Understanding Backdoor Attacks through the Adaptability Hypothesis

Xun Xian [*1] Ganghua Wang [*2] Jayanth Srinivasa [3] Ashish Kundu [3] Xuan Bi [4] Mingyi Hong [1] Jie Ding [2]

Abstract

A poisoning backdoor attack is a rising security concern for deep learning. This type of attack can result in the backdoored model functioning normally most of the time but exhibiting abnormal behavior when presented with inputs containing the backdoor trigger, making it difficult to detect and prevent. In this work, we propose the *adaptability hypothesis* to understand when and why a backdoor attack works for general learning models, including deep neural networks, based on the theoretical investigation of classical kernel-based learning models. The *adaptability hypothesis* postulates that for an effective attack, the effect of incorporating a new dataset on the predictions of the original data points will be small, provided that the original data points are distant from the new dataset. Experiments on benchmark image datasets and state-of-the-art backdoor attacks for deep neural networks are conducted to corroborate the hypothesis. Our finding provides insight into the factors that affect the attack's effectiveness and has implications for the design of future attacks and defenses.

1. Introduction

Recent years have seen significant growth in Deep Learning (DL) research, resulting in successful real-world applications such as autonomous driving and disease diagnosis (Grigorescu et al., 2020; Oh et al., 2020). However, studies have revealed that deep neural networks (DNNs) are susceptible to various adversarial attacks including adversarial examples (Szegedy et al., 2013; Goodfellow et al., 2014) and data poisoning attacks (Koh & Liang, 2017; Weber et al., 2020). As DL is increasingly used in safety-critical applications where a wrong decision can have severe consequences, the security and trustworthiness of these systems have become critical concerns (Dreossi et al., 2019).

The recent emergence of poisoning backdoor attacks, which exploit the learning capabilities of DNNs, has further highlighted the need to ensure DL systems are robust and trustworthy. Poisoning backdoor attacks involve the insertion of manipulated data samples, containing specific triggers such as a square patch, into the training data, which are then labeled with a specific target class. This causes a DNN trained on the poisoned dataset to learn both the patterns present in the original data and the artificial relationship between the manipulated samples and the target class. As a result, the DNN behaves maliciously when presented with test input that includes the trigger, but functions normally for input without the trigger, making the attack more difficult to detect as it does not impact normal inputs.

Empirical studies have demonstrated that poisoning backdoor attacks can be very successful with human-imperceptible triggers and a relatively small proportion of injected backdoor training data (Gu et al., 2017; Chen et al., 2017; Turner et al., 2019; Zhao et al., 2020; Nguyen & Tran, 2020; Bagdasaryan & Shmatikov, 2021; Doan et al., 2021a;b; Qi et al., 2022; Souri et al., 2022), such that poisoned models can attain good test accuracy on both clean and backdoor data. However, the understanding of when and how these attacks are effective is still limited. For instance, what is the minimum size of triggers to achieve the desired accuracy, and what is the most efficient pattern of the trigger? Gaining a deeper understanding of these attacks could provide principles for researchers and practitioners to build more secure, robust, and trustworthy DL applications.

1.1. Our contributions.

Theoretical analyses of backdoor attacks under classical machine learning context. We analyze the effectiveness of backdoor attacks for classical machine learning models, such as kernel smoothing methods (Audibert & Tsybakov, 2007; Devroye et al., 2013). Our findings indicate that larger triggers, as measured by certain norms, result in more effective attacks. Additionally, under certain conditions, we observed that the most effective directions for adding a backdoor trigger are those dimensions of the data with low variances.

[*] Equal contribution [1] Department of ECE, University of Minnesota [2] School of Statistics, University of Minnesota [3] Cisco Research [4] Carlson School of Management, University of Minnesota. Correspondence to: Jie Ding <dingj@umn.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

An adaptability hypothesis concerning when and how a backdoor attack works under general machine/deep learning context. Building on our theoretical understandings, we articulate the Adaptability hypothesis to explain the effectiveness of backdoor attacks on general learning models, including DNNs. Intuitively, the *adaptability hypothesis* states that when a learning model trained on dataset $\mathcal{D}$ is updated with another set of arbitrarily labeled dataset $\mathcal{D}'$, the change between the outputs of the pre-updated and post-updated models for a typical data point z from the distribution of $\mathcal{D}$ will be small if z is relatively distant from $\mathcal{D}'$ and substantial if z is located close to $\mathcal{D}'$.

The hypothesis implies that incorporating an additional dataset, $\mathcal{D}'$, will have a limited or substantial effect on the predictions of a majority of data from the original distribution, depending on their distance from $\mathcal{D}'$. It is crucial to note that the terms 'far' and 'close' should be determined using an appropriate metric, which may not necessarily be restricted to Euclidean distances. A natural outcome of the adaptability hypothesis is that to minimize the influence on the predictions of the majority of data from the original distribution, it is advisable to place the backdoor data as far away as possible. Additionally, if the additional dataset is manipulated to suit the attacker's objectives, it can result in a successful backdoor attack with high accuracy on both clean and backdoor test data.

Experiments on computer vision benchmarks for validating the Adaptability hypothesis. We performed experimental studies on computer vision datasets using state-of-the-art (SOTA) backdoor attacks (Gu et al., 2017; Qi et al., 2022) on CNNs to test the *Adaptability* hypothesis, as shown in Figure 1. Our findings suggest that the Adaptability hypothesis holds true across a range of different experimental configurations, thereby providing evidence for the validity and practical applicability of this hypothesis.

Implications on creating future backdoor attacks/defenses. We provide implications of the Adaptability hypothesis in the design of new, effective attacks/defenses. For instance, our theoretical findings have led to the proposal of a new data representation that may have the potential to improve the performance of existing defense mechanisms (Chen et al., 2018; Tran et al., 2018) against backdoor attacks. By utilizing popular visualization tools such as PCA and TSNE (Van der Maaten & Hinton, 2008), this representation allows for the clear differentiation between clean and backdoor data generated by state-of-the-art methods (Qi et al., 2022), which is not possible with the original data representation. While further research is required to fully evaluate the potential of this representation, it may lead to new defense strategies that are more effective in detecting and filtering out backdoor samples than existing methods (Chen et al., 2018; Tran et al., 2018) that operate on the original data space.

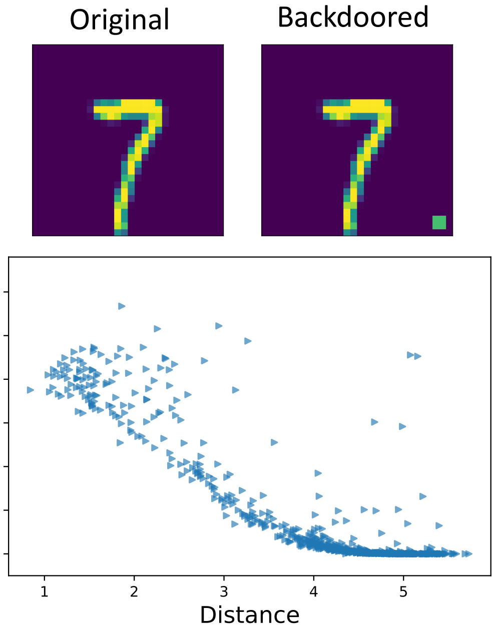


Figure 1. Bottom row: An illustration of the ResNet (He et al., 2016) satisfies the adaptability hypothesis with the MNIST dataset. A subset of original images labeled as seven is manipulated by adding a patch in the lower-right corner of the images (top row of figures) and re-labeled as 0. A pre-trained ResNet is then fine-tuned using these manipulated images. The distance (details in Section 5) between the clean training data points labeled as 7, and the backdoor data distribution, as well as the change in predicted values for the clean training data before and after fine-tuning the ResNet, are plotted. The results show that for data points that are close to the backdoor distribution, the changes in their predicted values are significant, while changes in predicted values are small for data points that are far from the backdoor distribution. These observations support the proposed Adaptability Hypothesis.

1.2. Related Work

Data poisoning attacks. Classical data poisoning attacks, such as the manipulation of training features (Biggio et al., 2012; Koh & Liang, 2017; Jagielski et al., 2018; Weber et al., 2020; Jagielski et al., 2021), aimed to impair the overall prediction performance of machine learning models by corrupting the training data. For instance, the contamination of training data was used to reduce the overall accuracy of supported vector machines (Biggio et al., 2012). However, while data poisoning is also employed in common backdoor attacks (Gu et al., 2017; Chen et al., 2017; Turner et al., 2019; Li et al., 2021a; Qi et al., 2022), there are distinctions between these attacks and traditional data poisoning attacks. Specifically, backdoor attacks target specific tasks and only manifest malicious behavior when triggered, while preserving overall test accuracy on primary tasks.

Backdoor attacks. Backdoor attacks in machine/deep learning aim to manipulate the predictions of a learning model on specific inputs while having no impact on normal inputs. There are various methods to implement backdoor attacks, the approach considered in this paper is via data poisoning, where the attacker manipulates the training data (Gu et al., 2017; Chen et al., 2017; Turner et al., 2019; Li et al., 2021a; Qi et al., 2022). Other methods involve the attacker having full control over the training pipeline and modifying the parameters of the learning models (Nguyen & Tran, 2020; Bagdasaryan & Shmatikov, 2021; Doan et al., 2021a;b).

Theoretical understandings towards backdoor attacks. To our best knowledge, (Manoj & Blum, 2021) is the only attempt to provide theoretical analysis for poisoning backdoor attacks. The authors quantify the vulnerability of machine learning models by the capacity to memorize out-of-distribution values, which is similar to VC-dimension (Vapnik et al., 1994). They proved that the necessary and sufficient condition to find a successful poisoning backdoor attack under their formulation is a non-zero memorization capacity. We note that the definition for a successful attack in (Manoj & Blum, 2021) is different from ours. In particular, (Manoj & Blum, 2021) require a perfect fitting for the training data, which is often unlikely, while we only require the models learned before and after the attack to have similar performance for clean data. Additionally, we consider general learning algorithms, including non-parametric methods that (Manoj & Blum, 2021) cannot deal with.

2. Backgrounds and Formulations

Notations. In this paper, we consider backdoor attacks in the context of binary classification problems. Let $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ be a random variable with probability distribution $\mathbb{P}_{XY}$. We observe the original/clean training data $\mathcal{D}^{\text{cl}} = \{(X_i, Y_i)\}_{i=1}^{n_{\text{cl}}}$, which are independently drawn from $\mathbb{P}_{XY}$. The loss function for evaluating the prediction performance is denoted by $\ell(\cdot, \cdot) : [0, 1] \times \{0, 1\} \to \mathbb{R}$. For two sequence of real positive numbers a_n and b_n, we use $a_n = O(b_n)$ to denote $\lim_{n\to\infty} a_n/b_n < \infty$. For a vector x, we use $\|x\|$ to denote its ℓ_2-norm $\|x\| = (\sum_{i=1}^d |x_i|^2)^{1/2}$. The indicator function is denoted as $\mathbf{1}\{\cdot\}$.

Threat Model. We assume that attackers can only modify the training data. This represents the minimal requirement for backdoor attackers and is relevant to various real-world scenarios (Li et al., 2020). Without loss of generality, we assume that the training data is rearranged so that the first $n_1 < n_{\text{cl}}$ data points $\{X_i\}_{i=1}^{n_1}$ have ground-truth label 1, denoted as $\mathcal{D}^1 \triangleq \{(X_i, Y_i = 1)\}_{i=1}^{n_1}$, and the remaining data points have ground-truth labels 0. To create the backdoor data, the attacker selects first $n_{\text{bd}} = n_1\alpha_{\text{poi}}$ points in $\mathcal{D}^1$, where $\alpha_{\text{poi}} \in (0, 1)$, adds a trigger $\eta \in \mathbb{R}^d$ and changes their label to 0, resulting in $\mathcal{D}_\eta^{\text{bd}} = \{(X_i + \eta, 0)\}_{i=1}^{n_{\text{bd}}}$. It is important to note that the added trigger has a carefully crafted pattern instead of random noise. Next, the attacker will inject the backdoor data into the clean data to obtain a joint dataset $\mathcal{D}^{\text{poi}} \triangleq \mathcal{D}^{\text{cl}} \cup \mathcal{D}_\eta^{\text{bd}}$ for future training, with sample size $n \triangleq n_{\text{cl}} + n_{\text{bd}}$.

A learner, e.g., a third-party cloud server, will apply a supervised learning procedure $\mathcal{A}$ on the joint dataset to obtain a prediction function $f_{\mathcal{A}}^{\mathcal{D}^{\text{poi}}}$ for its downstream tasks. A learning procedure $\mathcal{A}$ is a mapping from a dataset $\mathcal{D}$ to a prediction function $f_{\mathcal{A}}^{\mathcal{D}} : \mathbb{R}^d \to \mathbb{R}$. Common examples of $\mathcal{A}$ include logistic regressions, kernel smoothing estimators/k-nearest neighbors, and DNNs. We assume $f_{\mathcal{A}}^{\mathcal{D}}$ take value in $[0, 1]$, which is interpreted as the estimated probability of the input having label 1. To obtain the one-hot encoding label, one can apply the classical decision rule of $\mathbf{1}\{f_{\mathcal{A}}^{\mathcal{D}}(x) > 0.5\}$.

Whenever it is clear from the text, we use shorthand notions of f^{cl}, f^{poi}, and f^{pbd} to represent the prediction functions obtained on dataset $\mathcal{D}^{\text{cl}}$, $\mathcal{D}^{\text{poi}}$, and $\mathcal{D}_\eta^{\text{bd}}$, respectively. Note that $f^{\text{pbd}}(x) = 0$ for all $x \in \mathbb{R}^d$ since the backdoor target label is set to zero.

Backdoor Attacks Goals. Given a clean training dataset $\mathcal{D}^{\text{cl}}$ and a learning procedure $\mathcal{A}$, a poisoning backdoor attack aims to select a backdoor trigger $\eta \in \mathbb{R}^d$ to minimize the following:

- **Performance gap for clean test data:**

  $$R_n^{\text{cl}} \triangleq \mathbb{E}^{\text{poi}}\mathbb{E}_{(X,Y)\sim\mathbb{P}_{XY}}[\ell(f^{\text{poi}}(X), Y) - \ell(f^{\text{cl}}(X), Y)],$$

  where $\mathbb{E}^{\text{poi}}$ is taken with respect to $\mathcal{D}^{\text{poi}}$, and $\mathbb{E}_{(X,Y)}$ is taken over the clean test data, which is a future input independently drawn from $\mathbb{P}_{XY}$.

- **Performance gap for backdoor test data:**

  $$R_n^{\text{bd}} \triangleq \mathbb{E}^{\text{poi}}\mathbb{E}_{X\sim\mathbb{P}_1^\eta}[\ell(f^{\text{poi}}(X), 0) - \ell(f^{\text{pbd}}(X), 0)],$$

  where $\mathbb{P}_1^\eta$ is the distribution of $X|Y = 1$ shifted by η, namely the backdoor data. The zero in the equation is actually $f^{\text{pbd}}(x)$, the true label for backdoor data.

A successful backdoor attack means both R_n^{cl} and R_n^{bd} are small.

3. Theoretical Insights

In this section, we aim to identify the factors that contribute to the effectiveness of backdoor attacks in the realm of classical machine learning, which further leads to insights for general learning algorithms. In particular, we derive theoretical results for backdoor attacks under kernel smoothing algorithms.

We begin with a brief overview of kernel smoothing estimation. Essentially, it estimates the output at a point x

by taking the weighted average of nearby observed data points. The Nadaraya-Waston (NW) kernel smoothing estimator (Audibert & Tsybakov, 2007; Devroye et al., 2013). for a given point x using the $\mathcal{D}^{\text{poi}}$ is given by:

$$f^{\text{poi}}(x) = \frac{\sum_{(X,Y)\in\mathcal{D}^{\text{poi}}} K(\frac{x-X}{h_n}) \cdot Y}{\sum_{(X,Y)\in\mathcal{D}^{\text{poi}}} K(\frac{x-X}{h_n})},$$

where $K(u) = \mathbf{1}\{\|u\| \leq 1\}$ is the box kernel and $h_n > 0$ is the bandwidth. The bandwidth h_n regulates the level of smoothing applied. For instance, a larger bandwidth leads to the inclusion of more data points from the vicinity, yielding a more averaged prediction. Additionally, we follow the conventions to define $0/0 = 0$.

Kernel smoothing estimation admits effective backdoor attacks. Intuitively speaking, from the expressions above, kernel smoothing estimation relies only on labels of nearby points in making predictions. Thus, when the backdoor data is positioned far from the clean data, the predictions for typical clean data points will largely remain unchanged between f^{cl} and f^{poi}. Similarly, in the region of typical backdoor data, the function f^{poi} is likely to output 0 (as per definition) as it is substantially distant from clean data, thereby achieving both aims of the backdoor attack.

The following results quantify the above intuition. For illustrating purposes, we assume that the conditional distributions of $X|Y = 0$ and $X|Y = 1$ are d-dimensional normal distributions with parameters (μ_0, Σ) and (μ_1, Σ), respectively, where $\mu_1 + \mu_0 = 0$ and Σ is a diagonal matrix with diagonal elements $\lambda_{d-1} \geq \ldots \geq \lambda_0 > 0$. It is worth noting that these results can be extended to other distributions with well-behaved tails, such as exponential distributions. We provide a simplified statement as follows. Detailed results and proofs are included in Appendix A.

Theorem 3.1 (Effectiveness of backdoor attacks with kernel smoothing)**.** *Suppose that the loss function is Lipschitz. Given any sufficiently large $\|\eta\|$, with appropriate selections of the bandwidth h_n, we have*

$$R_n^{cl} \leq O(\exp{(-C_1\|\eta\|^2/\lambda_{d-1})} + \alpha_{poi}\exp{(-\|\eta\|^2/\lambda_0)}),$$

and

$$R_n^{bd} \leq O(\exp{(-C_2\|\eta\|^2/\lambda_{d-1})} + \alpha_{poi}^{-1}\exp{(-\|\eta\|^2/\lambda_0)}),$$

where $\alpha_{poi} \in (0, 1)$ is the backdoor poisoning rate, and C_1, C_2 are constants of $\mathbb{P}_0$ and $\mathbb{P}_1$.

We first explain the implications of the result and then provide a proof sketch. Theorem 3.1 implies that as the size of the backdoor perturbation η increases, the maximum performance gaps in clean and backdoor test data decreases. That is, the backdoor attack will be more successful as the poisoned data is positioned farther away. With a sufficiently large perturbation, the attacker can expect f^{poi} to provide near-perfect prediction performance on both clean and backdoor data, compared to their counterparts f^{cl} and f^{pbd}. The above results also depend on the poisoning rate $\alpha_{\text{poi}} = n_{\text{bd}}/n_1$. For instance, a larger value of α_{poi} leads to a smaller upper bound for R_n^{cl}. This makes sense because a small number of backdoor samples compared to the size of clean data may not significantly impact the predictions. On the other hand, a smaller α_{poi} results in a larger upper bound for R^{bd}, indicating that f^{poi} is far from f^{pbd} on average.

Overview of Proof. We outline the proof for R_n^{cl}, and the same reasoning can be used for R_n^{bd}. The key idea is that the kernel smoothing estimation relies on neighbor data points. We partition the input space $\mathbb{R}^d$ into two subregions: a region C_r around the mean of the test distribution $\mathbb{P}_X$ with radius $r > 0$ and its complement. The gap in the region $\mathbb{R}^d \setminus C_r$ is controlled by the probability of clean data falling into this region, which decreases quickly under the assumption that the tail of $\mathbb{P}_X$ is well-behaved. Determining the gap over C_r requires further reasoning. We show that the predicted value's change within this region is inversely related to its distance to the backdoor distribution. For narrative clarity, we use $\text{Diff}_x(f^{\text{cl}}, f^{\text{poi}}) \triangleq \mathbb{E}^{\text{poi}}|f^{\text{cl}}(x) - f^{\text{poi}}(x)|$ to track the difference between outputs of $f^{\text{cl}}/f^{\text{pbd}}$ and f^{poi} at x.

Lemma 3.2 (Change in predicted values inversely related to distance)**.** *We follow the same setup in Theorem 3.1. For any given $r, s > 0$, denote $C_r = \{x \in \mathbb{R}^d \mid \|x\| \leq r\}$ and $B_{s,\eta} = \{x \in \mathbb{R}^d \mid \|x - \mu_1 - \eta\| \leq s\}$ to be two sets representing the typical data from the clean and backdoor distribution, respectively. For each $x \in C_r$, we have*

$$\mathit{Diff}_x(f^{cl}, f^{poi}) \leq O(\frac{\exp{(-d(x, \mathbb{P}_1^\eta))}}{\exp{(-d(x, \mathbb{P}_1^\eta))} + (\alpha_{poi}^{-1} - 1)T_{\mu_1,r}}),$$

where $d(x, \mathbb{P}_1^\eta) \triangleq (x - \mu_1 - \eta)^\top\Sigma^{-1}(x - \mu_1 - \eta)$ and $T_{\mu_1,r} = \exp{(-2(r^2 + \|\mu_1\|^2)/\lambda_0)}$, and for each $x \in B_{s,\eta}$,

$$\mathit{Diff}_x(f^{pbd}, f^{poi}) \leq O(\frac{\exp(-d(x, \mathbb{P}_1))}{\exp(-d(x, \mathbb{P}_1)) + \alpha_{poi}Q_s}),$$

where $d(x, \mathbb{P}_1) \triangleq (x - \mu_1)^\top\Sigma^{-1}(x - \mu_1)$ and $Q_s = \exp{(-\|s\|^2/\lambda_0)}$.

We interpret the result from the perspective of clean data, but similar arguments apply to the perspective of backdoor data. Specifically, we observe that for each point in $x \in C_r$, the change between f^{cl} and f^{poi} decrease as its Mahalanobis distance (Mahalanobis, 1936) $d(x, \mathbb{P}_1^\eta)$ increases. We take the expectation of $\mathbb{E}^{\text{poi}}$ to ensure that this property holds for the majority of replications on the training data, thereby eliminating the possibility of this being an exceptional case due to a specific realization of the training data $\mathcal{D}^{\text{poi}}$.

Remark 3.3 (Effective Directions for Backdoor Triggers). For a given point $x \in C_r$, the larger Mahalanobis distance $d(x, \mathbb{P}_1^\eta) = (x-\eta)^\top \Sigma^{-1}(x-\eta)$ (assuming $\mu_1 = 0$), the smaller the difference between between the outputs of $f^{\mathrm{cl}}(x)$ and $f^{\mathrm{poi}}(x)$. Given the magnitude of the backdoor trigger, $\|\eta\|$, it can be observed that η concentrated on the direction of $\mathbb{P}_1$ with the lowest variance will maximize the Mahalanobis distance $d(x, \mathbb{P}_1^\eta)$. In other words, given a fixed magnitude of $\|\eta\|$, the most effective way to add a backdoor trigger is in the dimension of data with low variances.

4. The Adaptability Hypothesis

Inspired by the theoretical insights from Theorem 3.1, we propose the Adaptability hypothesis for DNNs in this section. Recall that the efficacy of kernel smoothing backdoor attacks relies on the backdoor data being sufficiently distant so that the updated prediction functions' outputs are similar to the pre-update outputs.

The Adaptability Hypothesis. *In a successful backdoor attack, if we update a DNN f^{cl} that has been trained on a dataset $\mathcal{D}^{cl}$ to f^{poi} with a backdoor dataset $\mathcal{D}_\eta^{bd}$, then the difference in the predictions between f^{cl} and the updated model f^{poi} at a typical point x from the distribution of $\mathcal{D}^{cl}$ should be small if x is relative 'far' from $\mathcal{D}_\eta^{bd}$, and substantial if x is located 'close' to $\mathcal{D}_\eta^{bd}$.*

Characterization the distance from a point to a distribution. The concept of 'far' and 'close' should be measured in appropriate metrics, because the classical Euclidean distance may not be meaningful for images, voice, and text data. For example, previous research has shown that human-imperceptible backdoor triggers can achieve similar levels of test accuracy on both clean and backdoor data as compared to visible triggers, such as a square patch. This can be explained by the fact that human-imperceptible triggers may have smaller magnitudes in terms of Euclidean distance when compared to visible triggers, but they may have similar or even larger distances in terms of the distance metric defined on the distribution of data. To account for this, we need to consider distance measurements that take the data distribution into account, such as the Mahalanobis distance in Lemma 3.2, which scales the Euclidean distance based on the covariance of the data. Using the Mahalanobis distance allows for a small trigger, when measured in Euclidean distance, to significantly alter the point's position in relation to the probability mass of the data, making the backdoor data distinct from the clean data, as illustrated in Figure 2, increasing the chances of a successful backdoor attack.

When and how does a DNN backdoor attack work? From the above discussion, an effective backdoor trigger alters an original input to create a backdoor version, where the Euclidean distance between them is small, but the probability distance between their distributions is generally large. This ensures that retraining the model with backdoor samples and its associated backdoor labels will not significantly affect its performance on the original data while learning the artificially created relationship between the backdoor data and attacker-specified labels, thus achieving the dual goals of a backdoor attack.

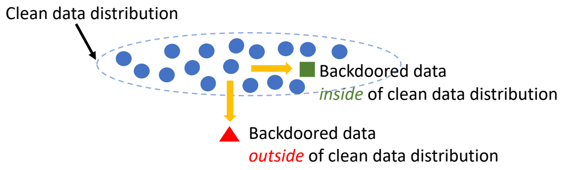

Clean data distribution
Backdoored data
inside of clean data distribution
Backdoored data
outside of clean data distribution

Figure 2. Illustration of the concept of proper distance measurement. A tiny trigger can push the input far away from its original probability mass. Thus, an effective backdoor should cause large changes in those input dimensions whose original variances are small.

Quantification of the Adaptability Hypothesis We first introduce a concept to quantify the Adaptability hypothesis. We then provide a general result on the effectiveness of a backdoor attack using this proposed concept.

Definition 4.1 (Adaptability)**.** Let $\mathcal{D}^{\mathrm{cl}}$ and $\mathcal{D}_\eta^{\mathrm{bd}}$ to be a set of clean and backdoor data with distributions $\mathbb{P}_X$, $\mathbb{P}_1^\eta$ of mean parameters μ_X, μ_1^η respectively. Let $r, s > 0$ be any fixed numbers and denote $C_r = \{x \in \mathbb{R}^d \mid \|X - \mu_X\| \leq r \text{ for } r > 0\}$ and $B_{s,\eta} = \{x \in \mathbb{R}^d \mid \|X - \mu_1^\eta\| \leq s \text{ for } s > 0\}$. The learning procedure $\mathcal{A}$ is said to have (g, h, d)-Adaptability with respect to $\mathcal{D}^{\mathrm{poi}} \triangleq \mathcal{D}^{\mathrm{cl}} \cup \mathcal{D}_\eta^{\mathrm{bd}}$ and a pair of positive values (r, s), if there exists two monotonically decreasing functions, $g, h : \mathbb{R}_+ \to \mathbb{R}_+$, and a distance measurement $d(q, P)$ of a point q to a distribution P taking values in $[0, \infty]$, such that,

1. for all $x \in C_r$, $\mathbb{E}^{\mathrm{poi}}|f^{\mathrm{cl}}(x) - f^{\mathrm{poi}}(x)| \leq g(d(x, \mathbb{P}_1^\eta))$,
2. for all $x \in B_{s,\eta}$, $\mathbb{E}^{\mathrm{poi}}|f^{\mathrm{pbd}}(x) - f^{\mathrm{poi}}(x)| \leq h(d(x, \mathbb{P}_1))$,

where the expectation $\mathbb{E}^{\mathrm{poi}}$ is taken with respect to $\mathcal{D}^{\mathrm{poi}}$, and $\mathbb{P}_1$ is the distribution of $X|Y=1$.

Theorem 4.2 (The effectiveness of a backdoor attack)**.** *Consider the threat model discussed in Section 2. Suppose that the loss function is Lipschitz and the learning procedure $\mathcal{A}$ satisfies adaptability with respect to $\mathcal{D}^{poi}$ and a pair of value (r, s), then we have*

$$R_n^{cl} \leq L_K \mathbb{P}_X(\mathbb{R}^d \setminus C_r) + L_K \max_{u \in C_r} g(d(u, \mathbb{P}_1^\eta)),$$

and

$$R_n^{bd} \leq L_K \mathbb{P}_1^\eta(\mathbb{R}^d \setminus B_{s,\eta}) + L_K \max_{u \in B_{s,\eta}} h(d(u, \mathbb{P}_1)).$$

Example 1 (Kernel Smoothing with $d-$dimensional Gaussian). Following the same setup in Section 3. We have

$$d(x, \mathbb{P}_1^\eta) = (x - \mu_1 - \eta)^\top \Sigma^{-1}(x - \mu_1 - \eta),$$

and

$$g(x) = \frac{\exp(-x) + C_3}{\exp(-x) + C_3 + C_4},$$

where C_3, C_4 are constants of n_{bd}, h_n and $\mathbb{P}_1$. By setting $(r, s) = (\|\eta\|/4, \|\eta\|/4)$, we recover the results in Theorem 3.1.

In general, we would expect that the learning procedure $\mathcal{A}$ is adaptable enough to handle both sufficiently large r and s, resulting in a small summation in the upper bound. In terms of the distance measurement, one should select appropriate distance metric for different tasks. Additionally, in practice, we can specify g and h using a particular parametric form, such as exponential functions, and determine the parameters through empirical fitting.

5. Experiments for validating the Adaptability Hypothesis

This section provides empirical studies to validate the proposed Adaptability hypothesis. We begin by introducing the experimental setting.

Datasets & Models We use 3 popular datasets: MNIST (LeCun & Cortes, 2010), CIFAR10 (Krizhevsky et al., 2009), and GTSRB (Stallkamp et al., 2012). We include the results for (1) MNIST with LetNet (LeCun et al., 2015), ResNet (He et al., 2016), (2) CIFAR10 with ResNet, VGG (Simonyan & Zisserman, 2014) in the main text, and defer the rest to the appendix. The detailed configurations, including model structure, training schedule, and tuning hyperparameters, are included in Appendix D.

Backdoor Data Generation The designated target label for the backdoor is set to 0 and a random class of images, known as the source class, with the exception of class 0, is chosen to incorporate backdoor triggers. For example, images from class 2 are chosen for adding backdoor triggers in Figure 3(a). Backdoor triggers are square patches, placed in the lower-right corner of the images, as demonstrated in the SOTA BadNets attacks (Gu et al., 2017). The default value for each pixel of the square patches is set to 255. We provide an ablation study on validating the Adaptability hypothesis with different source classes in Appendix E.

How to calculate the distances for images? A common way to assess the distance or similarity between images is to use the latent representations of a trained CNN (Chen et al., 2018; Hayase et al., 2021; Li et al., 2021b). In alignment with this method, we calculate all distances using the representations of the layer before the prediction layers and use the Mahalanobis distance as the metric throughout this section. The rationale for this choice is investigated in Section 5.2.

5.1. Experiments to corroborate the Adaptability

We now present empirical results for assessing the adaptability hypothesis on MNIST datasets using LetNet 5/ResNet 9. Similar results are observed on CIFAR10/GTSRB with ResNet20, and details are included in Appendix E. The empirical examinations on the validity of the adaptability are conducted under two varying factors: (1) the pixel values of the backdoor triggers, i.e., square patches, and (2) the poisoning ratio α_{poi}.

The Adaptability under backdoor triggers with different pixel values. In Figure 3(a) (with source class 2) and 3(b) (with source class 4), the $x-$axis value of each point represents the Mahalanobis distance of one clean data point z to the backdoor distribution, and the $y-$axis value represents the absolute change in the predicted probability between f^{cl} and f^{poi} for that clean data point's ground-truth class, namely $|f^{\text{cl}}(z) - f^{\text{poi}}(z)|$. In each figure, as the distance from a point to the backdoor distribution increase, the corresponding change in absolute value between f^{cl} and f^{poi} decreases, supporting the Adaptability hypothesis. Moreover, as the pixel values of the backdoor triggers increase, we observe that the average change f^{cl} and f^{poi} also decreases, which aligns with our theoretical intuitions in Theorem 3.1.

The Adaptability under different poisoning ratios α_{poi}. We fixed the pixel values of backdoor triggers and varied the poisoning ratios in our experiments, with results in Figures 10 and 12 in Appendix E. We found that the change in predicted values decreases as the distance to the backdoor distribution increases and that the average change in predicted values increases as the poisoning ratio increases, consistent with our theory.

5.2. Is the Mahalanobis distance reasonable?

Empirical results are presented to support the suitability of the Mahalanobis distance as a metric for backdoor attacks in CNNs. In light of the Adaptability hypothesis and Lemma 3.2, an effective backdoor trigger should aim to maximize the Mahalanobis distance between clean and backdoor data, or in other words, it should be added in dimensions with low variances.

To verify the above point, we consider three poisoning state-of-the-art attacks: (1) BadNets (Gu et al., 2017), which adds a patch at the lower-right corner of the clean images, (2) Adaptive Blend (Qi et al., 2022) (abbreviated as Ada-B), which embeds a portion of hello kitty into the clean images, and (3) Ad K-triggers (Qi et al., 2022) (abbreviated as Ada-K), which adds adaptive patches into the clean images. An example of clean images and their associated backdoor version is illustrated in Figure 9 in the appendix. We plot

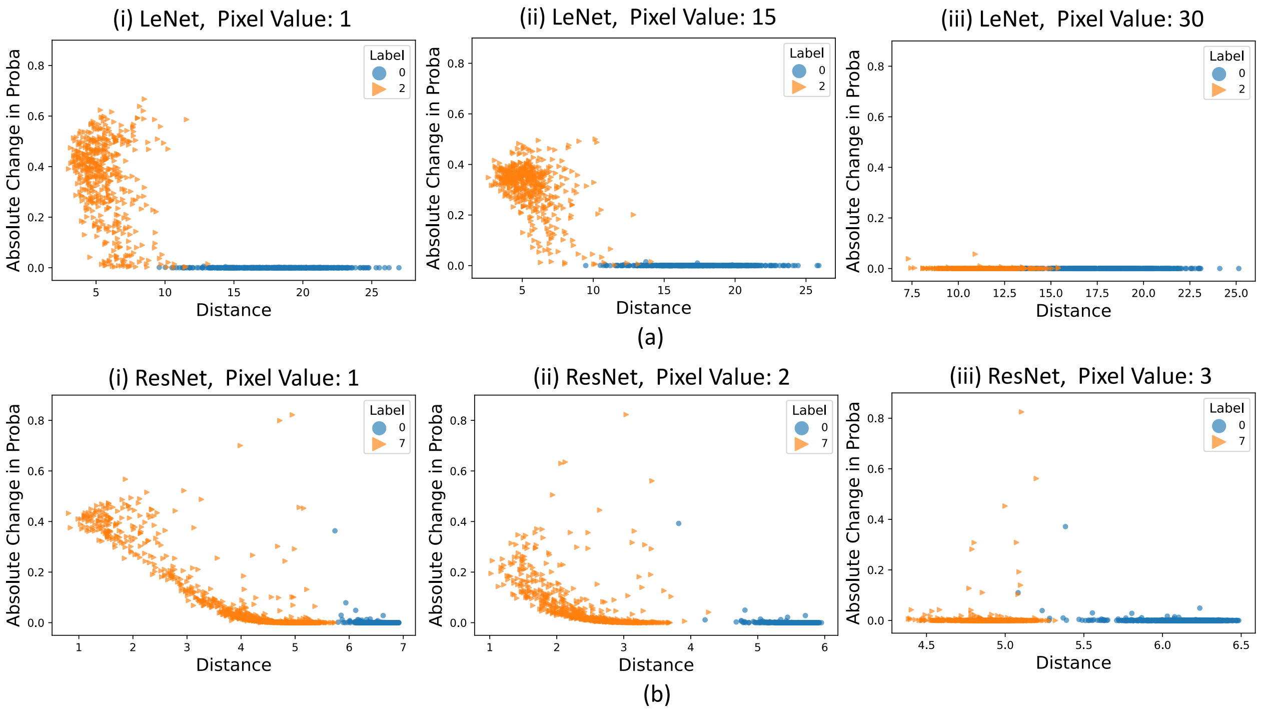


Figure 3. Figures (a) and (b) demonstrate the Adaptability Hypothesis on the MNIST dataset using LeNet and ResNet respectively. A subset of original images labeled as 2 in Figure (a) and 7 in Figure (b) is altered by adding a patch and re-labeled as 0. A pre-trained LeNet/ResNet is then fine-tuned using these manipulated images. The distance between the clean training data points (labeled as 0 and 2 for LeNet, and 0 and 7 for ResNet) and the backdoor data distribution, as well as the change in predicted values for the clean training data before and after fine-tuning the LeNet/ResNet, are plotted. The results show that for data points that are close to the backdoor distribution, the changes in their predicted values are significant, while changes in predicted values are small for data points that are far from the backdoor distribution. These observations support the proposed Adaptability Hypothesis.

the relative change, i.e., dimensional Mahalanbois distance, between clean data and backdoor data in Figure 4 for three attacks, where each point's $x-$axis value is the variance of the i-th dimension of backdoor data $\mathcal{D}_{\eta}^{\text{bd}}$, and the $y-$axis is the relative change along this dimension given by the absolute difference between clean and backdoor data at the i-th divided by the standard deviation of i-th dimension. We observe that for Ada-B and Ada-K attacks under CIFAR10 with ResNet 20, the relative distance tends to be greater on dimensions of data with small variance, which aligns with our theoretical results and also confirms the effectiveness of Ada-B and Ada-K attacks. More experimental results with different poisoning ratios and different model architectures are included in Section F, H in the appendix.

6. Implications of the Adaptability Hypothesis

We provide implications of the Adaptability of designing new effective backdoor attacks and defenses. In short, it is important to carefully choose a metric when measuring the similarity or distance between clean and backdoor data, such as incorporating the data distribution.

Effective methods for visually separating backdoor data Our theory suggests a novel data representation that enables distinct visual distinctions between clean and backdoor data when viewed through techniques like PCA and TSNE, whereas such distinctions are not apparent in the original space. This separation, made possible by our new metric, could potentially inform new defense mechanisms.

To start, we will briefly examine the importance of visually distinguishing between clean and backdoor data in defending against backdoor attacks, before delving into the specifics of our proposed method. In recent times, defense mechanisms have gained popularity as a means to protect learning models (Li et al., 2020). This has led to the development of sophisticated attacks (triggers) that can bypass standard defense methods. For example, (Qi et al., 2022) aims to design triggers that result in indistinguishable latent representations of CNNs between clean and backdoor data, thereby circumventing defenses (Chen et al., 2018; Tran et al., 2018) that rely on visually distinguishing between clean and backdoor data and filtering out the latter. The effectiveness of the triggers proposed in (Qi et al., 2022) is visually demonstrated in Figure 5(a) and Figure 6(a). For instance, in Figure 5(a), we apply PCA to visualize the latent space of a ResNet 20 model trained on CIFAR10 for three different attacks, and it can be observed that the latent rep-

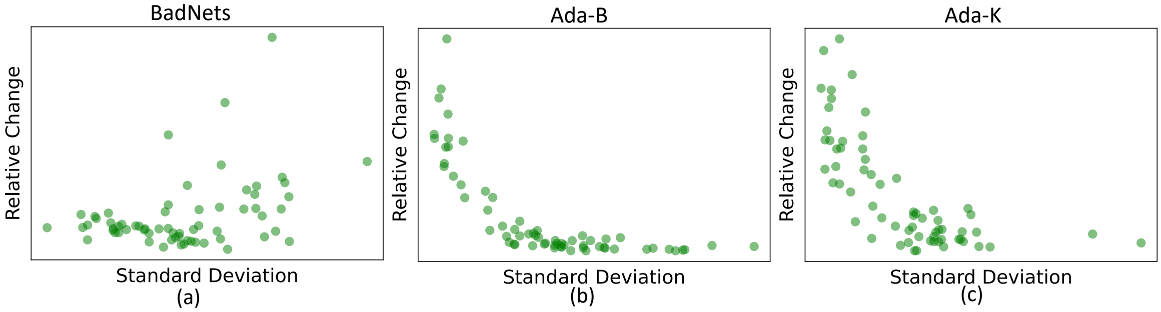


Figure 4. Illustrations of dimensional Mahalanbois distance for three attacks on CIFAR10 and ResNet 20. In each figure, each point's horizontal value represents the standard deviation (std) of one dimension of backdoor data, i.e., the standard deviation of the jth dimension of data, and its vertical value is the relative change along the same dimension, i.e., the difference (in absolute value) between clean and backdoor data along the jth dimension divided by the std of the jth dimension. We observed that for Ada-K and Ada-B attacks, they tend to have a larger (smaller) relative distance on the dimension of data with lower (higher) std, consistent with our theory.

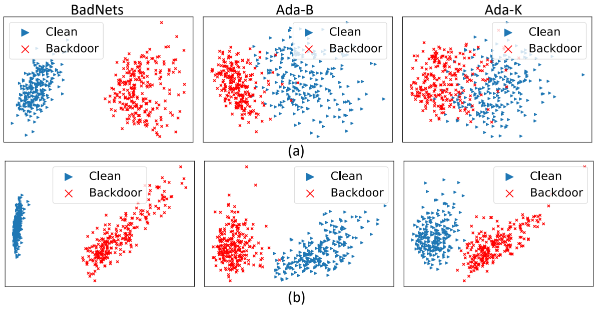


Figure 5. PCA visualizations of (a) the original and (b) transformed latent spaces of ResNet20 trained on CIFAR10 with $\alpha_{\text{poi}} = 0.5\%$ for three SOTA attacks. The latent spaces of Ada-B and Ada-K attacks do not exhibit a clear separation of clusters in their original form, suggesting the relative effectiveness of Ada-B/K compared with the BadNets. However, our theoretical results demonstrate that a transformed version of these spaces does display two distinct and separate clusters, as shown in (b).

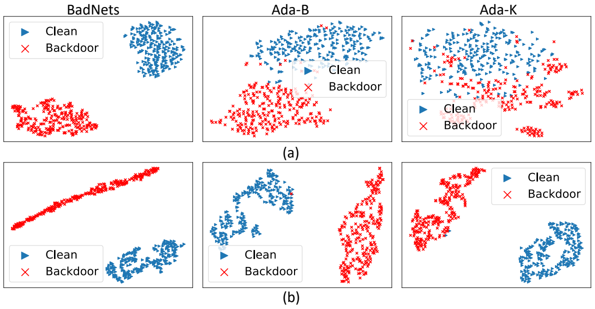


Figure 6. TSNE visualizations of (a) the original and (b) transformed latent spaces of ResNet 20 on trained CIFAR10 with $\alpha_{\text{poi}} = 0.5\%$ for three SOTA Attacks. Our proposed transformation separates clusters in the latent spaces of Ada-K attack, which were not separated in their original form.

resentations for Ada-K/B attacks do not form two separate clusters, unlike the well-separated clusters for BadNets. However, our results suggest that one possible reason for the visual indistinguishability could be that both PCA and TSNE operate on the original latent spaces of CNNs, without considering their distributions. To address this, we propose a method for visualizing the latent spaces of CNNs that first calculates the Mahalanobis distance of each data point to each class-specific distribution and then combines these distances to form a new representation, (see Algorithm 1 in Appendix G for details). The reasoning behind this transformation is illustrated in Figure 2, where the red triangle, which likely belongs to a different distribution, has a much greater Mahalanobis distance to the distribution (with a blue dashed circle) than the points within the dashed circle. By applying the aforementioned transformations, we demonstrate the ability to distinguish between clean and potentially backdoor data, as illustrated in Figure 5 and 6. The separation of clean and backdoor data allows for the identification and filtering of backdoor attacks, reducing the potential threat of such attacks. A potential limitation of this method as a defense strategy is the requirement of access to a small subset of both clean and backdoor data.

7. Conclusion

In this study, we studied the theoretical aspects of backdoor attacks. We first examine backdoor attacks in the context of classical machine learning and propose a hypothesis to explain their effectiveness in general learning models. Experiments are conducted to validate the hypothesis and implications for future defenses, and attacks are discussed. There are several potential areas for future research. These include deriving theoretical results for other types of machine learning models, such as feed-forward neural networks, and conducting experiments on different data forms, including text and voice data.

The appendix includes restatements and proofs of theoretical results and additional experimental results.

Acknowledgement

The work of Xun Xian, Xuan Bi, and Mingyi Hong was supported in part by Cisco Research. The work of Ganghua Wang and Jie Ding was supported in part by the Army Research Office under grant number W911NF-20-1-0222.

References

Audibert, J.-Y. and Tsybakov, A. B. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2): 608–633, 2007.

Bagdasaryan, E. and Shmatikov, V. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1505–1521, 2021.

Biggio, B., Nelson, B., and Laskov, P. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

Chen, B., Carvalho, W., Baracaldo, N., Ludwig, H., Edwards, B., Lee, T., Molloy, I., and Srivastava, B. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

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

Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Doan, K., Lao, Y., and Li, P. Backdoor attack with imperceptible input and latent modification. *Advances in Neural Information Processing Systems*, 34, 2021a.

Doan, K., Lao, Y., Zhao, W., and Li, P. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11966–11976, 2021b.

Dreossi, T., Donzé, A., and Seshia, S. A. Compositional falsification of cyber-physical systems with machine learning components. *Journal of Automated Reasoning*, 63(4): 1031–1053, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020.

Gu, T., Dolan-Gavitt, B., and Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Hayase, J., Kong, W., Somani, R., and Oh, S. Spectre: defending against backdoor attacks using robust statistics. *arXiv preprint arXiv:2104.11315*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., and Li, B. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 19–35. IEEE, 2018.

Jagielski, M., Severi, G., Pousette Harger, N., and Oprea, A. Subpopulation data poisoning attacks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3104–3122, 2021.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

LeCun, Y. et al. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 20(5):14, 2015.

Li, Y., Wu, B., Jiang, Y., Li, Z., and Xia, S.-T. Backdoor learning: A survey. *arXiv preprint arXiv:2007.08745*, 2020.

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

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021b.

Mahalanobis, P. C. On the generalized distance in statistics. National Institute of Science of India, 1936.

Manoj, N. and Blum, A. Excess capacity and backdoor poisoning. *Advances in Neural Information Processing Systems*, 34:20373–20384, 2021.

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

Oh, S. L., Hagiwara, Y., Raghavendra, U., Yuvaraj, R., Arunkumar, N., Murugappan, M., and Acharya, U. R. A deep learning approach for parkinson's disease diagnosis from eeg signals. *Neural Computing and Applications*, 32(15):10927–10933, 2020.

Qi, X., Xie, T., Mahloujifar, S., and Mittal, P. Circumventing backdoor defenses that are based on latent separability. *arXiv preprint arXiv:2205.13613*, 2022.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Souri, H., Fowl, L., Chellappa, R., Goldblum, M., and Goldstein, T. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. *Advances in Neural Information Processing Systems*, 35: 19165–19178, 2022.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32: 323–332, 2012.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tran, B., Li, J., and Madry, A. Spectral signatures in backdoor attacks. *arXiv preprint arXiv:1811.00636*, 2018.

Turner, A., Tsipras, D., and Madry, A. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vapnik, V., Levin, E., and Le Cun, Y. Measuring the vc-dimension of a learning machine. *Neural computation*, 6 (5):851–876, 1994.

Weber, M., Xu, X., Karlavs, B., Zhang, C., and Li, B. Rab: Provable robustness against backdoor attacks. *arXiv preprint arXiv:2003.08904*, 2020.

Zhao, S., Ma, X., Zheng, X., Bailey, J., Chen, J., and Jiang, Y.-G. Clean-label backdoor attacks on video recognition models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14443–14452, 2020.

Appendix for Understanding Backdoor Attacks through the Adaptability Hypothesis

The appendix includes restatements and proofs of the main results in Section A. In Section B, proofs for technical lemmas that were not previously included are provided. Section C lists all omitted technical lemmas that were informally referred to earlier. The configurations for the experiments are listed in Section D. Additional experiments to validate the hypothesis and theoretical justifications can be found in Section E and F. The pseudo-code for the visualization algorithms and additional experiments are included in Section G.

A. Missing Proof

In this section, we will formally state our theoretical results and provide proof.

Notations: We denote $\mathbb{P}_1$ and $\mathbb{P}_0$ to be the class-conditional distributions of $X|Y=1$ and $X|Y=0$, with mean parameter μ_1, μ_0 respectively. Without loss of generality, we assume that $\mu_0+\mu_1=0$. For any given $r,s>0$ and $\eta\in\mathbb{R}^d$, denote $C_r=\{x\in\mathbb{R}^d \mid \|x-(\mu_1+\mu_0)\|\le r\}$ and $B_{s,\eta}=\{x\in\mathbb{R}^d \mid \|x-(\mu_1+\eta)\|\le s\}$ to be two sets representing the typical data from the clean and backdoor distribution, respectively. We denote $\mathbb{B}_{c,r}=\{x\in\mathbb{R}^d \mid \|x-c\|\le r\}$ to be the ball centered at the point c with a radius $r>0$, and $V_d(r)$ to be the volume of a ball with radius $r>0$ in $\mathbb{R}^d$.

Before restating and proving the main results, we first list some technical assumptions and introduce useful lemmas.

A.1. General Assumptions.

Assumption 1. For a given poisoned dataset $\mathcal{D}^{\text{poi}}$, the bandwidth h_n is set to the order of $n^{-1/(2k+d)}$, where n and d are the sample size and dimension of $\mathcal{D}^{\text{poi}}$, and k is a parameter associated with the data generating distribution $\mathbb{P}_{X,Y}$.

Remark A.1. The bandwidth parameter signifies a trade-off between the variance and bias of the predicted generalization error. An increase in bandwidth will result in a kernel smoothing method with higher variance and lower bias. The selection of bandwidth, in this case, strikes a balance between bias and variance, resulting in f^{cl} having an optimal expected generalization error under distribution $\mathbb{P}_X$.

Assumption 2. The loss function is Lipschitz, namely, for every $K>0$ there exists a constant $L_K>0$ such that

$$|\ell(w_1,y)-\ell(w_2,y)|\le L_K\,|w_1-w_2|$$

for all $w_1,w_2\in[-K,K]$ and for all $y\in\{0,1\}$.

Remark A.2. Several loss functions are frequently utilized for classification tasks that satisfy the above assumption, e.g., (1) the square loss $\ell(w,y)=(w-y)^2$ with $L_K=2K+2$, (2) the logistic loss $\ell(w,y)=\ln\left(1+e^{-wy}\right)$ with $L_K=e^K/\left(1+e^K\right)\le 1$.

A.2. Useful technical preliminary results

The following result quantifies the change between f^{cl} and f^{poi} for the class-conditional distribution $\mathbb{P}_1$ being a general distribution. The results in the main text can easily be obtained from the below result by specifying $\mathbb{P}_1$ to be a multi-variate normal distribution.

Lemma A.3 (Kernel smoothing estimation under general distributions $\mathbb{P}_1$)**.** *Following the threat model described in Section 2, we consider using kernel smoothing as the learning procedure. Given any $\eta\in\mathbb{R}^d$ and $n>0$, by setting $h_n<0.5\|\eta\|$, we have for each $x\in\mathbb{R}^d$,*

$$\mathbb{E}^{poi}|f^{cl}(x)-f^{poi}(x)|\le 1-\frac{(n_1/n_{bd}-1)\mathbb{P}_1(\mathbb{B}_{x,h_n})}{(n_1/n_{bd}-1)\mathbb{P}_1(\mathbb{B}_{x,h_n})+\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})+1/n_{bd}}, \tag{1}$$

and for each $x\in\mathbb{R}^d$,

$$\mathbb{E}^{poi}|f^{pbd}-f^{poi}(x)|\le 1-\frac{\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})}{n_1/n_{bd}\mathbb{P}_1(\mathbb{B}_{x,h_n})+\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})+1/n_{bd}}. \tag{2}$$

A.3. Refined version of Lemma 3.2

In this section, we provide a refined statement of Lemma 3.2. The restatement and proof of Lemma 3.2 are included in the section subsection.

We make the following two assumptions regarding (1) the underlying data distributions and (2) the corresponding backdoor triggers based on the assumed data distributions.

Assumption 3. The class-conditional distributions of $X|Y = 0$ and $X|Y = 1$ are d-dimensional normal distributions with parameters (μ_0, Σ) and (μ_1, Σ), respectively, where $\mu_1 + \mu_0 = 0$ and Σ is a diagonal matrix with diagonal elements $\lambda_{d-1} \geq \ldots \geq \lambda_0 > 0$.

Assumption 4. The backdoor trigger η satisfies $\|\eta\| > 2h_n$, $\min(\|\eta + \mu_1\|, \|\eta + \mu_0\|) > 7/8\|\eta\|$, $\|\eta\|/4 > \max(\|\mu_1\|, \|\mu_0\|)$, and $0.5\|\eta\| > \sqrt{\|\eta\|}$.

Lemma A.4 (Refined Statements of Lemma 3.2)**.** *Following the threat model described in Section 2, we consider using kernel smoothing as the learning procedure. With Assumptions 1, 2, and 3, we have, for each $x \in C_r$,*

$$\lim_{n\to\infty} \mathbb{E}^{poi}|f^{cl}(x) - f^{poi}(x)| \leq \frac{\exp\left(-d(x, \mathbb{P}_1^{\eta})/2\right)}{\exp\left(-d(x, \mathbb{P}_1^{\eta})/2\right) + (\alpha_{poi}^{-1} - 1)\exp(-d(x, \mathbb{P}_1)/2)}, \tag{3}$$

where $d(x, \mathbb{P}_1^{\eta}) \triangleq (x - \mu_1 - \eta)^\top \Sigma^{-1}(x - \mu_1 - \eta)$ and for each $x \in B_{s,\eta}$,

$$\lim_{n\to\infty} \mathbb{E}^{poi}|f^{pbd} - f^{poi}(x)| \leq \frac{\exp(-d(x, \mathbb{P}_1)/2)}{\exp(-d(x, \mathbb{P}_1)/2) + \alpha_{poi}\exp(-d(x, \mathbb{P}_1^{\eta})/2)}, \tag{4}$$

where $d(x, \mathbb{P}_1) \triangleq (x - \mu_1)^\top \Sigma^{-1}(x - \mu_1)$ and recall $f^{pbd}(x) = 0$ for all $x \in \mathbb{R}^d$.

Proof. To prove the results, we invoke the above two Lemma A.3, for any $n > 0$, for each $x \in C_r$, we have

$$\begin{aligned}\mathbb{E}^{\text{poi}}|f^{\text{cl}}(x) - f^{\text{poi}}(x)| &\leq \frac{\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n}) + 1/n_{\text{bd}}}{(n_1/n_{\text{bd}} - 1)\mathbb{P}_1(\mathbb{B}_{x,h_n}) + \mathbb{P}_1(\mathbb{B}_{x-\eta,h_n}) + 1/n_{\text{bd}}}\\ &\leq \frac{\max_{z\in\mathbb{B}_{x,h_n}} \exp\left(-d(z, \mathbb{P}_1^{\eta})/2\right) + (2\pi)^{d/2}\sqrt{|\Sigma|}/(n_{\text{bd}}V_d(h_n))}{(\alpha_{\text{poi}}^{-1} - 1)\min_{z\in\mathbb{B}_{x,h_n}} \exp\left(-d(z, \mathbb{P}_1)/2\right) + \max_{z\in\mathbb{B}_{x,h_n}} \exp\left(-d(z, \mathbb{P}_1^{\eta})/2\right) + (2\pi)^{d/2}\sqrt{|\Sigma|}/(n_{\text{bd}}V_d(h_n))},\end{aligned}$$

where $\alpha_{\text{poi}} = n_{\text{bd}}/n_1$ is the ratio of backdoor sample size.

In particular, we have

$$\lim_{n\to\infty} \mathbb{E}^{\text{poi}}|f^{\text{cl}}(x) - f^{\text{poi}}(x)| \leq \frac{\exp\left(-d(x, \mathbb{P}_1^{\eta})/2\right)}{(\alpha_{\text{poi}}^{-1} - 1)\exp\left(-d(x, \mathbb{P}_1)/2\right) + \exp\left(-d(x, \mathbb{P}_1^{\eta})/2\right)},$$

since $n_{\text{bd}}V_d(h_n)$ is of the order $n^{1-d/(d+4)} = n^{4/(d+4)}$ which goes to ∞ from the Assumption 1.

Similarly, for any $n > 0$, for each $x \in B_{s,\eta}$, we have

$$\begin{aligned}\mathbb{E}^{\text{poi}}|0 - f^{\text{poi}}(x)| &\leq 1 - \frac{n_{\text{bd}}\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})}{n_1\mathbb{P}_1(\mathbb{B}_{x,h_n}) + n_{\text{bd}}\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n}) + 1},\\ &\leq \frac{\alpha_{\text{poi}}^{-1}\max_{z\in\mathbb{B}_{x,h_n}} \exp\left(-d(z, \mathbb{P}_1)/2\right) + (2\pi)^{d/2}\sqrt{|\Sigma|}/(n_{\text{bd}}V_d(h_n))}{\alpha_{\text{poi}}^{-1}\max_{z\in\mathbb{B}_{x,h_n}} \exp\left(-d(z, \mathbb{P}_1)/2\right) + \min_{z\in\mathbb{B}_{x,h_n}} \exp\left(-d(z, \mathbb{P}_1^{\eta})/2\right) + (2\pi)^{d/2}\sqrt{|\Sigma|}/(n_{\text{bd}}V_d(h_n))}.\end{aligned}$$

In particular, we have

$$\lim_{n\to\infty} \mathbb{E}^{\text{poi}}|0 - f^{\text{poi}}(x)| \leq \frac{\exp\left(-d(x, \mathbb{P}_1)/2\right)}{\exp\left(-d(x, \mathbb{P}_1)/2\right) + \alpha_{\text{poi}}\exp\left(-d(x, \mathbb{P}_1)/2\right)}.$$

□

A.4. Proof of Lemma 3.2

Lemma A.5 (Restatements of Lemma 3.2)**.** *Following the threat model described in Section 2, we consider using kernel smoothing as the learning procedure. With Assumptions 1, 2, and 3, we have, for each $x \in C_r$,*

$$\lim_{n\to\infty} \mathbb{E}^{poi}|f^{cl}(x) - f^{poi}(x)| \le \frac{\exp\left(-d(x, \mathbb{P}_1^{\eta})/2\right)}{\exp\left(-d(x, \mathbb{P}_1^{\eta})/2\right) + (\alpha_{poi}^{-1} - 1)\exp\left(-(r^2 + \|\mu_1\|^2)/\lambda_0\right)}, \tag{5}$$

where $d(x, \mathbb{P}_1^{\eta}) \triangleq (x - \mu_1 - \eta)^\top \Sigma^{-1}(x - \mu_1 - \eta)$ and for each $x \in B_{s,\eta}$,

$$\lim_{n\to\infty} \mathbb{E}^{poi}|0 - f^{poi}(x)| \le \frac{\exp(-d(x, \mathbb{P}_1)/2)}{\exp(-d(x, \mathbb{P}_1)/2) + \alpha_{poi}\exp\left(-\|s\|^2/2\lambda_0\right)}, \tag{6}$$

where $d(x, \mathbb{P}_1) \triangleq (x - \mu_1)^\top \Sigma^{-1}(x - \mu_1)$.

Proof. The proof is directly from combining the Lemma A.4 and the following result.

Lemma A.6 (Probability of Gaussian Balls)**.** *Let $\mathbb{P}_1$ a $d-$dimensional normal distribution with the diagonal covariance matrix Σ with diagonal elements $\lambda_{d-1} \ge \lambda_{d-1} \ge \ldots \ge \lambda_0 > 0$. For any $x \in C_r$, we have:*

- $\mathbb{P}_1(\mathbb{B}_{x,h_n}) \ge \frac{V_d(h_n)}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-((r + h_n)^2 + \|\mu_1\|^2)/\lambda_0\right)$,
- $\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n}) \le \frac{V_d(h_n)}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \max_{z\in\mathbb{B}_{x,h_n}} \exp\left(-d(z, \mathbb{P}_1^{\eta})/2\right)$.

For any $x \in B_{s,\eta}$, we have:

- $\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n}) \ge \frac{V_d(h_n)}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-(\|s\| + h_n)^2/2\lambda_0\right)$,
- $\mathbb{P}_1(\mathbb{B}_{x,h_n}) \le \frac{V_d(h_n)}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \max_{z\in\mathbb{B}_{x,h_n}} \exp\left(-d(z, \mathbb{P}_1)/2\right)$.

The proof of the above two Lemmas is included in Section B.1 and B.2, respectively. □

A.5. Proof of Theorem 3.1

Theorem A.7 (Restatement of Theorem 3.1)**.** *Following the threat model described in Section 2, we consider using kernel smoothing as the learning procedure. Under Assumptions 1, 2, 3, and 4, as $n \to \infty$, we have*

$$R_n^{cl} \le O(\exp\left(-C_1\|\eta\|^2/\lambda_{d-1}\right) + \alpha_{poi}\exp\left(-\|\eta\|^2/\lambda_0\right)),$$

and

$$R_n^{bd} \le O(\exp\left(-C_2\|\eta\|^2/\lambda_{d-1}\right) + \alpha_{poi}^{-1}\exp\left(-\|\eta\|^2/\lambda_0\right)),$$

where α_{poi} is the backdoor poisoning rate, and C_1, C_2 are some constants $\mathbb{P}_1$ and $\mathbb{P}_0$.

Proof. In the following, we provide detailed proof of the gap in the clean test error, i.e., R_n^{cl}. The proof of the gap in the backdoor test error follows similar reasoning, and thus the details are omitted. As $n \to \infty$, we have

$$\begin{aligned}
R_n^{\mathrm{cl}} &= \mathbb{E}^{\mathrm{poi}}\mathbb{E}_{(X,Y)\sim\mathbb{P}_{XY}}[\ell(f^{\mathrm{poi}}(X), Y) - \ell(f^{\mathrm{cl}}(X), Y)] \\
&\le \mathbb{E}^{\mathrm{poi}}\mathbb{E}_{X\sim\mathbb{P}_{\mathbb{X}}} L_K|f^{\mathrm{poi}}(X) - f^{\mathrm{cl}}(X)| \qquad (7)\\
&= \mathbb{E}_{X\sim\mathbb{P}_{\mathbb{X}}}\mathbb{E}^{\mathrm{poi}} L_K|f^{\mathrm{poi}}(X) - f^{\mathrm{cl}}(X)| \qquad (8)\\
&= \mathbb{E}_{X\sim\mathbb{P}_{\mathbb{X}}}\mathbb{E}^{\mathrm{poi}} L_K|f^{\mathrm{poi}}(X) - f^{\mathrm{cl}}(X)|\mathbf{1}\{X \notin C_r\} \\
&+ \mathbb{E}_{X\sim\mathbb{P}_{\mathbb{X}}}\mathbb{E}^{\mathrm{poi}} L_K|f^{\mathrm{poi}}(X) - f^{\mathrm{cl}}(X)|\mathbf{1}\{X \in C_r\} \\
&\le L_K\mathbb{P}_X(\mathbb{R}^d \setminus C_r) + L_K\mathbb{E}_{X\sim\mathbb{P}_{\mathbb{X}}} \max_{X\in C_r}(\mathbb{E}^{\mathrm{poi}}|f^{\mathrm{poi}}(X) - f^{\mathrm{cl}}(X)|)\mathbf{1}\{X \in C_r\}
\end{aligned}$$

$$
\begin{aligned}
&\leq L_K \mathbb{P}_X(\mathbb{R}^d \setminus C_r) + O(\max_{X \in C_r} \frac{L_K \exp{(-d(X, \mathbb{P}_1^{\eta})/2)}}{(\alpha_{\text{poi}}^{-1} - 1) \exp{(-d(X, \mathbb{P}_1)/2)} + \exp{(-d(X, \mathbb{P}_1^{\eta})/2)}}) && (9) \\
&= L_K \mathbb{P}_X(\mathbb{R}^d \setminus C_r) + O(\max_{X \in C_r} \frac{L_K}{(\alpha_{\text{poi}}^{-1} - 1) \exp{((\eta^\top \Sigma^{-1} \eta - \eta^\top \Sigma^{-1}(x - \mu_1))/2)} + 1}) \\
&\leq L_K \mathbb{P}_X(\mathbb{R}^d \setminus C_r) + O(\frac{L_K}{(\alpha_{\text{poi}}^{-1} - 1) \exp{(\eta^\top \Sigma^{-1} \eta / 16)} + 1}) && (10) \\
&= O(\exp{(-\|\eta\|^2 / 320 \lambda_{d-1})} + \alpha_{\text{poi}} \exp{(-\eta^\top \Sigma^{-1} \eta / 16)}), && (11)
\end{aligned}
$$

where the inequality (7) is from the assumption on the Lipschitz condition, the equality (8) is by Fubini's Theorem, and the inequality in (9) is from Lemma A.4. Additionally, the inequality in (10) holds from Assumption 2 (length of $\|\eta\|$) and the definition of C_r with $r = \|\eta\|/4$, and (11) follows from the standard concentration inequalities of Chi-square random variables, and the fact that $1/(x+1) \leq 1/x$, for $x > 0$. To minimize the upper bound in (11), we choose effective directions for η given $|\eta|$ by placing all the weight of η on the direction of λ_0, which leads to the result $O(\exp{(-\|\eta\|^2/320\lambda_{d-1})} + \alpha_{\text{poi}} \exp{(-\|\eta\|^2/(16\lambda_0))})$.

□

A.6. Proof of Theorem 4.2

Proof. The proof for this result follows the same arguments in Theorem 3.1. Similarly, we provide proof of the gap in the clean test error, i.e., R_n^{cl}. The proof of the gap in the backdoor test error follows similar reasoning, and thus the details are omitted.

$$
\begin{aligned}
R_n^{\text{cl}} &= \mathbb{E}^{\text{poi}} \mathbb{E}_{(X,Y) \sim \mathbb{P}_{XY}} [\ell(f^{\text{poi}}(X), Y) - \ell(f^{\text{cl}}(X), Y)] \\
&\leq \mathbb{E}^{\text{poi}} \mathbb{E}_{(X,Y) \sim \mathbb{P}_{XY}} [|\ell(f^{\text{poi}}(X), Y) - \ell(f^{\text{cl}}(X), Y)|] \\
&\leq \mathbb{E}^{\text{poi}} \mathbb{E}_{X \sim \mathbb{P}_{\mathbb{X}}} L_K |f^{\text{poi}}(X) - f^{\text{cl}}(X)| && (12) \\
&= \mathbb{E}_{X \sim \mathbb{P}_{\mathbb{X}}} \mathbb{E}^{\text{poi}} L_K |f^{\text{poi}}(X) - f^{\text{cl}}(X)| && (13) \\
&= \mathbb{E}_{X \sim \mathbb{P}_{\mathbb{X}}} \mathbb{E}^{\text{poi}} L_K |f^{\text{poi}}(X) - f^{\text{cl}}(X)| \mathbf{1}\{X \notin C_r\} \\
&+ \mathbb{E}_{X \sim \mathbb{P}_{\mathbb{X}}} \mathbb{E}^{\text{poi}} L_K |f^{\text{poi}}(X) - f^{\text{cl}}(X)| \mathbf{1}\{X \in C_r\} \\
&\leq L_K \mathbb{P}_X(\mathbb{R}^d \setminus C_r) + L_K \mathbb{E}_{X \sim \mathbb{P}_{\mathbb{X}}} g(d(X, \mathbb{P}_1^{\eta})) \mathbf{1}\{X \in C_r\} && (14) \\
&\leq L_K \mathbb{P}_X(\mathbb{R}^d \setminus C_r) + L_K \max_{u \in C_r} g(d(u, \mathbb{P}_1^{\eta})) \mathbb{P}_X(C_r),
\end{aligned}
$$

where the inequality (12) is from the assumption on the lipschitz condition, the equality (13) is by Fubini's Theorem, and the inequality (14) is from the definition of adaptability. □

B. Proofs of Lemmas

B.1. Proof of Lemma A.3

Proof. In this proof, we demonstrate the use of kernel smoothing with general distributions.

Proof for Gap between f^{cl} and f^{poi}. For the ease of narrative, we decompose the clean training data set $\mathcal{D}^{\text{cl}}$ into D^0, $\mathcal{D}^{1,\text{cl}}$, and $\mathcal{D}^{1,\text{bd}}$. $\mathcal{D}^0$ is the clean training data with ground-truth label 0, $\mathcal{D}^{1,\text{cl}}$ is the clean training data with label 1 and not being backdoored by attackers, and $\mathcal{D}^{1,\text{bd}}$ is the clean training data with label 1 and backdoored by attackers. The following results provide an upper bound on the absolute change between f^{cl} and f^{poi} by only involving data samples with original labels of one. We include the proof of Lemma B.1 and Lemma B.2 in Section B.3 and Section B.4, respectively.

Lemma B.1. *For any given poisoned dataset $\mathcal{D}^{poi}$, we have*

$$
|f^{cl}(x) - f^{poi}(x)| \leq \frac{\sum_{X \in \mathcal{D}^{1,bd}} K(\frac{x-X-\eta}{h_n})}{\sum_{X \in \mathcal{D}^{1,cl}} K(\frac{x-X}{h_n}) + \sum_{X \in \mathcal{D}^{1,bd}} K(\frac{x-X-\eta}{h_n})},
$$

for all $x \in \mathbb{R}^d$.

Lemma B.2. *For any $c > 0$, we have*

$$\mathbb{E}^{poi} \frac{\sum_{X \in \mathcal{D}^{1,cl}} K(\frac{x-X}{h_n})}{\sum_{X \in \mathcal{D}^{1,cl}} K(\frac{x-X}{h_n}) + c} \geq \frac{(n_1 - n_{bd})\mathbb{P}_1(\mathbb{B}_{x,h_n})}{(n_1 - n_{bd})\mathbb{P}_1(B_{x,h_n}) + c + 1}.$$

Back to the main proof, for any given $\mathcal{D}^{\text{poi}}$, invoking Lemma B.1, we have

$$\begin{aligned}
&\mathbb{E}^{\text{poi}}|f^{\text{cl}}(x) - f^{\text{poi}}(x)| \\
&\leq \mathbb{E}^{\text{poi}} \frac{\sum_{X \in \mathcal{D}^{1,\text{bd}}} K(\frac{x-X-\eta}{h_n})}{\sum_{X \in \mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n}) + \sum_{X \in \mathcal{D}^{1,\text{bd}}} K(\frac{x-X-\eta}{h_n})}, \\
&\leq \mathbb{E}^{\text{poi}} \frac{n_{\text{bd}} \mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})}{\sum_{X \in \mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n}) + n_{\text{bd}} \mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})}, && (15) \\
&= \mathbb{E}^{\text{poi}} 1 - \frac{\sum_{X \in \mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n})}{\sum_{X \in \mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n}) + n_{\text{bd}} \mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})}, \\
&\leq 1 - \frac{(n_1 - n_{\text{bd}})\mathbb{P}_1(\mathbb{B}_{x,h_n})}{(n_1 - n_{\text{bd}})\mathbb{P}_1(\mathbb{B}_{x,h_n}) + n_{\text{bd}} \mathbb{P}_1(\mathbb{B}_{x-\eta,h_n}) + 1}, && (16)
\end{aligned}$$

where the inequality in (15) is because of the Jensen's Inequality and (16) holds from the Lemma B.2.

Proof for Gap between f^{poi} and f^{pbd}

We need the following lemma, and we include the proof in Section B.5.

Lemma B.3. *Given any $c > 0$, for any $x \in \mathbb{R}^d$ we have*

$$\mathbb{E}^{poi} \frac{\sum_{X \in \mathcal{D}^{1,bd}} K(\frac{x-X-\eta}{h_n})}{c + \sum_{X \in \mathcal{D}^{1,bd}} K(\frac{x-X}{h_n}) + \sum_{X \in \mathcal{D}^{1,bd}} K(\frac{x-X-\eta}{h_n})} \geq \frac{n_{bd} \mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})}{(n_{bd})(\mathbb{P}_1(\mathbb{B}_{x,h_n}) + \mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})) + c + 1}.$$

Invoking the above result, we have:

$$\begin{aligned}
&\mathbb{E}^{\text{poi}}|0 - f^{\text{poi}}(x)| \\
&= \mathbb{E}^{\text{poi}} \frac{\sum_{X \in \mathcal{D}^0} K(\frac{x-X}{h_n}) \cdot 0 + \sum_{X \in \mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n}) \cdot 1 + \sum_{X \in \mathcal{D}^{1,\text{bd}}} K(\frac{x-X}{h_n}) \cdot 1}{\sum_{X \in \mathcal{D}^0} K(\frac{x-X}{h_n}) + \sum_{X \in \mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n}) + \sum_{X \in \mathcal{D}^{1,\text{bd}}} [K(\frac{x-X}{h_n}) + K(\frac{x-X-\eta}{h_n})]}, \\
&\leq \mathbb{E}^{\text{poi}} \frac{\sum_{X \in \mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n}) + \sum_{X \in \mathcal{D}^{1,\text{bd}}} K(\frac{x-X}{h_n})}{\sum_{X \in \mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n}) + \sum_{X \in \mathcal{D}^{1,\text{bd}}} [K(\frac{x-X}{h_n}) + K(\frac{x-X-\eta}{h_n})]}, \\
&= 1 - \mathbb{E}^{\text{poi}} \frac{\sum_{X \in \mathcal{D}^{1,\text{bd}}} K(\frac{x-X-\eta}{h_n})}{\sum_{X \in \mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n}) + \sum_{X \in \mathcal{D}^{1,\text{bd}}} [K(\frac{x-X}{h_n}) + K(\frac{x-X-\eta}{h_n})]}, \\
&\leq 1 - \mathbb{E}^{\text{poi}} \frac{\sum_{X \in \mathcal{D}^{1,\text{bd}}} K(\frac{x-X-\eta}{h_n})}{(n_1 - n_{\text{bd}})\mathbb{P}_1(\mathbb{B}_{x,h_n}) + \sum_{X \in \mathcal{D}^{1,\text{bd}}} [K(\frac{x-X}{h_n}) + K(\frac{x-X-\eta}{h_n})]}, && (17) \\
&\leq 1 - \frac{n_{\text{bd}} \mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})}{(n_1 - n_{\text{bd}})\mathbb{P}_1(\mathbb{B}_{x,h_n}) + (n_{\text{bd}})(\mathbb{P}_1(\mathbb{B}_{x,h_n}) + \mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})) + 1}, && (18) \\
&\leq 1 - \frac{n_{\text{bd}} \mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})}{n_1 \mathbb{P}_1(\mathbb{B}_{x,h_n}) + n_{\text{bd}} \mathbb{P}_1(\mathbb{B}_{x-\eta,h_n}) + 1},
\end{aligned}$$

where the inequality in (17) is because of the Jensen's inequality, and (18) holds from Lemma B.3.

□

B.2. Proof of Lemma A.6

Proof. Recall that $V_d(r)$ denotes the volume of a ball with radius $r > 0$ in $\mathbb{R}^d$.

For any $x \in C_r$ we have

$$\begin{aligned}
&\mathbb{P}_1(\mathbb{B}_{x,h_n}) \\
&= \int_{\mathbb{B}_{x,h_n}} \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-(t-\mu_1)^\top \Sigma^{-1}(t-\mu_1)/2\right)dt \\
&\geq \frac{V_d(h_n)}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \min_{z\in\mathbb{B}_{x,h_n}} \exp\left(-\|z-\mu_1\|^2/2\lambda_0\right) \\
&\geq \frac{V_d(h_n)}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \min_{z\in\mathbb{B}_{x,h_n}} \exp\left(-(\|z\|^2+\|\mu_1\|^2)/\lambda_0\right) \qquad (19) \\
&\geq \frac{V_d(h_n)}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-((r+h_n)^2+\|\mu_1\|^2)/\lambda_0\right), \qquad (20)
\end{aligned}$$

where the inequality in (19) is because of the fact that $\|a-b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any $a, b \in \mathbb{R}^d$, and the inequality in (20) holds from the definition of C_r and $\mathbb{B}_{x,h_n}$.

Similarly, for any $x \in C_r$,

$$\begin{aligned}
&\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n}) \\
&= \mathbb{P}_1^\eta(\mathbb{B}_{x,h_n}) \qquad (21) \\
&= \int_{\mathbb{B}_{x,h_n}} \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-(t-\mu_1-\eta)^\top \Sigma^{-1}(t-\mu_1-\eta)/2\right)dt \\
&\leq \frac{V_d(h_n)}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \max_{z\in\mathbb{B}_{x,h_n}} \exp\left(-d(z,\mathbb{P}_1^\eta)/2\right)
\end{aligned}$$

where the inequality in (21) follows from the fact $\mathbb{P}_1^\eta$ is shifted by η from $\mathbb{P}_1$.

Similarly, for any $x \in B_{s,\eta}$,

$$\begin{aligned}
&\mathbb{P}_1(\mathbb{B}_{x-\eta}, h_n) \\
&= \mathbb{P}_1^\eta(\mathbb{B}_{x,h_n}) \\
&= \int_{\mathbb{B}_{x,h_n}} \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-(t-\mu_1-\eta)^\top \Sigma^{-1}(t-\mu_1-\eta)/2\right)dt \\
&\geq \frac{V_d(h_n)}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \min_{z\in\mathbb{B}_{x,h_n}} \exp\left(-\|z-\mu_1-\eta\|^2/2\lambda_0\right), \\
&= \frac{V_d(h_n)}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-(\|s\|+h_n)^2/2\lambda_0\right), \qquad (22)
\end{aligned}$$

where the inequality in (22) is from the definition of $B_{s,\eta}$ and the fact that $x \in B_{s,\eta}$.

For any $x \in B_{s,\eta}$, we have

$$\begin{aligned}
&\mathbb{P}_1(\mathbb{B}_x, h_n) \\
&= \int_{\mathbb{B}_{x,h_n}} \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-(t-\mu_1)^\top \Sigma^{-1}(t-\mu_1)/2\right)dt \\
&\leq \frac{V_d(h_n)}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \max_{z\in\mathbb{B}_{x,h_n}} \exp\left(-d(x,\mathbb{P}_1)/2\right).
\end{aligned}$$

□

B.3. Proof of Lemma B.1

Proof. Given any poisoned dataset $\mathcal{D}^{\text{poi}}$, for all $x \in \mathbb{R}^d$, we consider the following two cases: **Case (i)** If $\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n}) = 0$, by the definitions of Nadaraya-Watson kernel estimate, we can verify that

$$f^{\text{cl}}(x) = f^{\text{poi}}(x) = 0.$$

Case (ii) If $\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n}) > 0$, we have

$$\begin{aligned}
&f^{\text{cl}}(x) - f^{\text{poi}}(x)\\
&= \frac{\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n})\cdot Y}{\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n})} - \frac{\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n})\cdot Y + \sum_{X\in\mathcal{D}^{\text{bd}}_\eta} K(\frac{x-X}{h_n})\cdot 0}{\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n}) + \sum_{X\in\mathcal{D}^{\text{bd}}_\eta} K(\frac{x-X}{h_n})},\\
&= (\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n})\cdot Y)\cdot \frac{\sum_{X\in\mathcal{D}^{\text{bd}}_\eta} K(\frac{x-X}{h_n})}{(\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n}))(\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n}) + \sum_{X\in\mathcal{D}^{\text{bd}}_\eta} K(\frac{x-X}{h_n}))},\\
&\leq (\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n})\cdot 1)\cdot \frac{\sum_{X\in\mathcal{D}^{\text{bd}}_\eta} K(\frac{x-X}{h_n})}{(\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n}))(\sum_{(X,Y)\in\mathcal{D}^{\text{cl}}} K(\frac{x-X}{h_n}) + \sum_{X\in\mathcal{D}^{\text{bd}}_\eta} K(\frac{x-X}{h_n}))},\\
&= \frac{\sum_{X\in\mathcal{D}^{\text{bd}}_\eta} K(\frac{x-X}{h_n})}{\sum_{(X,Y)\in\mathcal{D}^{0}} K(\frac{x-X}{h_n}) + \sum_{(X,Y)\in\mathcal{D}^{1}} K(\frac{x-X}{h_n}) + \sum_{X\in\mathcal{D}^{\text{bd}}_\eta} K(\frac{x-X}{h_n})}, \qquad (23)\\
&\leq \frac{\sum_{X\in\mathcal{D}^{1,\text{bd}}} K(\frac{x-X-\eta}{h_n})}{\sum_{X\in\mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n}) + \sum_{X\in\mathcal{D}^{1,\text{bd}}} K(\frac{x-X-\eta}{h_n})},
\end{aligned}$$

where (23) follows the fact that $\mathcal{D}^{\text{cl}} \triangleq \mathcal{D}^0 \cup \mathcal{D}^1$.

□

B.4. Proof of Lemma B.2

Proof.

$$\begin{aligned}
&\mathbb{E}^{\text{poi}} \frac{\sum_{X\in\mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n})}{\sum_{X\in\mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n}) + c},\\
&= \sum_{X\in\mathcal{D}^{1,\text{cl}}} \mathbb{E}_{\mathcal{D}^1\sim\mathbb{P}_1} \frac{K(\frac{x-X}{h_n})}{c + \sum_{X\in\mathcal{D}^{1,\text{cl}}} K(\frac{x-X}{h_n})}, \qquad (24)\\
&= (n_1 - n_{\text{bd}})\mathbb{E}_{\mathcal{D}^1\sim\mathbb{P}_1} \frac{K(\frac{x-X_1}{h_n})}{c + K(\frac{x-X_1}{h_n}) + \sum_{X\in\mathcal{D}^{1,\text{cl}}\setminus X_1} K(\frac{x-X}{h_n})}, \qquad (25)\\
&= (n_1 - n_{\text{bd}})\mathbb{E}_{\mathcal{D}^1\setminus X_1\sim\mathbb{P}_1}(\frac{0}{c + 0 + \sum_{X\in\mathcal{D}^{1,\text{cl}}\setminus X_1} K(\frac{x-X}{h_n})})\mathbb{P}_1(K(\frac{x-X_1}{h_n}) = 0)\\
&+ (n_1 - n_{\text{bd}})\mathbb{E}_{\mathcal{D}^1\setminus X_1\sim\mathbb{P}_1}(\frac{1}{c + 1 + \sum_{X\in\mathcal{D}^{1,\text{cl}}\setminus X_1} K(\frac{x-X}{h_n})})\mathbb{P}_1(K(\frac{x-X_1}{h_n}) = 1)\\
&= (n_1 - n_{\text{bd}})\mathbb{E}_{\mathcal{D}^1\setminus X_1\sim\mathbb{P}_1}(\frac{1}{c + 1 + \sum_{X\in\mathcal{D}^{1,\text{cl}}\setminus X_1} K(\frac{x-X}{h_n})})\mathbb{P}_1(K(\frac{x-X_1}{h_n}) = 1),\\
&= (n_1 - n_{\text{bd}})\mathbb{E}_{Z\sim\text{Bino}((n_1-n_{\text{bd}}-1),\mathbb{P}_1(\mathbb{B}_{x,h_n}))}(\frac{1}{Z + c + 1})\mathbb{P}_1(K(\frac{x-X_1}{h_n}) = 1), \qquad (26)\\
&\geq \frac{(n_1 - n_{\text{bd}})\mathbb{P}_1(\mathbb{B}_{x,h_n})}{(n_1 - n_{\text{bd}} - 1)\mathbb{P}_1(\mathbb{B}_{x,h_n}) + c + 1}, \qquad (27)\\
&\geq \frac{(n_1 - n_{\text{bd}})\mathbb{P}_1(\mathbb{B}_{x,h_n})}{(n_1 - n_{\text{bd}})\mathbb{P}_1(\mathbb{B}_{x,h_n}) + c + 1},
\end{aligned}$$

where (24) follows the linearity of expectation, and (25) is because $\mathcal{D}^1$ consist of i.i.d samples and the symmetric property. Finally, (26) follows from the fact that $\sum_{X\in\mathcal{D}^1_{\text{cl}}\backslash X_1} K(\frac{x-X}{h_n})+$ is a binomial random variable with no. of trials $(n_1-n_{\text{bd}}-1)$ and the success probability $\mathbb{P}_1(\mathbb{B}_{x,h_n})$, and the inequality in (27) is due to the Jensen's inequality.

□

B.5. Proof of Lemma B.3

Proof.

$$
\begin{aligned}
&\mathbb{E}^{\text{poi}}\frac{\sum_{X\in\mathcal{D}^{1,\text{bd}}} K(\frac{x-X-\eta}{h_n})}{c+\sum_{X\in\mathcal{D}^{1,\text{bd}}} K(\frac{x-X}{h_n})+\sum_{X\in\mathcal{D}^{1,\text{bd}}} K(\frac{x-X-\eta}{h_n})},\\
&=\sum_{X\in\mathcal{D}^{1,\text{bd}}}\mathbb{E}_{\mathcal{D}^1\sim\mathbb{P}_1}\frac{K(\frac{x-X-\eta}{h_n})}{c+\sum_{X\in\mathcal{D}^{1,\text{bd}}} K(\frac{x-X}{h_n})+\sum_{X\in\mathcal{D}^{1,\text{bd}}} K(\frac{x-X-\eta}{h_n})}, &(28)\\
&=n_{\text{bd}}\mathbb{E}_{\mathcal{D}^1\sim\mathbb{P}_1}\frac{K(\frac{x-X_1-\eta}{h_n})}{c+K(\frac{x-X_1}{h_n})+K(\frac{x-X_1-\eta}{h_n})+\sum_{X\in\mathcal{D}^{1,\text{bd}}\backslash X_1} K(\frac{x-X}{h_n})+\sum_{X\in\mathcal{D}^{1,\text{bd}}\backslash X_1} K(\frac{x-X-\eta}{h_n})}, &(29)\\
&=n_{\text{bd}}\mathbb{E}_{\mathcal{D}^1\backslash X_1\sim\mathbb{P}_1}(\frac{0}{c+0+\sum_{X\in\mathcal{D}^{1,\text{bd}}\backslash X_1}(K(\frac{x-X}{h_n})+K(\frac{x-X-\eta}{h_n}))})\mathbb{P}_1(K(\frac{x-X_1-\eta}{h_n})=0,K(\frac{x-X_1}{h_n})=0)\\
&+n_{\text{bd}}\mathbb{E}_{\mathcal{D}^1\backslash X_1\sim\mathbb{P}_1}(\frac{0}{c+1+\sum_{X\in\mathcal{D}^{1,\text{bd}}\backslash X_1}(K(\frac{x-X}{h_n})+K(\frac{x-X-\eta}{h_n}))})\mathbb{P}_1(K(\frac{x-X_1-\eta}{h_n})=0,K(\frac{x-X_1}{h_n})=1)\\
&+n_{\text{bd}}\mathbb{E}_{\mathcal{D}^1\backslash X_1\sim\mathbb{P}_1}(\frac{1}{c+1+\sum_{X\in\mathcal{D}^{1,\text{bd}}\backslash X_1}(K(\frac{x-X}{h_n})+K(\frac{x-X-\eta}{h_n}))})\mathbb{P}_1(K(\frac{x-X_1-\eta}{h_n})=1,K(\frac{x-X_1}{h_n})=0)\\
&+n_{\text{bd}}\mathbb{E}_{\mathcal{D}^1\backslash X_1\sim\mathbb{P}_1}(\frac{1}{c+2+\sum_{X\in\mathcal{D}^{1,\text{bd}}\backslash X_1}(K(\frac{x-X}{h_n})+K(\frac{x-X-\eta}{h_n}))})\mathbb{P}_1(K(\frac{x-X_1-\eta}{h_n})=1,K(\frac{x-X_1}{h_n})=1), &(30)\\
&=n_{\text{bd}}\mathbb{E}_{\mathcal{D}^1\backslash X_1\sim\mathbb{P}_1}(\frac{1}{c+1+\sum_{X\in\mathcal{D}^{1,\text{bd}}\backslash X_1}(K(\frac{x-X}{h_n})+K(\frac{x-X-\eta}{h_n}))})\mathbb{P}_1(K(\frac{x-X_1-\eta}{h_n})=1), &(31)\\
&=n_{\text{bd}}\mathbb{E}_{Z\sim\text{Bino}(n_{\text{bd}}-1,\mathbb{P}_1(\mathbb{B}_{x,h_n})+\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n}))}(\frac{1}{Z+c+1})\mathbb{P}_1(K(\frac{x-X_1-\eta}{h_n})=1), &(32)\\
&\geq n_{\text{bd}}\frac{1}{(n_{\text{bd}}-1)(\mathbb{P}_1(\mathbb{B}_{x,h_n})+\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n}))+c+1}\mathbb{P}_1(B_{x-\eta,h_n}), &(33)\\
&=\frac{n_{\text{bd}}\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n})}{(n_{\text{bd}}-1)(\mathbb{P}_1(\mathbb{B}_{x,h_n})+\mathbb{P}_1(\mathbb{B}_{x-\eta,h_n}))+c+1},
\end{aligned}
$$

where (28) follows the linearity of expectation, and (29) is because $\mathcal{D}^1$ consist of i.i.d samples and the symmetric property. Additionally, the term in (30) equals 0 because the event $X_1\in\mathbb{B}_{x-\eta,h_n}$ and $X_1\in\mathbb{B}_{x,h_n}$ contradict with each other since $2h_n<\|\eta\|_2$. Similarly, (31) holds from

$$
\begin{aligned}
&\mathbb{P}_1(K(\frac{x-X_1-\eta}{h_n})=1,K(\frac{x-X_1}{h_n})=0)=\\
&\mathbb{P}_1(K(\frac{x-X_1}{h_n})=0|K(\frac{x-X_1-\eta}{h_n})=1)\cdot\mathbb{P}_1(K(\frac{x-X_1-\eta}{h_n})=1)=\mathbb{P}_1(K(\frac{x-X_1-\eta}{h_n})=1).
\end{aligned}
$$

Finally, (32) follows from the fact that $\sum_{X\in\mathcal{D}^1_{\text{bd}}\backslash X_1} K(\frac{x-X}{h_n})+K(\frac{x-X-\eta}{h_n})$ is a binomial random variable with no. of trials $(n_{\text{bd}}-1)$ and the success probability $\mathbb{P}_1(\mathbb{B}_{x,h_n})+\mathbb{P}_1(B_{x-\eta,h_n})$, and the inequality in (33) is due to the Jensen's inequality. □

C. Omitted Lemmas

C.1. Proof of Example 1

Proof. Recall that η satisfies $\|\eta + \mu_1\| > 7/8\|\eta\|$, $\|\eta + \mu_0\| > 7/8\|\eta\|$, $\|\eta\|/4 > \|\mu_1\|, \|\mu_0\|$, and $0.5\|\eta\| > \sqrt{\|\eta\|}$. Also, we have $\mu_1 + \mu_0 = 0$.

We have

$$\begin{aligned}
&\mathbb{P}_X(\mathbb{R}^d \setminus C_r) \\
=&\mathbb{P}_X(\|X\|_2^2 \geq r^2) \\
=&k\mathbb{P}_0(\|X\|_2^2 \geq r^2) + (1-k)\mathbb{P}_1(\|X\|_2^2 \geq r^2) && (34) \\
\leq&k\mathbb{P}_0(\|X-\mu_0\|_2^2 \geq r^2/4) + (1-k)\mathbb{P}_1(\|X-\mu_1\|_2^2 \geq r^2/4) && (35) \\
\leq&\exp\left(-r^2/80\lambda_{d-1}\right), && (36)
\end{aligned}$$

where the equality in (34) is by assuming that class-conditional conditional distributions are Gaussians with prior k and the total law of probability. Additionally, the inequality in (35) holds because $\{x|\|x-\mu_1\| \leq \|\eta\|/4\} \subset \{x|\|x\| \leq \|\eta\|/2\}$ and $\{x|\|x-\mu_0\| \leq \|\eta\|/4\} \subset \{x|\|x\| \leq \|\eta\|/2\}$ provided that $\|\mu_0\|, \|\mu_1\| \leq \|\eta\|/4$. Finally the inequality in (36) follows standard tail inequalities of Chi-square distributions.

□

D. Experiment configurations

D.1. Computing Environments

All of our experiments are conducted on a workstation with one A100 GPU.

D.2. Data Descriptions

We employed three computer vision datasets of varying complexity, all of which have been utilized in prior research. This allows for a increased confidence in the validity of the proposed hypothesis.

MNIST: The MNIST dataset comprises of $70,000$ grayscale images with a resolution of 28×28, divided into a training set of $60,000$ images and a test set of $10,000$ images. To enhance the performance of the training process, we employed data augmentation techniques, such as random cropping and rotation. However, during the evaluation stage, no additional augmentation was applied. Some examples are shown below.

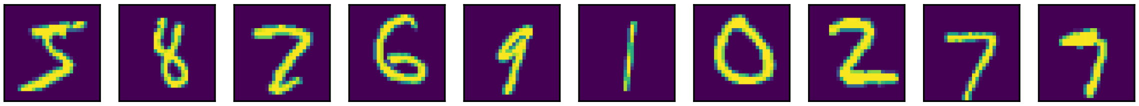


Figure 7. Examples of MNIST

CIFAR10: The CIFAR-10 dataset is one of the most widely used datasets for machine learning research. It contains $60,000$ color images with a resolution of 32×32, divided into 10 classes with $6,000$ images per class. The dataset is split into a training set of $50,000$ images and a test set of $10,000$ images. To enhance the performance of the training process, we employed data augmentation techniques, such as random cropping and rotation. However, during the evaluation stage, no additional augmentation was applied. Some examples are shown below.

Figure 8. Examples of CIFAR10

GTSRB The German Traffic Sign Recognition Benchmark (GTSRB) dataset is gaining popularity in the field of Backdoor Learning. The dataset comprises of $60,000$ images with 43 classes and varying resolution from 32×32 to 250×250. It is divided into a training set of $39,209$ images and a test set of $12,630$. To enhance the performance of the training process, we resized to 32×32 pixels and employed data augmentation techniques, such as random cropping and rotation. However, during the evaluation stage, no additional augmentation was applied.

D.3. Model Configurations & Training Schedule

We consider 3 representative CNN classifiers, namely LeNet5, ResNet 9/20 and VGG 16, and summarize their usage in Table 1.

Table 1. Summary of data and models

Dataset	Model
MNIST	LeNet 5, ResNet 9
CIFAR10	ResNet 20, VGG 16
GTSRB	ResNet 20

For ResNet and VGG models, we adopt the standard training pipeline of SGD with a momentum of 0.9, a weight decay of 10^{-4} , and a batch size of 128 for optimization. For LetNet, we adopt the standard training pipeline of SGD with the initial learning rate of $0.1/0.01$.

D.4. State-of-the-art Backdoor Attacks

Below are examples of three backdoor attacks on CIFAR10.

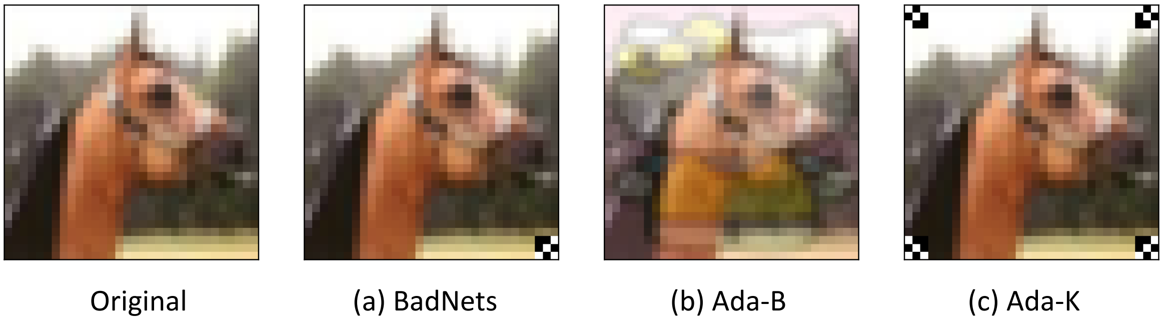


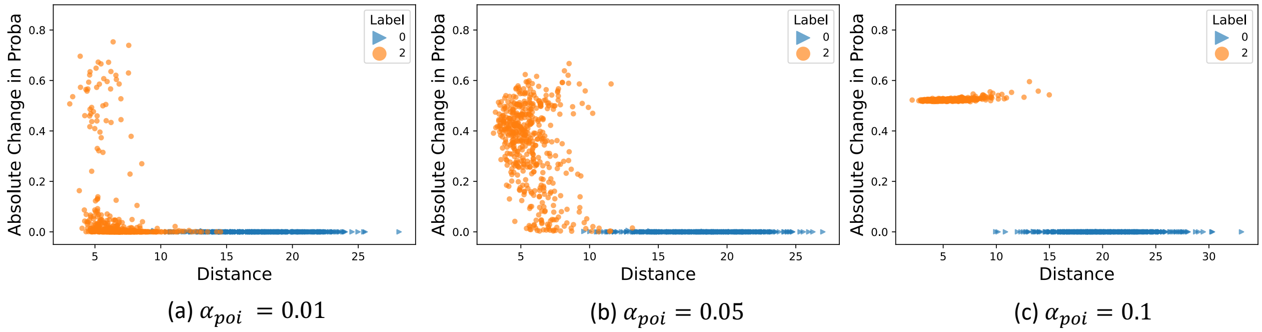
Figure 9. Illustration of an original image from CIFAR10 and its backdoor versions: (a) BadNets, (b) Adaptive Blend, and (c) Adaptive K-triggers

E. Additional Experiments for validating the Adaptability Hypothesis

E.1. Experimental Results for testing MNIST with LeNet/ResNet

The Adaptability under different poisoning ratios α_{poi}. In Figure 10, we observe that the change in predicted values decreases as the distance to the backdoor distribution increases, and that the average change in predicted values increases as the poisoning ratio increases, consistent with our theory.

The Adaptability under different source class for backdoor. We set the source class of images to be backdoored to be 4. Figures 11 and 12 show results on MNIST with LeNet with different backdoor trigger intensity and the poisoning ratio α_{poi}. We observe that that as the intensity of the backdoor triggers increases, the average change in predicted probabilities decreases, consistent with the hypothesis. Also, in each figure, as the distance from a point to the backdoor distribution increases, its corresponding change in f^{cl} and f^{poi} decreases, align with our results. Additionally, as the ratio increases, the average change in predicted probability also increases, which conforms to our results.

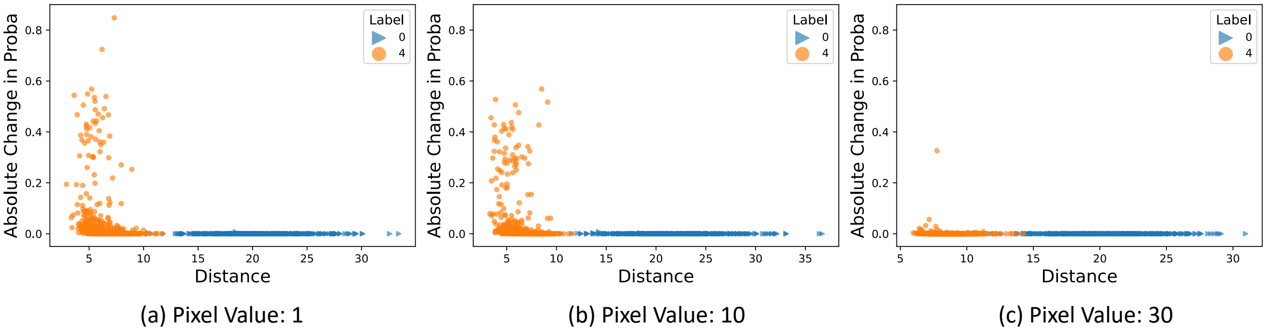


(a) α_{poi} = 0.01

(b) α_{poi} = 0.05

(c) α_{poi} = 0.1

Figure 10. Case studies on BadNets attacks with different backdoor poisoning ratio α_{poi}, on MNIST and LetNet 5.



(a) Pixel Value: 1

(b) Pixel Value: 10

(c) Pixel Value: 30

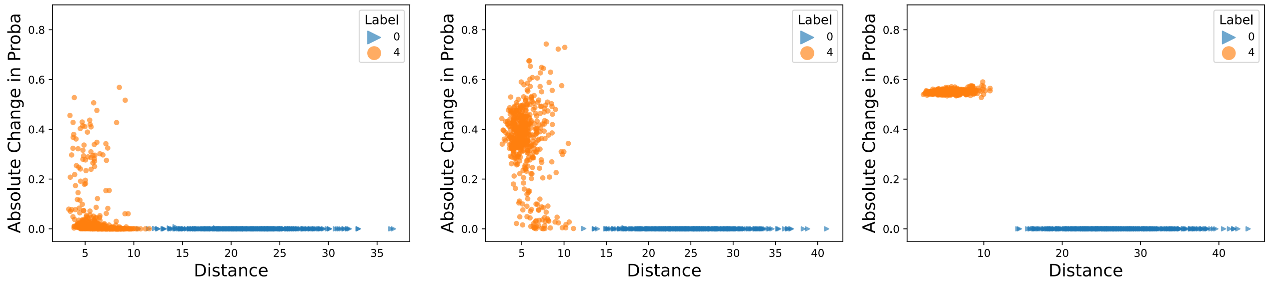
Figure 11. Case studies on BadNets attacks with different levels of patches, namely the pixel value, on MNIST and LetNet 5.

E.2. Additional Experimental Results for testing the Adaptability Hypothesis on CIFAR10/GTSRB with ResNet

We train ResNet on the GTSRB to demonstrate the adaptability.

F. Additional Experimental Results for verifying the effective directions for SOTA attacks

We plot the dimensional Mahalanbois distances between clean data and backdoor data for three attacks with different poisoning ratios in Figures 14 and 15. We observe that effective attacks tend to have a larger relative distance at the dimension of data with low variances, supporting our results.

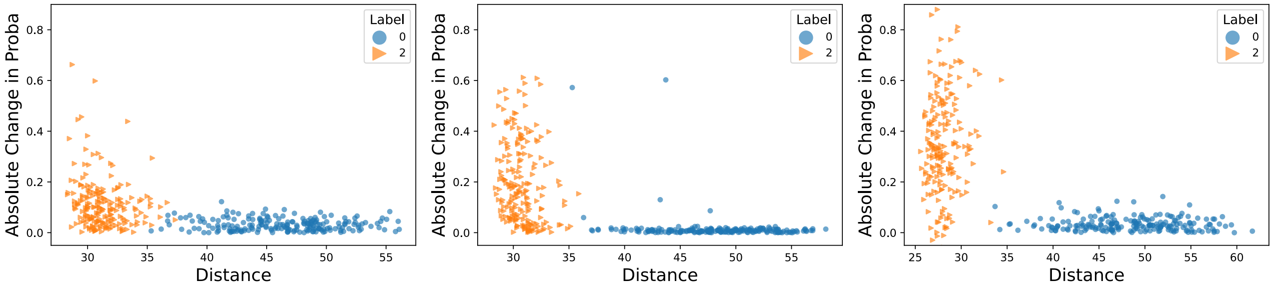


(a)α_{poi} = 0.01

(b) α_{poi} = 0.05

(c) α_{poi} = 0.1

Figure 12. Case studies on BadNets attacks with different backdoor poisoning ratio α_{poi}, on MNIST and LetNet 5.



(a) α_{poi} = 0.01

(b) α_{poi} = 0.05

(c) α_{poi} = 0.1

Figure 13. Case studies on BadNets attacks with different backdoor poisoning ratio α_{poi}, on GTSRB and ResNet 20.

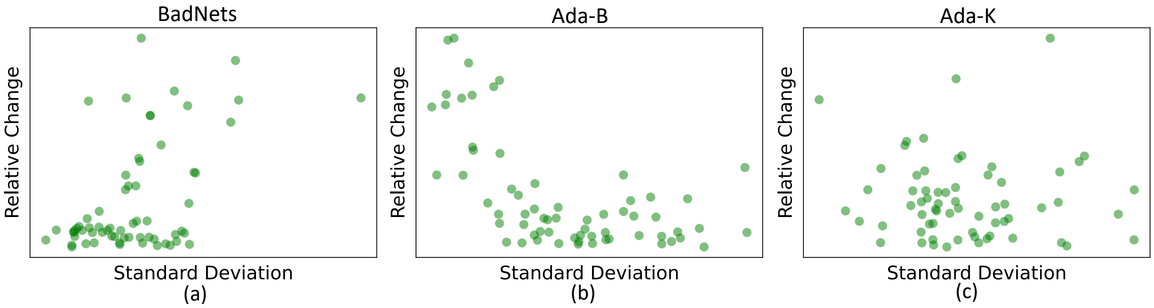


Figure 14. Illustrations of dimensional Mahalanbois distance for three attacks on CIFAR10 and ResNet 20 with $\alpha_{\text{poi}} = 0.01$. In each figure, each point's horizontal value represents the standard deviation (std) of one dimension of backdoor data, i.e., the standard deviation of the jth dimension of data, and its vertical value is the relative change along the same dimension, i.e., the difference (in absolute value) between clean and backdoor data along the jth dimension divided by the std of the jth dimension. We observe that for Ada-B attacks, they tend to have a larger (smaller) relative distance on the dimension of data with lower (higher) std, consistent with our theory.

G. Pseudo-code for visualisation algorithms and additional experimental results

We give the pseudo-code for our visualizing algorithm in this section. Additionally, we provide two extra empirical studies on the effect of backdoor sample size ratio α_{poi} as demonstrated in Figures 16 and 17. We observe that CNNs latent spaces are well separated under our transformation with Algorithm 1, yielding the effectiveness of our proposed method.

H. More results on VGG architectures

We apply our algorithms for visualizing to the latent embeddings of VGG 16, with a dimension of 1024, with CIFAR 10 Dataset and the Ada-k attack, as illustrated in Figure 18. We observe that our algorithms can still visually separate between clean and backdoor data.

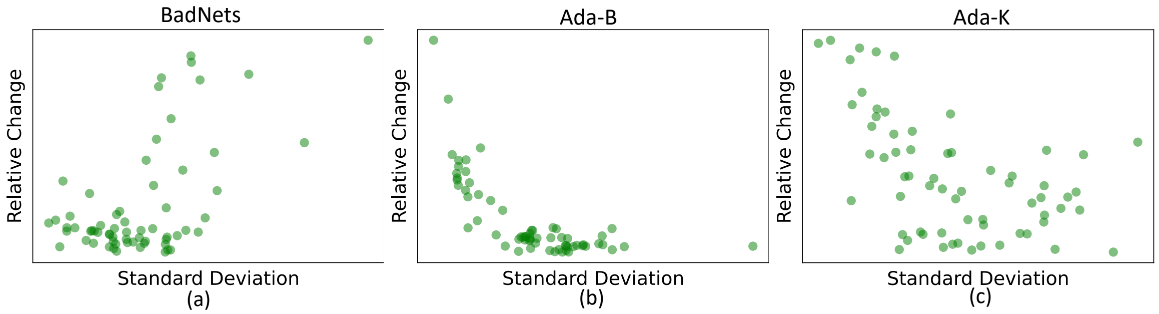


Figure 15. Illustrations of dimensional Mahalanbois distance for three attacks on CIFAR10 and ResNet 20 with $\alpha_{\text{poi}} = 0.05$. In each figure, each point's horizontal value represents the standard deviation (std) of one dimension of backdoor data, i.e., the standard deviation of the jth dimension of data, and its vertical value is the relative change along the same dimension, i.e., the difference (in absolute value) between clean and backdoor data along the jth dimension divided by the std of the jth dimension. We observe that for Ada-K and Ada-B attacks, they tend to have a larger (smaller) relative distance on the dimension of data with lower (higher) std, consistent with our theory.

Algorithm 1 Visualizing high-dimensional data

Input: Data $\{(x_i, y_i)\}_{i=1}^n$, Number of Classes K, An empty array $Q = [q_1, \dots, q_n]$ of dimension $n \times K$

```
for $c = 1$ to $K$ do
    Calculate class-conditional mean $\mu_c$ and covariance $\Sigma_c$
end for
for $i = 1$ to $n$ do
    for $c = 1$ to $K$ do
        Append $(x_i - \mu_c)^\top \Sigma_c^{-1} (x_i - \mu_c)$ to $q_i$ // Class-conditional Mahalanbois Distance
    end for
end for
```

Output: Transformed data Q

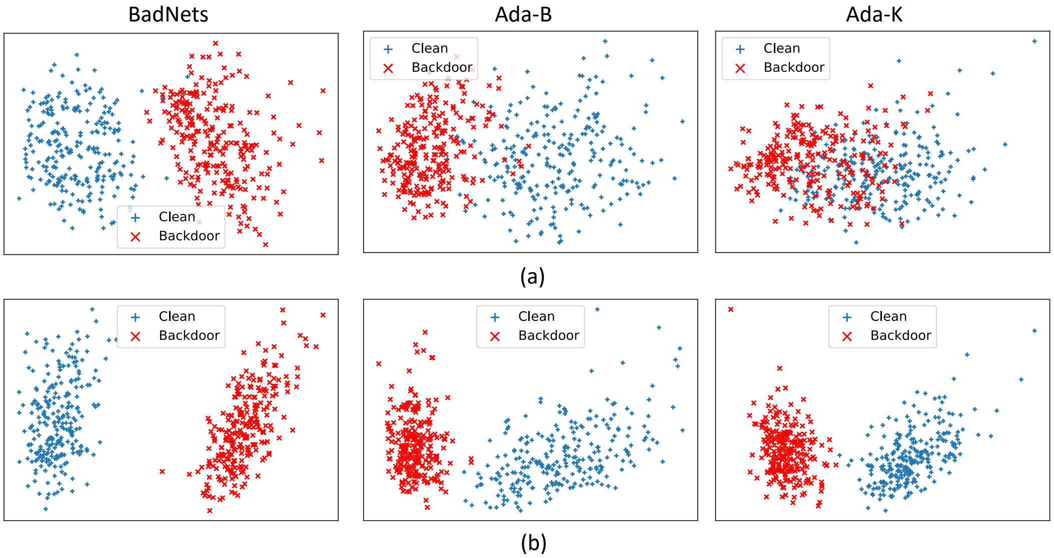
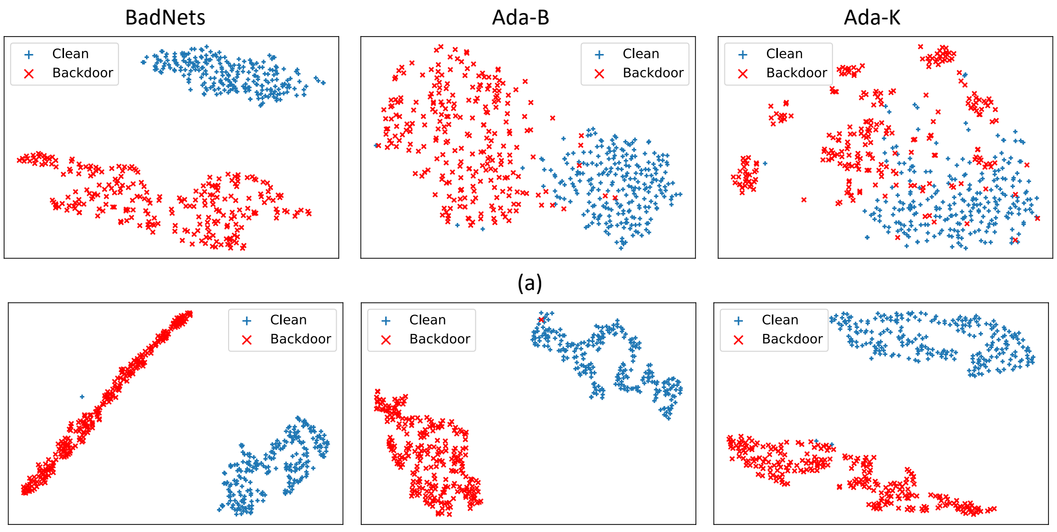


Figure 16. PCA visualizations of (a) the original and (b) our theoretically transformed (via Algorithm 1) latent spaces of ResNet 20 on CIFAR10 $\alpha_{\text{poi}} = 1\%$ three SOTA Attacks. The latent spaces of Ada-B and Ada-K attacks do not exhibit a clear separation of clusters in their original form. However, our theoretical results demonstrate that a transformed version of these spaces does display two distinct and separate clusters.


BadNets

Ada-B

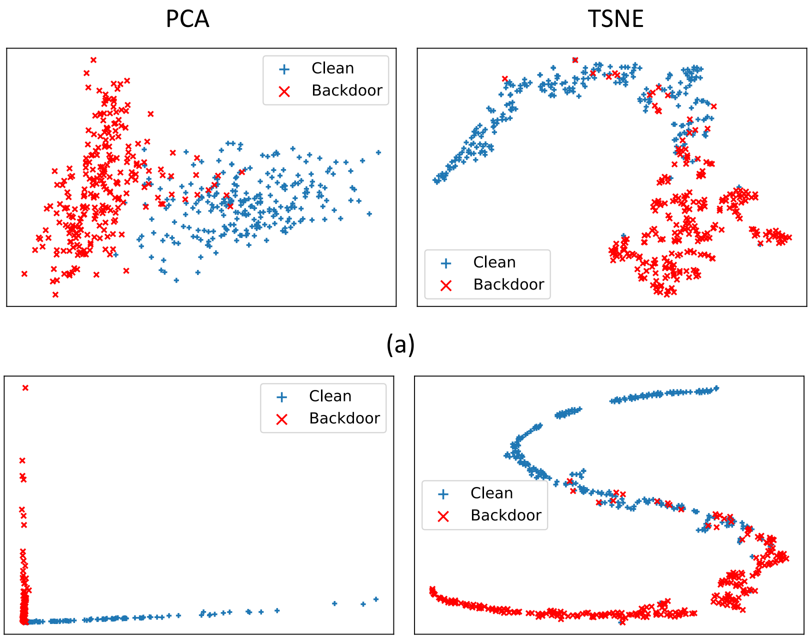
Ada-K

(a)

(b)

(b)

Figure 17. TSNE visualizations of (a) the original and (b) our theoretically transformed (via Algorithm 1) latent spaces of ResNet 20 on CIFAR10 $\alpha_{\text{poi}} = 1\%$ three SOTA Attacks. The latent spaces of Ada-B and Ada-K attacks do not exhibit a clear separation of clusters in their original form. However, our theoretical results demonstrate that a transformed version of these spaces does display two distinct and separate clusters.


PCA

TSNE

(a)

(b)

Figure 18. Visualizations on the latent spaces of VGG 16 on CIFAR10 with Ada-K attacks.