Mitigating Classifier Bias

- It is often desired to control the dependence of a classification algorithm on some protected feature of the data samples.
- In particle colliders, searches for resonant new physics require classifiers that do not sculpt a peak in an otherwise smooth background spectrum.
- Standard techniques aim to produce classifiers that are independent from the protected feature. This is sufficient to avoid localized structures, but is not necessary.

When a classifier turns cats into dogs (or quarks into W bosons)

- Jets produced from boosted W boson decays are interesting in many extensions of the Standard Model.
- Various features of the substructure of these jets can be used to distinguish the boosted bosons (signal) from generic quark and gluon jets (background.)
- A bump hunt (search for excess signal in the mass spectrum of a sample of events) is performed on the mass (protected attribute) of the candidate.
- The challenge with substructure classifiers is that they can introduce artificial bumps into the mass spectrum because substructure is correlated with mass (see central figures.)

MODe is a new, fast and simple regularization technique which produces unbiased classifiers with better classification power than state-of-the-art methods.

We regularize the standard classification loss by adding the following term

$$L_{\text{MODe}} = \sum_m \int |F_m(s) - \sum_{l=0}^{\ell} c_l(s) P_l(m)|^2 \, ds,$$

where $s$ is the classifier output. We decompose the empirical CDF of $s$ conditioned on the protected attribute $m$, $F_m(s)$, into its first $\ell$ Legendre modes. Higher order moments are penalized, resulting in a classifier with a simpler dependence on $m$.

To assess classification performance we use the background rejection factor $R_{50}$: the inverse false positive rate at 50% true positive rate.

To quantify peak sculpting we use the signal bias induced by the classifier selection. Values less than unity are consistent with no bias (since the true signal rate is zero), while values significantly larger indicate substantial bias that could result in false claims of observations.

The figure above shows that the flexibility to go beyond decorrelation provided by MODe[1,2] results in unbiased signal estimators at larger background-rejection power. This would directly translate to improved sensitivity in a real-world analysis.