On the (In)feasibility of ML Backdoor Detection as an Hypothesis Testing Problem

Georg Pichler
TU Wien

Marco Romanelli
New York University

Divya Prakash Manivannan
New York University

Prashanth Krishnamurthy
New York University

Farshad Khorrami
New York University

Siddharth Garg
New York University

Abstract

We introduce a formal statistical definition for the problem of backdoor detection in machine learning systems and use it to analyze the feasibility of such problems, providing evidence for the utility and applicability of our definition. The main contributions of this work are an impossibility result and an achievability result for backdoor detection. We show a no-free-lunch theorem, proving that universal (adversary-unaware) backdoor detection is impossible, except for very small alphabet sizes. Thus, we argue, that backdoor detection methods need to be either explicitly, or implicitly adversary-aware. However, our work does not imply that backdoor detection cannot work in specific scenarios, as evidenced by successful backdoor detection methods in the scientific literature. Furthermore, we connect our definition to the probably approximately correct (PAC) learnability of the out-of-distribution detection problem.

1 INTRODUCTION

The adoption of modern Machine Learning (ML) methods in a range of real-world tasks including navigation (Chen et al., 2022; Wang et al., 2022), medical diagnosis (Varoquaux and Cheplygina, 2022; Tchango et al., 2022), and system control (Zhang et al., 2023) has grown dramatically. However, safe and trustworthy ML systems remain elusive (Ilyas et al., 2019; Wu et al., 2022), for reasons including poor interpretability (Burkart and Huber, 2021; Roscher et al., 2020), test time adversarial inputs (Goodfellow et al., 2015) and, relevant to this paper, *training time poisoning and backdooring attacks* (Gu et al., 2019). As the scale, complexity and training data requirements of modern deep neural network architectures has grown, few can afford to train models from scratch. Many users therefore download and fine-tune pre-trained models, or deploy them as is. Consequently, purposefully implanted backdoors in pre-trained ML models pose a key security risk for future ML deployments.

In the classic backdoor threat model, a malicious actor trains a backdoored ML model by altering its training data. During inference, for certain, *backdoored* inputs, modified in an attacker chosen way, the model then provides erroneous predictions. For example, (Gu et al., 2019) demonstrate a backdoor in a traffic sign detector that misclassifies stop signs as speed limit signs when stop signs are modified with stickers or sticky notes. Here the sticker or sticky note serve as a trigger, misleading the model into making incorrect decisions. While there are many ways such a backdoor could be embedded into a model, prior work shows that altering even a small fraction of training data yields models with stealthy and effective backdoors (Qi et al., 2023).

In light of these attacks, substantial efforts have been devoted to *backdoor defenses* with the goal of identifying such backdoor attacks. To detect a backdoor,

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

the model user (i.e., the *defender*) has access to a, typically small, validation dataset of clean inputs. In the Model Backdoor Detection (MBD) problem (Chen et al., 2019, Sec. 4.2), (Wang et al., 2019; Shen et al., 2021), the defender wishes to detect if the model itself contains a backdoor. In the Sample Backdoor Detection (SBD) problem (Liu et al., 2023; Ma et al., 2023; Fu et al., 2023), the defender wants to detect if a specific test input is backdoored or not, assuming that models deployed in the field might be backdoored. We note that our backdooring threat model is part of the larger body of work on data poisoning threats. The latter encompasses all scenarios where a model trained on–partially–poisoned data is negatively impacted in some way, possibly, but necessarily, by implanting a backdoor. With the growing use of large pretrained foundation models, the backdooring threat (where the defender receives a model, not training data) is of increasing relevance.

Unfortunately, despite several years of research, the field is still *plagued by a cat-and-mouse game between attackers and defenders*. Certifiable defenses against backdooring attacks have remained elusive, and despite some recent progress in this direction for data poisoning attacks, those results don't translate to our setting as discussed in Section 3. In fact, despite the large body of work in the area, backdoor detection has *not been formally studied* from first principles.

Here, we undertake the first *formal* exposition of the ML backdooring problem for both the MBD and SBD settings. Although it has been observed that backdoor detection can work in specific scenarios, we are interested in the feasibility of detecting arbitrary, unknown backdoors. We thus ask the following questions. *First*, what are fundamental bounds lower bounds on backdoor detection—in fact, is backdoor detection even feasible, and if so, under what assumptions? *Second*, how is backdoor detection related to other statistical problems in ML? Recent work has leveraged Out-Of-Distribution (OOD) detection methods for backdoor detection; can this relationship be formalized? *Third*, how are MBD and SBD related? Both are separately addressed in literature, but their relationship has not been explicated. And *finally*, what are the implications for future progress in building practical backdoor defenses? We provide answers to all four questions using theoretical results and a "toy" example.

Contributions. In this paper, we present the first precise statistical formulation of the MBD and SBD problems (Section 2.1). This formulation enables several new insights on backdoor detection.

1. *Relationship to well-known statistical problems:* Our formulation unifies MBD, SBD and even OOD detection within a common framework and we reduce these problems to standard statistical hypothesis testing problems.
2. *Infeasibility:* Leveraging these reductions, we conclude that under realistic assumptions, universal (adversary-unaware) backdoor detection is not possible for an infinite alphabet of the training data.
3. *Bound for finite alphabet size:* For a finite data alphabet, we provide a bound on the achievable error probability given a fixed training set size. These bounds are evaluated for commonly used datasets in ML, showing that universal backdoor detection is only achievable for very small alphabets.
4. *Connections to Probably Asymptotically Correct (PAC) learning theory of OOD detection:* We show that detecting a backdoor in training data is equivalent to a binary Neyman-Pearson hypothesis test if OOD detection is PAC learnable as defined in (Fang et al., 2022).
5. *Methodological weakness in existing defenses:* we observe that almost all defense strategies only evaluate on backdoored models, and fail to report *false positive* rates in the likely case that models are actually clean.

2 THEORETICAL FORMULATION AND RESULTS

We focus on MBD and SBD as introduced above, in the case, where the attacker has limited control over the training data and is able to backdoor a certain portion of the dataset. The training itself is performed using a standard method, e.g., Stochastic Gradient Descent (SGD). For an extensive overview of other empirical backdoor problems, the reader is referred to, e.g., (Wu et al., 2022).

2.1 Formulating Model Backdoor Detection (MBD)

Overview. After N samples of training data are collected, the backdoor attacker has the option of backdooring a portion of the training data, by replacing each *clean* sample with a *backdoored* sample. This backdooring may alter, e.g., images as well as their labels. Subsequently, an Artificial Neural Network (ANN) is trained on the resulting training set. Given the resulting trained network (i.e., the network parameters), the task of the backdoor detector is to determine whether the training data had been backdoored. The detector may obtain M additional *clean* samples, e.g., by independently collecting additional data. We assume that the backdoor attacker has no access to these samples. Table 1 provides an overview of the notation used.

Dataset and training. Consider a, possibly stochastic, training algorithm $\mathcal{A}$ (e.g., SGD), that trains a model on training data[1] $\mathcal{D} = (X_1, X_2, \ldots, X_N)$, consisting of N i.i.d. random variables, distributed like $X \sim P$, as input and produces a parameter vector $\theta = \mathcal{A}(\mathcal{D})$ as output.

Clean data. Let $P_0 \in \mathcal{P}(\mathcal{X})$ be the probability distribution on $\mathcal{X}$ of clean samples and let $\mathcal{D}^{(0)} = (X_1^{(0)}, X_2^{(0)}, \ldots, X_N^{(0)})$ be a clean dataset, consisting of N i.i.d. random variables, drawn from P_0.

Backdoor. To *backdoor* a model, an adversary may replace some training samples with backdoored samples, drawn from a different distribution $P_b \in \mathcal{P}(\mathcal{X})$. This distribution may result from applying a backdoor function to the clean samples. Note, that as the training sample X includes the data and the label, both may be altered by the adversary.

Backdoored training data. Assuming that a fraction $\gamma \in (0, 1]$ of the training data is backdoored, the backdoored training dataset $\mathcal{D}^{(1)} = (X_1^{(1)}, X_2^{(1)}, \ldots, X_N^{(1)})$ is independently drawn according to $P_1 = \gamma P_b + (1-\gamma)P_0$, i.e., according to P_b with probability γ and from P_0 with probability $1-\gamma$.

Additional clean data. Furthermore, let $\mathcal{D}' = (X_1', X_2', \ldots, X_M') \sim P_0^M$ be M i.i.d. additional clean samples distributed according to P_0. These samples correspond to clean validation data or may have been collected by the backdoor detector prior to making a decision.

Model Backdoor Detection. The backdoor detector is a function g, that takes $\theta = \mathcal{A}(\mathcal{D}^{(j)})$ and additional data $\mathcal{D}'$ as its input and outputs 0 for *"backdoor"* and 1 for *"no backdoor"*. For MBD, we require the detector to determine j with high probability. For ease of notation, we use a Bernoulli-$\frac{1}{2}$ random variable $J \sim \mathcal{B}(\frac{1}{2})$ and define the input for the detector as $\mathbf{Q} = (\mathcal{A}(\mathcal{D}^{(J)}), \mathcal{D}')$, such that the error probability $\Pr\{g(\mathbf{Q}) \neq J\}$ of the detector is well-defined.

Possible data distributions. Finally, the last observation needed to obtain a well-defined backdoor detection problem is, that we need to avoid the possibility of $P_0 = P_b$. In the case where the clean and the backdoor distributions are identical, clearly, detection is impossible. We opt for the general approach of defining a suitable set $\mathcal{P} \subseteq \mathcal{P}(\mathcal{X})^2$ that contains all possible clean and backdoor distribution pairs $(P_0, P_b) \in \mathcal{P}$.

[1] The training sample X may be a vector that includes data and label.

These discussions then naturally lead to the following central definition.

Definition 1. *The MBD problem for a training algorithm $\mathcal{A}$ is determined by the following quantities: $\gamma \in (0, 1]$, $N \in \mathbb{N}$, $M \in \mathbb{N}$, and $\mathcal{P} \subseteq \mathcal{P}(\mathcal{X})^2$.*

Fixing these quantities, we define the risk *of a backdoor detector g associated with (P_0, P_b) as*

$$R(g; P_0, P_b) := \Pr\{g(\mathbf{Q}) \neq J\} \tag{1}$$

$$= \frac{1}{2} \sum_{j=0,1} \Pr\{g(\mathcal{A}(\mathcal{D}^{(j)}), \mathcal{D}') \neq j\}. \tag{2}$$

We say that a backdoor detector is α-error *for some $\alpha \in [0, \frac{1}{2}]$ if, for every pair $(P_0, P_b) \in \mathcal{P}$, the risk is bounded by*

$$R(g; P_0, P_b) \leq \alpha. \tag{3}$$

Remark 1. Instead of bounding the risk as in (3), it may seem more natural to require $\Pr\{g(\mathbf{Q}) \neq j | J = j\} \leq \alpha$ for $j = 0, 1$. But note that $\Pr\{g(\mathbf{Q}) \neq j | J = j\} \leq 2\alpha$ for $j = 0, 1$ immediately follows from (3).

2.2 (In)feasibility of Model Backdoor Detection

It will be useful to consider easier problems than α-error detection (Definition 1) and establish reductions. To this end, we consider four different *Types* of detectors, starting with Type 0 that corresponds to MBD. The other three detectors also seek to infer J, but are given access to progressively *more* information via oracles; as such, any subsequent detectors can improve on previous ones. The four detectors are:

Type 0: The default detector, $g_0(\mathbf{Q}_0)$, as used in Definition 1 with with $\mathbf{Q}_0 = \mathbf{Q} = (\mathcal{A}(\mathcal{D}^{(J)}), \mathcal{D}')$. This detector corresponds to our MBD problem.

Type 1: Detector $g_1(\mathbf{Q}_1)$ with $\mathbf{Q}_1 = (\mathcal{D}^{(J)}, \mathcal{D}')$, i.e., with access to the training dataset $\mathcal{D}^{(J)}$ instead of just the trained model, and M independent clean validation samples $\mathcal{D}'$.

Type 2: Detector $g_2(\mathbf{Q}_2)$ with $\mathbf{Q}_2 = (\mathcal{D}^{(J)}, P_0)$, i.e., with access to the clean data distribution instead of clean validation samples. This is an OOD detection problem: Is $\mathcal{D}^{(J)}$ OOD with respect to P_0?

Type 3: Detector $g_3(\mathbf{Q}_3)$ with $\mathbf{Q}_3 = (\mathcal{D}^{(J)}, P_0, P_b)$; i.e., the previous detector also now gets the backdoor distribution P_b, and must decide which distribution the training data came from. This is classical binary Neyman-Pearson hypothesis testing problem between P_0^N and P_1^N.

Symbol	Description
$X \sim P$	X distributed according to P
$\mathcal{X}$	Alphabet of X
$\lvert\mathcal{X}\rvert \in \mathbb{N} \cup \{\infty\}$	Alphabet size
$\mathcal{P}(\mathcal{X})$	Set of all probability distributions on $\mathcal{X}$
$X^{(0)} \sim P_0$	Clean sample/distribution
$X^{(1)} \sim P_b$	Backdoored sample/distribution
$\gamma \in [0, 1]$	Backdoor probability
$P_1 = \gamma P_b + (1-\gamma)P_0$	Mixture of P_0 and P_b
$\mathcal{D} = (X_1, \dots X_N) \sim P^N$	Dataset of N i.i.d. samples
$\mathcal{D}^{(0)} = (X_1^{(0)}, \dots X_N^{(0)}) \sim P_0^N$	Clean dataset of N i.i.d. samples
$\mathcal{D}^{(1)} = (X_1^{(1)}, \dots X_N^{(1)}) \sim P_1^N$	Backdoored dataset of N i.i.d. samples
$\mathcal{D}' = (X'_1, \dots X'_M) \sim P_0^M$	Dataset of M additional clean samples
$\theta = \mathcal{A}(\mathcal{D})$	Parameters resulting from training on $\mathcal{D}$
$\mathcal{P} \subseteq P(\mathcal{X})^2$	Set of allowed distributions $(P_0, P_b) \in \mathcal{P}$
$J \sim \mathcal{B}(\frac{1}{2})$	Bernoulli random variable indicating a backdoor
$\mathbf{Q} = (\mathcal{A}(\mathcal{D}^{(J)}), \mathcal{D}')$	Parameters θ and samples $\mathcal{D}'$ given to the detector
$g(\mathbf{Q}) \in \{0, 1\}$	Backdoor detector
$R(g; P_0, P_b) = \Pr\{g(\mathbf{Q}) \neq J\}$	Risk of detector g at (P_0, P_b)
$\alpha \in [0, \frac{1}{2}]$	Error probability
$g_0 = g$, g_1, g_2, g_3	Type 0,1,2,3 detectors
$\mathbf{Q}_0 = \mathbf{Q}$, $\mathbf{Q}_1$, $\mathbf{Q}_2$, $\mathbf{Q}_3$	Input for Type 0,1,2,3 detectors
$\mathrm{TV}(P, Q)$	Total variational distance between two distributions
$\beta \in [0, 1)$	Distance constraint $\mathrm{TV}(P_0, P_b) \geq 1 - \beta$

Table 1: Overview of Notation. This excludes Section 2.3, where some definitions are generalized.

We assume that detectors of Types 2 and 3 have access to P_0 (and P_b for a Type 3 detector) in terms of evaluation of the distribution, and also have the ability to sample from the distribution. We thus consider Types 2 and 3 as randomized detectors to account for sampling. The definitions of risk and α-error detection of g_2, g_3 apply mutatis mutandis as in Definition 1, where the probability in (1) is also taken over the randomness of g.

Remark 2 (Ordering of detector Types). Types 0 to 3 are listed in order of decreasing difficulty as, e.g., more information is provided to a Type 3 detector than to a Type 2 detector. Thus, an α-error detector g immediately provides an α-error Type 1 detector g_1, which in turn immediately provides an α-error Type 2 detector g_2, which yields an α-error detector g_3 of Type 3. Thus, we can define a total ordering on the different Types of detectors, using $A \prec B$ to signify that A can be derived from B: $\mathbf{Q}_0 \prec \mathbf{Q}_1 \prec \mathbf{Q}_2 \prec \mathbf{Q}_3$. The formal argument, showing this claim can be found in Lemma 5 in Appendix C.

In Section 2.2.1 we will show that for a reasonable $\mathcal{P}$, α-error Type 2 detection is impossible with $\alpha < \frac{1}{2}$. The reduction argument in Remark 2 thus ensures that α-error detection with $\alpha < \frac{1}{2}$ is also impossible for Type 0 and Type 1 detectors.

We can resolve the situation for a Type 3 detector using the Neyman-Pearson lemma.

Lemma 1. *Given a Type 3 backdoor detector $g_3(\mathcal{D}, P_0, P_b)$, for any pair $(P_0, P_b) \in \mathcal{P}(\mathcal{X})^2$ we have*

$$R(g_3; P_0, P_b) \geq \frac{1}{2} - \frac{1}{2}\,\mathrm{TV}(P_0^N, P_1^N) \tag{4}$$

$$\geq \frac{1}{2} - \frac{\gamma N}{2}\,\mathrm{TV}(P_0, P_b), \tag{5}$$

where the first equality in (4) *can be achieved by the Neyman-Pearson detector. Thus, an α-error detector of Type 3 can only exist if $\alpha \geq \frac{1}{2} - \frac{\gamma N}{2}\,\mathrm{TV}(P_0, P_b)$ for all $(P_0, P_b) \in \mathcal{P}$.*

Proof see Appendix A in the supplementary material.

Before we can analyze detectors of Types 1 and 2, we need to specify the set of allowable distributions $\mathcal{P}$. We do this, using Lemma 1.

First, we show that merely excluding the identity $P_0 \neq P_b$, i.e., $\mathcal{P} = \{(P_0, P_b) \in \mathcal{P}(\mathcal{X})^2 : P_0 \neq P_b\}$ is not sufficient.

Example 1. Let $g_3(\mathcal{D}, P_0, P_1)$ be an α-error Type 3 detector and assume that $\mathcal{X}$ is infinite, i.e., $\lvert\mathcal{X}\rvert = \infty$. Let $\mathcal{P}$ be given as above, ensuring only that $P_0 \neq P_b$. For any $\varepsilon > 0$, we can then choose[2] $(P_0, P_b) \in \mathcal{P}$ with

[2]Without loss of generality, we can assume $\mathcal{X}$ =

$0 < \mathrm{TV}(P_0, P_b) \leq \frac{2}{\gamma N}\varepsilon$. By Lemma 1, we have $\alpha \geq \frac{1}{2} - \frac{\gamma N}{2}\mathrm{TV}(P_0, P_b) \geq \frac{1}{2} - \varepsilon$. As $\varepsilon > 0$ was arbitrary, we have $\alpha = \frac{1}{2}$.

Lemma 1 and Example 1 show that even for a Type 3 detector, we need $\mathrm{TV}(P_0, P_b) > \frac{1-2\alpha}{\gamma N}$ for all $(P_0, P_b) \in \mathcal{P}$, in order for α-error detection to be achievable. In the following we will assume that $\mathcal{P}$ is the set of probability distributions P_0, P_b with $\mathrm{TV}(P_0, P_b) \geq 1 - \beta$, for some fixed $\beta \in [0, 1)$. This strong requirement is motivated by the fact that in this case, $\frac{1-\gamma+\gamma\beta}{2}$-error Type 3 detection is achievable with only one sample.

Remark 3. Thorough reasoning and examples, illustrating why total variation distance is the preferred distance measure for distribution hypothesis testing can be found in (Canonne, 2022, Section 1.2).

2.2.1 Impossibility

In the following we prove an impossibility result, which implies that *for an infinite alphabet $\mathcal{X}$, the error probability (as given in Definition 1) of any detector (of Type 0, Type 1 or Type 2) is $\frac{1}{2}$, the error probability of a random guess.* Additionally, for finite $\mathcal{X}$, we provide a lower bound on the size of the training set N, as a function of α.

Theorem 1. *Fix $N \in \mathbb{N}$, $\alpha \in (0, \frac{1}{2}]$, $\beta \in [0, 1]$, and $\mathcal{P} = \{(P_0, P_b) : \mathrm{TV}(P_0, P_b) \geq 1 - \beta\}$. Let $g_2(\mathcal{D}, P_0)$ be an α-error Type 2 detector. For $|\mathcal{X}| = \infty$, we then have necessarily $\alpha = \frac{1}{2}$, while for $|\mathcal{X}| < \infty$, we have*

$$N \geq \frac{\log 2\alpha}{2} + \sqrt{\frac{(\log 2\alpha)^2}{4} + (\beta|\mathcal{X}| - 1)\log\frac{1}{2\alpha}}. \quad (6)$$

Proof see Appendix A in the supplementary material.

It is important to notice that the bound (6) relates the number of training samples N with the alphabet size $|\mathcal{X}|$ and the risk α, while the number M of clean samples available to the defender does not appear. By the reduction argument in Lemma 1 the impossibility result in Theorem 1 also holds for detector Types 0 and 1 for all possible values $M \in \mathbb{N}$.

For a fixed dataset alphabet size $|\mathcal{X}|$ and allowed error probability α, the bound (6) gives the minimum size of the training set N for the error level α to be achievable. Note the following special cases in terms of α, β:

- For $\alpha = \frac{1}{2}$, the bound (6) is always satisfied as the RHS is 0, showing that $\frac{1}{2}$-error detection is always achievable. This coincides with the error probability of a random guess.
- The bound (6) is monotonically decreasing in α and for $\alpha \to 0$, it approaches $\beta|\mathcal{X}|$.
- In case $\beta = 0$, the bound (6) is always satisfied as the RHS is zero for $\alpha \in (0, \frac{1}{2}]$ in this case. This shows that α-error detection is always possible if P_0 and P_b have disjoint support, i.e., $\mathrm{TV}(P_0, P_b) = 1$.

For an infinite alphabet $\mathcal{X}$, (6) needs to be satisfied for arbitrarily large values of $|\mathcal{X}|$. For finite training set size N, this is only possible if $\alpha = \frac{1}{2}$ as then, $\log\frac{1}{2\alpha} = 0$. Thus, in this case, for any Type 2 detector, there is a particular clean distribution and backdoor strategy, such that this detector performs no better than random guessing.

For fixed α and β, we can use (6) to determine the minimum size of the training set N for popular datasets, for α error probability to be achievable by a Type 2 detector. To this end, we use the width W, height H, number of channels C and color depth D of an image dataset to compute $|\mathcal{X}| = D^{WHC}$. For categorical datasets, we may multiply the number of categories for all the properties recorded in the dataset to obtain $|\mathcal{X}|$. The resulting value for the bound in (6) is given in Table 2 for several popular datasets. As can be seen by these numbers, this universal backdoor detection is infeasible for all, but the smallest tabular datasets. Code for computing the values in Table 2 can be found at https://github.com/g-pichler/in_feasibility_of_ml_backdoor_detection.

Note also, that the impossibility of Type 2 backdoor detection automatically precludes the existence of Type 1 or Type 0 error detectors with equal performance by the reduction argument in Remark 2.

Illustrative Example

We noted previously the following consequence of Theorem 1, in case of an infinite alphabet: For any Type 2 backdoor detector, there exists an attacker, such that the detector is no better than a random guess. Here, we will showcase this on a toy example of a binary classification task, for a specific data distribution P_0, and any backdoor detector from a family of Type 2 detectors, parameterized by $\mathbf{v} \in \mathbb{R}^K$. For any parameter $\mathbf{v}$, we show how to construct a backdoor attack that is both effective in changing the decision regions of a classifier trained on backdoored data, and undetectable by the backdoor detector.

Data Distribution. We have data and label pairs $X = (Y, \mathbf{Z})$, where $Y \in \{-1, 1\}$ is a binary label and $\mathbf{Z} = Y\mathbf{1} + \sigma\mathbf{W}$ with $\sigma > 0$ and $\mathbf{W}$ is multivariate normal with dimension K. For $K = 2$ dimensions, the optimal classifier for this problem decides $\hat{y} = 1$

$\mathbb{N}$. Then, this can, e.g., be achieved by $P_0 = \mathcal{U}(\{0, 1, 2, \ldots, \lfloor\frac{\gamma N}{2\varepsilon}\rfloor\})$ and $P_b = \mathcal{U}(\{1, 2, \ldots, \lfloor\frac{\gamma N}{2\varepsilon}\rfloor\})$. We use $\mathcal{U}(\cdot)$ to denote a uniform distribution on a finite set.

if $z_1 + z_2 \geq 0$ and otherwise $\hat{y} = -1$, leading to the decision boundary $z_2 = -z_1$.

Backdoor Detector. The Type 2 detector $g_2(\mathcal{D}, P_0)$ is parameterized by the unit vector $\mathbf{v} \in \mathbb{R}^K$ with $\|\mathbf{v}\| = 1$. Using the fact that $f(X) := \mathbf{v} \cdot (Y\mathbf{Z}) = \mathbf{v} \cdot \mathbf{1} + \sigma W$ with a standard normal variable W, we see that applying the function $f(\cdot)$ to $\mathcal{D}$ yields N i.d.d. Gaussian random variables with mean $\mu := \mathbf{1} \cdot \mathbf{v}$ and variance σ^2. The detector then performs a statistical goodness-of-fit test on this dataset. We utilize the Kolmogorov-Smirnov test for this purpose.

Backdoor. Knowing $\mathbf{v}$, the attacker transforms an input sample $x = (\mathbf{z}, y)$ into a backdoored sample $b(x) := (\mathbf{z} + y\Delta, -y)$ with the opposite label, and shifted by $y\Delta$, where $\Delta = \frac{2}{\sqrt{K-\mu^2}}(\mathbf{1} - \mu\mathbf{v}) - 2\mu\mathbf{v}$. This transformation ensures that the statistics of $f(X)$ do not change when applying the backdoor.

After the attacker replaces clean samples with backdoored samples at a rate of $\gamma \in (0, 1]$, the Kolmogorov-Smirnov test is performed. Figure 1 showcases this strategy in $K = 2$ dimensions with $N = 150$, $\gamma = 0.5$, $\sigma = 0.5$, and $\mathbf{v} = (0.981, 0.196)$, resulting in $\mu = 1.177$. The Kolmogorov-Smirnov test obtained a p-value of $p_{\text{val}} = 0.2381$, thus not detecting the backdoor. The resulting histograms of $f(\mathcal{D})$ for clean and backdoored data are shown in Fig. 2. The code for this example can be found at https://github.com/g-pichler/in_feasibility_of_ml_backdoor_detection.

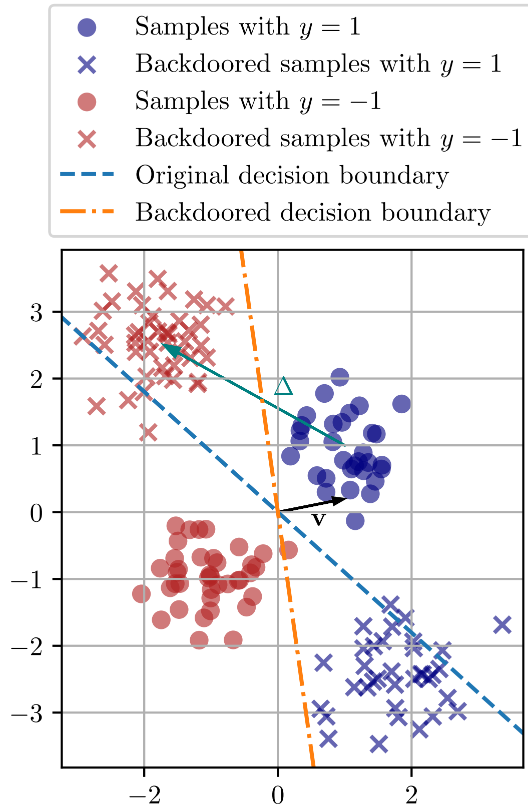

Figure 1: MBD example with $N = 150$ samples. The backdoor detector uses projection onto $\mathbf{v}$ to take a decision. The vector Δ is the additive backdoor trigger used by the attacker. The decision boundary changes when applying the backdoor.

2.2.2 Achievability

In this section we are going to show that α-error Type 2 detection is always achievable if the size of the alphabet $|\mathcal{X}|$ is small enough:

Theorem 2. *Considering the backdoor detection setup of Definition 1 with* $\mathcal{P} = \{(P_0, P_b) : \text{TV}(P_0, P_b) \geq 1 - \beta\}$ *and a finite alphabet* $|\mathcal{X}| < \infty$. *If*

$$\alpha > 2|\mathcal{X}| \exp\left(-\frac{2N\gamma^2(1-\beta)^2}{|\mathcal{X}|^2}\right), \qquad (7)$$

then there exists an α*-error Type 2 detector.*

Proof see Appendix A in the supplementary material.

Note the following special cases for α, β, γ, and $|\mathcal{X}|$:

- The bound on the RHS of (7) increases monotonically from 0 to ∞ for increasing $|\mathcal{X}|$. Thus, there is some fixed alphabet size, below which, α-error detection is guaranteed to be possible.
- For $\alpha = 0$, (7) cannot be satisfied.
- The case $\beta = 1$ allows for $P_0 = P_b$ and thus no α-error detector exists for $\alpha \in [0, \frac{1}{2})$ in this case and (7) cannot be satisfied.
- For $\gamma = 0$, the distributions $P_0 = P_b$ are identical and thus no α-error detector exists for $\alpha \in [0, \frac{1}{2})$ in this case and (7) cannot be satisfied.

2.2.3 Connections to PAC-Learnability of OOD Detection

Note that a Type 1 detector essentially needs to solve an OOD detection problem: The detector g_1 needs to determine if the N samples $\mathcal{D}$ were drawn from the same distribution as $\mathcal{D}'$.

The goal of this section is to prove Theorem 3. This theorem has an interesting implication in case the OOD detection problem is PAC-learnable: If an α-error Type 3 backdoor detector g_3 exists, then $(\alpha + \epsilon)$-error detection is also possible for a Type 1 detector for any $\epsilon > 0$. Thus, essentially Types 1 to 3 all become equivalent if OOD detection is PAC-learnable. Note that Type 3 detection is characterized by Lemma 1.

The PAC-learnability of the detector in Type 1 was analyzed in (Fang et al., 2022). We fist restate a spe-

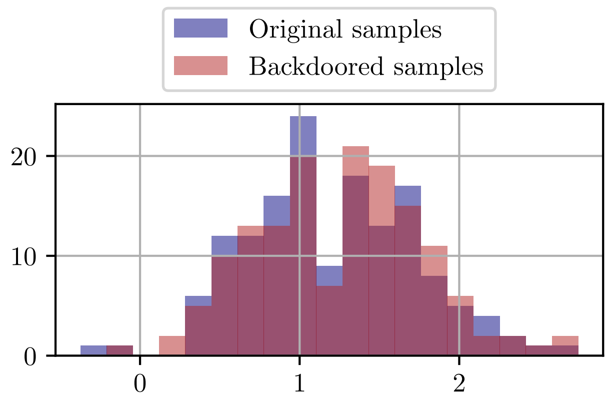

Figure 2: Histogram of the detector decision statistics clean and backdoored samples depicted in Fig. 1.

Table 2: Lower bound (6) on N evaluated for popular datasets with $\alpha = 0.1$ and $\beta = 0.001$.

Dataset	$\|\mathcal{X}\|$	N
Lisa Traffic Sign	256^{307200}	$\geq 10^{369904}$
ImageNet	256^{150528}	$\geq 10^{181252}$
CIFAR10	256^{3072}	$\geq 10^{3697}$
MNIST	256^{784}	$\geq 10^{942}$
B/W MNIST	2^{784}	$\geq 10^{116}$
Adult	$\geq 10^{21.86}$	$\geq 10^{9}$
Heart Disease	$\geq 10^{13.51}$	$\geq 10^{5}$
Iris	$\geq 10^{6.35}$	$\geq 10^{1}$

cial case of the definition of (weak) PAC-learnability as given in (Fang et al., 2022, Def. 1).

Definition 2. *For distributions P_0, P_b on $\mathcal{X}$, the* OOD-risk *of a function $f : \mathcal{X} \to \{0,1\}$, w.r.t. the Hamming distance, is defined as*

$$\bar{R}(f, P_0, P_b) := \Pr\{f(X^{(J)}) \neq J\} = \frac{1}{2}\Pr\{f(X^{(0)}) = 1\} + \frac{1}{2}\Pr\{f(X^{(1)}) = 0\}.$$

Given a space of probability function $\mathcal{P}$, OOD-detection is PAC-learnable *on $\mathcal{P}$ if there exists an algorithm $\mathcal{G}\colon \bigcup_{m=1}^{\infty} \mathcal{X}^m \to \{0,1\}^{\mathcal{X}}$ and a monotonically decreasing sequence $\epsilon(m)$ such that $\lim_{m\to\infty} \epsilon(m) = 0$ and for all $(P_0, P_b) \in \mathcal{P}$, and all $M \in \mathbb{N}$ we have*[3]

$$\mathbb{E}[\bar{R}(\mathcal{G}(\mathcal{D}'), P_0, P_b)] - \inf_f \bar{R}(f, P_0, P_b) \leq \epsilon(M), \quad (8)$$

where the expectation is taken w.r.t. $\mathcal{D}'$ and the infimum is over $\{0,1\}^{\mathcal{X}}$, i.e., all functions $f\colon \mathcal{X} \to \{0,1\}$.

Remark 4. Definition 2 is a special case of (Fang et al., 2022, Def. 1) in several ways[4]:

- The hypothesis space is the complete function space $\mathcal{H} = \{0,1\}^{\mathcal{X}}$, of functions $f\colon \mathcal{X} \to \{0,1\}$.
- The loss function, as used in (Fang et al., 2022, Eq. (1)) is the Hamming distance, i.e., $\ell(y, y') = 1$ if and only if $y \neq y'$.
- We are purely concerned with one-class novelty detection, i.e., $K = 1$ in (Fang et al., 2022, Sec. 2). Therefore we do not take Y_O and Y_I into account, as $Y_I \equiv 1$ and $Y_O \equiv 2$.
- Note that $(P_0, P_b) \in \mathcal{P}$ play the role of (D_{X_O}, D_{X_I}) and the complete domain space is then given by $\mathscr{D}_{XY} = \{D_{XY} : D_{XY} = \frac{1}{2}P_0 + \frac{1}{2}P_b, (P_0, P_b) \in \mathcal{P}\}$.

Note, that strong PAC-learnability (Fang et al., 2022, Def. 2) implies weak learnability.

To connect PAC-learnability of OOD detection to the learning of backdoor detectors, we consider PAC-learnability on the N-dimensional product space, i.e., on $\mathcal{X}^N$ with distributions P_0^N, P_b^N.

We can now connect PAC-learnability to the existence of α-error detectors of Types 1 and 3.

Theorem 3. *Consider the backdoor detection setup of Definition 1, with fixed $\gamma \in (0,1]$, $N \in \mathbb{N}$ and some set of possible distributions $\mathcal{P}$. Let $\mathcal{P}'$ be the set of N-fold products of (P_0, P_1), i.e., $\mathcal{P}' = \{(P_0^N, (\gamma P_b + (1-\gamma)P_0)^N) : (P_0, P_b) \in \mathcal{P}\}$. Then, OOD-detection is PAC-learnable on $\mathcal{P}'$ if and only if the following holds for any $\epsilon > 0$ and any Type 3 detector $g_3(\mathcal{D}, P_0, P_b)$: We can find $M \in \mathbb{N}$ and a Type 1 detector $g_1(\mathcal{D}, \mathcal{D}')$, which satisfies $R(g_1, P_0, P_b) \leq R(g_3, P_0, P_b) + \epsilon$ for every $(P_0, P_b) \in \mathcal{P}$.*

Proof see Appendix A in the supplementary material.

Corollary 1. *If OOD-detection is PAC-learnable on $\mathcal{P}'$, we have the following: If α-error backdoor detection is possible in the easier case of Type 3 detection, which is completely characterized by Lemma 1, then $(\alpha + \epsilon)$-error detection is also possible for a Type 1 detector for any $\epsilon > 0$.*

Consequently, up to topological closure, the same error probability is achievable for all detector Types 1 to 3, if OOD-detection is PAC-learnable on $\mathcal{P}'$.

2.3 Generalizing to Sample Backdoor Detection

We can generalize Definition 1 to SBD by providing a detector $g'(\mathbf{Q}')$ with input $\mathbf{Q}' = (\mathbf{Q}, X^{(I)}) =$

[3] Note that $\mathcal{D}'$ contains M samples.

[4] We use the notation of (Fang et al., 2022, Sec. 2) for the following symbols: $\mathcal{H}$, X_O, X_I, Y_O, Y_I, D_{X_O}, D_{X_I}, D_{XY}, $\mathscr{D}_{XY}$, $\ell(\cdot,\cdot)$, K.

$(\mathcal{A}(\mathcal{D}^{(J)}), \mathcal{D}', X^{(I)})$, where a random variable I on $\{0,1\}$ determines if a sample $X^{(I)}$ was drawn as $X^{(0)} \sim P_0$ $(I = 0)$ or as[5] $X^{(1)} \sim P_b$ $(I = 1)$.

We define a general target function $t(j,i) \in \{0,1\}$ and require that a backdoor detector satisfies $g'(\mathbf{Q}') = t(J,I)$ with high probability. In this case, it is beneficial to allow for an arbitrary probability distribution P_{JI} of (J,I) on $\{0,1\}^2$. This naturally leads to the following alternative definition of α-error detection, generalizing Definition 1.

Definition 3. *A backdoor detection problem for a training algorithm $\mathcal{A}$ is determined by the following quantities: $\gamma \in (0,1]$, $N \in \mathbb{N}$, $M \in \mathbb{N}$, $\mathcal{P} \subseteq \mathcal{P}(\mathcal{X})^2$, $P_{JI} \in \mathcal{P}(\{0,1\}^2)$, and $t\colon \{0,1\}^2 \to \{0,1\}$.*

Fixing these quantities, we define the risk *of a backdoor detector g' associated with (P_0, P_b) as $R(g'; P_0, P_b) := \Pr\{g'(\mathbf{Q}') \neq t(J,I)\}$, where the probability is w.r.t. $\mathbf{Q}' = (\mathcal{A}(\mathcal{D}^{(J)}), \mathcal{D}', X^{(I)})$ and (J,I).*

We say that a backdoor detector is α-error *for some $\alpha \in [0, \frac{1}{2}]$ if, for every pair $(P_0, P_b) \in \mathcal{P}$, the risk is bounded by $R(g'; P_0, P_b) \leq \alpha$.*

Then, even OOD can be modeled using our setup. Figure 3 presents an overview of the target function $t(j,i)$ for MBD, SBD and OOD.

Note that several cells in the diagrams in Fig. 3 are grayed out. This reflects the fact that for certain flavors of backdoor detection, specific combinations of (j,i) are not relevant. For MBD for instance, we are not interested in whether the target sample $X^{(I)}$ contains a backdoor and we can thus assume $I = 0$ in this case, effectively reducing this case to the problem introduced in Section 2.1 with $M+1$ samples being drawn from P_0, i.e., $(\mathcal{D}', X^{(0)})$, available to the detector. Conversely, the case of a clean model, i.e., $j = 0$ and a sample with a backdoor, i.e., $i = 1$ is not realistic for SBD and we set $P_{JI}(0,1) = 0$ in this case. By setting $J = 0$ (i.e., model is trained on clean data and $P_{JI}(1,0) = P_{JI}(1,1) = 0$) and using $t_{\mathrm{OOD}}(j,i) = t_{\mathrm{SBD}}(i,0) = i$, we obtain an OOD detection problem, where the detector has access to a model $\mathcal{A}(\mathcal{D}^{(0)})$ trained on clean data and additional clean data $\mathcal{D}'$. The detector then needs to determine whether $X^{(I)}$ is in or out of distribution.

To show how our result from Sections 2.2.1 and 2.2.2 carry over to other variants of backdoor detection, we will directly use Theorem 1 to derive a similar result for SBD. In analogy to the different Types of MBD detectors introduced in Section 2, we have a Type 2 detector $g_2'(\mathbf{Q}_2')$ with $\mathbf{Q}_2' = (\mathcal{D}^{(J)}, P_0, X^{(I)})$ for SBD.

[5]Note, that $X^{(1)}$ is distributed according to P_b and **not** according to $P_1 = (1-\gamma)P_0 + \gamma P_b$.

For such a detector we can leverage a reduction argument to obtain the following.

Corollary 2. *Let $g_2'(\mathcal{D}^{(J)}, P_0, X^{(I)})$ be a Type 2 detector for an SBD problem, where we have $r = \min\{P_{JI}(0,0), P_{JI}(1,1)\} > 0$ and $\mathcal{P} = \{(P_0, P_b) : \mathrm{TV}(P_0, P_b) \geq 1 - \beta\}$. Then, if g_2' is α-error, we have $\alpha \geq r$ if $|\mathcal{X}| = \infty$, and for $|\mathcal{X}| < \infty$, we obtain*

$$N \geq \frac{\log \frac{\alpha}{r}}{2} + \sqrt{\frac{(\log \frac{\alpha}{r})^2}{4} + (\beta|\mathcal{X}| - 1) \log \frac{r}{\alpha}}. \quad (9)$$

Proof see Appendix A in the supplementary material.

3 RELATED WORKS

Backdoor attacks. Early backdoor methods rely on triggers that are visible to the human eye, and generally consist of a local patch on the samples (Gu et al., 2019; Shafahi et al., 2018; Nguyen and Tran, 2020). Other attacks add a layer of stealthiness by using invisible triggers, which are commonly covering the whole sample and are not detectable by the human eye (Chen et al., 2017; Zeng et al., 2021; Li et al., 2021a). Additive attacks (Gu et al., 2019; Chen et al., 2017; Shafahi et al., 2018) fuse the triggers to the clean samples as additive noise. Conversely, non-additive attacks (Zeng et al., 2021; Li et al., 2021a; Nguyen and Tran, 2021) modify the samples by changing attributes such as the color of the pixels or applying spatial transformations. Additionally, some attacks add the same trigger to all samples and are therefore sample-agnostic (Gu et al., 2019; Chen et al., 2017), while others are sample-specific (Nguyen and Tran, 2020; Nguyen and Tran, 2021). Finally, recent attacks have been proposed to reduce the issue of the linear separability between clean and backdoored samples (Qi et al., 2023) which arises in many of the previously mentioned works.

Backdoor defenses. This work is concerned with "post-training" defenses, i.e. those methods that aim to remove or mitigate the backdoor effect from a backdoored model, as opposed to techniques that deal with the problem before (Udeshi et al., 2022; Gao et al., 2023) or during training (Huang et al., 2022). The solutions that mostly align with the proposed frameworks are those that are designed to detect whether the model is backdoored (Liu et al., 2017; Wang et al., 2019), or those that can detect backdoored samples (Li et al., 2021b; Huang et al., 2022; Liu et al., 2019). While some detector only work when the trigger is assumed to be sample-agnostic (Chou et al., 2020; Gao et al., 2019; Tao et al., 2022), others are reported to be effective on sample-specific triggers (Zeng et al., 2021; Liu et al., 2023). Moreover, recent detectors propose to replace the need for a set of clean samples with the generation of perturbed samples which may help

Sample Backdoored? (i)

Model Backdoored? (j)

	$0/N$	$1/Y$
$0/N$	0	–
$1/Y$	1	–

(a) $t_{\mathrm{MBD}}(j,i) = j$.

	$0/N$	$1/Y$
$0/N$	0	–
$1/Y$	0	1

(b) $t_{\mathrm{SBD}}(j,i) = i$.

	$0/N$	$1/Y$
$0/N$	0	1
$1/Y$	–	–

(c) $t_{\mathrm{OOD}}(j,i) = t_{\mathrm{SBD}}(j,i) = i$.

Figure 3: Target function $t(j,i)$ for different backdoor detection flavors. $j \in \{0,1\}$ signals if the training dataset is backdoored $(j=1)$ or not $(j=0)$, while $i \in \{0,1\}$ indicates if the test sample is backdoored.

to create a representation of the backdoored samples (Liu et al., 2023; Pang et al., 2023). In providing a theoretical analysis from first principles, and in suggesting a strong connection between OOD detection and backdoor detection, our work is complementary to (Ma et al., 2023). Finally, is important to notice that most literature on backdoor detection focuses on the SBD problem and on the mitigation of backdoor attacks, when a dataset is known to be backdoored (Udeshi et al., 2022; Huang et al., 2022; Tran et al., 2018; Wu and Wang, 2021).

Certifiable Defenses. For data poisoning attacks, where the attacker's goal is only to diminish the accuracy of the trained classifier, certifiable defenses do exist (Steinhardt et al., 2017; Koh et al., 2022). Also for backdoor attacks, smoothing strategies (Weber et al., 2023; Wang et al., 2020) were proposed, which allow for certified robustness against backdoor attacks. However, the threat model is severely restricted to very small (in size and amplitude) triggers, which can be successfully obscured by adding smoothing noise. The impossibility result in Section 2.2.1 showcases, why certifiable defenses cannot be mounted against a capable attacker in general.

4 CONCLUSIONS

We provided a formal statistical definition of backdoor detection and investigated the feasibility of backdoor detection. As the backdoor attack is usually not known to the defender, in our analysis we focused on universal (adversary-unaware) backdoor detection. This implies that such backdoor defense schemes must be robust against targeted attacks, which are crafted to fool the specific defense strategy, excluding any "security-by-obscurity" schemes, where defense only holds as long as it is not public knowledge. We concluded that under very general assumptions, universal (adversary-unaware) backdoor detection is not possible. Thus, backdoor detectors need to be adversary-aware to perform well at their task.

Ultimately, this work makes the claim that designing universal (adversary-unaware) backdoor detection methods is an exercise in futility. As did (Shokri and Tan, 2020), we make the case that backdoor detectors need to be adversary-aware or make specific assumptions on the data distribution and/or backdoor strategy employed. Unfortunately this is not the case for much published work, which implies that the proposed methods must fail on many untested instances of backdoor detection.

Furthermore, we note that when designing a backdoor detection algorithm, the advantage should be given to the attacker, which is able to adapt to a defense strategy, but not the other way around.

Acknowledgments

This work is supported in part by grants from the National Science Foundation (NSF) and the ARO 77191, in part by the Army Research Office under grant #W911NF-21-1-0155, in part by Intel, and in part by the NYUAD Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010.

References

Burkart, N. and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Canonne, C. L. (2022). Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends in Communications and Information Theory*, 19(6):1032–1198.

Chen, H., Fu, C., Zhao, J., and Koushanfar, F. (2019). DeepInspect: A black-box trojan detection and mitigation framework for deep neural networks. In *IJCAI*, volume 2, page 8.

Chen, P., Ji, D., Lin, K., Zeng, R., Li, T. H., Tan, M., and Gan, C. (2022). Weakly-supervised multi-granularity map learning for vision-and-language navigation. In *NeurIPS*.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv*, arXiv:1712.05526.

Chou, E., Tramer, F., and Pellegrino, G. (2020). SentiNet: Detecting localized universal attacks against deep learning systems. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 48–54, Los Alamitos, CA, USA.

Fang, Z., Li, Y., Lu, J., Dong, J., Han, B., and Liu, F. (2022). Is out-of-distribution detection learnable? In *Advances in Neural Information Processing Systems*, volume 35, pages 37199–37213.

Fu, H., Krishnamurthy, P., Garg, S., and Khorrami, F. (2023). Differential analysis of triggers and benign features for black-box dnn backdoor detection. *Trans. Info. For. Sec.*, 18:4668–4680.

Gao, K., Bai, Y., Gu, J., Yang, Y., and Xia, S. (2023). Backdoor defense via adaptively splitting poisoned dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 4005–4014.

Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D. C., and Nepal, S. (2019). STRIP: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 113–125.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Gu, T., Dolan-Gavitt, B., and Garg, S. (2019). BadNets: Identifying vulnerabilities in the machine learning model supply chain. *IEEE Access*, 7:47230–47244.

Huang, K., Li, Y., Wu, B., Qin, Z., and Ren, K. (2022). Backdoor defense via decoupling the training process. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32.

Koh, P. W., Steinhardt, J., and Liang, P. (2022). Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*, pages 1–47.

Li, Y., Li, Y., Wu, B., Li, L., He, R., and Lyu, S. (2021a). Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16463–16472.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. (2021b). Anti-backdoor learning: Training clean models on poisoned data. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 14900–14912.

Liu, X., Li, M., Wang, H., Hu, S., Ye, D., Jin, H., Wu, L., and Xiao, C. (2023). Detecting backdoors during the inference stage based on corruption robustness consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16363–16372.

Liu, Y., Lee, W.-C., Tao, G., Ma, S., Aafer, Y., and Zhang, X. (2019). ABS: Scanning neural networks for back-doors by artificial brain stimulation. In *ACM SIGSAC Conference on Computer and Communications Securi*, CCS '19, pages 1265–1282, New York, NY, USA.

Liu, Y., Xie, Y., and Srivastava, A. (2017). Neural trojans. In *2017 IEEE International Conference on Computer Design (ICCD)*, pages 45–48.

Ma, W., Wang, D., Sun, R., Xue, M., Wen, S., and Xiang, Y. (2023). The "Beatrix" resurrections: Robust backdoor detection via gram matrices. In *Network and Distributed System Security (NDSS) Symposium*.

Nguyen, T. A. and Tran, A. T. (2020). Input-aware dynamic backdoor attack. In *Advances in Neural Information Processing Systems*, volume 33, pages 3454–3464.

Nguyen, T. A. and Tran, A. T. (2021). WaNet - Imperceptible warping-based backdoor attack. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Pang, L., Sun, T., Ling, H., and Chen, C. (2023). Backdoor cleansing with unlabeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12218–12227.

Qi, X., Xie, T., Li, Y., Mahloujifar, S., and Mittal, P. (2023). Revisiting the assumption of latent separability for backdoor defenses. In *11th International Conference on Learning Representations, ICLR*.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216.

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., and Goldstein, T. (2018). Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Annual Conference on Neural Information Processing Systems, NeurIPS 2018*, pages 6106–6116.

Shen, G., Liu, Y., Tao, G., An, S., Xu, Q., Cheng, S., Ma, S., and Zhang, X. (2021). Backdoor scanning for deep neural networks through k-arm optimization. In *International Conference on Machine Learning*, pages 9525–9536. PMLR.

Shokri, R. and Tan, T. J. L. (2020). Bypassing backdoor detection algorithms in deep learning. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 175–183. IEEE.

Steinhardt, J., Koh, P. W. W., and Liang, P. S. (2017). Certified defenses for data poisoning attacks. *Advances in neural information processing systems*, 30.

Tao, G., Shen, G., Liu, Y., An, S., Xu, Q., Ma, S., Li, P., and Zhang, X. (2022). Better trigger inversion optimization in backdoor scanning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13358–13368.

Tchango, A. F., Goel, R., Wen, Z., Martel, J., and Ghosn, J. (2022). DDXPlus: A new dataset for automatic medical diagnosis. In *Advances in Neural Information Processing Systems*, volume 35, pages 31306–31318.

Tran, B., Li, J., and Madry, A. (2018). Spectral signatures in backdoor attacks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31.

Udeshi, S., Peng, S., Woo, G., Loh, L., Rawshan, L., and Chattopadhyay, S. (2022). Model agnostic defence against backdoor attacks in machine learning. *IEEE Transactions on Reliability*, 71(2):880–895.

Varoquaux, G. and Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digit. Medicine*, 5.

Villani, C. (2021). *Topics in optimal transportation*, volume 58. American Mathematical Soc.

Wang, B., Cao, X., jia, J., and Gong, N. Z. (2020). On certifying robustness against backdoor attacks via randomized smoothing. *CVPR Workshop on Adversarial Machine Learning in Computer Vision*.

Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., and Zhao, B. Y. (2019). Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723.

Wang, H., Liang, W., Gool, L. V., and Wang, W. (2022). Towards versatile embodied navigation. In *Advances in Neural Information Processing Systems*, volume 35, pages 36858–36874.

Weber, M., Xu, X., Karlaš, B., Zhang, C., and Li, B. (2023). Rab: Provable robustness against backdoor attacks. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1311–1328. IEEE.

Wu, B., Chen, H., Zhang, M., Zhu, Z., Wei, S., Yuan, D., and Shen, C. (2022). Backdoorbench: A comprehensive benchmark of backdoor learning. In *36th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wu, D. and Wang, Y. (2021). Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925.

Zeng, Y., Park, W., Mao, Z. M., and Jia, R. (2021). Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16473–16481.

Zhang, T., Ren, T., Xiao, C., Xiao, W., Gonzalez, J. E., Schuurmans, D., and Dai, B. (2023). Energy-based predictive representations for partially observed reinforcement learning. In *Uncertainty in Artificial Intelligence, UAI 2023*, volume 216 of *Proceedings of Machine Learning Research*, pages 2477–2487. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not Applicable

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

A Proofs

Lemma 1. *Given a Type 3 backdoor detector $g_3(\mathcal{D}, P_0, P_b)$, for any pair $(P_0, P_b) \in \mathcal{P}(\mathcal{X})^2$ we have*

$$R(g_3; P_0, P_b) \geq \frac{1}{2} - \frac{1}{2}\operatorname{TV}(P_0^N, P_1^N) \tag{4}$$

$$\geq \frac{1}{2} - \frac{\gamma N}{2}\operatorname{TV}(P_0, P_b), \tag{5}$$

where the first equality in (4) can be achieved by the Neyman-Pearson detector. Thus, an α-error detector of Type 3 can only exist if $\alpha \geq \frac{1}{2} - \frac{\gamma N}{2}\operatorname{TV}(P_0, P_b)$ for all $(P_0, P_b) \in \mathcal{P}$.

Proof of Lemma 1. Fix (P_0, P_b) and let $\mathcal{Q} = \{\mathbf{x} \in \mathcal{X}^N : g_3(\mathbf{x}, P_0, P_b) = 1\}$ to obtain

$$1 - R(g_3; P_0, P_b) = \frac{1}{2}\sum_{j \in \{0,1\}} \Pr\{g_3(\mathcal{D}^{(j)}, P_0, P_b) = j\}$$

$$= \frac{1}{2}\int \mathbb{1}_{\mathcal{Q}}\, dP_1^N + \frac{1}{2}\int \mathbb{1}_{\mathcal{Q}^c}\, dP_0^N \tag{10}$$

$$= \frac{1}{2} + \frac{1}{2}\int \mathbb{1}_{\mathcal{Q}}\, dP_1^N - \frac{1}{2}\int \mathbb{1}_{\mathcal{Q}}\, dP_0^N \tag{11}$$

$$= \frac{1}{2} + \frac{1}{2}\int \mathbb{1}_{\mathcal{Q}}\, d(P_1^N - P_0^N) \tag{12}$$

$$\leq \frac{1}{2} + \frac{1}{2}\operatorname{TV}(P_0^N, P_1^N) \tag{13}$$

$$\leq \frac{1}{2} + \frac{N}{2}\operatorname{TV}(P_0, P_1) \tag{14}$$

$$\leq \frac{1}{2} + \frac{\gamma N}{2}\operatorname{TV}(P_0, P_b), \tag{15}$$

where (13) is a consequence of (Villani, 2021, Exercise 1.17). Also using (Villani, 2021, Exercise 1.17), we see that equality in (13) is achieved for the Neyman-Pearson detector

$$g_3(\mathcal{D}, P_0, P_b) = \mathbb{1}\left\{\frac{dP_1^N}{dP_0^N}(\mathcal{D}) \geq 1\right\}. \tag{16}$$

The last two steps (14) and (15) follow from Lemma 2. □

Theorem 1. *Fix $N \in \mathbb{N}$, $\alpha \in (0, \frac{1}{2}]$, $\beta \in [0, 1]$, and $\mathcal{P} = \{(P_0, P_b) : \operatorname{TV}(P_0, P_b) \geq 1 - \beta\}$. Let $g_2(\mathcal{D}, P_0)$ be an α-error Type 2 detector. For $|\mathcal{X}| = \infty$, we then have necessarily $\alpha = \frac{1}{2}$, while for $|\mathcal{X}| < \infty$, we have*

$$N \geq \frac{\log 2\alpha}{2} + \sqrt{\frac{(\log 2\alpha)^2}{4} + (\beta|\mathcal{X}| - 1)\log\frac{1}{2\alpha}}. \tag{6}$$

Proof of Theorem 1. For brevity we assume P_0 to be given and drop it as an argument for $g_2(\mathcal{D}, P_0) = g_2(\mathcal{D})$. Assume that g_2 is an α-error detector. Without loss of generality, we will assume $|\mathcal{X}| = K \in \mathbb{N}$ and set $\mathcal{X} = \{1, \dots, K\}$. The case $|\mathcal{X}| = \infty$ will follow by letting $K \to \infty$.

Choose $P_0 = \mathcal{U}(\mathcal{X})$, the uniform distribution on $\mathcal{X} = \{1, \dots, K\}$. For an arbitrary, vector $\mathbf{y} = (y_1, y_2, \dots, y_M) \in \mathcal{X}^M$, let $\mathcal{Q}_{\mathbf{y}}$ be the discrete uniform distribution on the elements of $\mathbf{y}$. Note that this is only the uniform distribution on the set $\{y_m : m = 1, \dots, M\}$ if all components of $\mathbf{y}$ are different. Clearly, we have $\operatorname{TV}(P_0, \mathcal{Q}_{\mathbf{y}}) \geq 1 - \frac{M}{K}$. Thus, by choosing $M \leq \beta K$ it is ensured that $\operatorname{TV}(P_0, \mathcal{Q}_{\mathbf{y}}) \geq 1 - \beta$.

Let $\mathbf{Y} = (Y_1, Y_2, \dots Y_M)$ be a random vector with M elements, each drawn i.i.d. according to $Y_m \sim P_0$. We now draw another random vector $\mathbf{Z}$ with N elements $\mathbf{Z} = \{Z_n\}_{n=1,2,\dots,N}$ according to $Z_n = (1 - G_n)X_n^{(0)} + G_n Y_{V_n}$, where $V_n \sim \mathcal{U}(\{1, 2, \dots, M\})$ and $G_n \sim \mathcal{B}(\gamma)$ are all independently drawn for $n = 1, 2, \dots, N$. Thus, V_n is uniformly drawn from $\{1, 2, \dots, M\}$ and G_n satisfies $\Pr\{G_n = 1\} = \gamma$ and $\Pr\{G_n = 0\} = 1 - \gamma$.

We note the following two facts about this construction:

1. The marginal distribution of every $Z_n \in \mathbf{Z}$ is P_0, but the selection is non-i.i.d. as Z_n and $Z_{n'}$ depend on each other through $\mathbf{Y}$. However, when conditioning on the fact that all components of $\mathbf{V} = (V_1, V_2, \dots, V_N)$ are pairwise distinct, then the random variables Y_{V_n} and $Y_{V_{n'}}$ are independent for $n \neq n'$ and thus $\mathbf{Z}$ is a vector of i.i.d. variables distributed according to P_0.
2. When conditioning on $\mathbf{Y} = \mathbf{y}$, we have a different situation, where $Z_n \sim (1-\gamma)P_0 + \gamma\mathcal{Q}_{\mathbf{y}}$ are i.i.d., and by choosing $M \leq \beta K$, we have $(P_0, \mathcal{Q}_{\mathbf{y}}) \in \mathcal{P}$.

Let $|\mathbf{V}| = |\{V_1, V_2, \dots, V_N\}| = N$ be the event that $\mathbf{V}$ contains pairwise distinct elements, i.e., no repetitions occur. Using the first fact above, we calculate

$$\begin{aligned}
&\Pr\{g_2(\mathbf{Z}) = 1\} && (17)\\
&\quad\leq \Pr\left\{g_2(\mathbf{Z}) = 1 \middle| |\mathbf{V}| = N\right\} + \Pr\{|\mathbf{V}| \neq N\} \\
&\quad\leq \Pr\left\{g_2(\mathbf{Z}) = 1 \middle| |\mathbf{V}| = N\right\} + 1 - \frac{M!}{M^N(M-N)!} \\
&\quad\leq \Pr\left\{g_2(\mathbf{Z}) = 1 \middle| |\mathbf{V}| = N\right\} + 1 - \left(1 - \frac{N}{M}\right)^N \\
&\quad= \Pr\left\{g_2(\mathcal{D}^{(0)}) = 1\right\} + 1 - \left(1 - \frac{N}{M}\right)^N && (18)\\
&\quad= 2 - \Pr\{g_2(\mathcal{D}^{(0)}) = 0\} - \left(1 - \frac{N}{M}\right)^N && (19)\\
&\quad\leq 2 - \Pr\{g_2(\mathcal{D}^{(0)}) = 0\} - \exp\frac{-N^2}{M-N}, && (20)
\end{aligned}$$

where we used the union bound as well as the inequality $\log(1+x) \geq \frac{x}{1+x}$.

Using the second fact from above, we condition on $\mathbf{Y} = \mathbf{y}$ and then have $\mathbf{Z}$ i.i.d. according to $P_1 = (1-\gamma)P_0 + \gamma P_b$ for a valid backdoor distribution $P_b = \mathcal{Q}_{\mathbf{y}}$. We then write

$$\begin{aligned}
&\frac{1}{2}\Pr\{g_2(\mathcal{D}^{(0)}) = 0\} + \frac{1}{2}\Pr\{g_2(\mathbf{Z}) = 1\} && (21)\\
&\quad= \frac{1}{2}\Pr\{g_2(\mathcal{D}^{(0)}) = 0\} + \frac{1}{2}K^{-M}\sum_{\mathbf{y}\in\mathcal{X}^M}\Pr\left\{g_2(\mathbf{Z}) = 1 \middle| \mathbf{Y} = \mathbf{y}\right\} && (22)\\
&\quad= K^{-M}\sum_{\mathbf{y}\in\mathcal{X}^M}\left(\frac{1}{2}\Pr\left\{g_2(\mathcal{D}^{(0)}) = 0\right\} + \frac{1}{2}\Pr\left\{g_2(\mathbf{Z}) = 1 \middle| \mathbf{Y} = \mathbf{y}\right\}\right) && (23)\\
&\quad= K^{-M}\sum_{\mathbf{y}\in\mathcal{X}^M}\left(\frac{1}{2}\Pr\left\{g_2(\mathcal{D}^{(0)}) = 0\right\} + \frac{1}{2}\Pr\left\{g_2(\mathcal{D}^{(1)}) = 1 \middle| \mathbf{Y} = \mathbf{y}\right\}\right) && (24)\\
&\quad\geq K^{-M}\sum_{\mathbf{y}\in\mathcal{X}^M}(1-\alpha) && (25)\\
&\quad= 1-\alpha. && (26)
\end{aligned}$$

In total we have

$$\begin{aligned}
1-\alpha &\overset{(26)}{\leq} \frac{1}{2}\Pr\{g_2(\mathcal{D}^{(0)}) = 0\} + \frac{1}{2}\Pr\{g_2(\mathbf{Z}) = 1\} && (27)\\
&\overset{(20)}{\leq} \frac{1}{2}\left(\Pr\{g_2(\mathcal{D}^{(0)}) = 0\} + 2 - \Pr\{g_2(\mathcal{D}^{(0)}) = 0\} - \exp\frac{-N^2}{M-N}\right) && (28)\\
&= 1 - \frac{1}{2}\exp\frac{-N^2}{M-N} && (29)
\end{aligned}$$

and thus

$$\alpha \geq \frac{1}{2}\exp\frac{-N^2}{M-N}. \tag{30}$$

This already resolves the case $|\mathcal{X}| = \infty$ as we can then let $K \to \infty$ and $M = \lfloor \beta K \rfloor \to \infty$, showing that $\alpha = \frac{1}{2}$ for $|\mathcal{X}| = \infty$.

On the other hand, for $|\mathcal{X}| < \infty$, we choose $K = |\mathcal{X}|$, $M = \lfloor \beta K \rfloor$ and obtain (6) by

$$\alpha \geq \frac{1}{2} \exp \frac{-N^2}{M - N} \tag{31}$$

$$-\log 2\alpha \leq \frac{N^2}{\lfloor \beta K \rfloor - N} \tag{32}$$

$$0 \leq N^2 - N \log 2\alpha + \lfloor \beta K \rfloor \log 2\alpha \tag{33}$$

$$N \geq \frac{\log 2\alpha}{2} + \sqrt{\frac{(\log 2\alpha)^2}{4} - \lfloor \beta K \rfloor \log 2\alpha} \tag{34}$$

$$N \geq \frac{\log 2\alpha}{2} + \sqrt{\frac{(\log 2\alpha)^2}{4} + (\beta K - 1) \log \frac{1}{2\alpha}} \tag{35}$$

□

Theorem 2. *Considering the backdoor detection setup of Definition 1 with $\mathcal{P} = \{(P_0, P_b) : \mathrm{TV}(P_0, P_b) \geq 1 - \beta\}$ and a finite alphabet $|\mathcal{X}| < \infty$. If*

$$\alpha > 2|\mathcal{X}| \exp\left(-\frac{2N\gamma^2(1-\beta)^2}{|\mathcal{X}|^2}\right), \tag{7}$$

then there exists an α-error Type 2 detector.

In the proof of this theorem, the auxiliary Lemmas 2 and 3 are used, which are provided in Appendix B.

Proof of Theorem 2. In the following we will show that the detector

$$g(\mathcal{D}, P_0) = \begin{cases} 1 & \mathrm{TV}(P_0, S_N) \geq \gamma \frac{1-\beta}{2} \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

is α-error if (7) is satisfied. Here, the distribution S_N is the so-called *type* of $\mathcal{D}$, i.e.,

$$S_N(x) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_x(X_n), \tag{37}$$

where for any $x \in \mathcal{X}$, $\mathbb{1}_x(X_n)$ is the indicator function that takes value 1 if $X_n = x$ and 0 otherwise.

In Lemma 3 it is shown that the type S_N is close to the true distribution P with high probability. We can now analyze the error probability of the detector (36) for $P = P_1$, i.e.,

$$\Pr\{g(\mathcal{D}^{(1)}, P_0) = 0\} = \Pr\left\{\mathrm{TV}(S_N^{(1)}, P_0) \leq \gamma \frac{1-\beta}{2}\right\} \tag{38}$$

$$\leq \Pr\left\{\mathrm{TV}(S_N^{(1)}, P_0) \leq \frac{\mathrm{TV}(P_0, P_1)}{2}\right\} \tag{39}$$

$$\leq \Pr\left\{\mathrm{TV}(S_N^{(1)}, P_1) \geq \frac{\mathrm{TV}(P_0, P_1)}{2}\right\} \tag{40}$$

$$\leq \Pr\left\{\mathrm{TV}(S_N^{(1)}, P_1) \geq \gamma \frac{1-\beta}{2}\right\} \tag{41}$$

$$\leq 2|\mathcal{X}| \exp\left(-\frac{2N\gamma^2(1-\beta)^2}{|\mathcal{X}|^2}\right), \tag{42}$$

where we used Lemma 3 in (42) and the fact that $\mathrm{TV}(P_0, P_1) = \gamma \,\mathrm{TV}(P_0, P_b) \geq (1-\beta)\gamma$ by Lemma 2 in (39) and (41). Similarly, we obtain that the error probability for $j = 0$ is upper bounded by the same expression

$$\Pr\{g(\mathcal{D}^{(0)}, P_0) = 1\} = \Pr\left\{\mathrm{TV}(S_N^{(0)}, P_0) \geq \gamma \frac{1-\beta}{2}\right\} \tag{43}$$

$$\leq 2|\mathcal{X}| \exp\left(-\frac{2N\gamma^2(1-\beta)^2}{|\mathcal{X}|^2}\right), \tag{44}$$

applying Lemma 3 in (44).

Thus, we have shown that g, as defined in (36), is α-error, provided that (7) holds. □

Theorem 3. *Consider the backdoor detection setup of Definition 1, with fixed $\gamma \in (0, 1]$, $N \in \mathbb{N}$ and some set of possible distributions $\mathcal{P}$. Let $\mathcal{P}'$ be the set of N-fold products of (P_0, P_1), i.e., $\mathcal{P}' = \{(P_0^N, (\gamma P_b + (1-\gamma)P_0)^N) : (P_0, P_b) \in \mathcal{P}\}$. Then, OOD-detection is PAC-learnable on $\mathcal{P}'$ if and only if the following holds for any $\epsilon > 0$ and any Type 3 detector $g_3(\mathcal{D}, P_0, P_b)$: We can find $M \in \mathbb{N}$ and a Type 1 detector $g_1(\mathcal{D}, \mathcal{D}')$, which satisfies $R(g_1, P_0, P_b) \leq R(g_3, P_0, P_b) + \epsilon$ for every $(P_0, P_b) \in \mathcal{P}$.*

In the proof of this theorem, the auxiliary Lemma 4 is used, which is provided in Appendix B

Proof of Theorem 3. Assume first that OOD-detection is PAC-learnable on $\mathcal{P}'$, fix $\epsilon > 0$ and let g_3 be any Type 3 detector. By Lemma 4, we know that there is a Type 1 detector g_1^M with some M such that $\epsilon(M) \leq \epsilon$, satisfying (82). Noting that $\frac{1}{2} - \frac{1}{2}\mathrm{TV}(P_0^N, P_1^N) \leq R(g_3, P_0, P_b)$ by Lemma 1 completes this part of the proof.

On the other hand, let g_3 be the Type 3 Neyman-Pearson detector that satisfies $R(g_3, P_0, P_b) = \frac{1}{2} - \frac{1}{2}\mathrm{TV}(P_0^N, P_1^N)$, which exists by Lemma 1. By our assumptions, for any $k \in \mathbb{N}$, we set $\epsilon = \frac{1}{k}$ and find a Type 1 detector $\hat{g}_1^k$, operating on $\mathcal{D}$ with size $M = m(k)$ satisfying

$$\frac{1}{2}\mathrm{TV}(P_0^N, P_1^N) - \frac{1}{2} + R(\hat{g}_1^k, P_0, P_b) \leq \frac{1}{k}. \tag{45}$$

We can find a monotonically increasing sequence k_M for $M = 1, 2, \ldots$ with $\lim_{M\to\infty} k_M = \infty$, that satisfies $m(k_M) \leq M$. Using the sequence of Type 1 detectors[6] $g_1^M(\mathcal{D}, \mathcal{D}') = \hat{g}_1^{k_M}(\mathcal{D}, [\mathcal{D}']_1^{m(k_M)})$ and $\epsilon(M) = \frac{1}{k_M}$, we have for every $(P_0, P_b) \in \mathcal{P}$,

$$\epsilon(M) \geq \frac{1}{2}\mathrm{TV}(P_0^N, P_1^N) - \frac{1}{2} + R(\hat{g}_1^{k_M}, P_0, P_b) \tag{46}$$

$$= \frac{1}{2}\mathrm{TV}(P_0^N, P_1^N) - \frac{1}{2} + R(g_1^M, P_0, P_b). \tag{47}$$

This completes the proof as $\lim_{m\to\infty} \epsilon(m) = \lim_{m\to\infty} \frac{1}{k_m} = 0$ and thus, PAC learnability is guaranteed by Lemma 4. □

Corollary 2. *Let $g_2'(\mathcal{D}^{(J)}, P_0, X^{(I)})$ be a Type 2 detector for an SBD problem, where we have $r = \min\{P_{JI}(0,0), P_{JI}(1,1)\} > 0$ and $\mathcal{P} = \{(P_0, P_b) : \mathrm{TV}(P_0, P_b) \geq 1-\beta\}$. Then, if g_2' is α-error, we have $\alpha \geq r$ if $|\mathcal{X}| = \infty$, and for $|\mathcal{X}| < \infty$, we obtain*

$$N \geq \frac{\log\frac{\alpha}{r}}{2} + \sqrt{\frac{(\log\frac{\alpha}{r})^2}{4} + (\beta|\mathcal{X}| - 1)\log\frac{r}{\alpha}}. \tag{9}$$

Proof of Corollary 2. Assuming that this detector is α-error implies

$$\begin{aligned}\alpha \geq R(g_2', P_0, P_b) \geq\; & P_{JI}(0,0)\Pr\{g_2'(\mathbf{Q}_2') \neq 0 | J = I = 0\} \\ & + P_{JI}(1,1)\Pr\{g_2'(\mathbf{Q}_2') \neq 1 | J = I = 1\} \end{aligned} \tag{48}$$

$$\geq r\big(\Pr\{g_2'(\mathbf{Q}_2') \neq 0 | J = I = 0\} + \Pr\{g_2'(\mathbf{Q}_2') \neq 1 | J = I = 1\}\big). \tag{49}$$

[6] We use the notation $[\mathbf{x}]_k^l = [(x_1, x_2, \ldots, x_N)]_k^l = (x_k, x_{k+1}, \ldots, x_l)$ for slicing.

Now consider the MBD problem with $\gamma = 1$ and the training set size $N' = N + 1$. We can define a Type 2 detector[7] $g_2(\mathcal{D}, P_0) = g_2'(\mathcal{D}, P_0, X_{N'})$ with risk

$$R(g_2, P_0, P_b) = \frac{1}{2} \Pr\{g_2'(\mathbf{Q}_2') \neq 0 | J = I = 0\} + \frac{1}{2} \Pr\{g_2'(\mathbf{Q}_2') \neq 1 | J = I = 1\} \tag{50}$$

$$\leq \frac{1}{2r}\alpha. \tag{51}$$

From Theorem 1, we now know that $\frac{1}{2r}\alpha \geq \frac{1}{2}$ if $\mathcal{X} = \mathbb{N}$ and obtain (9) for $|\mathcal{X}| < \infty$. □

B Auxiliary Results

This appendix contains auxiliary results, which are utilized in the proofs provided in Appendix A.

Lemma 2 (Properties of Total Variation)**.** *The total variation between two probability distributions $P_0, P_1 \in \mathcal{P}(\mathcal{X})$, is given by*

$$\mathrm{TV}(P_0, P_1) = \|P_0 - P_1\|_{\mathrm{TV}} := \sup_A |P_0(A) - P_1(A)|, \tag{52}$$

where the supremum is over all measurable sets $A \subseteq \mathcal{X}$. We then have

$$\|P_0 - P_1\|_{\mathrm{TV}} = 2 \inf_{X_0, X_1 : P_{X_0} = P_0, P_{X_1} = P_1} \Pr\{X_0 \neq X_1\}, \tag{53}$$

where the infimum is over all random variables X_0, X_1 on $\mathcal{X}$, such that the marginal distributions satisfy $P_{X_0} = P_0$, $P_{X_1} = P_1$. For $P_0', P_1' \in \mathcal{P}(\mathcal{Y})$, we have

$$\|P_0 - P_1\|_{\mathrm{TV}} \leq \|P_0 \times P_0' - P_1 \times P_1'\|_{\mathrm{TV}} \leq \|P_0 - P_1\|_{\mathrm{TV}} + \|P_0' - P_1'\|_{\mathrm{TV}}. \tag{54}$$

and thus $\|P_0 - P_1\|_{\mathrm{TV}} \leq \|P_0^N - P_1^N\|_{\mathrm{TV}} \leq N\|P_0 - P_1\|_{\mathrm{TV}}$. Furthermore, for $\gamma \in [0, 1]$,

$$\|P_0 - (1-\gamma)P_0 - \gamma P_1\|_{\mathrm{TV}} = \gamma \|P_0 - P_1\|_{\mathrm{TV}} \tag{55}$$

Proof. The characterization (53) can be found in (Villani, 2021).

To show the first inequality in (54), observe that

$$\|P_0 \times P_0' - P_1 \times P_1'\|_{\mathrm{TV}} = \sup_B |[P_0 \times P_0'](B) - [P_1 \times P_1'](B)| \tag{56}$$

$$\geq \sup_A |[P_0 \times P_0'](A \times \mathcal{Y}) - [P_1 \times P_1'](A \times \mathcal{Y})| \tag{57}$$

$$= \|P_0 - P_1\|_{\mathrm{TV}}. \tag{58}$$

To show the second inequality in (54), we use (53) and for an arbitrary $\varepsilon > 0$, choose $(X_0, X_1) \perp (Y_0, Y_1)$ such that $P_{X_0} = P_0$, $P_{X_1} = P_1$, $P_{Y_0} = P_0'$, $P_{Y_1} = P_1'$, and

$$\|P_0 - P_1\|_{\mathrm{TV}} + \varepsilon \geq 2 \Pr\{X_0 \neq X_1\}, \tag{59}$$

$$\|P_0' - P_1'\|_{\mathrm{TV}} + \varepsilon \geq 2 \Pr\{Y_0 \neq Y_1\}. \tag{60}$$

Clearly $P_{X_0, Y_0} = P_0 \times P_0'$ as well as $P_{X_1, Y_1} = P_1 \times P_1'$ and thus by (53),

$$\|P_0 \times P_0' - P_1 \times P_1'\|_{\mathrm{TV}} \leq 2 \Pr\{(X_0, Y_0) \neq (X_1, Y_1)\} \tag{61}$$

$$\leq 2 \Pr\{X_0 \neq X_1\} + 2 \Pr\{Y_0 \neq Y_1\} \tag{62}$$

$$\leq \|P_0 - P_1\|_{\mathrm{TV}} + \|P_0' - P_1'\|_{\mathrm{TV}} + 2\varepsilon. \tag{63}$$

As $\varepsilon > 0$ was arbitrary, this proves (54).

[7] If $\gamma > 0$ for the SBD problem, randomly replace elements of $\mathcal{D}$ by independently drawn realizations of P_0.

To show (55), we use (52) and have

$$\|P_0 - (1-\gamma)P_0 - \gamma P_1\|_{\mathrm{TV}} = \sup_A |P_0(A) - (1-\gamma)P_0(A) - \gamma P_1(A)| \tag{64}$$

$$= \sup_A |\gamma P_0(A) - \gamma P_1(A)| \tag{65}$$

$$= \gamma \|P_0 - P_1\|_{\mathrm{TV}}. \tag{66}$$

□

Lemma 3. *Let S_N be the type of $\mathbf{X} = (X_1, X_2, \dots, X_N)$, distributed according to P^N. For any $t \in [0,1]$, we then have the bound*

$$\Pr\{\mathrm{TV}(S_N, P) \geq t\} \leq 2|\mathcal{X}| \exp\left(-\frac{8Nt^2}{|\mathcal{X}|^2}\right). \tag{67}$$

Proof. By using the Hoeffding's inequality we can bound the probability of the deviation of S_N from its expected value. In particular, we have that

$$\Pr\{|S_N(x) - P(x)| \geq t\} = \Pr\{|S_N(x) - \mathbb{E}[S_N(x)]| \geq t\} \tag{68}$$

$$\leq 2\exp\left(\frac{-2t^2}{\sum_{n=1}^{N}(\frac{1}{N} - 0)^2}\right) \tag{69}$$

$$= 2\exp\left(\frac{-2t^2}{\frac{1}{N}}\right) \tag{70}$$

$$= 2\exp\left(-2Nt^2\right), \tag{71}$$

where we note that $\mathbb{E}[S_N(x)] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\mathbb{1}_x[X_n]] = P(x)$.

The next and final step is to extend the bound to the whole alphabet $\mathcal{X}$. In order to do so, we define the event $\mathcal{A}_x = \{|S_N(x) - P(x)| \geq t\}$. We want to bound the probability of the event

$$\mathcal{A} = \bigcup_{x \in \mathcal{X}} \mathcal{A}_x = \{\exists x \in \mathcal{X} : \mathcal{A}_x\}. \tag{72}$$

By applying the union bound we obtain

$$\Pr \mathcal{A} = \Pr\left\{\bigcup_{x\in\mathcal{X}} \mathcal{A}_x\right\} \tag{73}$$

$$\leq \sum_{x\in\mathcal{X}} \Pr\{\mathcal{A}_x\} \tag{74}$$

$$\leq \sum_{x\in\mathcal{X}} 2\exp\left(-2Nt^2\right) \tag{75}$$

$$= 2|\mathcal{X}|\exp\left(-2Nt^2\right). \tag{76}$$

Let us consider the event $\mathcal{A} = \{\exists x \in \mathcal{X} : |S_N(x) - P(x)| \geq t\}$: this is the error event, i.e., the divergence between the observed samples frequency and its expected value diverges more than a given value $t > 0$ for at least one $x \in \mathcal{X}$. The complement of this event is the event that the divergence is less than t for all $x \in \mathcal{X}$, i.e., the event that the observed frequency is close to the expected value for all $x \in \mathcal{X}$. This can be written as

$$\mathcal{A}^c = \{\forall x \in \mathcal{X},\ |S_N(x) - P(x)| < t\}. \tag{77}$$

Now, $\mathcal{A}^c$ implies that

$$\sum_{x\in\mathcal{X}} |S_N(x) - P(x)| < t|\mathcal{X}| \tag{78}$$

$$\frac{1}{2}\sum_{x\in\mathcal{X}}|S_N(x)-P(x)|<\frac{1}{2}t|\mathcal{X}| \tag{79}$$

$$\mathrm{TV}(S_N,P)<t' \tag{80}$$

where $t'=\frac{1}{2}t|\mathcal{X}|$. Thus, $\Pr\mathcal{A}^c\le\Pr\{\mathrm{TV}(S_N,P)<t'\}$ and therefore

$$\Pr\{\mathrm{TV}(S_N,P)\ge t'\}\le\Pr\mathcal{A}\le 2|\mathcal{X}|\exp\left(-2Nt^2\right), \tag{81}$$

where we have used (76). By writing t in terms of t' in (81), we obtain (67). □

Lemma 4. *Given $\mathcal{P}$ and $N\in\mathbb{N}$ and letting $\gamma\in(0,1]$, OOD-detection is PAC-learnable on $\mathcal{P}'=\{(P_0^N,P_1^N):(P_0,P_b)\in\mathcal{P}\}$ with $P_1=(1-\gamma)P_0+\gamma P_b$ if and only if the following holds: For the MBD problem, there exists a sequence of Type 1 backdoor detectors $g_1^M(\mathcal{D},\mathcal{D}')$ for $M=1,2,\dots$ and a decreasing sequence $\epsilon(m)$ with $\lim_{m\to\infty}\epsilon(m)=0$ such that for any $M\in\mathbb{N}$ and any pair $(P_0,P_b)\in\mathcal{P}$, we have*

$$\frac{1}{2}\mathrm{TV}(P_0^N,P_1^N)-\frac{1}{2}+R(g_1^M,P_0,P_b)\le\epsilon(M). \tag{82}$$

Proof. Assume that OOD-detection is PAC-learnable on $\mathcal{P}'$. By definition we can find a function $\mathcal{G}\colon\bigcup_{m=1}^{\infty}\mathcal{X}^{Nm}\to\{0,1\}^{\mathcal{X}^N}$ and a monotonically decreasing sequence $\epsilon'(m)$ that tends to zero and satisfies for every $(P_0,P_b)\in\mathcal{P}$, $m\in\mathbb{N}$, that

$$\mathbb{E}[\bar{R}(\mathcal{G}(\mathcal{D}'),P_0^N,P_1^N)]-\inf_f\bar{R}(f,P_0^N,P_1^N)\le\epsilon'(m), \tag{83}$$

where we set the size of $\mathcal{D}'$ to be $M=mN$ and the infimum is over all functions $f\colon\mathcal{X}^N\to\{0,1\}$.

For any $M\in\mathbb{N}$, we define[8] $g_1^M(\mathcal{D},\mathcal{D}'):=\mathcal{G}([\mathcal{D}']_1^{mN})(\mathcal{D})$ as well as $\epsilon(M)=\epsilon'(m)$, where m is the largest integer such that $mN\le M$. Notice that $R(g_1^M,P_0,P_b)=\mathbb{E}[\bar{R}(\mathcal{G}([\mathcal{D}']_1^{mN}),P_0^N,P_1^N)]$ and that $\inf_f\bar{R}(f,P_0^N,P_1^N)=\frac{1}{2}-\frac{1}{2}\mathrm{TV}(P_0^N,P_1^N)$ by Lemma 1. We thus obtain from (83), that for any $M\in\mathbb{N}$,

$$\epsilon(M)=\epsilon'(m)\ge\mathbb{E}[\bar{R}(\mathcal{G}([\mathcal{D}']_1^{mN}),P_0^N,P_1^N)]-\frac{1}{2}+\frac{1}{2}\mathrm{TV}(P_0^N,P_1^N) \tag{84}$$

$$=R(g_1^M,P_0,P_b)-\frac{1}{2}+\frac{1}{2}\mathrm{TV}(P_0^N,P_1^N). \tag{85}$$

Noting that $\epsilon(M)$ approaches zero completes this part of the proof.

On the other hand, assume that $g_1^M(\mathcal{D},\mathcal{D}')$ and $\epsilon(M)$ satisfy the requirement (82). For any $m\in\mathbb{N}$, we then set $M=mN$ and define $\mathcal{G}(\mathcal{D}')(\mathcal{D}):=g_1^M(\mathcal{D},\mathcal{D}')$ as well as $\epsilon'(m)=\epsilon(M)$. We can now rewrite (82) using $\mathbb{E}[\bar{R}(\mathcal{G}(\mathcal{D}'),P_0^N,P_1^N)]=R(g_1^M,P_0,P_b)$ and Lemma 1 to obtain

$$\epsilon'(m)=\epsilon(mN)\ge\frac{1}{2}\mathrm{TV}(P_0^N,P_1^N)-\frac{1}{2}+R(g_1^M,P_0,P_b) \tag{86}$$

$$=\mathbb{E}[\bar{R}(\mathcal{G}(\mathcal{D}'),P_0^N,P_b^N)]-\inf_f\bar{R}(f,P_0^N,P_1^N). \tag{87}$$

Thus, we have shown that the algorithm $\mathcal{G}$ and the sequence ϵ' satisfy Definition 2 and OOD-detection is PAC-learnable on $\mathcal{P}'$. □

C Additional Results

Lemma 5. *Let g_l be a detector as listed in Section 2 with input $\mathbf{Q}_l$ for $l\in\{0,1,2\}$, where we set $g_0=g$ and $\mathbf{Q}_0=\mathbf{Q}$. If g_l is α-error in the sense of Definition 1, then for $m\in\{1,2,3\}$ and $m>l$ we can find a backdoor detector g_m with input $\mathbf{Q}_m$ that is also α-error.*

[8] We use the notation $[\mathbf{x}]_k^l=[(x_1,x_2,\dots,x_N)]_k^l=(x_k,x_{k+1},\dots,x_l)$ for slicing.

Proof. It is sufficient to show the lemma for $m = l+1$. The claim then follows by applying the result repeatedly.

In the case $l = 2$ (and $m = 3$) we obtain g_3 with $R(g_3, P_0, P_b) = R(g_2, P_0, P_b)$ by $g_3(\mathcal{D}, P_0, P_b) = g_2(\mathcal{D}, P_0)$.

For $l = 1$, we can define the randomized detector $g_2(\mathcal{D}, x, P_0)$ to first draw M i.i.d. samples $\mathcal{D}' \sim P_0^M$ and then yield $g_2(\mathcal{D}, P_0) = g_1(\mathcal{D}, \mathcal{D}')$.

Finally, for $l = 0$ we obtain g_1 with equal risk by defining $g_1(\mathcal{D}, \mathcal{D}') = g(\mathcal{A}(\mathcal{D}), \mathcal{D}')$. $\square$