Università di Genova | **DIBRIS** DIPARTIMENTO DI INFORMATICA, BIOINGEGNERIA, ROBOTICA E INGEGNERIA DEI SISTEMI | MaLGa

Efficient kernel methods for large scale problems in HEP

ACAT 2021

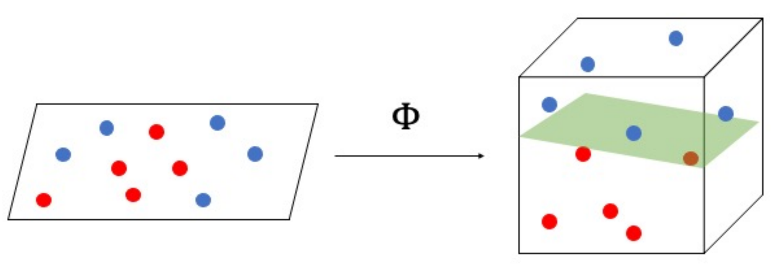
M. Letizia[ab], G. Losapio[a], M. Rando[a], L. Rosasco[acd]

[a]MaLGa – DIBRIS University of Genoa, [b]INFN Genoa, [c]MIT, [d]IIT - Genoa

Motivations

Kernel methods represent an elegant and mathematically sound approach to nonparametric learning, but so far could hardly be used in large scale problems, since naïve implementations scale poorly with data size. Recent improvements have shown the benefits of a number of algorithmic ideas, combining optimization, numerical linear algebra and random projections. These, combined with (multi-)GPU specific implementations, allow for great speedups on large scale datasets while delivering highly competitive performances. We demonstrate here their effectiveness on HEP specific problems such as signal-versus-background classification and model-independent new physics searches. We also compare kernel methods with with similar neural network based models, showing significant gains in terms of training times and computational costs while maintaining comparable performances.

Kernel methods



Fast and flexible nonlinear models

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

Empirical Risk Minimization: $\quad f_{\widehat{w}} = \arg\min_{f \in \mathcal{H}} L(f_w) + \lambda R$

Poor scaling with N, $K_{ij} = k(x_i, x_j)$ (1B points ≈ 10^{18} bytes RAM)

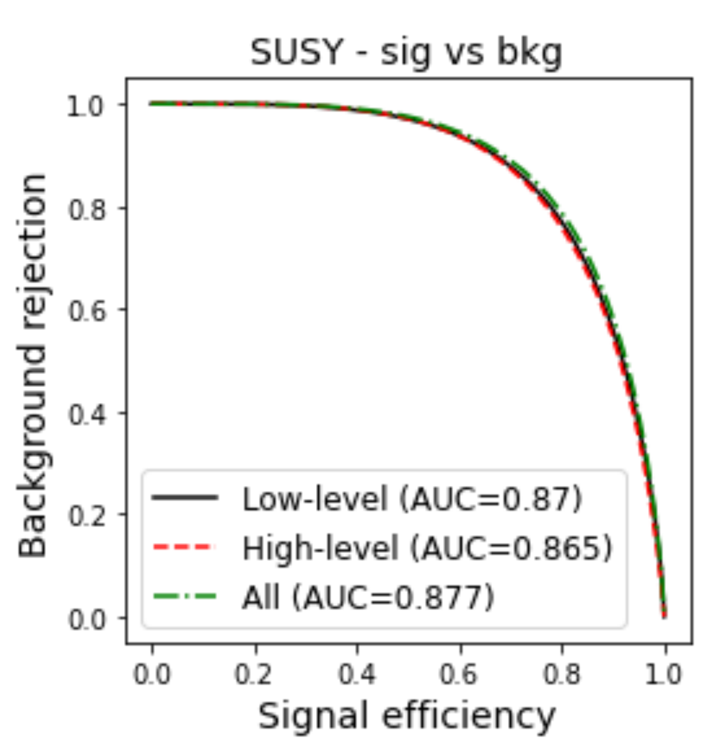
→ FALKON[1]:

- Gaussian $\quad k(x, x') = \exp -(\|x - x'\|^2 / 2\sigma^2)$
- L2 penalty $\quad R = \sum_{ij} w_i w_j k(x_i, x_j)$
- Nyström approx $\quad \{\tilde{x}_1, \dots, \tilde{x}_M\} \subset \{x_1, \dots, x_N\} \;\to f(x) = \sum_{i=1}^{M} w_i\, k(x, \tilde{x}_i)$
- (multi-)GPU implementation

Classification

Signal vs background classification (model-dependent)[1,2]

Data: SUSY (9/19d)
Training + valid : 2.6M + 100k
Test: 500k



Model	HIGGS - $d=28$, $n \approx 10^7$		SUSY - $d=19$, $n \approx 10^6$	
	1-AUC	time	c-error	time
Falkon	0.1804±0.0003	443±2 s	19.67±0.02 %	22±0 s
LogFalkon	0.1787±0.0002	2267±5 s	19.58±0.03 %	83±1 s

Accuracy and running-time comparisons on large scale HEP datasets[1].

Model-independent searches

Learning the (extended) likelihood ratio test statistics[3,4]

$$\mathcal{L}(y, D) = e^{-N(y)} \prod_{x \in D} n(x|y), \qquad \text{two hyp. } \; y = \{0,1\}$$

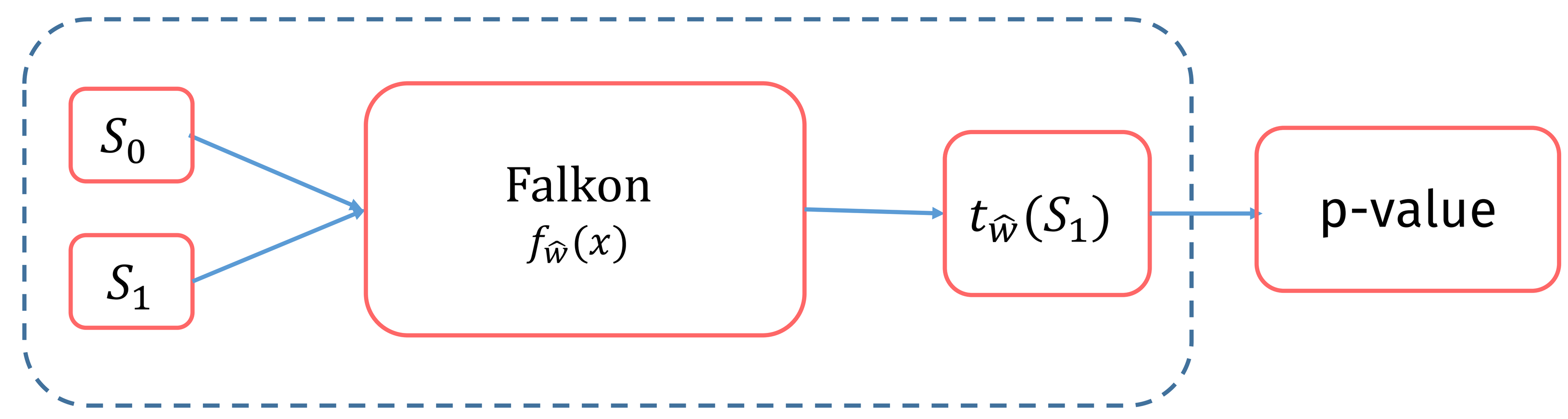
$$n(x|y) = N(y)p(x|y), \qquad N(y) = \int dx\, n(x|y) \quad \text{exp. \# events}$$

$$\boxed{t(D) = 2\log\frac{\mathcal{L}(1)}{\mathcal{L}(0)} = 2\left[N(0) - N(1) + \sum_{x \in D} \log\frac{n(x|1)}{n(x|0)}\right]}$$

Data:

- Standard model MC $\quad S_0 = \{x_i\}_{i=1}^{N_0}, \quad x_i \sim p(x|0)$
- Measurements $\quad S_1 = \{x_i\}_{i=1}^{N_1}, \quad x_i \sim p(x|1), \quad N_0 \gg N_1$

Model and results



Re-weighted logistic loss

$$L = \sum_{(x,y)} \left[(1-y)\frac{N(0)}{N_0}\log\left(1 + e^{f_w(x)}\right) + y\log\left(1 + e^{-f_w(x)}\right)\right],$$

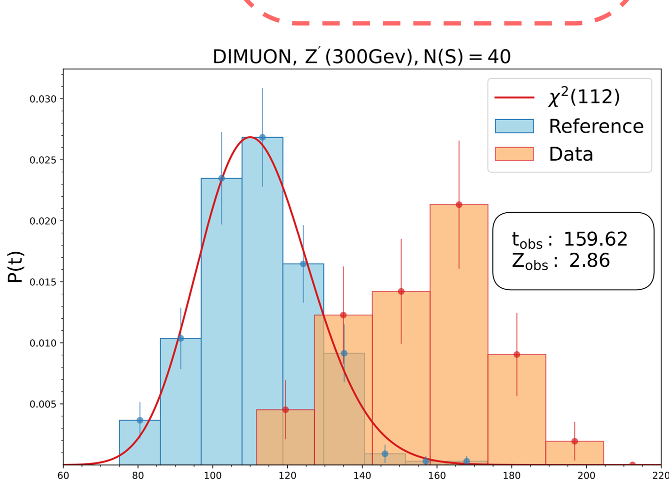
$$f_{\widehat{w}}(x) \approx \log\frac{n(x|1)}{n(x|0)}, \qquad N_{\widehat{w}}(1) \approx \frac{N(0)}{N_0}\sum_{x \in S_0} e^{f_{\widehat{w}}(x)}$$

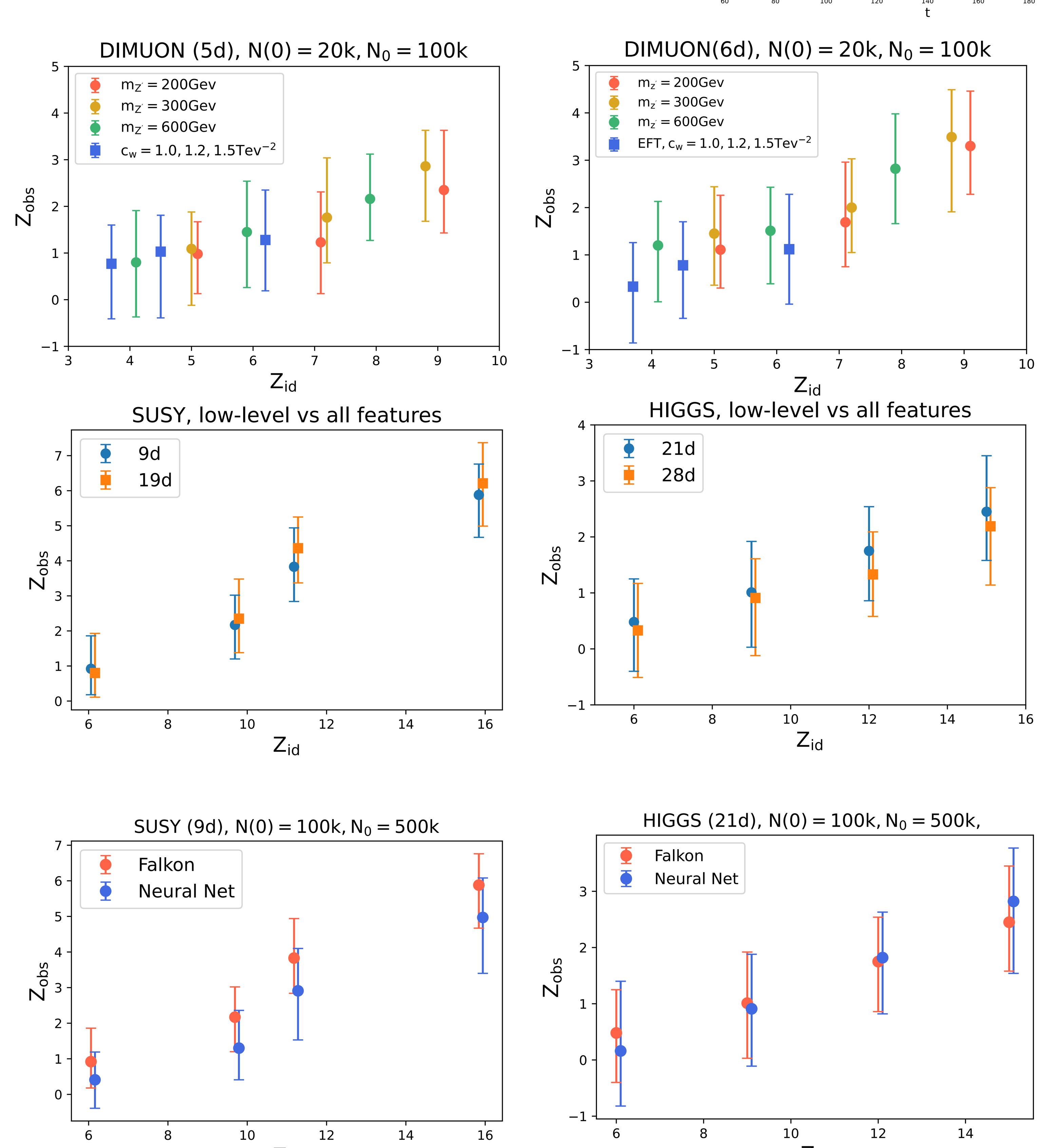
$$\to \boxed{t_{\widehat{w}}(S_1) = 2\left[\frac{N(0)}{N_0}\sum_{x \in S_0}\left(1 - e^{f_{\widehat{w}}(x)}\right) + \sum_{x \in S_1} f_{\widehat{w}}(x)\right]}$$

Estimate $p(t|0)$ with SM toys
↓
p-value

Experiments on simulated data:

- DIMUON[4]: $pp \to \mu^+\mu^-$
  $x_{in} = [p_{T1}, p_{T2}, \eta_1, \eta_2, \Delta\Phi, m_{\ell\ell}]$
- SUSY and HIGGS[2]:
  low-level (9/21d), low+high-level (19/28d)





Model	DIMUON	SUSY	HIGGS
Falkon	(53.8 ± 1.9) s	(44.8 ± 1.5) s	(88.7 ± 2.2) s
Neural Net	(4.23 ± 0.73) h	(73.1 ± 10) h	(112 ± 9) h

Average training times per single run with standard deviations.

REFERENCES

1.Meanti, et al. "Kernel methods through the roof: handling billions of points efficiently." *arXiv:2006.10350* (2020).
2.Baldi, et al. "Searching for exotic particles in high-energy physics with deep learning." *Nature communications* 5.1 (2014): 1-9.
3.D'Agnolo, et al. "Learning new physics from a machine." *Physical Review D* 99.1 (2019): 015014.
4.D'Agnolo, et al. "Learning multivariate new physics." *The European Physical Journal C* 81.1 (2021): 1-21.

CONTACTS

Marco Letizia - marco.letizia@edu.unige.it