Learning new physics with a (kernel) machine

Marco Letizia – Machine Learning Genoa Center

In collaboration with:

G. Grosso (IAIFI), M. Pierini (CERN), L. Rosasco (MaLGa), A. Wulzer (IFAE), M. Zanetti (UniPd).

UniGe | MaLGa

INFN

erc

SLING

- Based on: arXiv:2204.02317, arXiv:2303.05413, arXiv:2305.14137.
- Code: https://github.com/FalkonHEP, https://github.com/mletizia/FalkonNPLM_1D.

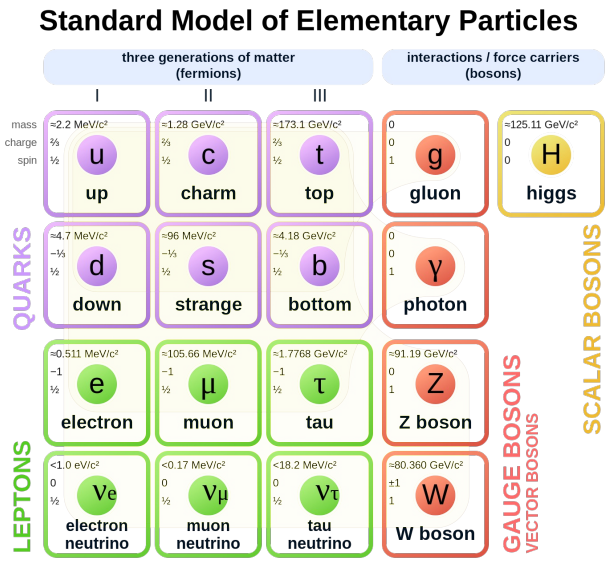
Outline

- Motivation: beyond model-driven analyses in HEP
- Learning new physics with a kernel machine
  - The New Physics Learning Machine
  - Fast kernel methods
- Use cases
- Current developments
- Outlook

Beyond model-driven analyses in HEP

Standard Model not the ultimate theory of elementary particles and their interactions:

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i\bar{\psi}\not{D}\psi + h.c. + \bar{\psi}_i y_{ij} \psi_j \phi + h.c. + |D_\mu \phi|^2 - V(\phi)$$



- Dark matter
- Dark energy
- Gravity
- Neutrinos
- CP
- Hierarchy
- …

However, no conclusive evidence of physics at the LHC!

→ change how we look at data to maximise discovery potential! (HL-LHC)

Beyond model-driven analyses in HEP

Data $\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}, \qquad x_i \overset{\text{iid}}{\sim} p_{\text{true}}(x)$

Traditional analyses are model-driven:

alternative models *(Beyon the Standard Model)* are tested individually against the SM.

- ✓ Highest power (NP lemma)
- ✓ Highly optimized: effective feature selection and cuts
- ✓ Low dimensions: standard statistical techniques
- x One test for every proposal
- x Each test is blind to other BSM scenarios
- x We might not have the correct BSM model
- x We don't let the data tell us what's wrong with the model

$$t(\mathcal{D}) = -2 \log \prod_{x \in \mathcal{D}} \frac{\mathcal{L}(x|SM)}{\mathcal{L}(x|BSM_i)}$$

Beyond model-driven analyses in HEP

Design model-independent approaches: is the SM a good fit to the data?

→ detect generic departures from the SM with minimal theoretical bias.

- New physics effect are *small* or *rare.*
- Sensitive to a large family of possible effects.
- No aggressive data selection and compression: operate in large *N* and *d.*
- Interpretable.

x Less power against any specific new physics effect.

Typical model independence: weaken hypothesis on alternative scenarios, e.g.

simplified BSM models, bump hunts, effective field theories.

→ Learn how data deviates from SM.

The New Physics Learning Machine

D'Agnolo et al, arXiv:1806.02350

Data $\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}, \quad x_i \overset{\text{iid}}{\sim} p_{\text{true}}(x), \quad N(\text{true})$

Reference sample $\mathcal{R} = \{x_i\}_{i=1}^{N_\mathcal{R}}, \quad x_i \overset{\text{iid}}{\sim} p(x|R), \quad N(R), \quad (N_\mathcal{R} \gg N_\mathcal{D})$

$\rightarrow$ Test H_0: $n_{\text{true}}(x) = n(x|R)$

$n(x|\cdot) = N(\cdot)p(x|\cdot)$

$$\mathcal{L}(\mathcal{D}|\cdot) = \frac{e^{-N(\cdot)}}{N_\mathcal{D}!}\prod_{x=1}^{N_\mathcal{D}} n(x|\cdot)$$

Anomaly: *distribution or normalization shift*.

Use machine learning to design a model-independent likelihood-ratio test.

The New Physics Learning Machine

D'Agnolo et al, arXiv:1806.02350

Data $\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}, \quad x_i \overset{\text{iid}}{\sim} p_{\text{true}}(x), \quad N(\text{true})$

Ref $\mathcal{R} = \{x_i\}_{i=1}^{N_\mathcal{R}}, \quad x_i \overset{\text{iid}}{\sim} p(x|R), \quad N(R), \quad (N_\mathcal{R} \gg N_\mathcal{D})$

Advantages of ML for hyp. testing:
- Data driven
- Multivariate maps
- Rich hypothesis spaces
- Can be physics-informed
- …

$n(x|R)$

$n(x|\cdot) = N(\cdot)p(x|\cdot)$

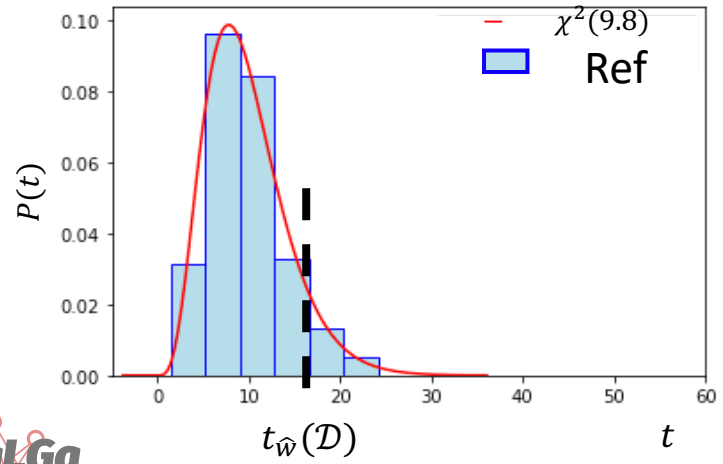
$\mathcal{L}(\mathcal{D}|\cdot) = \frac{e^{-N(\cdot)}}{N_\mathcal{D}!}\prod_{x=1}^{N_\mathcal{D}} n(x|\cdot)$

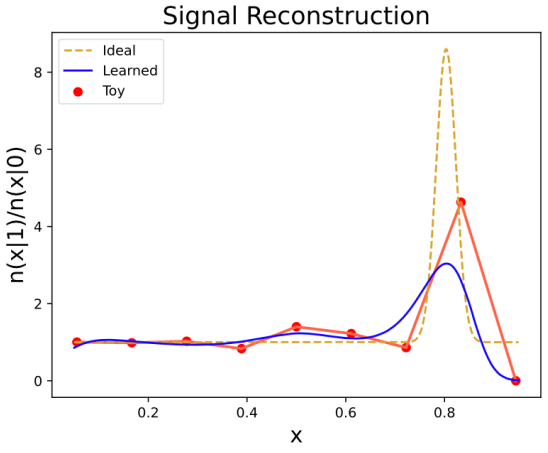
alization shift.

Use machine learning to design a model-independent likelihood-ratio test.

The New Physics Learning Machine

- Estimate the true density ratio with a classifier $\mathcal{R}$ vs $\mathcal{D}$ $\quad f_{\hat{w}} \approx \log \frac{n_{\text{true}}(x)}{n(x|R)}$
- Evaluate the likelihood ratio (in-sample) $\quad t_{\hat{w}}(\mathcal{D}) = -2\left[\frac{N(R)}{N_{\mathcal{R}}}\sum_{x\in\mathcal{R}}\left(e^{f_{\hat{w}}(x)}-1\right) - \sum_{x\in\mathcal{D}} f_{\hat{w}}(x)\right]$
- Estimate the null distribution and compute p-value/Z score





The New Physics Learning Machine

ML et al (2022), arXiv:2204.02317

- Classifier to learn the density ratio

Data: $\{(x_i, y_i)\}_{i=1}^{N_{\mathcal{D}}+N_{\mathcal{R}}}$, with $\begin{cases} y_i = 0 \text{ if } x_i \in \mathcal{R} \\ y_i = 1 \text{ if } x_i \in \mathcal{D} \end{cases}$

Weighted logistic: $\ell(f_w(x), y) = (1-y)\frac{N(R)}{N_{\mathcal{R}}}\log\left(1+e^{f_w(x)}\right) + y\log\left(1+e^{-f_w(x)}\right)$

Target $f_{\hat{w}} \approx f^* = \log\frac{n_{\text{true}}(x)}{n(x|R)}$

Data as a local deformation of the reference

The New Physics Learning Machine

ML et al (2022), arXiv:2204.02317

- Likelihood-ratio test statistic

$$\text{Extended likelihood:}\quad \mathcal{L}(\mathcal{D}|R) = \frac{e^{-N(R)}}{N_{\mathcal{D}}!}\prod_{x=1}^{N_{\mathcal{D}}} n(x|R)$$

$$\text{LR: } t(\mathcal{D}) = -2\log\prod_{x\in\mathcal{D}}\frac{\mathcal{L}(\mathcal{D}|R)}{\mathcal{L}_{\text{true}}(\mathcal{D})} = -2\left[\frac{N(R)}{N_{\mathcal{R}}}\sum_{x\in\mathcal{R}}\left(e^{f(x)}-1\right)-\sum_{x\in\mathcal{D}} f(x)\right], \qquad f(x) = \log\frac{n_{\text{true}}(x)}{n(x|R)}$$

$$\Rightarrow t_{\hat{w}}(\mathcal{D}) = -2\left[\frac{N(R)}{N_{\mathcal{R}}}\sum_{x\in\mathcal{R}}\left(e^{f_{\hat{w}}(x)}-1\right)-\sum_{x\in\mathcal{D}} f_{\hat{w}}(x)\right].$$

The New Physics Learning Machine

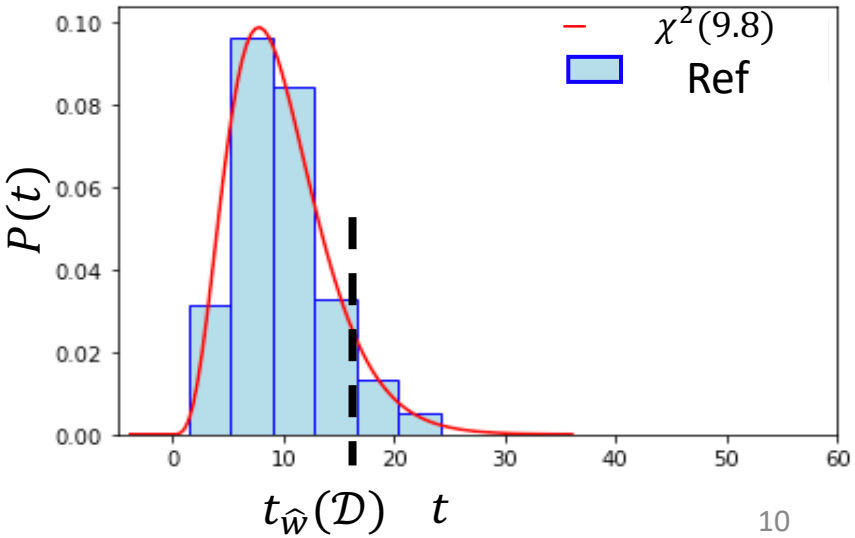
- Estimation of the null hypothesis

Large $t_{\hat{w}}(\mathcal{D}) \rightarrow$ disagreement with the reference model.

How large? We need to calibrate.

- Train the model on $\mathcal{R}$ against multiple R-distributed *toys* $\mathcal{D}_i^{(R)}$.
- Permutations.
- Bootstrap.
- Analytic.

$$\rightarrow p_{\text{value}} = \int_{t_{\hat{w}}(\mathcal{D})}^{\infty} dt\, p(t), \qquad Z = \Phi^{-1}(1 - p_{\text{value}})$$



Learning new physics with a kernel machine

- Function space

Kernel methods: $$f_w(x) = \sum_{i=1}^{N} w_i \, k_\sigma(x, x_i)\,, \qquad k_\sigma(x, x') = \exp - \frac{\|x - x'\|^2}{2\sigma^2}.$$

- Universal approximators.
- Convex optimization with guarantees.

→ Falkon: a SOTA solver for kernel methods G. Meanti et al, arXiv:2006.10350

Learning new physics with a kernel machine

Kernel methods are expensive:

$\mathcal{O}(n^2)$ in space and $\mathcal{O}(n^3)$ in time $-$ store and invert $K_{nn} \in \mathbb{R}^{n \times n}$.

Some approximation is needed.

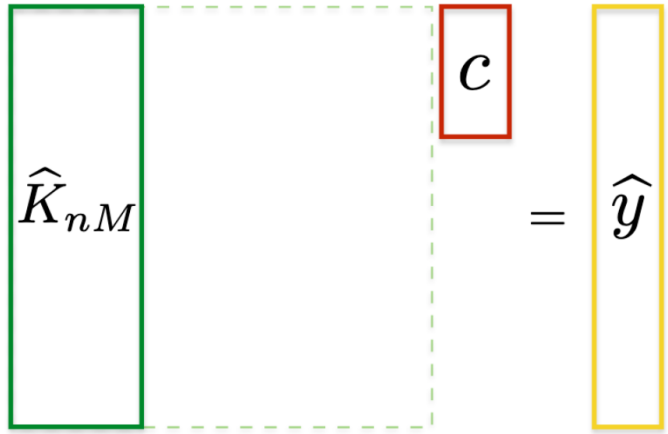
Falkon makes use of:

- Random projections (Nyström)
  - To reduce the size of the problem – $\mathcal{O}(n)$ in space
  - Efficient preconditioned conjugate gradient – $\mathcal{O}(n\sqrt{n}\log n)$ in time
- Efficient (multi-)GPU implementation

Learning new physics with a kernel machine

- Random projections (Nyström)

$$f_w(x) = \sum_{i=1}^{N} w_i \, k(x, x_i) \quad \rightarrow \quad \sum_{i=1}^{M} w_i \, k(x, x_i) \,,$$



$\{x_1, \dots, x_M\} \subset \{x_1, \dots, x_n\}$ sampled uniformly at random (centers)

Optimal statistical bounds can be obtained with $M = \mathcal{O}(\sqrt{n})$

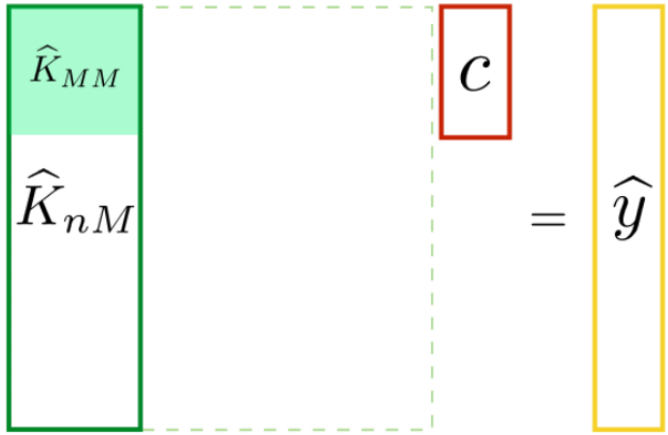
Theorem (Rudi, Camoriano, R. '15)

Let $(\tilde{x}_i)_{i=1}^{M} \subseteq (x_i)_{i=1}^{n}$ *picked uniformly at random, if* $\lambda = 1/\sqrt{n}$ *and* $M \geqslant \sqrt{n}$ *then*

$$\mathbb{E}L(\widehat{f}_{\lambda,M}) - \min_{f \in \mathcal{H}} L(f) \lesssim \frac{1}{\sqrt{n}}$$

Learning new physics with a kernel machine

- Conjugate gradient with efficient preconditioning



Preconditioner: $\mathrm{P}^{\mathrm{T}}\mathrm{P} = (K_{nM}K_{nM} + \lambda n K_{MM})^{-1}$

Nyström → $P^T P \approx \left(\frac{n}{M} K_{MM}^2 + \lambda n K_{MM}\right)^{-1}$

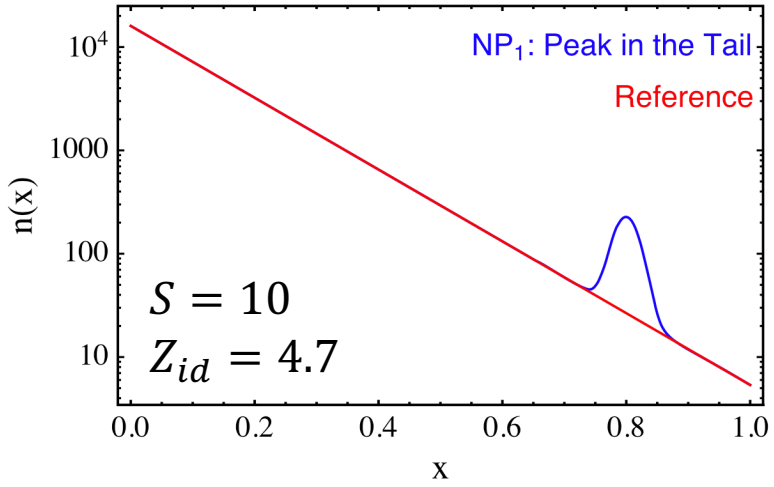
Theorem (Rudi, Carratino, Rosasco '17)

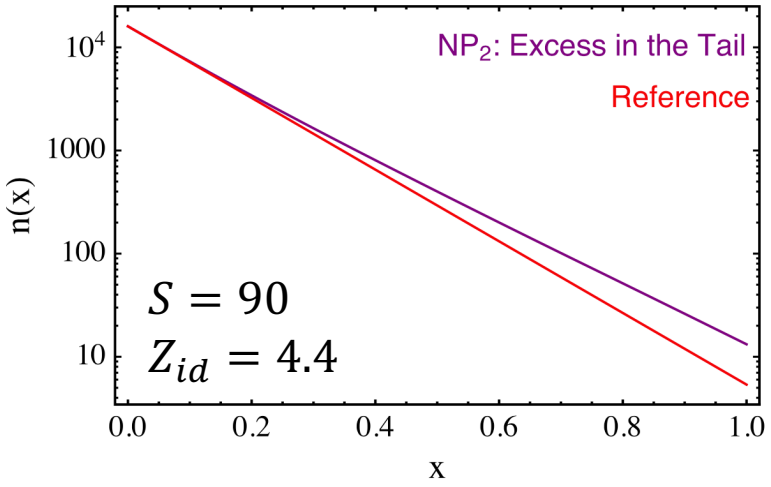
Let $(\tilde{x}_i)_{i=1}^{M} \subseteq (x_i)_{i=1}^{n}$ *uniformly at random, then if* $\lambda = 1/\sqrt{n}$, $M \geqslant \sqrt{n}$ *and* $t \geqslant \log(n)$

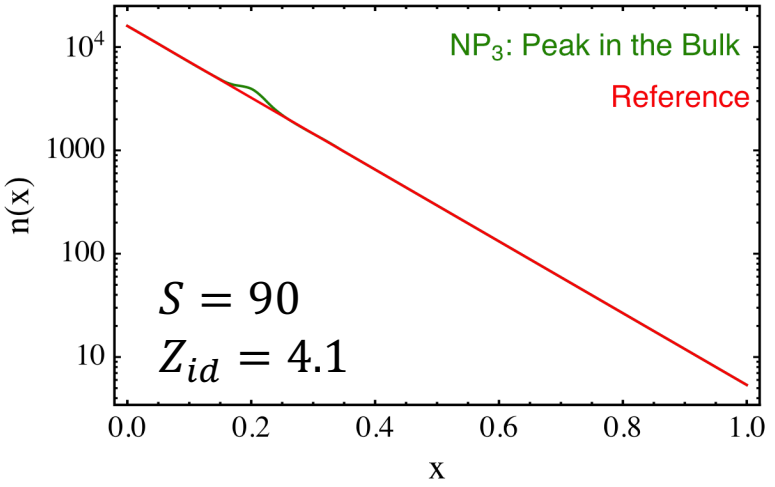
$$\mathbb{E}L(\widehat{f}_{\lambda,M,t}) - \min_{f\in\mathcal{H}} L(f) \lesssim \frac{1}{\sqrt{n}}$$

Univariate example

$N_{\mathcal{R}} = 2{\times}10^5, \quad N(R) = 2000, \quad N_{\mathcal{D}} = N(R) + S$



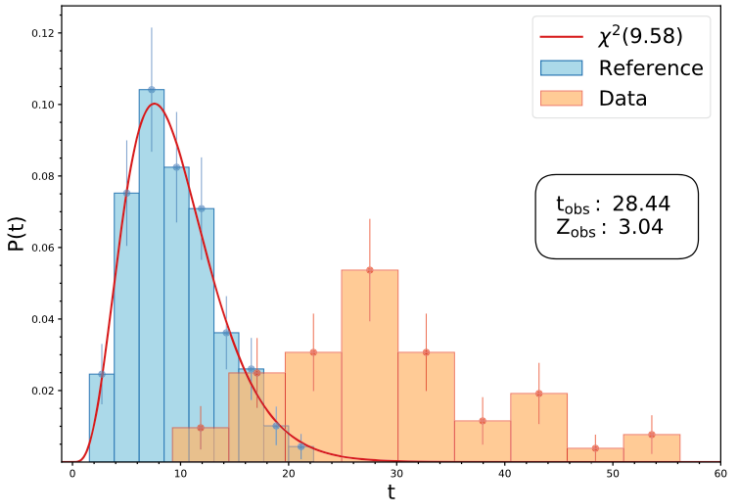


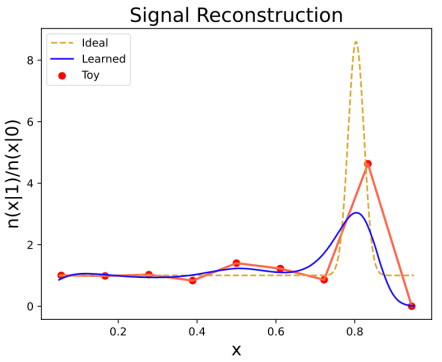


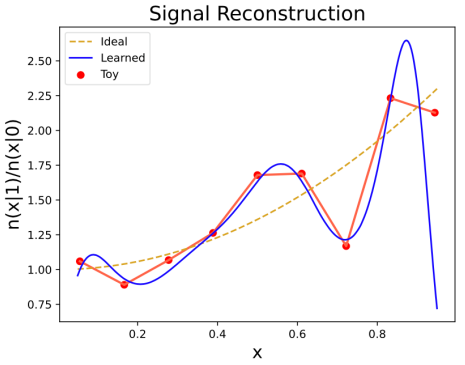
300 R-toys
100 D-toys

$Z_{obs} = (2.43, 3.04, 2.82)$

$\bar{t}_{tr} = 2.11$ sec



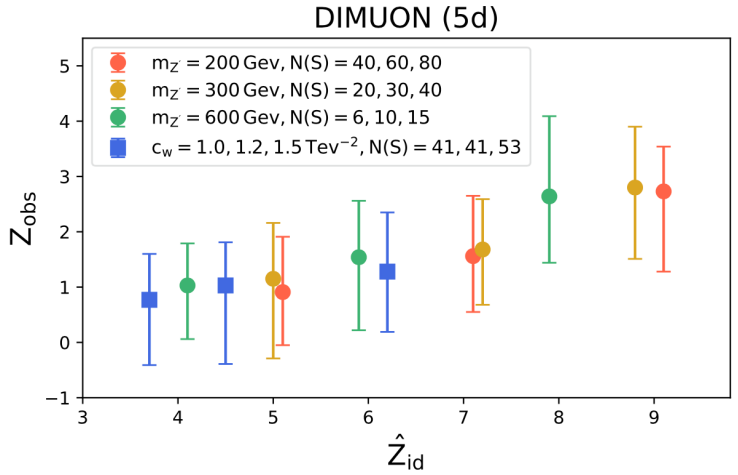


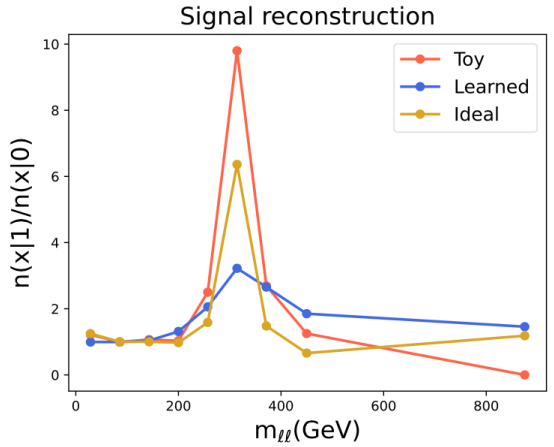


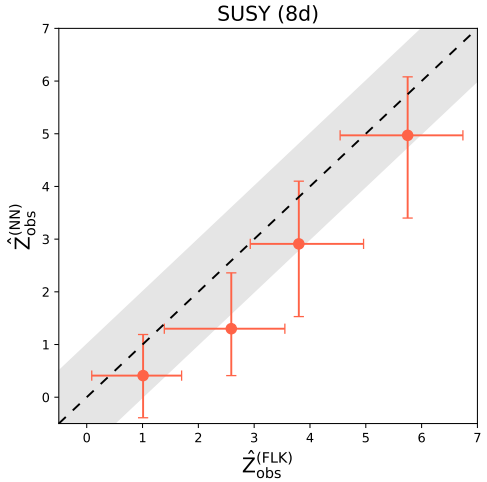
Multivariate

$pp \to \mu^+\mu^-$: SM vs SM+Z'/EFT $[p_{T1}, p_{T2}, \eta_1, \eta_2, \Delta\phi]$, SUSY (8d), HIGGS (21d)

$N(R) = 2\times10^4$, $N_B = 10^5$ $N(R) = 10^5$, $N_R = 5\times10^5$







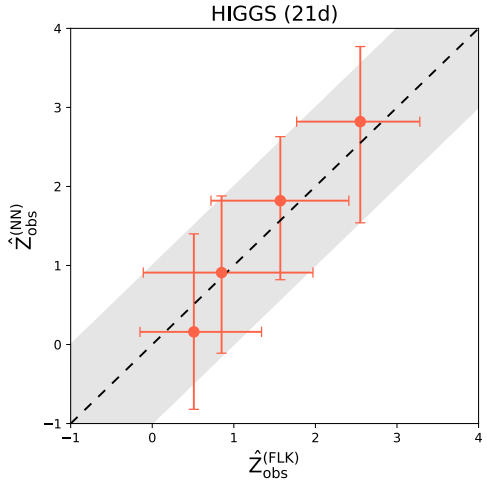


Table 1 Average training times per single run with standard deviations (low level features and reference toys). Note that time measured in hours (for NN) and seconds (for Falkon)

Model	DIMUON	SUSY	HIGGS
FLK	**(44.9 ± 3.4) s**	**(18.2 ±1.2) s**	**(22.7 ± 0.4) s**
NN	(4.23 ± 0.73) h	(73.1 ± 10) h	(112 ± 9) h

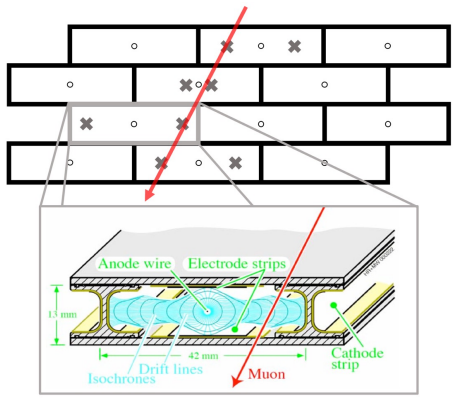
Bold values indicate the lowest for each column (lower is better)

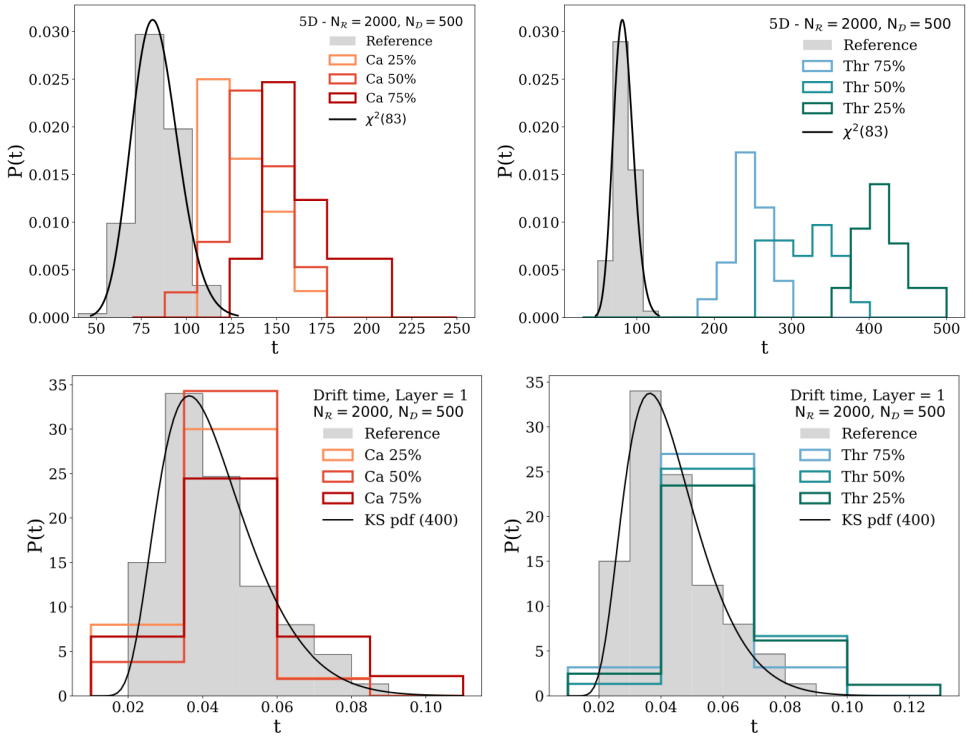
Data: https://zenodo.org/records/4442665

Data Quality Monitoring

G. Grosso et al, arXiv:2303.05413

Drift tube chambers from Legnaro INFN National Laboratory.





DATASET:

- Drift times (t_i): the four drift times of the muon track.
- Slope (ϕ): the angle with respect to the vertical axis.
- Reference data is collected in a controlled regime.
- Anomalies:
  - reduced voltage of cathodic strips to 75%, 50%, and 25% of their nominal value (-1.2 kV)
  - lowered front-end thresholds to 75%, 50%, and 25% of nominal value (100 mV)

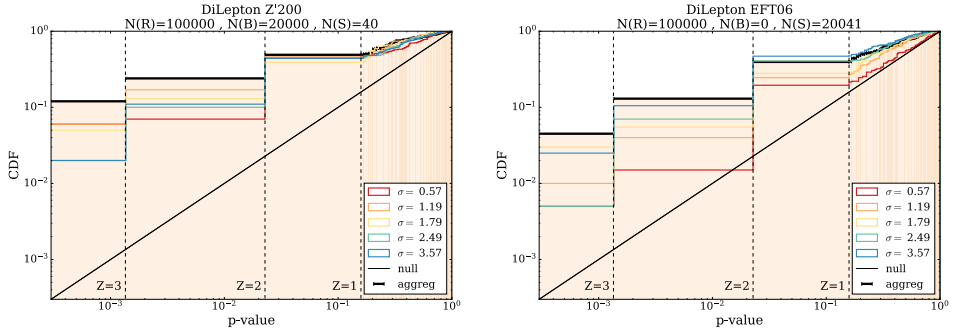
$\bar{t}_{tr} \approx 0.5$ sec

Data: https://zenodo.org/records/7128223

Current developments

- Multiple testing

Sensitivity to any given signal depends on model hyperparameters → combine multiple tests



- Systematic uncertainties

Extension to profile likelihood formalism D'Agnolo et al arXiv:1912.12155

$$\text{nuisance parameters} \quad f_w(x) \to f_w(x) + \log r_\nu(x)\,, \qquad r_\nu(x) = \exp\left[\nu\,\delta_1(x) + \frac{\nu^2}{2}\,\delta_2(x) + \cdots\right]$$

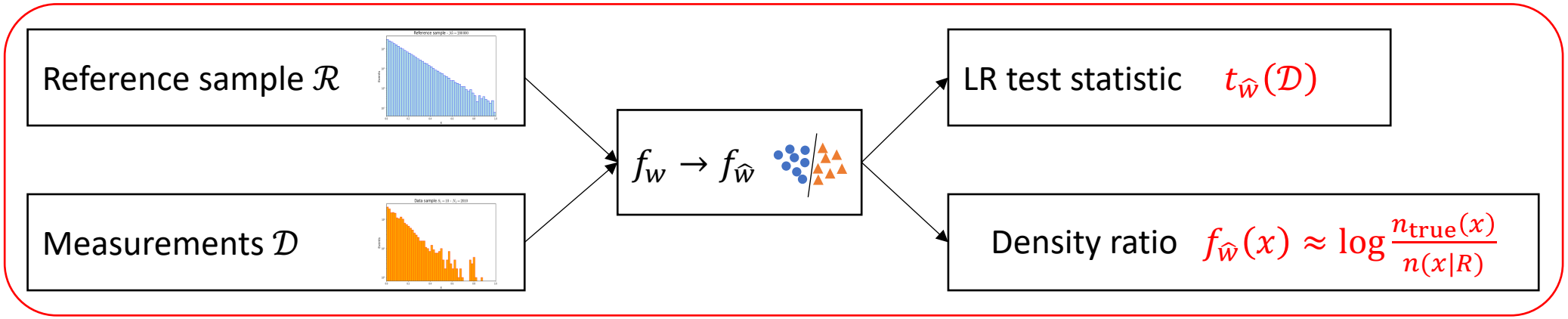
$$t(\mathcal{D}) = 2\log\frac{\max_{w,\nu}\mathcal{L}\left(H_{w,\nu}\middle|\mathcal{D},\mathcal{A}\right)}{\max_{\nu}\mathcal{L}(R_\nu|\mathcal{D},\mathcal{A})}$$

Summary

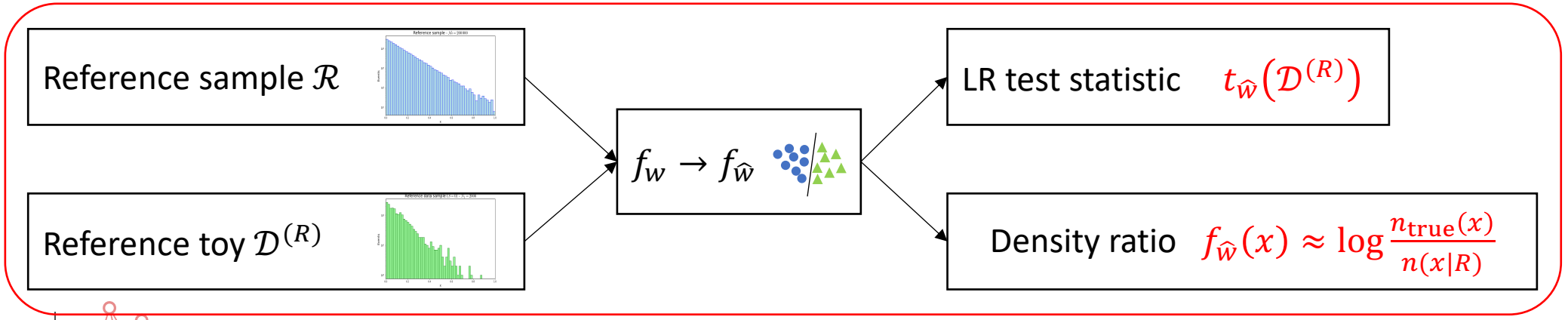
- *New Physics Learning Machine*: methodology to compare a model to the data.
- Developed for new physics searches (ongoing CMS analysis).
- Implementation based on SOTA large-scale kernel methods.
- Many developments/applications
  - Systematics
  - Algorithmic ideas combining optimization and statistics
  - Data quality monitoring
  - Background re-weighting
  - Evaluation of generative models
  - Comparison of MC simulators
  - Connections with foundation models for HEP

Calibration

Observed test statistic



Null distribution

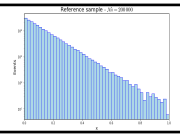


$\times N_{toys}$

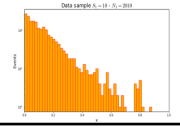
Calibration

Observed test statistic

Reference sample $\mathcal{R}$



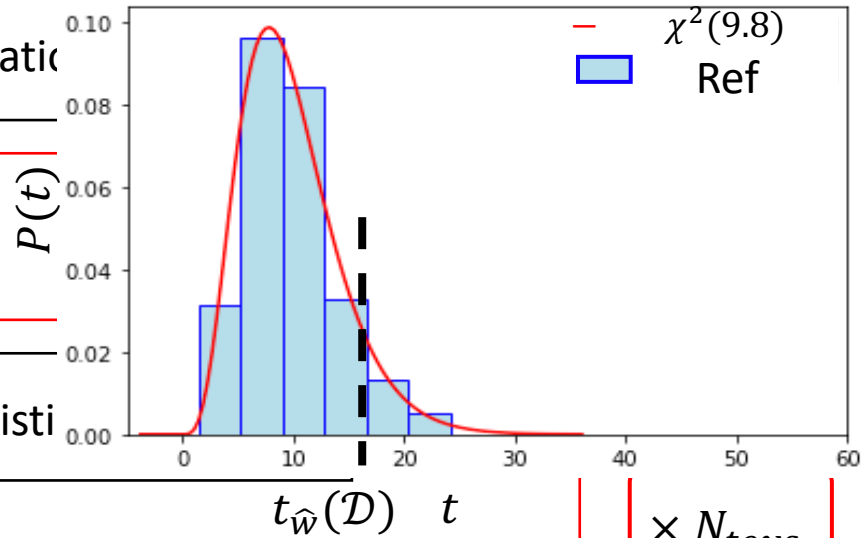
Measurements $\mathcal{D}$



$f_w \rightarrow f_{\hat{w}}$

LR test statistic $t_{\hat{w}}(\mathcal{D})$

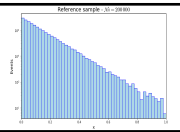
Density ratio



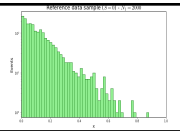
$\times N_{toys}$

Null distribution

Reference sample $\mathcal{R}$



Reference toy $\mathcal{D}^{(R)}$



$f_w \rightarrow f_{\hat{w}}$

LR test statisti

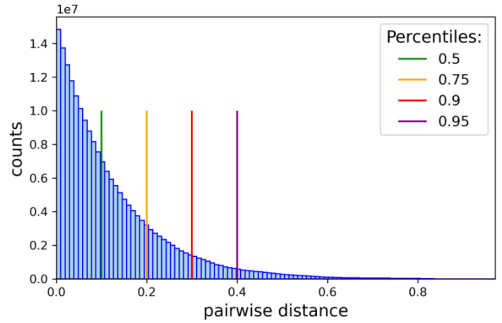
Density ratio $f_{\hat{w}}(x) \approx \log \frac{n_{\text{true}}(x)}{n(x|R)}$

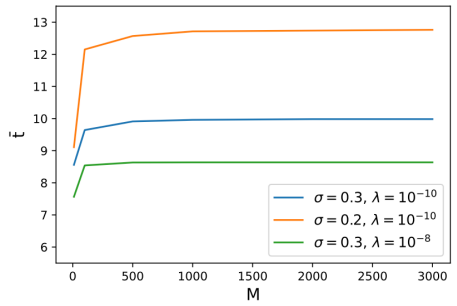
Backup

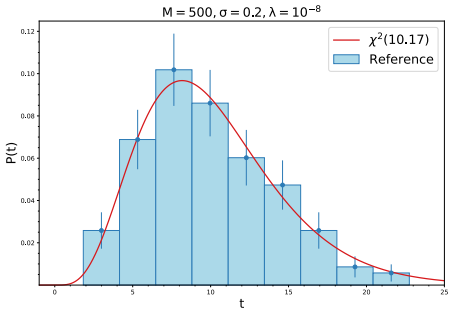
Falkon has three main hyperparameters (M, σ, λ)

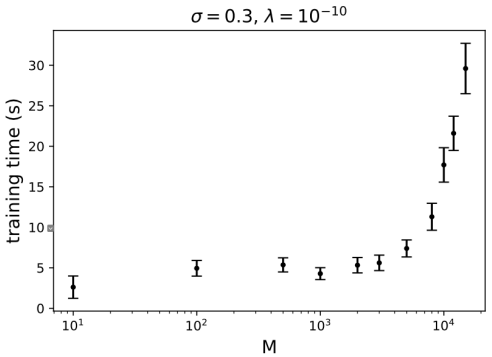
No cross-validation to preserve model-independence.

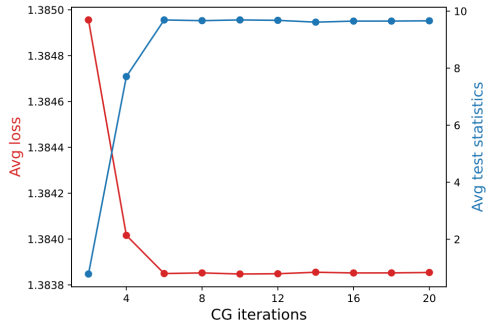
→ mix of heuristics, statistical considerations and efficency





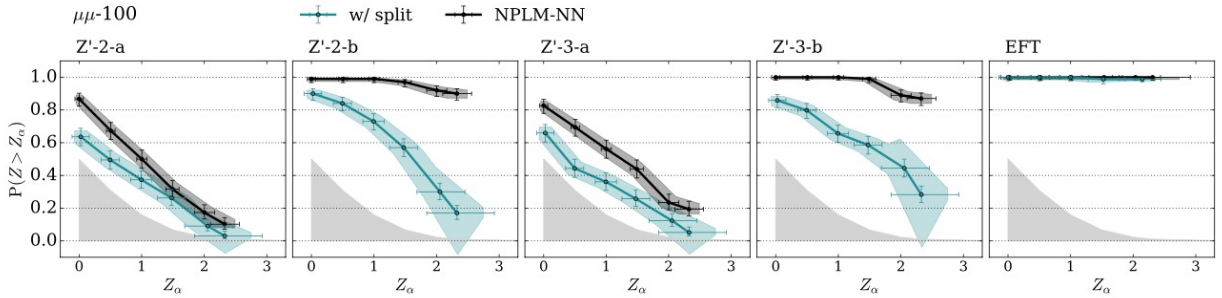




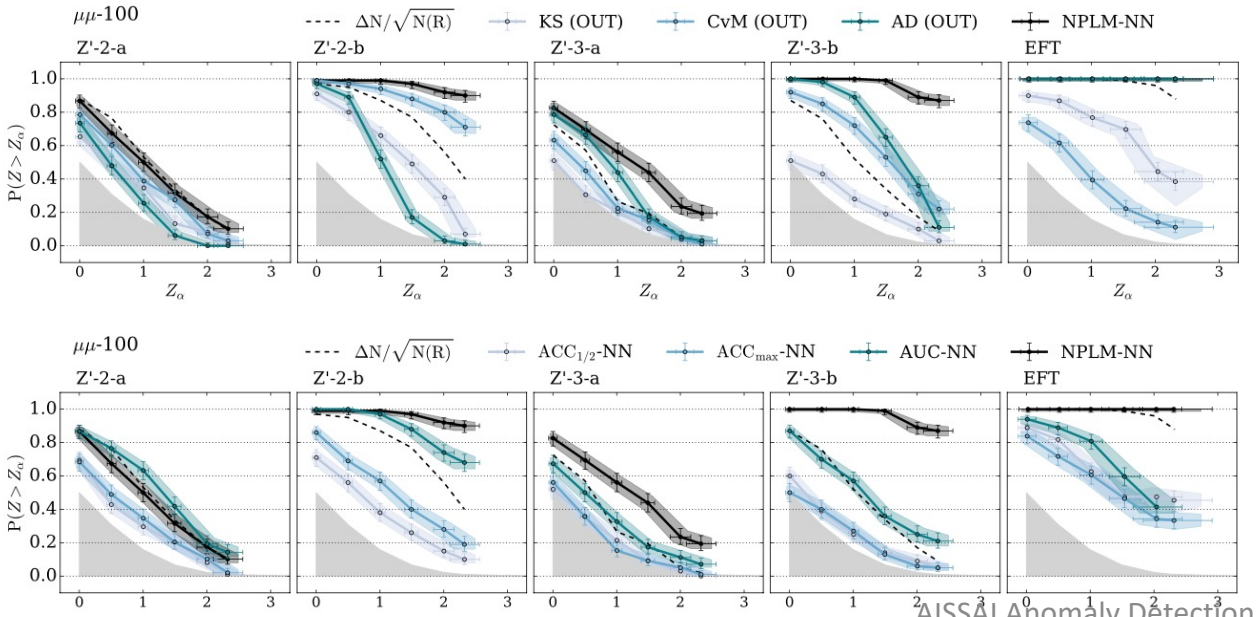


Backup G. Grosso, ML, M. Pierini, A. Wulzer arXiv:2305.14137

Train-test split



Different metrics



Backup – deriving the test statistic

Model data as local deformation of the reference

$$n_{\text{true}}(x) = e^{f(x)}\, n(x|R) \quad \rightarrow \quad \text{Likelihood: } \mathcal{L}(\mathcal{D}|\cdot) = \frac{e^{-N(\cdot)}}{N_{\mathcal{D}}!}\prod_{x=1}^{N_{\mathcal{D}}} n(x|\cdot)$$

$$\text{LR: } -2\log\frac{\mathcal{L}(\mathcal{D}|R)}{\mathcal{L}_{\text{true}}(\mathcal{D})} = -2\left[N(\text{true}) - N(R) - \sum_{x\in D}\log\frac{n_{true}(x)}{n(x|R)}\right]$$

$$N(\text{true}) = \int n_{true}(x)\,dx = \int e^{f(x)}\, n(x|R)\,dx = N(\text{R})\int e^{f(x)}p(x|R)\,dx \approx \frac{N(R)}{N_{\mathcal{R}}}\sum_{x\in\mathcal{R}} e^{f(x)}$$

Take the parametrized form $f(x) \rightarrow f_w(x)$:

$$\text{Likelihood ratio test: } \quad t_w(\mathcal{D}) = -2\left[\frac{N(R)}{N_{\mathcal{R}}}\sum_{x\in\mathcal{R}}\left(e^{f_w(x)} - 1\right) - \sum_{x\in\mathcal{D}} f_w(x)\right]$$

The New Physics Learning Machine

Data $\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}, \quad x_i \overset{\text{iid}}{\sim} p_{\text{true}}(x), \quad N(\text{true})$

Reference sample $\mathcal{R} = \{x_i\}_{i=1}^{N_\mathcal{R}}, \quad x_i \overset{\text{iid}}{\sim} p(x|R), \quad N(R), \quad N_\mathcal{R} \gg N_\mathcal{D}$

→ Test $H_0\colon\ n_{\text{true}}(x) = n(x|R)$

$n(x|R) = N(R)p(x|R)$

Introduce a local deformation of the reference $f_w\colon\ e^{f_{\hat{w}}(x)} \approx \dfrac{n_{\text{true}}(x)}{n(x|R)}$.

→ Likelihood ratio test: $t_w(\mathcal{D}) = -2\left[\dfrac{N(R)}{N_\mathcal{R}}\sum_{x\in\mathcal{R}}\left(e^{f_{\hat{w}}(x)} - 1\right) - \sum_{x\in\mathcal{D}} f(x)\right]$

The New Physics Learning Machine

Data $\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}, \quad x_i \overset{\text{iid}}{\sim} p_{\text{true}}(x), \quad N(\text{true})$

Reference sample $\mathcal{R} = \{x_i\}_{i=1}^{N_\mathcal{R}}, \quad x_i \overset{\text{iid}}{\sim} p(x|R), \quad N(R)$

$\Rightarrow$ Test $H_0: \; n_{\text{true}}(x) = n(x|R)$

$$n(x|R) = N(R)p(x|R)$$

Likelihood: $L(\mathcal{D}|\cdot) = \dfrac{e^{-N(\cdot)}}{N_\mathcal{D}!}\prod_{x=1}^{N_\mathcal{D}} n(x|\cdot)$

Likelihood ratio test: $t_w(\mathcal{D}) = -2\left[\dfrac{N(R)}{N_\mathcal{R}}\sum_{x\in\mathcal{R}}\left(e^{f(x)}-1\right) - \sum_{x\in\mathcal{D}} f(x)\right]$

The New Physics Learning Machine

Choose $\hat{w}$ from the data as a supervised learning problem

Data: $\{(x_i, y_i)\}_{i=1}^{N_{\mathcal{D}}+N_{\mathcal{R}}}$, with $\begin{cases} y_i = 0 \text{ if } x_i \in \mathcal{R} \\ y_i = 1 \text{ if } x_i \in \mathcal{D} \end{cases}$

Loss $\ell(f_w(x), y)$: minimum $f_{\hat{w}} \approx f^* = \log \frac{n(x|1)}{n(x|0)} = \log \frac{n_{\text{true}}(x)}{n(x|R)}$

$$\Rightarrow \quad t_{\hat{w}}(\mathcal{D}) = -2\left[\frac{N(R)}{N_{\mathcal{R}}} \sum_{x \in \mathcal{R}} \left(e^{f_{\hat{w}}(x)} - 1\right) - \sum_{x \in \mathcal{D}} f_{\hat{w}}(x)\right]$$