.
    - Malin: <Too slow to capture the response, fill in later>
    - Lukas: The general machinery is there, but there are some issues with merging the custom modifiers to make sure that we can avoid having a custom modifier per analysis. There were attempts in 2023 at string based morphing approach, but this and other attempts can hopefully be revisited.
        - Lukas PR draft from 2025: https://github.com/scikit-hep/pyhf/pull/2579

General Discussion

x