A Theoretical Analysis of Backdoor Poisoning Attacks in Convolutional Neural Networks

Boqi Li [1] **Weiwei Liu** [1]

Abstract

The rising threat of backdoor poisoning attacks (BPAs) on Deep Neural Networks (DNNs) has become a significant concern in recent years. In such attacks, the adversaries strategically target a specific class and generate a poisoned training set. The neural network (NN), well-trained on the poisoned training set, is able to predict any input with the trigger pattern as the targeted label, while maintaining accurate outputs for clean inputs. However, why the BPAs work remains less explored. To fill this gap, we employ a dirty-label attack and conduct a detailed analysis of BPAs in a two-layer convolutional neural network. We provide theoretical insights and results on the effectiveness of BPAs. Our experimental results on two real-world datasets validate our theoretical findings.

1. Introduction

The security of DNNs has become a significant concern in recent years (Zhang et al., 2019; Nguyen et al., 2023; Zhou & Liu, 2023; Li & Liu, 2023). Most state-of-the-art models require huge training data (Schmidt et al., 2018; Dosovitskiy et al., 2021; Wang et al., 2023b), but the training data from unreliable data sources is vulnerable to data poisoning attacks (Shafahi et al., 2018; Cinà et al., 2021; Koh et al., 2022), for example, BPA. BPA (Gu et al., 2017; Jha et al., 2023) is a training-time attack, which embeds backdoors into the NN by providing poisoned data to users. In BPA, the adversary firstly targets a class and generates the poisoned data with a special pattern, called a trigger pattern. The two primary categories of BPA algorithms include clean-label attacks (Shafahi et al., 2018; Turner et al., 2019; Barni et al., 2019) and dirty-label attacks (Gu et al., 2017; Chen et al., 2017; Nguyen & Tran, 2021). The clean-label attacks modify only the inputs, while the dirty-label attacks modify both inputs and labels at the same time. The adversary adds a small partition of poisoned data to the training set, and the NN, well-trained on the poisoning training set, is consequently endowed with hidden backdoors. The backdoored NN predicts all clean data as the same as a normal model but misbehaves when a specific trigger pattern appears.

BPA, especially dirty-label attack, can successfully compromise a model by merely adding a small partition of poisoned data to the training set. The security of DNNs has been widely studied from a theoretical perspective in recent years (Xu & Liu, 2022; Ma et al., 2022; Zou & Liu, 2023). However, the theoretical reasons behind the effectiveness of BPA algorithms remain less explored. To fill this gap, we investigate the dirty-label attacks in a two-layer convolutional neural network utilizing a multi-view data model in this paper. To the best of our knowledge, this work is the first to theoretically analyze the effectiveness of BPA by studying the learning process in a convolutional neural network.

In this paper, we provide theoretical insights on backdoor learning, which trains the model over poisoned data. Specifically, we analyze the outputs of a backdoored network, to identify the sufficient conditions for successful BPA algorithms. We compare the update rules for standard and backdoor learning to understand the difference in the training process over clean and poisoned data. Moreover, motivated by the technique of Shen et al. (2022), we study the dynamic of backdoor learning, and investigate the time cost associated with learning the main features and trigger patterns of training data. We also show the effectiveness of dirty-label attack. A formal theoretical result is presented regarding the effectiveness of BPA. Our results indicate that the success of BPA relies on three key components: the number of feature vectors in the datasets, the norm ratio of trigger pattern and feature vector, and the percentage of poisoned data in the training set.

Empirically, we present experimental results on two real-world datasets to validate our theoretical insights and findings. Our analysis involves a comparison of the loss of

[1] School of Computer Science, National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan, China. Correspondence to: Weiwei Liu <liuweiwei863@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

poisoned and clean data. Additionally, we investigate the gradients of weights with respect to (w.r.t.) the training loss. We also study the poisoned and clean data with the representation of a backdoored NN. Lastly, we discuss which components affect the success of BPA.

2. Related Work

Theoretical Analysis on Backdoor Poisoning Attack To our best knowledge, there are only a few works theoretically analyzing the effectiveness of BPA algorithms. Manoj & Blum (2021) present a property of the function class, called memorization capacity. They show that non-zero memorization capacity implies the existence of a BPA to succeed. Xian et al. (2023) show that if the hypothesis class is adaptive w.r.t. the distribution of poisoned data, the BPA can successfully embed the backdoor into the NN. Wang et al. (2023a) study the backdoor attacks from a statistical perspective, and they focus on the statistical risk of backdoored model on both clean and poioned data. All of these works study this problem from the perspective of hypothesis class, while our work aims to investigate the BPA by analyzing the process of backdoor learning in a two-layer convolution neural network.

Theoretical Analysis on Learning Process Recently, a lot of works (Allen-Zhu & Li, 2021; Jelassi & Li, 2022; Shen et al., 2022) analyze deep learning as a feature learning process. Allen-Zhu & Li (2021) and Wen & Li (2021) investigate the effectiveness of adversarial training and self-supervised contrastive learning in a two-layer ReLU neural network based on the sparse coding model. Shen et al. (2022) and Allen-Zhu & Li (2023) use the multi-view model to study the knowledge distillation and data augmentation, respectively. In this paper, inspired by Shen et al. (2022); Allen-Zhu & Li (2023), we use a multi-view data model for BPA.

3. Preliminaries

Consider a binary classification problem. Let $\mathcal{D}_z$ be the distribution of $Z = (X, Y)$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We follow Shen et al. (2022) to use a multi-view data model. In a multi-view data model, each data point $\mathbf{x}$ consists of P non-overlapped patches $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^P) \in \mathbb{R}^{d\times P}$, and each patch is a vector with dimension d. We assume that there exists K orthogonal features $\mathbf{u}^1, \ldots, \mathbf{u}^K$ useful for classification with the same norm. Let $\mathcal{D}_u$ be a discrete distribution over these features. Additionally, there exists a main noise vector $\boldsymbol{\xi} \in \mathbb{R}^d$ and background noise vectors $\{\boldsymbol{\zeta}^p\}_{p=1}^{P-2}$ in $\mathbf{x}$, and the distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are denoted by $\mathcal{D}_\xi$ and $\mathcal{D}_\zeta$, respectively. Let $[n] = \{1, \ldots, n\}$, We define the feature-noise multi-view data model as follows:

Definition 3.1. Given feature distribution $\mathcal{D}_u$, and noise distributions $\mathcal{D}_\xi$ and $\mathcal{D}_\zeta$, a data point $\mathbf{z} = (\mathbf{x}, y)$ is drawn from the distribution $\mathcal{D}_{\mathbf{z}}^n$ which is defined as follows:

1. Draw the label $y \in \{+1, -1\}$ uniformly.
2. Given y, arbitrarily choose two patches p_u, p_ξ, where $p_\xi \neq p_u$. The feature patch $\mathbf{x}^{p_u}$ is set as $\mathbf{x}^{p_u} = y\mathbf{u}$, where $\mathbf{u} \sim \mathcal{D}_u$, and the noise patch is set as $\mathbf{x}^{p_\xi} = \boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{D}_\xi$.
3. Each remaining background patch $p_\zeta \in [P] \setminus \{p_u, p_\xi\}$ of $\mathbf{x}$ is set as $\mathbf{x}^{p_\zeta} = \boldsymbol{\zeta}$, where $\boldsymbol{\zeta} \sim \mathcal{D}_\zeta$.

In this paper, we assume $\mathcal{D}_u$ is a discrete uniform distribution, i.e. $\forall k \in [K], \mathbb{P}[\mathbf{u} = \mathbf{u}^k] = 1/K$, and $\mathcal{D}_\xi$ and $\mathcal{D}_\zeta$ are two zero-mean Gaussian distributions $\mathcal{N}(0, \frac{\sigma_\xi^2}{d}\mathbf{I}_d)$, $\mathcal{N}(0, \frac{\sigma_\zeta^2}{d}\mathbf{I}_d)$. We set $\sigma_\zeta = \sigma_\xi P^{-1}$ to limit the variance of the background noise. We use $\mathcal{S}$ to denote the set of data points, and $\mathcal{I}$ to denote the index set of the set $\mathcal{S}$. The clean training set $\mathcal{S}_{cl}^{tr}$ consists of n independent and identically distributed (i.i.d.) data points drawn from $\mathcal{D}_{\mathbf{z}}^n$. Consider a BPA algorithm $\mathfrak{P} = (\mathfrak{P}^X, \mathfrak{P}^Y) : \mathcal{Z} \to \mathcal{Z}$, where $\mathfrak{P}^X : \mathcal{X} \to \mathcal{X}$ generates the poisoned input from a clean input, and $\mathfrak{P}^Y : \mathcal{Y} \to \mathcal{Y}$ modifies the label to the targeted label. We then construct the poisoned training set $\mathcal{S}_{po}^{tr}$ by randomly applying the backdoor attack to the data points in $\mathcal{S}_{cl}^{tr}$ as follows:

Definition 3.2. Given a set $\mathcal{S}_{cl}^{tr}$, the number of poisoned data n_{po} and a BPA algorithm $\mathfrak{P}$, The poisoned training set $\mathcal{S}_{po}^{tr}$ is constructed as below:

1. Randomly select a set of an index set $\mathcal{S}_b$ with $|\mathcal{S}_b| = n_{po}$ from the non-targeted class, ensuring each data point is chosen uniformly.
2. Given $\mathcal{S}_b$, let $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ and $\hat{\mathbf{z}}_i = (\hat{\mathbf{x}}_i, \hat{y}_i)$ be the i-th data point in $\mathcal{S}_{cl}^{tr}$ and $\mathcal{S}_{po}^{tr}$, respectively, then for any $i \in [n]$,

$$\hat{\mathbf{z}}_i = \begin{cases} \mathfrak{P}(\mathbf{z}_i) & \text{if } i \in \mathcal{I}_b, \\ \mathbf{z}_i & \text{if } i \notin \mathcal{I}_b. \end{cases}$$

In this paper, we study the dirty-label backdoor attack. In a dirty-label backdoor attack, the adversary firstly chooses a label y^p, and adds the trigger pattern to inputs with $\mathfrak{P}^X$ for data points that have a different label with y^p, the adversary also flips the labels of the chosen data points to y^p. Patch attack (Gu et al., 2017; Chen et al., 2017) is one of dirty-label backdoor attacks, which chooses a fixed patch p_v, and uses a specific trigger vector to replace p_v of clean data.

Definition 3.3 (Patch attack)**.** Given a trigger $\mathbf{v}$, a user-defined backdoor patch p_v, and the targeted label y^p. The Patch attack $\mathfrak{P}_{patch}(\cdot; p_v, \mathbf{v}, y^p) : \mathcal{Z} \to \mathcal{Z}$ is defined as:

$$\mathfrak{P}_{patch}^X(\mathbf{x}; p_v, \mathbf{v})^p = \begin{cases} \mathbf{x}^p & \text{if } p \neq p_v, \\ \mathbf{v} & \text{if } p = p_v, \end{cases} \quad \mathfrak{P}_{patch}^Y(y; y^p) = y^p.$$

y^p, $\mathbf{v}$, and p_v are all user-specific in practice, and the adversary aims to choose patch p_v that is not related to the main feature, for example, the corner of an image. To simplify the problem, we assume p_v is chosen to be one of the background patches, i.e., $p_v \in [P] \setminus \{p_u, p_\xi\}$. We use $\mathcal{P}^\zeta$ to denote the set of background patches, and $\mathcal{P}^\zeta = [P] \setminus \{p_u, p_\xi\}$ for clean data while $\mathcal{P}^\zeta = [P] \setminus \{p_u, p_\xi, p_v\}$ for poisoned data.

We use a patch-wise convolutional neural network architecture $F(\mathbf{x})$ with C channels, which is defined as

$$F(\mathbf{x}) = \sum_{c=1}^{C} \lambda_c \sum_{p=1}^{P} \phi\left(\langle \mathbf{w}_c, \mathbf{x}^p \rangle\right),$$

where

$$\phi(z) = \begin{cases} \frac{1}{q} z^q & \text{if } |z| \leq 1 \\ z + \frac{1}{q} & \text{if } z > 1 \\ z - \frac{1}{q} & \text{if } z < 1, \end{cases}$$

is the activation function. This activation function is a smoothed version of symmetrized ReLU, and has been adopted in a line of theoretical works (Shen et al., 2022; Yang et al., 2023). We follow Yang et al. (2023) to use the activation function with $q = 3$, and our results are easily extended to any $q > 3$. For F, We follow Shen et al. (2022) to fix the weights of the second layer as an all-one vector, i.e.,$\forall c \in [C], \lambda_c = 1$, and only consider the change of trainable parameters $\{\mathbf{w}_1, \ldots, \mathbf{w}_C\}$ of the first layer.

We use the logistic loss $\ell(F(\mathbf{x}), y) = \log\left(1 + e^{-yF(\mathbf{x})}\right)$ as the loss function, and use gradient descent (GD) to optimize the parameters. The network predicts label with $y' = \operatorname{sign}(F(\mathbf{x}))$, where $\operatorname{sign}(\cdot)$ denotes the sign function.

Gaussian initialization is used to initialize the weights of the model, i.e. $\mathbf{w}_c(0) \sim \mathcal{N}(0, \sigma_0 \mathbf{I}_d)$. Given a learning rate η, at round t, the parameters of the network are updated by

$$\begin{aligned} \mathbf{w}_c(t+1) &= \mathbf{w}_c(t) - \frac{\eta}{n} \sum_{i=1}^{n} \ell'(F(\hat{\mathbf{x}}_i), y_i) \nabla_{\mathbf{w}_c} \ell(F(\hat{\mathbf{x}}_i), y_i) \\ &= \mathbf{w}_c(t) - \frac{\eta}{n} \sum_{i=1}^{n} \sum_{p=1}^{P} y_i \ell'(F(\hat{\mathbf{x}}_i), y_i) \phi'(\langle \mathbf{w}_c(t), \hat{\mathbf{x}}_i^p \rangle) \hat{\mathbf{x}}_i^p. \quad (1) \end{aligned}$$

The process of backdoor learning is that the attacker firstly generates the poisoned training set $\mathcal{S}_{po}^{tr}$, the user then runs GD algorithm on the poisoned training set $\mathcal{S}_{po}^{tr}$ with T rounds to obtain $\hat{F}_T$. The attacker's goal is that $\hat{F}_T$ achieves both high clean accuracy and high attack success rate of poisoned data. The clean accuracy is defined as

$$\operatorname{Acc}(\hat{F}_T; \mathcal{D}_{\mathbf{z}}) = \mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}_{\mathbf{z}}}[\hat{F}_T(\mathbf{x}) = y],$$

and the attack success rate, is defined as

$$\operatorname{ASR}(\hat{F}_T; \mathcal{D}_{\mathbf{z}}, \mathfrak{P}, y^p) = \mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}_{\mathbf{z}}}[\hat{F}_T(\mathfrak{P}^X(\mathbf{x})) = y^p | y \neq y^p].$$

We use the standard asymptotic notations O, Θ, Ω in this paper. Given $f : \mathcal{R} \to \mathcal{R}_+$ and $g : \mathcal{R} \to \mathcal{R}_+$, we denote $f \leq O(g)$ if there exists $x_0, \alpha \in \mathcal{R}$ such that for all $x > x_0$, we have $f(x) \leq \alpha g(x)$. We denote $f \geq \Omega(g)$ if there exists $x_0, \alpha \in \mathcal{R}$ such that for all $x > x_0$, we have $f(x) \geq \alpha g(x)$. The notation $f = \Theta(g)$ means that $f \geq \Omega(g)$ and $f \leq O(g)$. We use $f \leq o(g)$ to denote that for every $\alpha > 0$, there exists x_0 such that for all $x > x_0$ we have $f(x) \leq \alpha g(x)$. We use $f \geq \omega(g)$ to denote that for every $\alpha > 0$, there exists x_0 such that for all $x > x_0$ we have $f(x) \geq \alpha g(x)$. Finally, we use $\widetilde{O}, \widetilde{\Theta}, \widetilde{\Omega}$ to hide the log factors in O, Θ, Ω, respectively.

4. Theoretical Insights on Backdoor Learning

In this section, we show theoretical insights on backdoor learning in two aspects: The outputs of a backdoored network F and the difference of update rules between standard and backdoor learning. We then analyze the dynamic of backdoor learning and the effectiveness of dirty-label attack.

4.1. The Outputs of a Backdoored Network

In this subsection, we assume $\sigma_\xi = \sigma_\zeta = 0$, which means the image only contains feature and trigger vectors. We further assume that $\mathbf{v}$ is orthogonal to all feature vectors $\left\{\mathbf{u}^k\right\}_{k=1}^{K}$. Consider a model F, and two data points $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$, where $\mathbf{x}_1$ belongs to the targeted class, and $\mathbf{x}_2$ belongs to another class, i.e., $y_2 \neq y_1 = y^p$. We suppose the feature vectors in $\mathbf{x}_1$ and $\mathbf{x}_2$ are $y_1 \mathbf{u}^k$ and $y_2 \mathbf{u}^k$, respectively. If F is well-trained on the poisoned set $\mathcal{S}_{po}^{tr}$ with rounds T, we intuitively have the following results: (1) F can correctly classify $\mathbf{x}_1$ and $\mathbf{x}_2$ with a high probability; (2) The trigger vector has been effectively captured by F, and F predicts $\mathfrak{P}(\mathbf{x}_1)$ to the targeted class, i.e., $y_1 F(\mathbf{x}_1) + y_1 F(\mathfrak{P}(\mathbf{x}_1)) > y_1 F(\mathbf{x}_1) > 0$; (3) F predicts $\mathfrak{P}(\mathbf{x}_1)$ to the targeted class, as well. Since $y^p \neq y_2$, we have $y_2 F(\mathbf{x}_2) - y_2 F(\mathfrak{P}(\mathbf{x}_2)) > y_2 F(\mathbf{x}_2) > 0$. Recall the decomposition of data points, and vectors in $\left\{\mathbf{u}^k\right\}_{k \in [K]} \cup \{\mathbf{v}\}$ are mutually orthogonal in pairs, $\forall (\mathbf{x}, y) \in \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\}$, we yield that

$$0 < y \sum_{c=1}^{C} \phi\left(\left\langle \mathbf{w}_c(T), y \mathbf{u}^k \right\rangle\right) \left\|\mathbf{u}^k\right\|_2^2 < y^p \sum_{c=1}^{C} \phi(\langle \mathbf{w}_c(T), \mathbf{v} \rangle) \|\mathbf{v}\|_2^2. \quad (2)$$

$\phi(\cdot)$ is an odd function, which implies that $y\phi\left(\left\langle \mathbf{w}_c(T), y\mathbf{u}^k \right\rangle\right) = \phi\left(\left\langle \mathbf{w}_c(T), \mathbf{u}^k \right\rangle\right)$. Moreover, for any $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_C)$, we have $\max_{c \in [C]} \mathbf{z}_c \leq \sum_{c \in [C]} \mathbf{z}_c \leq C \max_{c \in [C]} \mathbf{z}_c$, we can focus on the max element of the output of F's first layer. A sufficient condition of the inequality (2) holds is:

$$\begin{aligned} 0 &< \max_{c \in [C]} \phi\left(\left\langle \mathbf{w}_c(T), \mathbf{u}^k \right\rangle\right) \left\|\mathbf{u}^k\right\|_2^2 < C \max_{c \in [C]} \phi\left(\left\langle \mathbf{w}_c(T), \mathbf{u}^k \right\rangle\right) \left\|\mathbf{u}^k\right\|_2^2 \\ &< y^p \max_{c \in [C]} \phi\left(\langle \mathbf{w}_c(T), \mathbf{v} \rangle\right) \|\mathbf{v}\|_2^2. \end{aligned}$$

This result shows that if $\max_c \mathbf{w}_c(T)$ has at least C times larger projection in the directions of $\mathbf{v}$ than the projection in the directions of $\mathbf{u}^k$, the network is successfully attacked by the adversary. Additionally, this result illustrates that it is theoretically possible for the backdoor, which is activated only for data points with trigger patterns from the non-targeted class, to maintain the predicted class of the network for data points with trigger patterns from the targeted class. We then provide theoretical insights into the effectiveness of backdoor learning in the following subsection.

4.2. The Update Rule in Backdoor Learning

We initially illustrate the difference between standard learning and backdoor learning. We take a close look at the update rule of $\mathbf{w}_c$ in these two scenarios. In standard learning, we can rewrite $-\frac{n}{\eta}(\mathbf{w}_c(t+1)-\mathbf{w}_c(t))$ due to Equation (1) as

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{p=1}^{P} y_i\ell'(F(\mathbf{x}_i),y_i)\phi'\left(\langle \mathbf{w}_c(t),\mathbf{x}_i^p\rangle\right)\mathbf{x}_i^p\\
&= \sum_{i\in\mathcal{I}_{cl}^{tr}} \ell'(F(\mathbf{x}_i),y_i)\phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i\rangle\right)\mathbf{u}_i \qquad (3)\\
&+ \sum_{i\in\mathcal{I}_{cl}^{tr}} y_i\ell'(F(\mathbf{x}_i),y_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i\rangle\right)\boldsymbol{\xi}_i, \qquad (4)\\
&+ \sum_{i\in\mathcal{I}_{cl}^{tr}}\sum_{p\in\mathcal{P}_i^{\zeta}} y_i\ell'(F(\mathbf{x}_i),y_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p\rangle\right)\boldsymbol{\zeta}_i^p. \qquad (5)
\end{aligned}$$

The update of $\mathbf{w}_c$ can be decomposed into two parts: along with the directions of the feature vector and noise vectors. Since the distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are both isotropic Gaussian, Equation (4) and Equation (5) represent a linear combination of vectors pointing in different random directions, in contrast, Equation (3) represents an average across K orthonormal directions. Intuitively, since feature vectors $\left\{\mathbf{u}^k\right\}_{k\in[K]}$ are orthogonal, if K is finite, and n is large enough, all feature vectors can be captured by the model. In the case of backdoor learning, we can rewrite $-\frac{n}{\eta}(\mathbf{w}_c(t+1)-\mathbf{w}_c(t))$ due to Equation (1) as

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{p=1}^{P} \hat{y}_i\ell'(F(\hat{\mathbf{x}}_i),\hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t),\hat{\mathbf{x}}_i^p\rangle\right)\hat{\mathbf{x}}_i^p\\
&= \sum_{i\in\mathcal{I}_{po}^{tr}} y_i\hat{y}_i\ell'(F(\hat{\mathbf{x}}_i),\hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i\rangle\right)\mathbf{u}_i\\
&+ \sum_{i\in\mathcal{I}_b} \hat{y}_i\ell'(F(\hat{\mathbf{x}}_i),\hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \mathbf{v}\rangle\right)\mathbf{v}\\
&+ \sum_{i\in\mathcal{I}_{po}^{tr}} \hat{y}_i\ell'(F(\hat{\mathbf{x}}_i),\hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i\rangle\right)\boldsymbol{\xi}_i,\\
&+ \sum_{i\in\mathcal{I}_{po}^{tr}}\sum_{p\in\mathcal{P}_i^{\zeta}} \hat{y}_i\ell'(F(\hat{\mathbf{x}}_i),\hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p\rangle\right)\boldsymbol{\zeta}_i^p.
\end{aligned}$$

Recall that $\mathcal{S}_b$ is the collection only contains poisoned data, there are two main differences compared with standard learning. Firstly, The update of $\mathbf{w}_c$ can be decomposed into three parts. Except along with the directions of the feature vectors and noise vectors, in backdoor learning, there exists an extra direction, which is the direction of the trigger vector from n_{po} poisoned data. To ensure that the NN can capture the trigger vector and feature vectors simultaneously, the inner product $\langle \mathbf{u}, \mathbf{v}\rangle$ should be bounded. For example, $\left\{\mathbf{u}^k\right\}_{k\in[K]} \cup \{\mathbf{v}\}$ are orthogonal, then the inner product is 0. To guarantee that the NN can effectively capture the trigger vector, n_{po} can not be too small, i.e., $n_{po} \geq \Omega(1)$. Secondly, since the dirty-label attack flips the labels of poisoned data points, the update along the directions of the feature vector can be decomposed into two components: one aligned with $\mathbf{u}_i$ and the other with $-\mathbf{u}_i$. This implies that the poisoned data exhibits harmful effects on the learning of the feature vectors. Even worse, $\mathcal{S}_{po}^{tr}$ contains less clean data than $\mathcal{S}_{cl}^{tr}$, which may also hurt the learning of the feature vectors. Inspired by this, n_{po} can not be too large to safely neglect the harmful impact of poisoned data, for example, $n_{po} \leq o(n)$. The adversary only adds a small partition of poisoned data practically to avoid these two influences.

4.3. The Dynamics of Backdoor Learning

In this subsection, we study the dynamics of backdoor learning. We firstly show a lemma about the norms of $\langle \mathbf{w}_c, hatx^p\rangle$ for different patch at the initialization:

Lemma 4.1. *Given the weights of network* $\mathbf{w}_c$ *initialized as* $\mathbf{w}_c(0) \sim \mathcal{N}(0,\sigma_0)$. *For any* $k \in [K]$, *with a probability of* $1 - \frac{2KC}{d} - \frac{2K}{e^{C/4}}$, *we have:*

$$\|\mathbf{v}\|_2\,\sigma_0/2 \leq \max_{c\in[C]} \left|\langle \mathbf{w}_c(0), \mathbf{v}\rangle\right| \leq \sqrt{\log d}\,\|\mathbf{v}\|_2\,\sigma_0,$$

$$\|\mathbf{u}\|_2\,\sigma_0/2 \leq \max_{c\in[C]} \left|\left\langle \mathbf{w}_c(0), \mathbf{u}^k\right\rangle\right| \leq \sqrt{\log d}\,\|\mathbf{u}\|_2\,\sigma_0.$$

The proof of Lemma 4.1 can be found in Appendix A. Lemma 4.1 implies that $\max_{c\in[C]}|\langle \mathbf{w}_c(0),\mathbf{v}\rangle| = \widetilde{\Theta}(\|\mathbf{v}\|_2\,\sigma_0)$ and $\max_{c\in[C]}|\langle \mathbf{w}_c(0),\mathbf{u}\rangle| = \widetilde{\Theta}(\|\mathbf{u}\|_2\,\sigma_0)$. We assume that $\sigma_\xi = \sigma_\zeta = 0$ to avoid the effect of noises. At the beginning of the training process, the weights of F are closer to the initialization, and we have

$$\begin{aligned}
\frac{d\langle \mathbf{w}_c, \mathbf{u}^k\rangle}{dt} =& -\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i\ell_i'\phi'(\langle \mathbf{w}_c, \hat{\mathbf{x}}^p\rangle)\left\langle \hat{\mathbf{x}}^p, \mathbf{u}^k\right\rangle\\
=& -\frac{1}{n}\sum_{i\in\mathcal{I}_{u^k}\setminus\mathcal{I}_b} y_i y_i\ell_i'\phi'(\langle \mathbf{w}_c, y_i\mathbf{u}_i\rangle)\,\|\mathbf{u}_i\|_2^2 \qquad (6)\\
&+\frac{1}{n}\sum_{i\in\mathcal{I}_{u^k}\cap\mathcal{I}_b} y_i y_i\ell_i'\phi'(\langle \mathbf{w}_c, y_i\mathbf{u}_i\rangle)\,\|\mathbf{u}_i\|_2^2 \qquad (7)\\
&+\frac{1}{n}\sum_{i\in\mathcal{I}_b} y_i\ell_i'\phi'(\langle \mathbf{w}_c, \mathbf{v}\rangle)\left\langle \mathbf{v}, \mathbf{u}^k\right\rangle, \qquad (8)
\end{aligned}$$

where $\mathcal{S}_{u^k}$ denotes the set of training data with feature vector $\mathbf{u}^k$ or $-\mathbf{u}^k$. At the initialization, $F(\hat{\mathbf{x}}_i) = o(1)$, and $\ell_i' = \ell'(y_i F(\hat{\mathbf{x}}_i)) \approx -1/2$. Then we have (6) $\approx (n - n_{po})K^{-1}n^{-1}\sigma_0^2 \left\|\mathbf{u}^k\right\|_2^4$, (7) $\approx\ -n_{po}K^{-1}n^{-1}\sigma_0^2 \left\|\mathbf{u}^k\right\|_2^4$ and (8) $\approx -n_{po}K^{-1}n^{-1}\sigma_0^2 \left\|\mathbf{v}\right\|_2^2 \left\langle \mathbf{v}, \mathbf{u}^k \right\rangle$. If $n_{po} \ll n$, and $\left\langle \mathbf{v}, \mathbf{u}^k \right\rangle < (n - n_{po}) \left\|\mathbf{u}^k\right\|_2^4 / n_{po} \left\|\mathbf{v}\right\|_2^2$, then (7) and (8) can be ignored. Then the dynamic reduces to an ODE. Ignoring the constants, let $g(t) = \left\langle \mathbf{w}, \mathbf{u}^k \right\rangle$, we have $g'(t) \approx (n - n_{po})K^{-1}n^{-1}\phi'(g(t))$. When $g(t) \leq 1$, we have $(g(t)^{-1})' = (n - n_{po})K^{-1}n^{-1} \left\|\mathbf{u}^k\right\|_2^2$ due to the definition of ϕ. Then at $T_u = \frac{nK}{(n - n_{po})\|\mathbf{u}^k\|_2^2 g(0)} = \frac{nK}{(n - n_{po})(\sigma_0 \|\mathbf{u}^k\|_2^3)}$, we yield $\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle \geq \Omega(1)$, which implies that $\mathbf{u}^k$ has been captured by the NN. Moreover, since the feature vectors $\left\{\mathbf{u}^k\right\}_{k\in[K]}$ are orthogonal with the same norm, and appear in the data points uniformly, the NN can fit all feature vectors at T_u. Furthermore, for the trigger vector $\mathbf{v}$, we have

$$
\begin{aligned}
\frac{d\left\langle \mathbf{w}_c, \mathbf{v} \right\rangle}{dt} &= -\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \ell_i' \phi'(\langle \mathbf{w}_c, \hat{\mathbf{x}}^p \rangle) \left\langle \hat{\mathbf{x}}^p, \mathbf{v} \right\rangle \\
&= -\frac{1}{n}\sum_{i\in\mathcal{I}_b} \hat{y}_i \ell_i' \phi'(\langle \mathbf{w}_c, \mathbf{v} \rangle) \left\|\mathbf{v}\right\|_2^2 \qquad (9) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i y_i \ell_i' \phi'(\langle \mathbf{w}_c, \mathbf{u}_i \rangle) \left\langle \mathbf{v}, \mathbf{u}_i \right\rangle. \qquad (10)
\end{aligned}
$$

Similarly, we have (9) $\approx n_{po}n^{-1}\sigma_0^2 \left\|\mathbf{v}\right\|_2^4$, and $|$(10)$| \approx \sigma_0^2 \left\|\mathbf{u}\right\|_2^2 \left|\left\langle \mathbf{v}, \mathbf{u}^k \right\rangle\right|$. If $n_{po} \left\|\mathbf{v}\right\|_2^4 \gg n \left\|\mathbf{u}\right\|_2^2 \left|\langle \mathbf{v}, \mathbf{u} \rangle\right|$, then (10) can be ignored. Then the dynamics reduces to an ODE. Let $g(t) = \langle \mathbf{w}, \mathbf{v} \rangle$, we have $g'(t) \approx n_{po}n^{-1} \left\|\mathbf{v}\right\|_2^2 \phi'(g(t))$. When $g(t) \leq 1$, we have $(g^{-1}(t))' = n_{po}n^{-1}$ due to the definition of ϕ. Then at $T_v = \frac{n}{n_{po}\|\mathbf{v}\|_2^2 g(0)} = \frac{n}{n_{po}\sigma_0\|\mathbf{v}\|_2^3}$, we yield $\langle \mathbf{w}_c, y^p\mathbf{v} \rangle \geq \Omega(1)$, which implies that $\mathbf{v}$ has been captured by the NN. Consequently, if $\max_{k\in[K]} \left|\left\langle \mathbf{v}, \mathbf{u}^k \right\rangle\right| \leq \min\left\{(n - n_{po}) \left\|\mathbf{u}\right\|_2^4 / n_{po} \left\|\mathbf{v}\right\|_2^2, n_{po} \left\|\mathbf{v}\right\|_2^4 / n \left\|\mathbf{u}\right\|_2^2\right\}$ and $n_{po} \ll n$, the feature vectors and trigger vector are both captured by the backdoored NN. A special case of the first condition is that $\left\{\mathbf{u}^k\right\}_{k\in[K]} \cup \{\mathbf{v}\}$ are orthogonal. Although n_{po} is required to be small, the order of n_{po} should have a lower bound, since n_{po} affects T_v. If n_{po} is too small, then T_v is too large. Furthermore, if $\frac{\|\mathbf{v}\|_2^3}{\|\mathbf{u}\|_2^3} \gg \frac{n}{n_{po}K}$, which implies that $T_v \ll T_u$, and the network firstly fits the trigger vector and then fits the feature vector. After that, $\langle \mathbf{w}, \mathbf{v} \rangle$ increases continuously, and keeps greater than $\left\langle \mathbf{w}, \mathbf{u}^k \right\rangle$ at least until time T_u.

The trigger pattern is similar to the spurious feature, but they are essentially different. Firstly, the label of poisoned data points is flipped by the adversary in the dirty-label attack, while the data with spurious features has the correct label. Secondly, the number of poisoned data is limited, which is to avoid being detected by the user and the spurious features may appear in all data points. Finally, in a dirty-label attack, the trigger vector is only added to the data points from the non-targeted class, which is user-specific, while the data with spurious features is not user-specific. Shen et al. (2022) show that if the spurious feature vector appears predominantly in one class, the network can overfit the spurious feature, and use the spurious feature to classify the data points. This result does not conflict with our analysis, Moreover, we emphasise that after training the model with only a small fraction of the poisoned data, the attacker can successfully manipulate the outputs of the NN with the trigger vectors.

4.4. The Effectiveness of Dirty-Label Attack

The dirty-label attack, compared with clean-label attack, only requires a small partition of poisoned data, can efficiently injure the trigger into NN. The analysis in Section 4.3 shows that it requires a large norm of trigger vector, and NN can firstly fit the trigger vector and then fit the feature vector. However, when $\max_c \langle \mathbf{w}_c, \mathbf{u} \rangle$ and $\max_c \langle \mathbf{w}_c, y^p\mathbf{v} \rangle$ both achieves the order of $\Omega(1)$, it is challenging to show which vector primarily influences the outputs of NN. We call this stage as the late stage. In this subsection, we show another effectiveness from dirty-label attack, which guarantee that the trigger vector still primarily influences the outputs of NN. We still assume that $\sigma_\xi = \sigma_\zeta = 0$ to simplify the problem. In the late stage, $-\ell'(\hat{y}_i F(\mathbf{x}_i)) \leq O(e^{-\hat{y}_i F(\mathbf{x}_i)}), \forall i \notin \mathcal{I}_b$. The updates of $\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle > 0$ when

$$
\frac{(n - n_{po})e^{-\sum_c \phi(\langle \mathbf{w}_c, \mathbf{u}^k \rangle)}}{n_{po}\left(1 + e^{-\sum_c \phi(\langle \mathbf{w}_c, \mathbf{u}^k \rangle) + \sum_c \phi(\langle \mathbf{w}_c, y^p\mathbf{v} \rangle)}\right)^{-1}} > \Omega(1). \quad (11)
$$

Suppose $\sum_c \phi(\langle \mathbf{w}_c, y^p\mathbf{v} \rangle) > \Omega\left(\sum_c \phi(\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle)\right)$, we yield $\sum_c \phi(\langle \mathbf{w}_c, y^p\mathbf{v} \rangle) - \sum_c \phi(\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle) > \Omega\left(\frac{1}{\log(n)} \sum_c \phi(\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle)\right)$. Otherwise, if $\sum_c \phi(\langle \mathbf{w}_c, y^p\mathbf{v} \rangle) < O\left(\sum_c \phi(\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle)\right)$, we have $\sum_c \phi(\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle) < O(\log n) \leq \widetilde{O}(1)$, which shows the effectiveness of dirty-label attack. Additionally, the effect will be more significant if $n_{po} = \Theta(n)$. As a result, $\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle$ can not achieves a higher order than $\langle \mathbf{w}_c, y^p\mathbf{v} \rangle$ when $\langle \mathbf{w}_c, y^p\mathbf{v} \rangle \geq \Omega(1)$, i.e., $\sum_c \phi(\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle) \leq \widetilde{O}(\sum_c \phi(\langle \mathbf{w}_c, y^p\mathbf{v} \rangle))$. Moreover, it can be proved that $\langle \mathbf{w}_c, y^p\mathbf{v} \rangle$ still primarily influences the outputs of NN during the learning process. When $\frac{\|\mathbf{v}\|_2^3}{\|\mathbf{u}\|_2^3} \gg \frac{n}{n_{po}K}$, although $\langle \mathbf{w}_c, y^p\mathbf{v} \rangle$ and $\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle$ achieves the same order, the update of $\langle \mathbf{w}_c, y^p\mathbf{v} \rangle$ is still larger than $\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle$. The formal result is shown in Lemma 6.6.

5. Main Results

We show our main results in this section. The proofs of Theorems 5.3 and 5.4 can be found in Appendix D. Our results depend on the following conditions about parameters.

Condition 5.1. We suppose the following conditions hold:

1. The vectors in $\{\mathbf{u}^k\}_{k\in[K]} \cup \{\mathbf{v}\}$ are mutually orthogonal in pairs. $\forall k \in [K]$, $\|\mathbf{u}^k\|_2 = 1$.
2. The number of channels C is as the order of logarithm of d, i.e., $C = \Theta(\log d)$.
3. The network is over-parameterized, and n, P, K satisfies $nP^2K \le o(\sqrt{d})$.
4. The standard deviation σ_0 satisfies $\sigma_0 \le o(1)$, and σ_ξ satisfies $\omega(1) \le \sigma_\xi \le o\left(\frac{1}{\sigma_0}\right)$.
5. The size of training set is larger than the number of useful features, i.e., $n \ge \Omega\left(\sigma_0^{-3}K\right)$. Moreover, in backdoor learning, the training set only contains a small proportion of poisoned data points, i.e., $\Omega(1) \le n_{po} \le o(n)$. The norm of trigger vectors satisfies $\Omega(1) \le \|\mathbf{v}\|_2 \le O(\sigma_\xi)$.

Remark 5.2. In Condition 5.1, we ignore the effect from the angle between feature vectors $\{\mathbf{u}^k\}_{k\in[K]}$ and trigger vector $\mathbf{v}$. We assume vectors in $\{\mathbf{u}^k\}_{k\in[K]} \cup \{\mathbf{v}\}$ are mutually orthogonal in pairs, and this condition is possible to be replaced by a weaker condition that the inner product of feature vectors $\{\mathbf{u}^k\}_{k\in[K]}$ and trigger vector $\mathbf{v}$ are bounded. We suppose that C, n increases with d in different orders. Since $C = \Theta(\log d)$, and n has a lower order than $\sqrt{d}$, this problem is studied in an over-parameterized case. We suppose the upper and lower bounds for the variances σ_ξ^2 and σ_ζ^2, then the output of the network cannot be easily dominated by the feature vectors at initialization. We also assume $n \ge \Omega\left(\sigma_0^{-3}K\right)$ such that each feature vector can be successfully captured by NN.

Given a clean training set $\mathcal{S}_{cl}^{tr}$, the following theorem shows that under mild conditions, there exists round T_u such that the well-trained NN achieves high clean accuracy and low attack success rate.

Theorem 5.3. *[standard learning] Under the Condition 5.1, given a clean training set $\mathcal{S}_{cl}^{tr}$ with size n, there exists $T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$ such that for $T_1 \ge T_u$, the network $\hat{F}_{T_1}$ fits all clean data points with a high probability:*

$$\mathbb{P}(\forall i \in [n], y_i\hat{F}_{T_1}(\mathbf{x}_i) \geq \widetilde{\Omega}(1)) \geq 1 - O\left(\frac{n^2P^2KC}{poly(d)}\right). \quad (12)$$

Moreover, $\hat{F}_{T_1}$ achieves a high clean accuracy but leaves a low attack success rate at T_1:

$$Acc(\hat{F}_{T_1}; \mathcal{D}_{\mathbf{z}}) \geq 1 - O\left(\frac{nP^2KC}{poly(d)}\right), \quad (13)$$

$$ASR(\hat{F}_{T_1}; \mathcal{D}_{\mathbf{z}}, \mathfrak{P}) \leq O\left(\frac{nP^2KC}{poly(d)}\right). \quad (14)$$

Theorem 5.3 shows that under the data model defined in Definition 3.1, NN can achieve a high clean accuracy after T_u rounds of standard training. Since $\mathbf{v}$ has a smaller norm than σ_ξ at the initialization, NN can not capture the trigger vector since poisoned data is not included in the training set, and the feature vectors finally achieve a larger signal than the trigger vector $\mathbf{v}$ and other noise vectors. When the poisoned data is added to the training set, we have the following result:

Theorem 5.4. *[Backdoor Learning] Under the Condition 5.1, given a poisoned training set $\mathcal{S}_{po}^{tr}$ with size n, if $n_{po}\|\mathbf{v}\|_2^2 > \omega(nK^{-1})$, there exists $T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$ such that for $T_2 \ge T_u$ the network $\hat{F}_{T_2}$ fits both clean and poisoned training data points with a high probability:*

$$\mathbb{P}(\forall i \in [n], \hat{y}_i\hat{F}_{T_2}(\hat{\mathbf{x}}_i) \geq \widetilde{\Omega}(1)) \geq 1 - O\left(\frac{n^2P^2KC}{poly(d)}\right). \quad (15)$$

Furthermore, there exists $T_v = \widetilde{\Theta}\left(\frac{n}{\eta n_{po}\|\mathbf{v}\|_2^3\sigma_0}\right)$ such that $\hat{F}$ achieves high attack success rate at $T_2' \ge T_v$ and achieves high clean accuracy at $T_2 \ge T_u > T_v$:

$$Acc(\hat{F}_{T_2}; \mathcal{D}_{\mathbf{z}}) \geq 1 - O\left(\frac{nP^2KC}{poly(d)}\right), \quad (16)$$

$$ASR(\hat{F}_{T_2'}; \mathcal{D}_{\mathbf{z}}, \mathfrak{P}) \geq 1 - O\left(\frac{nP^2KC}{poly(d)}\right). \quad (17)$$

In Theorem 5.4, we show the model, well-trained on a poisoned training set, can simultaneously achieve high clean accuracy and attack success rate. Moreover, note that T_u in Theorem 5.3 and in Theorem 5.4 have the same order, which implies that the poisoned data has limited effects on the learning of the feature vectors for the model. The results also indicate that the success of the backdoor attack is influenced by the number of the feature vectors K, the poisoning rate n_{po}/n, and the norm ratio $\|\mathbf{v}\|_2 / \|\mathbf{u}\|_2$.

6. Analysis of Standard and Backdoor Learning

In this section, we show some key techniques used in our proofs for the main results. We additionally denote $\langle \mathbf{w}_c(t+1), \mathbf{u}^k\rangle - \langle \mathbf{w}_c(t), \mathbf{u}^k\rangle$ and $\langle \mathbf{w}_c(t+1), y^p\mathbf{v}\rangle - \langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle$ as $\Delta_c^t(\mathbf{u}^k)$ and $\Delta_c^t(\mathbf{v})$, respectively. Consider the early stage of the standard learning, the updates of $\langle \mathbf{w}_c, \boldsymbol{\xi}\rangle$ and $\langle \mathbf{w}_c, \boldsymbol{\zeta}\rangle$ are both small, while for any k, $\langle \mathbf{w}_c, \mathbf{u}^k\rangle$ has a significant update. The following lemma shows that $\langle \mathbf{w}_c, \mathbf{u}^k\rangle$ is increasing:

Lemma 6.1. *Under the Condition 5.1. In both standard and backdoor learning, suppose there exists t such that $\forall k \in [K]$, $\langle \mathbf{w}_c(t), \mathbf{u}^k\rangle \le O(C^{-1})$, $|\langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle| \le \widetilde{O}(\sigma_0\sigma_\xi)$ and $|\langle \mathbf{w}_c(t), \boldsymbol{\zeta}\rangle| \le \widetilde{O}(\sigma_0\sigma_\zeta)$ for some $0 \le t \le T$. We then* *yield*

$$\forall k \in [K], \Delta_c^t(\mathbf{u}^k) \geq \widetilde{\Omega}\left(\frac{\eta \|\mathbf{u}\|_2^2 \phi'\left(\left|\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle\right|\right)}{K + Ke^{C^{-2}}}\right) \tag{18}$$

is increasing. Furthermore, since $-\ell' \leq 1$, we have

$$\forall k \in [K], \Delta_c^t(\mathbf{u}^k) \leq \widetilde{O}\left(\eta K^{-1} \|\mathbf{u}\|_2^2 \, \phi'\left(\left|\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle\right|\right)\right). \tag{19}$$

As the model does not fit the noise vectors, the feature vectors primarily influence the outputs of the model. $\max_c \left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle$ keeps increasing until $\max_c \left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle$ reaches the order of $\widetilde{\Omega}(1)$. After that, $\phi'(\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle)$ is also of the order of $\widetilde{\Omega}(1)$, and the update of $\max_c \left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle$ is small as shown in the following lemma:

Lemma 6.2. *Under the Condition 5.1. In standard learning, suppose there exists $0 \leq t \leq T$ such that $\forall k \in [K]$, $\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle \geq \widetilde{\Omega}(1)$), we have*

$$\forall k \in [K], \Delta_c^t(\mathbf{u}^k) \leq \widetilde{O}\left(\eta K^{-1} \|\mathbf{u}\|_2^2 e^{-\max_c \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle}\right). \tag{20}$$

Remark 6.3. In the early stage of the learning process, for each patch $\hat{\mathbf{x}}^p$, the inner product $\langle \mathbf{w}_c, \hat{\mathbf{x}}^p \rangle$ is small, which implies that $-\ell' = \Theta(1)$. The increment of $\langle \mathbf{w}_c, \hat{\mathbf{x}}^p \rangle$ strongly depends on $\phi'(\langle \mathbf{w}_c, \hat{\mathbf{x}}^p \rangle)$. In the late stage, $F(\mathbf{x})$ achieves the order of $\widetilde{\Omega}(1)$, which means $-\ell' \leq \widetilde{O}(1)$. $\max_c \left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle$ has a relatively large increment compared to $\max_c \langle \mathbf{w}_c, \hat{\mathbf{x}}^p \rangle$ for $p \neq p_u$, since $\max_c \phi'(\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle) \geq \widetilde{\Omega}(1)$.

We continue to analyze the process of backdoor learning. The update of $\left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle$ can be divided into two groups since a small part of data points has the flipped label. As we mentioned in Section 4.2, $n_{po} \leq o(n)$ implies that the increment from groups of poisoned data points, which have the flipped label, can be ignored in the early stage, and Lemma 6.1 holds in backdoor learning. Moreover, the trigger vector, which is orthogonal to the feature vectors, can be captured by NN in backdoor learning as $\Delta_c^t(\mathbf{v})$ is larger than $\Delta_c^t(\mathbf{u}^k)$.

Lemma 6.4. *Under the Condition 5.1. In backdoor learning, suppose $\langle \mathbf{w}_c(t), \mathbf{v} \rangle \leq O(C^{-1/3})$, $|\langle \mathbf{w}_c(t), \boldsymbol{\xi} \rangle| \leq \widetilde{O}\left(\sigma_0 \sigma_\xi\right)$ and $|\langle \mathbf{w}_c(t), \boldsymbol{\zeta} \rangle| \leq \widetilde{O}\left(\sigma_0 \sigma_\zeta\right)$ for some $0 \leq t \leq T$, we have*

$$\Delta_c^t(\mathbf{v}) = \widetilde{\Theta}\left(n_{po} n^{-1} \eta \|\mathbf{v}\|_2^2 \, \phi'\left(|\langle \mathbf{w}_c(t), \mathbf{v} \rangle|\right)\right) \tag{21}$$

is increasing.

Remark 6.5. To analyze the learning process of feature vectors, we divide the feature vectors into K groups. We can regard the learning of feature and trigger vectors as a race. The number of data points containing $\mathbf{u}^k$ is n/K while the number of poisoned data points is n_{po}. When K increases, the adversary can use $\mathbf{v}$ with a small norm to successfully attack the model.

Note that $\Delta_c^t(\mathbf{v})$ has a higher order than $\Delta_c^t(\mathbf{u})$, which implies that $\Delta_c^t(\mathbf{v})$ can first achieve the order of $\widetilde{\Omega}(1)$. In addition, in the early stage of the process, we yield $\langle \mathbf{w}_c(0), y^p \mathbf{v} \rangle / \left\langle \mathbf{w}_c(0), \mathbf{u}^k \right\rangle = \widetilde{\Theta}(\|\mathbf{v}\|_2 / \|\mathbf{u}\|_2)$ from Lemma 4.1. The ratio of updates $\Delta_c^t(\mathbf{v})/\Delta_c^t(\mathbf{u}^k) \geq \widetilde{\Omega}\left(\frac{n_{po} K \|\mathbf{v}\|_2^2 \phi'(|\langle \mathbf{w}_c(t), y^p \mathbf{v} \rangle|)}{n \|\mathbf{u}\|_2^2 \phi'(|\langle \mathbf{w}_c(t), \mathbf{u}^k \rangle|)}\right) \geq \widetilde{\Omega}\left(\|\mathbf{v}\|_2^2\right)$. In the early stage of learning process, we have $\langle \mathbf{w}_c(t), y^p \mathbf{v} \rangle / \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle \geq \widetilde{\Omega}\left(\|\mathbf{v}\|_2^2\right)$, i.e., $\mathbf{v}$ primarily influences the outputs of the model.

After that, when $\max_c \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle$ and $\max_c \langle \mathbf{w}_c(t), y^p \mathbf{v} \rangle$ both achieve the order of $\widetilde{\Omega}(1)$, we yield $-\ell' \leq \widetilde{O}(1)$. The neglect effect of the groups of poisoned data may not be ignored, which causes that the learning process of $\mathbf{u}$ in backdoor learning is different with in standard learning. Note that $\max_c \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle$ increases only if $\sum_{i \in \mathcal{I}_{u^k} \setminus \mathcal{I}_b} \ell_i' \geq \Omega\left(\sum_{i \in \mathcal{I}_{u^k} \cap \mathcal{I}_b} \ell_i'\right)$, and $\max_c \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle$ can not achieve a higher order than $\max_c \langle \mathbf{w}_c(t), y^p \mathbf{v} \rangle$. $\mathbf{v}$ continues to primarily influence the outputs of the model in the late stage.

Lemma 6.6. *Under the Condition 5.1, in backdoor learning, suppose there exists $0 \leq t \leq T$ such that $\forall k \in [K], \max_{c \in [C]} \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle \geq \widetilde{\Omega}(1)$ and $\max_{c \in [C]} \langle \mathbf{w}_c(t), y^p \mathbf{v} \rangle \geq \widetilde{\Omega}(1)$, we have*

$$\max_{c \in [C]} \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle \leq \widetilde{O}\left(\max_{c \in [C]} \langle \mathbf{w}_c(t), y^p \mathbf{v} \rangle\right). \tag{22}$$

Furthermore, the trigger vector $\mathbf{v}$ primarily influence the outputs of NN:

$$\forall k \in [K], \sum_{c \in [C]} \phi(\langle \mathbf{w}_c(t), y^p \mathbf{v} \rangle) - \phi(\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle) \geq \widetilde{\Omega}(1). \tag{23}$$

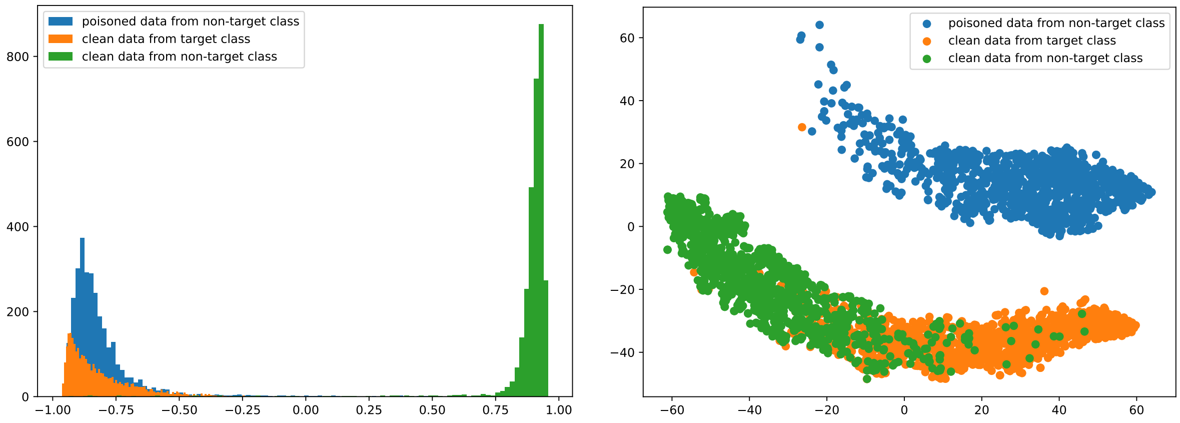
Remark 6.7. Lemma 6.6 indicates the relationship between $\langle \mathbf{w}_c(t), y^p \mathbf{v} \rangle$ and $\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle$. It also shows that even if $\max_c \langle \mathbf{w}_c(t), y^p \mathbf{v} \rangle$ has the same order as $\max_c \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle$, the outputs of NN are manipulated by the trigger vector $\mathbf{v}$. It is challenging to study the order of $\langle \mathbf{w}_c(t), y^p \mathbf{v} \rangle - \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle$ when both feature and trigger vectors are captured by NN.

Finally, $\Delta_c^t(\mathbf{u}^k)$ and $\Delta_c^t(\mathbf{v})$ are both upper-bounded in the late stage of backdoor learning, as shown in the following lemma.

Lemma 6.8. *Under the Condition 5.1. Suppose there exists $0 \leq t \leq T$ such that $\forall k \in [K]$, $\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle \geq \widetilde{\Omega}(1)$) and $\langle \mathbf{w}_c(t), y^p \mathbf{v} \rangle \geq \widetilde{\Omega}(1)$, we have*

$$\forall k \in [K], \Delta_c^t(\mathbf{u}^k) \leq \widetilde{O}\left(K^{-1} \eta \|\mathbf{u}\|_2^2 \, e^{-\max_c \left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle}\right), \tag{24}$$

$$\Delta_c^t(\mathbf{v}) \leq \widetilde{O}\left(n_{po} n^{-1} \eta \|\mathbf{v}\|_2^2 \, e^{-\max_c \langle \mathbf{w}_c, y^p \mathbf{v} \rangle}\right). \tag{25}$$



(a) Cosine Similarity

(b) T-SNE

Figure 1. Results about the representation vectors on CIFAR-10 under the BadNets attack. (a) The cosine similarities of the representation vectors and the top singular vector. (b) The T-SNE plot of representation vectors. The representation vectors are centered by the average representation vector.

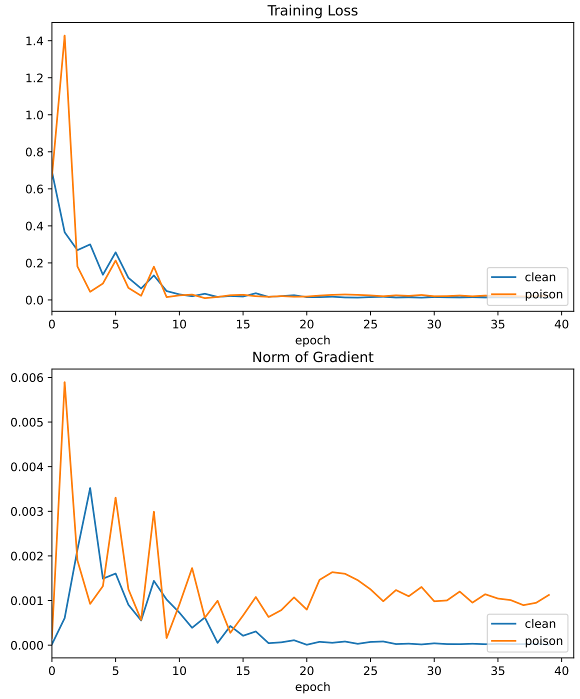


Figure 2. Top: loss of clean and poisoned data. Bottom: the norms of the gradients of weights w.r.t. the loss of clean and poisoned data. We evaluate the result for the initialized model at epoch 0.

7. Experiments

In the experiments, we empirically study the BPA. We use two dirty-label backdoor attacks: BadNets (Gu et al., 2017) and four-corner attack (Turner et al., 2019) on two real-world datasets, MNIST and CIFAR-10 (Krizhevsky et al., 2009). The details can be found in Appendix E.

7.1. Empirical Study about Poisoned Data

Theoretical results led us to expect that the loss of poisoned data decreases fast, coupled with a large norm of the gradients in the early stages of training, and empirical validation supports these expectations. In Figure 2, we use the BadNets attack to generate the poisoned training set from CIFAR-10 with a poisoning rate of 0.05, and the results show that the loss of poisoned data decreases faster than the loss of clean data in the first 5 epochs. Furthermore, we record the norm of gradients of weights $\nabla_{\mathbf{w}}\bar{\ell}^{po} = \frac{1}{|\mathcal{S}_{po}|}\sum_{i\in\mathcal{I}_{po}} \nabla_{\mathbf{w}}\ell(F(\hat{\mathbf{x}}_i, y_i))$ and $\nabla_{\mathbf{w}}\bar{\ell}^{cl} = \frac{1}{|\mathcal{S}_{cl}|}\sum_{i\in\mathcal{I}_{cl}} \nabla_{\mathbf{w}}\ell(F(\hat{\mathbf{x}}_i, y_i))$ with poisoned and clean test set. The norm $\nabla_{\mathbf{w}}\bar{\ell}^{po}$ maintains a larger norm of gradients than $\nabla_{\mathbf{w}}\bar{\ell}^{cl}$ in the first 5 epochs, as shown in Figure 2.

Finally, we use SVD decomposition to analyze the representations of clean and poisoned data. For a l-layer NN, the layers from the first layer to the $l-1$-th layer of the NN are utilized as a feature extractor. Tran et al. (2018) study the spectral signatures for poisoned data with representations and find that the matrix of poisoned representations has a larger spectral norm. Apart from their analysis, we study the cosine similarities of the maximum singular vector and representation vectors. We collect the representations of both clean and poisoned test data as a matrix and obtain the maximum singular vector. The cosine similarities of the maximum singular vector and representation vectors are shown in the results in Figure 1(a). The results show that most of the representation vectors of poisoned data align with the negative direction of the maximum singular vector, and most of the representation vectors of clean data from the targeted class have the same direction with the maximum singular vector. Besides, the representation vectors of clean

Table 1. The effects from the size of the training set and poisoning rate in MNIST. We evaluate the accuracy and attack success rate at the last epoch. We use **Bold** to denote the results with ASR $> 95\%$, which means the attacker successfully embeds the backdoor in the model.

MNIST		Poisoning rate									
Size		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
2000	ACC	99.25	99.25	99.25	99.31	99.31	99.31	99.31	99.20	99.20	99.09
	ASR	0.78	0.78	0.78	0.78	0.78	1.12	1.35	3.03	11.32	68.95
	Time	–	–	–	–	–	–	–	–	–	–
4000	ACC	99.41	99.47	99.41	99.41	99.57	99.57	99.47	99.47	99.52	99.52
	ASR	0.78	1.01	85.76	**96.3**	**99.22**	**99.22**	**99.55**	**99.78**	**99.66**	**99.66**
	Time	–	–	–	28	26	23	23	19	19	19
6000	ACC	99.68	99.63	99.57	99.57	99.57	99.63	99.68	99.68	99.68	99.68
	ASR	0.34	84.3	**95.52**	**98.32**	**98.77**	**98.88**	**99.55**	**99.66**	**99.55**	**99.66**
	Time	–	–	64	24	24	21	20	17	17	13
8000	ACC	99.73	99.73	99.68	99.79	99.79	99.79	99.79	99.79	99.79	99.73
	ASR	6.95	**98.88**	**99.66**	**99.78**	**99.78**	**99.89**	**99.89**	**99.89**	**99.89**	**99.89**
	Time	–	25	19	19	14	13	12	11	11	11
10000	ACC	99.73	99.79	99.84	99.84	99.84	99.84	99.79	99.84	99.84	99.84
	ASR	79.15	**99.33**	**99.44**	**99.78**	**99.89**	**99.89**	**99.89**	**99.89**	**99.89**	**99.89**
	Time	–	18	15	14	11	10	10	9	9	9

data from the targeted and non-targeted classes are in opposite directions. It is important to note that the representation vectors of clean data from the targeted class and poisoned data are not clustered together. To illustrate this, the T-SNE is used to show the relationship of poisoned and clean data. and the results are shown in Figure 1(b). To summarize, the changes caused by the trigger pattern are two aspects: value and direction. Our theoretical results also consider these two aspects.

7.2. Key Components for Backdoor Attacks

Furthermore, we delve into the key components of BPA algorithms. We firstly adjust the size of the training set and the poisoning rate within the range of 2000 to 10000 and 0.01 to 0.1, respectively. We keep other hyper-parameters unchanged and show the results in Table 1. Table 1 indicate that with an increase in the size of the training set, the lowest poisoning rate required for a successful attack decreases. This suggests that as the training set size grows, less poisoned data is needed, validating our condition regarding n_{po}. Additionally, given a fixed size of the training set, as the poisoning rate increases, the accuracy remains a slight change, implying that the negative impact of poisoned data can be negligible. We also study the time $T^\star$ such that for any $t > T$, the attack success rate is always greater than 95%, and the results show that the $T^\star$ decreases as the poisoning rate increases. We also study the effectiveness from the norm of the trigger vector, and the results can be found in Appendix E. Additionally, the experimental results of the four-corner attack can be also found in Appendix E.

8. Conclusion

To comprehend the effectiveness of BPA, we conduct a theoretical and empirical analysis in this paper. We provide theoretical insights on backdoor learning and further show theoretical results in a two-layer convolutional neural network with the multi-view model. Empirically, we investigate the curve of training loss and norms of gradients w.r.t. loss for both clean and poisoned data. Moreover, we study the components affecting the lowest poisoning rate to succeed in BPA, and the experimental results support our theoretical findings.

Impact Statement

The primary goal of our study is to comprehensively understand the effectiveness of BPA instead of proposing a novel BPA algorithm. A potential negative impact of our research is that malicious attackers could design BPA algorithms with the theoretical analysis in this paper. However, our work can also help users to understand poisoned data points and detect them. Our study emphasizes the importance of enhancing the security of deep learning models as well.

Acknowledgments

This work is supported by the National Key R&D Program of China under Grant 2023YFC3604702, and the Fundamental Research Fund Program of LIESMARS.

References

Allen-Zhu, Z. and Li, Y. Feature purification: How adversarial training performs robust deep learning. In *FOCS*, pp. 977–988, 2021.

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *ICLR*, 2023.

Barni, M., Kallas, K., and Tondi, B. A new backdoor attack in CNNS by training set corruption without label poisoning. In *ICIP*, pp. 101–105, 2019.

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

Cinà, A. E., Grosse, K., Vascon, S., Demontis, A., Biggio, B., Roli, F., and Pelillo, M. Backdoor learning curves: Explaining backdoor poisoning beyond influence functions. *CoRR*, abs/2106.07214, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

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

Jelassi, S. and Li, Y. Towards understanding how momentum improves generalization in deep learning. In *ICML*, volume 162, pp. 9965–10040, 2022.

Jha, R. D., Hayase, J., and Oh, S. Label poisoning is all you need. In *NeurIPS*, 2023.

Koh, P. W., Steinhardt, J., and Liang, P. Stronger data poisoning attacks break data sanitization defenses. *Mach. Learn.*, 111(1):1–47, 2022.

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

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.

Li, B. and Liu, W. WAT: improve the worst-class robustness in adversarial training. In *AAAI*, pp. 14982–14990, 2023.

Ma, X., Wang, Z., and Liu, W. On the tradeoff between robustness and fairness. In *NeurIPS*, 2022.

Manoj, N. and Blum, A. Excess capacity and backdoor poisoning. In *NeurIPS*, pp. 20373–20384, 2021.

Nguyen, D. T., Nguyen, T. M., Tran, A. T., Doan, K. D., and WONG, K. S. IBA: Towards irreversible backdoor attacks in federated learning. In *NeurIPS*, 2023.

Nguyen, T. A. and Tran, A. T. Wanet - imperceptible warping-based backdoor attack. In *ICLR*, 2021.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *NeurIPS*, pp. 5019–5031, 2018.

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., and Goldstein, T. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *NeurIPS*, pp. 6106–6116, 2018.

Shen, R., Bubeck, S., and Gunasekar, S. Data augmentation as feature manipulation. In *ICML*, volume 162, pp. 19773–19808, 2022.

Tran, B., Li, J., and Madry, A. Spectral signatures in backdoor attacks. In *NeurIPS*, pp. 8011–8021, 2018.

Turner, A., Tsipras, D., and Madry, A. Clean-label backdoor attacks, 2019. URL https://openreview.net/forum?id=HJg6e2CcK7.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wang, G., Xian, X., Srinivasa, J., Kundu, A., Bi, X., Hong, M., and Ding, J. Demystifying poisoning backdoor attacks from a statistical perspective. *CoRR*, abs/2310.10780, 2023a.

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., and Yan, S. Better diffusion models further improve adversarial training. In *ICML*, volume 202, pp. 36246–36263, 2023b.

Wen, Z. and Li, Y. Toward understanding the feature learning process of self-supervised contrastive learning. In *ICML*, 2021.

Xian, X., Wang, G., Srinivasa, J., Kundu, A., Bi, X., Hong, M., and Ding, J. Understanding backdoor attacks through the adaptability hypothesis. In *ICML*, volume 202, pp. 37952–37976, 2023.

Xu, J. and Liu, W. On robust multiclass learnability. In *NeurIPS*, 2022.

Yang, Y., Steinhardt, J., and Hu, W. Are neurons actually collapsed? on the fine-grained structure in neural representations. In *ICML*, volume 202, pp. 39453–39487, 2023.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *ICML*, volume 97, pp. 7472–7482, 2019.

Zhou, Z. and Liu, W. Sample complexity for distributionally robust learning under chi-square divergence. *Journal of Machine Learning Research*, 24:230:1–230:27, 2023.

Zou, X. and Liu, W. Generalization bounds for adversarial contrastive learning. *Journal of Machine Learning Research*, 24:114:1–114:54, 2023.

A. Theoretical Results at Initialization

Initialization is the same in both standard and backdoor learning. We first show the important lemmas, which are useful in our proof.

Lemma A.1 (Lemma 1 in Laurent & Massart (2000))**.** *Suppose $X_i \ldots, X_n$ are n i.i.d. Gaussian random variables with mean 0 and variance 1. Let $a_1, \ldots, a_n$ be non-negative. We set*

$$|a|_\infty = \sup_{i=1,\ldots,n} |a_i|\,,\ |a|_2^2 = \sum_{i=1}^{n} a_i^2. \tag{26}$$

Let

$$Z = \sum_{i=1}^{n} a_i(X^2 - 1). \tag{27}$$

Then, the following inequalities hold for any positive t:

$$\mathbb{P}\left[Z \geq 2\,|a|_2\,\sqrt{t} + 2\,|a|_\infty\,t\right] \leq \exp(-t). \tag{28}$$

$$\mathbb{P}\left[Z \leq -2\,|a|_2\,\sqrt{t}\right] \leq \exp(-t). \tag{29}$$

Lemma A.2 (Lemma 4 in Shen et al. (2022))**.** *Consider independently sampled Gaussian vectors $\mathbf{z}_1 \sim \mathcal{N}(0, \sigma_1^2 \mathbf{I}_d)$ and $\mathbf{z}_2 \sim \mathcal{N}(0, \sigma_2^2 \mathbf{I}_d)$. For any $\delta \in (0, 1)$ and a large enough d, there exists constants c_1, c_2 such that*

$$\mathbb{P}\left[|\langle \mathbf{z}_1, \mathbf{z}_2 \rangle| \leq c_1 \sigma_1 \sigma_2 \sqrt{d \log(2/\delta)}\right] \geq 1 - \delta, \tag{30}$$

$$\mathbb{P}\left[\langle \mathbf{z}_1, \mathbf{z}_2 \rangle \geq c_2 \sigma_1 \sigma_2 \sqrt{d}\right] \geq 1/4. \tag{31}$$

Lemma A.3 (Proposition 2.5 in Wainwright (2019))**.** *Suppose that the variables $X_i, i = 1, \ldots, n$, are independent, and X_i has mean μ_i and sub-Gaussian parameter σ_i. Then for all $r \geq 0$, we have*

$$\mathbb{P}\left[\sum_{i=1}^{n} (X_i - \mu_i) \geq t\right] \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^{n} \sigma_i^2}\right\} \tag{32}$$

Proposition A.4. *Given a standard Gaussian variable $Z \sim \mathcal{N}(0, 1)$, then we have $\mathbb{P}\left[Z \geq 1/2\right] \geq 1/4$.*

A.1. Inner Product of Different Patches

With the lemmas shown in Appendix A, we can individually analyze the inner product of different components. As the feature vectors $\left\{\mathbf{u}^k\right\}_{k\in[K]} \cup \{\mathbf{v}\}$ are orthogonal in our assumption, we have $\forall i, j \in [K], i \neq j, \left\langle \mathbf{u}^i, \mathbf{u}^j \right\rangle = 0$ and $\left\langle \mathbf{u}^i, \mathbf{v} \right\rangle = 0$. We then analyze the inner product of ξ and $\mathbf{v}$.

Lemma A.5. *Given $\mathcal{S}_{cl} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ are i.i.d. drawn from the distribution $\mathcal{D}_\mathbf{z}$ defined in Definition 3.1, and a trigger vector $\mathbf{v}$, with a probability of $1 - \frac{2n(P-1)}{d}$, we have*

$$\forall i \in [n], -\sqrt{\log d/d}\,\|\mathbf{v}\|_2\,\sigma_\xi \leq \langle \boldsymbol{\xi}_i, \mathbf{v} \rangle \leq \sqrt{\log d/d}\,\|\mathbf{v}\|_2\,\sigma_\xi. \tag{33}$$

$$\forall i \in [n], p \in \mathcal{P}_i^\zeta, -\sqrt{\log d/d}\,\|\mathbf{v}\|_2\,\sigma_\zeta \leq \langle \boldsymbol{\zeta}_i, \mathbf{v} \rangle \leq \sqrt{\log d/d}\,\|\mathbf{v}\|_2\,\sigma_\zeta. \tag{34}$$

Proof. Since the distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are spherically symmetric, and we have $\forall i, \langle \boldsymbol{\xi}_i, \mathbf{v} \rangle \sim \mathcal{N}(0, \|\mathbf{v}\|_2\,\sigma_\xi/\sqrt{d})$. Due to Lemma A.3, given $i \in [n]$, we have

$$\mathbb{P}\left[|\langle \boldsymbol{\xi}_i, \mathbf{v} \rangle| \geq \sqrt{\log d/d}\,\|\mathbf{v}\|_2\,\sigma_\xi\right] \leq \frac{2}{d}. \tag{35}$$

Equation (34) can be immediately obtained by using the union bound. Similarly, given i, p, we have

$$\mathbb{P}\left[|\langle \boldsymbol{\zeta}_i^p, \mathbf{v} \rangle| \geq \sqrt{\log d/d}\,\|\mathbf{v}\|_2\,\sigma_\zeta\right] \leq \frac{2}{d}. \tag{36}$$

By using the union bound, we conclude our proof. □

The Lemma A.5 shows that $|\langle \boldsymbol{\xi}_i, \mathbf{v}\rangle| \le \widetilde{O}(\sigma_\xi \left\|\mathbf{v}\right\|_2/\sqrt{d})$ and $|\langle \boldsymbol{\zeta}_i^p, \mathbf{v}\rangle| \le \widetilde{O}(\sigma_\zeta \left\|\mathbf{v}\right\|_2/\sqrt{d})$. Similarly, $|\langle \boldsymbol{\xi}_i, \mathbf{u}_j\rangle|$ and $|\langle \boldsymbol{\zeta}_i^p, \mathbf{u}_j\rangle|$ can be also bounded as shown in the following lemma.

Lemma A.6. *Given $\mathcal{S}_{cl} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ are i.i.d. drawn from the distribution $\mathcal{D}_\mathbf{z}$ defined in Definition 3.1, with a probability of $1 - \frac{2nK(P-1)}{d}$, we have*

$$\forall i \in [n], k \in [K] - \sqrt{\log d/d}\left\|\mathbf{u}^k\right\|_2 \sigma_\xi \le \left\langle \boldsymbol{\xi}_i, \mathbf{u}^k\right\rangle \le \sqrt{\log d/d}\left\|\mathbf{u}^k\right\|_2 \sigma_\xi. \tag{37}$$

$$\forall i \in [n], p \in \mathcal{P}_i^\zeta, k \in [K] - \sqrt{\log d/d}\left\|\mathbf{u}^k\right\|_2 \sigma_\zeta \le \left\langle \boldsymbol{\zeta}_i, \mathbf{u}^k\right\rangle \le \sqrt{\log d/d}\left\|\mathbf{u}^k\right\|_2 \sigma_\zeta. \tag{38}$$

Proof. Since the distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are spherically symmetric, and we have $\forall i, \left\langle \boldsymbol{\xi}_i, \mathbf{u}^k\right\rangle \sim \mathcal{N}(0, \left\|\mathbf{u}^k\right\|_2 \sigma_\xi/\sqrt{d})$. Due to Lemma A.3, given $i \in [n], k \in [K]$, we have

$$\mathbb{P}\left[\left|\left\langle \boldsymbol{\xi}_i, \mathbf{u}^k\right\rangle\right| \ge \sqrt{\log d/d}\left\|\mathbf{u}^k\right\|_2 \sigma_\xi\right] \le \frac{2}{d}. \tag{39}$$

Equation (38) can be immediately obtained by using the union bound. Similarly, given i, p, k, we have

$$\mathbb{P}\left[\left|\left\langle \boldsymbol{\zeta}_i^p, \mathbf{u}^k\right\rangle\right| \ge \sqrt{\log d/d}\left\|\mathbf{u}^k\right\|_2 \sigma_\zeta\right] \le \frac{2}{d}. \tag{40}$$

By using the union bound, we conclude our proof. □

We then analyze the inner product of two noise vectors.

Lemma A.7. *Given $\mathcal{S}_{cl} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ are i.i.d. drawn from the distribution $\mathcal{D}_\mathbf{z}$ defined in Definition 3.1, with a probability of $1 - \frac{2n^2(P-1)^2 + 2n^2P}{d}$, we have*

$$\forall i, i' \in [n], i \ne i', |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle| \le a_1 \sigma_\xi^2 \sqrt{\log(2d)/d}, \tag{41}$$

$$\forall i \in [n], \sigma_\xi^2(1 + 2\sqrt{\log d}) \le \|\boldsymbol{\xi}_i\|_2 \le \sigma_\xi^2(1 + 2\sqrt{\log d} + 2\log d). \tag{42}$$

$$\forall i, i' \in [n], p, p' \in \mathcal{P}_i^\zeta, i \ne i' \textit{ or } p \ne p', \left|\left\langle \boldsymbol{\zeta}_i^p, \boldsymbol{\zeta}_{i'}^{p'}\right\rangle\right| \le a_1 \sigma_\zeta^2 \sqrt{\log(2d)/d}, \tag{43}$$

$$\forall i \in [n], p \in \mathcal{P}_i^\zeta, \sigma_\zeta^2(1 + 2\sqrt{\log d}) \le \|\boldsymbol{\zeta}_i^p\|_2 \le \sigma_\zeta^2(1 + 2\sqrt{\log d} + 2\log d). \tag{44}$$

$$\forall i, i' \in [n], p \in \mathcal{P}_i^\zeta, |\langle \boldsymbol{\xi}_i, \boldsymbol{\zeta}_{i'}^p\rangle| \le a_1 \sigma_\xi \sigma_\zeta \sqrt{\log(2d)/d}, \tag{45}$$

Proof. These results are the immediate result by using union bound and Lemmas A.1 to A.3. Specifically, $\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle$, $\left\langle \boldsymbol{\zeta}_i^p, \boldsymbol{\zeta}_{i'}^{p'}\right\rangle$ and $\langle \boldsymbol{\xi}_i, \boldsymbol{\zeta}_{i'}^p\rangle$ are all gaussian variables, and Lemma A.2 shows that for any $i, i' \in [n]$, there exists a constant a_1 such that with a probability of $1 - \frac{2}{d}$, we have

$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle| \le a_1 \sigma_\xi^2 \sqrt{\log d/d}. \tag{46}$$

Equation (41) can be obtained by using the union bound. Similarly, with a probability of $1 - \frac{2n(P-2)}{d}$, we have

$$\forall i, i' \in [n], p, p' \in \mathcal{P}_i^\zeta, i \ne i' \text{ or } p \ne p', \left|\left\langle \boldsymbol{\zeta}_i^p, \boldsymbol{\zeta}_{i'}^{p'}\right\rangle\right| \le \sigma_\zeta^2 \sqrt{\log(2d)/d}. \tag{47}$$

Furthermore, with a probability of $1 - \frac{2n(P-2)}{d}$, we have

$$\forall i, i' \in [n], p \in \mathcal{P}_i^\zeta, |\langle \boldsymbol{\xi}_i, \boldsymbol{\zeta}_{i'}^p\rangle| \le a_1 \sigma_\xi \sigma_\zeta \sqrt{\log(2d)/d} \tag{48}$$

Since $\|\boldsymbol{\xi}_i\|_2^2$ and $\|\boldsymbol{\zeta}_i^p\|_2^2$ are both the sum of squares of d i.i.d. Gaussian variable, $\frac{\|\boldsymbol{\xi}_i\|_2^2}{\sigma_\xi^2}$ and $\frac{\left\|\boldsymbol{\zeta}_i^p\right\|_2^2}{\sigma_\zeta^2}$ are both χ^2 random variables. given $i \in [n], p \in \mathcal{P}_i^\zeta$, Lemma A.1 implies that

$$\mathbb{P}\left[\frac{\|\boldsymbol{\xi}_i\|_2^2}{\sigma_\xi^2} - 1 \ge 2\sqrt{\log d} + 2\log d\right] \le \frac{1}{d}, \mathbb{P}\left[\frac{\|\boldsymbol{\xi}_i\|_2^2}{\sigma_\xi^2} - 1 \le 2\sqrt{\log d}\right] \le \frac{1}{d}. \tag{49}$$

$$\mathbb{P}\left[\frac{\|\boldsymbol{\zeta}_i^p\|_2^2}{\sigma_\zeta^2} - 1 \ge 2\sqrt{\log d} + 2\log d\right] \le \frac{1}{d}, \mathbb{P}\left[\frac{\|\boldsymbol{\zeta}_i^p\|_2^2}{\sigma_\zeta^2} - 1 \le 2\sqrt{\log d}\right] \le \frac{1}{d}. \tag{50}$$

By using the union bound, we conclude our proof. □

A.2. Inner Products of Weight Vectors and Patch Vectors

Since $\mathbf{w}$ is initialized as $\mathbf{w}_c(0) \sim \mathcal{N}(0, \sigma_0)$, The analysis of the inner product of weight vector and patch vector is similar to the proofs of Appendix A.1.

Lemma A.8. *Given the weights of network $\mathbf{w}_c$, which is initialized as $\mathbf{w}_c(0) \sim \mathcal{N}(0, \sigma_0)$, $\mathcal{S}_{cl} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ are i.i.d. drawn from the distribution $\mathcal{D}_{\mathbf{z}}$ defined in Definition 3.1. For a trigger vector $\mathbf{v}$, with a probability of $1 - \frac{2KC+2nC+2C}{d} - \frac{2K+2n+2}{e^{C/4}}$, we have for any $k \in [K]$,*

$$\max_{c\in[C]} |\langle \mathbf{w}_c(0), \mathbf{v}\rangle| \le \sqrt{\log d}\, \|\mathbf{v}\|_2\, \sigma_0,\ \max_{c\in[C]} \langle \mathbf{w}_c(0), \mathbf{v}\rangle \ge \|\mathbf{v}\|_2\, \sigma_0/2. \tag{51}$$

$$\forall k \in [K], \max_{c\in[C]} \left|\left\langle \mathbf{w}_c(0), \mathbf{u}^k\right\rangle\right| \le \sqrt{\log d}\, \|\mathbf{u}\|_2\, \sigma_0,\ \max_{c\in[C]} \langle \mathbf{w}_c(0), \mathbf{u}_i\rangle \ge \|\mathbf{u}\|_2\, \sigma_0/2, \tag{52}$$

$$\forall i \in [n], \max_{c,k} |\langle \mathbf{w}_c(0), \boldsymbol{\xi}_i\rangle| \le a_1\sigma_0\sigma_\xi\sqrt{\log(2d)},\ \max_{c\in[C]} \langle \mathbf{w}_c(0), \boldsymbol{\xi}_i\rangle \ge a_2\sigma_0\sigma_\xi. \tag{53}$$

$$\forall i \in [n], p \in \mathcal{P}_i^\zeta, \max_{c\in[C]} |\langle \mathbf{w}_c(0), \boldsymbol{\zeta}_i^p\rangle| \le a_1\sigma_0\sigma_\xi\sqrt{\log(2d)},\ \max_{c\in[C]} \langle \mathbf{w}_c(0), \boldsymbol{\zeta}_i^p\rangle \ge a_2\sigma_0\sigma_\xi. \tag{54}$$

Proof. Since the distribution of $\mathbf{w}_c(0)$ is spherically symmetric, for any i, we have $\langle \mathbf{w}_c(0), \mathbf{u}_i\rangle \sim \mathcal{N}(0, \sigma_0)$. Due to Lemma A.3, we have

$$\mathbb{P}\left[\exists k, c, \left|\left\langle \mathbf{w}_c(0), \mathbf{u}^k\right\rangle\right| \ge \sqrt{\log d}\, \|\mathbf{u}\|_2\, \sigma_0\right] \le \sum_{c=1}^{C}\sum_{k=1}^{K} \mathbb{P}\left[\left|\left\langle \mathbf{w}_c(0), \mathbf{u}^k\right\rangle\right| \ge \sqrt{\log d}\, \|\mathbf{u}\|_2\, \sigma_0\right] \le \frac{2KC}{d}. \tag{55}$$

Meanwhile, Proposition A.4 implies that for any k,

$$\mathbb{P}\left[\max_{c\in[C]} \left\langle \mathbf{w}_c(0), \mathbf{u}^k\right\rangle \le \frac{\|\mathbf{u}\|_2\, \sigma_0}{2}\right] \le \prod_{c=1}^{C} \mathbb{P}\left[\|\mathbf{u}\|_2 \left\langle \mathbf{w}_c(0), \mathbf{u}^k\right\rangle \le \frac{\sigma_0}{2}\right] \le 2\left(\frac{3}{4}\right)^C \le 2K\mathbf{e}^{-C/4}. \tag{56}$$

Similarly, $\langle \mathbf{w}_c(0), \mathbf{v}\rangle \sim \mathcal{N}(0, \|\mathbf{v}\|_2\, \sigma_0)$, and we have

$$\mathbb{P}\left[\exists c, |\langle \mathbf{w}_c(0), \mathbf{v}\rangle| \ge \sqrt{\log d}\, \|\mathbf{v}\|_2\, \sigma_0\right] \le \sum_{c=1}^{C} \mathbb{P}\left[|\langle \mathbf{w}_c(0), \mathbf{v}\rangle| \ge \sqrt{\log d}\, \|\mathbf{v}\|_2\, \sigma_0\right] \le \frac{2C}{d}, \tag{57}$$

$$\mathbb{P}\left[\max_{c\in[C]} \langle \mathbf{w}_c(0), \mathbf{v}\rangle \le \frac{\|\mathbf{v}\|_2\, \sigma_0}{2}\right] \le \prod_{c=1}^{C} \mathbb{P}\left[\langle \mathbf{w}_c(0), \mathbf{v}\rangle \le \frac{\|\mathbf{v}\|_2\, \sigma_0}{2}\right] \le 2r\left(\frac{3}{4}\right)^C \le 2\mathbf{e}^{-C/4}. \tag{58}$$

Recall that for any i, $\boldsymbol{\xi}_i \sim \mathcal{N}(0, \sigma_\xi/\sqrt{d}\mathbf{I}_d)$, with combining the Union bound and Lemma A.2, there exists constants a_1 such that with a probability of 1- $\frac{2nC}{d}$, we have

$$\max_{i\in[n], c\in[C]} |\langle \mathbf{w}_c(0), \boldsymbol{\xi}_i\rangle| \le a_1\sigma_0\sigma_\xi\sqrt{\log(2d)}, \tag{59}$$

For any i, there exists a constant a_2 such that

$$\mathbb{P}\left[\max_{c\in[C]} \langle \mathbf{w}_c(0), \boldsymbol{\xi}_i\rangle \le a_2\sigma_0\sigma_\xi\right] \le \prod_{c=1}^{C} \mathbb{P}\left[\langle \mathbf{w}_c(0), \boldsymbol{\xi}_i\rangle \le a_2\sigma_0\sigma_\xi\right] \le 2\left(\frac{3}{4}\right)^C \le 2\mathbf{e}^{-C/4}. \tag{60}$$

Using the Union bound, we have

$$\mathbb{P}\left[\min_{i\in[n]}\max_{c\in[C]} \langle \mathbf{w}_c(0), \boldsymbol{\xi}_i\rangle \ge a_2\sigma_0\sigma_\xi\right] = 1 - \mathbb{P}\left[\exists i, \max_{c\in[C]} \langle \mathbf{w}_c(0), \boldsymbol{\xi}_i\rangle \le a_2\sigma_0\sigma_\xi\right] \ge 1 - 2n\mathbf{e}^{-C/4}. \tag{61}$$

Similarly, for any $i \in [n], p \in \mathcal{P}_i^\zeta$, $\boldsymbol{\zeta}_i^p \sim \mathcal{N}(0, \sigma_\zeta/\sqrt{d}\mathbf{I}_d)$, and with a probability of 1- $\frac{2n(P-2)C}{d}$, we have

$$\max_{i\in[n], p\in\mathcal{P}_i^\zeta, c\in[C]} |\langle \mathbf{w}_c(0), \boldsymbol{\zeta}_i^p\rangle| \le a_1\sigma_0\sigma_\zeta\sqrt{\log(2d)}, \tag{62}$$

We further have

$$\mathbb{P}\left[\min_{i\in[n],p\in\mathcal{P}_i^{\zeta}}\max_{c\in[C]}\langle\mathbf{w}_c(0),\boldsymbol{\zeta}_i^p\rangle\geq a_2\sigma_0\sigma_\zeta\right]=1-\mathbb{P}\left[\exists i\in[n],p\in\mathcal{P}_i^{\zeta},\max_{c\in[C]}\langle\mathbf{w}_c(0),\boldsymbol{\zeta}_i^p\rangle\leq a_2\sigma_0\sigma_\zeta\right] \tag{63}$$

$$\geq 1-2n(P-2)\mathrm{e}^{-C/4}. \tag{64}$$

We conclude our proof. □

B. Standard and Backdoor Learning in Early Stage

Our techniques used in this section are inspired by (Shen et al., 2022). Note that standard learning is a special case that $n_{po}=0$. In this section, we first focus on backdoor learning, and then extend our results to standard learning. In backdoor learning, given a learning rate η, the parameters are optimized as

$$\begin{aligned}
\mathbf{w}_c(t+1)=\mathbf{w}_c(t)&-\frac{\eta}{n}\sum_{i=1}^{n}\nabla\ell(F(\hat{\mathbf{x}}_i),\hat{y}_i)\\
=\mathbf{w}_c(t)&-\frac{\eta}{n}\sum_{i=1}^{n}\sum_{p=1}^{P}\hat{y}_i\ell'(F(\hat{\mathbf{x}}_i),\hat{y}_i)\phi'\left(\langle\mathbf{w}_c(t),\mathbf{x}_i^p\rangle\right)\mathbf{x}_i^p\\
=\mathbf{w}_c(t)&-\frac{\eta}{n}\sum_{i\in\mathcal{I}_{po}^{tr}}\ell'(F(\mathbf{x}_i),\hat{y}_i)\phi'\left(\langle\mathbf{w}_c(t),\hat{y}_i\mathbf{u}_i\rangle\right)\mathbf{u}_i-\frac{\eta}{n}\sum_{i\in\mathcal{I}_{po}^{tr}}\hat{y}_i\ell'(F(\mathbf{x}_i),\hat{y}_i)\phi'\left(\langle\mathbf{w}_c(t),\mathbf{v}\rangle\right)\mathbf{v}\\
&-\frac{\eta}{n}\sum_{i\in\mathcal{I}_{po}^{tr}}\hat{y}_i\ell'(F(\mathbf{x}_i),\hat{y}_i)\phi'\left(\langle\mathbf{w}_c(t),\boldsymbol{\xi}_i\rangle\right)\boldsymbol{\xi}_i-\frac{\eta}{n}\sum_{i\in\mathcal{I}_{po}^{tr}}\sum_{p\in\mathcal{P}_i^{\zeta}}\hat{y}_i\ell'(F(\mathbf{x}_i),\hat{y}_i)\phi'\left(\langle\mathbf{w}_c(t),\boldsymbol{\zeta}_i^p\rangle\right)\boldsymbol{\zeta}_i^p \qquad (65)\\
=\mathbf{w}_c(t)&-\frac{\eta}{n}\sum_{i\notin\mathcal{I}_b}\ell'(F(\hat{\mathbf{x}}_i),y_i)\phi'\left(\langle\mathbf{w}_c(t),y_i\mathbf{u}_i\rangle\right)\mathbf{u}_i+\frac{\eta}{n}\sum_{i\in\mathcal{I}_b}\ell'(F(\mathbf{x}_i),-y_i)\phi'\left(\langle\mathbf{w}_c(t),y_i\mathbf{u}_i\rangle\right)\mathbf{u}_i\\
&-\frac{\eta}{n}\sum_{i\notin\mathcal{I}_b}y_i\ell'(F(\hat{\mathbf{x}}_i),y_i)\phi'\left(\langle\mathbf{w}_c(t),\boldsymbol{\xi}_i\rangle\right)\boldsymbol{\xi}_i+\frac{\eta}{n}\sum_{i\in\mathcal{I}_b}y_i\ell'(F(\hat{\mathbf{x}}_i),-y_i)\phi'\left(\langle\mathbf{w}_c(t),\boldsymbol{\xi}_i\rangle\right)\boldsymbol{\xi}_i\\
&-\frac{\eta}{n}\sum_{i\notin\mathcal{I}_b}\sum_{p\in\mathcal{P}_i^{\zeta}}y_i\ell'(F(\hat{\mathbf{x}}_i),y_i)\phi'\left(\langle\mathbf{w}_c(t),\boldsymbol{\zeta}_i^p\rangle\right)\boldsymbol{\zeta}_i^p+\frac{\eta}{n}\sum_{i\in\mathcal{I}_b}\sum_{p\in\mathcal{P}_i^{\zeta}}y_i\ell'(F(\mathbf{x}_i),-y_i)\phi'\left(\langle\mathbf{w}_c(t),\boldsymbol{\zeta}_i^p\rangle\right)\boldsymbol{\zeta}_i^p
\end{aligned}$$

$$\tag{66}$$

Equation (65) due to the decomposition of $\mathbf{x}$, and Equation (66) due to the fact that $\forall i\in\mathcal{I}_b,\hat{y}_i=-y_i$.

These two versions of update rules Equations (65) and (66) have individual advantages in our analysis. For standard learning, the update rules can be immediately obtained by setting $\mathcal{S}_b=\emptyset$, and $\forall i,\hat{y}_i=y_i$. The parameters are optimized in standard learning as

$$\begin{aligned}
\mathbf{w}_c(t+1)=\mathbf{w}_c(t)&-\frac{\eta}{n}\sum_{i=1}^{n}\nabla\ell(F(\hat{\mathbf{x}}_i),\hat{y}_i)=\mathbf{w}_c(t)-\frac{\eta}{n}\sum_{i=1}^{n}\sum_{p=1}^{P}\hat{y}_i\ell'(F(\hat{\mathbf{x}}_i),\hat{y}_i)\phi'\left(\langle\mathbf{w}_c(t),\mathbf{x}_i^p\rangle\right)\mathbf{x}_i^p\\
=\mathbf{w}_c(t)&-\frac{\eta}{n}\sum_{i\in\mathcal{I}_{cl}^{tr}}\ell'(F(\mathbf{x}_i),\hat{y}_i)\phi'\left(\langle\mathbf{w}_c(t),\hat{y}_i\mathbf{u}_i\rangle\right)\mathbf{u}_i-\frac{\eta}{n}\sum_{i\in\mathcal{I}_{cl}^{tr}}\hat{y}_i\ell'(F(\mathbf{x}_i),\hat{y}_i)\phi'\left(\langle\mathbf{w}_c(t),\boldsymbol{\xi}_i\rangle\right)\boldsymbol{\xi}_i\\
&-\frac{\eta}{n}\sum_{i\in\mathcal{I}_{cl}^{tr}}\sum_{p\in\mathcal{P}_i^{\zeta}}\hat{y}_i\ell'(F(\mathbf{x}_i),\hat{y}_i)\phi'\left(\langle\mathbf{w}_c(t),\boldsymbol{\zeta}_i^p\rangle\right)\boldsymbol{\zeta}_i^p
\end{aligned}\tag{67}$$

Next, due to Condition 5.1, we have the following results as a corollary of Condition 5.1:

Condition B.1. Under the Condition 5.1, we further have the following results:

1. Since $n>\Omega(\sigma_0^{-3}K)$, $\sigma_\xi\leq o(\frac{1}{\sigma_0})$ and $nP^2K\leq o(\sqrt{d})$, we immediately have $K\sigma_\xi^3\leq o(n)$ and $K\sigma_\xi^3\leq o(\sqrt{d})$.

2. Since $K\sigma_\xi^3 \leq o(\sqrt{d})$, if $n_{po} \left\|\mathbf{v}\right\|_2^3 > w(nK^{-1})$, we have $n_{po} \geq w\left(\frac{n\sigma_\xi^3}{\|\mathbf{v}\|_2^3\sqrt{d}}\right)$.

3. $n_{po} \left\|\mathbf{v}\right\|_2^2 > w(nK^{-1})$ and $\left\|\mathbf{v}\right\|_2 \geq \Omega(1)$ imply that $n_{po} \left\|\mathbf{v}\right\|_2^3 > w(nK^{-1})$

4. If $n_{po} \left\|\mathbf{v}\right\|_2^3 > w(nK^{-1})$, since $K \leq K + Ke^{C^{-2}}$, we yield $\left(\frac{n}{\eta n_{po}\|\mathbf{v}\|_2^3\sigma_0}\right) \leq o\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$.

5. Since $n > \Omega(\sigma_0^{-3}K)$ and $\sigma_\xi \leq o(\frac{1}{\sigma_0})$, we yield $K \leq o(n\sigma_0^2\sigma_\xi^{-1})(1+e^{C^{-2}})^{-1}$ which can be rewrite as $\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right) \leq o\left(\frac{n\sigma_0}{\eta\sigma_\xi}\right)$.

6. Since $n_{po} \left\|\mathbf{v}\right\|_2^2 > w(nK^{-1})$, and $K\sigma_\xi^3 \leq o(\sqrt{d})$, we have $n\sigma_\xi^3 \leq o(n_{po}\sqrt{d}\left\|\mathbf{v}\right\|_2^2)$.

These conditions are important in our proofs.

B.1. Theoretical Analysis on $\langle \mathbf{w}_c(t), \boldsymbol{\xi} \rangle$ and $\langle \mathbf{w}_c(t), \boldsymbol{\zeta} \rangle$

In our analysis, we emphasis the effects from feature vectors and trigger patterns, and we assume that the noise vectors are not fitted by NN in the whole process. To show the time T that the noise vectors are barely fitted, we show that the increase of noise vectors is slow in backdoor learning.

Lemma B.2. *Under the Condition 5.1. In both standard and backdoor learning, for* $t \leq o\left(\frac{n\sigma_0}{\eta\sigma_\xi}\right)$ *and* $i' \in [n]$, *we have*

$$\forall i' \in [n], |\langle \mathbf{w}_c(t), \boldsymbol{\xi}_{i'} \rangle| \leq \widetilde{O}\left(\sigma_0\sigma_\xi\right). \tag{68}$$

Proof. Consider that standard learning is a special case, we first show the results for backdoor learning. Due to Equation (65), we can upper bound $|\langle \mathbf{w}_c(t+1), \boldsymbol{\xi}_{i'} \rangle - \langle \mathbf{w}_c(t), \boldsymbol{\xi}_{i'} \rangle|$ as

$$\begin{aligned}
&|\langle \mathbf{w}_c(t+1), \boldsymbol{\xi}_{i'} \rangle - \langle \mathbf{w}_c(t), \boldsymbol{\xi}_{i'} \rangle| \\
\leq&\frac{\eta}{n}\sum_{i\in\mathcal{I}_{po}^{tr}} |\ell'(F(\mathbf{x}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \mathbf{u}_i \rangle\right)\langle \mathbf{u}_i, \boldsymbol{\xi}_{i'} \rangle| && (69)\\
-&\frac{\eta}{n}\sum_{i\in\mathcal{I}_{b}} |\ell'(F(\mathbf{x}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \mathbf{v} \rangle\right)\langle \mathbf{v}, \boldsymbol{\xi}_{i'} \rangle| && (70)\\
-&\frac{\eta}{n}\sum_{i\in\mathcal{I}_{po}^{tr}} |\ell'(F(\mathbf{x}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i \rangle\right)\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| && (71)\\
-&\frac{\eta}{n}\sum_{i\in\mathcal{I}_{po}^{tr}}\sum_{p\in\mathcal{P}_i^{\zeta}} |\ell'(F(\mathbf{x}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p \rangle\right)\langle \boldsymbol{\zeta}_i^p, \boldsymbol{\xi}_{i'} \rangle| && (72)
\end{aligned}$$

Due to $\phi' \leq 1$ and $-\ell' \leq 1$, Lemmas A.5 to A.7 imply that

$$|(69)| \leq \widetilde{O}\left(\frac{\eta\sigma_\xi}{\sqrt{d}}\right), |(70)| \leq \widetilde{O}\left(\frac{\eta n_{po}\left\|\mathbf{v}\right\|_2\sigma_\xi}{n\sqrt{d}}\right), |(71)| \leq \widetilde{O}\left(\frac{\eta\sigma_\xi^2}{\sqrt{d}} + \frac{\eta\sigma_\xi^2}{n}\right), |(72)| \leq \widetilde{O}\left(\frac{\eta P\sigma_\xi\sigma_\zeta}{\sqrt{d}}\right).$$

Since $|\langle \mathbf{w}_c(0), \boldsymbol{\xi}_{i'} \rangle| \leq \widetilde{O}(\sigma_0\sigma_\xi)$, $n \leq o(\sqrt{d})$, $\left\|\mathbf{v}\right\|_2 < O\left(\sigma_\xi\right)$, and $\sigma_\xi = P\sigma_\zeta$, when $T \leq o\left(\frac{n\sigma_0}{\eta\sigma_\xi}\right)$, for $0 \leq t \leq T$, we have

$$|\langle \mathbf{w}_c(t), \boldsymbol{\xi}_{i'} \rangle - \langle \mathbf{w}_c(0), \boldsymbol{\xi}_{i'} \rangle| \leq o\left(\sigma_0\sigma_\xi\right) \tag{73}$$

. As for standard learning. These results also holds for standard learning by setting $\mathcal{S}_b = \emptyset$. Since $|(70)| = 0$ in standard learning, the condition $\left\|\mathbf{v}\right\|_2 < O\left(\sigma_\xi\right)$ can be dropped in standard learning. We conclude our proof. □

The analysis of $\langle \mathbf{w}_c(t), \boldsymbol{\zeta} \rangle$ is similar to Lemma B.2.

Lemma B.3. *Under the Condition 5.1. In both standard and backdoor learning, for any $t \leq o\left(\frac{nP\sigma_0}{\eta\sigma_\zeta}\right)$, $i' \in \mathcal{S}_{po}^{tr}$, $p' \in [P]$, we have*

$$\left|\left\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_{i'}^{p'} \right\rangle\right| \leq \widetilde{O}\left(\sigma_0\sigma_\zeta\right). \tag{74}$$

Proof. We first show the results for backdoor learning. Due to Equation (65), we can upper bound $\left|\left\langle \mathbf{w}_c(t+1), \boldsymbol{\zeta}_{i'}^{p'} \right\rangle - \left\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_{i'}^{p'} \right\rangle\right|$ as

$$\begin{aligned}
&|\langle \mathbf{w}_c(t+1), \boldsymbol{\zeta}_{i'}^{p} \rangle - \langle \mathbf{w}_c(t), \boldsymbol{\zeta}_{i'}^{p} \rangle| \\
\leq &\frac{\eta}{n} \sum_{i \in \mathcal{I}_{po}^{tr}} \sum_{p \in \mathcal{P}_i^\zeta} \left| \ell'(F(\mathbf{x}_i), \hat{y}_i) \phi'\left(\langle \mathbf{w}_c(t), \mathbf{u}_i \rangle\right) \left\langle \mathbf{u}_i, \boldsymbol{\zeta}_{i'}^{p'} \right\rangle \right| && (75) \\
- &\frac{\eta}{n} \sum_{i \in \mathcal{I}_b} \sum_{p \in \mathcal{P}_i^\zeta} \left| \ell'(F(\mathbf{x}_i), \hat{y}_i) \phi'\left(\langle \mathbf{w}_c(t), \mathbf{v} \rangle\right) \left\langle \mathbf{v}, \boldsymbol{\zeta}_{i'}^{p'} \right\rangle \right| && (76) \\
- &\frac{\eta}{n} \sum_{i \in \mathcal{I}_{po}^{tr}} \sum_{p \in \mathcal{P}_i^\zeta} \left| \ell'(F(\mathbf{x}_i), \hat{y}_i) \phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i \rangle\right) \left\langle \boldsymbol{\xi}_i, \boldsymbol{\zeta}_{i'}^{p'} \right\rangle \right| && (77) \\
- &\frac{\eta}{n} \sum_{i \in \mathcal{I}_{po}^{tr}} \sum_{p \in \mathcal{P}_i^\zeta} \left| \ell'(F(\mathbf{x}_i), \hat{y}_i) \phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p \rangle\right) \left\langle \boldsymbol{\zeta}_i^p, \boldsymbol{\zeta}_{i'}^{p'} \right\rangle \right| && (78)
\end{aligned}$$

Due to $\phi' \leq 1$ and $-\ell' \leq 1$, Lemmas A.5 to A.7 imply that

$$|(75)| \leq \widetilde{O}\left(\frac{\eta P \sigma_\zeta}{\sqrt{d}}\right), |(76)| \leq \widetilde{O}\left(\frac{\eta n_{po} P \left\|\mathbf{v}\right\|_2 \sigma_\zeta}{n\sqrt{d}}\right), |(77)| \leq \widetilde{O}\left(\frac{\eta P \sigma_\xi \sigma_\zeta}{\sqrt{d}}\right), |(78)| \leq \widetilde{O}\left(\frac{\eta P \sigma_\zeta^2}{\sqrt{d}} + \frac{\eta\sigma_\zeta^2}{n}\right).$$

Since $|\langle \mathbf{w}_c(0), \boldsymbol{\xi}_{i'} \rangle| \leq \sigma_0\sigma_\xi$, $nP^2 \leq o(\sqrt{d})$, $\left\|\mathbf{v}\right\|_2 < O\left(\sigma_\xi\right)$, $\sigma_\xi = P\sigma_\zeta$, when $T \leq o\left(\frac{n\sigma_0}{\eta\sigma_\zeta}\right)$, for $0 \leq t \leq T$, we have

$$\left|\left\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_{i'}^{p'} \right\rangle - \left\langle \mathbf{w}_c(0), \boldsymbol{\zeta}_{i'}^{p'} \right\rangle\right| \leq o\left(\sigma_0\sigma_\zeta\right). \tag{79}$$

Note that these results also hold for standard learning by setting $\mathcal{S}_b = \emptyset$, and the condition $\left\|\mathbf{v}\right\|_2 < O\left(\sigma_\xi\right)$ can be dropped in standard learning since $|(76)| = 0$. We conclude our proof. □

To summerize, note that $\sigma_\xi = P\sigma_\zeta$, which imples $\frac{n\sigma_0}{\eta\sigma_\zeta} = \frac{nP\sigma_0}{\eta\sigma_\xi}$, and $\frac{nP\sigma_0}{\eta\sigma_\xi} > \frac{n\sigma_0}{\eta\sigma_\xi}$, therefore, for $0 \leq t \leq T \leq o(\frac{n\sigma_0}{\eta\sigma_\xi})$, both $\left|\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}_{i'}^{p'} \right\rangle\right|$ and $\left|\left\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_{i'}^{p'} \right\rangle\right|$ are at the order of $o(1)$ in both standard and backdoor learning.

B.2. Theoretical Analysis on $\langle \mathbf{w}_c(t), \mathbf{u} \rangle$

In backdoor learning, due to the orthogonality of main features and trigger vector, the updates of main features and trigger vector can be analyzed separately if the effects from noise vectors can be ignored, then $\forall k$, $\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle$ is increasing in both standard and backdoor learning.

Lemma 6.1. *Under the Condition 5.1. In both standard and backdoor learning, suppose there exists t such that $\forall k \in [K]$, $\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle \leq O(C^{-1})$, $|\langle \mathbf{w}_c(t), \boldsymbol{\xi} \rangle| \leq \widetilde{O}\left(\sigma_0\sigma_\xi\right)$ and $|\langle \mathbf{w}_c(t), \boldsymbol{\zeta} \rangle| \leq \widetilde{O}\left(\sigma_0\sigma_\zeta\right)$ for some $0 \leq t \leq T$. We then yield*

$$\forall k \in [K], \Delta_c^t(\mathbf{u}^k) \geq \widetilde{\Omega}\left(\frac{\eta \|\mathbf{u}\|_2^2 \phi'\left(\left|\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle\right|\right)}{K + Ke^{C^{-2}}}\right) \tag{18}$$

is increasing. Furthermore, since $-\ell' \leq 1$, we have

$$\forall k \in [K], \Delta_c^t(\mathbf{u}^k) \leq \widetilde{O}\left(\eta K^{-1} \|\mathbf{u}\|_2^2 \, \phi'\left(\left|\langle \mathbf{w}_c(t), \mathbf{u}^k \rangle\right|\right)\right). \tag{19}$$

Proof. In backdoor learning, due to the update rule, we can rewrite $\left\langle \mathbf{w}_c(t+1), \mathbf{u}^k \right\rangle - \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle$ as

$$\begin{aligned}
&\left\langle \mathbf{w}_c(t+1), \mathbf{u}^k \right\rangle - \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle \\
=& -\frac{\eta}{n}\sum_{i=1}^{n} \hat{y}_i y_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i\rangle\right)\left\langle \mathbf{u}_i, \mathbf{u}^k \right\rangle - \frac{\eta}{n}\sum_{i=1}^{n} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i\rangle\right)\left\langle \boldsymbol{\xi}_i, \mathbf{u}^k \right\rangle \\
&- \frac{\eta}{n}\sum_{i=1}^{n}\sum_{p\in\mathcal{P}_i^{\zeta}} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p\rangle\right)\left\langle \boldsymbol{\zeta}_i^p, \mathbf{u}^k \right\rangle + \frac{\eta}{n}\sum_{i\in\mathcal{I}_b} y_i \ell'(F(\mathbf{x}_i), -y_i)\phi'\left(\langle \mathbf{w}_c(t), \mathbf{v}\rangle\right)\left\langle \mathbf{v}, \mathbf{u}^k \right\rangle \\
=& -\frac{\eta}{n}\sum_{i\notin\mathcal{I}_b, \mathbf{u}_i=\mathbf{u}^k} \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i\rangle\right)\left\langle \mathbf{u}_i, \mathbf{u}^k \right\rangle + \frac{\eta}{n}\sum_{i\in\mathcal{I}_b, \mathbf{u}_i=\mathbf{u}^k} \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i\rangle\right)\left\langle \mathbf{u}_i, \mathbf{u}^k \right\rangle \\
&- \frac{\eta}{n}\sum_{i=1}^{n} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i\rangle\right)\left\langle \boldsymbol{\xi}_i, \mathbf{u}^k \right\rangle - \frac{\eta}{n}\sum_{i=1}^{n}\sum_{p\in\mathcal{P}_i^{\zeta}} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p\rangle\right)\left\langle \boldsymbol{\zeta}_i^p, \mathbf{u}^k \right\rangle \\
&+ \frac{\eta}{n}\sum_{i\in\mathcal{I}_b} y_i \ell'(F(\mathbf{x}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \mathbf{v}\rangle\right)\left\langle \mathbf{v}, \mathbf{u}^k \right\rangle.
\end{aligned} \tag{80}$$

Equation (80) = 0 due to the orthogonality of feature vectors and the trigger vector. Since Lemma B.2 implies $|\langle \mathbf{w}_c(t), \boldsymbol{\xi}_{i'}\rangle| \le \widetilde{O}(\sigma_0\sigma_\xi)$ and Lemma B.3 implies $|\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_{i'}^p\rangle| \le \widetilde{O}(\sigma_0\sigma_\zeta)$, we have

$$\left|\frac{\eta}{n}\sum_{i=1}^{n} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i\rangle\right)\left\langle \boldsymbol{\xi}_i, \mathbf{u}^k \right\rangle\right| \le \widetilde{O}\left(\eta\sigma_0^2\sigma_\xi^3 d^{-1/2}\left\|\mathbf{u}^k\right\|_2\right),$$

$$\left|\frac{\eta}{n}\sum_{i=1}^{n}\sum_{p\in\mathcal{P}_i^{\zeta}} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p\rangle\right)\left\langle \boldsymbol{\zeta}_i^p, \mathbf{u}^k \right\rangle\right| \le \widetilde{O}\left(\eta P\sigma_0^2\sigma_\zeta^3 d^{-1/2}\left\|\mathbf{u}^k\right\|_2\right).$$

When $\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle \le O(C^{-1})$, we have $\max_i \hat{y}_i F(\hat{\mathbf{x}}_i) \le O(C^{-2})$, and $\min_i -\ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i) \ge \Omega(\frac{1}{1+e^{C^{-2}}})$. We then have

$$-\frac{\eta}{n}\sum_{i\notin\mathcal{I}_b, \mathbf{u}_i=\mathbf{u}^k} \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i\rangle\right)\left\langle \mathbf{u}_i, \mathbf{u}^k \right\rangle \ge \widetilde{\Omega}\left(\frac{1}{K+Ke^{C^{-2}}}\eta\left\|\mathbf{u}\right\|_2^2\phi'\left(\left|\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle\right|\right)\right), \tag{81}$$

$$-\frac{\eta}{n}\sum_{i\notin\mathcal{I}_b, \mathbf{u}_i=\mathbf{u}^k} \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i\rangle\right)\left\langle \mathbf{u}_i, \mathbf{u}^k \right\rangle \le \widetilde{O}\left(K^{-1}\eta\left\|\mathbf{u}\right\|_2^2\phi'\left(\left|\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle\right|\right)\right), \tag{82}$$

and

$$\begin{aligned}
\left|\frac{\eta}{n}\sum_{i\in\mathcal{I}_b, \mathbf{u}_i=\mathbf{u}^k} \phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i\rangle\right)\left\langle \mathbf{u}_i, \mathbf{u}^k \right\rangle\right| \le& \widetilde{O}\left(n_{po}n^{-1}K^{-1}\eta\left\|\mathbf{u}\right\|_2^2\phi'\left(\left|\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle\right|\right)\right) \\
\le& o\left(\frac{1}{K+Ke^{C^{-2}}}\eta\left\|\mathbf{u}\right\|_2^2\phi'\left(\left|\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle\right|\right)\right),
\end{aligned} \tag{83}$$

where Equation (83) due to the condition $n_{po} \le o(n)$. If $K\sigma_\xi^3 \le o(\sqrt{d})$, we have

$$\left\langle \mathbf{w}_c(t+1), \mathbf{u}^k \right\rangle \ge \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle + \widetilde{\Omega}\left(\frac{1}{K+Ke^{C^{-2}}}\eta\left\|\mathbf{u}\right\|_2^2\phi'\left(\left|\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle\right|\right)\right), \tag{84}$$

which shows $\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle$ is increasing. Furthermore, we have

$$\left\langle \mathbf{w}_c(t+1), \mathbf{u}^k \right\rangle \le \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle + \widetilde{O}\left(K^{-1}\eta\left\|\mathbf{u}\right\|_2^2\phi'\left(\left|\left\langle \mathbf{w}_c(t). \mathbf{u}^k \right\rangle\right|\right)\right). \tag{85}$$

Moreover, for standard learning, we set $\mathcal{S}_b = \emptyset$, and drop the condition $n_po \le o(n)$, Equation (83) still holds since the LHS of Equation (83) is 0. We conclude our proof.

We conclude our proof. □

Lemma B.4. *Under the Condition 5.1. In both standard and backdoor learning, for any $k \in [K]$, suppose $\langle \mathbf{w}_c(t), \mathbf{u}^k \rangle \leq O(C^{-1})$ and $|\langle \mathbf{w}_c(t), \boldsymbol{\xi} \rangle| \leq \widetilde{O}(\sigma_0 \sigma_\xi)$ for some $0 \leq t \leq T$, there exists*

$$\widetilde{\Omega}\left(\frac{K}{\eta\sigma_0}\right) \leq T_u \leq \widetilde{O}\left(\frac{K + Ke^{C^{-2}}}{\eta\sigma_0}\right) \tag{86}$$

such that $\max_{c\in[C]} \langle \mathbf{w}_c(T_u), \mathbf{u}^k \rangle = \Theta(C^{-1})$.

Proof. Lemma 6.1 shows $\langle \mathbf{w}_c(t), \mathbf{u}^k \rangle$ is increasing, and starting from $\Theta(\sigma_0 \|\mathbf{u}\|_2)$, it takes

$$T_u \leq \widetilde{O}\left(\sum_{i=0}^{r-1} \frac{(K + Ke^{C^{-2}})2^i\sigma_0}{\eta(2^i\sigma_0)^2}\right) \leq \widetilde{O}\left(\sum_{i=0}^{\infty} \frac{(K + Ke^{C^{-2}})2^i\sigma_0}{\eta(2^i\sigma_0)^2}\right) \leq \widetilde{O}\left(\frac{K + Ke^{C^{-2}}}{\eta\sigma_0}\right) \tag{87}$$

time steps to reach $\max_{c\in[C]} \langle \mathbf{w}_c(T), \mathbf{u}^k \rangle = 2^r \max_{c\in[C]} \langle \mathbf{w}_c(0), \mathbf{u}^k \rangle$, note that

$$T_u \geq \widetilde{\Omega}\left(\frac{2\sigma_0 K}{\eta(2\sigma_0)^2}\right) \geq \widetilde{\Omega}\left(\frac{K}{\eta\sigma_0}\right). \tag{88}$$

Thus it takes T_u to reach $\max_{c\in[C]} \langle \mathbf{w}_c(T_u), \mathbf{u}^k \rangle = \Theta(C^{-1})$. The probability of Equation (86) holds can be immediately obtained by using a union bound combining Lemmas A.5 to A.8. □

Since $\mathcal{D}_u$ is a discrete uniform distribution, and $\left\|\mathbf{u}^k\right\|_2 = \|\mathbf{u}\|_2$ holds for any k. The time that NN captures each feature vector is of the same order. Note that the lemmas shown in this subsection both hold for standard and backdoor learning, the main reason is that $n_{po} \leq o(n)$, which implies that the update rate of $\langle \mathbf{w}_c(t), \mathbf{u} \rangle$ are in the same order in both standard and backdoor learning. Practically, it is hard for users to detect the poisoned data by just comparing the loss of clean data.

B.3. Theoretical Analysis on $\langle \mathbf{w}_c(t), \mathbf{v} \rangle$

In backdoor learning, the increment of $\langle \mathbf{w}_c(t), \mathbf{v} \rangle$ is small. The following lemma shows that $\mathbf{v}$ is not learnt in standard learning.

Lemma B.5. *Under the Condition 5.1. In standard learning, for $t \leq o\left(\frac{n\sigma_0}{\eta\sigma_\xi}\right)$ and $i' \in [n]$, we have*

$$\max_{c\in[C]} |\langle \mathbf{w}_c(0), \mathbf{v} \rangle| \leq o(1). \tag{89}$$

Proof. Due to the update rule, we can rewrite $\langle \mathbf{w}_c(t+1), \mathbf{v} \rangle - \langle \mathbf{w}_c(t), \mathbf{v} \rangle$ as

$$\begin{aligned}
&\langle \mathbf{w}_c(t+1), \mathbf{v} \rangle - \langle \mathbf{w}_c(t), \mathbf{v} \rangle \\
=& -\frac{\eta}{n}\sum_{i=1}^{n} \hat{y}_i y_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i \rangle\right)\langle \mathbf{u}_i, \mathbf{v} \rangle - \frac{\eta}{n}\sum_{i=1}^{n} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i \rangle\right)\langle \boldsymbol{\xi}_i, \mathbf{v} \rangle \\
&-\frac{\eta}{n}\sum_{i=1}^{n}\sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p \rangle\right)\langle \boldsymbol{\zeta}_i^p, \mathbf{v} \rangle \\
=&\frac{\eta}{n}\sum_{i=1}^{n} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i \rangle\right)\langle \boldsymbol{\xi}_i, \mathbf{v} \rangle - \frac{\eta}{n}\sum_{i=1}^{n}\sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p \rangle\right)\langle \boldsymbol{\zeta}_i^p, \mathbf{v} \rangle.
\end{aligned} \tag{90}$$

We can upper bound $|\langle \mathbf{w}_c(t+1), \mathbf{v} \rangle - \langle \mathbf{w}_c(t), \mathbf{v} \rangle|$ as

$$|\langle \mathbf{w}_c(t+1), \mathbf{v} \rangle - \langle \mathbf{w}_c(t), \mathbf{v} \rangle| \tag{91}$$

$$\leq\frac{\eta}{n}\sum_{i=1}^{n} |\ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i \rangle\right)\langle \boldsymbol{\xi}_i, \mathbf{v} \rangle| + \frac{\eta}{n}\sum_{i=1}^{n}\sum_{p\in\mathcal{P}_i^\zeta} |\ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p \rangle\right)\langle \boldsymbol{\zeta}_i^p, \mathbf{v} \rangle| \tag{92}$$

$$\leq\widetilde{O}\left(\eta\sigma_\xi d^{-1/2}\|\mathbf{v}\|_2\right) + \widetilde{O}\left(\eta P\sigma_\zeta d^{-1/2}\|\mathbf{v}\|_2\right) \tag{93}$$

When $T \leq o\left(\frac{n\sigma_0}{\eta\sigma_\xi}\right)$, since Condition 5.1 shows that $\|\mathbf{v}\|_2 < O\left(\sigma_\xi\right)$, we have

$$\begin{aligned}|\langle \mathbf{w}_c(T), \mathbf{v}\rangle - \langle \mathbf{w}_c(0), \mathbf{v}\rangle| \leq & \widetilde{O}\left(\eta T \sigma_\xi d^{-1/2} \|\mathbf{v}\|_2\right) + \widetilde{O}\left(\eta T P \sigma_\zeta d^{-1/2} \|\mathbf{v}\|_2\right) & (94)\\ \leq & o\left(n^{-1}\eta T \sigma_\xi^2\right) + o\left(n^{-1}\eta T P \sigma_\zeta \sigma_\xi\right) \leq o(\sigma_0\sigma_\xi) & (95)\end{aligned}$$

Recall that at the beginning of the learning process, $\max_{c\in[C]} |\langle \mathbf{w}_c(0), \mathbf{v}\rangle| \leq \widetilde{O}(\sigma_0 \|\mathbf{v}\|_2)$, since $\|\mathbf{v}\|_2 < O\left(\sigma_\xi\right)$, we have $\|\mathbf{v}\|_2 \leq o\left(\frac{1}{\sigma_0}\right)$, we then conclude our proof. □

We then analyze the update of $\langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle$ in backdoor learning. In backdoor learning, the analysis on $\langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle$ is similar to $\langle \mathbf{w}_c(t), \mathbf{u}\rangle$, since $\mathbf{v}$ and $\mathbf{u}$ are both d-dimensional vectors having the different norm. In this subsection, we assume a different condition that $\langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle \leq O(C^{-1/3})$ rather than $\langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle \leq O(C^{-1})$. The motivation behind this is that we aim to find the time that $\langle \mathbf{w}_c(t), \mathbf{u}\rangle$ increases to at least $\widetilde{\Omega}(1)$, while $\langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle$ has a larger norm than $\langle \mathbf{w}_c(t), \mathbf{u}\rangle$. We call this stage the early stage. After that, in the late stage, $\langle \mathbf{w}_c(t), \mathbf{u}\rangle$ and $\langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle$ both achieves the order $\widetilde{\Omega}(1)$, and we study the late stage in Appendix C.

We firstly show that $\langle \mathbf{w}_c(t), \mathbf{v}\rangle$ is also increasing if the effects from noise vectors can be ignored.

Lemma 6.4. *Under the Condition 5.1. In backdoor learning, suppose $\langle \mathbf{w}_c(t), \mathbf{v}\rangle \leq O(C^{-1/3})$, $|\langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle| \leq \widetilde{O}\left(\sigma_0\sigma_\xi\right)$ and $|\langle \mathbf{w}_c(t), \boldsymbol{\zeta}\rangle| \leq \widetilde{O}\left(\sigma_0\sigma_\zeta\right)$ for some $0 \leq t \leq T$, we have*

$$\Delta_c^t(\mathbf{v}) = \widetilde{\Theta}\left(n_{po} n^{-1}\eta \|\mathbf{v}\|_2^2 \phi'\left(|\langle \mathbf{w}_c(t), \mathbf{v}\rangle|\right)\right) \tag{21}$$

is increasing.

Proof. Due to the update rule, we can rewrite $\langle \mathbf{w}_c(t+1), \mathbf{v}\rangle - \langle \mathbf{w}_c(t), \mathbf{v}\rangle$ as

$$\begin{aligned}&\langle \mathbf{w}_c(t+1), \mathbf{v}\rangle - \langle \mathbf{w}_c(t), \mathbf{v}\rangle\\ =& -\frac{\eta}{n}\sum_{i=1}^n \hat{y}_i y_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i\rangle\right)\langle \mathbf{u}_i, \mathbf{v}\rangle - \frac{\eta}{n}\sum_{i=1}^n \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i\rangle\right)\langle \boldsymbol{\xi}_i, \mathbf{v}\rangle\\ &-\frac{\eta}{n}\sum_{i=1}^n \sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p\rangle\right)\langle \boldsymbol{\zeta}_i^p, \mathbf{v}\rangle + \frac{\eta}{n}\sum_{i\in\mathcal{I}_b} y_i \ell'(F(\mathbf{x}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \mathbf{v}\rangle\right)\langle \mathbf{v}, \mathbf{v}\rangle\\ =&\frac{\eta}{n}\sum_{i=1}^n \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i\rangle\right)\langle \boldsymbol{\xi}_i, \mathbf{v}\rangle - \frac{\eta}{n}\sum_{i=1}^n \sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p\rangle\right)\langle \boldsymbol{\zeta}_i^p, \mathbf{v}\rangle\\ &+\frac{\eta}{n}\sum_{i\in\mathcal{I}_b} y_i \ell'(F(\mathbf{x}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \mathbf{v}\rangle\right)\langle \mathbf{v}, \mathbf{v}\rangle\end{aligned} \tag{96}$$

Since Lemma B.2 implies $|\langle \mathbf{w}_c(t), \boldsymbol{\xi}_{i'}\rangle| \leq \widetilde{O}(\sigma_0\sigma_\xi)$ and Lemma B.3 implies $|\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_{i'}^p\rangle| \leq \widetilde{O}(\sigma_0\sigma_\zeta)$, we have

$$\left|\frac{\eta}{n}\sum_{i=1}^n \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i\rangle\right)\langle \boldsymbol{\xi}_i, \mathbf{v}\rangle\right| \leq \widetilde{O}\left(\eta\sigma_0^2\sigma_\xi^3 d^{-1/2}\|\mathbf{v}\|_2\right) \tag{97}$$

$$\left|\frac{\eta}{n}\sum_{i=1}^n \sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p\rangle\right)\langle \boldsymbol{\zeta}_i^p, \mathbf{v}\rangle\right| \leq \widetilde{O}\left(\eta P\sigma_0^2\sigma_\zeta^3 d^{-1/2}\|\mathbf{v}\|_2\right) \tag{98}$$

When $\langle \mathbf{w}_c(t), \mathbf{v}\rangle \leq O(C^{-1/3})$, since $C = \log d$, we have $\hat{y}_i F(\hat{\mathbf{x}}_i) \leq \widetilde{O}(1)$, and $-\ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i) \geq \widetilde{\Omega}(1)$. We then have

$$-\frac{\eta}{n}\sum_{i\in\mathcal{I}_b} y_i \ell'(F(\mathbf{x}_i), -y_i)\phi'\left(\langle \mathbf{w}_c(t), \mathbf{v}\rangle\right)\langle \mathbf{v}, \mathbf{v}\rangle = \widetilde{\Theta}\left(n_{po} n^{-1}\eta \|\mathbf{v}\|_2^2 \phi'\left(|\langle \mathbf{w}_c(t), \mathbf{v}\rangle|\right)\right). \tag{99}$$

If $n\sigma_\xi^3 \leq o(n_{po}\sqrt{d}\left\|\mathbf{v}\right\|_2^3)$, note that $\forall i \in \mathcal{I}_b, \hat{y} = y^p$ we have

$$\langle \mathbf{w}_c(t+1), y^p\mathbf{v}\rangle = \langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle + \widetilde{\Theta}\left(n_{po}n^{-1}\eta \left\|\mathbf{v}\right\|_2^2 \phi'\left(\langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle\right)\right) \tag{100}$$

which shows $\langle \mathbf{w}_c(t), \mathbf{v}\rangle$ is increasing. Finally, by using a union bound combining Lemmas A.5 to A.8 we conclude the proof. □

Lemma B.6. *Under the Condition 5.1. In backdoor learning, suppose* $\max_c \langle \mathbf{w}_c(t), \mathbf{v}\rangle \leq O(C^{-1/3})$ *and* $|\langle \mathbf{w}_c(t), \boldsymbol{\xi}_{i'}\rangle| \leq \widetilde{O}\left(\sigma_0\sigma_\xi\right)$ *for some* $0 \leq t \leq T$. *If* $n_{po} \geq w\left(\frac{n\sigma_\xi^3}{\|\mathbf{v}\|_2^3\sqrt{d}}\right)$, *then with a probability of at least* $1 - O(\frac{n^2P^2KC}{d})$, *there exists*

$$\widetilde{\Omega}\left(\frac{n}{\eta n_{po}\left\|\mathbf{v}\right\|_2^2\sigma_0}\right) \leq T_v \leq \widetilde{O}\left(\frac{n}{\eta n_{po}\left\|\mathbf{v}\right\|_2^2\sigma_0}\right) \tag{101}$$

such that $\max_c \langle \mathbf{w}_c(T_v), \mathbf{v}\rangle = \Theta(C^{-1/3})$.

Proof. Lemma 6.1 shows $\max_c \langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle$ is increasing, and starting from $\Theta(\sigma_0\left\|\mathbf{v}\right\|_2)$, it takes

$$T_v \leq \widetilde{O}\left(\sum_{i=0}^{r-1}\frac{2^i\sigma_0 n\left\|\mathbf{v}\right\|_2}{\eta n_{po}\left\|\mathbf{v}\right\|_2^2(2^i\sigma_0\left\|\mathbf{v}\right\|_2)^2}\right) \leq \widetilde{O}\left(\sum_{i=0}^{\infty}\frac{2^i\sigma_0 n\left\|\mathbf{v}\right\|_2}{\eta n_{po}\left\|\mathbf{v}\right\|_2^2(2^i\sigma_0\left\|\mathbf{v}\right\|_2)^2}\right) \leq \widetilde{O}\left(\frac{n}{\eta n_{po}\left\|\mathbf{v}\right\|_2^3\sigma_0}\right) \tag{102}$$

time steps to reach $\max_c \langle \mathbf{w}_c(T), y^p\mathbf{v}\rangle = 2^r \max_c \langle \mathbf{w}_c(0), \mathbf{v}\rangle$, note that

$$T_v \geq \widetilde{\Omega}\left(\frac{2n\sigma_0\left\|\mathbf{v}\right\|_2}{\eta n_{po}\left\|\mathbf{v}\right\|_2^2(2\sigma_0\left\|\mathbf{v}\right\|_2)^2}\right) \geq \widetilde{\Omega}\left(\frac{n}{\eta n_{po}\left\|\mathbf{v}\right\|_2^3\sigma_0}\right). \tag{103}$$

Thus it takes T_v to reach $\max_c \langle \mathbf{w}_c(T_v), y^p\mathbf{v}\rangle = \Theta(C^{-1/3})$. The probability of Equation (101) holds can be immediately obtained by using a union bound combining Lemmas A.5 to A.8. □

When $\frac{n}{n_{po}K\|\mathbf{v}\|_2^3} \leq o(1+e^{C^{-2}})$, which means $T_u = \left(\frac{n}{\eta n_{po}\|\mathbf{v}\|_2^3\sigma_0}\right) \leq o\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right) < T_v$, and $\langle \mathbf{w}_c, y^p\mathbf{v}\rangle$ firstly achieves the order of $\Theta\left(C^{-1/3}\right)$ while $\max_c \langle \mathbf{w}_c, \mathbf{u}\rangle$ still has a small order. Since both $\max_c \langle \mathbf{w}_c(t), \mathbf{u}\rangle$ and $\max_c \langle \mathbf{w}_c(t), \mathbf{v}\rangle$ are increasing. At time $T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$, $\langle \mathbf{w}_c, \mathbf{u}\rangle$ achieves the order of $\Theta\left(C^{-1}\right)$ while $\max_c \langle \mathbf{w}_c, y^p\mathbf{v}\rangle$ is of the order at least $\Theta\left(C^{-1/3}\right)$. Moreover, Condition B.1 shows $T_v = \left(\frac{n}{\eta n_{po}\|\mathbf{v}\|_2^3\sigma_0}\right) \leq o\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right) < o\left(\frac{n\sigma_0}{\eta\sigma_\xi}\right)$, which means in the whole early stage, NN does not fit noise vectors.

C. Standard and Backdoor Learning in the Late Stage

We go on to analyze standard and backdoor learning in the stage that the network fits all training data points.

C.1. Standard Learning in the Late Stage

After time $T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$, $\max_c \langle \mathbf{w}_c, \mathbf{u}\rangle \geq \widetilde{\Omega}(1)$, then $-\ell' \leq O(1)$. $\mathbf{u}$ primarily influences the outputs of the network, and the increment of $\langle \mathbf{w}_c, \mathbf{u}\rangle$ decreases.

Lemma 6.2. *Under the Condition 5.1. In standard learning, suppose there exists* $0 \leq t \leq T$ *such that* $\forall k \in [K]$, $\left\langle \mathbf{w}_c(t), \mathbf{u}^k\right\rangle \geq \widetilde{\Omega}(1))$, *we have*

$$\forall k \in [K], \Delta_c^t(\mathbf{u}^k) \leq \widetilde{O}\left(\eta K^{-1}\|\mathbf{u}\|_2^2 e^{-\max_c\left\langle \mathbf{w}_c(t), \mathbf{u}^k\right\rangle}\right). \tag{20}$$

Proof. We have $\hat{y}F(\mathbf{x}) \geq \widetilde{\Omega}(1)$ due to the condition $\max_c \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle \geq \widetilde{\Omega}(1)$. Moreover, $-\ell'(F(\mathbf{x}), \hat{y}) = \frac{1}{1+e^{\hat{y}F(\mathbf{x})}} = \Theta(e^{-\hat{y}F(\mathbf{x})})$, which implies that

$$\left| \frac{\eta}{n} \sum_{i=1}^{n} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i) \phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i \rangle\right) \left\langle \boldsymbol{\xi}_i, \mathbf{u}^k \right\rangle \right| \leq \widetilde{O}\left(\eta \sigma_\xi d^{-1/2} \left\| \mathbf{u}^k \right\|_2 \sum_{i=1}^{n} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)} \right) \tag{104}$$

$$\left| \frac{\eta}{n} \sum_{i=1}^{n} \sum_{p \in \mathcal{P}_i^\zeta} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i) \phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p \rangle\right) \left\langle \boldsymbol{\zeta}_i^p, \mathbf{u}^k \right\rangle \right| \leq \widetilde{O}\left(\eta P \sigma_\zeta d^{-1/2} \left\| \mathbf{u}^k \right\|_2 \sum_{i=1}^{n} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)} \right). \tag{105}$$

Note that in Equations (104) and (105), we use the bound $\phi'(\cdot) \leq 1$ to avoid to discuss the order of $\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i \rangle$ and $\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p \rangle$. On the other hand, we have

$$-\frac{\eta}{n} \sum_{\mathbf{u}_i = \mathbf{u}^k} \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i) \phi'\left(\langle \mathbf{w}_c(t), y_i \mathbf{u}_i \rangle\right) \left\langle \mathbf{u}_i, \mathbf{u}^k \right\rangle = \widetilde{\Theta}\left(n^{-1} \eta \left\| \mathbf{u} \right\|_2^2 \sum_{\mathbf{u}_i = \mathbf{u}^k} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)} \right). \tag{106}$$

$\max_c \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle$ is still increasing for any $k \in [K]$ as $K\sigma_\xi \leq o(K\sigma_\xi^3) \leq o(\sqrt{d})$, which implies that

$$\forall i', \max_c \left\langle \mathbf{w}_c(t+1), \mathbf{u}^k \right\rangle - \max_c \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle = \widetilde{\Theta}\left(n^{-1} \eta \left\| \mathbf{u} \right\|_2^2 \sum_{\mathbf{u}_i = \mathbf{u}^k} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)} \right). \tag{107}$$

We should to discuss the order of $\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)$. It is clear that $\hat{y}_i F(\mathbf{x}_i, \hat{y}_i)$ has the same order for any i. Specifically, for a sample point $\hat{\mathbf{z}} = (\hat{\mathbf{x}}_{i'}, \hat{y}_{i'})$ with feature $\mathbf{u}_i = \mathbf{u}^k$. Recall the update rule of $\langle \mathbf{w}_c, \boldsymbol{\xi}_{i'} \rangle$ as shown in Equations (69) to (72), we have

$$\left| \langle \mathbf{w}_c(t+1), \boldsymbol{\xi}_{i'} \rangle - \langle \mathbf{w}_c(t), \boldsymbol{\xi}_{i'} \rangle \right| \leq \widetilde{O}\left(\eta n^{-1} \sigma_\xi^2 \ell'\left(\hat{y}_{i'} F(\mathbf{x}_i') \right) \right). \tag{108}$$

We then yield that

$$\frac{\max_c \left| \langle \mathbf{w}_c(t+1), \boldsymbol{\xi}_{i'} \rangle - \langle \mathbf{w}_c(t), \boldsymbol{\xi}_{i'} \rangle \right|}{\max_c \langle \mathbf{w}_c(t+1), \mathbf{u}^k \rangle - \max_c \langle \mathbf{w}_c(t), \mathbf{u}^k \rangle} \leq \widetilde{O}\left(\frac{\eta n^{-1} \sigma_\xi^2}{K^{-1} \eta \left\| \mathbf{u} \right\|_2^2} \right) \leq o\left(\frac{1}{\sigma_\xi} \right) \leq o(1). \tag{109}$$

Similarly, for the update rule of $\left\langle \mathbf{w}_c, \boldsymbol{\zeta}_{i'}^{p'} \right\rangle$ as shown in Equations (75) to (78), we have

$$\left| \left\langle \mathbf{w}_c(t+1), \boldsymbol{\zeta}_{i'}^{p'} \right\rangle - \left\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_{i'}^{p'} \right\rangle \right| \leq \widetilde{O}\left(\eta P n^{-1} \sigma_\zeta^2 \ell'\left(\hat{y}_{i'} F(\mathbf{x}_i') \right) \right), \tag{110}$$

and

$$\frac{\max_c \left| \left\langle \mathbf{w}_c(t+1), \boldsymbol{\zeta}_{i'}^{p'} \right\rangle - \left\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_{i'}^{p'} \right\rangle \right|}{\max_c \langle \mathbf{w}_c(t+1), \mathbf{u}^k \rangle - \max_c \langle \mathbf{w}_c(t), \mathbf{u}^k \rangle} \leq \widetilde{O}\left(\frac{\eta n^{-1} \sigma_\zeta^2}{K^{-1} \eta \left\| \mathbf{u} \right\|_2^2} \right) \leq o\left(\frac{1}{P^2 \sigma_\xi} \right) \leq o(1). \tag{111}$$

As a result, in the late stage, both $\langle \mathbf{w}_c(t), \boldsymbol{\xi} \rangle$ and $\langle \mathbf{w}_c(t), \boldsymbol{\zeta} \rangle$ increase slower than $\langle \mathbf{w}_c(t), \mathbf{u} \rangle$, for a sample point $\hat{\mathbf{z}} = (\hat{x}_i, \hat{y}_i)$ contains feature $\mathbf{u}^k$, we then have $\hat{y}F(\mathbf{x}) = \widetilde{\Theta}(\max_c \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle)$ and

$$(107) = \widetilde{\Theta}\left(K^{-1} \eta \left\| \mathbf{u} \right\|_2^2 e^{-\max_c \left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle} \right) \tag{112}$$

We conclude our proof. □

C.2. Backdoor Learning in the Late Stage

In the early stage, since $\Delta_c^t(\mathbf{u}^k) < O(\Delta_c^t(\mathbf{v}))$, $\max_c \langle \mathbf{w}_c, y^p\mathbf{v}\rangle$ firstly achieves the order of $\widetilde{\Omega}(1)$, and $\max_c \langle \mathbf{w}_c, y^p\mathbf{v}\rangle$ has a higher order than $\max_c \langle \mathbf{w}_c, \mathbf{u}\rangle$ at least until $T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$. After T_u, $\max_c \langle \mathbf{w}_c, \mathbf{u}\rangle$ also achieves the order of $\widetilde{\Omega}(1)$, and the following lemma shows trigger vector primarily influences the outputs of the model.

Lemma 6.6. *Under the Condition 5.1, in backdoor learning, suppose there exists* $0 \le t \le T$ *such that* $\forall k \in [K], \max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{u}^k\rangle \ge \widetilde{\Omega}(1)$ *and* $\max_{c\in[C]} \langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle \ge \widetilde{\Omega}(1)$*, we have*

$$\max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{u}^k\rangle \le \widetilde{O}\left(\max_{c\in[C]} \langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle\right). \tag{22}$$

Furthermore, the trigger vector $\mathbf{v}$ *primarily influence the outputs of NN:*

$$\forall k\in[K], \sum_{c\in[C]} \phi(\langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle) - \phi(\langle \mathbf{w}_c(t), \mathbf{u}^k\rangle) \ge \widetilde{\Omega}(1). \tag{23}$$

Proof. Recall the decomposition as shown in Equation (96), we have

$$\left|\frac{\eta}{n}\sum_{i=1}^{n} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i\rangle\right)\langle \boldsymbol{\xi}_i, \mathbf{v}\rangle\right| \le \widetilde{O}\left(\eta\sigma_\xi d^{-1/2}\left\|\mathbf{v}\right\|_2 \sum_{i=1}^{n} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)}\right) \tag{113}$$

$$\left|\frac{\eta}{n}\sum_{i=1}^{n}\sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^p\rangle\right)\langle \boldsymbol{\zeta}_i^p, \mathbf{v}\rangle\right| \le \widetilde{O}\left(\eta P\sigma_\zeta d^{-1/2}\left\|\mathbf{v}\right\|_2 \sum_{i=1}^{n} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)}\right) \tag{114}$$

Since $nP\sigma_\zeta = n\sigma_\xi \le o(n\sigma_\xi^3) \le o(n_{po}\sqrt{d}\left\|\mathbf{v}\right\|_2^3)$, we have

$$\Delta_c^t(\mathbf{v}) = \langle \mathbf{w}_c(t+1), y^p\mathbf{v}\rangle - \langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle = \Theta\left(n_{po}n^{-1}\eta\left\|\mathbf{v}\right\|_2^2 \sum_{i\in\mathcal{I}_b} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)}\right) \tag{115}$$

As for $\langle \mathbf{w}_c(t), \mathbf{u}^k\rangle$, since $n_{po} \le o(n)$, Equations (104) and (105) still hold, and we yield

$$-\frac{\eta}{n}\sum_{i\notin\mathcal{I}_b, \mathbf{u}_i=\mathbf{u}^k} \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i\rangle\right)\langle \mathbf{u}_i, \mathbf{u}^k\rangle = \widetilde{\Theta}\left(n^{-1}\eta\left\|\mathbf{u}\right\|_2^2 \sum_{i\notin\mathcal{I}_b, \mathbf{u}_i=\mathbf{u}^k} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)}\right) \tag{116}$$

and

$$-\frac{\eta}{n}\sum_{i\in\mathcal{I}_b, \mathbf{u}_i=\mathbf{u}^k} \ell'(F(\hat{\mathbf{x}}_i), \hat{y}_i)\phi'\left(\langle \mathbf{w}_c(t), y_i\mathbf{u}_i\rangle\right)\langle \mathbf{u}_i, \mathbf{u}^k\rangle = \widetilde{\Theta}\left(n^{-1}\eta\left\|\mathbf{u}\right\|_2^2 \sum_{i\in\mathcal{I}_b, \mathbf{u}_i=\mathbf{u}^k} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)}\right) \tag{117}$$

For $\widetilde{\Omega}\left(\frac{n}{\eta n_{po}\|\mathbf{v}\|_2^3\sigma_0}\right) \le T \le \widetilde{O}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$, $\max_c \langle \mathbf{w}_c, y^p\mathbf{v}\rangle \ge \widetilde{\Omega}(1)$ while $\max_c \langle \mathbf{w}_c(t), \mathbf{u}^k\rangle \le \widetilde{O}(1)$, it is easy to check that $\max_c \langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle$ primarily influences the outputs of the NN, and we study this problem when $\max_c \langle \mathbf{w}_c(t), \mathbf{u}\rangle$ and $\max_c \langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle$ both achieve the order of $\widetilde{\Omega}(1)$.

After $T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$, $\max_c \langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle$ achieve the order of $\widetilde{\Omega}(1)$ and $\Delta_c^t(\mathbf{v})$ decreases. Given k, when $\max_c \langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle > o(\max_c \langle \mathbf{w}_c(t), \mathbf{u}^k\rangle)$, the outputs are manipulated by $\mathbf{v}$. We discuss when $\max_c \langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle \le O(\max_c \langle \mathbf{w}_c(t), \mathbf{u}^k\rangle)$.

We prove that $\forall k \in [K], \sum_{c\in[C]} \phi(\langle \mathbf{w}_c(t), y^p\mathbf{v}\rangle) - \sum_{c\in[C]} \phi(\langle \mathbf{w}_c(t), \mathbf{u}^k\rangle) \ge \widetilde{\Omega}(1)$ using an induction. Given k, we suppose that there exists T such that (1) $\max_c \langle \mathbf{w}_c(T), y^p\mathbf{v}\rangle = \widetilde{\Theta}(\max_c \langle \mathbf{w}_c(T), \mathbf{u}^k\rangle)$, (2) $\sum_{c\in[C]} \phi(\langle \mathbf{w}_c(T), y^p\mathbf{v}\rangle) - \sum_{c\in[C]} \phi(\langle \mathbf{w}_c(T), \mathbf{u}^k\rangle) \ge \widetilde{\Omega}(1)$, (3) $\max_c \langle \mathbf{w}_c(T), \mathbf{u}^k\rangle \ge \widetilde{\Omega}(1)$. We show that at $T+1$, these conditions still hold.

At this time, $n^{-1}\eta\left\|\mathbf{u}\right\|_2^2\sum_{i\notin\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)} < o\left(n^{-1}\eta\left\|\mathbf{u}\right\|_2^2\sum_{i\in\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}\right)$, if $K\sigma_\xi^3 \le o(\sqrt{d})$, we have

$$\Delta_c^T(\mathbf{u}^k) = \left\langle\mathbf{w}_c(T+1),\mathbf{u}^k\right\rangle - \left\langle\mathbf{w}_c(T),\mathbf{u}^k\right\rangle = \widetilde{\Theta}\left(n^{-1}\eta\left\|\mathbf{u}\right\|_2^2\left(\sum_{i\notin\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)} - \sum_{i\in\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}\right)\right). \tag{118}$$

The ratio of updates can be written as

$$\begin{aligned}\frac{\Delta_c^T(\mathbf{u}^k)}{\Delta_c^T(\mathbf{v})} =&\widetilde{\Theta}\left(\frac{n^{-1}\eta\left\|\mathbf{u}^k\right\|_2^2\left(\sum_{i\notin\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)} - \sum_{i\in\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}\right)}{n^{-1}\eta\left\|\mathbf{v}\right\|_2^2\sum_{i\in\mathcal{I}_b}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}}\right) &(119)\\ \le&\widetilde{O}\left(\frac{n^{-1}\eta\left\|\mathbf{u}^k\right\|_2^2\sum_{i\notin\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}}{n^{-1}\eta\left\|\mathbf{v}\right\|_2^2\sum_{i\in\mathcal{I}_b}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}}\right) + \widetilde{O}\left(\frac{n^{-1}\eta\left\|\mathbf{u}^k\right\|_2^2\left(\sum_{i\in\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}\right)}{n^{-1}\eta\left\|\mathbf{v}\right\|_2^2\sum_{i\in\mathcal{I}_b}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}}\right) &(120)\\ \le&O\left(\frac{n\left\|\mathbf{u}\right\|_2^2}{n_{po}K\left\|\mathbf{v}\right\|_2^2}\right) + O\left(\frac{\left\|\mathbf{u}\right\|_2^2}{K\left\|\mathbf{v}\right\|_2^2}\right) \le o\left(1\right) + o\left(\frac{n_{po}}{n}\right) \le o(1). &(121)\end{aligned}$$

The last two inequalities due to the conditions $n_{po}\left\|\mathbf{v}\right\|_2^2 \ge \omega(nK^{-1})$, and $n_{po} \le o(n)$. Equation (121) shows that at $T+1$, the three conditions still hold. Furthermore, Equations (109) and (111) hold when $\sum_{c\in[C]}\phi(\langle\mathbf{w}_c(T),y^p\mathbf{v}\rangle) - \sum_{c\in[C]}\phi(\left\langle\mathbf{w}_c(T),\mathbf{u}^k\right\rangle) \ge \widetilde{\Omega}(1)$, which means both $\langle\mathbf{w}_c(T),\boldsymbol{\xi}\rangle$ and $\langle\mathbf{w}_c(T),\boldsymbol{\zeta}\rangle$ increase slower than $\max_{c\in[C]}\left\langle\mathbf{w}_c(T),\mathbf{u}^k\right\rangle$, $\max_{c\in[C]}\left\langle\mathbf{w}_c(T),\mathbf{u}^k\right\rangle$ increase slower than $\max_{c\in[C]}\langle\mathbf{w}_c,y^p\mathbf{v}\rangle$. We then have $\forall i\in\mathcal{I}_b,\hat{y}_iF(\hat{\mathbf{x}}_i) = y^pF(\hat{\mathbf{x}}_i) \ge \widetilde{\Omega}(1)$, and $\forall i\notin\mathcal{I}_b,\hat{y}_iF(\hat{\mathbf{x}}_i) = y_iF(\hat{\mathbf{x}}_i) \ge \widetilde{\Omega}(1)$.

Finally, at $T = T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$, $\sum_{c\in[C]}\phi(\left\langle\mathbf{w}_c(T),\mathbf{u}^k\right\rangle) > \max_c\phi(\left\langle\mathbf{w}_c(T),\mathbf{u}^k\right\rangle) \ge \Omega\left(\frac{1}{C}\right)$ while $\sum_{c\in[C]}\phi(\left\langle\mathbf{w}_c(T),\mathbf{u}^k\right\rangle) < C\phi(\max_c\left\langle\mathbf{w}_c(T),\mathbf{u}^k\right\rangle) < O\left(\frac{1}{C^2}\right)$ for any $k\in[K]$. Therefore, using the induction, after $T \ge T_u$, we have $\forall k\in[K],\sum_{c\in[C]}\phi(\langle\mathbf{w}_c(T),y^p\mathbf{v}\rangle) - \phi(\left\langle\mathbf{w}_c(T),\mathbf{u}^k\right\rangle) \ge \widetilde{\Omega}(1)$, and $\mathbf{v}$ primarily influences the outputs of the NN. Equation (115) implies that there exists $T_u' \ge \Omega\left(\frac{ne^{\text{poly}(d)}}{\eta n_{po}\|\mathbf{v}\|_2^2}\right)$ to reach that $\max_{c\in[C]}\langle\mathbf{w}_c,y^p\mathbf{v}\rangle > \omega(1)$, and $\max_c\langle\mathbf{w}_c,y^p\mathbf{v}\rangle > o(\max_c\left\langle\mathbf{w}_c,\mathbf{u}^k\right\rangle)$, which means $\mathbf{v}$ still primarily influences the outputs in the late stage.

Additionally, $\max_c\left\langle\mathbf{w}_c(t),\mathbf{u}^k\right\rangle$ continues to increase when $\Delta_c^t(\mathbf{u}^k) \ge 0$, which means

$$\left(\sum_{i\notin\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}\right) \ge \Omega\left(\sum_{i\in\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}\right), \tag{122}$$

$\sum_{c\in[C]}\phi(\langle\mathbf{w}_c(t),y^p\mathbf{v}\rangle) - \phi(\left\langle\mathbf{w}_c(t),\mathbf{u}^k\right\rangle) \ge \widetilde{\Omega}(1)$ implies that

$$\frac{\sum_{i\notin\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}}{\sum_{i\in\mathcal{I}_b,\mathbf{u}_i=\mathbf{u}^k}e^{-\hat{y}_iF(\hat{\mathbf{x}}_i,\hat{y}_i)}} = \Theta\left(\frac{(n-n_{po})e^{-\max_c\left\langle\mathbf{w}_c(t),\mathbf{u}^k\right\rangle}}{n_{po}e^{-\max_c\langle\mathbf{w}_c(t),y^p\mathbf{v}\rangle}}\right). \tag{123}$$

By rewriting Equation (122), we have

$$\max_c\left\langle\mathbf{w}_c(t),\mathbf{u}^k\right\rangle \le \widetilde{O}\left(\max_c\langle\mathbf{w}_c(t),y^p\mathbf{v}\rangle\right). \tag{124}$$

The inequality holds due to $\Omega\left(1\right) \le n_{po} \le o(n)$. We conclude our proof. □

We immediately have the following lemma, which shows that both $\Delta_c^t(\mathbf{u}^k)$ and $\Delta_c^t(\mathbf{v})$ are upper bounded in the late stage.

Lemma 6.8. *Under the Condition 5.1. Suppose there exists* $0 \leq t \leq T$ *such that* $\forall k \in [K]$, $\left\langle \mathbf{w}_c(t), \mathbf{u}^k \right\rangle \geq \widetilde{\Omega}(1))$ *and* $\langle \mathbf{w}_c(t), y^p \mathbf{v}\rangle \geq \widetilde{\Omega}\,(1)$*, we have*

$$\forall k \in [K], \Delta_c^t(\mathbf{u}^k) \leq \widetilde{O}\left(K^{-1}\eta \left\|\mathbf{u}\right\|_2^2 e^{-\max_c \left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle}\right), \tag{24}$$

$$\Delta_c^t(\mathbf{v}) \leq \widetilde{O}\left(n_{po} n^{-1}\eta \left\|\mathbf{v}\right\|_2^2 e^{-\max_c \langle \mathbf{w}_c, y^p \mathbf{v}\rangle}\right). \tag{25}$$

Proof. Equation (118) can be further bounded as

$$\begin{aligned}
\Delta_c^t(\mathbf{u}^k) =& \widetilde{\Theta}\left(n^{-1}\eta \left\|\mathbf{u}\right\|_2^2 \left(\sum_{i \notin \mathcal{I}_b, \mathbf{u}_i = \mathbf{u}^k} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)} - \sum_{i \in \mathcal{I}_b, \mathbf{u}_i = \mathbf{u}^k} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)}\right)\right) \\
\leq & \widetilde{O}\left(n^{-1}\eta \left\|\mathbf{u}\right\|_2^2 \left(\sum_{\mathbf{u}_i = \mathbf{u}^k} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)}\right)\right) \\
\leq & \widetilde{O}\left(n^{-1}\eta \left\|\mathbf{u}\right\|_2^2 \left(\sum_{\mathbf{u}_i = \mathbf{u}^k} e^{-\max_c \left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle}\right)\right) = \widetilde{O}\left(K^{-1}\eta \left\|\mathbf{u}\right\|_2^2 e^{-\max_c \left\langle \mathbf{w}_c, \mathbf{u}^k \right\rangle}\right).
\end{aligned}$$

The last inequality due to Lemma 6.6. Similarly, Equation (115) can be upper bounded as

$$\Delta_c^t(\mathbf{v}) \leq O\left(n^{-1}\eta \left\|\mathbf{v}\right\|_2^2 \sum_{i \in \mathcal{I}_b} e^{-\hat{y}_i F(\hat{\mathbf{x}}_i, \hat{y}_i)}\right) \leq \widetilde{O}\left(n_{po} n^{-1}\eta \left\|\mathbf{v}\right\|_2^2 e^{-\max_c \langle \mathbf{w}_c, y^p \mathbf{v}\rangle}\right). \tag{125}$$

□

D. Proofs for Main Results

Theorem 5.3. *[standard learning] Under the Condition 5.1, given a clean training set* $\mathcal{S}_{cl}^{tr}$ *with size* n*, there exists* $T_u = \widetilde{\Theta}\left(\frac{K + K e^{C-2}}{\eta \sigma_0}\right)$ *such that for* $T_1 \geq T_u$*, the network* $\hat{F}_{T_1}$ *fits all clean data points with a high probability:*

$$\mathbb{P}(\forall i \in [n], y_i \hat{F}_{T_1}(\mathbf{x}_i) \geq \widetilde{\Omega}(1)) \geq 1 - O\left(\frac{n^2 P^2 K C}{poly(d)}\right). \tag{12}$$

Moreover, $\hat{F}_{T_1}$ *achieves a high clean accuracy but leaves a low attack success rate at* T_1*:*

$$Acc(\hat{F}_{T_1}; \mathcal{D}_{\mathbf{z}}) \geq 1 - O\left(\frac{n P^2 K C}{poly(d)}\right), \tag{13}$$

$$ASR(\hat{F}_{T_1}; \mathcal{D}_{\mathbf{z}}, \mathfrak{P}) \leq O\left(\frac{n P^2 K C}{poly(d)}\right). \tag{14}$$

Proof. We can regard standard learning as a special case for backdoor learning with $n_{po} = 0$. Condition B.1 shows that

$\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right) \leq o\left(\frac{n\sigma_0}{\eta\sigma_\xi}\right)$, at $T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$, for a clean training sample, the output can be rewritten as:

$$\begin{aligned}
\hat{y}_i F(\hat{\mathbf{x}}_i; T_u) &= \sum_{c\in[C]} \hat{y}_i \phi\left(\langle \mathbf{w}_c(T_u), \hat{\mathbf{x}}_i^p\rangle\right) \\
&= \sum_{c\in[C]} \langle \mathbf{w}_c(T_u), \mathbf{u}_i\rangle + \sum_{c\in[C]} \hat{y}_i \phi\left(\langle \mathbf{w}_c(T_u), \boldsymbol{\xi}_i\rangle\right) + \sum_{c\in[C]} \sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i \phi\left(\langle \mathbf{w}_c(T_u), \boldsymbol{\zeta}_i^p\rangle\right) \\
&\geq \sum_{c\in[C]} \phi\left(\langle \mathbf{w}_c(T_u), \mathbf{u}_i\rangle\right) - \sum_{c\in[C]} \phi\left(|\langle \mathbf{w}_c(T_u), \boldsymbol{\xi}_i\rangle|\right) - \sum_{c\in[C]} \sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i \phi\left(|\langle \mathbf{w}_c(T_u), \boldsymbol{\zeta}_i^p\rangle|\right) \\
&\geq \max_{c\in[C]} \phi\left(\langle \mathbf{w}_c(T_u), \mathbf{u}_i\rangle\right) - C \max_{c\in[C]} \phi\left(|\langle \mathbf{w}_c(T_u), \boldsymbol{\xi}_i\rangle|\right) - CP \max_{c\in[C], p\in\mathcal{P}_i^\zeta} \phi\left(|\langle \mathbf{w}_c(T_u), \boldsymbol{\zeta}_i^p\rangle|\right) \\
&\geq \Omega\left(\frac{1}{C^3}\right) - CO\left(\sigma_0^3\sigma_\xi^3\right) - CP^{-2}O\left(\sigma_0^3\sigma_\xi^3\right) \geq \widetilde{\Omega}\left(1\right) \qquad (126)
\end{aligned}$$

Using the union bound, we obtain Equation (12). Equation (126) shows that all the training data points have been fit by hypothesis F at time $T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$. We further evaluate the performance of F on the population distribution. Given a data point $\mathbf{z} = (\mathbf{x}, y)$ sampled from $\mathcal{D}_{\mathbf{z}}$, since the training data is i.i.d. drawn from $\mathcal{D}_{\mathbf{z}}$, we have a similar result with Equation (131). We can regard $\{(\mathbf{x}, y)\}$ as another clean set $\mathcal{S}'_{cl}$ with size 1.By using a union bound combining Lemmas A.5 to A.8, with a probability of $1 - O\left(\frac{nP^2K^2C}{\text{poly}(d)}\right)$, we have

$$yF_{T_u}(\mathbf{x}) = \sum_{c\in[C]} y_i \phi\left(\langle \mathbf{w}_c(T_u), \hat{\mathbf{x}}^p\rangle\right) \geq \Omega\left(\frac{1}{C^3}\right) - CO\left(\sigma_0^3\sigma_\xi^3\right) - CP^{-2}O\left(\sigma_0^3\sigma_\xi^3\right) \geq \widetilde{\Omega}\left(1\right) \qquad (127)$$

The probability of Equation (127) holds is $1 - O\left(\frac{nP^2KC}{\text{poly}(d)}\right)$ rather than $1 - O\left(\frac{n^2P^2KC}{\text{poly}(d)}\right)$ since the size of $\mathcal{S}'_{cl}$ is 1 instead of n.

In the late stage, for $T \geq T_u$, $\max_{c\in[C]} \langle \mathbf{w}_c(T), \mathbf{u}_i\rangle \geq \widetilde{\Omega}\left(1\right)$, Equations (109) and (111) imply that $\hat{y}_i F(\hat{\mathbf{x}}_i; T) = \sum_{c\in[C]} \hat{y}_i \phi\left(\langle \mathbf{w}_c(T), \hat{\mathbf{x}}_i^p\rangle\right) \geq \widetilde{\Omega}\left(1\right)$ for each training data $\mathbf{z}_i = (\hat{x}_i, \hat{y}_i)$ and $yF_{T_u}(\mathbf{x}) \geq \widetilde{\Omega}\left(1\right)$ for a clean test data $\mathbf{z} = (\mathbf{x}, y)$.

We then evaluate the attack fail rate for poisoned data $\mathfrak{P}(\mathbf{z}) = (\mathfrak{P}^X(\mathbf{x}), \mathfrak{P}^Y(y))$ in both the early and late stages. For true label y, the output of NN can be rewritten as:

$$\begin{aligned}
yF(\mathfrak{P}^X(\mathbf{x}); T) &= \sum_{c\in[C]} y\phi\left(\langle \mathbf{w}_c(T), \hat{\mathbf{x}}^p\rangle\right) + \sum_{c\in[C]} \sum_{p\in\mathcal{P}_i^\zeta} y\phi\left(\langle \mathbf{w}_c(T), \boldsymbol{\zeta}^p\rangle\right) \\
&= \sum_{c\in[C]} \langle \mathbf{w}_c(T), \mathbf{u}\rangle + \sum_{c\in[C]} y\phi\left(\langle \mathbf{w}_c(T), \mathbf{v}\rangle\right) + \sum_{c\in[C]} y\phi\left(\langle \mathbf{w}_c(T), \boldsymbol{\xi}\rangle\right) - \sum_{c\in[C]} \sum_{p\in\mathcal{P}_i^\zeta} y\phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\zeta}^p\rangle|\right) \\
&\geq \sum_{c\in[C]} \phi\left(\langle \mathbf{w}_c(T), \mathbf{u}\rangle\right) - \sum_{c\in[C]} \phi\left(|\langle \mathbf{w}_c(T), \mathbf{v}\rangle|\right) - \sum_{c\in[C]} \phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\xi}\rangle|\right) \\
&\geq \max_{c\in[C]} \phi\left(\langle \mathbf{w}_c(T), \mathbf{u}\rangle\right) - C\max_{c\in[C]} \phi\left(|\langle \mathbf{w}_c(T), \mathbf{v}\rangle|\right) - C\max_{c\in[C]} \phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\xi}\rangle|\right) - CP \max_{c\in[C], p\in\mathcal{P}_i^\zeta} \phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\zeta}^p\rangle|\right) \quad (128)
\end{aligned}$$

For $0 \leq T \leq T_u \leq o\left(\frac{n\sigma_0}{\eta\sigma_\xi}\right)$, Lemma B.2 implies $\max_{c\in[C]} |\langle \mathbf{w}_c(T), \boldsymbol{\xi}_{i'}\rangle| \leq \widetilde{O}(\sigma_0\sigma_\xi)$, Lemma B.3 implies $\max_{c\in[C]} |\langle \mathbf{w}_c(T), \boldsymbol{\zeta}_{i'}^p\rangle| \leq \widetilde{O}(\sigma_0\sigma_\zeta)$, Lemma B.5 implies that $\max_{c\in[C]} \phi\left(|\langle \mathbf{w}_c(T_u), \mathbf{v}\rangle|\right) \leq o(1)$, we have

$$(128) \geq \Omega\left(\frac{1}{C^3}\right) - Co(1) - CP^{-2}O\left(\sigma_0^3\sigma_\xi^3\right) \geq \widetilde{\Omega}\left(1\right) \qquad (129)$$

This implies that $\mathfrak{P}^Y(y)F(\mathfrak{P}^X(\mathbf{x}); T) = -yF(\mathfrak{P}^X(\mathbf{x}); T) \leq -\widetilde{\Omega}\left(1\right)$. For $T > T_u$, Equations (109) and (111) imply that for any k, $\mathbf{u}^k$ has a faster rate than noise vectors. As for $\mathbf{v}$, recall that $\|\mathbf{v}\|_2 \leq O(\sigma_\xi)$, Equation (93) implies that for any

$k \in [K]$,

$$\frac{\max_c |\langle \mathbf{w}_c(t+1),\mathbf{v}\rangle - \langle \mathbf{w}_c(t),\mathbf{v}\rangle|}{\max_c \langle \mathbf{w}_c(t+1),\mathbf{u}^k\rangle - \max_c \langle \mathbf{w}_c(t),\mathbf{u}^k\rangle} \le \widetilde{O}\left(\frac{\eta d^{-1/2}\sigma_\xi \|\mathbf{v}\|_2 + \eta d^{-1/2} P\sigma_\zeta \|\mathbf{v}\|_2}{K^{-1}\eta\|\mathbf{u}\|_2^2}\right) \le o(1). \tag{130}$$

Consequently, $\mathbf{u}$ primarily influences the outputs of NN, and we have $\mathfrak{P}^Y(y)F(\mathfrak{P}^X(\mathbf{x});T) = -yF(\mathfrak{P}^X(\mathbf{x});T) \le -\widetilde{\Omega}\,(1)$ for $T > T_u$. We conclude our proof. □

Theorem 5.4. *[Backdoor Learning] Under the Condition 5.1, given a poisoned training set $\mathcal{S}_{po}^{tr}$ with size n, if $n_{po}\|\mathbf{v}\|_2^2 > \omega(nK^{-1})$, there exists $T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$ such that for $T_2 \ge T_u$ the network $\hat{F}_{T_2}$ fits both clean and poisoned training data points with a high probability:*

$$\mathbb{P}(\forall i \in [n], \hat{y}_i \hat{F}_{T_2}(\hat{\mathbf{x}}_i) \ge \widetilde{\Omega}(1)) \ge 1 - O\left(\frac{n^2P^2KC}{poly(d)}\right). \tag{15}$$

Furthermore, there exists $T_v = \widetilde{\Theta}\left(\frac{n}{\eta n_{po}\|\mathbf{v}\|_2^3\sigma_0}\right)$ such that $\hat{F}$ achieves high attack success rate at $T_2' \ge T_v$ and achieves high clean accuracy at $T_2 \ge T_u > T_v$:

$$Acc(\hat{F}_{T_2}; \mathcal{D}_{\mathbf{z}}) \ge 1 - O\left(\frac{nP^2KC}{poly(d)}\right), \tag{16}$$

$$ASR(\hat{F}_{T_2'}; \mathcal{D}_{\mathbf{z}}, \mathfrak{P}) \ge 1 - O\left(\frac{nP^2KC}{poly(d)}\right). \tag{17}$$

Proof. Condition B.1 shows that $T_v = \widetilde{\Theta}\left(\frac{n}{\eta n_{po}\|\mathbf{v}\|_2^3\sigma_0}\right) \le o\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$ and $T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right) \le o\left(\frac{n\sigma_0}{\eta\sigma_\xi}\right)$, which means that the NN firstly fits the trigger vector, and then fits all feature vectors. The effects of noise vectors can be always ignored. Based on the results in Appendix B, for a clean training sample, the output can be rewritten as:

$$\begin{aligned}
&\hat{y}_i F_T(\hat{\mathbf{x}}_i) = \sum_{c\in[C]} \hat{y}_i \phi\left(\langle \mathbf{w}_c(T), \hat{\mathbf{x}}_i^p\rangle\right) \\
&= \sum_{c\in[C]} \langle \mathbf{w}_c(T), \mathbf{u}_i\rangle + \sum_{c\in[C]} \hat{y}_i \phi\left(\langle \mathbf{w}_c(T), \boldsymbol{\xi}_i\rangle\right) + \sum_{c\in[C]}\sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i \phi\left(\langle \mathbf{w}_c(T), \boldsymbol{\zeta}_i^p\rangle\right) \\
&\ge \sum_{c\in[C]} \phi\left(\langle \mathbf{w}_c(T), \mathbf{u}_i\rangle\right) - \sum_{c\in[C]} \phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\xi}_i\rangle|\right) - \sum_{c\in[C]}\sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i \phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\zeta}_i^p\rangle|\right) \\
&\ge \max_{c\in[C]} \phi\left(\langle \mathbf{w}_c(T), \mathbf{u}_i\rangle\right) - C\max_{c\in[C]} \phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\xi}_i\rangle|\right) - CP \max_{c\in[C], p\in\mathcal{P}_i^\zeta} \phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\zeta}_i^p\rangle|\right)
\end{aligned} \tag{131}$$

In the early stage, for $T_u = \widetilde{\Theta}\left(\frac{K+Ke^{C^{-2}}}{\eta\sigma_0}\right)$, we have

$$(131) \ge \Omega\left(\frac{1}{C^3}\right) - CO\left(\sigma_0^3\sigma_\xi^3\right) - CP^{-2}O\left(\sigma_0^3\sigma_\xi^3\right) \ge \widetilde{\Omega}\,(1)\,.$$

In the late stage, for $T \ge T_u$, Equations (109) and (111) imply that both $\langle \mathbf{w}_c(t), \boldsymbol{\xi}_i\rangle$ and $\left\langle \mathbf{w}_c(t), \boldsymbol{\zeta}_i^{p'}\right\rangle$ increase slower than $\max_{c\in[C]}\langle \mathbf{w}_c(t), \mathbf{u}\rangle$, and we have $\hat{y}_i F_T(\hat{\mathbf{x}}_i) \ge \widetilde{\Omega}\,(1)$.

For a backdoor training sample $(\hat{\mathbf{x}}_i, \hat{y}_i)$, the output can be rewritten as:

$$\begin{aligned}
&\hat{y}_i F_T(\hat{\mathbf{x}}_i) = \sum_{c\in[C]} \hat{y}\phi\left(\langle \mathbf{w}_c(T), \hat{\mathbf{x}}_i^p\rangle\right) + \sum_{c\in[C]}\sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i\phi\left(\langle \mathbf{w}_c(T), \boldsymbol{\zeta}_i^p\rangle\right) \\
&= \sum_{c\in[C]} \langle \mathbf{w}_c(T), \mathbf{u}_i\rangle + \sum_{c\in[C]} \hat{y}_i\phi\left(\langle \mathbf{w}_c(T), \mathbf{v}\rangle\right) + \sum_{c\in[C]} \hat{y}_i\phi\left(\langle \mathbf{w}_c(T), \boldsymbol{\xi}_i\rangle\right) - \sum_{c\in[C]}\sum_{p\in\mathcal{P}_i^\zeta} \hat{y}_i\phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\zeta}_i^p\rangle|\right) \\
&\geq \sum_{c\in[C]} \hat{y}_i\phi\left(\langle \mathbf{w}_c(T), \mathbf{v}\rangle\right) - \sum_{c\in[C]} \phi\left(\langle \mathbf{w}_c(T), \mathbf{u}_i\rangle\right) - \sum_{c\in[C]} \phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\xi}_i\rangle|\right) \\
&\geq \max_{c\in[C]} \hat{y}_i\phi\left(\langle \mathbf{w}_c(T), \mathbf{v}\rangle\right) - C\max_{c\in[C]} \phi\left(\langle \mathbf{w}_c(T), \mathbf{u}_i\rangle\right) - C\max_{c\in[C]} \phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\xi}_i\rangle|\right) - CP\max_{c\in[C],p\in\mathcal{P}_i^\zeta} \phi\left(|\langle \mathbf{w}_c(T), \boldsymbol{\zeta}_i^p\rangle|\right).
\end{aligned} \tag{132}$$

In the early stage, for $T_v \leq T \leq T_u$, $\max_{c\in[C]} \phi\left(\langle \mathbf{w}_c(T), \mathbf{u}_i\rangle\right)$ is at most of the order of $O\left(\frac{1}{C^3}\right)$, we have

$$(132) \geq \Omega\left(\frac{1}{C}\right) - O\left(\frac{1}{C^2}\right) - CO\left(\sigma_0^3\sigma_\xi^3\right) - CP^{-2}O\left(\sigma_0^3\sigma_\xi^3\right) \geq \widetilde{\Omega}\left(1\right)$$

while in the late stage, for $T > T_u$,when $\max_{c\in[C]} \langle \mathbf{w}_c(T), \mathbf{u}_i\rangle$ and $\max_{c\in[C]} \langle \mathbf{w}_c(T), y^p\mathbf{v}\rangle$ reach the same order of $\widetilde{\Theta}(1)$, Lemma 6.6 implies that $\sum_{c\in[C]} \langle \mathbf{w}_c(T), y^p\mathbf{v}\rangle - \langle \mathbf{w}_c(T), \mathbf{u}_i\rangle \geq \widetilde{\Omega}(1)$ until $\max_{c\in[C]} \langle \mathbf{w}_c(T), y^p\mathbf{v}\rangle$ reach a higher order than $\max_{c\in[C]} \langle \mathbf{w}_c(T), \mathbf{u}_i\rangle$, and we also have $\hat{y}_i F_T(\hat{\mathbf{x}}_i) \geq \widetilde{\Omega}\left(1\right)$ for $T > T_u$.

By using a union bound combining Lemmas A.5 to A.8, we obtain Equation (12). Equations (131) and (132) shows that all the training data points have been fit by hypothesis F at time $T \geq T_u$. We further evaluate the performance of F on the population distribution. Given a data point $\mathbf{z} = (\mathbf{x}, y)$ sampled from $\mathcal{D}_\mathbf{z}$, since the training data is i.i.d. drawn from $\mathcal{D}_\mathbf{z}$, we have a similar result with Equation (131). We can regard $\{(\mathbf{x}, y)\}$ as another clean training set with size 1, so with a probability of $1 - O\left(\frac{nP^2KC}{\text{poly}(d)}\right)$, we have

$$yF_T(\mathbf{x}) = \sum_{c\in[C]} y_i\phi\left(\langle \mathbf{w}_c(T), \hat{\mathbf{x}}^p\rangle\right) \geq \Omega\left(\frac{1}{C^3}\right) - CO\left(\sigma_0^3\sigma_\xi^3\right) - CP^{-2}O\left(\sigma_0^3\sigma_\xi^3\right) \geq \widetilde{\Omega}\left(1\right) \tag{133}$$

For a poisoned data $\mathfrak{P}(\mathbf{z}) = (\mathfrak{P}^X(\mathbf{x}), \mathfrak{P}^Y(y))$, with a probability of $1 - O\left(\frac{nP^2KC}{\text{poly}(d)}\right)$, we have

$$\mathfrak{P}^Y(y) F_T(\mathfrak{P}^X(\mathbf{x})) = \sum_{c\in[C]} \mathfrak{P}^Y(y)\phi\left(\langle \mathbf{w}_c(T), \mathbf{x}^p\rangle\right) \tag{134}$$

$$\geq \begin{cases} \Omega\left(\frac{1}{C}\right) - o\left(1\right) - CO\left(\sigma_0^3\sigma_\xi^3\right) - CP^{-2}O\left(\sigma_0^3\sigma_\xi^3\right) \geq \widetilde{\Omega}\left(1\right) & \text{if } T_v \leq T < T_u \\ \Omega\left(\frac{1}{C}\right) - O\left(\frac{1}{C^2}\right) - CO\left(\sigma_0^3\sigma_\xi^3\right) - CP^{-2}O\left(\sigma_0^3\sigma_\xi^3\right) \geq \widetilde{\Omega}\left(1\right) & \text{if } T = T_u \\ \widetilde{\Omega}\left(1\right) & \text{if } T > T_u. \end{cases} \tag{135}$$

we conclude our proof. □

E. Description of Experiments and More Empirical Results

In our experiments, we use two datasets, MNIST and CIFAR-10. MNIST contains grayscale handwritten digits with 10 classes, while CIFAR-10 contains color images with 10 classes. For MNIST, we collect the data points from classes 0 and 5 as a binary classification task. The background of data in MNIST is all black, which means the background noise is absent. For CIFAR-10, we sample the data points from classes ’Airplane’ and ’Bird’ as a binary classification task. Different from MNIST, the features and backgrounds in CIFAR-10 are complex. We train a LeNet-5 with 80 epochs on the MNIST and a ResNet-18 with 100 epochs on the CIFAR-10. The clean images and the poisoned images with trigger pattern are shown in Figure 3.

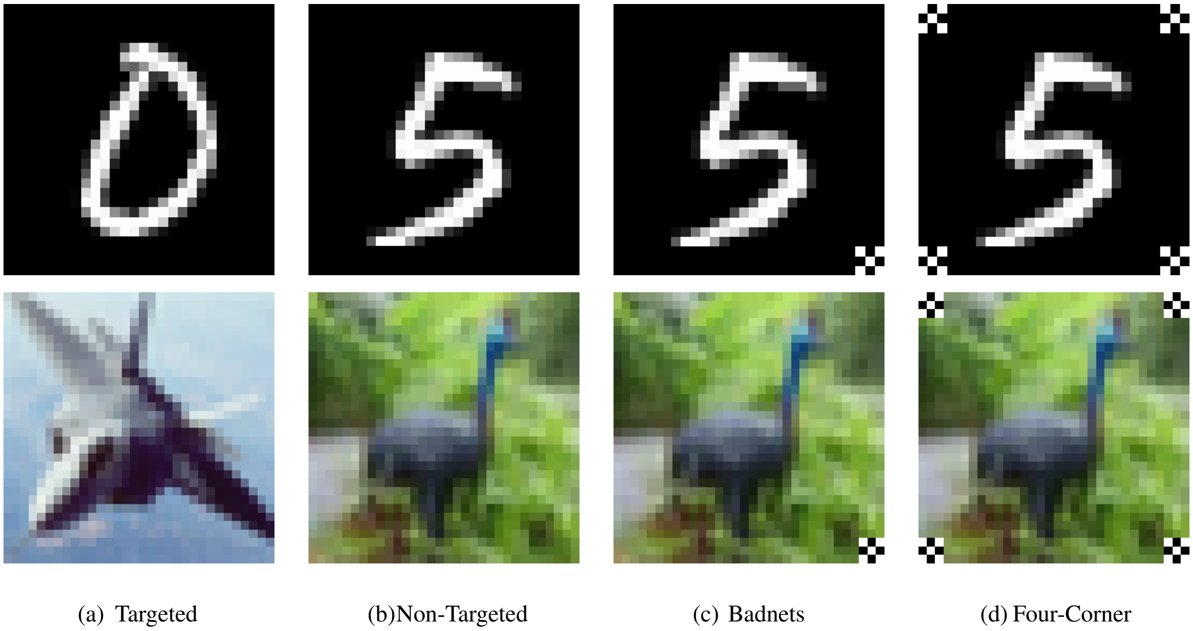


Figure 3. Clean and Poisoned Data in MNIST and CIFAR-10. (a) A clean image from the targeted class. (b) A clean image from the non-targeted class. (c) The Poisoned image generated by BadNets (d) The Poisoned image generated by the four-corner attack. We show the images from MNIST in the first row while the images from CIFAR-10 are in the second row.

E.1. More Analysis about Poisoned Data

We show the results of the cosine similarities of the maximum singular vector and representation vectors under the four-corner attack in Figure 4(a). We observe a similar result that the direction of poisoned data from the non-targeted class is closer to the clean data from the target class than from the non-target class. Moreover, we show the results of the visualization of representation vectors with T-SNE in Figure 4(b). It has a similar distribution to the results under the BadNets attack. We further show the results under the BadNets and four-corner attack on MNIST in Appendix E.1 and Appendix E.1, respectively. The results in MNIST are similar to that in CIFAR-10.

E.2. More Results on the Key Components for Backdoor Attacks

We change the norm of the trigger vector. We employ a linear combination of the original patch and the trigger pattern instead of simply reducing the norm $\mathbf{v}$, as it is a more natural approach for color images. Decreasing the norm alone may result in the patch becoming closer to a pure black patch. However, a black patch can also be perceived as a specific trigger, especially when the background of the image is not entirely black. The generalized Patch attack is defined as follows:

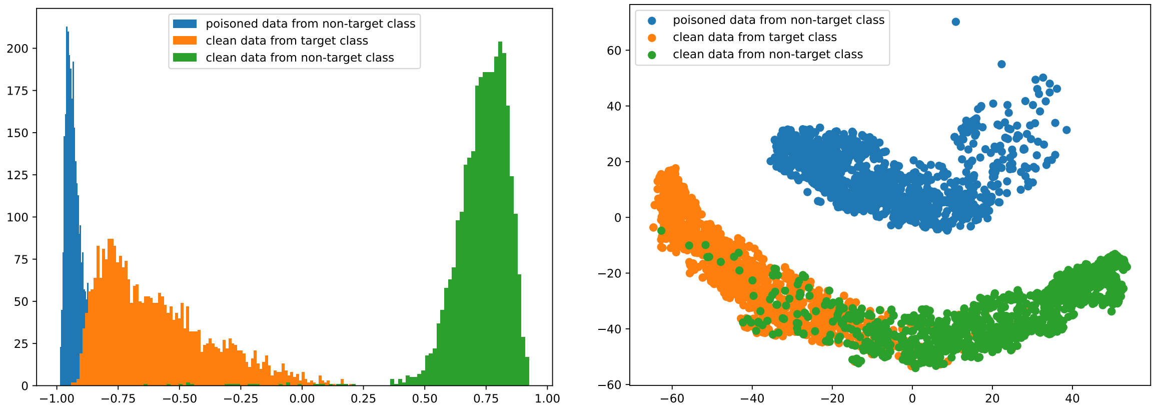
Definition E.1 (Generalized Patch attack)**.** Given a trigger $\mathbf{v}$, a user-defined backdoor patch p_v, and the targeted label y^p. The generalized patch attack $\mathfrak{P}_{patch}(\cdot; p_v, \mathbf{v}, y^p, \alpha) : \mathcal{Z} \to \mathcal{Z}$ is defined as:

$$\mathfrak{P}^X_{patch}(\mathbf{x}; p_v, \mathbf{v}, \alpha)^{(p)} = \begin{cases} \mathbf{x}^{(p)} & \text{if } p \neq p_v, \\ \alpha\mathbf{v}+(1-\alpha)\mathbf{x}^{(p)} & \text{if } p = p_v, \end{cases}$$

and $\mathfrak{P}^Y_{patch}(y; y^p) = y^p$.

We manipulate the hyper-parameter α to control the norm of the trigger pattern. We use 6000 training data and fix the poisoning rate as 0.1 in both MNIST and CIFAR-10. We use BadNets attack, and adjust α within the range $\{0.0, 0.25, 0.5, 0.75, 1.0\}$. As shown in Table 2, as α grows up, the accuracy of the model remains a minimal change. Beyond $\alpha > 0.5$, the attacker successfully embeds the backdoors in NN in MNIST and CIFAR-10. The time $T^\star$ decreases as α increases, which shows that α significantly influences the effectiveness of the trigger pattern, playing a vital role in the backdoor attack. The complete results of Table 2 can be found in Tables 3 and 8

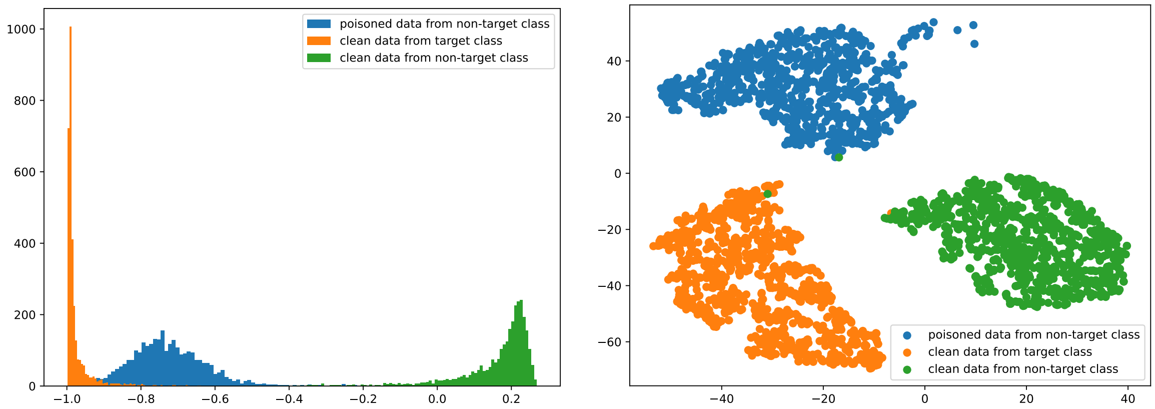
Next, we also evaluate the relationship between the size of the training set and poisoning rate and the relationship between α



(a) Cosine Similarity

(b) T-SNE

Figure 4. Results about the representation vectors on CIFAR-10 under the four-corner attack. (a) The cosine similarities of the representation vectors and the top singular vector. (b) The T-SNE plot of representation vectors. The representation vectors are centered by the average representation vector.



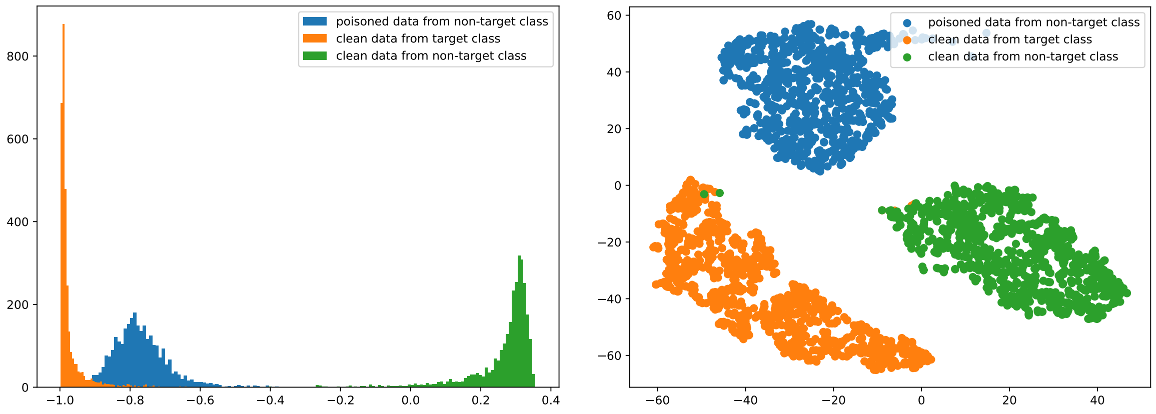
(a) Cosine Similarity

(b) T-SNE

Figure 5. Results about the representation vectors on MNIST under the BadNets attack. (a) The cosine similarities of the representation vectors and the top singular vector. (b) The T-SNE plot of representation vectors. The representation vectors are centered by the average representation vector.

and poisoning rate under the four-corner attack on MNIST, the results are shown in Tables 4 and 5, respectively. In Table 4 the time at the poisoning rate is used with 0.09 and size equals 2000, is larger than the results of the time at the poisoning rate used with 0.08 and size equals 2000. We find that this is due to that we choosing a high threshold and strict condition that for any $t \geq T$, the ASR should be always greater than 95%. If we use a smaller threshold, for example, 80%, this phenomenon is absent. Similar phenomena are caused by this threshold as well.

Finally, we show the results about the relationship between the size of the training set and poisoning rate under the BadNets attack and four-corner attack on CIFAR-10 in Table 6 and Table 7, respectively. The relationship between α and poisoning rate under the BadNets attack and four-corner attack on CIFAR-10 are shown in Table 8 and Table 9. In these tables, we find similar phenomenons with that in MNIST.



(a) Cosine Similarity

(b) T-SNE

Figure 6. Results about the representation vectors on MNIST under the four-corner attack. (a) The cosine similarities of the representation vectors and the top singular vector. (b) The T-SNE plot of representation vectors. The representation vectors are centered by the average representation vector.

Table 2. Ablation study on the norm of trigger pattern. We evaluate the accuracy and attack success rate at the last epoch. We show the time when the attacker succeeds.

	MNIST			CIFAR-10		
α	ACC	ASR	Time	ACC	ASR	Time
0.0	99.52	0.56	–	86.20	16.10	–
0.25	99.57	0.56	–	88.60	87.90	–
0.5	99.52	98.88	25	89.25	99.90	5
0.75	99.63	99.66	17	87.60	99.90	4
1.0	99.68	99.44	13	89.10	100.00	3

Table 3. The effects from α and poisoning rate in MNIST. We use BadNets attack and evaluate the accuracy and attack success rate at the last epoch.

MNIST		Poisoning rate									
α		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
0.0	ACC	99.63	99.68	99.57	99.57	99.52	99.52	99.52	99.47	99.52	99.52
	ASR	0.34	0.34	0.45	0.45	0.56	0.56	0.56	0.67	0.56	0.56
	Time	–	–	–	–	–	–	–	–	–	–
0.25	ACC	99.63	99.68	99.57	99.57	99.52	99.52	99.57	99.47	99.52	99.57
	ASR	0.34	0.34	0.45	0.45	0.56	0.56	0.56	0.67	0.56	0.67
	Time	–	–	–	–	–	–	–	–	–	–
0.5	ACC	99.63	99.63	99.57	99.57	99.52	99.47	99.52	99.52	99.52	99.52
	ASR	0.34	0.34	0.56	0.56	0.90	88.12	96.86	98.32	98.88	99.22
	Time	–	–	–	–	–	–	52	30	30	25
0.75	ACC	99.63	99.68	99.52	99.52	99.52	99.57	99.63	99.63	99.63	99.63
	ASR	0.34	0.56	75.11	97.09	97.87	98.09	98.77	99.55	99.66	99.66
	Time	–	–	–	36	31	24	24	22	18	17
1.0	ACC	99.63	99.57	99.57	99.57	99.57	99.63	99.68	99.68	99.68	99.68
	ASR	0.34	85.20	94.06	98.43	98.88	98.88	99.22	99.44	99.44	99.66
	Time	–	–	–	27	20	18	15	15	13	13

Table 4. The effects from the size of the training set and poisoning rate in MNIST. We use four-corner attack and evaluate the accuracy and attack success rate at the last epoch.

MNIST		Poisoning rate									
Size		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
2000	ACC	99.25	99.25	99.31	99.36	99.41	99.36	99.36	99.36	99.31	99.36
	ASR	0.90	2.47	44.51	84.30	93.39	96.41	97.76	98.54	99.22	99.44
	Time	–	–	–	–	–	35	32	27	30	22
4000	ACC	99.63	99.57	99.57	99.52	99.63	99.52	99.52	99.52	99.52	99.52
	ASR	5.94	91.14	97.20	97.65	99.10	99.22	99.44	99.55	99.66	99.66
	Time	–	–	29	25	25	14	14	13	23	12
6000	ACC	99.52	99.57	99.57	99.52	99.52	99.47	99.47	99.41	99.41	99.41
	ASR	73.32	93.61	96.86	98.77	98.99	99.22	99.33	99.44	99.44	99.55
	Time	–	–	29	23	14	11	9	9	9	9
8000	ACC	99.57	99.63	99.63	99.63	99.63	99.57	99.63	99.63	99.63	99.63
	ASR	87.89	97.42	99.10	98.99	99.89	99.89	100.00	100.00	100.00	100.00
	Time	–	20	20	9	9	8	8	7	7	7
10000	ACC	99.68	99.68	99.68	99.63	99.68	99.79	99.73	99.73	99.68	99.79
	ASR	94.84	98.77	98.99	99.66	99.78	99.89	99.89	100.00	100.00	100.00
	Time	61	18	13	13	10	7	6	6	10	6

Table 5. The effects from α and poisoning rate in MNIST. We use four-corner attack and evaluate the accuracy and attack success rate at the last epoch.

MNIST		Poisoning rate									
α		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
0.0	ACC	99.68	99.63	99.57	99.57	99.47	99.52	99.52	99.47	99.41	99.47
	ASR	0.34	0.34	0.45	0.45	0.67	0.56	0.56	0.67	0.78	0.67
	Time	–	–	–	–	–	–	–	–	–	–
0.25	ACC	99.63	99.68	99.52	99.57	99.52	99.52	99.52	99.57	99.41	99.47
	ASR	0.34	0.34	0.56	0.56	0.78	32.29	77.91	93.61	96.52	98.54
	Time	–	–	–	–	–	–	–	–	41	29
0.5	ACC	99.68	99.57	99.63	99.57	99.57	99.57	99.57	99.52	99.57	99.57
	ASR	0.45	24.55	93.61	96.86	97.76	98.65	99.10	99.55	99.66	99.55
	Time	–	–	–	26	23	22	19	18	18	15
0.75	ACC	99.63	99.63	99.57	99.52	99.57	99.52	99.47	99.47	99.52	99.47
	ASR	20.96	90.47	95.85	98.65	98.65	99.10	99.33	99.33	99.55	99.55
	Time	–	–	30	27	18	16	14	14	18	14
1.0	ACC	99.52	99.57	99.57	99.52	99.52	99.47	99.47	99.41	99.41	99.41
	ASR	73.32	93.61	96.86	98.77	98.99	99.22	99.33	99.44	99.44	99.55
	Time	–	–	29	23	14	11	9	9	9	9

Table 6. The effects from the size of the training set and poisoning rate in CIFAR-10. We use BadNets attack and evaluate the accuracy and attack success rate at the last epoch.

CIFAR-10		Poisoning rate									
Size		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
2000	ACC	84.65	86.65	85.85	84.00	83.80	85.10	84.05	86.10	83.70	84.45
	ASR	15.50	33.20	82.50	98.90	94.00	99.00	99.70	99.80	99.70	99.90
	Time	–	–	–	21	–	21	5	5	5	4
4000	ACC	87.95	87.10	87.55	88.05	87.25	87.45	87.95	87.30	87.25	87.60
	ASR	16.00	93.70	96.70	99.60	99.50	99.90	99.90	99.90	100.00	99.90
	Time	–	–	28	12	5	5	4	4	3	3
6000	ACC	88.90	88.85	88.60	88.50	88.60	89.35	88.60	88.40	89.00	89.10
	ASR	41.10	90.40	99.10	99.80	99.70	99.90	99.70	100.00	100.00	100.00
	Time	–	–	19	4	7	3	4	3	1	3
8000	ACC	90.20	90.55	90.30	90.50	89.40	89.85	89.95	90.50	89.40	89.55
	ASR	83.70	99.70	99.70	99.60	100.00	100.00	99.80	100.00	100.00	100.00
	Time	–	6	4	12	4	3	3	1	1	1
10000	ACC	91.15	90.15	91.00	91.10	90.70	91.15	90.65	90.90	90.85	90.40
	ASR	98.70	99.60	99.80	100.00	100.00	100.00	100.00	99.90	100.00	100.00
	Time	24	9	4	4	3	3	3	3	1	1

Table 7. The effects from the size of the training set and poisoning rate in CIFAR-10. We use four-corner attack and evaluate the accuracy and attack success rate at the last epoch.

CIFAR-10		Poisoning rate									
Size		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
2000	ACC	86.00	86.20	84.70	85.50	85.05	85.40	87.15	86.35	84.90	85.65
	ASR	27.00	94.50	99.70	100.00	99.80	99.90	100.00	99.70	99.90	99.90
	Time	–	–	11	4	5	4	4	6	4	4
4000	ACC	87.45	87.25	88.45	87.30	87.15	87.45	86.90	87.60	88.45	88.05
	ASR	80.30	99.20	100.00	99.80	100.00	100.00	99.90	100.00	100.00	100.00
	Time	–	4	29	3	3	3	3	3	3	3
6000	ACC	89.15	87.95	88.50	88.95	89.70	89.15	89.20	89.75	88.55	89.60
	ASR	94.00	99.90	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	Time	–	11	1	3	1	1	1	3	1	1
8000	ACC	90.15	90.00	91.10	90.00	90.60	90.70	90.40	90.10	89.55	90.25
	ASR	97.90	100.00	100.00	100.00	99.90	100.00	100.00	100.00	100.00	100.00
	Time	18	8	1	1	1	1	1	1	1	1
10000	ACC	90.95	90.50	90.95	91.45	90.70	90.95	90.45	91.05	91.40	90.80
	ASR	96.10	100.00	100.00	99.80	100.00	100.00	100.00	100.00	100.00	100.00
	Time	40	9	1	1	1	1	1	1	1	1

Table 8. The effects from α and poisoning rate in CIFAR-10. We use BadNets attack and evaluate the accuracy and attack success rate at the last epoch.

CIFAR-10		Poisoning rate									
α		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
0.0	ACC	89.15	88.85	88.60	88.45	87.05	87.85	87.25	87.20	87.90	86.20
	ASR	11.50	11.20	11.70	12.20	14.80	14.50	16.00	15.90	15.40	16.10
	Time	–	–	–	–	–	–	–	–	–	–
0.25	ACC	88.80	87.70	88.45	88.60	87.55	88.10	88.65	88.05	88.05	88.60
	ASR	11.30	13.00	12.10	15.40	26.50	30.50	55.20	90.30	87.00	87.90
	Time	–	–	–	–	–	–	–	–	–	–
0.5	ACC	89.05	88.25	88.50	88.95	88.75	89.00	89.55	89.00	87.90	89.25
	ASR	12.30	14.90	56.50	96.40	95.30	98.80	99.30	99.70	99.90	99.90
	Time	–	–	–	20	27	14	8	5	6	5
0.75	ACC	88.65	89.50	89.45	89.25	89.70	89.80	90.05	89.55	88.85	87.60
	ASR	19.40	44.80	96.20	97.80	99.10	99.70	99.60	99.90	100.00	99.90
	Time	–	–	20	10	9	6	4	4	4	4
1.0	ACC	88.90	88.85	88.60	88.50	88.60	89.35	88.60	88.40	89.00	89.10
	ASR	41.10	90.40	99.10	99.80	99.70	99.90	99.70	100.00	100.00	100.00
	Time	–	–	19	4	7	3	4	3	1	3

Table 9. The effects from α of the training set and poisoning rate in CIFAR-10. We use four-corner attack and evaluate the accuracy and attack success rate at the last epoch.

CIFAR-10		Poisoning rate									
α		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
0.0	ACC	87.60	88.30	88.65	87.85	87.30	87.20	86.25	87.40	87.15	86.10
	ASR	12.50	12.80	14.90	12.90	13.80	14.40	18.10	16.10	16.70	18.50
	Time	–	–	–	–	–	–	–	–	–	–
0.25	ACC	88.75	88.05	87.60	89.85	89.20	88.90	88.50	88.45	89.45	90.15
	ASR	12.60	18.40	37.80	92.60	99.40	99.30	98.60	99.60	99.80	100.00
	Time	–	–	–	–	8	5	20	4	5	3
0.5	ACC	89.55	89.05	89.25	88.90	89.75	88.65	88.55	89.00	88.30	89.40
	ASR	33.00	98.00	99.50	99.90	99.70	99.80	100.00	99.90	100.00	99.90
	Time	–	11	9	3	4	18	3	3	3	3
0.75	ACC	90.25	88.75	90.00	89.40	89.75	89.80	90.55	89.15	89.20	89.55
	ASR	57.70	99.90	99.70	99.80	99.60	100.00	99.90	100.00	100.00	100.00
	Time	–	17	7	3	7	3	3	1	1	3
1.0	ACC	89.15	87.95	88.50	88.95	89.70	89.15	89.20	89.75	88.55	89.20
	ASR	94.00	99.90	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.90
	Time	–	11	1	3	1	1	1	3	1	1