CERN openlab

IML Summary:
GANplifying event samples
Topological data anaysis

Students' meeting, October 27

Kristina Jaruskova

GANplifying event samples

Anja Butter, Sascha Diefenbacher, Gregor Kasieczka, Benjamin Nachman, Tilman Plehn

(Universit¨at Heidelberg, Germany, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, DESY, Germany)

IML presentation

https://indico.cern.ch/event/852553/contributions/4061795/attachments/2126569/3580420/CERNiml_GANplify.pdf

Article

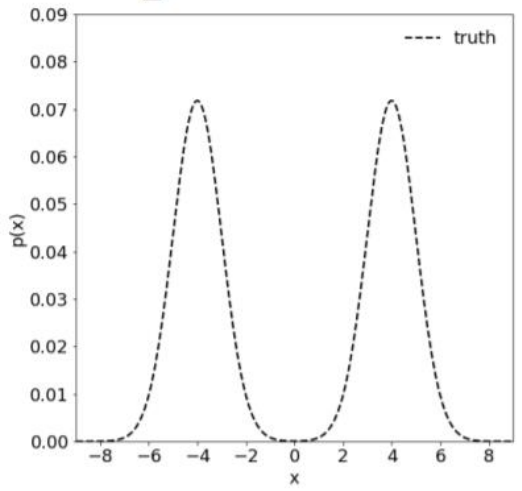
https://arxiv.org/pdf/2008.06545.pdf

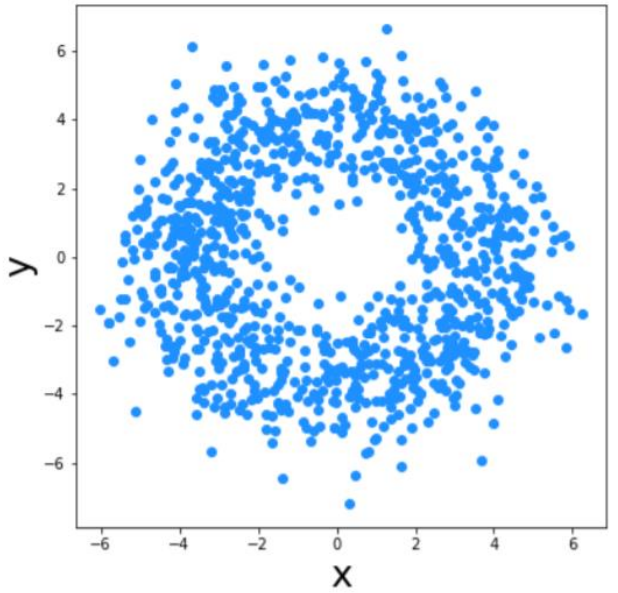
Intro

- Can generated events add statistical precision beyond the training sample?
- i.e. If GAN is trained on N data points, is it possible to draw from the GAN more data?

- Test using toy example
  - 1-D model: camel back function
  - 2-D model: ring
    (2D analogue to camel back)

$$p(X) = \frac{1}{2}(N_{-4,1}(x) + N_{4,1}(x)$$

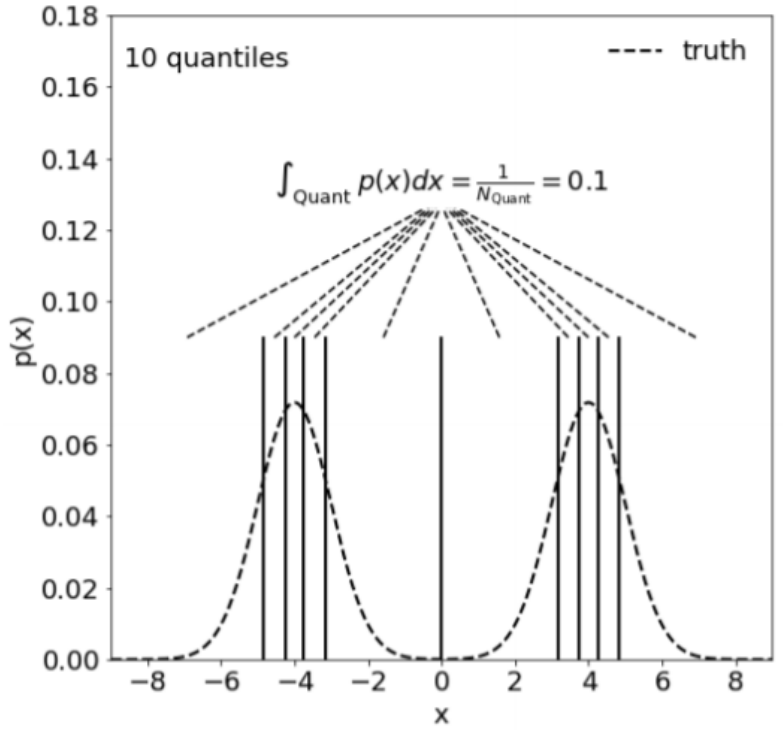




1-D model: Camel back function

Quantiles

- Measurement how well function is described
- Define N quantiles on true distribution
- Each quantile contains equal probability



1-D model: Camel back function

- Draw 100 points from true distribution – training data
  - Empirical quantiles – baseline comparison
- Use training data to fit 5 parameter camelback function

$$p(X) = a\, N_{\mu_1,\sigma_1}(x) + (1-a) N_{\mu_2,\sigma_2}(x)$$

- Train GAN on training data and generate new samples
- Calculate fraction of samples in each quantile

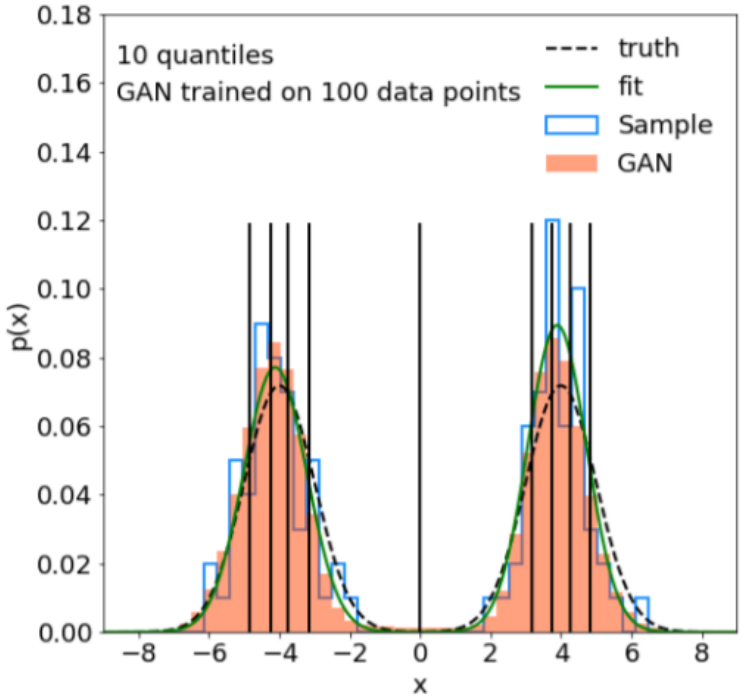
  -> compare error for training samples, parameter fitting and GAN

1-D model: Camel back function

Generative Network

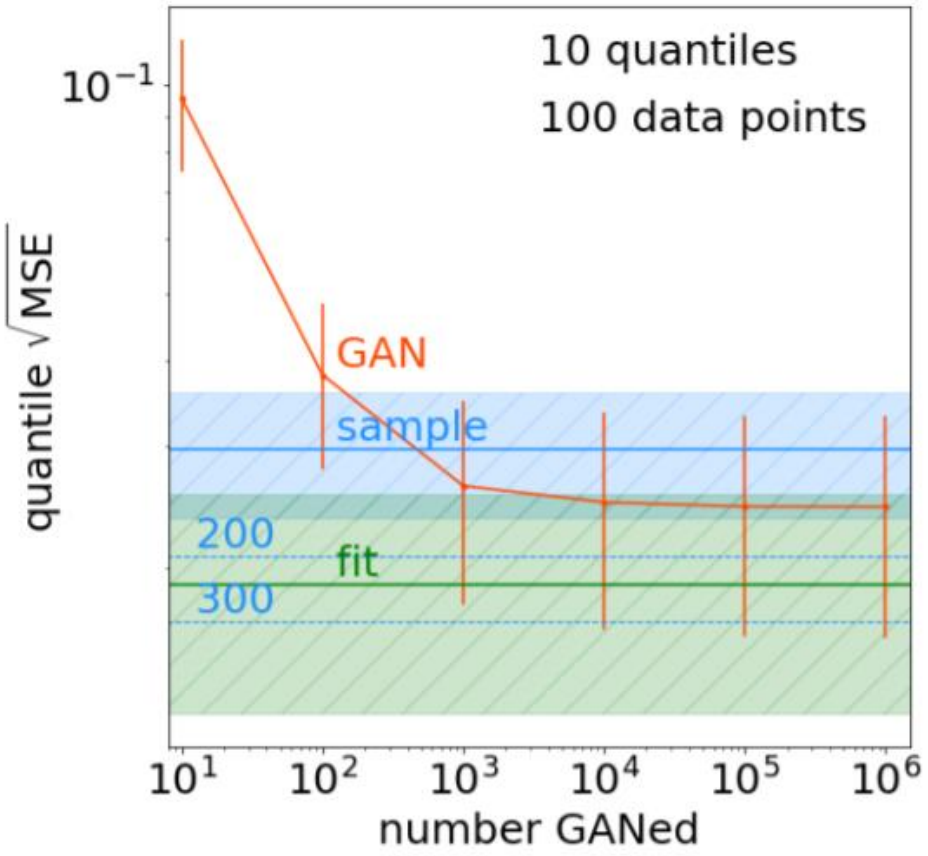
- Generate $O(10^7)$ data points using GAN
- Calculate fraction of points in each quantile
- Define quantile MSE:

$$\mathrm{MSE} = \frac{1}{N_{\mathrm{quant}}} \sum_{j=1}^{N_{\mathrm{quant}}} \left(x_j - \frac{1}{N_{\mathrm{quant}}} \right)^2$$



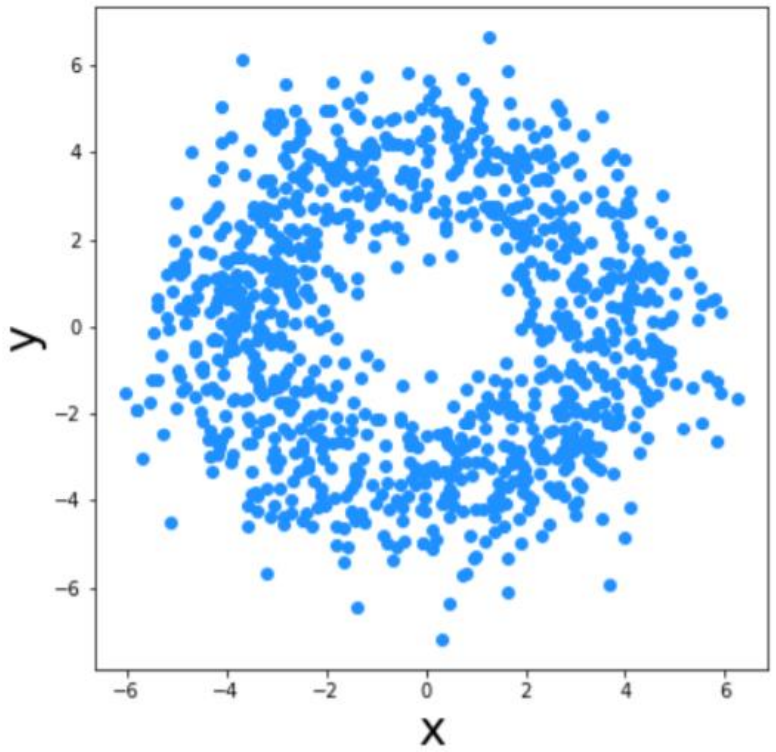
Results

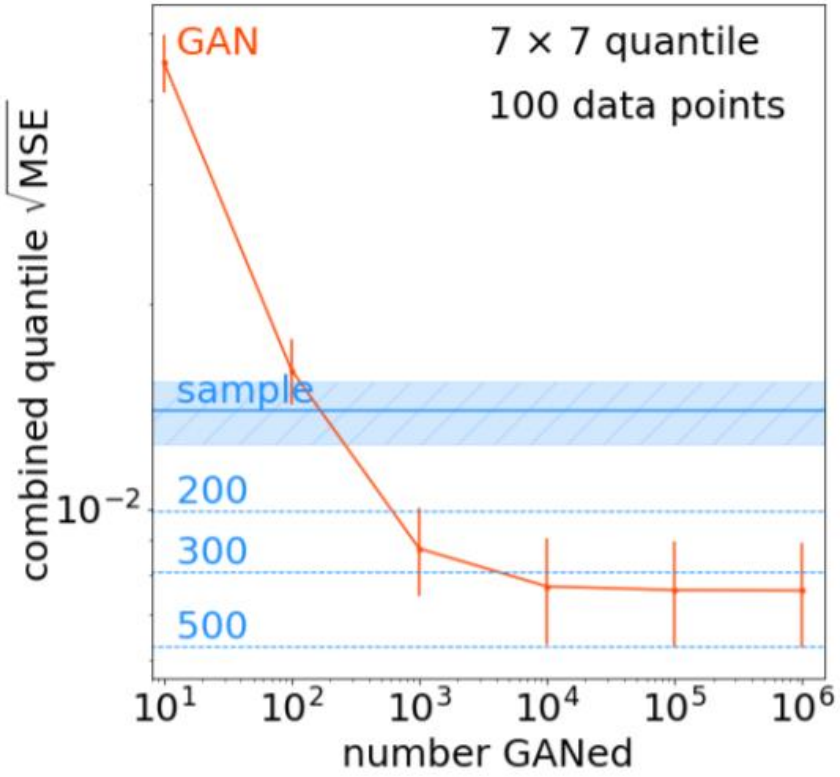
- GAN better than training data
- Needs 10 000 GAN generated point to match 150 true samples
- Explanation:
  - Sample: only data points
  - Fit: data + true function
  - GAN: data + smooth continuous function -> interpolation
- If dataset allows for smooth interpolation -> more points than in training set can be drawn from GAN



2D model: Ring

- Similar behaviour -> GAN interpolates in the 2D space as well





Using Topological Data Analysis to Disentangle Complex Datasets

Maurizio Sanarico

IML presentation

https://indico.cern.ch/event/852553/contributions/4065109/attachments/2126042/3579483/Cern_Sanarico-Mapper_1.pdf

General paper about TDA

https://arxiv.org/pdf/1710.04019.pdf

Intro to TDA

- **SDG Group** – management consulting group specialized in Business Analytics and Corporate Performance Management
- **Topology** – branch of mathematics, formalizes notions of proximity and continuity
- Informally: topology on a set is a description of how elements in the set are spatially related
- **TDA** – combines topology, geometry, statistics, computing
  - "a collection of powerful tools that can quantify shape and structure in data in
  - order to answer questions from the data's domain." [Elizabeth Munch]
- **Topological signature** – simplified representation of the topology of a given space

THE GOAL OF TDA

- Basic Idea: **Data has shape**
- This shape can be rigorously quantified with topology:

  Using topological signatures
- Such signatures act as **summaries of the data**
- The Goal of TDA:

  Use tools from topology to make meaningful signatures of the data

  Topological signatures lead to topological **invariants**, and such invariants ***enable greater understanding of the relationships in—and transformations of—real data***

sdg

7

TDA

- Suitable for high-dimensional and complex data.
- Interpretable
- Giotto TDA – topological ML toolbox in Python

- Presented usecase – Amazon Reviews
- Applications
  - Discover sequences of alarms/warnings, identifying the first one and characterize the impact on performance on a packaging line with 1800 types of alarms and warnings (Alarm flood problem).
  - Cluster WES data (Whole Exome Sequence) to characterize the genetic contribution to Covid19 severity (still to be submitted for publication).
  - Use in a NLP pipeline to classify CRM (Customer Relationship Management) messages into categories based on the content.