QBDT, a new boosting decision tree method with systematical uncertainties into training for High Energy Physics

Li-Gang Xia
Department of Physics, Warwick University, CV4 7AL, UK

Abstract

A new boosting decision tree (BDT) method, QBDT, is proposed for the classification problem in the field of high energy physics (HEP). In many HEP researches, great efforts are made to increase the signal significance with the presence of huge background and various systematical uncertainties. Why not develop a BDT method targeting the statistical significance (denoted by $Q$) directly? Indeed, the statistical significance plays a central role in this new method. It is used to split a node in building a tree and to be also the weight contributing to the BDT score. As the systematical uncertainties can be easily included in the significance calculation, this method is able to learn about reducing the effect of the systematical uncertainties via training. Taking the search of the rare radiative higgs decay in proton-proton collisions $pp \rightarrow h + X \rightarrow \gamma \tau^+ \tau^- + X$ as example, QBDT and the popular Gradient BDT (GradBDT) method are compared. QBDT is found to reduce the correlation between the signal strength and systematical uncertainty sources and thus to give a better statistical significance. The contribution to the signal strength uncertainty from the systematical uncertainty sources using the new method is 50-85 % of that using the GradBDT method.

PACS numbers: 29.85.Fj, 02.50.Sk
I. INTRODUCTION

In the field of high energy physics (HEP), many machine-learning (ML) methods are used for object identification [1–5] and to search for rare signals [6–15]. Especially, artificial neutral network (ANN) and boosting decision tree (BDT) are two of the most popular methods. The latter one is found to be more robust and stable in some analyses [16, 17]. On the other hand, various systematical uncertainties are involved in HEP (for examples, 13 main systematical uncertainties are considered in the analysis [6]) and they will inevitably affect the distribution of the input variables in training. However, there is no mature ML algorithm on training with systematical uncertainties as far as I know. This issue will be a big concern for either physicists or computer scientists in the near future as stated in Ref. [18]. In this paper, we propose a new BDT method, QBDT, to improve the sensitivity of probing rare signal with the existence of systematical uncertainties. Intuitively, the method pays more attention to the parameter space (spanned by the input variables) with higher signal purity and smaller background uncertainties at the same time. In HEP, the statistical significance is a quantity suitable to balance the different requirements on signal purity and background uncertainty. One rough idea is including systematical uncertainties into the significance (denoted by $Q$) calculation, and using $Q$ to build decision trees and to act as (or at least part of) the BDT output.

In this paper we only focus on the application of BDT methods in the classification problem in HEP. We will review current BDT methods in Sec. II and present the QBDT algorithm in Sec. III. Comparison of this method and the Gradient BDT will be performed in three successive examples described in Sec. IV B, Sec. IV C and the Appendix A respectively. The conclusions will be summarized in Sec. V.

II. REVIEW OF THE BDT METHODS

In the first place, let us introduce some necessary concepts. For any classification problem in HEP, we have two categories, signal and background. We can assign them different values called truth value $Y$. It is 1 for signal events and -1 for background events by convention. Some observables are selected as input in the BDT training. They could be four-momenta of the final particles or various mass variables, whose distribution is usually peaky for signal
while smooth for background. They are denoted by a vector \( \vec{x} \equiv (x_1, x_2, \cdots) \).

There are two popular BDT methods. One is the Adaptive BDT (AdaBDT) and the other is the Gradient BDT (GradBDT). Both methods use decision trees as the weak learners. Usually, hundreds of trees are used and each tree contains several terminal nodes (denoted by \( R \)) which are regions defined with the input variables \( \vec{x} \). The terminal nodes do not overlap and one event will fall into one of them. Each tree provides an output, which is the classification result, and a tree weight, which reflects the probability of this classification being correct. The final output, called as BDT score \( y(\vec{x}) \), is a combination of the outputs from all trees. They are boosting algorithms because the subsequent decision tree will pay more attention to the events which are misclassified by previous trees. Let us call it “tree update” for convenience throughout this paper.

For AdaBDT, the output of each tree is denoted by \( k \), which is 1 if an instance is classified as a signal event (namely if it falls in a terminal node dominated by the signal events) and -1 if classified as a background event (namely if it falls in a terminal node dominated by the background events). The confidence of the classification is measured by a function of the misclassification rate. It is \( \alpha = \frac{1}{2} \ln \frac{1-\epsilon}{\epsilon} \) where \( \epsilon \) is the misclassification rate for a tree. \( \alpha \) is big if the misclassification rate is low. The final score is a combination, \( y(\vec{x}) = \sum_{i=1}^{m} k_i(\vec{x})\alpha_i \) for a \( m \)-tree training. In building a tree, the split of a node into two is determined by the Gini index. It is defined as \( \sum_{\vec{x} \in N} p(1-p) \) where \( N \) denotes a node in a tree (keep in mind that we call it \( R \) instead if it is a terminal node) and \( p \) is the fraction of signal events (purity) in that node. The tree update in AdaBDT is realized by applying a big weight, \( e^{\omega} \), to those misclassified events and taking the weighted events as input in subsequent tree. It can be shown that AdaBDT is to minimize the loss function \( L(y) = \sum_{\vec{x}} e^{-y(\vec{x})Y(\vec{x})} \) \cite{19, 20}.

For GradBDT, we can use arbitrary differentiable loss function like \( L(y) = \sum_{\vec{x}} \frac{1}{2} (y(\vec{x}) - Y(\vec{x}))^2 \) or \( L(y) = \sum_{\vec{x}} \frac{1}{1+e^{y(\vec{x})Y(\vec{x})}} \). GradBDT is to minimize the loss function in a stage-wise way. Let us describe the process in the following example. We start with a random guess, like 0 for all events. The first tree is trained to fit the difference between 0 and the truth values. The output of the first tree is denoted by \( w_1(\vec{x}) \). This is done by maximizing \( L(y_0) - L(y_1) \) (thus the loss function will reduce after the first tree), where \( y_1(\vec{x}) = 0 + w_1(\vec{x}) = w_1(\vec{x}) \) for the first tree and \( y_0(\vec{x}) = 0 \) as the initial guess. It can be shown that \( w_1(\vec{x}) \) is roughly negative gradient \cite{21} of the loss function evaluated at \( y_0 = 0, \frac{\partial L(y)}{\partial y} |_{y=y_0} = 0 \). The second tree is trained to fit the difference between \( y_1(\vec{x}) \) and the truth values. The output of the
TABLE I. Comparison of different BDT methods. Here \( m \) denotes a \( m \)-tree training. \( \epsilon \) is the misidentification rate in AdaBDT. \( w \) is negative gradient of the loss function in GradBDT. \( p \) and \( Q \) is the purity and significance respectively of a terminal node in QBDT.

| Quantity        | AdaBDT value | GradBDT value | QBDT value |
|-----------------|--------------|---------------|------------|
| Input variables | \( \vec{x} = (x_1, x_2, \cdots) \) | same          | same       |
| True value \( Y \) | -1, 1        | same          | none       |
| Tree output \( k \) | \( -1, 1 \)  | negative gradient \( w \) | \( 2p - 1 \) |
| Tree weight \( \alpha \) | \( \frac{1}{2} \ln \frac{1-\epsilon}{\epsilon} \) | 1            | \( \sum_{j=1}^{J} Q_j \) |
| Tree update     | apply \( e^\alpha \) to wrong guess fit the residues | apply \( e^{\pm p \sum_{j} Q_j} \) |
| Node split     | Gini index reduction | loss function reduction significance \( Q \) increaseamnt |
| BDT score \( y_m \) | \( y_m = y_{m-1} + \alpha_m \hat{k}_m \) | \( y_m = y_{m-1} + w_m \) | \( y_m = y_{m-1} + (2p - 1) \sum_{j} Q_j \) |

Loss function \( L_m(\vec{x}, y_m) = \sum_{\vec{x}} e^{-Y(\vec{x})y_m(\vec{x})} \) any form statistical significance

second tree is \( w_2(\vec{x}) \), which is also roughly negative gradient of the loss function evaluated at \( y_1 \). So \( y_2(\vec{x}) = y_1(\vec{x}) + w_2(\vec{x}) = w_1(\vec{x}) + w_2(\vec{x}) \). This tree update can be repeated and the final score \( y_m(\vec{x}) = \sum_{i=1}^{m} w_i(\vec{x}) \) for a \( m \)-tree training. In GradBDT, the node split is naturally determined by minimizing the loss function. For example, a node \( N \) may give two disjoint daughter nodes \( N_L \) and \( N_R \) depending on \( x < x_0 \) or \( x > x_0 \). The split position \( x_0 \) is determined by maximizing \( \sum_{\vec{x} \in N} l(y(\vec{x})) - \sum_{\vec{x} \in N_L} l(y(\vec{x})) - \sum_{\vec{x} \in N_R} l(y(\vec{x})) \), where \( l(y(\vec{x})) \) is the loss function for one event. Recent development about GradBDT can be found in Ref. [22].

Table I compares the two methods as well as the new one as presented in next section. The first column lists the basic elements constituting a BDT algorithm. This table does not only show the features of different algorithms, but also helps to construct a new algorithm as long as all elements are defined reasonably.

III. QBDT ALGORITHM

The new BDT method is based on the concept of statistical significance in HEP. For a measurement with the number of signal and background events being \( s \) and \( b \) respectively,
the expected significance is the likelihood ratio. It is denoted by $Q$ (that is why the new method is named as “QBDT”) and defined as

$$Q \equiv 2 \ln \frac{P(s + b|s + b)}{P(s + b|b)} = 2 \left[ (s + b) \ln \left(1 + \frac{s}{b}\right) - s \right],$$

(1)

where $P(n|\nu) \equiv \frac{\nu^n}{n!} e^{-\nu}$ is the Poisson probability distribution function (PDF) with the expectation value $\nu$. If $s \ll b$ as in the case of searching for rare signals in HEP, we have $Q \approx \frac{s^2}{b}$.

Let us explain the various elements for this algorithm with the help of Table I.

- **Input variables ($\vec{x}$):** same as other BDT methods
- **True value:** it is irrelevant in the new algorithm
- **Tree output:** this depends upon the terminal node. It is $2p_j - 1$ if an event fall into the terminal node $R_j$. Here $p_j$ is the purity and defined as $p_j \equiv \frac{s_j}{s_j + b_j}$ with $s_j$ ($b_j$) being the number of signal (background) events in $R_j$. We use $2p - 1$ as the output so that it is 1 for a region with purely signal events and -1 for a region with purely background events. Purity is the right variable to maximize the significance in the following sense. Suppose we have two regions $R_1$ and $R_2$, the total significance is approximately $\frac{s_1^2}{b_1} + \frac{s_2^2}{b_2}$ which is greater than or equal to the significance of merging two regions, namely, $\frac{s_1^2}{b_1} + \frac{s_2^2}{b_2} \geq \frac{(s_1 + b_1)^2}{b_1 + b_2}$. The equal sign holds if and only if $\frac{s_1}{b_1} = \frac{s_2}{b_2}$, which means two regions have the same purity. As each tree output is part of the BDT score and the score will be used as the observable to extract the signal strength, we should assign different scores to events falling into regions with different purities. Otherwise, the statistical significance will not increase.
- **Tree weight:** it measures the probability of correct classification based on the present tree. It is defined as the total significance in all terminal nodes, $\sum_{j=1}^{J} Q_j$. This is consistent with our intuitive understanding. The trees with higher total significance should contribute more to the final score.
- **Tree update:** this is similar to AdaBDT. We apply a weight of $e^{p\sum_s Q_j}$ to the background events and a weight of $e^{-p\sum_s Q_j}$ to the signal events. It means that a background event happening to fall in a high-purity region or a signal event happening to fall in a low-purity region will be given a bigger weight in subsequent tree.
• Node split: this is similar to GradBDT. For any possible split about a variable \( x \) with the split position \( c \), let \( Q \) be the significance of a node before splitting and \( Q_L \) (\( Q_R \)) be the significance of the daughter node corresponding to \( x < c \) (\( x > c \)). We search for the variable \( x \) and the split position \( c \) to maximize the significance increasement \( \Delta Q(x, c) \equiv Q_L(x, c) + Q_R(x, c) - Q(x) \).

\[
x, c = \arg \max_{x, c} \Delta Q(x, c)
\]  

(2)

Figure 1 illustrates the growth of a tree. We starts with the whole set of events, denoted by “\( N_1: \text{no cut} \)” in the top. Performing the significance maximization procedure above gives the variable \( x_1 \) and the split position \( c_1 \). Then we have two nodes “\( N_2 : x_1 < c_1 \)” and “\( N_3 : x_1 > c_1 \)” For each of them, we repeat the significance maximization procedure to find the variable and split position, namely, \( x_2 \) and \( c_2 \) for \( N_2 \) with the significance increasement \( \Delta Q_2 \) and \( x_3 \) and \( c_3 \) for \( N_3 \) with the significance increasement \( \Delta Q_3 \). If \( \Delta Q_2 > \Delta Q_3 \), the next split will happen to \( N_2 \) and the daughter nodes are denoted by “\( N_4 : x_1 < c_1, x_2 < c_2 \)” and “\( N_5 : x_1 < c_1, x_2 > c_2 \)”. If we require a tree to have only 3 terminal nodes, the splitting stops and the terminal nodes are \( N_3 \), \( N_4 \) and \( N_5 \), which are renamed as \( R_1 \), \( R_2 \) and \( R_3 \) in Fig. 1. For each of them, we can calculate the purity \( p_j \) and the significance \( Q_j \) (\( j = 1, 2, 3 \)). \( 2p_j - 1 \) is the output of the tree while \( \sum_j Q_j \) is the tree weight. If we allow more terminal nodes, the next split will happen to one of \( N_4 \), \( N_5 \) and \( N_3 \), which gives the largest \( \Delta Q \). Note that for each split, we allow to use the same variable, like \( x_2 = x_1 \).

• BDT score: the definition is shown below

\[
y_m(\vec{x}) = \sum_{i=1}^{m} \sum_{j=1}^{J} \delta_{\vec{x}, R_j^{(i)}}(2p_j^{(i)} - 1) \sum_{k=1}^{J} Q_k^{(i)}
\]  

(3)

for \( m \) trees, where \( p_j^{(i)} \) is the purity of the terminal node \( R_j^{(i)} \) in the \( i \)-th tree; \( Q_k^{(i)} \) is the significance of the terminal node \( R_k^{(i)} \) in the \( i \)-th tree; and \( \delta_{\vec{x}, R_j^{(i)}} = 1 \) if \( \vec{x} \in R_j^{(i)} \) and 0 otherwise.

The algorithm of QBDT should be clear after all elements are defined. Here is the workflow.

1. Reweight both signal and background samples so that each gives 1 event. This reweighting is not necessary, but very helpful in practice. It is because the significance is tiny
for a rare signal (but this is usually where BDT is used) and the boosting is very slow. Many more trees would be needed without reweighting.

2. According to the node split procedure, we can build the first tree. Supposing only $J$ terminal nodes are allowed, we can calculate the purity $p_j^{(1)}$ and the significance $Q_j^{(1)}$ for each terminal node $R_j^{(1)}$. The tree weight is the total significance, $\sum_{j=1}^{J} Q_j^{(1)}$. If an event falls into $R_x^{(1)}$, then the output is $(2p_x^{(1)} - 1) \sum_{j=1}^{J} Q_j^{(1)}$.

3. Before building the second tree, we apply a weight of $\exp(p_j^{(1)} \sum_{j=1}^{J} Q_j^{(1)})$ to every background event (supposing it falls into $R_{j_b}^{(1)}$) and similarly apply a weight of $\exp(-p_j^{(1)} \sum_{j=1}^{J} Q_j^{(1)})$ to every signal event (supposing it falls into $R_{j_s}^{(1)}$). Reweight both signal and background samples to 1 event again.

4. Build the second tree based on the reweighed samples using the same node split procedure. We have $J$ terminal nodes, $R_j^{(2)}$ and the corresponding purity $p_j^{(2)}$ and significance $Q_j^{(2)}$. If an event falls into $R_x^{(2)}$, then the output is $(2p_x^{(2)} - 1) \sum_{j=1}^{J} Q_j^{(2)}$.

5. We can repeat the steps 2–4 to build many more trees. The final output (BDT score) is given by Eq. 3.

---

FIG. 1. Illustration diagram for the node splitting in QBDT method.
The key feature of QBDT is that it is able to consider systematical uncertainties into training. In HEP, various factors will affect the input variables. For example, if the muon transverse momentum is one input variable, we have to evaluate the uncertainty due to muon momentum calibration. Furthermore, the identification efficiency and isolation efficiency of the muon lepton may be dependent upon the transverse momentum. Thus the efficiency uncertainties will also affect the transverse momentum distribution. To include these uncertainties into training, the definition of $Q$ can be extended to be

$$Q = 2 \left[ \frac{(s + b)(b + \sigma_b^2)}{b^2 + (s + b)\sigma_b^2} - \frac{b^2}{\sigma_b^2} \ln \left[ 1 + \frac{s\sigma_b^2}{b(b + \sigma_b^2)} \right] \right], \quad (4)$$

where $\sigma_b$ is the systematical uncertainty on the background number of events. If $s \ll b$ and $s\sigma_b \ll b$, $Q \approx s^2 b + \sigma_b^2$. The purity in the tree output will be extended to be $\frac{s}{s + b + \sigma_b^2}$ accordingly.

In the case of multiple systematical uncertainties, we may need to consider the possible correlations. The total background uncertainty will be

$$\sigma_b^2 = \sum_{i=1}^{N_{\text{sys}}} \left( \frac{\partial b}{\partial \theta_i} \sigma_{\theta_i} \right)^2 + \sum_{i \neq j} \frac{\partial b}{\partial \theta_i} \frac{\partial b}{\partial \theta_j} \sigma_{\theta_i} \sigma_{\theta_j} \rho_{ij}, \quad (5)$$

where each $\theta_i$ is a nuisance parameter associated with one systematical uncertainty source; $\sigma_{\theta_i}$ is the size of the uncertainty which is usually estimated by other independent measurements; $\frac{\partial b}{\partial \theta_i} \sigma_{\theta_i}$ is the size of the background number uncertainty due to this systematical uncertainty source; $\rho_{ij}$ is the correlation coefficient between two systematical uncertainty sources. In most of HEP cases, we perform a likelihood fit to the distribution of the final BDT score. The exact uncertainty size and correlation can only be obtained after the fit.

Fortunately, explicit expression of the uncertainty on the signal strength, $\sigma_s$, is derived in Ref. [24] assuming small correlations in the maximum likelihood estimation method. Noting that $\sigma_s^2 = b + \sigma_b^2 + \delta^2$ (Eq.(6) in Ref. [24]) where $\delta$ is the Monte Carlo (MC) statistical uncertainty (it exists if any background component is estimated by MC simulation) and using the expression on $\sigma_s$ (Eq.(26) in Ref. [24]), we have

$$\sigma_b^2 = b \left( \sum_{j=1}^{M} \frac{(\Delta^j)^2}{b + (\Delta^j)^2} - \sum_{i \neq j} \frac{(\Delta^i)^2(\Delta^j)^2}{(b + (\Delta^i)^2)(b + (\Delta^j)^2)} \right). \quad (6)$$

Here $M$ is the number of systematical uncertainty sources; $\Delta^i$ is the uncertainty on the background number of events due to the $i$-th systematical uncertainty source; and the second term in the big brackets represents the contribution due to the correlation between different
systematical uncertainty sources. It deserves mentioning that we can consider MC statistical uncertainty in training as well by replacing $\sigma_b^2$ by $\sigma_b^2 + \delta^2$. In practice, $\sigma_b^2$ is found to be negative sometimes. This means the correlation term has very large negative contribution. But small correlation is assumed in the derivation in Ref. [24] and the expression is not valid. We set it to 0 in this case.

In the following sections, we will compare the performance of GradBDT and QBDT based on an HEP example. It is worth mentioning that we do not use the BDT score directly, but a function of it. The GradBDT algorithm in the package TMVA [25] uses tanh($y$) as the final score. The advantage of this function is that it maps $(-\infty, +\infty)$ to a bounded region $(-1, 1)$. We will use a similar form tanh($cy$), where $c$ is a constant and is optimized to be 0.65 for GradBDT (0.65 is better than 1 used in the TMVA package at least for the example presented below) while 0.25 for QBDT.

IV. ONE EXAMPLE TO COMPARE GRADBDT AND QBDT

A. Description of the example

Here we present an example. It is to search for the rare radiative decay of the higgs boson in the proton-proton collisions at the center-of-mass energy $\sqrt{s} = 13$ TeV assuming a dataset of 80 fb$^{-1}$. The signal process is $pp \rightarrow h + X \rightarrow \gamma\tau\tau + X$, while the background process is $pp \rightarrow \gamma\tau\tau + X$ with the higgs contribution vetoed. We further require one tau decays leptonically, $\tau^- \rightarrow l^-\nu_\tau\bar{\nu_\tau} + c.c.$ (here $l$ is $e/\mu$ and c.c. means the charge-conjugated channel), and the other decays hadronically, $\tau^- \rightarrow h^-\nu_\tau + c.c.$ The events with one charged lepton candidate $e/\mu$, one hadronic tau candidate $\tau_{\text{had}}$ and one photon candidate $\gamma$ are selected. The transverse momentum, $p_T$, is required to be greater than 20 GeV for the lepton and tau candidates and 10 GeV for the photon candidate. The missing transverse energy $E_T^{\text{miss}}$ is defined as the negative vector sum of transverse momenta of all other visible objects. It reflects the information of the neutrinos. Eight variables are used in the training. Table II summarizes the variables. They are the transverse momentum/energy of the final objects ($p_T(l)$, $p_T(\tau_{\text{had}})$, $p_T(\gamma)$ and $E_T^{\text{miss}}$) and four mass variables constructed from the
TABLE II. Definition of the eight variables used in the training.

| Variable          | Explanation                                           |
|-------------------|-------------------------------------------------------|
| $p_T(l)$          | $p_T$ of the charged lepton $e/\mu$                   |
| $p_T(\tau_{\text{had}})$ | $p_T$ of the hadronic tau                         |
| $p_T(\gamma)$    | $p_T$ of the photon                                   |
| $E_T^{\text{miss}}$ | missing transverse momentum                          |
| $m(l\tau_{\text{had}})$ | invariance mass of the lepton $e/\mu$ and the hadronic tau |
| $m(l\tau_{\text{had}}\nu)$ | invariance mass of the lepton $e/\mu$, hadronic tau and neutrinos |
| $m(l\tau_{\text{had}}\gamma)$ | invariance mass of the lepton $e/\mu$, hadronic tau and photon |
| $m(l\tau_{\text{had}}\gamma\nu)$ | invariance mass of the lepton $e/\mu$, hadronic tau, photon and neutrinos |

FIG. 2. The distribution of eight variables used in the training. In top row, it is $p_T(l)$, $p_T(\tau_{\text{had}})$, $p_T(\gamma)$ and $E_T^{\text{miss}}$ from left to right. In the bottom row, it is $m(l\tau_{\text{had}})$, $m(l\tau_{\text{had}}\nu)$, $m(l\tau_{\text{had}}\gamma)$ and $m(l\tau_{\text{had}}\gamma\nu)$ from left to right. The blue histogram represents the signal while the red histogram represents the background. All distributions are renormalized to unit area.

For both GradBDT and QBDT, half of the samples are used for training and the other
| Systematical Source | Uncertainty Size | Nuisance Parameter |
|---------------------|------------------|-------------------|
| $\tau_{\text{had}} p_T$ calibration | $\Delta p_T(\tau_{\text{had}}) = \pm 2\% p_T(\tau_{\text{had}})$ [26] | $\theta(p_T)$ |
| $\tau_{\text{had}}$ ID efficiency | $\pm 5\%$ [26] | $\theta(\tau_{\text{ID}})$ |
| $E_T^{\text{miss}}$ resolution | $\pm 10\%$ [27] | $\theta(E_T^{\text{miss}})$ |
| Lepton $p_T$ calibration | $\Delta p_T(l) = \pm 0.1\% p_T(l)$ [28] | $\theta(l_T)$ |
| Lepton ID efficiency | $\pm 0.2\%$ for $p_T(l) < 100$ GeV and $\pm 0.5\%$ otherwise [28] | $\theta(l_{\text{ID}})$ |
| Photon $p_T$ calibration | $\Delta p_T(\gamma) = \pm 0.3\% p_T(\gamma)$ [29] | $\theta(\gamma_T)$ |
| Photon ID efficiency | $\pm 2\%$ for $p_T(\gamma) < 40$ GeV and $\pm 1\%$ otherwise [29] | $\theta(\gamma_{\text{ID}})$ |

Half used for testing. For GradBDT, we are using 1000 trees, learning rate of 0.1 and maximal depth of 3. These numbers are optimized. For QBDT, we are using 100–200 trees and 7 terminal nodes for each tree. Increasing either the number of trees or the number of terminal nodes leads to little improvement. We have four setups for QBDT training:

- **QBDT0**: training with nominal samples, i.e., no systematical uncertainty considered in the training.
- **QBDTX**: $X = 1, 3, 7$ is an integer denoting the number of systematical uncertainty sources considered in the training.

Different QBDT setups as well as the GradBDT are compared in three cases, namely, the case of one systematical uncertainty source in Sec. IV.B, three sources in Sec. IV.C and seven sources in the Appendix A. Table III summarizes all systematical uncertainties considered in this paper. Here the uncertainty sizes are selected according to the performance [26–29] of the ATLAS detector at the Large Hadron Collider (LHC).

To get an overall impression about the performance of different BDT setups, we can look at the so-called ROC curves. A ROC curve describes the background rejection rate as a function of the signal acceptance if we apply any cut on the BDT score. Figure 3 shows the ROC curves for all BDT setups. The dashed curves are from the training samples while the solid curves are from the testing samples. The gap between the training and testing results reflects the degree of overtraining, which is unavoidable. We can see that all BDT setups have very similar performances. All comparisons will be based on the testing samples in the
FIG. 3. Comparison of ROC curves from the testing samples.

following sections.

B. Case I: one systematical uncertainty source

In the first case, we introduce the systematical uncertainty on the transverse momentum calibration of the hadronic tau. The corresponding nuisance parameter is denoted by $\theta(p_T(\tau_{\text{had}}))$. The uncertainty size is $\Delta p_T(\tau_{\text{had}}) = \pm 2\% p_T(\tau_{\text{had}})$. Figure 4 shows the envelope plots to illustrate the effect of the systematical uncertainty on the BDT score distribution. But it seems not easy to draw conclusions on how different algorithms treat with the systematical uncertainty from these plots.

Both the statistics-only (stat.-only) fit, not including any systematical uncertainty, and the full fit with this systematical uncertainty are performed. The expected signal significance and the fitted uncertainty of the nuisance parameter are shown in Table IV. We can see that the stat.-only significance is very close to each other. If the systematical uncertainty is included, QBDT0 has a similar significance as GradBDT, but QBDT1, the one considering
TABLE IV. Expected significance and post-fit uncertainty of the nuisance parameters.

| Significance | GradBDT | QBDT0 | QBDT1 |
|--------------|---------|-------|-------|
| Stat.-only fit | 0.89    | 0.90  | 0.87  |
| Full fit      | 0.72    | 0.73  | 0.78  |
| $\sigma(\tau p_T)$ | 0.47    | 0.32  | 0.39  |

FIG. 4. Envelope plot for the systematical uncertainty of $p_T(\tau_{\text{had}})$. Left: GradBDT, middle: QBDT0, and right: QBDT1.

the systematical uncertainty in training, gives the highest significance. However, we have to notice that the post-fit uncertainty of $\theta(\tau p_T)$ is smaller in QBDT1 than in GradBDT. This may lead to a concern that the significance is better just because the nuisance parameter is more constrained in QBDT1. We can look at more cases (next section and Appendix A) to have a better understanding.

One interesting and important feature happens to the correlation matrix shown in Fig. 5. The correlation coefficient between the signal strength (parameter of interest or POI) and the systematical uncertainty source $\theta(\tau p_T)$ is close between QBDT0 and GradBDT, but reduced in QBDT1. This is a good sign that QBDT1 is really learning about how to reduce the effect of the systematical uncertainty by decreasing the correlation between the signal strength and this systematical uncertainty source [24]. As the correlation reduces, it is not surprising that the statistical significance is better.
In this case, two more systematical uncertainty sources are introduced. One is the $\tau_{\text{had}}$ identification (ID) efficiency uncertainty. It is 5% independent upon the transverse momentum. The other is $E_T^{\text{miss}}$ resolution uncertainty of 10%. The envelope plots for these systematical uncertainties are shown in Fig. 6, 7 and 8, respectively. Note that for the tau $p_T$ calibration uncertainty, the envelope plots from GradBDT and BDT0 in Fig. 6 are the same as in Case I (namely, those in Fig. 4).

The fit results are summarized in Table V. We can draw the same conclusion as in Case I that GradBDT and QBDT0 have similar performance, and QBDT3 is the best. In Case I, we see the unique nuisance parameter $\theta(\tau p_T)$ is a bit constrained in QBDT1 than in GradBDT. But in this case, we do not see all nuisance parameters are constrained in any specific BDT training. Therefore, we do not think the significance improvement is purely due to the over-constraining. The correlation matrices are compared in Fig. 9. We can also see that the correlation between the signal strength and other nuisance parameters is reduced in QBDT3. In Appendix A 7 systematical uncertainty sources are introduced in total. The similar correlation reduction is seen and QBDT7 has a better performance than the GradBDT and QBDT0. Therefore, this correlation reduction is indeed an evidence that the QBDT method works with systematical uncertainties into training.

To compare QBDT and GradBDT quantitatively, we propose to measure the performance using the contribution to the uncertainty of the signal strength from the systematical uncertainty sources. Let $\Delta \mu_0$ denote the uncertainty of the signal strength from the stat-only fit.
FIG. 6. Envelope plot for the systematical uncertainty of tau energy scale. Left: GradBDT, middle: QBDT0, and right: QBDT3.

FIG. 7. Envelope plot for the systematical uncertainty of tau ID efficiency. Left: GradBDT, middle: QBDT0, and right: QBDT3.

FIG. 8. Envelope plot for the systematical uncertainty of missing transverse energy resolution. Left: GradBDT, middle: QBDT0, and right: QBDT3.
TABLE V. Expected significance and post-fit uncertainty of the nuisance parameters.

| Significance                  | GradBDT | QBDT0 | QBDT3 |
|-------------------------------|---------|-------|-------|
| Stat.-only fit                | 0.89    | 0.90  | 0.89  |
| Full fit                      | 0.71    | 0.70  | 0.75  |
| \( \sigma_{\theta}(\tau_{pT}) \) | 0.54    | 0.54  | 0.61  |
| \( \sigma_{\theta}(\tau_{ID}) \) | 0.28    | 0.46  | 0.36  |
| \( \sigma_{\theta}(E_{T}^{\text{miss}}) \) | 0.87    | 0.87  | 0.80  |

FIG. 9. Correlation matrix obtained from the fits. Left: GradBDT, middle: QBDT0, and right: QBDT3.

\( \Delta \mu_1 \) denote that from the full fit. The contribution from the systematical uncertainties can be represented by \( \Delta \mu \) defined as \( \Delta \mu \equiv \sqrt{\Delta \mu_1^2 - \Delta \mu_0^2} \) and the numerical results are summarized in Table VI for the two cases presented above and the case in Appendix A. We can see that the \( \Delta \mu \) for QBDT0 is 50–85 % of that for GradBDT. Two things should be emphasized from these numbers. One is about the systematical uncertainty on the \( \tau \) ID efficiency. It only affects the background normalization and it should be impossible to reduce its effect by changing the score distribution, or in other word, by QBDT. This is why we do not see much improvement in Case II. The other is about \( \Delta \mu \) variation among different cases. For GradBDT, \( \Delta \mu \) is greater with more systematical uncertainty sources. But this behavior is not seen in QBDT. This is because we perform QBDT training specifically for each set of systematical uncertainty sources and the correlation matrix is evolving in different cases.
TABLE VI. Contributions to the uncertainty of the signal strength from the systematical uncertainties. The rightmost column is the ratio of the contribution in QBDTX and that in GradBDT.

| $\Delta \mu$ | $N_{\text{syst}}$ | GradBDT | QBDTX | $\frac{\Delta \mu_{\text{QBDTX}}}{\Delta \mu_{\text{GradBDT}}}$ |
|------------|----------------|---------|-------|----------------|
| Case I     | 1              | 0.79    | 0.56  | 71%            |
| Case II    | 3              | 0.82    | 0.70  | 85%            |
| Case III   | 7              | 0.98    | 0.49  | 50%            |

V. SUMMARY

In summary, a new boosting decision tree method, QBDT, is proposed to reduce the effect of the systematical uncertainties for the classification problem in high energy physics. The concept of statistical significance in HEP plays two important roles in this method. One is that the node split in building a tree is determined by maximizing the significance. This is similar to the split criteria in GradBDT, where the node split is determined by minimizing the loss function. The other is to act as the tree weight, which is part of the BDT score, and used to update the event weight for subsequent tree. This point is similar to the tree update rule in AdaBDT, where heavier weight is applied to the misclassified events. As the various systematical uncertainties can be conveniently included in the calculation of the statistical significance, QBDT is able to perform training with the systematical uncertainties. Taking a typical HEP search for the rare radiative higgs decay $pp \rightarrow h+X \rightarrow \gamma\tau^+\tau^-+X$ as example, it turns out that the correlation between the signal strength and the systematical uncertainty sources is reduced in QBDT. Therefore, QBDT gives a better signal significance. Based on the example, the contribution to the signal strength uncertainty from the systematical uncertainty sources in the QBDT method is 50–85% of that in the GradBDT method.

VI. ACKNOWLEDGEMENT

I would like to thank Fang Dai for encouraging words.
### Appendix A: Case III: seven systematical uncertainty sources

Here in this case 7 systematical uncertainty sources are considered. The definitions have been already summarized in Table III. Except those introduced in the main text, we have the lepton $p_T$ calibration uncertainty, lepton ID efficiency uncertainty, photon $p_T$ calibration uncertainty and photon ID efficiency uncertainty.

The envelope plots are compared between GradBDT and QBDT7 in Fig. 10 and 11. The fit results are summarized in Table VII. The same as in Case I (Sec. IV B) and Case II (Sec. IV C), the correlation between signal and the other nuisance parameters is reduced and QBDT7 gives the best statistical significance. This case is to show that QBDT can take effect with the presence of many systematical uncertainty sources.

---

**TABLE VII.** Expected significance and post-fit uncertainty of the nuisance parameters.

| Significance        | GradBDT | QBDT0 | QBDT7 |
|---------------------|---------|-------|-------|
| Stat.-only fit      | 0.89    | 0.90  | 0.91  |
| Full fit            | 0.66    | 0.69  | 0.83  |
| $\sigma_\theta(\tau p_T)$ | 0.59    | 0.56  | 0.59  |
| $\sigma_\theta(\tau ID)$ | 0.44    | 0.54  | 0.49  |
| $\sigma_\theta(E_{\text{miss}})$ | 0.89    | 0.88  | 0.77  |
| $\sigma_\theta(l p_T)$ | 0.92    | 0.95  | 0.91  |
| $\sigma_\theta(l ID)$ | 0.99    | 0.99  | 0.99  |
| $\sigma_\theta(\gamma p_T)$ | 0.85    | 0.94  | 0.86  |
| $\sigma_\theta(\gamma ID)$ | 0.93    | 0.93  | 0.93  |

---

[1] ATLAS Collaboration, ATLAS-CONF-2017-029, [http://cds.cern.ch/record/2261772](http://cds.cern.ch/record/2261772)
[2] ATLAS Collaboration, Eur.Phys.J. C76 (2016) no.12, 666, arXiv: 1606.01813
[3] ATLAS Collaboration, JINST 11 (2016) no.04, P04008, arXiv: 1512.01094
[4] CMS Collaboration, CMS-PAS-TAU-16-002, [http://cds.cern.ch/record/2196972](http://cds.cern.ch/record/2196972)
[5] CMS Collaboration, JINST 13 (2018) no.05, P05011, arXiv: 1712.07158
[6] ATLAS Collaboration, CERN-EP-2018-212, arXiv:1808.09054
FIG. 10. Envelope plot for the systematical uncertainties. Left column: GradBDT, right column: QBDT7. From top to bottom, the systematical uncertainty source is tau $p_T$ calibration, tau ID efficiency and $E_{T}^{miss}$ resolution.

[7] ATLAS Collaboration, CERN-EP-2018-168, arXiv:1808.03599
[8] ATLAS Collaboration, CERN-EP-2018-164, arXiv:1808.00336
[9] ATLAS Collaboration, Phys.Rev. D98 (2018) no.5, 052003, arXiv:1807.08639
[10] ATLAS Collaboration, JHEP 1810 (2018) 031, arXiv:1806.07355
FIG. 11. Envelope plot for the systematical uncertainties. Left column: GradBDT, right column: QBDT7. From top to bottom, the systematical uncertainty source is lepton $p_T$ calibration, lepton ID efficiency, photon $p_T$ calibration and photon ID efficiency.
FIG. 12. Correlation matrix obtained from the fits. Left: GradBDT, middle: QBDT0 and right: QBDT7.

[11] CMS Collaboration, CMS-HIG-17-019, CERN-EP-2018-165, arXiv:1807.06325

[12] CMS Collaboration, CMS-HIG-16-042, CERN-EP-2018-141, FERMILAB-PUB-18-352-CMS, arXiv:1806.05246

[13] CMS Collaboration, CMS-HIG-17-008, CERN-EP-2017-343, arXiv:1806.00408

[14] CMS Collaboration, CMS-HIG-16-040, CERN-EP-2018-060, arXiv:1804.02716

[15] CMS Collaboration, JHEP 1808 (2018) 066, arXiv:1803.05485

[16] B. P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor, Nucl.Instrum.Meth. A543 (2005) no.2-3, 577-584, arXiv:physics/0408124

[17] H.-J. Yang, B. P. Roe, J. Zhu, Nucl.Instrum.Meth. A574 (2007) 342-349

[18] K. Albertsson et al., FERMILAB-PUB-18-318-CD-DI-PPD, arXiv: 1807.02876

[19] J. Friedman, T. Hastie, and R. Tibshirani, Ann. Statist. 28 (2000) 337-407

[20] R. Rojas, http://www.inf.fu-berlin.de/inst/ag-ki/adaboost4.pdf

[21] J. Friedman, Ann. Statist. 29 (2001) 1189-1232

[22] T. Chen and C. Guestrin, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794, arXiv:1603.02754

[23] G. Cowan, https://www.pp.rhul.ac.uk/~cowan/stat/medsig/medsigNote.pdf

[24] L. G. Xia, arXiv:1805.03961

[25] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, PoS ACAT 040 (2007), arXiv:physics/0703039

[26] ATLAS Collaboration, ATLAS-CONF-2017-029, http://cdsweb.cern.ch/record/2261772

[27] ATLAS Collaboration, CERN-EP-2017-274, arXiv: 1802.08168

[28] ATLAS Collaboration, Eur. Phys. J. C76 (2016) 292
[29] ATLAS Collaboration, ATL-PHYS-PUB-2017-022, https://cds.cern.ch/record/2298955; CERN-EP-2018-216, arXiv: 1810.05087