Adversarial Training against Systematic Uncertainty

Gilles Louppe,[1] **Michael Kagan**,[2] and **Kyle Cranmer**[1]
[1] *New York University*
[2] *SLAC National Accelerator Laboratory*

In high energy physics, systematic uncertainties represent our incomplete knowledge in the theory of physical processes or in the properties of the experimental detection apparatus. In effect, these uncertainties typically cause variations in the conditional distributions of the data, thereby altering the decisions and resulting efficacy of a classifier trained from it. To contain this issue, we propose to repurpose adversarial training as a means to learn a pivotal classifier which is invariant to variations of the nuisance parameters describing the effects of systematic uncertainties. In particular, we show and derive theoretical conditions under which classification optimality and invariance with respect to systematics can be achieved, which we confirm experimentally on a couple of illustrative examples. In machine learning, this problem is closely related to those of domain adaptation and enforcing fairness in classification. Following that line of work, the proposed method can be regarded as generalization that also supports the continuous case, which can be viewed as handling infinitely many domains or as enforcing fairness over continuous features.

I. INTRODUCTION

The discovery of new particles or physical phenomena in high energy physics experiments, like those at the Large Hadron Collider, requires the observation of statistically significant deviations from the predictions of the Standard Model, the current model of the known fundamental particles and their interactions. Typically, discovery requires satisfying the *5-sigma rule*, whereby the deviation must be at least five standard deviations from predictions, i.e. having a p-value of $p \lesssim 3 \times 10^{-7}$. The evaluation of the statistical significance must not only account for statistical uncertainties due to random fluctuations, but also for systematic uncertainties that represent our incomplete knowledge of the theory of particle interactions or of the properties of the experimental detection apparatus. Systematic uncertainties can alter the expected rate of signal and background classes, e.g. their priors, as well as the distributions of features. As such, the signal class true-positive rate and the background class false-positive rate of a classifier can change with variations in the systematic uncertainty thereby decreasing the sensitivity of an analysis. In this paper, we present a new strategy for training a classifier, based on adversarial training [16], which aims at building a classifier whose decision function is invariant under variations due to systematic uncertainties. In this way, uncertainties on predictions from simulations will have a reduced impact when computing the significance of potential deviations after application of the discriminant model and thereby increase the sensitivity of the analysis to signals from potential new physics sources.

Given data X and associated labels Y taking values $y \in \mathcal{Y} = \{s, b_{i=1\dots k}\}$ where s is the signal class and $b_{i=1\dots k}$ are the background classes, the probability density of the data can be written as a mixture

$$p(X) = \pi_s p(X|Y=s) + \sum_{i=1}^{k} \pi_{b_i} p(X|Y=b_i), \quad (1)$$

where π_y represents the prior for the signal or a given background, and $p(X|Y)$ is the conditional distribution of the data for the signal or a given background. Systematic uncertainties can affect this distribution in the following ways [7, 21]:

- Uncertainties on the priors π_y. These arise from uncertainties on the rates we expect a given process to occur as predicted from theoretical calculations or from alternative measurements of the rate of a given class.
- Uncertainties in the simulators of the physical interactions, giving rise to uncertainties on the distributions $p(X|Y)$. These uncertainties are often determined by comparing the predictions of different simulations using differing models of the same physical process.
- Uncertainties on the detection apparatus and its effect on the measurement of particle properties, giving rise to uncertainties on the distributions $p(X|Y)$. For instance, this includes uncertainties on our knowledge of the mean and variance of the measured energy distribution of a particle.

In this paper, we will not address uncertainties of the first type mentioned above (uncertainties on the priors). Rather, we will focus on mitigating the effect of systematic uncertainties that alter the conditional distributions $p(X|Y)$. [GL: Remove this comment?]

Systematic uncertainties are parameterized by nuisance parameters in the statistical analysis of data (see e.g., [7, 9]). Nuisance parameters often, though not always, have a suitable prior coming from alternative measurements in data. For instance, the average measured energy of an electron in simulation may be compared with those in data using a relatively pure sample of Z boson decays to electrons. The average measured electron energy can not be measured without uncertainty, and the measured uncertainty can be used in other analyses

to estimate the uncertainty deriving from our imperfect knowledge of the electron's energy. A nuisance parameter is assigned to the electron energy uncertainty, which controls the variation of the electron energy in the statistical analysis of data. Nuisance parameters are typically, though not always, constrained by a suitable prior, e.g. a Gaussian distribution with a standard deviation equal to the measured uncertainty. [GL: This paragraph could be improved, it is not very clear what we want to say...]

[GL: Add paper outline.]

II. PROBLEM STATEMENT

Let assume a probability space $(\Omega, \mathcal{F}, P)$, where Ω is a sample space, $\mathcal{F}$ is a set of events and P is a probability measure. Let consider the multivariate random variables $X_z : \Omega \mapsto \mathbb{R}^p$ and $Y : \Omega \mapsto \mathcal{Y}$, where X_z denotes a dependence of the functional X on a nuisance parameter Z, whose values $z \in \mathcal{Z}$ define a parameterized family of its systematic uncertainties. That is, X_z and Y induce together a joint probability distribution $p(X, Y|Z = z)$, where the conditional denotes X_z. For training, let further assume a finite set $\{x_i, y_i, z_i\}_{i=1}^N$ of realizations $X_{z_i}(\omega_i), Y(\omega_i)$, for $\omega_i \in \Omega$ and known values z_i of the nuisance parameter. Our goal is to learn a score function $f(\cdot; \theta_f) : \mathbb{R}^p \mapsto \mathcal{S}$ of parameters θ_f (e.g., a neural network-based probabilistic classifier) and minimizing a loss $\mathcal{L}_f(\theta_f)$ (e.g., the cross-entropy). In addition, we require that $f(X_z; \theta_f)$ should be robust to the value z of the nuisance parameter – which remains unknown at test time. More specifically, we aim at building f such that in the ideal case

$$f(X_z(\omega); \theta_f) = f(X_{z'}(\omega); \theta_f) \tag{2}$$

for all samples $\omega \in \Omega$ and all z, z' pairs of values of the nuisance parameter.

Since in general we do not have training tuples $(X_z(\omega), X_{z'}(\omega))$ (for the same unknown ω), we propose instead to solve the closely related problem of finding a predictive function f such that

$$\begin{aligned} &P(\{\omega | f(X_z(\omega); \theta_f) = s\}) \\ &\quad = P(\{\omega' | f(X_{z'}(\omega'); \theta_f) = s\}) \end{aligned} \tag{3}$$

for all values $s \in \mathcal{S}$ taken by the score function f. In words, we are looking for a predictive function f which is a pivotal quantity [11] with respect to the nuisance parameter. That is, such that the distribution of $f(X_z; \theta_f)$ is invariant with respect to the value z of the nuisance. Note that a function f for which Eqn. 2 is true necessarily satisfies Eqn. 3. In general, the converse is however not true, since the sets of samples $\{\omega | f(X_z(\omega); \theta_f) = s\}$ and $\{\omega' | f(X_{z'}(\omega'); \theta_f) = s\}$ do not need to be the same for the equality to hold. In order to simplify notations, and as only Eqn. 3 is of direct interest in this work, we denote from here on the pivotal quantity criterion as

$$p(f(X; \theta_f) = s|z) = p(f(X; \theta_f) = s|z') \tag{4}$$

for all $z, z' \in \mathcal{Z}$ and all values $s \in \mathcal{S}$ of $f(X; \theta_f)$.

III. METHOD

Joint training of adversarial networks was first proposed by [16] as a way to build a generative model capable of producing samples from random noise z. More specifically, the authors pit a generative model $g : \mathbb{R} \mapsto \mathbb{R}^p$ against an adversary classifier $d : \mathbb{R}^p \mapsto [0, 1]$ whose antagonistic objective is to recognize real data X from generated data $g(Z)$. Both models g and d are trained simultaneously, in such a way that g learns to produce samples that are difficult to identify by d, while d incrementally adapts to changes in g. At the equilibrium, g models a distribution whose samples can be identified by d only by chance. That is, assuming enough capacity in d and g, the distribution of $g(Z)$ eventually converges towards the real distribution of X.

In this work, we repurpose adversarial networks as a means to constraint the predictive model f in order to satisfy Eqn. 4. As illustrated in Fig. 1, we pit f against an adversary model $r := p_{\theta_r}(z|f(X; \theta_f) = s)$ of parameters θ_r and associated loss $\mathcal{L}_r(\theta_f, \theta_r)$. This model takes as input realizations s of $f(X; \theta_f)$, for the current value θ_f of f parameters, and produces as output a function $p_{\theta_r}(z|f(X; \theta_f) = s)$ modeling the posterior probability density that z parameterizes the sample observed as s. Intuitively, if $p(f(X; \theta_f) = s|z)$ varies with z, then the corresponding correlation can be captured by r. By contrast, if $p(f(X; \theta_f) = s|z)$ is invariant with z, as we require, then r should perform poorly and be close to random guessing. Training f such that it additionally minimizes the performance of r therefore acts as a regularization towards Eqn. 4.

If Z takes discrete values, then p_{θ_r} can be represented e.g. as a probabilistic classifier $\mathbb{R} \mapsto \mathbb{R}^{|\mathcal{Z}|}$ whose output j (for $j = 1, \ldots, |\mathcal{Z}|$) is the estimated probability mass $p_{\theta_r}(z_j|f(X; \theta_f) = s)$. Similarly, if Z takes continuous values and if we assume some parameteric distribution $\mathcal{P}$ for $Z|f(X; \theta_f) = s$ (e.g., a mixture of gaussians, as modeled with a mixture density network [5]), then p_{θ_r} can be represented e.g. as network whose output j is the estimated value of the corresponding parameter γ_j of that distribution (e.g., the mean, variance and mixing coefficients of its components). As in [5, 19], the estimated probability density $p_{\theta_r}(z|f(X; \theta_f) = s)$ can then be evaluated for any $z \in \mathcal{Z}$ and any score $s \in \mathcal{S}$. As further explained in the next section, let us note that the adversary r may take any form, i.e. it does need to be a neural network, as long as it exposes a differentiable function $p_{\theta_r}(z|f(X; \theta_f) = s)$ of sufficient capacity to represent the true distribution.

As for generative adversarial networks, we propose to train f and r simultaneously, which we carry out by considering the value function

$$E(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) - \mathcal{L}_r(\theta_f, \theta_r) \tag{5}$$

that we optimize by finding the saddle point $(\hat{\theta}_f, \hat{\theta}_r)$ such

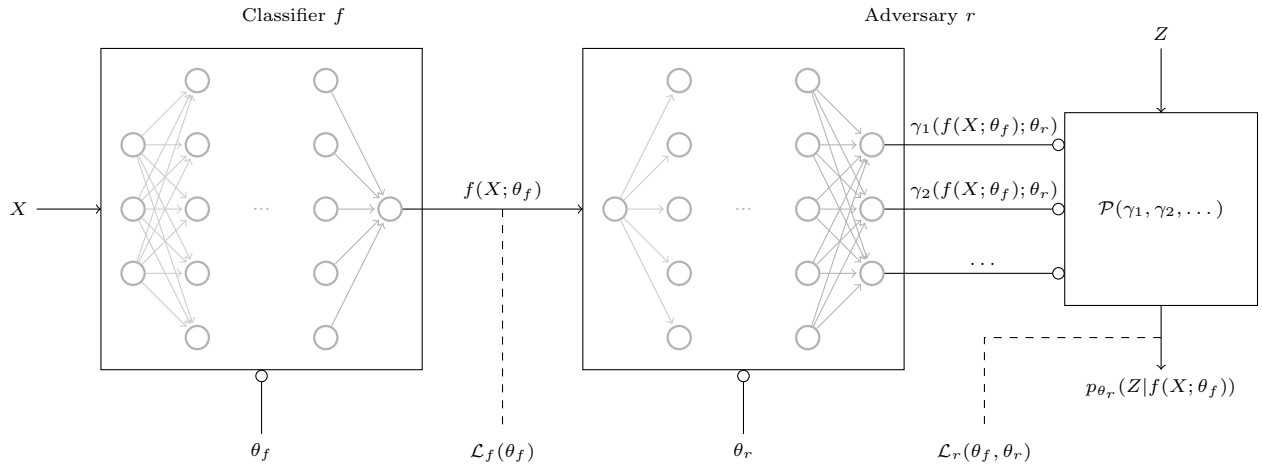

FIG. 1. Architecture for the adversarial training of a binary classifier f against a nuisance parameter Z. The adversary r models the distribution $p(z|f(X;\theta_f)=s)$ of the nuisance as observed only through the output $f(X;\theta_f)$ of the classifier. By maximizing the antagonistic objective $\mathcal{L}_r(\theta_f,\theta_r)$ (as part of minimizing $\mathcal{L}_f(\theta_f)-\lambda\mathcal{L}_r(\theta_f,\theta_r)$), the classifier f forces $p(z|f(X;\theta_f)=s)$ towards the prior $p(z)$, which happens when $f(X;\theta_f)$ is independent of the nuisance parameter Z and therefore pivotal.

that

$$\hat{\theta}_f = \arg\min_{\theta_f} E(\theta_f, \hat{\theta}_r), \tag{6}$$

$$\hat{\theta}_r = \arg\max_{\theta_r} E(\hat{\theta}_f, \theta_r). \tag{7}$$

Without loss of generality, the adversarial training procedure to obtain $(\hat{\theta}_f, \hat{\theta}_r)$ is formally presented in Algorithm 1 in the case of a binary classifier $f : \mathbb{R}^p \mapsto [0,1]$ modeling $p(Y=1|X)$. For reasons further explained in Section IV, $\mathcal{L}_f$ and $\mathcal{L}_r$ are respectively set to the expected value of the negative log-likelihood of $Y|X$ under f and of $Z|f(X;\theta_f)$ under r:

$$\mathcal{L}_f(\theta_f) = \mathbb{E}_{x\sim X}\mathbb{E}_{y\sim Y|x}[-\log p_{\theta_f}(y|x)], \tag{8}$$

$$\mathcal{L}_r(\theta_f,\theta_r) = \mathbb{E}_{s\sim f(X;\theta_f)}\mathbb{E}_{z\sim Z|s}[-\log p_{\theta_r}(z|s)]. \tag{9}$$

The optimization algorithm consists in using stochastic gradient descent alternatively for solving Eqn. 6 and 7.

IV. THEORETICAL RESULTS

In this section, we show that in the setting of Algorithm 1 where $\mathcal{L}_f$ and $\mathcal{L}_r$ are respectively set to expected value of the negative log-likelihood of $Y|X$ under f and of $Z|f(X;\theta_f)$ under r, the procedure converges to a classifier f which is a pivotal quantity in the sense of Eqn. 4.

In this setting, the nuisance parameter Z is considered as a random variable of prior $p(z)$ (for $z \in \mathcal{Z}$), and our goal is to find a function $f(\cdot;\theta_f)$ such that $f(X;\theta_f)$ and Z are independent random variables. Importantly, classification of Y with respect to X is considered in the context where Z is marginalized out, which means that the classifier minimizing $\mathcal{L}_f$ is optimal with respect to $Y|X$, but not necessarily with $Y|X,Z$ (unless Z is made explicit and is included among the input variables in X). Results hold for a nuisance parameter Z taking either categorical or continuous values. By abuse of notation, $H(Z)$ denotes the differential entropy in this latter case. Finally, the proposition below is derived in a non-parametric setting, by assuming that both f and r have enough capacity.

Proposition 1. *If there exists a saddle point $(\hat{\theta}_f, \hat{\theta}_r)$ for Eqn. 6 and 7 such that $E(\hat{\theta}_f, \hat{\theta}_r) = H(Y|X) - H(Z)$, then $f(\cdot;\hat{\theta}_f)$ is both an optimal classifier and a pivotal quantity.*

Proof. For fixed θ_f, the adversary r is optimal at

$$\hat{\theta}_r = \arg\max_{\theta_r} E(\theta_f,\theta_r) = \arg\min_{\theta_r} \mathcal{L}_r(\theta_f,\theta_r),$$

in which case $p_{\hat{\theta}_r}(z|f(X;\theta_f)=s) = p(z|f(X;\theta_f)=s)$ for all z and all s, and $\mathcal{L}_r$ reduces to the expected entropy $\mathbb{E}_{s\sim f(X;\theta_f)}[H(Z|f(X;\theta_f)=s)]$ of the conditional distribution of the nuisance. This expectation is nothing else than the conditional entropy of the random variables Z and $f(X;\theta_f)$ and can be written as $H(Z|f(X;\theta_f))$. Accordingly, the value function E can be restated as a function depending on θ_f only:

$$E'(\theta_f) = \mathcal{L}_f(\theta_f) - H(Z|f(X;\theta_f)).$$

In particular, we have the lower bound

$$H(Y|X) - H(Z) \leq \mathcal{L}_f(\theta_f) - H(Z|f(X;\theta_f))$$

where the equality holds at $\hat{\theta}_f = \arg\min_{\theta_f} E'(\theta_f)$ when:

- $\hat{\theta}_f$ minimizes the negative log-likelihood of $Y|X$ under f, which happens when $\hat{\theta}_f$ are the parameters of an optimal classifier. In this case, $\mathcal{L}_f$ reduces to its minimum value $H(Y|X)$.

Algorithm 1 Adversarial training of a classifier f against an adversary r.

Inputs: training data $\{x_i, y_i, z_i\}_{i=1}^N$;
Outputs: $\hat{\theta}_f, \hat{\theta}_r$;
Hyper-parameters: Number T of training iterations, Number K of gradient steps to update r.

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

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

  end for
  Sample minibatch $\{x_m, y_m, z_m, s_m = f(x_m; \theta_f)\}_{m=1}^M$ of size M; ▷ Update f
  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 $1(y_m = 0)(1 - s_m) + 1(y_m = 1)s_m$;
end for

- $\hat{\theta}_f$ maximizes the conditional entropy $H(Z|f(X;\theta_f))$ since $H(Z|f(X;\theta)) \leq H(Z)$. Note that this latter inequality holds for both the discrete and the differential definitions of entropy.

When the lower bound is active, we have $H(Z|f(X;\theta_f)) = H(Z)$ because of the second condition, which happens exactly when Z and $f(X;\theta_f)$ are independent variables. In other words, the optimal classifier $f(\cdot;\hat{\theta}_f)$ is also a pivotal quantity.

□

Proposition 1 suggests that if at each step of Algorithm 1 the adversary r is allowed to reach its optimum given f (e.g., by setting K sufficiently high) and if f is updated to improve $\mathcal{L}_f(\theta_f) - H(Z|f(X;\theta_f))$, then f should converge to a classifier which is both optimal and pivotal, provided such a classifier exists. A formal proof of convergence of the alternating stochastic gradient descent procedure of Algorithm 1 remains however to be proven.

On many practical problems, let us note that such a classifier may not exist because the nuisance parameter directly shapes the decision boundary. In this case, the lower bound $H(Y|X) - H(Z) < \mathcal{L}_f(\theta_f) - H(Z|f(X;\theta_f))$ is strict: f can either be an optimal classifier or a pivotal quantity, but not both simultaneously. In this situation, it is natural to rewrite the value function E as

$$E_\lambda(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) - \lambda \mathcal{L}_r(\theta_f, \theta_r), \tag{10}$$

where $\lambda \geq 0$ 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 will preferably enforces f to be pivotal while setting λ close to 0 will rather constraint f to be optimal.

Interestingly, let us finally emphasize that these results hold using only the (1D) output s of $f(\cdot;\theta_f)$ (in the case of binary classification) as input to the adversary. We could similarly enforce an intermediate representation of the data to be pivotal, e.g. as in [14], but this is in fact not necessary.

V. EXPERIMENTS

A. Toy example

As a guiding toy example, let us consider the binary classification of 2D data drawn from multivariate gaussians with equal priors, such that

$$x \sim \mathcal{N}((0,0), \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}) \qquad \text{when } Y = 0, \tag{11}$$

$$x \sim \mathcal{N}((1, 1+Z), \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}) \qquad \text{when } Y = 1. \tag{12}$$

The continuous nuisance parameter Z represents in this case our uncertainty about the exact location of the mean of the second gaussian. Our goal is to build a classifier $f(\cdot;\theta_f)$ for predicting Y given X, but such that the probability distribution of $f(X;\theta_f)$ is invariant with respect to the nuisance parameter Z.

Assuming a gaussian prior $z \sim \mathcal{N}(0,1)$, we start by generating training data $\{x_i, y_i, z_i\}_{i=1}^N$, from which we train a neural network classifier f minimizing $\mathcal{L}_f(\theta_f)$ without considering its adversary r. The network architecture comprises 2 dense hidden layers of 20 nodes with ReLU activations, followed by a dense output layer with a single node with a sigmoid activation. As shown

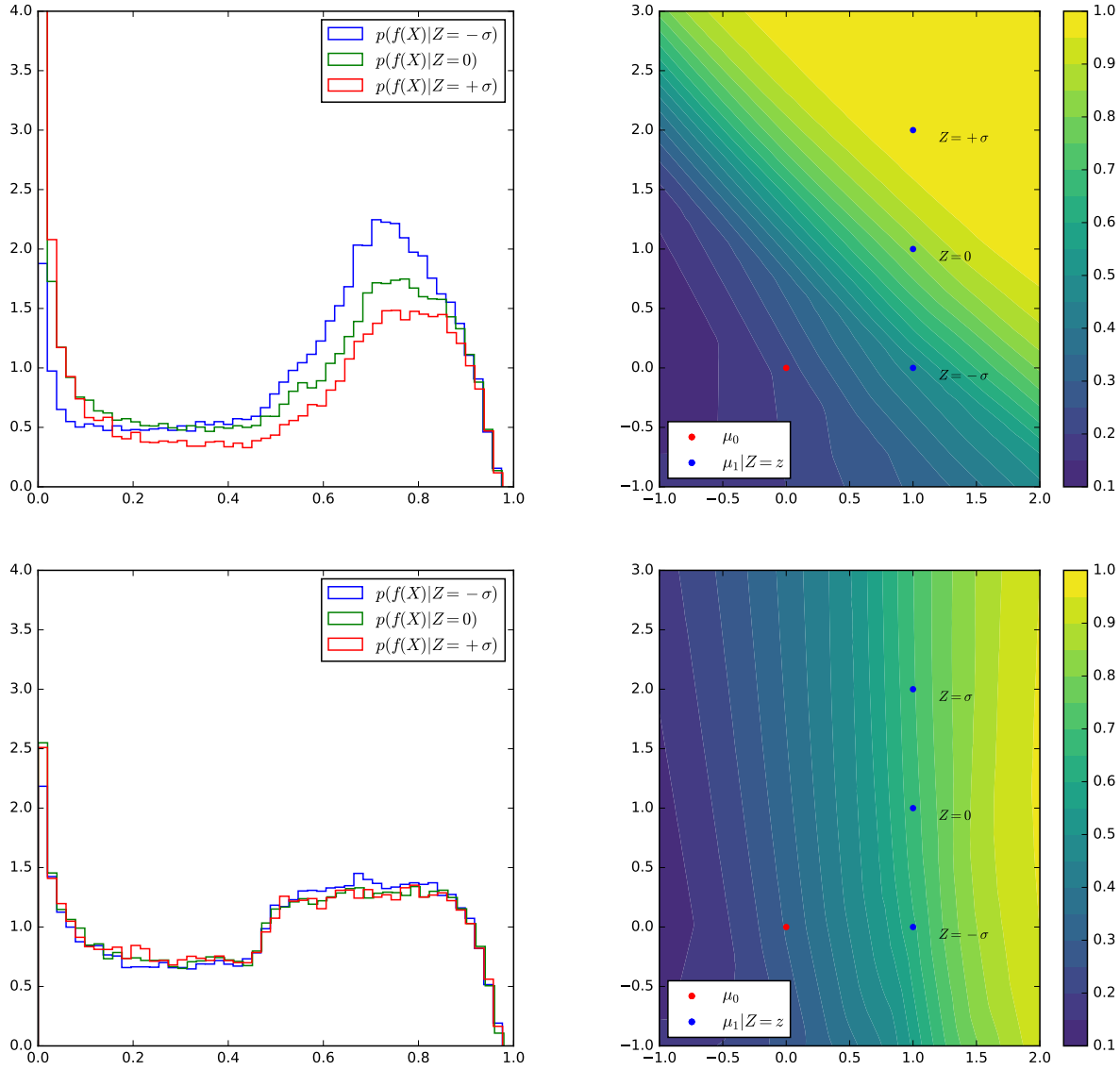

FIG. 2. Toy example. (Upper left) Conditional probability densities of the decision scores at $Z = -\sigma, 0, \sigma$ when f is built without adversarial training. The resulting densities are clearly dependent on the continuous parameter Z, indicating that f is not pivotal. (Upper right) The corresponding decision surface, highlighting the fact that samples are easier to classify for values of Z above to σ, hence partially explaining the dependency. (Lower left) Conditional probability densities of the decision scores at $Z = -\sigma, 0, \sigma$ when f is built with adversarial training, as outlined in Section III. The resulting densities are now almost identical to each other, indicating only a small dependency on Z. (Lower right) The corresponding decision surface, illustrating how adversarial training bends the decision function vertically to erase the dependency on Z.

in the upper plots of Fig. 2, the resulting classifier is not pivotal, as the conditional probability densities of its decision scores $f(X; \theta_f)$ show large discrepancies between values z of the nuisance. While not shown here, a classifier trained only from data generated at the nominal value $Z = 0$ would also not be pivotal.

Let us now consider the joint training of f against an adversary r implemented as a mixture density network modeling $Z|f(X; \theta_f)$ as a mixture of five gaussians. As for f, the network architecture comprises 2 dense hidden layers of 20 nodes with ReLU activations, but is followed by an output layer of 15 nodes corresponding to the means, standard deviations and mixture coefficients of the five gaussians. Output nodes for the mean values come with linear activations, output nodes for the standard deviations with exponential activations to ensure positivity, while output nodes for the mixture coefficients implement the softmax function to ensure positivity and normalization. When running Algorithm 1 as initialized with the classifier f obtained previously, ad-

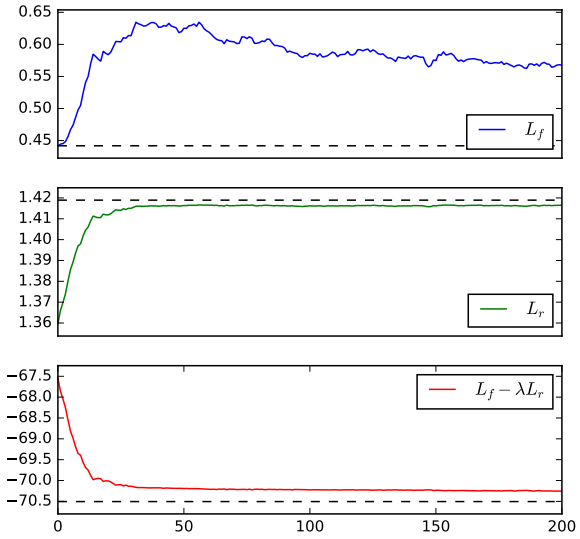

FIG. 3. Toy example. Training curves for $\mathcal{L}_f(\theta_f)$, $\mathcal{L}_r(\theta_f, \theta_r)$ and $\mathcal{L}_f(\theta_f) - \lambda\mathcal{L}_r(\theta_f, \theta_r)$. Adversarial training was performed for 200 iterations, mini-batches of size $M = 128$, $K = 500$ and $\lambda = 50$.

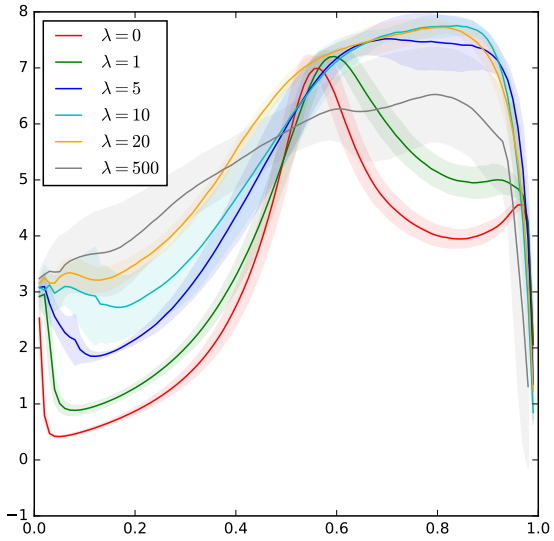

FIG. 4. Approximate median significance as a function of the decision threshold on the output of f. As shown at $\lambda = 10$, trading classification accuracy for independence with respect to pileup results in a positive total net benefit in terms of statistical significance.

versarial training effectively reshapes the decision function so it that becomes almost independent on the nuisance parameter, as shown in the lower plots of Fig. 2. In particular, the conditional probability densities of the decision scores $f(X;\theta_f)$ are now very similar to each other, indicating only a small residual dependency on the nuisance, as theoretically expected. The dynamics of adversarial training is illustrated in Fig. 3, where the losses $\mathcal{L}_f$, $\mathcal{L}_r$ and $\mathcal{L}_f - \lambda\mathcal{L}_r$ are evaluated after each iteration of Algorithm 1. In the first iterations, we observe that the global objective $\mathcal{L}_f - \lambda\mathcal{L}_r$ is minimized by making the classifier less accurate, hence the corresponding increase of $\mathcal{L}_f$, but which results in a classifier that is more pivotal, hence the corresponding increase of $\mathcal{L}_r$ and the total net benefit. As learning goes, minimizing E then requires making predictions that are more accurate, hence decreasing $\mathcal{L}_f$, or that are even less dependent on Z, hence shaping p_{θ_r} towards the prior $p(z)$. Indeed, $\mathcal{L}_f$ eventually starts to slightly decrease, while remaining lower bounded by $\min_{\theta_f} \mathcal{L}_f(\theta_f)$ as approximated by the dashed line in the first plot. Similarly, $\mathcal{L}_r$ tends towards the differential entropy $H(Z)$ of the prior (where $H(Z) = \log(\sigma\sqrt{2\pi e}) = 1.419$ in the case of a gaussian with unit variance), as shown by the dashed line in the second plot. Finally, let us note that the ideal situation of a classifier that is both optimal and pivotal appears to be unreachable for this problem, as shown in the third plot by the offset between $\mathcal{L}_f - \lambda\mathcal{L}_r$ and the dashed line approximating $H(Y|X) - \lambda H(Z)$.

B. High energy physics example

In collisions at the Large Hadron Collider, massive particles can be produced at such high energies that their decay products are collimated and the resulting jets overlap. Still, deducing whether an observed jet is either due to a single hadronic particle or to the decay of a massive object into multiple hadronic particles is critical for understanding the nature of the particles produced in the collision [3, 8]. Critically, this classification problem is often made difficult by the presence of simultaneous interactions, known as pileup, which contributes significant energy depositions unrelated to the particles of interest.

In this setting, we reuse the classification problem of [3] between single jets produced in quark or gluon fragmentation and two overlapping jets produced when a high-velocity W boson decays to a collimated pair of quarks. As nuisance parameter, we consider events without pileup ($Z = 0$) and events with pileup ($Z = 1$), for which an average number of $\langle\mu\rangle = 50$ unrelated interactions are overlaid. Our goal is to build an accurate classifier, for which we also want to minimize the effects due to the uncertainties on the nuisance. More specifically, we choose to recast the classification problem as a hypothesis test between signal+background (jets originating from W bosons or from single quarks and gluons) and background only (jets originating from single quarks and gluons only) and tune the decision threshold of our classifier by maximizing its approximate median significance (AMS), when uncertainties in the background are taken into account (see Eqn. 20 of [1]). Our motivation

is that reducing the effects of uncertainties by requiring independence of Z with the classifier output $f(X;\theta_f)$ should allow for a larger maximum significance.

To minimize the effects of Z in the background events, we train a classifier using Algorithm 1 but consider the adversarial term $\mathcal{L}_r$ conditioned on $Y = 0$ only. Both f and r are neural networks with 3 hidden layers of 64 nodes and ReLU activations, each terminated by a single final output node with a sigmoid activation. Experiments are performed on the high-level features data described in [3], on a subset of 150000 samples for training while AMS is evaluated on an independent test set of 5000000 samples. Results reported below are averages over 5 runs.

As Fig. 4 illustrates, without adversarial training (at $\lambda = 0$), the maximum significance peaks at 7. By contrast, as the independence constraint is made stronger (for $\lambda > 0$) the AMS peak moves higher, with a maximum value around 7.8 for $\lambda = 10$. In other words, trading classification accuracy for independence with respect to pileup results in a positive total net benefit in terms of statistical significance. Setting λ too high however (e.g. $\lambda = 500$) results in a decrease of the maximum significance, by focusing the capacity of f too strongly on independence, at the expense of classification accuracy. As demonstrated in this example, controlling the classification versus pivot trade-off through λ therefore gives us a principled and effective approach for maximizing significance by desensitizing the classifier output $f(X;\theta_f)$ in the most beneficial way.

VI. RELATED WORK

To account for systematic uncertainties, experimentalists in high energy physics typically take as fixed a classifier f built from training data for a nominal value z_0 of the nuisance parameter, and then propagate uncertainty by estimating $p(f(x)|z)$ with a parameterized calibration procedure. Clearly, this classifier is however not optimal for $z \neq z_0$. To overcome this issue, the classifier f is sometimes built instead on a mixture of training data generated from several nominal values $z_0, z_1, \ldots$ of the nuisance. While this certainly improves with respect to classification performance, there is however no guarantee that the resulting classifier is pivotal, as shown previously in Section V A. As an alternative, parameterized classifiers [4, 10] directly take (nuisance) parameters as additional input variables, hence ultimately providing the most statistically powerful approach for incorporating the effect of systematics on the underlying classification task. As argued in [18], such classifiers can however not be used on real data since the correct value z of the nuisance often remains unknown. This is typically not an issue in the context of parameter inference [10], where nuisance parameters are marginalized out, but otherwise often limits the range of their applications. In practice, parameterized classifiers are also computationally expensive to build and evaluate. In particular, calibrating their decision function, i.e. approximating $p(f(x,z)|z)$ as a continuous function of z, remains an open challenge. By contrast, constraining f to be pivotal yields a classifier which may not be optimal with respect to $Y|X,Z$, as discussed in Section IV, but that can otherwise be used in a wider range of applications, since knowing the correct value z of the nuisance is not necessary. Similarly, calibration needs to be carried out only once, since the dependence on the nuisance is now built-in.

In machine learning, learning a pivotal quantity can be related to the problem of domain adaptation [2, 6, 14, 15, 17, 20], where the goal is often stated as trying to learn a domain-invariant representation of the data. Likewise, our method also relates to the problem of enforcing fairness in classification [12, 13, 22, 23], which is stated as learning a classifier that is independent of some chosen attribute such as gender, color or age. For both families of methods, the problem can equivalently be stated as learning a classifier which is a pivotal quantity with respect to either the domain or the selected feature. In this context, [12, 14] are certainly among the closest to our work, in which domain invariance and fairness are enforced through an adversarial minimax setup composed of a classifier and an adversary discriminator. Following this line of work, our method can be regarded as a generalization that also supports the continuous case, which can be viewed as handling infinitely many domains, provided they can be continuously parameterized, or as enforcing fairness over continuous attributes.

VII. CONCLUSIONS

ACKNOWLEDGMENTS

[1] Adam-Bourdarios, C., Cowan, G., Germain, C., Guyon, I., Kégl, B., and Rousseau, D. The higgs boson machine learning challenge. In *NIPS 2014 Workshop on High-energy Physics and Machine Learning* (2014), vol. 42, p. 37.

[2] Baktashmotlagh, M., Harandi, M., Lovell, B., and Salzmann, M. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 769–776.

[3] Baldi, P., Bauer, K., Eng, C., Sadowski, P., and Whiteson, D. Jet substructure classification in high-energy physics with deep neural networks. *Physical Review D 93*, 9 (2016), 094034.

[4] Baldi, P., Cranmer, K., Faucett, T., Sadowski, P., and Whiteson, D. Parameterized Machine Learning for

High-Energy Physics. *arXiv preprint arXiv:1601.07913* (2016).

[5] Bishop, C. M. Mixture density networks.

[6] Blitzer, J., McDonald, R., and Pereira, F. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (2006), Association for Computational Linguistics, pp. 120–128.

[7] Bohm, G., and Zech, G. *Introduction to statistics and data analysis for physicists*. DESY, 2010.

[8] Butterworth, J. M., Davison, A. R., Rubin, M., and Salam, G. P. Jet substructure as a new higgs-search channel at the large hadron collider. *Physical review letters 100*, 24 (2008), 242001.

[9] Cowan, G., Cranmer, K., Gross, E., and Vitells, O. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C 71*, 2 (2011), 1–19.

[10] Cranmer, K., Pavez, J., and Louppe, G. Approximating likelihood ratios with calibrated discriminative classifiers.

[11] Degroot, M. H., and Schervish, M. J. *Probability and statistics*, 4 ed. 2010.

[12] Edwards, H., and Storkey, A. J. Censoring representations with an adversary.

[13] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), ACM, pp. 259–268.

[14] Ganin, Y., and Lempitsky, V. Unsupervised Domain Adaptation by Backpropagation. *ArXiv e-prints* (Sept. 2014).

[15] Gong, B., Grauman, K., and Sha, F. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of The 30th International Conference on Machine Learning* (2013), pp. 222–230.

[16] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems* (2014), pp. 2672–2680.

[17] Gopalan, R., Li, R., and Chellappa, R. Domain adaptation for object recognition: An unsupervised approach. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (2011), IEEE, pp. 999–1006.

[18] Neal, R. M. Computing likelihood functions for high-energy physics experiments when distributions are defined by simulators with nuisance parameters. In *Proceedings of PhyStat2007, CERN-2008-001* (2007), pp. 111–118.

[19] Nix, D. A., and Weigend, A. S. Estimating the mean and variance of the target probability distribution. In *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on* (1994), vol. 1, IEEE, pp. 55–60.

[20] Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. Domain adaptation via transfer component analysis. *Neural Networks, IEEE Transactions on 22*, 2 (2011), 199–210.

[21] Sinervo, P. K. Definition and treatment of systematic uncertainties in high energy physics and astrophysics. In *Statistical Problems in Particle Physics, Astrophysics, and Cosmology* (2003), vol. 1, Citeseer, p. 122.

[22] Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness constraints: A mechanism for fair classification. *arXiv preprint arXiv:1507.05259* (2015).

[23] Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations.