Uncertainties associated with GAN-generated datasets in high energy physics

Prasanth Shyamsundar, Fermilab Quantum Institute

based on arXiv:2002.06307 [hep-ph] by K. Matchev, A. Roman, P. Shyamsundar

ML4Jets 2021, Universität Heidelberg

July 8, 2021

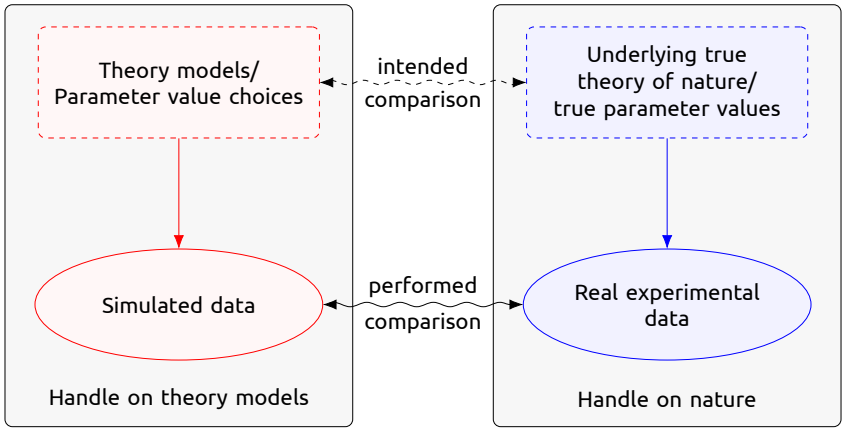
Quick introduction

- Monte Carlo production pipeline is resource intensive. Projected resource requirements for simulations for future experiments may not be met.
- Recently, several ML ideas, primarily Generative Adversarial Networks (GANs), have been proposed to solve this problem.
- Naively, the idea is to
  - Train an ML model (GAN) on a set of events produced by the true simulators.
  - Use the trained ML model (GAN) to mimic the true simulator and produce additional events.

Quick introduction

- Monte Carlo production pipeline is resource intensive. Projected resource requirements for simulations for future experiments may not be met.
- Recently, several ML ideas, primarily Generative Adversarial Networks (GANs), have been proposed to solve this problem.
- Naively, the idea is to
  - Train an ML model (GAN) on a set of events produced by the true simulators.
  - Use the trained ML model (GAN) to mimic the true simulator and produce additional events.
- There are some information theoretic arguments that **heavily** limit the possibilities in this domain. We'll look at these arguments in this talk.
- **Note:** Unless otherwise stated, our claims are not opinions but things we believe to be true.

Why do we need simulations?



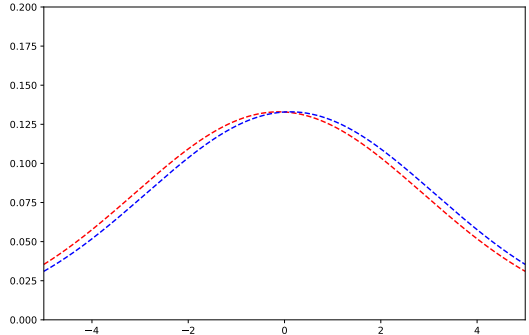
Important: The simulations need to be faithful to their corresponding theory models.

Why do we need more simulations?

Because the amount of simulations we have is not enough.

Not enough

- to reach a target sensitivity to the presence of a signal,
- to reach a target uncertainty for parameter estimation, etc.



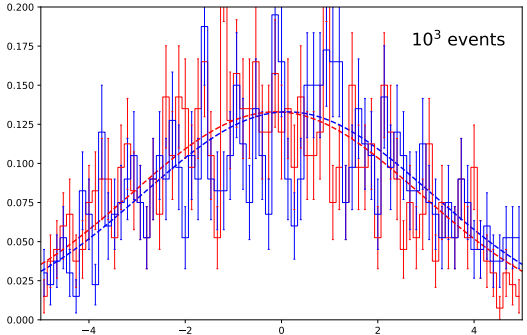
- Hypotheses become more distinguishable with increasing data.
- Same principle for real data and simulations.
- The histograms should match the underlying truth **within the errorbars**!
- **Visual simularity** is not a meaningful metric here…

Why do we need more simulations?

Because the amount of simulations we have is not enough.

Not enough

- to reach a target sensitivity to the presence of a signal,
- to reach a target uncertainty for parameter estimation, etc.



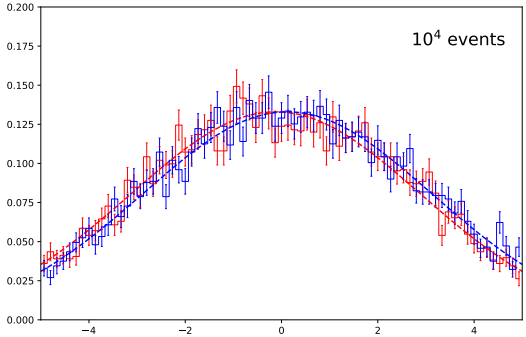
- Hypotheses become more distinguishable with increasing data.
- Same principle for real data and simulations.
- The histograms should match the underlying truth **within the errorbars**!
- **Visual simularity** is not a meaningful metric here…

Why do we need more simulations?

Because the amount of simulations we have is not enough.

Not enough

- to reach a target sensitivity to the presence of a signal,
- to reach a target uncertainty for parameter estimation, etc.



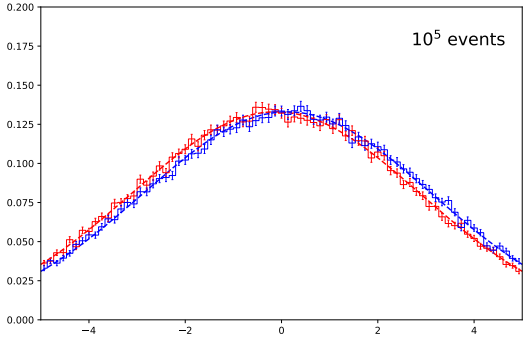
- Hypotheses become more distinguishable with increasing data.
- Same principle for real data and simulations.
- The histograms should match the underlying truth **within the errorbars**!
- **Visual simularity** is not a meaningful metric here…

Why do we need more simulations?

Because the amount of simulations we have is not enough.

Not enough

- to reach a target sensitivity to the presence of a signal,
- to reach a target uncertainty for parameter estimation, etc.



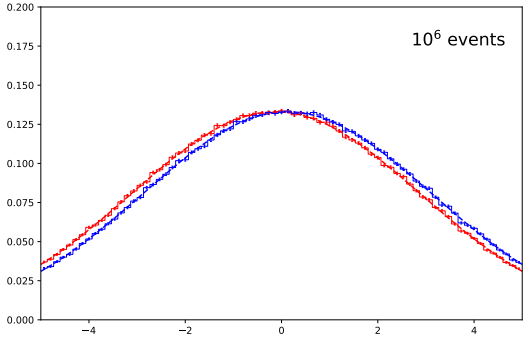
- Hypotheses become more distinguishable with increasing data.
- Same principle for real data and simulations.
- The histograms should match the underlying truth **within the errorbars**!
- **Visual simularity** is not a meaningful metric here…

Why do we need more simulations?

Because the amount of simulations we have is not enough.

Not enough

- to reach a target sensitivity to the presence of a signal,
- to reach a target uncertainty for parameter estimation, etc.



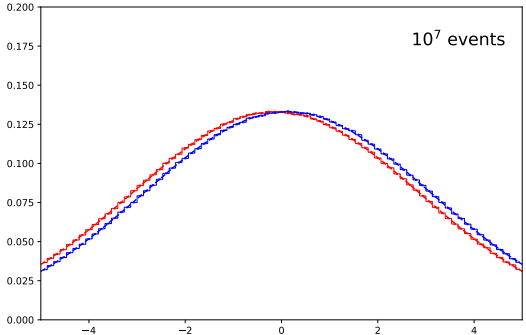
- Hypotheses become more distinguishable with increasing data.
- Same principle for real data and simulations.
- The histograms should match the underlying truth **within the errorbars**!
- **Visual simularity** is not a meaningful metric here…

Why do we need more simulations?

Because the amount of simulations we have is not enough.

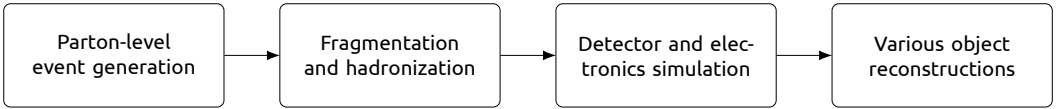
Not enough

- to reach a target sensitivity to the presence of a signal,
- to reach a target uncertainty for parameter estimation, etc.



- Hypotheses become more distinguishable with increasing data.
- Same principle for real data and simulations.
- The histograms should match the underlying truth **within the errorbars**!
- **Visual simularity** is not a meaningful metric here…

Monte Carlo production + reconstruction pipeline



This talk will focus on the case when a GAN will be used to replace the entire production pipeline from parton-level generation up to

- detector hits (step 3), or
- reconstructed objects (step 4)

Theorem 1

Under the following two assumptions, the model (or model parameter) discriminating power *achievable* using only the available true-simulated data cannot be improved by augmenting the true-simulated data with additional GAN-generated data.

Assumptions:

1. The GAN will be used to circumvent the MC pipeline, starting from the parton-level event generation and going all the way up to either detector hits or reconstructed high-level objects.
2. There are no restrictions on the types of analyses a practitioner is allowed to perform using the real and the (GAN/true) simulated data. For example, the practitioner should not be restricted only to a particular technique, which may or may not be optimal in its usage of the available data

Arguments as a dialogue

Person A proposes to use GANs to replace the entire MC pipeline.
Person B is a doing an analysis using the real and simulated data. In particular, their analysis is agreed to be optimal for the task at hand.

B: *I wish I had access to more simulated data :(*

A: *Oh, I can get you more datapoints with a GAN :)*

B: *I don't think that'll help me. I'm already doing the best one can with the available true-sim data.*

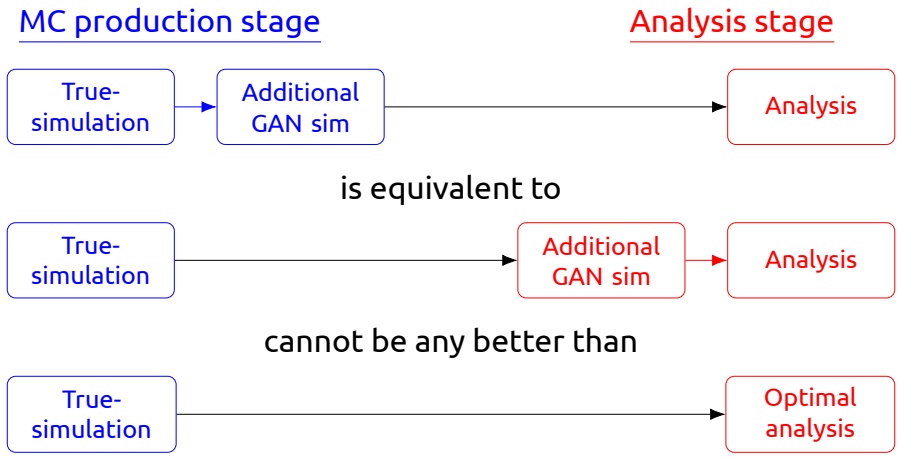
A: *But a GAN can learn the correlations in the high-dimensional data and approximate the underlying distribution.*

B: *My analysis already does that! We use machine learning and other techniques to squeeze out all the information contained in the data.*

Let's say person A is still not convinced…

Proof

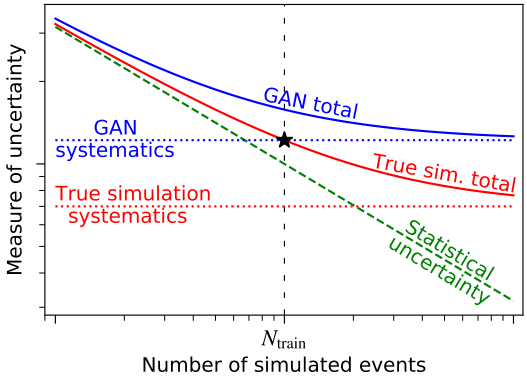
The GAN training can always be performed at the analysis stage.



The theorem is a consequence of the **data processing inequality**.

Some technical details…

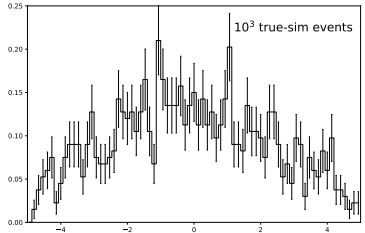
- Top mass measurement…
  Let's say we have N true-simulated events produced with m_t set to 173 GeV.
- Let's say this dataset is similarly consistent with $m_t = 173$ GeV and $m_t = 175$ GeV.
- Train a GAN on this dataset. Will the learned distribution correspond to 173 GeV or 175 GeV? (answer: neither)



The uncertainties from the training data get frozen into the GAN generated data as a systematic uncertainty.

Amplification for sub-optimal analyses? Maybe.

Now, **Person B** is doing a sub-optimal analysis—histogramming some event variable.



A: *You're missing out on all the high-dim information! That's sad.*

B: *I know, but I'm just too lazy to improve my analyses :(*

A: *Don't worry! Let me train a GAN, and produce a large number of events from it.*
You can do the exact same analysis, but your sensitivity will be better!

B: *Fantastic! Let's do it!*

Amplification for sub-optimal analyses? Maybe.

Now, **Person B** is doing a sub-optimal analysis—histogramming some event variable.

A: *You're missing out on all the high-dim information! That's sad.*

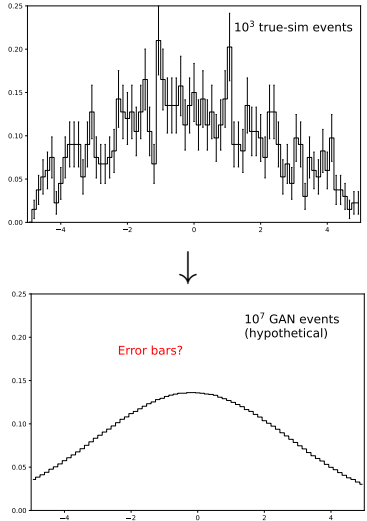
B: *I know, but I'm just too lazy to improve my analyses :(*

A: *Don't worry! Let me train a GAN, and produce a large number of events from it.*
You can do the exact same analysis, but your sensitivity will be better!

B: *Fantastic! Let's do it!*

B: *Okay, so what are the error bars on the GAN-generated histograms?*

There is no precedent in the literature for providing these uncertainties



Amplification for sub-optimal analyses? Maybe.

- Without knowing what the uncertainties are, the amplification (if any) cannot be tapped into.
- Saying the uncertainties are probably better than before using the GAN is not good enough. We need to reduce the reported uncertainties!
- One technique in the literature to get the error bars is to compare the GAN generated histograms against the true distribution, **using the functional form** of the distribution. The functional form is not available in realistic situations, hence the need for simulations in the first place.
- Another approach would be to produce more and more true-sim data and see where the agreement breaks down. But there's no speed-up anymore!
- Although leverageable amplification is yet to be demonstrated, it is not impossible.

Amplification for sub-optimal analyses? Maybe.

- **For a given, sub-optimal analysis**, it is not impossible to *predict* the amplification provided by the GAN.
  (bootstrapping techniques, Bayesian networks, p(arton)df literature, etc. might help)
- Even then, GANs can only help us utilize **the available true-sim data** better. **It would be an analysis technique, not a data generation technique.**
- GANs are not the only techniques we have for tapping into high-dimensional information
  (event selection and categorization, weighted histograms, multivariate techniques, matrix element method, machine learning, other curve fitting techniques, etc.)
- The reason for requiring higher simulation statistics is to **go beyond what can be achieved with interpolation**.
  GANs cannot satisfy this need.
  not to mention, the risk of interpolating (at the analysis stage) from an interpolation (in the MC stage)

Additional content in the paper

- Math: We show that the GANned dataset can be no better than the training dataset in terms of Mutual Information, Fisher Information, and Kullback–Leibler Divergence (Section 3).
- We reconcile our claims with the seemingly incompatible results in the literature (Section 6).
- We **identify some potential uses of GANs**, where the applicability is not completely ruled out by the data processing inequality (Section 7).
- We demonstrate our arguments with a toy example (Appendix A).

Outlook:

- The computational resource crunch is a very real, upcoming crisis in HEP.
- It is not entirely clear whether the technical and technological breakthroughs required to overcome it will even happen.
- In my opinion, we are probably putting a few too many eggs in the GAN basket.

Thank you!