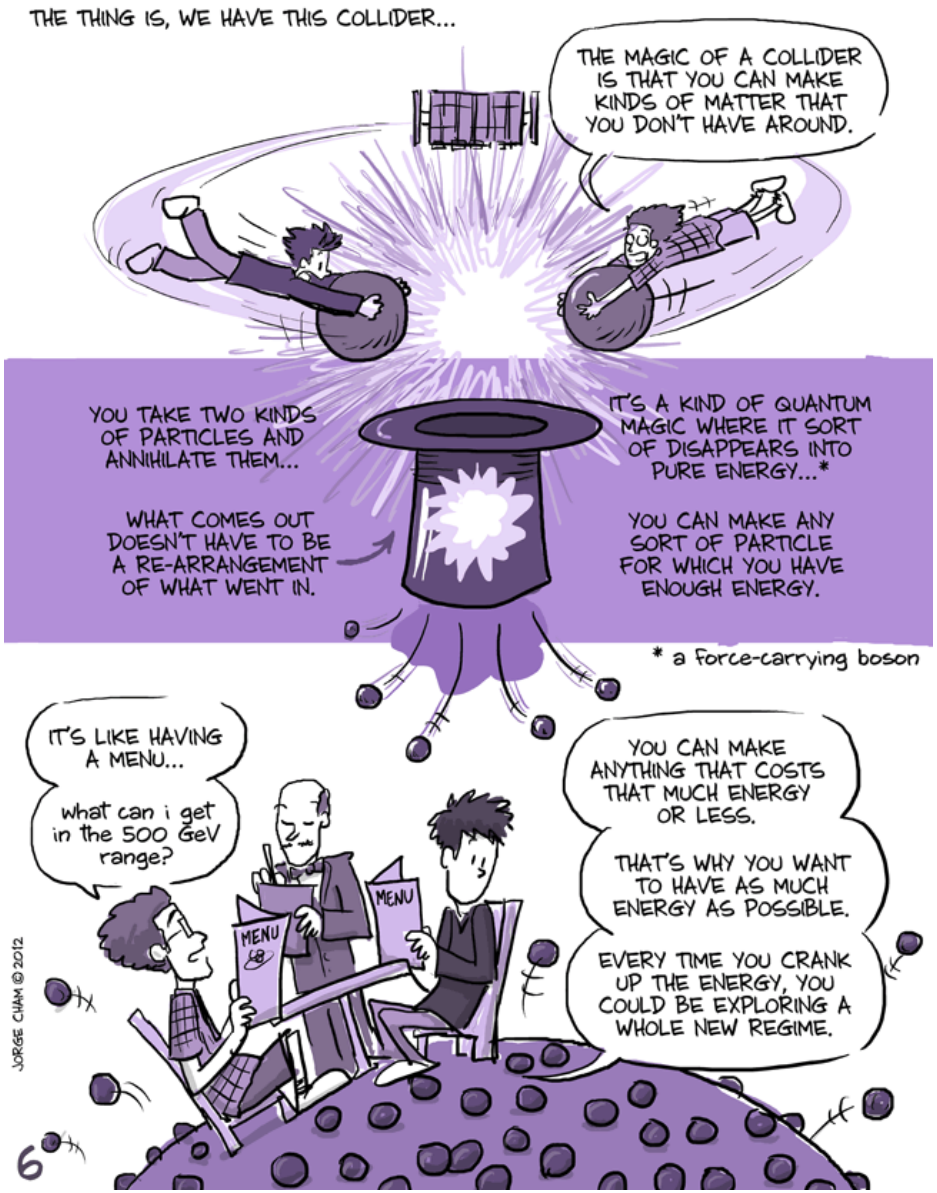
Learning to Pivot with Adversarial Networks

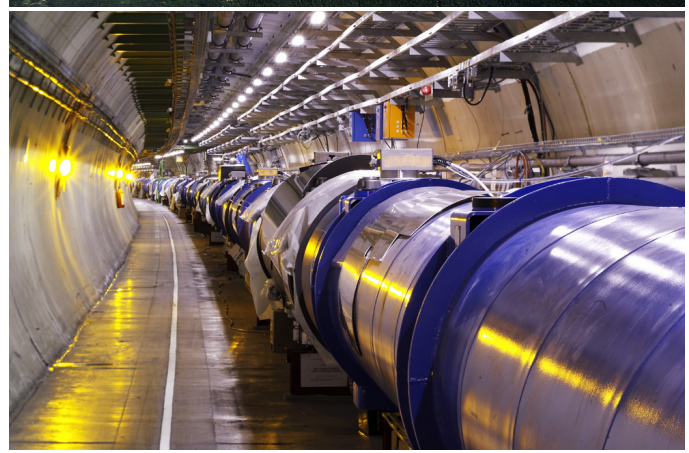
arXiv:1611.01046

Gilles Louppe, Michael Kagan, Kyle Cranmer

Credits: Jorge Cham

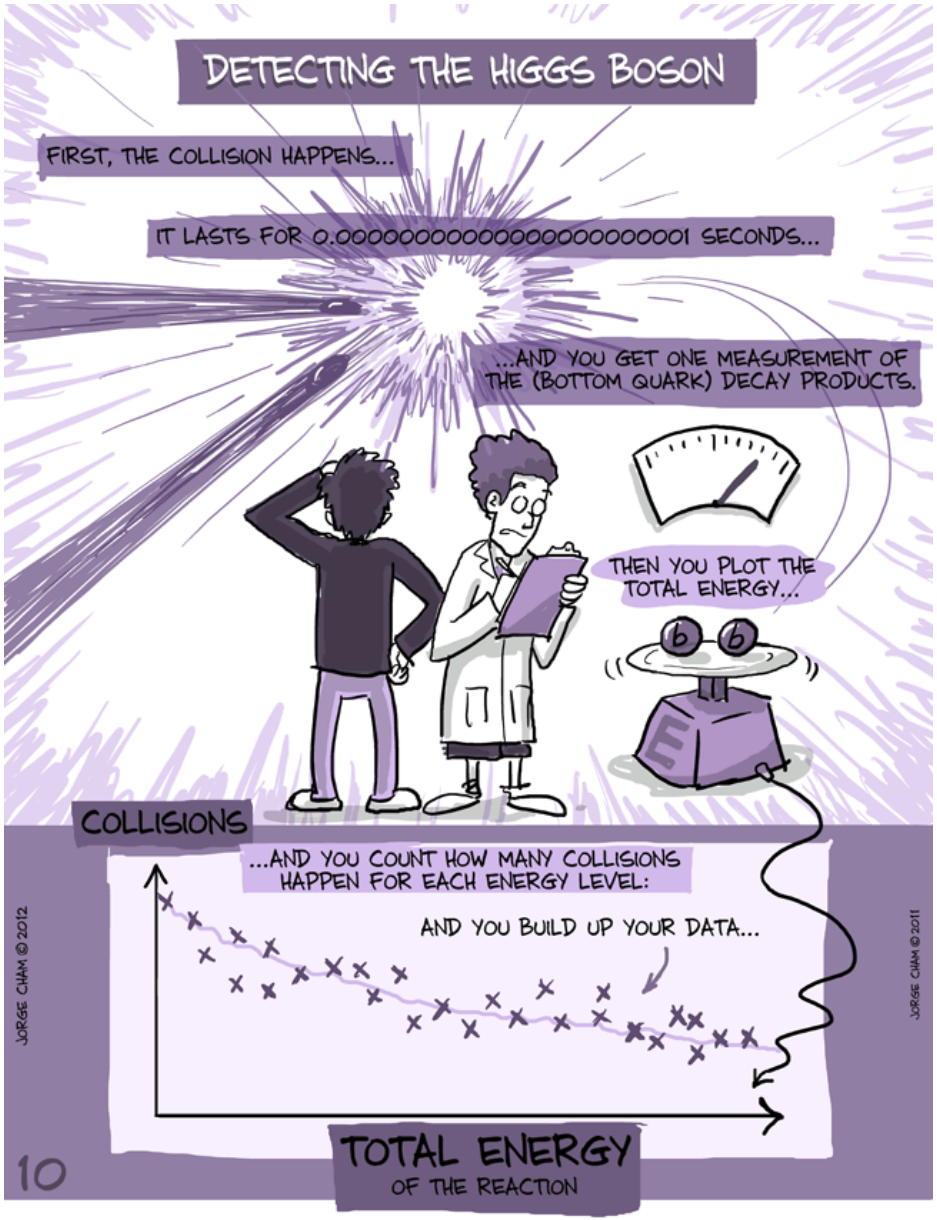
Testing for new physics

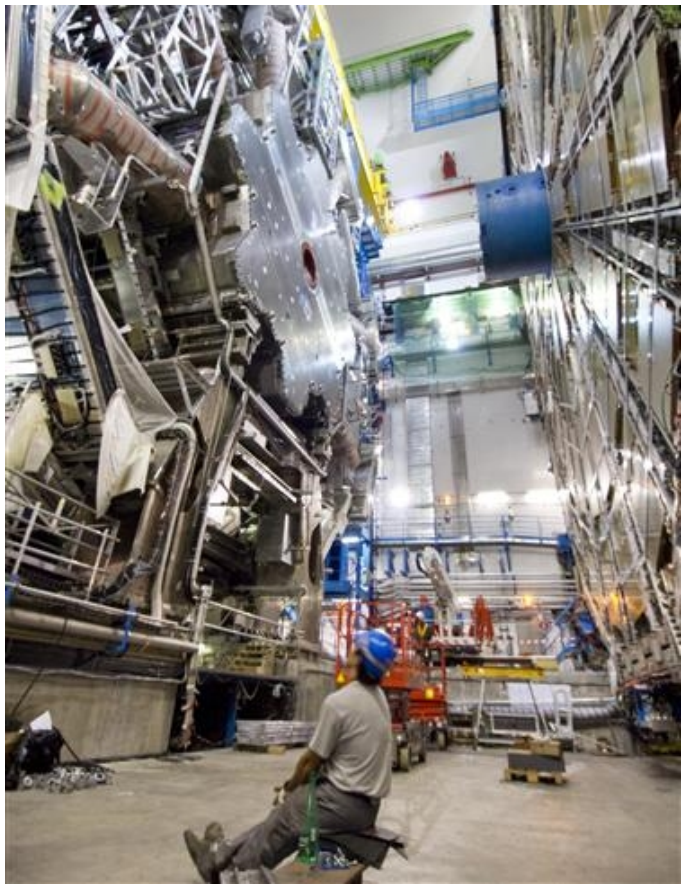




Credits: Jorge Cham

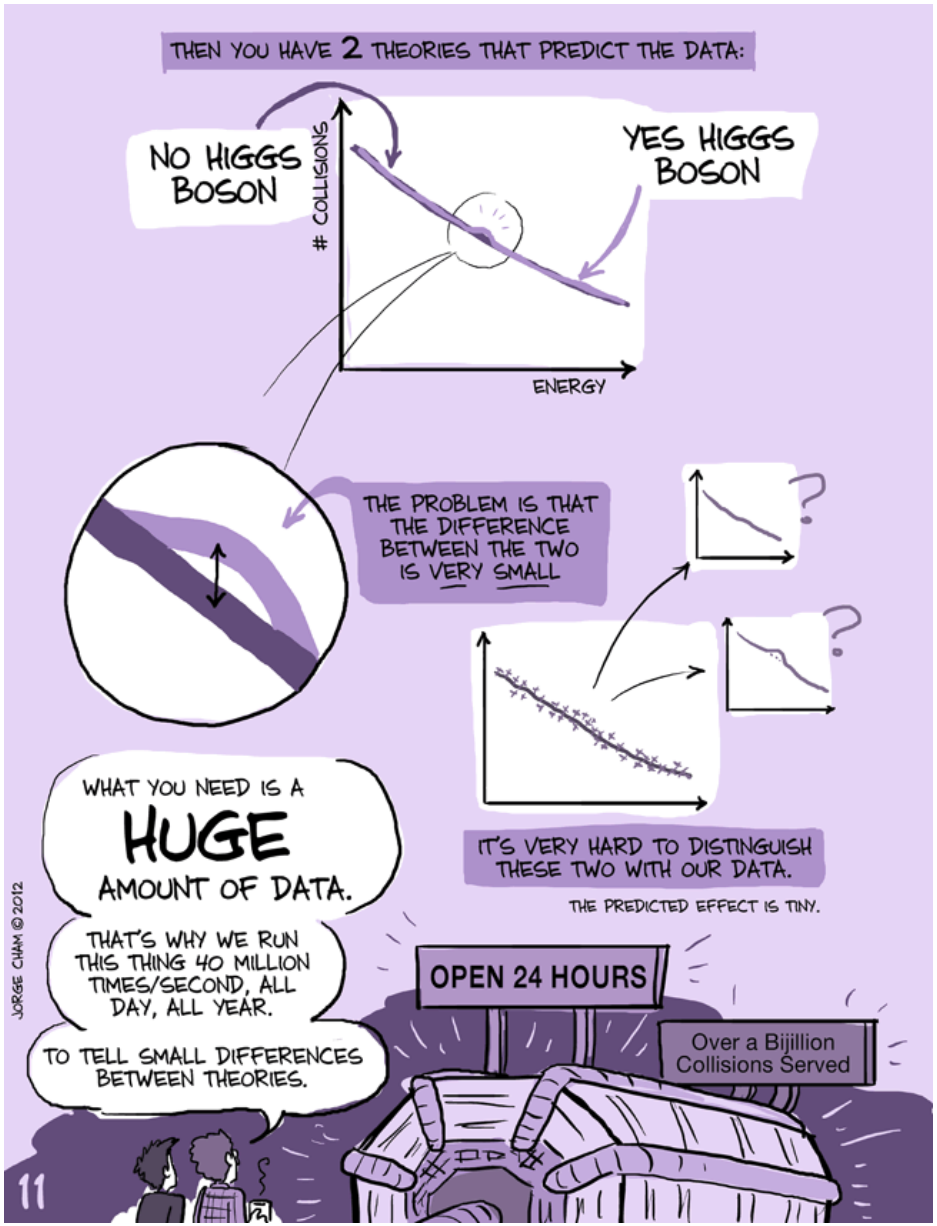
Testing for new physics

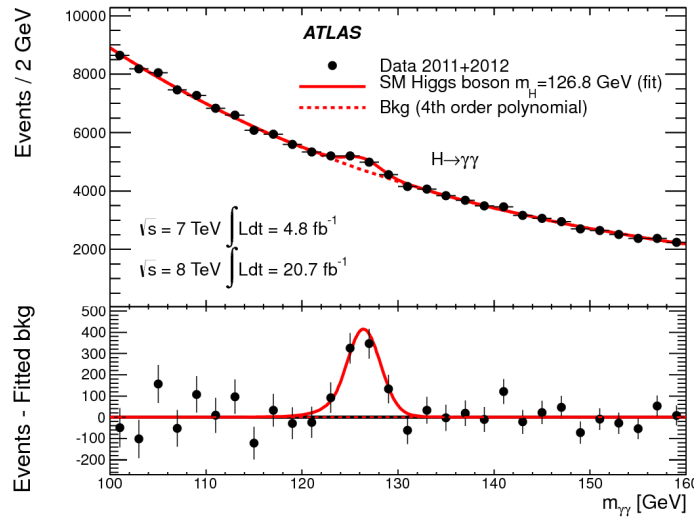




Credits: Jorge Cham

Testing for new physics



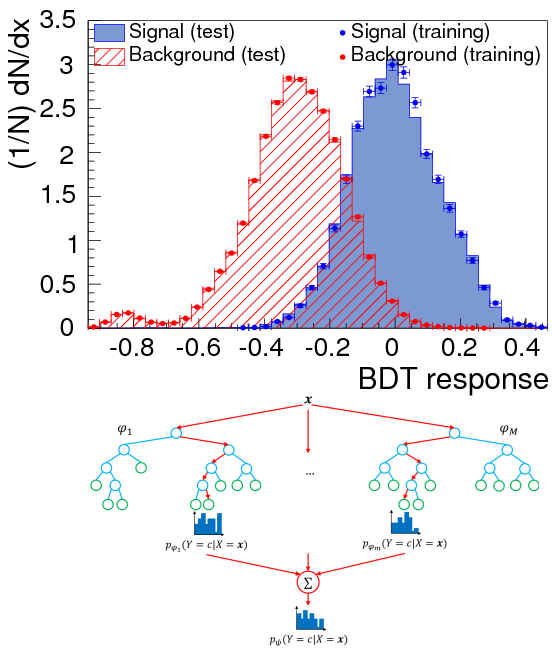


$$\frac{p(\text{data}|\text{background} + \text{signal})}{p(\text{data}|\text{background})}$$

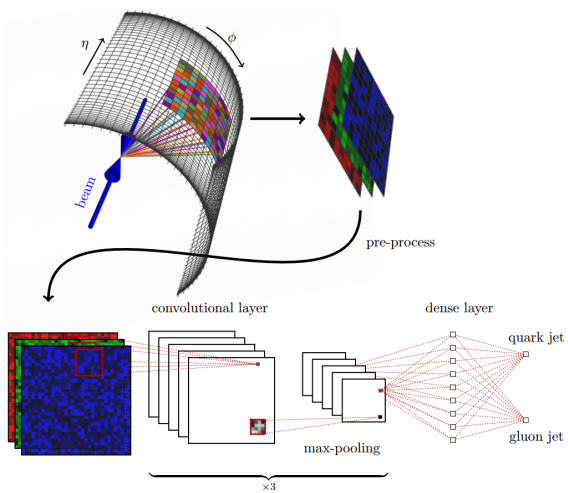
Supervised learning

$$\frac{p(\text{data}|\text{background} + \text{signal})}{p(\text{data}|\text{background})} \Leftrightarrow \text{Classifying background vs. signal}$$

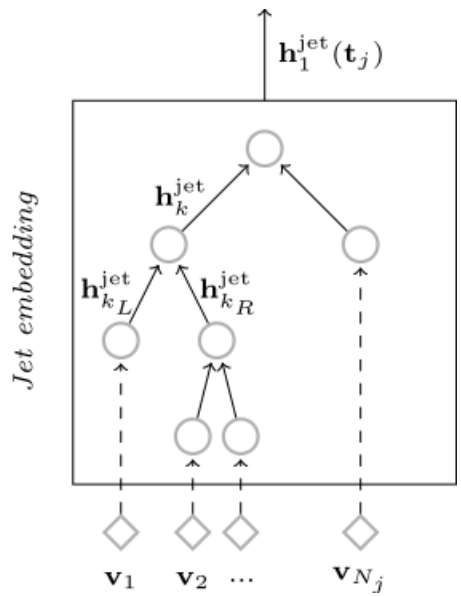
Boosted decision trees



Conv. nets



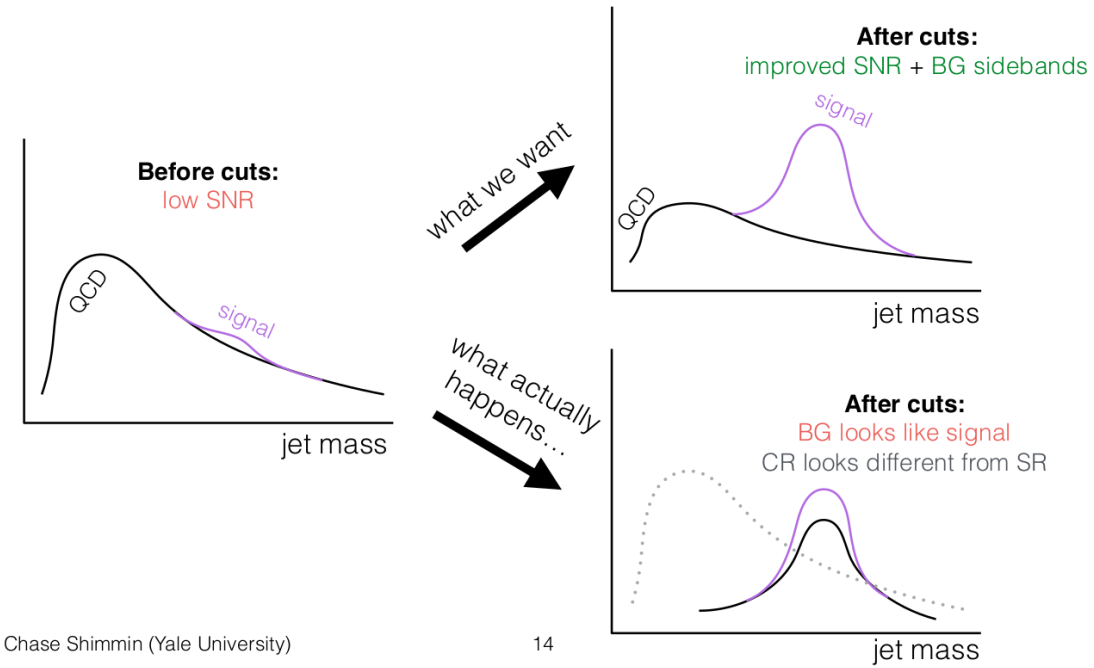
Recursive nets

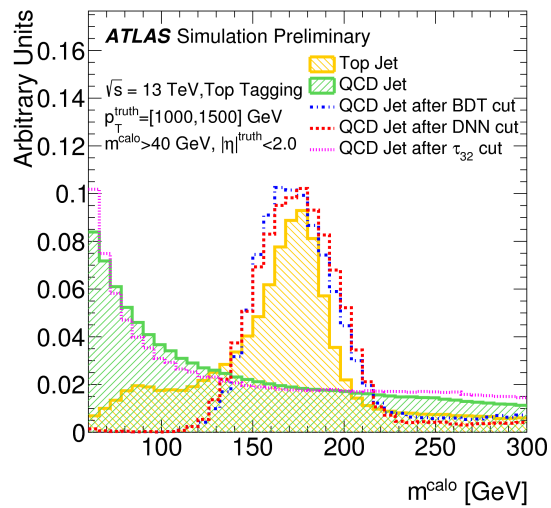


Credits: 1703.03507, ATL-PHYS-PUB-2017-004

Independence from physics variates

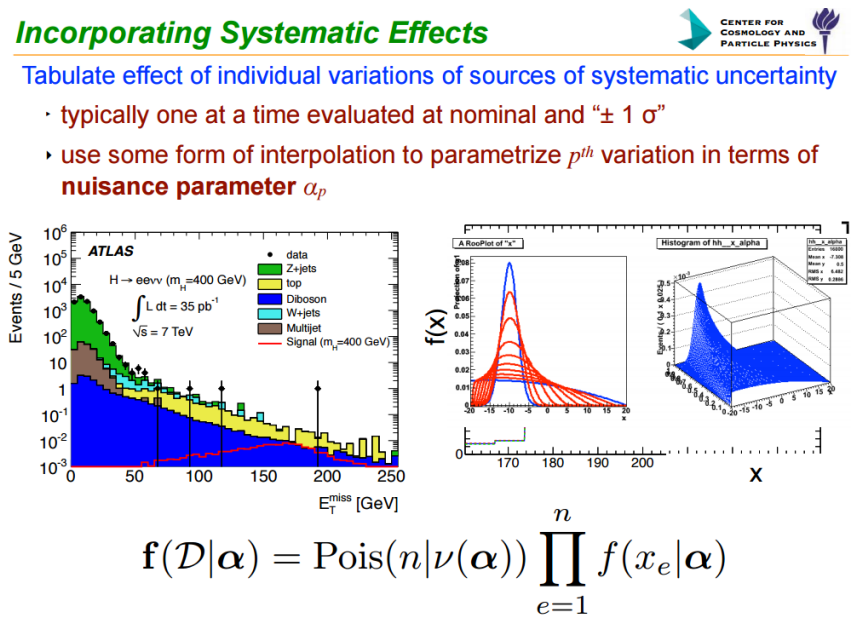
Analysis often rely on the assumption that the classifier is independent from some physics variates (e.g., mass).
Correlation with these variates leads to systematic uncertainties that cannot easily be characterized and controlled.





Independence from known unknowns

- The data generation process is often not uniquely specified or known exactly, hence the presence of systematic uncertainties.
- Data generation processes are formulated as a family of data generation processes parametrized by nuisance parameters.
- Ideally, we would like a classifier that is robust to nuisance parameters.



Problem statement

- Let us assume a family of data generation processes $p(X, Y, Z)$ where
  - X are the data (taking values $x \in \mathcal{X}$),
  - Y are the target labels (taking values $y \in \mathcal{Y}$),
  - Z is an auxiliary random variable (taking values $z \in \mathcal{Z}$).
    - Z corresponds to physics variates or nuisance parameters.

- We want to learn a regression function $f(\cdot; \theta_f) : \mathcal{X} \mapsto \mathcal{Y}$.

- We want inference based on $f(X; \theta_f)$ to be robust to the value $z \in \mathcal{Z}$.
  - E.g., we want a classifier that does not change with systematic variations, even though the data might.

Pivot

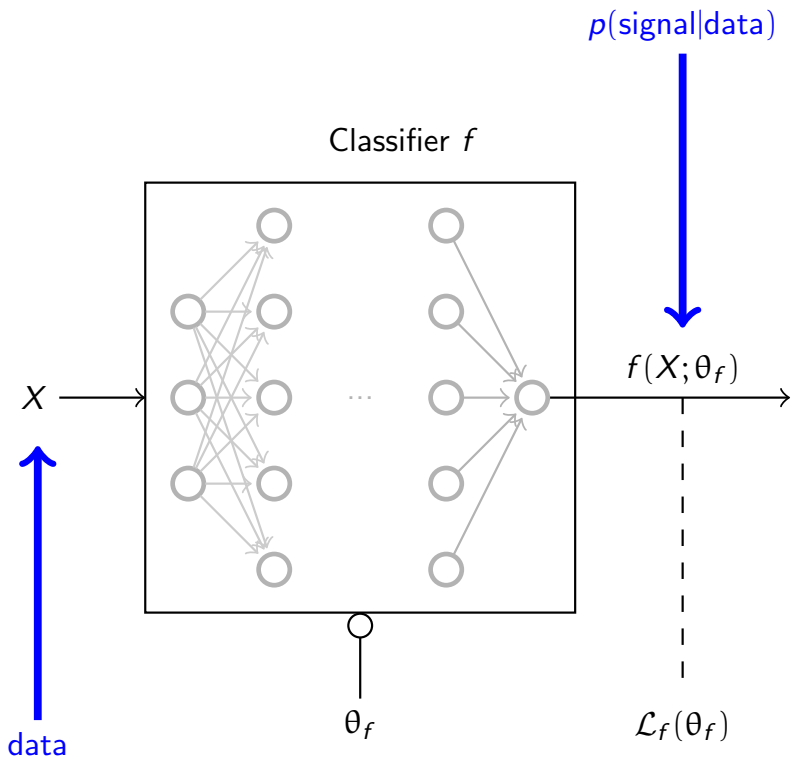
- We define robustness as requiring the distribution of $f(X;\theta_f)$ conditional on Z to be invariant with Z. That is, such that

$$p(f(X;\theta_f) = s|z) = p(f(X;\theta_f) = s|z')$$

for all $z, z' \in \mathcal{Z}$ and all values $s \in \mathcal{S}$ of $f(X;\theta_f)$.
If f satisfies this criterion, then f is known as a pivotal quantity.

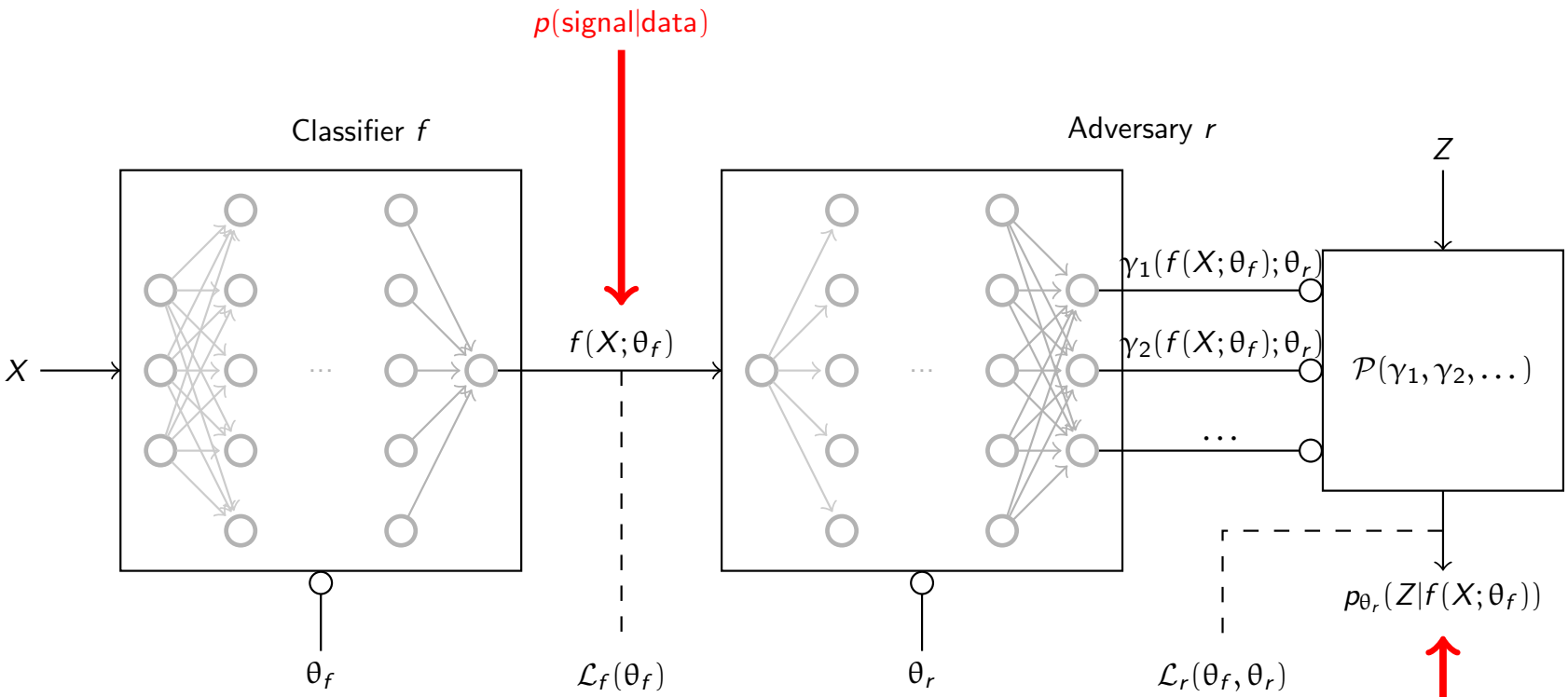
- Alternatively, this amounts to find f such that $f(X;\theta_f)$ and Z are independent random variables.

Adversarial Networks



Let consider a classifier f built as usual, minimizing the cross-entropy $\mathcal{L}_f(\theta_f) = \mathbb{E}_{x\sim X}\mathbb{E}_{y\sim Y|x}[-\log p_{\theta_f}(y|x)]$.

Adversarial Networks

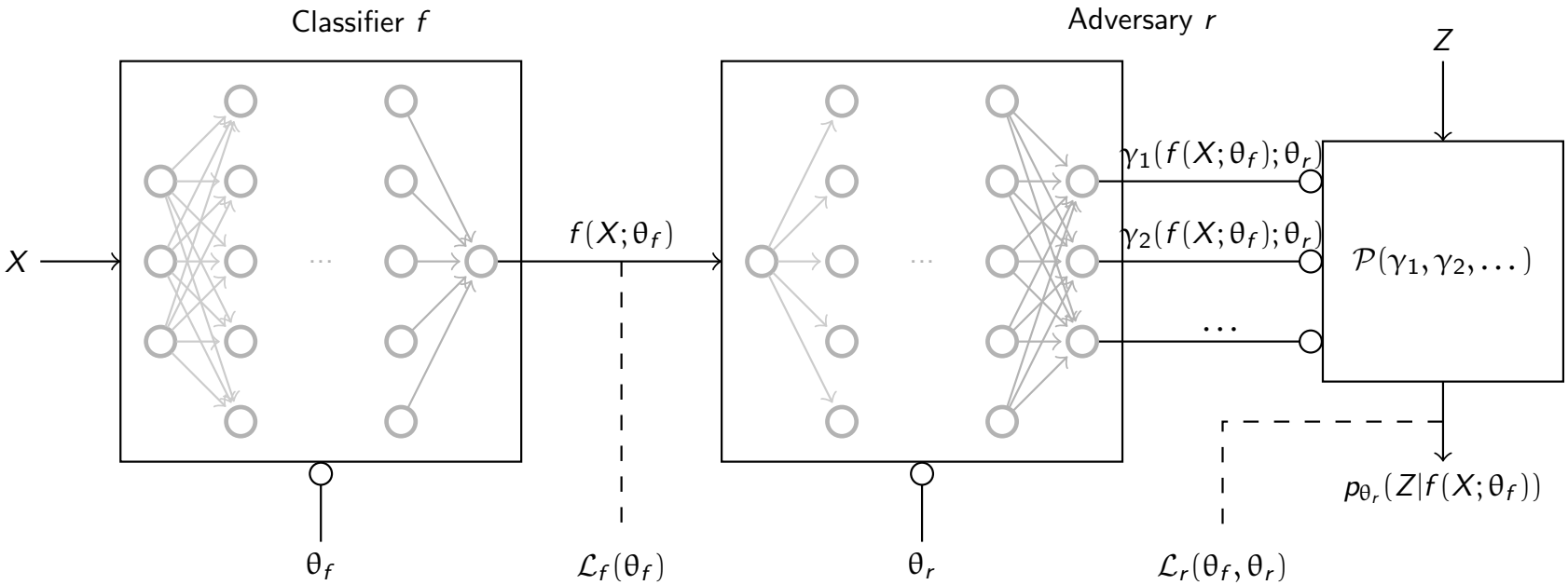


We pit f against an adversary network r producing as output the posterior $p_{\theta_r}(z|f(X;\theta_f) = s)$.

We set r to minimize the cross entropy $\mathcal{L}_r(\theta_f, \theta_r) = \mathbb{E}_{s\sim f(X;\theta_f)}\mathbb{E}_{z\sim Z|s}[-\log p_{\theta_r}(z|s)]$.

Regression of Z from f's output

Adversarial Networks



Goal is to solve: $\hat{\theta}_f, \hat{\theta}_r = \arg\min_{\theta_f} \max_{\theta_r} \mathcal{L}_f(\theta_f) - \mathcal{L}_r(\theta_f, \theta_r)$

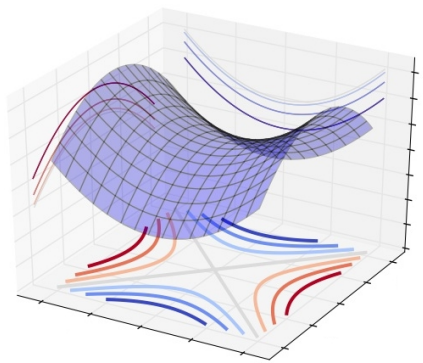
Intuitively, r penalizes f for outputs that can be used to infer Z.

Theoretical motivation

Proposition. *If there exists a minimax solution* $(\hat{\theta}_f, \hat{\theta}_r)$ *such that* $\mathcal{L}_f(\theta_f) - \mathcal{L}_r(\theta_f, \theta_r) = H(Y|X) - H(Z)$*, then* $f(\cdot; \hat{\theta}_f)$ *is both an optimal classifier and a pivotal quantity.*

Proof (sketch):

$$
\begin{aligned}
&\min_{\theta_f} \max_{\theta_r} \mathcal{L}_f(\theta_f) - \mathcal{L}_r(\theta_f, \theta_r) \\
=&\min_{\theta_f} \mathcal{L}_f(\theta_f) - \mathbb{E}_{s \sim f(X;\theta_f)}[H(Z|f(X;\theta_f) = s)] \\
=&\min_{\theta_f} \mathcal{L}_f(\theta_f) - H(Z|f(X;\theta_f)) \\
\geq& H(Y|X) - H(Z)
\end{aligned}
$$



where the equality holds when

- f is an optimal classifier (in which case $\mathcal{L}_f(\theta_f) = H(Y|X)$);
- $f(X;\theta_f)$ and Z are independent random variables (in which case $H(Z|f(X;\theta_f)) = H(Z)$).

Alternating stochastic gradient descent

1: **for** $t = 1$ to T **do**
2: **for** $k = 1$ to K **do** ▷ Update r
3: Sample minibatch $\{x_m, z_m, s_m = f(x_m; \theta_f)\}_{m=1}^{M}$ of size M;
4: With θ_f fixed, update r by ascending its stochastic gradient $\nabla_{\theta_r} E(\theta_f, \theta_r) :=$

$$\nabla_{\theta_r} \sum_{m=1}^{M} \log p_{\theta_r}(z_m|s_m);$$

5: **end for**
6: Sample minibatch $\{x_m, y_m, z_m, s_m = f(x_m; \theta_f)\}_{m=1}^{M}$ of size M; ▷ Update f
7: With θ_r fixed, update f by descending its stochastic gradient $\nabla_{\theta_f} E(\theta_f, \theta_r) :=$

$$\nabla_{\theta_f} \sum_{m=1}^{M} \left[-\log p_{\theta_f}(y_m|x_m) + \log p_{\theta_r}(z_m|s_m) \right],$$

where $p_{\theta_f}(y_m|x_m)$ denotes $\mathbf{1}(y_m = 0)(1 - s_m) + \mathbf{1}(y_m = 1)s_m$;
8: **end for**

Accuracy versus robustness trade-off

- The assumption of existence of a classifier that is both optimal and pivotal may not hold.

- However, the minimax objective can be rewritten as

$$E_\lambda(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) - \lambda \mathcal{L}_r(\theta_f, \theta_r)$$

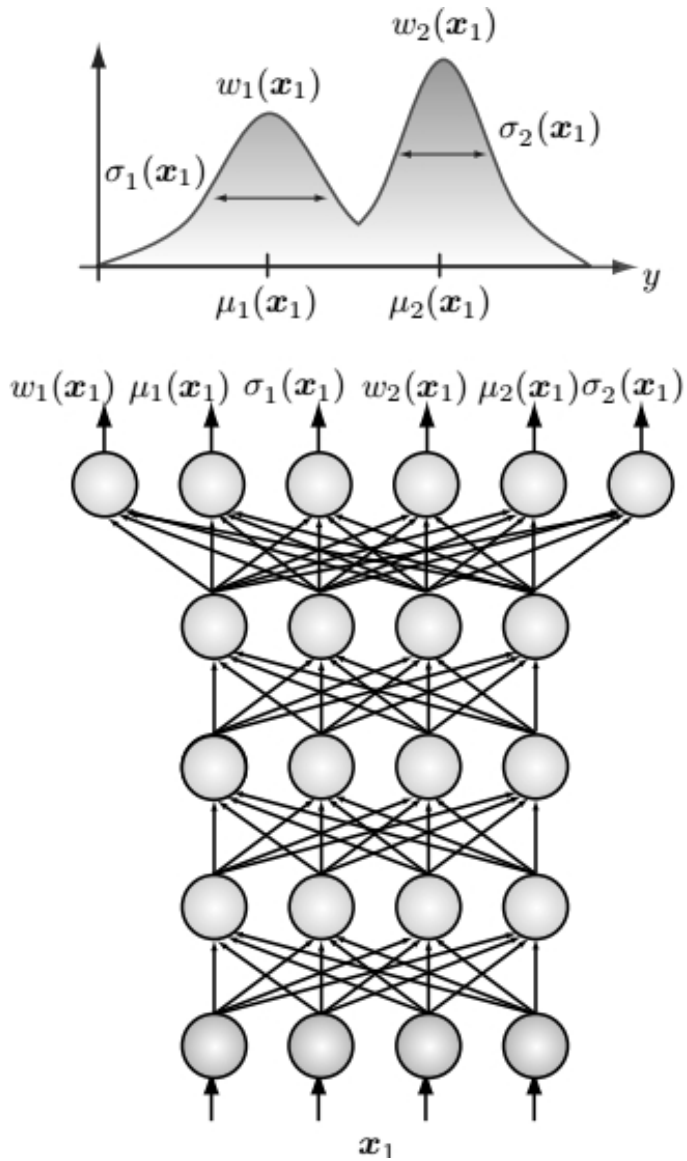
  where λ is a hyper-parameter controlling the trade-off between the performance of f and its independence with respect to the nuisance parameter.

  - Setting λ to a large value enforces f to be pivotal.
  - Setting λ close to 0 constraints f to be optimal.

- Tuning λ is guided by a higher-level objective (e.g., statistical significance).

Architecture for the adversary

- If Z is categorical, then the posterior can be modeled with a standard classifier (e.g., a NN with a softmax output layer).
- If Z is continuous, then the posterior can be modeled with a *mixture density network*.
- No constraint on the prior $p(Z)$.



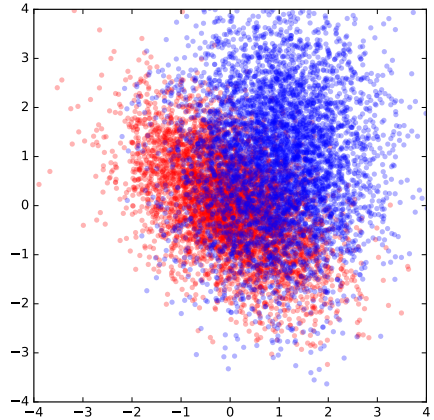
Mixture density network

Toy example

- Binary classification of 2D data drawn from multivariate gaussians with equal priors, such that

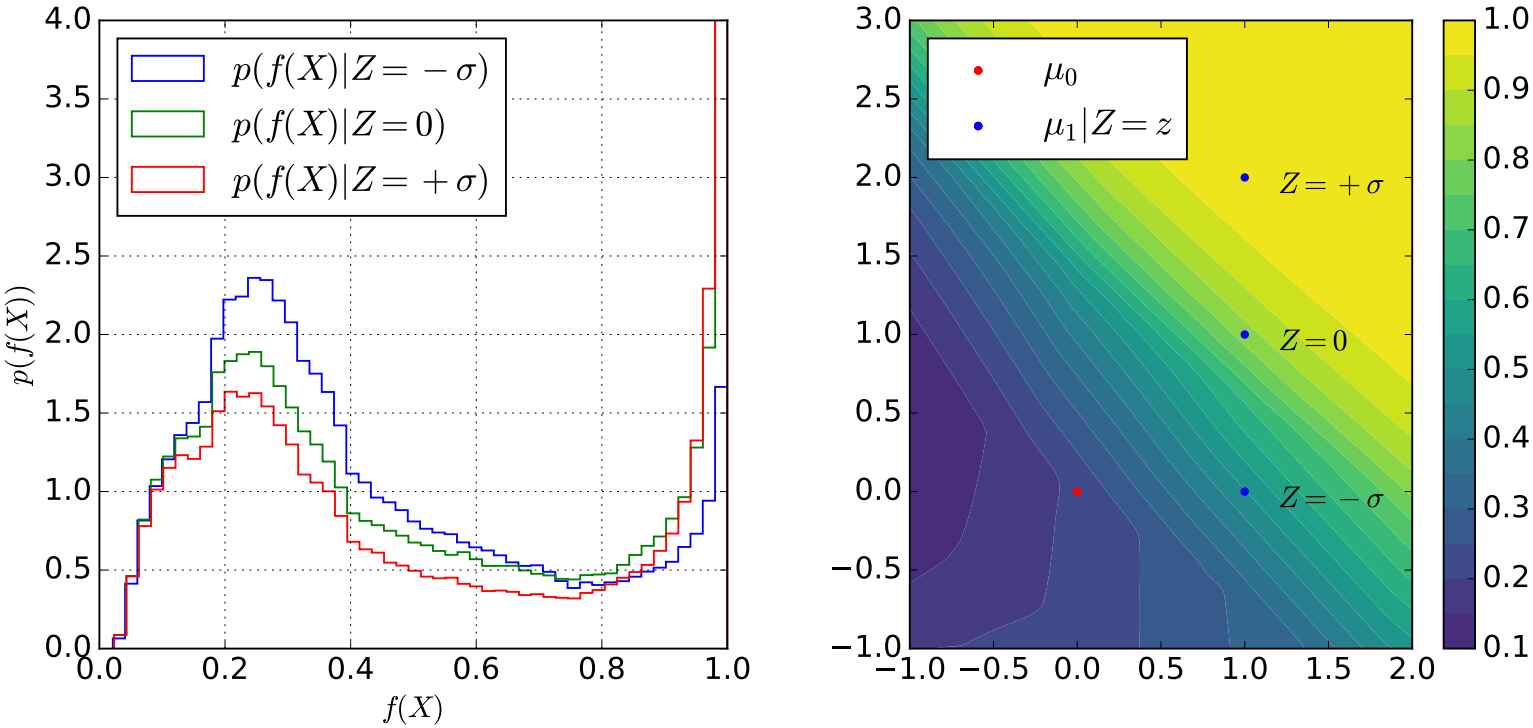
$$x \sim \mathcal{N}\left((0,0), \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}\right) \qquad \text{when } Y = 0,$$

$$x \sim \mathcal{N}\left((1, 1+Z), \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \qquad \text{when } Y = 1.$$



- The continuous nuisance parameter Z represents in this case our uncertainty about the exact location of the mean of the second gaussian. We assume a gaussian prior $z \sim \mathcal{N}(0,1)$.
- We assume training data $\{x_i, y_i, z_i\}_{i=1}^{N} \sim p(X, Y, Z)$.

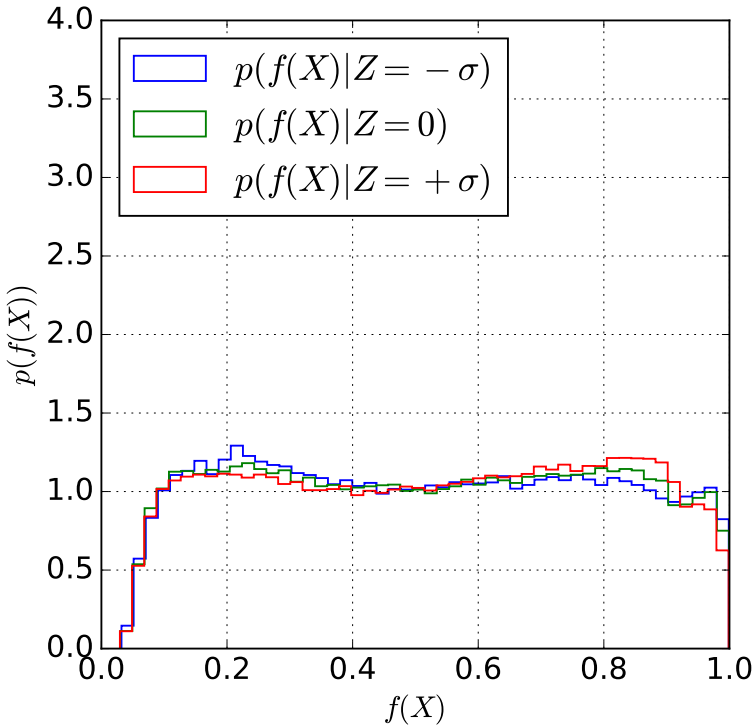
Standard training without the adversary r

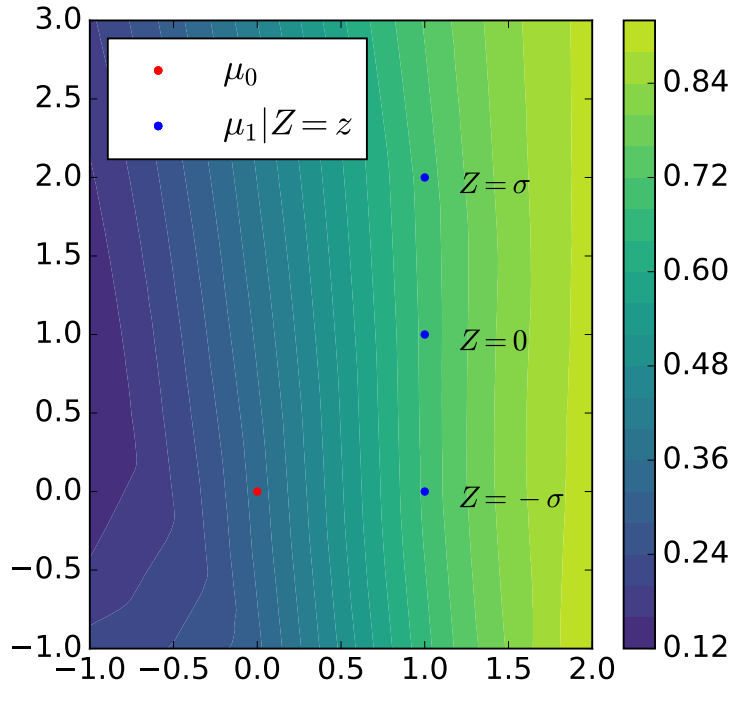


(Left) The conditional probability distributions of $f(X;\theta_f)|Z=z$ changes with z.

(Right) The decision surface strongly depends on X_2.

Reshaping f with adversarial training

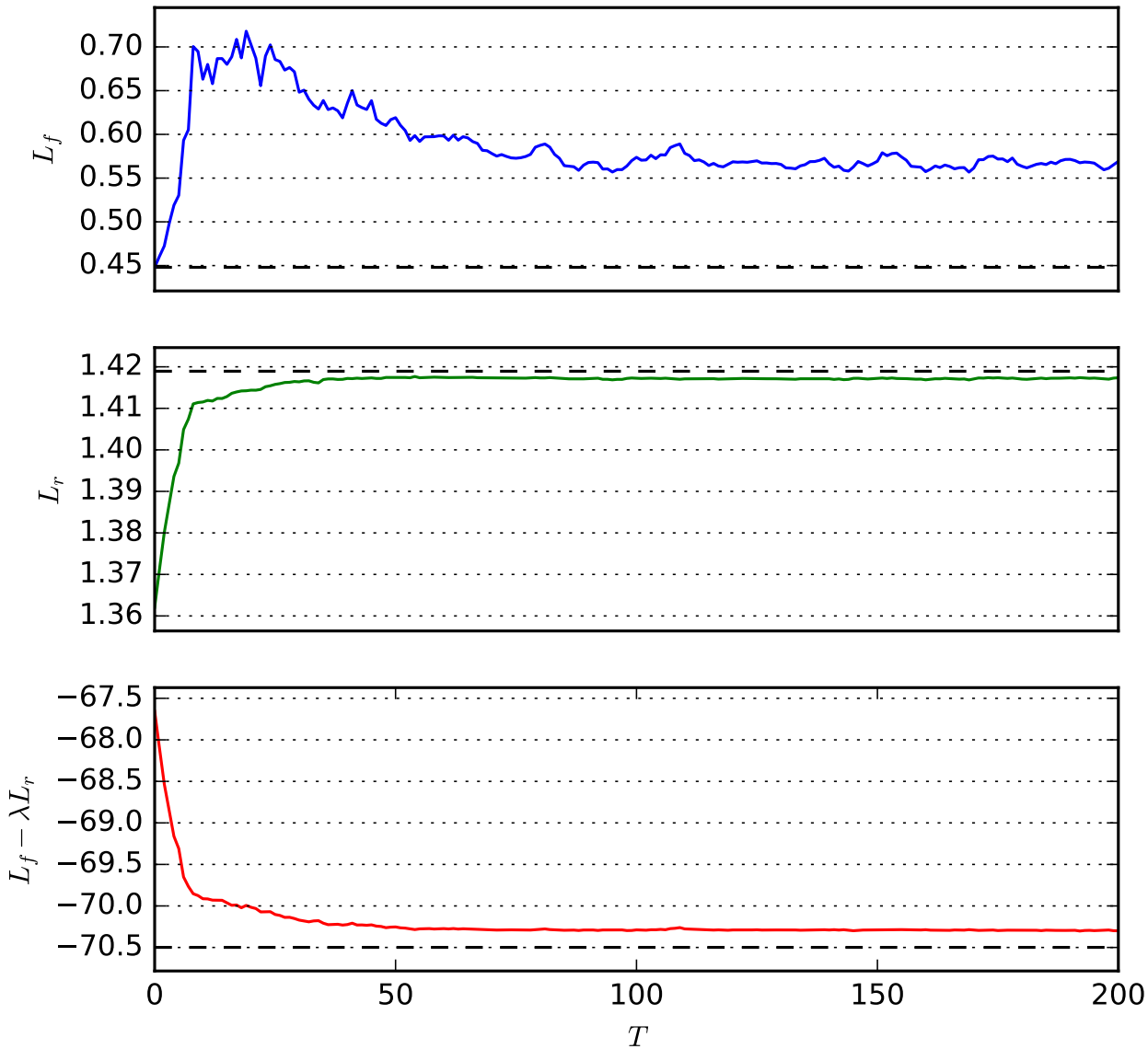




(Left) The conditional probability distributions of $f(X;\theta_f)|Z=z$ are now (almost) invariant with z!

(Right) The decision surface is now independent of X_2.

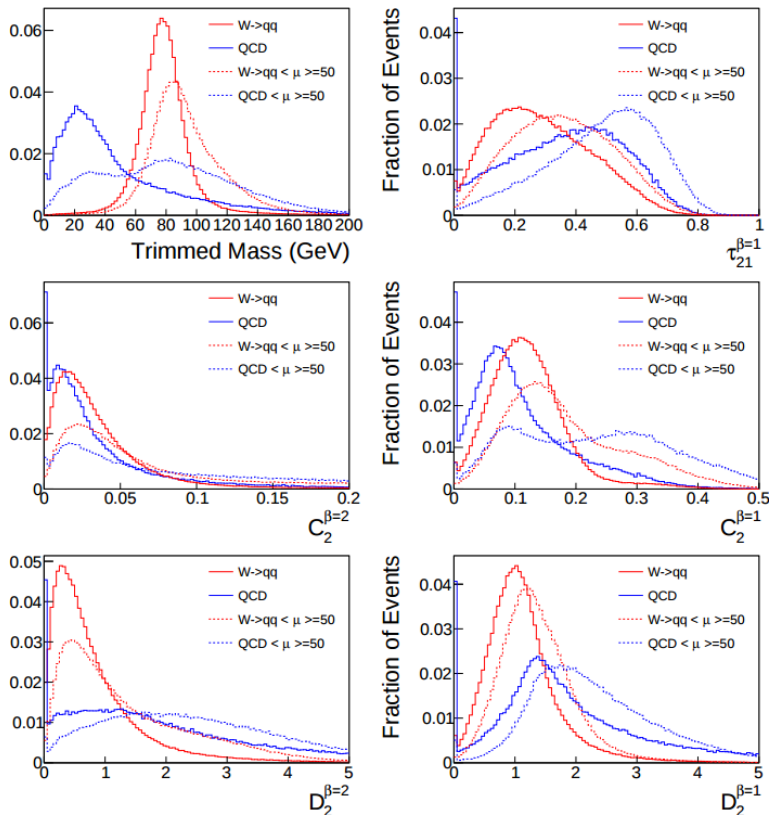
Dynamics of adversarial training



Physics example: pileup independence

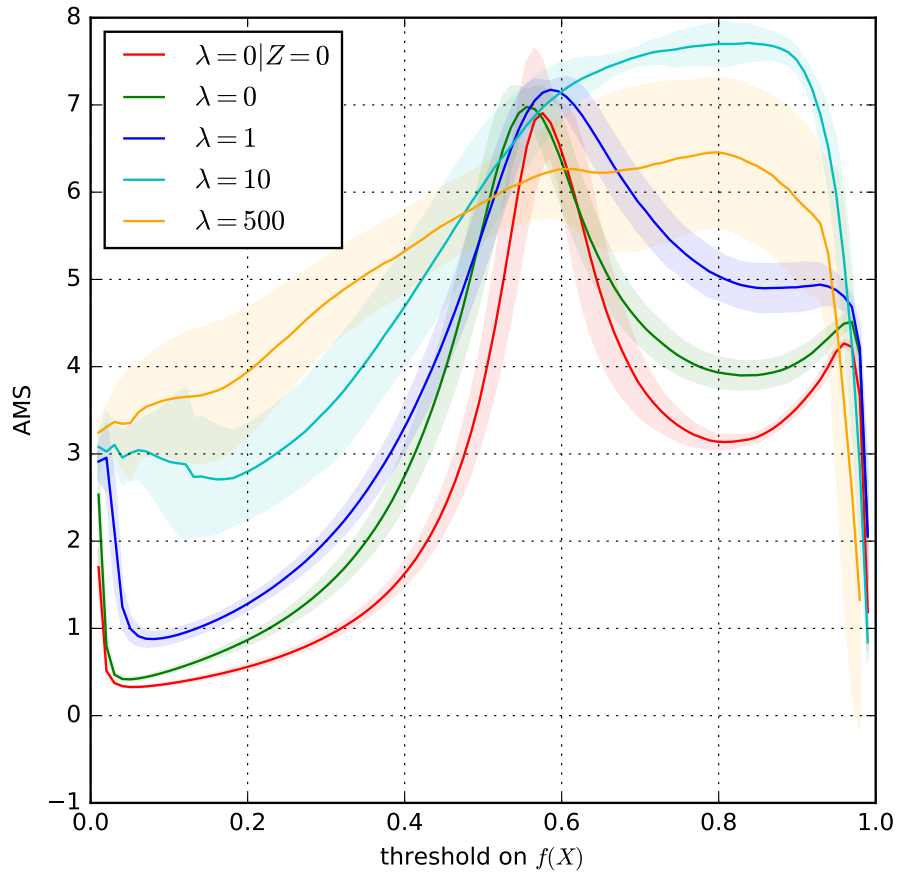
- Discriminate between QCD jets ($Y = 0$) and W-jets ($Y = 1$) from high-level features (data from Baldi et al, arXiv:1603.09349).

- Taking some liberty, we consider an extreme categorical nuisance parameter where
  - $Z = 0$ corresponds to events without pileup,
  - $Z = 1$ corresponds to events with pileup, for which there are an average of 50 independent pileup interactions overlaid.



Maximizing significance by tuning λ

- We optimize the accuracy-independence trade-off by tuning λ with respect to a higher level objective.
- Cut and count analysis: Hypothesis test of a null with no signal events against an alternate hypothesis that is a mixture of signal and background events.
  - Background = 1000 weighted QCD jets, Signal = 100 weighted boosted W's.
  - Without systematics, optimizing $\mathcal{L}_f$ indirectly optimizes the power of a classical hypothesis test.
  - With systematics, we need to balance classification performance against robustness to the nuisance parameter.
  - To this end, we use the Approximate Median Significance (AMS) as higher-level objective.
  - Note that since we are performing a hypothesis test of the null, we only wish to impose the pivotal property on background events.

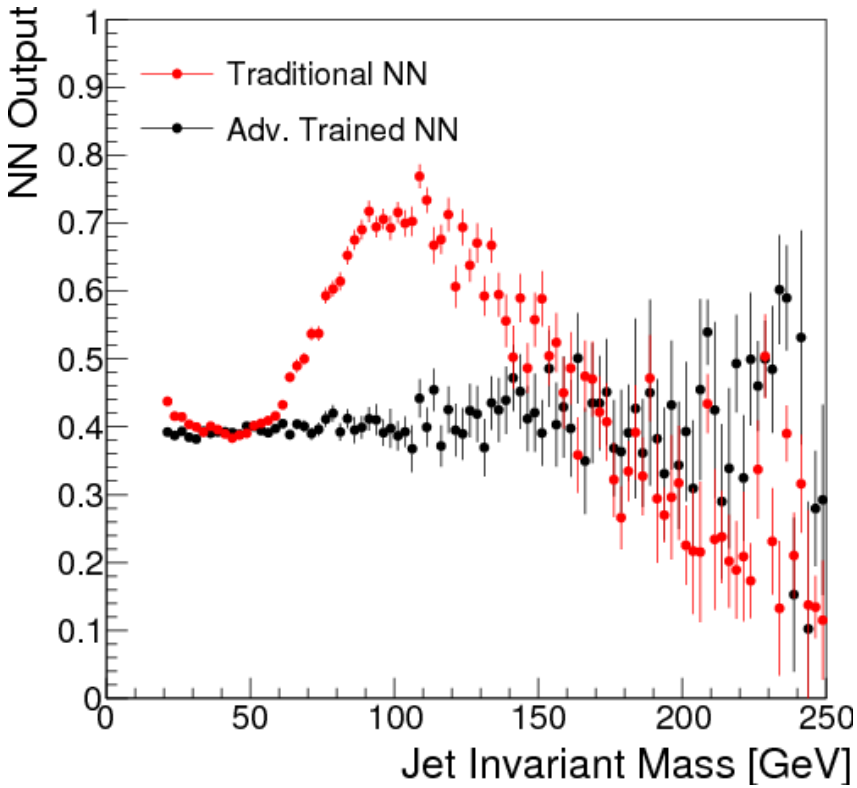


$\lambda = 0|Z = 0$: standard training from $p(X, Y, Z = 0)$.

$\lambda = 0$: standard training from $p(X, Y, Z)$.

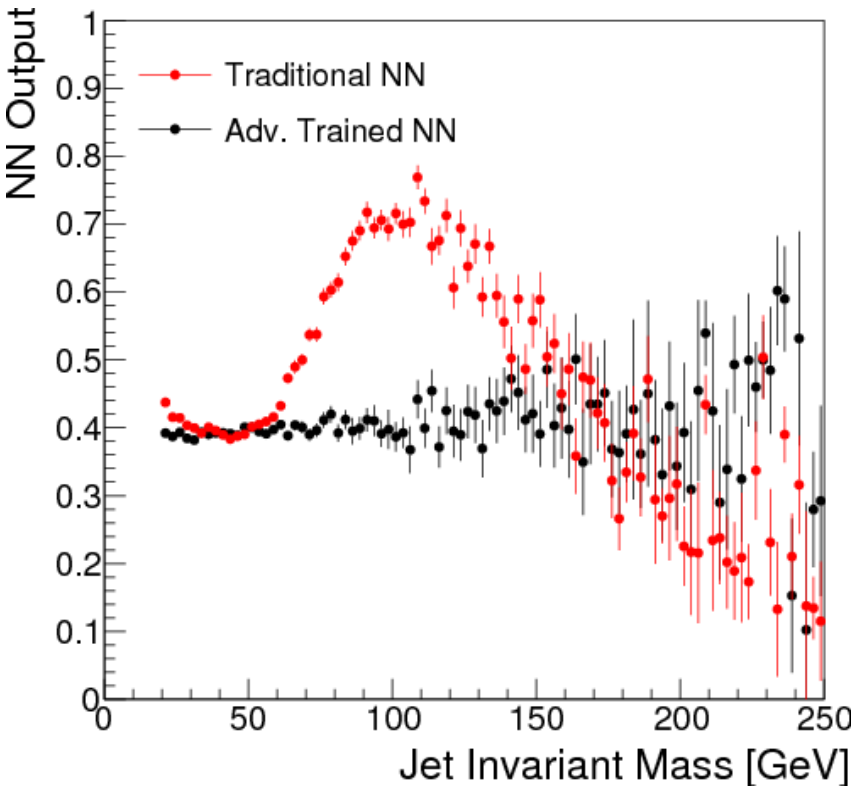
$\lambda = 10$: trading accuracy for robustness wrt pileup results in a net benefit in terms of maximum statistical significance.

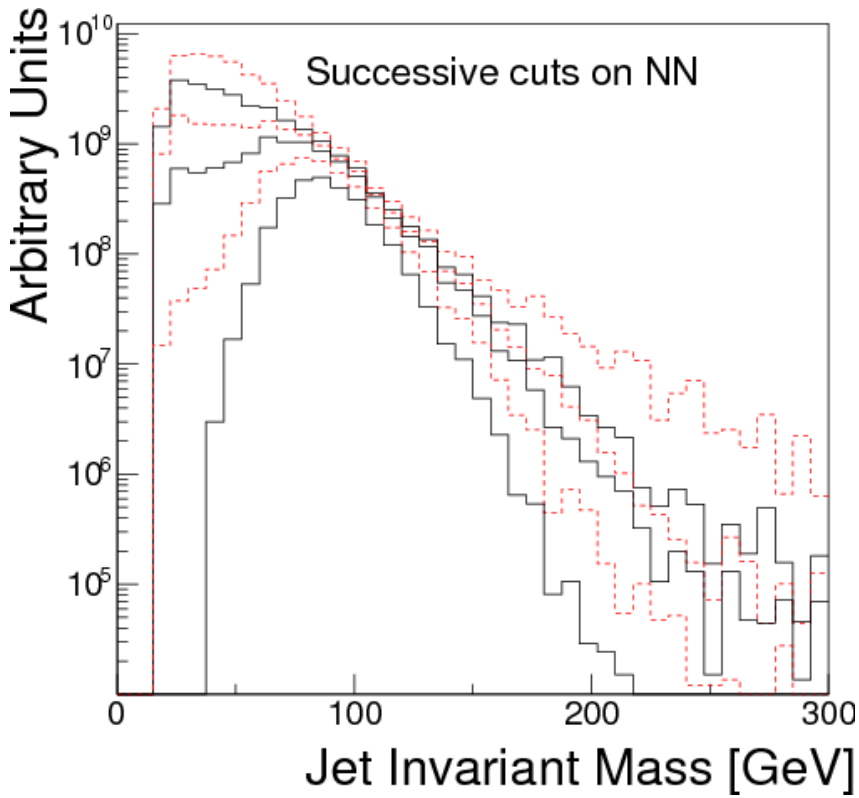
Decorrelated Jet Substructure Tagging using Adversarial Neural Networks (Shimmin et al, 1703.03507)



✓ Tagger profile is flatter

Decorrelated Jet Substructure Tagging using Adversarial Neural Networks (Shimmin et al, 1703.03507)

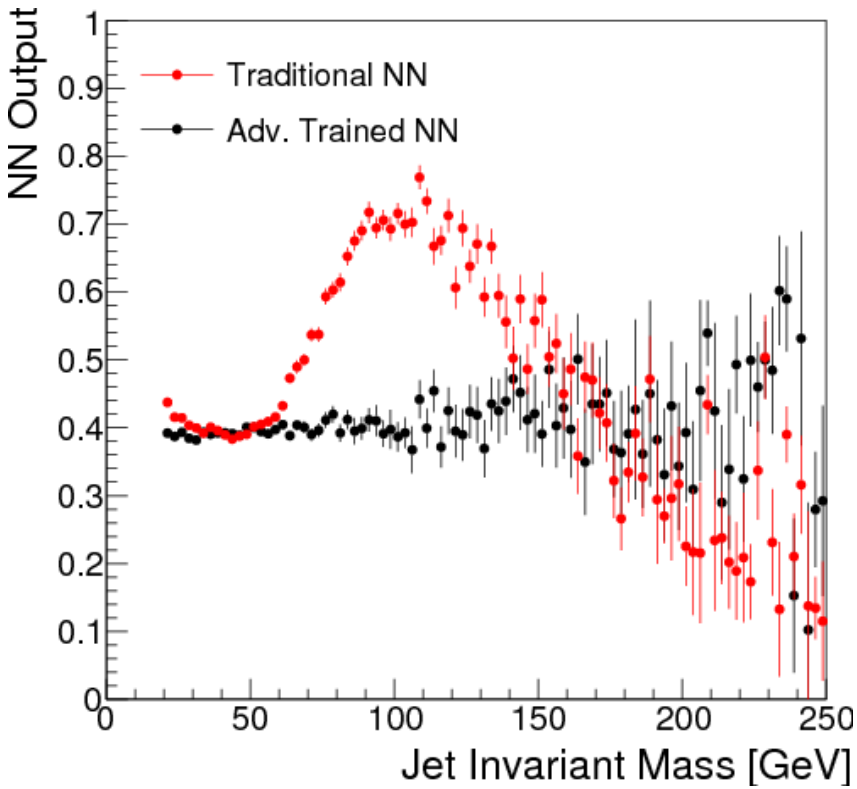


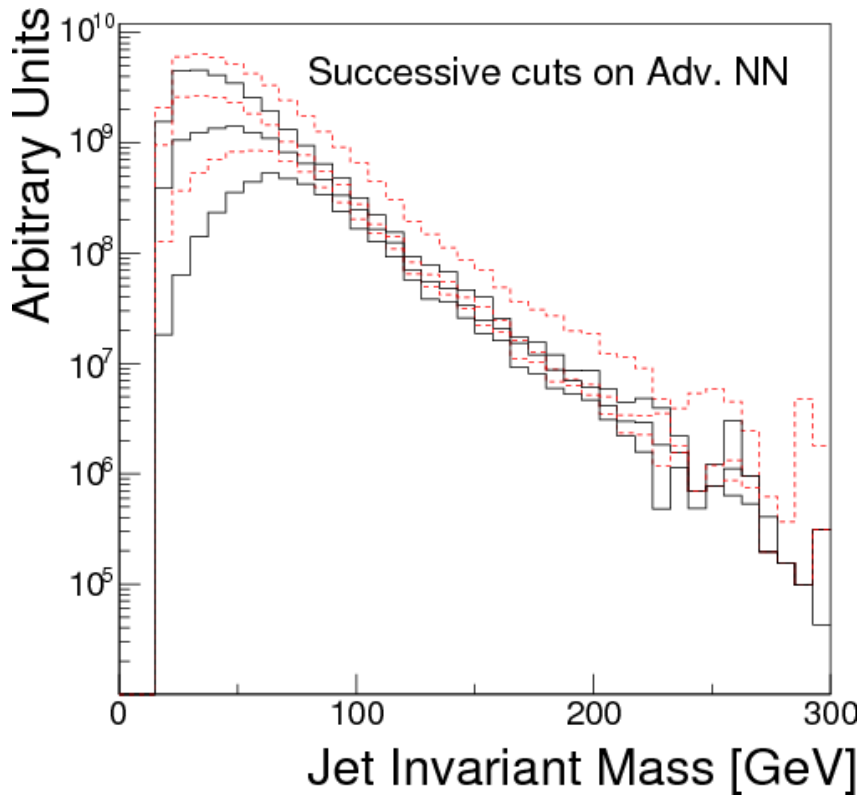


✓ Tagger profile is flatter

X Background distortion (standard neural net)

Decorrelated Jet Substructure Tagging using Adversarial Neural Networks (Shimmin et al, 1703.03507)





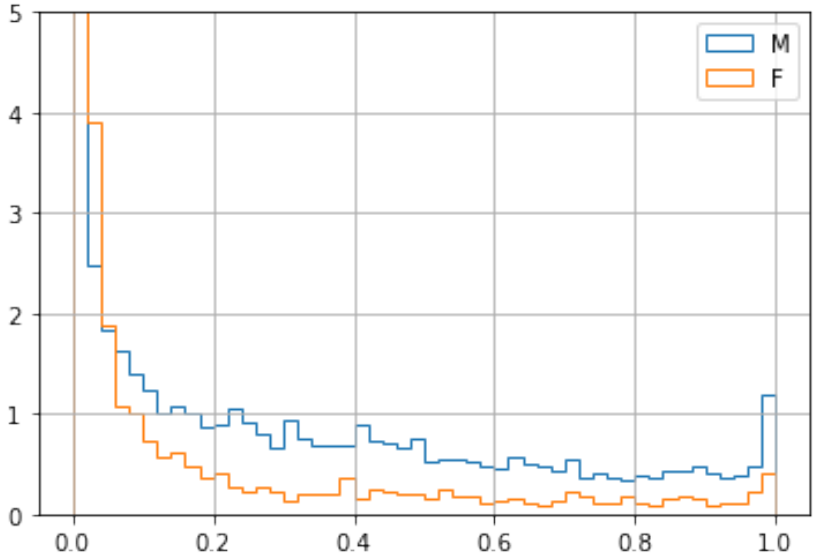
✓ Tagger profile is flatter

✓ Background distortion is reduced (adversarial net)

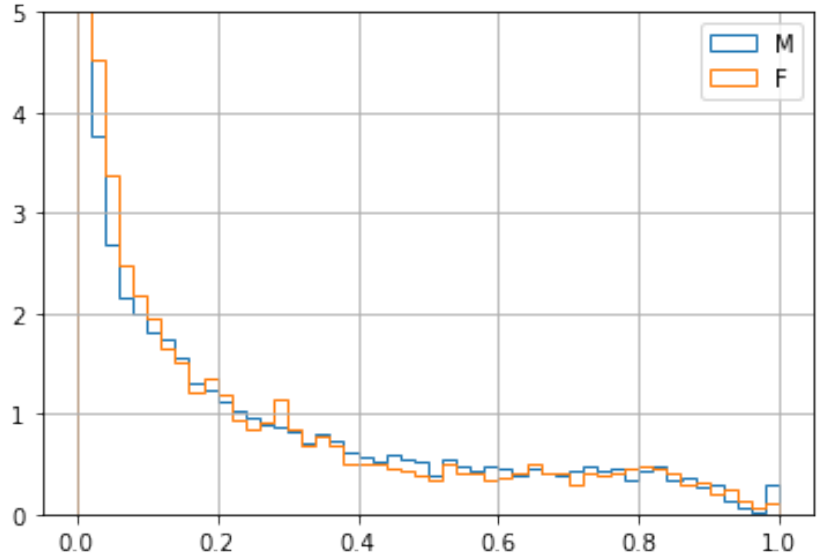
See Jupyter notebook

Fairness in machine learning

- Learning to pivot extends beyond high energy physics.
- Example: predict whether someone makes over 50,000$ a year from demographic data. We want to build a *fair* classifier, that is independent of gender.



X Women are less likely than men to make more than 50,000$ a year, because of gender bias in the data.



✓ With adversarial training, gender bias is corrected.

Summary

- We proposed a principled approach based on adversarial networks for building a model whose output can be constrained to be independent of a chosen random variable. E.g.:
  - a specific (physics) variate such as mass
  - a nuisance parameter

- The method supports both the categorical and continuous cases.

- Control is offered to tune the accuracy versus robustness trade-off in order to maximize a higher-level objective.