Controlling Classifier Bias with Moment Decomposition: A Method to Enhance Searches for Resonances

Ouail Kitouni
Massachusetts Institute of Technology
kitouni@mit.edu

Benjamin Nachman
Lawrence Berkeley National Laboratory
bpnachman@lbl.gov

Constantin Weisser
Massachusetts Institute of Technology
weisser@mit.edu

Mike Williams
Massachusetts Institute of Technology
mwill@mit.edu

Abstract

A key challenge in searches for resonant new physics is that classifiers trained to detect and enhance potential signals must not induce localized structures, *i.e.* they must not sculpt a peak in an otherwise smooth background spectrum. Such structures could result in a false signal when the background is estimated from data using sideband methods. A variety of techniques have been developed to construct classifiers which are independent from the resonant feature (often a mass). Such strategies are sufficient to avoid localized structures, but are not necessary. We develop a new set of tools using a novel moment loss function (Moment Decomposition or MODE) which relax the assumption of independence without creating structures in the background. By allowing classifiers to be more flexible, we enhance the sensitivity to new physics without compromising the fidelity of the background estimation.

1 Introduction

Searching for new phenomena associated with localized excesses in otherwise featureless spectra, often referred to as bump hunting, is one of the most widely used approaches in particle and nuclear physics. A key feature of these searches is that they are relatively model-agnostic since sidebands in data can be used to estimate the background under a potential localized excess. These sideband fits are possible because the background data can be well-approximated either with simple parametric functions or smooth non-parametric techniques such as Gaussian processes [1]. Sideband methods for background estimation are often combined with relatively simple and robust event selections in order to ensure broad coverage. However, there is a growing use of modern machine learning to enhance signal sensitivity [2, 3, 4, 5, 6]. For example, both ATLAS [7] and CMS [8] have developed W jet taggers using deep learning models to improve the sensitivity of searches involving Lorentz-boosted and hadronically decaying W bosons.

A key challenge with complex event selections like those involved in boosted W tagging is that they can invalidate the smoothness assumption of the background. In particular, if classifiers can infer the mass of the parent resonance, then selecting signal-like events will simply pick out background events with a reconstructed mass near the target resonance mass. Many techniques have been developed that modify or simultaneously optimize classifiers so that their responses are independent of a given resonance feature. For machine learning classifiers, the proposed solutions include modifications to loss functions that implicitly or explicitly enforce independence. A variety of similar proposals under

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

the monikers of domain adaptation and fairness have been proposed in the machine learning literature (see e.g. Ref. [9, 10] and Ref. [11, 12]).

Ensuring that a classifier is independent from a given resonant feature is sufficient for mitigating sculpting, but it is not necessary. The original requirement is simply that a selection using the classifier does not introduce localized features in the background spectrum, which is a much looser requirement than enforcing independence. For example, if a classifier has a linear dependence on the resonant feature, then there would be a strong correlation. However, a threshold requirement on such a classifier would not sculpt any bumps in the background-only case. This example motivates a new class of techniques that allow classifiers to depend on the resonant feature in a controlled way. In the limit that constant dependence is required, then the classifier and the resonant feature will be independent. The advantage of relaxing the independence requirement is that the resulting classifiers can achieve superior performance because they are allowed to be more flexible.

We present a new set of tools that allow for controlled dependence on a resonant feature.This new approach is called *Moment Decomposition* or MODE (see Ref. [13] for more details.) Using MODE, analysts can require independence, linear dependence, quadratic dependence, *etc.* In addition, analysts can place bounds on the slope of the linear dependence, and restrict quadratic dependence to be monotonic.

2 Methods

2.1 Existing decorrelation methods

We will consider the binary classification setting in which examples are given by the triplet (X, Y, M), where $X \in \mathcal{X}$ is a feature vector, $Y \in \mathcal{Y} := \{0, 1\}$ is the target label, and finally, $M \in \mathcal{M}$ is the resonant feature (or protected attribute) whose spectrum will be used in the bump hunting. The feature vector X can either contain M directly as one of its elements or contain other features that are arbitrarily indicative of M. Decorrelation methods are interested in finding a mapping $f : \mathcal{X} \to \mathcal{S}$ where $s \in \mathcal{S}$ are scores used to obtain predictions $\hat{y} \in \mathcal{Y}$ with the additional constraint that f be conditionally independent of (or uniform with) M in the sense that

$$p(f(X) = s|M = m, Y = y) = p(f(X) = s|Y = y) \; \forall \; m \in \mathcal{M} \text{ and } \forall \; s \in \mathcal{S}, \tag{1}$$

for one or more values y, *e.g.*, independence could be required for the background, the signal, or both. Decorrelation methods include planing [14, 15], adversaries [16, 17, 18, 19], distance correlation (DISCO) [20, 21], and flatness [22].

2.2 Beyond decorrelation: Moment decomposition

Rather than decorrelation, our method allows for controllable mass dependence in the form of an ℓ^{th} order polynomial, where ℓ is a hyperparameter chosen by the analyst. The total loss is

$$\mathcal{L}[f] = L_{\text{class}} + \lambda L^{\ell}_{\text{MODE},} \tag{2}$$

where λ is a tradeoff paramter, $L_{\text{class}}[f]$ is some classification loss (e.g., cross-entropy) and the MODE loss is given by

$$L^{\ell}_{\text{MoDe}} \equiv \sum_m \int |F_m(s) - \sum_{i=0}^{\ell} c_i(s) P_i(\tilde{m})|^2 \mathrm{d}s, \tag{3}$$

where a transformation is performed on m such that $\mathcal{M} \to [-1, 1]$. Here, $F_m(s)$ is the conditional cumulative distribution function of scores, s, in mass bin m (central mass value $\tilde{m}$), P_i are the Legendre polynomials, and the Legendre moments are given by

$$c_i(s) = \left[\frac{2i+1}{2}\right] \int_{-1}^{1} P_i(m') F(s|m') \mathrm{d}m'. \tag{4}$$

The MODE loss in Eq. (3) (which we will denote by MODE[ℓ]) is optimal when $F_m(s)$ is an ℓ^{th} order polynomial $\forall \, s$. It can be shown that the minimizer for $\ell = 0$ satisfies the independence condition in Eq. (1). More interestingly, choosing $\ell = 1$ leads to linear mass dependence, $\ell = 2$ quadratic dependence, *etc.*

3 The boosted W tagging challenge

Highly lorentz boosted, hadronically decaying W bosons commonly arise in extensions of the Standard Model. The boost causes the decay products of these bosons to be mostly captured by a single large-radius jet. Various features of the substructure of these jets can be used to distinguish the boosted bosons from generic quark and gluon jets.

A bump hunt is performed either in the mass of the W candidate jet, m_J, or the mass of one W candidate jet and another (possibly W candidate) jet, m_{JJ}. The challenge with substructure classifiers is that they can introduce artificial bumps into the mass spectrum because substructure is correlated with the jet mass and the jet kinematic properties (which are related to m_{JJ}). For this reason, boosted W tagging has become a benchmark process for studying decorrelation methods at the LHC.

The mass of the simulated samples used in this section and shown in the left of Fig. 1 is the same as in Ref. [20] (intended to emulate the study in Ref. [23]). An implementation for MoDe in PyTorch as well as Tensorflow/Keras is available at `https://github.com/okitouni/MoDe`, along with other examples.

3.1 Classifier Details

MoDe and DisCo: We use a simple 3-layer neural network with a similar architecture to that described in Ref. [20]. However, unlike Refs. [20] and [23], after each of the 3 fully connected 64-node layers, we use Swish activation [24] as it provides a notable performance increase. We also use a batch normalization layer after the first fully connected layer. The output layer has a single node with a sigmoid activation. Both MoDe and DisCo are trained with the ADAM optimizer [25] using a 1cycle learning rate policy [26] with a starting learning rate of 10^{-3} and a maximum learning rate of 10^{-2}, which is reached using a cosine annealing strategy [27] and decayed to 10^{-5} during the last few iterations. Momentum is cycled in the inverse direction from 0.95 to a minimum of 0.85. These hyperparameters were selected through a learning rate range test.

Adversarial Decorrelation: The same classifier used for MoDe and DisCo is trained against a Gaussian Mixture Network (GMN) [28] that parametrizes a Gaussian mixture model with 20 components, *i.e.* its outputs are the means, variances, and mixing coefficients of 20 normal distributions. We follow a similar adversarial setup to that referenced in Refs. [23] and [20]. We use one hidden layer with 64 nodes with ReLu activation connected to 60 output nodes. These outputs model the posterior probability density function $p(M|f(X;\theta_{\text{class}}))$ which is used to define the adversarial loss $L_{\text{adv}} = \mathbb{E}_{s\sim f(X)}\mathbb{E}_{m\sim M|s}\left[-\log p_{\theta_{\text{adv}}}(m|s)\right]$ where θ_{class} and θ_{adv} are the classifier and adversary parameters, respectively. Decorrelation is obtained by finding the minimax solution to

$$\arg\min_{\theta_{\text{class}}} \max_{\theta_{\text{adv}}} \left[L_{\text{class}(\theta_{\text{class}})} - \lambda L_{\text{adv}}(\theta_{\text{class}}, \theta_{\text{adv}})\right].$$

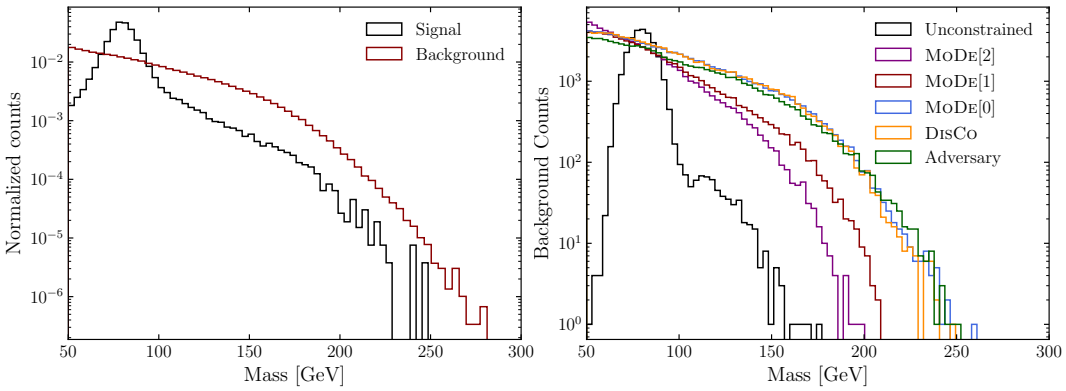

Figure 1: Left: (True) distributions of signal and background events. Right: Distributions of events predicted to be backgrounds at 50% signal efficiency (true positive rate) for different classifiers. The unconstrained classifier sculpts a peak at the W-boson mass, while other classifiers do not.

3.2 Example Results

The left of Fig. 1 shows the signal and background distributions, while the right shows that, as expected, without imposing a strong constraint on mass decorrelation, the classifier learns to select samples near the W-boson mass, which sculpts a fake peak in the background.

Following Ref. [20], we quantify the classification performance using the background rejection factor, R_{50}, which is defined as the inverse false positive rate at 50% signal efficiency (true positive rate.) To assess the danger of sculpting a peak, we use the signal bias induced by the classifier selection, which is what actually matters when searching for resonant new physics.

Specifically, we use the signal estimators obtained by fitting samples of backgrounds classified as signal (false positives) to a simple polynomial function as proxies for the signal biases. The maximum likelihood estimators $\hat{s}$ obtained for different samples are divided by their uncertainties such that values of roughly unity are consistent with no bias (since the true signal rate is zero), while values significantly larger than unity indicate substantial bias that could result in false claims of observations or invalid confidence intervals on observed W-boson rates.

Figure 2 shows that the DisCo and MoDe[0] decorrelation methods provide signal estimators that are consistent with the true value of 0 for $R_{50} \lesssim 9$. Figure 2 also shows that the flexibility to go beyond decorrelation provided by MoDe[1] and MoDe[2] results in achieving unbiased signal estimators at larger background-rejection power. This would directly translate to improved sensitivity in a real-world analysis.

4 Conclusions and Outlooks

In summary, a key challenge in searches for resonant new physics is that classifiers trained to enhance potential signals must not induce localized structures. We presented a new set of tools using a novel moment loss function (Moment Decomposition or MoDe) which relax the assumption of independence from a resonant feature (often a mass) without creating structures in the background. Using MoDe, analysts can require independence, linear dependence, quadratic dependence, *etc*. By allowing classifiers to be more flexible, we enhance the sensitivity to new physics without compromising the fidelity of the background estimation. In addition, our method is simple, fast (more details are in Ref. [13]) and requires no further training (unlike adversaries which are notoriously difficult to train) and introduces only one hyperparameter (the number of m bins to use) in addition to the decorrelation-classification tradeoff parameter λ.

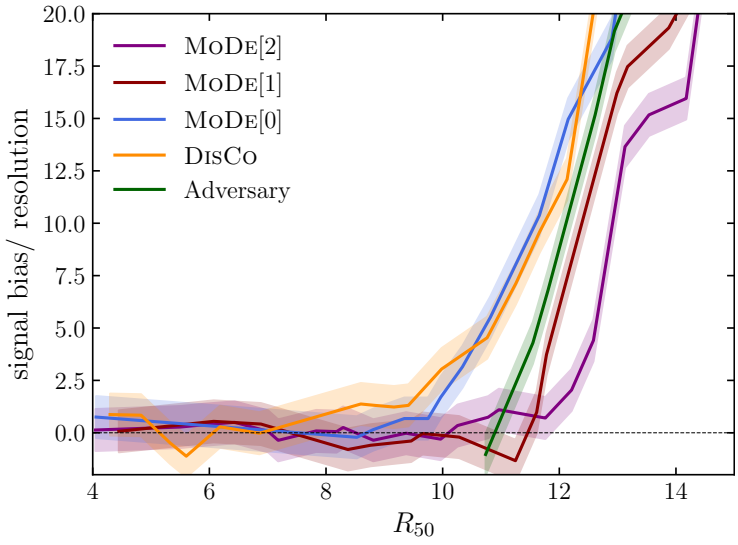

Figure 2: Signal bias (zero is ideal) relative to resolution versus background-rejection power (larger is better.) The flexibility beyond simple decorrelation provided by MoDe[1] and MoDe[2] results in improved performance. i.e. classifiers have better classification power and introduce no signal bias.

Broader Impact

While the MODE loss proposed here was developed for particle physics, it can be applied to fairness and explainability in AI problems in a variety of fields. This can lead to increased transparency of machine learning models and strengthened trust in their use. One example would be to develop a disease risk model for individual patients with a monotonic dependence on age despite a non-trivial age dependence of the the input features. Furthermore, using the MODE loss as a decorrelation method can decrease discrimination against protected groups in decisions ranging from issuing job offers to granting prison parole. Finally, the MODE loss can make modelling tasks simpler and their solutions more sensitive in other scientific domains, leading to heightened output.

References

[1] M. Frate, K. Cranmer, S. Kalia, A. Vandenberg-Rodes and D. Whiteson, *Modeling Smooth Backgrounds and Generic Localized Signals with Gaussian Processes*, 1709.05681.

[2] A. J. Larkoski, I. Moult and B. Nachman, *Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning*, *Phys. Rept.* **841** (2020) 1–63, [1709.04464].

[3] D. Guest, K. Cranmer and D. Whiteson, *Deep Learning and its Application to LHC Physics*, 1806.11484.

[4] K. Albertsson et al., *Machine Learning in High Energy Physics Community White Paper*, 1807.02876.

[5] A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel et al., *Machine learning at the energy and intensity frontiers of particle physics*, *Nature* **560** (2018) 41–48.

[6] D. Bourilkov, *Machine and Deep Learning Applications in Particle Physics*, *Int. J. Mod. Phys. A* **34** (2020) 1930019, [1912.08245].

[7] ATLAS collaboration, M. Aaboud et al., *Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC*, *Eur. Phys. J. C* **79** (2019) 375, [1808.07858].

[8] CMS collaboration, A. M. Sirunyan et al., *Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques*, *JINST* **15** (2020) P06005, [2004.08262].

[9] H. Edwards and A. J. Storkey, *Censoring representations with an adversary*, in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2016, http://arxiv.org/abs/1511.05897.

[10] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette et al., *Domain-adversarial training of neural networks*, *Journal of Machine Learning Research* **17** (2016) 1–35.

[11] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, *A survey on bias and fairness in machine learning*, 2019.

[12] A. Chouldechova and A. Roth, *The frontiers of fairness in machine learning*, 2018.

[13] O. Kitouni, B. Nachman, C. Weisser and M. Williams, *Enhancing searches for resonances with machine learning and moment decomposition*, 2010.09745.

[14] S. Chang, T. Cohen and B. Ostdiek, *What is the Machine Learning?*, *Phys. Rev. D* **97** (2018) 056009, [1709.10106].

[15] L. M. B. N. L. de Oliveira, M. Kagan and A. Schwartzman, *Jet-Images – Deep Learning Edition.*, *JHEP* **07** (2016) 069, [1511.05190].

[16] G. Louppe, M. Kagan and K. Cranmer, *Learning to Pivot with Adversarial Networks*, 1611.01046.

[17] C. Shimmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul et al., *Decorrelated Jet Substructure Tagging using Adversarial Neural Networks*, 1703.03507.

[18] C. Englert, P. Galler, P. Harris and M. Spannowsky, *Machine Learning Uncertainties with Adversarial Neural Networks*, *Eur. Phys. J.* **C79** (2019) 4, [1807.08763].

[19] J. M. Clavijo, P. Glaysher and J. M. Katzy, *Adversarial domain adaptation to reduce sample bias of a high energy physics classifier*, 2005.00568.

[20] G. Kasieczka and D. Shih, *DisCo Fever: Robust Networks Through Distance Correlation*, 2001.05310.

[21] G. Kasieczka, B. Nachman, M. D. Schwartz and D. Shih, *ABCDisCo: Automating the ABCD Method with Machine Learning*, 2007.14400.

[22] A. Rogozhnikov, A. Bukva, V. Gligorov, A. Ustyuzhanin and M. Williams, *New approaches for boosting to uniformity*, *JINST* **10** (2015) T03002, [1410.4140].

[23] ATLAS COLLABORATION collaboration, *Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS*, Tech. Rep. ATL-PHYS-PUB-2018-014, CERN, Geneva, Jul, 2018.

[24] P. Ramachandran, B. Zoph and Q. V. Le, *Searching for activation functions*, 2017.

[25] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015, http://arxiv.org/abs/1412.6980.

[26] L. N. Smith and N. Topin, *Super-convergence: Very fast training of neural networks using large learning rates*, 2018.

[27] I. Loshchilov and F. Hutter, *Sgdr: Stochastic gradient descent with warm restarts*, 2017.

[28] C. M. Bishop, *Mixture density networks*, tech. rep., 1994.