X. Supplementary Materials

A. Experiment Setup

Data Sets. We use four datasets to compare the performance of all methods:

(1) **MNIST**[5]: The MNIST is an image dataset of handwritten digits. It contians 10 classes and 784 features;

(2) **20 Newsgroups**[6]: The dataset collates approximately 20,000 newsgroup documents partitioned across 20 different newsgroups. The Word2vec is used to prepossess text data.

(3) **CIFAR-10**[7]: It is a standard classification dataset consisting of 32×32 color images belonging to 10 different object classes.

(4) **VGGFace**[8]: The dataset consists of the crawled images of celebrities on the Web. There are 2622 celebrities in the dataset.

Evaluation Metrics. We use **F-measure** to measure the performance. This measure produces a combined effect of precision (P) and recall (R) of the auditing performance,

$$F\text{-}measure = \frac{2 * P * R}{P + R}.$$

F-measure = 1 if the method identifies all training data with no false positives.

We also employ **AUC**[9] ("Area under the ROC Curve") to access performance. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots True Positive Rate (TPR) and False Positive Rate (FPR).

B. Results on synthetic data

1) SVM-based model: SVM is a discriminative classifier which classifies new data points by calculating an optimal separating hyperplane[10]. In two dimensional space this hyperplane is a line dividing a plane into two parts each defining a class for data points within it. In this paper, we illustrate with a least squares SVM classifier[11]. Given a set of instance-label pairs (x_i, y_i), $i = 1, \cdots, l$, $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$, it solves the following optimization problem: $\min_w \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi^2$, subject to the equality constraints: $y_i(wx_i + b) = 1 - \xi_i$, where $C > 0$ is a penalty parameter. We set the partial derivatives of x on the cost:

$$\frac{\partial J(w)}{\partial x} = 2[1 - wx)]w.$$

Then, we can calculate η by Eqn.(5).

2) Tree-based model: We discuss random forest model with completely-random trees as target model in this part[12]. random forest model is usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. Each tree in the classifications takes input from samples in the initial dataset. Features are then randomly selected, which are used in growing the tree at each node. Every tree in the forest should not be pruned until the end of the exercise when the prediction is reached decisively.

Note that because the random forest is a non-linear form, it is not easy to use the gradient to calculate η. In this paper, we attempt to set Gaussian noise perturbation as η to observe this result. Figure 9 shows the complete illustrations.

3) Neural network-baesd: A neural network (NN) is a technique that uses a hierarchical composition of n parametric functions to model an input x. Each function f_i for $i \in 1, \cdots, n$ is modeled using a layer of neurons, which are elementary computing units applying an activation function to the previous layer's weighted representation of the input to generate a new representation. Each layer is parameterized by a weight vector θ_i impacting each neuron's activation. Such weights hold the knowledge of a NN model and are evaluated during its training phase, as detailed below. Thus, a NN defines and computes:

$$M(x) = f_n(\theta_n, f_{n-1}(\theta_{n-1}, \cdots f_2(\theta_2, f_1(\theta_1, x)))) \quad (11)$$

At each layer: $y_j = f(\sum_{i=1}^{3} w_{ij} x_i + b)$, where W_{ij}, x_i and y_j are the weights, input and output respectively. We show a two-linear-layer NN example as follow:

$$\begin{aligned} M(x) &= w_2(w_1 x + b_1) + b_2 \\ &= w_2 w_1 x + w_2 b_1 + b_2 \end{aligned} \quad (12)$$

The above equation can be view as an SVM-based model. We, therefore, apply the same way to calculate η. Figure 10 shows NN-based model results.

Figure 11 shows the results of multiplicative implementation on synthetic data data, the experiment setup is the same as Section VII-B. (a)-(c) are results under black-box condition.

[5]http://yann.lecun.com/exdb/mnist/

[6]http://qwone.com/ jason/20Newsgroups/

[7]https://www.cs.toronto.edu/ kriz/cifar.html

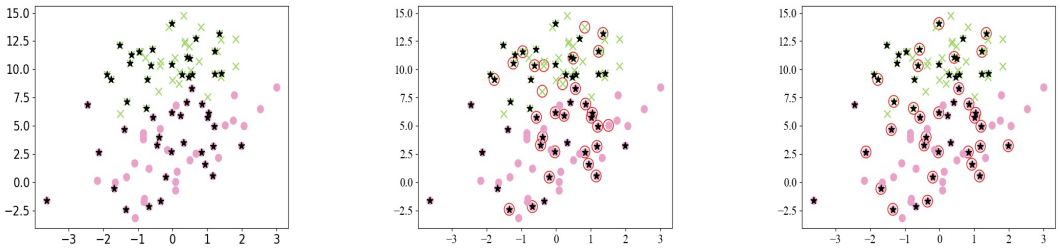
[8]https://www.robots.ox.ac.uk/ vgg/data/vgg_face/

[9]https://en.wikipedia.org/wiki/Receiver_operating_characteristic

[10]Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. Machine Learning 20, 3 (1995), 273–297.

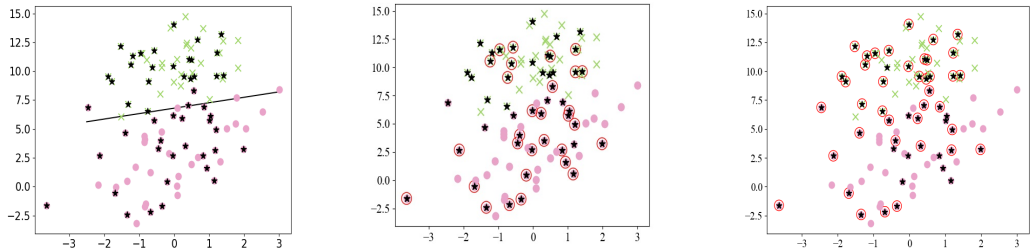
[11]Johan A. K. Suykens and Joos Vandewalle. 1999. Least Squares Support VectorMachine Classifiers.Neural Processing Letters 9, 3 (1999), 293–300.

[12]Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, et. al. 2008. Top 10 algorithms in data mining. Knowledge and Information Systems 14, 1 (2008), 1–37.



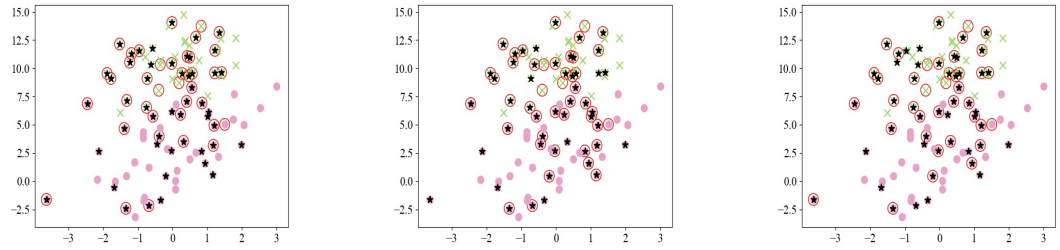
(a) Auditing data. (b) Results on white-box condition. (c) Results on black-box condition.

Fig. 9. An illustration on the Tree-based model. Training data are marked by black star marks training data. Red circle is the auditing results. (b) is result under white-box condition, (c) is result under black-box condition.



(a) Auditing data and hyperplane. (b) Results on white-box condition. (c) Results on black-box condition.

Fig. 10. An illustration on the NN-based model. Training data are marked by black star marks training data. Red circle is the auditing results. (b) is result under white-box condition, (c) is result on black-box condition.



(a) Results on SVM-based model. (b) Results on Tree-based model. (c) Results on NN-based model.

Fig. 11. An illustration on the results of multiplicative implementation. Black star marks training data. red circle marks auditing results. (a) is SVM-based model result, (b) is tree-based model result, (c) is NN-based model result.