Efficient kernel methods for model-independent new physics searches

Marco Letizia
MaLGa Center - DIBRIS
Università di Genova, Genoa, Italy
marco.letizia@edu.unige.it

Gianvito Losapio
MaLGa Center - DIBRIS
Università di Genova, Genoa, Italy
s4803867@studenti.unige.it

Marco Rando
MaLGa Center - DIBRIS
Università di Genova, Genoa, Italy
marco.rando@edu.unige.it

Gaia Grosso
Dipartimento di Fisica e Astronomia
Università di Padova, Padua, Italy
CERN, Experimental Physics Department
Geneva, Switzerland
gaia.grosso@cern.ch

Lorenzo Rosasco
MaLGa Center - DIBRIS, IIT, MIT
Università di Genova, Genoa, Italy
lorenzo.rosasco@unige.it

Abstract

We present a novel kernel-based anomaly detection algorithm for model-independent new physics searches. The model is based on a re-weighted version of kernel logistic regression and it aims at learning the likelihood ratio test statistics from simulated anomaly-free background data and experimental data. Model-independence is enforced by avoiding any prior assumption about the presence or shape of new physics components in the data. This is made possible by kernel methods being non-parametric models that, given enough data, can approximate any continuous function and adapt to potentially any type of anomaly. This model shows dramatic advantages compared to similar neural network implementations in terms of training times and computational resources, while showing comparable performances. We test the model on datasets of different dimensionalities showing that modern implementations of kernel methods are competitive options for large scale problems.

1 Introduction

Signals of new physics (NP) manifest themselves as discrepancies (excesses or deficits of events) in collected experimental data with respect to the predictions of a “reference model", such as the standard model of particle physics (SM). In a fully model-independent approach, data analysis should remain agnostic about the type and nature of potential NP signatures in the data. In practice, this is very difficult to achieve given the complexity of the experimental data in modern experiments.

We here introduce a novel anomaly detection algorithm for model-independent NP searches in high energy physics. The basic strategy is to train a binary classifier on simulated SM data and experimental data to reconstruct the likelihood ratio test statistics. The algorithm is based on a re-weighted version of logistic regression with Gaussian kernels and model-independence is enforced by not making

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

any prior assumption about the presence or shape of NP signatures in the data. This approach shows dramatic advantages in efficiency, in terms of both training time and computational resources, compared to similar implementations based on neural networks (NN) [1, 2], with comparable performances.

The area of machine learning algorithms for model-independent searches in high energy physics is a fast growing field of research, see for instance [3–8]. The NN model in [1, 2] share the same statistical framework of our proposal and a similar methodology. We therefore reconstructed their model and used it for comparison. In the machine learning literature, several approaches to anomaly detection have been developed with different levels of supervision [9–12].

2 Designing a classifier for hypothesis testing

In this section, we present the different aspects of the algorithm proposed in this work. The basic strategy is to train a binary classifier on a *reference sample* S_0 of SM simulated data representing anomaly-free behavior and on a *data sample* S_1 of experimental measurements, to construct a hypothesis test based on the (maximum) likelihood ratio test statistics. The goal of the test is to determine whether to accept or reject the null hypothesis that the experimental data are drawn from the reference distribution and do not exhibit significant anomalies.

To enforce model independence, we introduce a parameterized class of distributions $p_w(x|1)$ to represent the alternative hypothesis. We then select the point $w = \hat{w}$ that maximizes the likelihood with respect to the experimental data at hand. The distribution $p_{\hat{w}}(x|1)$ represents the specific, data driven, alternative hypothesis while the null hypothesis is represented by the SM prediction, $p(x|0)$. The extended likelihood of the data corresponding to the reference distribution is given by

$$\mathcal{L}(S_1, y=0) = \frac{e^{-N(0)} N(0)^{\mathcal{N}_1}}{\mathcal{N}_1!} \prod_{x=1}^{\mathcal{N}_1} p(x|0) = \frac{e^{-N(0)}}{\mathcal{N}_1!} \prod_x n(x|0), \tag{1}$$

where $n(x|0) = N(0)p(x|0)$ is the data distribution normalized to the expected number of events

$$N(0) = \int dx\, n(x|0). \tag{2}$$

Compared to the usual likelihood, the extended likelihood has a Poisson factor that takes into account the fluctuations in the number of collected events $\mathcal{N}_1$ (see [13]). Then, the likelihood ratio takes the following form

$$\begin{aligned} t(S_1) &= 2 \log \frac{\mathcal{L}_{\hat{w}}(S_1, y=1)}{\mathcal{L}(S_1, y=0)} \\ &= 2 \max_w \left[N(0) - N_w(1) + \sum_{x \in S_1} f_w(x) \right], \quad N_w(1) = \int dx\, n_w(x|1), \end{aligned} \tag{3}$$

where we defined $f_w(x) = \log \frac{n_w(x|1)}{n(x|0)}$. The test statistics t is itself a random variable and knowing its distribution under the null hypothesis $p(t|0)$, we could compute a p-value measuring the tension between the reference data and the experimental measurements.

In order to have an effective model-independent algorithm, the model should be able to explore a large family of distributions $p_w(x|1)$. In this work we rely on kernel methods, which are universal approximators [14, 15] with guaranteed convergence and generalization properties [16]. They consider functions of the type $f_c(x) = \sum_i c_i k(x, x_i)$, where $k(x, x_i)$ is the kernel function which measures similarity between any pair of inputs. The coefficients c_i are computed via empirical risk minimization [17] with iterative methods. The specific solver that we use in this work is part of a public library known as Falkon [18].[1] Specifically, it is a Nyström-based kernel method with Gaussian kernels $k(x, x') = \exp(\|x - x'\|^2 / 2\sigma^2)$ and $L2$ regularization.

The model is trained as a classifier with a loss function given by a re-weighted logistic loss that reads as

$$L(f) = \sum_{(x,y) \in S} \left[\frac{N(0)}{\mathcal{N}_0} (1-y) \log \left(1 + e^{f(x)} \right) + y \log \left(1 + e^{-f(x)} \right) \right], \tag{4}$$

[1] https://falkonml.github.io/falkon/

with $\mathcal{N}_y$ being the size of the class y. We choose this objective function because its target function is directly the density ratio

$$f_{\hat{w}} \approx f^* = \log \frac{n(x|1)}{n(x|0)}, \tag{5}$$

which is learned with a maximum likelihood approach without having to estimate the two (possibly multivariate) distributions individually and it is classification calibrated, therefore it outputs reliable probabilities. In order to accurately represent the reference distribution, it is preferable to consider a large reference sample while the size of the data sample is determined by the parameters of the experiment, specifically its luminosity [19]. By re-weighting the loss function as in Eq.(4), we avoid the issues related to imbalanced datasets ($\mathcal{N}_0 > \mathcal{N}_1$) while keeping the statistical advantage of having a large reference sample.[2] In order to reconstruct the test statistics in Eq.(3) once $f_{\hat{w}}$ has been learned, the number of expected events in the alternative hypothesis needs to be computed. From Eq.(3) and $f_{\hat{w}}$, one can estimate $N_{\hat{w}}(1)$ using a Monte Carlo approximation over S_0 as follows

$$N_{\hat{w}}(1) \approx \frac{N(0)}{\mathcal{N}_0} \sum_{x \in S_0} e^{f_{\hat{w}}(x)}. \tag{6}$$

Hence by learning $f_{\hat{w}}(x)$ we can estimate $N_{\hat{w}}(1)$ and therefore the test statistics $t(S_1)$.

The training strategy goes as follows:

- We train the model once on reference data S_0 and experimental data S_1 to estimate the log-likelihood ratio and reconstruct the test statistics $t(S_1)$ for the experimental data.
- The model is re-trained on reference data S_0 against $N_{\text{toy}} \approx 300$ toy anomaly-free samples following the SM, to reconstruct the distribution of the test statistics under the null hypothesis $p(t|0)$ and compute the p-value p_{S_1}.
- We further rewrite the p-value as a Z-score, $Z_{obs}(S_1) = \Phi^{-1}(1 - p_{S_1})$, where Φ^{-1} is the quantile of a Normal distribution with zero mean and unit variance.

Hyper-parameters The algorithm has three main hyper-parameters: the number of Nyström centers M, the bandwidth of the Gaussian kernel σ and the regularization parameter λ. The Nyström method is a low rank approximation which is based on selecting a subset of size M of data points $\{\tilde{x}_1, ..., \tilde{x}_M\} \subset \{x_1, ..., x_\mathcal{N}\}$ on which the kernels are "centered". Selecting a small M brings a lower storage and computational cost but it it could affect the accuracy of the likelihood ratio estimation. In practice, we find that a typical value that works well is of the order of the number of expected events, $M = \mathcal{O}(N(0))$.[3] By varying the parameters σ and λ more or less complex functions can be selected. In particular for large λ and/or σ the model simplifies and tends to become linear while for small values it tends to fit the data. Concretely, the effect of σ is more aggressive and we find that a good strategy is to tune σ first and then λ. To select σ, we search for values that return flexible models while avoiding overfitting the noise in the samples. We then look at the distribution of the pairwise (euclidean) distance of the examples in the dataset as a proxy for the relevant scales in the data and we then select a value at least as large as the median, typically around the 75th percentile. Once σ is chosen, we take λ as small as possible while maintaining a stable algorithm. We train the model up to a maximum of 10^6 iterations while imposing a threshold on the variation of the loss function of $\Delta L = 10^{-7}$ below which training is halted. These are both extremely conservative criteria as the number of iterations required for convergence is $\mathcal{O}(10)$ and reducing ΔL does not bring any observable benefit.
Note that, to preserve model-independence, any step required for hyper-parameter tuning is performed the reference data only. We find this strategy to be robust and reproducible as "reasonable" variations in the hyper-parameters do not significantly affect the final results.

3 Experiments and discussion

Datasets We report the results of the algorithm on three simulated datasets of increasing dimensionality: DIMUON (d=5), SUSY (d=9) and HIGGS (d=21). They are all characterised by two classes:

[2] Note that using Eq.(4) can be seen as minimizing the sum of false positives and false negatives rather than the overall error.

[3] There are studies (see [20, 21]) showing that optimal statistical bounds can be achieved with $M = \mathcal{O}(\sqrt{\mathcal{N}})$.

SM only reference data and NP data (SM background combined with a NP component). In the first dataset we test both a resonant and a non-resonant signal with varying mass and coupling constant.

The ratio between the size of the reference sample and the number of expected background events is fixed as $\mathcal{N}_0/N(0) = 5$ for all datasets. We then considered $N(0) = 2 \times 10^5$ for the DIMUON case and $N(0) = 10^6$ for SUSY and HIGGS. The simulated NP data are also characterized by the expected number of signal events $N(S)$, so that the actual number of events is $\mathcal{N}_1 \sim \text{Pois}(N(0) + N(S))$. We repeat our experiments at different values of the ratio $N(S)/N(0)$ between the expected number of signal and background events to test the performance of the model. We refer the reader to [2, 22] for additional details about the data. Being simulated datasets, we also reconstruct the distribution of the test statistics under the alternative hypothesis $p(t|1)$ by training the model on different toy samples ($N_{\text{toy}} \approx 100$) following the NP distribution.

We report the median observed significance Z_{obs} against an estimated ideal significance $\hat{Z}_{id}$ computed with traditional model-dependent techniques, e.g., cut-and-count analyses. We also show that the distribution $p(t|0)$ follows closely a χ^2 distribution with a number of degrees of freedom that depends on the complexity of the underlying model in agreement with Wilk's theorem [23], see Figure 1. We use this fact to obtain an estimate of the p-value but we also report examples of the maximum reach of the reconstructed distributions.

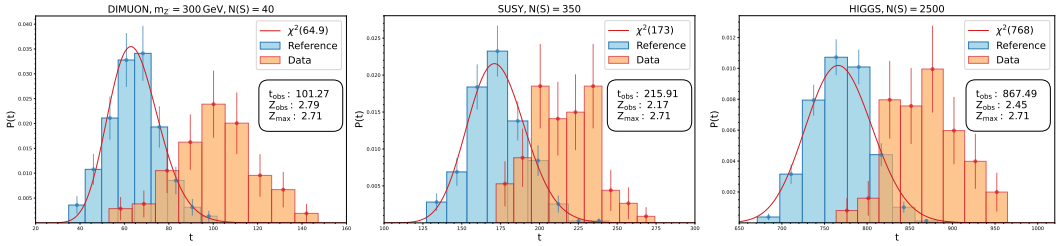

Figure 1: Examples of distributions of the test statistics under the null and alternative hypotheses.

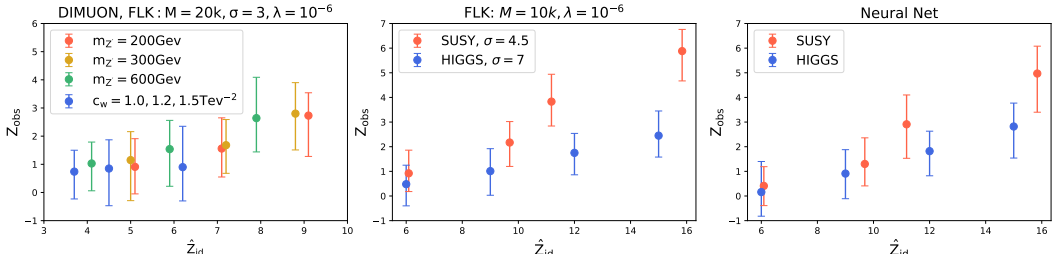

Figure 2: Median observed significance against estimated ideal significance.

Results In Figure 1 we show typical examples of reconstructed distributions of the test statistics under both hypotheses for all datasets. The corresponding χ^2 distribution is determined using a Kolmogorov-Smirnof test. In Figure 2 we show the median observed Z-score for the three datasets for different amounts of signal injection corresponding to different values of the ideal significance. The error bars represent the 16th and 84th percentile. We can observe that in the DIMUON case, the observe significance seems to depend mostly on the ideal significance and weakly on the nature of the signal. We also compare our results with the ones obtainable with the NN-based model presented in [1, 2] and trained according the the authors' guidelines. The NN results regarding the DIMUON dataset can be found in [2]. Overall, we can see that the two models return quite comparable results. However, in Table 1 we show the dramatic difference in training times between the two models. Note that in order to reconstruct the distribution of t under the null hypothesis, a large number of training rounds are typically required. For the NN implementation this means that one cannot rely on sequential experiments performed on single GPU systems, while this is not a problem for our algorithm. The main limiting factor seems to be the number dimensions, as the ratio between the the observed significance and estimated ideal one deteriorates with the number of features. Nonetheless,

this limitation is not associated with the specific model since it is also observed in the NN case. In our experiments, we also observed that the median observed significance quickly approaches $\hat{Z}_{id}$ when performing appropriate cuts on the input features, such as removing the Z boson peak from the DIMUON dataset. At the same time we also noticed that using additional engineered features with higher discriminative power does not improve the results in a considerable way.

Remarks: Experiments were performed on a single GPU system with a NVIDIA Titan Xp (12 GB RAM). We want to thank the authors of [2] for giving us access to their dataset for our tests.

Table 1: Average training times per single run with standard deviations.

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

Discussion We discussed a new approach for model-independent new physics searches in particle physics. The main focus of our work is on the efficiency of the proposed model which is based on a re-weighted version of kernel logistic regression with Gaussian kernels and $L2$ regularization. The core machine learning algorithm has been developed to extend the use of kernel methods to large scale problems by leveraging different algorithmic ideas such as random projections and taking full advantage of GPU architectures. We obtain results on simulated realistic data that rival similar NN-based implementations while drastically reducing training times and the required computational resources. Nonetheless, we see different directions for improving our proposal. These include: a treatment of systematic uncertainties associated with the imperfect knowledge of the reference model; a more principled procedure for hyper-parameter tuning; alternative approaches to estimate the distribution of the test statistics under the null hypothesis; testing the algorithm on real data. On the algorithmic side, possible developments include: different strategies for the selection of Nyström centers; taking advantage of NN architectures as feature extractors in conjunction with the efficiency of kernel methods.

Acknowledgements: M.L., G.L., M.R. and L.R. acknowledge the financial support of the European Research Council (grant SLING 819789), the AFOSR projects FA9550-18-1-7009, FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), the EU H2020-MSCA-RISE project NoMADS - DLV-777826, and the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. We gratefully acknowledge the support of NVIDIA Corporation for the donation of the Titan Xp GPUs and the Tesla k40 GPU used for this research. G.G. is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n^o 772369).

References

[1] Raffaele Tito D'Agnolo and Andrea Wulzer. Learning new physics from a machine. *Physical Review D*, 99(1), Jan 2019.

[2] Raffaele Tito D'Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti. Learning multivariate new physics. *Eur. Phys. J. C*, 81(1):89, 2021.

[3] Purvasha Chakravarti, Mikael Kuusela, Jing Lei, and Larry Wasserman. Model-independent detection of new physics signals using interpretable semi-supervised classifier tests, 2021.

[4] Benjamin Nachman and David Shih. Anomaly detection with density estimation. *Physical Review D*, 101(7):075042, 2020.

[5] Andrea De Simone and Thomas Jacques. Guiding new physics searches with unsupervised learning. *The European Physical Journal C*, 79(4):1–15, 2019.

[6] Mikael Kuusela, Tommi Vatanen, Eric Malmi, Tapani Raiko, Timo Aaltonen, and Yoshikazu Nagai. Semi-supervised anomaly detection–towards model-independent searches of new physics. In *Journal of Physics: Conference Series*, volume 368, page 012032. IOP Publishing, 2012.

[7] Konstantin T Matchev, Prasanth Shyamsundar, and Jordan Smolinsky. A quantum algorithm for model independent searches for new physics. *arXiv preprint arXiv:2003.02181*, 2020.

[8] Adrian Alan Pol, Victor Berger, Cecile Germain, Gianluca Cerminara, and Maurizio Pierini. Anomaly detection with conditional variational autoencoders. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 1651–1657. IEEE, 2019.

[9] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[10] Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60:19–31, 2016.

[11] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[12] Caleb C Noble and Diane J Cook. Graph-based anomaly detection. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, 2003.

[13] Roger Barlow. Extended maximum likelihood. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 297(3):496–506, 1990.

[14] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(95):2651–2667, 2006.

[15] *Support Vector Machines*. Information Science and Statistics. Springer New York, New York, NY, 2008. ISSN: 1613-9011.

[16] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. *arXiv preprint arXiv:1705.10958*, 2017.

[17] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[18] Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: handling billions of points efficiently, 2020.

[19] P.A. Zyla et al. (Particle Data Group). Review of Particle Physics. *Progress of Theoretical and Experimental Physics*, 2020(8), August 2020.

[20] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *NIPS*, pages 1657–1665, 2015.

[21] Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Globally convergent newton methods for ill-conditioned generalized self-concordant losses. *arXiv preprint arXiv:1907.01771*, 2019.

[22] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5(1), Jul 2014.

[23] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sections 3.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We use a standard statistical analysis and our applications are on high energy physics data only. The main focus of our paper is on the efficiency of the specific implementation we propose.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In Section 2 we give instructions on the training of the model, the analysis of the results and we include a link to the repository of the library we use.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported standard devations and percentiles to show the range of the reconstructed distributions.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the end of Section 2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We refer to the original paper where the solver we use (Falkon) was introduced.
   (b) Did you mention the license of the assets? [Yes] We mention that the library we use is public.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] We perform a standard statistical analysis which is easily reproducible with standard Python libraries. The training of the model can be easily performed with the library we refer to in the text.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The data we are using/curating does not contain personally identifiable information or offensive content
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]