Published for SISSA by Springer

Received: January 5, 2023
Revised: May 8, 2023
Accepted: July 9, 2023
Published: July 25, 2023

Resonant anomaly detection with multiple reference datasets

Mayee F. Chen,[a] **Benjamin Nachman**[b,c] **and Frederic Sala**[d]

[a] *Computer Science Department, Stanford University, 353 Jane Stanford Way, Stanford, CA 94305, U.S.A.*
[b] *Physics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, U.S.A.*
[c] *Berkeley Institute for Data Science, University of California, 190 Doe Library, Berkeley, CA 94720, U.S.A.*
[d] *Department of Computer Sciences, University of Wisconsin, 1210 W. Dayton Street, Madison, WI 53706, U.S.A.*

E-mail: mfchen@stanford.edu, bpnachman@lbl.gov, fredsala@cs.wisc.edu

Abstract: An important class of techniques for resonant anomaly detection in high energy physics builds models that can distinguish between reference and target datasets, where only the latter has appreciable signal. Such techniques, including Classification Without Labels (CWoLa) and Simulation Assisted Likelihood-free Anomaly Detection (SALAD) rely on a single reference dataset. They cannot take advantage of commonly-available multiple datasets and thus cannot fully exploit available information. In this work, we propose generalizations of CWoLa and SALAD for settings where multiple reference datasets are available, building on weak supervision techniques. We demonstrate improved performance in a number of settings with realistic and synthetic data. As an added benefit, our generalizations enable us to provide finite-sample guarantees, improving on existing asymptotic analyses.

Keywords: Jets and Jet Substructure, Other Weak Scale BSM Models, Specific BSM Phenomenology

ArXiv ePrint: 2212.10579

https://doi.org/10.1007/JHEP07(2023)188

JHEP07(2023)188

Contents

JHEP07(2023)188

1 Introduction

Due to the vast parameter space of Standard Model extensions and to the lack of significant evidence for new particles or forces of nature, a new model-agnostic search paradigm has emerged. Many of these *anomaly detection* (AD) strategies are enabled by machine learning (see e.g. refs. [1–4]) and the first results with collision data are now available [5, 6]. One way to characterize AD methods is based on their physics assumption of the new phenomena [2]. Strategies that assume the new physics is "rare" [7] estimate (explicitly or implicitly) the data probability density and focus on events with low density. In contrast, techniques that assume the new physics will manifest as an overdensity in phase space use likelihood ratio methods to compare a reference dataset to a target dataset. The latter approach has been extensively studied in the context of *resonant anomaly detection* [8], where one resonant feature (usually a mass) is used to create a sideband region (reference dataset) nearly devoid of any anomalous events and a signal region (target dataset) that may contain anomalies. The reference dataset is used to estimate the presence of anomalies in the target dataset via interpolation.

Many existing approaches are defined using one reference dataset and one target dataset [9–18, 18–24]. However, in practice one can have access to or construct *multiple* references. First, there may exist multiple resonant features that can be used to construct sideband and signal regions. For instance, when a particle decays into two new particles, the decay products can be used to construct all three intermediate resonances, a setting present in the LHC Olympics Dataset [3]. Second, there may also exist multiple independent Standard Model simulators available for producing a dataset (e.g. Pythia [25], Herwig [26], or Sherpa [27]). Using multiple reference datasets may improve performance, but it is not clear how to incorporate all of their information when using existing methods designed for a single set.

We explore two generalizations of resonant AD to multiple reference datasets. First, we consider Classification Without Labels (CWoLa) [9, 10, 28], in which the reference is simply the sideband region — a form of weak supervision where the noisy label of "signal" is assigned to events in the signal region and the noisy label of 'background' to events in the sideband region. We propose a new method, Multi-CWoLa, that builds multiple reference datasets by constructing signal and sideband regions along different resonant features. We consider a point's membership in each feature's signal region as a noisy vote for anomaly, learn weights on each vote, and aggregate them to produce a higher-quality noisy label. We demonstrate Multi-CWoLa's performance on the LHC Olympics Dataset [3].

Next, we study Simulation Assisted Likelihood-free Anomaly Detection (SALAD) [14]. In this method, a reweighting function between a reference simulation dataset and a target dataset is learned in the sideband conditioned on the resonant feature. The simulated events in the signal region are reweighted by interpolating this function and then are used to distinguish anomalies in the target dataset. We extend this to the case of multiple simulated datasets, each of which may make different approximation choices and thus provide complementary accuracy when using SALAD. We introduce Multi-SALAD, which combines the simulated datasets accordingly and then reweights, with the key find-

JHEP07(2023)188

ing that combining data helps when each simulator approximates different components of the background well. We demonstrate MULTI-SALAD's performance on synthetic data. Reweighting simulations is a widely-used procedure in high energy physics, well beyond anomaly detection. While we focus on anomaly detection here because of the need for model-independent bounds, the approach here is generally applicable.

Finally, we study the finite sample guarantees of our proposed methods. Many resonant AD methods have optimality guarantees in some asymptotic limit, but there is no first-principles understanding of the methods' performance with finite samples. In particular, approaches like the ones described above that use classifiers to distinguish a reference dataset from a target dataset approximate the signal-to-background likelihood ratio. When the reference (physics) model is correct, this approach will converge to the optimal[1] Neyman-Pearson likelihood ratio test in the limit of infinite statistics, complex enough classifier architecture, and flexible enough training procedure [15, 29]. However, a finite sample understanding of these approaches is lacking. We draw on results from statistical theory to begin a formal study of resonant AD methods with limited data. Our results lay a foundation for future investigations into the finite sample properties of AD and related methods.

This paper is organized as follows. Section 2 briefly set up the resonant AD setting and then MULTI-CWOLA and MULTI-SALAD are introduced in sections 3 and 4, respectively. The paper ends with conclusions and outlook in section 5.

2 Problem setup

We have an input space of *discriminating features* $x \in \mathcal{X}$ and k *resonant features* $m = [m^1, \ldots, m^k] \in \mathbb{R}^k$. Associated with a point (x, m) is an unknown label $y \in \mathcal{Y}$ for $\mathcal{Y} = \{0, 1\}$ (background vs. signal). Points (x, m, y) are drawn from a distribution $\mathcal{P}$ with density $p(\cdot)$. For a resonant feature $m^i \in \mathbb{R}$, an interval $\mathcal{I}_{m^i} \subset \mathbb{R}$ is used to define a signal region $SR_i = \{(x, m) : m^i \in \mathcal{I}_{m^i}\}$ and a sideband region $SB_i = \{(x, m) : m^i \notin \mathcal{I}_{m^i}\}$ (when the resonant feature is obvious, the i is dropped and we use SR and SB). We assume that the sideband region contains little to no signal, i.e., $p(y = 1|(x, m) \in SB) \approx 0$. Our goal is to construct a predictor $f : \mathcal{X} \to \mathcal{Y}$ to perform anomaly detection.

3 Multi-CWoLa: learning from multiple resonant features

We introduce MULTI-CWOLA, an approach to anomaly detection that uses multiple reference datasets and is built using principles from the area of weak supervision [30, 31].

Standard CWoLa. We have one unlabeled dataset $\mathcal{D} = \{(x_i, m_i)\}_{i=1}^n$ with one resonant feature ($k = 1$) that we want to use to learn f. We use m to construct the signal and sideband regions, $\mathcal{D}_{SR}, \mathcal{D}_{SB} \subset \mathcal{D}$ where $\mathcal{D}_{SR} = \mathcal{D} \cap SR$ and $\mathcal{D}_{SB} = \mathcal{D} \cap SB$, with distributions p_{SR} and p_{SB} respectively. With the intuition that there are more anomalies in the signal

[1] Optimal for a given model is not achievable as we do not posit a particular signal hypothesis. Instead, we are optimal for the following hypothesis test where the null hypothesis is that data is drawn from the reference dataset and the alternative hypothesis is that the data is drawn from the target dataset. See appendix A in ref. [15] for details.

JHEP07(2023)188

region, we express each distribution as a mixture of signal and background components with weight $0 \leq \eta_{SR}, \eta_{SB} \leq 1$:

$$p_{SR}(x) = \eta_{SR} p(x|y=1) + (1-\eta_{SR}) p(x|y=0) \tag{3.1}$$

$$p_{SB}(x) = \eta_{SB} p(x|y=1) + (1-\eta_{SB}) p(x|y=0) \tag{3.2}$$

Under this construction, the density ratio of the mixtures $\frac{p_{SR}(x)}{p_{SB}(x)}$ can be written in terms of the ratio of the signal and background components, $r(x) = \frac{p(x|y=1)}{p(x|y=0)}$, as $\frac{p_{SR}(x)}{p_{SB}(x)} = \frac{\eta_{SR} r(x) + 1 - \eta_{SR}}{\eta_{SB} r(x) + 1 - \eta_{SB}}$. Assuming $\eta_{SR} > \eta_{SB}$ (e.g. more signal in the signal region), the mixture ratio is monotonically increasing in $r(x)$. Therefore, we train a classifier f to learn $\frac{p_{SR}(x)}{p_{SB}(x)}$ by distinguishing between $\mathcal{D}_{SR}$ and $\mathcal{D}_{SB}$, and this f provides information about $r(x)$ and can be used for anomaly detection. Note that CWoLa requires that x and m are independent of each other.

3.1 Multi-CWoLa method

Intuitively, CWoLa uses the resonant feature m as a noisy label that identifies the signal versus sideband region and then trains a classifier using these. This idea leads to a simple question — if more than one such feature is available ($k > 1$), how can the multiple noisy labels best be utilized? We tackle this question using principles from weak supervision [30–33]. The idea is that some features that were used to train the weakly supervised classifier could instead be used to improve the weak labels. In many cases, it is possible to know how to pick labeling functions with these 'resonant' features,[2] such as when a known particle mass is reconstructed. If the additional resonant features correspond to unknown particle masses, it may not be possible to know how to use the features to inform accurate weak labels. Scanning over possible intervals could result in a large trails factor. In the numerical examples below, we include both cases where the features are useful (we know the right interval) and not very useful (we do not know the right interval) for the weak labeling.

3.1.1 Model

In our approach, we split $\mathcal{D}$ along each resonant feature m^i to produce pairs of datasets $\mathcal{D}_{SB_i}$ and $\mathcal{D}_{SR_i}$ for each $i \in [k]$ based on membership in I_{m^i}. A straightforward way to use all datasets $(\mathcal{D}_{SB_1}, \mathcal{D}_{SR_1}), \ldots, (\mathcal{D}_{SB_k}, \mathcal{D}_{SR_k})$ is to apply standard CWoLa k times by training k classifiers that we can then ensemble or average. Instead, in Multi-CWoLa, we construct a binary vector per x consisting of k noisy membership labels, $\mathbf{M}(m) = \{M_1(m), \ldots, M_k(m)\} \in \{0,1\}^k$, where $M_i(m) = 1$ if $(x,m) \in \mathcal{D}_{SR_i}$ and $M_i(m) = 0$ if $(x,m) \in \mathcal{D}_{SB_i}$. We propose to directly aggregate these labels $\mathbf{M}(m)$ into an estimate of y, $\hat{y}$, and train a classifier on the aggregated $\hat{y}$ along with the discriminative features x. Since each $M_i(m)$'s "vote" can have different correlation with the true y, we aim to combine the votes in a weighted fashion. We cannot directly measure each membership label's accuracy since the true y is unknown, so we draw on methods from weak supervision.

[2] By resonant, we really mean that we can isolate a region of high signal-to-background — it does not have to be a closed interval.

JHEP07(2023)188

We model the distribution $p(y, \mathbf{M}(m))$ as a probabilistic graphical model with the following parametrization:

$$p(y, \mathbf{M}(m); \theta) = \frac{1}{Z} \exp\left(\theta_y \widetilde{y} + \sum_{i=1}^{k} \theta_i \widetilde{M}_i(m) \widetilde{y}\right), \tag{3.3}$$

where $\theta = \{\theta_y, \theta_i \ \forall i \in [k]\}$ are the canonical parameters of the distribution, Z is for normalization, and $\widetilde{y}$ and $\widetilde{M}_i(m)$ are y and $M_i(m)$ scaled from $\{0, 1\}$ to $\{-1, 1\}$. Intuitively, θ_i represents the (unobserved) strength of the correlation between $M_i(m)$ and y and thus captures a notion of M_i's accuracy. This model also implies, for simplicity, that $M_i(m) \perp M_j(m)|y$; that is, the resonant features are conditionally independent given y in addition to m and x being independent as in the standard CWoLa case.[3]

Our goal is to estimate the parameters of the graphical model and use them to perform inference, producing aggregated weak labels $\hat{y}$ from the distribution $p(y = 1|\mathbf{M}(m); \theta)$ given a vector of noisy labels $\mathbf{M}(m)$.

3.1.2 Parameter estimation

We first learn the parameters of $p(y, \mathbf{M}(m); \theta)$ as defined in (3.3). Of key interest is the accuracy parameter $\alpha_i = p(M_i(m) = 1|y = 1) = p(M_i(m) = 0|y = 0)$ of the ith resonant feature, which corresponds to the canonical parameter θ_i (see [35] for more background on probabilistic graphical models). We estimate the accuracy parameters by adapting the *triplet* approach from [31]. First, we draw triplets of resonant features $a, b, c \in [k]$.[4] If the distribution on $y, \mathbf{M}(m)$ follows the graphical model in (3.3), it holds that $yM_a(m) \perp yM_b(m)$ if $M_a(m) \perp M_b(m)|y$. Then, we have that $\mathbb{E}[\widetilde{y}\widetilde{M}_a(m)]\mathbb{E}[\widetilde{y}\widetilde{M}_b(m)] = \mathbb{E}[\widetilde{M}_a(m)\widetilde{M}_b(m)]$ since $\widetilde{y}^2 = 1$. Writing one such equation for each pair in the triplet (a, b, c), we have that

$$\begin{aligned}
\mathbb{E}[\widetilde{y}\widetilde{M}_a(m)]\mathbb{E}[\widetilde{y}\widetilde{M}_b(m)] &= \mathbb{E}[\widetilde{M}_a(m)\widetilde{M}_b(m)] \\
\mathbb{E}[\widetilde{y}\widetilde{M}_a(m)]\mathbb{E}[\widetilde{y}\widetilde{M}_c(m)] &= \mathbb{E}[\widetilde{M}_a(m)\widetilde{M}_c(m)] \\
\mathbb{E}[\widetilde{y}\widetilde{M}_b(m)]\mathbb{E}[\widetilde{y}\widetilde{M}_c(m)] &= \mathbb{E}[\widetilde{M}_b(m)\widetilde{M}_c(m)].
\end{aligned}$$

Solving this system, we obtain

$$|\mathbb{E}[\widetilde{y}\widetilde{M}_a(m)]| = \sqrt{\left|\frac{\mathbb{E}[\widetilde{M}_a(m)\widetilde{M}_b(m)]\mathbb{E}[\widetilde{M}_a(m)\widetilde{M}_c(m)]}{\mathbb{E}[\widetilde{M}_b(m)\widetilde{M}_c(m)]}\right|},$$

and similarly for b and c. We assume that each signal region is positively correlated with the true signal, which allows for us to ignore the absolute value and uniquely recover $\mathbb{E}[\widetilde{y}\widetilde{M}_a(m)]$. Next, we can use $\mathbb{E}[\widetilde{y}\widetilde{M}_a(m)] = 2p(\widetilde{y} = \widetilde{M}_a(m)) - 1$ to obtain α_i using properties of the graphical model in (3.3). Note that in practice, all of these quantities are empirical estimates, with terms such as $\hat{\mathbb{E}}[\widetilde{M}_a(m)\widetilde{M}_b(m)] = \frac{1}{n}\sum_{i=1}^{n} \widetilde{M}_a(m_i)\widetilde{M}_b(m_i)$.

[3] We can model some dependencies among resonant features if desired (see [31] for a method and see [34] for how to learn if resonant features are not conditionally independent). However, we need at least three conditionally independent subsets of resonant features in $\mathbf{M}(m)$ in order for the estimation method from [31] to recover the correct parameters.

[4] We assume that $k \geq 3$. In Lemma 1, we discuss why having $k = 1$ or $k = 2$ resonant features does not recover a unique model.

JHEP07(2023)188

Algorithm 1 MULTI-CWOLA.

Input: dataset $\mathcal{D} = \{(x_i, m_i)\}_{i=1}^n$; thresholds $\mathcal{I}_{m^i}$ that split $\mathcal{D}$ into signal and sideband regions, $\mathcal{D}_{SR_i}$ and $\mathcal{D}_{SB_i}$ respectively, for each m^i; class balance probability of anomaly $p(y=1)$

Construct noisy label $M_i(m) = \begin{cases} 1 & (x,m) \in \mathcal{D}_{SR_i} \\ 0 & (x,m) \in \mathcal{D}_{SB_i} \end{cases}$ for each resonant feature m^i.

for each triplet $a, b, c \in [k]$ **do**

$$\beta_a := \sqrt{|\hat{\mathbb{E}}[\widetilde{M}_a(m)\widetilde{M}_b(m)]\hat{\mathbb{E}}[\widetilde{M}_a(m)\widetilde{M}_c(m)]/\hat{\mathbb{E}}[\widetilde{M}_b(m)\widetilde{M}_c(m)]|} \tag{3.4}$$

$$\beta_b := \sqrt{|\hat{\mathbb{E}}[\widetilde{M}_a(m)\widetilde{M}_b(m)]\hat{\mathbb{E}}[\widetilde{M}_b(m)\widetilde{M}_c(m)]/\hat{\mathbb{E}}[\widetilde{M}_a(m)\widetilde{M}_c(m)]|} \tag{3.5}$$

$$\beta_c := \sqrt{|\hat{\mathbb{E}}[\widetilde{M}_a(m)\widetilde{M}_c(m)]\hat{\mathbb{E}}[\widetilde{M}_b(m)\widetilde{M}_c(m)]/\hat{\mathbb{E}}[\widetilde{M}_a(m)\widetilde{M}_b(m)]|}, \tag{3.6}$$

where $\hat{\mathbb{E}}$ is an empirical estimate of the expectation over $\mathcal{D}$, and $\widetilde{M}(m)$ indicates $M(m)$ scaled to $\{-1, 1\}$.

end for

Set accuracy parameter $\alpha_i = \hat{p}(M_i(m) = 1|y=1) = \hat{p}(M_i(m) = 0|y=0) = \hat{p}(M_i(m) = y) = \frac{\beta_i+1}{2}$.

Compute estimate $\hat{p}(y=1|\mathbf{M}(m)) \propto \prod_{i=1}^m \hat{p}(M_i(m)|y=1)p(y=1)$.

Construct $\hat{y} \sim \hat{p}(y=1|\mathbf{M}(m))$ for each $(x,m) \in \mathcal{D}$.

Output: classifier $\hat{f}$ for anomaly detection trained on $\{(x_i, m_i, \hat{y}_i)\}_{i=1}^n$.

3.1.3 Inference and training

After we learn the accuracy parameters, we use them to estimate $p(y=1|\mathbf{M}(m))$ for a given $\mathbf{M}(m)$. We use Bayes' rule and the conditional independence among $\mathbf{M}(m)$ to write $p(y|\mathbf{M}(m)) = \frac{\prod_{i=1}^m p(M_i(m)|y=1)p(y=1)}{p(\mathbf{M}(m))}$. We assume that the class balance $p(y=1)$ is known; otherwise, it can be estimated via tensor decomposition [33]. $p(M_i(m)|y=1)$ is either equal to α_i if $M_i(m) = 1$ or $1-\alpha_i$ if $M_i(m) = 0$, and the denominator $p(\mathbf{M}(m))$ can be either directly estimated since all quantities are observable or computed as $\prod_{i=1}^m p(M_i(m)|y=1)p(y=1) + \prod_{i=1}^m p(M_i(m)|y=0)p(y=0)$ using the estimated accuracies and class balance.

Once $p(y=1|\mathbf{M}(m))$ is estimated for all $\mathbf{M}(m) \in \{0,1\}^k$, the aggregated weak label $\hat{y}$ is drawn from such distribution. With labels $\hat{y}$ for each $(x,m) \in \mathcal{D}$, we train a classifier $\hat{f}$ on the weakly labeled dataset $\{(x, \hat{y})\}_{i=1}^n$. This procedure is summarized in Algorithm 1.

3.2 Theoretical results

Under (3.3), MULTI-CWOLA offers *finite-sample generalization* guarantees. Suppose the downstream model $\hat{f}$ trained on $\hat{y}$ belongs to class $\mathcal{F}$. Define a loss function $\ell_C : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and let the expected loss of f be $L_C(f) := \mathbb{E}\left[\ell_C(f(x), y)\right]$ on true labels. Then, the optimal classifier is $f^\star = \text{argmin}_{f\in\mathcal{F}} L_C(f)$, which is achieved with unlimited labeled data. Let the empirical loss of f on $\hat{y}$ be $\hat{L}_C(f) := \frac{1}{n}\sum_{i=1}^n \ell_C(f(x_i), \hat{y}_i)$. Then, the $\hat{f}$ we learn is constructed from $\hat{f} = \text{argmin}_{f\in\mathcal{F}} \hat{L}_C(f)$, which is learned on finite and noisily labeled

JHEP07(2023)188

data. Note that this construction is different from the standard empirical risk minimization (ERM) loss on labeled data, and thus $\hat{L}_C(f)$ does not asymptotically equal $L_C(f)$. We aim to minimize the *generalization error* $L_C(\hat{f}) - L_C(f^\star)$.

We now present our result on an upper bound for $L_C(\hat{f}) - L_C(f^\star)$. Define the Rademacher complexity of $\mathcal{F}$ as $\mathfrak{R}_n(\ell \circ \mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \ell(f(x_i), y_i)\right]$ with random variables $\Pr(\varepsilon = 1) = \Pr(\varepsilon = -1) = \frac{1}{2}$. Define $e_{\min}$ as the minimum eigenvalue of the covariance matrix on $[y, M_1(m), \ldots, M_k(m)]$, and let $a_{\min}$ be the minimum value of $\mathbb{E}[\widetilde{M}_i(m)y]$ over all i.

Theorem 1. *Assume that $p(y, \mathbf{M}(m))$ can be parametrized according to (3.3) and that ℓ is scaled to be bounded in $[0, 1]$. Assume that the class balance $p(y)$ is known (if not, there are ways to estimate it [33]), and that $k \geq 3$. Then, with probability at least $1 - \delta$, the generalization error of* Multi-CWoLa *on $\mathcal{D}$ is at most*

$$L_C(\hat{f}) - L_C(f^\star) \leq 4\mathfrak{R}_n(\ell \circ \mathcal{F}) + 2\sqrt{\frac{\log 2/\delta}{2n}} + \frac{c_1}{e_{\min} a_{\min}^5}\left(\sqrt{\frac{k}{n}} + \frac{c_2 k}{\sqrt{n}}\right),$$

where c_1, c_2 are positive constants.

The proof of Theorem 1 is provided in appendix E. We observe that there are several quantities controlling the above bound:

- The *Rademacher complexity* of $\mathcal{F}$: this term describes the model's expressivity. Smaller Rademacher complexity means that the model is easier to learn and that our $\hat{f}$ will be closer to the best model in $\mathcal{F}$. This quantity can be readily computed for a variety of function classes $\mathcal{F}$, such as decision trees, linear models, and two-layer feedforward networks, which makes our bound in Theorem 1 tractable. See appendix D.2 for exact values.

- Using n finite samples: as the amount of data increases, the error decreases in $\mathcal{O}(n^{-1/2})$.

- Using noisy labels $\hat{y}$ instead of y: for our weak supervision algorithm and graphical model, using $\hat{y}$ rather than y contributes an additional $\mathcal{O}(n^{-1/2})$ error. Asymptotically, our approach thus does no worse than training with labeled data.

- The number of resonant features k: as k increases, the expression increases. This is due to the fact that larger k results in more accuracy parameters to estimate.

By contrast, the standard CWoLa approach with $k = 1$ does not utilize any aggregation or weak supervision, which requires $k \geq 3$. For standard CWoLa, the second term in the generalization error is irreducible due to the fact that using any single resonant feature in place of y is biased; see Theorem 3 in appendix D for the exact generalization bound. On the other hand, Multi-CWoLa corrects for some of this bias; the second term asymptotically approaches 0 with more data.

JHEP07(2023)188

3.3 Empirical results

In figure 1, we compare Multi-CWoLa with standard CWoLa as well as three other baselines. For this study, we use datasets from the LHC Olympics [3]. This data challenge was created to develop and test resonant anomaly detection methods. The setting is events with a pair of hadronic jets whose invariant mass is in excess of 1 TeV. Inclusive bump hunts are broadly sensitive to localized excess in the dijet invariant mass spectrum, but are not particularly sensitive when the jets have complex substructure. For example, if the jets are formed from a new resonance A that decays to two other resonances B and C with $m_B, m_C \ll m_A$ and $B, C \to qq'$, then B and C particles will form two-prong jets that are readily distinguishable from the one-prong quark and gluon jets in the background. This signal scenario is exactly what is present in the LHC Olympics dataset. All events are simulated with Pythia 8 [25] and for simplicity, are summarized by five features: the invariant mass of the two jets ($k = 1$), the masses of the two jets, and a measure of the two-pronginess of the two jets (via the n-subjettiness ratio τ_{21} [36, 37]). In the standard CWoLa setup, we use one thresholded resonant feature ($k = 1$) and use 4 discriminative features as x. For Multi-CWoLa, we have generated $k = 3$ mixtures by varying how the 3 resonant features (the jet masses in addition to the dijet mass) are thresholded and use 2 discriminative features as x. Since we do not know the values of the masses, we pick windows in the middle of the spectrum for the individual jet masses (correct for one jet and not the other), as shown in figure 7. The proportion of anomalies in the training dataset is 0.149, while the proportion in the test dataset is 0.289. We have three other baselines that utilize 3 resonant features:

- CWoLa + intersect defines the signal region as the intersection of the resonant features' signal regions, e.g. $SR = SR_1 \cap SR_2 \cap SR_3$, but this can be overly conservative.
- CWoLa $+x$ thresholding has one resonant feature as the noisy label $\hat{y} = M_1(m)$, and includes the remaining thresholded features as discriminative $\{M_2(m), M_3(m), x\}$.
- CWoLa + average runs standard CWoLa three times, once per resonant feature and with the 2 discriminative features. The three model scores are averaged to produce the final output.

We vary the number of samples available on a logarithmic scale from $n = 59$ to 6003 and plot the AUC averaged over 5 runs per sample size in 1. We find that Multi-CWoLa offers a higher AUC and lower variance, especially when there is limited data. We also plot the (Significance Improvement) SI curves averaged over 5 runs for $n = 59, 530, 6003$ in 2. The SI corresponds to a multiplicative factor by which the significance of the anomaly increases with a corresponding threshold set by the x-axis. Numerically, the SI is the true positive rate divided by the square root of the false positive rate. The anomaly detection is useful only if the max SI is above 1. In practice, we do not know what threshold to make (which is signal-model dependent), so a set of cuts could be applied. Scanning over the cut would introduce a trials factor, but it is non-trivial to calculate given the correlation between different thresholds. Recent experimental results have used a couple of thresholds that are widely different so that these correlations and the trials factor are both small [5]. More experimental details and results are in appendix F.

JHEP07(2023)188

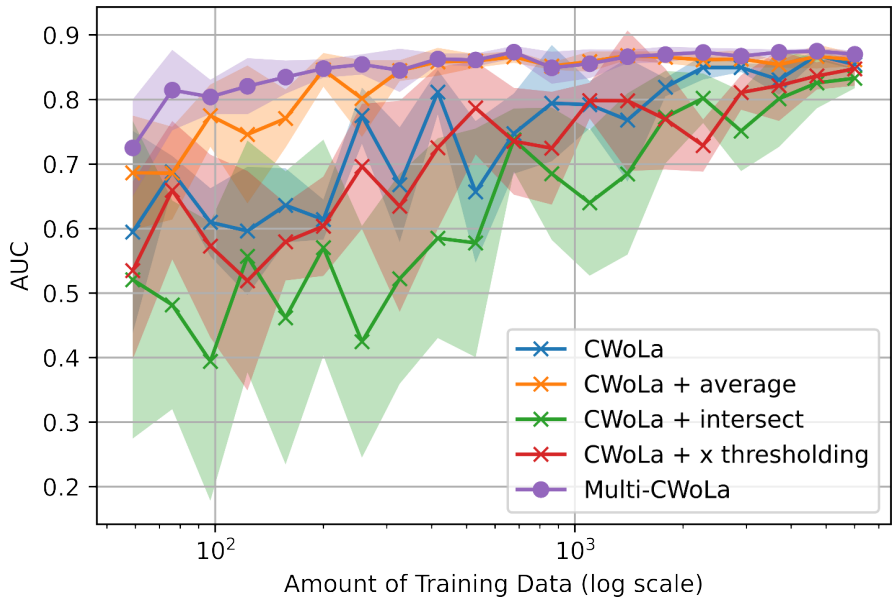

Figure 1. Comparison between CWoLa and Multi-CWoLa. Using multiple mixed samples helps performance across a range of dataset sizes. Access to multiple weak sources enables better AUC and lower variance compared to the single-feature version.

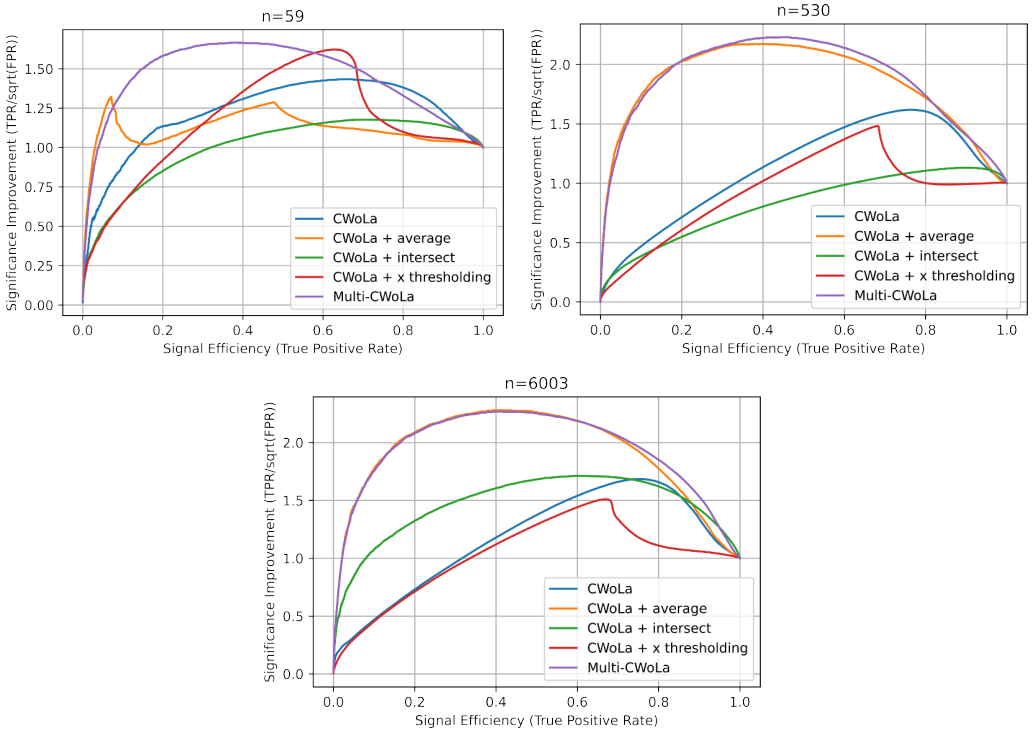

Figure 2. Significance Improvement (SI) curve for Multi-CWoLa at sizes $n = 59, 530$, and 6003.

JHEP07(2023)188

4 Multi-SALAD: learning from multiple simulations

We often have access to a(n approximate) simulation of the background process. We first provide an overview of SALAD, which reweighs samples from the simulation to better assist with classification on the real dataset. Then, we present MULTI-SALAD, a variant of SALAD that uses multiple simulations.

Standard SALAD. We have a background simulation dataset $\mathcal{D}^{\text{sim}} = \{(x_i, m_i)\}_{i=1}^{n_{\text{sim}}}$ with $y_i = 0$ for all i in addition to one true dataset $\mathcal{D} = \{(x_i, m_i)\}_{i=1}^{n}$. $\mathcal{D}^{\text{sim}}$ is drawn from some distribution $\mathcal{P}_{\text{sim}}$ with density p_{sim}. While CWoLA learns the likelihood ratio between the signal and sideband regions of $\mathcal{D}$ alone, SALAD utilizes $\mathcal{D}^{\text{sim}}$ as well. Note that if p_{sim} is equal to $p(\cdot|y = 0)$, we could directly train a model to distinguish between $\mathcal{D}$ and $\mathcal{D}^{\text{sim}}$ in the signal region to get a classifier that could detect anomalies. However, since $\mathcal{D}^{\text{sim}}$ may not match the true background data, we instead first need to learn a reweighting function that captures the differences between $\mathcal{D}^{\text{sim}}$ and $\mathcal{D}$'s background data, and then we train a model to distinguish between $\mathcal{D}$ and the reweighted $\mathcal{D}^{\text{sim}}$ in the signal region. Formally, given fixed SR and SB for both datasets, the method can be broken into two steps:

1. **Reweighting:** a classifier $\hat{g}$ is trained to distinguish between $\mathcal{D}^{\text{sim}}_{SB} = \mathcal{D}^{\text{sim}} \cap SB$ and $\mathcal{D}_{SB}$. Assuming that the sideband region has no anomalies, this $\hat{g}$ is able to produce an estimate of the weight ratio[5] $w(x, m) = \frac{p(x,m|y=0)}{p_{\text{sim}}(x,m|y=0)} \approx \frac{\hat{g}(x,m)}{1-\hat{g}(x,m)}$, assuming that the datasets are the same size ($|\mathcal{D}^{\text{sim}}_{SB}| = |\mathcal{D}_{SB}|$).

2. **Detection:** using a loss function L_S with estimated $\hat{w}(x, m)$ applied to $\mathcal{D}^{\text{sim}}_{SR} = \mathcal{D}^{\text{sim}} \cap SR$, a classifier $\hat{h}$ is trained to distinguish between $\mathcal{D}_{SR}$ and $\mathcal{D}^{\text{sim}}_{SR}$.

If the estimate $\hat{w}(x, m)$ is exactly equal to $w(x, m)$ (e.g. $\hat{g}$ is Bayes-optimal), then the second step will be equivalent in expectation to learning the ratio $\frac{p(x)}{p(x|y=0)}$ (see Lemma 2 in appendix D.4), from which one can detect anomalies.

4.1 Multi-SALAD method

Now, we have multiple simulation datasets $\mathcal{D}^{\text{sim}}_1, \ldots, \mathcal{D}^{\text{sim}}_k$. One approach would be to maintain distinctions among simulations by reweighing each pair to learn k weight functions $w_i(x, m)$, and then using one overall loss function that weights points from each $\mathcal{D}^{\text{sim}}_{SR,i}$ with w_i. However, it has been shown that importance reweighting, despite working in expectation, can be highly unstable and result in poor performance of tasks on the target data $\mathcal{D}$ [42]. To understand why, ref. [43] showed that the generalization error of an empirical loss function with importance weights w depends on the magnitude of w. Applied to our setting, it suggests that the more inaccurate the simulation is, the less the reweighted loss recovers the true $\frac{p(x)}{p(x|y=0)}$, and the model may instead pick up on differences between $\mathcal{D}_{SR}$ and the reweighted $\mathcal{D}^{\text{sim}}_{SR}$ that are noise rather than the anomaly. As a result, aggregating individual SALAD outputs can be equivalent to ensembling many poor classifiers.

[5] This is with the binary cross entropy loss function (also works for other functions [38]). This likelihood-ratio trick is well-known (see e.g. refs. [39, 40]), also in high-energy physics (see e.g. ref. [41]).

JHEP07(2023)188

Given these observations, MULTI-SALAD uses multiple simulation datasets in a very simple yet theoretically principled way: control the magnitude of the overall w by combining all the $\mathcal{D}_i^{\text{sim}}$ to produce one large simulation dataset $\widetilde{\mathcal{D}}^{\text{sim}}$ whose distribution best approximates the true background $p(x|y=0)$, and then use standard SALAD with $\widetilde{\mathcal{D}}^{\text{sim}}$ and $\mathcal{D}$. Note that this approach both improves sample complexity and can "suppress" a simulation that on its own has high w, while the approach of learning k weight functions would not offer such improvements. In Algorithm 2 and appendix C.1, we write this procedure out where we simply concatenate all $\mathcal{D}_i^{\text{sim}}$ together. However, with domain knowledge on the strengths and weaknesses of each simulation across features, one could produce $\widetilde{D}^{\text{sim}}$ by sampling accordingly from each. We leave this direction for future work.

4.2 Theoretical results

We now present a finite sample generalization error bound on MULTI-SALAD that also applies to SALAD. To measure the generalization error, recall $w(x,m) = \frac{p(x,m|y=0)}{p_{\text{sim}}(x,m|y=0)}$ and let $\hat{w}$ be the classifier g's estimate. We denote h as the reweighted classifier. Let $h^\star = \text{argmin}_{h\in\mathcal{H}}\ L_S(h,w)$ and let $\hat{h} = \text{argmin}_{h\in\mathcal{H}}\ \hat{L}_S(h,\hat{w})$. We aim to bound $L_S(\hat{h},\hat{w}) - L_S(h^\star,w)$.

We first set up some definitions. Define n^{SR} as the number of points from $\mathcal{D}$ and $\widetilde{\mathcal{D}}^{\text{sim}}$ belonging to the signal region, and n^{SB} as the number of points belonging to the sideband. Let n_{sim}^{SR} be the number of points in $\widetilde{\mathcal{D}}^{\text{sim}}$ belonging to the signal region. Let $\hat{g}(x) \in [\hat{g}_{\min}, \hat{g}_{\max}]$ and $g^\star(x) \in [g^\star_{\min}, g^\star_{\max}]$, where $g^\star$ is the optimal classifier. Let $\mathfrak{R}_{n^{SR}}(\ell_S \circ \{H,G\})$ be the Rademacher complexity of the overall loss $L_S(h,w)$ across function classes $h\in\mathcal{H}, g\in\mathcal{G}$. Define $W = \max_{x,m} w(x,m)$ as the maximum ratio between the simulation and true background. Let $B_1 = \max\{-\log h^\star(x,m), -\log(1-h^\star(x,m))\}$ be based on the most extreme value of $h^\star$ (i.e. how far apart p and $p(\cdot|y=0)$ can be). Let $\eta = \max(-\log(1-h^\star(x,m)))$ for $x,m \in \mathcal{D}_{SR}^{\text{sim}}$. Let $\mathfrak{R}_{n^{SB}}(\ell\circ\mathcal{G})$ is the Rademacher complexity of the loss function class used for learning the reweighting, where ℓ is point-wise cross-entropy. Finally, let $B_2 = -\log(\min\{\hat{g}_{\min}, g^\star_{\min}\})$.

Theorem 2. *With probability at least $1-\delta$, there exists a constant $c>0$ such that the generalization error of MULTI-SALAD on $\widetilde{D}^{sim}$ and $\mathcal{D}$ is at most*

$$L_S(\hat{h},\hat{w}) - L_S(h^\star,w) \le 2\mathfrak{R}_{n^{SR}}(\ell_S\circ\{\mathcal{H},\mathcal{G}\}) + (1+WB_1)\sqrt{\frac{\log 8/\delta}{2n^{SR}}} \tag{4.1}$$
$$+ \frac{\eta n_{\text{sim}}^{SR}}{(1-\hat{g}_{\max})(1-g^\star_{\max})n^{SR}}\left(4c\mathfrak{R}_{n^{SB}}(\ell\circ\mathcal{G}) + 2c\sqrt{\frac{\log 4/\delta}{2n^{SB}}} + B_2\sqrt{\frac{\log 8/\delta}{2n_{\text{sim}}^{SR}}}\right).$$

We make several observations about this bound:

- The bound scales in $(n^{SB})^{-1/2}$ and $(n_{\text{sim}}^{SR})^{-1/2}$, where the former comes from the initial reweighting step while the latter comes from the weighted classification step.
- The bound is also dependent on the Rademacher complexities of both classifiers g and h used.

JHEP07(2023)188

- The bound depends on the difference between the simulation and data distributions through quantities W, $B_1, B_2, \eta, \hat{g}_{\max}, g_{\max}$. If the distributions have very different densities, these quantities will all be large, increasing the generalization error.

We comment how this bound is different when instantiated for SALAD versus Multi-SALAD. The following example shows how SALAD with one simulation can result in a large W (and other large constants), while Multi-SALAD with two simulations combined can reduce W in the bound.

Example 1. *Let $\mathcal{P}^1_{\text{sim}}(x|y=0) = \mathcal{N}(\mu, \sigma^2)$, $\mathcal{P}^2_{\text{sim}}(x|y=0) = \mathcal{N}(-\mu, \sigma^2)$ be Gaussian distributions on x with $\mu, \sigma^2 \in \mathbb{R}$, and let the true background distribution $\mathcal{P}(\cdot|y=0)$ be a mixture of the Gaussians on x, $\mathcal{P}(x|y=0) = \frac{1}{2}\mathcal{P}^1_{\text{sim}} + \frac{1}{2}\mathcal{P}^2_{\text{sim}}$. Let $\mathcal{P}^1_{\text{sim}}, \mathcal{P}^2_{\text{sim}}$, and $\mathcal{P}$ have the same marginal distribution over m with $x \perp\!\!\!\perp m|y$. Then, if we only use one simulation $\mathcal{P}^1_{\text{sim}}$,*

$$\begin{aligned} w(x,m) &= \frac{p(x,m|y=0)}{p^1_{\text{sim}}(x,m|y=0)} = \frac{p(x|y=0)}{p^1_{\text{sim}}(x|y=0)} \\ &= \frac{\frac{1}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) + \frac{1}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{(x+\mu)^2}{2\sigma^2}\right)}{\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} \\ &= \frac{1}{2} + \frac{1}{2}\exp\left(\frac{(x-\mu)^2}{2\sigma^2} - \frac{(x+\mu)^2}{2\sigma^2}\right) = \frac{1}{2} + \frac{1}{2}\exp\left(\frac{-2x\mu}{\sigma^2}\right). \end{aligned}$$

Therefore, as $x \to -\infty$, $W \to \infty$. However, if we define $\mathcal{P}_{\text{sim}}$ as the distribution of the two simulation datasets concatenated, we have that $p_{\text{sim}}(x|y=0) = p(x|y=0)$, and as a result, $W \to 1$, making the generalization error bound smaller.

From this example, we can see that significantly differing simulation and data distributions can result in large, unbounded weight ratios, which are correlated with poor performance.[6] This concretely motivates our algorithmic objective to combine multiple simulation datasets as to closely approximate the true data.

4.3 Empirical results

To demonstrate how Multi-SALAD can improve over using only one simulation and over using simulations separately, we consider a synthetic experiment with two simulation datasets.[7] The true background is $\mathcal{P}(\cdot|y=0) = \frac{1}{2}\mathcal{N}(-1, 0.2) + \frac{1}{2}\mathcal{N}(1, 0.2)$, and the anomaly is $\mathcal{P}(\cdot|y=1) = \frac{1}{2}\mathcal{N}(-1.5, 0.2) + \frac{1}{2}\mathcal{N}(1.5, 0.2)$. Simulation 1 is $\mathcal{P}^1_{\text{sim}} = \frac{1}{2}\mathcal{N}(1, 0.2) + \frac{1}{2}\mathcal{N}(0, 1)$, and simulation 2 is $\mathcal{P}^2_{\text{sim}} = \frac{1}{2}\mathcal{N}(-1, 0.2) + \frac{1}{2}\mathcal{N}(0, 1)$. We generate 2000 points from the true background and 100 points that are anomalies to form $\mathcal{D}$, and 2000 points each from $\mathcal{P}^1_{\text{sim}}$ and $\mathcal{P}^2_{\text{sim}}$ to form $\mathcal{D}^{\text{sim}}_1$ and $\mathcal{D}^{\text{sim}}_2$. We construct signal and sideband regions from these

[6] The bound in Theorem 2 is meant to provide a general understanding of SALAD's performance. It can be made tighter by replacing terms that are maxima like M and B_2 with terms that are based on the overall data distributions (e.g. variance, as in ref. [43]). Variance-based bounds are less likely to be vacuous, but will still demonstrate how performance is dependent on the intrinsic differences between the two distributions.

[7] We find that the differences between the simulations in the LHC Olympics are not enough to see a noticeable gain from Multi-SALAD over SALAD.

JHEP07(2023)188

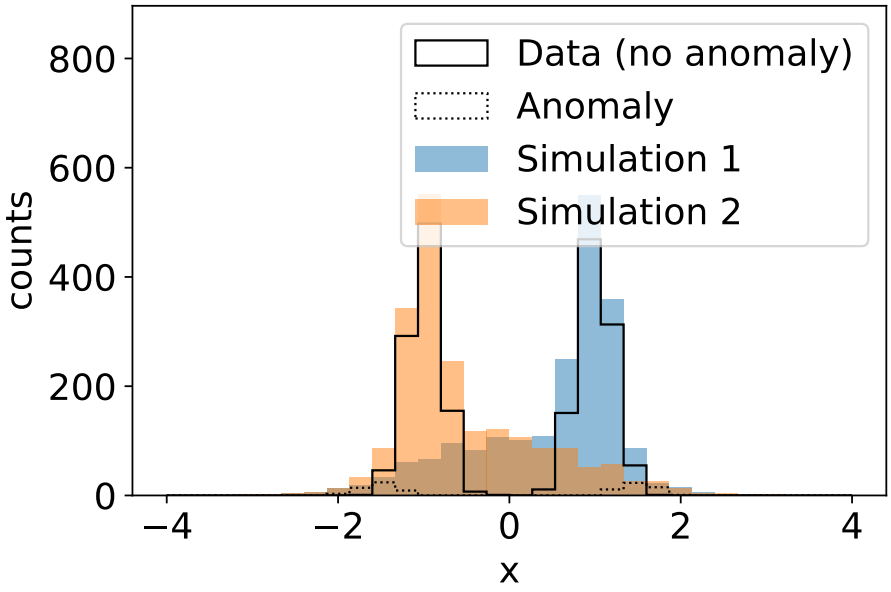

Figure 3. Synthetic data for evaluating Multi-SALAD.

by splitting datasets in half randomly, assuming they follow the same distribution over x (i.e., m is independent of x) except that there is no anomaly in the sideband regions. A visualization is shown in figure 3.

Intuitively, the anomaly is only slightly different from the background data, which makes it important to learn a good reweighting function from the simulations. Because each simulation alone diverges greatly from the data for one mode, each individual reweighting may not approximate the true $\mathcal{P}(\cdot|y = 1)$ well. On the other hand, if we combine both simulation datasets together, the aggregate distribution has smaller weights with lower variance, which can allow for more accurate reweighting. This is demonstrated in figure 4, which depicts the reweighting in the sideband region. We have plotted the true density of the data distribution in the sideband region, as well as the simulation's distribution and the reweighted simulation's distribution. In red, we plot the absolute value of the difference in the densities between the true data and the reweighted simulation. We find that the mean difference when using simulation 1 only is 0.1150, when using simulation 2 only is 0.1027, and when using simulation 1 and 2 is 0.0640.

Figure 5 depicts the reweighting's interpolation into the signal region, where we introduce an additional baseline SALAD-Switch, which uses k separate weight functions $w_i(x, m)$ and switches among them in the reweighted loss function L_S. We have plotted the true density of the data distribution in the signal region, which consists of both background and anomaly, and the reweighted estimate of the background data. In red, we plot the absolute value of the difference in the densities between the true data and the reweighted simulation. We find that the absolute difference in the densities is lower in regions of no anomaly (e.g., away from $\mathcal{N}(-1.5, 0.2)$ and $\mathcal{N}(1.5, 0.2)$) when using Multi-SALAD. Note that the reweighting does not remove discrepancies caused by the signal. Therefore,

JHEP07(2023)188

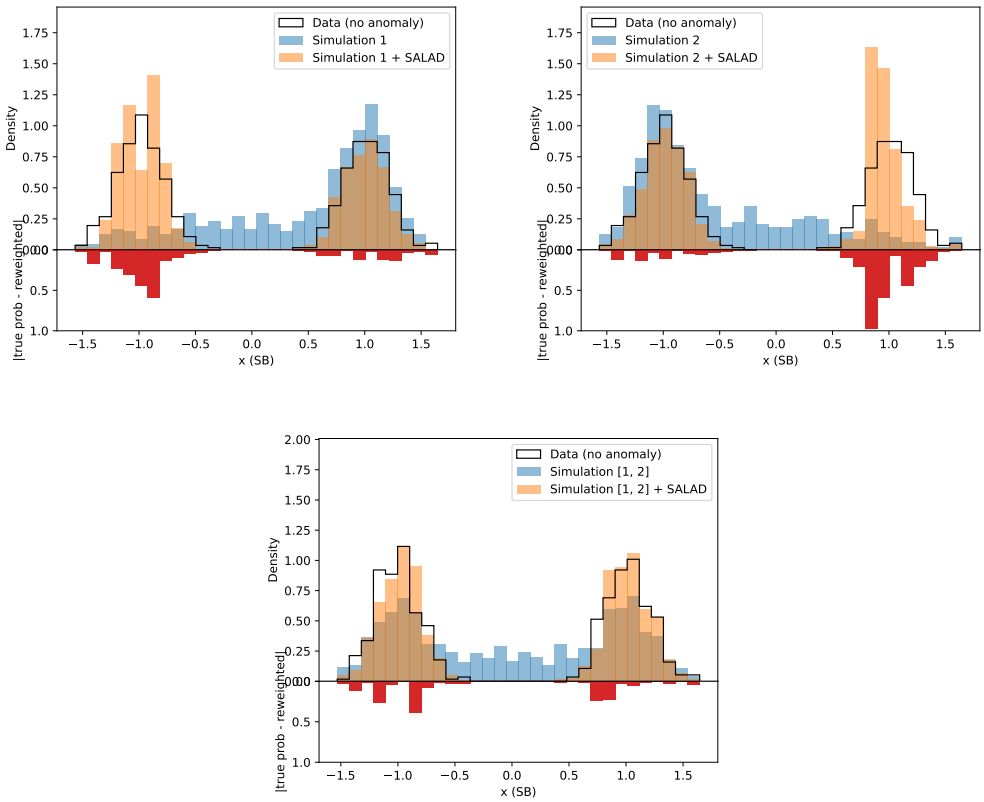

Figure 4. Top left: SALAD reweighting using simulation 1 on sideband region. Top right: reweighting using simulation 2. Bottom: reweighting using simulation 1 and 2 combined.

Multi-SALAD approximates the background data well, and so a classifier trained on this reweighted simulation will be able to distinguish differences between background and anomaly; on the other hand, a classifier trained on a high-variance reweighted simulation will more likely learn distinctions coming from poor approximation, rather than anomaly.

With these observations, we present the signal efficiency to rejection rate of each method in figure 6, where we compare Multi-SALAD against SALAD using simulation 1 only, SALAD using simulation 2 only, and SALAD-Switch. Table 1 contains the accuracy and AUC scores for each method. Averaged over 10 random seeds, Multi-SALAD outperforms other methods. The signal efficiency to rejection rate for each of the 10 runs is available in appendix F.

5 Conclusions and outlook

We extend two resonant AD approaches to incorporate multiple reference datasets. For Multi-CWoLa, we draw from weak supervision models to handle multiple resonant features. For Multi-SALAD, we combine multiple simulation datasets to best approximate the background process. Future work includes 1) exploring Multi-SALAD's applicability

JHEP07(2023)188

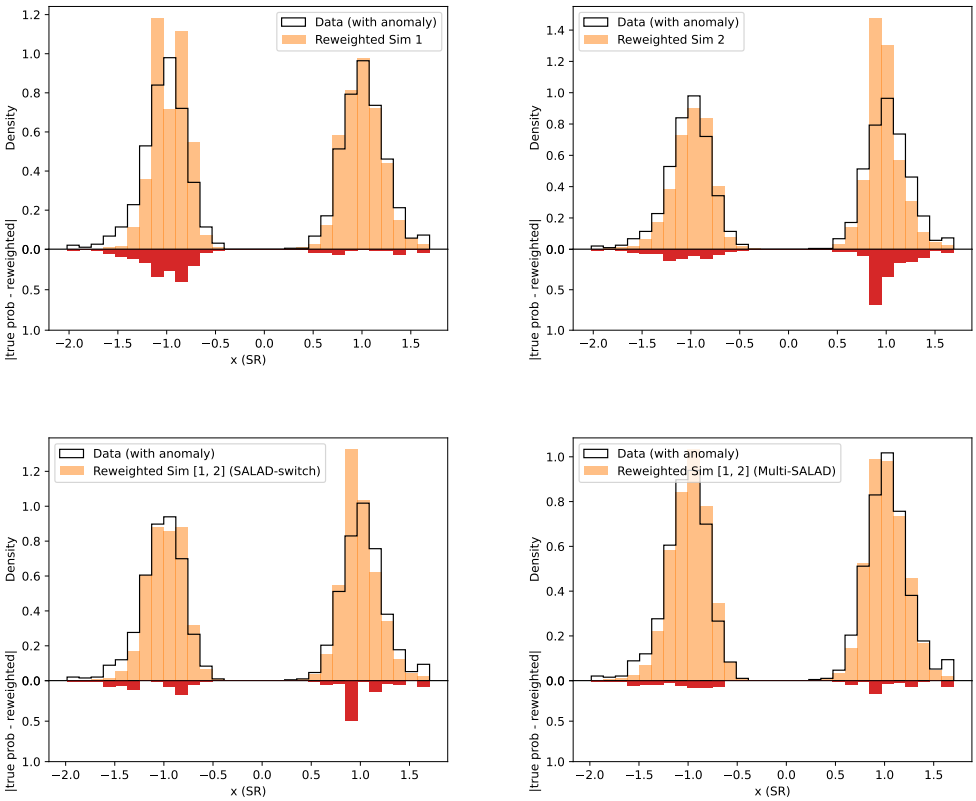

Figure 5. Top left: SALAD reweighting using simulation 1 on signal region. Top right: reweighting using simulation 2. Bottom left: using both simulation 1 and 2 weights separately. Bottom right: reweighting using simulation 1 and 2 combined.

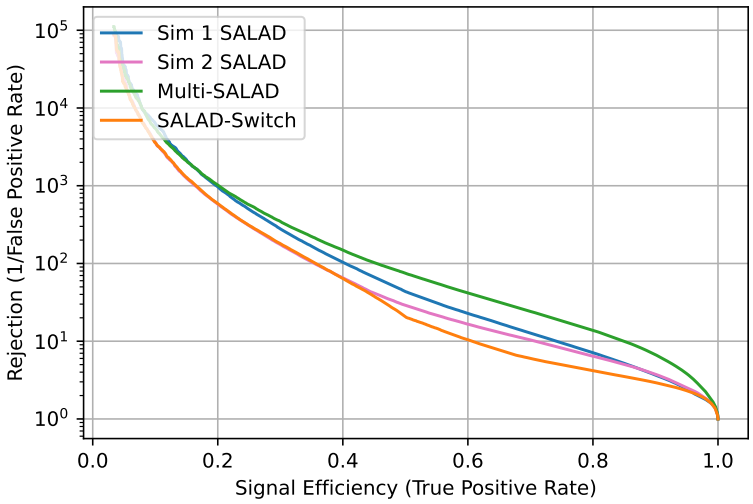

Figure 6. Signal efficiency to rejection of Multi-SALAD versus other baselines (weighted and unweighted).

JHEP07(2023)188

	Simulation 1		Simulation 2		Simulation 1 and 2		
Method	None	SALAD	None	SALAD	None	SALAD-Switch	Multi-SALAD
Accuracy	$45.6_{\pm 1.9}$	$58.3_{\pm 5.7}$	$44.7_{\pm 3.1}$	$62.1_{\pm 9.8}$	$50.0_{\pm 0.0}$	$54.8_{\pm 4.7}$	$\mathbf{65.7}_{\pm 8.2}$
AUC	$30.2_{\pm 3.8}$	$82.8_{\pm 12.0}$	$29.3_{\pm 3.2}$	$80.9_{\pm 13.1}$	$12.6_{\pm 5.1}$	$76.5_{\pm 15.2}$	$\mathbf{90.2}_{\pm 8.3}$

Table 1. Accuracy and AUC scores (%) for Multi-SALAD on two simulation datasets. We compare to SALAD-Switch (different reweighting), as well as standard SALAD on individual simulations and no reweighting. Performance is averaged over 20 random runs with one standard deviation reported.

on real data and algorithms for sampling from simulation datasets 2) extending Multi-CWoLa to model more complex relationships among resonant features and 3) using such approaches together over multiple simulations and resonant features, effectively utilizing as much information as possible.

Acknowledgments

We thank David Shih and Jesse Thaler for useful discussions and comments about the manuscript. BN was supported by the Department of Energy, Office of Science under contract number DE-AC02-05CH11231. FS is grateful for the support of the NSF under CCF2106707 and the Wisconsin Alumni Research Foundation (WARF). We gratefully acknowledge the support of NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML); ARL under No. W911NF-21-2-0251 (Interactive Human-AI Teaming); ONR under No. N000141712266 (Unifying Weak Supervision); ONR N00014-20-1-2480: Understanding and Applying Non-Euclidean Geometry in Machine Learning; N000142012275 (NEPTUNE); NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, Google Cloud, Salesforce, Total, the HAI-GCP Cloud Credits for Research program, the Stanford Data Science Initiative (SDSI), and members of the Stanford DAWN project: Facebook, Google, and VMWare. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of NIH, ONR, or the U.S. Government.

A Appendix organization

We provide a glossary of notation in B. We provide algorithmic details for Multi-SALAD in section C. We present additional theoretical results on Rademacher complexities and the asymptotic behavior of SALAD in section D. In section E, we provide proofs for our theoretical results. In section F, we provide additional experimental details.

JHEP07(2023)188

B Glossary

The glossary is given in table 2.

C Additional algorithmic details

C.1 Multi-SALAD algorithm

Multi-SALAD is described in Algorithm 2. We have simulation datasets $\mathcal{D}_1^{\text{sim}}, \dots \mathcal{D}_k^{\text{sim}}$, where $\mathcal{D}_i^{\text{sim}} = \{(x_j, m_j)\}_{j=1}^{n_{\text{sim}}}$ and all points belong to the background ($y = 0$). As discussed in section 4, we propose using these simulation datasets by aggregating them into a single simulation dataset $\mathcal{D}^{\text{sim}}$, whether it be by concatenating the datasets, stratified sampling — taking a fixed number of samples from each dataset — or something more advanced like importance sampling or weighting. Then the rest of this section proceeds as follows and is a review of the standard SALAD method.

Reweighting. First, we learn weights to correct for the bias of the simulated background data. We split the both simulation and true data along m to produce sets $\mathcal{D}_{SR}^{\text{sim}}, \mathcal{D}_{SB}^{\text{sim}}$ and $\mathcal{D}_{SR}$ and $\mathcal{D}_{SB}$. We train a classifier over $\mathcal{D}_{SB}^{\text{sim}}$ and $\mathcal{D}_{SB}$ to distinguish between simulation and real data in the sideband region. That is, we train a binary classifier $\hat{g}$ over points (x, m, z) in the sideband where x, m is either from $p_{\text{sim}}(\cdot|y = 0)$ $(z = 0)$ or $p(\cdot|y = 0)$ $(z = 1)$, where we recall that simulation data only contains $y = 0$, and no anomalies are present in the sideband. Denote q as the joint density of (x, m, z). We define the weight as the estimated likelihood ratio

$$\begin{aligned}\hat{w}(x, m) &= \frac{\hat{g}(x, m)}{1 - \hat{g}(x, m)} \approx \frac{q(z = 1|x, m)}{q(z = 0|x, m)} = \frac{q(x, m|z = 1)}{q(x, m|z = 0)} \cdot \frac{q(z = 1)}{q(z = 0)} \\ &= \frac{q(x, m|z = 1)}{q(x, m|z = 0)} = \frac{p(x, m|y = 0)}{p_{\text{sim}}(x, m|y = 0)}. \end{aligned} \tag{C.1}$$

Here, we assume that $q(z = 1) = q(z = 0)$ (i.e. balanced simulation and real dataset, which we can always ensure by generating more or less simulation data). Equality is obtained in the expression above when $\hat{g}$ is Bayes-optimal.

Training. The above $\hat{w}(x, m)$ is defined on the sideband region. Next, we interpolate and correct the bias of the simulation in the signal region. Let $\mathcal{D}_{SR}^{\text{sim}}$ be the set of simulation data in the signal region of size n_{sim}^{SR}, and let $\mathcal{D}_{SR}$ be the set of true data in the signal region of size n_{data}^{SR}, for a total of n^{SR} points. We train a classifier h to distinguish between the reweighted simulated data, which approximates true background data, and the true data. In particular, the loss function used is

$$\hat{L}_S(h, \hat{w}) = -\frac{1}{n^{SR}} \left(\sum_{x \in \mathcal{D}_{SR}} \log h(x, m) + \sum_{x \in \mathcal{D}_{SR}^{\text{sim}}} \hat{w}(x, m) \log(1 - h(x, m)) \right). \tag{C.2}$$

In expectation with an optimal w, we can see that minimizing this loss is equivalent to minimizing the cross-entropy loss on a task that distinguishes between points drawn from p and points drawn from $p(\cdot|y = 0)$ in the signal region. Therefore, h can be used for anomaly detection. The procedure is summarized in Algorithm 2.

JHEP07(2023)188

Symbol	Used for
x	Discriminative feature $x \in \mathcal{X}$.
m	Resonant feature vector of length k, $m = [m^1, \ldots, m^k] \in \mathbb{R}^k$.
y	True unknown label $y \in \mathcal{Y} = \{0, +1\}$, where 0 is background and 1 is signal.
$\mathcal{P}, p$	Distribution and density of data (x, m, y).
I_{m^i}	Interval along which ith resonant feature m^i is thresholded to produce signal region and sideband.
SR, SB	Signal region and sideband. For an interval I_{m^i}, $SR_i = \{(x,m) : m^i \in I_{m^i}\}$ and $SB_i = \{(x,m) : m^i \notin I_{m^i}\}$.
f	Classifier $f : \mathcal{X} \to \mathcal{Y}$ used for anomaly detection.
$\mathcal{D}$	Unlabeled dataset $\mathcal{D} = \{(x_i, m_i)\}_{i=1}^n$ of discriminative and resonant features.
$\mathcal{D}_{SR}, \mathcal{D}_{SB}$	Signal region and sideband of $\mathcal{D}$, $\mathcal{D}_{SR} = \mathcal{D} \cap SR$, $\mathcal{D}_{SB} = \mathcal{D} \cap SB$.
η_{SR}, η_{SB}	Mixture weights corresponding to $p(y=1 \mid x \in SR)$ and $p(y=1 \mid x \in SB)$. It is assumed that $\eta_{SR} > \eta_{SB}$.
$M_i(m)$	Noisy membership label for the ith resonant feature, equal to 0 if $x \in \mathcal{D}_{SB_i}$ and 1 if $x \in \mathcal{D}_{SR_i}$. $\mathbf{M}(m) = M_1(m), \ldots, M_k(m)$.
$\hat{y}$	Weak label drawn from estimated distribution on $p(y \mid \mathbf{M}(m))$.
θ_y, θ_i	Canonical parameters of graphical model on $y, \mathbf{M}(m)$ in (3.3). θ_y scales with the class balance of y and θ_i scales with the accuracy of $M_i(m)$.
Z	Partition function used for normalizing distribution $p(y, \mathbf{M}(m))$ in (3.3).
$\widetilde{y}, \widetilde{\mathbf{M}}(m)$	y and $\mathbf{M}(m)$ scaled from $\{0,1\}$ to $\{-1,1\}$.
α_i	Accuracy parameter $\alpha_i = p(M_i(m) = 1 \mid y = 1)$ for the membership label of the ith resonant feature.
ℓ_C	Loss function $\ell_C : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ for training classifier f.
$L_C(f)$	Expected loss on labeled data using f, $L_C(f) = \mathbb{E}\left[\ell_C(f(x), y)\right]$.
$f^\star$	Optimal classifier trained on infinite labeled data, $f^\star = \operatorname{argmin}_{f \in \mathcal{F}} L_C(f)$.
$\hat{L}_C(f)$	Empirical loss on $\mathcal{D}$ with weak labels using f, $\hat{L}_C(f) = \frac{1}{n} \sum_{i=1}^n \ell_C(f(x_i), \hat{y}_i)$.
$\hat{f}$	Classifier learned using Multi-CWoLa, $\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \hat{L}_C(f)$.
$\mathcal{D}^{\text{sim}}$	Simulation dataset used in standard SALAD, $\mathcal{D}^{\text{sim}} = \{(x_i, m_i)\}_{i=1}^{n_{\text{sim}}}$. Has distribution $\mathcal{P}_{\text{sim}}$ and density $p_{\text{sim}}(\cdot)$.
$\mathcal{D}^{\text{sim}}_{SB}$, $\mathcal{D}^{\text{sim}}_{SR}$	$\mathcal{D}^{\text{sim}}_{SB} = \mathcal{D}^{\text{sim}} \cap SB$, $\mathcal{D}^{\text{sim}}_{SR} = \mathcal{D}^{\text{sim}} \cap SR$.
$w(x,m)$	Density ratio between $\mathcal{D}^{\text{sim}}_{SB}$ and $\mathcal{D}_{SB}$ used for reweighting, $w(x,m) = \frac{p(x,m \mid y=0)}{p_{\text{sim}}(x,m \mid y=0)}$.
$\hat{g}$	Classifier trained to classify $\mathcal{D}^{\text{sim}}_{SB}$ vs $\mathcal{D}_{SB}$, used for approximating $w(x,m)$ when $\lvert\mathcal{D}^{\text{sim}}_{SB}\rvert = \lvert\mathcal{D}_{SB}\rvert$.
$L_S(h,w)$	Cross-entropy loss function used to classify $\mathcal{D}^{\text{sim}}_{SR}$ reweighted with w vs $\mathcal{D}_{SR}$.
$\hat{h}$	Classifier trained using L_S.
$\mathcal{D}^{\text{sim}}_1, \ldots, \mathcal{D}^{\text{sim}}_k$	k multiple simulation datasets used in Multi-SALAD.
$\widetilde{\mathcal{D}}_{\text{sim}}$	Dataset aggregated from $\mathcal{D}^{\text{sim}}_1, \ldots, \mathcal{D}^{\text{sim}}_k$.
n^{SR}	$n^{SR} = \lvert\mathcal{D}_{SR}\rvert$.
n^{SB}	$n^{SB} = \lvert\mathcal{D}_{SB}\rvert$.
n^{SR}_{sim}	$n^{SR}_{\text{sim}} = \lvert\mathcal{D}^{\text{sim}}_{SR}\rvert$.
$h^\star$	The optimal classifier $h^\star = \operatorname{argmin}_{h \in \mathcal{H}} L_S(h,w)$.
W	The maximum ratio between the simulation and true background, $W = \max_{x,m} w(x,m)$.

Table 2. Glossary of variables and symbols used in this paper.

JHEP07(2023)188

Algorithm 2 MULTI-SALAD.

1: **Input:** simulation datasets $\mathcal{D}_1^{\text{sim}}, \ldots, \mathcal{D}_k^{\text{sim}}$ and real dataset $\mathcal{D}$.
2: Construct overall simulation dataset $\mathcal{D}^{\text{sim}} = \bigcup_{i=1}^{k} \mathcal{D}_i^{\text{sim}}$.
3: Split each dataset into signal region and sideband region using resonant feature m to get $\{\mathcal{D}_{SR}^{\text{sim}}, \mathcal{D}_{SB}^{\text{sim}}\}$ and $\{\mathcal{D}_{SR}, \mathcal{D}_{SB}\}$.
4: Learn weight $\hat{w}(x,m) = \frac{\hat{g}(x,m)}{1-\hat{g}(x,m)}$, where $\hat{g}$ is a classifier that distinguishes data $\mathcal{D}_{SB}$ from simulation $\mathcal{D}_{SB}^{\text{sim}}$ in the sideband region.
5: Train a new classifier $\hat{h}$ on the signal region to distinguish between points in $\mathcal{D}_{SR}$ and points in $\mathcal{D}_{SR}^{\text{sim}}$ reweighted by $\hat{w}$, using the following loss:

$$\hat{L}_S(h,\hat{w}) = -\frac{1}{n^{SR}}\left(\sum_{x\in\mathcal{D}_{SR}} \log h(x,m) + \sum_{x\in\mathcal{D}_{SR}^{\text{sim}}} \hat{w}(x,m)\log(1-h(x,m))\right). \tag{C.3}$$

6: **Output:** classifier output $\hat{h}(x,m)$, which yields a score that is thresholded for anomaly detection.

SALAD-Switch. Here, we formally describe the baseline SALAD-SWITCH. Rather than combining the reference datasets, we keep them separate and learning a reweighting function on each. That is, the ith weight ratio function is $w_i(x,m) = \frac{p(x,m|y=0)}{p_{\text{sim}^i}(x,m|y=0)}$ for $x, m \in \mathcal{D}_i^{\text{sim}}$ and $i \in [k]$. The SALAD-SWITCH objective function is

$$\hat{L}_{\text{switch}}(h,\hat{w}) = -\frac{1}{n^{SR}}\left(\sum_{x\in\mathcal{D}_{SR}} \log h(x,m) + \sum_{i=1}^{k}\sum_{x\in\mathcal{D}_{SR,i}^{\text{sim}}} \hat{w}_i(x,m)\log(1-h(x,m))\right). \tag{C.4}$$

These two steps, learning w_i and the new objective function, correspond to lines 4 and 5 in Algorithm 2.

D Additional theoretical results

D.1 The need for 3 resonant features

We show that to identify the model (3.3), we need at least $k = 3$ resonant features.

Lemma 1. *If $k = 1$ or $k = 2$ in model (3.3), the parameters θ_1 and θ_2 cannot be recovered from the observable quantities.*

Proof. The strategy we use to show that the model cannot be identified for $k = 1$ or $k = 2$ is to prove that the observable distributions $P(\widetilde{M}_1(m), \ldots, \widetilde{M}_k(m))$ are consistent with multiple values of θ. We do so by direct calculation.

First, consider the case of $k = 1$. Set $\theta_y = 0$ for simplicity. Then, the model is $\frac{1}{Z}\exp(\theta\widetilde{M}_1(m)\widetilde{y})$. Then $Z = 2\exp(\theta) + 2\exp(-\theta)$, and

$$P(\widetilde{M}_1(m) = 1) = \frac{\exp(\theta)+\exp(-\theta)}{2\exp(\theta)+2\exp(-\theta)} = \frac{1}{2}.$$

Thus, any θ value produces the same observable distribution, so that we cannot identify θ.

JHEP07(2023)188

Next, we consider $k = 2$. Again, set $\theta_y = 0$. The model is now $\frac{1}{Z}\exp(\theta_1 \widetilde{M}_1(m)\widetilde{y} + \theta_2 \widetilde{M}_2(m)\widetilde{y})$. We similarly compute

$$Z = 2(\exp(\theta_1 + \theta_2) + \exp(-\theta_1 + \theta_2) + \exp(\theta_1 - \theta_2) + \exp(-\theta_1 - \theta_2)).$$

The observable distribution is now $P(\widetilde{M}_1(m), \widetilde{M}_2(m))$. We have that

$$P(\widetilde{M}_1(m) = 1, \widetilde{M}_2(m) = 1) = \frac{1}{Z}(\exp(\theta_1 + \theta_2) + \exp(-\theta_1 - \theta_2)),$$

and

$$P(\widetilde{M}_1(m) = 1, \widetilde{M}_2(m) = -1) = \frac{1}{Z}(\exp(\theta_1 - \theta_2) + \exp(-\theta_1 + \theta_2)).$$

Note that we have $P(\widetilde{M}_1(m) = -1, \widetilde{M}_2(m) = -1) = P(\widetilde{M}_1(m) = 1, \widetilde{M}_2(m) = 1)$ and $P(\widetilde{M}_1(m) = -1, \widetilde{M}_2(m) = 1) = P(\widetilde{M}_1(m) = 1, \widetilde{M}_2(m) = -1)$.

As a result, we have the same distribution $P(\widetilde{M}_1(m), \widetilde{M}_2(m))$ for the parameters $\theta_1, \theta_2 = a, b$ and for $\theta_1, \theta_2 = b, a$, where a, b are some non-negative values. If $a \neq b$, we end up with at least two solutions that cannot be distinguished, completing the proof. □

D.2 Rademacher complexity bounds

We present bounds on the Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$ of various models $\mathcal{F}$. For all of the $\mathcal{F}$ below, we obtain $\mathfrak{R}_n(\ell \circ \mathcal{F})$ by computing $\mathfrak{R}_n(\mathcal{F})$. These two Rademacher complexities are equal when we assume that ℓ is 1-Lipschitz and apply Talagrand's lemma.

- **Linear models:** we define $f_\theta(x) = \theta^\top x$ with $\|\theta\|_2 \leq B$ and $E[\|x\|_2^2] \leq C^2$, $\mathfrak{R}_n(\mathcal{F}) \leq \frac{BC}{\sqrt{n}}$ [44, Theorem 5.5].
- **Two-layer feed-forward neural networks (MLPs):** we define $f_\theta(x)$ where $\theta = (U, w)$ are the parameters for the weights for the two layers of an MLP. Here $U \in \mathbb{R}^{m \times d}$ and $w \in \mathbb{R}^m$. Suppose ReLU is the activation function, $\|w\|_2 \leq B_w$, $\|u_i\|_2 \leq B_u$ for all $1 \leq i \leq m$, and that $\mathbb{E}[\|x\|_2^2 \leq C^2$. Then, $\mathfrak{R}_n(\mathcal{F}) \leq 2B_w B_u C\sqrt{\frac{m}{n}}$ [44, Theorem 5.9].
- **Kernels:** let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous symmetric function so that for $x_1, \ldots, x_n$, the matrix given by $K_{ij} = k(x_i, x_j)$ is positive semidefinite. The class of kernel estimators consists of functions $f(x) = \sum_{i=1}^n \alpha_i k(X_i, x)$. Suppose that $\sum_{i,j} \alpha_i \alpha_j k(X_i, X_j) \leq B^2$; then, from [45], $\mathfrak{R}_n(\mathcal{F}) \leq 2B\sqrt{\frac{E[k(X,X)]}{n}}$. For particular kernels it is easy to bound the term in the numerator above. For example, we consider the RBF kernel which has maximum one, yielding $\mathfrak{R}_n(\mathcal{F}) \leq \frac{2B}{\sqrt{n}}$.

D.3 Bound on CWoLa's generalization error

We present a result on CWoLa's generalization error showing that for $k = 1$, there exists an irreducible error due to the noise in using the resonant feature as the label.

For CWoLa, we train a classifier on noisy labels. We describe this objective function as $\hat{L}_{\text{noisy}}(f) = \frac{1}{n}\sum_{i=1}^n \ell_C(f(x_i), \hat{y}_i)$, where $\hat{y}_i$ is the resonant feature of the ith sample, $M_1(x_i)$, and has error $p := p(M_1(x) \neq y)$. The CWoLa classifier is equal

JHEP07(2023)188

to $\hat{f} = \text{argmin}_{f\in\mathcal{F}}\hat{L}_{\text{noisy}}(f)$. We distinguish this from the clean objective function, $\hat{L}_{\text{clean}}(f) = \frac{1}{n}\sum_{i=1}^{n}\ell_C(f(x_i), y_i)$ over true labels, and its population-level equivalent, $L_{\text{clean}}(f) = \mathbb{E}\left[\ell_C(f(x), y)\right]$, which is equal to the $L_C(f)$ used in the main text. Let $\bar{y}$ be the flipped binary value of y, so that $\bar{y} = 1$ when $y = 0$ (and vice-versa). Let $\hat{L}_{\text{flipped}}(f) = \frac{1}{n}\sum_{i=1}^{n}\ell_C(f(x_i), \bar{y}_i)$ be the empirical loss of f on flipped labels.

Theorem 3. *For all $f \in \mathcal{F}$, assume that $|\hat{L}_{\text{clean}}(f) - \hat{L}_{\text{flipped}}(f)| \leq \Delta$, the penalty incurred from trying to predict the flipped label $\bar{y}$ rather than the true label y. Then, with probability at least $1-\delta$, the generalization error of* Multi-CWoLa *on $\mathcal{D}$ is at most*

$$L_{\text{clean}}(\hat{f}) - L_{\text{clean}}(f^\star) \leq 2p\Delta + 4\mathfrak{R}_n(\ell\circ\mathcal{F}) + 2\sqrt{\frac{\log 2/\delta}{2n}}.$$

Proof. We can write $\hat{L}_{\text{noisy}}(f)$ as the mixture of $\hat{L}_{\text{clean}}$ and $\hat{L}_{\text{flipped}}$; for all f, we have that

$$\hat{L}_{\text{noisy}}(f) = p\hat{L}_{\text{flipped}}(f) + (1-p)\hat{L}_{\text{clean}}(f), \tag{D.1}$$

so that

$$\hat{L}_{\text{clean}}(f) = \hat{L}_{\text{noisy}}(f) + p(\hat{L}_{\text{clean}}(f) - \hat{L}_{\text{flipped}}(f)) \tag{D.2}$$

We can think of the expression $\hat{L}_{\text{clean}}(f) - \hat{L}_{\text{flipped}}(f)$ as the average penalty from trying to predict the flipped label $\bar{y}$ rather than the true label y. This penalty varies with the function f; let us consider an upper bound so that for all $f \in \mathcal{F}$,

$$|\hat{L}_{\text{clean}}(f) - \hat{L}_{\text{flipped}}(f)| \leq \Delta.$$

Now we can apply this idea to the generalization bound. We aim to bound $L_{\text{clean}}(\hat{f}) - L_{\text{clean}}(f^\star)$. Using the exact same decomposition as in the proof of Theorem 1, we have

$$\begin{aligned} &L_{\text{clean}}(\hat{f}) - L_{\text{clean}}(f^\star) \\ &\quad = (L_{\text{clean}}(\hat{f}) - \hat{L}_{\text{clean}}(\hat{f})) + (\hat{L}_{\text{clean}}(\hat{f}) - \hat{L}_{\text{clean}}(f^\star)) + (\hat{L}_{\text{clean}}(f^\star) - L_{\text{clean}}(f^\star)) \end{aligned}$$

We can apply a standard Rademacher complexity bound to the first and third terms in parentheses as was done in Theorem 1. The middle term $\hat{L}_{\text{clean}}(\hat{f}) - \hat{L}_{\text{clean}}(f^\star)$ can be written as

$$\hat{L}_{\text{clean}}(\hat{f}) - \hat{L}_{\text{clean}}(f^\star) \leq \hat{L}_{\text{noisy}}(f) - \hat{L}_{\text{noisy}}(f^\star) + 2p\Delta \leq 2p\Delta. \tag{D.3}$$

and so our overall generalization bound is

$$L_{\text{clean}}(\hat{f}) - L_{\text{clean}}(f^\star) \leq 2p\Delta + 4\mathfrak{R}_n(\ell\circ\mathcal{F}) + 2\sqrt{\frac{\log(2/\delta)}{2n}} \tag{D.4}$$

Note that unlike in the $k \geq 3$ case, there is no way to reduce the $2p\Delta$ term. □

JHEP07(2023)188

D.4 Asymptotic behavior of SALAD's $\hat{L}_S(h, w)$

Lemma 2. *Assume that the reweighting function is Bayes-optimal, meaning that* $\hat{w}(x, m) = w(x, m)$. *Then,*

$$\lim_{n^{SR}\to\infty} \hat{L}(h, \hat{w}) \propto L_{CE}(h),$$

where $L_{CE}(h) = \mathbb{E}_{x,m,z'=1}\left[-\log h(x, m)\right] + \mathbb{E}_{x,m,z'=0}\left[-\log(1 - h(x, m))\right]$ *is the cross entropy loss on label* $z' = \begin{cases} 1 & x, m \sim \mathcal{P} \\ 0 & x, m \sim p(\cdot|y = 0) \end{cases}$.

Proof. Let n^{SR}_{data} be the number of points from $\mathcal{D}$ that belong to the signal region. Under our assumptions, the empirical loss function can be written as

$$\begin{aligned}\hat{L}(h, \hat{w}) \propto &-\frac{n^{SR}_{\text{data}}}{n^{SR}} \cdot \frac{1}{n^{SR}_{\text{data}}} \sum_{x\in\mathcal{D}_{SR}} \log h(x, m) \\ &- \frac{n^{SR}_{\text{sim}}}{n^{SR}} \cdot \frac{1}{n^{SR}_{\text{sim}}} \sum_{x\in\mathcal{D}^{\text{sim}}_{SR}} \frac{p(x, m|y = 0)}{p_{\text{sim}}(x, m|y = 0)} \log(1 - h(x, m)).\end{aligned}$$

As $n^{SR} \to \infty$, the first term approaches $-\Pr(z' = 1)\cdot\mathbb{E}_{x,m\sim\mathcal{P}}\left[\log h(x, m)\right] = -\Pr(z' = 1)\cdot\mathbb{E}_{x,m|z'=1}\left[\log h(x, m)\right]$. For the second term, we can construct $n^{SR,0}_{\text{data}}$, the amount of data where x is from $p(\cdot|y = 0)$, to be equal to n^{SR}_{sim} such that the expression asymptotically approaches $-\Pr(z' = 0)\cdot\mathbb{E}_{x,m\sim\mathcal{P}_{\text{sim}}}\left[\frac{p(x,m|y=0)}{p_{\text{sim}}(x,m|y=0)}\log(1 - h(x, m))\right]$. Performing a change of expectation, this is equal to $-\Pr(z' = 0)\cdot\mathbb{E}_{x,m|z'=0}\left[\log(1 - h(x, m))\right]$. Putting this together, we have that

$$\begin{aligned}\lim_{n^{SR}\to\infty} \hat{L}(h, \hat{w}) &\propto -\Pr(z' = 1)\mathbb{E}_{x,m|z'=1}\left[\log h(x, m)\right] - \Pr(z' = 0)\mathbb{E}_{x,m|z'=0}\left[\log(1 - h(x, m))\right] \\ &= L_{CE}(h).\end{aligned}$$

□

E Proofs

E.1 Proof of Theorem 1

Proof. From Theorem 3 of [31], we have that $L_C(\hat{f}) - L_C(f^\star)$ is bounded by the traditional ERM generalization gap of $L_C(\bar{f}) - L_C(f^\star)$, where $\bar{f} = \text{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i, m_i), y_i)$ is the classifier learned on labeled data, plus the term $\frac{c_1}{e_{\min}a^5_{\min}}\left(\sqrt{\frac{k}{n}} + \frac{c_2 k}{\sqrt{n}}\right)$.

We can apply standard learning theory bounds on $L_C(\bar{f}) - L_C(f^\star)$. In particular, this quantity is equal to

$$\begin{aligned}L_C(\bar{f}) - L_C(f^\star) &= (L_C(\bar{f}) - \hat{L}_C(\bar{f})) + (\hat{L}_C(\bar{f}) - \hat{L}_C(f^\star)) + (\hat{L}_C(f^\star) - L_C(f^\star)) \\ &\le L_C(\bar{f}) - \hat{L}_C(\bar{f}) + \hat{L}_C(f^\star) - L_C(f^\star) \\ &\le 2\sup_{f\in\mathcal{F}} |L_C(f) - \hat{L}_C(f)|,\end{aligned}$$

JHEP07(2023)188

where we have used the fact that $\hat{L}_C(\bar{f}) \leq \hat{L}_C(f^\star)$. Then, using Theorem 3.3 of [46], we have that with probability $1-\delta$,

$$L_C(\bar{f}) - L_C(f^\star) \leq 2\left(2\mathfrak{R}_n(\ell \circ \mathcal{F}) + \sqrt{\frac{\log 2/\delta}{2n}}\right).$$

Combining these two terms gives us our desired result.

□

E.2 Proof of Theorem 2

Proof. We define the true (cross-entropy) loss as

$$L_S(h,w) = -\Pr(z'=1)\mathbb{E}_{z'=1}\left[\log h(x,m)\right] - \Pr(z'=0)\mathbb{E}_{x,m\in\mathcal{P}^{SR}_{\text{sim}}}\left[w(x,m)\log(1-h(x,m))\right],$$

where $z'=1$ for $x,m\sim\mathcal{P}$ and 0 for $x,m\sim\mathcal{P}(\cdot|y=0)$. Next, define $w(x,m)=\frac{q(x,m|z=1)}{q(x,m|z=0)}$ and let $\hat{w}$ be the weight ratio learned by our model. Let $\hat{h}=\text{argmin}_{h\in\mathcal{H}}\ \hat{L}_S(h,\hat{w})$, and let $h^\star=\text{argmin}_{h\in\mathcal{H}}\ L(h,w^\star)$. Intuitively, $h^\star$ corresponds to the true difference between $\mathcal{P}^{SR}_{\text{data}}$ and $\mathcal{P}^{SR}_{\text{data}}(\cdot|y=0)$. We can first decompose the generalization error as

$$L_S(\hat{h},\hat{w}) - L_S(h^\star,w) = [L_S(\hat{h},\hat{w}) - \hat{L}_S(\hat{h},\hat{w})] + [\hat{L}_S(\hat{h},\hat{w}) - \hat{L}_S(h^\star,\hat{w})] \tag{E.1}$$

$$+ [\hat{L}_S(h^\star,\hat{w}) - \hat{L}_S(h^\star,w)] + [\hat{L}_S(h^\star,w) - L_S(h^\star,w)]. \tag{E.2}$$

We know that $\hat{L}_S(\hat{h},\hat{w}) \leq \hat{L}_S(h^\star,\hat{w})$, so

$$\begin{aligned} L_S(\hat{h},\hat{w}) - L_S(h^\star,w) &\leq |L_S(\hat{h},\hat{w}) - \hat{L}_S(\hat{h},\hat{w})| + |\hat{L}_S(h^\star,w) - L_S(h^\star,w)| \\ &\quad + \hat{L}_S(h^\star,\hat{w}) - \hat{L}_S(h^\star,w) \\ &\leq \sup_{h,w}|L_S(h,w) - \hat{L}_S(h,w)| + |\hat{L}_S(h^\star,w) - L_S(h^\star,w)| \\ &\quad + \hat{L}_S(h^\star,\hat{w}) - \hat{L}_S(h^\star,w). \end{aligned}$$

We first bound $\sup_{h,w}|L_S(h,w) - \hat{L}_S(h,w)|$. For notation, we rewrite $L_S(h,w)$ as $L_S(h,g)$, where $w(x,m) = \frac{g(x,m)}{1-g(x,m)}$ and g belongs to some function class $\mathcal{G}$. Then, using Theorem 3.3 from [46], we get that $\sup_{h,w}|L_S(h,w) - \hat{L}_S(h,w)| \leq 2\mathfrak{R}_{n^{SR}}(\ell_S \circ \{\mathcal{H},\mathcal{G}\}) + \sqrt{\frac{\log 1/\delta}{2n^{SR}}}$ with probability at least $1-\delta$, where $\ell_S \circ \{\mathcal{H},\mathcal{G}\}$ is defined as satisfying $\ell_S(h(x,m),g(x,m),y) = -y\log h(x,m) - (1-y)\frac{g(x,m)}{1-g(x,m)}\log(1-h(x,m))$ for $h\in\mathcal{H}, g\in\mathcal{G}$.

Next, we bound $|\hat{L}_S(h^\star,w) - L_S(h^\star,w)|$. Let $W=\max w(x,m) < \infty$ be the maximum density ratio, and let $B_1 = \max_{x,m}\{-\log h^\star(x,m), -\log(1-h^\star(x,m))\}$. Assume that $B_1 < \infty$. We can apply standard concentration inequalities here (Hoeffding) to get that $|\hat{L}_S(h^\star,w) - L_S(h^\star,w)| \leq WB_1\sqrt{\frac{\log 2/\delta}{2n^{SR}}}$ with probability at least $1-\delta$.

Finally, we bound $\hat{L}_S(h^\star,\hat{w}) - \hat{L}_S(h^\star,w)$. We can write $\hat{L}_S(h^\star,\hat{w}) - \hat{L}_S(h^\star,w)$ as

$$\hat{L}_S(h^\star,\hat{w}) - \hat{L}_S(h^\star,w) = \frac{1}{n^{SR}}\sum_{x\in\mathcal{D}^{\text{sim}}_{SR}}(\hat{w}(x,m) - w(x,m))\cdot(-\log(1-h^\star(x,m))). \tag{E.3}$$

JHEP07(2023)188

Define $\eta = \max(-\log(1 - h^\star(x,m))) \geq 0$ for $x, m \in \mathcal{D}^{\text{sim}}_{SR}$, which is small as long as $h^\star(x,m)$ sufficiently classifies x and is hence a property of how separated the reweighted simulation and true data is. Then,

$$|\hat{L}_S(h^\star, \hat{w}) - \hat{L}_S(h^\star, w)| \leq \frac{\eta}{n^{SR}} \sum_{x,m \in \mathcal{D}^{\text{sim}}_{SR}} |\hat{w}(x,m) - w(x,m)|. \tag{E.4}$$

Recall that $\hat{w}(x,m) = \frac{\hat{g}(x,m)}{1-\hat{g}(x,m)}$ and $w(x,m) = \frac{g^\star(x,m)}{1-g^\star(x,m)}$ where $g^\star(x,m) = \Pr(z = 1|x,m)$, so $|\hat{w}(x,m) - w(x,m)| = \frac{|\hat{g}(x,m)-g^\star(x,m)|}{(1-\hat{g}(x,m))(1-g^\star(x,m))}$. This denominator is greater than $(1-\hat{g}_{\max})(1-g^\star_{\max})$. Then,

$$|\hat{L}_S(h^\star, \hat{w}) - \hat{L}_S(h^\star, w)| \leq \frac{\eta}{(1-\hat{g}_{\max})(1-g^\star_{\max})n^{SR}} \sum_{x,m \in \mathcal{D}^{\text{sim}}_{SR}} |\hat{g}(x,m) - g^\star(x,m)|. \tag{E.5}$$

We now look at the classifier for training g. The per-point cross entropy loss for (x,m,z) is $\ell(g(x,m),z) = -\log g(x,m)$ for $z = 1$ and $-\log(1-g(x,m))$ for $z = 0$. WLOG, assume for some x and m, $g^\star(x,m) > \hat{g}(x,m)$. Then $|\ell(g^\star(x,m),1) - \ell(\hat{g}(x,m),1)| = \log \frac{g^\star(x,m)}{\hat{g}(x,m)} = \log\left(1 + \left(\frac{g^\star(x,m)}{\hat{g}(x,m)} - 1\right)\right) \geq \frac{g^\star(x,m)/\hat{g}(x,m)-1}{g^\star(x,m)/\hat{g}(x,m)} = \frac{g^\star(x,m)-\hat{g}(x,m)}{g^\star(x,m)} \geq |g^\star(x,m) - \hat{g}(x,m)|$ and $|\ell(g^\star(x,m),0) - \ell(\hat{g}(x,m),0)| = \log \frac{1-\hat{g}(x,m)}{1-g^\star(x,m)} = \log\left(1 + \left(\frac{1-\hat{g}(x,m)}{1-g^\star(x,m)} - 1\right)\right) \geq \frac{(1-\hat{g}(x,m))/(1-g^\star(x,m))-1}{(1-\hat{g}(x,m))/(1-g^\star(x,m))} = \frac{g^\star(x,m)-\hat{g}(x,m)}{1-\hat{g}(x,m)} \geq |g^\star(x,m) - \hat{g}(x,m)|$, where we use the inequality $\log(1+x) \geq \frac{x}{1+x}$ for $x > -1$. Therefore, with probability $1-\delta$,

$$\begin{aligned}
|\hat{L}_S(h^\star, \hat{w}) - \hat{L}_S(h^\star, w)| &\leq \frac{\eta}{(1-\hat{g}_{\max})(1-g^\star_{\max})n^{SR}} \sum_{x,m \in SR} |\ell(\hat{g}(x,m),z) - \ell(g^\star(x,m),z)| \\
&\leq \frac{\eta n^{SR}_{\text{sim}}}{(1-\hat{g}_{\max})(1-g^\star_{\max})n^{SR}} \left(\mathbb{E}\left[|\ell(\hat{g}(x,m),z) - \ell(g^\star(x,m),z)|\right] + B_2 \sqrt{\frac{\log 2/\delta}{2n^{SR}_{\text{sim}}}} \right),
\end{aligned}$$

where $B_2 = \max_{x,y}\{\ell(\hat{g}(x,m),z), \ell(g^\star(x,m),z)\} = -\log(\min\{\hat{g}_{\min}, g^\star_{\min}\})$. We assume that B_2 is finite, so there exists a constant c such that

$$|\hat{L}_S(h^\star, \hat{w}) - \hat{L}_S(h^\star, w)| \leq \frac{\eta n^{SR}_{\text{sim}}}{(1-\hat{g}_{\max})(1-g^\star_{\max})n^{SR}} \left(c|L(\hat{g}) - L(g^\star)| + B_2 \sqrt{\frac{\log 2/\delta}{2n^{SR}_{\text{sim}}}} \right),$$

where $L(g) = \mathbb{E}_{x,m \in SR}[\ell(g(x,m),z)]$. Since $g^\star(x,m)$ is Bayes optimal, $|L(\hat{g}) - L(g^\star)| = L(\hat{g}) - L(g^\star) = L(\hat{g}) - \hat{L}(\hat{g}) + \hat{L}(\hat{g}) - \hat{L}(g^\star) + \hat{L}(g^\star) - L(g^\star) \leq 2\sup_{g \in \mathcal{G}} |L(g) - \hat{L}(g)|$. From Theorem 3.3 in [46], this is bounded by $2\mathfrak{R}_{n^{SB}}(\ell \circ \mathcal{G}) + \sqrt{\frac{\log 1/\delta}{2n^{SB}}}$ with probability at least $1-\delta$. Then, applying a union bound, with probability $1-\delta$, we have

$$\begin{aligned}
&|\hat{L}_S(h^\star, \hat{w}) - \hat{L}_S(h^\star, w)| \\
&\qquad \leq \frac{\eta n^{SR}_{\text{sim}}}{(1-\hat{g}_{\max})(1-g^\star_{\max})n^{SR}} \left(4c\mathfrak{R}_{n^{SB}}(\ell \circ \mathcal{G}) + 2c\sqrt{\frac{\log 2/\delta}{2n^{SB}}} + B_2 \sqrt{\frac{\log 4/\delta}{2n^{SR}_{\text{sim}}}} \right).
\end{aligned}$$

JHEP07(2023)188

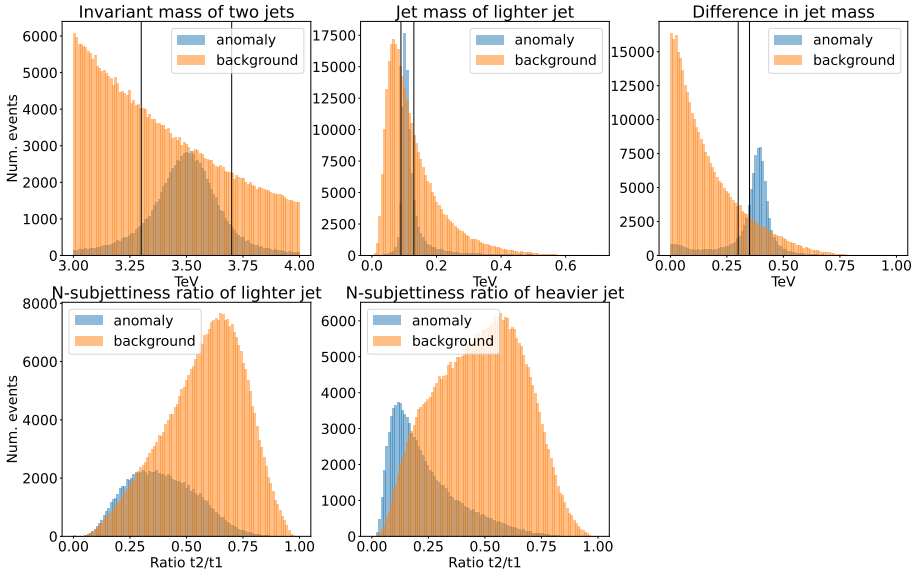

Figure 7. Distribution of each feature (background vs anomaly) for the MULTI-CWOLA experiments. The first three features are used in MULTI-CWOLA as resonant features while the latter two are used as discriminative features. The first feature (invariant mass of the dijet) is used as the resonant feature in CWOLA.

Putting everything together with another union bound, with probability $1-\delta$, the generalization error is at most

$$L_S(\hat{h},\hat{w}) - L_S(h^\star, w) \leq 2\mathfrak{R}_{n^{SR}}(\ell_S \circ \{\mathcal{H},\mathcal{G}\}) + (1+WB_1)\sqrt{\frac{\log 8/\delta}{2n^{SR}}} \tag{E.6}$$

$$+ \frac{\eta n^{SR}_{\text{sim}}}{(1-\hat{g}_{\max})(1-g^\star_{\max})n^{SR}}\left(4c\mathfrak{R}_{n^{SB}}(\ell\circ\mathcal{G}) + 2c\sqrt{\frac{\log 4/\delta}{2n^{SB}}} + B_2\sqrt{\frac{\log 8/\delta}{2n^{SR}_{\text{sim}}}}\right), \tag{E.7}$$

where we combine like terms $\sqrt{\frac{\log(4/\delta)}{2n_{SR}}}$ and $WB_1\sqrt{\frac{\log(8/\delta)}{2n_{SR}}}$ into being upper bounded by $(1+WB_1)\sqrt{\frac{\log(8/\delta)}{2n_{SR}}}$.

□

F Experiment details

F.1 Multi-CWoLa experiments

For the MULTI-CWOLA experiment, we used the anomaly and simulation data from the Pythia 8 simulations in the LHC Olympics Dataset to create an unlabeled dataset we want to perform anomaly detection on [3]. We have $k = 3$, and construct $M_i(m)$ based on the thresholds $[[3.3, 3.7], [0.09, 0.13], [0.3, 0.35]]$ on the first three features. For standard CWOLA, only the first feature is regarded as the resonant feature, and it is thresholded with the interval $[3.3, 3.7]$ (see figure 7). We constructed training datasets of varying sizes

JHEP07(2023)188

Method	SALAD 1	SALAD 2	SALAD-Switch	Multi-SALAD
AUC	87.5 ± 9.7	72.4 ± 18.9	93.8 ± 2.1	$\mathbf{94.6 \pm 0.9}$

Table 3. AUC for imbalanced classifier, averaged over 5 runs.

with class balance $\Pr(y = 1) = 0.149$. We used one test dataset with 65755 randomly sampled anomaly points and 161658 randomly sampled background points.

All methods were trained using scikit-learn's MLPClassifier with max_iter=5000. For Multi-CWoLa's weak supervision step, we learn the parameters of the graphical model using SGD and PyTorch [47] with 30000 steps and learning rate $= 1e-6$. We do not assume the class balance $\Pr(y = 1) = 0.149$ is known, and instead set a prior estimate $\hat{p}(y = 1) = 0.25$ to be used in the algorithm.

F.2 Multi-SALAD experiments

Setup. We use MLPs from Keras [48], each with 3 hidden layers of dimension 32, ReLu activation, and trained with cross-entropy loss and the Adam optimizer. We train for 50 epochs, batch size 200, and default parameters otherwise. Finally, we evaluate our approach on a new test set containing 200000 background points and 200000 anomaly points. This test set is used to produce the signal efficiency to rejection rate. All experiments were run on a personal laptop.

Additional results. In figure 8, we show our results on individual runs. This is because computing the confidence intervals of these curves averaged across the 10 random runs is too noisy due to the magnitude of the reciprocal 1/FPR.

Unbalanced data. When the simulation and true dataset are imbalanced, we adjust our reweighting by $\frac{q(z=1)}{q(z=0)}$ in (C.1), the ratio of real to simulated data. To examine our approach in this setting, we use our synthetic experiment with 1000 points from the true background, 50 points that are anomalies, and 5000 points each for $\mathcal{D}_1^{\text{sim}}$ and $\mathcal{D}_2^{\text{sim}}$. Note that the ratio of anomalies in the true data is the same as our original setting, but we have increased the amount of simulation data by 5 times and adjust weights by $\frac{1}{5}$. Table 3 reports results over 5 random runs.

JHEP07(2023)188

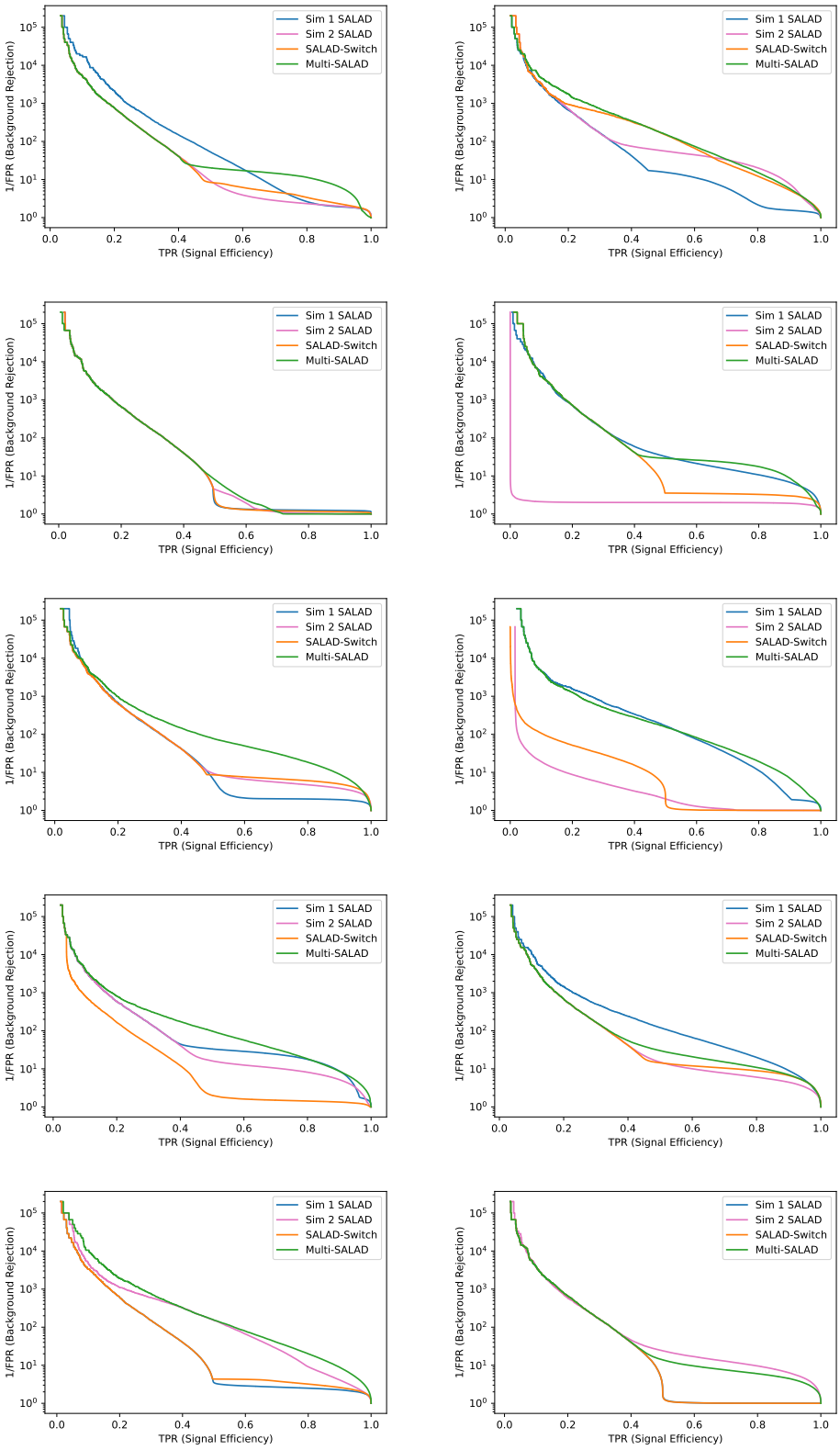

Figure 8. Results on individual runs.

JHEP07(2023)188

References

[1] H.E.P.M.L. community, *A living review of machine learning for particle physics*, https://iml-wg.github.io/HEPML-LivingReview/.

[2] G. Karagiorgi et al., *Machine learning in the search for new fundamental physics*, arXiv:2112.03769 [INSPIRE].

[3] G. Kasieczka et al., *The LHC olympics* 2020 *a community challenge for anomaly detection in high energy physics*, *Rept. Prog. Phys.* **84** (2021) 124201 [arXiv:2101.08320] [INSPIRE].

[4] T. Aarrestad et al., *The dark machines anomaly score challenge: benchmark data and model independent event classification for the Large Hadron Collider*, *SciPost Phys.* **12** (2022) 043 [arXiv:2105.14027] [INSPIRE].

[5] ATLAS collaboration, *Dijet resonance search with weak supervision using* $\sqrt{s} = 13$ *TeV pp collisions in the ATLAS detector*, *Phys. Rev. Lett.* **125** (2020) 131801 [arXiv:2005.02983] [INSPIRE].

[6] ATLAS Collaboration collaboration, *Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states using* $\sqrt{s} = 13$ *TeV pp collisions with the ATLAS detector*, ATLAS-CONF-2022-045, CERN, Geneva, Switzerland (2022).

[7] G. Kasieczka et al., *Anomaly detection under coordinate transformations*, *Phys. Rev. D* **107** (2023) 015009 [arXiv:2209.06225] [INSPIRE].

[8] G. Kasieczka, B. Nachman and D. Shih, *New methods and datasets for group anomaly detection from fundamental physics*, in the proceedings of the *Conference on knowledge discovery and data mining*, (2021) [arXiv:2107.02821] [INSPIRE].

[9] J.H. Collins, K. Howe and B. Nachman, *Anomaly detection for resonant new physics with machine learning*, *Phys. Rev. Lett.* **121** (2018) 241803 [arXiv:1805.02664] [INSPIRE].

[10] J.H. Collins, K. Howe and B. Nachman, *Extending the search for new resonances with machine learning*, *Phys. Rev. D* **99** (2019) 014038 [arXiv:1902.02634] [INSPIRE].

[11] R.T. D'Agnolo and A. Wulzer, *Learning new physics from a machine*, *Phys. Rev. D* **99** (2019) 015014 [arXiv:1806.02350] [INSPIRE].

[12] R.T. D'Agnolo et al., *Learning multivariate new physics*, *Eur. Phys. J. C* **81** (2021) 89 [arXiv:1912.12155] [INSPIRE].

[13] K. Benkendorfer, L.L. Pottier and B. Nachman, *Simulation-assisted decorrelation for resonant anomaly detection*, *Phys. Rev. D* **104** (2021) 035003 [arXiv:2009.02205] [INSPIRE].

[14] A. Andreassen, B. Nachman and D. Shih, *Simulation assisted likelihood-free anomaly detection*, *Phys. Rev. D* **101** (2020) 095004 [arXiv:2001.05001] [INSPIRE].

[15] B. Nachman and D. Shih, *Anomaly detection with density estimation*, *Phys. Rev. D* **101** (2020) 075042 [arXiv:2001.04990] [INSPIRE].

[16] O. Amram and C.M. Suarez, *Tag N' Train: a technique to train improved classifiers on unlabeled data*, *JHEP* **01** (2021) 153 [arXiv:2002.12376] [INSPIRE].

JHEP07(2023)188

[17] A. Hallin et al., *Classifying anomalies through outer density estimation*, *Phys. Rev. D* **106** (2022) 055006 [arXiv:2109.00546] [INSPIRE].

[18] J.A. Raine, S. Klein, D. Sengupta and T. Golling, *CURTAINs for your sliding window: constructing unobserved regions by transforming adjacent intervals*, *Front. Big Data* **6** (2023) 899345 [arXiv:2203.09470] [INSPIRE].

[19] R.T. d'Agnolo et al., *Learning new physics from an imperfect machine*, *Eur. Phys. J. C* **82** (2022) 275 [arXiv:2111.13633] [INSPIRE].

[20] P. Chakravarti, M. Kuusela, J. Lei and L. Wasserman, *Model-independent detection of new physics signals using interpretable semi-supervised classifier tests*, arXiv:2102.07679 [INSPIRE].

[21] B.M. Dillon, R. Mastandrea and B. Nachman, *Self-supervised anomaly detection for new physics*, *Phys. Rev. D* **106** (2022) 056005 [arXiv:2205.10380] [INSPIRE].

[22] M. Letizia et al., *Learning new physics efficiently with nonparametric methods*, *Eur. Phys. J. C* **82** (2022) 879 [arXiv:2204.02317] [INSPIRE].

[23] K. Krzyżańska and B. Nachman, *Simulation-based anomaly detection for multileptons at the LHC*, *JHEP* **01** (2023) 061 [arXiv:2203.09601] [INSPIRE].

[24] S. Alvi, C.W. Bauer and B. Nachman, *Quantum anomaly detection for collider physics*, *JHEP* **02** (2023) 220 [arXiv:2206.08391] [INSPIRE].

[25] T. Sjostrand, S. Mrenna and P.Z. Skands, *A brief introduction to PYTHIA* 8.1, *Comput. Phys. Commun.* **178** (2008) 852 [arXiv:0710.3820] [INSPIRE].

[26] J. Bellm et al., *Herwig* 7.0/*Herwig++* 3.0 *release note*, *Eur. Phys. J. C* **76** (2016) 196 [arXiv:1512.01178] [INSPIRE].

[27] SHERPA collaboration, *Event generation with Sherpa* 2.2, *SciPost Phys.* **7** (2019) 034 [arXiv:1905.09127] [INSPIRE].

[28] E.M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: learning from mixed samples in high energy physics*, *JHEP* **10** (2017) 174 [arXiv:1708.02949] [INSPIRE].

[29] J. Neyman and E.S. Pearson, *On the problem of the most efficient tests of statistical hypotheses*, *Phil. Trans. Roy. Soc. Lond. A* **231** (1933) 289 [INSPIRE].

[30] A. Ratner et al., *Snorkel: rapid training data creation with weak supervision*, in the *Proceedings of the of the* 44$^{\text{th}}$ *international conference on Very Large Data Bases (VLDB)*, Rio de Janeiro, Brazil (2018).

[31] D.Y. Fu et al., *Fast and three-rious: speeding up weak supervision with triplet methods*, in *International conference on machine learning*, (2020) [arXiv:2002.11955].

[32] A. Ratner et al., *Data programming: creating large training sets, quickly*, in the *Proceedings of the of the* 30$^{\text{th}}$ *International Conference on Neural Information Processing Systems*, Red Hook, NY, U.S.A. (2016), p. 3574.

[33] A. Ratner et al., *Training complex models with multi-task weak supervision*, in the *Proceedings of the of the AAAI Conference on Artificial Intelligence*, (2019).

[34] P. Varma et al., *Learning dependency structures for weak supervision models*, in the *Proceedings of the of the* 36$^{\text{th}}$ *International Conference on Machine Learning*, (2019).

[35] M.J. Wainwright and M.I. Jordan, *Graphical models, exponential families, and variational inference*, *Foundations and Trends® in Machine Learning* **1** (2007) 1.

JHEP07(2023)188

[36] J. Thaler and K. Van Tilburg, *Identifying boosted objects with N-subjettiness*, *JHEP* **03** (2011) 015 [arXiv:1011.2268] [INSPIRE].

[37] J. Thaler and K. Van Tilburg, *Maximizing boosted top identification by minimizing N-subjettiness*, *JHEP* **02** (2012) 093 [arXiv:1108.2701] [INSPIRE].

[38] B. Nachman and J. Thaler, *Learning from many collider events at once*, *Phys. Rev. D* **103** (2021) 116013 [arXiv:2101.07263] [INSPIRE].

[39] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning*, Springer, New York, NY, U.S.A. (2009) [DOI:10.1007/978-0-387-84858-7].

[40] M. Sugiyama, T. Suzuki and T. Kanamori, *Density ratio estimation in machine learning*, Cambridge University Press, Cambridge, U.K. (2012) [DOI:10.1017/cbo9781139035613].

[41] K. Cranmer, J. Pavez and G. Louppe, *Approximating likelihood ratios with calibrated discriminative classifiers*, arXiv:1506.02169 [INSPIRE].

[42] S. Dasgupta and P.M. Long, *Boosting with diverse base classifiers*, in the *Proceedings of the Learning Theory and Kernel Machines*, Berlin, Heidelberg, Germany (2003), p. 273.

[43] C. Cortes, Y. Mansour and M. Mohri, *Learning bounds for importance weighting*, in the *Proceedings of the Advances in Neural Information Processing Systems*, Curran Associates Inc. (2010).

[44] T. Ma, *Lecture notes for machine learning theory (CS229M/STATS214)*, https://web.stanford.edu/class/stats214/ June 2022

[45] P.L. Bartlett and S. Mendelson, *Rademacher and Gaussian complexities: risk bounds and structural results*, *J. Mach. Learn. Res.* **3** (2002) 463.

[46] M. Mohri, A. Rostamizadeh and A. Talwalkar, *Foundations of machine learning*, MIT Press, Cambridge, MA, U.S.A. (2018).

[47] A. Paszke et al., *PyTorch: an imperative style, high-performance deep learning library*, in *Advances in neural information processing systems* 32, Curran Associates Inc. (2019), p. 8024.

[48] F. Chollet et al., *Keras*, https://github.com/fchollet/keras.

JHEP07(2023)188