Data Science applications to accelerate High Energy Physics

DQE presentation (05/07/2021)

Irina Espejo (NYU)

Overview

I have been funded by FULBRIGHT Google DeepMind

and last 2 years by the NSF iris hep

on the grants NSF (OAC-1841471)
NSF (ACI-1450310)

Scalable CyberInfrastructure for Artificial Intelligence and Likelihood Free Inference (SCAILFIN)

The SCAILFIN Project

IRIS-HEP poster session for the NSF 2020 at Princeton

Scalable cyberinfrastructure applications

Team: J. Brehmer[1 2], K. Cranmer[1 2], **Irina Espejo**[1], S. Macaluso[1 2] and H. Müller[1]
Institutions: [1] Center for Data Science, New York University
[2] Department of Physics, New York University

The SCAILFIN Project

MadMiner on REANA

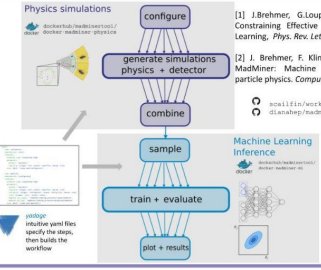
Simulators provide a *high-fidelity description* of phenomena. If they are too complex the likelihood is intractable leading to *poor statistical analysis*

MadMiner (Machine learning-based inference for particle physics [1], [2]) .
Its goal is to estimate parameters using advances in *Neural Networks*.

To deploy MadMiner in a scalable way we need:
- ✓ containerization
- ✓ parallelization
- ✓ reusability

reana

Physics simulations



yadage
intuitive yaml files specify the steps, then builds the workflow

[1] J.Brehmer, G.Louppe, J.Pavez and K.Cranmer. Constraining Effective Field Theories with Machine Learning, *Phys. Rev. Lett. 2018* 121, 111801

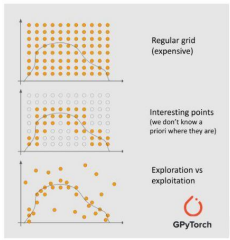
[2] J. Brehmer, F. Kling, I. Espejo and K. Cranmer. MadMiner: Machine learning-based inference for particle physics. *Comput.Softw.Big Sci.* 4 3 (2020)

scailfin/workflow-madminer
dianahep/madminer

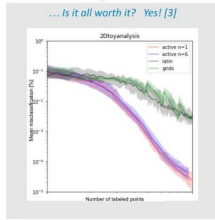
Excursion

in collaboration with G. Louppe[3] and L. Heinrich[3]
[3]University of Liège, [4]CERN

- Goal is to find *level sets of black-box functions* that are expensive to evaluate. Examples: test statistics from complex simulations.
- Evaluate the black box function at *interesting points only* instead of evaluating at whole regular grid. We use a *Gaussian process* to: interpolate between samples and model uncertainty in the knowledge of the black box function.
  The *acquisition function* regulates the exploration vs exploitation tradeoff. Select one that *minimizes global uncertainty* of the location of the excursion set.
- Future: efforts will focus on *scalability* wrt the dimensionality of the function domain space. Example, likelihood ratio as function of mass, charge, spin,...



diana-hep/excursion
irinaespejo/excursion

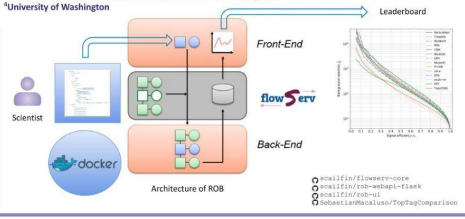


[3] L. Heinrich, G. Louppe, K. Cranmer, *Excursion Set Estimation using Sequential Entropy Reduction for Efficient Searches for New Physics at the LHC*, ACAT 2019

ROB
Reproducible Open Benchmark Platform

in collaboration with S. Hsu[4], A. Maritz[4], A. Rawat[4] and C. Suaysom[4]
[4]University of Washington

- ROB is an experimental prototype for enabling *community benchmarks* of data analysis algorithms. The goal of ROB is to allow user communities to evaluate the performance of their different data analysis algorithms in a *reproducible competition-style* format.
- The *workflow template* and input data are defined by a coordinator. The template contains placeholders for workflow steps that are implemented by the participants (e.g., with Docker containers). The backend processes the submission workflows. The user interface allows participants to *submit new runs* and to view the results.



scailfin/flowserv-core
scailfin/rob-webapi-flask
scailfin/rob-ui
SebastianMacaluso/TopTagComparison

National Science Foundation (OAC-1841471)
National Science Foundation (ACI-1450310)

dianahep NYU Center for Data Science MSDSE DeepMind FULBRIGHT

This project is supported by National Science Foundation under Cooperative Agreement OAC-1836650. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Acknowledgements


Kyle Cranmer (NYU)


Lukas Heinrich (CERN, former NYU)


Johann Brehmer (Qualcomm, former Moore-Sloan NYU)


Felix Kling (SLAC)


Sinclert Pérez (NYU DS3)


Gilles Louppe (U. Liège, former Moore-Sloan NYU)


Juan Pavez (U. Tecnica Federico Santa Maria)


Leonora Vesterbacka (former NYU)


Heiko Müller (NYU DS3)


Eleni Skorda (Lund University)


Janik von Ahnen (DESY)


Patrick Rieck (MPI)


Paul Gadow (DESY)


Chris Hollowell (BNL)


Carl Evans (NYU HPC)


Ben Galewsky (NCSA)

reana

Developer Team, special thanks to Tibor Simko

Roadmap

1. Introduction
2. **MadMiner deployment on REANA**
3. **Active Learning for Excursion Set estimation**
4. Impact
5. Summary
6. Further work

1. Introduction

Statement for my main line of research

I want to develop new techniques and cyber-infrastructure to streamline the task of testing particle physics theories and address bottlenecks within the current approach.

Challenges encountered: a) a vast number of theories to test b) an expensive simulation-based inference pipeline and c) complex generative model implemented with legacy software.

Overcoming these challenges will help me understand how to bridge between Natural Sciences and Data Science where the end goal is scientific discovery.

1. Introduction: general context

Data at the LHC consists of many iid collisions with associated high dimensional continuous features $x \in \mathbb{R}^d$

Different categories (indexed by c) are defined according to an indicator function $1^{(c)}_{\{(h(x^{(i)})=0,\, g(x^{(i)})>0)_{i=1\ldots d}\}}(x)$

Let n be the number of observed collisions that satisfy the indicator function and $\nu = \mathbb{E}[n]$

The dataset consists of $\mathcal{D} = \{x_1, .., x_n\}$

The **probability model** is a Poisson point process $P(\mathcal{D}|\nu) = \text{Pois}(n|\nu) \prod_{e=1}^{n} p(x_e)$

discrete continuous

Furthermore, the total intensity decomposes in different physical processes (Mixture model)

$$p(x_e) = \frac{1}{\nu} \sum_{i \in \text{processes}} \nu_i p_i(x_e)$$

with $\nu = \sum_{i \in \text{processes}} \nu_i$

1. Introduction: adding physics

Different theories will predict different distributions and coefficients for the mixture components

with reparametrization $\nu_i \rightarrow \nu_i(\theta, \mu),\ p_i(x) \rightarrow p_i(x|\theta, \mu),$

where are θ parameters of interest and μ are nuisance parameters

Taking multiple categories into account the model becomes

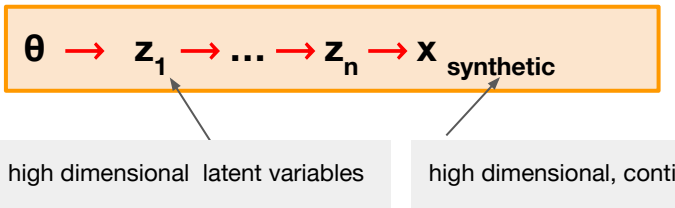
The model

$$p(\mathcal{D}|\theta, \mu) = \prod_{c \in \text{categories}} \text{Pois}(n_c|\nu_c(\theta, \mu)) \prod_{e=1}^{n_c} p_c(x_{e,c}|\theta, \mu)$$

1. Introduction: intractability

However the probability density and nu/total intensity are intractable and implicitly defined by the simulation !

Generative model
(forward mode only)



$$p(x|\theta) = \int dz_\cdot \ldots dz_n \, p(x|z_n) \ldots p(z_i|z_{i-1}) \ldots p(z_1|\theta)$$

intractable likelihood

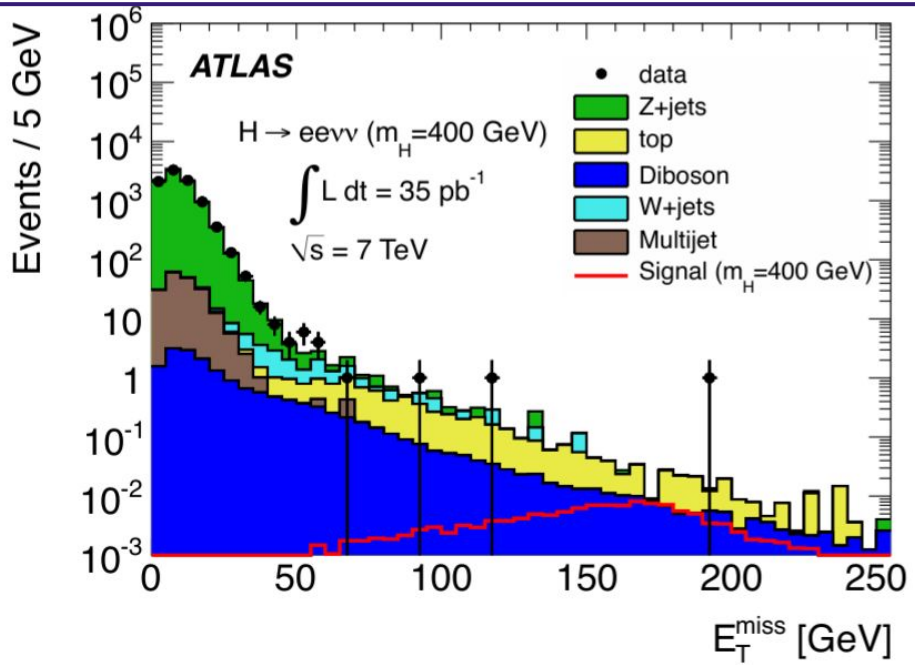
Therefore physicists work with low (1,2) dimensional summary statistics: $x \longrightarrow s(x)$ using feature engineering based on domain-knowledge.

They estimate the $p_i(s)$ and ν_i with density estimation (histogram) and samples from the simulation $\hat{p}_i(s)$, $\hat{\nu}_i$

This is expensive! In order to do this we need to run the simulation and analysis pipeline which includes the code to implement the indicator functions as well as the summary statistic.

1. Introduction: traditional inference strategy

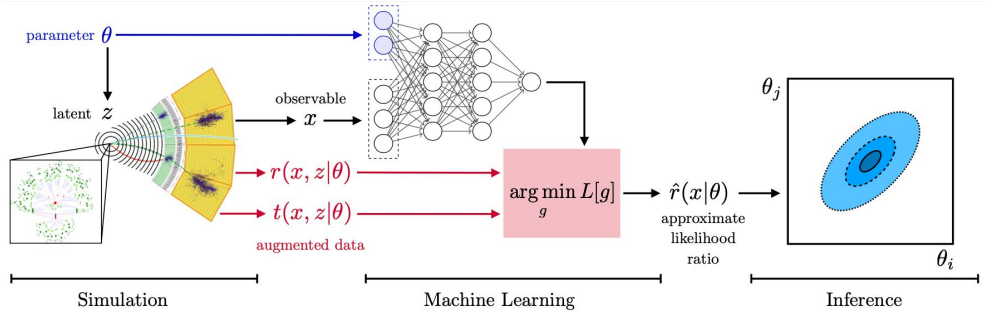
$$\hat{p}(\mathcal{D}|\theta,\mu) = \prod_{c\in\text{categories}} \text{Pois}(n_c|\hat{\nu}_c(\theta,\mu)) \prod_{e=1}^{n_c} \hat{p}_c(s_{e,c}|\theta,\mu)$$



1. Introduction: Two strategies for discovery in High Energy Physics

Indirect searches - MadMiner

- The null $\boldsymbol{\theta} = 0$ is nested in the family $p(x|\boldsymbol{\theta})$.
- Perform inference for a parameter $\boldsymbol{\theta}$ and hope to find an inconsistency.
- Use ML allows to get rid of summary statistics.



Direct searches - Excursion

- Select a theoretical candidate and proceed with a classic hypothesis testing pipeline.
- $$\hat{p}(s|\theta) = \frac{1}{\hat{\nu}(\theta)} \sum_{i \in \text{processes}} \hat{\nu}_i \hat{p}_i(s) + \hat{\nu}_{\text{new}}(\theta) \hat{p}_{\text{new}}(s|\theta)$$

The null is not nested in the class of alternative models.

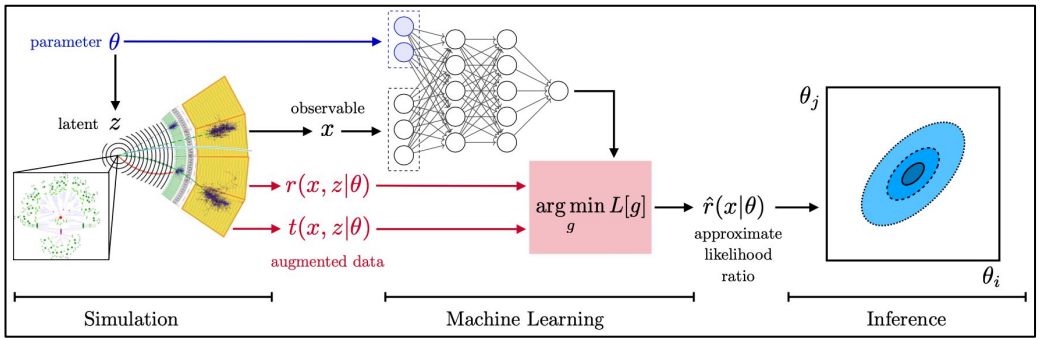
RECAST is a framework that adds the additional term to the mixture model for testing and works with the same summary statistic and indicator functions.

2. MadMiner deployment on REANA

2. MadMiner deployment on REANA

MadMiner Brehmer et al., *PRL*, (2018) is an extensive set of new simulation-based techniques based on machine learning that removes the need for summary statistics and improves the power of the resulting statistical tests. MadMiner is a step forward from traditional histogram-based analysis. One study showed it obtains similar constraints with 90% less data. diana-hep/madminer

from Brehmer et al., *PRL*, (2018)



- Integrate different software components
- Complicated environments (simulator)
- Complicated to use despite its results.

Challenge (my work): **successful deployment requires addressing scalability and reproducibility issues.**

J. Brehmer, F. Kling, I. Espejo, K. Cranmer. Madminer: Machine learning-based inference for particle Physics. *Computing and Software for Big Science* 4, 3. (2020)

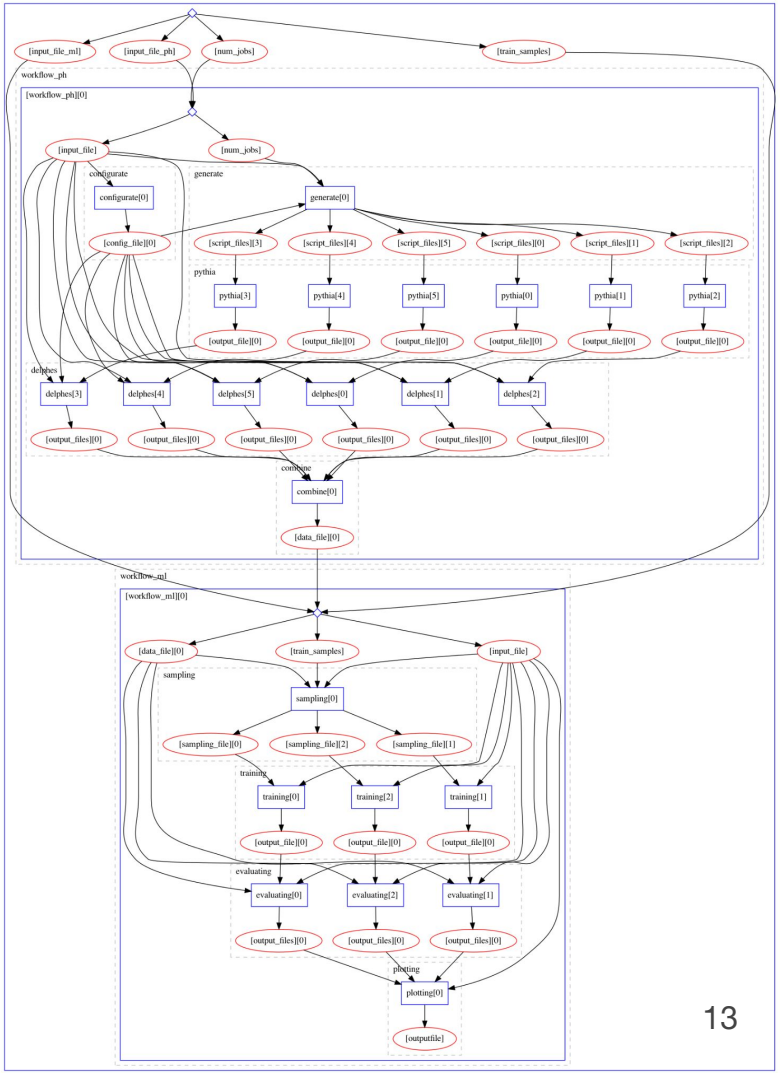
2. MadMiner deployment on REANA

REANA is a reproducible research data analysis platform. **Scalable**, **reusable**, **open software** and **flexible**.

scailfin/madminer-workflow

MadMiner is now deployed at BNL and NYU HPC besides CERN

This has generated interest in the community so I created a tutorial followed by 100+ grad students.



2. MadMiner deployment on REANA at NYU

REANA — Mozilla Firefox

REANA

https://reana.hpc.nyu.edu/details/6...

reana

Help

madminer-workflow #11
Finished 15 days ago

finished in 6 min 9 sec
step 21/1

Logs | Workspace | Specification

Name	Modified	Size
workflow.yml	2021-02-16T03:05:09	908
ph/input.yml	2021-02-16T03:05:07	1421
ml/input.yml	2021-02-16T03:05:08	2425
_yadage/yadage_snapshot_w...	2021-02-16T03:11:17	149428
_yadage/yadage_workflow_in...	2021-02-16T03:11:18	19984
_yadage/yadage_workflow_in...	2021-02-16T03:11:18	563370
_yadage/yadage_workflow_in...	2021-02-16T03:11:18	30782
ph/yadage/steps.yml	2021-02-16T03:05:07	2029
ph/yadage/workflow.yml	2021-02-16T03:05:07	1740
ml/yadage/steps.yml	2021-02-16T03:05:08	1987
ml/yadage/workflow.yml	2021-02-16T03:05:08	2027
_yadage/adage/adagesnap.txt	2021-02-16T03:11:17	40298
_yadage/adage/workflow.gif	2021-02-16T03:11:17	1699515
workflow_ph/generate/py.py	2021-02-16T03:05:42	8778
workflow_ph/generate/folde...	2021-02-16T03:05:47	5409816

REANA — Mozilla Firefox

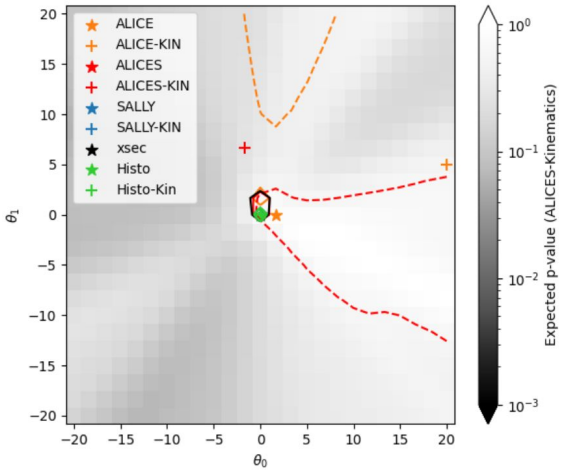
REANA

https://reana.hpc.nyu.edu/details/6ade3...

madminer-workflow #11
Finished 15 days ago

finished in 6 min 9 sec
step 21/1

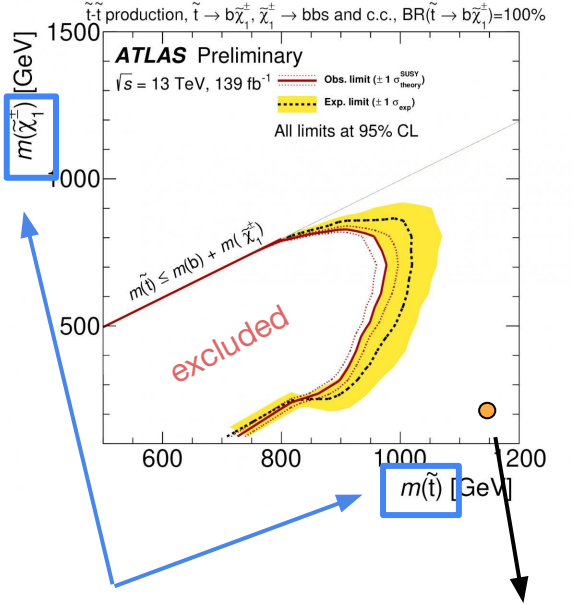
workflow_ml/plotting/plots/all_methods.png



Download

3. Active Learning for Excursion Set Estimation

Motivation

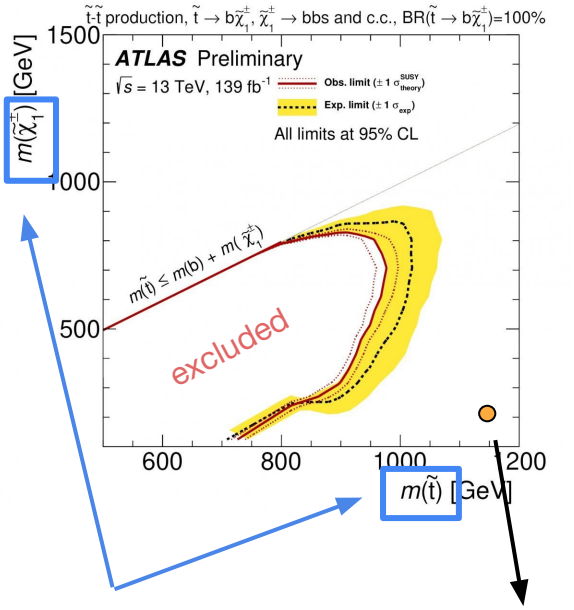


$$\hat{p}(s|\theta) = \frac{1}{\hat{\nu}(\theta)} \sum_{i \in \text{processes}} \hat{\nu}_i \hat{p}_i(s) + \hat{\nu}_{\text{new}}(\theta)\hat{p}_{\text{new}}(s|\theta)$$

parameters θ

high latency: computing each point takes days, weeks

Motivation

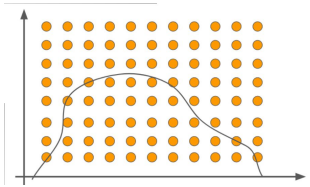


parameters θ

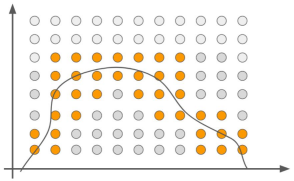
high latency: computing each point takes days, weeks

The naive approach suffers from the curse of dimensionality.
This is a bottleneck for use case theories with 5 or 19 dim.
It is <u>infeasible</u> to make exclusion plots in higher dimensions using ATLAS current approach, *we need a more efficient approach*.

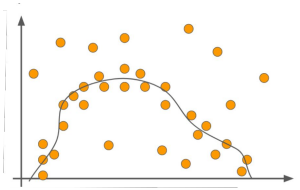
Idea



Current ATLAS approach



Better queries



Excursion method

From Lukas' Talk at ACAT 2019

Active Learning

Statement

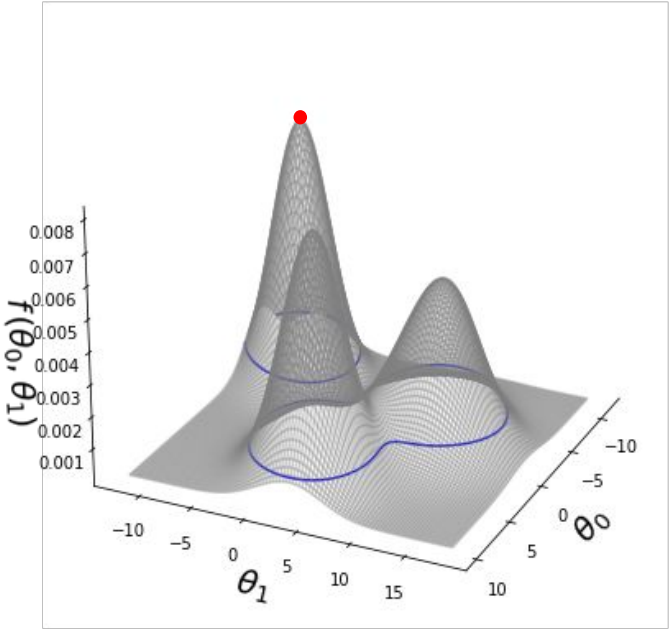
The black box function is $f(\boldsymbol{\theta}) \rightarrow$ *p-value*.

The goal is to find the excursion set $E_t(f) = \{\boldsymbol{\theta} | f(\boldsymbol{\theta}) = t\}$ for a given a threshold *t* with as few queries as possible.

Method

1. Start with dataset $\mathcal{D} = \{\boldsymbol{\theta}_i, f(\boldsymbol{\theta}_i)\}$
2. Train a Gaussian process $Y|\boldsymbol{\theta}, \mathcal{D}$ with predictive mean $\mu_{Y|\mathcal{D}}(\boldsymbol{\theta})$ and covariance $\Sigma^2_{Y|\mathcal{D}}(\boldsymbol{\theta}, \boldsymbol{\theta}')$
3. Evaluate the acquisition function $U_t(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ (cheap) using $Y|\boldsymbol{\theta}, \mathcal{D}$
4. Select new point $\boldsymbol{\theta}^* = \text{argmax} \quad U_t(\boldsymbol{\theta})$
5. Query the simulator at $f(\boldsymbol{\theta}^*)$ and update dataset $\mathcal{D} \leftarrow \mathcal{D} \cup (\boldsymbol{\theta}^*, f(\boldsymbol{\theta}^*))$

Bayesian Optimization vs Level Set Estimation



Bayesian Optimization	Level Set Estimation
Find a point	Find a curve
Extensive literature	Little literature

Acquisition functions

Consider the level set estimation problem as a classification problem for the parameter points over a subjacent grid

$$Z|\boldsymbol{\theta} \sim Bernoulli(S(\boldsymbol{\theta})) \quad \text{with} \quad S(\boldsymbol{\theta}) = \int_{-\infty}^{t} p(Y = y|\boldsymbol{\theta}, \mathcal{D})dy \quad \text{and consider the entropy}$$

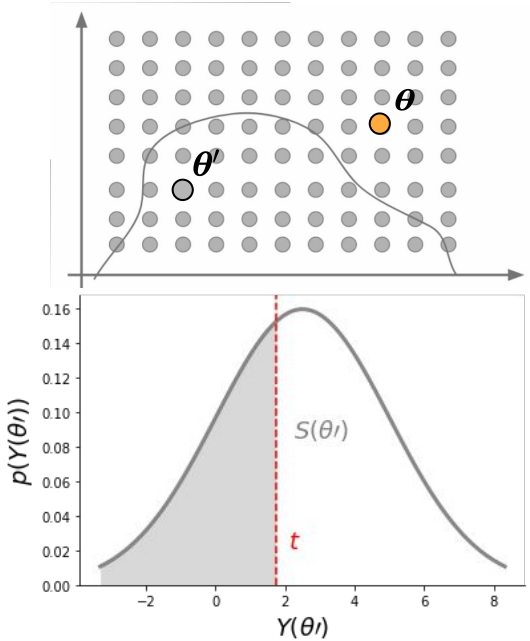
$$H[Z|\boldsymbol{\theta}] = -S(\boldsymbol{\theta}) \log S(\boldsymbol{\theta}) + (S(\boldsymbol{\theta}) - 1) \log(1 - S(\boldsymbol{\theta}))$$

- ***Predictive Entropy Search (PES)*** (Hernandez-Lobato 2014)

$$U_{PES}(\boldsymbol{\theta}) = H[Z|\boldsymbol{\theta}] - \int d\boldsymbol{\theta}' \mathbb{E}_{Z'|\boldsymbol{\theta}'} H\left[Y|\boldsymbol{\theta}, Z', \boldsymbol{\theta}'\right]$$

- ***Maximum Entropy Search (MES)***

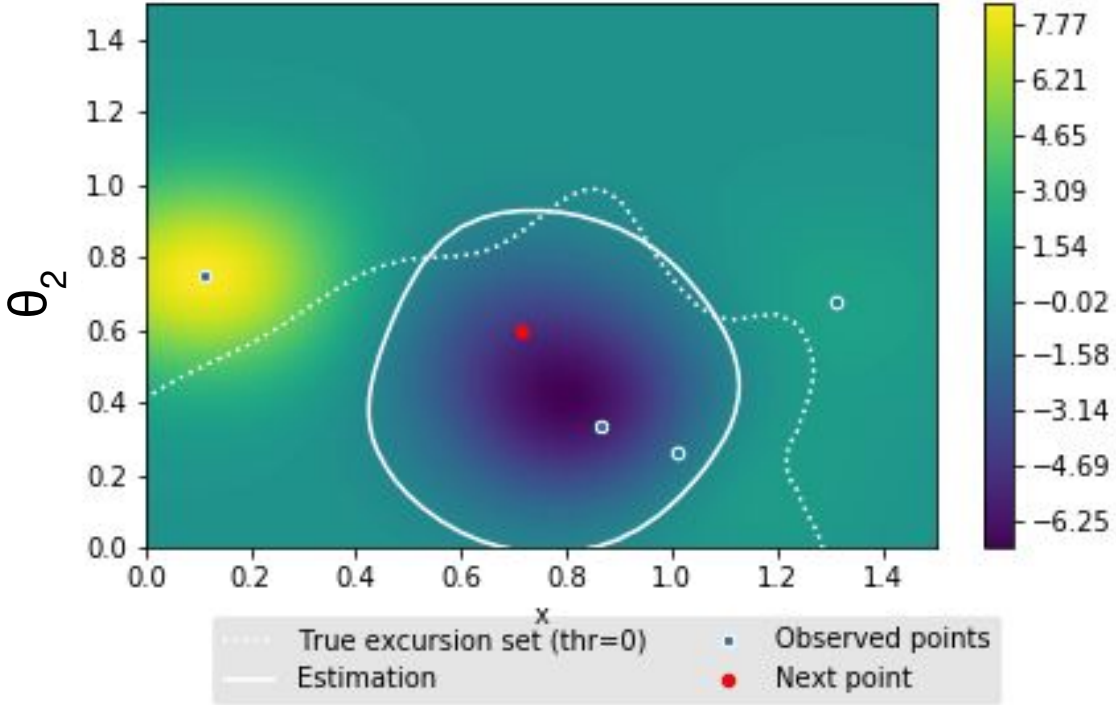
$$U_{MES}(\boldsymbol{\theta}) = H[Z|\boldsymbol{\theta}]$$



Acquisition functions

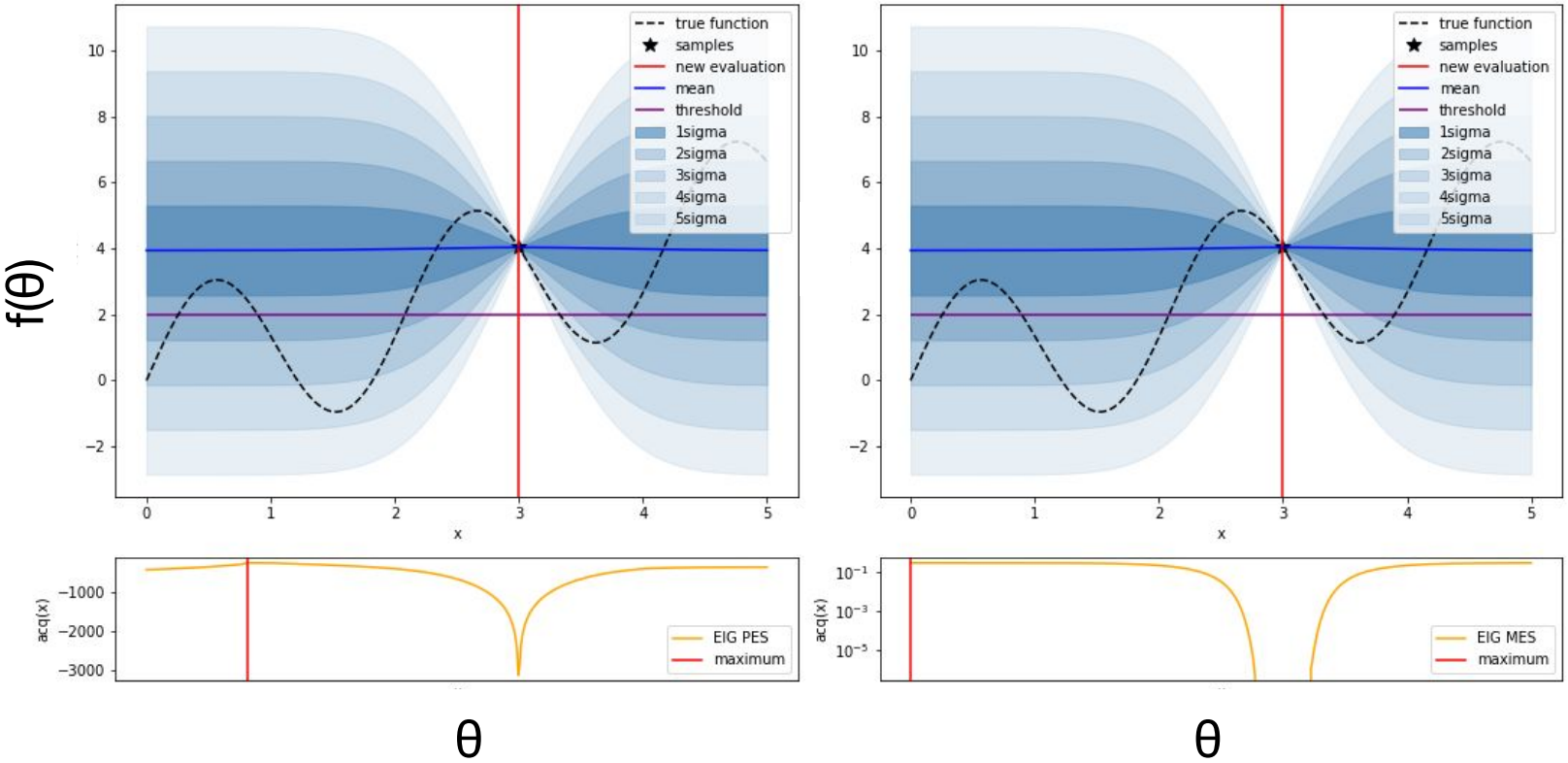
	PES	MES
Drawbacks	- The integral $\int d\boldsymbol{\theta}'$ does not scale well - The term $\mathbb{E}_{Z'\|\theta'} H\left[Y\|\boldsymbol{\theta}, Z', \boldsymbol{\theta}'\right]$ favours points far from edges - Need to approximate the entropy of $Y\|\boldsymbol{\theta},\ Z', \boldsymbol{\theta}'$ (bivariate gaussian truncated in one dimension) by a gaussian with same first and second moments.	- Not as general purpose as PES - Losing the global pairwise term might not favour level sets that do not touch the edges.

Excursion 2D



Excursion 1D

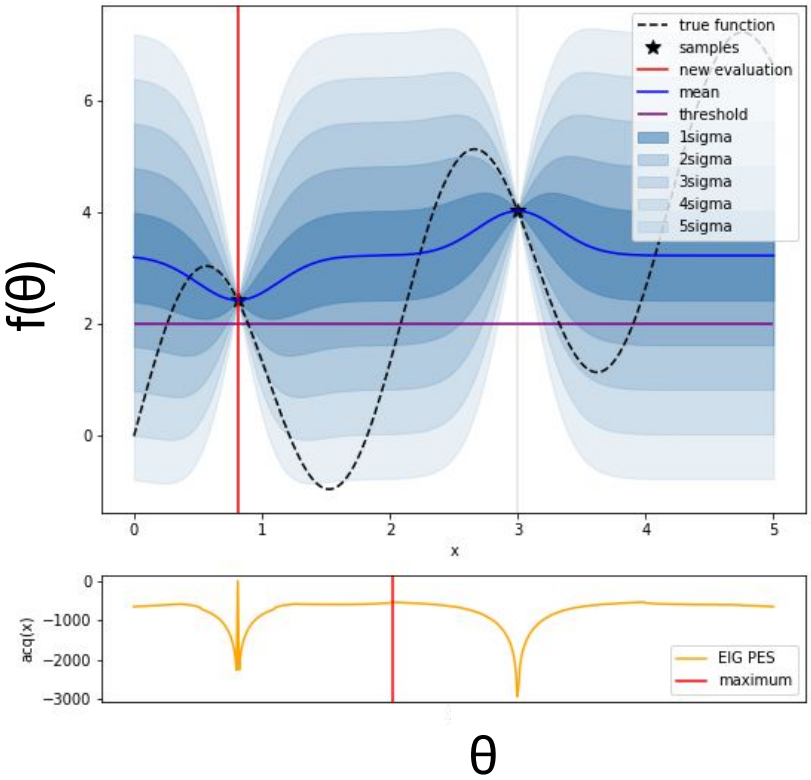
PES



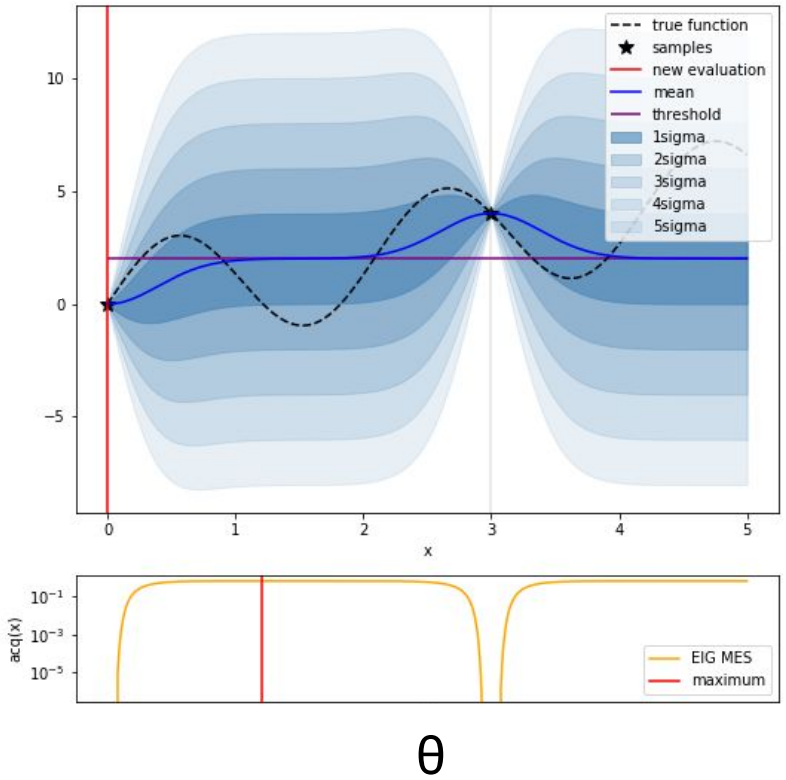
MES

Excursion 1D

PES

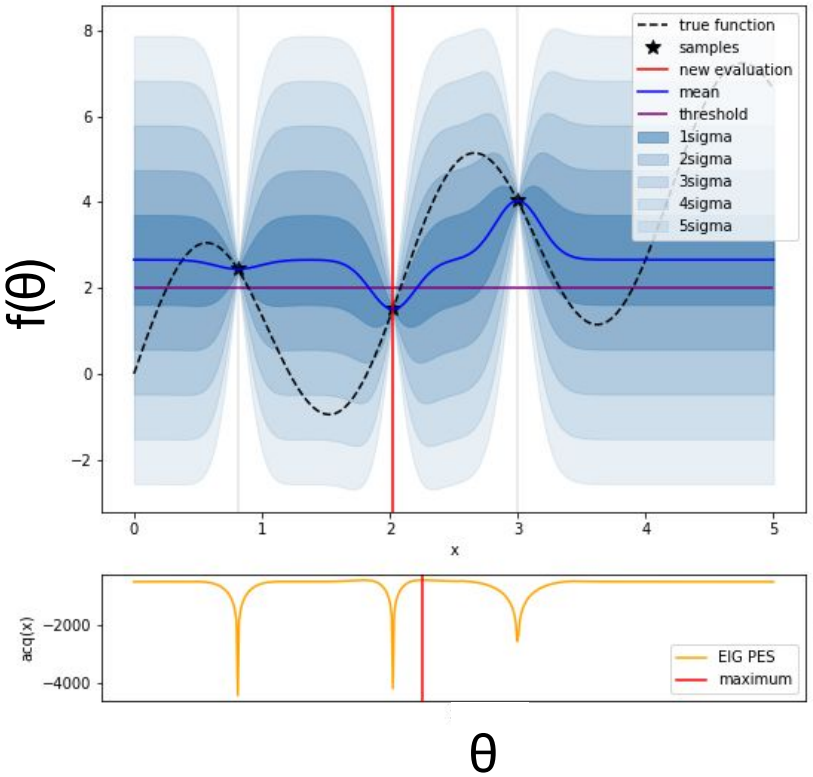


MES

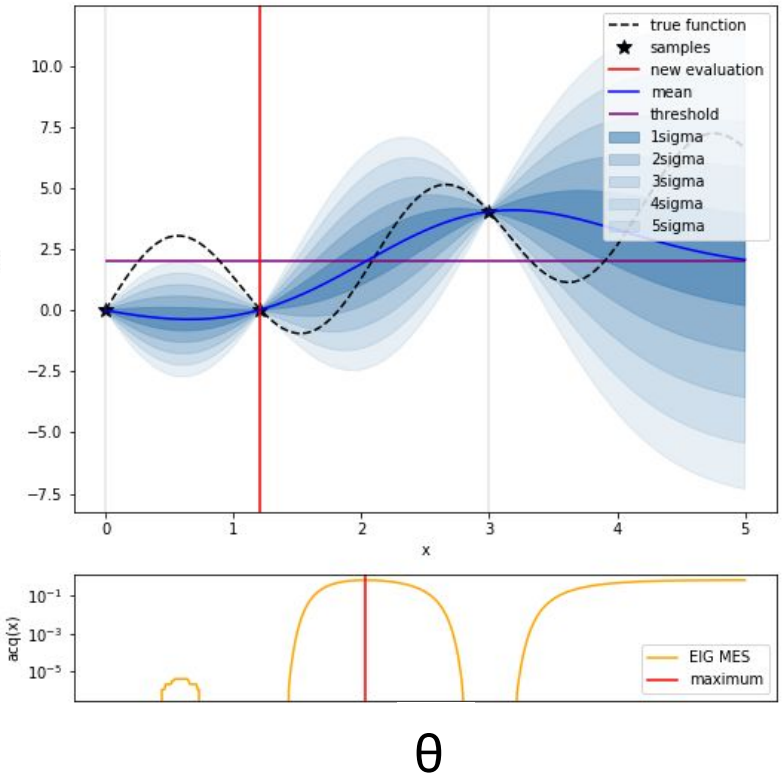


Excursion 1D

PES

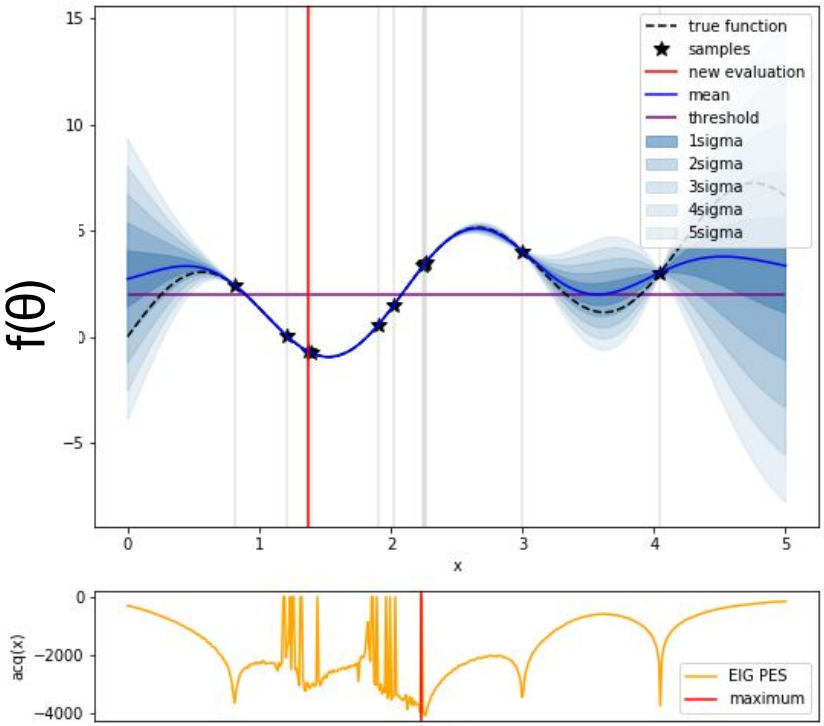


MES

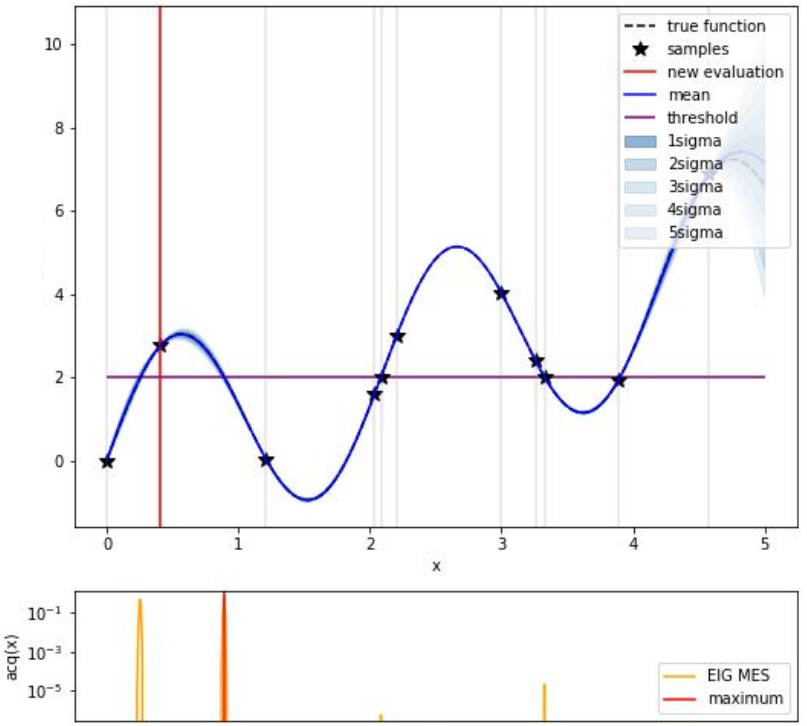


Excursion 1D

PES

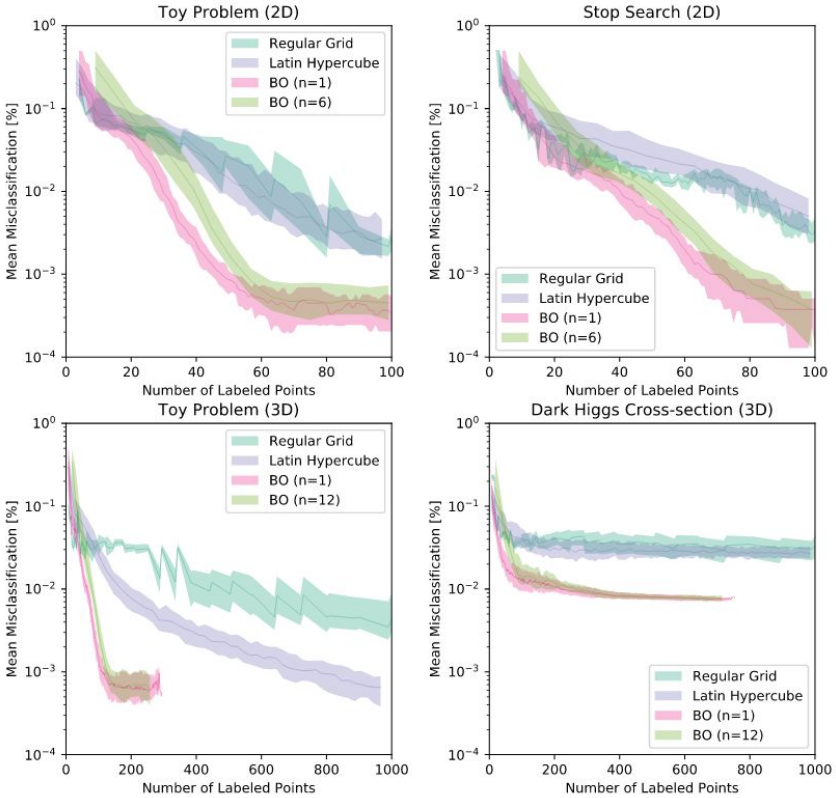


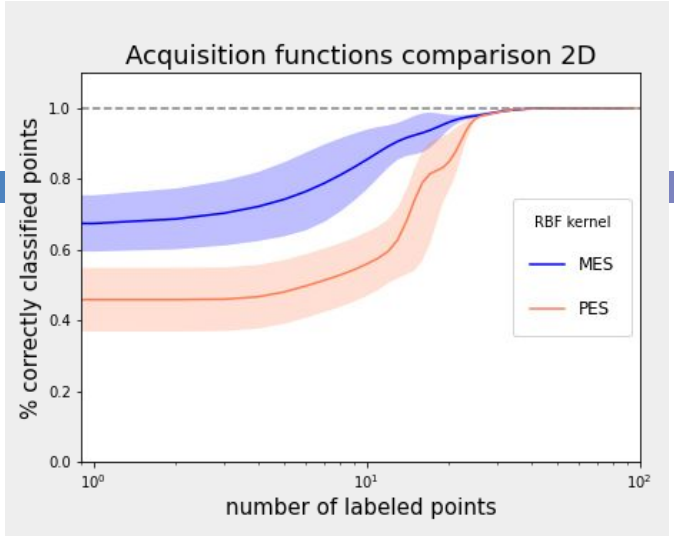
MES

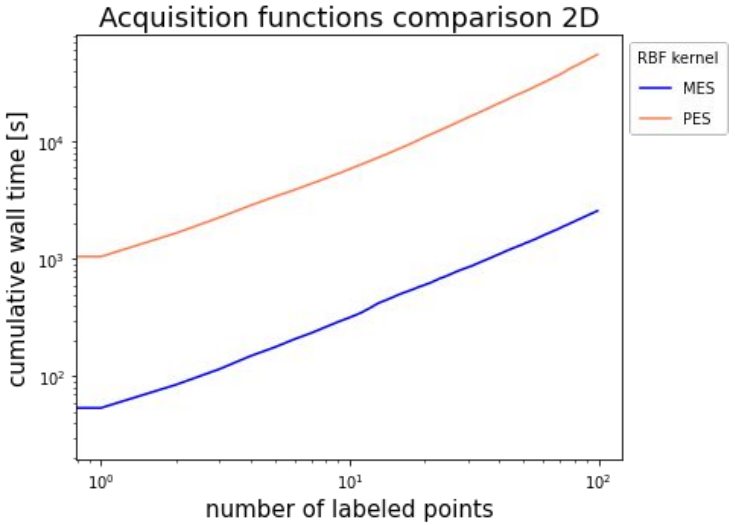


Results

From Lukas' Talk at ACAT 2019



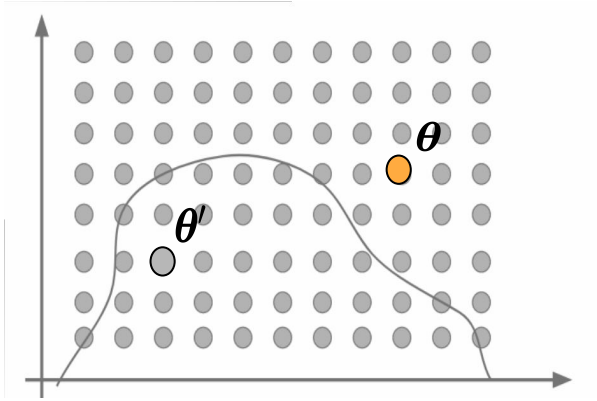


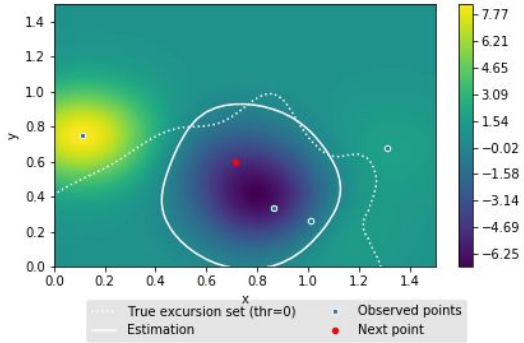


GPyTorch support

GPyTorch features

- Fast state-of-the-art posterior fitting techniques Gardner et al., *NeurIPS* (2018) ,
  Pleiss et al., *ICML*, (2019)
- Often our data has the format of a grid, so a structure-exploiting kernel such as GridRegressionKernel is faster for 2D-4D. (Wilson and Nickish, 2015)
- Lazy Tensor Evaluation and GPU support

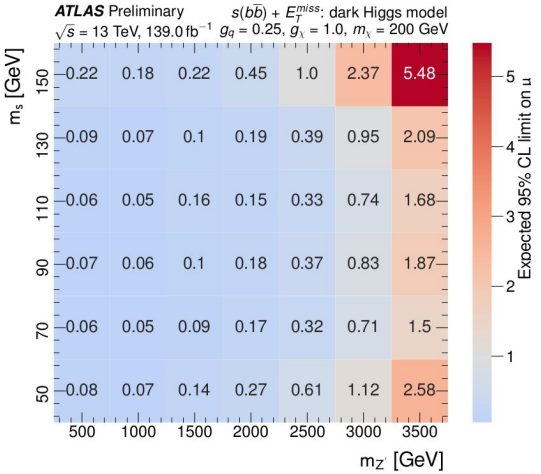




Actively Learning Exotic Physics

Application to pMSSM analysis (ongoing project in ATLAS)

- pMSSM is a popular theoretical candidate for New Physics with in 19 dimensions.
- To make it manageable we often study submodels of 2, 5 or 12 dimensions.
- Our goal is to apply Excursion to estimate exclusion contours in a scalable way (never done before).
- The black box has high latency (computing each point takes days or weeks) **so we generalize the Excursion acquisition function to optimize for batches of queries.**
- Instead of a cold start, we initialize the mean of the GP with less expensive data and apply Active Learning with the full-blown simulator as black box.



by Eleni Skorda

4. Impact

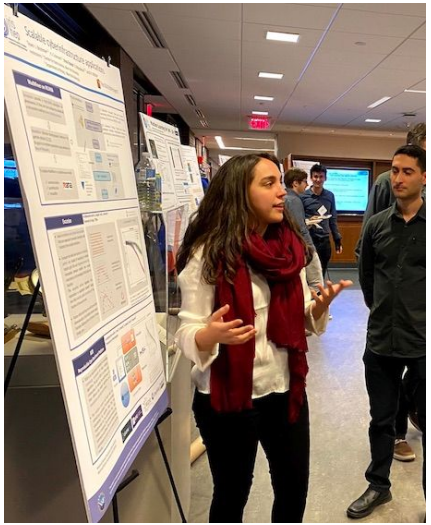
Scientific output

- Two published papers

  J. Brehmer, F. Kling, I. Espejo, K. Cranmer. Madminer: Machine learning-based inference for particle Physics. *Computing and Software for Big Science* 4, 3. (2020)

  J. Brehmer, K. Cranmer, F. Kling, G. Louppe, J. Pavez. Effective LHC measurements with matrix elements and machine learning. *Journal of Physics*: Conference Series 1525, 012022 (2020)

- Collaborated in the effort to build infrastructure at NYU and BNL to deploy MadMiner on REANA
- Invited to talks and poster sessions.



IRIS-HEP poster session 2020 at Princeton

Outreach

- Created a tutorial for MadMiner deployment on REANA followed by over 100+ grad students.
- Our work on Active Learning has been featured in the ATLAS General Physics Coordination meeting as worth pursuing.
- There are two ongoing projects in ATLAS started by this research. One of them aiming for publication by this summer.
- Joined the ATLAS Collaboration.

5. Summary

During the 3-year trajectory we have accomplished

- The MadMiner on REANA project is initiating scalability tests.

- The Active Learning project is at the stage of producing results with real data within the ATLAS experiment.

- There is a growing interest in the particle physics community about MadMiner and Active Learning.

6. Further work

- Investigate alternatives to Gaussian Processes. Up to 5 dimensions, Active Learning is feasible but with 19 dimensions it looks challenging. We would like to investigate if we can use the level set method by Osher et al. e.g it might be less probabilistic but more scalable.

- In the Active Learning line, it is interesting to think how to generalize the acquisition function when there is more than one simulator with different resolutions and costs. Which one should we query first? and after?

- Thesis proposal: I would like to perform a systematic search over the space of theories that strikes a balance between a data-driven approach and a knowledge-based approach. The outcome would be a list of the most promising theories that experimentalists could test.

Thank you. Are there any questions?

Backups

3. Smart sampling for pMSSM parameters

Motivation

pMSSM is a 19-dimensional theoretical model which is a candidate for New Physics. Often we study submodels, for this case 12-dimensional $\boldsymbol{\theta}$.

We measure a set of 12 observable variables $\{\mathcal{O}_i\}_{i=1..12}$ we know that $\mathcal{O}_i(\boldsymbol{\theta})$

so we fit $f(\boldsymbol{\theta}) = \mathcal{O}$ using a NN or another end-to-end differentiable method.

Goal: select specific tuples of parameters that are as informative as possible given the observables we observe.

3. Smart sampling for pMSSM parameters

Method

- Data: tabular with ~ 5000 samples consisting on tuples of values for the 12 parameters and the 12 observable variables.

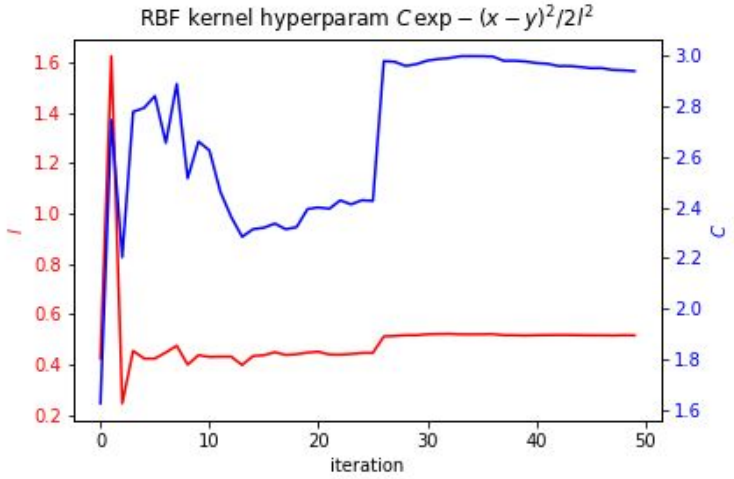
- Calculate *Jeffrey's prior* $\pi(\boldsymbol{\theta}) \propto \frac{1}{\sqrt{\det I(\boldsymbol{\theta})}}$ where $I(\boldsymbol{\theta})$ is the information matrix of the regression fit $f(\boldsymbol{\theta})$. The prior is used to reweight the space of parameters, the higher the weight the more important a parameter region is with respect to the observables.

- For that we need to estimate the derivatives $J_{ij} = \frac{\partial f_i(\boldsymbol{\theta}_{test})}{\partial \theta_j}$ with autograd and then

$$I(\boldsymbol{\theta_0}) = J^T(\boldsymbol{\theta_0})\, J(\boldsymbol{\theta_0})$$

Backup

Hyperparameter evolution as diagnostics



Backup

Wang et al., *NeurIPS*, 2019

In practice, however, exact GP inference can be intractable for large datasets, as it naïvely requires $\mathcal{O}(n^3)$ computations and $\mathcal{O}(n^2)$ storage for n training points [32]. Many approximate methods

$$\mathbb{E}\left[f(\mathbf{x}^*)\,|\,X,\mathbf{y}\right] = \mu(\mathbf{x}^*) + \mathbf{k}_{X\mathbf{x}^*}^\top \widehat{K}_{XX}^{-1}\mathbf{y}$$
$$\mathrm{Var}\left[f(\mathbf{x}^*)\,|\,X,\mathbf{y}\right] = k(\mathbf{x}^*,\mathbf{x}^*) - \mathbf{k}_{X\mathbf{x}^*}^\top \widehat{K}_{XX}^{-1}\mathbf{k}_{X\mathbf{x}^*}$$

prediction

$$\mathcal{L} = \log p(\mathbf{y}\,|\,X,\theta) \propto -\mathbf{y}^\top \widehat{K}_{XX}^{-1}\mathbf{y} - \log|\widehat{K}_{XX}|,$$
$$\frac{\partial \mathcal{L}}{\partial \theta} \propto \mathbf{y}^\top \widehat{K}_{XX}\frac{\partial \widehat{K}_{XX}^{-1}}{\partial \theta}\widehat{K}_{XX}\mathbf{y} - \mathrm{tr}\left\{\widehat{K}_{XX}^{-1}\frac{\partial \widehat{K}_{XX}}{\partial \theta}\right\}.$$

hyperparameter tuning

Backup

4.3.1 Statement of the result

Theorem 4.3.1 (Marginals and conditionals of an MVN). *Suppose* $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ *is jointly Gaussian with parameters*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} \tag{4.67}$$

Then the marginals are given by

$$\begin{aligned} p(\mathbf{x}_1) &= \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ p(\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \end{aligned} \tag{4.68}$$

and the posterior conditional is given by

$$\begin{aligned} p(\mathbf{x}_1 | \mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1} \end{aligned} \tag{4.69}$$

From Murphy

Backup

Derivation of the first and second moments of the truncated bivariate gaussian needed in PES:

$$E_j := \mathbb{E}[Y(x_{cand} \mid S(x', c_j))] = \mu(x_{cand}) - \sigma(x_{cand})\rho_j \frac{\phi(\beta_j) - \phi(\alpha_j)}{\Phi(\beta_j) - \Phi(\alpha_j)}$$

$$E_j^2 := \mathbb{E}[Y(x_{cand} \mid S(x', c_j))^2] = \sigma(x_{cand})^2 - \sigma(x_{cand})\rho_j E_j$$

$$- \mu(x_{cand})^2 + 2\mu(x_{cand})E_j$$

with $\beta_j = \frac{t_{j+1} - \mu_S}{\sigma_S}$, $\alpha_j = \frac{t_j - \mu_S}{\sigma_S}$, $\phi(x)$ is standard normal pdf, $\Phi(x)$ is the standard normal cdf and ρ_j correlation between $Y(x_{cand})$ and $S(x', c_j)$.

The entropy $H[Y(x_{cand}) \mid S(x', c_j)]$ is approximated by that of a gaussian with variance E^2 that is $H[Y(x_{cand}) \mid S(x', c_j)] \approx -\frac{1}{2}\log(2\pi e E^2)$

Backup

- **Does PES favour points away from the edges?**

Yes, here is a theoretical argument.

Let x_1 be the midpoint of the grid and x_2 another different point such that $Y(x_1) = Y(x_2)$, that is, the Gaussian posterior is the same at both points. Take the expectation term of $u_{PES}(x)$: $\mathbb{E}_{S(x',c_j)}H[Y(x_{cand})|S(x',j)]$. This expectation has all the terms equal for $x_{cand} = x_1$ and $x_{cand} = x_2$ except to ρ_j.

$$\mathbb{E}_{S(x',c_j)}H[Y(x_{cand})|S(x',j)] \propto E_j^2 \propto \rho_j$$

ρ_j is the correlation of a bivariate gaussian, with components

$$\begin{bmatrix} \mathcal{N}\left(\mu(x_{cand}), k(x_{cand}, x_{cand})\right) \\ \mathcal{N}\left(\mu(x'), k(x', x')\right) \end{bmatrix}$$

truncated in one dimension from t_j to t_{j+1}. The type of GP we are using is smooth and x_1 is the midpoint of the grid so the correlation is higher for x_1 than for x_2 for most of the points in the grid x'. Thus, we have $\int dx' \rho(x_1, x') > \int dx' \rho(x_2, x')$. The expectation term is negative in $u_{PES}(x)$. Therefore, $u_{PES}(x_1) > u_{PES}(x_2)$.

Backup

Backup

Backup

Backup